On regression and classification with possibly missing response variables in the data

Majid Mojirsheibani[a] [1], William Pouliot[b], and Andre Shakhbandaryan[c]

[a,c] *Department of Mathematics, California State University, Northridge, CA, USA*

[b] *Department of Economics, University of Birmingham, Birmingham, UK*

Abstract

This paper considers the problem of kernel regression and classification with possibly unobservable response variables in the data, where the mechanism that causes the absence of information can depend on both predictors and the response variables. Our proposed approach involves two steps: First we construct a family of models (possibly infinite dimensional) indexed by the unknown parameter of the missing probability mechanism. In the second step, a search is carried out to find the empirically optimal member of an appropriate cover (or subclass) of the underlying family in the sense of minimizing the mean squared prediction error. The main focus of the paper is to look into some of the theoretical properties of these estimators. The issue of identifiability is also addressed. Our methods use a data-splitting approach which is quite easy to implement. We also derive exponential bounds on the performance of the resulting estimators in terms of their deviations from the true regression curve in general L_p norms, where we allow the size of the cover or subclass to diverge as the sample size n increases. These bounds immediately yield various strong convergence results for the proposed estimators. As an application of our findings, we consider the problem of statistical classification based on the proposed regression estimators and also look into their rates of convergence under different settings. Although this work is mainly stated for kernel-type estimators, it can also be extended to other popular local-averaging methods such as nearest-neighbor and histogram estimators.

MSC2020 subject classifications: Primary 62G05; secondary 62G08

Keywords and phrases: Regression, partially observed data, kernel, convergence, classification, margin condition.

1 Introduction

During the past decade, there has been a steady growing interest in developing appropriate procedures to perform estimation and inference in the presence of incomplete data under the complex regime where the data is not missing at random (NMAR). The NMAR setup is generally acknowledged to be a difficult problem in incomplete data literature due to identifiability issues; this is

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

This work was supported by the National Science Foundation (NSF) under Grant DMS-1916161 of Majid Mojirsheibani

significantly different from the simpler missing at random model where the absence of Y depends on $\mathbf{X}$ only (and not Y itself).

The focus of this paper is on the theoretical performance of kernel regression and classification under the realistic assumption that many response values in the data may be unavailable or missing. Unobservable or incomplete data occur frequently in medical data, survey data, public opinion polls, as well as the data collected in many areas of scientific activities. 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 fully observable, the classical Nadaraya-Watson kernel estimator of $m(\mathbf{x})$ (Nadaraya (1964), Watson (1964)) 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 $h \equiv h_n > 0$. 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

$$I_n(p) = \int |\widehat{m}_n(\mathbf{x}) - m(\mathbf{x})|^p \, \mu(d\mathbf{x}), \quad 1 \leq p < \infty\,,$$

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 (1996; Sec. 6.2) and Devroye and Krzyżak (1989). In fact, in the cited paper, Devroye and Krzyżak 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* (see Definition 1) 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.

Now, suppose that the response variable Y is allowed to be missing according to the NMAR mechanism. Define the indicator random variable $\Delta = 0$ if Y is missing, and $\Delta = 1$ otherwise. Similarly, for $i = 1, \cdots, n$, let $\Delta_i = 0$ if 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 $m_n^{\text{cc}}(\mathbf{x}) := \sum_{i=1}^n \Delta_i Y_i \, \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h) \big/ \sum_{i=1}^n \Delta_i \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h)$. Unfortunately, $m_n^{\text{cc}}(\mathbf{x})$ turns out to be the estimator of the quantity $E(\Delta Y|\mathbf{X} = \mathbf{x})/E(\Delta|\mathbf{X} = \mathbf{x})$ which, in general, is not equal to the regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ under a NMAR response mechanism.

For the important case of predictive models (such as regression), Kim and Yu (2011) considered a highly versatile logistic type missing probability mechanism that works as follows. Let $\pi(\mathbf{x}, y) := E\left[\Delta \,\middle|\, \mathbf{X} = \mathbf{x}, Y = y\right]$ be the *selection probability*, also called the *nonresponse propensity*. Then,

Kim and Yu (2011) considered the flexible model

$$\pi_\gamma(\mathbf{x}, y) \,:=\, E_\gamma\big[\Delta \,\big|\mathbf{X} = \mathbf{x}, Y = y\big] \,=\, P_\gamma\big\{\Delta = 1 \,\big|\mathbf{X} = \mathbf{x}, Y = y\big\} \,=\, \frac{1}{1 + \exp\big\{g(\mathbf{x}) + \gamma y\big\}}\,, \qquad (2)$$

indexed by the real parameter γ, where g is a completely unknown function of the predictor $\mathbf{x}$ which also depends on γ. The true value of the unknown parameter γ will be denoted by γ^*.

It is well-understood in the framework of NMAR missing data that imposing parametric models on both $\pi_\gamma(\mathbf{x}, y)$ and the distribution of $(\mathbf{X}, Y)$ is too strong to be useful in practice (Molenberghs and Kenward (2007)). In fact, fully parametric models are sensitive to failure of the model assumptions (Little (1985)). On the other hand, in a fully nonparametric setup where $\pi_\gamma(\mathbf{x}, y)$ and the distribution of $(\mathbf{X}, Y)$ are unknown, one faces the issue of non-identifiability when estimating the function π_γ (Shao and Wang (2016)). Some authors have assumed a fully parametric model for $\pi_\gamma(\mathbf{x}, y)$ only, but not the underlying distributions (Qin et al. (2002) and Wang et al (2014)), but this is also considered to be too strong in practice. To deal with these issues, Kim and Yu (2011) considered the semi-parametric model (2) as a reasonable compromised solution.

The missing probability mechanism (2) has been used and studied extensively in the literature; see, for example, Zhao and Shao (2015), Shao and Wang (2016), Morikawa et al (2017), Uehara and Kim (2018), Morikawa and Kim (2018), Morikawa and Kano (2018), Fang et al (2018), O'Brien et al (2018), Maity et al (2019), Sadinle and Reiter (2019), Zhao et al (2019), Yuan et al (2020), Chen et al (2020), Mojirsheibani (2021), and Liu and Yau (2021). In fact, in view of the recent widespread use of model (2) in the literature, there appears to be the tacit consensus that (2) is versatile enough to be used in predictive models such as regression and classification, and this will also be the direction of the current paper. We observe that if $\gamma = 0$, then (2) reduces to the simpler case of missing at random assumption (MAR).

The problem of regression function estimation with NMAR missing data is generally considered to be challenging. In fact, to the best of our knowledge, there are only a few results available in the literature in this direction that also address the theoretical validity of their proposed methods. These include the results of Niu et al (2014) and Guo et al (2019) for the case of linear regression, those of Bindele et al (2018) to estimate β in the model $E(Y|\mathbf{X} = \mathbf{x}) = g(\mathbf{x}, \beta)$, where g is completely known, and the results of Li et al (2018) for parameter estimation in functional linear regression. In the case of nonparametric regression, Mojirsheibani (2022) studied the asymptotic distribution of the maximal deviation of kernel regression estimators. However, all of these results assume the availability of either an independent validation sample or an independent follow-up survey for estimating the parameters in (2). Furthermore, the current work does not assume linearity of the underlying regression model.

Our contributions in this paper are three-fold. (i) We develop two types of easy-to-implement

estimators of the regression curve $m(\mathbf{x})$ in the presence of NMAR missing data. Additionally, we consider a more general version of model (2) where the quantity $\exp\{\gamma y\}$ will be replaced by a more general positive function $\varphi(y)$. We also propose estimators of $\varphi(y)$. The new estimators, which are based on the approximation theory of totally bounded classes of functions, are constructed using a data-splitting approach. (ii) We will carefully explore and study the global properties of the proposed regression estimators in general L_p norms; these results parallel those of Devroye and Krzyżak (1989) for the simpler case of no missing data. More specifically, we provide exponential performance bounds on the L_p norms of the proposed regression estimators that are valid under rather standard assumption. Such bounds in conjunction with the Borel-Cantelli lemma immediately yield various strong convergence and optimality results. Exploiting these bounds further, we also look into the rates of convergence of the proposed estimators (in L_p). (iii) A study of the applications of our proposed estimators to the problem of nonparametric classification in the presence of partially observed data is also considered.

As an important application of our results to the field of machine learning and statistical classification, we note that in the so-called semi-supervised learning one usually has to deal with large amounts of missing responses (or missing labels) in the data. In such setups, researchers in machine learning have made efforts to develop procedures for utilizing the unlabeled cases (i.e., the data points with missing Y_i's) in order to construct more effective classification rules; see, for example, Wang and Shen (2007). But most such results assume that the response variable is missing completely at random; see, for example, Azizyan et al (2013). Our results in Section 3 make it possible to develop classification rules in the presence of NMAR response variables for the semi-supervised setup, where we also study the rates of convergence of such classifiers.

The rest of the paper is organized as follows. Section 2 presents the main results, where in Subsection 2.1 the estimation of the true γ^* can be based on any available method. Subsection 2.1 also proposes a generalization of the model (2), as given by (8), where new estimation methods based on the theory of totally bounded classes of functions are employed. Subsection 2.2 uses a Horvitz-Thompson type inverse weighting approach to estimate the underlying regression function. Section 3 focuses on the applications of our estimators to the problem of nonparametric classification with partially observed data. All proofs are deferred to Section 4.

Throughout this paper, we denote by $C, C', C_0, C_1, \cdots, c, c', c_0, c_1, \cdots$ some real constants that are strictly positive; also, for reals a and b, we use the notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For the ease of notation, when the events or random variables of interest involve the random variable Δ or Δ_i's, the notations $P_\gamma\{\cdot\}$ and $E_\gamma[\,\cdot\,]$ (resp. $P_\varphi\{\cdot\}$ and $E_\varphi[\,\cdot\,]$) will be used only when γ (resp. φ) is different from the true parameter value γ^* (resp. φ^*). Furthermore, for real sequences a_n and $b_n > 0$, the notation $a_n = \mathcal{O}(b_n)$, as $n \to \infty$, means that $|a_n|/b_n$ is a bounded sequence in the sense

that there are positive constants M and n_o such that $|a_n| \leq M \cdot b_n$ for all $n \geq n_o$.

2 Main results

2.1 The first estimator and a more general missing mechanism

Consider the missing probability mechanism (2) and let $\mathbb{D}_n = \{(\mathbf{X}_1, Y_1, \Delta_1), \ldots, (\mathbf{X}_n, Y_n, \Delta_n)\}$ be independent and identically distributed (iid) observations, i.e., the data. Then, clearly the estimator $\widehat{m}_n$ in (1) is no longer available due to the presence of missing Y_i's. Furthermore, as discussed in the introduction, the complete-case estimator that only uses the fully observable data is not necessarily the correct estimator under model (2) anymore. In the following two sections, we propose some alternative estimators instead. To justify our first estimator, we start by constructing an initial naive plug-in type estimator which works as follows. Define the quantity

$$\eta_k(\mathbf{x}, t) = E\left[\Delta Y^{2-k} \exp\{t\, Y\} \middle| \mathbf{X} = \mathbf{x}\right], \quad k = 1, 2, \quad t \in \mathbb{R}, \tag{3}$$

and observe that when $P\{\Delta = 1\} \neq 1$, i.e., when Y is allowed to be missing, one can use Lemma 1 (upon replacing $\varphi^*(y)$ by $\exp\{\gamma^* y\}$ in this lemma) to express the regression curve $m(\mathbf{x})$ as

$$m(\mathbf{x}) \equiv m_{\gamma^*}(\mathbf{x}) = \eta_1(\mathbf{x}, 0) + \frac{\eta_1(\mathbf{x}, \gamma^*)}{\eta_2(\mathbf{x}, \gamma^*)}\big(1 - \eta_2(\mathbf{x}, 0)\big). \tag{4}$$

Now, let $\widehat{\gamma}$ be any estimator of γ^* and consider the following simple kernel-type estimator of (4)

$$\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x}) = \widehat{\eta}_1(\mathbf{x}, 0) + \frac{\widehat{\eta}_1(\mathbf{x}, \widehat{\gamma})}{\widehat{\eta}_2(\mathbf{x}, \widehat{\gamma})}\big(1 - \widehat{\eta}_2(\mathbf{x}, 0)\big), \tag{5}$$

where

$$\widehat{\eta}_k(\mathbf{x}, t) = \frac{\sum_{i=1}^n \Delta_i Y_i^{2-k} \exp\{t\, Y_i\}\, \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h)}{\sum_{i=1}^n \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h)}, \quad k = 1, 2, \quad t \in \mathbb{R}, \tag{6}$$

and, as in (1), $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}_+$ is the kernel used with bandwidth h. In passing, we also point out that although we are considering a kernel type estimator in (5), virtually all our results in this paper continue to hold for other popular local-averaging estimators such as nearest-neighbor estimators, cubic histograms, as well as general partitioning estimators. However, to avoid making this work unnecessarily long and tedious, the paper is confined to kernel estimators only.

How good of an estimator is $\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x})$ in (5)? To answer this question, we start by assuming that the kernel $\mathcal{K}$ is *regular*:

Definition 1 *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.*

For more on this, see Devroye and Krzyżak (1989). We also require the following condition regarding the selection probability $\pi_\gamma(\mathbf{x}, y) := P_\gamma\{\Delta = 1|\mathbf{X} = \mathbf{x}, Y = y\}$, which is quite standard in missing data literature:

Assumption (A). The selection probability, $\pi_\gamma(\mathbf{x}, y)$, satisfies $\inf_{\mathbf{x},y} \pi_\gamma(\mathbf{x}, y) =: \pi_{\min} > 0$, for some $\pi_{\min}$.

Assumption (A) essentially states that the response Y can always be observed with a non-zero probability for any values of $\mathbf{x}$ and y. The following basic result gives upper bounds on the performance of the L_p norms of the estimator $\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x})$ under standard assumptions.

Theorem 1 *Let $\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x})$ be the estimator of $m(\mathbf{x})$ defined in (5), where $\widehat{\gamma}$ may be any estimator of γ^* in (4), and suppose that assumption (A) holds. Suppose that the kernel $\mathcal{K}$ in (6) is regular and that its bandwidth satisfies $h \to 0$ and $nh^d \to \infty$, as $n \to \infty$. Then, for every $\epsilon > 0$, every $1 \leq p < \infty$, any distribution of $(\mathbf{X}, Y) \in \mathbb{R}^d \times [-L, L]$, $L < \infty$, and n large enough,*

$$P\left\{\int \left|\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x}) - m(\mathbf{x})\right|^p \mu(d\mathbf{x}) > \epsilon\right\} \;\leq\; c_1\, e^{-c_2 n} + \, c_3\, P\big\{\, |\widehat{\gamma} - \gamma^*| > C_0\big\}, \tag{7}$$

where μ is the probability measure of $\mathbf{X}$ and c_1, c_2, c_3, and C_0 are positive constants not depending on n; here, c_2 also depends on ϵ.

In passing, we note that the bound in Theorem 1 is in the spirit of the classical result of Devroye and Krzyżak (1989) for kernel regression estimators with no missing data (modulo the term $P\{|\widehat{\gamma}-\gamma^*| > C_0\}$ on the right side of (7)).

Remark 1 *The bound in Theorem 1 shows that the consistency of $\widehat{\gamma}$, as an estimator of γ^*, is needed in order for the proposed regression estimator to converge in the L_p norm. Unfortunately, due to parameter identifiability issues, consistent estimation of γ^* can be a serious challenge unless one either has access to additional external data, as in Kim and Yu (2011), or one can correctly assume that the function $g(\mathbf{x})$ in (2) is independent/free of certain components of $\mathbf{x} = (x_1, \cdots, x_d)^T$; see, for example, Shao and Wang (2016) or Uehara and Kim (2018). Here, we consider a different estimation procedure based on the approximation theory of totally bound class of functions.*

In what follows, we consider a more general version of the missing probability model (2) given by

$$\pi_\varphi(\mathbf{x}, y) \;:=\; E_\varphi\big[\Delta \,\big|\mathbf{X} = \mathbf{x}, Y = y\big] \;=\; P_\varphi\big\{\Delta = 1 \,\big|\mathbf{X} = \mathbf{x}, Y = y\big\} \;=\; \frac{1}{1 + \exp\big\{g(\mathbf{x})\big\} \cdot \varphi(y)}\,, \tag{8}$$

where the model is indexed by the functional parameter $\varphi > 0$; the true φ will be denoted by φ^*. Clearly the function $\exp\{\gamma y\}$ in (2) is a special case of $\varphi(y)$. Our approach to estimate the function φ^* here is based on the approximation theory of totally bounded function spaces. More

specifically, consider the situation where φ^* belongs to a totally bounded class of functions in the following sense: Let $\mathcal{F}$ be a given class of function $\varphi : [-L, L] \to (0, B]$, for some $B < \infty$. Fix $\varepsilon > 0$ and suppose that the finite collection of functions $\mathcal{F}_\varepsilon = \{\varphi_1, \ldots, \varphi_{N(\varepsilon)}\}$, $\varphi_i : [-L, L] \to (0, B]$, is an ε-cover of $\mathcal{F}$, i.e., for each $\varphi \in \mathcal{F}$, there is a $\bar{\varphi} \in \mathcal{F}_\varepsilon$ such that $\|\varphi - \bar{\varphi}\|_\infty < \varepsilon$; 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 monograph by van der Vaart and Wellner (1996; p. 83) provides more details on such concepts.

To present our methods, we employ a data splitting approach that works as follows. Let $\mathbb{D}_n = \{(\boldsymbol{X}_1, Y_1, \Delta_1), \ldots, ((\boldsymbol{X}_n, Y_n, \Delta_n)\}$ be the data (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 in our main results. Also, define the index sets

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

Next, for each fixed $\varphi \in \mathcal{F}$, consider the kernel-type estimator of $m(\mathbf{x})$ constructed based on the training set $\mathbb{D}_m$ alone, given by

$$\widehat{m}_m(\mathbf{x}; \varphi) \;=\; \widehat{\eta}_{m,1}(\mathbf{x}) + \frac{\widehat{\psi}_{m,1}(\mathbf{x}; \varphi)}{\widehat{\psi}_{m,2}(\mathbf{x}; \varphi)} \big(1 - \widehat{\eta}_{m,2}(\mathbf{x})\big), \tag{9}$$

where $\widehat{\psi}_{m,k}(\mathbf{x}; \varphi)$ and $\widehat{\eta}_{m,k}(\mathbf{x})$, $k = 1, 2$, are the quantities

$$\widehat{\psi}_{m,k}(\mathbf{x}; \varphi) \;=\; \frac{\sum_{i \in \boldsymbol{\mathcal{I}}_m} \Delta_i Y_i^{2-k} \varphi(Y_i)\, \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h)}{\sum_{i \in \boldsymbol{\mathcal{I}}_m} \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h)}, \quad k = 1, 2, \quad \varphi \in \mathcal{F}, \tag{10}$$

$$\widehat{\eta}_{m,k}(\mathbf{x}) \;=\; \frac{\sum_{i \in \boldsymbol{\mathcal{I}}_m} \Delta_i Y_i^{2-k} \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h)}{\sum_{i \in \boldsymbol{\mathcal{I}}_m} \mathcal{K}((\mathbf{x} - \mathbf{X}_i)/h)}, \quad k = 1, 2. \tag{11}$$

Of course, (9) is not quite an estimator because φ itself must also be estimated. To this end, we first observe that in view of the results of Kim and Yu (2011), the term $\exp\{g(\mathbf{x})\}$ that appears in (8) can also be expressed as

$$\exp\{g(\mathbf{x})\} = E_\varphi\big[1 - \Delta \big| \mathbf{X} = \mathbf{x}\big] \big/ E_\varphi\big[\Delta\, \varphi(Y) \big| \mathbf{X} = \mathbf{x}\big]. \tag{12}$$

To appreciate (12), first observe that by (8), $\frac{1}{E_\varphi(\Delta|\mathbf{X},Y)} - 1 = \frac{1}{\pi_\varphi(\mathbf{X},Y)} - 1 = \exp\{g(\mathbf{X})\} \cdot \varphi(Y)$. Therefore, $E_\varphi\big[\Delta\big(\frac{1}{E_\varphi(\Delta|\mathbf{X},Y)} - 1\big)|\mathbf{X}\big] = E_\varphi[1 - \Delta|\mathbf{X}] = \exp\{g(\mathbf{X})\} \cdot E_\varphi[\Delta\varphi(Y)|\mathbf{X}]$, from which (12) follows. Estimating the right side of (12) can be challenging due to identifiability issues, and a sufficient condition for model identification is (see, for example, Uehara and Kim (2018)) 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 $\mathbf{X} = (\mathbf{Z}, \mathbf{V})$; see assumption (G) on the next page. Under this assumption, the selection probability model in (8) becomes

$$\pi_\varphi(\mathbf{z}, y) \;:=\; E_\varphi\big[\Delta \,\big|\mathbf{Z} = \mathbf{z}, Y = y\big] \;=\; P_\varphi\big\{\Delta = 1 \,\big|\mathbf{Z} = \mathbf{z}, Y = y\big\} \;=\; \frac{1}{1 + \exp\{g(\mathbf{z})\} \cdot \varphi(y)}\,. \tag{13}$$

It is not hard to see that under (13) the expression in (12) becomes

$$\exp\{g(\mathbf{z})\} = E_\varphi\big[1 - \Delta\big|\mathbf{Z} = \mathbf{z}\big] / E_\varphi\big[\Delta\,\varphi(Y)\big|\mathbf{Z} = \mathbf{z}\big]. \tag{14}$$

Next, we propose the following two-step procedure to estimate the function φ in (9):

Step 1. For each given φ, the selection probability in (13) is estimated, based on $\mathbb{D}_m$ alone, by

$$\widehat{\pi}_\varphi(\mathbf{z}, y) = \Big[1 \,+\, \widehat{\exp\{g(\mathbf{z})\}} \cdot \varphi(y)\Big]^{-1}, \tag{15}$$

where $\widehat{\exp\{g(\mathbf{z})\}}$ is the kernel regression estimator of (14) based on $\mathbb{D}_m$, i.e.,

$$\widehat{\exp\{g(\mathbf{z})\}} = \frac{\sum_{i \in \boldsymbol{\mathcal{I}_m}} (1 - \Delta_i) \mathcal{K}_0((\mathbf{z} - \mathbf{Z}_i)/h)}{\sum_{i \in \boldsymbol{\mathcal{I}}_m} \Delta_i \varphi(Y_i)\, \mathcal{K}_0((\mathbf{z} - \mathbf{Z}_i)/h)}\,, \tag{16}$$

and $\mathcal{K}_0$ is the kernel used with bandwidth h.

Step 2. Let $\varepsilon_n > 0$ be a decreasing sequence $\varepsilon_n \downarrow 0$, 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 proposed estimator of φ in (9) is then defined by

$$\widehat{\varphi}_n \;:=\; \underset{\varphi \in \mathcal{F}_{\varepsilon n}}{\operatorname{argmin}}\, \widehat{L}_{m,\ell}(\varphi) \quad \text{where} \quad \widehat{L}_{m,\ell}(\varphi) \;=\; \ell^{-1} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i}{\widehat{\pi}_\varphi(\mathbf{Z}_i, Y_i)} \big|\widehat{m}_m(\mathbf{X}_i; \varphi) - Y_i\big|^2, \tag{17}$$

where $\widehat{m}_m(\mathbf{x}; \varphi)$ is as in (9). The subscript n at $\widehat{\varphi}_n$ reflects the fact that the entire data of size n has been used here. Finally, our estimator of the regression function $m(\mathbf{x})$ is given by

$$\widehat{m}(\mathbf{x}; \widehat{\varphi}_n) \;:=\; \widehat{m}_m(\mathbf{x}; \varphi)\big|_{\varphi = \widehat{\varphi}_n}, \quad \text{with } \widehat{m}_m(\mathbf{x}; \varphi) \text{ as in (9).} \tag{18}$$

The estimator in (17) may be viewed as the empirical version of the minimizer of the mean squared error, i.e., the empirical version of

$$\varphi_{\varepsilon_n} \;:=\; \underset{\varphi \in \mathcal{F}_{\varepsilon_n}}{\operatorname{argmin}}\, E\big|m(\mathbf{X}; \varphi) - Y\big|^2, \tag{19}$$

where $m(\mathbf{X}; \varphi)$ is the regression function $m(\mathbf{X}; \varphi^*)$ evaluated at an arbitrary $\varphi \in \mathcal{F}_{\varepsilon_n}$ (see Lemma 1). We also note that φ_{ε_n} in (19) is an approximation to the true function φ^* based on the cover $\mathcal{F}_{\varepsilon_n}$ of $\mathcal{F}$. In fact, we have

$$\varphi^* \;:=\; \underset{\varphi:\, [-L,L] \to \mathbb{R}_+}{\operatorname{argmin}}\, E\big|m(\mathbf{X}; \varphi) - Y\big|^2. \tag{20}$$

How good is $\widehat{m}(\mathbf{x};\widehat{\varphi}_n)$ in (18) as an estimator of the true regression curve $m(\mathbf{x})$? To answer this, we first state a number of assumptions.

Assumption (A′). For all $\varphi \in \mathcal{F}$, the selection probability $\pi_\varphi(\mathbf{z}, y)$ in (13) satisfies $\inf_{\mathbf{z},y}\, \pi_\varphi(\mathbf{z}, y) =: \pi_{\min} > 0$ for some $\pi_{\min}$, where $\mathcal{F}$ is a totally bounded class of functions $\varphi : [-L, L] \to (0, B]$, for some $B < \infty$ and $L < \infty$.

Assumption (B). The kernel $\mathcal{K}$ 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 $nh^d \to \infty$, as $n \to \infty$.

Assumption (C). The density function $f(\mathbf{z})$ of $\mathbf{Z}$ 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). $E_\varphi[\Delta\, \varphi(Y)|\mathbf{X} = \mathbf{x}] \geq \varrho_0$, for μ–a.e. $\mathbf{x}$ and each $\varphi \in \mathcal{F}$, for some constant $\varrho_0 > 0$.

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

Assumption (F). The deviation $A_{m,\ell}(\varphi) = |\widehat{L}_{m,\ell}(\varphi) - E[|\widehat{m}_m(\mathbf{X};\varphi) - Y|^2|\mathbb{D}_m]|$, where $\widehat{L}_{m,\ell}(\varphi)$ and $\widehat{m}_m(\mathbf{x};\varphi)$ are as in (17) and (9) satisfies $P\{A_{m,\ell}(\varphi) > t\} \leq \sup_{\varphi\in\mathcal{F}} P_\varphi\{A_{m,\ell}(\varphi) > t\}$, $\forall\, t > 0$.

Assumption (G). *[Identifiability]* There is a part of $\mathbf{X}$, say $\mathbf{V}$, which is conditionally independent of Δ, given Y and $\mathbf{Z}$, where $\mathbf{X} = (\mathbf{Z}, \mathbf{V})$.

Assumption (B) is not restrictive at all because the choice of the kernel $\mathcal{K}$ 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 because $E_\varphi[\Delta\, \varphi(Y)|\mathbf{X}] = E_\varphi[\varphi(Y)E_\varphi(\Delta|\mathbf{X}, Y)|\mathbf{X}] \geq \pi_{\min} E[\varphi(Y)|\mathbf{X}]$ and the fact that $\varphi(y) > 0$ for all y. Assumption (E) has already been used in the literature, whereas assumption (F) is technical. Assumption (G) is a sufficient condition for model identifiability; see, for example, Uehara and Kim (2018).

The following result gives exponential upper bounds on the performance of the L_2 norms of the estimator defined via (18) and (17). This result readily extends to more general L_p norms ($p \geq 1$); see Remark 2 below.

Theorem 2 *Let $\widehat{m}(\mathbf{x};\widehat{\varphi}_n)$ be as in (18) and suppose that assumptions (A′), (B)–(G) hold. Also let the missing probability mechanism π_φ be as in (13). Then for every $\varepsilon_n > 0$ satisfying $\varepsilon_n \downarrow 0$, as $n \to \infty$, every $t > 0$, any distribution of $(\mathbf{X}, Y) \in \mathbb{R}^d \times [-L, L]$, $L < \infty$, and n large enough,*

$$P\left\{\int \left|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\right|^2 \mu(d\mathbf{x}) > t\right\} \leq\ c_4|\mathcal{F}_{\varepsilon_n}|e^{-c_5\ell t^2} + c_6\ell\,|\mathcal{F}_{\varepsilon_n}|\left(e^{-c_7 m h^d} + e^{-c_8 m h^d\, t^2}\right), \quad (21)$$

whenever $\varphi^* \in \mathcal{F}$, where $|\mathcal{F}_{\varepsilon_n}|$ is the cardinality of the set $\mathcal{F}_{\varepsilon_n}$ and c_4–c_8 are positive constants not depending on m, ℓ, n, or t.

Remark 2 *Although the above theorem is stated in the L_2 sense, the theorem continues to hold for all $p \geq 2$. To appreciate this, observe that in the case of $p > 2$ one can always write*

$$\big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\big|^p \leq \Big(\big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n)\big| + \big|m(\mathbf{x})\big|\Big)^{p-2}\big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\big|^2 \leq (3L)^{p-2}\big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\big|^2.$$

On the other hand, if $p \in [1,2)$ then by Hölder's inequality we have

$$P\left\{\int \Big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\Big|^p \mu(d\mathbf{x}) > t\right\} \;\leq\; P\left\{\int \Big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\Big|^2 \mu(d\mathbf{x}) > t^{2/p}\right\}.$$

In passing, we note that the bound in (21) may be viewed as a generalization of the classical results of Devroye and Krzyżak (1989) for kernel regression estimators with fully observable data, where they also assumed $|Y| \leq L < \infty$. This assumption readily allows one to establish exponential performance bounds for general L_p norms, $p \geq 1$, of kernel regression estimators (and not just for $p = 1$ or 2). It is also justified by the fact that our main application is to the problem of classification where Y is bounded. A more desirable result would be obtained if the boundedness of Y could be relaxed to the moment condition $E|Y|^c < \infty$, for some $c \geq 1$. This has been achieved in the case of fully observable data; see, for example Krzyżak (1992) and Györfi et al (1998). However, so far we have not been able to extend our results in this direction. In fact, to the best of our knowledge, such extensions are not available even for the simpler case of data missing at random (MAR) setups, where the probability that Y is missing depends on $\mathbf{X}$, but not Y itself.

The following simple corollary shows that the above theorem can be used to establish strong convergence results.

Corollary 1 *Let $\widehat{m}(\mathbf{X};\widehat{\varphi}_n)$ be the estimator in (18). If, as $n \to \infty$,*

$$\varepsilon_n \downarrow 0, \quad \frac{\log \ell}{mh^d} \to 0, \quad \frac{\log|\mathcal{F}_{\varepsilon_n}|}{mh^d} \to 0, \text{ and } \quad \frac{\log|\mathcal{F}_{\varepsilon_n}|}{\ell} \to 0, \tag{22}$$

then under the conditions of Theorem 2 we have

$$E\Big[\big|\widehat{m}(\mathbf{X};\widehat{\varphi}_n) - m(\mathbf{X})\big|^p \Big| \mathbb{D}_n\Big] \longrightarrow^{a.s.} 0, \quad \text{for all } p \in [2,\infty).$$

Clearly, by Lebesgue dominated convergence theorem, under the conditions of Corollary 1 and without further ado,

$$E\big|\widehat{m}(\mathbf{X};\widehat{\varphi}_n) - m(\mathbf{X})\big|^p \longrightarrow 0, \quad \text{for all } p \in [2,\infty).$$

Unfortunately, this result does not provide a rate of convergence. The following theorem sheds more light on the convergence properties of the estimator in (18).

Theorem 3 *Consider the estimator $\widehat{m}(\mathbf{X};\widehat{\varphi}_n)$ in (18). Then, under the conditions of Theorem 2, for n large enough,*

$$E\Big|\widehat{m}(\mathbf{X};\widehat{\varphi}_n) - m(\mathbf{X})\Big|^p$$
$$\leq \sqrt{\frac{c_9 + \log \ell + \log|\mathcal{F}_{\varepsilon_n}|}{c_{10}\cdot(\ell \wedge mh^d)}} + \sqrt{\frac{1}{c_{11}\cdot(\ell \wedge mh^d)\big[c_9 + \log \ell + \log|\mathcal{F}_{\varepsilon_n}|\big]}} + c_{12}\,\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell\, e^{-c_{13}\,mh^d},$$

for all $p \in [2,\infty)$, where c_9–c_{13} are positive constants not depending on m, ℓ, or n.

The following result, which is an immediate corollary to Theorem 3, looks into the rate of convergence of the proposed regression estimator.

Corollary 2 *Let $\widehat{m}(\mathbf{X};\widehat{\varphi}_n)$ be the estimator in (18) and suppose that (22) holds. Then, under the conditions of Theorem 2, for all $p \geq 2$,*

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

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

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

An Example.
To compare and contrast the asymptotic performance of our estimation approach with the existing methods, consider the class $\mathcal{F}$ of functions φ of the form:

$$\varphi(y) = \exp\{\gamma y\}\,, \quad |\gamma| \leq M, \;\; |y| \leq L, \;\; \text{for some } M, L < \infty, \tag{23}$$

which is similar to the selection probability model used by Kim and Yu (2011). It is straightforward to see that for every $\varepsilon > 0$, the finite collection of functions

$$\mathcal{F}_\varepsilon = \left\{\exp\{\gamma y\},\ |y| \leq L \,\middle|\, \gamma \in \left\{\left\{2\,i\varepsilon/\big(L\exp(ML)\big) \,\middle|\, |i| \leq \lfloor ML\exp\{ML\}/\varepsilon\rfloor\right\} \cup \{-M\} \cup \{M\}\right\}\right\} \tag{24}$$

is an ε-cover of $\mathcal{F}$ and its covering number is bounded by $(2\,ML\exp\{ML\}\varepsilon^{-1}+3)$; see the Appendix for details. Since this bound grows like ε^{-1} (as $\varepsilon \downarrow 0$), one obtains strong L_p consistency results for the regression estimator (18) under the conditions of Theorem 2 for any sequence $\varepsilon_n \downarrow 0$ (as $n \to \infty$) for which $\log(1/\varepsilon_n)/(mh^d \vee \ell)) \to 0$. Similarly, the conclusions of Theorem 3 and Corollary 2 continue to hold for such a sequence.

2.2 A Horvitz-Thompson type estimator

Our estimators in this section are based on a Horvitz-Thompson type inverse weighting approach (Horvitz and Thompson (1952)). This method works by scaling each observed response variable Y with the inverse of the estimate of the *selection* probability, $\pi_{\varphi^*}(\mathbf{Z}, Y)$, as given by (13), where φ^* is the true function φ in (13) in the sense that

$$m(\mathbf{X}; \pi_{\varphi^*}) := E\big[\Delta Y / \pi_{\varphi^*}(\mathbf{X}, Y) \big| \mathbf{X}\big] = E[Y|\mathbf{X}] = m(\mathbf{X}). \tag{25}$$

To motivate this approach, consider the hypothetical (and unrealistic) situation where the true function π_{φ^*} is completely known. Then in view of (25) a kernel-type estimator of the regression curve $m(\mathbf{x})$ is simply

$$\widetilde{m}_n(\mathbf{x}; \pi_{\varphi^*}) = \sum_{i=1}^{n} \frac{\Delta_i Y_i}{\pi_{\varphi^*}(\mathbf{Z}_i, Y_i)} \mathcal{K}\big((\mathbf{x} - \mathbf{X}_i)/h\big) \Big/ \sum_{i=1}^{n} \mathcal{K}\big((\mathbf{x} - \mathbf{X}_i)/h\big). \tag{26}$$

Since π_{φ^*} is unknown, we proceed as follows. For each $\varphi \in \mathcal{F}$, consider the estimate of the selection probability π_φ of (13), based on $\mathbb{D}_m$, given by

$$\widetilde{\pi}_\varphi(\mathbf{Z}_i, Y_i) = \left[1 \,+\, \frac{1 - \widetilde{\eta}_m(\mathbf{Z}_i)}{\widetilde{\psi}_m(\mathbf{Z}_i; \varphi)} \cdot \varphi(Y_i)\right]^{-1}, \tag{27}$$

where

$$\begin{cases} \widetilde{\psi}_m(\mathbf{Z}_i; \varphi) \,=\, \sum_{j \in \boldsymbol{\mathcal{I}}_m,\, j \neq i} \Delta_j \varphi(Y_j)\, \mathcal{K}_0((\mathbf{Z}_i - \mathbf{Z}_j)/h) \Big/ \sum_{j \in \boldsymbol{\mathcal{I}}_m,\, j \neq i} \mathcal{K}_0((\mathbf{Z}_i - \mathbf{Z}_j)/h) \\ \widetilde{\eta}_m(\mathbf{Z}_i) \,=\, \sum_{j \in \boldsymbol{\mathcal{I}_m},\, j \neq i} \Delta_j \mathcal{K}_0((\mathbf{Z}_i - \mathbf{Z}_j)/h) \Big/ \sum_{j \in \boldsymbol{\mathcal{I}_m},\, j \neq i} \mathcal{K}_0((\mathbf{Z}_i - \mathbf{Z}_j)/h). \end{cases} \tag{28}$$

Since $\pi_\varphi > \pi_{\min} > 0$ (by assumption (A)) and since $\widehat{\psi}_m(\mathbf{Z}_i; \varphi)$ in (28) is the estimator of the conditional expectation $E_\varphi[\Delta_i \varphi(Y_i) | \mathbf{Z}_i] \geq \varrho_0 > 0$ (by assumption (D)), we also consider the following truncated-type version of the estimator in (27)

$$\breve{\pi}_\varphi(\mathbf{Z}_i, Y_i) = \left[1 \,+\, \frac{1 - \widetilde{\eta}_m(\mathbf{Z}_i)}{\pi_0 \vee \widetilde{\psi}_m(\mathbf{Z}_i; \varphi)} \cdot \varphi(Y_i)\right]^{-1}, \tag{29}$$

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 (29) can be viewed as a one-sided winsorized estimator of π_φ (compare this with $\widetilde{\pi}_\varphi$ in (27)). In applications with either simulated or real data, π_0 is chosen to be a small positive number such as $10^{-\nu}$, $\nu \geq 3$. Next, let $\varepsilon_n > 0$ be a decreasing sequence $\varepsilon_n \downarrow 0$, 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 $\mathcal{F}$. Then, depending on whether (27) or (29) is used, an estimator of the unknown function φ^* based on the ε_n-cover $\mathcal{F}_{\varepsilon_n}$ is given by

$$\begin{cases} \widetilde{\varphi}_n \;:=\; \operatorname{argmin}_{\varphi \in \mathcal{F}_{\varepsilon_n}} \ell^{-1} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i}{\widetilde{\pi}_\varphi(\mathbf{Z}_i, Y_i)} \big|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - Y_i\big|^2, & \text{if (27) is used,} \\ \breve{\varphi}_n \;:=\; \operatorname{argmin}_{\varphi \in \mathcal{F}_{\varepsilon_n}} \ell^{-1} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i}{\breve{\pi}_\varphi(\mathbf{Z}_i, Y_i)} \big|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \breve{\pi}_\varphi) - Y_i\big|^2, & \text{if (29) is used,} \end{cases} \tag{30}$$

where

$$\widehat{m}_m^{\mathrm{HT}}(\mathbf{x};\widetilde{\pi}_\varphi) \;=\; \sum_{i\in\boldsymbol{\mathcal{I}}_m} \frac{\Delta_i Y_i}{\widetilde{\pi}_\varphi(\mathbf{Z}_i,Y_i)}\,\mathcal{K}\big((\mathbf{x}-\mathbf{X}_i)/h\big)\Big/ \sum_{i\in\boldsymbol{\mathcal{I}}_m}\mathcal{K}\big((\mathbf{x}-\mathbf{X}_i)/h\big), \tag{31}$$

and $\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\breve{\pi}_\varphi)$ is obtained by replacing $\widetilde{\pi}_\varphi$ with $\breve{\pi}_\varphi$ in (31). Finally, our proposed Horvitz-Thompson type estimator of the regression function $m(\mathbf{x})$ is given by

$$\begin{cases} \widehat{m}^{\mathrm{HT}}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_n}) \;:=\; \widehat{m}_m^{\mathrm{HT}}(\mathbf{x};\pi_\varphi)\Big|_{\pi_\varphi=\widetilde{\pi}_{\widetilde{\varphi}_n}} & \text{if (27) is used,} \\ \widehat{m}^{\mathrm{HT}}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n}) \;:=\; \widehat{m}_m^{\mathrm{HT}}(\mathbf{x};\pi_\varphi)\Big|_{\pi_\varphi=\breve{\pi}_{\breve{\varphi}_n}} & \text{if (29) is used,} \end{cases} \tag{32}$$

where $\widehat{m}_m^{\mathrm{HT}}(\mathbf{x};\pi_\varphi)$ is as in (31) but with $\widetilde{\pi}_\varphi$ replaced by π_φ.

Next, we compare and study the asymptotic performance of the two estimators in (32). It turns out, as in Theorem 2 and its corollary (i.e., Corollary 1), that exponential upper bounds along with strong consistency results are available for both estimators. However, in the case of the winsorized-type estimator $\widehat{m}^{\mathrm{HT}}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})$, one can also study the rates of convergence. To present these results, we start by stating the following theorem.

Theorem 4 *Consider the two regression function estimators defined via (32) and let the missing probability mechanism π_φ be as in (13).*

(i) *Let $\widehat{m}^{\mathrm{HT}}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_n})$ be the top estimator in (32) and suppose that assumptions (A′),(B)–(G) hold. Then for every $\varepsilon_n > 0$ satisfying $\varepsilon_n \downarrow 0$, as $n\to\infty$, every $t>0$, any distribution of $(\mathbf{X},Y)\in\mathbb{R}^d\times[-L,L]$, $L<\infty$, and n large enough,*

$$P\left\{\int\Big|\widehat{m}^{\mathrm{HT}}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_n})-m(\mathbf{x})\Big|^2\mu(d\mathbf{x}) \,>\, t\right\} \,\le\, \big|\mathcal{F}_{\varepsilon_n}\big|\left(c_{13}\,e^{-c_{14}\ell t^2}+c_{15}\ell m\,e^{-c_{16}mh^dt^2}+c_{17}\,\ell m\,e^{-c_{18}mh^d}\right), \tag{33}$$

whenever $\varphi^*\in\mathcal{F}$, where $|\mathcal{F}_{\varepsilon_n}|$ is the cardinality of the set $\mathcal{F}_{\varepsilon_n}$ and c_{13}–c_{18} are positive constants not depending on m, ℓ, n, or t.

(ii) *Let $\widehat{m}^{\mathrm{HT}}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})$ be the second estimator in (32) and suppose that assumptions (A′), (B)–(G) hold. If the truncation constant π_0 in (29) is any constant satisfying $0<\pi_0\le\pi_{_{min}}$ then, under the conditions of part **(i)** of the theorem, the bound in (33) continues to hold (with different constants c_{13}–c_{18}) for the probability $P\big\{\int|\widehat{m}^{\mathrm{HT}}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})-m(\mathbf{x})|^2\mu(d\mathbf{x})>t\big\}$.*

Remark 3 *As in Remark 2, it is straightforward to show that Part (ii) of the above theorem holds more generally for all $p\ge 2$. I.e., the bound in (33) holds for $P\big\{\int|\widehat{m}^{HT}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})-m(\mathbf{x})|^p\mu(d\mathbf{x})>t\big\}$, for all $p\ge 2$.*

The following result may be viewed as the counterpart of Corollary 1 for the two regression function estimators in (32).

Corollary 3 *Consider the two estimators in (32). If, as* $n \to \infty$,

$$\varepsilon_n \downarrow 0, \quad \frac{\log(m \vee \ell)}{mh^d} \to 0, \quad \frac{\log|\mathcal{F}_{\varepsilon_n}|}{mh^d} \to 0, \text{ and } \quad \frac{\log|\mathcal{F}_{\varepsilon_n}|}{\ell} \to 0, \tag{34}$$

then, under the conditions of Theorem 4, the top estimator in (32) satisfies the strong convergence property, $E\Big[\big|\widehat{m}^{\mathrm{HT}}(\mathbf{X}; \widetilde{\pi}_{\widetilde{\varphi}_n}) - m(\mathbf{X})\big|^2 \Big| \mathbb{D}_n\Big] \to^{a.s.} 0$. *However, for the second estimator in (32),*

$$E\Big[\big|\widehat{m}^{\mathrm{HT}}(\mathbf{X}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\big|^p \Big| \mathbb{D}_n\Big] \longrightarrow^{a.s.} 0\,, \quad \text{for all } p \geq 2.$$

We also note that under the conditions of Corollary 3, by Lebesgue dominated convergence theorem, and without further ado, one has $E\big|\widehat{m}^{\mathrm{HT}}(\mathbf{X}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\big|^p \to 0$, for all $p \in [2, \infty)$. However, to study the rates of convergence here, we state the following theorem which is the counterpart of Theorem 3 for the estimator $\widehat{m}^{\mathrm{HT}}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n})$.

Theorem 5 *Let* $\widehat{m}^{HT}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n})$ *be the second estimator in (32). Then, under the conditions of Theorem 4, for all* $p \in [2, \infty)$ *and* n *large enough,*

$$\begin{aligned} &E\Big|\widehat{m}^{HT}(\mathbf{X}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\Big|^p \\ &\quad \leq \sqrt{\frac{c_{19} + \log \ell + \log m + \log|\mathcal{F}_{\varepsilon_n}|}{c_{20} \cdot (\ell \wedge mh^d)}} + \sqrt{\frac{1}{c_{21} \cdot (\ell \wedge mh^d)\big[c_{19} + \log \ell + \log m + \log|\mathcal{F}_{\varepsilon_n}|\big]}} \\ &\qquad + \; c_{22}\,\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell\, m\, e^{-c_{23}\, mh^d}, \end{aligned}$$

where c_{19}–c_{23} *are positive constants not depending on* m, ℓ, *or* n.

The following is an immediate corollary to Theorem 5.

Corollary 4 *Let* $\widehat{m}^{HT}(\mathbf{x}; \breve{\pi}_{\breve{\varphi}_n})$ *be the second estimator in (32) and suppose that (34) holds. Then, under the conditions of Theorem 4, for all* $p \geq 2$,

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

Once again, 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

$$E\Big|\widehat{m}^{\mathrm{HT}}(\mathbf{X}; \breve{\pi}_{\breve{\varphi}_n}) - m(\mathbf{X})\Big|^p \;=\; \mathcal{O}\left(\sqrt{\frac{\log(n \vee |\mathcal{F}_{\varepsilon_n}|)}{nh^d}}\right), \quad \text{for all } p \geq 2.$$

Remark 4 *The rates of convergence derived in Corollaries 2 and 4 are 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 cardinality of the* ε_n*-cover satisfies* $\log|\mathcal{F}_{\varepsilon_n}| = \mathcal{O}(n)$.

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 NMAR missing data where the estimation process involves many steps and many components. The rates in Corollaries 2 and 4 also show 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 $m = \alpha \cdot n$ and $\ell = (1-\alpha) \cdot n$ for any $\alpha \in (0,1)$.

3 Applications to classification with partially labeled data

Consider the following standard two-group classification problem. Let $(\mathbf{X}, Y)$ be a random pair, where $\mathbf{X} \in \mathbb{R}^d$ is a vector of covariates and $Y \in \{0,1\}$, called the class variable or class label, has to be predicted based on $\mathbf{X}$. More specifically, the aim of classification is to find a map/function $g : \mathbb{R} \to \{0,1\}$ for which the misclassification error, i.e.,

$$L(g) := P\{g(\mathbf{X}) \neq Y\}, \tag{35}$$

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

$$g_{\text{B}}(\mathbf{x}) = \begin{cases} 1 & \text{if } \; m(\mathbf{x}) \; := \; E\big[Y \big| \, \mathbf{X} = \mathbf{x}\big] \; > \; \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{36}$$

see, for example, Devroye et al (1996; Ch. 2). Since the distribution of $(\mathbf{X}, Y)$ is virtually always unknown, finding the best classifier g_{B} is impossible. However, 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

$$L_n(\widehat{g}_n) = P\big\{\widehat{g}_n(\mathbf{X}) \neq Y \big| \, \mathbb{D}_n\big\} \tag{37}$$

be the conditional misclassification error of $\widehat{g}_n$. Then $\widehat{g}_n$ is said to be weakly (strongly) Bayes consistent if $L_n(\widehat{g}_n) \to L(g_{\text{B}})$ in probability (almost surely). Now, let $\widehat{m}(\mathbf{x})$ be any estimator of the regression function $m(\mathbf{x}) := E\big[Y|\mathbf{X} = \mathbf{x}\big]$ and consider the plug-in type classifier

$$\widehat{g}_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \; \widehat{m}(\mathbf{x}) \; > \; \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{38}$$

Then, one has (see Lemma 6.1 of Devroye et al (1996))

$$L_n(\widehat{g}_n) - L(g_{\text{B}}) \; \leq \; 2E\big[\big|\widehat{m}(\mathbf{X}) - m(\mathbf{X})\big| \, \big| \mathbb{D}_n\big] \, , \tag{39}$$

and by the dominated convergence theorem, $E\big[L_n(\widehat{g}_n)\big] - L(g_{\rm B}) \;\leq\; 2E\big|\widehat{m}(\mathbf{X}) - m(\mathbf{X})\big|$. Next, suppose that some of the Y_i's may be missing not at random (NMAR) and consider the regression estimator $\widehat{m}(\mathbf{x};\widehat{\varphi}_n)$ in (18). Denote the plug-in classifier corresponding to $\widehat{m}(\mathbf{x};\widehat{\varphi}_n)$ by

$$\widehat{g}_n(\mathbf{x};\widehat{\varphi}_n) \;:=\; \begin{cases} 1 & \text{if } \; \widehat{m}(\mathbf{x};\widehat{\varphi}_n) \;>\; \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

To study the asymptotic performance of the classifier in (40), we first state the following so-called margin condition (see, for example, Audibert and Tsybakov(2007)).

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

$$P\left\{0 \;<\; \left|m(\mathbf{X}) - \frac{1}{2}\right| \;\leq\; t\right\} \;\leq\; c\,t^{\alpha}, \quad \text{for all } t > 0. \tag{41}$$

Applications of the margin condition to classification has been studied by many authors; see, for example, Mammen and Tsybakov (1999), Massart and Nédélec (2006), Audibert and Tsybakov(2007), Tsybakov and van de Geer (2005), Kohler and Krzyżak (2007), and Döring et al (2016).

Theorem 6 *Consider the classifier $\widehat{g}_n(\mathbf{x};\widehat{\varphi}_n)$ given by (40). If (22) holds then, under the conditions of Theorem 2, we have*

(i) $\displaystyle P\left\{\widehat{g}_n(\mathbf{X};\widehat{\varphi}_n) \neq Y \middle| \mathbb{D}_n\right\} \longrightarrow^{a.s.} P\{g_{B}(\mathbf{X}) \neq Y\}.$

(ii) $\displaystyle P\left\{\widehat{g}_n(\mathbf{X};\widehat{\varphi}_n) \neq Y\right\} - P\{g_{B}(\mathbf{X}) \neq Y\} = \mathcal{O}\left(\left(\tfrac{\log(\ell\vee|\mathcal{F}_{\varepsilon_n}|)}{\ell\wedge(mh^d)}\right)^{1/4}\right).$

(iii) *If the margin condition (41) holds then*

$$P\left\{\widehat{g}_n(\mathbf{X};\widehat{\varphi}_n) \neq Y\right\} - P\{g_{B}(\mathbf{X}) \neq Y\} = \mathcal{O}\left(\left(\frac{\log(\ell \vee |\mathcal{F}_{\varepsilon_n}|)}{\ell \wedge (mh^d)}\right)^{\frac{1+\alpha}{2(2+\alpha)}}\right),$$

where α is as in (41).

Part (iii) of the above theorem shows that for large values of α we can obtain rates closer to $\big(\log(\ell \vee |\mathcal{F}_{\varepsilon_n}|)/\big[\ell \wedge (mh^d)\big]\big)^{1/2}$ which is the same as that of the actual regression estimator (see Corollary 2).

Next, consider the Horvitz-Thompson type regression estimators given by (32) and denote the corresponding plug-in classifiers by

$$\widetilde{g}_n^{\rm HT}(\mathbf{x};\widetilde{\pi}) = \begin{cases} 1 & \text{if } \; \widehat{m}^{\rm HT}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_n}) \;>\; \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \breve{g}_n^{\rm HT}(\mathbf{x};\breve{\pi}) = \begin{cases} 1 & \text{if } \; \widehat{m}^{\rm HT}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n}) \;>\; \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{42}$$

where $\widehat{m}^{\rm HT}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_n})$ and $\widehat{m}^{\rm HT}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})$ are as in (32). As for the asymptotic performance of the two classifiers in (42), we have the following counterpart of Theorem (6).

Theorem 7 *Let $\widetilde{g}_n^{HT}$ and $\breve{g}_n^{HT}$ be the two classifiers in (42). If (34) holds then, under the conditions of Theorem 4, we have*

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

(ii) $$P\left\{\breve{g}_n^{HT}(\mathbf{X};\breve{\varphi}_n) \neq Y\right\} - P\{g_B(\mathbf{X}) \neq Y\} = \mathcal{O}\left(\left(\tfrac{\log(\ell \vee m \vee |\mathcal{F}_{\varepsilon_n}|)}{\ell \wedge (mh^d)}\right)^{1/4}\right).$$

(iii) *If the margin condition (41) holds then*

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

where α is as in (41).

Here, we observe that for large values of α in part (iii) of the above theorem, one can obtain rates closer to $\big(\log(\ell \vee m \vee |\mathcal{F}_{\varepsilon_n}|)/\big[\ell \wedge (mh^d)\big]\big)^{1/2}$ which is similar to that of the winsorized-type regression estimator $\widehat{m}^{\mathrm{HT}}(\mathbf{x};\breve{\pi}_{\breve{\varphi}_n})$ in (32); see Corollary 4.

4 Proofs of the main results

We start by stating a number of lemmas whose proofs appear in the Appendix. Using the notation of Section 2.1, let $\mathcal{F}$ be 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 2.1). Next, for each $\varphi \in \mathcal{F}$, put

$$\psi_k(\mathbf{x};\varphi) := E\left[\Delta Y^{2-k}\varphi(Y)\middle|\mathbf{X}=\mathbf{x}\right] \quad \text{and} \quad \eta_k(\mathbf{x}) := E\big[\Delta Y^{2-k}\big|\mathbf{X}=\mathbf{x}\big], \quad \text{for } k=1,2, \tag{43}$$

and define

$$m(\mathbf{x};\varphi) \;=\; \eta_1(\mathbf{x}) + \frac{\psi_1(\mathbf{x};\varphi)}{\psi_2(\mathbf{x};\varphi)} \cdot (1-\eta_2(\mathbf{x}))\,. \tag{44}$$

Also, define

$$\widehat{L}_{m,\ell}(\varphi) \;:=\; \frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i}{\widehat{\pi}_\varphi(\mathbf{Z}_i,Y_i)}\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2, \tag{45}$$

where $\widehat{\pi}_\varphi(\mathbf{z},y)$ and $\widehat{m}_m(\mathbf{x};\varphi)$ are as in (15) and (9), and put

$$\varphi_\varepsilon \;:=\; \operatorname*{argmin}_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\varphi) - Y\big|^2 \quad \text{and} \quad \widehat{\varphi}_\varepsilon \;:=\; \operatorname*{argmin}_{\varphi\in\mathcal{F}_\varepsilon} \widehat{L}_{m,\ell}(\varphi). \tag{46}$$

Lemma 1 *Let φ^* be the true (unknown) version of the function φ in (13). Also, let $m(\mathbf{x};\varphi)$ be as defined in model (44). Then the regression function $m(\mathbf{x}) = E[Y|\mathbf{X}=\mathbf{x}]$ can be represented as*

$$m(\mathbf{x}) \;=\; m(\mathbf{x};\varphi^*) \;=\; \eta_1(\mathbf{x}) + \frac{\psi_1(\mathbf{x};\varphi^*)}{\psi_2(\mathbf{x};\varphi^*)} \cdot (1-\eta_2(\mathbf{x}))\,. \tag{47}$$

where the functions ψ_k and η_k, $k=1,2$, are given by (43).

Lemma 2 *Let $m(\mathbf{x};\varphi_j)$, $j=1,2$, be as in (44), where $\varphi_j : [-L,L] \to (0,B]$ for some positive number B. Then, under assumption (D), one has*

$$E\Big|m(\mathbf{X};\varphi_1) - m(\mathbf{X};\varphi_2)\Big| \;\le\; C \cdot \sup_{-L\le y\le L} \big|\varphi_1(y) - \varphi_2(y)\big|,$$

where the constant $C>0$ can be taken to be $C = 2L/\varrho_0$, with ϱ_0 as in assumption (D).

Lemma 3 *Let $m(\mathbf{x};\varphi)$, $\widehat{L}_{m,\ell}(\varphi)$, φ_ε, and $\widehat{\varphi}_\varepsilon$ be as in (44), (45), and (46), respectively. Then, under the conditions of Theorem 2, we have*

$$\begin{aligned} E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\Big|^2 \Big| \mathbb{D}_n\right] \;\le\; & \sup_{\varphi\in\mathcal{F}_\varepsilon} \left| E\left[\Big|\widehat{m}_m(\mathbf{X};\varphi) - Y\Big|^2 \Big|\mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\varphi)\right| \\ & + \sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widehat{L}_{m,\ell}(\varphi) - E\Big|m(\mathbf{X};\varphi) - Y\Big|^2\right| \;+\; C_1\,\varepsilon^{1/2}, \quad (48) \end{aligned}$$

where C_1 is a positive constant not depending on n or ε, and $\widehat{m}_m(\mathbf{X};\varphi)$ is as in (9).

Lemma 4 *Let $\mathcal{K}$ be a regular kernel. Also, let μ be any probability measure on the Borel sets of $\mathbb{R}^d$. Then there is a positive constant $\rho(\mathcal{K})$, depending on the kernel $\mathcal{K}$ but not n, such that for every $h>0$,*

$$\sup_{\mathbf{u}\in\mathbb{R}^d} \int \frac{\mathcal{K}((\mathbf{x}-\mathbf{u})/h)}{E[\mathcal{K}((\mathbf{x}-\mathbf{X})/h)]}\,\mu(d\mathbf{x}) \;\le\; \rho(\mathcal{K})\,.$$

Lemma 5 *Let $(\mathbf{U},Y),(\mathbf{U}_1,Y_1),\ldots,(\mathbf{U}_n,Y_n)$ be iid $\mathbb{R}^d\times[-A,A]$-valued random vectors for some $A\in[0,\infty)$. Also, let $m(\mathbf{u}) = E[Y|\mathbf{U}{=}\mathbf{u}]$ be the regression function and define quantity $\widetilde{m}_n(\mathbf{u}) = \sum_{i=1}^n Y_i\,\mathcal{K}\big((\mathbf{u}-\mathbf{U}_i)/h\big)\big/\big\{nE[\mathcal{K}((\mathbf{u}-\mathbf{U})/h)]\big\}$, where $\mathcal{K}$ is a regular kernel. If $h\to 0$ and $nh^d\to\infty$, as $n\to\infty$, then for every $t>0$ and large enough n,*

$$P\left\{\int \Big|\widetilde{m}_n(\mathbf{u}) - m(\mathbf{u})\Big|\,\mu(d\mathbf{u}) > t\right\} \;\le\; e^{-nt^2/(64A^2\,\rho^2(\mathcal{K}))}$$

where μ is the probability measure of $\mathbf{U}$, and $\rho(\mathcal{K})$ is as in Lemma 4.

PROOF OF THEOREM 1

First observe that for every $p\ge 1$, $|\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x}) - m(\mathbf{x})|^p \le \{|\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x})| + |m(\mathbf{x})|\}^{p-1}|\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x}) - m(\mathbf{x})| \le (3L)^{p-1}|\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x}) - m(\mathbf{x})|$, where the term $(3L)$ follows from the observations that $|m(\mathbf{x})|\le L$ and $|\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x})| \le |\widehat{\eta}_1(\mathbf{x},0)| + \big(|\widehat{\eta}_1(\mathbf{x},\widehat{\gamma})/\widehat{\eta}_2(\mathbf{x},\widehat{\gamma})|\big)\cdot|1-\widehat{\eta}_2(\mathbf{x},0)| \le L + \big(L\cdot|\widehat{\eta}_2(\mathbf{x},\widehat{\gamma})|/|\widehat{\eta}_2(\mathbf{x},\widehat{\gamma})|\big)\cdot 1 = 2L$. Therefore, we only need to prove the theorem for the case of $p=1$. To this end, let $\eta_k(\mathbf{x},t)$ and $\widehat{\eta}_k(\mathbf{x},t)$, $k=1,2$, $t\in\mathbb{R}$, be the quantities defined in (3) and (6), respectively. Then it is straightforward to show that in view of (5) and (4), and the fact that $|\widehat{\eta}_1(\mathbf{x},\widehat{\gamma})/\widehat{\eta}_2(\mathbf{x},\widehat{\gamma})| \le L\cdot|\widehat{\eta}_2(\mathbf{x},\widehat{\gamma})|/|\widehat{\eta}_2(\mathbf{x},\widehat{\gamma})| = L$, one has

$$\big|\widehat{m}_{n,\widehat{\gamma}}(\mathbf{x}) - m(\mathbf{x})\big| \;\le\; \big|\widehat{\eta}_1(\mathbf{x},0) - \eta_1(\mathbf{x},0)\big| + \left|\frac{\widehat{\eta}_1(\mathbf{x},\widehat{\gamma})}{\widehat{\eta}_2(\mathbf{x},\widehat{\gamma})} - \frac{\eta_1(\mathbf{x},\gamma^*)}{\eta_2(\mathbf{x},\gamma^*)}\right| + L\cdot\big|\widehat{\eta}_2(\mathbf{x},0) - \eta_2(\mathbf{x},0)\big|. \quad (49)$$

But the first and third terms on the right side of (49) can be immediately bounded using the classical result of Devroye and Krzyżak (1989). More specifically, for every $t > 0$ and n large enough,

$$P\Big\{ \int \big|\widehat{\eta}_1(\mathbf{x},0) - \eta_1(\mathbf{x},0)\big|\mu(d\mathbf{x}) > t\Big\} \leq e^{-c_{24}n} \text{ and } P\Big\{ \int L\big|\widehat{\eta}_2(\mathbf{x},0) - \eta_2(\mathbf{x},0)\big|\mu(d\mathbf{x}) > t\Big\} \leq e^{-c_{25}n} \tag{50}$$

where c_{24} and c_{25} are positive constants that depend on t but not n. To deal with the middle term on the right side of (49), we note that it can be written as

$$\begin{aligned}
&\left|\frac{1}{\eta_2(\mathbf{x},\gamma^*)}\left[\frac{\widehat{\eta}_1(\mathbf{x},\widehat{\gamma})}{\widehat{\eta}_2(\mathbf{x},\widehat{\gamma})}\cdot\big(\eta_2(\mathbf{x},\gamma^*) - \widehat{\eta}_2(\mathbf{x},\widehat{\gamma})\big) + \big(\widehat{\eta}_1(\mathbf{x},\widehat{\gamma}) - \eta_1(\mathbf{x},\gamma^*)\big)\right]\right| \\
&\leq \frac{1}{\pi_{\min}\exp\{-L|\gamma^*|\}}\Big[\big|\widehat{\eta}_1(\mathbf{x},\widehat{\gamma}) - \widehat{\eta}_1(\mathbf{x},\gamma^*)\big| + \big|\widehat{\eta}_1(\mathbf{x},\gamma^*) - \eta_1(\mathbf{x},\gamma^*)\big| + L\big|\widehat{\eta}_2(\mathbf{x},\widehat{\gamma}) - \widehat{\eta}_2(\mathbf{x},\gamma^*)\big| \\
&\qquad + L\big|\widehat{\eta}_2(\mathbf{x},\gamma^*) - \eta_2(\mathbf{x},\gamma^*)\big|\Big],
\end{aligned}$$

where the above inequality follows from assumption (A) with the simple fact that $\eta_2(\mathbf{X},\gamma^*) = E\left(E\left[\Delta\exp\{\gamma^*Y\}\big|\mathbf{X},Y\right]\big|\mathbf{X}\right) = E\left[\exp\{\gamma^*Y\}\pi_{\gamma^*}(\mathbf{X},Y)\big|\mathbf{X}\right] \geq \inf_{\mathbf{z},y}\pi_{\gamma^*}(\mathbf{z},y)\exp(-|\gamma^*|L)$, together with the observation that $\widehat{\eta}_1(\mathbf{x},\widehat{\gamma})/\widehat{\eta}_2(\mathbf{x},\widehat{\gamma}) \leq L$. Consequently, for every $t > 0$, the integral of the middle term on the right side of (49) can be dealt with as follows

$$\begin{aligned}
P\left\{\int\left|\frac{\widehat{\eta}_1(\mathbf{x},\widehat{\gamma})}{\widehat{\eta}_2(\mathbf{x},\widehat{\gamma})} - \frac{\eta_1(\mathbf{x},\gamma^*)}{\eta_2(\mathbf{x},\gamma^*)}\right|\mu(d\mathbf{x}) > t\right\} \;\leq\; & P\left\{\int\big|\widehat{\eta}_1(\mathbf{x},\widehat{\gamma}) - \widehat{\eta}_1(\mathbf{x},\gamma^*)\big|\mu(d\mathbf{x}) > \frac{\pi_{\min}t}{4\exp\{L|\gamma^*|\}}\right\} \\
&+P\left\{\int\big|\widehat{\eta}_1(\mathbf{x},\gamma^*) - \eta_1(\mathbf{x},\gamma^*)\big|\mu(d\mathbf{x}) > \frac{\pi_{\min}t}{4\exp\{L|\gamma^*|\}}\right\} \\
&+P\left\{\int\big|\widehat{\eta}_2(\mathbf{x},\widehat{\gamma}) - \widehat{\eta}_2(\mathbf{x},\gamma^*)\big|\mu(d\mathbf{x}) > \frac{\pi_{\min}t}{4L\exp\{L|\gamma^*|\}}\right\} \\
&+P\left\{\int\big|\widehat{\eta}_2(\mathbf{x},\gamma^*) - \eta_2(\mathbf{x},\gamma^*)\big|\mu(d\mathbf{x}) > \frac{\pi_{\min}t}{4L\exp\{L|\gamma^*|\}}\right\} \\
&:= \sum_{k=1}^{4}\mathcal{P}_{nk}(t).
\end{aligned} \tag{51}$$

To deal with the first term in (51), i.e., the term $\mathcal{P}_{n1}(t)$, put

$$\begin{aligned}
\Gamma_n'(\mathbf{x}) &= \sum_{i=1}^{n}\Delta_i Y_i\left(\exp\{\widehat{\gamma}\,Y_i\} - \exp\{\gamma^*\,Y_i\}\right)\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)\Big/ nE\big[\mathcal{K}((\mathbf{x}-\mathbf{X})/h)\big] \\
\Gamma_n''(\mathbf{x}) &= \frac{\sum_{i=1}^{n}\Delta_i Y_i\left(\exp\{\widehat{\gamma}\,Y_i\} - \exp\{\gamma^*\,Y_i\}\right)\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)}{\sum_{i=1}^{n}\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)}\cdot\left[\frac{\sum_{i=1}^{n}\mathcal{K}((\mathbf{x}-\mathbf{X}_i)/h)}{nE\big[\mathcal{K}((\mathbf{x}-\mathbf{X})/h)\big]} - 1\right]
\end{aligned}$$

and observe that

$$\mathcal{P}_{n1}(t) \;\leq\; P\left\{\int\big|\Gamma_n'(\mathbf{x})\big|\mu(d\mathbf{x}) > \frac{\pi_{\min}t}{8\exp\{L|\gamma^*|\}}\right\} + P\left\{\int\big|\Gamma_n''(\mathbf{x})\big|\mu(d\mathbf{x}) > \frac{\pi_{\min}t}{8\exp\{L|\gamma^*|\}}\right\}. \tag{52}$$

Furthermore,

$$\begin{aligned}
\int \big|\Gamma_n'(\mathbf{x})\big|\mu(d\mathbf{x}) &\leq \sup_{\mathbf{z}} \int \frac{\mathcal{K}((\mathbf{x}-\mathbf{z})/h)}{E\big[\mathcal{K}((\mathbf{x}-\mathbf{X})/h)\big]}\mu(d\mathbf{x}) \cdot \frac{1}{n}\sum_{i=1}^{n}\Big|\Delta_i Y_i\Big(\exp\{\widehat{\gamma}\,Y_i\} - \exp\{\gamma^* Y_i\}\Big)\Big| \\
&\leq \frac{L\rho(\mathcal{K})}{n}\sum_{i=1}^{n}\Big|\exp\{\widehat{\gamma}\,Y_i\} - \exp\{\gamma^* Y_i\}\Big|, \quad \text{(by Lemma 4)} \\
&\leq n^{-1}L\,\rho(\mathcal{K})\sum_{i=1}^{n}\Big|\big(\widehat{\gamma}-\gamma^*\big)Y_i\,\exp\big\{\overline{\gamma}Y_i\big\}\Big|, \quad \text{(via a one-term Taylor expansion)}, \\
&\leq n^{-1}L^2\,\rho(\mathcal{K})\big|\widehat{\gamma}-\gamma^*\big|\cdot\sum_{i=1}^{n}\exp\Big\{\big|\overline{\gamma}-\gamma^*\big|L+\gamma^* Y_i\Big\}, \qquad (53)
\end{aligned}$$

because $|Y_i| \leq L$ and $\overline{\gamma}Y_i = (\overline{\gamma}-\gamma^*)Y_i + \gamma^* Y_i \leq |\overline{\gamma}-\gamma^*|\cdot L + \gamma^* Y_i$, where $\overline{\gamma}$ is a point in the interior of the line segment joining $\widehat{\gamma}$ and γ^*. Therefore, using the fact that $|\overline{\gamma}-\gamma^*| \leq |\widehat{\gamma}-\gamma^*|$, one finds, for every constants $t > 0$ and $C_0 > 0$,

$$\begin{aligned}
&P\left\{\int \big|\Gamma_n'(\mathbf{x})\big|\mu(d\mathbf{x}) > \frac{\pi_{\min}t}{8\exp\{L|\gamma^*|\}}\right\} \\
&\leq P\left\{\big|\widehat{\gamma}-\gamma^*\big|\exp\big\{\big|\widehat{\gamma}-\gamma^*\big|L\big\}\cdot\frac{1}{n}\sum_{i=1}^{n}\exp\{\gamma^* Y_i\} > \frac{\pi_{\min}t}{8L^2\,\rho(\mathcal{K})\exp\{L|\gamma^*|\}}\right\} \\
&\leq P\Bigg\{\left[\big|\widehat{\gamma}-\gamma^*\big|\exp\big\{\big|\widehat{\gamma}-\gamma^*\big|L\big\}\cdot\frac{1}{n}\sum_{i=1}^{n}\exp\{\gamma^* Y_i\} > \frac{\pi_{\min}t}{8L^2\rho(\mathcal{K})\exp\{L|\gamma^*|\}}\right] \\
&\qquad \cap\Big\{|\widehat{\gamma}-\gamma^*| \leq C_0\Big\}\Bigg\} + P\Big\{|\widehat{\gamma}-\gamma^*| > C_0\Big\} \\
&\leq n\cdot P\left\{\exp\{\gamma^* Y_1\} > \frac{\pi_{\min}t}{8L^2C_0\,\rho(\mathcal{K})\exp\big\{(|\gamma^*|+C_0)L\big\}}\right\} + P\Big\{|\widehat{\gamma}-\gamma^*| > C_0\Big\} \qquad (54) \\
&= 0 + P\Big\{|\widehat{\gamma}-\gamma^*| > C_0\Big\}, \qquad (55)
\end{aligned}$$

for any C_0 satisfying $4L^2C_0\,\rho(\mathcal{K})\exp\big\{(|\gamma^*|+C_0)L\big\} < \pi_{\min}t$. Here, the last line follows because the random variable $\exp\{\gamma^* Y_1\}$ is bounded by $\exp\{|\gamma^*|L\}$, which implies that taking C_0 small enough forces the first probability statement in (54) to become zero. As for the term $\Gamma_n''(\mathbf{x})$, we note that in view of (53) and the observation that $|\overline{\gamma}-\gamma^*| \leq |\widehat{\gamma}-\gamma^*|$, one obtains

$$\int \big|\Gamma_n''(\mathbf{x})\big|\mu(d\mathbf{x}) \;\leq\; L^2\big|\widehat{\gamma}-\gamma^*\big|\max_{1\leq i\leq n}\exp\Big\{\big|\widehat{\gamma}-\gamma^*\big|L+|\gamma^*|L\Big\}\cdot\int\left|\frac{\sum_{j=1}^{n}\mathcal{K}((\mathbf{x}-\mathbf{X}_j)/h)}{nE\big[\mathcal{K}((\mathbf{x}-\mathbf{X})/h)\big]}-1\right|\mu(d\mathbf{x}).$$

Now, observe that

$$\begin{aligned}
&P\left\{\int \big|\Gamma_n''(\mathbf{x})\big|\mu(d\mathbf{x}) \;>\; \frac{\pi_{\min}t}{8\exp\{L|\gamma^*|\}}\right\} \\
&\leq P\left\{\big|\widehat{\gamma}-\gamma^*\big|\;\exp\big\{\big|\widehat{\gamma}-\gamma^*\big|L\big\}\cdot\int\left|\frac{\sum_{j=1}^{n}\mathcal{K}((\mathbf{x}-\mathbf{X}_j)/h)}{nE\big[\mathcal{K}((\mathbf{x}-\mathbf{X})/h)\big]}-1\right|\mu(d\mathbf{x}) \;>\; \frac{\pi_{\min}t}{8L^2\exp\{2L|\gamma^*|\}}\right\}
\end{aligned}$$

$$\leq P\left\{\int\left|\frac{\sum_{j=1}^{n}\mathcal{K}((\mathbf{x}-\mathbf{X}_j)/h)}{nE\big[\mathcal{K}((\mathbf{x}-\mathbf{X})/h)\big]}-1\right|\mu(d\mathbf{x})>\frac{\pi_{\min}t}{8L^2C_0\exp\big\{(2|\gamma^*|+C_0)L\big\}}\right\}+P\big\{|\widehat{\gamma}-\gamma^*|>C_0\big\}$$

$$\leq \exp\left\{\frac{-n\,\pi_{\min}^2t^2}{64^2L^4C_0^2\rho^2(\mathcal{K})\cdot\exp\big\{2\big(2|\gamma^*|+C_0\big)L\big\}}\right\}+P\Big\{|\widehat{\gamma}-\gamma^*|>C_0\Big\}, \tag{56}$$

for large n, by Lemma 5, where C_0 is as in (55); here, we have used Lemma 5 with $m(\mathbf{u})=1$ and $Y_i=1$ for all $i=1,\cdots,n$. Putting together (52), (55), and (56), we find

$$\mathcal{P}_{n1}(t)\;\leq\;\exp\{-C_2n^2t^2\}+2P\Big\{|\widehat{\gamma}-\gamma^*|>C_0\Big\}, \tag{57}$$

for n large enough, where C_2 is a positive constant not depending on n. It is a simple exercise to show that the term $\mathcal{P}_{n3}(t)$ in (51) can also be bounded by the right side of (57). Furthermore, as in (50), once again we can invoke the result of Devroye and Krzyżak (1989) to conclude that $\mathcal{P}_{n2}(t)\leq e^{-c_{26}n}$ and $\mathcal{P}_{n4}(t)\leq e^{-c_{27}n}$, for n large enough, where c_{26} and c_{27} are positive constants not depending on n. These observations in conjunction with (57), (51), (50), and (49) complete the proof of Theorem 1.

□

PROOF OF THEOREM 2

To proceed with the proof, first note that for each $i\in\boldsymbol{\mathcal{I}}_\ell$, we have

$$\frac{\Delta_i\big|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\big|^2}{\widehat{\pi}_\varphi(\mathbf{Z}_i,Y_i)}\;=\;\frac{\Delta_i\,\big|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\big|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)}-\;\Delta_i\,\big|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\big|^2\left[\frac{1}{\pi_\varphi(\mathbf{Z}_i,Y_i)}-\frac{1}{\widehat{\pi}_\varphi(\mathbf{Z}_i,Y_i)}\right].$$

Therefore, by the definition of $\widehat{L}_{m,\ell}(\varphi)$ in (45), one finds for every $\beta>0$

$$\begin{aligned}
&P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\varphi)-E\left[\left|\widehat{m}_m(\mathbf{X};\varphi)-Y\right|^2\Big|\mathbb{D}_m\right]\right|>\beta\right\}\\
&\leq\;P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\frac{\Delta_i\,\big|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\big|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)}-E\left[\left|\widehat{m}_m(\mathbf{X};\varphi)-Y\right|^2\Big|\mathbb{D}_m\right]\right|>\frac{\beta}{2}\right\}\\
&\quad+\;P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Delta_i\,\big|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\big|^2\left[\frac{1}{\pi_\varphi(\mathbf{Z}_i,Y_i)}-\frac{1}{\widehat{\pi}_\varphi(\mathbf{Z}_i,Y_i)}\right]\right|>\frac{\beta}{2}\right\}\\
&:=\;S_n(1)+S_n(2).
\end{aligned}\tag{58}$$

But with $\pi_\varphi(\mathbf{z},y)$ modeled as (13), where $\varphi\in\mathcal{F}$ is the free functional parameter, for each $i\in\boldsymbol{\mathcal{I}}_\ell$

$$E_\varphi\left[\frac{\Delta_i}{\pi_\varphi(\mathbf{Z}_i,Y_i)}\left|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\right|^2\Big|\mathbb{D}_m\right]\;\;=\;\;E_\varphi\left[\frac{\left|\widehat{m}_m(\mathbf{X}_i;\varphi)-Y_i\right|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)}\,E_\varphi\Big(\Delta_i\Big|\mathbb{D}_m,\mathbf{X}_i,Y_i\Big)\Big|\mathbb{D}_m\right]$$

$$= \quad E_\varphi\Big[\Big|\widehat{m}_m(\mathbf{X};\varphi) - Y\Big|^2 \,\Big|\mathbb{D}_m\Big].$$

Furthermore, conditional on $\mathbb{D}_m$, the terms $\Delta_i \big|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\big|^2/\pi_\varphi(\mathbf{Z}_i, Y_i)$, $i \in \boldsymbol{\mathcal{I}}_\ell$, are independent bounded random variables, taking values in $\big[0,\, (3L)^2/\pi_{\min}\big]$. Therefore,

$$\begin{aligned}
S_n(1) \;&\le\; \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\,\in\mathcal{F}_\varepsilon} P\Bigg\{\Bigg|\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i\,\big|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\big|^2}{\pi_\varphi(\mathbf{Z}_i, Y_i)} - E\left[\Big|\widehat{m}_m(\mathbf{X};\varphi) - Y\Big|^2\Big|\mathbb{D}_m\right]\Bigg| > \frac{\beta}{2}\Bigg\} \\
&\le\; \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\,\in\mathcal{F}} E_\varphi\Bigg[P_\varphi\Bigg\{\Bigg|\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i\big|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\big|^2}{\pi_\varphi(\mathbf{Z}_i, Y_i)} - \; E_\varphi\Big[\Big|\widehat{m}_m(\mathbf{X};\varphi) - Y\Big|^2\Big|\mathbb{D}_m\Big]\Bigg| > \frac{\beta}{2}\,\Big|\,\mathbb{D}_m\Bigg\}\Bigg] \\
&\le\; 2\,\big|\mathcal{F}_\varepsilon\big|\,\exp\big\{-\pi^2_{\min}\ell\beta^2/(162L^4)\big\} \quad \text{(via Hoeffding's inequality)}, \qquad (59)
\end{aligned}$$

where the line above (59) follows from conditioning in conjunction with assumption (F). To deal with the term $S_n(2)$ in (58), let $\widetilde{\psi}_m(\mathbf{Z}_i;\varphi)$ and $\widetilde{\eta}_m(\mathbf{Z}_i)$ be as in (28) and observe that in view of (13), (14), (15), (16), and the fact that $|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i| \le 3L$, we can write

$$\begin{aligned}
S_n(2) \;&\le\; \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}_\varepsilon} P\Bigg\{\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Bigg|\frac{1}{\widehat{\pi}_\varphi(\mathbf{Z}_i, Y_i)} - \frac{1}{\pi_\varphi(\mathbf{Z}_i, Y_i)}\Bigg| > \frac{\beta}{18L^2}\Bigg\} \\
&\le\; \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}_\varepsilon} \sum_{i\in\boldsymbol{\mathcal{I}}_\ell} E\Bigg[P\Bigg\{\Bigg|\frac{1-\widetilde{\eta}_m(\mathbf{Z}_i)}{\widetilde{\psi}_m(\mathbf{Z}_i;\varphi)} - \frac{E[1-\Delta_i|\mathbf{Z}_i]}{E[\Delta_i\,\varphi(Y_i)|\mathbf{Z}_i]}\Bigg|\,\varphi(Y_i) > \frac{\beta}{18L^2}\,\Bigg|\,\mathbf{Z}_i, Y_i\Bigg\}\Bigg], \qquad (60)
\end{aligned}$$

where the last line follows upon replacing the term $\exp\{g(\mathbf{z})\}$ in (13) by the right side of (14). Now, to bound (60), we note that

$$\begin{aligned}
&\Bigg|\frac{1-\widetilde{\eta}_m(\mathbf{Z}_i)}{\widetilde{\psi}_m(\mathbf{Z}_i;\varphi)} - \frac{1-E\big[\Delta_i\big|\mathbf{Z}_i\big]}{E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big]}\Bigg| \\
&\qquad=\; \Bigg|-\frac{1-\widetilde{\eta}_m(\mathbf{Z}_i)}{\widetilde{\psi}_m(\mathbf{Z}_i;\varphi)}\cdot\frac{\widetilde{\psi}_m(\mathbf{Z}_i;\varphi) - E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big]}{E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big]} + \frac{E\big[\Delta_i\big|\mathbf{Z}_i\big] - \widetilde{\eta}_m(\mathbf{Z}_i)}{E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big]}\Bigg| \\
&\qquad\qquad\le\; \Bigg|\frac{1-\widetilde{\eta}_m(\mathbf{Z}_i)}{\widetilde{\psi}_m(\mathbf{Z}_i;\varphi)}\Bigg|\cdot\Bigg|\frac{\widetilde{\psi}_m(\mathbf{Z}_i;\varphi) - E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big]}{E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big]}\Bigg| + \Bigg|\frac{E\big[\Delta_i\big|\mathbf{Z}_i\big] - \widetilde{\eta}_m(\mathbf{Z}_i)}{E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big]}\Bigg|.
\end{aligned}$$

Therefore, in view of (43), the inner conditional probability in (60) becomes

$$\begin{aligned}
&P\left\{\left|\frac{1-\widetilde{\eta}_m(\mathbf{Z}_i)}{\widetilde{\psi}_m(\mathbf{Z}_i;\varphi)} - \frac{1-E\big[\Delta_i\big|\mathbf{Z}_i\big]}{E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big]}\right|\,\varphi(Y_i) \;>\; \frac{\beta}{18L^2}\,\middle|\,\mathbf{Z}_i, Y_i\right\} \\
&\qquad\le\; P\left\{\left|\frac{1-\widetilde{\eta}_m(\mathbf{Z}_i)}{\widetilde{\psi}_m(\mathbf{Z}_i;\varphi)}\right|\cdot\Big|\widetilde{\psi}_m(\mathbf{Z}_i;\varphi) - E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big]\Big| > \frac{\varrho_0\beta}{36BL^2}\,\middle|\,\mathbf{Z}_i, Y_i\right\} \\
&\qquad\quad+\; P\left\{\Big|\widetilde{\eta}_m(\mathbf{Z}_i) - E\big[\Delta_i\big|\mathbf{Z}_i\big]\Big| > \frac{\varrho_0\beta}{36BL^2}\,\middle|\,\mathbf{Z}_i, Y_i\right\} \\
&\qquad:=\; P_{n,1}(i) + P_{n,2}(i), \qquad (61)
\end{aligned}$$

where we have used the facts that $\varphi(y) \in (0, B]$, $B > 0$, and $E\big[\Delta\varphi(Y)\big|\mathbf{Z} = \mathbf{z}\big] \geq \varrho_0$ (by assumption D). But, using standard arguments, it is not difficult to show that, under assumptions (B)–(E) and m large enough, one has

$$P_{n,2}(i) \;\leq\; C_{12}\, e^{-C_{13}\, mh^d\beta^2} \tag{62}$$

where c_{12} and c_{13} are positive constants not depending om m, ℓ, or β. Furthermore, it is also shown in the Appendix that

$$P_{n,1}(i) \;\leq\; C_{14}\, e^{-C_{15}\, mh^d\beta^2} + C_{16}\, e^{-C_{17}\, mh^d}, \tag{63}$$

where C_{16} and C_{17} are positive constant not depending on m or ℓ. Therefore, in view of (58) – (63), for every $\beta > 0$ and n large enough, we have

$$\begin{aligned} P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\varphi) - E\left[\left|\widehat{m}_m(\mathbf{X};\varphi) - Y\right|^2\middle|\mathbb{D}_m\right]\right| > \beta\right\} \;&\leq\; \ell\,\big|\mathcal{F}_\varepsilon\big|\left[C_{21}\, e^{-C_{22}mh^d} + C_{23}\, e^{-C_{24}\, mh^d\beta^2}\right] \\ &\quad +\; 2\big|\mathcal{F}_\varepsilon\big|\, e^{-\pi^2_{\min}\ell\beta^2/(162L^4)}, \end{aligned} \tag{64}$$

Next, we deal with the second term on the right side of (48). To this end, first note that since for $\varphi \in \mathcal{F}$, $E\big|m(\mathbf{X};\varphi) - Y\big|^2 = E_\varphi\big|m(\mathbf{X};\varphi) - Y\big|^2 = E_\varphi\big[\Delta\big|m(\mathbf{X};\varphi) - Y\big|^2/\pi_\varphi(\mathbf{Z},Y)\big]$, one obtains

$$\begin{aligned} \left|\widehat{L}_{m,\ell}(\varphi) - E\Big|m(\mathbf{X};\varphi) - Y\Big|^2\right| \;&\leq\; \left|\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\frac{\Delta_i\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)} - \frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\frac{\Delta_i\left|m(\mathbf{X}_i;\varphi) - Y_i\right|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)}\right| \\ &\quad + \left|\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\frac{\Delta_i\left|m(\mathbf{X}_i;\varphi) - Y_i\right|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)} - E\varphi\left[\frac{\Delta\big|m(\mathbf{X};\varphi) - Y\big|^2}{\pi_\varphi(\mathbf{Z},Y)}\right]\right| \\ &\quad + \left|\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Delta_i\left|\widehat{m}_m(\mathbf{X}_i;\varphi) - Y_i\right|^2\left[\frac{1}{\pi_\varphi(\mathbf{Z}_i,Y_i)} - \frac{1}{\widehat{\pi}_\varphi(\mathbf{Z}_i,Y_i)}\right]\right| \\ &:=\; |U_{n,1}(\varphi)| + |U_{n,2}(\varphi)| + |U_{n,3}(\varphi)|. \end{aligned} \tag{65}$$

Therefore, for every $\beta > 0$,

$$P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\varphi) - E\Big|m(\mathbf{X};\varphi) - Y\Big|^2\right| > \beta\right\} \;\leq\; \sum_{k=1}^{3} P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\big|U_{n,k}(\varphi)\big| > \frac{\beta}{3}\right\}.$$

But using assumption (A) and the simple fact that $a^2 - b^2 \leq |a-b||a+b|$, one can write

$$\begin{aligned} &P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\big|U_{n,1}(\varphi)\big| > \beta/3\right\} \\ &\leq\; P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Big[\big|\widehat{m}_m(\mathbf{X}_i;\varphi) - m(\mathbf{X}_i;\varphi)\big|\cdot\big|\widehat{m}_m(\mathbf{X}_i;\varphi) + m(\mathbf{X}_i;\varphi) - 2Y_i\big|\Big] > \frac{\beta\pi_{\min}}{3}\right\} \\ &\leq\; \big|\mathcal{F}_\varepsilon\big|\sup_{\varphi\in\mathcal{F}_\varepsilon}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell} P\left\{\big|\widehat{m}_m(\mathbf{X}_i;\varphi) - m(\mathbf{X}_i;\varphi)\big| > \frac{\beta\pi_{\min}}{15L}\right\}, \end{aligned}$$

where we have used the fact that $\big|\widehat{m}_m(\mathbf{X}_i;\varphi) + m(\mathbf{X}_i;\varphi) - 2Y_i\big| \le 5L$. Now, using standard arguments, it is not hard to show that under assumptions (B)–(E) and m large enough, one has

$$P\Big\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} |U_{n,1}(\varphi)| > \beta/3\Big\} \le \ell\,\big|\mathcal{F}_\varepsilon\big| C_{25}\,\exp\{-C_{26}\,mh^d\beta^2\}, \tag{66}$$

for positive constants C_{25} and C_{26} not depending on m, ℓ, or β. Next, since the iid random variables $\Delta_i|m(\mathbf{X}_i;\varphi) - Y_i|^2/\pi_\varphi(\mathbf{Z}_i, Y_i)$, $i \in \boldsymbol{\mathcal{I}}_\ell$, take values in $(0,\, 4L^2/\pi_{\min})$, an application of Hoeffding's inequality yields

$$P\Big\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} |U_{n,2}(\varphi)| > \beta/3\Big\} \le \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}} P_\varphi\big\{|U_{n,2}(\varphi)| > \beta/3\big\} \le 2\,\big|\mathcal{F}_\varepsilon\big| \exp\big\{ -\ell\pi_{\min}^2\beta^2/(72L^4)\big\}. \tag{67}$$

Furthermore, the same arguments that were used to deal with the term $S_n(2)$ in (58) can be employed to show that

$$P\Big\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} |U_{n,3}(\varphi)| > \beta/3\Big\} \le \ell\,\big|\mathcal{F}_\varepsilon\big| \Big[C_{27}\,\exp\big\{ - C_{28}\,mh^d\big\} + C_{29}\,\exp\big\{ - C_{30}\,mh^d\beta^2\big\}\Big], \tag{68}$$

for n large enough and positive constants C_{27} – C_{30} that do not depend on m, ℓ, or β. Putting together (65), (66), (67), and (68), one finds, for every $\beta > 0$,

$$\begin{aligned} P\left\{ \sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widehat{L}_{m,\ell}(\varphi) - E\Big|m(\mathbf{X};\varphi) - Y\Big|^2\right| > \beta\right\} \;\le\; & C_{31}\,\ell\,\big|\mathcal{F}_\varepsilon\big| \Big[e^{-C_{32}\,mh^d} + e^{-C_{33}\,mh^d\beta^2}\Big] \\ & + 2\,\big|\mathcal{F}_\varepsilon\big|\, e^{-C_{34}\,\ell\,\beta^2}, \end{aligned} \tag{69}$$

for n large enough, where C_{31} – C_{34} are positive constants not depending on m or ℓ. Now to complete the proof of the theorem, let $0 < \varepsilon_n \downarrow 0$ be as in the statement of the theorem and let φ_{ε_n} be as in (19). Then, (47) in conjunction with the arguments used in the proof of Lemma 3 (in particular (94), (95)), and the C_p-inequality (with $p = 2$), one has

$$\int \Big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\Big|^2 \mu(d\mathbf{x}) \;\le\; 2\int \Big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x};\varphi_{\varepsilon_n})\Big|^2 \mu(d\mathbf{x}) + 4LC\,\varepsilon_n\,, \tag{70}$$

where $C > 0$ is the constant in Lemma 2. Finally, observe that (70) in conjunction with Lemma 3 implies that, for every $t > 0$,

$$\begin{aligned} & P\left\{\int \Big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\Big|^2 \mu(d\mathbf{x}) > t\right\} \le P\left\{\int \Big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x};\varphi_{\varepsilon_n})\Big|^2 \mu(d\mathbf{x}) > \frac{t}{2} - 2LC\,\varepsilon_n\right\} \\ & \le\; P\left\{ \sup_{\varphi\in\mathcal{F}_{\varepsilon_n}} \left| E\left[\Big|\widehat{m}_m(\mathbf{X};\varphi) - Y\Big|^2 \Big| \mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\varphi)\right| > \frac{t/2 - 2LC\,\varepsilon_n - C_1\sqrt{\varepsilon_n}}{2}\right\} \\ & \quad + P\left\{ \sup_{\varphi\in\mathcal{F}_{\varepsilon_n}} \left|\widehat{L}_{m,\ell}(\varphi) - E\Big|m(\mathbf{X};\varphi) - Y\Big|^2\right| > \frac{t/2 - 2LC\,\varepsilon_n - C_1\sqrt{\varepsilon_n}}{2}\right\}. \end{aligned}$$

Now, since $\varepsilon_n \downarrow 0$, as $n \to \infty$, we can choose n large enough so that $t/2 - 2LC\,\varepsilon_n - C_1\sqrt{\varepsilon_n} > t/4$. Therefore, in view of (64) and (69), for every $t > 0$ and for n large enough, one finds

$$P\left\{\int \left|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\right|^2 \mu(d\mathbf{x}) > t\right\} \;\leq\; 4\,\big|\mathcal{F}_{\varepsilon_n}\big|\, e^{-c_{35}\ell t^2} + c_{36}\,\ell\big|\mathcal{F}_{\varepsilon_n}\big|\, e^{-c_{37}mh^d} + c_{38}\,\ell\big|\mathcal{F}_{\varepsilon_n}\big|\, e^{-c_{39}mh^d t^2},$$

which completes the proof of Theorem 2.

□

PROOF OF COROLLARY 1

Corollary 1 follows from an application of the Borel-Cantelli lemma in conjunction with (22), the bound in Theorem 2, and Remark 2.

□

PROOF OF THEOREM 3

We first note that by Remark 2 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

$$\begin{aligned}
E\big|\widehat{m}(\mathbf{X};\widehat{\varphi}_n) - m(\mathbf{X})\big|^2 &= E\left[\int_{\mathbb{R}^d} \left|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\right|^2 \mu(d\mathbf{x}\right] \\
&= \int_0^\infty P\left\{\int_{\mathbb{R}^d} \left|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\right|^2 \mu(d\mathbf{x}) > t\right\} dt \\
&= \int_0^{9L^2} P\left\{\int_{\mathbb{R}^d} \left|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\right|^2 \mu(d\mathbf{x}) > t\right\} dt\,, \qquad (71)
\end{aligned}$$

where the last line follows because by the definition of the estimator $\widehat{m}(\mathbf{x};\widehat{\varphi}_n)$ in (18), one has

$$\begin{aligned}
\left|\widehat{m}(\mathbf{x};\widehat{\varphi}_n) - m(\mathbf{x})\right|^2 &\leq \big(\big|\widehat{m}(\mathbf{x};\widehat{\varphi}_n)\big| + \big|m(\mathbf{x})\big|\big)^2 \\
&\leq \left(\big|\widehat{\eta}_{m,1}(\mathbf{x})\big| + \left|\frac{\widehat{\psi}_{m,1}(\mathbf{x};\widehat{\varphi}_n)}{\widehat{\psi}_{m,2}(\mathbf{x};\widehat{\varphi}_n)}\right| \cdot \big|1 - \widehat{\eta}_{m,2}(\mathbf{x})\big| + L\right)^2 \;\leq\; (L + L\cdot 1 + L)^2.
\end{aligned}$$

Therefore, by Theorem 2, for n large enough, we have

$$\begin{aligned}
&\big(\text{right side of (71)}\big) \\
&\leq \int_0^u dt \;+\; \big(c_4 \vee c_6\big)\big|\mathcal{F}_{\varepsilon_n}\big| \cdot \left[\int_u^{9L^2} e^{-c_5\ell t^2}\, dt \;+\; \ell \int_u^{9L^2} e^{-c_8\, mh^d t^2}\, dt \;+\; \ell\, e^{-c_7\, mh^d} \int_u^{9L^2} dt\right], \\
&\qquad \text{(where } c_4\text{–}c_8 \text{ are as in Theorem 2)} \\
&\leq u \;+\; 2(c_4 \vee c_6)\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell \int_u^{9L^2} e^{-(c_5\wedge c_8)(\ell \wedge mh^d)\, t^2}\, dt \;+ (c_4 \vee c_6)(9L^2)\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell\, e^{-c_7\, mh^d} \\
&\leq u \;+\; \frac{2(c_4 \vee c_6)\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell}{\sqrt{(c_5 \wedge c_8)(\ell \wedge mh^d)}} \cdot \int_{u\sqrt{(c_5\wedge c_8)(\ell\wedge mh^d)}}^\infty e^{-v^2/2}\, dv \;+\; (c_4 \vee c_6)(9L^2)\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell\, e^{-c_7\, mh^d}
\end{aligned}$$

(which follows from the change of variable $v = \sqrt{(c_5 \wedge c_8)(\ell \wedge mh^d)}\, t$

$$\leq\; u \,+\, \frac{2(c_4 \vee c_6)\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell}{\sqrt{(c_5 \wedge c_8)(\ell \wedge mh^d)}} \cdot \frac{e^{-(c_5 \wedge c_8)(\ell \wedge mh^d)\,u^2/2}}{\sqrt{(c_5 \wedge c_8)(\ell \wedge mh^d)}\;u} \,+\, (c_4 \vee c_6)(9L^2)\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell\, e^{-c_7\, mh^d}, \quad (72)$$

where the last line follows from the upper bound on Mill's ratio; see, for example, Mitrinovic (1970; p. 177). Now, put

$$c = 2(c_4 \vee c_6)\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell \quad \text{and} \quad N = (c_5 \wedge c_8)(\ell \wedge mh^d)/4$$

and observe that the right side of (72) can be written as

$$u + \frac{c}{4Nu}\, e^{-2Nu^2} + (c_4 \vee c_6)(9L^2)\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell\, e^{-c_7\, mh^d}. \quad (73)$$

But the term $u + \frac{c}{4Nu}\, e^{-2Nu^2}$ in (73) is approximately minimized by taking $u = \sqrt{\log(c)/(2N)}$, and the corresponding minimum value of (73) is

$$\begin{aligned}
&\sqrt{\frac{\log(c)}{2N}} + \sqrt{\frac{1}{8N\log(c)}} + (c_4 \vee c_6)(9L^2)\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell\, e^{-c_7\, mh^d} \\
&= \sqrt{\frac{c_{41} + \log\ell + \log|\mathcal{F}_{\varepsilon_n}|}{c_{42}\,(\ell \wedge mh^d)}} + \sqrt{\frac{1}{c_{43}\,(\ell \wedge mh^d)\big[c_{41} + \log\ell + \log|\mathcal{F}_{\varepsilon_n}|\big]}} + c_{44}\,\big|\mathcal{F}_{\varepsilon_n}\big|\,\ell\, e^{-c_7\, mh^d},
\end{aligned}$$

where c_{41}–c_{44} are positive constants not depending on m, ℓ, and n.

□

PROOF OF THEOREM 4

Let $\widehat{m}_m^{\mathrm{HT}}(\mathbf{x}; \widetilde{\pi}_\varphi)$, $m(\mathbf{x}, \pi_{\varphi^*})$, and φ_ε be as in (31), (25), and (46) respectively. Also, define

$$\widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi) = \ell^{-1} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i}{\widetilde{\pi}_\varphi(\mathbf{Z}_i, Y_i)} \Big|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - Y_i\Big|^2, \quad (74)$$

where $\widetilde{\pi}_\varphi(\mathbf{x}, y)$ is given by (27), and put

$$\widetilde{\varphi}_\varepsilon = \operatorname*{argmin}_{\varphi \in \mathcal{F}_\varepsilon} \widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi).$$

Then, using the arguments that led to (90) and (91), yield

$$\begin{aligned}
&\int \Big|\widehat{m}_m^{\mathrm{HT}}(\mathbf{x}; \widetilde{\pi}_{\widetilde{\varphi}_\varepsilon}) - m(\mathbf{x}; \pi_{\varphi_\varepsilon})\Big|^2 \mu(d\mathbf{x}) \\
&\leq \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| E\left[\Big|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}; \widetilde{\pi}_\varphi) - Y\Big|^2 \Big| \mathbb{D}_m\right] - \widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi)\right| + \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| \widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi) - E\Big|m(\mathbf{X}; \pi_\varphi) - Y\Big|^2\right| \\
&\quad + \; 2\, E\left[\Big|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}; \widetilde{\pi}_{\widetilde{\varphi}_\varepsilon}) - m(\mathbf{X}; \pi_{\varphi_\varepsilon})\Big| \cdot \Big|m(\mathbf{X}; \pi_{\varphi_\varepsilon}) - m(\mathbf{X}; \pi_{\varphi^*})\Big| \Big| \mathbb{D}_n\right] \quad (75)
\end{aligned}$$

where, as before, φ^* is the true φ. But by Cauchy-Schwarz inequality, the last line on the right side of (75) is bounded by

$$2\sqrt{\int \left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_\varepsilon}) - m(\mathbf{x};\pi_{\varphi_\varepsilon})\right|^2 \mu(d\mathbf{x})} \cdot \sqrt{E\left|m(\mathbf{X};\pi_{\varphi_\varepsilon}) - m(\mathbf{X};\pi_{\varphi^*})\right|^2}$$

$$\leq \; C_3 \sqrt{\int \left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_\varepsilon}) - m(\mathbf{x};\pi_{\varphi_\varepsilon})\right|^2 \mu(d\mathbf{x})} \cdot \sqrt{\varepsilon}\,, \tag{76}$$

where (76) follows from arguments similar to those used to arrive at (94) and (95); here C_3 is a positive constant not depending on n or ε. Therefore, in view of (75) and (76), for any $t > 0$

$$P\left\{\int \left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_\varepsilon}) - m(\mathbf{x};\pi_{\varphi_\varepsilon})\right|^2 \mu(d\mathbf{x}) > t\right\} - P\left\{\int \left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_\varepsilon}) - m(\mathbf{x};\pi_{\varphi_\varepsilon})\right|^2 \mu(d\mathbf{x}) > \frac{t^2}{c_4\varepsilon}\right\}$$

$$\leq \; P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon} \left|E\left[\left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X};\widetilde{\pi}_\varphi) - Y\right|^2 \middle| \mathbb{D}_m\right] - \widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi)\right| > \frac{t}{3}\right\}$$

$$+ P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi) - E\left|m(\mathbf{X};\pi_\varphi) - Y\right|^2\right| > \frac{t}{3}\right\}, \tag{77}$$

where $c_4 = (3C_3)^2$ with C_3 as in (76). Therefore, for every constant $\beta > 0$

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

$$\leq \; P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i\left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi) - Y_i\right|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)} - E\left[\left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X};\widetilde{\pi}_\varphi) - Y\right|^2 \middle| \mathbb{D}_m\right]\right| > \frac{\beta}{2}\right\}$$

$$+ \; P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon} \left|\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \Delta_i\left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi) - Y_i\right|^2 \left[\Big(\pi_\varphi(\mathbf{Z}_i,Y_i)\Big)^{-1} - \Big(\widetilde{\pi}_\varphi(\mathbf{Z}_i,Y_i)\Big)^{-1}\right]\right| > \frac{\beta}{2}\right\}$$

$$:= \; T_n(1) + T_n(2). \tag{78}$$

However, for every $i \in \boldsymbol{\mathcal{I}}_\ell$ and $\varphi \in \mathcal{F}$, one finds that $E_\varphi\big[\Delta_i|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi) - Y_i|^2/\pi_\varphi(\mathbf{Z}_i,Y_i)\,\big|\mathbb{D}_m\big] = E_\varphi\big[E_\varphi\big\{\Delta_i|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi) - Y_i|^2/\pi_\varphi(\mathbf{Z}_i,Y_i)\,\big|\mathbb{D}_m,\mathbf{X}_i,Y_i\big\}\big|\mathbb{D}_m\big] = E_\varphi\big[|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi) - Y_i|^2\,\big|\mathbb{D}_m\big]$, which follows from the definition of π_φ in (13). Moreover, by the definition of $\widetilde{\pi}_\varphi(\mathbf{Z},Y)$ in (27), one finds

$$|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi)| \;\leq\; \max_{k\in\boldsymbol{\mathcal{I}}_m}\left|\Delta_k Y_k/\widetilde{\pi}_\varphi(\mathbf{Z}_k,Y_k)\right| \;\leq\; L\cdot\left(1 + \max_{k\in\boldsymbol{\mathcal{I}}_m}\left|\frac{1}{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)}\right|\cdot B\right), \tag{79}$$

where the function $\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)$ is as given by (28). Consequently, conditional on $\mathbb{D}_m$, the terms $\big[\Delta_i|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi) - Y_i|^2\big]/\pi_\varphi(\mathbf{Z}_i,Y_i)$, $i \in \boldsymbol{\mathcal{I}}_\ell$, are independent nonnegative random variables bounded by $2L^2\big\{4 + B^2\max_{k\in\boldsymbol{\mathcal{I}}_m}^2\big|1/\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)\big|\big\}/\pi_{\min}$. Therefore, using the arguments that lead to (59), the term $T_n(1)$ in (78) can be handled as follows

$$T_n(1) \leq \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}_\varepsilon} P\left\{\left|\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i\left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi) - Y_i\right|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)} - E\left[\left|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X};\widetilde{\pi}_\varphi) - Y\right|^2 \middle| \mathbb{D}_m\right]\right| > \frac{\beta}{2}\right\}$$

$$\leq \big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}} E_\varphi\Bigg[P_\varphi\Bigg\{\Bigg|\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\frac{\Delta_i\big|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi)-Y_i\big|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)} - E_\varphi\Big[\Big|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X};\widetilde{\pi}_\varphi)-Y\Big|^2\Big|\mathbb{D}_m\Big]\Bigg| > \frac{\beta}{2}\Big|\mathbb{D}_m\Bigg\}\Bigg]$$

$$\leq 2\big|\mathcal{F}_\varepsilon\big| \sup_{\varphi\in\mathcal{F}} E_\varphi\Bigg[\exp\Bigg\{\frac{-2\ell^2\,(\beta/2)^2}{2L^2\pi_{\min}^{-1}\Big\{4+B^2\max_{k\in\boldsymbol{\mathcal{I}}_m}^2\big|1/\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)\big|\Big\}}\Bigg\}\Bigg], \tag{80}$$

via Hoeffding's inequality. Now let ϱ_0 be the constant in assumption (D) and observe that since the exponential function in (80) is always bounded by 1, the expectation on the right side of (80) is bounded by

$$\begin{aligned}
&E_\varphi\Bigg[\exp\Bigg\{\frac{-2\ell^2\,(\beta/2)^2}{2L^2\pi_{\min}^{-1}\Big\{4+\max_{k\in\boldsymbol{\mathcal{I}}_m}^2\big|B/\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)\big|\Big\}}\Bigg\}\,\mathbb{I}\Big\{\bigcap_{k\in\boldsymbol{\mathcal{I}}_m}\Big[\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)\geq\frac{\varrho_0}{2}\Big]\Big\}\Bigg]\\
&\quad + E_\varphi\Bigg[\mathbb{I}\Big\{\bigcup_{k\in\boldsymbol{\mathcal{I}}_m}\Big[\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)<\varrho_0/2\Big]\Big\}\Bigg]\\
&\leq \exp\Bigg\{\frac{-2\ell^2\,(\beta/2)^2}{2L^2\pi_{\min}^{-1}\{4+B^2(2/\varrho_0)^2\}}\Bigg\}\cdot P_\varphi\Big\{\bigcap_{k\in\boldsymbol{\mathcal{I}}_m}\Big[\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)\geq\varrho_0/2\Big]\Big\}\\
&\quad + \sum_{k\in\boldsymbol{\mathcal{I}}_m} P_\varphi\Big\{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)<\varrho_0/2\Big\}\\
&\leq \exp\Bigg\{\frac{-2\ell^2\,(\beta/2)^2}{2L^2\pi_{\min}^{-1}\{4+B^2(2/\varrho_0)^2\}}\Bigg\} + \sum_{k\in\boldsymbol{\mathcal{I}}_m} P_\varphi\Big\{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)<\varrho_0/2\Big\}.
\end{aligned}\tag{81}$$

If we put $\psi(\mathbf{Z}_k;\varphi) := E_\varphi\big[\Delta_k\varphi(Y_k)\big|\mathbf{Z}_k\big]$, then we find $P_\varphi\big\{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)<\varrho_0/2\,\big|\mathbf{Z}_k\big\}\leq P_\varphi\big\{-\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)+\psi(\mathbf{Z}_k;\varphi)>\varrho_0-\varrho_0/2\,\big|\mathbf{Z}_k\big\}\leq P_\varphi\big\{\big|\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)-\psi(\mathbf{Z}_k;\varphi)\big|>\varrho_0/2\,\big|\mathbf{Z}_k\big\}\leq C_{16}\exp\{-C_{17}mh^d\}$, for n large enough and positive constants C_{16} and C_{17} not depending on n or φ. Thus, by (80) and (81),

$$T_n(1) \;\leq\; 2\big|\mathcal{F}_\varepsilon\big|\left(\exp\Bigg\{\frac{-2\ell^2\,(\beta/2)^2}{2L^2\pi_{\min}^{-1}\{4+B^2(2/\varrho_0)^2\}}\Bigg\} + C_{16}\,m\,\exp\Big\{-C_{17}mh^d\Big\}\right). \tag{82}$$

As for the term $T_n(2)$ that appears in (78), one can use the fact that $|\widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i;\widetilde{\pi}_\varphi)-Y_i|^2\leq 2L^2\big\{4+B^2\max_{k\in\boldsymbol{\mathcal{I}}_m}^2\big|1/\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)\big|\big\}$ to write

$$\begin{aligned}
&T_n(2)\\
&\leq \big|\mathcal{F}_\varepsilon\big|\sup_{\varphi\in\mathcal{F}_\varepsilon}\Bigg(P\Bigg\{\Bigg[\frac{2L^2}{\ell}\Big\{4+\Big(\max_{k\in\boldsymbol{\mathcal{I}}_m}\Big|B/\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)\Big|\Big)^2\Big\}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Big|\frac{1}{\pi_\varphi(\mathbf{Z}_i,Y_i)}-\frac{1}{\widetilde{\pi}_\varphi(\mathbf{Z}_i,Y_i)}\Big|>\frac{\beta}{2}\Bigg]\\
&\qquad\qquad\cap\Bigg[\bigcap_{k\in\boldsymbol{\mathcal{I}}_m}\Big\{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)\geq\varrho_0/2\Big\}\Bigg]\Bigg\}+\sum_{k\in\boldsymbol{\mathcal{I}}_m}P\Big\{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)<\varrho_0/2\Big\}\Bigg)\\
&\leq \big|\mathcal{F}_\varepsilon\big|\sup_{\varphi\in\mathcal{F}_\varepsilon}\Bigg(P\Bigg\{\ell^{-1}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\Big|\frac{1}{\pi_\varphi(\mathbf{Z}_i,Y_i)}-\frac{1}{\widetilde{\pi}_\varphi(\mathbf{Z}_i,Y_i)}\Big|>\frac{\beta}{4L^2[4+B^2(2/\varrho_0)^2]}\Bigg\}
\end{aligned}$$

$$+ \sum_{k \in \boldsymbol{\mathcal{I}}_m} P\Big\{ \widetilde{\psi}_m(\mathbf{Z}_k; \varphi) < \varrho_0/2 \Big\} \Bigg) .$$

Employing the arguments that were used in (60), (61), (62), and (63), one arrives at

$$T_n(2) \;\le\; |\mathcal{F}_\varepsilon| \left(C_{46}\, \ell \exp\Big\{ -C_{47}\, m h^d \beta^2 \Big\} + C_{48}\, \ell \exp\Big\{ -C_{49} m h^d \Big\} + C_{50}\, \ell \exp\Big\{ -C_{51} m h^d \Big\} \right). \quad (83)$$

Now, putting together (78), (82), and (83), we find that for every $\beta > 0$ and n large enough (and thus m and ℓ),

$$\begin{aligned} & P\left\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| \widetilde{L}_{m,\ell}(\varphi) - E\left[\left| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}; \widetilde{\pi}_\varphi) - Y \right|^2 \middle| \mathbb{D}_m \right] \right| > \beta \right\} \\ & \quad \le\; |\mathcal{F}_\varepsilon| \left(\exp\left\{ -C_{52}\, \ell^2 \beta^2 \right\} + C_{46}\, \ell \exp\Big\{ -C_{47}\, m h^d \beta^2 \Big\} + C_{53}\, (\ell \vee m) \exp\Big\{ -C_{54}\, m h^d \Big\} \right). \quad (84) \end{aligned}$$

To wrap up the proof, we also need to deal with the last probability statement on the right side of (77). To that end, it is shown in the Appendix that for every $\beta > 0$

$$\begin{aligned} P\left\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| \widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi) - E\Big| m(\mathbf{X}; \pi_\varphi) - Y \Big|^2 \right| > \beta \right\} \;\; &\le \;\; |\mathcal{F}_\varepsilon| \Big(C_{50} \ell m\, e^{-C_{51} m h^d \beta^2} + C_{52} \ell m\, e^{-C_{53} m h^d} \\ & \qquad + \; 2\, e^{-C_{55} \ell \beta^2} \Big) \qquad (85) \end{aligned}$$

for positive constants C_{55}–C_{59} not depending on m, ℓ, or β. Now, for any decreasing sequence $0 < \varepsilon_n \downarrow 0$, let φ_{ε_n} be as in (19). Then, employing arguments similar to those used in the proof of Lemma 3 (in particular those used to arrive at (94) and (95)), we find

$$\begin{aligned} \int \left| \widehat{m}^{\mathrm{HT}}(\mathbf{x}; \widetilde{\varphi}_n) - m(\mathbf{x}) \right|^2 \mu(d\mathbf{x}) \;\; &= \;\; \int \left| \widehat{m}^{\mathrm{HT}}(\mathbf{x}; \widetilde{\varphi}_n) - m(\mathbf{x}; \varphi_{\varepsilon_n}) + m(\mathbf{x}; \varphi_{\varepsilon_n}) - m(\mathbf{x}) \right|^2 \mu(d\mathbf{x}) \\ & \le \;\; 2 \int \left| \widehat{m}^{\mathrm{HT}}(\mathbf{x}; \widetilde{\varphi}_n) - m(\mathbf{x}; \varphi_{\varepsilon_n}) \right|^2 \mu(d\mathbf{x}) \;+\; 4LC\, \varepsilon_n \,, \qquad (86) \end{aligned}$$

where $C > 0$ is the constant in Lemma 2. Therefore, in view of (86) and (77), for every constant $t > 0$ we have

$$\begin{aligned} & \frac{1}{2}\, P\left\{ \int \left| \widehat{m}^{\mathrm{HT}}(\mathbf{x}; \widetilde{\pi}_{\widetilde{\varphi}_n}) - m(\mathbf{x}) \right|^2 \mu(d\mathbf{x}) \;>\; t \right\} \\ & \quad \le \;\; \frac{1}{2}\, P\left\{ \int \left| \widehat{m}^{\mathrm{HT}}(\mathbf{x}; \widetilde{\pi}_{\widetilde{\varphi}_n}) - m(\mathbf{x}; \varphi_{\varepsilon_n}) \right|^2 \mu(d\mathbf{x}) > \; t/2 - 2LC\varepsilon_n \right\} \\ & \quad \le \;\; P\left\{ \int \left| \widehat{m}^{\mathrm{HT}}(\mathbf{x}; \widetilde{\pi}_{\widetilde{\varphi}_n}) - m(\mathbf{x}; \varphi_{\varepsilon_n}) \right|^2 \mu(d\mathbf{x}) > \; t/2 - 2LC\varepsilon_n \right\} \\ & \qquad\quad - P\left\{ \int \left| \widehat{m}^{\mathrm{HT}}(\mathbf{x}; \widetilde{\pi}_{\widetilde{\varphi}_n}) - m(\mathbf{x}; \varphi_{\varepsilon_n}) \right|^2 \mu(d\mathbf{x}) > \; (t/2 - 2LC\varepsilon_n)^2/(c_4\, \varepsilon_n^2) \right\} \\ & \qquad\quad \text{(for } n \text{ large enough, where } c_4 > 0 \text{ is as in the first line of (77))} \\ & \quad \le \;\; P\left\{ \sup_{\varphi \in \mathcal{F}_{\varepsilon_n}} \left| E\left[\left| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}; \widetilde{\pi}_\varphi) - Y \right|^2 \middle| \mathbb{D}_m \right] - \widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi) \right| > \frac{t/2 - 2LC\varepsilon_n}{3} \right\} \end{aligned}$$

$$+ P\left\{\sup_{\varphi\in\mathcal{F}_{\varepsilon_n}}\left|\widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi) - E\Big|m(\mathbf{X};\pi_\varphi) - Y\Big|^2\right| > \frac{t/2 - 2LC\varepsilon_n}{3}\right\}.$$

Finally, choosing n large enough so that $(t/2-2LC\varepsilon_n)/3 > t/12$, and using the bounds in (84) and (85), we find

$$P\left\{\int\left|\widehat{m}^{\mathrm{HT}}(\mathbf{x};\widetilde{\pi}_{\widetilde{\varphi}_n}) - m(\mathbf{x})\right|^2\mu(d\mathbf{x}) > t\right\} \;\leq\; |\mathcal{F}_{\varepsilon_n}|\Big(C_{65}\,e^{-C_{66}\ell t^2} + C_{67}e^{-C_{68}\ell^2t^2} + C_{69}\ell\,e^{-C_{70}mh^dt^2} + C_{71}\,\ell m\,e^{-C_{72}mh^d(t^2\vee 1)} + C_{73}(\ell\vee m)e^{-C_{74}mh^d}\Big),$$

for n large enough where C_{65}–C_{74} 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) of the theorem is virtually the same and, in fact, easier and therefore will not be given.

□

PROOF OF COROLLARY 3

The corollary follows from the Borel-Cantelli lemma in conjunction with (34), the bound in Theorem 4, and Remark 3.

□

PROOF OF THEOREM 5

The proof of this theorem is similar to that of Theorem 3 and therefore will not be given.

□

PROOF OF THEOREM 6

Part (i).

By (39), we have

$$P\left\{\widehat{g}_n(\mathbf{X};\widehat{\varphi}_n)\neq Y\Big|\mathbb{D}_n\right\} - P\{g_{\mathrm{B}}(\mathbf{X})\neq Y\} \;\leq\; 2E\left[\left|\widehat{m}(\mathbf{X};\widehat{\varphi}_n) - m(X)\right|\Big|\mathbb{D}_n\right]. \tag{87}$$

Now, Part (i) of the theorem follows from (87) and Corollary 1 in conjunction with the Cauchy-Schwarz inequality.

Part (ii).

Taking the expectation of both sides of (87), the result follows from Corollary 2 together with the Cauchy-Schwarz inequality.

Part (iii).

By a result of Audibert and Tsybakov (2007; Lemma 5.2), under the margin assumption (H), we have

$$P\left\{\widehat{g}_n(\mathbf{X};\widehat{\varphi}_n)\neq Y\right\}-P\{g_{\mathrm{B}}(\mathbf{X})\neq Y\}\;\leq\;\left(E\Big|\widehat{m}(\mathbf{X};\widehat{\varphi}_n)-m(X)\Big|^2\right)^{\frac{1+\alpha}{2+\alpha}}, \tag{88}$$

where α is as in (41). The result now follows from Corollary 2.

□

PROOF OF THEOREM 7

The proof uses Corollaries 3 and 4 and is virtually the same as that of Theorem 7, and thus will not be given.

□

Appendix: auxiliary proofs

PROOF OF LEMMA 1.

Let $\pi_\varphi(\mathbf{x},y)$ be as in (8) and note that

$$1-\pi_{\varphi^*}(\mathbf{X},Y)=\frac{\exp\{g(\mathbf{X})\}\,\varphi^*(Y)}{1+\exp\{g(\mathbf{X})\}\,\varphi^*(Y)}=\exp\{g(\mathbf{X})\}\,\varphi^*(Y)\,\pi_{\varphi^*}(\mathbf{X},Y). \tag{89}$$

Now, writing $m(\mathbf{x})=E[Y|\mathbf{X}=\mathbf{x}]=E[Y\Delta|\mathbf{X}=\mathbf{x}]+\frac{E[Y(1-\Delta)|\mathbf{X}=\mathbf{x}]}{E[1-\Delta|\mathbf{X}=\mathbf{x}]}\cdot E[1-\Delta|\mathbf{X}=\mathbf{x}]$, one finds

$$\begin{aligned}\frac{E[Y(1-\Delta)|\mathbf{X}]}{E[1-\Delta|\mathbf{X}]} &= \frac{E\big[E\big\{Y(1-\Delta)|\mathbf{X},Y\big\}\big|\mathbf{X}\big]}{E\big[E\big\{1-\Delta|\mathbf{X},Y\big\}\big|\mathbf{X}\big]}=\frac{E\big[Y\big\{1-\pi_{\varphi^*}(\mathbf{X},Y)\big\}\big|\mathbf{X}\big]}{E\big[1-\pi_{\varphi^*}(\mathbf{X},Y)\big|\mathbf{X}\big]}\\ &\overset{\text{by (89)}}{=} \frac{E\big[Y\exp\{g(\mathbf{X})\}\,\varphi^*(Y)\,\pi_{\varphi^*}(\mathbf{X},Y)\big|\mathbf{X}\big]}{E\big[\exp\{g(\mathbf{X})\}\,\varphi^*(Y)\,\pi_{\varphi^*}(\mathbf{X},Y)\big|\mathbf{X}\big]}=\frac{E\big[Y\varphi^*(Y)\,\Delta\big|\mathbf{X}\big]}{E\big[\varphi^*(Y)\,\Delta\big|\mathbf{X}\big]}=\frac{\psi_1(\mathbf{X};\varphi^*)}{\psi_2(\mathbf{X};\varphi^*)}.\end{aligned}$$

The proof of the lemma now follows from this and the definitions of ψ_k and η_k, $k=1,2,$ in (43).

□

PROOF OF LEMMA 2.

Let $\psi_k(\mathbf{x};\varphi)$, $k=1,2$, be as in (43) and observe that

$$\begin{aligned}\Big|m(\mathbf{x};\varphi_1)-m(\mathbf{x};\varphi_2)\Big| &= \left|\frac{-\psi_1(\mathbf{x};\varphi_1)}{\psi_2(\mathbf{x};\varphi_1)}\cdot\frac{\psi_2(\mathbf{x};\varphi_1)-\psi_2(\mathbf{x};\varphi_2)}{\psi_2(\mathbf{x};\varphi_2)}+\frac{\psi_1(\mathbf{x};\varphi_1)-\psi_1(\mathbf{x};\varphi_2)}{\psi_2(\mathbf{x};\varphi_2)}\right|\\ &\quad\times E\big[1-\Delta|\mathbf{X}=\mathbf{x}\big]\\ &\leq \frac{1}{\psi_2(\mathbf{x};\varphi_2)}\big\{L\big|\psi_2(\mathbf{x};\varphi_1)-\psi_2(\mathbf{x};\varphi_2)\big|+\big|\psi_1(\mathbf{x};\varphi_1)-\psi_1(\mathbf{x};\varphi_2)\big|\big\},\end{aligned}$$

where we used the fact $|\psi_1(\mathbf{x};\varphi_1)|/|\psi_2(\mathbf{x};\varphi_1)|\leq L|\psi_2(\mathbf{x};\varphi_1)|/|\psi_2(\mathbf{x};\varphi_1)|=L$ (because $\varphi_k>0$). But, since $|Y\Delta|\leq L$, one finds $|\psi_1(\mathbf{x};\varphi_1)-\psi_1(\mathbf{x};\varphi_2)|\leq E\left[|\Delta Y|\cdot\big|\varphi_1(Y)-\varphi_2(Y)\big|\,\big|\mathbf{X}=\mathbf{x}\right]\leq$

$L\sup_{-L\le y\le L}\big|\varphi_1(y)-\varphi_2(y)\big|$. Similarly, $|\psi_2(\mathbf{x};\varphi_1)-\psi_2(\mathbf{x};\varphi_2)|\le \sup_{-L\le y\le L}\big|\varphi_1(y)-\varphi_2(y)\big|$. On the other hand, by assumption (D), we have $\psi_2(\mathbf{x};\varphi_2)\ge \varrho_0>0$, for μ–a.e. $\mathbf{x}$. Therefore

$$\Big|m(\mathbf{x};\varphi_1)-m(\mathbf{x};\varphi_2)\Big| \le (2L/\varrho_0)\sup_{-L\le y\le L}\big|\varphi_1(y)-\varphi_2(y)\big|.$$

The lemma follows now by integrating both sides of this inequality with respect to $\mu(d\mathbf{x})$.

□

PROOF OF LEMMA 3.

Observe that $E\big[|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)-Y|^2\big|\mathbb{D}_n\big] = E\big[|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)-m(\mathbf{X};\varphi_\varepsilon)|^2\big|\mathbb{D}_n\big]+E|m(\mathbf{X};\varphi_\varepsilon)-Y|^2$ $+2E\big[\big(\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)-m(\mathbf{X};\varphi_\varepsilon)\big)\big(m(\mathbf{X};\varphi_\varepsilon)-Y\big)\big|\mathbb{D}_n\big]$. Also, let φ^* be as in (20) and note that

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

where we have used the fact that in view of (47), $E[Y|\mathbf{X}=\mathbf{x}]:=m(\mathbf{x})=m(\mathbf{x};\varphi^*)$. Therefore

$$\begin{aligned}
&E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)-m(\mathbf{X};\varphi_\varepsilon)\Big|^2\Big|\mathbb{D}_n\right]\\
&=\ \left\{E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)-Y\Big|^2\Big|\mathbb{D}_n\right]-E\Big|m(\mathbf{X};\varphi_\varepsilon)-Y\Big|^2\right\}\\
&\quad-2E\left[\Big(\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)-m(\mathbf{X};\varphi_\varepsilon)\Big)\Big(m(\mathbf{X};\varphi_\varepsilon)-m(\mathbf{X};\varphi^*)\Big)\Big|\mathbb{D}_n\right]\\
&:=\mathbf{I}_n+\mathbf{II}_n.
\end{aligned}\tag{90}$$

Now, observe that

$$\begin{aligned}
\mathbf{I}_n &= E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)-Y\Big|^2\Big|\mathbb{D}_n\right]-\inf_{\varphi\in\mathcal{F}_\varepsilon}E\Big|m(\mathbf{X};\varphi)-Y\Big|^2\\
&= \sup_{\varphi\in\mathcal{F}_\varepsilon}\left\{E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)-Y\Big|^2\Big|\mathbb{D}_n\right]-\widehat{L}_{m,\ell}(\varphi)+\widehat{L}_{m,\ell}(\varphi)-\widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon)\right.\\
&\qquad\left.+\ \widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon)-E\Big|m(\mathbf{X};\varphi)-Y\Big|^2\right\},\qquad\text{(where }\widehat{L}_{m,\ell}(\varphi)\text{ is as in (45))}\\
&\le \left(E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)-Y\Big|^2\Big|\mathbb{D}_n\right]-\widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon)\right)+\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\varphi)-E\Big|m(\mathbf{X};\varphi)-Y\Big|^2\right|,
\end{aligned}$$

where the last line follows since $\widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon)\le\widehat{L}_{m,\ell}(\varphi)$ holds for all $\varphi\in\mathcal{F}_\varepsilon$ (because of the definition of $\widehat{\varphi}_\varepsilon$ in (46)). Therefore,

$$|\mathbf{I}_n|\le\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|E\left[\Big|\widehat{m}_m(\mathbf{X};\varphi)-Y\Big|^2\Big|\mathbb{D}_m\right]-\widehat{L}_{m,\ell}(\varphi)\right|+\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\varphi)-E\Big|m(\mathbf{X};\varphi)-Y\Big|^2\right|,\tag{91}$$

where 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 the definitions of $\widehat{m}_m(\mathbf{x};\varphi)$, $\widehat{\psi}_{m,k}$, and $\widehat{\eta}_{m,k}$ in (9), (10), and (11), respectively, and the fact that $|m(\mathbf{X};\varphi)| \le |Y| \le L$ holds for each φ, one finds

$$\big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon)\big| \le |\widehat{\eta}_{m,1}(\mathbf{X})| + \big(|\widehat{\psi}_{m,1}(\mathbf{X};\widehat{\varphi}_\varepsilon)|/|\widehat{\psi}_{m,2}(\mathbf{X};\widehat{\varphi}_\varepsilon)|\big)\cdot\big|1-\widehat{\eta}_{m,2}(\mathbf{X})\big| \le L + L\cdot 1 = 2L. \quad (92)$$

Therefore, one can bound $\mathbf{II}_n$ in (90) by

$$\begin{aligned}\big|\mathbf{II}_n\big| &\le 2E\left[\Big|\widehat{m}_m(\mathbf{X};\widehat{\varphi}_\varepsilon) - m(\mathbf{X};\varphi_\varepsilon)\Big|\cdot\Big|m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*)\Big|\Big|\mathbb{D}_n\right]\\ &\le 6L\cdot E\Big|m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*)\Big| \le 6L\sqrt{E\big|m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*)\big|^2}, \quad (93)\end{aligned}$$

via Cauchy-Schwarz inequality. Next, consider the identity $E\big|m(\mathbf{X};\varphi_\varepsilon)-Y\big|^2 = E\big|m(\mathbf{X};\varphi^*)-Y\big|^2 + E\big|m(\mathbf{X};\varphi_\varepsilon)-m(\mathbf{X};\varphi^*)\big|^2$, which holds because $E\big[E\big\{\big(m(\mathbf{X};\varphi_\varepsilon)-m(\mathbf{X};\varphi^*)\big)\big(m(\mathbf{X};\varphi^*)-Y\big)\big|\mathbf{X}\big\}\big] = E\big[\big(m(\mathbf{X};\varphi_\varepsilon)-m(\mathbf{X};\varphi^*)\big)E\big\{m(\mathbf{X};\varphi^*)-Y\big|\mathbf{X}\big\}\big] = 0$ (since $E(Y|\mathbf{X}) = m(\mathbf{X}) = m(\mathbf{X};\varphi^*)$). Using this identity, one finds

$$\begin{aligned}E\big|m(\mathbf{X};\varphi_\varepsilon) - m(\mathbf{X};\varphi^*)\big|^2 &= \inf_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\varphi) - Y\big|^2 - E\big|m(\mathbf{X};\varphi^*) - Y\big|^2\\ &= \inf_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\varphi) - m(\mathbf{X};\varphi^*)\big|^2\\ &\le 2L\inf_{\varphi\in\mathcal{F}_\varepsilon} E\big|m(\mathbf{X};\varphi) - m(\mathbf{X};\varphi^*)\big|, \quad (94)\end{aligned}$$

where the last line in (94) follows because $\big|m(\mathbf{X};\varphi) - m(\mathbf{X};\varphi^*)\big|^2 \le \big|m(\mathbf{X};\varphi) - m(\mathbf{X};\varphi^*)\big| \times (2L)$. 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, by Lemma 2 and the fact that the right side of (94) is an infimum, one finds

$$\begin{aligned}\text{(Right side of (94))} &\le 2L\cdot E\big|m(\mathbf{X};\varphi^\dagger) - m(\mathbf{X};\varphi^*)\big| \le 2LC\sup_{-L\le y\le L}\big|\varphi^\dagger(y) - \varphi^*(y)\big|\\ &\le 2LC\cdot\varepsilon \quad \big(\text{because } \varphi^* \in B(\varphi^\dagger,\varepsilon)\big), \quad (95)\end{aligned}$$

where C is as in Lemma 2. Therefore, by (93) and (94), we have

$$\big|\mathbf{II}_n\big| \le 6L\sqrt{2LC\cdot\varepsilon} =: C_1\sqrt{\varepsilon}\,. \quad (96)$$

Lemma 3 now follows from (90), (91), and (96).

□

PROOF OF LEMMA 4.

The proof of this lemma appears in Devroye and Krzyżak (1989; Lemma 1).

□

PROOF OF LEMMA 5.

The proof can be found in Györfi et al. (2002; Sec. 23).

□

Proof of (24).

Start by letting

$$\Omega_\varepsilon = \left\{ 2\,i\varepsilon/(L\exp(ML)) \,\middle|\, -\lfloor ML\exp(ML)/\varepsilon \rfloor \,\le\, i \,\le\, \lfloor ML\exp(ML)/\varepsilon \rfloor \right\} \cup \{-M\} \cup \{M\}.$$

Also, let $\gamma \in [-M, M]$ be given and put $\varphi(y) = e^{\gamma y} \in \mathcal{F}$. If $\widetilde{\gamma} \in \Omega_\varepsilon$ is the closest value to γ, then

$$\begin{aligned} \sup_{|y|\le L} \left| e^{\gamma y} - e^{\widetilde{\gamma} y} \right| &= \sup_{|y|\le L} \left| y\,\exp\{\gamma^\dagger y\} \right| \cdot \left| \widetilde{\gamma} - \gamma \right|, \quad \big(\text{where } \gamma^\dagger \in (\widetilde{\gamma} \wedge \gamma\,,\, \widetilde{\gamma} \vee \gamma)\big) \\ &\le L\exp\{ML\} \cdot \left| \widetilde{\gamma} - \gamma \right| \,\le\, L\exp\{ML\} \cdot \frac{\varepsilon}{L\exp\{ML\}} \,=\, \varepsilon\,, \end{aligned}$$

where the last line follows from the fact that the distance between γ and its nearest value in Ω_ε is bounded by $\varepsilon/(L\exp\{ML\})$. Therefore, the class $\mathcal{F}$ is totally bounded. Moreover, a count of the number of terms in Ω_ε shows that the ε-covering number of $\mathcal{F}$ is bounded by the quantity $2\big\lfloor ML\exp\{ML\}\varepsilon^{-1} \big\rfloor + 3$.

□

Proof of (63).

Put $\mathcal{B}_m(\mathbf{Z}_i) = \big\{ \widetilde{\psi}_m(\mathbf{Z}_i;\varphi) \ge \varrho_0/2 \big\}$, where ϱ_0 is as in assumption D, and note that

$$\begin{aligned} P_{n,1}(i) \;&\le\; P\left\{ \left[\left| \frac{1-\widetilde{\eta}_m(\mathbf{Z}_i)}{\widetilde{\psi}_m(\mathbf{Z}_i;\varphi)} \right| \cdot \left| \widetilde{\psi}_m(\mathbf{Z}_i;\varphi) - E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big] \right| > \frac{\varrho_0\beta}{36BL^2} \right] \cap \mathcal{B}_m(\mathbf{Z}_i) \,\middle|\, \mathbf{Z}_i, Y_i \right\} \\ &\quad + \; P\Big\{ \mathcal{B}_m^c(\mathbf{Z}_i) \big| \mathbf{Z}_i, Y_i \Big\} \\ &:= \; P'_{n,1}(i) + P''_{n,1}(i). \end{aligned}$$

However, straightforward but tedious arguments show that

$$P'_{n,1}(i) \;\le\; P\left\{ \left| \widetilde{\psi}_m(\mathbf{Z}_i;\varphi) - E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big] \right| > \frac{\varrho_0^2\beta}{72BL^2} \,\middle|\, \mathbf{Z}_i, Y_i \right\} \;\le\; C_{14}\, e^{-C_{15}\, mh^d\beta^2},$$

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 $P''_{n,1}(i)$, we have $P''_{n,1}(i) = P\big\{ \widetilde{\psi}_m(\mathbf{Z}_i;\varphi) - E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big] < \varrho_0/2 - E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big] \big| \mathbf{Z}_i, Y_i \big\} \le P\big\{ \big| \widetilde{\psi}_m(\mathbf{Z}_i;\varphi) - E\big[\Delta_i\varphi(Y_i)\big|\mathbf{Z}_i\big] \big| > \varrho_0/2 \,\big|\mathbf{Z}_i, Y_i \big\} \le C_{16}\, \exp\{-C_{17}\, mh^d\}$, where we have used the fact that ψ_2 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

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

□

Proof of (85).

Start by defining the quantities

$$Q_{n,1}(\varphi) = \left| \frac{1}{\ell} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i \left| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - Y_i \right|^2}{\pi_\varphi(\mathbf{Z}_i, Y_i)} - \frac{1}{\ell} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i \left| m(\mathbf{X}_i; \pi_\varphi) - Y_i \right|^2}{\pi_\varphi(\mathbf{Z}_i, Y_i)} \right| \tag{97}$$

$$Q_{n,2}(\varphi) = \left| \frac{1}{\ell} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \frac{\Delta_i \left| m(\mathbf{X}_i; \pi_\varphi) - Y_i \right|^2}{\pi_\varphi(\mathbf{Z}_i, Y_i)} - E\left[\frac{\Delta \left| m(\mathbf{X}; \pi_\varphi) - Y \right|^2}{\pi_\varphi(\mathbf{Z}, Y)} \right] \right| \tag{98}$$

$$Q_{n,3}(\varphi) = \left| \frac{1}{\ell} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \Delta_i \left| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - Y_i \right|^2 \left[\Big(\pi_\varphi(\mathbf{Z}_i, Y_i) \Big)^{-1} - \Big(\widetilde{\pi}_\varphi(\mathbf{Z}_i, Y_i) \Big)^{-1} \right] \right|, \tag{99}$$

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

$$\begin{aligned} & P\left\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| \widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi) - E\Big| m(\mathbf{X}; \pi_\varphi) - Y \Big|^2 \right| > \beta \right\} \\ & \leq P\Big\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \big| Q_{n,1}(\varphi) \big| > \frac{\beta}{3} \Big\} + P\Big\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \big| Q_{n,2}(\varphi) \big| > \frac{\beta}{3} \Big\} + P\Big\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \big| Q_{n,3}(\varphi) \big| > \frac{\beta}{3} \Big\}. \\ & := P_{n,1} + P_{n,2} + P_{n,3}. \end{aligned} \tag{100}$$

However, in view of (79) and the fact that $|m(\mathbf{X}_i; \pi_\varphi)| \leq L/\pi_{\min}$, one obtains

$$\begin{aligned} & P_{n,1} \\ & \leq P\Big\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \frac{1}{\ell} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \Big[\Delta_i \big| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - m(\mathbf{X}_i; \pi_\varphi) \big| \big| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) + m(\mathbf{X}_i; \pi_\varphi) - 2Y_i \big| \Big] > \frac{\beta \pi_{\min}}{3} \Big\} \\ & \leq P\left\{ \sup_{\varphi \in \mathcal{F}_\varepsilon} \left| \frac{1}{\ell} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \left[\big| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - m(\mathbf{X}_i; \pi_\varphi) \big| \left(3 + \frac{1}{\pi_{\min}} + \max_{k \in \boldsymbol{\mathcal{I}}_m} \left| \frac{B}{\widetilde{\psi}_m(\mathbf{Z}_k; \varphi)} \right| \right) \right] \right| > \frac{\beta \pi_{\min}}{3L} \right\} \\ & \leq \big| \mathcal{F}_\varepsilon \big| \sup_{\varphi \in \mathcal{F}_\varepsilon} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} \left(P\left\{ \left[\Big| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - m(\mathbf{X}_i; \pi_\varphi) \Big| \Big(3 + \pi_{\min}^{-1} + B/(\varrho_0/2) \Big) > \frac{\beta \pi_{\min}}{3L} \right] \right. \right. \\ & \qquad \left. \left. \cap \left[\bigcap_{k \in \boldsymbol{\mathcal{I}}_m} \Big\{ \widetilde{\psi}_m(\mathbf{Z}_k; \varphi) \geq \frac{\varrho_0}{2} \Big\} \right] \right\} + \sum_{k \in \boldsymbol{\mathcal{I}}_m} P\Big\{ \widetilde{\psi}_m(\mathbf{Z}_k; \varphi) < \frac{\varrho_0}{2} \Big\} \right) \\ & \leq \big| \mathcal{F}_\varepsilon \big| \left[\sup_{\varphi \in \mathcal{F}_\varepsilon} \sum_{i \in \boldsymbol{\mathcal{I}}_\ell} P\left\{ \Big| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - m(\mathbf{X}_i; \pi_\varphi) \Big| > C_\beta \right\} + \ell \sum_{k \in \boldsymbol{\mathcal{I}}_m} P\Big\{ \widetilde{\psi}_m(\mathbf{Z}_k; \varphi) < \frac{\varrho_0}{2} \Big\} \right] \end{aligned} \tag{101}$$

where

$$C_\beta = \pi_{\min} \beta / 3L \left(3 + \pi_{\min}^{-1} + B/(\varrho_0/2) \right).$$

But the first probability statement in (101) can be bounded as follows. First, observe that

$$\begin{aligned} & P\left\{ \Big| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - m(\mathbf{X}_i; \pi_\varphi) \Big| > C_\beta \right\} \\ & \leq P\left\{ \Big| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \widetilde{\pi}_\varphi) - \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \pi_\varphi) \Big| > \frac{C_\beta}{2} \right\} + P\left\{ \Big| \widehat{m}_m^{\mathrm{HT}}(\mathbf{X}_i; \pi_\varphi) - m(\mathbf{X}_i; \pi_\varphi) \Big| > \frac{C_\beta}{2} \right\} \end{aligned}$$

$$:= \mathcal{P}_{n1}(\beta) + \mathcal{P}_{n1}(\beta). \tag{102}$$

On the other hand,

$$\begin{aligned}
\mathcal{P}_{n1}(\beta) &= P\left\{\left|\sum_{k\in\boldsymbol{\mathcal{I}}_m}\left[\Big(\widetilde{\pi}_\varphi(\mathbf{Z}_k,Y_k)\Big)^{-1}-\Big(\pi_\varphi(\mathbf{Z}_k,Y_k)\Big)^{-1}\right]\frac{\Delta_k Y_k \mathcal{K}((\mathbf{X}_i-\mathbf{X}_k)/h)}{\sum_{j\in\boldsymbol{\mathcal{I}}_m}\mathcal{K}((\mathbf{X}_i-\mathbf{X}_j)/h)}\right|>\frac{C_\beta}{2}\right\}\\
&\le P\left\{\max_{k\in\boldsymbol{\mathcal{I}}_m}\left|\Big(\widetilde{\pi}_\varphi(\mathbf{Z}_k,Y_k)\Big)^{-1}-\Big(\pi_\varphi(\mathbf{Z}_k,Y_k)\Big)^{-1}\right|>\frac{C_\beta}{2L}\right\}\\
&\le \sum_{k\in\boldsymbol{\mathcal{I}}_m} P\left\{\left|\Big(\widetilde{\pi}_\varphi(\mathbf{Z}_k,Y_k)\Big)^{-1}-\Big(\pi_\varphi(\mathbf{Z}_k,Y_k)\Big)^{-1}\right|>\frac{C_\beta}{2L}\right\}.
\end{aligned}$$

Therefore, using arguments similar to those leading to (60), (61), (62), and (63), we find, for every $\beta>0$ and n large enough,

$$\mathcal{P}_{n1}(\beta) \le C_{39}\, me^{-C_{40}mh^d\beta^2} + C_{41}\, me^{-C_{42}mh^d},$$

where C_{39}–C_{42} are positive constants not depending on m, ℓ, or β. Furthermore, tedious but standard arguments can be used to show that for n large enough, there are positive constants C_{43} and C_{44}, not depending on m, ℓ, or β, such that

$$\mathcal{P}_{n2}(\beta) \le C_{43}\, e^{-C_{44}mh^d\beta^2}.$$

As for the last probability statement on the right side of (101), our earlier arguments (see the paragraph after equation (81)) yield $P\{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi) < \varrho_0/2\} \le C_{16}\exp\{-C_{17}\,mh^d\}$, for n large enough, where C_{16} and C_{17} are positive constants not depending on n. Therefore, in view of (101) we arrive at

$$P_{n,1} \le \ell\,\big|\mathcal{F}_\varepsilon\big|\left(C_{39}\,me^{-C_{40}mh^d\beta^2} + C_{43}\,e^{-C_{44}mh^d\beta^2} + C_{55}\,me^{-C_{56}mh^d}\right), \tag{103}$$

for n large enough, where $P_{n,1}$ is as in (100). To deal with $P_{n,2}$, we first note that the terms $\Delta_i\big|m(\mathbf{X}_i;\pi_\varphi)-Y_i\big|^2/\pi_\varphi(\mathbf{Z}_i,Y_i)$, $i\in\boldsymbol{\mathcal{I}}_\ell$, are iid bounded random variables taking values in the interval $\big[0,\, L^2(1+1/\pi_{\min})^2/\pi_{\min}\big]$. Therefore an application of Hoeffding's inequality (in conjunction with the union bound) immediately yields

$$\begin{aligned}
P_{n,2} &\le \big|\mathcal{F}_\varepsilon\big|\sup_{\varphi\in\mathcal{F}_\varepsilon} P\left\{\left|\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\frac{\Delta_i\big|m(\mathbf{X}_i;\pi_\varphi)-Y_i\big|^2}{\pi_\varphi(\mathbf{Z}_i,Y_i)} - E\left[\frac{\Delta\big|m(\mathbf{X};\pi_\varphi)-Y\big|^2}{\pi_\varphi(\mathbf{Z},Y)}\right]\right|>\frac{\beta}{3}\right\}\\
&\le 2\,\big|\mathcal{F}_\varepsilon\big|\exp\Big\{-2\pi_{\min}^2\ell\,(\beta/3)^2/[L^4(1+1/\pi_{\min})^4]\Big\}.
\end{aligned} \tag{104}$$

Finally, to deal with the term $P_{n,3}$ in (100), we observe that in view of (79), and with ϱ_0 as in Assumption (D), one has

$$P_{n,3} \le \big|\mathcal{F}_\varepsilon\big|\sup_{\varphi\in\mathcal{F}_\varepsilon}\left(P\left\{\left[\left(2+\max_{k\in\boldsymbol{\mathcal{I}}_m}\left|\frac{B}{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)}\right|\right)^2 L^2\cdot\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\left|\frac{1}{\widetilde{\pi}_\varphi(\mathbf{Z}_i,Y_i)}-\frac{1}{\pi_\varphi(\mathbf{Z}_i,Y_i)}\right|>\frac{\beta}{3}\right]\right.\right.$$

$$\cap\Big[\bigcap_{k\in\boldsymbol{\mathcal{I}}_m}\Big\{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)\geq\varrho_0/2\Big\}\Big]\Bigg\}+\sum_{k\in\boldsymbol{\mathcal{I}}_m}P\Big\{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)<\varrho_0/2\Big\}\Bigg)$$

$$\leq\ |\mathcal{F}_\varepsilon|\sup_{\varphi\in\mathcal{F}_\varepsilon}\left(P\left\{\frac{1}{\ell}\sum_{i\in\boldsymbol{\mathcal{I}}_\ell}\left|\frac{1}{\widetilde{\pi}_\varphi(\mathbf{Z}_i,Y_i)}-\frac{1}{\pi_\varphi(\mathbf{Z}_i,Y_i)}\right|>d_\beta\right\}+\sum_{k\in\boldsymbol{\mathcal{I}}_m}P\Big\{\widetilde{\psi}_m(\mathbf{Z}_k;\varphi)<\varrho_0/2\Big\}\right)$$

where $d_\beta=\big[3L^2(2+2B/\varrho_0)^2\big]^{-1}\beta$. Now, employing the arguments used to bound the term $S_n(2)$ in (58), (see (60), (61), (62), (63)), it is straightforward to show that for n large enough

$$P_{n,3}\ \leq\ |\mathcal{F}_\varepsilon|\Big(C_{58}\ell\,e^{-C_{59}mh^d\beta^2}+C_{60}m\,e^{-C_{61}mh^d}\Big),\tag{105}$$

for positive constants C_{58}–C_{61} not depending on ℓ, m, or β. Putting together (100), (103), (104), and (105), one finds that for each $\beta>0$ and n large enough,

$$P\left\{\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widetilde{L}_{m,\ell}(\widetilde{\pi}_\varphi)-E\Big|m(\mathbf{X};\pi_\varphi)-Y\Big|^2\right|>\beta\right\}\ \leq\ |\mathcal{F}_\varepsilon|\Big(C_{39}\ell m\,e^{-C_{40}mh^d\beta^2}+C_{55}\ell m\,e^{-C_{56}mh^d}+\,2\,e^{-C_{64}\ell\beta^2}\Big).$$

□

Acknowledgements

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

Conflict of interest

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

References

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

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

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

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.

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

Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of kernel regression estimate. *Journal of Statistical Planning and Inference,* **23**, 71–82.

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.

Fang, F., Zhao, J., and Shao, J. (2018). Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values. *Statistica Sinica.* **28** 1677–1701.

Guo, X., Song, Y., and Zhu, L. (2019). Model checking for general linear regression with nonignorable missing response. *Computational Statistics & Data Analysis*, **138**, 1–12.

Györfi, L., Kohler, M., Walk, H. (1998). Weak and strong universal consistency of semi-recursive partitioning and kernel regression estimate. *Statistics and Decisions*, **16** 1–18.

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

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.

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.

Krzyżak, A. (1992). Global convergence of the recursive kernel regression estimates with applications in classification and nonlinear system estimation.*IEEE Trans. Inform. Theory*, **38** 1323–1338.

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

Liu, Z. and Yau, C.-Y. (2021). Fitting time series models for longitudinal surveys with nonignorable missing data. *J. Statist. Plann. Inference.* **214** 1–12.

Little, R. (1985). A note about models for selectivity bias. *Econometrica* **53** 1469–74.

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.

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

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

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

Mojirsheibani, M. (2021). On classification with nonignorable missing data. *J. Multivariate Anal.* **184** 104755.

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

Molenberghs, G. and Kenward, M. (2007). Missing Data in Clinical Studies. New York, Wiley.

Morikawa, K., Kim, J. K., and Kano, Y. (2017). Semiparametric maximum likelihood estimation with data missing not at random. *Can. J. Statist.* **45** 393–409.

Morikawa, K. and Kim, J. K. (2018). A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse. *Stat. & Probab. Lett.* **140** 1–6.

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

Niu, C., Guo, X., Xu, W., and Zhu, L. (2014). Empirical likelihood inference in linear regression with nonignorable missing response. *Computational Statistics & Data Analysis*, **79** 91–112.

O'Brien, J., Gunawardena, H., Paulo, J., Chen, X., Ibrahim, J., Gygi, S., and Qaqish, B. (2018).

The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *Ann. Appl. Statist.* **12** 2075–2095.

Qin, J., Leung, D., Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *J. Am. Statist. Assoc.* **97**, 193–200.

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

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

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.

Uehara, M. and Kim, J.K. (2018). Semiparametric response model with nonignorable nonresponse. Preprint on arXiv:1810.12519. https://arxiv.org/abs/1810.12519v1

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

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

Wang, L., Shao, J., and Fang, F. (2021). Propensity model selection with nonignorable nonresponse and instrument variable. *Statistica Sinica* **31** 647–671.

Wang, S., Shao, J., and Kim, J.K. (2014). Identifiability and estimation in problems with nonignorable nonresponse. Statistica Sinica 24, 1097 - 1116.

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

Yuan, C., Hedeker, D., Mermelstein, R., Xie, H. (2020). A tractable method to account for high-dimensional nonignorable missing data in intensive longitudinal data. *Stat. Med.* **39** 2589–2605.

Zhao, J., Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with

nonignorable missing data. *J. Am. Statist. Assoc* 110, 1577-1590.

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.