On nonparametric regression with incomplete data via inverse-weighting and their convergence in L_p norms

Majid Mojirsheibani [1] **and Tatiana Shahinian**[2]

Department of Mathematics, California State University, Northridge, CA, 91330, USA

Abstract

This paper deals with the problem of nonparametric regression when the response variable may be missing but not necessarily at random. Here, we propose a new approach to construct kernel-type estimators of an unknown regression function based on Horvitz-Thompson inverse weighting when the data suffers from missing response values. The proposed approach may be viewed as a two-step procedure: the first step involves constructing a family of kernel-type regression estimators based on inverse weighting where the members of this family are indexed by the unknown parameters of the missing probability mechanism (the selection probability). In the second step, a search will be carried out to find the member of a cover of this family that has the smallest mean-squared prediction error. Furthermore, we establish exponential performance bounds on the deviations of the proposed estimators from the true regression curve in general L_p norms; these bounds yield various strong convergence results. We also study the rates of convergence of these estimators. As an important application of our results, we consider the problem of statistical classification with incomplete data.

Keywords: kernel regression, convergence in L_p norms, classification, partially observed data.

1 INTRODUCTION

This paper considers the problem of kernel regression estimation in the presence of missing response variables, Y, for the *missing not at random* (MNAR) setup where the mechanism that causes Y to be missing is allowed to depend on both the predictor $\mathbf{X}$ and the real-valued response variable Y. The MNAR setup is generally acknowledged to be a difficult problem in incomplete data literature due to identifiability issues; this is significantly different from the simpler *missing at random* model where the absence of Y depends on $\mathbf{X}$ only (and not on both $\mathbf{X}$ and Y).

More specifically, let $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ be a random vector and consider the problem of estimating the regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, based on n independent and identically distributed (iid) observations $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, drawn from the distribution of $(\mathbf{X}, Y)$. When the data is

[1]Corresponding author. Email: majid.mojirsheibani@csun.edu

[2]Email: tatiana.shahinian.475@my.csun.edu

fully observable, the classical Nadaraya-Watson kernel estimator of $m(\mathbf{x})$ (Nadaraya [26], Watson [36]) is given by

$$\widehat{m}_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i \, \mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)}{\sum_{i=1}^n \mathcal{K}(\mathbf{x}-\mathbf{X}_i)/h)}\,, \tag{1}$$

where the function $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}_+$ is the kernel used with the bandwidth $h > 0$. A global measure of the accuracy of $\widehat{m}_n(\cdot)$ as an estimator of $m(\cdot)$ is given by its L_p-type statistic $I_n(p) = \int |\widehat{m}_n(\mathbf{x}) - m(\mathbf{x})|^p \, \mu(d\mathbf{x}),\ 1 \leq p < \infty$, where μ is the probability measure of $\mathbf{X}$. The quantity $I_n(1)$ plays an important role in statistical classification; see for example Devroye et al ([7], Sec. 6.2). For the strong convergence of $I_n(1)$ to zero see, for example, Devroye and Krzyżak [8]. In fact, in the cited paper, Devroye and Krzyźak obtain a number of equivalent results under the assumption that $|Y| \leq L < \infty$, one of which states that if the kernel $\mathcal{K}$ is *regular* (the definition will be given later) then for every $\epsilon > 0$ and n large enough, one has $P\{I_n(1) > \epsilon\} \leq \exp\{-c\,n\}$, where c is a positive constant depending on ϵ but not on n. For some more recent interesting results along these lines one may refer to Berrett et al [3].

To appreciate the difficulties involved in constructing (1) in the current setup, suppose that the response variable Y is allowed to be missing according to the MNAR mechanism and define the indicator random variable $\Delta = 0$ if Y is missing, and $\Delta = 1$ otherwise. Similarly, for $i = 1, \cdots, n$, let $\Delta_i = 0$ when Y_i is missing (and $\Delta_i = 1$ otherwise). Then, it is not hard to see that the estimator $\widehat{m}_n(\mathbf{x})$ in (1) is no longer available. Of course, one might decide (incorrectly) to use the kernel estimator based on the complete cases only, i.e., the estimator

$$m_n^{\rm cc}(\mathbf{x}) \ := \ \sum_{i=1}^n \Delta_i Y_i \, \mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h) \big/ \sum_{i=1}^n \Delta_i \mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h). \tag{2}$$

However, $m_n^{\rm cc}(\mathbf{x})$ turns out to be the kernel estimator of the quantity $E(\Delta Y|\mathbf{X}=\mathbf{x})\big/E(\Delta|\mathbf{X}=\mathbf{x})$ which, in general, is not equal to the regression function $m(\mathbf{x}) = E(Y|\mathbf{X}=\mathbf{x})$ under the MNAR response mechanism.

Regression function estimation under the MNAR setup is challenging, yet some progress has been made in the literature. For example, one can refer to the results of Bindele et al [4] to estimate β in the model $E(Y|\mathbf{X}=\mathbf{x}) = g(\mathbf{x},\beta)$, where g is completely known, those of Niu et al [27] and Guo et al [11] for linear regression, and the results of Li et al [15] for functional linear models. Unlike these results, here we do not assume a linear model. Another relevant result is that of Mojirsheibani [22] on the limiting distribution of the maximal deviation of certain kernel-type regression estimators with MNAR missing responses. In the case of functional covariates one can refer to the work of Ferraty et al [10], Ling et al [16], and Bouzebda et al [5], under the missing at random assumption.

To present our Horvitz-Thompson type estimators of the unknown regression function $m(\mathbf{x}) = E(Y|\mathbf{X}=\mathbf{x})$, we also need to discuss the choice of the *selection probability*. Let $\pi(\mathbf{x}, y) = E\big[\Delta\big|\mathbf{X} = \mathbf{x}, Y = y\big]$ be the selection probability, also called the *nonresponse propensity*, where the random variable $\Delta = 1$ if Y is observable (and $\Delta = 0$ otherwise). Classical methods such as those of Greelees et al (1982) assumed a fully parametric model for both the outcome and the selection probability. Such model assumptions have been relaxed in recent years, and semiparametric methods that assume a parametric model for either the outcome or the selection probability (but not both) have been proposed. Relevant work along these lines includes Shao and Wang [31] who use instrument variables to estimate the selection probability without specifying the outcome model for the distribution of $y|\mathbf{x}$ when $\mathbf{x} \in \mathbb{R}^d$, $d \geq 2$. Similarly, Morikawa et al [25] used kernel regression estimators to avoid parametric outcome model assumptions, while Tang et al [32] and Zhao and Ma [38] estimated the outcome model without specifying the selection probability. Miao and Li [20] employ follow-up subsamples to deal with identifiability without using instrumental variables; Morikawa and Kim [24] use a parametric model for the selection probability and a fully nonparametric outcome model for $y \,|\, \mathbf{x}, \Delta{=}1$, and develop a nonparametric test procedure for model identification. Maity et al [17] propose a likelihood-based method to improve the bias in estimating logistic regression coefficients; Uehara et al [34] consider a semiparametric response model and use instrument variables to estimate the selection probability, whereas Sadinle and Reiter [30] propose a class of nonignorable missingness mechanisms for modeling multivariate data and use auxiliary information on marginal distributions to identify the underlying models. Zhao et al [39] propose an adjusted empirical likelihood method in the presense of nonignorable nonresponses for which the Wilks' theorem continues to hold. Chen et al [6] study a semiparametric model with unspecified missingness mechanism and develop maximum pseudo likelihood estimation procedures when the response conditional density is an exponential family.

For the important case of predictive models (such as regression and classification), in this paper we consider a versatile logistic-type selection probability model that works as follows. For any function $\varphi > 0$ on $\mathbb{R}$, let

$$\pi_{\varphi}(\mathbf{x}, y) \;=\; \frac{1}{1 + \exp\big\{g(\mathbf{x})\big\} \cdot \varphi(y)}.$$

Then we have the following selection probability which is in the spirit of Kim and Yu [13]

$$E\left[\Delta\,\middle|\mathbf{X} = \mathbf{x}, Y = y\right] \;=\; P\left\{\Delta = 1\,\middle|\mathbf{X} = \mathbf{x}, Y = y\right\} \;=\; \frac{1}{1 + \exp\big\{g(\mathbf{x})\big\} \cdot \varphi^*(y)} \;:=\; \pi_{\varphi^*}(\mathbf{x}, y), \tag{3}$$

where φ^* represents the true function φ that can depend on unknown parameters and g is a completely unspecified function. A popular choice of φ is $\varphi(y) = \exp(\gamma y)$ for some unknown

parameter γ (Kim and Yu [13]). It is also well known that estimating the unknown quantities in (3) can be challenging due to identifiability issues, and a sufficient condition for model identification is (see, for example, Shao and Wang [31] to assume that there is a part of $\mathbf{X}$, say $\mathbf{V}$, which is conditionally independent of Δ, given Y and $\mathbf{Z}$, where $\mathbf{X} = (\mathbf{Z}, \mathbf{V}) \in \mathbb{R}^d$ and $\mathbf{Z} \in \mathbb{R}^p$, with $1 \leq p < d$. On the other hand, this approach fails for the important case where the covariate $\mathbf{X}$ is in $\mathbb{R}^1$. Furthermore, finding consistent estimators of φ^* based on the above assumption on $\mathbf{X}$ does not necessarily yield strongly optimal (in L_p norms) of kernel regression estimators in general. To deal with these issues, and as in Kim and Yu [13] and Miao and Li [20], we consider the case where one has access to a small follow-up subsample of response values selected from the set of non-respondents. From an applied point of view, an attractive feature of our approach is that the follow-up subsample size can be negligibly small; we have pressed this issue here to emphasize that the seemingly unpleasant need for a follow-up subsample can in practice be a non-issue. This fact is further asserted in our numerical studies where sometimes a subsample of size as small as 2, 1, or even 0 will do!

Our contributions in this paper are as follows. (i) We develop a Horvitz-Thompson-type inverse weighting approach to estimate a regression curve $m(\mathbf{x})$, nonparametrically, in the presence of MNAR response variables. The proposed method uses a data-splitting approach to estimate φ^* in (3). (ii) We study convergence properties of the proposed regression estimators in general L_p norms (as well as their rates of convergence). (iv) We apply our results to the problem of classification where we construct asymptotically optimal nonparametric classification rules in the presence of MNAR missing data. This constitutes an important application to the field of machine learning and statistical classification in semi-supervised learning where one usually has to deal with large amounts of missing labels in the data. Researchers in machine learning have made efforts to develop procedures for utilizing the unlabeled cases (i.e., missing Y_i's) in order to construct better classification rules; see Wang and Shen [37]. However, most such results assume that the response variable is missing completely at random; see, for example, Azizyan et al [1]. Our estimators can also be used in *ensemble* methods such as those of Fischer and Mougeot [40].

The rest of the paper is organized as follows. The main results are presented in Section 2 where we construct our regression estimators and study their asymptotic properties in general L_p norms. The presence of missing response variables is handles using a Horvitz-Thompson type inverse weighting approach. Section 3 deals with the applications of our estimators to the problem of nonparametric classification with partially observed data, where we also look into the rates of convergence of the proposed classifiers under different conditions. Section 4 presents some numerical studies; the results here confirm the good finite-sample performance of the proposed estimators under different conditions. All proofs are deferred to Section 5.

2 MAIN RESULTS

2.1 *The proposed estimator*

To present our estimator, we employ a data splitting approach that works as follows. Let $\mathbb{D}_n = \{(\boldsymbol{X}_1, Y_1, \Delta_1), \ldots, (\boldsymbol{X}_n, Y_n, \Delta_n)\}$ represent the sample of size n (iid), where $\Delta_i = 0$ if Y_i is missing (and $\Delta_i = 1$ otherwise). Now, randomly split the data into a training sample $\mathbb{D}_m$ of size m and a validation sequence $\mathbb{D}_\ell$ of size $\ell = n - m$, where $\mathbb{D}_m \cup \mathbb{D}_\ell = \mathbb{D}_n$ and $\mathbb{D}_m \cap \mathbb{D}_\ell = \varnothing$. Here, it is assumed that $\ell \to \infty$ and $m \to \infty$, as $n \to \infty$; the choices of m and ℓ will be discussed later. Also, define the index sets

$$\boldsymbol{\mathcal{I}}_m = \Big\{ i \in \{1, \cdots, n\} \,\Big|\, (\mathbf{X}_i, Y_i, \Delta_i) \in \mathbb{D}_m \Big\} \quad \text{and} \quad \boldsymbol{\mathcal{I}}_\ell = \Big\{ i \in \{1, \cdots, n\} \,\Big|\, (\mathbf{X}_i, Y_i, \Delta_i) \in \mathbb{D}_\ell \Big\}.$$

Let $\mathcal{F}$ be the class of functions to which the unknown function φ^* in (3) belongs. In the hypothetical case where the function π_{φ^*} in (3) is completely known (unrealistic), a Horvitz-Thompson type (Horvitz and Thompson [12]) kernel estimator of the regression curve $m(\mathbf{x})$ is given by

$$\widehat{m}(\mathbf{x}; \pi_{\varphi^*}) \;\; := \;\; \sum_{i=1}^{n} \frac{\Delta_i Y_i}{\pi_{\varphi^*}(\mathbf{X}_i, Y_i)} \mathcal{K}\big((\mathbf{x} - \mathbf{X}_i)/h\big) \Big/ \sum_{i=1}^{n} \mathcal{K}\big((\mathbf{x} - \mathbf{X}_i)/h\big), \tag{4}$$

Here, (4) is justified (as an estimator of $m(\mathbf{x})$) by the fact that it is the kernel estimator of

$$m(\mathbf{X}; \pi_{\varphi^*}) := E\Big[\frac{\Delta Y}{\pi_{\varphi^*}(\mathbf{X}, Y)}\Big|\mathbf{X}\Big] = E\Big[E\Big(\frac{\Delta Y}{\pi_{\varphi^*}(\mathbf{X}, Y)}\Big|\mathbf{X}, Y\Big)\Big|\mathbf{X}\Big] \overset{\text{via (3)}}{=} E[Y|\mathbf{X}] \;=\; m(\mathbf{X}). \tag{5}$$

Of course, (4) is not quite the right estimator because π_{φ^*} is unknown and must be estimated itself. To present our proposed estimators, we start as follows. For each $\varphi \in \mathcal{F}$, define

$$\pi_\varphi(\mathbf{x}, y) \;=\; \frac{1}{1 + \exp\big\{g(\mathbf{x})\big\} \cdot \varphi(y)} \quad \text{and} \quad m(\mathbf{x}; \pi_\varphi) = E\Big[\frac{\Delta Y}{\pi_\varphi(\mathbf{X}, Y)}\Big|\mathbf{X} = \mathbf{x}\Big], \tag{6}$$

and consider the class of kernel-type estimators of $m(\mathbf{x}; \pi_\varphi)$ constructed based on the training sample $\mathbb{D}_m$ alone, and indexed by $\varphi \in \mathcal{F}$, given by

$$\widehat{m}_m(\mathbf{x}; \widehat{\pi}_\varphi) \;\; = \;\; \sum_{i \in \boldsymbol{\mathcal{I}}_m} \frac{\Delta_i Y_i}{\widehat{\pi}_\varphi(\mathbf{X}_i, Y_i)} \mathcal{K}\big((\mathbf{x} - \mathbf{X}_i)/h\big) \Big/ \sum_{i \in \boldsymbol{\mathcal{I}}_m} \mathcal{K}\big((\mathbf{x} - \mathbf{X}_i)/h\big), \quad \varphi \in \mathcal{F}, \tag{7}$$

where, for each fixed $\varphi \in \mathcal{F}$, the quantity $\widehat{\pi}_\varphi$ is the estimate of the selection probability π_φ based on the training sample $\mathbb{D}_m$, given by

$$\widehat{\pi}_\varphi(\mathbf{X}_i, Y_i) = \left[1 \,+\, \frac{1 - \widehat{\eta}_m(\mathbf{X}_i)}{\widehat{\psi}_m(\mathbf{X}_i; \varphi)} \cdot \varphi(Y_i)\right]^{-1}, \tag{8}$$

where the quantities $\widehat{\psi}_m(\mathbf{X}_i;\varphi)$ and $\widehat{\eta}_m(\mathbf{X}_i)$ are the kernel estimators of the functions $\psi(\mathbf{x};\varphi) := E\big[\Delta\,\varphi(Y)\big|\mathbf{X}=\mathbf{x}\big]$ and $\eta(\mathbf{x}) = E\big[\Delta\big|\mathbf{X}=\mathbf{x}\big]$, respectively, and given by

$$\begin{cases} \widehat{\psi}_m(\mathbf{X}_i;\varphi) \;=\; \displaystyle\sum_{j\in\boldsymbol{\mathcal{I}}_m,\,j\neq i} \Delta_j\varphi(Y_j)\,\mathcal{K}((\mathbf{X}_i-\mathbf{X}_j)/h) \;\div\; \sum_{j\in\boldsymbol{\mathcal{I}}_m,\,j\neq i} \mathcal{K}((\mathbf{X}_i-\mathbf{X}_j)/h), \quad \varphi\in\mathcal{F} \\ \widehat{\eta}_m(\mathbf{X}_i) \;=\; \displaystyle\sum_{j\in\boldsymbol{\mathcal{I}_m},\,j\neq i} \Delta_j\mathcal{K}((\mathbf{X}_i-\mathbf{X}_j)/h) \;\div\; \sum_{j\in\boldsymbol{\mathcal{I}_m},\,j\neq i} \mathcal{K}((\mathbf{X}_i-\mathbf{X}_j)/h), \end{cases} \tag{9}$$

with the usual convention $0/0=0$, where $\mathcal{K}:\mathbb{R}^p\to\mathbb{R}_+$ with bandwidth h. The estimator $\widehat{\pi}_\varphi$ in (8) is justified by the fact that the term $\exp\{g(\mathbf{x})\}$ in (3) can be expressed as (Kim and Yu [13])

$$\exp\{g(\mathbf{X})\} \;=\; \frac{1-\eta(\mathbf{x})}{\psi(\mathbf{x};\varphi)}, \qquad \text{whrere} \quad \begin{cases} \eta(\mathbf{x}) \;=\; E\big[\Delta\big|\mathbf{X}=\mathbf{x}\big] \\ \psi(\mathbf{x};\varphi) \;=\; E\big[\Delta\,\varphi(Y)\big|\mathbf{X}=\mathbf{x}\big], \end{cases} \tag{10}$$

which in turn means that one can re-write $\pi_\varphi(\mathbf{x},y)$ as

$$\pi_\varphi(\mathbf{x},y) \;=\; \left[1+\frac{1-\eta(\mathbf{x})}{\psi(\mathbf{x};\varphi)}\cdot\varphi(y)\right]^{-1}. \tag{11}$$

Furthermore, since $\pi_\varphi>\pi_{\min}>0$ by assumption (A) and since $\widehat{\psi}_m(\mathbf{X}_i;\varphi)$ in (9) is the estimator of the strictly positive quantity $E[\Delta_i\varphi(Y_i)|\mathbf{X}_i]\geq\varrho_0$, where $\varrho_0>0$ (see assumption (D)), we also consider the following truncated-type version of the estimator in (8)

$$\breve{\pi}_\varphi(\mathbf{X}_i,Y_i) = \left[1\;+\;\frac{1-\widehat{\eta}_m(\mathbf{X}_i)}{\pi_0\vee\widehat{\psi}_m(\mathbf{X}_i;\varphi)}\cdot\varphi(Y_i)\right]^{-1}, \tag{12}$$

where $\pi_0>0$ is a fixed constant whose choice will be discussed later under assumption (A′). Here, we note that $\breve{\pi}_\varphi$ in (12) may be viewed as a one-sided winsorized estimator of π_φ (compare this with $\widehat{\pi}_\varphi$ in (8)). In applications with either simulated or real data, π_0 can be zero (as in (11)) or is chosen to be a small positive number such as $10^{-\nu}$, $\nu\geq 10$; this will also become clear later in our numerical studies. In fact, the presence of π_0 in 12) is only for theoretical purposes.

Next, one has to estimate φ^*. To this end, we employ approaches based on the approximation theory of totally bounded function spaces. This turns out to be a suitable approach when studying the global performance of our proposed regression estimators via their L_p norms. More specifically, let $\mathcal{F}$ be a given class of functions $\varphi\colon[-L,L]\to(0,B]$, for some $B<\infty$. Fix $\varepsilon>0$ and suppose that the finite collection of functions $\mathcal{F}_\varepsilon=\{\varphi_1,\ldots,\varphi_{N(\varepsilon)}\}$, $\varphi_i\colon[-L,L]\to(0,B]$, is an ε-cover of $\mathcal{F}$, i.e., for each $\varphi\in\mathcal{F}$, there is a $\varphi'\in\mathcal{F}_\varepsilon$ such that $\|\varphi-\varphi'\|_\infty<\varepsilon$; here, $\|\cdot\|_\infty$ is the usual supnorm. The cardinality of the smallest ε-cover of $\mathcal{F}$ is called the *covering number* of the family $\mathcal{F}$ and will be denoted by $\mathcal{N}_\varepsilon(\mathcal{F})$. If $\mathcal{N}_\varepsilon(\mathcal{F})<\infty$ holds for every $\varepsilon>0$, then the family $\mathcal{F}$ is said to be *totally bounded* (with respect to $\|\cdot\|_\infty$). The quantity $\log(\mathcal{N}_\varepsilon(\mathcal{F}))$ is called Kolmogorov's

ϵ-entropy of the set $\mathcal{F}$. The monograph by van der Vaart and Wellner ([35]; p. 83) provides more details on such concepts.

Now, let $0 < \varepsilon_n \downarrow 0$ be a decreasing sequence, as $n \to \infty$, and let $\mathcal{F}_{\varepsilon_n} = \{\varphi_1, \ldots, \varphi_{N(\varepsilon_n)}\} \subset \mathcal{F}$ be any ε_n-cover of $\mathcal{F}$; the choice of ε_n will be discussed later in Corollary 1. Also, as explained in the introduction, here we follow Kim and Yu [13], and Miao and Li [20], and consider the setup in which one has access to response values for a small follow-up subsample selected from the set of non-respondents. More formally, let δ_i, $i = 1, \cdots, \ell$, be iid Bernoulli random variables, independent of the data $\mathbb{D}_n$, with the probability of success

$$p_n = P\{\delta_i = 1\}, \quad i = 1, \cdots, \ell, \quad \text{with } p_n \to 0, \text{ as } n \to \infty. \tag{13}$$

Also recall that $\Delta_i = 0$ if Y_i is missing and $\Delta_i = 1$ otherwise. Then we select a non-respondent to be in the small follow-up subsample only if $(1 - \Delta_i)\,\delta_i = 1$. Next, for each $\varphi \in \mathcal{F}_{\varepsilon_n}$ define the empirical L_2 error of $\widehat{m}_m(\mathbf{x}; \widehat{\pi}_\varphi)$ in (7) based on the testing sequence $\mathbb{D}_\ell$ by

$$\begin{aligned} \widehat{L}_{m,\ell}(\widehat{\pi}_\varphi) &= \frac{1}{\ell}\Big[\sum_{i \in \boldsymbol{\mathcal{I}}_{\boldsymbol{\ell}}} \Delta_i \big|\widehat{m}_m(\mathbf{X}_i; \widehat{\pi}_\varphi) - Y_i\big|^2 + \sum_{i \in \boldsymbol{\mathcal{I}}_{\boldsymbol{\ell}}} (1 - \Delta_i)(\delta_i / p_n) \big|\widehat{m}_m(\mathbf{X}_i; \widehat{\pi}_\varphi) - Y_i\big|^2 \Big] \\ &= \frac{1}{\ell} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \Big(\Delta_i + \frac{(1 - \Delta_i)\delta_i}{p_n} \Big) \big|\widehat{m}_m(\mathbf{X}_i; \widehat{\pi}_\varphi) - Y_i\big|^2. \end{aligned} \tag{14}$$

Then we use the following two-step procedure to estimate φ^* which will then be plugged into (7) for the free functional parameter φ.

Step 1. For each $\varphi \in \mathcal{F}_{\varepsilon_n}$, use the training sample $\mathbb{D}_m$ to compute $\widehat{m}_m(\mathbf{x}; \widehat{\pi}_\varphi)$ in (7), where $\widehat{\pi}_\varphi$ is as in (8). Alternatively, if $\breve{\pi}_\varphi$ in (12) is used instead of $\widehat{\pi}_\varphi$, then one computes $\widehat{m}_m(\mathbf{x}; \breve{\pi}_\varphi)$, which is obtained by replacing $\widehat{\pi}_\varphi$ with $\breve{\pi}_\varphi$ in (7).

Step 2. The proposed estimator of φ^* is then defined by

$$\begin{cases} \widehat{\varphi}_n := \operatorname{argmin}_{\varphi \in \mathcal{F}_{\varepsilon_n}} \widehat{L}_{m,\ell}(\widehat{\pi}_\varphi), & \text{if (8) is used,} \\ \breve{\varphi}_n := \operatorname{argmin}_{\varphi \in \mathcal{F}_{\varepsilon_n}} \widehat{L}_{m,\ell}(\breve{\pi}_\varphi), & \text{if (12) is used,} \end{cases} \tag{15}$$

where $\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)$ is as in (14) and $\widehat{L}_{m,\ell}(\breve{\pi}_\varphi)$ is obtained upon replacing $\widehat{\pi}_\varphi$ by $\breve{\pi}_\varphi$ in (14). The subscript n at $\widehat{\varphi}_n$ and $\breve{\varphi}_n$ reflects the fact that the entire data of size n has been used here. Finally, our proposed Horvitz-Thompson type estimator of $m(\mathbf{x})$ is given by

$$\begin{cases} \widehat{m}(\mathbf{x}; \widehat{\pi}_{\widehat{\varphi}_n}) := \widehat{m}_m(\mathbf{x}; \widehat{\pi}_\varphi)\Big|_{\varphi = \widehat{\varphi}_n} & \text{if (8) is used,} \\ \widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n}) := \widehat{m}_m(\mathbf{x}; \breve{\pi}_\varphi)\Big|_{\varphi = \breve{\varphi}_n} & \text{if (12) is used,} \end{cases} \tag{16}$$

where $\widehat{m}_m(\mathbf{x}; \widehat{\pi}_\varphi)$ is as in (7).

2.2 *Theoretical goodness of the estimators*

How good are the estimators $\widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n})$ and $\widehat{m}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})$ in (16)? To address and answer this question, we start by assuming that the kernel $\mathcal{K}$ is *regular*:

Definition *A nonnegative kernel $\mathcal{K}$ is said to be regular if there are real constants $b > 0$ and $r > 0$ such that $\mathcal{K}(\mathbf{u}) \geq b\, I\{\mathbf{u} \in S_{0,r}\}$ and $\int \sup_{\mathbf{y}\in\mathbf{u}+S_{0,r}} \mathcal{K}(\mathbf{y})\, d\mathbf{u} < \infty$, where $S_{0,r}$ is the ball of radius r centered at the origin.*

See Devroye and Krzyżak [8] for more on this. In order to study and assess the performance of the estimators $\widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n})$ and $\widehat{m}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})$ in (16), we first state a number of assumptions.

Assumption (A). The selection probability, $\pi_\varphi(\mathbf{x}, y)$, in (3) satisfies $\inf_{\mathbf{x},y}\, \pi_\varphi(\mathbf{x}, y) =: \pi_{\min} > 0$ for all $\varphi \in \mathcal{F}$.

Assumption (B). The kernel $\mathcal{K}$ in (9) satisfies $\int_{\mathbb{R}^d} \mathcal{K}(\mathbf{x})\, d\mathbf{x} = 1$ and $\int_{\mathbb{R}^d} |x_i|\mathcal{K}(\mathbf{x})\, d\mathbf{x} < \infty$ for $x_i \in (x_1, \cdots, x_d)^T = \mathbf{x}$. Also, the smoothing parameter h satisfies $h \to 0$ and $mh^d \to \infty$, as n (and thus m) $\to \infty$.

Assumption (C). The density function $f(\mathbf{x})$ of $\mathbf{X}$ is compactly supported and is bounded away from zero and infinity on its compact support. Additionally, the first-order partial derivatives of f exist and are bounded on the interior of its support.

Assumption (D). $E[\Delta\, \varphi(Y)|\mathbf{X} = \mathbf{x}] \geq \varrho_0$, for μ–a.e. $\mathbf{x}$ and each $\varphi \in \mathcal{F}$, for some constant $\varrho_0 > 0$. Furthermore, $E\big[\exp\{2g(\mathbf{X})\}\big] < \infty$, where $g(\mathbf{x})$ is as in (3).

Assumption (E). The partial derivatives $\frac{\partial}{\partial x_i} E[\Delta|\mathbf{X} = \mathbf{x}]$ and $\frac{\partial}{\partial x_i} E[\Delta\, \varphi(Y)|\mathbf{X} = \mathbf{x}]$ exist for $i = 1, \ldots, \dim(\mathbf{x})$, and are bounded on the compact support of f.

Assumption (F). The class $\mathcal{F}$ is a totally bounded class of functions $\varphi : [-L, L] \to (0, B]$, for some $B < \infty$ and $L < \infty$.

Assumption (A′). The selection probability, $\pi_\varphi(\mathbf{x}, y)$, in (3) satisfies $\inf_{\mathbf{x},y}\, \pi_\varphi(\mathbf{x}, y) =: \pi_{\min} > 0$ for all $\varphi \in \mathcal{F}$, and the truncation constant π_0 in (12) can be any number satisfying $0 < \pi_0 \leq \pi_{\min}$.

In passing we note that Assumption (B) is not restrictive because the choice of the kernel is at our discretion. The first part of Assumption (C) is usually imposed in the literature on nonparametric regression to avoid unstable estimates of $m(\mathbf{x})$ in the tails of the density, f. The second part of this assumption is technical. Assumption (D) is quite mild and is justified by the fact that $E[\Delta\, \varphi(Y)|\mathbf{X}] = E[\varphi(Y)E(\Delta|\mathbf{X}, Y)|\mathbf{X}] \geq \pi_{\min} E[\varphi(Y)|\mathbf{X}]$ together with the fact that

$\varphi(y) > 0$ for all y. Assumption (E) is technical and has already been used in the literature.

Theorem 1 *Consider the two estimators of the regression function $m(\mathbf{x})$ defined by (16).*

(i) *Let $\widehat{m}(\mathbf{x}; \widehat{\pi}_{\widehat{\varphi}_n})$ be the top estimator in (16). Let the kernel $\mathcal{K}$ in (7) be regular and suppose that assumptions (A)–(F) hold. Then, for every $\varepsilon_n > 0$ satisfying $\varepsilon_n \downarrow 0$, as $n \to \infty$, every $t > 0$, and n large enough,*

$$P\left\{\int \left|\widehat{m}(\mathbf{x}; \widehat{\pi}_{\widehat{\varphi}_n}) - m(\mathbf{x})\right|^2 \mu(d\mathbf{x}) > t\right\} \;\le\; \left|\mathcal{F}_{\varepsilon_n}\right| \left(c_9 e^{-c_{10}\,\ell\, p_n^2 t^2} + c_{11}\ell m\, e^{-c_{12}\, m h^d\, p_n^2 t^2} + c_{13}\ell m\, e^{-c_{14} m h^d}\right) \tag{17}$$

whenever $\varphi^* \in \mathcal{F}$, where $|\mathcal{F}_{\varepsilon_n}|$ is the cardinality of the set $\mathcal{F}_{\varepsilon_n}$ and c_9–c_{14} are positive constants not depending on m, ℓ, n, or t.

(ii) *Let $\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n})$ be the second estimator in (16) and suppose that assumptions (A′), (B)–(F) hold. Then, under the conditions of part **(i)** of the theorem, the bound in (17) continues to hold (with different constants c_9–c_{14}) for the probability $P\left\{\int |\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) > t\right\}$.*

Remark 1 *Although the above theorem is stated in the L_2 sense, it is straightforward to show that Part (ii) continues to hold for all $p \ge 2$. To appreciate this, first observe that*

$$\left|\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n})\right| \;\le\; L \cdot \max_{1\le k\le m} \left|1 + \frac{1 - \widehat{\eta}_m(\mathbf{X}_k)}{\pi_0 \vee \widehat{\psi}_m(\mathbf{X}_k; \breve{\pi}_{\breve{\varphi}_n})}\, \breve{\varphi}_n(Y_k)\right| \;\le\; L + LB/\pi_0, \tag{18}$$

which holds because $|1 - \widehat{\eta}_m(\mathbf{X}_k)| \le 1$ and $\breve{\varphi}_n(Y_k) < B$, where $\widehat{\eta}_m(\mathbf{X}_k)$ and $\widehat{\psi}_m(\mathbf{X}_k; \breve{\pi}_{\breve{\varphi}_n})$ are as in (9). Therefore, for any $p > 2$ one can always write

$$\left|\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{x})\right|^p \le \left(\left|\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n})\right| + \left|m(\mathbf{x})\right|\right)^{p-2} \left|\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{x})\right|^2 \le\; C_0 \left|\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{x})\right|^2,$$

where $C_0 = \left(2L + LB/\pi_0\right)^{p-2}$, which in turn implies that the upper bound in (17) continues to hold for $P\left\{\int |\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{x})|^p \mu(d\mathbf{x}) > t\right\}$, for all $p \ge 2$, but with different constants. On the other hand, if $p \in [1, 2)$, it is not hard to show that $P\left\{\int \left|\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{x})\right|^p \mu(d\mathbf{x}) > t\right\} \le P\left\{\int \left|\widehat{m}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{x})\right|^2 \mu(d\mathbf{x}) > t^{2/p}\right\}$.

The following result is an immediate Corollary to Theorem 1.

Corollary 1 *Consider the two estimators in (16) and let p_n be as in (13). If, as $n \to \infty$,*

$$\varepsilon_n \downarrow 0, \quad \frac{\log|\mathcal{F}_{\varepsilon_n}|}{\ell p_n^2} \longrightarrow 0, \quad \frac{\log(|\mathcal{F}_{\varepsilon_n}| \vee m \vee \ell)}{m h^d p_n^2} \longrightarrow 0, \quad \textit{and} \quad \frac{(\ell p_n^2) \wedge (m h^d p_n^2)}{\log n} \longrightarrow \infty, \tag{19}$$

then, under the conditions of Theorem 1, the top estimator in (16) satisfies the strong convergence property, $E\left[\left|\widehat{m}(\mathbf{X}; \widehat{\pi}_{\widehat{\varphi}_n}) - m(\mathbf{X})\right|^2 \middle| \mathbb{D}_n\right] \to^{a.s.} 0$. However, for the second estimator in (16),

$$E\left[\left|\widehat{m}(\mathbf{X}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\right|^p \middle| \mathbb{D}_n\right] \longrightarrow^{a.s.} 0\,, \quad \textit{for all } p \ge 2.$$

In passing, we also note that under the conditions of Corollary 1, by Lebesgue dominated convergence theorem one has $E\big|\widehat{m}(\mathbf{X};\breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\big|^p \to 0,$ for all $p \in [2,\infty)$. However, to study the rates of convergence here, we state the following theorem.

Theorem 2 *Under the conditions of Theorem 1, for all $p \in [2,\infty)$ and n large enough,*

$$\begin{aligned} & E\big|\widehat{m}(\mathbf{X};\breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\big|^p \\ & \leq \sqrt{\frac{c_{17} + \log \ell + \log m + \log |\mathcal{F}_{\varepsilon_n}|}{c_{18}\,(\ell \wedge mh^d)\,p_n^2}} + \sqrt{\frac{1}{c_{19}\left(c_{17} + \log \ell + \log m + \log |\mathcal{F}_{\varepsilon_n}|\right) \cdot (\ell \wedge mh^d)\,p_n^2}}, \\ & \quad + \; c_{16}\,\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell m\, e^{-c_{20}\,mh^d}, \end{aligned}$$

where c_{16}–c_{20} are positive constants not depending on m, ℓ, and n.

The following is an immediate corollary to Theorem 2.

Corollary 2 *Suppose that (19) holds. Then, under the conditions of Theorem 1, for all $p \geq 2$,*

$$E\Big|\widehat{m}(\mathbf{X};\breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\Big|^p \;=\; \mathcal{O}\left(\sqrt{\frac{\log(\ell \vee m \vee |\mathcal{F}_{\varepsilon_n}|)}{(\ell \wedge mh^d)\cdot p_n^2}}\right).$$

Here we note that for the special case where $m = \alpha \cdot n$ and $\ell = (1-\alpha)\cdot n$, where $\alpha \in (0,1)$, under the above conditions one finds that for all $p \geq 2$,

$$E\Big|\widehat{m}(\mathbf{X};\breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\Big|^p \;=\; \mathcal{O}\left(\sqrt{\frac{\log(n \vee |\mathcal{F}_{\varepsilon_n}|)}{nh^d\,p_n^2}}\right). \tag{20}$$

Corollary 2 shows that choosing ℓ and m to satisfy either $\ell/n \to 0$ or $m/n \to 0$ can generally result in estimators with convergence rates worse than the case where $m = \lfloor \alpha n \rfloor$ for any $\alpha \in (0,1)$.

Remark 2 *The rates of convergence in Corollary 2 are generally not optimal as compared to those of kernel regression estimators based on no missing data. A better rate would be of order $\mathcal{O}\big(\sqrt{\log n/nh^d}\,\big)$ which is achievable if the following conditions hold: (i) $p_n = c \in (0,1]$ for some fixed probability c instead of $p_n = o(1)$, (ii) the cardinality of the ε_n-cover satisfies $\log|\mathcal{F}_{\varepsilon_n}| = \mathcal{O}(n)$, and (iii) m is chosen as $m = \lfloor \alpha n \rfloor$ for some $\alpha \in (0,1)$. It is also well-understood in the framework of kernel regression (with no missing data) that under additional assumptions such as the Lipschitz continuity of the regression function $m(\mathbf{x})$, one can establish rates as fast as $\mathcal{O}\big((nh^d)^{-1} + h^2\big)$ for the usual kernel estimator in (1) based on the naive kernel; see, for example, Györfi et al (2002; Sec. 5.3). Unfortunately, such rates do not seem to be available for our estimators with MNAR missing data where the estimation process involves many steps and many components. In fact, to the best of our knowledge, such fast rates are not available even for the simpler case of MAR*

missing data. Additionally, the dependence of the rate of convergence on p_n in Corollary 2 shows that if obtaining a follow-up subsample is convenient, then one can have good rates by taking p_n to be a fixed percentage, such as 15%, of the entire data (as in Kim and Yu [13]). Otherwise, by choosing $p_n = o(1)$ appropriately, one requires a much smaller subsample size while still retaining the convergence in (20), but at rates slower than $\mathcal{O}\big(\sqrt{\log n/nh^d}\,\big)$.

3 APPLICATIONS TO CLASSIFICATION WITH POSSIBLY MISSING LABELS

Here we consider the following two-group classification problem. Let $(\mathbf{X}, Y) \in \mathbb{R}^d \times \{0,1\}$ be a random pair where the class label Y has to be predicted based on the covariate $\mathbf{X}$. More specifically, the goal is to find a function $g : \mathbb{R}^d \to \{0,1\}$ for which the misclassification error, i.e.,

$$L(g) := P\{g(\mathbf{X}) \neq Y\}, \tag{21}$$

is as small as possible. The optimal classifier, also referred to as the Bayes classifier, is given by

$$g_{\text{B}}(\mathbf{x}) = \begin{cases} 1 & \text{if } \; m(\mathbf{x}) \; > \; \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

where $m(\mathbf{x}) := E[Y|\,\mathbf{X} = \mathbf{x}]$ is the regression function; see, for example, Chapter 2 of Devroye et al [7]. Of course, in practice, the distribution of $(\mathbf{X}, Y)$ is almost always unknown and therefore finding the best classifier g_{B} is impossible. But suppose that we have access to n iid observations (the data), $\mathbb{D}_n := \{(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)\}$, where $(\mathbf{X}_i, Y_i) \stackrel{\text{iid}}{=} (\mathbf{X}, Y),\ i = 1, \cdots, n$, and let $\widehat{g}_n$ be any classifier constructed based on the data $\mathbb{D}_n$. Also, let

$$L_n(\widehat{g}_n) = P\big\{\widehat{g}_n(\mathbf{X}) \neq Y \big|\, \mathbb{D}_n\big\} \tag{23}$$

be the conditional misclassification error of $\widehat{g}_n$. Now, let $\widehat{m}(\mathbf{x})$ be any estimator of the regression function $m(\mathbf{x}) := E[Y|\mathbf{X} = \mathbf{x}]$ and consider the plug-in type classifier

$$\widehat{g}_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \; \widehat{m}(\mathbf{x}) \; > \; \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

Then the following bound follows from Devroye et al ([7]; Lemma 6.1)

$$L_n(\widehat{g}_n) - L(g_{\text{B}}) \;\leq\; 2E\big[\big|\widehat{m}(\mathbf{X}) - m(\mathbf{X})\big|\,\big|\mathbb{D}_n\big]\,, \tag{25}$$

and thus $E\big[L_n(\widehat{g}_n)\big] - L(g_{\text{B}}) \;\leq\; 2E\big|\widehat{m}(\mathbf{X}) - m(\mathbf{X})\big|$. Now, suppose that some of the Y_i's may be missing not at random (MNAR). Consider our proposed regression estimators given by (16) and

denote the corresponding plug-in classifiers by

$$\widehat{g}_n(\mathbf{x};\widehat{\pi}) = \begin{cases} 1 & \text{if } \widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n}) > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \breve{g}_n(\mathbf{x};\breve{\pi}) = \begin{cases} 1 & \text{if } \widehat{m}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n}) > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{26}$$

where $\widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n})$ and $\widehat{m}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})$ are as in (16). To study the asymptotic performance of the classifier in (26), we also state the following so-called margin condition which can be found in [2].

Assumption (G) *[Margin condition.]* There exist constants $c > 0$ and $\alpha > 0$ such that

$$P\left\{0 < \left|m(\mathbf{X}) - \frac{1}{2}\right| \leq t\right\} \leq c\,t^{\alpha}, \quad \text{for all } t > 0. \tag{27}$$

A number of authors have studied applications of (27) to classification; these include Mammen and Tsybakov [18], Massart and Nédélec [19], Audibert and Tsybakov [2], Tsybakov and van de Geer [33], Kohler and Krzyżak [14], and Döring et al [9].

Theorem 3 *Let $\widehat{g}_n$ and $\breve{g}_n$ be the two classifiers in (26). If (19) holds then, under the conditions of Theorem 1, we have*

(i) $P\left\{\widehat{g}_n(\mathbf{X};\widehat{\pi}) \neq Y \middle| \mathbb{D}_n\right\} \xrightarrow{a.s.} P\{g_B(\mathbf{X}) \neq Y\}$ *and* $P\left\{\breve{g}_n(\mathbf{X};\breve{\pi}) \neq Y \middle| \mathbb{D}_n\right\} \xrightarrow{a.s.} P\{g_B(\mathbf{X}) \neq Y\}$.

(ii) *If the margin condition (27) holds then*

$$P\left\{\breve{g}_n(\mathbf{X};\breve{\pi}) \neq Y\right\} - P\{g_B(\mathbf{X}) \neq Y\} = \mathcal{O}\left(\left(\frac{\log(\ell \vee m \vee |\mathcal{F}_{\varepsilon_n}|)}{(\ell \wedge mh^d)\cdot p_n^2}\right)^{\frac{1+\alpha}{2(2+\alpha)}}\right),$$

where α is as in (27).

In passing we also note that for large values of α in part (ii) above, one obtains rates closer to $\big(\log(\ell \vee m \vee |\mathcal{F}_{\varepsilon_n}|)/\big[(\ell \wedge mh^d)\cdot p_n^2\big]\big)^{1/2}$ which is similar to that of the winsorized-type regression estimator in (20); see Corollary 2.

4 NUMERICAL STUDIES

This section provides some numerical examples to illustrate the performance of our proposed estimator $\widehat{m}$ defined via (16), under different settings. To carry out numerical work, we generated $n = 50$ and $n = 100$ observations from each of the following two models:

Model A. $\mathbf{X} \sim N_5(\mathbf{1}, \mathbf{\Sigma})$ and $Y = \mu_y - X_1 + X_3X_4 - X_2^2 + \exp(-X_5) + N(0, \sigma_y^2)$

Model B. $\mathbf{X} \sim N_4(\mathbf{0}, \mathbf{\Sigma})$ and $Y = X_1 + (2X_2 - 1)^2 + \frac{\sin(2\pi X_3)}{2 - \sin(2\pi X_3)} + \sin(2\pi X_4) + 2\cos(2\pi X_4)$
$+ 3\sin^2(2\pi X_4) + 4\cos^2(2\pi X_4) + N(0, \sigma_y^2)$,

where $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j\geq 1}$ with $\sigma_{ij} = 2^{-|i-j|+1}$ in Model A and $\sigma_{ij} = 2^{-|i-j|}$ in Model B. As for σ_y, two values are considered, 0.5 and 4 (high variance model); in Model A we used two values of μ_y: 1 and 2.6. Here, Model B is as in Meier et al. (2009). We also considered two choices for the function φ^* in (3), $\varphi^*(y) = \exp(\gamma^* y)$ as in Kim and Yu [13] and $\varphi^*(y) = \left[0.1 + (\gamma^* y)^2\right]^{-1}$. Next, we consider some models for the selection probability.

The following choice of coefficients result in approximately 50% missing rate in Model A:

(A1) $\pi(\mathbf{x}, y) = \left(1 + \exp\left\{\beta_0 + \sum_{j=1}^{5} \beta_j x_j\right\} \cdot \exp\{\gamma y\}\right)^{-1}$

with $(\beta_0, \cdots, \beta_5) = (0.6,\ 0.8,\ 0.25,\ -0.35,\ -0.3,\ 0.75)$, $\gamma = -0.98$, and $\mu_y = 2.6$.

(A2) $\pi(\mathbf{x}, y) = \left(1 + \exp\left\{\beta_0 + \sum_{j=1}^{4} \beta_j x_j + \beta_5 \gamma\, x_5\right\} \cdot \exp\{\gamma y\}\right)^{-1}$

with $(\beta_0, \beta_1, \cdots, \beta_5) = (0.50,\ 0.75,\ -0.25,\ 0.25,\ -0.25,\ 0.75)$, $\gamma = -0.98$, and $\mu_y = 1$.

The following choices of coefficients result in approximately 50% missing rate in Model B:

(B1) $\pi(\mathbf{x}, y) = \left(1 + \exp\left\{\beta_0 \gamma + \sum_{j=1}^{4} \beta_j x_j\right\} \cdot \left(0.1 + \gamma^2 y^2\right)^{-1}\right)^{-1}$

with $(\beta_0, \cdots, \beta_4) = (0.85,\ 0.6,\ 0.35,\ -0.45,\ 0.55)$ and $\gamma = 0.16$.

(B2) $\pi(\mathbf{x}, y) = \left(1 + \exp\left\{\beta_0 + \sum_{j=1}^{3} \beta_j x_j + \beta_4 \gamma\, x_4\right\} \cdot \exp\{\gamma y\}\right)^{-1}$

with $(\beta_0, \beta_1, \cdots, \beta_4) = (2.6,\ 0.6,\ 0.35,\ -0.45,\ 0.4)$ and $\gamma = -0.36$.

To estimate γ, we used the data-splitting approach outlined earlier in Sec. 2 with $m = 0.7n$ and $\ell = n - m = 0.3n$, where γ is selected to minimize (14) over a grid of equally-spaced values of γ in $[-M, M]$. Here, we took $M = 15$ but a smaller value such as $M = 5$ would have been sufficient. A small follow-up subsample was selected from the set of non-respondents in $\mathbb{D}_\ell$ with each non-respondent having probability p_n of being selected where $p_n = \left((\log n)^{0.25}/(n\lambda^d)^{1-\alpha}\right)^{1/2}$ with $\lambda = 0.95$ and $\alpha = 0.01$. This choice of p_n guarantees that the follow-up subsample sizes will be very small on average (see the results in tables 1 and 2). Here we used the Gaussian kernel where the smoothing parameters were selected using the cross-validation method of Racine and Li [29] available from the R package "np"; see Racine and Hayfield [2008].

To assess the performance of the proposed estimators we computed their empirical L_2 errors committed on a validation set of 1000 additional observations generated from the distribution of the data under each of models A and B. We also computed the empirical L_1 errors of our estimators. The above process was repeated a total of 400 times, each time using a sample of size n (50 and

then 100) and a validation set of size 1000, and the average empirical errors of the estimators in (16) were computed for π_0=0 (the top classifier in (16)) as well as $\pi_0 = 10^{-10}, 10^{-20}, \cdots, 10^{-300}$ (for the bottom classifier in (16)). The results were virtually the same. We also found the corresponding results for the complete-case estimator $m_n^{cc}(\mathbf{x})$ in (2). Finally, we computed the same errors for the usual estimator with no missing data; this allows us to see how different the results could have been with no missing values. Table 1 illustrates the results for the case of low noise ($\sigma_y = 0.5$) in

Table 1: Empirical L_1 and L_2 errors when $\sigma_y = 0.5$ (low noise) in models A and B. Here, the proposed estimator $\widehat{m}(\mathbf{x}; \widehat{\pi}_{\widehat{\varphi}_n})$ is as in (16), the complete-case estimator $m_n^{cc}(\mathbf{x})$ is given by (2), and the estimator $\widehat{m}_n(\mathbf{x})$ based on no missing data is given by (1). The numbers in parentheses are standard errors and those in square brackets are the average follow-up subsample sizes drawn from the set of non-respondents in $\mathbb{D}_\ell$.

Model	n	$\pi(\mathbf{x}, y)$	Errors	$\widehat{m}(\mathbf{x}; \widehat{\pi}_{\widehat{\varphi}_n})$	$m_n^{cc}(\mathbf{x})$	$\widehat{m}_n(\mathbf{x})$
A	50	A1	L_2	22.82 (0.3580), [1.62]	36.34 (0.4673)	15.92 (0.3983)
			L_1	3.09 (0.0191)	4.15 (0.0327)	2.54 (0.0282)
		A2	L_2	28.72 (0.3994), [1.60]	37.07 (0.5000)	15.71 (0.4314)
			L_1	3.46 (0.0186)	4.21 (0.0345)	2.38 (0.0240)
	100	A1	L_2	20.98 (0.2684), [1.99]	30.61 (0.3775)	11.89 (0.2305)
			L_1	2.88 (0.0196)	3.68 (0.0314)	2.07 (0.0153)
		A2	L_2	25.39 (0.3463), [2.08]	32.70 (0.4472)	13.31 (0.3420)
			L_1	3.16 (0.0198)	3.83 (0.0353)	2.14 (0.0218)
B	50	B1	L_2	51.64 (0.4477), [1.46]	55.64 (0.7276)	45.72 (0.5857)
			L_1	4.76 (0.0238)	5.88 (0.0584)	4.61 (0.0384)
		B2	L_2	53.31 (0.6894), [1.49]	57.69 (1.0273)	43.88 (0.6160)
			L_1	4.83 (0.0320)	6.09 (0.0749)	4.55 (0.0419)
	100	B1	L_2	48.77 (0.4280), [2.05]	55.30 (0.6627)	42.78 (0.6404)
			L_1	4.59 (0.0199)	5.92 (0.0538)	4.45 (0.0425)
		B2	L_2	49.00 (0.7387), [2.21]	55.56 (0.6890)	42.38 (0.6419)
			L_1	4.62 (0.0239)	6.04 (0.0550)	4.44 (0.0417)

Models A and B; the numbers in parentheses are the standard errors and those in square brackets are the average follow-up subsample sizes drawn from the set of non-respondents in $\mathbb{D}_\ell$ over 400 Monte Carlo runs. Table 2 gives the same results for the high noise setup ($\sigma_y = 4$) in models A and B. Figures 1 and 2 illustrate the boxplots of the L_2 errors.

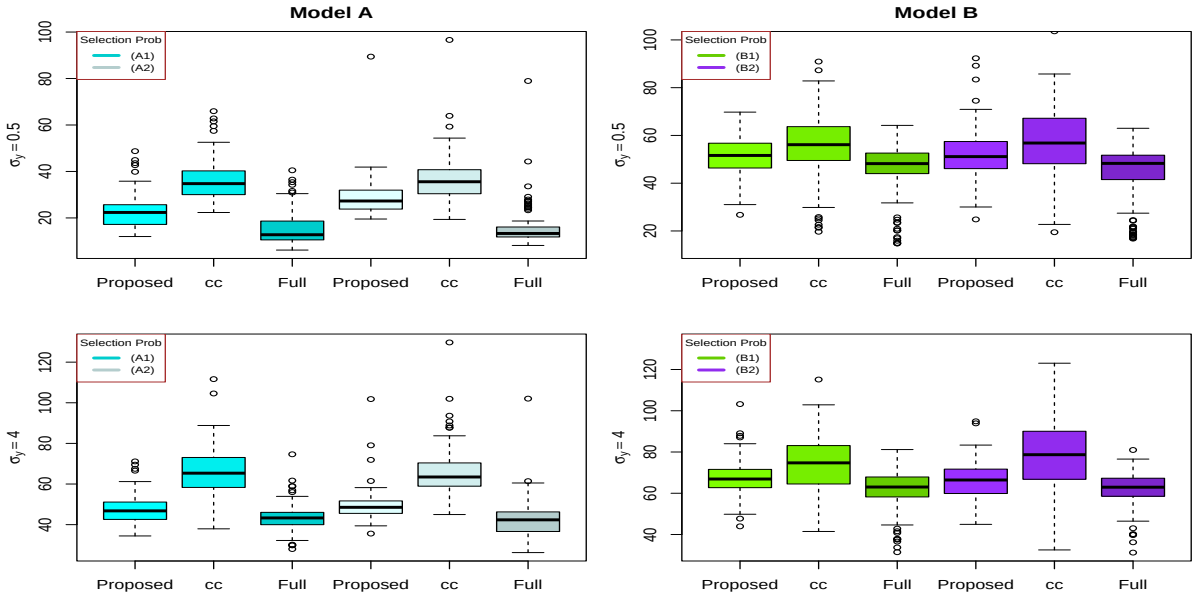

Figure 1: Box plots of the empirical L_2 error rates for the case of n=50.

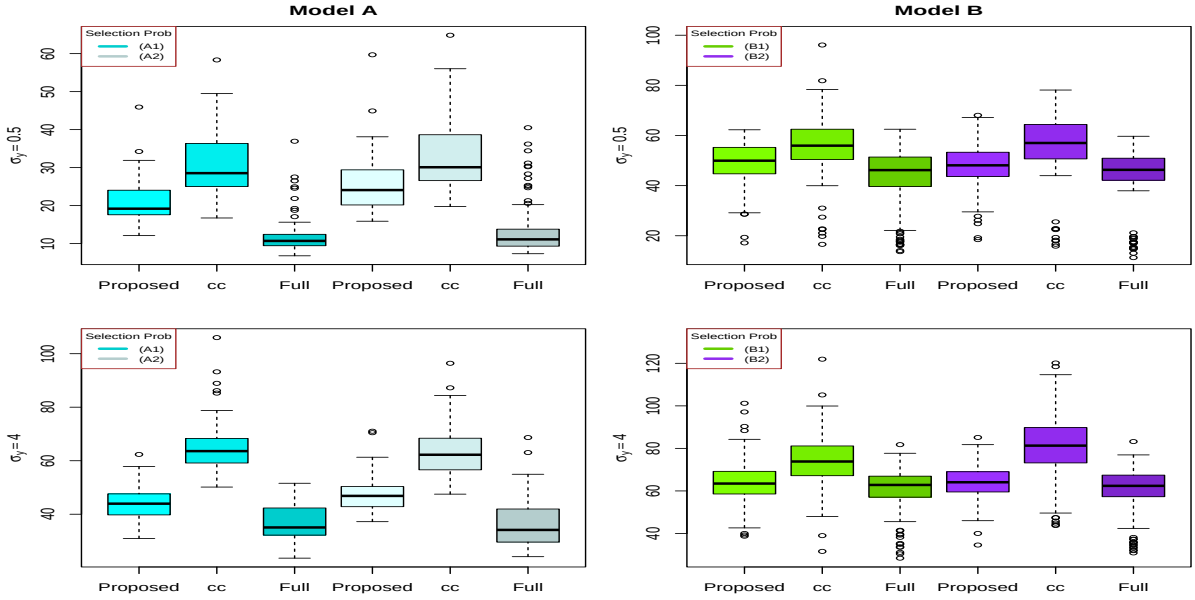

Figure 2: Box plots of the empirical L_2 error rates for the case of n=100.

As these results show, the proposed estimator shows good performance with errors that are significantly better than those of the complete case estimators, in fact, these tables show that the

error of the proposed estimator can sometimes be quite close to the one based on no missing data; see, for example, Table 2, rows 1,2, 3, 5, 6, 11, 12, 14, 15, and 16. Perhaps more importantly, the tables show that the average follow-up subsample sizes used for our proposed estimators is around 2 when $n = 100$ and about 1.5 when $n = 50$. In other words, the undesirable need for a follow-up subsample here is virtually a non-issue in practice.

Table 2: Empirical L_1 and L_2 errors when $\sigma_y = 4$ (high noise) in models A and B. Here, the proposed estimator $\widehat{m}(\mathbf{x}; \widehat{\pi}_{\widehat{\varphi}_n})$ is as in (16), the complete-case estimator $m_n^{cc}(\mathbf{x})$ is given by (2), and the estimator $\widehat{m}_n(\mathbf{x})$ based on no missing data is given by (1). The numbers in parentheses are standard errors and those in square brackets are the average follow-up subsample sizes drawn from the set of non-respondents in $\mathbb{D}_\ell$.

Model	n	$\pi(\mathbf{x}, y)$	Errors	$\widehat{m}(\mathbf{x}; \widehat{\pi}_{\widehat{\varphi}_n})$	$m_n^{cc}(\mathbf{x})$	$\widehat{m}_n(\mathbf{x})$
A	50	A1	L_2	47.28 (0.3584), [1.56]	66.72 (0.5866)	43.29 (0.3593)
			L_1	5.14 (0.0161)	6.30 (0.0306)	4.94 (0.0156)
		A2	L_2	49.40 (0.4065), [1.63]	65.81 (0.6055)	42.15 (0.4552)
			L_1	5.24 (0.0129)	6.23 (0.0253)	4.85 (0.0168)
	100	A1	L_2	43.66 (0.2802), [2.35]	64.58 (0.4400)	36.75 (0.3297)
			L_1	4.96 (0.0150)	6.23 (0.0236)	4.59 (0.0146)
		A2	L_2	47.29 (0.3230), [1.85]	63.39 (0.4634)	36.13 (0.4264)
			L_1	5.16 (0.0137)	6.14 (0.0234)	4.53 (0.0205)
B	50	B1	L_2	68.68 (0.4144), [1.35]	73.91 (0.4299)	61.07 (0.3908)
			L_1	6.01 (0.1225)	6.83 (0.1307)	5.72 (0.1195)
		B2	L_2	66.27 (0.4097), [1.32]	79.36 (0.9226)	61.82 (0.4473)
			L_1	5.90 (0.0191)	7.21 (0.0557)	5.74 (0.0214)
	100	B1	L_2	65.09 (0.6243), [2.09]	74.09 (0.6603)	60.23 (0.5538)
			L_1	5.80 (0.0222)	6.85 (0.0397)	5.63 (0.0245)
		B2	L_2	63.96 (0.4172), [2.11]	81.30 (0.8187)	59.47 (0.5854)
			L_1	5.84 (0.0197)	7.38 (0.0466)	5.64 (0.0270)

5 PROOFS OF THE MAIN RESULTS

We start by stating a number of lemmas. In what follows, we use the notation of Section 2, where $\mathcal{F}$ is a totally bounded class of functions $\varphi : [-L, L] \to (0, B]$, for some $B < \infty$. Also, for any $\varepsilon > 0$, let $\mathcal{F}_\varepsilon$ be any ε-cover of $\mathcal{F}$ (see Section 2). Let $\widehat{\pi}_\varphi(\mathbf{x}, y)$, $m(\mathbf{x}; \pi_\varphi)$, and $\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)$ be as in (8), (6), and (14), and for each $\varphi \in \mathcal{F}$ define the quantities

$$m(\mathbf{x}; \pi_\varphi) = E\Big[\frac{\Delta Y}{\pi_\varphi(\mathbf{X}, Y)}\Big|\mathbf{X} = \mathbf{x}\Big], \qquad \varphi_\varepsilon = \operatorname*{argmin}_{\varphi \in \mathcal{F}_\varepsilon} E\big|m(\mathbf{X}; \pi_\varphi) - Y\big|^2, \qquad \widehat{\varphi}_\varepsilon = \operatorname*{argmin}_{\varphi \in \mathcal{F}_\varepsilon} \widehat{L}_{m,\ell}(\widehat{\pi}_\varphi). \tag{28}$$

Lemma 1 *Let $m(\mathbf{x}; \pi_\varphi)$ be as in (28), where π_φ is as in (3), and suppose that assumption (G) holds. Then, for every φ_1 and φ_2 in $\mathcal{F}$,*

$$\big|m(\mathbf{x}; \pi_{\varphi_1}) - m(\mathbf{x}; \pi_{\varphi_2})\big| \;\le\; L \cdot \exp\big\{g(\mathbf{x})\big\} \sup_{-L \le y \le L} \big|\varphi_1(y) - \varphi_2(y)\big|,$$

where $g(\mathbf{x})$ is the function in (3) and L is as in assumption (F).

Proof of Lemma 1.

First observe that for each φ in $\mathcal{F}$,

$$m(\mathbf{x}; \pi_\varphi) = E\left[E\left(\frac{\Delta Y}{\pi_\varphi(\mathbf{X}, Y)}\,\middle|\,\mathbf{X}, Y\right)\middle|\mathbf{X}\right] = E\left[\frac{Y \cdot E(\Delta|\mathbf{X}, Y)}{\pi_\varphi(\mathbf{X}, Y)}\middle|\mathbf{X}\right] \overset{\text{by (3)}}{=} E\left[Y \cdot \frac{\pi_{\varphi^*}(\mathbf{X}, Y)}{\pi_\varphi(\mathbf{X}, Y)}\,\middle|\,\mathbf{X}\right].$$

Therefore, in view of the definition of π_φ in (3),

$$\begin{aligned}
\big|m(\mathbf{x}; \pi_{\varphi_1}) - m(\mathbf{x}; \pi_{\varphi_2})\big| &\le E\left[\big|Y \cdot \pi_{\varphi^*}(\mathbf{X}, Y)\big| \cdot \Big|\frac{1}{\pi_{\varphi_1}(\mathbf{X}, Y)} - \frac{1}{\pi_{\varphi_2}(\mathbf{X}, Y)}\Big|\,\middle|\,\mathbf{X} = \mathbf{x}\right] \\
&\le L \cdot E\left[\Big|1 + \exp\big\{g(\mathbf{X})\big\}\varphi_1(Y) - 1 - \exp\big\{g(\mathbf{X})\big\}\varphi_2(Y)\Big|\,\middle|\,\mathbf{X} = \mathbf{x}\right] \\
&\le L \cdot \exp\big\{g(\mathbf{x})\big\} \sup_{-L \le y \le L} \big|\varphi_1(y) - \varphi_2(y)\big|.
\end{aligned}$$

□

Lemma 2 *Let $\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)$ and $\widehat{m}_m(\mathbf{x}; \widehat{\pi}_\varphi)$ be as in (14) and (7), respectively. Also, let $m(\mathbf{x}; \pi_\varphi)$, φ_ε, and $\widehat{\varphi}_\varepsilon$ be as in (28). Then, under the conditions of Theorem 1, for every $\varepsilon > 0$ we have*

$$\begin{aligned}
E\left[\Big|\widehat{m}_m(\mathbf{X}; \widehat{\pi}_{\widehat{\varphi}_\varepsilon}) - m(\mathbf{X}; \pi_{\varphi_\varepsilon}))\Big|^2\Big|\mathbb{D}_n\right] \;\le\; & \sup_{\varphi \in \mathcal{F}_\varepsilon} \left|E\left[\Big|\widehat{m}_m(\mathbf{X}; \widehat{\pi}_\varphi) - Y\Big|^2\Big|\mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)\right| \\
& + \sup_{\varphi \in \mathcal{F}_\varepsilon} \left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi) - E\Big|m(\mathbf{X}; \pi_\varphi) - Y\Big|^2\right| \\
& + \varepsilon \cdot C_1 \sqrt{E\left[\Big|\widehat{m}_m(\mathbf{X}; \widehat{\pi}_{\widehat{\varphi}_\varepsilon}) - m(\mathbf{X}; \pi_{\varphi_\varepsilon})\Big|^2\Big|\mathbb{D}_n\right]},
\end{aligned} \tag{29}$$

where C_1 is a positive constant not depending on n or ε.

Proof of Lemma 2.

Observe that $E\big[|\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-Y|^2\big|\mathbb{D}_n\big] = E\big[|\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-m(\mathbf{X};\pi_{\varphi_\varepsilon})|^2\big|\mathbb{D}_n\big] + E|m(\mathbf{X};\pi_{\varphi_\varepsilon})-Y|^2$ $+2E\big[\big(\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-m(\mathbf{X};\pi_{\varphi_\varepsilon})\big)\big(m(\mathbf{X};\pi_{\varphi_\varepsilon})-Y\big)\big|\mathbb{D}_n\big]$. Also, let φ^* be as in (3) and note that

$$\begin{aligned}
&E\left[\Big(\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-m(\mathbf{X};\pi_{\varphi_\varepsilon})\Big)\Big(m(\mathbf{X};\pi_{\varphi_\varepsilon})-Y\Big)\Big|\mathbb{D}_n\right]\\
&= E\left[\Big(\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-m(\mathbf{X};\pi_{\varphi_\varepsilon})\Big)\Big(m(\mathbf{X};\pi_{\varphi_\varepsilon})-m(\mathbf{X};\pi_{\varphi^*})+m(\mathbf{X};\pi_{\varphi^*})-Y\Big)\Big|\mathbb{D}_n\right]\\
&= E\left[\Big(\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-m(\mathbf{X};\pi_{\varphi_\varepsilon})\Big)\Big(m(\mathbf{X};\pi_{\varphi_\varepsilon})-m(\mathbf{X};\pi_{\varphi^*})\Big)\Big|\mathbb{D}_n\right],
\end{aligned}$$

where we have used the fact that in view of (5), $E[Y|\mathbf{X}=\mathbf{x}] := m(\mathbf{x}) = m(\mathbf{x};\pi_{\varphi^*})$. Therefore

$$\begin{aligned}
&E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-m(\mathbf{X};\pi_{\varphi_\varepsilon})\Big|^2\Big|\mathbb{D}_n\right]\\
&= \left\{E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-Y\Big|^2\Big|\mathbb{D}_n\right]-E\Big|m(\mathbf{X};\varphi_\varepsilon)-Y\Big|^2\right\}\\
&\quad -2E\left[\Big(\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-m(\mathbf{X};\pi_{\varphi_\varepsilon})\Big)\Big(m(\mathbf{X};\pi_{\varphi_\varepsilon})-m(\mathbf{X};\pi_{\varphi^*})\Big)\Big|\mathbb{D}_n\right]\\
&:= \mathbf{I}_n+\mathbf{II}_n. \qquad (30)
\end{aligned}$$

To deal with the term $\mathbf{I}_n$, let $\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)$ be as in (30) and observe that

$$\begin{aligned}
\mathbf{I}_n &= E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-Y\Big|^2\Big|\mathbb{D}_n\right]-\inf_{\varphi\in\mathcal{F}_\varepsilon}E\Big|m(\mathbf{X};\pi_\varphi)-Y\Big|^2\\
&= \sup_{\varphi\in\mathcal{F}_\varepsilon}\left\{E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-Y\Big|^2\Big|\mathbb{D}_n\right]-\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)+\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)-\widehat{L}_{m,\ell}(\widehat{\pi}_{\widehat{\varphi}_\varepsilon})\right.\\
&\qquad \left.+\,\widehat{L}_{m,\ell}(\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-E\Big|m(\mathbf{X};\pi_\varphi)-Y\Big|^2\right\}\\
&\le \left(E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-Y\Big|^2\Big|\mathbb{D}_n\right]-\widehat{L}_{m,\ell}(\widehat{\pi}_{\widehat{\varphi}_\varepsilon})\right)+\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)-E\Big|m(\mathbf{X};\pi_\varphi)-Y\Big|^2\right|,
\end{aligned}$$

where the last line follows since $\widehat{L}_{m,\ell}(\widehat{\pi}_{\widehat{\varphi}_\varepsilon}) \le \widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)$ holds for all $\varphi\in\mathcal{F}_\varepsilon$ (because of the definition of $\widehat{\varphi}_\varepsilon$ in (28)). Therefore

$$|\mathbf{I}_n| \le \sup_{\varphi\in\mathcal{F}_\varepsilon}\left|E\left[\Big|\widehat{m}_m(\mathbf{X};\pi_\varphi)-Y\Big|^2\Big|\mathbb{D}_m\right]-\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)\right|+\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)-E\Big|m(\mathbf{X};\pi_\varphi)-Y\Big|^2\right|, \quad (31)$$

where the conditioning on $\mathbb{D}_m$ in the above expression reflects the fact that $\widehat{m}_m(\mathbf{X};\varphi)$ depends on $\mathbb{D}_m$ only (and not the entire data $\mathbb{D}_n$). Furthermore, by Cauchy-Schwarz inequality

$$|\mathbf{II}_n| \le 2\sqrt{E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon})-m(\mathbf{X};\pi_{\varphi_\varepsilon})\Big|^2\Big|\mathbb{D}_n\right]}\cdot\sqrt{E\Big|m(\mathbf{X};\pi_{\varphi_\varepsilon})-m(\mathbf{X};\pi_{\varphi^*})\Big|^2}. \quad (32)$$

Now, let $\varphi^\dagger \in \mathcal{F}_\varepsilon$ be such that $\varphi^* \in B(\varphi^\dagger, \varepsilon)$; such a $\varphi^\dagger \in \mathcal{F}_\varepsilon$ exists because $\varphi^* \in \mathcal{F}$ and $\mathcal{F}_\varepsilon$ is an ε-cover of $\mathcal{F}$. Then, using the fact that $E\big|m(\mathbf{X};\pi_{\varphi_\varepsilon}) - Y\big|^2 = E\big|m(\mathbf{X};\pi_{\varphi^*}) - Y\big|^2 + E\big|m(\mathbf{X};\pi_{\varphi_\varepsilon}) - m(\mathbf{X};\pi_{\varphi^*})\big|^2$, one finds

$$\begin{aligned} E\big|m(\mathbf{X};\pi_{\varphi_\varepsilon}) - m(\mathbf{X};\pi_{\varphi^*})\big|^2 &= \inf_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\pi_\varphi) - Y\big|^2 - E\big|m(\mathbf{X};\pi_{\varphi^*}) - Y\big|^2 \\ &= \inf_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\pi_\varphi) - m(\mathbf{X};\pi_{\varphi^*})\big|^2 \;\le\; E\big|m(\mathbf{X};\pi_{\varphi^\dagger}) - m(\mathbf{X};\pi_{\varphi^*})\big|^2 \\ &\le L^2\, E\big(\exp\{2g(\mathbf{X})\}\big)\Big[\sup_{-L\le y\le L}\big|\varphi^\dagger(y) - \varphi^*(y)\big|\Big]^2, \quad \text{(by Lemma 1)} \\ &\le C\cdot\varepsilon^2\,, \end{aligned} \tag{33}$$

because $\varphi^* \in B(\varphi^\dagger, \varepsilon)$, where $C = L^2 E\big(\exp\{2g(\mathbf{X})\}\big) < \infty$ is a positive constant not depending on n or ε (see the second part of assumption (D)). Therefore, in view of (32) and (33), one finds

$$\big|\mathbf{II}_n\big| \;\le\; \varepsilon\cdot C_1\sqrt{E\Big[\big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_\varepsilon}) - m(\mathbf{X};\pi_{\varphi_\varepsilon})\big|^2\Big|\mathbb{D}_n\Big]}, \tag{34}$$

for a constant $C_1 > 0$ not depending on n or ε. Now Lemma 2 follows from (30), (31), and (34).

□

Proof of Theorem 1

Part (i). First observe that by Lemma 2 one can write

$$\begin{aligned} &\int \big|\widehat{m}_m(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_\varepsilon}) - m(\mathbf{x};\pi_{\varphi_\varepsilon})\big|^2\mu(d\mathbf{x}) \\ &\quad\le \sup_{\varphi\in\mathcal{F}_\varepsilon}\left|E\left[\big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_\varphi) - Y\big|^2\Big|\mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)\right| + \sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi) - E\Big|m(\mathbf{X};\pi_\varphi) - Y\Big|^2\right| \\ &\qquad+\; \varepsilon\cdot C_1\sqrt{\int\big|\widehat{m}_m(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_\varepsilon}) - m(\mathbf{x};\pi_{\varphi_\varepsilon})\big|^2\mu(d\mathbf{x})}\,, \end{aligned} \tag{35}$$

where, φ_ε and $\widehat{\varphi}_\varepsilon$ are as in (28). Therefore, in view of (35), for every constant $t > 0$

$$\begin{aligned} &P\left\{\int\Big|\widehat{m}_m(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_\varepsilon}) - m(\mathbf{x};\pi_{\varphi_\varepsilon})\Big|^2\mu(d\mathbf{x}) > t\right\} - P\left\{\int\Big|\widehat{m}_m(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_\varepsilon}) - m(\mathbf{x};\pi_{\varphi_\varepsilon})\Big|^2\mu(d\mathbf{x}) > \frac{t^2}{\varepsilon^2 c_4}\right\} \\ &\qquad\le\; P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_\varphi) - Y\Big|^2\Big|\mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)\right| > \frac{t}{3}\right\} \\ &\qquad\qquad + P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi) - E\Big|m(\mathbf{X};\pi_\varphi) - Y\Big|^2\right| > \frac{t}{3}\right\}, \end{aligned} \tag{36}$$

where $c_4 = (3C_1)^2$ with C_1 as in (35). But observe that for every constant $\beta > 0$

$$P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi) - E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_\varphi) - Y\Big|^2\Big|\mathbb{D}_m\right]\right| > \beta\right\} \tag{37}$$

$$\leq P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Big(\Delta_i+\frac{(1-\Delta_i)\delta_i}{p_n}\Big)\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)-Y_i\Big|^2-E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_\varphi)-Y\Big|^2\Big|\mathbb{D}_m\right]\right|>\beta\right\}.$$

On the other hand, for every $i\in\boldsymbol{\mathcal{I}}_\ell$, we have

$$\begin{aligned}
&E\left[\Big(\Delta_i+\frac{(1-\Delta_i)\delta_i}{p_n}\Big)\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)-Y_i\Big|^2\Big|\mathbb{D}_m\right]\\
&=E\left[\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)-Y_i\Big|^2\left(E\left\{\Delta_i\Big|\mathbb{D}_m,\mathbf{X}_i,Y_i,\delta_i\right\}+\frac{\delta_i}{p_n}E\left\{1-\Delta_i\Big|\mathbb{D}_m,\mathbf{X}_i,Y_i,\delta_i\right\}\right)\Big|\mathbb{D}_m\right]\\
&=E\left[\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)-Y_i\Big|^2\pi_{\varphi^*}(\mathbf{X}_i,Y_i)\Big|\mathbb{D}_m\right]+\frac{E(\delta_i)}{p_\ell}E\left[\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)-Y_i\Big|^2\big(1-\pi_{\varphi^*}(\mathbf{X}_i,Y_i)\big)\Big|\mathbb{D}_m\right]\\
&=E\left[\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)-Y_i\Big|^2\Big|\mathbb{D}_m\right],
\end{aligned}$$

because δ_i is independent of the data $\mathbb{D}_n$ (with $E(\delta_i)=p_n$), and the fact that Δ_i is independent of $\mathbb{D}_m$ for all $i\in\boldsymbol{\mathcal{I}}_\ell$. Moreover, by the definition of $\widehat{\pi}_\varphi(\mathbf{X},Y)$ in (8), one finds

$$\big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)\big| \;\leq\; \max_{k\in\boldsymbol{\mathcal{I}}_m}\big|\Delta_kY_k/\widehat{\pi}_\varphi(\mathbf{X}_k,Y_k)\big| \;\leq\; L\cdot\left(1+\max_{k\in\boldsymbol{\mathcal{I}}_m}\left|\frac{1}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)}\right|\cdot B\right). \tag{38}$$

Thus, conditional on $\mathbb{D}_m$, the terms $\big(\Delta_i+(1-\Delta_i)\delta_i/p_n\big)\big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)-Y_i\big|^2$, $i\in\boldsymbol{\mathcal{I}}_\ell$, are independent nonnegative random variables bounded by $(2L^2/p_n)\big\{4+B^2\max^2_{k\in\boldsymbol{\mathcal{I}}_m}\big|1/\widehat{\psi}_m(\mathbf{X}_k;\varphi)\big|\big\}$. Therefore

$$\begin{aligned}
(37) \;&\leq\; \big|\mathcal{F}_\varepsilon\big|\sup_{\varphi\in\mathcal{F}_\varepsilon}E\left[P\left\{\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)-E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_\varphi)-Y\Big|^2\Big|\mathbb{D}_m\right]\right|>\beta\Big|\mathbb{D}_m\right\}\right]\\
&\leq\; 2\big|\mathcal{F}_\varepsilon\big|\sup_{\varphi\in\mathcal{F}_\varepsilon}E\left[\exp\left\{\frac{-2\ell\,\beta^2\,p_n^2}{4L^4\left\{4+B^2\max^2_{k\in\boldsymbol{\mathcal{I}}_m}\big|1/\widehat{\psi}_m(\mathbf{X}_k;\varphi)\big|\right\}^2}\right\}\right],
\end{aligned} \tag{39}$$

via Hoeffding's inequality. Now let ϱ_0 be the constant in Assumption (D) and observe that since the exponential function in (39) is always bounded by 1, the expectation on the right side of (39) is bounded by

$$\begin{aligned}
&E\left[\exp\left\{\frac{-2\ell\,\beta^2\,p_n^2}{4L^4\left\{4+\max^2_{k\in\boldsymbol{\mathcal{I}}_m}\big|B/\widehat{\psi}_m(\mathbf{X}_k;\varphi)\big|\right\}^2}\right\}\mathbb{I}\Big\{\bigcap_{k\in\boldsymbol{\mathcal{I}}_m}\Big[\widehat{\psi}_m(\mathbf{X}_k;\varphi)\geq\frac{\varrho_0}{2}\Big]\Big\}\right]\\
&\quad+E\left[\mathbb{I}\Big\{\bigcup_{k\in\boldsymbol{\mathcal{I}}_m}\Big[\widehat{\psi}_m(\mathbf{X}_k;\varphi)<\varrho_0/2\Big]\Big\}\right]\\
&\leq\;\exp\left\{\frac{-2\ell\,\beta^2\,p_n^2}{4L^4\left\{4+B^2(2/\varrho_0)^2\right\}^2}\right\}+\sum_{k\in\boldsymbol{\mathcal{I}}_m}P\Big\{\widehat{\psi}_m(\mathbf{X}_k;\varphi)<\varrho_0/2\Big\}.
\end{aligned} \tag{40}$$

Now, as in (10), let $\psi(\mathbf{X}_k;\varphi)=E\big[\Delta_k\varphi(Y_k)\big|\mathbf{X}_k\big]$ and observe that by assumption (D) one can write $P\big\{\widehat{\psi}_m(\mathbf{X}_k;\varphi)<\frac{\varrho_0}{2}\big\}\leq P\big\{-\widehat{\psi}_m(\mathbf{X}_k;\varphi)+\psi(\mathbf{X}_k;\varphi)>\varrho_0-\frac{\varrho_0}{2}\big\}\leq P\big\{\big|\widehat{\psi}_m(\mathbf{X}_k;\varphi)-\psi(\mathbf{X}_k;\varphi)\big|>$

$\frac{\varrho_0}{2}\big\}$. However, straightforward but tedious arguments (as in Mojirsheibani [23]) show that, under assumptions (B), (C), and (E)

$$P\big\{\big|\widehat{\psi}_m(\mathbf{X}_k;\varphi)-\psi(\mathbf{X}_k;\varphi)\big|>\varrho_0/2\big\}\;\le\;C_{16}\,\exp\big\{-C_{17}mh^d\big\}, \tag{41}$$

for n (and thus m) large enough and positive constants C_{16} and C_{17} not depending on n. Thus, in view of (37), (39), (40), and (41), for every $\beta>0$ and n large enough one finds

$$\begin{aligned} &P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)-E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\pi}_\varphi)-Y\Big|^2\Big|\mathbb{D}_m\right]\right|>\beta\right\}\\ &\qquad\le\;2\,\big|\mathcal{F}_\varepsilon\big|\,\left(\exp\left\{\frac{-2\ell\,\beta^2\,p_n^2}{4L^4\,\{4+B^2(2/\varrho_0)^2\}^2}\right\}+C_{16}\,m\,\exp\big\{-C_{17}mh^d\big\}\right). \end{aligned}\tag{42}$$

To complete the proof, we also need to deal with the last probability statement on the right side of (36). To that end, let $\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)$ be as in (14) and define the quantities

$$Q_{n,1}(\varphi)\;=\;\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)-\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Big(\Delta_i+\frac{(1-\Delta_i)\delta_i}{p_n}\Big)\Big|m(\mathbf{X}_i;\pi_\varphi)-Y_i\Big|^2\right| \tag{43}$$

$$Q_{n,2}(\varphi)\;=\;\left|\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Big(\Delta_i+\frac{(1-\Delta_i)\delta_i}{p_n}\Big)\Big|m(\mathbf{X}_i;\pi_\varphi)-Y_i\Big|^2-E\big|m(\mathbf{X};\pi_\varphi)-Y\big|^2\right| \tag{44}$$

and observe that for every $\beta>0$,

$$\begin{aligned} &P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)-E\Big|m(\mathbf{X};\pi_\varphi)-Y\Big|^2\right|>\beta\right\}\\ &\qquad\le\;P\Big\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\big|Q_{n,1}(\varphi)\big|>\frac{\beta}{2}\Big\}+P\Big\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\big|Q_{n,2}(\varphi)\big|>\frac{\beta}{2}\Big\}\;=:\;Q_{n,1}+Q_{n,2}\,. \end{aligned}\tag{45}$$

But $\big(\Delta_i+((1-\Delta_i)\delta_i)/p_n\big)\le 1/p_n$ and

$$\big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)\big|\;\le\;L\max_{1\le k\le m}\left|1+\frac{1-\widehat{\eta}_m(\mathbf{X}_k)}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)}\,\varphi(Y_k)\right|\;\le\;L+L\max_{1\le k\le m}\left|\frac{B}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)}\right|.$$

Therefore, by the definition of $\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)$ in (14) and the fact that $|a^2-b^2|\le|a-b||a+b|$, one has

$$\begin{aligned} Q_{n,1}\;&\le\;P\Bigg\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Big[\Big(\Delta_i+\frac{(1-\Delta_i)\delta_i}{p_n}\Big)\\ &\qquad\times\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)-m(\mathbf{X}_i;\pi_\varphi)\Big|\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)+m(\mathbf{X}_i;\pi_\varphi)-2Y_i\Big|\Big]>\frac{\beta}{2}\Bigg\}\\ &\le\;P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\left[\frac{1}{p_n}\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi)-m(\mathbf{X}_i;\pi_\varphi)\Big|\left(4+\max_{k\in\boldsymbol{\mathcal{I}}_m}\left|\frac{B}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)}\right|\right)\right]\right|>\frac{\beta}{2L}\right\} \end{aligned}$$

$$
\begin{aligned}
&\leq \ \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}_\varepsilon} \sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \Bigg(P\Bigg\{ \Bigg[\Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi) - m(\mathbf{X}_i;\pi_\varphi)\Big|\big(4+B/(\varrho_0/2)\big) > \frac{\beta\, p_n}{2L} \Bigg] \\
&\qquad\qquad \cap \Bigg[\bigcap_{k\in\boldsymbol{\mathcal{I}}_m} \Big\{ \widehat{\psi}_m(\mathbf{X}_k;\varphi) \geq \frac{\varrho_0}{2} \Big\} \Bigg] \Bigg\} + \sum_{k\in\boldsymbol{\mathcal{I}}_m} P\Big\{ \widehat{\psi}_m(\mathbf{X}_k;\varphi) < \frac{\varrho_0}{2} \Big\} \Bigg) \\
&\leq \ \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}_\varepsilon} \sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \Bigg(P\Big\{ \Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi) - m(\mathbf{X}_i;\pi_\varphi)\Big| > C_\ell(\beta) \Big\} + \sum_{k\in\boldsymbol{\mathcal{I}}_m} P\Big\{ \widehat{\psi}_m(\mathbf{X}_k;\varphi) < \frac{\varrho_0}{2} \Big\} \Bigg),
\end{aligned}
\tag{46}
$$

where

$$
C_\ell(\beta) = \beta\, p_n \big/ \big(2L\,(4+2B/\varrho_0)\,\big). \tag{47}
$$

But the first probability statement in (46) can be bounded as follows. First, observe that

$$
\begin{aligned}
&P\Big\{ \Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi) - m(\mathbf{X}_i;\pi_\varphi)\Big| > C_\ell(\beta) \Big\} \\
&\leq P\Big\{ \Big|\widehat{m}_m(\mathbf{X}_i;\widehat{\pi}_\varphi) - \widehat{m}_m(\mathbf{X}_i;\pi_\varphi)\Big| > \frac{C_\ell(\beta)}{2} \Big\} + P\Big\{ \Big|\widehat{m}_m(\mathbf{X}_i;\pi_\varphi) - m(\mathbf{X}_i;\pi_\varphi)\Big| > \frac{C_\ell(\beta)}{2} \Big\} \\
&:= \mathcal{P}_{n1}(\beta) + \mathcal{P}_{n2}(\beta).
\end{aligned}
\tag{48}
$$

However, in view of (8) and (11) we can write

$$
\begin{aligned}
\mathcal{P}_{n1}(\beta) \ &= \ P\Bigg\{ \Bigg| \sum_{k\in\boldsymbol{\mathcal{I}}_m} \Big[\Big(\widehat{\pi}_\varphi(\mathbf{X}_k,Y_k)\Big)^{-1} - \Big(\pi_\varphi(\mathbf{X}_k,Y_k)\Big)^{-1} \Big] \frac{\Delta_K Y_k \mathcal{K}((\mathbf{X}_i-\mathbf{X}_k)/h)}{\sum_{j\in\boldsymbol{\mathcal{I}}_m}\mathcal{K}((\mathbf{X}_i-\mathbf{X}_j)/h)} \Bigg| > \frac{C_\ell(\beta)}{2} \Bigg\} \\
&\leq \ P\Bigg\{ \max_{k\in\boldsymbol{\mathcal{I}}_m} \Big| \Big(\widehat{\pi}_\varphi(\mathbf{X}_k,Y_k)\Big)^{-1} - \Big(\pi_\varphi(\mathbf{X}_k,Y_k)\Big)^{-1} \Big| > \frac{C_\ell(\beta)}{2L} \Bigg\} \\
&\leq \ \sum_{k\in\boldsymbol{\mathcal{I}}_m} P\Big\{ \Big| \Big(\widehat{\pi}_\varphi(\mathbf{X}_k,Y_k)\Big)^{-1} - \Big(\pi_\varphi(\mathbf{X}_k,Y_k)\Big)^{-1} \Big| > \frac{C_\ell(\beta)}{2L} \Big\} \\
&\leq \ \sum_{k\in\boldsymbol{\mathcal{I}}_m} E\Bigg[P\Bigg\{ \Bigg| \frac{1-\widehat{\eta}_m(\mathbf{X}_k)}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)} - \frac{E\big[1-\Delta\big|\mathbf{X}_k\big]}{E\big[\Delta\varphi(Y)\big|\mathbf{X}_k\big]} \Bigg| \cdot \varphi(Y_k) > \frac{C_\ell(\beta)}{2L} \Bigg| \mathbf{X}_k, Y_k \Bigg\} \Bigg],
\end{aligned}
\tag{49}
$$

where $\widehat{\psi}_m(\mathbf{X}_k;\varphi)$ and $\widehat{\eta}_m(\mathbf{X}_k)$ are as in (9). Next to bound (49), let $\eta(\mathbf{X}) = E[\Delta|\mathbf{X}]$ and $\psi(\mathbf{X};\varphi) = E[\Delta\varphi(Y)|\mathbf{X}]$ be as in (10) and observe that

$$
\begin{aligned}
\left| \frac{1-\widehat{\eta}_m(\mathbf{X}_k)}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)} - \frac{E\big[1-\Delta\big|\mathbf{X}_k\big]}{E\big[\Delta\varphi(Y)\big|\mathbf{X}_k\big]} \right| \ &= \ \left| -\frac{1-\widehat{\eta}_m(\mathbf{X}_k)}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)} \cdot \frac{\widehat{\psi}_m(\mathbf{X}_k;\varphi) - \psi(\mathbf{X}_k;\varphi)}{\psi(\mathbf{X}_k;\varphi)} + \frac{\eta(\mathbf{x}_k) - \widehat{\eta}_m(\mathbf{X}_k)}{\psi(\mathbf{X}_k;\varphi)} \right| \\
&\leq \ \left| \frac{1-\widehat{\eta}_m(\mathbf{X}_k)}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)} \right| \cdot \left| \frac{\widehat{\psi}_m(\mathbf{X}_k;\varphi) - \psi(\mathbf{X}_k;\varphi)}{\psi(\mathbf{X}_k;\varphi)} \right| + \left| \frac{\eta(\mathbf{x}) - \widehat{\eta}_m(\mathbf{X}_k)}{\psi(\mathbf{X}_k;\varphi)} \right|.
\end{aligned}
$$

Therefore, the conditional probability in (49) can be bounded as follows

$$
P\Bigg\{ \Bigg| \frac{1-\widehat{\eta}_m(\mathbf{X}_k)}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)} - \frac{E\big[1-\Delta\big|\mathbf{X}_k\big]}{E\big[\Delta\varphi(Y)\big|\mathbf{X}_k\big]} \Bigg| \cdot \varphi(Y_k) \ > \ \frac{C_\ell(\beta)}{2L} \ \Bigg| \mathbf{X}_k, Y_k \Bigg\}
$$

$$
\begin{aligned}
&\leq\ P\left\{\left|\frac{1-\widehat{\eta}_m(\mathbf{X}_k)}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)}\right|\cdot\left|\widehat{\psi}_m(\mathbf{X}_k;\varphi)-\psi(\mathbf{X}_k;\varphi)\right|>\frac{C_\ell(\beta)\varrho_0}{4BL}\middle|\mathbf{X}_k,Y_k\right\}\\
&\qquad+\ P\left\{\left|\eta(\mathbf{X}_k)-\widehat{\eta}_m(\mathbf{X}_k)\right|>\frac{C_\ell(\beta)\varrho_0}{4BL}\middle|\mathbf{X}_k,Y_k\right\}\\
&=:\ P_{n,1}(k)+P_{n,2}(k),
\end{aligned}
\tag{50}
$$

where we have used the facts that $\varphi(y)\in(0,B]$, $B>0$, and $\psi(\mathbf{x};\varphi)\geq\varrho_0$ holds by assumption (D). But using standard arguments it is not difficult to show that (B), (C), and (E), and n (and thus m) large enough

$$
P_{n,2}(k)\ \leq\ C_{12}\,e^{-C_{13}\,mh^d\,p_n^2\,\beta^2},
\tag{51}
$$

where C_{12} and C_{13} are positive constants not depending on k, m, ℓ, or β. As for the term $P_{n,1}(k)$ in (50), put

$$
\mathcal{B}_m(\mathbf{X}_k)=\big\{\widehat{\psi}_m(\mathbf{X}_k;\varphi)\geq\varrho_0/2\big\},
$$

where ϱ_0 is as in assumption (D), and note that

$$
\begin{aligned}
P_{n,1}(k)\ &\leq\ P\Bigg\{\left[\left|\frac{1-\widehat{\eta}_m(\mathbf{X}_k)}{\widehat{\psi}_m(\mathbf{X}_k;\varphi)}\right|\cdot\left|\widehat{\psi}_m(\mathbf{X}_k;\varphi)-\psi(\mathbf{X}_k;\varphi)\right|>\frac{C_\ell(\beta)\varrho_0}{4BL}\right]\cap\mathcal{B}_m(\mathbf{X}_k)\Big|\mathbf{X}_k,Y_k\Bigg\}\\
&\qquad+\ P\Big\{\mathcal{B}_m^c(\mathbf{X}_k)\big|\mathbf{X}_k,Y_k\Big\}\\
&:=\ P'_{n,1}(k)+P''_{n,1}(k).
\end{aligned}
$$

However, since $|1-\widehat{\eta}_m(\mathbf{X}_k)|\leq 1$, straightforward but tedious arguments show that

$$
P'_{n,1}(k)\ \leq\ P\Bigg\{\left|\widehat{\psi}_m(\mathbf{X}_k;\varphi)-\psi(\mathbf{X}_k;\varphi)\right|>\frac{C_\ell(\beta)\varrho_0^2}{8BL}\middle|\mathbf{X}_k,Y_k\Bigg\}\ \leq\ C_{14}\,e^{-C_{15}\,mh^d\,p_n^2\,\beta^2},
$$

for n (and thus m) large enough, where C_{14} and C_{15} are positive constants not depending on m, ℓ, or β. As for the term $P''_{n,1}(k)$, we have $P''_{n,1}(k)=P\big\{\widehat{\psi}_m(\mathbf{X}_k;\varphi)-\psi(\mathbf{X}_k;\varphi)<\varrho_0/2-\psi(\mathbf{X}_k;\varphi)\big|\mathbf{X}_k,Y_k\big\}\leq P\big\{\big|\widehat{\psi}_m(\mathbf{X}_k;\varphi)-\psi(\mathbf{X}_k;\varphi)\big|>\varrho_0/2\,\big|\mathbf{X}_k,Y_k\big\}\leq C_{16}\,\exp\{-C_{17}\,mh^d\}$, where we have used the fact that ψ is bounded by assumption (D); here C_{16} and C_{17} are positive constant not depending on m or ℓ. Putting these bounds together, we find

$$
P_{n,1}(k)\ \leq\ P'_{n,1}(k)+P''_{n,1}(i)\ \leq\ C_{14}\,e^{-C_{15}\,mh^d\,p_n^2\,\beta^2}+C_{16}\,e^{-C_{17}\,mh^d}.
\tag{52}
$$

Therefore, putting together (52), (51), (50), and (49), one finds for n large enough

$$
\mathcal{P}_{n1}(\beta)\ \leq\ \sum_{j\in\boldsymbol{\mathcal{I}}_m}\big(P_{n,1}(k)+P_{n,2}(k)\big)\ \leq\ C_{18}\,me^{-C_{19}\,mh^d\,p_n^2\,\beta^2}+C_{16}\,me^{-C_{17}\,mh^d},
$$

where C_{16}–C_{19} are positive constants not depending on m, ℓ, or β. Regarding the term $\mathcal{P}_{n2}(\beta)$ in (48), tedious but straightforward arguments can be used to show that for n (and thus m) large

enough there are positive constants C_{20} and C_{21}, not depending on m, ℓ, or β, such that

$$\mathcal{P}_{n2}(\beta) \;\le\; C_{20}\, e^{-C_{21}\, mh^d\, p_n^2\, \beta^2}.$$

It was shown earlier (see the paragraph before (52)) that $P\big\{\widehat{\psi}_m(\mathbf{X}_k;\varphi) < \varrho_0/2 \big| \mathbf{X}_k, Y_k\big\} = P''_{n,1}(k) \le C_{16}\, \exp\{-C_{17}\, mh^d\}$, where C_{16} and C_{17} are positive constant not depending on m or ℓ. Therefore in view of (46) and (48) one finds

$$Q_{n,1} \;\le\; \ell\, \big|\mathcal{F}_\varepsilon\big| \left(C_{18}\, me^{-C_{19}\, mh^d\, p_n^2\, \beta^2} + C_{20}\, e^{-C_{21}\, mh^d\, p_n^2\, \beta^2} + 2C_{16}\, me^{-C_{17} m\, h^d} \right), \tag{53}$$

for n large enough, where $Q_{n,1}$ is as (45). To deal with the term $Q_{n,2}$ in (45), first we note that for $i \in \boldsymbol{\mathcal{I}}_\ell$ the terms $\big(\Delta_i + (1-\Delta_i)\delta_i/p_n\big) \cdot \big|m(\mathbf{X}_i;\pi_\varphi) - Y_i\big|^2$ are independent bounded random variables taking values in $\big[0\,,\, L^2\big(1+1/\pi_{\min}\big)^2/p_n\big]$. Therefore an application of Hoeffding's inequality gives

$$\begin{aligned} Q_{n,2} \;&\le\; \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}_\varepsilon} P\left\{ \left| \frac{1}{\ell} \sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \Big(\Delta_i + \frac{(1-\Delta_i)\delta_i}{p_n} \Big) \Big| m(\mathbf{X}_i;\pi_\varphi) - Y_i \Big|^2 - E\big| m(\mathbf{X};\pi_\varphi) - Y \big|^2 \right| > \frac{\beta}{2} \right\} \\ &\le\; 2\, \big|\mathcal{F}_\varepsilon\big| \exp\Big\{ -2\ell\, p_n^2\, (\beta/2)^2 / \big[L^4 (1+1/\pi_{\min})^4 \big] \Big\}. \end{aligned} \tag{54}$$

Putting together (45), (53), and (54), for every $\beta > 0$ and n large enough, one has

$$\begin{aligned} P\left\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} \Big| \widehat{L}_{m,\ell}(\widehat{\pi}_\varphi) - E\Big| m(\mathbf{X};\pi_\varphi) - Y \Big|^2 \Big| > \beta \right\} \;&\le\; \big|\mathcal{F}_\varepsilon\big| \Big(C_{18} \ell m e^{-C_{19} m h^d\, p_n^2\, \beta^2} + C_{22} \ell m e^{-C_{23} m h^d} \\ &\qquad + \; 2\, e^{-C_{24} \ell\, p_n^2\, \beta^2} \Big). \end{aligned} \tag{55}$$

Now, for any decreasing sequence $0 < \varepsilon_n \downarrow 0$, let $m(\mathbf{x};\pi_{\varphi_{\varepsilon_n}})$ and φ_{ε_n} be as in (28). Then arguments similar to those that led to (33) gives

$$\begin{aligned} \int \Big| \widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n}) - m(\mathbf{x}) \Big|^2 \mu(d\mathbf{x}) \;&=\; \int \Big| \widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n}) - m(\mathbf{x};\pi_{\varphi_{\varepsilon_n}}) + m(\mathbf{x};\pi_{\varphi_{\varepsilon_n}}) - m(\mathbf{x}) \Big|^2 \mu(d\mathbf{x}) \\ &\le\; 2 \int \Big| \widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n}) - m(\mathbf{x};\pi_{\varphi_{\varepsilon_n}}) \Big|^2 \mu(d\mathbf{x}) \\ &\qquad + \; 2 \int \Big| m(\mathbf{x};\pi_{\varphi_{\varepsilon_n}}) - m(\mathbf{x};\pi_{\varphi^*}) \Big|^2 \mu(d\mathbf{x}) \\ &\le\; 2 \int \Big| \widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n}) - m(\mathbf{x};\varphi_{\varepsilon_n}) \Big|^2 \mu(d\mathbf{x}) \;+\; 2C\, \varepsilon_n^2\,, \end{aligned} \tag{56}$$

because $m(\mathbf{x}) = m(\mathbf{x};\pi_{\varphi^*})$, where $C = L^2 E\big(\exp\{2g(\mathbf{X})\}\big) < \infty$. Therefore, in view of (56) and (36), for every constant $t > 0$ we have

$$\begin{aligned} &\frac{1}{2}\, P\left\{ \int \Big| \widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n}) - m(\mathbf{x}) \Big|^2 \mu(d\mathbf{x}) \;>\; t \right\} \\ &\qquad \le\; \frac{1}{2}\, P\left\{ \int \Big| \widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n}) - m(\mathbf{x};\varphi_{\varepsilon_n}) \Big|^2 \mu(d\mathbf{x}) > \; t/2 - C\varepsilon_n^2 \right\} \end{aligned}$$

$$
\begin{aligned}
&\leq\ P\left\{\int\left|\widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n})-m(\mathbf{x};\varphi_{\varepsilon_n})\right|^2\mu(d\mathbf{x})>\ t/2-C\varepsilon_n^2\right\}\\
&\quad-P\left\{\int\left|\widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n})-m(\mathbf{x};\varphi_{\varepsilon_n})\right|^2\mu(d\mathbf{x})>\ (t/2-C\varepsilon_n^2)^2/(c_4\,\varepsilon_n^2)\right\}\\
&\qquad\text{(for } n \text{ large enough, where } c_4>0 \text{ is as in the first line of the l.h.s. of (36))}\\
&\leq\ P\left\{\sup_{\varphi\in\mathcal{F}_{\varepsilon_n}}\left|E\left[\left|\widehat{m}_m(\mathbf{X};\widehat{\pi}_\varphi)-Y\right|^2\middle|\mathbb{D}_m\right]-\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)\right|>\frac{t/2-C\varepsilon_n^2}{3}\right\}\\
&\qquad+P\left\{\sup_{\varphi\in\mathcal{F}_{\varepsilon_n}}\left|\widehat{L}_{m,\ell}(\widehat{\pi}_\varphi)-E\middle|m(\mathbf{X};\pi_\varphi)-Y\right|^2\right|>\frac{t/2-C\varepsilon_n^2}{3}\right\}.
\end{aligned}
$$

Finally, choosing n large enough so that $(t/2-C\varepsilon_n^2)/3>t/12$ and using the bounds in (42) and (55), we find

$$
\begin{aligned}
P\left\{\int\left|\widehat{m}(\mathbf{x};\widehat{\pi}_{\widehat{\varphi}_n})-m(\mathbf{x})\right|^2\mu(d\mathbf{x})>\ t\right\}\ \leq\ &|\mathcal{F}_{\varepsilon_n}|\Big(C_{25}\,e^{-C_{26}\ell^2\,p_n^2\,t^2}+C_{27}\,\ell m\,e^{-C_{28}\,mh^d\,p_n^2\,t^2}\\
&+\ C_{29}\,\ell m\,e^{-C_{30}mh^d}+C_{31}\,e^{-C_{32}\,\ell p_n^2\,t^2}\Big),
\end{aligned}
$$

for n large enough where C_{25}–C_{32} are positive constants not depending on m, ℓ, or t. This completes the proof of part (i) of the theorem.

Part (ii). The proof of part (ii) is virtually the same and, in fact, easier and therefore will not be given here.

□

Proof of Corollary 1

The corollary follows from the Borel-Cantelli lemma in conjunction with (19), the bound in Theorem 1, and Remark 1.

□

Proof of Theorem 2

We first note that, by Remark 1, it is sufficient to prove the theorem for the case of $p=2$. The proof is along standard arguments and goes as follows. Observe that

$$
\begin{aligned}
E\left|\widehat{m}(\mathbf{X};\breve{\pi}_{\breve{\varphi}_n})-m(\mathbf{X})\right|^2\ &=\ E\left[\int_{\mathbb{R}^d}\left|\widehat{m}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})-m(\mathbf{x})\right|^2\mu(d\mathbf{x}\right]\\
&=\ \int_0^\infty P\left\{\int_{\mathbb{R}^d}\left|\widehat{m}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})-m(\mathbf{x})\right|^2\mu(d\mathbf{x})>t\right\}dt\\
&=\ \int_0^{(2+B/\pi_0)^2L^2}P\left\{\int_{\mathbb{R}^d}\left|\widehat{m}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})-m(\mathbf{x})\right|^2\mu(d\mathbf{x})>t\right\}dt\,, \qquad (57)
\end{aligned}
$$

where the last line follows because in view of (18) one has

$$\big|\widehat{m}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{x})\big|^2 \le \big(\big|\widehat{m}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})\big| + \big|m(\mathbf{x})\big|\big)^2 \le \left(\Big(1+\frac{B}{\pi_0}\Big)L + L\right)^2 = (2+B/\pi_0)^2 L^2.$$

Therefore, by Theorem 1, for n large enough we have

$$\begin{aligned}
(57) \;\le\;& \int_0^u dt \;+\; c_{15}\big|\mathcal{F}_{\varepsilon_n}\big| \cdot \Bigg[\int_u^{(2+B/\pi_0)^2L^2} e^{-c_{10}\ell\, p_n^2\, t^2}\, dt \;+\; \ell m \int_u^{(2+B/\pi_0)^2L^2} e^{-c_{12}\, mh^d\, p_n^2\, t^2}\, dt \\
& + \ell m\, e^{-c_{14}\, mh^d} \int_u^{(2+B/\pi_0)^2L^2} dt \Bigg], \quad \text{where } c_{15} = c_9 \vee c_{11} \vee c_{13} \text{ and } c_9\text{–}c_{14} \text{ are as in (17)} \\
\le\;& u + 2c_{15}\big|\mathcal{F}_{\varepsilon_n}\big|\ell m \int_u^{(2+B/\pi_0)^2L^2} e^{-(c_{10}\wedge c_{12})(\ell\wedge mh^d)p_n^2 t^2} dt + c_{15}((2+B/\pi_0)L)^2\big|\mathcal{F}_{\varepsilon_n}\big|\ell m\, e^{-c_{14} mh^d} \\
\le\;& u + \frac{2\,c_{15}\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell m}{\sqrt{(c_{10}\wedge c_{12})(\ell\wedge mh^d)p_n^2}} \cdot \int_{u\sqrt{(c_{10}\wedge c_{12})(\ell\wedge mh^d)p_n^2}}^{\infty} e^{-v^2/2} dv \;+\; c_{16}\big|\mathcal{F}_{\varepsilon_n}\big|\ell m\, e^{-c_{14}\, mh^d} \\
& \qquad \text{(which follows from the change of variable } v = t\,\sqrt{(c_{10}\wedge c_{12})(\ell\wedge mh^d)p_n^2}\,\text{)} \\
\le\;& u + \frac{2\,c_{15}\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell m}{\sqrt{(c_{10}\wedge c_{12})(\ell\wedge mh^d)p_n^2}} \cdot \frac{e^{-(c_{10}\wedge c_{12})(\ell\wedge mh^d)p_n^2\, u^2/2}}{\sqrt{(c_{10}\wedge c_{12})(\ell\wedge mh^d)p_n^2}\cdot u} \;+\; c_{16}\big|\mathcal{F}_{\varepsilon_n}\big|\ell m\, e^{-c_{14}\, mh^d}, \qquad (58)
\end{aligned}$$

where the last line follows from the upper bound of the Mill's ratio; see, for example, Mitrinovic ([21]; p. 177). Now, put $c = 2\,c_{15}\big|\mathcal{F}_{\varepsilon_n}\big|\ell m$ and $N = (c_{10}\wedge c_{12})(\ell\wedge mh^d)p_n^2/4$, and observe that the right side of (58) can be written as

$$u + \frac{c}{4Nu}\, e^{-2Nu^2} + c_{16}\big|\mathcal{F}_{\varepsilon_n}\big|\ell m\, e^{-c_{14}\, mh^d}. \qquad (59)$$

But the term $u + \frac{c}{4Nu}\, e^{-2Nu^2}$ in (59) is approximately minimized by taking $u = \sqrt{\log(c)/(2N)}$, and the corresponding minimum of (59) becomes

$$\begin{aligned}
& \sqrt{\frac{\log(c)}{2N}} + \sqrt{\frac{1}{8N\log(c)}} + \; c_{16}\big|\mathcal{F}_{\varepsilon_n}\big|\ell m\, e^{-c_{12}\, mh^d} \\
&= \sqrt{\frac{c_{17} + \log\ell + \log m + \log|\mathcal{F}_{\varepsilon_n}|}{c_{18}\,(\ell\wedge mh^d)\,p_n^2}} + \sqrt{\frac{1}{c_{19}\,\big(c_{17} + \log\ell + \log m + \log|\mathcal{F}_{\varepsilon_n}|\big)\cdot(\ell\wedge mh^d)\,p_n^2}}, \\
&\quad + \; c_{16}\,\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell m\, e^{-c_{14}\, mh^d},
\end{aligned}$$

where c_{14}, c_{16}–c_{19} are positive constants not depending on m, ℓ, and n.

□

Proof of Theorem 3

Part (i). By (25), we have

$$P\left\{\widehat{g}_n(\mathbf{X};\widehat{\pi}) \neq Y \middle| \mathbb{D}_n\right\} - P\{g_{\mathrm{B}}(\mathbf{X}) \neq Y\} \;\le\; 2E\left[\big|\widehat{m}(\mathbf{X};\widehat{\pi}_{\widehat{\varphi}_n}) - m(\mathbf{X})\big| \,\middle|\, \mathbb{D}_n\right] \qquad (60)$$

$$P\left\{\breve{g}_n(\mathbf{X};\breve{\pi}) \neq Y \middle| \mathbb{D}_n\right\} - P\{g_{\mathrm{B}}(\mathbf{X}) \neq Y\} \;\leq\; 2E\left[\left|\widehat{m}(\mathbf{X};\breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\right| \middle| \mathbb{D}_n\right]. \tag{61}$$

Now part (i) follows from (60), (61), and Corollary 1 with $p=2$, in conjunction with the Cauchy-Schwarz inequality.

Part (ii). By a result of Audibert and Tsybakov [2] (Lemma 5.2), under assumption (G) we have

$$P\left\{\breve{g}_n(\mathbf{X};\breve{\pi}) \neq Y\right\} - P\{g_{\mathrm{B}}(\mathbf{X}) \neq Y\} \;\leq\; \left(E\left|\widehat{m}(\mathbf{X};\breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\right|^2\right)^{\frac{1+\alpha}{2+\alpha}}, \tag{62}$$

where α is as in (27). The result now follows from Corollary 2 with $p=2$.

□

FUNDING

This work was supported by NSF grant DMS-2310504 of Majid Mojirsheibani.

CONFLICTS OF INTEREST

The authors of this work declare that they have no conflicts of interest.

CREDIT AUTHOR STATEMENT

Both authors contributed equally to methodology, proofs, applications, numerical studies, editing, and write-up.

References

1. Azizyan, M., Singh, A., Wasserman, L., et al. (2013) Density-sensitive semisupervised inference. *Ann. Statist.* **41** 751–771.

2. Audibert, J. Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers under the margin condition.*Ann. Statist.* **35** 608–633.

3. Berrett, T., Györfi, L., and Walk, H. (2020). Strongly universally consistent nonparametric regression and classification with privatised data. *Electron. J. Stat.* **15**, 2430–2453.

4. Bindele, H. and Zhao, Y. (2018). Rank-based estimating equation with non-ignorable missing responses via empirical likelihood. *Statist. Sinica*, **28**, 1787–1820.

5. Bouzebda, S., Souddi, Y., and Madani, F. (2024). Weak convergence of the conditional set-indexed empirical process for missing at random functional ergodic data. *Mathematics.* 12, 1–23

6. Chen, X., Diao, G., and Qin, J. (2020). Pseudo likelihood-based estimation and testing of missingness mechanism function in nonignorable missing data problems. *Scand. J. Stat.* **47** 1377–1400.

7. Devroye, L., Györfi, L., and Lugosi, G. (1996) A probabilistic theory of pattern recognition. Springer-Verlag, New York.

8. Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of kernel regression estimate. *J. Statist. Plann. Inference,* **23** 71-82.

9. Döring, M., Györfi, L., and Walk, H. Exact rate of convergence of kernel-based classification rule. Challenges in computational statistics and data mining, 71–91, Stud. Comput. Intell., 605, Springer, Cham, 2016.

10. Ferraty, F., Sued, M., and Vieu, P. (2013). Mean estimation with data missing at random for functional covariables, *Statistics*, *47*, 688–706.

11. Guo, X., Song, Y., and Zhu, L. (2019). Model checking for general linear regression with nonignorable missing response. *Comput. Statist. Data Anal*, **138** 1–12.

12. Horvitz D. G. and Thompson D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47** 663–685

13. Kim, J.K. and Yu, C.L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *J. Am. Statist. Assoc.* **106** 157–65.

14. Kohler, M. and Krzyżak, A. (2007). On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Trans. Inform. Theory* **53** 1735–1742.

15. Li, T., Xie, F., Feng, X., Ibrahim, J, and Zhu, H. (2018). Functional Linear Regression Models for Nonignorable Missing Scalar Responses. *Statist. Sinica*, **28** 1867–1886.

16. Ling, N., Liang, L., and Vieu, P. (2015). Nonparametric regression estimation for functional stationary ergodic data with missing at random, *J. Stat. Plan. Inference*, *162*, 75–87.

17. Maity, A., Pradhan, V., and Das, U. (2019). Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. *Amer. Statist.* **73** 340–349.

18. Mammen, E. and Tsybakov, A.B. (1999) Smooth discriminant analysis. *Ann. Statist.* **27** 1808–1829.

19. Massart, P. and E. Nédélec, E. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326–2366.

20. Miao, W., Li, X., and Sun, B. (2023). A stableness of resistance model for nonresponse adjustment with callback data. https://doi.org/10.48550/arXiv.2112.02822

21. Mitrinovic, D. S. Analytic Inequalities. New York. Springer-Verlag, 1970.

22. Mojirsheibani, M. (2022). On the maximal deviation of kernel regression estimators with MNAR response variables. *Statist. Papers*, **63** 1677–1705.

23. Mojirsheibani, M. (2007). Nonparametric curve estimation with missing covariates: A general empirical process approach. *J. Stat. Plan. Inference*, **137** 2733–2758

24. Morikawa, K. and Kim, J.K. (2021). Semiparametric optimal estimation with nonignorable nonresponse data. *Ann. Statist.*, **49** 2991–3014.

25. Morikawa, K., Kim, J.K., and Kano, Y. (2017). Semiparametric maximum likelihood estimation with data missing not at random. *J. Am. Statist. Assoc.* **106** 157–65.

26. Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

27. Niu, C., Guo, X., Xu, W., and Zhu, L. (2014). Empirical likelihood inference in linear regression with nonignorable missing response. *Comput. Statist. Data Anal*, **79** 91–112.

28. Racine, J. and Hayfield, T. (2008). Nonparametric Econometrics: The np Package. J. Statist. Software, **27**, 1–32

29. Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *J. Econometrics,* **119** 99–130.

30. Sadinle, M. and Reiter, J. (2019). Sequentially additive nonignorable missing data modelling using auxiliary marginal information. *Biometrika*. **106** 889–911.

31. Shao, J. and Wang, L. (2016) Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika.* **103** 175–187.

32. Tang, G., Little, R.J.A., and Raghunathan, T.E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*. **90** 747–764.

33. Tsybakov, A.B. and van de Geer, S. (2005). Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Statist.* **33** 1203–1224.

34. Uehara, M., Lee, D., and Kim, J.K. (2023). Statistical inference with semiparametric nonignorable nonresponse models. *Scan J. Stat.* **50** 1795–1817.

35. van der Vaart, A., Wellner, J. (1996) Weak Convergence and Empirical Processes with Applications to Statistics. Springer, New York.

36. Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Ser. A.* **26** 359–372.

37. Wang, J. and Shen, X. (2007) Large margin semi-supervised learning. *J. Mach. Learn. Res.*, **8** 1867–1891.

38. Zhao, J. and Ma, Y. (2022). A versatile estimation procedure without estimating the nonignorable missingness mechanism. *J. Am. Statist. Assoc.* **117** 1916–1930.

39. Zhao, P., Wang, L., and Shao, J. (2019). Empirical likelihood and Wilks phenomenon for data with nonignorable missing values. *Scand. J. Stat.* **46** 1003–1024.

40. Fischer, A. and Mougeot, M. (2019). Aggregation using input–output trade-off. *J. Stat. Plan. Inference*, *200*, 1–19