Statistical Papers (2024) 65:5615–5648
https://doi.org/10.1007/s00362-024-01611-8

REGULAR ARTICLE

Strong optimality of kernel functional regression in L^p norms with partial response variables and applications

Majid Mojirsheibani[1]

Received: 5 March 2023 / Revised: 1 June 2024 / Accepted: 10 September 2024 /
Published online: 7 October 2024

Abstract

This paper proposes kernel-type estimators of a regression function, with possibly unobservable response variables in a functional covariate setting, along with their rates of convergence in general L^p norms. Here, the mechanism that causes the absence of information (in the sense of having unobservable responses) is allowed to depend on both predictors and the response variables; this makes the problem particularly more challenging in those cases where model identifiability is an issue. As an immediate byproduct of these results, we propose asymptotically optimal classification rules for the challenging problem of semi-supervised learning based on the proposed estimators. Our proposed approach involves two steps: in the first step, 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 a weighted mean squared prediction error. The main focus of the paper is to look into the rates of *almost complete* convergence of the L^p norms of these estimators. The issue of identifiability is also addressed. As an application of our findings, we consider the classical problem of statistical classification based on the proposed regression estimators when there are a large number of missing labels in the data.

Keywords Nonparametric · Functional regression · Rates of convergence · Classification · Partially observed response

Mathematics Subject Classification Primary 62G05; Secondary 62G08

✉ Majid Mojirsheibani
majid.mojirsheibani@csun.edu

[1] Department of Mathematics, California State University Northridge, 18111 Nordhoff Street, Northridge, Los Angeles, CA 91330, USA

1 Introduction

This paper deals with the problem of kernel regression estimation with possibly unobservable response variables, Y, for the setup where the mechanism that causes the absence of information (i.e., causes Y to be possibly unobservable) is allowed to depend on both the predictor $\boldsymbol{\chi}$, which may be infinite dimensional, and the real-valued response variable Y. This is generally considered to be a challenging problem in incomplete data literature and is very different from the simpler popular missing at random model where the absence of Y depends on $\boldsymbol{\chi}$ only (but not Y itself).

More specifically, let $(\boldsymbol{\chi}, Y) \in \mathbb{X} \times \mathbb{R}$ be a random pair, where $\mathbb{X}$ may be an abstract space, and consider the problem of estimating the regression function $m(\chi) = \mathbb{E}(Y|\boldsymbol{\chi} = \chi)$, based on a sample of n independent and identically distributed (iid) data points $(\boldsymbol{\chi}_i, Y_i)$, $i = 1, \ldots, n$, drawn from the distribution of $(\boldsymbol{\chi}, Y)$. Let $d(\cdot,\cdot)$ be a semi-metric on $\mathbb{X}$ and observe that when the data is fully observable, the popular Nadaraya–Watson kernel estimator of $m(\boldsymbol{\chi})$ (Nadaraya 1964; Watson 1964) is given by

$$\widehat{m}_n(\chi) = \frac{\sum_{i=1}^n Y_i\, \mathcal{K}(d(\chi, \boldsymbol{\chi}_i)/h)}{\sum_{i=1}^n \mathcal{K}(d(\chi, \boldsymbol{\chi}_i)/h)}, \tag{1}$$

where the function $\mathcal{K} : \mathbb{R}_+ \to \mathbb{R}_+$ is the kernel used with the bandwidth $h \equiv h(n) > 0$. For detailed studies of some theoretical properties (point-wise and uniform) of the estimator in (1), one can refer to Ferraty and Vieu (2006) and Ferraty et al. (2010).

The focus of this paper is on the setup where the response variable Y could be unobservable for some cases. But more importantly, the probability that Y is observable is allowed to depend on Y itself, as well as $\boldsymbol{\chi}$ (and not just $\boldsymbol{\chi}$ alone). It is straightforward to see that under this setup the nonparametric estimator $\widehat{m}_n(\chi)$ in (1) is no longer available. This is a challenging case, yet some progress has been made in the literature. For example, in the case of linear regression models, one can refer to the developments and results of Niu et al. (2014), Guo et al. (2019), and Li et al. (2018) . Unlike the results of these three paper, here we do not assume a linear model. Mojirsheibani (2022) derives the limiting distribution of the maximal deviation of a particular kernel regression estimator based on auxiliary random variables where Y is missing NMAR and, unlike the current seup, the covariate $\boldsymbol{\chi}$ is in $\mathbb{R}$. However, in the cited paper, it is assumed that one already has available an estimator $\widehat{\gamma}$ of γ [corresponding to $\varphi(y) = exp\{\gamma y\}$] satisfying certain consistency properties. To the best of our knowledge, the nonparametric case with functional covariates was first studied in Kim and Yu (2011) and then by Ling et al. (2015), under the assumption that Y_i's were missing at random.

In passing, we also note that the estimator based on the complete cases only, i.e., the estimator

$$m_n^{\text{cc}}(\chi) = \frac{\sum_{i=1}^n Y_i \Delta_i\, \mathcal{K}(d(\chi, \boldsymbol{\chi}_i)/h)}{\sum_{i=1}^n \Delta_i \mathcal{K}(d(\chi, \boldsymbol{\chi}_i)/h)}, \tag{2}$$

where $\Delta_i = 1$ if Y_i is observable ($\Delta_i = 0$ otherwise), turns out to be the "wrong" estimator in the sense that it estimates the quantity $\mathbb{E}(\Delta Y|\boldsymbol{\chi} = \chi)/\mathbb{E}(\Delta|\boldsymbol{\chi} = \chi)$ which is not in general equal to the regression function $m(\chi) = \mathbb{E}(Y|\boldsymbol{\chi} = \chi)$ for the current setup where the probability that Y is observable could depend on both Y and $\boldsymbol{\chi}$. Let $\pi(\chi, y) = \mathbb{P}\{\Delta = 1|\boldsymbol{\chi} = \chi, Y = y\} = \mathbb{E}[\Delta|\boldsymbol{\chi} = \chi, Y = y]$, be the *selection probability*, also called the *nonresponse propensity*, where the random variable $\Delta = 1$ if Y is observable (and $\Delta = 0$ otherwise). For the important case of predictive models (as in regression or classification), we consider the following versatile logistic-type selection probability model, which is a generalization of the popular model proposed by Kim and Yu (2011),

$$\pi_{\varphi}(\chi, y) := \mathbb{E}\left[\Delta \,\middle|\, \boldsymbol{\chi} = \chi, Y = y\right] = \frac{1}{1 + \exp\left\{g(\chi)\right\} \cdot \varphi(y)}, \tag{3}$$

where $\varphi > 0$ is a given function that could depend on unknown parameter(s) and g is an unknown function; both φ and g are real-valued. In what follows, the true but unknown function φ will be denoted by φ^*.

The generalization $\varphi(y)$ in (3) and its estimation is not new and has been considered in the literature before. For example, Wang et al. (2021) replace $\exp(\gamma y)$ by a general function $q(y, \gamma)$ and propose estimators based on the generalized method of moments. Similarly, Mojirsheibani (2021) considers the generalization $\varphi(y)$ for the problem of classification when identifiability is not an issue.

We observe that when $\varphi(y) = e^{\gamma y}$ for an unknown parameter γ, then (3) reduces to the model proposed by Kim and Yu (2011), which has been studied and used extensively in the literature; see, for example, Zhao and Shao (2015), Shao and Wang (2016), Morikawa et al. (2017), Uehara et al. (2018), Morikawa and Kim (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), and Wang et al. (2014, 2021).

Of course, one may decide to consider more general nonparametric models instead of (3), but the estimation of such general models will become a difficult (if not impossible) issue. In fact, in view of the recent widespread use of the model proposed by Kim and Yu (2011) in the literature, the model in (3) 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.

One aim of this paper is to explore the construction of counterparts of the kernel estimator in (1) for the case where the response variable Y can be missing, but not necessarily at random (NMAR). Another aim is to apply our results to the problem of classification where we construct asymptotically optimal nonparametric classification rules in the presence of NMAR response variables. Our contributions here may be summarized as follows. (i) We develop an easy-to-implement estimators of the regression curve $m(\chi)$, with functional covariates, in the presence of NMAR data. (ii) We will carefully explore and study the global properties of the proposed regression estimators in general L^p norms. More specifically, we study the rates of convergence (in L^p) of the proposed estimator under the NMAR setups. (iii) We look into the applications of our proposed regression estimator to the problem of nonparametric classification in the presence of partially observed data.

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 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 Sect. 3 make it possible to develop asymptotically correct 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 2.1 presents the main results of the paper. Theorem 1 gives the rates of convergence (in L^p) of the proposed estimators. Section 3 explores the applications of the results of Sect. 2.1 to the problem of statistical classification with incomplete covariates (also known as semi-supervised classification). Strong optimality of the proposed classifiers is addressed in Theorem 2. Numerical examples are presented in Sect. 4; our numerical findings confirm the good finite-sample performance of our estimators. All proofs are deferred to Sect. 5.

2 Main results

2.1 The estimator

To present our results, let $\mathbb{D}_n$ represent n independent and identically distributed (iid) data values,

$$\mathbb{D}_n = \{(\boldsymbol{\chi}_1, Y_1, \Delta_1), \ldots, (\boldsymbol{\chi}_n, Y_n, \Delta_n)\},$$

where $\Delta_i = 0$ if Y_i is missing (and $\Delta_i = 1$ otherwise). Next, 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 $\mathbb{D}_m \cap \mathbb{D}_\ell = \varnothing$. It is assumed that ℓ (and m) $\to \infty$, as $n \to \infty$. Here, one can of course take $m = \lfloor \frac{n}{2} \rfloor$, but more general choices of m and ℓ will be discussed later in our main results. Let $\mathcal{F}$ be the class of functions to which the unknown function φ of (3) belongs. For each $\varphi \in \mathcal{F}$, put

$$\begin{aligned} \psi_k(\chi; \varphi) &:= \mathbb{E}\left[\Delta Y^{2-k} \varphi(Y) \middle| \boldsymbol{\chi} = \chi\right] \quad \text{and} \\ \eta_k(\chi) &:= \mathbb{E}\left[\Delta Y^{2-k} \middle| \boldsymbol{\chi} = \chi\right], \quad \text{for } k = 1, 2, \end{aligned} \tag{4}$$

and define

$$m(\chi; \varphi) = \eta_1(\chi) + \frac{\psi_1(\chi; \varphi)}{\psi_2(\chi; \varphi)} \cdot (1 - \eta_2(\chi)). \tag{5}$$

It will be noted in Lemma 1 that the true underlying regression function $m(\chi)$ is equal to $m(\chi; \varphi^*)$, where φ^* is the true (but unknown) function φ in (3). Also, define the index sets

$$\mathcal{I}_m = \left\{ i \in \{1, \ldots, n\} \,\middle|\, (\boldsymbol{\chi}_i, Y_i, \Delta_i) \in \mathbb{D}_m \right\} \quad \text{and}$$
$$\mathcal{I}_\ell = \left\{ i \in \{1, \ldots, n\} \,\middle|\, (\boldsymbol{\chi}_i, Y_i, \Delta_i) \in \mathbb{D}_\ell \right\}.$$

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

$$\widehat{m}_m(\chi; \varphi) = \widehat{\eta}_{m,1}(\chi) + \frac{\widehat{\psi}_{m,1}(\chi; \varphi)}{\widehat{\psi}_{m,2}(\chi; \varphi)} \big(1 - \widehat{\eta}_{m,2}(\chi)\big), \tag{6}$$

where $\widehat{\psi}_{m,k}(\chi; \varphi)$ and $\widehat{\eta}_{m,k}(\chi)$, $k = 1, 2$, are the kernel estimators of $\psi_k(\chi; \varphi)$ and $\eta_k(\chi)$ in (4), i.e.,

$$\widehat{\psi}_{m,k}(\chi; \varphi) = \frac{\sum_{i \in \mathcal{I}_m} \Delta_i Y_i^{2-k} \varphi(Y_i)\, \mathcal{K}\big(h^{-1} d(\chi, \boldsymbol{\chi}_i)\big)}{\sum_{i \in \mathcal{I}_m} \mathcal{K}\big(h^{-1} d(\chi, \boldsymbol{\chi}_i)\big)}, \quad k = 1, 2, \quad \varphi \in \mathcal{F}, \tag{7}$$

$$\widehat{\eta}_{m,k}(\chi) = \frac{\sum_{i \in \mathcal{I}_m} \Delta_i Y_i^{2-k} \mathcal{K}\big(h^{-1} d(\chi, \boldsymbol{\chi}_i)\big)}{\sum_{i \in \mathcal{I}_m} \mathcal{K}\big(h^{-1} d(\chi, \boldsymbol{\chi}_i)\big)}, \quad k = 1, 2. \tag{8}$$

Clearly (6) is not quite an estimator of the regression function $m(\chi)$ because one still needs to replace φ by an estimator of the unknown function φ^*. Our approach to estimate the function φ^* 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, $\| \|_\infty$ 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 $\| \|_\infty$). The quantity $\log(\mathcal{N}_\varepsilon(\mathcal{F}))$ is called Kolmogorov's ϵ-entropy of the set $\mathcal{F}$. The monograph by van der Vaart and Wellner (1996, p. 83) provides more details on such concepts.

Now, to estimate the function φ^*, we first observe that in view of the results of Kim and Yu (2011), the term $\exp\{g(\chi)\}$ that appears in (3) can also be expressed as

$$\exp\{g(\chi)\} = \frac{\mathbb{E}\big[1 - \Delta \big| \boldsymbol{\chi} = \chi\big]}{\mathbb{E}\big[\Delta\, \varphi(Y) \big| \boldsymbol{\chi} = \chi\big]} \overset{\text{via (4)}}{=} \frac{1 - \eta_2(\chi)}{\psi_2(\chi; \varphi)}. \tag{9}$$

However, estimating the right side of (9) can be challenging due to identifiability problems. The issue of model identifiability arises when different sets of parameters do no yield distinct models. In the context of this paper on regression function estimation, we follow Shao and Wang (2016) and consider a population $\mathcal{P}$ to be identifiable if for two sets of distince parameters, the corresponding versions of $\mathcal{P}$, say $\mathcal{P}_1$ and $\mathcal{P}_2$, do not give the same $\pi_\varphi(\chi, y) f(y|\chi)$, where π_φ is as in (3) and $f(y|\chi)$ is the conditional density of the random variable Y given χ. As an example, let $f(y|\chi)$ be a the normal distribution $N(\mu(\chi),\ \sigma^2(\chi))$, where both $\mu(\chi)$ and $\sigma^2(\chi)$ are unspecified. Additionally, consider the special case of $\varphi(y)=\exp(\gamma y)$ as in Kim and Yu (2011), and let $\{g_1(\chi),\mu_1(\chi),\sigma_1(\chi),\gamma_1\}$ and $\{g_2(\chi),\mu_2(\chi),\sigma_2(\chi),\gamma_2\}$ be two distinct sets of parameters corresponding to populations $\mathcal{P}_1$ and $\mathcal{P}_2$. Then, it follows from the work of Shao and Wang (2016) that both $\mathcal{P}_1$ and $\mathcal{P}_2$ will have the same $\pi_\varphi(\chi,y)f(y|\chi)$ whenever $\gamma_1=-\gamma_2$, $g_1(\chi)=-g_2(\chi)$, $\mu_1(\chi)=\mu_2(\chi)-\gamma_1\sigma_2^2(\chi)$ and $g_2(\chi)=\gamma_2^2\sigma_2^2(\chi)/2-\gamma_2\mu_2(\chi)$. At the same time, in view of (refexpgx), one also needs $g_2(\chi)=\log\left\{\mathbb{E}\left[1-\Delta\middle|\boldsymbol{\chi}=\chi\right]\div\mathbb{E}\left[\Delta\ \exp(\gamma_2 Y)\ \middle|\boldsymbol{\chi}=\chi\right]\right\}$, which makes it virtually impossible to check all the above conditions in practice.

To deal with identifiability in the current setup where $\boldsymbol{\chi}$ is a functional predictor, let $(\Omega,\mathcal{A},\mathbb{P})$ be the underlying probability space. We take $\mathbb{X}$ (see the introduction for the notation) to be the space of square-integrable functions defined on an interval of the real line, i.e., $\boldsymbol{\chi}$ is a random function on $(\Omega,\mathcal{A},\mathbb{P})$ with values (sample paths) in $L^2(\mathcal{I})$, where $\mathcal{I}$ is an interval on the real line; in fact, we take $\mathcal{I}=[a,b]$, for some finite $a<b$. A sufficient condition for identifiability is that there is a segment of $\boldsymbol{\chi}$, say $\boldsymbol{\zeta}=\boldsymbol{\chi}\big|_s$, which is independent of Δ, given Y and $\boldsymbol{\chi}\big|_{s^c}$, where $s=[a,t_o]$ for some $t_o\in(a,b)$, and $\boldsymbol{\chi}\big|_s$ represents the restriction of $\boldsymbol{\chi}(t)$ to $t\in s$.

To justify this identifiability condition [which also appears under Assumption (A1) later in this section], observe that since $\boldsymbol{\chi}\in L^2([a,b])$, i.e., a separable Hilbert space, one can write $\boldsymbol{\chi}(t)\big|_{t\in s}=\sum_{j=1}^{\infty}U_j\,\psi_j(t)$, $t\in s=[a,t_o]$, where $\{\psi_1,\psi_2,\ldots\}$ is a complete orthonormal basis for $L^2(s)$ and $U_j=\int_s\boldsymbol{\chi}(t)\psi_j(t)\,dt$, $j=1,2,\ldots$ Here, the infinite sum $\sum_{j=1}^{\infty}U_j\psi_j(t)$ converges in $L^2(s)$. Similarly, one can write $\boldsymbol{\chi}(t)\big|_{t\in s^c}=\sum_{j=1}^{\infty}V_j\,\phi_j(t)$, $t\in s^c=(t_0,b]$, for any complete orthonormal basis $\{\phi_1,\phi_2,\ldots\}$ of $L^2(s^c)$, where $V_j=\int_{s^c}\boldsymbol{\chi}(t)\phi_j(t)\,dt$, $j=1,2,\ldots$ Since any infinite dimensional Hilbert space is isomorphic to the space $\ell_2=\left\{(x_1,x_2,\ldots)\middle|\sum_{i=1}^{\infty}|x_i|^2<\infty\right\}$ the segments $\boldsymbol{\chi}(t)\big|_{t\in s}$ and $\boldsymbol{\chi}(t)\big|_{t\in s^c}$ of the covariate curve $\boldsymbol{\chi}$ can be represented by the *surrogate* vectors $\mathbf{U}=(U_1,U_2,\ldots)$ and $\mathbf{V}=(V_1,V_2,\ldots)$ in the sense that knowing $\mathbf{U}$ (respectively $\mathbf{V}$) is the same as knowing the curve $\boldsymbol{\chi}(t)\big|_{t\in s}$ (respectively $\boldsymbol{\chi}(t)\big|_{t\in s^c}$). Therefore, the assumption that the curve segment $\boldsymbol{\zeta}$ $(:=\boldsymbol{\chi}\big|_s)$ is independent of Δ, given Y and $\boldsymbol{\chi}\big|_{s^c}$, is equivalent to $\mathbf{U}$ being independent of Δ, given Y and $\mathbf{V}$, which is a sufficient condition for model identification; see, for example, Uehara et al. (2018) or Shao and Wang (2016). Under this assumption, the selection probability in (3) becomes

$$\pi_\varphi(\zeta,y):=\mathbb{E}\left[\Delta\,\middle|\boldsymbol{\zeta}=\zeta,Y=y\right]=\frac{1}{1+\exp\{g(\zeta)\}\cdot\varphi(y)},\tag{10}$$

where the true φ is denoted by φ^* (as before), and therefore (9) reduces to

$$\exp\{g(\zeta)\}=\frac{1-\eta_o(\zeta)}{\psi_o(\zeta;\varphi)},\quad\text{where}\quad\psi_o(\zeta;\varphi)=\mathbb{E}\left[\Delta\,\varphi(Y)\middle|\boldsymbol{\zeta}=\zeta\right]\quad\text{and}$$
$$\eta_o(\zeta)=\mathbb{E}\left[\Delta\middle|\boldsymbol{\zeta}=\zeta\right].\tag{11}$$

Also, for each given $\varphi\in\mathcal{F}$, consider the following estimators of $\psi_o(\zeta;\varphi)$ and $\eta_o(\zeta)$

$$\begin{cases} \widehat{\psi}_{m,o}(\zeta;\varphi) = \sum_{j\in\mathcal{I}_m} \Delta_j \varphi(Y_j)\, \mathcal{K}\big(h^{-1}d_o(\zeta,\, \boldsymbol{\zeta}_j)\big) \div \sum_{j\in\mathcal{I}_m} \mathcal{K}\big(h^{-1}d_o(\zeta,\, \boldsymbol{\zeta}_j)\big), \\ \widehat{\eta}_{m,o}(\zeta) = \sum_{j\in\mathcal{I}_m} \Delta_j\, \mathcal{K}\left(h^{-1}d_o(\zeta,\, \boldsymbol{\zeta}_j)\right) \div \sum_{j\in\mathcal{I}_m} \mathcal{K}\left(h^{-1}d_o(\zeta,\, \boldsymbol{\zeta}_j)\right), \end{cases} \tag{12}$$

where d_o is the usual metric induced by the $L^2(s)$ norm with $s=[a, t_o]$. Finally, our proposed estimator of the unknown function φ^* is obtained in the following two steps. Let $\varepsilon_m > 0$ be a decreasing sequence $\varepsilon_m \downarrow 0$, as $m \to \infty$, and let $\mathcal{F}_{\varepsilon_m} = \{\varphi_1, \ldots, \varphi_{N(\varepsilon_m)}\} \subset \mathcal{F}$ be any ε_m-cover of $\mathcal{F}$; the choice of ε_m will be discussed later in Theorem 1. Then

Step 1. For each fixed (given) $\varphi \in \mathcal{F}_{\varepsilon_m}$, use the training sample $\mathbb{D}_m$ to compute $\widehat{m}_m(\chi;\varphi)$, which is given by (6), and also to estimate the selection probability $\pi_\varphi(\zeta, y)$ in (10) by

$$\widehat{\pi}_\varphi(\boldsymbol{\zeta}, y) = \left[1 + \widehat{\exp\{g(\boldsymbol{\zeta})\}} \cdot \varphi(y)\right]^{-1}, \tag{13}$$

where in view of (11), $\widehat{\exp\{g(\boldsymbol{\zeta})\}}$ is given by

$$\widehat{\exp\{g(\boldsymbol{\zeta})\}} = \frac{1 - \widehat{\eta}_{m,o}(\boldsymbol{\zeta})}{\widehat{\psi}_{m,o}(\boldsymbol{\zeta};\varphi)}. \tag{14}$$

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

$$\widehat{\varphi}_n := \underset{\varphi\,\in\mathcal{F}_{\varepsilon_m}}{\operatorname{argmin}}\ \ell^{-1} \sum_{i\in\mathcal{I}_\ell} \frac{\Delta_i}{\widehat{\pi}_\varphi(\boldsymbol{\zeta}_i, Y_i)} \left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2, \tag{15}$$

where $\widehat{m}_m(\chi;\varphi)$ is as in (6). The subscript n at $\widehat{\varphi}_n$ reflects the fact that the entire data of size n has been used here. Finally, the corresponding estimator of the regression function $m(\chi)$ is given by

$$\widehat{m}(\chi;\widehat{\varphi}_n) := \widehat{m}_m(\chi;\varphi)\big|_{\varphi=\widehat{\varphi}_n}, \text{ with } \widehat{m}_m(\chi;\varphi) \text{ as in (6).} \tag{16}$$

Remark 1 The estimator in (15) may be viewed as the empirical version of the minimizer of the mean squared error, i.e., the empirical version of

$$\varphi_{\varepsilon_m} := \underset{\varphi\in\mathcal{F}_{\varepsilon_m}}{\operatorname{argmin}}\ \mathbb{E}\big|m(\boldsymbol{\chi};\varphi) - Y\big|^2, \tag{17}$$

where $m(\boldsymbol{\chi};\varphi)$ is the regression function $m(\boldsymbol{\chi};\varphi^*)$ evaluated at an arbitrary $\varphi \in \mathcal{F}_{\varepsilon_m}$. To appreciate this, observe that upon conditioning on Y and $\boldsymbol{\chi}$, for each $\varphi \in \mathcal{F}_{\varepsilon_m}$ one finds $\mathbb{E}\{\frac{\Delta}{\pi_{\varphi^*}(\boldsymbol{\zeta}, Y)}\,\big|m(\boldsymbol{\chi};\varphi) - Y\big|^2\} = \mathbb{E}\big[\mathbb{E}\{\frac{\Delta}{\pi_{\varphi^*}(\boldsymbol{\chi}|_s, Y)}\,\big|m(\boldsymbol{\chi};\varphi) - Y\big|^2\big|\boldsymbol{\chi}, Y\}\big] = \mathbb{E}\big[\big|m(\boldsymbol{\chi};\varphi) - Y\big|^2 \frac{1}{\pi_{\varphi^*}(\boldsymbol{\chi}|_s, Y)} \cdot \mathbb{E}\{\Delta\big|\,\boldsymbol{\chi}, Y\}\big] = \mathbb{E}|m(\boldsymbol{\chi};\varphi) - Y|^2$, where the last equality follows from the definition of $\pi_\varphi(\zeta, y)$ in (10) with φ^* being the true value of φ, and the fact that $\zeta = \chi|_s$. We also note that φ_{ε_m} in (17) is an approximation to the true function φ^* based on the cover $\mathcal{F}_{\varepsilon_m}$ of $\mathcal{F}$. In fact, with L as in Assumption (A9), we have

$$\varphi^* := \underset{\varphi:[-L,L]\to\mathbb{R}_+}{\operatorname{argmin}} \mathbb{E}\big|m(\boldsymbol{\chi};\varphi) - Y\big|^2. \tag{18}$$

How good is the regression estimator $\widehat{m}(\chi; \widehat{\varphi}_n)$ in (16)? To answer this question, we first state a number of assumptions. In what follows, $\mathbb{X}$ is the space $L^2([a,b])$, $-\infty < a < b < \infty$ as before, and d (respectively d_0) is the metric induced by the usual $L^2([a,b])$-norm (respectively $L^2([a,t_o])$-norm). Furthermore, $\forall\chi \in \mathbb{X}$, we define $B(\chi,h) = \big\{\chi' \in \mathbb{X}\,\big|\, d(\chi',\chi) < h\big\}$. Similarly, $\forall\zeta \in \mathbb{X}|_s$, where $s = [a,\ t_o]$ for any $t_o \in (a,b)$, we define $B_o(\zeta,h) = \big\{\zeta' \in \mathbb{X}|_s\,\big|\, d_0(\zeta',\zeta) < h\big\}$.

Assumption (A0) There is a subset $\mathcal{S}_{\mathbb{X}} \subset \mathbb{X}$ satisfying $\mathbb{P}\big\{\boldsymbol{\chi} \in \mathcal{S}_{\mathbb{X}}\big\} = 1$.

Assumption (A1) (*Identifiability*) There is a segment of $\boldsymbol{\chi}$, say $\boldsymbol{\zeta} = \boldsymbol{\chi}\big|_s$, which is independent of Δ, given Y and $\boldsymbol{\chi}\big|_{s^c}$, where $\boldsymbol{\chi}\big|_s$ is the restriction of $\boldsymbol{\chi}(t)$ to $t \in s := [a, t_o]$, for some $t_o \in (a,b)$.

Assumption (A2) Let $\mathcal{S}^o_{\mathbb{X}} = \big\{\zeta\big|\, \zeta = \chi|_s,\ \chi \in \mathcal{S}_{\mathbb{X}},\ s = [a,\ t_o]\big\}$. There exist functions ϕ_1 and ϕ_0 such that $\forall\chi \in \mathcal{S}_{\mathbb{X}}$, and $\forall\zeta \in \mathcal{S}^o_{\mathbb{X}}$, and for all $h > 0$,

$$\begin{aligned}
&0 < C\phi_1(h) \le \mathbb{P}\big\{\boldsymbol{\chi} \in B(\chi,h)\big\} \le C'\phi_1(h) \quad \text{and}\\
&0 < C_0\,\phi_0(h) \le \mathbb{P}\big\{\boldsymbol{\zeta} \in B_o(\zeta,h)\big\} \le C_0'\phi_0(h)
\end{aligned}$$

for positive constants C, C', C_0, C_0'.

Assumption (A3) (*Lipschitz conditions on ψ_k*) Let ψ_k be as in (4) and ψ_o as in (11). There are constants $\beta_0, \beta_1, \beta_2 > 0$ such that $\forall\,\chi_1,\ \chi_2 \in \mathcal{S}_{\mathbb{X}}$, $\forall\,\zeta_1,\ \zeta_2 \in \mathcal{S}^o_{\mathbb{X}}$, and $\forall\,\varphi \in \mathcal{F} \cup \{1\}$

$$\begin{aligned}
&\big|\psi_k(\chi_1;\varphi) - \psi_k(\chi_2;\varphi)\big| \le C_k\, d^{\beta_k}(\chi_1,\chi_2), \quad k = 1,2, \quad \text{and}\\
&\big|\psi_o(\zeta_1;\varphi) - \psi_o(\zeta_2;\varphi)\big| \le C_0\, d_0^{\beta_0}(\zeta_1,\zeta_2),
\end{aligned}$$

where $C_0,\ C_1,\ C_2$ are positive constants.

Assumption (A4) The kernel $\mathcal{K}$ is nonnegative, bounded and Lipschitz on its support $[0,1)$, and with $\mathcal{K}(1) = 0$ satisfying $-\infty < C < \mathcal{K}'(t) < C' < \infty$, for all $t \in [0,1)$, for constants C and C'.

Assumption (A5)

(A5a) The function ϕ_1 in Assumption (A1) is such that $\exists C > 0$, $\exists\eta_0 > 0$ such that $\forall\eta < \eta_0$, $\phi_1'(\eta) < C$. Furthermore, with $\mathcal{K}(1) = 0$, $\exists C > 0$, $\exists\eta_0 > 0$ such that $\forall 0 < \eta < \eta_0$, the function ϕ_1 satisfies $\int_0^\eta \phi_1(t)\,dt > C\eta\phi_1(\eta)$.

(A5b) Similarly, the function ϕ_0 in Assumption (A1) satisfies the same requirements with possibly different constants C and η_0.

Assumption (A6) (*Assumptions on* ϕ_1, ϕ_0, *and the covering number of* $\mathcal{S}_{\mathbb{X}}$) For any $\tau > 0$, let $\mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})$ be the τ-covering number of $\mathcal{S}_{\mathbb{X}}$, i.e., the smallest number of open balls of d-radius equal to τ needed to cover $\mathcal{S}_{\mathbb{X}}$.
(i) Let $\tau_m := \log m/m$. For n (and thus m) large enough, $(\log m)^2/\big(m\phi_1(h)\big) < \log[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})] < m\phi_1(h)/\log m$, and also $(\log m)^2/\big(m\phi_0(h)\big) < \log[\mathcal{N}_{\tau_m}(\mathcal{S}^o_{\mathbb{X}})] < m\phi_0(h)/\log m$, where $\mathcal{S}^o_{\mathbb{X}}$ is as in Assumption (A2). Furthermore, for $k = 0, 1$, $mh\sqrt{\phi_k(h)} \to 0$ as $m \to \infty$.
(ii) The Kolmogorov's τ_m-entropy of $\mathcal{S}_{\mathbb{X}}$ and ε_n-entropy of $\mathcal{F}$ satisfy the summability condition $\sum_{m=1}^{\infty} \exp\{(1-\beta)\log\big[\mathcal{N}_{\varepsilon_m}(\mathcal{F}) \vee \mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\big] < \infty$, for some $\beta > 1$.

Assumption (A7) There is a constant $\pi_{\min} > 0$ such that $\pi_\varphi(\varsigma, y) > \pi_{\min}$, for all $\varsigma \in \mathcal{S}^o_{\mathbb{X}}$ and all y, where π_φ is the selection probability in (10). Furthermore, $\mathbb{E}[\Delta\, \varphi(Y)|\boldsymbol{\chi} = \chi] \geq \varrho_0$, for all $\chi \in \mathcal{S}_{\mathbb{X}}$ and each $\varphi \in \mathcal{F}$, for some constant $\varrho_0 > 0$.

Assumption (A8) The deviation $A_{m,\ell}(\varphi) = |\widehat{L}_{m,\ell}(\varphi) - E[|\widehat{m}_m(\boldsymbol{\chi};\varphi) - Y|^2|\mathbb{D}_m]|$, $\varphi \in \mathcal{F}_{\varepsilon_m}$, where $\widehat{L}_{m,\ell}(\varphi)$ and $\widehat{m}_m(\chi;\varphi)$ are as in (27) and (6), satisfies $\mathbb{P}\{A_{m,\ell}(\varphi) > t\} \leq \sup_{\varphi\in\mathcal{F}} \mathbb{P}_\varphi\{A_{m,\ell}(\varphi) > t\}$, $\forall\, t > 0$, where $\mathbb{P}_\varphi$ denotes the probability computed when φ is the true function, and $\mathbb{P}$ means $\mathbb{P}_{\varphi^*}$.

Assumption (A9) The function φ^* belongs to a totally bounded class $\mathcal{F}$ of functions $\varphi : [-L, L] \to (0, B]$, for some $B < \infty$ and $L < \infty$, where φ^* is the true φ in (10).

Assumption (A0) is not new and has already been considered in the literature; see, for example, Ferraty et al. (2013). As an example, a particularly interesting subset $\mathcal{S}_{\mathbb{X}}$ of $\mathbb{X}$ is the class of functions in $L^2([a, b])$ satisfying the following classical Kolmogorov–Riesz sufficient conditions for $\mathcal{S}_{\mathbb{X}}$ to be totally bounded: (i) $\mathcal{S}_{\mathbb{X}}$ is bounded, and (ii) for every $\epsilon > 0$ there is a $\rho > 0$ such that, for every $\boldsymbol{\chi} \in \mathcal{S}_{\mathbb{X}}$ and $|y| < \rho$, one has $\int \big|\boldsymbol{\chi}_{{}_{[a,b]}}(t+y) - \boldsymbol{\chi}_{{}_{[a,b]}}(t)\big|^2 dt < \epsilon^2$, where $\boldsymbol{\chi}_{{}_{[a,b]}}(t) := \boldsymbol{\chi}(t) \cdot \mathbb{1}_{\{a \leq t \leq b\}} \in L^2(\mathbb{R})$. Assumption (A1) deals with the identifiability issue (as discussed earlier in Sect. 2.1). Assumptions (A2)–(A6) are standard in functional kernel regression; see, for example, Ferraty et al. (2010). The first part of assumption (A7) is common in missing data literature (as in Cheng and Chu 1996 or Ferraty et al. 2013); this assumption essentially states that Y can be observed (i.e., $\Delta = 1$) with a non-zero probability for all value of (ς, y). The second part of Assumption (A7) is rather mild and can be justified by noticing that $\mathbb{E}[\Delta\, \varphi(Y)|\boldsymbol{\chi}] = \mathbb{E}[\varphi(Y)\mathbb{E}(\Delta\,|\boldsymbol{\chi}, Y)\,|\boldsymbol{\chi}] \geq \pi_{\min}\mathbb{E}[\varphi(Y)|\boldsymbol{\chi}]$ together with the fact that $\varphi(y) > 0$ for all y. The last two assumptions are technical.

The following result explores the almost complete (a. co.) convergence of the L^p norm of $\widehat{m}(\chi;\widehat{\varphi}_n)$. In passing, we recall (see, for example, Ferraty et al. 2010) that a sequence of real-valued random variables Z_n is said to converge a. co. to a constant c if for every $t > 0$, $\sum_{n\geq 1} P\{|Z_n - c| > t\} < \infty$.

Theorem 1 *Let* $\widehat{m}(\chi;\widehat{\varphi}_n)$ *be the estimator in* (16) *and suppose that Assumptions* (A0)–(A9) *hold. Also let the selection probability* π_φ *be as in* (10). *Let* $\varepsilon_m \downarrow 0$ *be any sequence of positive constants satisfying* $\ell^{-1}\log[\mathcal{N}_{\varepsilon_m}(\mathcal{F})] \to 0$, *as* n (*thus* m *and* ℓ) $\to \infty$. *Then, for any* $p \in [2, \infty)$, *one has*

$$\begin{aligned}&\mathbb{E}\Big[\big|\widehat{m}(\boldsymbol{\chi};\widehat{\varphi}_n)-m(\boldsymbol{\chi})\big|^p\Big|\mathbb{D}_n\Big]\\&\quad=\mathcal{O}(h^\alpha)\\&\qquad+\mathcal{O}_{a.co.}\left(\sqrt{\frac{\log[\mathcal{N}_{\varepsilon_m}(\mathcal{F})]}{\ell}}\right)+\mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F})\vee\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m\cdot\phi(h)}}\right)+\sqrt{\varepsilon_m},\end{aligned}$$

where $\alpha>0$ *is a constant not depending on* n, $\tau_m=\log(m)/m$, *and* $\phi(h)=\phi_0(h)\wedge\phi_1(h)$ *with* ϕ_0 *and* ϕ_1 *as in Assumption* (A2).

It is also desirable to find the limiting distribution of the L^2 norm of the proposed estimator as it can provide a tool to perform inferences for the unknown regression curve in a global sense. However, such results are quite difficult (if not impossible) to establish when $\boldsymbol{\chi}$ is functional and Y may be missing NMAR. In fact, to the best of our knowledge, such results are not available even for the simpler case of missing at random (MAR) response values with Euclidean $\boldsymbol{\chi}$'s. The only related result we are aware of is that of Mojirsheibani (2022) which derives the limiting distribution of the maximal deviation of a very particular kernel-type regression estimator (under some rather stringent conditions).

Remark 2 Consider the class $\mathcal{F}$ of functions φ of the form:

$$\varphi(y)=\exp\{\gamma y\},\quad|\gamma|\le M,\quad|y|\le L,\quad\text{for any } M,L<\infty,\tag{19}$$

which is similar to the function used by Kim and Yu (2011) in their version of the selection probability model (10).

It is straightforward to see that for every $\varepsilon>0$, the finite collection of functions

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

is an ε-cover of $\mathcal{F}$ and the covering number of $\mathcal{F}$ is bounded by $(2\,ML\exp\{ML\}\varepsilon^{-1}+3)$. Since this bound grows like ε^{-1} (as $\varepsilon\downarrow 0$), one obtains the strong L^p, $p\in[2,\infty)$, convergence results for the regression estimator (16), under the conditions of Theorem 1, for any sequence $\varepsilon_m\downarrow 0$ (as $m\to\infty$) that satisfies $\ell^{-1}\log(1/\varepsilon_m)\to 0$.

2.2 Estimation of *s*

The methodology proposed in the previous section assumes that s, the interval related to identifiability Assumption (A1), was known from prior information. In practice s is often unknown and estimating it is particularly complicated. In fact, it is complicated even for the simpler case where $\boldsymbol{\chi}\in\mathbb{R}^d$; see Wang et al. (2021) for three methods of finding the subset $\mathbf{U}$ of $\boldsymbol{\chi}$, where $\boldsymbol{\chi}=(\mathbf{U},\mathbf{Z})$, such that $P(\Delta=1\big|\boldsymbol{\chi},Y)=P(\Delta=1|\mathbf{U},Y)$. Our proposed approach to estimate s is more in the spirit of the third method discussed in the cited paper. To motivate our approach, let $m(\chi)$ be the regression function and observe that in view of (5) (and Lemma 1), $m(\chi)=m(\chi;\varphi^*)$, where φ^*

is the true function φ. Furthermore, by the definition of the regression function,

$$\min_{T:L^2([a,b])\to\mathbb{R}} E\big[Y-T(\boldsymbol{\chi})\big]^2 \overset{\text{def}}{=} E[Y-m(\boldsymbol{\chi})]^2 = E\big[Y-m(\boldsymbol{\chi};\varphi^*)\big]^2$$
$$= E\left[\frac{\Delta\cdot\big[Y-m(\boldsymbol{\chi};\varphi^*)\big]^2}{\pi_{\varphi^*}(\boldsymbol{\chi}|_s, Y)}\right], \tag{20}$$

where s is of the form $[a, t_0]$ for some unknown t_0, one can consider estimating s by minimizing an empirical version of the far right side of (20). Since φ^* is also unknown, we consider the following joint estimation of s and φ^* in our actual numerical studies

$$(\widehat{\varphi}_n, \widehat{s}) = \underset{s\in\{s_1,\ldots,s_q\},\ \varphi\in\mathcal{F}_{\varepsilon_n}}{\operatorname{argmin}} \ell^{-1}\sum_{i\in\mathcal{I}_\ell}\frac{\Delta_i}{\widehat{\pi}_\varphi(\boldsymbol{\chi}_i|_s, Y_i)}\left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi)-Y_i\right|^2, \tag{21}$$

where $s_j=[a,t_0^j]$, $j=1,\ldots,q$, for a grid of values of $a<t_0^1<t_0^2<\cdots<t_0^q=b$. Of course, the finer the grid is, the more accurate the estimator of s will be. Although this is a rough estimate of s, we believe that it can still provide good results for numerical work.

3 Applications to classification with partially labeled data

Here, we consider the following standard two-group classification problem. Let $(\boldsymbol{\chi}, Y)\in\mathbb{X}\times\{0,1\}$ be a random pair, where $\mathbb{X}$ may be an abstract space and Y, called the class label (or class variable), has to be predicted based on $\boldsymbol{\chi}$. More precisely, the aim of classification is to find a function $g:\mathbb{X}\to\{0,1\}$ for which the misclassification error probability,

$$L(g) := \mathbb{P}\{g(\boldsymbol{\chi})\neq Y\}, \tag{22}$$

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

$$g_{\text{B}}(\chi) = \begin{cases} 1 \text{ if } m(\chi) := \mathbb{E}\big(Y|\,\boldsymbol{\chi}=\chi\big) > \frac{1}{2}, \\ 0 \text{ otherwise,}\end{cases} \tag{23}$$

i.e., g_{B} has the smallest error probability given by $L(g_{\text{B}}) = \inf_{g:\mathbb{X}\to\{0,1\}} P\{g(\boldsymbol{\chi})\neq Y\}$; see, for example, Cérou and Guyader (2006), Abraham et al. (2006), and Devroye et al. (1996, Chap. 2). Since the distribution of $(\boldsymbol{\chi}, Y)$ is almost always unknown, finding the classifier g_{B} is virtually impossible. However, suppose that one has access to a random sample (the data) $\mathbb{D}_n=\{(\boldsymbol{\chi}_1, Y_1, \Delta_1),\ldots,(\boldsymbol{\chi}_n, Y_n, \Delta_n)\}$, where $\Delta_i=0$ if Y_i is missing (and $\Delta_i=1$ otherwise); here, Y_i's may be missing but not necessarily at random. Now, consider the regression estimator $\widehat{m}(\chi;\widehat{\varphi}_n)$ defined in (16), and denote the corresponding plug-in type version of (23) by

$$\widehat{g}_n(\chi\,;\widehat{\varphi}_n) := \begin{cases} 1 \text{ if } \widehat{m}(\chi;\widehat{\varphi}_n) \;>\; \frac{1}{2}, \\ 0 \text{ otherwise.} \end{cases} \tag{24}$$

The following result shows that the classifier defined via (24) is *almost completely* (and thus strongly) optimal in the usual sense that its misclassification error converges, almost completely, to that of the best classifier.

Theorem 2 *Consider the classifier* $\widehat{g}_n(\chi;\widehat{\varphi}_n)$ *given by* (24). *Then, under the conditions of Theorem* 1, *we have*

$$\mathbb{P}\left\{\widehat{g}_n(\boldsymbol{\chi};\widehat{\varphi}_n) \neq Y \middle| \mathbb{D}_n\right\} \xrightarrow{a.co.} \mathbb{P}\{g_{B}(\boldsymbol{\chi}) \neq Y\}, \quad as\, n \to \infty.$$

4 Numerical examples

In this section, we present some numerical examples in order to study and assess the finite-sample performance of our proposed regression estimators and their corresponding plug-in-type classification rules.

4.1 Example 1: regression

In what follows, we consider the proposed estimator $\widehat{m}(\chi;\widehat{\varphi}_n)$ in (16), the complete-case regression estimator $m_n^{cc}(\chi)$ in (2) that only uses the fully observed part of the data, as well as the estimator $m_n(\chi)$ in (1) based on the full data. The estimator $m_n(\chi)$ is included merely to see how much better the results would have been, had we not had any missing values. To perform the numerical studies, random samples of curves were generated according to $\boldsymbol{\chi}_i(t) = (t-0.5)^2 A_i + B_i$, $i = 1,\ldots,n$, where $t \in [0, 1]$, $A_i \sim N(5,\, 2^2)$ and $B_i \sim N(1,\, 0.5^2)$. For the purpose of simulations, each initial discretized curve was generated from 500 equispaced points $t \in [0, 1]$. A sample of 20 of these curves is provided in Fig. 1.

The response variable Y_i corresponding to $\boldsymbol{\chi}_i$, is taken to follow two possible models

Model A: $Y_i = \log\left(\int_0^1 \boldsymbol{\chi}_i^2(t)\,dt\right) + e_i$, where $e_i \sim N(0,\ 1)$.

Model B: $Y_i = \log\left(\int_0^1 \boldsymbol{\chi}_i^2(t)\,dt\right) + e_i$, where $e_i \sim N(0,\ 2^2)$, (High noise).

With respect to the choice of the functions φ and g in the missing probability mechanism (10), we considered $\varphi(y) = \exp(\gamma y)$, which is in the spirit of Kim and Yu (2011), and

$$g(\zeta) = \left\{\gamma_0 + \gamma_1 \log\left(\int_0^{0.6} \zeta^2(t)\,dt\right)\right\}, \quad \text{where} \quad \zeta = \chi\big|_{[0,0.6]} = \chi(t)\cdot \mathbb{1}_{\{0\le t\le 0.6\}}, \tag{25}$$

see Remark 3 at the end of this example for more on the choice of g. As for the choice of the coefficients $(\gamma_0, \gamma_1, \gamma)$ we considered $(-4.9,\ 0.05,\ 0.98)$ to produce

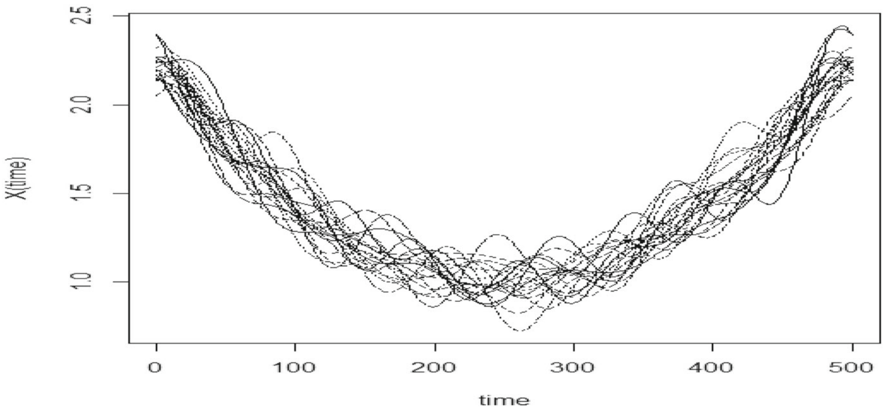

Fig. 1 A sample of covariate curves $\boldsymbol{\chi}_i$ for Example 4.1

Table 1 Empirical L^2 errors for models A and B with 50% missing response

Estimator	Model A			Model B (High Noise)		
	$n = 50$	$n = 100$	$n = 300$	$n = 50$	$n = 100$	$n = 300$
Complete-case	1.297	1.283	1.280	6.072	5.833	5.921
estimator: $m_n^{\text{cc}}(\chi)$	(0.0090)	(0.0061)	(0.0041)	(0.0494)	(0.0326)	(0.0189)
Proposed $\widehat{m}(\chi; \widehat{\varphi}_n)$	1.090	1.058	1.039	4.598	4.516	4.395
estimator: $\widehat{m}(\chi; \widehat{\varphi}_n)$	(0.0058)	(0.0037)	(0.0024)	(0.0248)	(0.0202)	(0.0173)
No missing data	1.019	1.018	1.003	4.113	4.044	4.025
estimator: $m_n(\chi)$	(0.0025)	(0.0022)	(0.0020)	(0.0124)	(0.0097)	(0.0088)

The numbers in parentheses are the standard errors computed over 400 Monte Carlo runs

approximately 25% missing response values, $(-6,\ 0.2,\ 1.5)$ that results in 50% missing vales, and $(-5,\ 0.14,\ 1.9)$ to produce 80% missing values. For each of the three sample sizes $n = 50, 100, 300$, the three regression estimators $\widehat{m}(\chi; \widehat{\varphi}_n)$, $m_n^{\text{cc}}(\chi)$, and $m_n(\chi)$ were constructed using a data-splitting ratio of $0.7n$ for the training sample and $0.30n$ for the testing sequence. Next, these three regression estimators were used to predict the response Y for a validation set of 1000 additional observations from the underlying distribution of the data. Here, we used the Epanechnikov-type kernel $\mathcal{K}(s) = \frac{3}{2}(1 - s^2)\mathbb{1}_{\{0 \leq s \leq 1\}}$, however, as in general nonparametric kernel regression estimation, the shape of the kernel is of little importance here. For our estimators and their corresponding smoothing parameters we employed the R package "fda.usc" developed by Febrero-Bande and Oviedo de la Fuente (2012), where the cross-validation option was used to estimate the smoothing parameters. Finally to assess the performance of these three regression estimators, we computed the empirical L^2 error of each estimator (for each sample size n) committed on the validation set of size 1000. This entire process was repeated a total of 400 times (each time using a sample of size n and a validation set of size 1000) and the average empirical L^2 error was computed. The results appear in Table 1.

Table 2 Empirical L^2 errors for models A and B with 80% missing data

Estimator	Model A			Model B (High Noise)		
	$n = 50$	$n = 100$	$n = 300$	$n = 50$	$n = 100$	$n = 300$
Complete-case estimator: $m_n^{\text{cc}}(\chi)$	2.151 (0.0246)	2.079 (0.0187)	2.041 (0.0124)	8.429 (0.0764)	8.539 (0.0499)	8.392 (0.0328)
Proposed estimator: $\widehat{m}(\chi; \widehat{\varphi}_n)$	1.623 (0.0271)	1.396 (0.0165)	1.224 (0.0062)	5.942 (0.0598)	5.724 (0.0418)	5.266 (0.0312)
No missing data estimator: $m_n(\chi)$	1.019 (0.0025)	1.018 (0.0022)	1.003 (0.0020)	4.113 (0.0124)	4.044 (0.0097)	4.025 (0.0088)

The numbers in parentheses are the standard errors computed over 400 Monte Carlo runs

Table 3 Empirical L^2 errors for models A and B with 25% missing data

Estimator	Model A			Model B (High Noise)		
	$n = 50$	$n = 100$	$n = 300$	$n = 50$	$n = 100$	$n = 300$
Complete-case estimator: $m_n^{\text{cc}}(\chi)$	1.061 (0.0043)	1.078 (0.0032)	1.053 (0.0026)	4.605 (0.0229)	4.592 (0.0197)	4.622 (0.0141)
Proposed estimator: $\widehat{m}(\chi; \widehat{\varphi}_n)$	1.030 (0.0030)	1.036 (0.0026)	1.006 (0.0022)	4.203 (0.0199)	4.089 (0.0108)	4.058 (0.0102)
No missing data estimator: $m_n(\chi)$	1.019 (0.0025)	1.018 (0.0022)	1.003 (0.0020)	4.113 (0.0124)	4.044 (0.0097)	4.025 (0.0088)

The numbers in parentheses are the standard errors computed over 400 Monte Carlo runs

The numbers appearing in parentheses are the standard errors computed over 400 Monte Carlo runs. The last row of this table gives the errors for the estimator $m_n(\chi)$ in (1) that involves no missing data. The second row of Table 1 shows that the proposed regression estimator $\widehat{m}(\chi; \widehat{\varphi}_n)$ performs quite well as its error rates follow closely those of the optimal case in the third row (with no missing values) under both models A and B, regardless of the sample size, and despite the fact that the data suffer from a 50% missing rate. The performance of the complete-case regression estimator $m_n^{\text{cc}}(\chi)$ in the first row of the table is rather poor; this should not be surprising because, as discussed earlier [immediately after Eq. (2)], this estimator is in general the "wrong" estimator.

Table 2 gives the same results for the case with 80% missing data. Once again the proposed estimator $\widehat{m}(\chi; \widehat{\varphi}_n)$ performs well in terms of its error being relatively close to the one with no missing values (the third row of the table) despite the 80% missing rate.

Next, Table 3 gives the same results for the case with 25% missing data. As this table shows, the performance of the proposed regression estimator $\widehat{m}(\chi; \widehat{\varphi}_n)$ for this case comes quite close to that of the one based on no missing data, i.e., $m_n(\chi)$.

Remark 3 The function g in (25) is clearly free of the portion $\chi|_{[0,0.4]}$ of the curve $\chi(t)$, $t \in [0, 1]$, which ensures that the missing probability mechanism (10) does not depend on $\chi|_{[0,0.4]}$. That is, the missingness of Y is not influenced by the portion

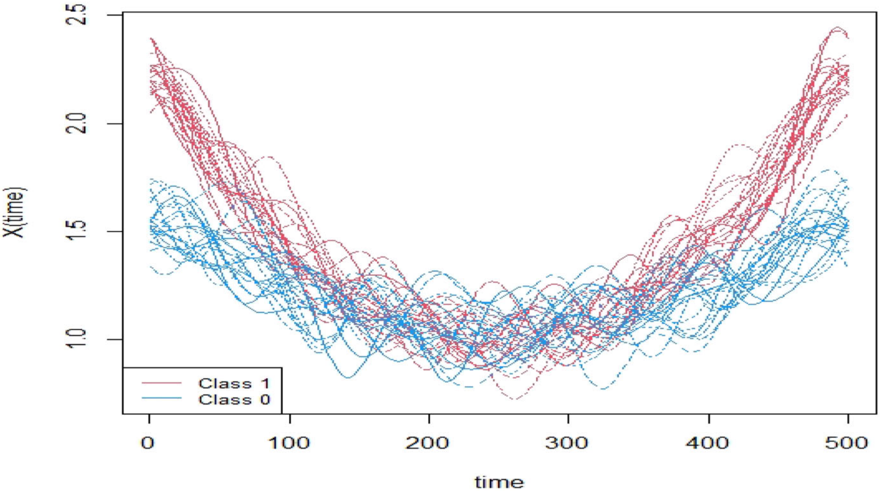

Fig. 2 A sample of curves $\boldsymbol{\chi}_i$ from each of the two classes for Example 4.2

$\chi|_{[0,0.4]}$ of the covariate curve. Although requiring g to be free of a portion of the curve $\chi(t)$, $t \in [0, 1]$, is part of the identifiability Assumption (A1), our additional numerical work (not reported here) shows that the final numerical results are not much influenced by the choice of $t_o = 0.4$ in the interval $[0, t_o]$; in fact, our numerical results (i.e., the reported empirical error rates) remain virtually the same for other choices of $t_o \in (0, 1)$. This is particularly important from an applied point of view where the true value of t_o may not be exactly known in advance.

4.2 Example 2: semi-supervised classification (partially observed labels)

In this example we consider the prediction of the class membership, $Y = 1$ or $Y = 0$, of any entity based on the functional predictor $\boldsymbol{\chi}(t)$, $t \in [0, 1]$, where $\boldsymbol{\chi}(t) = (t - 0.5)^2 A + B$, with $A \sim N(5, 2^2)$ and $B \sim N(1, 0.5^2)$ for class 1 (i.e., $Y = 1$), and $A \sim \text{Unif}(0, 4)$ and $B \sim \text{Unif}(0, 2.1)$ for class 0 (i.e., $Y = 0$). These distributions are similar to those in Rachdi and Vieu (2007). The parameters of the distributions of A and B are deliberately chosen in such a way that would make the task of classification rather difficult in this case; in fact, a sample of these curves and their class memberships in Fig. 2 reveals a significant overlap in large segments of the two sets of curves, thus making classification more challenging here.

Here, the class probabilities are taken to be $P(Y = 1) = P(Y = 0) = 0.5$. Furthermore, the function g in the missing probability mechanism (10) is the same as that in Example 4.1 given by (25), and the function $\varphi(y) = \exp(\exp(y\sqrt{\gamma}))$. Regarding the choice of the coefficients $(\gamma_0, \gamma_1, \gamma)$, we considered $(-1.3, 0.10, 0.5)$ and $(-3.3, 0.15, 0.95)$ to produce 60% and 30% missing rates, respectively. Next, using two different sample sizes, $n = 100$ and $n = 300$, we constructed the classifier $\widehat{g}_n(\chi\,; \widehat{\varphi}_n)$ that appears in (24). Additionally, we constructed the complete cases classifier, denoted by $g_n^{\text{cc}}(\chi)$, that replaces $\widehat{m}(\chi; \widehat{\varphi}_n)$ with (2) in (24), as well as the classifier based on no missing data,

Table 4 Misclassification errors of the three classifiers

Classifier	60% Missing rate		30% Missing rate	
	$n = 100$	$n = 300$	$n = 100$	$n = 300$
Complete-case	0.448	0.495	0.352	0.462
classifier: $g_n^{\text{cc}}(\chi)$	(0.0063)	(0.0015)	(0.0094)	(0.0046)
Proposed	0.216	0.206	0.209	0.197
classifier: $\widehat{g}_n(\chi; \widehat{\varphi}_n)$	(0.0090)	(0.0085)	(0.0086)	(0.0080)
No missing data	0.196	0.185	0.196	0.185
classifier: $g_n(\chi)$	(0.0025)	(0.0013)	(0.0025)	(0.0013)

The numbers in parentheses are the standard errors computed over 400 Monte Carlo runs

denoted by $g_n(\chi)$, which replaces $\widehat{m}(\chi; \widehat{\varphi}_n)$ by (1) in (24); the reason for including the classifier based on no missing data [i.e., $g_n(\chi)$] is to see (i) how far off the other two classifiers are in terms of their error rates, and (ii) what the results would have been, had we not had any missing Y_i's in the data. As in Example 4.1, here we used the Epanechnikov kernel, where once again the cross-validation option of the R package "fda.usc" of Febrero-Bande and Oviedo de la Fuente (2012) was used to estimate the smoothing parameter of the kernel.

Next, these three classifiers were used to classify a validation data set of 1000 additional observations generated from the underlying distribution of the data (with 500 from each class); this was done for each of the two sample sizes. The entire above process was repeated a total of 400 times (each time using a sample of size n and a validation set of size 1000) and the average misclassification errors were computed. The results appear in Table 4 along with their standard errors in parentheses. Table 4 shows that the classifier $\widehat{g}_n(\chi\,; \widehat{\varphi}_n)$ has the ability to perform well, as compared to the classifier with no missing data, $g_n(\chi)$, despite the fact that it suffers from huge missing rates. In passing, we note that the complete-case classifier is performing quite poorly; as noted after Eq. (2), this is because $m_n^{\text{cc}}(\chi)$ in (2) is actually the "wrong" estimator of $m(\chi) = \mathbb{P}\{Y = 1 | \boldsymbol{\chi} = \chi\}$ in the sense that $m_n^{\text{cc}}(\chi)$ estimates the quantity $\mathbb{E}(\Delta Y | \boldsymbol{\chi} = \chi)/\mathbb{E}(\Delta | \boldsymbol{\chi} = \chi)$ which is not in general equal to the regression function $m(\chi)$ and can, in fact, be larger that 1 for a probability! As a result, as n increases from 100 to 300, the error of the complete-case classifier tends to get worse (with larger standard errors), as it tends to get closer to the error of a completely incorrect classifier.

4.3 Example 3: spectrometric data

In the previous two examples we assumed that s, the interval related to identifiability Assumption (A1), was known from prior information or studies; see (25) where we had $s = [0, 0.6]$. Here, we consider estimating s based on the method discussed in Sect. 2.2.

This data set consists of pairs of measurements $(\boldsymbol{\chi}_i, Y_i)$, $i = 1, \ldots, 215$, on 215 pieces of finely chopped meat, where $\boldsymbol{\chi}_i$ is the spectrometric curve corresponding to the absorbance measured at 100 wavelengths (thus $\boldsymbol{\chi}_i = (\chi_i(t_1), \ldots, \chi_i(t_{100}))$ for the ith meat sample. Here, Y_i is a measure of the fat content of the ith meat sample.

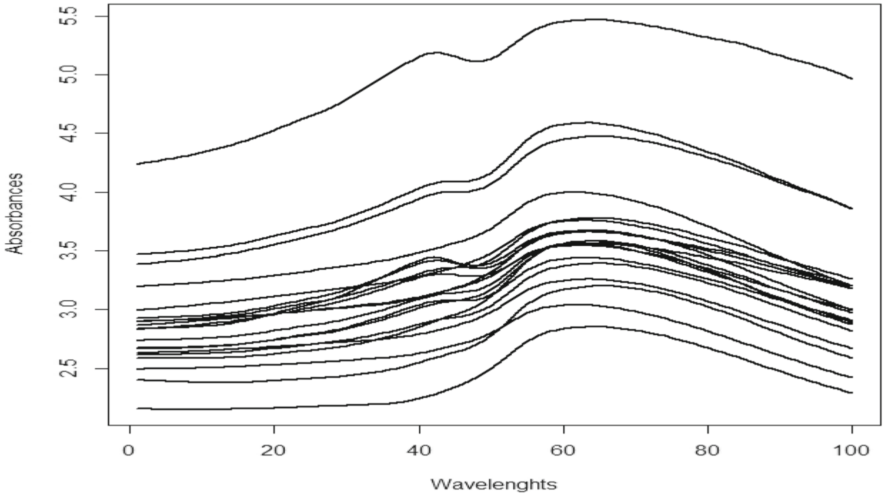

Fig. 3 A sample of 20 data curves $\boldsymbol{\chi}_i$ from Example 4.3

This real data set and information about its origin can be found at https://www.math.univ-toulouse.fr/~ferraty/SOFTWARES/NPFDA/ under *Datasets*. A sample of 20 of these curves appears in Fig. 3.

The question of interest is to predict Y based on $\boldsymbol{\chi}$. To proceed, we randomly select $n = 170$ of the measurement pairs $(\boldsymbol{\chi}_i, Y_i)$ to be used as the data while the remaining 45 pairs are used as the validation set. For missing values, we generated Δ_i's using (10) with

$$g(\zeta) = \gamma_0 + \gamma_1 \int_0^{t_{50}} \zeta^2(t)\, dt, \quad \text{where} \quad \zeta = \chi\big|_{[0,t_{50}]} = \chi(t) \cdot \mathbb{1}_{\{0 \le t \le t_{50}\}}, \qquad (26)$$

and $\varphi(y) = \exp(\gamma y)$. Thus the true s for $\zeta = \chi\big|_s$ is $[0, t_{50}]$ in (10). We took $\gamma_0 = 4$, $\gamma_1 = 0.16$, and $\gamma = -0.5$ corresponding to approximately 50% missing data. For the purpose of presentation, we re-scaled the frequencies to fall in the interval $[0,\ 1]$, thus $t_1 = [0,\ 0.01]$, $t_2 = [0, 0.02], \ldots, t_{100} = [0, 1]$. Therefore, $s = [0, t_{50}] = [0, 0.50]$ in (26). Finally, the three regression estimators $\widehat{m}(\chi; \widehat{\varphi}_n)$, $m_n^{\text{cc}}(\chi)$, and $m_n(\chi)$ were constructed using the Epanechnikov-type kernel and a data-splitting ratio of $0.7n$ to $0.3n$ as in the previous examples. These three regression estimators were used to predict the response Y in the validation set of size 45, and the empirical L^2 error was computed. Repeating this process 400 times, each time dividing the 215 observations randomly into a data set of size $n = 170$ and a validation set of size 45, we computed the average empirical L^2 error (over 400 runs). The results appear in the top row of Table 5 where $t = 0.50$. Once again, the proposed regression estimator $\widehat{m}(\chi; \widehat{\varphi}_n)$ tends to perform relatively well. Representing the estimate of s by $\widehat{s} = [0, \hat{t}\,]$, the table also gives the average value of $\hat{t}$, over 400 runs, where in each run $\widehat{s}$ was selected from the set $\{s_1, \ldots, s_{20}\}$ to minimize (21); here $s_1 = [0,\ 0.05]$, $s_2 = [0,\ 0.10], \ldots, s_{19} = [0,\ 0.95]$, $s_{20} = [0,\ 1]$. In passing, we observe that the average value of $\hat{t}$ $(= 0.48)$ over 400 runs is relatively close to the true $t = 0.50$. The bottom row of the table

Table 5 Empirical L^2 errors for the data in Example 4.3

$s = [0, t]$	$\widehat{m}(\chi; \widehat{\varphi}_n)$	$\hat{t}$	$m_n^{\mathrm{cc}}(\chi)$	$m_n(\chi)$
$t = 0.50$	136.65	0.48	182.89	101.90
	(1.600)	(0.015)	(2.118)	(1.681)
$t = 0.20$	140.08	0.27	177.12	101.90
	(2.262)	(0.011)	(2.408)	(1.681)

The numbers in parentheses are the standard errors computed over 400 Monte Carlo runs

gives the same result for the case where $s = [0, 0.20]$ with $\gamma_0 = 4.84$, $\gamma_1 = 0.18$, and $\gamma = -0.5$ for 50% missing data. In this case, the average value of $\hat{t}$ over 400 runs is about 0.27. The values appearing in parentheses are the standard errors.

5 Proofs of the main results

We start by stating a number of lemmas. In what follows, we use the notation of Sect. 2.1 and let $\mathcal{F}_\varepsilon$ be any ε-cover of $\mathcal{F}$ (as defined in Sect. 2.1). Next, for each $\varphi \in \mathcal{F}$, define

$$\widehat{L}_{m,\ell}(\varphi) := \frac{1}{\ell} \sum_{i \in \mathcal{I}_\ell} \frac{\Delta_i}{\widehat{\pi}_\varphi(\boldsymbol{\zeta}_i, Y_i)} \left|\widehat{m}_m(\boldsymbol{\chi}_i; \varphi) - Y_i\right|^2, \tag{27}$$

where $\widehat{\pi}_\varphi(\zeta, y)$ is as in (13) and $\widehat{m}_m(\chi; \varphi)$ is given by (6). Also, let $m(\chi; \varphi)$ be as in (5) and put

$$\varphi_\varepsilon := \underset{\varphi \in \mathcal{F}_\varepsilon}{\operatorname{argmin}}\ \mathbb{E}\left|m(\boldsymbol{\chi}; \varphi) - Y\right|^2 \quad \text{and} \quad \widehat{\varphi}_\varepsilon := \underset{\varphi \in \mathcal{F}_\varepsilon}{\operatorname{argmin}}\ \widehat{L}_{m,\ell}(\varphi). \tag{28}$$

Lemma 1 *Let φ^* be the true (unknown) version of the function φ in (10). Also, let $m(\chi; \varphi)$ be as in (5). Then the regression function $m(\chi) = \mathbb{E}[Y | \boldsymbol{\chi} = \chi]$ can be represented as*

$$m(\chi) := m(\chi; \varphi^*) = \eta_1(\chi) + \frac{\psi_1(\chi; \varphi^*)}{\psi_2(\chi; \varphi^*)} \cdot (1 - \eta_2(\chi)). \tag{29}$$

where the functions ψ_k and η_k, $k = 1, 2$, are given by (4).

Proof of Lemma 1. The proof of this lemma is straightforward and therefore omitted. □

Lemma 2 *Let $m(\chi; \varphi_j)$, $j = 1, 2$, be defined as in (5), where $\varphi_j : [-L, L] \to (0, B]$ for some positive number B. Then, under Assumptions* (A7) *and* (A9), *one has*

$$\mathbb{E}\left|m(\boldsymbol{\chi}; \varphi_1) - m(\boldsymbol{\chi}; \varphi_2)\right| \leq C \cdot \sup_{-L \leq y \leq L} \left|\varphi_1(y) - \varphi_2(y)\right|,$$

where the constant $C>0$ *can be taken to be* $C=2L/\varrho_0$, *with* ϱ_0 *as in Assumption* (A7).

Proof of Lemma 2. Let $S_j(\chi)=\mathbb{E}[\Delta Y\varphi_j(Y)|\boldsymbol{\chi}=\chi]$ and $T_j(\chi)=\mathbb{E}[\Delta\varphi_j(Y)|\boldsymbol{\chi}=\chi]$, $j=1,2$, and observe that

$$\begin{aligned}\Big|m(\chi;\varphi_1)-m(\chi;\varphi_2)\Big| &= \left|\frac{-S_1(\chi)}{T_1(\chi)}\cdot\frac{T_1(\chi)-T_2(\chi)}{T_2(\chi)}+\frac{S_1(\chi)-S_2(\chi)}{T_2(\chi)}\right|\cdot\mathbb{E}\big[1-\delta|\boldsymbol{\chi}=\chi\big]\\ &\le \frac{1}{T_2(\chi)}\big\{L\big|T_1(\chi)-T_2(\chi)\big|+\big|S_1(\chi)-S_2(\chi)\big|\big\}.\end{aligned}$$

But, $|S_1(\chi)-S_2(\chi)|\le\ \mathbb{E}\left[|\Delta Y|\cdot\big|\varphi_1(Y)-\varphi_2(Y)\big|\,\big|\boldsymbol{\chi}=\chi\right]\le\ L\ \sup_{-L\le y\le L}\big|\varphi_1(y)-\varphi_2(y)\big|$. Similarly, $|T_1(\chi)-T_2(\boldsymbol{x})|\ \le\ \sup_{-L\le y\le L}\big|\varphi_1(y)-\varphi_2(y)\big|$. On the other hand, by the second part of Assumption (A7), we have $T_2(\chi)\ge\varrho_0>0$. Therefore

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

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

Lemma 3 *Let* $m(\chi;\varphi)$, $\widehat{L}_{m,\ell}(\varphi)$, φ_ε, *and* $\widehat{\varphi}_\varepsilon$ *be as in* (5), (27), *and* (28), *respectively. Then, under the conditions of Theorem* 1, *we have*

$$\begin{aligned}\mathbb{E}\left[\Big|\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-m(\boldsymbol{\chi};\varphi_\varepsilon)\Big|^2\Big|\mathbb{D}_n\right]\le&\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\mathbb{E}\left[\Big|\widehat{m}_m(\boldsymbol{\chi};\varphi)-Y\Big|^2\Big|\mathbb{D}_m\right]-\widehat{L}_{m,\ell}(\varphi)\right|\\ &+\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\varphi)-\mathbb{E}\Big|m(\boldsymbol{\chi};\varphi)-Y\Big|^2\right|\\ &+C_1\,\varepsilon^{1/2},\end{aligned}\tag{30}$$

where C_1 *is a positive constant not depending on n or* ε, *and* $\widehat{m}_m(\boldsymbol{\chi};\varphi)$ *is as in* (6).

Proof of Lemma 3. Start with the simple decomposition $\mathbb{E}\big[|\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-Y|^2\big|\mathbb{D}_n\big]=\mathbb{E}\big[|\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-m(\boldsymbol{\chi};\varphi_\varepsilon)|^2\big|\mathbb{D}_n\big]+\mathbb{E}|m(\boldsymbol{\chi};\varphi_\varepsilon)-Y|^2+2\mathbb{E}\big[\big(\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-m(\boldsymbol{\chi};\varphi_\varepsilon)\big)\big(m(\boldsymbol{\chi};\varphi_\varepsilon)-Y\big)\big|\mathbb{D}_n\big]$. Also, let φ^* be as in (18) and observe that

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

where we have used the fact that in view of (29), $\mathbb{E}[Y|\boldsymbol{\chi}=\chi]:=m(\chi)=m(\chi;\varphi^*)$. Therefore

$$\mathbb{E}\left[\Big|\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-m(\boldsymbol{\chi};\varphi_\varepsilon)\Big|^2\Big|\mathbb{D}_n\right]=\left\{\mathbb{E}\left[\Big|\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-Y\Big|^2\Big|\mathbb{D}_n\right]-\mathbb{E}\Big|m(\boldsymbol{\chi};\varphi_\varepsilon)-Y\Big|^2\right\}$$

$$
\begin{aligned}
&-2\mathbb{E}\left[\big(\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-m(\boldsymbol{\chi};\varphi_\varepsilon)\big)\big(m(\boldsymbol{\chi};\varphi_\varepsilon)-m(\boldsymbol{\chi};\varphi^*)\big)\Big|\mathbb{D}_n\right]\\
&:=\mathbf{I}_n+\mathbf{II}_n. \qquad (31)
\end{aligned}
$$

Now, observe that

$$
\begin{aligned}
\mathbf{I}_n &= \mathbb{E}\left[\left|\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-Y\right|^2\Big|\mathbb{D}_n\right]-\inf_{\varphi\in\mathcal{F}_\varepsilon}\mathbb{E}\left|m(\boldsymbol{\chi};\varphi)-Y\right|^2\\
&=\sup_{\varphi\in\mathcal{F}_\varepsilon}\Bigg\{\mathbb{E}\left[\left|\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-Y\right|^2\Big|\mathbb{D}_n\right]-\widehat{L}_{m,\ell}(\varphi)+\widehat{L}_{m,\ell}(\varphi)-\widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon)\\
&\qquad +\widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon)-\mathbb{E}\left|m(\boldsymbol{\chi};\varphi)-Y\right|^2\Bigg\}, \quad (\text{where}\widehat{L}_{m,\ell}(\varphi)\text{ is as in }(27))\\
&\le\left(\mathbb{E}\left[\left|\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-Y\right|^2\Big|\mathbb{D}_n\right]-\widehat{L}_{m,\ell}(\widehat{\varphi}_\varepsilon)\right)+\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\varphi)-\mathbb{E}\left|m(\boldsymbol{\chi};\varphi)-Y\right|^2\right|,
\end{aligned}
$$

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 (28)]. Therefore,

$$
\left|\mathbf{I}_n\right|\le\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\mathbb{E}\left[\left|\widehat{m}_m(\boldsymbol{\chi};\varphi)-Y\right|^2\Big|\mathbb{D}_m\right]-\widehat{L}_{m,\ell}(\varphi)\right|+\sup_{\varphi\in\mathcal{F}_\varepsilon}\left|\widehat{L}_{m,\ell}(\varphi)-\mathbb{E}\left|m(\boldsymbol{\chi};\varphi)-Y\right|^2\right|, \tag{32}
$$

where the conditioning on $\mathbb{D}_m$ in the above expression reflects the fact that $\widehat{m}_m(\boldsymbol{\chi};\varphi)$ depends on $\mathbb{D}_m$ only (and not the entire data $\mathbb{D}_n$). Furthermore, the term $\mathbf{II}_n$ in (31) can be bounded as follows.

$$
\begin{aligned}
\left|\mathbf{II}_n\right|&\le 2\mathbb{E}\left[\left|\widehat{m}_m(\boldsymbol{\chi};\widehat{\varphi}_\varepsilon)-m(\boldsymbol{\chi};\varphi_\varepsilon)\right|\cdot\left|m(\boldsymbol{\chi};\varphi_\varepsilon)-m(\boldsymbol{\chi};\varphi^*)\right|\Big|\mathbb{D}_n\right]\\
&\le 6L\cdot\mathbb{E}\left|m(\boldsymbol{\chi};\varphi_\varepsilon)-m(\boldsymbol{\chi};\varphi^*)\right| \le 6L\sqrt{\mathbb{E}\left|m(\boldsymbol{\chi};\varphi_\varepsilon)-m(\boldsymbol{\chi};\varphi^*)\right|^2}. \qquad (33)
\end{aligned}
$$

But, using the identity $\mathbb{E}\left|m(\boldsymbol{\chi};\varphi_\varepsilon)-Y\right|^2=\mathbb{E}\left|m(\boldsymbol{\chi};\varphi^*)-Y\right|^2+\mathbb{E}\left|m(\boldsymbol{\chi};\varphi_\varepsilon)-m(\boldsymbol{\chi};\varphi^*)\right|^2$, we have

$$
\begin{aligned}
\mathbb{E}\left|m(\boldsymbol{\chi};\varphi_\varepsilon)-m(\boldsymbol{\chi};\varphi^*)\right|^2&=\inf_{\varphi\in\mathcal{F}_\varepsilon}\mathbb{E}\left|m(\boldsymbol{\chi};\varphi)-Y\right|^2-\mathbb{E}\left|m(\boldsymbol{\chi};\varphi^*)-Y\right|^2\\
&=\inf_{\varphi\in\mathcal{F}_\varepsilon}\mathbb{E}\left|m(\boldsymbol{\chi};\varphi)-m(\boldsymbol{\chi};\varphi^*)\right|^2\\
&\le 2L\inf_{\varphi\in\mathcal{F}_\varepsilon}\mathbb{E}\left|m(\boldsymbol{\chi};\varphi)-m(\boldsymbol{\chi};\varphi^*)\right|. \qquad (34)
\end{aligned}
$$

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, in view of Lemma 2 and the fact that the right side of (34) is an infimum, one finds

$$
(\text{Rght side of }(34))\le 2L\cdot\mathbb{E}\left|m(\boldsymbol{\chi};\varphi^\dagger)-m(\boldsymbol{\chi};\varphi^*)\right|\le 2LC\sup_{-L\le y\le L}\left|\varphi^\dagger(y)-\varphi^*(y)\right|
$$

$$\le 2LC \cdot \varepsilon \quad \big(\text{because}\, \varphi^{*} \in B(\varphi^{\dagger}, \varepsilon)\big), \tag{35}$$

where C is as in Lemma 2. Therefore, by (33) and (34), we have

$$\big|\mathbf{I\!I}_n\big| \le 6L\sqrt{2LC \cdot \varepsilon} =: C_1\sqrt{\varepsilon}. \tag{36}$$

Now Lemma 3 follows from (31), (32), and (36). □

Lemma 4 *Let $\mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})$ and $\mathcal{N}_\tau(\mathcal{S}^{o}_{\mathbb{X}})$ be the τ-covering numbers of $\mathcal{S}_{\mathbb{X}}$ and $\mathcal{S}^{o}_{\mathbb{X}}$, $\tau > 0$, where $\mathcal{S}^{o}_{\mathbb{X}}$ is as in Assumption (A2). Then for all $\tau > 0$, we have $\mathcal{N}_\tau(\mathcal{S}^{o}_{\mathbb{X}}) \le \mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})$.*

Proof of Lemma 4. Let $\chi_1, \ldots, \chi_{\mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})}$ be a τ-cover for $\mathcal{S}_{\mathbb{X}}$, i.e., $\mathcal{S}_{\mathbb{X}} \subset \bigcup_{j=1}^{\mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})} B(\chi_j, \tau)$. Now, observe that, with $\mathbb{X} = L^2([a,b])$, $\infty < a < b < \infty$ and $s = [a, t_o]$, $t_o \in [a,b]$, for any $\chi \in \mathcal{S}_{\mathbb{X}}$, we have

$$\begin{aligned}\min_{1 \le j \le \mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})} \big\|(\chi_j - \chi)|_s\big\|_{L^2(s)} &= \min_{1 \le j \le \mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})} \left[\int_s \big|(\chi_j - \chi)|_s\big|^2\right]^{1/2} \\ &= \min_{1 \le j \le \mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})} \left[\int_{[a,b]} \big|\chi_j - \chi\big|^2 \cdot \mathbb{1}_s\right]^{1/2} \le \tau,\end{aligned}$$

where $\mathbb{1}_s$ is the indicator function of the set s; thus the restrictions, $\chi_j\big|_s$, $j = 1, \ldots, \mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})$, form a τ-cover of $\mathcal{S}^{o}_{\mathbb{X}}$, and this completes the proof of the lemma. □

Lemma 5 *Suppose that Assumptions (A0) and (A2)–(A9) hold, and let $\tau_m = \log m/m$.* **(i)** *Let ψ_k and η_k be as in (4). Also, let $\widehat{\psi}_k$ and $\widehat{\eta}_k$ be as in (7) and (8). Then, for $k = 1, 2$,*

$$\sup_{\varphi \in \mathcal{F}_{\varepsilon m}} \sup_{\chi \in \mathcal{S}_{\mathbb{X}}} \left|\widehat{\psi}_{m,k}(\chi;\varphi) - \psi_k(\chi;\varphi)\right| = \mathcal{O}\big(h^{\beta_k}\big) + \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F}) \vee \mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m \cdot \phi_1(h)}}\right) \tag{37}$$

$$\sup_{\chi \in \mathcal{S}_{\mathbb{X}}} \left|\widehat{\eta}_{m,k}(\chi) - \eta_k(\chi)\right| = \mathcal{O}\big(h^{\beta_k}\big) + \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m \cdot \phi_1(h)}}\right), \tag{38}$$

where β_1 and β_2 are the positive constants in Assumption (A3) and $\mathcal{N}_\tau$ is as in Assumption (A6).
(ii) *Let ψ_o and η_o be as in (11); also, let $\widehat{\psi}_{m,o}$ and $\widehat{\eta}_{m,o}$ be as in (12). Then*

$$\sup_{\varphi \in \mathcal{F}_{\varepsilon m}} \sup_{\zeta \in \mathcal{S}^{o}_{\mathbb{X}}} \left|\widehat{\psi}_{m,o}(\zeta;\varphi) - \psi_o(\zeta;\varphi)\right| = \mathcal{O}\big(h^{\beta_o}\big)$$

$$+\mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F})\vee\mathcal{N}_{\tau_m}(\mathcal{S}^o_{\mathbb{X}})\right]}{m\cdot\phi_0(h)}}\right), \tag{39}$$

$$\sup_{\zeta\in\mathcal{S}^o_{\mathbb{X}}}\left|\widehat{\eta}_{m,o}(\zeta)-\eta_o(\zeta)\right| = \mathcal{O}\big(h^{\beta_o}\big)+\mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\tau_m}(\mathcal{S}^o_{\mathbb{X}})\right]}{m\cdot\phi_0(h)}}\right), \tag{40}$$

where β_o is as in Assumption (A3) and $\mathcal{S}^o_{\mathbb{X}}$ is as in Assumption (A2).

Proof of Lemma 5 We start with the proof of (37). Here, we employ some of the arguments used in Ferraty et al. (2010). Le ψ_k and $\widehat{\psi}_k$ be as in (4) and (7), respectively, and observe that

$$\begin{aligned}\widehat{\psi}_{m,1}(\chi;\varphi)-\psi_1(\chi;\varphi) = \frac{1}{\widehat{f}_m(\chi)}\Big\{&\Big[\widehat{g}_m(\chi;\varphi)-\mathbb{E}\big(\widehat{g}_m(\chi;\varphi)\big)\Big]\\ &+\Big[\mathbb{E}\big(\widehat{g}_m(\chi;\varphi)\big)-\psi_1(\chi;\varphi)\Big]+\Big[1-\widehat{f}_m(\chi)\Big]\cdot\psi_1(\chi;\varphi)\Big\},\end{aligned} \tag{41}$$

where

$$\begin{aligned}\widehat{f}_m(\chi) &= \frac{\sum_{i\in\mathcal{I}_m}\mathcal{K}\big(h^{-1}d(\chi,\ \boldsymbol{\chi}_i)\big)}{m\mathbb{E}\big[\mathcal{K}\big(h^{-1}d(\chi,\ \boldsymbol{\chi}_1)\big)\big]} \quad \text{and}\\ \widehat{g}_m(\chi;\varphi) &= \frac{\sum_{i\in\mathcal{I}_m}\Delta_i Y_i\varphi(Y_i)\,\mathcal{K}\big(h^{-1}d(\chi,\ \boldsymbol{\chi}_i)\big)}{m\mathbb{E}\big[\mathcal{K}\big(h^{-1}d(\chi,\ \boldsymbol{\chi}_1)\big)\big]}.\end{aligned}$$

Let $\widetilde{\chi}_j$, $j=1,\ldots,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})$ be a τ_m-cover for $\mathcal{S}_{\mathbb{X}}$, i.e., $\mathcal{S}_{\mathbb{X}}\subset\bigcup_{j=1}^{\mathcal{N}_\tau(\mathcal{S}_{\mathbb{X}})} B\left(\widetilde{\chi}_j,\tau_m\right)$, where $\tau_m=\log m/m$ as before, and start with the basic decomposition

$$\begin{aligned}\sup_{\varphi\in\mathcal{F}_{\varepsilon_m}}\sup_{\chi\in\mathcal{S}_{\mathbb{X}}}\left|\widehat{g}_m(\chi;\varphi)-\mathbb{E}\big(\widehat{g}_m(\chi;\varphi)\big)\right| \le{}& \sup_{\varphi\in\mathcal{F}_{\varepsilon_m}}\max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})}\\ &\sup_{\chi\in B(\widetilde{\chi}_j,\tau_m)}\left|\widehat{g}_m(\chi;\varphi)-\widehat{g}_m(\widetilde{\chi}_j;\varphi)\right|\\ &+\sup_{\varphi\in\mathcal{F}_{\varepsilon_m}}\max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})}\sup_{\chi\in B(\widetilde{\chi}_j,\tau_m)}\left|\mathbb{E}\left[\widehat{g}_m(\chi;\varphi)\right]-\mathbb{E}\left[\widehat{g}_m(\widetilde{\chi}_j;\varphi)\right]\right|\\ &+\sup_{\varphi\in\mathcal{F}_{\varepsilon_m}}\max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})}\left|\widehat{g}_m(\widetilde{\chi}_j;\varphi)-\mathbb{E}\left[\widehat{g}_m(\widetilde{\chi}_j;\varphi)\right]\right|\\ :={}& I_m+II_m+III_m.\end{aligned} \tag{42}$$

It can be shown (see Ferraty and Vieu 2006, Lemma 4.4) that in view of Assumptions (A2) and (A5a) there are constants $0<C'<C''<\infty$ such that

$$\forall\ \chi\in\mathcal{S}_{\mathbb{X}},\quad C'\phi_1(h)<\mathbb{E}\big[\mathcal{K}\big(h^{-1}d(\chi,\ \boldsymbol{\chi}_1)\big)\big]<C''\phi_1(h).$$

Now, this observation together with the fact that $|\Delta_i Y_i \varphi(Y_i)| \leq LB$ for all $i = 1, \ldots, n$, implies that

$$\begin{aligned}
&\sup_{\chi \in B(\tilde{\chi}_j,\, \tau_m)} \left| \widehat{g}_m(\chi; \varphi) - \widehat{g}_m(\tilde{\chi}_j; \varphi) \right| \\
&= \sup_{\chi \in B(\tilde{\chi}_j,\, \tau_m)} \frac{1}{m} \left| \frac{\sum_{i \in \mathcal{I}_m} \Delta_i Y_i \varphi(Y_i)\, \mathcal{K}\big(h^{-1} d(\chi,\, \boldsymbol{\chi}_i)\big)}{\mathbb{E}\big[\mathcal{K}\big(h^{-1} d(\chi,\, \boldsymbol{\chi}_1)\big)\big]} \right. \\
&\quad \left. - \frac{\sum_{i \in \mathcal{I}_m} \Delta_i Y_i \varphi(Y_i)\, \mathcal{K}\big(h^{-1} d(\tilde{\chi}_j,\, \boldsymbol{\chi}_i)\big)}{\mathbb{E}\big[\mathcal{K}\big(h^{-1} d(\tilde{\chi}_j,\, \boldsymbol{\chi}_1)\big)\big]} \right| \\
&\leq \sup_{\chi \in B(\tilde{\chi}_j,\, \tau_m)} \frac{CLB}{\phi_1(h)} \cdot \frac{1}{m} \sum_{i \in \mathcal{I}_m} \left| \mathcal{K}\big(h^{-1} d(\chi,\, \boldsymbol{\chi}_i)\big) - \mathcal{K}\big(h^{-1} d(\tilde{\chi}_j,\, \boldsymbol{\chi}_i)\big) \right| \\
&\quad \cdot \mathbb{1}_{\left\{\boldsymbol{\chi}_i \in \left[B(\chi,\, h) \cup B(\tilde{\chi}_j,\, h)\right]\right\}} \\
&\leq \sup_{\chi \in B(\tilde{\chi}_j,\, \tau_m)} \frac{CLB}{\phi_1(h)} \cdot \frac{1}{m} \cdot \frac{\tau_m}{h} \sum_{i \in \mathcal{I}_m} \mathbb{1}_{\left\{\boldsymbol{\chi}_i \in \left[B(\chi,\, h) \cup B(\tilde{\chi}_j,\, h)\right]\right\}},
\end{aligned} \tag{43}$$

where the last line follows because $\mathcal{K}$ is Lipschitz on [0, 1] which implies that

$$\left| \mathcal{K}\big(h^{-1} d(\chi,\, \boldsymbol{\chi}_i)\big) - \mathcal{K}\big(h^{-1} d(\tilde{\chi}_j,\, \boldsymbol{\chi}_i)\big) \right| \leq \frac{1}{h} d(\boldsymbol{\chi},\, \tilde{\chi}_j) \leq \frac{\tau_m}{h}, \quad \forall\ \boldsymbol{\chi} \in B(\tilde{\chi}_j,\, \tau_m).$$

However, if $\chi \in B(\tilde{\chi}_j,\, \tau_m)$, where $\tau_m := \log m/m \leq h$, then one finds $B(\chi,\, h) \cup B(\tilde{\chi}_j,\, h) \subset B(\tilde{\chi}_j,\, 2h)$. Consequently

$$\text{(Right side of (43))} \leq \frac{CLB}{\phi_1(h)} \cdot \frac{1}{m} \cdot \frac{\tau_m}{h} \sum_{i \in \mathcal{I}_m} \underbrace{\mathbb{1}_{\left\{\boldsymbol{\chi}_i \,\in B(\tilde{\chi}_j,\, 2h)\right\}}}_{\text{free of}\chi} := \frac{C_1}{m} \sum_{i \in \mathcal{I}_m} Z_{ij}, \tag{44}$$

where

$$Z_{ij} = \frac{\tau_m}{h \cdot \phi_1(h)}\, \mathbb{1}_{\left\{\boldsymbol{\chi}_i \,\in B(\tilde{\chi}_j,\, 2h)\right\}}. \tag{45}$$

Furthermore, using Assumption (A2), one immediately finds

$$\mathbb{E}(Z_{ij}) = C_2 \tau_m \phi_1(2h)/[h\phi_1(h)] \quad \text{and} \quad \mathbb{E}(Z_{ij}^2) = C_2 \tau_m^2 \phi_1(2h)/[h^2 \phi_1^2(h)],$$

where C_2 is a positive constant not depending on n. Also, one finds $\mathrm{Var}(Z_{ij}) = \mathbb{E}(Z_{ij}^2) - [\mathbb{E}(Z_{ij})]^2 = C_2 \tau_m^2 \phi_1(2h) \cdot [1 - C_2 \phi_1(2h)]/[h^2 \phi_1^2(h)]$. Therefore, in view of (43) and (44) (and upon replacing Z_{ij} by $Z_{ij} - \mathbb{E}(Z_{ij}) + \mathbb{E}(Z_{ij})$ in (44)), one arrives at

$$\sup_{\chi\in B(\tilde{\chi}_j,\,\tau_m)}\left|\widehat{g}_m(\chi;\varphi)-\widehat{g}_m(\tilde{\chi}_j;\varphi)\right|\le\frac{1}{m}\left|\sum_{i\in\boldsymbol{\mathcal{I}}_m}Z'_{ij}\right|+\mathcal{O}\left(\frac{\tau_m\phi_1(2h)}{h\phi_1(h)}\right),\tag{46}$$

where $Z'_{ij}=C_1[Z_{ij}-\mathbb{E}(Z_{ij})]$ and where the big-O term does not depend on χ, $\tilde{\chi}_j$, or φ. Furthermore, in view of the last part of Assumption (A6)(ii), it is not hard to see that for m large enough,

$$\mathbb{E}\left|Z'_{ij}\right|^k\ \le\ C_k\left(\frac{\tau_m\sqrt{\phi_1(2h)}}{h\,\phi_1(h)}\right)^{2(k-1)},\quad\text{for all } k\ge 2.$$

Therefore, by Corollary A.8 of Ferraty and Vieu (2006), for any $t>0$

$$\begin{aligned}&\mathbb{P}\left\{\sup_{\varphi\in\mathcal{F}_{\varepsilon m}}\ \max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})}\frac{1}{m}\left|\sum_{i\in\boldsymbol{\mathcal{I}}_m}Z'_{ij}\right|\ge\ t\right\}\\&\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})}\mathbb{P}\left\{\frac{1}{m}\left|\sum_{i\in\boldsymbol{\mathcal{I}}_m}Z'_{ij}\right|\ >\ t\right\}\\&\qquad\text{(because } Z'_{ij}\text{ does not depend on }\varphi)\\&\qquad\le 2\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\exp\left\{\frac{-mh^2\phi_1^2(h)\,t^2}{2(1+t)\tau_m^2\phi_1(2h)}\right\}.\end{aligned}\tag{47}$$

Now, for any constant $t_0>0$, take $t=t_0\sqrt{\tau_m^2\phi_1(2h)\log[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})]/\big(mh^2\phi_1^2(h)\big)}$ and observe that in view of (47),

$$\begin{aligned}P(m):=\mathbb{P}&\left\{\sup_{\varphi\in\mathcal{F}_{\varepsilon m}}\ \max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})}\frac{1}{m}\left|\sum_{i\in\boldsymbol{\mathcal{I}}_m}Z'_{ij}\right|\ \ge\ t_0\sqrt{\frac{\tau_m^2\phi_1(2h)\log[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})]}{mh^2\phi_1^2(h)}}\right\}\\&\le\ 2\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\cdot\exp\left\{\frac{-t_0^2\log[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})]}{2\left(1+t_0\sqrt{\tau_m^2\phi_1(2h)\log[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})]/\big(mh^2\phi_1^2(h)\big)}\right)}\right\}\\&\le\ 2\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\cdot\big[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\big]^{-ct_0}\ =\ 2\big[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\big]^{1-ct_0},\end{aligned}\tag{48}$$

for m large enough, where $P(m)$ is as in (48) and c is a positive constant not depending on n (or m). Consequently, choosing t_0 suitably, one finds

$$\sum_{m=1}^{\infty}P(m)\le 2\sum_{m=1}^{\infty}[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})]^{1-Ct_0^2}\ <\infty,$$

which holds due to Assumption (A6)(ii). Therefore, in view of (46), Assumption (A6)(ii), and the fact that $\phi_1(2h)/\phi_1(h) = \mathcal{O}(1)$, one finds

$$I_m \;=\; \mathcal{O}_{a.co.}\left(\sqrt{\frac{\tau_m^2\phi_1(2h)\log[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})]}{mh^2\phi_1^2(h)}}\right) + \mathcal{O}\left(\frac{\tau_m\phi_1(2h)}{h\phi_1(h)}\right). \tag{49}$$

As for the term $I\!I_m$ in (42), first observe that by (43), (44), and (46)

$$\begin{aligned} I\!I_m &\le \mathbb{E}\left[\sup_{\varphi\in\mathcal{F}_{\varepsilon_m}} \max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})} \sup_{\chi\in B(\widetilde{\chi}_j,\tau_m)} \left|\widehat{g}_m(\chi;\varphi) - \widehat{g}_m(\widetilde{\chi}_j;\varphi)\right|\right] \\ &\le \mathbb{E}\left[\max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})} \frac{1}{m}\left|\sum_{i\in\boldsymbol{\mathcal{I}}_m} Z'_{ij}\right|\right] + \mathcal{O}\left(\frac{\tau_m\phi_1(2h)}{h\phi_1(h)}\right), \end{aligned} \tag{50}$$

where as before, $Z'_{ij} = C_1[Z_{ij} - \mathbb{E}(Z_{ij})]$ with Z_{ij} as in (45). On the other hand, since

$$\left|Z'_{ij}\right| \;\le\; C_1[1 + C_2\phi_1(2h)]\,\tau_m/\big(h\phi_1(h)\big) \;=:\; A(m), \tag{51}$$

one can proceed as follows

$$\mathbb{E}\left[\max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})} \frac{1}{m}\left|\sum_{i\in\boldsymbol{\mathcal{I}}_m} Z'_{ij}\right|\right] \;=\; \int_0^\infty \mathbb{P}\left\{\max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})} \frac{1}{m}\left|\sum_{i\in\boldsymbol{\mathcal{I}}_m} Z'_{ij}\right| \;>\; t\right\}dt$$

$$\le \int_0^u dt \;+\; \int_u^{A(m)} \mathbb{P}\left\{\max_{1\le j\le\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})} \frac{1}{m}\left|\sum_{i\in\boldsymbol{\mathcal{I}}_m} Z'_{ij}\right| \;>\; t\right\}dt$$

(becuase $|Z'_{ij}| \le A(m)$, where $A(m)$ is as in (51))

$$\le u + 2\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}) \int_u^{A(m)} \exp\left\{-mh^2\phi_1^2(h)\,t^2/\big(2(1 + A(m))\tau_m^2\phi_1(2h)\big)\right\}dt,$$

(via the exponential bound in (47) and the fact that $m^{-1}|\sum_{i\in\boldsymbol{\mathcal{I}}_m} Z'_{ij}| \le A(m)$)

$$\le\; u + \frac{2\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})}{\sqrt{mh^2\phi_1^2(h)/\big((1 + A(m))\tau_m^2\phi_1(2h)\big)}} \int_{u\sqrt{mh^2\phi_1^2(h)/((1+A(m))\tau_m^2\phi_1(2h))}}^{\infty} e^{-v^2/2}\,dv$$

$\Big($by the change of variable, $v = t\sqrt{mh^2\phi_1^2(h)/\big((1 + A(m))\tau_m^2\phi_1(2h)\big)}\,\Big)$

$$\le\; u + \frac{2\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\cdot\exp\big\{-\big[mh^2\phi_1^2(h)/\big((1 + A(m))\tau_m^2\phi_1(2h)\big)\big]u^2/2\big\}}{\big[mh^2\phi_1^2(h)/\big((1 + A(m))\tau_m^2\phi_1(2h)\big)\big]\cdot u}$$

$\big($via the upper bound in Mills ratio (see Mitrinovic 1970, p. 177)$\big)$

$$=:\; u + \frac{2\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})}{4Nu}\,e^{-2Nu^2}, \quad \text{where } N = mh^2\phi_1^2(h)/\big[4(1 + A(m))\tau_m^2\phi_1(2h)\big]. \tag{52}$$

But the expression $u + \left[2\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})/(4Nu)\right] e^{-2Nu^2}$ in (52) is approximately minimized by taking $u = \sqrt{\log(2\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}))/(2N)}$, and the corresponding value of the right side of (52) becomes

$$\begin{aligned}
&\sqrt{\frac{\log(2\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}))}{2N}} + \sqrt{\frac{1}{8N\log(2\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}))}} \\
&= \sqrt{\frac{\log(2\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}))[2(1+A(m))\tau_m^2\phi_1(2h)]}{mh^2\phi_1^2(h)}} + \sqrt{\frac{(1+A(m))\tau_m^2\phi_1(2h)}{2mh^2\phi_1^2(h)\cdot\log(2\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}))}} \\
&= \mathcal{O}\left(\sqrt{\frac{\tau_m^2\log(\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}))}{mh^2\phi_1(h)}}\right).
\end{aligned}$$

This last bound together with (52) and (50) implies that the term $I\!I_m$ in (42) satisfies

$$I\!I_m = \mathcal{O}\left(\sqrt{\frac{\tau_m^2\log(\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}))}{mh^2\phi_1(h)}}\right) + \mathcal{O}\left(\frac{\tau_m\phi_1(2h)}{h\phi_1(h)}\right). \tag{53}$$

To deal with $I\!I\!I_m$, i.e., the last term in (42), define the quantity

$$\mathbb{U}_{ij}(\varphi) = \frac{\Delta_i Y_i\varphi(Y_i)\,\mathcal{K}\big(h^{-1}d(\tilde{\chi}_j,\,\boldsymbol{\chi}_i)\big) - \mathbb{E}\Big[\Delta_i Y_i\varphi(Y_i)\,\mathcal{K}\big(h^{-1}d(\tilde{\chi}_j,\,\boldsymbol{\chi}_i)\big)\Big]}{\mathbb{E}\big[\mathcal{K}\big(h^{-1}d(\tilde{\chi}_j,\,\boldsymbol{\chi}_1)\big)\big]}$$

and observe that for every $t > 0$

$$\mathbb{P}\left\{I\!I\!I_m \geq t\right\} > \mathcal{N}_{\varepsilon_m}(\mathcal{F})\,\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\cdot \sup_{\varphi\in\mathcal{F}_{\varepsilon_m}}\ \max_{1\leq j\leq\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})} \mathbb{P}\left\{\frac{1}{m}\left|\sum_{i\in\mathcal{I}_m}\mathbb{U}_{ij}(\varphi)\right| > t\right\}. \tag{54}$$

But, for each fixed φ, it can be shown that $\mathbb{E}\big|\mathbb{U}_{ij}(\varphi)\big|^k = \mathcal{O}\big((\phi_1(h))^{-k+1}\big)$, for all $k \geq 2$; see Ferraty et al. (2010, p. 347) as well as Ferraty and Vieu (2006, p. 66). Therefore, by Corollary A.8 of Ferraty and Vieu (2006), for any arbitrary $t_0 > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{m}\left|\sum_{i\in\mathcal{I}_m}\mathbb{U}_{ij}(\varphi)\right| > t_0\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F})\vee\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m\phi_1(h)}}\right\} \\
&\leq 2\big[\mathcal{N}_{\varepsilon_m}(\mathcal{F})\vee\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\big]^{-c\,t_0^2}, \quad c > 0.
\end{aligned}$$

Therefore, in view of (54),

$$\mathbb{P}\left\{ \mathit{III}_m > t_0\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F}) \vee \mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m\phi_1(h)}} \right\} \leq 2\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F}) \vee \mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]^{2-c\,t_0^2}.$$

Choosing t_0 suitably so that $2 - c\,t_0^2 \leq 1 - \beta$, where $\beta > 1$ is as in Assumption (A6)(ii), one finds

$$\sum_{m=1}^{\infty}\mathbb{P}\left\{ \mathit{III}_m > t_0\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F}) \vee \mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m\phi_1(h)}} \right\} \leq 2\sum_{m=1}^{\infty}\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F}) \vee \mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]^{1-\beta} < \infty,$$

which then yields

$$\mathit{III}_m = \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F}) \vee \mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m\phi_1(h)}}\right). \tag{55}$$

Putting together (42), (49), (53), and (55), one finds

$$\sup_{\varphi\in\mathcal{F}_{\varepsilon m}}\sup_{\chi\in\mathcal{S}_{\mathbb{X}}}\left|\widehat{g}_m(\chi;\varphi) - \mathbb{E}\big(\widehat{g}_m(\chi;\varphi)\big)\right| = \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F}) \vee \mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m\phi_1(h)}}\right). \tag{56}$$

Regarding the term $\left[\mathbb{E}\big(\widehat{g}_m(\chi;\varphi)\big) - \psi_1(\chi;\varphi)\right]$ in (41), one can argue as in the proof of Lemma 10 of Ferraty et al. (2010) and Lemma 4.4 of Ferraty and Vieu (2006) that, under Assumptions (A2), (A3), (A4), and (A5a),

$$\begin{aligned}\left|\mathbb{E}\big(\widehat{g}_m(\chi;\varphi)\big) - \psi_1(\chi;\varphi)\right| &\leq \frac{1}{\mathbb{E}\big[\mathcal{K}\big(h^{-1}d(\chi,\,\boldsymbol{\chi}_1)\big)\big]} \\ &\quad \mathbb{E}\Big[\mathcal{K}\big(h^{-1}d(\chi,\,\boldsymbol{\chi}_1)\big)\cdot\underbrace{\left|\psi_1(\boldsymbol{\chi}_1;\varphi) - \psi_1(\chi;\varphi)\right|}_{\leq C_1 d^{\beta_1}(\chi,\,\boldsymbol{\chi}_1)}\Big] \\ &\leq \frac{C_1}{\mathbb{E}\big[\mathcal{K}\big(h^{-1}d(\chi,\,\boldsymbol{\chi}_1)\big)\big]}\mathbb{E}\Big[\mathcal{K}\big(h^{-1}d(\chi,\,\boldsymbol{\chi}_1)\big)\,\mathbb{1}_{\{\boldsymbol{\chi}_1\in B(\chi,h)\}}\cdot d^{\beta_1}(\chi,\,\boldsymbol{\chi}_1)\Big] \\ &\leq C_1 h^{\beta_1},\end{aligned}$$

where β_1 and C_1 are the positive constants in Assumption (A3). Since C_1 does not depend on χ or φ, we find

$$\sup_{\varphi\in\mathcal{F}_{\varepsilon m}}\sup_{\chi\in\mathcal{S}_{\mathbb{X}}}\left|\mathbb{E}\big(\widehat{g}_m(\chi;\varphi)\big) - \psi_1(\chi;\varphi)\right| = \mathcal{O}\big(h^{\beta_1}\big). \tag{57}$$

Furthermore, Lemma 8 and Corollary 9 of Ferraty et al. (2010) imply that under Assumptions (A2) and (A4)–(A6), one has

$$\sup_{\chi\in\mathcal{S}_{\mathbb{X}}}\left|1-\widehat{f}_m(\chi)\right| = \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})]}{m\phi_1(h)}}\right) \quad \text{and} \quad \mathbb{P}\left\{\inf_{\chi\in\mathcal{S}_{\mathbb{X}}}\widehat{f}_m(\chi) < \frac{1}{2}\right\} < \infty. \tag{58}$$

Putting together (41), (56), (57), and (58), one finds

$$\sup_{\varphi\in\mathcal{F}_{\varepsilon_m}}\sup_{\chi\in\mathcal{S}_{\mathbb{X}}}\left|\widehat{\psi}_{m,1}(\chi;\varphi)-\psi_1(\chi;\varphi)\right| = \mathcal{O}\big(h^{\beta_1}\big) + \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F})\vee\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m\phi_1(h)}}\right). \tag{59}$$

This completes the proof of (37) of Lemma 5 for the case of $k=1$. The case of $k=2$ is easier since it amounts to using (7) with $k=2$. The proofs of (38), (39), and (40) of Lemma 5 are similar (and in fact easier) and will not be given. □

Lemma 6 *Let $\psi_o(\zeta;\varphi)$ and $\widehat{\psi}_{m,o}(\zeta;\varphi)$ be as in part* (ii) *of Lemma 5. Then, under the conditions of Lemma 5,*

$$\sum_{m=1}^{\infty}\mathbb{P}\left\{\inf_{\varphi\in\mathcal{F}_{\varepsilon_m}}\inf_{\zeta\in\mathcal{S}^o_{\mathbb{X}}}\widehat{\psi}_{m,o}(\zeta;\varphi) \le \frac{\varrho_0}{2}\right\} < \infty, \tag{60}$$

where the constant $\varrho_0 > 0$ is as in Assumption (A7).

Proof of Lemma 6. First observe that

$$\begin{aligned}
&\inf_{\varphi\in\mathcal{F}_{\varepsilon m}}\inf_{\zeta\in\mathcal{S}^o_{\mathbb{X}}}\widehat{\psi}_{m,o}(\zeta;\varphi) \le \frac{\varrho_0}{2}\\
&\Leftrightarrow \exists\,\varphi'\in\mathcal{F}_{\varepsilon_m} \quad \text{and} \quad \zeta'\in\mathcal{S}^o_{\mathbb{X}} \quad \text{such that } \widehat{\psi}_{m,o}(\zeta';\varphi') \le \frac{\varrho_0}{2}\\
&\Leftrightarrow \exists\,\varphi'\in\mathcal{F}_{\varepsilon_m} \quad \text{and} \quad \zeta'\in\mathcal{S}^o_{\mathbb{X}} \quad \text{such that } \psi_o(\zeta';\varphi')\\
&-\widehat{\psi}_{m,o}(\zeta';\varphi') \ge \psi_o(\zeta';\varphi')-\frac{\varrho_0}{2}\\
&\Rightarrow \sup_{\varphi\in\mathcal{F}_{\varepsilon m}}\sup_{\zeta\in\mathcal{S}^o_{\mathbb{X}}}\left|\widehat{\psi}_{m,o}(\zeta;\varphi)-\psi_o(\zeta;\varphi)\right| \ge \frac{\varrho_0}{2},\\
&(\text{since } \psi_o(\zeta';\varphi') \ge \varrho_0 \text{ by Assumption } (A7)).
\end{aligned} \tag{61}$$

Now let $C > 0$ be any arbitrary constant and let $m_0 > 0$ be such that $Ch^{\beta_o} \le \frac{\varrho_0}{4}$ for all $m > m_0$ [which is possible because $h \to 0$ as n (and m) $\to \infty$], where β_o is as in Assumption (A3). Then, (61) implies that for $m > m_0$, one has $\sup_{\varphi \in \mathcal{F}_{\varepsilon_m}} \sup_{\zeta \in \mathcal{S}_{\mathbb{X}}^o} \left|\widehat{\psi}_{m,o}(\zeta;\varphi) - \psi_o(\zeta;\varphi)\right| - Ch^{\beta_o} \geq \frac{\varrho_0}{4}$. Therefore,

$$
\begin{aligned}
&\sum_{m=1}^{\infty} \mathbb{P}\left\{\inf_{\varphi \in \mathcal{F}_{\varepsilon_m}} \inf_{\zeta \in \mathcal{S}_{\mathbb{X}}^o} \widehat{\psi}_{m,o}(\zeta;\varphi) \leq \frac{\varrho_0}{2}\right\} \\
&\quad \leq \sum_{m=1}^{n_0} \mathbb{P}\left\{\sup_{\varphi \in \mathcal{F}_{\varepsilon_m}} \sup_{\zeta \in \mathcal{S}_{\mathbb{X}}^o} \left|\widehat{\psi}_{m,o}(\zeta;\varphi) - \psi_o(\zeta;\varphi)\right| \geq \frac{\varrho_0}{2}\right\} \\
&\qquad + \sum_{m=n_0+1}^{\infty} \mathbb{P}\left\{\sup_{\varphi \in \mathcal{F}_{\varepsilon_m}} \sup_{\zeta \in \mathcal{S}_{\mathbb{X}}^o} \left|\widehat{\psi}_{m,o}(\zeta;\varphi) - \psi_o(\zeta;\varphi)\right| - Ch^{\beta_o} \geq \frac{\varrho_0}{4}\right\} \\
&\quad < \infty, \quad (\text{by}(39)).
\end{aligned}
$$

□

Proof of Theorem 1 It is sufficient to prove the theorem for the case of $p=2$. To appreciate this, simply observe that in view of the definition of $\widehat{m}(\chi;\widehat{\varphi}_n)$ one finds

$$
\begin{aligned}
&\left|\widehat{m}(\chi;\widehat{\varphi}_n) - m(\chi)\right|^p \leq \left(\left|\widehat{m}(\chi;\widehat{\varphi}_n)\right| + \left|m(\chi)\right|\right)^{p-2}\left|\widehat{m}(\chi;\widehat{\varphi}_n) - m(\chi)\right|^2 \\
&\leq (3L)^{p-2}\left|\widehat{m}(\chi;\widehat{\varphi}_n) - m(\chi)\right|^2.
\end{aligned}
$$

To proceed with the proof of the theorem, we first note that by Lemmas 1, 2, and 3,

$$
\begin{aligned}
\mathbb{E}\left[\left|\widehat{m}(\boldsymbol{\chi};\widehat{\varphi}_n) - m(\boldsymbol{\chi})\right|^2 \Big|\mathbb{D}_n\right] &\leq 2\mathbb{E}\left[\left|\widehat{m}(\boldsymbol{\chi};\widehat{\varphi}_n) - m(\boldsymbol{\chi};\varphi_{\varepsilon_m})\right|^2\Big|\mathbb{D}_n\right] \\
&\quad +2\mathbb{E}\left|m(\boldsymbol{\chi};\varphi_{\varepsilon_m}) - m(\boldsymbol{\chi};\varphi^*)\right|^2 \\
&\leq 2\mathbb{E}\left[\left|\widehat{m}(\boldsymbol{\chi};\widehat{\varphi}_n) - m(\boldsymbol{\chi};\varphi_{\varepsilon_m})\right|^2\Big|\mathbb{D}_n\right] + 4LC\,\varepsilon_m, \\
&\quad \text{(via (34) and (35), where } C \text{ is as in Lemma 2)} \\
&\leq 2\sup_{\varphi \in \mathcal{F}_{\varepsilon_m}}\left|\mathbb{E}\left[\left|\widehat{m}_m(\boldsymbol{\chi};\varphi) - Y\right|^2\Big|\mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\varphi)\right| \\
&\quad +2\sup_{\varphi \in \mathcal{F}_{\varepsilon_m}}\left|\widehat{L}_{m,\ell}(\varphi) - E\left|m(\boldsymbol{\chi};\varphi) - Y\right|^2\right| \\
&\quad +2C_1\sqrt{\varepsilon_m} + 8LC\varepsilon_m.
\end{aligned}
\tag{62}
$$

On the other hand, the two supremum terms on the right side of (62) can be bounded as follows.

$$
\begin{aligned}
&\sup_{\varphi \in \mathcal{F}_{\varepsilon_m}}\left|\mathbb{E}\left[\left|\widehat{m}_m(\boldsymbol{\chi};\varphi) - Y\right|^2\Big|\mathbb{D}_m\right] - \widehat{L}_{m,\ell}(\varphi)\right| \\
&\leq \sup_{\varphi \in \mathcal{F}_{\varepsilon_m}}\left|\ell^{-1}\sum_{i \in \boldsymbol{\mathcal{I}}_\ell}\frac{\Delta_i\left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2}{\pi_\varphi(\boldsymbol{\zeta}_i, Y_i)} - \mathbb{E}\left[\left|\widehat{m}_m(\boldsymbol{\chi};\varphi) - Y\right|^2\Big|\mathbb{D}_m\right]\right|
\end{aligned}
$$

$$
\begin{aligned}
&+ \sup_{\varphi\in\mathcal{F}_{\varepsilon_m}} \left| \ell^{-1} \sum_{i\in\mathcal{I}_\ell} \Delta_i \left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2 \left[\frac{1}{\pi_\varphi(\boldsymbol{\zeta}_i, Y_i)} - \frac{1}{\widehat{\pi}_\varphi(\boldsymbol{\zeta}_i, Y_i)} \right] \right| \\
&:= \mathbf{I}_{n,1} + \mathbf{I}_{n,2}. \qquad (63)
\end{aligned}
$$

Similarly, one has

$$
\begin{aligned}
&\sup_{\varphi\in\mathcal{F}_{\varepsilon_m}} \left| \widehat{L}_{m,\ell}(\varphi) - \mathbb{E}\left| m(\boldsymbol{\chi};\varphi) - Y \right|^2 \right| \\
&\leq \sup_{\varphi\in\mathcal{F}_{\varepsilon_m}} \left| \frac{1}{\ell} \sum_{i\in\mathcal{I}_\ell} \frac{\Delta_i \left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2}{\pi_\varphi(\boldsymbol{\zeta}_i, Y_i)} - \frac{1}{\ell} \sum_{i\in\mathcal{I}_\ell} \frac{\Delta_i \left|m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2}{\pi_\varphi(\boldsymbol{\zeta}_i, Y_i)} \right| \\
&\quad + \sup_{\varphi\in\mathcal{F}_{\varepsilon_m}} \left| \frac{1}{\ell} \sum_{i\in\mathcal{I}_\ell} \frac{\Delta_i \left|m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2}{\pi_\varphi(\boldsymbol{\zeta}_i, Y_i)} - \mathbb{E}\left[\frac{\Delta\left|m(\boldsymbol{\chi};\varphi) - Y\right|^2}{\pi_\varphi(\boldsymbol{\zeta}, Y)} \right] \right| \\
&\quad + \mathbf{I}_{n,2} \text{ (where } \mathbf{I}_{n,2} \text{ is as in (63))} \\
&:= \mathbf{I}_{n,3} + \mathbf{I}_{n,4} + \mathbf{I}_{n,2}, \qquad (64)
\end{aligned}
$$

where $\mathbf{I}_{n,2}$ is as in (63). Therefore, in view of (62), one finds

$$
\mathbb{E}\left[\left|\widehat{m}(\boldsymbol{\chi};\widehat{\varphi}_n) - m(\boldsymbol{\chi})\right|^2 \middle| \mathbb{D}_n \right] \leq 2\left\{ \mathbf{I}_{n,1} + 2\mathbf{I}_{n,2} + \mathbf{I}_{n,3} + \mathbf{I}_{n,4} \right\} + 2C_1\sqrt{\varepsilon_m} + 8LC\,\varepsilon_m, \qquad (65)
$$

where the terms $\mathbf{I}_{n,1}, \mathbf{I}_{n,2}, \mathbf{I}_{n,3}$, and $\mathbf{I}_{n,4}$ are as in (63), and (64). To deal with the term $\mathbf{I}_{n,1}$, observe that conditional on $\mathbb{D}_m$, the terms $\Delta_i \left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2/\pi_\varphi(\boldsymbol{\zeta}_i, Y_i)$, $i \in \mathcal{I}_\ell$, are independent bounded random variables taking values in the interval $\left[0,\ (3L)^2/\pi_{\min}\right]$. Therefore, for every $t > 0$

$$
\begin{aligned}
&\mathbb{P}\{\mathbf{I}_{n,1} \geq t\} \\
&\leq \mathcal{N}_{\varepsilon_m}(\mathcal{F}) \sup_{\varphi\in\mathcal{F}_{\varepsilon_m}} \mathbb{P}\left\{ \left| \frac{1}{\ell} \sum_{i\in\mathcal{I}_\ell} \frac{\Delta_i \left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2}{\pi_\varphi(\boldsymbol{\zeta}_i, Y_i)} - \mathbb{E}\left[\left|\widehat{m}_m(\boldsymbol{\chi};\varphi) - Y\right|^2 \middle| \mathbb{D}_m \right] \right| \geq t \right\} \\
&\leq \mathcal{N}_{\varepsilon_m}(\mathcal{F}) \sup_{\varphi\in\mathcal{F}} \mathbb{E}_\varphi\left[\mathbb{P}_\varphi\left\{ \left| \frac{1}{\ell} \sum_{i\in\mathcal{I}_\ell} \frac{\Delta_i \left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2}{\pi_\varphi(\boldsymbol{\zeta}_i, Y_i)} - \mathbb{E}_\varphi\left[\left|\widehat{m}_m(\boldsymbol{\chi};\varphi) - Y\right|^2 \middle| \mathbb{D}_m \right] \right| \geq t \middle| \mathbb{D}_m \right\} \right] \\
&\leq 2\mathcal{N}_{\varepsilon_m}(\mathcal{F}) \cdot \exp\left\{ -\pi^2_{\min} \ell t^2/(81L^4) \right\},
\end{aligned}
$$

where the last line follows via Hoeffding's inequality in conjunction with Assumption (A8). Since the above bound holds for all $t > 0$, taking $t = t_0 \cdot \sqrt{\log(\mathcal{N}_{\varepsilon_m}(\mathcal{F}))/\ell}$, for any $t_0 > 0$, yields $\mathbb{P}\left\{\mathbf{I}_{n,1} \geq t\right\} \leq 2\left(\mathcal{N}_{\varepsilon_m}(\mathcal{F})\right)^{1-ct_0}$, where $c > 0$ is a constant not depending on n. Choosing t_0 large enough, we find $\sum_{n=1}^{\infty} \mathbb{P}\left\{\mathbf{I}_{n,1} \geq t\right\} \leq 2\sum_{n=1}^{\infty} \left(\mathcal{N}_{\varepsilon_m}(\mathcal{F})\right)^{1-ct_0} < \infty$. Therefore

$$\mathbf{I}_{n,1} = \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log(\mathcal{N}_{\varepsilon_m}(\mathcal{F}))}{\ell}}\right). \tag{66}$$

Next, to handle the term $\mathbf{I}_{n,2}$ in (65), let $\psi_o(\zeta;\varphi)$ and $\eta_o(\zeta)$ be as in (11). Also let $\widehat{\psi}_{m,o}(\zeta;\varphi)$ and $\widehat{\eta}_{m,o}(\zeta)$ be as in (12) and observe that since $0 \le \Delta_i \left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi) - Y_i\right|^2 \le 3L$, one finds

$$\begin{aligned}
\mathbf{I}_{n,2} &\le 9L^2 \sup_{\varphi\in\mathcal{F}_{\varepsilon m}} \left|\frac{1}{\ell}\sum_{i\in\mathcal{I}_\ell}\left[\frac{1}{\widehat{\pi}_\varphi(\boldsymbol{\zeta}_i, Y_i)} - \frac{1}{\pi_\varphi(\boldsymbol{\zeta}_i, Y_i)}\right]\right| \\
&\le \frac{9L^2}{\ell}\sum_{i\in\mathcal{I}_\ell}\sup_{\varphi\in\mathcal{F}_{\varepsilon m}}\left|\frac{1-\widehat{\eta}_{m,o}(\boldsymbol{\zeta}_i)}{\widehat{\psi}_{m,o}(\boldsymbol{\zeta}_i;\varphi)} - \frac{1-\eta_o(\boldsymbol{\zeta}_i)}{\psi_o(\boldsymbol{\zeta}_i;\varphi)}\right|\cdot\varphi(Y_i) \\
&\qquad \text{(where the inequality above follows from (10), (11), (13), (14), and (12))} \\
&= \frac{9L^2}{\ell}\sum_{i\in\mathcal{I}_\ell}\sup_{\varphi\in\mathcal{F}_{\varepsilon m}}\left| -\frac{1-\widehat{\eta}_{m,o}(\boldsymbol{\zeta}_i)}{\widehat{\psi}_{m,o}(\boldsymbol{\zeta}_i;\varphi)}\cdot\frac{\widehat{\psi}_{m,o}(\boldsymbol{\zeta}_i;\varphi)-\psi_o(\boldsymbol{\zeta}_i;\varphi)}{\psi_o(\boldsymbol{\zeta}_i;\varphi)} - \frac{\widehat{\eta}_{m,o}(\boldsymbol{\zeta}_i)-\eta_o(\boldsymbol{\zeta}_i)}{\psi_o(\boldsymbol{\zeta}_i;\varphi)}\right|\cdot\varphi(Y_i) \\
&\le \frac{9BL^2}{\varrho_0}\cdot\frac{1}{\ell}\sum_{i\in\mathcal{I}_\ell}\left[\frac{\left|1-\widehat{\eta}_{m,o}(\boldsymbol{\zeta}_i)\right|}{\inf_{\varphi\in\mathcal{F}_{\varepsilon m}}\widehat{\psi}_{m,o}(\boldsymbol{\zeta}_i;\varphi)}\cdot\sup_{\varphi\in\mathcal{F}_{\varepsilon m}}\left|\widehat{\psi}_{m,o}(\boldsymbol{\zeta}_i;\varphi)-\psi_o(\boldsymbol{\zeta}_i;\varphi)\right|\right] \\
&\quad + \frac{9BL^2}{\varrho_0}\cdot\frac{1}{\ell}\sum_{i\in\mathcal{I}_\ell}\left|\widehat{\eta}_{m,o}(\boldsymbol{\zeta}_i)-\eta_o(\boldsymbol{\zeta}_i)\right|, \quad \text{(where } \varrho_0 \text{ is as in Assumption } (A7)\text{)} \\
&\le C_\varrho\left[\frac{\sup_{\boldsymbol{\zeta}\in\mathcal{S}_{\mathbb{X}}^o}\left|1-\widehat{\eta}_{m,o}(\boldsymbol{\zeta})\right|}{\inf_{\varphi\in\mathcal{F}_{\varepsilon m}}\inf_{\boldsymbol{\zeta}\in\mathcal{S}_{\mathbb{X}}^o}\widehat{\psi}_{m,o}(\boldsymbol{\zeta};\varphi)}\sup_{\varphi\in\mathcal{F}_{\varepsilon m}}\sup_{\boldsymbol{\zeta}\in\mathcal{S}_{\mathbb{X}}^o}\left|\widehat{\psi}_{m,o}(\boldsymbol{\zeta};\varphi)-\psi_o(\boldsymbol{\zeta};\varphi)\right| + \sup_{\boldsymbol{\zeta}\in\mathcal{S}_{\mathbb{X}}^o}\left|\widehat{\eta}_{m,o}(\boldsymbol{\zeta})-\eta_o(\boldsymbol{\zeta})\right|\right]
\end{aligned} \tag{67}$$

by Assumption (A0), where $C_\varrho = 9BL^2/\varrho_0$. But the first term in the square brackets above satisfies

$$\frac{\sup_{\zeta\in\mathcal{S}_{\mathbb{X}}^o}\left|1-\widehat{\eta}_{m,o}(\zeta)\right|}{\inf_{\varphi\in\mathcal{F}_{\varepsilon_m}}\inf_{\zeta\in\mathcal{S}_{\mathbb{X}}^o}\widehat{\psi}_{m,o}(\zeta;\varphi)} = \mathcal{O}_p(1), \tag{68}$$

which follows from Lemma 6 and the fact that by part (ii) of Lemma 5 and the definition of $\eta_o(\zeta)$ in (11), one has

$$\begin{aligned}
\sup_{\zeta\in\mathcal{S}_{\mathbb{X}}^o}\left|1-\widehat{\eta}_{m,o}(\zeta)\right| \le 2 + \sup_{\zeta\in\mathcal{S}_{\mathbb{X}}^o}\left|\widehat{\eta}_{m,o}(\zeta)-\eta_o(\zeta)\right| &= 2 + \mathcal{O}\big(h^{\beta_o}\big) \\
+\mathcal{O}_{a.co.}&\left(\sqrt{\frac{\log\left[\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}^o)\right]}{m\cdot\phi_0(h)}}\right).
\end{aligned} \tag{69}$$

Therefore by (67), (68), and part (ii) of Lemma 5, one finds

$$\mathbf{I}_{n,2} = \mathcal{O}\big(h^{\beta_o}\big) + \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F})\vee\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}}^o)\right]}{m\cdot\phi_0(h)}}\right). \tag{70}$$

To deal with the term $\mathbf{I}_{n,3}$ in (65), first observe that in view of the definitions of $m(\chi;\varphi)$ and $\widehat{m}_m(\chi;\varphi)$ in (5) and (6), respectively, one has

$$\begin{aligned}\mathbf{I}_{n,3} &\leq \frac{1}{\pi_{\min}} \sup_{\varphi\in\mathcal{F}_{\varepsilon_m}} \frac{1}{\ell}\sum_{i\in\mathcal{I}_\ell}\left[\left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi)-m(\boldsymbol{\chi}_i;\varphi)\right|\cdot\left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi)+m(\boldsymbol{\chi}_i;\varphi)-2Y_i\right|\right]\\ &\leq \frac{5L}{\pi_{\min}} \sup_{\varphi\in\mathcal{F}_{\varepsilon_m}} \frac{1}{\ell}\sum_{i\in\mathcal{I}_\ell}\left|\widehat{m}_m(\boldsymbol{\chi}_i;\varphi)-m(\boldsymbol{\chi}_i;\varphi)\right|. \qquad (71)\end{aligned}$$

Furthermore, for each $\varphi\in\mathcal{F}$, and with $\widehat{\psi}_{m,k}(\chi;\varphi)$ and $\widehat{\eta}_{m,k}(\chi)$ as in (7) and (8), $k=1,2$, it is straightforward to see that

$$\begin{aligned}\left|\widehat{m}_m(\chi;\varphi)-m(\chi;\varphi)\right| &\leq \left|\widehat{\eta}_{m,1}(\chi)-\eta_1(\chi)\right| + \left|\frac{\widehat{\psi}_{m,1}(\chi;\varphi)}{\widehat{\psi}_{m,2}(\chi;\varphi)}-\frac{\psi_1(\chi;\varphi)}{\psi_2(\chi;\varphi)}\right|\\ &\quad +L\left|\widehat{\eta}_{m,2}(\chi)-\eta_2(\chi)\right|, \qquad (72)\end{aligned}$$

where $\psi_k(\chi;\varphi)$ and $\eta_2(\chi)$, $k=1,2$, are as in (4). On the other hand, one finds

$$\begin{aligned}&\left|\frac{\widehat{\psi}_{m,1}(\chi;\varphi)}{\widehat{\psi}_{m,2}(\chi;\varphi)}-\frac{\psi_1(\chi;\varphi)}{\psi_2(\chi;\varphi)}\right|\\ &=\frac{1}{\psi_2(\chi;\varphi)}\left|\frac{\widehat{\psi}_{m,1}(\chi;\varphi)}{\widehat{\psi}_{m,2}(\chi;\varphi)}\big(\widehat{\psi}_{m,2}(\chi;\varphi)-\psi_2(\chi;\varphi)\big)+\big(\widehat{\psi}_{m,1}(\chi;\varphi)-\psi_1(\chi;\varphi)\big)\right|\\ &\leq\frac{1}{\varrho_0}\left[L\cdot\left|\widehat{\psi}_{m,2}(\chi;\varphi)-\psi_2(\chi;\varphi)\right|+\left|\widehat{\psi}_{m,1}(\chi;\varphi)-\psi_1(\chi;\varphi)\right|\right], \qquad (73)\end{aligned}$$

where we used the facts that $\left|\widehat{\psi}_{m,1}(\chi;\varphi)/\widehat{\psi}_{m,2}(\chi;\varphi)\right|\leq L$ and $\psi_2(\chi;\varphi):=\mathbb{E}[\Delta\,\varphi(Y)|\boldsymbol{\chi}=\chi]\geq\varrho_0$ [by Assumption (A7)]. Therefore, combining (71), (72), (73), (37), and (38), one arrives at

$$\begin{aligned}\mathbf{I}_{n,3} &\leq \frac{5L}{\pi_{\min}}\sup_{\varphi\in\mathcal{F}_{\varepsilon_m}}\sup_{\chi\in\mathcal{S}_{\mathbb{X}}}\left|\widehat{m}_m(\chi;\varphi)-m(\chi;\varphi)\right| = \mathcal{O}\big(h^\beta\big)\\ &\quad+\mathcal{O}_{a.co.}\left(\sqrt{\frac{\log\left[\mathcal{N}_{\varepsilon_m}(\mathcal{F})\vee\mathcal{N}_{\tau_m}(\mathcal{S}_{\mathbb{X}})\right]}{m\cdot\phi_1(h)}}\right), \qquad (74)\end{aligned}$$

where $\beta=\beta_1\wedge\beta_2$, and β_k is as in (37). Finally, to deal with the term $\mathbf{I}_{n,4}$ in (65), we first note that the terms $\Delta_i\left|m(\boldsymbol{\chi}_i;\varphi)-Y_i\right|^2/\pi_\varphi(\boldsymbol{\zeta}_i,Y_i)$, $i\in\mathcal{I}_\ell$, are iid bounded random variables taking values in the interval $[0,\ 4L^2/\pi_{\min}]$. Therefore, by Hoeffding's inequality, for every $t>0$,

$$\begin{aligned}&\mathbb{P}\{\mathbf{I}_{n,4}>t\}\\ &\leq\mathcal{N}_{\varepsilon_m}(\mathcal{F})\sup_{\varphi\in\mathcal{F}_{\varepsilon_m}}\mathbb{P}\left\{\left|\frac{1}{\ell}\sum_{i\in\mathcal{I}_\ell}\frac{\Delta_i\left|m(\boldsymbol{\chi}_i;\varphi)-Y_i\right|^2}{\pi_\varphi(\boldsymbol{\zeta}_i,Y_i)}-\mathbb{E}\left[\frac{\Delta\left|m(\boldsymbol{\chi};\varphi)-Y\right|^2}{\pi_\varphi(\boldsymbol{\zeta},Y)}\right]\right|>t\right\}\\ &\leq 2\,\mathcal{N}_{\varepsilon_m}(\mathcal{F})\exp\left\{-\pi_{\min}^2\ell t^2/(8L^4)\right\}.\end{aligned}$$

Therefore, using the arguments that led to (66), one can show

$$\mathbf{I}_{n,4} = \mathcal{O}_{a.co.}\left(\sqrt{\frac{\log[\mathcal{N}_{\varepsilon_m}(\mathcal{F})]}{\ell}}\right). \tag{75}$$

Now, Theorem 1 follows from (65), (66), (70), (74), and (75), where β in this theorem can be taken to be $\min(\beta_o, \beta_1, \beta_2)$. □

Proof of Theorem 2 Let $g_{\text{B}}(\chi)$ and $\widehat{g}_n(\chi\,;\widehat{\varphi}_n)$ be as in (23) and (24), respectively. Then, it is not hard to show that

$$\left|\mathbb{P}\left\{\widehat{g}_n(\boldsymbol{\chi};\widehat{\varphi}_n) \neq Y\middle|\mathbb{D}_n\right\} - \mathbb{P}\left\{g_{\text{B}}(\boldsymbol{\chi}) \neq Y\right\}\right| \leq 2\mathbb{E}\left[\left|\widehat{m}(\boldsymbol{\chi};\widehat{\varphi}_n) - m(\boldsymbol{\chi})\right|\mathbb{D}_n\right]; \tag{76}$$

see, for example, Lemma 6.1 of Devroye et al. (1996). The proof of Theorem 2 now follows from Theorem 1 in conjunction with the Cauchy–Schwarz inequality. □

Acknowledgements This work was supported by the National Science Foundation Grants DMS-1916161 and DMS-2310504 of Majid Mojirsheibani.

References

Abraham C, Biau G, Cadre B (2006) On the kernel rule for functional classification. AISM 58:619–633

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

Cérou F, Guyader A (2006) Nearest neighbor classification in infinite dimensions. ESAIM-Probab Stat 10:340–355

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

Cheng PE, Chu CK (1996) Kernel estimation of distribution functions and quantiles with missing data. Stat Sin 6:63–78

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

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

Febrero-Bande M, Oviedo de la Fuente M (2012) Statistical computing in functional data analysis: the R package fda.usc. J Stat Softw 51:1–28

Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer, New York

Ferraty F, Laksaci A, Tadj A, Vieu P (2010) Rate of uniform consistency for nonparametric estimates with functional variables. J Stat Plan Inference 140:335–352

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

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

Kim JK, Yu CL (2011) A semiparametric estimation of mean functionals with nonignorable missing data. J Am Stat Assoc 106:157–65

Li T, Xie F, Feng X, Ibrahim J, Zhu H (2018) Functional linear regression models for nonignorable missing scalar responses. Stat Sin 28:1867–1886

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

Liu Z, Yau CY (2021) Fitting time series models for longitudinal surveys with nonignorable missing data. J Stat Plan Inference 214:1–12

Maity A, Pradhan V, Das U (2019) Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. Am Stat 73:340–349

Mitrinovic DS (1970) Analytic inequalities. Springer, New York

Mojirsheibani M (2021) On classification with nonignorable missing data. J Multivar Anal 184:104755

Mojirsheibani M (2022) On the maximal deviation of kernel regression estimators with MNAR response variables. Stat Pap 63:1677–1705

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

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

Nadaraya EA (1964) On estimating regression. Theory Probab Appl 9:141–142

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

O'Brien J, Gunawardena H, Paulo J, Chen X, Ibrahim J, Gygi S, Qaqish B (2018) The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. Ann Appl Stat 12:2075–2095

Rachdi M, Vieu P (2007) Nonparametric regression for functional data: automatic smoothing parameter selection. J Stat Plan Inference 137:2784–2801

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

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

Uehara M, Kim JK (2018) Semiparametric response model with nonignorable nonresponse (Preprint). arXiv:1810.12519v1

van der Vaart A, Wellner J (1996) Weak convergence and empirical processes with applications to statistics. Springer, New York

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

Wang S, Shao J, Kim JK (2014) Identifiability and estimation in problems with nonignorable nonresponse. Stat Sin 24:1097–1116

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

Wang L, Zhao P, Shao J (2021) Dimension-reduced semiparametric estimation of distribution functions and quantiles with nonignorable nonresponse. Comput Stat Data Anal 156:107142

Watson GS (1964) Smooth regression analysis. Sankhya A 26:359–372

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 Stat Assoc 110:1577–1590

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