Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 2733–2758
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Nonparametric curve estimation with missing data: A general empirical process approach☆

Majid Mojirsheibani*

Carleton University, Ottawa, Ont., Canada K1S 5B6

Received 3 February 2006; accepted 23 February 2006
Available online 12 March 2007

Abstract

A general nonparametric imputation procedure, based on kernel regression, is proposed to estimate points as well as set- and function-indexed parameters when the data are missing at random (MAR). The proposed method works by imputing a specific function of a missing value (and not the missing value itself), where the form of this specific function is dictated by the parameter of interest. Both single and multiple imputations are considered. The associated empirical processes provide the right tool to study the uniform convergence properties of the resulting estimators. Our estimators include, as special cases, the imputation estimator of the mean, the estimator of the distribution function proposed by Cheng and Chu [1996. Kernel estimation of distribution functions and quantiles with missing data. Statist. Sinica 6, 63–78], imputation estimators of a marginal density, and imputation estimators of regression functions.

Keywords: Missing data; Nonparametric; Imputation; Empirical process; Kernel

1. Introduction

In many nonparametric estimation problems, a drawback to the imputation of missing covariates and their substitution for the actual missing values in the statistic of interest is that the theoretical (asymptotic) properties of such statistics can become intractable. Examples include the empirical distribution function and nonparametric estimation of a marginal density function. In the case of the empirical distribution function, the method used by Cheng and Chu (1996) overcomes such difficulties by imputing the indicator functions directly, and not the missing values themselves.

Our contributions may be summarized as follows. In the first place, we extend the approach adopted by Cheng and Chu (1996), in a natural way, to the estimation of points as well as set- and function-indexed parameters. The associated empirical processes provide the right theoretical tools to study asymptotic properties of the proposed estimators. Our contributions also include the derivation of some new exponential performance bounds on the uniform deviations of the resulting statistics from their expectations, in the presence of missing covariates. These bounds will be used to assess the strong uniform consistency properties of a number of important curve estimators. In particular, we apply our results to nonparametric estimators of regression and density functions, in the presence of missing data. In addition to

☆ Research supported by an NSERC Discovery Grant of M. Mojirsheibani at Carleton University, Ottawa, Canada.
* Tel.: +1 613 520 2600x2134; fax: +1 613 520 3822.
E-mail address: majidm@math.carleton.ca.

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.02.016

single imputation, we also propose a multiple imputation procedure, that can be applied to many problems in practice. Both mechanics and the asymptotic validity of these procedures are discussed.

Consider the following setup. Let $\mathbf{Z}, \mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be iid R^{d+p}-valued random vectors with distribution function F, where $\mathbf{Z}=(\mathbf{X}^{\mathrm{T}}, \mathbf{Y}^{\mathrm{T}})^{\mathrm{T}}$. Here it is assumed that $\mathbf{X} \in R^d$, $d \geqslant 1$, is always observable but $\mathbf{Y} \in R^p$, $p \geqslant 1$, could be missing. Also, define $\delta_i = 0$ if $\mathbf{Y}_i$ is missing; otherwise $\delta_i = 1$. Note that the full data may be represented by

$$\mathscr{D}_n = \{(\mathbf{Z}_1, \delta_1), \ldots, (\mathbf{Z}_n, \delta_n)\} = \{(\mathbf{X}_1, \mathbf{Y}_1, \delta_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n, \delta_n)\}.$$

There are many different missing mechanisms in the literature; see the monograph by Little and Rubin (2002). For example, the missing mechanism is sometimes called MCAR (missing completely at random), when $P(\delta = 1|\mathbf{X}, \mathbf{Y}) = P(\delta{=}1)$. A more realistic assumption widely used in the literature is the so-called MAR (missing at random) assumption, where, $P(\delta = 1|\mathbf{X}, \mathbf{Y}) = P(\delta = 1|\mathbf{X})$, and this will be our assumption in the rest of this article. Let $\mathscr{G}$ be a class of functions $g : R^{d+p} \to R$ and for each $g \in \mathscr{G}$ define the "mean"

$$\mathscr{V}(g) = E(g(\mathbf{Z})),$$

and its empirical version, based on the complete cases only, by $\mathscr{V}_n(g) = (1/r(n))\sum_{i=1}^n \delta_i g(\mathbf{Z}_i)$, where $r(n) = \sum_{i=1}^n \delta_i$ is the count of the complete cases. Unfortunately, if a large proportion of the observed data (say 70%) have missing $\mathbf{Y}_i$'s then, from a practical point of view, it makes sense to somehow revise $\mathscr{V}_n(g)$ to take into account the information which is available from the $\mathbf{X}_i$'s corresponding to those missing $\mathbf{Y}_i$'s. There are also fundamental theoretical reasons for revising the estimator $\mathscr{V}_n(g)$. For example, in general, under the MAR assumption, $E(\mathscr{V}_n(g))$ is different from $\mathscr{V}(g)$, which implies that the resulting empirical process $\{\mathscr{V}_n(g) - \mathscr{V}(g) : g \in \mathscr{G}\}$ is not centered (not even asymptotically). To motivate our estimation approach, consider the hypothetical situation where the regression function $E[g(\mathbf{Z})|\mathbf{X}]$ is available, and define the revised estimator

$$\tilde{\mathscr{V}}_n(g) = \frac{1}{n}\sum_{i=1}^n \{\delta_i g(\mathbf{Z}_i) + (1-\delta_i)E[g(\mathbf{Z}_i)|\mathbf{X}_i]\}. \tag{1}$$

It is a simple exercise to show that, under the MAR assumption, $\tilde{\mathscr{V}}_n(g)$ is unbiased for $\mathscr{V}(g)$. Of course, in practice, the regression function $E[g(\mathbf{Z}_i)|\mathbf{X}_i]$, used in (1), is not available and must therefore be estimated. Here we consider a Nadaraya–Watson-type kernel estimate of $\mathscr{V}(g)$:

$$\overline{\mathscr{V}}_n(g) = \frac{1}{n}\left[\sum_{i=1}^n \delta_i g(\mathbf{Z}_i) + \sum_{i=1}^n (1-\delta_i)\left\{\sum_{j=1,\neq i}^n \varphi_{n,j}(\mathbf{X}_i) g(\mathbf{Z}_j)\right\}\right], \tag{2}$$

where

$$\varphi_{n,j}(\mathbf{X}_i) = \frac{\delta_j \mathscr{K}((\mathbf{X}_j - \mathbf{X}_i)/h_n)}{\sum_{k=1,\neq i}^n \delta_k \mathscr{K}((\mathbf{X}_k - \mathbf{X}_i)/h_n)},$$

with the convention $0/0 = 0$, and $\mathscr{K} : R^d \to R$ is a kernel with the smoothing parameter h_n ($\to 0$ as $n \to \infty$). There are a number of important examples of the function g in the literature. The case where $Y \in R^1$ and $g(\mathbf{Z}) = g(Y, \mathbf{X}) \equiv Y$ has been studied in the literature extensively; see, for example, Cheng (1994). In this case $\overline{\mathscr{V}}_n(g) = \widehat{EY}$ $= n^{-1}\sum_{i=1}^n [\delta_i Y_i + (1-\delta_i)\hat{Y}_i]$, where $\hat{Y}_i = \sum_{j=1,\neq i}^n \varphi_{n,j}(\mathbf{X}_i) Y_j$. In fact, one can consider the more general class of estimators $n^{-1}\sum_{i=1}^n [\delta_i Y_i / p_n(\mathbf{X}_i) + (1 - \delta_i / p_n(\mathbf{X}_i))\hat{Y}_i]$, for some weight functions $p_n(\mathbf{X}_i)$, $i = 1, \ldots, n$; see, for example, Wang et al. (2004). When $p_n(\mathbf{X}_i) = 1$, this reduces to Cheng's imputation estimator, whereas $p_n(\mathbf{X}_i) = \infty$ gives Cheng's (1994) other estimator. The case where $p_n(\mathbf{X}_i)$ is the Nadaraya–Watson kernel regression estimator of $E(\delta_i|\mathbf{X}_i)$ corresponds to the so-called propensity score estimator of EY. These three estimators of EY turn out to be asymptotically equivalent to the estimator of Hirano and Ridder (2003), defined by $\widehat{EY} = n^{-1}\sum_{i=1}^n Y_i \delta_i / p_n(\mathbf{X}_i)$. This last estimator is essentially a *corrected* version of the naive estimator based on the complete cases, where the correction is done by weighting the complete cases by the inverse of the estimated missing data probabilities. The case where the missing data probabilities are known has been studied by Robins et al. (1994). In a more recent article, Wang et al. (2004) studied the situation where Y follows the semiparametric model, $Y_i = \mathbf{X}_i\beta + g(\mathbf{T}_i) + \varepsilon_i$; here the covariates $(\mathbf{X}_i, \mathbf{T}_i) \in R^d \times R^{d*}$ are observable and Y_i could be missing at random, (ε_i are independent with $E(\varepsilon_i|\mathbf{X}_i, \mathbf{T}_i) = 0$).

These authors show that their estimator is asymptotically more efficient under the semiparametric assumption. Another important example of the function g is the indicator function $g(\mathbf{Z}) = g_y(Y, \mathbf{X}) = I\{Y \leqslant y\}$. This corresponds to the estimation of the marginal cdf of Y and has already been considered by Cheng and Chu (1996). In this case the class $\mathscr{G}$ may be identified by the collection $\{I_C : C \in \{(-\infty, \mathbf{x}), \mathbf{x} \in R^d\}\}$. Unlike the usual approaches to imputation where one replaces each missing Y_i with some data-based version $\hat{Y}_i$, the approach adopted by Cheng and Chu focuses directly on the imputation of the indicator function itself. In addition to its intuitive appeal, this approach can also be more tractable from a theoretical standpoint for kernel-based estimators (clearly the indicator function of the event $\{\hat{Y}_i \leqslant y\}$, where $\hat{Y}$ is the imputed value of Y, is not easy to work with in general).

How good is $\tilde{\mathscr{V}}_n$ as an estimator of $\mathscr{V}$? Recall that a class of function $\mathscr{G}$ is said to be *totally bounded* w.r.t. the L_p-norm, $1 \leqslant p \leqslant \infty$, if for every $\varepsilon > 0$ there is a set $\mathscr{G}_\varepsilon = \{g_1, \ldots, g_{N(\varepsilon)}\}$ such that for every $g \in \mathscr{G}$, there is a $g^* \in \mathscr{G}_\varepsilon$ satisfying $\|g - g^*\|_p < \varepsilon$. Here $\mathscr{G}_\varepsilon$ is called an ε-cover of $\mathscr{G}$. The following exponential bounds on the uniform (in g) deviations of $\tilde{\mathscr{V}}(g)$ from $\mathscr{V}(g)$ are readily available.

Theorem 1.1. *Suppose that $\|g\|_\infty < B$, for every $g \in \mathscr{G}$. If $\mathscr{G}$ is totally bounded with respect to the $\|\|_\infty$-norm, then for all $\varepsilon > 0$ and all $n \geqslant 1$,*

$$P\left\{\sup_{g\in\mathscr{G}} |\tilde{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon\right\} \leqslant 2\mathscr{N}_\infty\left(\frac{\varepsilon}{3}, \mathscr{G}\right) \mathrm{e}^{-2n\varepsilon^2/(9B^2)},$$

where $\mathscr{N}_\infty(\varepsilon, \mathscr{G})$ is the size of the smallest ε-cover of $\mathscr{G}$.

Bounds such as the one in the above theorem have many applications in probability and statistics. For example, they can be used to establish strong uniform consistency results in nonparametric curve estimation problems; here a curve can be a regression function, a density function, a distribution function, etc.

To study the more practical estimator $\overline{\mathscr{V}}_n(g)$ that appear in (2), we first state a number of regularity conditions. In order to avoid having unstable estimates (in the tails of the pdf f of $\mathbf{X}$), it is often assumed in the literature on nonparametric regression that f is compactly supported and is bounded away from zero, and so we shall make this assumption. More specifically, it is assumed that

($\mathscr{F}$) *The pdf f of $\mathbf{X}$ satisfies* $\inf_{\mathbf{x}} f(\mathbf{x}) =: f_{\min} > 0$. *Also, both f and its first-order partial derivatives are uniformly bounded on the compact support of f.*

We shall further assume that the conditional probability of $\delta = 1$ is nonzero:

(p) $\inf_{\mathbf{x}} P\{\delta = 1 | \mathbf{X} = \mathbf{x}\} =: p_{\min} > 0$.

As for the choice of the kernel, we shall require

($\mathscr{K}$) *The kernel $\mathscr{K}$ is a pdf and satisfies* $\int |u_i| \mathscr{K}(\mathbf{u})\, \mathrm{d}\mathbf{u} < \infty$, $i = 1, \ldots, d$ *and* $\|\mathscr{K}\|_\infty < \infty$.

We assume in addition

($\mathscr{V}$) *The first-order partial derivatives of the function (of* $\mathbf{x}$*)* $E(\delta g(\mathbf{Z}) | \mathbf{X} = \mathbf{x})$ *exist and are bounded on the compact support of f, uniformly in* $\mathbf{x}$ *and g.*

Conditions ($\mathscr{F}$), (p), and ($\mathscr{K}$) are standard in the literature. Condition ($\mathscr{V}$), which has also been imposed by Cheng and Chu (1996), is technical. The following result gives exponential performance bounds on the uniform closeness of $\overline{\mathscr{V}}_n(g)$ to $\mathscr{V}(g)$ (cf. Remark 1.1 for measurability conditions):

Theorem 1.2. *Let $\mathscr{G}$ be a totally bounded class of functions $g : R^{d+p} \to R$, with $\|g\|_\infty < B$ for every g in $\mathscr{G}$. Also, let $\overline{\mathscr{V}}_n(g)$ be as in* (2). *If, $h_n \to 0$ as $n \to \infty$, then, under the stated conditions, for every $\varepsilon > 0$, $\exists\, n_0$ such that $\forall n > n_0$*

$$P\left\{\sup_{g\in\mathscr{G}} |\overline{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon\right\} \leqslant 10n\mathscr{N}_\infty\left(\frac{\varepsilon}{3}, \mathscr{G}\right) \mathrm{e}^{-Cnh_n^d\varepsilon^2},$$

where $C = (f_{\min}^2 p_{\min}^2)/[2(54B)\|\mathscr{K}\|_\infty(\|f\|_\infty + 2p_{\min} f_{\min}/3)]$ does not depend on n.

Quite often, with more efforts one can study the uniform almost sure properties of various curve estimates. The following theorem, which is not a corollary of Theorem 1.2, is one such result. Suppose that there are constants $\alpha \geqslant 0$,

$s > 0$, and $M > 0$ such that for every $x > 0$,

$$\log \mathcal{N}_\infty(x, \mathcal{G}) \leqslant H_\alpha(x) := \begin{cases} Mx^{-\alpha} & \text{if } \alpha > 0, \\ \log(Mx^{-s}) & \text{if } \alpha = 0. \end{cases} \tag{3}$$

Such conditions on the rates of growth of the entropy are common in the literature; see, for example, Polonik and Yao (2002). Since the case of $\alpha = 0$ is trivial (and not interesting in practice), the following result focuses on the case where $\alpha > 0$.

Theorem 1.3 (*Not a consequence of Theorem 1.2*). *Suppose that the bound* (3) *holds for some* $\alpha > 0$, *and that as* $n \to \infty$,

$$\frac{\log n}{nh_n^d} \longrightarrow 0 \quad \textit{and} \quad \left(\frac{nh_n^d}{\log n}\right)^{1/(2+\alpha)} h_n \longrightarrow 0. \tag{4}$$

Then, *under the conditions of Theorem* 1.2,

$$\lim_{n\to\infty} \left(\frac{nh_n^d}{\log n}\right)^{1/(2+\alpha)} \sup_{g\in\mathcal{G}} |\overline{\mathscr{V}}_n(g) - \mathscr{V}(g)| \overset{\text{a.s.}}{=} 0.$$

Remark 1.1. When the class $\mathcal{G}$ is uncountable, the measurability of the supremum in the above theorems can become an important issue. One way to handle the measurability problem is to work with the outer probability (see the book by van der Vaart and Wellner (1996)). Alternatively, one may be able to avoid measurability difficulties by imposing the so-called *universal separability* assumption on the class $\mathcal{G}$, (cf. Pollard, 1984, p. 38). That is, there is a countable subclass $\mathcal{G}_0 \subset \mathcal{G}$ such that each $g \in \mathcal{G}$ can be written as a pointwise limit of a sequence in $\mathcal{G}_0$. In the sequel, we shall assume that the supremum functionals do satisfy measurability conditions.

2. More flexible estimators and multiple imputation

Recall the estimator $\overline{\mathscr{V}}_n(g)$ in (2):

$$\overline{\mathscr{V}}_n(g) = \frac{1}{n}\left[\sum_{i=1}^n \delta_i g(\mathbf{Z}_i) + \sum_{i=1}^n (1-\delta_i)\left\{\sum_{j=1,\neq i}^n \varphi_{n,j}(\mathbf{X}_i) g(\mathbf{Z}_j)\right\}\right],$$

where $\varphi_{n,j}(\mathbf{X}_i) = \delta_j \mathscr{K}((\mathbf{X}_j - \mathbf{X}_i)/h_n) \div \sum_{k=1,\neq i}^n \delta_k \mathscr{K}((\mathbf{X}_k - \mathbf{X}_i)/h_n)$. Observe that the term $\sum_{j=1,\neq i}^n \varphi_{n,j}(\mathbf{X}_i) g(\mathbf{Z}_j)$ in the above expression, which is the kernel regression estimate of $E[g(\mathbf{Z}_i)|\mathbf{X}_i]$ based on the complete cases, is our imputed "value" of $g(\mathbf{Z}_i)$, for the case where $\delta_i = 0$. Since each missing $g(\mathbf{Z}_i)$ is imputed once only, the estimator $\overline{\mathscr{V}}_n$ above is a *single imputation estimator*. Another popular estimator in the literature is based on *multiple imputation*; see Kolmogorov and Tikhomirov (1959). This works by constructing many, say $N \geqslant 2$, different imputed values of $g(\mathbf{Z}_i)$, which would then result in N estimates of $\mathscr{V}(g)$: $\widehat{\mathscr{V}}_{n,1}(g), \ldots, \widehat{\mathscr{V}}_{n,N}(g)$. The multiple imputation (MI) estimator corresponding to these N single estimators is simply the average

$$\widehat{\mathscr{V}}_{\text{MI}}(g) = \frac{1}{N}\sum_{k=1}^N \widehat{\mathscr{V}}_{n,k}(g).$$

To construct our proposed multiple imputation estimator of $\mathscr{V}$ start by randomly splitting the sample $\mathscr{D}_n$ into a subsample of size ℓ, say $\mathscr{D}_\ell$, and the remaining part $\mathscr{D}_m = \mathscr{D}_n - \mathscr{D}_\ell$, of size $m = n - \ell$. Also, let

$$\widehat{\mathscr{V}}_n(g) = \frac{1}{n}\left\{\sum_{i=1}^n \delta_i g(\mathbf{Z}_i) + \frac{n}{m}\sum_{i:(\mathbf{Z}_i,\delta_i)\in\mathscr{D}_m} (1-\delta_i)\left[\sum_{j:(\mathbf{Z}_j,\delta_j)\in\mathscr{D}_\ell} \varphi_{\ell,j}(\mathbf{X}_i) g(\mathbf{Z}_j)\right]\right\}, \tag{5}$$

where

$$\varphi_{\ell,j}(\mathbf{x}) = \delta_j \mathscr{K}\left(\frac{\mathbf{X}_j - \mathbf{x}}{h_\ell}\right) \Big/ \sum_{k:(\mathbf{X}_k,\delta_k)\in\mathscr{D}_\ell} \delta_k \mathscr{K}\left(\frac{\mathbf{X}_k - \mathbf{x}}{h_\ell}\right). \tag{6}$$

Repeating this process a total of N times results in the sample splits $(\mathscr{D}_\ell^{(1)}, \mathscr{D}_m^{(1)}), \ldots, (\mathscr{D}_\ell^{(N)}, \mathscr{D}_m^{(N)})$ and the corresponding estimators $\widehat{\mathscr{V}}_{n,1}(g), \ldots, \widehat{\mathscr{V}}_{n,\mathrm{N}}(g)$ of $\mathscr{V}(g)$. Finally, define the imputation estimator $\widehat{\mathscr{V}}_{\mathrm{MI}}(g) = N^{-1}\sum_{r=1}^{N} \widehat{\mathscr{V}}_{n,r}(g)$.

How good is $\widehat{\mathscr{V}}_{\mathrm{MI}}(g)$? The following performance bounds show that $\widehat{\mathscr{V}}_{\mathrm{MI}}(g)$ is uniformly (in g) close to $\mathscr{V}$, for large n. First we need the following condition:

$$(\mathscr{ML}) \qquad h_n \to 0, \quad nh_n^d \to \infty, \quad \ell \to \infty, \quad \text{and} \quad \ell/m \to 0 \quad \text{as } n \to \infty.$$

Theorem 2.1. *Let $\mathscr{G}$ be as in Theorem* 1.2 *and suppose that conditions $(\mathscr{F})$, (p), $(\mathscr{K})$, $(\mathscr{V})$, and $(\mathscr{ML})$ hold. Then, for every $\varepsilon > 0$ there is a n_0 such that for all $n > n_0$,*

$$P\left\{\sup_{g\in\mathscr{G}} |\widehat{\mathscr{V}}_{\mathrm{MI}}(g) - \mathscr{V}(g)| > \varepsilon\right\} \leqslant C_1 N n \mathscr{N}_\infty\left(\frac{\varepsilon}{3}, \mathscr{G}\right) \mathrm{e}^{-C_2 \ell h_\ell^d \varepsilon^2},$$

where C_1 and C_2 are positive constants not depending on n.

Note that strong convergence results for $\widehat{\mathscr{V}}_{\mathrm{MI}}(g)$, (uniformly in g), follows from the Borel–Cantelli lamma, under the minimal condition that $\log(N \vee n)/(\ell h_\ell^d) \to 0$, as $n \to \infty$.

Supnorm covers, or equivalently $\mathscr{N}_\infty(\varepsilon, \mathscr{G})$, can be very large. Of course, for $1 \leqslant p < \infty$, L_p-norm covers are weaker; unfortunately, they depend on the underlying distribution which is unknown. In what follows we consider data-based covers of $\mathscr{G}$ based on weighted empirical norms. More specifically, fixed $(\mathbf{z}_1, \delta_1), \ldots, (\mathbf{z}_n, \delta_n)$, $\mathbf{x}$, and consider the weighted seminorm

$$\|g\|_{\mathscr{W}_n(\mathbf{x})} = \sum_{j=1}^{n} \mathscr{W}_{n,j}(\mathbf{x})\delta_j |g(\mathbf{z}_j)|,$$

where $\mathscr{W}_{n,j}(\mathbf{x}) = \mathscr{W}_{n,j}(\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n) \in R$, $j = 1, \ldots, n$ are weight functions depending on $\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n$, but not $\mathbf{y}_1, \ldots, \mathbf{y}_n$, (recall that $\mathbf{z} = (\mathbf{x}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}})^{\mathrm{T}}$), and satisfying $\sum_{j=1}^{n} \mathscr{W}_{n,j}(\mathbf{x}) = 1$. Note that if $\mathscr{W}_{n,j}(\mathbf{x}) = 1/n$ for all j's then $\|g\|_{\mathscr{W}_n(\mathbf{x})} = \|g\|_{1,n} := (1/n)\sum_{j=1}^{n} \delta_j |g(\mathbf{z}_j)|$. For any $\varepsilon > 0$ define $\mathscr{N}(\varepsilon, \mathscr{G}, \|\|_{\mathscr{W}_n(\mathbf{x})})$ to be the cardinality of the smallest collection of functions $\mathscr{G}_{n,\varepsilon} = \{g_1, \ldots, g_{n_{\varepsilon,\mathbf{x}}}\}$ with the property that for every $g \in \mathscr{G}$, $\exists g^* \in \mathscr{G}_{n,\varepsilon}$ such that $\|g - g^*\|_{\mathscr{W}_n(\mathbf{x})} < \varepsilon$. Clearly, $\mathscr{N}(\varepsilon, \mathscr{G}, \|\|_{\mathscr{W}_n(\mathbf{x})}) \leqslant \mathscr{N}_\infty(\varepsilon, \mathscr{G})$. There are many choices for the weight function $\mathscr{W}_n$. Here we have in mind the kernel-type weight

$$\mathscr{W}_{n,j}(\mathbf{x}) = \frac{\mathscr{K}((\mathbf{X}_j - \mathbf{x})/h_n)}{\sum_{k=1}^{n} \mathscr{K}((\mathbf{X}_k - \mathbf{x})/h_n)},$$

for the kernel $\mathscr{K} : R^d \to R$. The following theorem provides exponential performance bounds for the uniform deviations of $\widehat{\mathscr{V}}_{\mathrm{MI}}(g)$ from $\mathscr{V}(g)$.

Theorem 2.2. *Let $\mathscr{G}$ be a class of functions $g : R^{d+p} \to R$, with $\|g\|_\infty < B$ for every g in $\mathscr{G}$, and put $m = n - \ell$. Suppose that conditions $(\mathscr{F})$, (p), $(\mathscr{K})$, $(\mathscr{V})$, and $(\mathscr{ML})$ hold. Then, for every $\varepsilon > 0$ there is a n_0 such that for all $n > n_0$,*

$$\begin{aligned} P\left\{\sup_{g\in\mathscr{G}} |\widehat{\mathscr{V}}_{\mathrm{MI}}(g) - \mathscr{V}(g)| > \varepsilon\right\} \leqslant\ & 8NE\left[\mathscr{N}\left(\frac{\varepsilon}{16}, \mathscr{G}, \|\|_{1,n}\right)\right] \mathrm{e}^{-n\varepsilon^2/(512B^2)} \\ & + 8NmE\left[\mathscr{N}\left(\frac{p_{\min}\varepsilon}{32}, \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X})}\right)\right] \mathrm{e}^{-m\varepsilon^2/(512B^2)} \\ & + 8Nm\Delta_\ell(\mathscr{G}, \varepsilon)[5\mathrm{e}^{-\ell h_\ell^d c_1} + 3\mathrm{e}^{-\ell h_\ell^d c_2 \varepsilon^2}], \end{aligned}$$

where c_1 and c_2 are constants not depending on n and

$$\Delta_\ell(\mathscr{G}, \varepsilon) = \sqrt{E\left[\mathscr{N}\left(\frac{p_{\min} f_{\min}\varepsilon}{864\|f\|_\infty}, \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X})}\right)\right]^2}. \tag{7}$$

Remark 2.1. The constants c_1 and c_2 that appear in the bound of Theorem 2.2 depend on many unknown parameters, and no attempts have been made to obtain their optimal values. However, for later sections, they may be taken to be $c_1 = b_1 \wedge b_2 \wedge b_3 \wedge b_4$ and $c_2 = b_5 \wedge b_6 \wedge b_7$, where

$$b_1 = 1/(4\|\mathscr{K}\|_\infty(\|f\|_\infty + 1)), \tag{8}$$

$$b_2 = \|f\|_\infty/(12\|\mathscr{K}\|_\infty), \tag{9}$$

$$b_3 = \frac{f_{\min}^2}{32\|\mathscr{K}\|_\infty(\|f\|_\infty + f_{\min}/4)}, \tag{10}$$

$$b_4 = \frac{p_{\min}^2 f_{\min}^2}{128\|\mathscr{K}\|_\infty(\|f\|_\infty + p_{\min} f_{\min}/8)}, \tag{11}$$

$$b_5 = \frac{p_{\min}^2 f_{\min}^2}{8(54)^2 B^2\|\mathscr{K}\|_\infty(\|f\|_\infty + p_{\min} f_{\min}/3)}, \tag{12}$$

$$b_6 = \frac{p_{\min}^2 f_{\min}^2}{8(108)^2 B^2\|\mathscr{K}\|_\infty(\|f\|_\infty + p_{\min} f_{\min}/36)}, \tag{13}$$

$$b_7 = \frac{p_{\min}^2 f_{\min}^2}{512(54)^2 B^2\|\mathscr{K}\|_\infty(\|f\|_\infty + 1)}. \tag{14}$$

In practice, when the members of the class $\mathscr{G}$ have certain desirable functional properties, one may then revise the estimator in (5) accordingly. Consider the following setup. Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \mathbf{X}^{(2)} \\ \mathbf{Y} \end{pmatrix} \quad \text{where } \mathbf{X}^{(1)} \in R^{d_1}, \ \mathbf{X}^{(2)} \in R^{d_2}, \ \text{and } \ d_1 + d_2 = d.$$

Also, suppose that $g = g_1 * g_2$ for every $g \in \mathscr{G}$, where $g_1 : R^{d_1} \to R^1$ and $g_2 : R^{d_2} \to R^1$. When the operation '$*$' is either '$\times$' or '$+$', or '$-$', one obtains $E[g(\mathbf{Z})|\mathbf{X} = \mathbf{x}] = g_1(\mathbf{x}^{(1)}) * E[g_2(\mathbf{U})|\mathbf{X} = \mathbf{x}]$, and thus the estimator in (5) may be changed to

$$\widehat{\mathscr{V}}_n(g) = \frac{1}{n}\left\{\sum_{i=1}^n \delta_i g(\mathbf{Z}_i) + \frac{n}{m}\sum_{i:(\mathbf{Z}_i,\delta_i)\in\mathscr{D}_m}(1-\delta_i)\left[g_1(\mathbf{X}_i^{(1)}) * \sum_{j:(\mathbf{Z}_j,\delta_j)\in\mathscr{D}_\ell}\varphi_{\ell,j}(\mathbf{X}_i)g_2(\mathbf{U}_j)\right]\right\}.$$

In this case, one may revise Theorem 2.2 as well. In fact, the following result is the counterpart of Theorem 2.2 for the case where g is multiplicative: $g(\mathbf{z}) = g((\mathbf{x}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}})^{\mathrm{T}}) = g_1(\mathbf{x}^{(1)}) \times g_2(\mathbf{u})$, $\exists g_1 \in \mathscr{G}_1$, $g_2 \in \mathscr{G}_2$, for some classes of functions $\mathscr{G}_1$ and $\mathscr{G}_2$. Of course, the choice of the classes $\mathscr{G}_1$ and $\mathscr{G}_2$ are determined by the statement of the problem of interest. (An example along these lines is the problem of nonparametric least-squares estimation in the presence of missing covariates; see Section 3.)

Theorem 2.3. *Let $\mathscr{G} \equiv \mathscr{G}_1 \times \mathscr{G}_2$ be a multiplicative class of functions, as stated above, with the property that $\|g_1\|_\infty \leqslant B_1 < \infty$ and $\|g_2\|_\infty \leqslant B_2 < \infty$, for every $g_1 \in \mathscr{G}_1$ and every $g_2 \in \mathscr{G}_2$. Then, under the conditions of*

Theorem 2.2, *for every* $\varepsilon > 0$ *there is a* n_0 *such that, for all* $n > n_0$,

$$P\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}_{\mathrm{MI}}(g)-\mathscr{V}(g)|>\varepsilon\right\}$$
$$\leqslant c_3 N E\left[\mathscr{N}\left(\frac{\varepsilon}{16},\mathscr{G},\|\|_{1,n}\right)\right]\mathrm{e}^{-n\varepsilon^2/(512B^2)}+c_4 N m E[\mathscr{N}(c_5\varepsilon,\mathscr{G}_2,\|\|_{\mathscr{W}_\ell(\mathbf{X})})]\mathrm{e}^{-c_6 m\varepsilon^2}$$
$$+c_7 N m\sqrt{E[\mathscr{N}(c_8\varepsilon,\mathscr{G}_2,\|\|_{\mathscr{W}_\ell(\mathbf{X})})]^2}\times[c_9\mathrm{e}^{-c_{10}\ell h_\ell^d}+c_{11}\mathrm{e}^{-c_{12}\ell h_\ell^d\varepsilon^2}],$$

where $c_3,\ldots,c_{12}$ *are positive constants not depending on n.*

The following result, which is the counterpart of Theorem 1.3 for the multiple imputation estimator $\widehat{\mathscr{V}}_{\mathrm{MI}}(g)$, is useful when studying the uniform asymptotic properties of the least-squares regression (as well as density) estimators, with missing data.

Theorem 2.4. *Let* $\widehat{\mathscr{V}}_{\mathrm{MI}}(g)$ *be as before and suppose that the bound* (3) *holds for some* $\alpha\geqslant 0$. *Also, suppose that conditions* $(\mathscr{F})$, (p), $(\mathscr{K})$, $(\mathscr{V})$, *and* $(\mathscr{ML})$ *hold. If, as* $n\to\infty$, $N\equiv N(n)\to\infty$ *and*

$$\frac{\log(m\vee N)}{\ell h_\ell^d}\longrightarrow 0,\quad \left(\frac{\ell h_\ell^d}{\log(m\vee N)}\right)^{1/(2+\alpha)}h_\ell\longrightarrow 0,\quad \textit{and}\quad \frac{\ell}{nh_\ell}\longrightarrow 0,$$

then

$$\lim_{n\to\infty}\left(\frac{\ell h_\ell^d}{\log(m\vee N)}\right)^{1/(2+\alpha)}\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}_{\mathrm{MI}}(g)-\mathscr{V}(g)|\overset{\text{a.s.}}{=}0.$$

Proof of Theorem 2.2. Start with the simple bound

$$P\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}_{\mathrm{MI}}(g)-\mathscr{V}(g)|>\varepsilon\right\}\leqslant\sum_{r=1}^{N}P\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}_{n,r}(g)-\mathscr{V}(g)|>\varepsilon\right\}. \tag{15}$$

Let $\widehat{\mathscr{V}}_n(g)$ be as in (5) and observe that $\widehat{\mathscr{V}}_{n,r}(g)\overset{\mathrm{d}}{=}\widehat{\mathscr{V}}_n(g)$, $r=1,\ldots,N$. Furthermore, since the data are iid, we may assume w.o.l.g that $\mathscr{D}_\ell=\{(\mathbf{Z}_1,\delta_1),\ldots,(\mathbf{Z}_\ell,\delta_\ell)\}$, which can always be achieved by a re-indexing of the observations in $\mathscr{D}_\ell$. Thus, one may take

$$\widehat{\mathscr{V}}_n(g)=\frac{1}{n}\left\{\sum_{i=1}^{n}\delta_i g(\mathbf{Z}_i)+\frac{n}{m}\sum_{i=\ell+1}^{n}(1-\delta_i)\left[\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}_i)g(\mathbf{Z}_j)\right]\right\}. \tag{16}$$

We will use a nonstandard symmetrization argument in the rest of the proof.

1. *A nonstandard first symmetrization* (*w.r.t. a hypothetical sample*): Let $\mathscr{D}'_n=\{(\mathbf{Z}'_1,\delta'_1),\ldots,(\mathbf{Z}'_n,\delta'_n)\}$ be a *ghost* sample, i.e., $(\mathbf{Z}'_i,\delta'_i)\overset{\text{iid}}{=}(\mathbf{Z}_1,\delta_1)$. Also define

$$\widehat{\mathscr{V}}'_n(g)=\frac{1}{n}\left[\sum_{i=1}^{n}\delta'_i g(\mathbf{Z}'_i)+\frac{n}{m}\sum_{i=\ell+1}^{n}(1-\delta'_i)\left\{\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}'_i)g(\mathbf{Z}_j)\right\}\right], \tag{17}$$

where $\varphi_{\ell,j}(\cdot)$ is as in (6). Note that (17) is not exactly the counterpart of (16); this is because unlike (16), the expression $\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}'_i)g(\mathbf{Z}_j)$ in (17) is a function of both the true and the hypothetical samples. (In fact, the term $\varphi_{\ell,j}(\mathbf{X}'_i)$ in (17) is a function of both $\mathbf{X}'_i$ and $(\delta_1,\mathbf{X}_1),\ldots,(\delta_\ell,\mathbf{X}_\ell)$.) Here, $\widehat{\mathscr{V}}'_n(g)$ does not have any direct applications as an estimator of $\mathscr{V}(g)$; it is only a symmetrization device to deal with the empirical process corresponding to $\widehat{\mathscr{V}}_n(g)$. To this end, fix the data $\mathscr{D}_n$ and observe that if $\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}_n(g)-\mathscr{V}(g)|>\varepsilon$, then there is at least some $g_\varepsilon\in\mathscr{G}$, which will

depend on $\mathscr{D}_n$ (but not the ghost sample $\mathscr{D}'_n$), such that $|\widehat{\mathscr{V}}_n(g_\varepsilon) - \mathscr{V}(g_\varepsilon|\mathscr{D}_n)| > \varepsilon$, where $\mathscr{V}(g_\varepsilon|\mathscr{D}_n) = E[g_\varepsilon(\mathbf{Z})|\mathscr{D}_n]$. Now, put

$$\beta_n(\mathscr{D}_n) := P\left\{|\widehat{\mathscr{V}}'_n(g_\varepsilon) - \mathscr{V}(g_\varepsilon|\mathscr{D}_n)| < \frac{\varepsilon}{2}\Big|\,\mathscr{D}_n\right\}$$

and note that

$$\begin{aligned}\beta_n(\mathscr{D}_n) &\leqslant P\left\{-|\widehat{\mathscr{V}}'_n(g_\varepsilon) - \widehat{\mathscr{V}}_n(g_\varepsilon)| + |\widehat{\mathscr{V}}_n(g_\varepsilon) - \mathscr{V}(g_\varepsilon|\mathscr{D}_n)| < \frac{\varepsilon}{2}\Big|\,\mathscr{D}_n\right\}\\ &\leqslant P\left\{|\widehat{\mathscr{V}}'_n(g_\varepsilon) - \widehat{\mathscr{V}}_n(g_\varepsilon)| > \frac{\varepsilon}{2}\Big|\,\mathscr{D}_n\right\}\\ &\leqslant P\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}'_n(g) - \widehat{\mathscr{V}}_n(g)| > \frac{\varepsilon}{2}\Big|\,\mathscr{D}_n\right\}. \end{aligned} \tag{18}$$

Next, let

$$I_{n,i}(g) = \frac{\sum_{j=1}^{\ell}\delta_j g(\mathbf{Z}_j)\mathscr{K}((\mathbf{X}_j - \mathbf{X}'_i)/h_\ell)}{\sum_{j=1}^{\ell}\delta_j \mathscr{K}((\mathbf{X}_j - \mathbf{X}'_i)/h_\ell)} - \frac{E[\delta'_i g(\mathbf{Z}'_i)|\mathbf{X}'_i]}{p(\mathbf{X}'_i)}, \tag{19}$$

and define

$$\pi_n(\mathscr{D}_n) := P\left\{\sum_{i=\ell+1}^{n}\sup_{g\in\mathscr{G}}|I_{n,i}(g)| \geqslant \frac{n\varepsilon}{18}\Big|\,\mathscr{D}_n\right\}, \tag{20}$$

and

$$\pi_n = E[\pi_n(\mathscr{D}_n)] = P\left\{\sum_{i=\ell+1}^{n}\sup_{g\in\mathscr{G}}|I_{n,i}(g)| \geqslant \frac{n\varepsilon}{18}\right\}. \tag{21}$$

It will be shown in Section 4 that

$$\pi_n \leqslant 12m\Delta_\ell(\mathscr{G},\varepsilon)[\mathrm{e}^{-\ell h_\ell^d b_8\varepsilon^2} + \mathrm{e}^{-\ell h_\ell^d b_9}], \tag{22}$$

and

$$\beta_n(\mathscr{D}_n) \geqslant 1 - 4\mathrm{e}^{-n\varepsilon^2/(288B^2)} - \pi_n(\mathscr{D}_n), \tag{23}$$

where $\Delta_\ell(\mathscr{G},\varepsilon)$ is as in Theorem 2.2, $b_8 = b_6 \wedge b_6 \wedge b_7$, and $b_9 = b_1 \wedge b_2$, and the constants $b_1,\ldots,b_7$ are as in (8)–(14). Now, observe that the above lower bound on $\beta_n(\mathscr{D}_n)$ and the upper bound on the far right side of (18) do not depend on any specific g_ε, and that the chain of inequalities between them (in (18)) remains valid on the set $\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon\}$. Therefore, integrating the two sides with respect to $\mathscr{D}_n$, over this set, one finds

$$E\left[(1 - 4\mathrm{e}^{-n\varepsilon^2/288B^2} - \pi_n(\mathscr{D}_n))I\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon\right\}\right] \leqslant P\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}'_n(g) - \widehat{\mathscr{V}}_n(g)| > \frac{\varepsilon}{2}\right\}. \tag{24}$$

On the other hand, since

$$E\left[\pi_n(\mathscr{D}_n)I\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon\right\}\right] \leqslant E[\pi_n(\mathscr{D}_n)] \overset{(2.21)}{:=} \pi_n,$$

one concludes that, for large n,

$$P\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon\right\} \leqslant \frac{1}{1 - 4\mathrm{e}^{-n\varepsilon^2/(288B^2)}}\left[P\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}'_n(g) - \widehat{\mathscr{V}}_n(g)| > \frac{\varepsilon}{2}\right\} + \pi_n\right]. \tag{25}$$

To complete the proof of the theorem, it remains to bound the probability statement that appears on the r.h.s. of (25). First note that, with $\widehat{\mathscr{V}}_n$ and $\widehat{\mathscr{V}}'_n$ given by (5) and (17), respectively, one may write

$$P\left\{\sup_{g\in\mathscr{G}}|\widehat{\mathscr{V}}'_n(g)-\widehat{\mathscr{V}}_n(g)|>\frac{\varepsilon}{2}\right\}\leqslant P\left\{\sup_{g\in\mathscr{G}}\frac{1}{n}\left|\sum_{i=1}^{n}[\delta_i g(\mathbf{Z}_i)-\delta'_i g(\mathbf{Z}'_i)]\right|>\frac{\varepsilon}{4}\right\}$$
$$+P\left\{\sup_{g\in\mathscr{G}}\frac{1}{m}\left|\sum_{i=\ell+1}^{n}\left[(1-\delta_i)\left(\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}_i)g(\mathbf{Z}_j)\right)\right.\right.\right.$$
$$\left.\left.\left.-(1-\delta'_i)\left(\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}'_i)g(\mathbf{Z}_j)\right)\right]\right|>\frac{\varepsilon}{4}\right\}$$
$$:=\pi_1(n)+\pi_2(n)\quad\text{(say)}. \tag{26}$$

2. *Second symmetrization* (*w.r.t. an independent Rademacher sequence*): Let $R_1,\ldots,R_n$ be an iid Rademacher sequence (i.e., $P(R_i=+1)=1/2=P(R_i=-1)$), independent of $\mathscr{D}_n$ and $\mathscr{D}'_n$. Since the joint distribution of $(\delta_1 g(\mathbf{Z}_1),\ldots,\delta_n g(\mathbf{Z}_n))$ and that of $(\delta'_1 g(\mathbf{Z}'_1),\ldots,\delta'_n g(\mathbf{Z}'_n))$ are not affected by randomly interchanging their corresponding components, one can write

$$\pi_1(n)=P\left\{\sup_{g\in\mathscr{G}}\frac{1}{n}\left|\sum_{i=1}^{n}R_i[\delta_i g(\mathbf{Z}_i)-\delta'_i g(\mathbf{Z}'_i)]\right|>\frac{\varepsilon}{4}\right\}\leqslant 2P\left\{\sup_{g\in\mathscr{G}}\frac{1}{n}\left|\sum_{i=1}^{n}R_i\delta_i g(\mathbf{Z}_i)\right|>\frac{\varepsilon}{8}\right\}$$
$$\leqslant 4E\left[\mathscr{N}\left(\frac{\varepsilon}{16},\mathscr{G},\|\|_{1,n}\right)\right]\mathrm{e}^{-n\varepsilon^2/(512B^2)}. \tag{27}$$

As for the term $\pi_2(n)$ in (26), first observe that

$$(1-\delta_i)\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}_i)g(\mathbf{Z}_j)=(1-\delta_i)\frac{\sum_{j=1}^{\ell}\delta_j g(\mathbf{Z}_j)\mathscr{K}((\mathbf{X}_j-\mathbf{X}_i)/h_\ell)}{\sum_{k=1}^{\ell}\delta_k\mathscr{K}((\mathbf{X}_k-\mathbf{X}_i)/h_\ell)}$$
$$:=H_g(\mathscr{D}_\ell,(\mathbf{X}_i,\delta_i))\quad\text{(say)}\quad i=\ell+1,\ldots,n.$$

Similarly,

$$(1-\delta'_i)\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}'_i)g(\mathbf{Z}_j)=(1-\delta'_i)\frac{\sum_{j=1}^{\ell}\delta_j g(\mathbf{Z}_j)\mathscr{K}((\mathbf{X}_j-\mathbf{X}'_i)/h_\ell)}{\sum_{k=1}^{\ell}\delta_k\mathscr{K}((\mathbf{X}_k-\mathbf{X}'_i)/h_\ell)}$$
$$:=H_g(\mathscr{D}_\ell,(\mathbf{X}'_i,\delta'_i)),\quad i=\ell+1,\ldots,n.$$

Therefore

$$\pi_2(n)=P\left\{\sup_{g\in\mathscr{G}}m^{-1}\left|\sum_{i=\ell+1}^{n}[H_g(\mathscr{D}_\ell,(\mathbf{X}_i,\delta_i))-H_g(\mathscr{D}_\ell,(\mathbf{X}'_i,\delta'_i))]\right|>\frac{\varepsilon}{4}\right\}.$$

Furthermore, $(\mathbf{X}_i,\delta_i)\overset{\text{iid}}{=}(\mathbf{X}'_i,\delta'_i)$, $i=\ell+1,\ldots,n$ and $\mathscr{D}_\ell$ is independent of both $(\mathbf{X}_i,\delta_i)_{i=\ell+1}^n$ and $(\mathbf{X}'_i,\delta'_i)_{i=\ell+1}^n$. Consequently, the joint distribution of the vector

$$(H_g(\mathscr{D}_\ell,(\mathbf{X}_{\ell+1},\delta_{\ell+1})),\ldots,H_g(\mathscr{D}_\ell,(\mathbf{X}_n,\delta_n)))$$

is the same as that of the vector

$$(H_g(\mathscr{D}_\ell,(\mathbf{X}'_{\ell+1},\delta'_{\ell+1})),\ldots,H_g(\mathscr{D}_\ell,(\mathbf{X}'_n,\delta'_n))),$$

and, more importantly, this joint distribution is not affected if one randomly interchanges the corresponding components of these two vectors. Therefore, for an iid Rademacher sequence $R_1, \ldots, R_n$, (which is taken independent of $\mathscr{D}_\ell$, $(\mathbf{X}_i, \delta_i)$, and $(\mathbf{X}'_i, \delta'_i)$, $i = \ell+1, \ldots, n$), one deduces

$$\begin{aligned}
\pi_2(n) &= P\left\{\sup_{g\in\mathscr{G}} m^{-1}\left|\sum_{i=\ell+1}^{n} R_i[H_g(\mathscr{D}_\ell, (\mathbf{X}_i, \delta_i)) - H_g(\mathscr{D}_\ell, (\mathbf{X}'_i, \delta'_i))]\right| > \frac{\varepsilon}{4}\right\} \\
&\leqslant P\left\{\sup_{g\in\mathscr{G}} m^{-1}\left|\sum_{i=\ell+1}^{n} R_i(1-\delta_i)\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}_i)g(\mathbf{Z}_j)\right| > \frac{\varepsilon}{8}\right\} \\
&\quad + P\left\{\sup_{g\in\mathscr{G}} m^{-1}\left|\sum_{i=\ell+1}^{n} R_i(1-\delta'_i)\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}'_i)g(\mathbf{Z}_j)\right| > \frac{\varepsilon}{8}\right\}. \\
&:= I_n + II_n = 2I_n
\end{aligned} \tag{28}$$

(since I_n and II_n are precisely the same). To bound the term I_n, fix $\mathscr{D}_n$ and for any $\varepsilon > 0$ put

$$\varepsilon' = \varepsilon p_{\min}/32. \tag{29}$$

Let $\mathscr{G}^{(i)}_{\ell,\varepsilon'}$ be an ε'-cover of $\mathscr{G}$, w.r.t. $\|\|_{\mathscr{W}_\ell(\mathbf{X}_i)}$, i.e., for every $g \in \mathscr{G}$, there is a $g^{(i)}_* \in \mathscr{G}^{(i)}_{\ell,\varepsilon'}$ such that,

$$\|g - g^{(i)}_*\|_{\mathscr{W}_\ell(\mathbf{X}_i)} < \varepsilon'. \tag{30}$$

Also, let $\mathscr{N}(\varepsilon', \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X}_i)})$ be the ε'-covering number of $\mathscr{G}$, with respect to $\|\|_{\mathscr{W}_\ell(\mathbf{X}_i)}$. For each $i = \ell+1, \ldots, n$, choose $g^{(i)}_* \in \mathscr{G}^{(i)}_{\ell,\varepsilon'}$ such that $\|g - g^{(i)}_*\|_{\mathscr{W}_\ell(\mathbf{X}_i)} < \varepsilon'$, and write

$$\begin{aligned}
&\left|\sum_{i=\ell+1}^{n}\left[R_i(1-\delta_i)\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}_i)g(\mathbf{Z}_j)\right]\right| \\
&\leqslant \left|\sum_{i=\ell+1}^{n}\left[R_i(1-\delta_i)\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}_i)g^{(i)}_*(\mathbf{Z}_j)\right]\right| \\
&\quad + \left|\sum_{i=\ell+1}^{n}\left[R_i(1-\delta_i)\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}_i)(g(\mathbf{Z}_j) - g^{(i)}_*(\mathbf{Z}_j))\right]\right|.
\end{aligned} \tag{31}$$

For $i = \ell+1, \ldots, n$, let $\hat{p}(\mathbf{X}_i) = \sum_{j=1}^{\ell}\delta_j \mathscr{K}((\mathbf{X}_j - \mathbf{X}_i)/h_\ell)/\sum_{j=1}^{\ell}\mathscr{K}((\mathbf{X}_j - \mathbf{X}_i)/h_\ell)$ be the kernel estimate of $E(\delta_i|\mathbf{X}_i)$. Then the 2nd term on the right hand side of (31) is bounded by

$$\begin{aligned}
\sum_{i=\ell+1}^{n}\left[|R_i(1-\delta_i)|\frac{1}{\hat{p}(\mathbf{X}_i)}\|g - g^{(i)}_*\|_{\mathscr{W}_\ell(\mathbf{X}_i)}\right] &\leqslant \sum_{i=\ell+1}^{n}\frac{\varepsilon'}{\hat{p}(\mathbf{X}_i)} \qquad \text{(by (30))} \\
&= \frac{p_{\min}\varepsilon}{32}\sum_{i=\ell+1}^{n}\frac{1}{\hat{p}(\mathbf{X}_i)}.
\end{aligned} \tag{32}$$

Define the events Ω_n and $\mathbb{E}_{n,\varepsilon}$ according to:

$$\Omega_n = \bigcap_{i=\ell+1}^{n}\left\{\hat{p}(\mathbf{X}_i) \geqslant \frac{1}{2}p_{\min}\right\},$$

and

$$\mathbb{E}_{n,\varepsilon} = \left\{ \sup_{g\in\bigcup_{i=\ell+1}^{n}\mathscr{G}_{\ell,\varepsilon'}^{(i)}} m^{-1}\left|\sum_{i=\ell+1}^{n}\left[R_i(1-\delta_i)\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}_i)g(\mathbf{Z}_j)\right]\right| + m^{-1}\left(\frac{p_{\min}\varepsilon}{32}\right)\sum_{i=\ell+1}^{n}\frac{1}{\hat{p}(\mathbf{X}_i)} > \frac{\varepsilon}{8}\right\},$$

and observe that (32) in conjunction with (31) and (28) imply that

$$\begin{aligned} I_n &\leqslant E[I_{\{\Omega_n\}}P\{\mathbb{E}_{n,\varepsilon}|\mathscr{D}_n\}] + P\{\overline{\Omega}_n\} \\ &\leqslant E\left[P\left\{\sup_{g\in\bigcup_{i=\ell+1}^{n}\mathscr{G}_{\ell,\varepsilon'}^{(i)}} m^{-1}\left|\sum_{i=\ell+1}^{n}\left[R_i(1-\delta_i)\sum_{j=1}^{\ell}\varphi_{\ell,j}(\mathbf{X}_i)g(\mathbf{Z}_i)\right]\right| + \frac{\varepsilon}{16} > \frac{\varepsilon}{8}\Big|\,\mathscr{D}_n\right\}\right] \\ &\quad + \sum_{i=\ell+1}^{n} P\left\{\hat{p}(\mathbf{X}_i) < \frac{1}{2}p_{\min}\right\} \\ &:= I_n^{(1)} + I_n^{(2)}, \end{aligned} \tag{33}$$

where $\overline{A}$ denotes the complement of an event A. Using standard arguments, one can show that

$$\begin{aligned} I_n^{(1)} &\leqslant 2\sum_{i=\ell+1}^{n} E[\mathscr{N}(\varepsilon', \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X}_i)})]\mathrm{e}^{-m\varepsilon^2/(512B^2)} \\ &= 2mE[\mathscr{N}(\varepsilon', \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X})})]\mathrm{e}^{-m\varepsilon^2/(512B^2)}. \end{aligned} \tag{34}$$

To deal with the term $I_n^{(2)}$, first let $\hat{f}_\ell(\mathbf{X}_i) = (\ell h_\ell^d)^{-1}\sum_{j=1}^{\ell}\mathscr{K}((\mathbf{X}_j - \mathbf{X}_i)/h_\ell)$ be the usual kernel density estimator of f, based on $\mathbf{X}_1, \ldots, \mathbf{X}_\ell$, at the point $\mathbf{X}_i$, $i = \ell+1, \ldots, n$. It is a simple exercise to show that $P\{\hat{f}_\ell(\mathbf{X}_i) < 1/2 f_{\min}\} = P\{f(\mathbf{X}_i) - \hat{f}_\ell(\mathbf{X}_i) > f(\mathbf{X}_i) - 1/2 f_{\min}\} \leqslant P\{|\hat{f}_\ell(\mathbf{X}_i) - f(\mathbf{X}_i)| > 1/2 f_{\min}\} \leqslant 2\exp\{-\ell h_\ell^d f_{\min}^2/[32\|\mathscr{K}\|_\infty(\|f\|_\infty + f_{\min}/4)]\}$. Consequently, the fact that $P\{\hat{p}(\mathbf{X}_i) < 1/2 p_{\min}\} \leqslant P\{|\hat{p}(\mathbf{X}_i) - p(\mathbf{X}_i)| > 1/2 p_{\min}\}$, together with (43)–(45), and (52) imply that $P\{\hat{p}(\mathbf{X}_i) < 1/2 p_{\min}\} \leqslant 4\exp\{-\ell h_\ell^d f_{\min}^2 p_{\min}^2/[128\|\mathscr{K}\|_\infty(\|f\|_\infty + p_{\min} f_{\min}/8)]\}$. This latter bound in conjunction with (33) and (34) lead to

$$I_n \leqslant 2mE[\mathscr{N}(\varepsilon', \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X})})]\mathrm{e}^{-m\varepsilon^2/512B^2} + 4m\mathrm{e}^{-\ell h_\ell^d(b_3\wedge b_4)}, \tag{35}$$

where b_3 and b_4 are given by (10) and (11). Now Theorem 2.2 follows from (15), (25), (22), (26), (28), and (35). □

3. Applications to nonparametric least-squares regression and density estimation

3.1. Least-squares regression in the presence of missing covariates

In this section we consider the problem of nonparametric estimation of a regression function when some of the covariates may be missing at random. More specifically, consider the random pair $(\mathbf{Z}, Y)$, where $\mathbf{Z} = (\mathbf{V}^{\mathrm{T}}, \mathbf{W}^{\mathrm{T}})^{\mathrm{T}} \in R^{d+p}$, $\mathbf{V} \in R^d$, and $Y \in R$. Let $\psi^*(\mathbf{z}) = E(Y|\mathbf{Z} = \mathbf{z})$ be the least-squares solution of the regression of Y on $\mathbf{Z}$ in the sense that $E|\psi^*(\mathbf{Z}) - Y|^2 = \inf_{\psi: R^{d+p}\to R} E|\psi(\mathbf{Z}) - Y|^2$. Let $\mathscr{D}_n = \{(\mathbf{Z}_1, Y_1, \delta_1), \ldots, (\mathbf{Z}_n, Y_n, \delta_n)\}$ be an iid sample, where $\delta_i = 0$ or 1, according to whether $\mathbf{W}_i$ is missing or not. Let Ψ be any class of candidate regression functions, (this could be, for example, a particular nonlinear class, a linear class, or a partially linear class). For any $\psi \in \Psi$, put $L(\psi) = E|\psi(\mathbf{Z}) - Y|^2$ and define its kernel-based estimator by

$$\widehat{L}_n(\psi) = \frac{1}{n}\left[\sum_{i=1}^{n}\delta_i|\psi(\mathbf{Z}_i) - Y_i|^2 + \frac{n}{m}\sum_{i:(\mathbf{Z}_i,\delta_i)\in\mathscr{D}_m}(1-\delta_i)\tilde{L}_{\ell,i}(\psi)\right], \tag{36}$$

where $\mathscr{D}_m$ is any random subset of $\mathscr{D}_n$, of size m, and

$$\tilde{L}_{\ell,i}(\psi) = \widehat{E}[\psi^2(\mathbf{Z}_i)|\mathbf{X}_i] - 2Y_i\widehat{E}[\psi(\mathbf{Z}_i)|\mathbf{X}_i] + Y_i^2, \quad \mathbf{X}_i = \begin{pmatrix} Y_i \\ \mathbf{V}_i \end{pmatrix} \in R^{1+d}, \quad \text{and}$$

$$\widehat{E}[\psi^k(\mathbf{Z}_i)|\mathbf{X}_i] = \sum_{j:(\mathbf{Z}_j,\delta_j)\in\mathscr{D}_\ell} \delta_j\psi^k(\mathbf{Z}_j)\mathscr{K}((\mathbf{X}_j - \mathbf{X}_i)/h_\ell) \div \sum_{j:(\mathbf{Z}_j,\delta_j)\in\mathscr{D}_\ell} \delta_j\mathscr{K}((\mathbf{X}_j - \mathbf{X}_i)/h_\ell),$$

where $\mathscr{D}_\ell = \mathscr{D}_n - \mathscr{D}_\ell$. Here $\tilde{L}_{\ell,i}(\psi)$ may be viewed as the imputed value of $L_i(\psi) := E(|\psi(\mathbf{Z}_i) - Y_i|^2|\mathbf{X}_i) = E[\psi^2(\mathbf{Z}_i)|\mathbf{X}_i] - 2Y_iE[\psi(\mathbf{Z}_i)|\mathbf{X}_i] + Y_i^2$. Repeating the entire procedure N (>1) times yields N copies of $\widehat{L}_n(\psi)$: $\widehat{L}_n^{(1)}(\psi), \ldots, \widehat{L}_n^{(N)}(\psi)$, where $\widehat{L}_n^{(r)}(\psi)$ is computed based on the rth sample split $(\mathscr{D}_\ell^{(r)}, \mathscr{D}_m^{(r)})$, $r = 1, \ldots, N$. Also, let $\widehat{L}_{\mathrm{MI}}(\psi) = N^{-1}\sum_{r=1}^N \widehat{L}_n^{(r)}(\psi)$. The proposed multiple imputation least-squares estimator of the regression function is given by

$$\psi_n = \operatorname*{argmin}_{\psi\in\Psi} \widehat{L}_{\mathrm{MI}}(\psi).$$

Note that in the hypothetical situation where $L_i(\psi)$ is available for all i's for which $\mathbf{W}_i$ is missing, one would choose ψ_n as the minimizer of the empirical error function $n^{-1}[\sum_{i=1}^n \delta_i|\psi(\mathbf{Z}_i) - Y_i|^2 + \sum_{i=1}^n(1-\delta_i)L_i(\psi)]$; cf. (1) and Theorem 1.1. To study the properties of the L_2-error of ψ_n, we first state the following fundamental lemma.

Lemma 3.1. *Let ψ^* and ψ_n be as before. Then*

$$E[|\psi_n(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2|\mathscr{D}_n] \leqslant 2\sup_{\psi\in\Psi}|\widehat{L}_{\mathrm{MI}}(\psi) - E|\psi(\mathbf{Z}) - Y|^2| + \inf_{\psi\in\Psi} E|\psi(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2. \tag{37}$$

The following result gives exponential performance bounds on the difference between the L_2-error of ψ_n (as an estimator of ψ^*) and that of the best member of Ψ, i.e., the difference

$$E[|\psi_n(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2|\mathscr{D}_n] - \inf_{\psi\in\Psi} E|\psi(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2.$$

Note that if Ψ is large enough so that $\psi^* \in \Psi$, then the above infimum is zero.

Theorem 3.1. *Suppose that $|Y| \leqslant B < \infty$. Let Ψ be a class of functions $\psi : R^{d+p} \to [-C, C]$, $C \geqslant B > 0$. Suppose that conditions $(\mathscr{F})$, (p), $(\mathscr{K})$, $(\mathscr{V})$, and $(\mathscr{ML})$ are satisfied. Then for every $\varepsilon > 0$, there is a n_0 such that, for all $n > n_0$,*

$$P\left\{\left|E[|\psi_n(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2|\mathscr{D}_n] - \inf_{\psi\in\Psi} E|\psi(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2\right| > \varepsilon\right\}$$

$$\leqslant c_{14}NE[\mathscr{N}(c_{15}\varepsilon, \Psi, \|\|_{1,n})]\mathrm{e}^{-c_{16}n\varepsilon^2} + NmE^{1/2}[\mathscr{N}(c_{21}\varepsilon, \Psi, \|\|_{\mathscr{W}_\ell(\mathbf{X})})]^2[c_{17}\mathrm{e}^{-c_{18}\ell h_\ell^d\varepsilon^2} + c_{19}\mathrm{e}^{-c_{20}\ell h_\ell^d}],$$

where $c_{14}, \ldots, c_{21}$ are positive constants not depending on n.

An immediate consequence of the above theorem is the strong consistency of the L_2 error of ψ_n. More specifically, if, as n (and thus ℓ and m) $\to \infty$,

$$\frac{\log(m \vee N)}{\ell h_\ell^d} \to 0 \quad \text{and} \quad \frac{\log \Delta_\ell(\Psi, \varepsilon)}{\ell h_\ell^d} \to 0,$$

then an application of the Borel–Cantelli lemma yields $E[|\psi_n(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2|\mathscr{D}_n] \overset{\text{a.s.}}{\to} \inf_{\psi\in\Psi} E|\psi(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2$. In fact, in many cases, more is true: suppose that the δ-entropy, $\log \mathscr{N}_\infty(\delta, \Psi)$, of the class Ψ satisfies condition (3) for some $\alpha \geqslant 0$. If, as $n \to \infty$, ($N \equiv N(n) \to \infty$),

$$\frac{\log(m \vee N)}{\ell h_\ell^d} \longrightarrow 0, \quad \left(\frac{\ell h_\ell^d}{\log(m \vee N)}\right)^{1/(2+\alpha)} h_\ell \longrightarrow 0, \quad \text{and} \quad \frac{\ell}{nh_\ell} \longrightarrow 0,$$

then, under the conditions of Theorem 3.1, one can show that

$$\lim_{n\to\infty}\left(\frac{\ell h_\ell^d}{\log(m\vee N)}\right)^{1/(2+\alpha)}\left|E[|\psi_n(\mathbf{Z})-\psi^*(\mathbf{Z})|^2|\mathscr{D}_n]-\inf_{\psi\in\Psi}E|\psi(\mathbf{Z})-\psi^*(\mathbf{Z})|^2\right|\stackrel{\text{a.s.}}{=}0.$$

Example 1 (*Differentiable functions*). For $i=1,\ldots,s$, let $\kappa_i\geqslant 0$ be nonnegative integers and put $\kappa=\kappa_1+\cdots+\kappa_s$. Also, for any $g: R^s\to R$, define

$$D^{(\kappa)}g(\mathbf{u})=\frac{\partial^\kappa}{\partial u_1^{\kappa_1},\ldots,\partial u_s^{\kappa_s}}g(\mathbf{u}).$$

Consider the class of functions with bounded partial derivatives of order r:

$$\Psi=\left\{\psi:[0,1]^{d+p}\to R^1\left|\sum_{\kappa\leqslant r}\sup_{\mathbf{u}}|D^{(\kappa)}g(\mathbf{u})|\leqslant A<\infty\right.\right\}.$$

Then, for every $\varepsilon>0$, $\log\mathscr{N}_\infty(\varepsilon,\Psi)\leqslant M\varepsilon^{-\alpha}$, where $\alpha=(d+p)/r$ and $M\equiv M(p,d,r)$; this is due to Rubin (1987).

Example 2. Consider the class Ψ of all convex functions $\psi:\mathscr{C}\to[0,1]$, where $\mathscr{C}\subset R^{d+p}$ is compact and convex. If ψ satisfies $|\psi(\mathbf{z}_1)-\psi(\mathbf{z}_2)|\leqslant L|\mathbf{z}_1-\mathbf{z}_2|$, for all $\mathbf{z}_1,\mathbf{z}_2\in\mathscr{C}$, then $\log\mathscr{N}_\infty(\varepsilon,\Psi)\leqslant M\varepsilon^{-(d+p)/2}$, for every $\varepsilon>0$, where $M\equiv M(p,d,L)$; see van der Vaart and Wellner (1996).

3.2. Maximum likelihood density estimation

Let $\mathbf{Z}=(\mathbf{X}^{\mathrm{T}},\mathbf{Y}^{\mathrm{T}})^{\mathrm{T}}\in R^{d+p}$, where $\mathbf{X}\in R^d$. Here $\mathbf{Y}$ could be missing (MAR) but $\mathbf{X}$ is always observable. We consider the problem of estimating the marginal probability density of $\mathbf{Y}$, in the presence of missing data: $(\mathbf{X}_1,\mathbf{Y}_1,\delta_1),\ldots,(\mathbf{X}_n,\mathbf{Y}_n,\delta_n)$, where as usual $\delta_i=0$ if $\mathbf{Y}_i$ is missing, (otherwise, $\delta_i=1$). In what follows, it is assumed that the true density p_0 belongs to a class of densities, say $\mathscr{P}$. Note that when there are no missing data, the usual maximum likelihood estimator of p_0 is simply $\hat{p}_n=\operatorname{argmax}_{p\in\mathscr{P}}\sum_{i=1}^n\log p(\mathbf{Y}_i)$. Next, let $\mathscr{D}_\ell$ be a random subset of $\mathscr{D}_n$, of size ℓ, and put $\mathscr{D}_m=\mathscr{D}_n-\mathscr{D}_\ell$. For any function $p:R^p\to(0,\infty)$ define

$$\widehat{L}_n(p)=\frac{1}{n}\left[\sum_{i=1}^n\delta_i\log p(\mathbf{Y}_i)+\frac{n}{m}\sum_{i:(\mathbf{Z}_i,\delta_i)\in\mathscr{D}_m}(1-\delta_i)\tilde{L}_{\ell,i}(p)\right],\tag{38}$$

where

$$\tilde{L}_{\ell,i}(p)=\sum_{j:(\mathbf{Z}_j,\delta_j)\in\mathscr{D}_\ell}\delta_j\log p(\mathbf{Y}_j)\mathscr{K}((\mathbf{X}_j-\mathbf{X}_i)/h_\ell)\div\sum_{j:(\mathbf{Z}_j,\delta_j)\in\mathscr{D}_\ell}\delta_j\mathscr{K}((\mathbf{X}_j-\mathbf{X}_i)/h_\ell).$$

Also, let $\widehat{L}_n^{(1)}(p),\ldots,\widehat{L}_n^{(\mathrm{N})}(p)$ be copies of (38), based on N independent sample splits $(\mathscr{D}_\ell^{(r)},\mathscr{D}_m^{(r)})$, $r=1,\ldots,N$. Put $\widehat{L}_{\mathrm{MI}}(p)=N^{-1}\sum_{r=1}^N\widehat{L}_n^{(r)}(p)$, and consider the MLE-type density estimator

$$\hat{p}_n=\operatorname*{argmax}_{p\in\mathscr{P}}\widehat{L}_{\mathrm{MI}}(p).$$

To study $\hat{p}_n$, let d_{H} be the Hellinger distance between two densities, i.e., for $p_1,p_2\in\mathscr{P}$,

$$d_{\mathrm{H}}(p_1,p_2)=\sqrt{\frac{1}{2}\int|p_1^{1/2}(\mathbf{y})-p_2^{1/2}(\mathbf{y})|^2\,\mathrm{d}\mathbf{y}}.$$

Also, put $\bar{p}_n = (\hat{p}_n + p_0)/2$. Then by Lemma 4.2 of van de Geer (2000),

$$d_{\mathrm{H}}^2(\hat{p}_n, p_0) \leqslant 16 d_{\mathrm{H}}^2(\bar{p}_n, p_0). \tag{39}$$

The following lemma may be viewed as a counterpart of Lemma 4.1 of van de Geer (2000), tailored to fit our missing data setup. Let $\widehat{L}_{\mathrm{MI}}$ be as before and define

$$L_n\left(\frac{\hat{p}_n + p_0}{2p_0} I\{p_0 > 0\}\right) = E\left[\log \frac{\hat{p}_n(\mathbf{Y}) + p_0(\mathbf{Y})}{2p_0(\mathbf{Y})} I\{p_0(\mathbf{Y}) > 0\} \,\middle|\, \mathscr{D}_n\right]. \tag{40}$$

$$L\left(\frac{p + p_0}{2p_0} I\{p_0 > 0\}\right) = E\left[\log \frac{p(\mathbf{Y}) + p_0(\mathbf{Y})}{2p_0(\mathbf{Y})} I\{p_0(\mathbf{Y}) > 0\}\right]. \tag{41}$$

Lemma 3.2. *Let* $\bar{p}_n = (\hat{p}_n + p_0)/2$. *Then*

$$d_{\mathrm{H}}^2(\bar{p}_n, p_0) \leqslant \frac{1}{2}\left[\widehat{L}_{\mathrm{MI}}\left(\frac{\bar{p}_n}{p_0} I\{p_0 > 0\}\right) - L_n\left(\frac{\bar{p}_n}{p_0} I\{p_0 > 0\}\right)\right].$$

Putting together (39) and Lemma 3.2, one obtains

$$d_{\mathrm{H}}^2(\hat{p}_n, p_0) \leqslant 8 \sup_{p \in \mathscr{P}} \left|\widehat{L}_{\mathrm{MI}}\left(\frac{p + p_0}{2p_0} I\{p_0 > 0\}\right) - L\left(\frac{p + p_0}{2p_0} I\{p_0 > 0\}\right)\right|,$$

where L is given by (41). Now let $\mathscr{G}$ be the class

$$\mathscr{G} = \left\{\frac{1}{2}\log \frac{p + p_0}{2p_0} I\{p_0 > 0\} \,\middle|\, p \in \mathscr{P}\right\}.$$

The above results (in conjunction with Theorem 2.2) can be summarized in the following theorem.

Theorem 3.2. *Define the class* $\mathscr{G}$ *as above. Then, under conditions* $(\mathscr{F})$, (p), $(\mathscr{K})$, $(\mathscr{V})$, *and* $(\mathscr{ML})$, *for every* $\varepsilon > 0$ *there is a* n_0 *such that for all* $n > n_0$,

$$\begin{aligned} P\{d_{\mathrm{H}}(\hat{p}_n, p_0) > \varepsilon\} \leqslant\ & c_{22} N E[\mathscr{N}(c_{23}\varepsilon^2, \mathscr{G}, \|\|_{1,n})] \mathrm{e}^{-c_{24} n \varepsilon^4} \\ & + N m E^{1/2}[\mathscr{N}(c_{25}\varepsilon^2, \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X})})]^2 [c_{26}\mathrm{e}^{-c_{27}\ell h_\ell^d \varepsilon^4} + c_{28}\mathrm{e}^{-c_{29}\ell h_\ell^d}], \end{aligned}$$

where $c_{22}, \ldots, c_{29}$ *are positive constants not depending on n.*

The above theorem can be used to study strong convergence results for the density estimate $\hat{p}_n$. Suppose that the δ-entropy, $\log \mathscr{N}_\infty(\delta, \mathscr{G})$, of the class $\mathscr{G}$ satisfies condition (3) for some $\alpha \geqslant 0$. If, as $n \to \infty$,

$$\frac{\log(m \vee N)}{\ell h_\ell^d} \longrightarrow 0, \quad \left(\frac{\ell h_\ell^d}{\log(m \vee N)}\right)^{1/(4+2\alpha)} h_\ell \longrightarrow 0, \quad \text{and} \quad \frac{\ell}{n h_\ell} \longrightarrow 0,$$

then, under the conditions of Theorem 3.2, one has

$$\lim_{n\to\infty} \left(\frac{\ell h_\ell^d}{\log(m \vee N)}\right)^{1/(4+2\alpha)} d_{\mathrm{H}}(\hat{p}_n, p_0) \overset{\text{a.s.}}{=} 0.$$

3.3. Kernel density estimation

As an alternative to maximum likelihood estimation, which assumes the knowledge of the underlying class of densities $\mathscr{P}$, one can also consider the popular kernel density estimate $\hat{p}_0$ of p_0 defined by (in the presence of missing data)

$$\hat{p}_0(\mathbf{y}) = \frac{1}{n}\left\{\sum_{i=1}^{n}\delta_i\frac{1}{a_n^p}\mathscr{H}\left(\frac{\mathbf{Y}_i-\mathbf{y}}{a_n}\right)+\sum_{i=1}^{n}(1-\delta_i)\left[\sum_{j=1}^{n}\varphi_{n,j}(\mathbf{X}_i)\frac{1}{a_n^p}\mathscr{H}\left(\frac{\mathbf{Y}_j-\mathbf{y}}{a_n}\right)\right]\right\}$$
$$= \frac{1}{na_n^p}\sum_{i=1}^{n}\left[\delta_i\mathscr{H}\left(\frac{\mathbf{Y}_i-\mathbf{y}}{a_n}\right)+(1-\delta_i)\frac{\sum_{j=1}^{n}\delta_j\mathscr{K}((\mathbf{X}_j-\mathbf{X}_i)/h_n)\mathscr{H}((\mathbf{Y}_j-\mathbf{y})/a_n)}{\sum_{j=1}^{n}\delta_j\mathscr{K}((\mathbf{X}_j-\mathbf{X}_i)/h_n)}\right], \tag{42}$$

where $\mathscr{H}: R^p \to R$ is the kernel with smoothing parameter a_n. When $p = d = 1$, (i.e., when (X, Y) is bivariate), and upon taking $\mathscr{K} = \mathscr{H}$ with possibly different smoothing parameters h_n and a_n, the estimator in (42) coincides with Hazelton's (2000) estimator. For this special case, it is shown in the cited paper that if $nh_na_n/\log n \to \infty$ then $\sup_{y\in R}|\hat{p}_0(y) - p_0(y)| \xrightarrow{\text{a.s.}} 0$ holds under classical assumptions (i.e., Lipschitz continuity and symmetricity of the common kernel, with a compact support, and the uniform continuity of the density $p_0(\mathbf{y})$). In fact, it is rather straightforward to show that if $nh_n^da_n^p/\log n \to \infty$ then $\sup_{\mathbf{y}\in R^p}|\hat{p}_0(\mathbf{y}) - p_0(\mathbf{y})| \xrightarrow{\text{a.s.}} 0$ for the estimator defined in (42), under similar assumptions on the analytic properties of $\mathscr{H}$ and p_0. The estimator (42) is essentially a single imputation estimator. One can, alternatively, consider a multiple imputation estimator.

4. Proofs of auxiliary results

Before proving various auxiliary results of this paper we state a number of technical lemmas:

Lemma 4.1. *Let $\mathscr{G}$ be a totally bounded class of functions $g: R^{d+p} \to R$, with $\|g\|_\infty < B$ for every g in $\mathscr{G}$. Also, let $\overline{\mathscr{V}}_n(g)$ be as in* (2). *Then, under conditions* $(\mathscr{F})$, (p), $(\mathscr{K})$, *and* $(\mathscr{V})$, *for every $\varepsilon > 0$ and $n \geqslant 2$*

$$P\left\{\sup_{g\in\mathscr{G}}|\overline{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon\right\} \leqslant 2n\mathscr{N}_\infty\left(\frac{\varepsilon}{3},\mathscr{G}\right)[2e^{-n\varepsilon^2/288B^2} + e^{-d_1(\varepsilon-d_2h_n)^2(n-1)h_n^d}$$
$$+ e^{-d_3(\varepsilon-d_4h_n)^2(n-1)h_n^d} + e^{-d_5(\varepsilon-d_6h_n)^2(n-1)h_n^d}],$$

where the positive constants $d_1, \ldots, d_6$ do not depend on n or ε and are given by

$$d_1 = d_3 = \left(\frac{p_{\min}f_{\min}}{27B}\right)^2 \Big/ (2\|\mathscr{K}\|_\infty(\|f\|_\infty + 2p_{\min}f_{\min}/3)),$$

$$d_5 = \left(\frac{p_{\min}f_{\min}}{54B}\right)^2 \Big/ (2\|\mathscr{K}\|_\infty(\|f\|_\infty + p_{\min}f_{\min}/3)),$$

$$d_2 = \frac{27}{p_{\min}f_{\min}}\left(\bigvee_{i=1}^{d}\sup_{g\in\mathscr{G}}\sup_{\mathbf{x}}\left|\frac{\partial\mathscr{V}(g|\mathbf{x})}{\partial x_i}\right|\|f\|_\infty + Bd\|f'\|_\infty\right)\sum_{j=1}^{d}\int|y_j|\mathscr{K}(\mathbf{y})\,d\mathbf{y},$$

$$d_4 = Bd_2,$$

$$d_6 = \frac{54B}{p_{\min}f_{\min}}\left(d\|f'\|_\infty\sum_{j=1}^{d}\int|y_j|\mathscr{K}(\mathbf{y})\,d\mathbf{y}\right).$$

Here, $\mathscr{V}(g|\mathbf{x}) = E[\delta g(\mathbf{Z})|\mathbf{X} = \mathbf{x}]$.

In order to state the next two lemmas, define

$$\phi(g|\mathbf{X}) = f(\mathbf{X})E[\delta g(\mathbf{Z})|\mathbf{X}], \tag{43}$$

$$\hat{\phi}(g|\mathbf{X}) = \frac{1}{\ell h_\ell^d}\sum_{j=1}^{\ell}\delta_j g(\mathbf{Z}_j)\mathscr{K}\left(\frac{\mathbf{X}_j - \mathbf{X}}{h_\ell}\right). \tag{44}$$

Lemma 4.2. *Suppose that conditions* $(\mathscr{F})$, $(\mathscr{K})$, *and* $(\mathscr{V})$ *hold. Then*

$$\sup_{g\in\mathscr{G}} |E[\hat{\phi}(g|\mathbf{X})|\mathbf{X}] - \phi(g|\mathbf{X})| \leqslant |\text{cons.}|h_\ell.$$

Lemma 4.3. *Suppose that* $\|f\|_\infty < \infty$ *and* $\|\mathscr{K}\|_\infty < \infty$. *If* $h_\ell \to 0$ *and* $\ell h_\ell^d \to \infty$, *as* $\ell \to \infty$, *then for every* $\gamma > 0$ *there is a* n_0 *such that for all* $n > n_0$

$$P\left\{\sup_{g\in\mathscr{G}} |\hat{\phi}(g|\mathbf{X}) - E[\hat{\phi}(g|\mathbf{X})|\mathbf{X}]| \geqslant \frac{\gamma}{2}\Big|\,\mathbf{X}\right\}$$
$$\leqslant 4\sqrt{E[\mathscr{N}(\gamma/(32\|f\|_\infty), \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X})})]^2}[\mathrm{e}^{-\gamma^2\ell h_\ell^d/512B^2b_{10}} + 2\mathrm{e}^{-\ell h_\ell^d/4b_{10}}] + 4\mathrm{e}^{-\ell h_\ell^d b_{11}}, \tag{45}$$

where $b_{10} = \|\mathscr{K}\|_\infty(\|f\|_\infty + 1)$ *and* $b_{11} = \|f\|_\infty/(12\|\mathscr{K}\|_\infty)$.

Proof of Theorem 1.2. This is an immediate consequence of Lemma 4.1. □

Proof of Theorem 1.3. Put $\delta_n = (n^{-1}h_n^{-d}\log n)^{1/(2+\alpha)}$. Then, for any $\varepsilon > 0$, Lemma 4.1 in conjunction with the entropy bound (3) lead to

$$P\left\{\sup_{g\in\mathscr{G}} \delta_n^{-1}|\overline{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon\right\}$$
$$\leqslant 2\mathrm{e}^{\log n + (\varepsilon\delta_n/3)^{-\alpha}} M\left[2\mathrm{e}^{-n\varepsilon^2\delta_n^2/288B^2} + \sum_{i=1}^{3}\mathrm{e}^{-d_{2i-1}(\varepsilon\delta_n - d_{2i}h_n)^2(n-1)h_n^d}\right]$$
$$= 4\exp\left\{-n\delta_n^2\left[\frac{\varepsilon^2}{288B^2} - \frac{(\varepsilon/3)^{-\alpha}M}{n\delta_n^{2+\alpha}} - \frac{\log n}{n\delta_n^2}\right]\right\}$$
$$+ 2\sum_{i=1}^{3}\exp\{-d_{2i-1}(\varepsilon\delta_n - d_{2i}h_n)^2(n-1)h_n^d + (\varepsilon/3)^{-\alpha}M\delta_n^{-\alpha} + \log n\}$$
$$:= 4I_{n,0}(\varepsilon) + 2\sum_{i=1}^{3} I_{n,i}(\varepsilon) \quad \text{(say)}.$$

Given conditions (4), it is straightforward to show that $\sum_{n=1}^{\infty} I_{n,0}(\varepsilon) < \infty$. Furthermore, for $i = 1, 2, 3$,

$$I_{n,i}(\varepsilon) \leqslant \exp\left\{-\delta_n^2 n h_n^d\left[\frac{d_{2i-1}}{2}\left(\varepsilon - \frac{d_{2i}h_n}{\delta_n}\right)^2 - \frac{(\varepsilon/3)^{-\alpha}M}{\delta_n^{2+n}nh_n^d} - \frac{\log n}{\delta_n^2 n h_n^d}\right]\right\}.$$

Therefore, $\sum_{n=1}^{\infty} I_{n,i}(\varepsilon) < \infty$, (since $h_n/\delta_n \to 0$), which completes the proof of Theorem 1.3, via the Borel–Cantelli lemma. □

Proof of (22). Put

$$\pi_{n,i} = P\left\{\sup_{g\in\mathcal{G}} |I_{n,i}(g)| \geqslant \frac{n\varepsilon}{18m}\right\}, \quad \text{and note that } \pi_n \leqslant \sum_{i=\ell+1}^{n} \pi_{n,i}, \tag{46}$$

where $I_{n,i}(g)$ is as in (19). Next, define

$$\mathscr{V}(g|\mathbf{x}) = E(\delta' g(\mathbf{Z}')|\mathbf{X}' = \mathbf{x}), \quad \widehat{\mathscr{V}}(g|\mathbf{x}) = \frac{\sum_{j=1}^{\ell}\delta_j g(\mathbf{Z}_j)\mathscr{K}((\mathbf{X}_j - \mathbf{x})/h_\ell)}{\sum_{j=1}^{\ell}\mathscr{K}((\mathbf{X}_j - \mathbf{x})/h_\ell)}, \quad \text{and}$$

$$\hat{p}(\mathbf{x}) = \frac{\sum_{j=1}^{\ell}\delta_j \mathscr{K}((\mathbf{X}_j - \mathbf{x})/h_\ell)}{\sum_{j=1}^{\ell}\mathscr{K}((\mathbf{X}_j - \mathbf{x})/h_\ell)},$$

and note that for $i = \ell+1, \ldots, n$,

$$\pi_{n,i} \leqslant P\left\{\sup_{g\in\mathcal{G}} \left|\frac{\widehat{\mathscr{V}}(g|\mathbf{X}_i')}{\hat{p}(\mathbf{X}_i')} - \frac{\mathscr{V}(g|\mathbf{X}_i')}{p(\mathbf{X}_i')}\right| \geqslant \frac{\varepsilon}{18}\right\}. \tag{47}$$

At the same time, since $|\widehat{\mathscr{V}}(g|\mathbf{X})/\hat{p}(\mathbf{X})| \leqslant \|g\|_\infty \leqslant B$, elementary algebra yields

$$\begin{aligned}\left|\frac{\hat{\mathscr{V}}(g|\mathbf{X})}{\hat{p}(\mathbf{X})} - \frac{\mathscr{V}(g|\mathbf{X})}{p(\mathbf{X})}\right| &= \left|\frac{-\widehat{\mathscr{V}}(g|\mathbf{X})/\hat{p}(\mathbf{X})}{p(\mathbf{X})}(\hat{p}(\mathbf{X}) - p(\mathbf{X})) + \frac{\widehat{\mathscr{V}}(g|\mathbf{X}) - \mathscr{V}(g|\mathbf{X})}{p(\mathbf{X})}\right| \\ &\leqslant \frac{|\hat{p}(\mathbf{X}) - p(\mathbf{X})|}{p_{\min}/B} + \frac{|\widehat{\mathscr{V}}(g|\mathbf{X}) - \mathscr{V}(g|\mathbf{X})|}{p_{\min}}.\end{aligned} \tag{48}$$

Also, let $\hat{f}_\ell$ be the usual kernel density estimator (of the distribution of $\mathbf{X}$), based on $\mathbf{X}_1, \ldots, \mathbf{X}_\ell$, i.e., $\hat{f}_\ell(\mathbf{x}) = (\ell h_\ell^d)^{-1}\sum_{i=1}^{\ell}\mathscr{K}((\mathbf{X}_i - \mathbf{x})/h_\ell)$. Then, it is straightforward to see that the second term on the r.h.s. of (48) can be bounded by

$$\begin{aligned}\left|\frac{\widehat{\mathscr{V}}(g|\mathbf{X}) - \mathscr{V}(g|\mathbf{X})}{p_{\min}}\right| &= \frac{1}{p_{\min}}\left|\frac{\hat{\phi}(g|\mathbf{X})}{\hat{f}_\ell(\mathbf{X})} - \frac{\phi(g|\mathbf{X})}{f(\mathbf{X})}\right| \\ &\leqslant \frac{1}{p_{\min} f_{\min}}[B|\hat{f}_\ell(\mathbf{X}) - f(\mathbf{X})| + |\hat{\phi}(g|\mathbf{X}) - \phi(g|\mathbf{X})|],\end{aligned} \tag{49}$$

where $\hat{\phi}(g|\mathbf{X})$ and $\phi(g|\mathbf{X})$ are given by (44) and (43). Furthermore, for every $\gamma > 0$,

$$\begin{aligned}&P\left\{\sup_{g\in\mathcal{G}} |\hat{\phi}(g|\mathbf{X}) - \phi(g|\mathbf{X})| \geqslant \gamma\right\} \\ &\leqslant E\left[P\left\{\sup_{g\in\mathcal{G}} |\hat{\phi}(g|\mathbf{X}) - E[\hat{\phi}(g|\mathbf{X})|\mathbf{X}]| + \sup_{g\in\mathcal{G}} |E[\hat{\phi}(g|\mathbf{X})|\mathbf{X}] - \phi(g|\mathbf{X})| \geqslant \gamma|\mathbf{X}\right\}\right] \\ &\leqslant E\,P\left\{\sup_{g\in\mathcal{G}} |\hat{\phi}(g|\mathbf{X}) - E[\hat{\phi}(g|\mathbf{X})|\mathbf{X}]| \geqslant \frac{\gamma}{2}\middle|\mathbf{X}\right\} \quad \text{for large } n \text{ (by Lemma 4.2)} \\ &\leqslant 4\sqrt{E[\mathscr{N}(\gamma/(32\|f\|_\infty), \mathcal{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X})})]^2}[\mathrm{e}^{-\gamma^2\ell h_\ell^d/512B^2 b_{10}} + 2\mathrm{e}^{-\ell h_\ell^d/4b_{10}}] + 4\mathrm{e}^{-\ell h_\ell^d b_{11}} \\ &\quad \text{(by Lemma 4.3).}\end{aligned} \tag{50}$$

Next, the term $p_{\min}^{-1}B|\hat{p}(\mathbf{X})-p(\mathbf{X})|$ that appears on the right-hand side of (48) can be handled as follows. Let $\phi(1|\mathbf{X})$ and $\hat{\phi}(1|\mathbf{X})$ be as in (43) and (44) with $g\equiv 1$, and note that

$$|\hat{p}(\mathbf{X})-p(\mathbf{X})| = \left|\frac{\hat{\phi}(1|\mathbf{X})}{\hat{f}(\mathbf{X})}-\frac{\phi(1|\mathbf{X})}{f(\mathbf{X})}\right| \leqslant \left|\frac{\hat{f}(\mathbf{X})-f(\mathbf{X})}{f(\mathbf{X})}\right| + \left|\frac{\hat{\phi}(1|\mathbf{X})-\phi(1|\mathbf{X})}{f(\mathbf{X})}\right|. \tag{51}$$

By straightforward arguments, (and in fact much simpler than those leading to (50)), one also finds $\forall\gamma>0$, (and for n large),

$$P\{|\hat{\phi}(1|\mathbf{X})-\phi(1|\mathbf{X})|\geqslant\gamma\}\leqslant 2\mathrm{e}^{-\ell h_\ell^d\gamma^2/8\|\mathscr{K}\|_\infty(\|f\|_\infty+\gamma/2)}. \tag{52}$$

It is also trivial to show that $\forall\gamma>0$, (and for n large) $P\{|\hat{f}_\ell(\mathbf{X})-f(\mathbf{X})|\geqslant\gamma|\mathbf{X}\}$ is also bounded by the r.h.s. of (52). Putting together (47)–(52), and the fact that $\mathbf{X}_i'$ is independent of $(\mathbf{Z}_j,\delta_j)$'s, one finds for n large enough,

$$\begin{aligned}\pi_{n,i} &\leqslant P\left\{\sup_{g\in\mathscr{G}}|\hat{\phi}(g|\mathbf{X}_i')-\phi(g|\mathbf{X}_i')|\geqslant\frac{\varepsilon p_{\min}f_{\min}}{54}\right\}+P\left\{|\hat{\phi}(1|\mathbf{X}_i')-\phi(1|\mathbf{X}_i')|\geqslant\frac{\varepsilon p_{\min}f_{\min}}{54B}\right\}\\ &\quad+P\left\{|\hat{f}_\ell(\mathbf{X}_i')-f(\mathbf{X}_i')|\geqslant\frac{\varepsilon p_{\min}f_{\min}}{108B}\right\}\\ &\leqslant 2[2\{\mathrm{e}^{-\ell h_\ell^d b_7\varepsilon^2}+2\mathrm{e}^{-\ell h_\ell^d b_1}\}\Delta_\ell(\mathscr{G},\varepsilon)+2\mathrm{e}^{-\ell h_\ell^d b_2}+\mathrm{e}^{-\ell h_\ell^d q_1(\varepsilon)\varepsilon^2}+\mathrm{e}^{-\ell h_\ell^d q_2(\varepsilon)\varepsilon^2}],\end{aligned}$$

where b_1, b_2,, b_7, and $\Delta_\ell(\mathscr{G},\varepsilon)$ are as in (8), (9), (14), and (7), respectively, and

$$q_1(\varepsilon)=\frac{p_{\min}^2f_{\min}^2}{8(54)^2B\|\mathscr{K}\|_\infty(B\|f\|_\infty+(\varepsilon p_{\min}f_{\min})/108)}, \tag{53}$$

$$q_2(\varepsilon)=\frac{p_{\min}^2f_{\min}^2}{8(108)^2B^2\|\mathscr{K}\|_\infty(\|f\|_\infty+(\varepsilon p_{\min}f_{\min})/216B)}. \tag{54}$$

Referring back to the definition of π_{ni} in (46), it is clear that one only needs to consider $\varepsilon\leqslant 36B$, (because $\pi_{ni}=0$, for $\varepsilon>36B$). In other words, one only needs to consider $q_1(\varepsilon)\geqslant b_5$ and $q_2(\varepsilon)\geqslant b_6$, where b_5 and b_6 are given by (12) and (13). Thus one finds the loose upper bound

$$\pi_{n,i}\leqslant 4[3\mathrm{e}^{-\ell h_\ell^d(b_5\wedge b_6\wedge b_7)\varepsilon^2}+3\mathrm{e}^{-\ell h_\ell^d(b_1\wedge b_2)}]\Delta_\ell(\mathscr{G},\varepsilon).$$

Now, (22) follows from the above bound on $\pi_{n,i}$ in conjunction with (46). □

Proof of (23).

$$\begin{aligned}\beta_n(\mathscr{D}_n) &\geqslant 1-P\left\{\left|\sum_{i=1}^n g_\varepsilon(\mathbf{Z}_i')-n\mathscr{V}(g_\varepsilon|\mathscr{D}_n)\right|\geqslant\frac{n\varepsilon}{6}\,\middle|\,\mathscr{D}_n\right\}\\ &\quad-P\left\{\left|\sum_{i=1}^n(1-\delta_i')(g_\varepsilon(\mathbf{Z}_i')-E[g_\varepsilon(\mathbf{Z}_i')|\mathbf{X}_i',\mathscr{D}_n])\right|\geqslant\frac{n\varepsilon}{6}\,\middle|\,\mathscr{D}_n\right\}\\ &\quad-P\left\{\left|\sum_{i=1}^n(1-\delta_i')E[g_\varepsilon(\mathbf{Z}_i')|\mathbf{X}_i',\mathscr{D}_n]-\left(1+\frac{\ell}{m}\right)\right.\right.\\ &\quad\left.\left.\times\sum_{i=\ell+1}^n(1-\delta_i')\left[\sum_{j=1}^\ell\varphi_{\ell,j}(\mathbf{X}_i')g_\varepsilon(\mathbf{Z}_j)\right]\right|\geqslant\frac{n\varepsilon}{6}\,\middle|\,\mathscr{D}_n\right\}\\ &:= 1-\Delta_{n,1}-\Delta_{n,2}-\Delta_{n,3}.\end{aligned} \tag{55}$$

The term $\Delta_{n,1}$ can be bounded using Hoeffding's inequality (recall: $\|g\|_\infty < B, \forall g \in \mathscr{G}$):

$$\Delta_{n,1} \leqslant 2e^{-n\varepsilon^2/(72B^2)}. \tag{56}$$

The term $\Delta_{n,2}$. First observe that, conditional on $\mathscr{D}_n$, $(\mathbf{X}'_1, \delta'_1), \ldots, (\mathbf{X}'_n, \delta'_n)$, and under the MAR assumption, the terms

$$W_i = (1-\delta'_i)(g_\varepsilon(\mathbf{Z}'_i) - E[g_\varepsilon(\mathbf{Z}'_i)|\mathbf{X}'_i, \mathscr{D}_n]), \quad i = 1, \ldots, n$$

are independent, zero-mean random variables (due to MAR assumption), which are bounded by $-2B$ and $+2B$. Invoking Hoeffding's inequality once again,

$$\Delta_{n,2} = E\left[P\left\{ \left| \sum_{i=1}^n W_i \right| \geqslant \frac{n\varepsilon}{6} \,\middle|\, \mathscr{D}_n, (\mathbf{X}'_1, \delta'_1), \ldots, (\mathbf{X}'_n, \delta'_n) \right\} \,\middle|\, \mathscr{D}_n \right] \leqslant 2e^{-n\varepsilon^2/(288B^2)}. \tag{57}$$

The term $\Delta_{n,3}$. Start with the bound

$$\begin{aligned}
\Delta_{n,3} &\leqslant P\left\{ \frac{1}{n} \left| \sum_{i=\ell+1}^n (1-\delta'_i) \left[\sum_{j=1}^\ell \varphi_{\ell,j}(\mathbf{X}'_i) g_\varepsilon(\mathbf{Z}_j) \right] - \sum_{i=\ell+1}^n (1-\delta'_i) E[g_\varepsilon(\mathbf{Z}'_i)|\mathbf{X}'_i, \mathscr{D}_n] \right| \right. \\
&\quad + \left| \frac{\ell}{mn} \sum_{i=\ell+1}^n (1-\delta'_i) \left[\sum_{j=1}^\ell \varphi_{\ell,j}(\mathbf{X}'_i) g_\varepsilon(\mathbf{Z}_j) \right] \right| \\
&\quad \left. + \left| \frac{1}{n} \sum_{i=1}^\ell (1-\delta'_i) E[g_\varepsilon(\mathbf{Z}'_i)|\mathbf{X}'_i, \mathscr{D}_n] \right| \geqslant \frac{\varepsilon}{6} \,\middle|\, \mathscr{D}_n \right\} \\
&\leqslant P\left\{ \frac{1}{n} \left| \sum_{i=\ell+1}^n (1-\delta'_i) \left[\sum_{j=1}^\ell \varphi_{\ell,j}(\mathbf{X}'_i) g_\varepsilon(\mathbf{Z}_j) \right] - \sum_{i=\ell+1}^n (1-\delta'_i) E[g_\varepsilon(\mathbf{Z}'_i)|\mathbf{X}'_i, \mathscr{D}_n] \right| \right. \\
&\quad \left. + \frac{\varepsilon}{18} + \frac{\varepsilon}{18} \geqslant \frac{\varepsilon}{6} \,\middle|\, \mathscr{D}_n \right\} \quad \text{(for large } n\text{)}, \qquad (58) \\
&\leqslant P\left\{ \sum_{i=\ell+1}^n \sup_{g\in\mathscr{G}} \left| \sum_{j=1}^\ell \varphi_{\ell,j}(\mathbf{X}'_i) g(\mathbf{Z}_j) - E[g(\mathbf{Z}'_i)|\mathbf{X}'_i] \right| \geqslant \frac{n\varepsilon}{18} \,\middle|\, \mathscr{D}_n \right\}, \qquad (59)
\end{aligned}$$

where (58) follows from the two trivial bounds $|\sum_{j=1}^\ell \varphi_{\ell,j}(\mathbf{X}'_i) g_\varepsilon(\mathbf{Z}_j)| \leqslant \|g\|_\infty < B$ and $|E[g_\varepsilon(\mathbf{Z}'_i)|\mathbf{X}'_i, \mathscr{D}_n]| \leqslant \|g\|_\infty < B$, and the fact that $\ell/n \to 0$. Furthermore, we may replace $E[g(\mathbf{Z}'_i)|\mathbf{X}'_i]$ in (59) by $(1/p(\mathbf{X}'_i))E[\delta'_i g(\mathbf{Z}'_i)|\mathbf{X}'_i]$, where $p(\mathbf{x}) = P(\delta' = 1|\mathbf{X}' = \mathbf{x})$. This is because $E[\delta'_i g(\mathbf{Z}'_i)|\mathbf{X}'_i] = E[E(\delta'_i g(\mathbf{Z}'_i)|\mathbf{X}'_i, \mathbf{Y}'_i)|\mathbf{X}'_i] \overset{\text{by MAR}}{=} E[g(\mathbf{Z}'_i) p(\mathbf{X}'_i)|\mathbf{X}'_i]$. Thus one obtains

$$\Delta_{n,3} \leqslant P\left\{ \sum_{i=\ell+1}^n \sup_{g\in\mathscr{G}} |I_{n,i}(g)| \geqslant \frac{n\varepsilon}{18} \,\middle|\, \mathscr{D}_n \right\} := \pi_n(\mathscr{D}_n),$$

where (as before)

$$I_{n,i}(g) := \frac{\sum_{j=1}^\ell \delta_j g(\mathbf{Z}_j) \mathscr{K}((\mathbf{X}_j - \mathbf{X}'_i)/h_\ell)}{\sum_{j=1}^\ell \delta_j \mathscr{K}((\mathbf{X}_j - \mathbf{X}'_i)/h_\ell)} - \frac{E[\delta'_i g(\mathbf{Z}'_i)|\mathbf{X}'_i]}{p(\mathbf{X}'_i)}.$$

This last bound on $\Delta_{n,3}$ together (55)–(57) gives

$$\beta_n(\mathscr{D}_n) \geqslant 1 - 4e^{-n\varepsilon^2/(288B^2)} - \pi_n(\mathscr{D}_n),$$

which completes the proof of (23). □

Proof of Theorem 3.1. First note that in view of Lemma 3.1

$$P\left\{\left|E[|\psi_n(\mathbf{Z})-\psi^*(\mathbf{Z})|^2|\mathscr{D}_n]-\inf_{\psi\in\Psi}E|\psi(\mathbf{Z})-\psi^*(\mathbf{Z})|^2\right|>\varepsilon\right\}$$

$$\leqslant P\left\{\sup_{\psi\in\Psi}|\widehat{L}_{\mathrm{MI}}(\psi)-E|\psi(\mathbf{Z})-Y|^2|>\frac{\varepsilon}{2}\right\}$$

$$\leqslant NP\left\{\sup_{\psi\in\Psi}|\widehat{L}_n(\psi)-E|\psi(\mathbf{Z})-Y|^2|>\frac{\varepsilon}{2}\right\}\quad\text{where } \widehat{L}_n \text{ is as in (36).}$$

But

$$P\left\{\sup_{\psi\in\Psi}|\widehat{L}_n(\psi)-E|\psi(\mathbf{Z})-Y|^2|>\frac{\varepsilon}{2}\right\}$$

$$\leqslant P\left\{\sup_{\psi\in\Psi}\left|\frac{1}{n}\left[\sum_{i=1}^{n}\delta_i\psi^2(\mathbf{Z}_i)+\frac{n}{m}\sum_{i=\ell+1}^{n}(1-\delta_i)\widehat{E}(\psi^2(\mathbf{Z}_i)|\mathbf{X}_i)\right]-E(\psi^2(\mathbf{Z}))\right|>\frac{\varepsilon}{6}\right\}$$

$$+P\left\{\left|\frac{1}{n}\left[\sum_{i=1}^{n}\delta_iY_i^2+\frac{n}{m}\sum_{i=\ell+1}^{n}(1-\delta_i)Y_i^2\right]-E(Y^2)\right|>\frac{\varepsilon}{6}\right\}$$

$$+P\left\{2\sup_{\psi\in\Psi}\left|\frac{1}{n}\left[\sum_{i=1}^{n}\delta_iY_i\psi(\mathbf{Z}_i)+\frac{n}{m}\sum_{i=\ell+1}^{n}(1-\delta_i)Y_i\widehat{E}(\psi(\mathbf{Z}_i)|\mathbf{X}_i)\right]-E(Y\psi(\mathbf{Z}))\right|>\frac{\varepsilon}{6}\right\}$$

$$:=\mathbf{I}_n+\mathbf{II}_n+\mathbf{III}_n\quad\text{(say)},\tag{60}$$

where $\widehat{E}(\psi^k(\mathbf{Z}_i)|\mathbf{X}_i)=\sum_{j=1}^{\ell}\delta_j\psi^k(\mathbf{Z}_i)\mathscr{K}((\mathbf{X}_j-\mathbf{X}_i)/h_\ell)\div\sum_{j=1}^{\ell}\delta_j\mathscr{K}((\mathbf{X}_j-\mathbf{X}_i)/h_\ell)$, for $k=1,2$, and $\mathbf{X}_i=(Y_i,\mathbf{V}_i^{\mathrm{T}})^{\mathrm{T}}$. We may bound $\mathbf{I}_n$ as follows. Let $\mathscr{G}=\{\psi^2|\psi\in\Psi\}$ and observe that $\forall g', g''\in\mathscr{G}$

$$\|g'-g''\|_{\mathscr{W}_\ell(\mathbf{X})}=\sum_{j=1}^{\ell}\delta_j\mathscr{W}_{\ell,j}(\mathbf{x})|g'(\mathbf{Z}_j)-g''(\mathbf{Z}_j)|$$

$$\leqslant\sum_{j=1}^{\ell}\delta_j\mathscr{W}_{\ell,j}(\mathbf{x})[|\psi'(\mathbf{Z}_j)-\psi''(\mathbf{Z}_j)|\times|\psi'(\mathbf{Z}_j)+\psi''(\mathbf{Z}_j)|]$$

$$\leqslant 2C\|\psi'-\psi''\|_{\mathscr{W}_\ell(\mathbf{X})}\quad\text{where } C=\|\psi'\|_\infty=\|\psi''\|_\infty.$$

In other words, if $\{\psi_1,\ldots,\psi_N\}$ is a minimal $\varepsilon/2C$-cover for Ψ, w.r.t. $\|\|_{\mathscr{W}_\ell(\mathbf{X})}$ then the class of functions $\{g_{\psi_1},\ldots,g_{\psi_N}\}$, where $g_{\psi_i}(\mathbf{Z})=\psi_i^2(\mathbf{Z})$, is an ε-cover of $\mathscr{G}$. Therefore, $\mathscr{N}(\varepsilon,\mathscr{G},\|\|_{\mathscr{W}_\ell(\mathbf{X})})\leqslant\mathscr{N}(\varepsilon/(2C),\Psi,\|\|_{\mathscr{W}_\ell(\mathbf{X})})$. Now this fact together with Theorem 2.2 give the bound

$$\mathbf{I}_n\leqslant c_{30}E[\mathscr{N}(c_{31}\varepsilon,\psi,\|\|_{1,n})]\mathrm{e}^{-c_{32}n\varepsilon^2}+c_{33}mE[\mathscr{N}(c_{34}\varepsilon,\psi,\|\|_{\mathscr{W}_\ell(\mathbf{X})})]\mathrm{e}^{-c_{35}m\varepsilon^2}$$

$$+m\sqrt{E[\mathscr{N}(c_{36}\varepsilon,\psi,\|\|_{\mathscr{W}_\ell(\mathbf{X})})]^2}\times[c_{37}\mathrm{e}^{-c_{38}\ell h_\ell^d}+c_{39}\mathrm{e}^{-c_{40}\ell h_\ell^d\varepsilon^2}],$$

where $c_{30},\ldots,c_{40}$ are positive constants not depending on n. The term $\mathbf{II}_n$ is rather trivial to deal with:

$$\mathbf{II}_n\leqslant P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\delta_iY_i^2-E(\delta Y^2)\right|>\frac{\varepsilon}{12}\right\}+P\left\{\left|\frac{1}{m}\sum_{i=\ell+1}^{n}(1-\delta_i)Y_i^2-E((1-\delta)Y^2)\right|>\frac{\varepsilon}{12}\right\}$$

$$\leqslant 2\mathrm{e}^{-c_{41}n\varepsilon^2}+2\mathrm{e}^{-c_{42}m\varepsilon^2}\quad\text{(via two applications of Hoeffding's inequality).}$$

As for the last term, $\mathbf{III}_n$, Theorem 2.3 implies that for large n,

$$\mathbf{III}_n \leqslant c_{43}E[\mathscr{N}(c_{44}\varepsilon, \psi, \|\|_{1,n})]e^{-c_{39}n\varepsilon^2} + c_{45}mE[\mathscr{N}(c_{46}\varepsilon, \psi, \|\|_{\mathscr{W}_\ell(\mathbf{X})})]e^{-c_{47}m\varepsilon^2}$$
$$+ m\sqrt{E[\mathscr{N}(c_{48}\varepsilon, \psi, \|\|_{\mathscr{W}_\ell(\mathbf{X})})]^2} \times [c_{49}e^{-c_{46}\ell h_\ell^d} + c_{50}e^{-c_{51}\ell h_\ell^d\varepsilon^2}],$$

for positive constants, $c_{43}, \ldots, c_{51}$, that do not depend on n. This completes the proof of Theorem 3.1. □

Proof of Lemma 3.1. Using the decomposition $E(|\psi_n(\mathbf{Z}) - Y|^2|\mathscr{D}_n) = E(|\psi_n(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2|\mathscr{D}_n) + E|\psi^*(\mathbf{Z}) - Y|^2$, one may write

$$E(|\psi_n(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2|\mathscr{D}_n) = \left[E(|\psi_n(\mathbf{Z}) - Y|^2|\mathscr{D}_n) - \inf_{\psi\in\Psi} E|\psi(\mathbf{Z}) - Y|^2\right]$$
$$+ \left[\inf_{\psi\in\Psi} E|\psi(\mathbf{Z}) - Y|^2 - E|\psi^*(\mathbf{Z}) - Y|^2\right].$$

However, the second square-bracketed term above is equal to $\inf_{\psi\in\Psi}E|\psi(\mathbf{Z}) - \psi^*(\mathbf{Z})|^2$, and

$$\text{the first square-bracketed term} = \sup_{\psi\in\Psi}\{E(|\psi_n(\mathbf{Z}) - Y|^2|\mathscr{D}_n) - \widehat{L}_{\mathrm{MI}}(\psi_n)$$
$$+ \widehat{L}_{\mathrm{MI}}(\psi_n) - \widehat{L}_{\mathrm{MI}}(\psi) + \widehat{L}_{\mathrm{MI}}(\psi) - E|\psi(\mathbf{Z}) - Y|^2\}$$
$$\leqslant 2\sup_{\psi\in\Psi}|\widehat{L}_{\mathrm{MI}}(\psi) - E|\psi(\mathbf{Z}) - Y|^2| \quad \text{since (61)} \leqslant 0. \qquad \square \tag{61}$$

Proof of Lemma 3.2. By the definition of $\hat{p}_n$, for every $0 < p \in \mathscr{P}$, one has $\widehat{L}_{\mathrm{MI}}(\hat{p}_n/p) = \widehat{L}_{\mathrm{MI}}(\hat{p}_n) - \hat{L}_{\mathrm{MI}}(p) \geqslant 0$. Furthermore, by the concavity of the logarithmic function

$$\frac{1}{2}\log\frac{\hat{p}_n}{p_0}I\{p_0 > 0\} \leqslant \log\frac{\hat{p}_n + p_0}{2p_0}I\{p_0 > 0\}.$$

Therefore,

$$0 \leqslant \widehat{L}_{\mathrm{MI}}\left(\frac{\hat{p}_n}{p_0}I\{p_0 > 0\}\right)$$
$$\leqslant 2\widehat{L}_{\mathrm{MI}}\left(\frac{\hat{p}_n + p_0}{2p_0}I\{p_0 > 0\}\right)$$
$$= 2\left[\widehat{L}_{\mathrm{MI}}\left(\frac{\hat{p}_n + p_0}{2p_0}I\{p_0 > 0\}\right) - L_n\left(\frac{\hat{p}_n + p_0}{2p_0}I\{p_0 > 0\}\right)\right] + 2L_n\left(\frac{\hat{p}_n + p_0}{2p_0}I\{p_0 > 0\}\right),$$

where L_n is given by (40). But $(\hat{p}_n + p_0)/2$ is also a density and hence, by Lemma 1.3 of van de Geer (2000),

$$L_n\left(\frac{\hat{p}_n + p_0}{2p_0}I\{p_0 > 0\}\right) \leqslant -2d_{\mathrm{H}}^2\left(\frac{\hat{p}_n + p_0}{2}, p_0\right),$$

which completes the proof of the lemma (since $\bar{p}_n := (\hat{p}_n + p_0)/2$). □

Proof of Lemma 4.1. Fix $\varepsilon > 0$ and let $\mathscr{G}_{\varepsilon/3}$ be an $\varepsilon/3$-cover of $\mathscr{G}$, w.r.t. the $\|\|_\infty$-norm, (recall that $\mathscr{G}$ is totally bounded). Also, let $\mathscr{N}_\infty(\varepsilon/3, \mathscr{G})$ be the cardinality of the smallest such covers. Then for every $g \in \mathscr{G}$ and every $\varepsilon > 0$, there is $g^* \in \mathscr{G}_{\varepsilon/3}$ such that $|\overline{\mathscr{V}}_n(g) - \mathscr{V}(g)| \leqslant 2\varepsilon/3 + |\overline{\mathscr{V}}_n(g^*) - \mathscr{V}(g^*)|$, where we have used the fact that $\sum_{j=1,\neq i}^n \varphi_{n,j}(\mathbf{X}) \leqslant 1$. Hence

$$P\left\{\sup_{g\in\mathscr{G}}|\overline{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon\right\} \leqslant \mathscr{N}_\infty\left(\frac{\varepsilon}{3}, \mathscr{G}\right)\max_{g\in\mathscr{G}_{\varepsilon/3}} P\{|\mathscr{V}_n(g) - \mathscr{V}(g)| > \varepsilon/3\}. \tag{62}$$

However, $P\{|\overline{\mathscr{V}}_n(g) - \mathscr{V}(g)| > \varepsilon/3\} \leqslant D_{n,1} + D_{n,2} + D_{n,3}$, where

$$D_{n,1} = P\left\{n^{-1}\left|\sum_{i=1}^{n} g(\mathbf{Z}_i) - n\mathscr{V}(g)\right| > \frac{\varepsilon}{9}\right\}$$

$$D_{n,2} = P\left\{n^{-1}\left|\sum_{i=1}^{n}(1-\delta_i)[g(\mathbf{Z}_i) - E(g(\mathbf{Z}_i)|\mathbf{X}_i)]\right| > \frac{\varepsilon}{9}\right\}$$

$$D_{n,3} = P\left\{n^{-1}\left|\sum_{i=1}^{n}(1-\delta_i)\left[\sum_{j=1,\neq i}^{n}\varphi_{n,j}(\mathbf{X}_i)g(\mathbf{Z}_j) - E(g(\mathbf{Z}_i)|\mathbf{X}_i)\right]\right| > \frac{\varepsilon}{9}\right\}.$$

Using Hoeffding's inequality, (also, cf. (56) and (57)) one immediately finds $D_{n,1} \leqslant 2\exp\{-n\varepsilon^2/(72B^2)\}$ and $D_{n,2} \leqslant 2\exp\{-n\varepsilon^2/(288B^2)\}$. Furthermore, arguments similar to (and in fact simpler than) those used in the proof of (22) lead to the bound: $D_{n,3} \leqslant \sum_{i=1}^{n} q_{n,i}$, with

$$0 \leqslant q_{n,i} \leqslant P\left\{2g|\hat{f}_{n-1}(\mathbf{X}_i) - f(\mathbf{X}_i)| + B^{-1}|\hat{\phi}_{n-1}(g|\mathbf{X}_i) - \phi(g|\mathbf{X}_i)|\right.$$
$$\left. + |\hat{\phi}_{n-1}(1|\mathbf{X}_i) - \phi(1|\mathbf{X}_i)| > \frac{p_{\min} f_{\min}\varepsilon}{9B}\right\}.$$

Here, $\hat{f}_{n-1}(\mathbf{X}_i) = (n-1)^{-1}h_n^{-d}\sum_{j=1,\neq i}^{n}\mathscr{K}((\mathbf{X}_j - \mathbf{X}_i)/h_n)$ and $\phi(g|\mathbf{X}_i) = f(\mathbf{X}_i)E[\delta_i g(\mathbf{Z}_i)|\mathbf{X}_i]$ and $\hat{\phi}_{n-1}(g|\mathbf{X}_i) = (n-1)^{-1}h_n^{-d}\sum_{j=1,\neq i}^{n}\delta_j g(\mathbf{Z}_j)\mathscr{K}((\mathbf{X}_j - \mathbf{X}_i)/h_n)$. On the other hand, under the assumptions of Lemma 4.1, $|E[\hat{\phi}_{n-1}(g|\mathbf{X}_i)|\mathbf{X}_i] - \phi(g|\mathbf{X}_i)| \leqslant h_n d_2 p_{\min} f_{\min}/27$, where d_2 is defined in Lemma 4.1, (the proof is similar to, and in fact simpler than, the proof of Lemma 4.2). Using this fact, it is straightforward to show (via Bennett's inequality) that for every $\gamma > 0$,

$$P\left\{|\hat{\phi}_{n-1}(g|\mathbf{X}_i) - \phi(g|\mathbf{X}_i)| > \gamma g|\mathbf{X}_i\right\} \leqslant 2\exp\left\{-\frac{(\gamma - c_0 h_n)^2(n-1)h_n^d}{2B\|\mathscr{K}\|_\infty[B\|f\|_\infty + (\gamma - c_0 h_n)]}\right\} \wedge 1,$$

where $c_0 = d_2 p_{\min} f_{\min}/27$. Similarly, one can also show that $|E[\hat{f}_{n-1}(\mathbf{X}_i)|\mathbf{X}_i] - f(\mathbf{X}_i)| \leqslant c_{00}h_n$, where $c_{00} = d_6 p_{\min} f_{\min}/(54B)$. Consequently, for every $\gamma > 0$,

$$P\{|\hat{f}_{n-1}(\mathbf{X}_i) - f(\mathbf{X}_i)| > \gamma|\mathbf{X}_i\} \leqslant 2\exp\left\{-\frac{(\gamma - c_{00}h_n)^2(n-1)h_n^d}{2\|\mathscr{K}\|_\infty[\|f\|_\infty + (\gamma - c_{00}h_n)]}\right\} \wedge 1.$$

The rest of the proof is trivial. □

Proof of Lemma 4.2. Since $\mathbf{X}$ is independent of $(\mathbf{Z}_1, \delta_1), \ldots, (\mathbf{Z}_\ell, \delta_\ell)$,

$$\begin{aligned} E[\hat{\phi}(g|\mathbf{X})|\mathbf{X}] - \phi(g|\mathbf{X}) &= h_\ell^{-d}E[\delta_1 g(\mathbf{Z}_1)\mathscr{K}((\mathbf{X}_1 - \mathbf{X})/h_\ell)|\mathbf{X}] - f(\mathbf{X})E[\delta g(\mathbf{Z})|\mathbf{X}] \\ &= h_\ell^{-d}E[\mathscr{K}((\mathbf{X}_1 - \mathbf{X})/h_\ell)E(\delta_1 g(\mathbf{Z}_1)|\mathbf{X}, \mathbf{X}_1)|\mathbf{X}] + f(\mathbf{X})\mathscr{V}(g|\mathbf{X}), \end{aligned}$$

where $\mathscr{V}(g|\mathbf{X}) = E(\delta g(\mathbf{Z})|\mathbf{X})$. Once again, from the independence of $(\mathbf{Z}_1, \delta_1)$ and $\mathbf{X}$, one obtains $E(\delta_1 g(\mathbf{Z}_1)|\mathbf{X}, \mathbf{X}_1) = E(\delta_1 g(\mathbf{Z}_1)|\mathbf{X}_1) =: \mathscr{V}(g|\mathbf{X}_1)$. Consequently, one finds

$$
\begin{aligned}
&E[\hat{\phi}(g|\mathbf{X})|\mathbf{X}] - \phi(g|\mathbf{X}) \\
&\quad = h_\ell^{-d} E[(\mathscr{V}(g|\mathbf{X}_1) - \mathscr{V}(g|\mathbf{X}))\mathscr{K}((\mathbf{X}_1 - \mathbf{X})/h_\ell)|\mathbf{X}] \\
&\qquad + E[\mathscr{V}(g|\mathbf{X})(h_\ell^{-d}\mathscr{K}((\mathbf{X}_1 - \mathbf{X})/h_\ell) - f(\mathbf{X}))|\mathbf{X}] \\
&\quad = T_1(\mathbf{X}) + T_2(\mathbf{X}) \quad \text{(say)}.
\end{aligned} \tag{63}
$$

A one-term Taylor expansion gives

$$
T_1(\mathbf{X}) = h_\ell^{-d} E\left[\left(\sum_{i=1}^{d} \frac{\partial \mathscr{V}(g|\mathbf{X}^*)}{\partial x_i}(X_{1,i} - X_i)\right)\mathscr{K}\left(\frac{\mathbf{X}_1 - \mathbf{X}}{h_\ell}\right)\middle|\mathbf{X}\right],
$$

where $X_{1,i}$ and X_i are the ith components of $\mathbf{X}_1$ and $\mathbf{X}$, respectively, and $\mathbf{X}^*$ is on the interior of the line segment joining $\mathbf{X}$ and $\mathbf{X}_1$. Thus

$$
\begin{aligned}
|T_1(\mathbf{X})| &\leqslant C_g \sum_{i=1}^{d} E\left[|X_{1,i} - X_i| \cdot h_\ell^{-d}\mathscr{K}\left(\frac{\mathbf{X}_1 - \mathbf{X}}{h_\ell}\right)\middle|\mathbf{X}\right] \quad \text{where } C_g = \bigvee_{i=1}^{d} \sup_g \sup_{\mathbf{x}} \left|\frac{\partial \mathscr{V}(g|\mathbf{x})}{\partial x_i}\right| \\
&= C_g \sum_{i=1}^{d} \int_{R^d} |x_i - X_i| h_\ell^{-d}\mathscr{K}\left(\frac{\mathbf{x} - \mathbf{X}}{h_\ell}\right) f(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&\leqslant C_g \|f\|_\infty \sum_{i=1}^{d} \int_{R^d} h_\ell |y_i| \mathscr{K}(\mathbf{y})\,\mathrm{d}\mathbf{y}.
\end{aligned} \tag{64}
$$

As for $T_2(\mathbf{X})$ that appears in (63), note that

$$
\begin{aligned}
T_2(\mathbf{X}) &= \mathscr{V}(g|\mathbf{X}) \int_{R^d} h_\ell^{-d}\mathscr{K}\left(\frac{\mathbf{u} - \mathbf{X}}{h_\ell}\right)[f(\mathbf{u}) - f(\mathbf{X})]\,\mathrm{d}\mathbf{u} \\
&= \mathscr{V}(g|\mathbf{X}) \int_{R^d} [f(\mathbf{X} + h_\ell \mathbf{y}) - f(\mathbf{X})]\mathscr{K}(\mathbf{y})\,\mathrm{d}\mathbf{y}.
\end{aligned}
$$

Since $|\mathscr{V}(g|\mathbf{X})| \leqslant B$, a one-term Taylor expansion results in

$$
|T_2(\mathbf{X})| \leqslant \left(Bd\|f'\|_\infty \sum_{i=1}^{d} \int |y_i| \mathscr{K}(\mathbf{y})\,\mathrm{d}\mathbf{y}\right) h_\ell. \tag{65}
$$

This last bound together with (64) and (63) complete the proof of Lemma 4.2. □

Proof of Lemma 4.3. We will use standard symmetrization arguments as follows. Let

$$
\hat{\mu}(g|\mathbf{X}) = \ell^{-1} \sum_{j=1}^{\ell} \delta_j g(\mathbf{Z}_j)\mathscr{K}\left(\frac{\mathbf{X}_j - \mathbf{X}}{h_\ell}\right) \quad \text{and} \quad \mu(g|\mathbf{X}) = E\left(\delta_0 g(\mathbf{Z}_0)\mathscr{K}\left(\frac{\mathbf{X}_0 - \mathbf{X}}{h_\ell}\right)\middle|\mathbf{X}\right), \tag{66}
$$

where $(\mathbf{Z}_0, \delta_0) = (\mathbf{X}_0, \mathbf{Y}_0, \delta_0) \overset{\text{iid}}{=} (\mathbf{X}_1, \mathbf{Y}_1, \delta_1)$. Now, clearly

$$
P\left\{\sup_{g\in\mathscr{G}} |\hat{\phi}(g|\mathbf{X}) - E[\hat{\phi}(g|\mathbf{X})|\mathbf{X}]| \geqslant \frac{\gamma}{2}\middle|\mathbf{X}\right\} = P\left\{\sup_{g\in\mathscr{G}} h_\ell^{-d} g|\hat{\mu}(g|\mathbf{X}) - \mu(g|\mathbf{X})g| \geqslant \frac{\gamma}{2}\middle|\mathbf{X}\right\}. \tag{67}
$$

Fix $\mathscr{D}_\ell$ and $\mathbf{X}$, and note that if $\sup_{g\in\mathscr{G}} h_\ell^{-d}|\hat{\mu}(g|\mathbf{X}) - \mu(g|\mathbf{X})| \geqslant \gamma/2$ then there is at least some $g^* \in \mathscr{G}$ (which will depend on $\mathscr{D}_\ell$ and $\mathbf{X}$), such that $h_\ell^{-d}|\hat{\mu}(g^*|\mathbf{X}) - \mu(g^*|\mathbf{X}, \mathscr{D}_\ell)| \geqslant \gamma/2$, where

$$
\mu(g^*|\mathbf{X}, \mathscr{D}_\ell) = E\left(\delta_0 g^*(\mathbf{Z}_0)\mathscr{K}\left(\frac{\mathbf{X}_0 - \mathbf{X}}{h_\ell}\right)\middle|\mathbf{X}, \mathscr{D}_\ell\right).
$$

Let $\mathscr{D}'_\ell = \{(\mathbf{X}'_i, \mathbf{Y}'_i, \delta'_i), i = 1, \ldots, \ell\}$ be a hypothetical sample independent of $\mathscr{D}_\ell$ and $\mathbf{X}$. Define the counterpart of $\hat{\mu}(g|\mathbf{X})$ by

$$\hat{\mu}'(g|\mathbf{X}) = \ell^{-1} \sum_{j=1}^{\ell} \delta'_j g(\mathbf{Z}'_j) \mathscr{K}\left(\frac{\mathbf{X}'_j - \mathbf{X}}{h_\ell}\right)$$

and observe that

$$P\left\{h_\ell^{-d}|\hat{\mu}'(g^*|\mathbf{X}) - \mu(g^*|\mathbf{X}, \mathscr{D}_\ell)| < \frac{\gamma}{4}\middle| \mathbf{X}, \mathscr{D}_\ell\right\} \geqslant 1 - \sup_{g\in\mathscr{G}} P\left\{\ell^{-1}\left|\sum_{j=1}^{\ell} W_j\right| \geqslant \frac{\gamma}{4}\middle| \mathbf{X}\right\}, \tag{68}$$

where, for $j = 1, \ldots, \ell$,

$$W_j = h_\ell^{-d}\left[\delta_j g(\mathbf{Z}_j)\mathscr{K}\left(\frac{\mathbf{X}_j - \mathbf{X}}{h_\ell}\right) - E\left(\delta_0 g(\mathbf{Z}_0)\mathscr{K}\left(\frac{\mathbf{X}_0 - \mathbf{X}}{h_\ell}\right)\middle| \mathbf{X}\right)\right].$$

However, conditional on $\mathbf{X}$, the terms W_j are independent, zero-mean random variables, bounded by $-h_\ell^{-d} B\|\mathscr{K}\|_\infty$ and $+h_\ell^{-d} B\|\mathscr{K}\|_\infty$. Furthermore, $\mathrm{Var}(W_j|\mathbf{X}) = E(W_j^2|\mathbf{X}) \leqslant h_\ell^{-2d} B^2 E[\mathscr{K}^2((\mathbf{X}_j - \mathbf{X})/h_\ell)|\mathbf{X}] \leqslant h_\ell^{-d} B^2 \|\mathscr{K}\|_\infty \int_{R^d} \mathscr{K}(\mathbf{u}) f(\mathbf{X} + \mathbf{u}h_\ell)\, d\mathbf{u} \leqslant h_\ell^{-d} B^2 \|\mathscr{K}\|_\infty \|f\|_\infty$. Therefore, by Bennett's inequality, and the fact that $\ell h_\ell^d \to \infty$, as $n \to \infty$, one finds

$$P\left\{\ell^{-1}\left|\sum_{j=1}^{\ell} W_j\right| \geqslant \frac{\gamma}{4}\middle| \mathbf{X}\right\} \leqslant 2\exp\left\{\frac{-\ell h_\ell^d \gamma^2}{32B\|\mathscr{K}\|_\infty(B\|f\|_\infty + \gamma/4)}\right\} \overset{\text{for large } n}{\leqslant} \frac{1}{2}. \tag{69}$$

Combining the above results, one finds for large n (and thus ℓ),

$$\begin{aligned}
\frac{1}{2} &\leqslant P\left\{h_\ell^{-d} g|\hat{\mu}'(g^*|\mathbf{X}) - \mu(g^*|\mathbf{X}, \mathscr{D}_\ell)g| < \frac{\gamma}{4}\middle| \mathbf{X}, \mathscr{D}_\ell\right\} \\
&\leqslant P\left\{-h_\ell^{-d} g|\hat{\mu}'(g^*|\mathbf{X}) - \hat{\mu}(g^*|\mathbf{X})| + \underbrace{h_\ell^{-d} g|\hat{\mu}(g^*|\mathbf{X}) - \mu(g^*|\mathbf{X}, \mathscr{D}_\ell)|}_{\geqslant \gamma/2} < \frac{\gamma}{4}\middle| \mathbf{X}, \mathscr{D}_\ell\right\} \\
&\leqslant P\left\{\sup_{g\in\mathscr{G}} |\hat{\mu}'(g|\mathbf{X}) - \hat{\mu}(g|\mathbf{X})| > \frac{\gamma h_\ell^d}{4}\middle| \mathbf{X}, \mathscr{D}_\ell\right\}.
\end{aligned} \tag{70}$$

Note that the far left and far right sides of (70) do not depend on any particular $g^* \in \mathscr{G}$. Multiplying both sides by $I\{\sup_{g\in\mathscr{G}}|\hat{\mu}(g|\mathbf{X}) - \mu(g|\mathbf{X})| > \gamma h_\ell^d/2\}$ and taking expectation w.r.t. the distribution of $\mathscr{D}_\ell$, one finds (for large n),

$$P\left\{\sup_{g\in\mathscr{G}} |\hat{\mu}(g|\mathbf{X}) - \mu(g|\mathbf{X})| > \frac{\gamma h_\ell^d}{2}\middle| \mathbf{X}\right\} \leqslant 2P\left\{\sup_{g\in\mathscr{G}} |\hat{\mu}'(g|\mathbf{X}) - \hat{\mu}(g|\mathbf{X})| > \frac{\gamma h_\ell^d}{4}\middle| \mathbf{X}\right\}. \tag{71}$$

Now, observe that the joint distribution of the vector

$$\left(\delta_1 g(\mathbf{Z}_1)\mathscr{K}\left(\frac{\mathbf{X}_1 - \mathbf{X}}{h_\ell}\right), \ldots, \delta_\ell g(\mathbf{Z}_\ell)\mathscr{K}\left(\frac{\mathbf{X}_\ell - \mathbf{X}}{h_\ell}\right)\right)$$

is the same as that of the vector

$$\left(\delta'_1 g(\mathbf{Z}'_1)\mathscr{K}\left(\frac{\mathbf{X}'_1 - \mathbf{X}}{h_\ell}\right), \ldots, \delta'_\ell g(\mathbf{Z}'_\ell)\mathscr{K}\left(\frac{\mathbf{X}'_\ell - \mathbf{X}}{h_\ell}\right)\right),$$

and this joint distribution is not affected if the corresponding components of these two vectors are randomly interchanged. Thus, for an independent Rademacher sequence, $\sigma_1, \cdots, \sigma_\ell$, independent of $\mathbf{X}$, $(\mathbf{X}_i, \mathbf{Y}_i, \delta_i)$, and $(\mathbf{X}_i', \mathbf{Y}_i', \delta_i')$, $i = 1, \ldots, \ell$, one has

$$\sup_{g\in\mathscr{G}} \frac{1}{\ell}\left|\sum_{i=1}^{\ell}\left[\delta_i g(\mathbf{Z}_i)\mathscr{K}\left(\frac{\mathbf{X}_i-\mathbf{X}}{h_\ell}\right)-\delta_i' g(\mathbf{Z}_i')\mathscr{K}\left(\frac{\mathbf{X}_i'-\mathbf{X}}{h_\ell}\right)\right]\right|$$

$$\stackrel{\mathrm{d}}{=}\sup_{g\in\mathscr{G}} \frac{1}{\ell}\left|\sum_{i=1}^{\ell}\sigma_i\left[\delta_i g(\mathbf{Z}_i)\mathscr{K}\left(\frac{\mathbf{X}_i-\mathbf{X}}{h_\ell}\right)-\delta_i' g(\mathbf{Z}_i')\mathscr{K}\left(\frac{\mathbf{X}_i'-\mathbf{X}}{h_\ell}\right)\right]\right|,$$

which leads to

$$\text{R.H.S. of (71)} \leqslant 2P\left\{\sup_{g\in\mathscr{G}}\left|\sum_{i=1}^{\ell}\sigma_i\left[\delta_i g(\mathbf{Z}_i)\mathscr{K}\left(\frac{\mathbf{X}_i-\mathbf{X}}{h_\ell}\right)\right]\right| > \frac{\gamma\ell h_\ell^d}{8}\,\middle|\,\mathbf{X}\right\}. \tag{72}$$

Next, fix $\mathscr{D}_\ell$, put $\varepsilon''=\gamma/(32\|f\|_\infty)$, and let $\mathscr{G}_{\ell,\varepsilon''}$ be an ε''-cover of $\mathscr{G}$ with respect to $\|\|_{\mathscr{W}_\ell(\mathbf{x})}$. Also, let $\mathscr{N}(\varepsilon'', \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{x})})$ be the ε'' covering number of $\mathscr{G}$ w.r.t $\|\|_{\mathscr{W}_\ell(\mathbf{x})}$. Then, for some $g^* \in \mathscr{G}_{\ell,\varepsilon''}$,

$$\left|\sum_{i=1}^{\ell}\sigma_i\left[\delta_i g(\mathbf{Z}_i)\mathscr{K}\left(\frac{\mathbf{X}_i-\mathbf{X}}{h_\ell}\right)\right]\right| \leqslant \left|\sum_{i=1}^{\ell}\sigma_i\delta_i g^*(\mathbf{Z}_i)\mathscr{K}\left(\frac{\mathbf{X}_i-\mathbf{X}}{h_\ell}\right)\right| + \ell h_\ell^d\,\widehat{f}_\ell(\mathbf{X})\|g-g^*\|_{\mathscr{W}_\ell(\mathbf{X})}$$

$$\leqslant |\text{the first term above}| + \ell h_\ell^d\,\widehat{f}_\ell(\mathbf{X})\varepsilon''.$$

Therefore, one finds

$$\begin{aligned}
&\text{(R.H.S. of (72))}\\
&\leqslant 2E\left[I\{\widehat{f}_\ell(\mathbf{X}) < 2\|f\|_\infty\}\right.\\
&\quad\left.\times P\left\{\sup_{g\in\mathscr{G}_{\ell,\varepsilon''}}\frac{1}{\ell}\left|\sum_{i=1}^{\ell}\sigma_i\delta_i g(\mathbf{Z}_i)\mathscr{K}\left(\frac{\mathbf{X}_i-\mathbf{X}}{h_\ell}\right)\right| > \frac{\gamma h_\ell^d}{8} - h_\ell^d\,\widehat{f}_\ell(\mathbf{X})\varepsilon''\,\middle|\,\mathbf{X}, \mathscr{D}_\ell\right\}\,\middle|\,\mathbf{X}\right]\\
&\quad + 2P\{\widehat{f}_\ell(\mathbf{X}) \geqslant 2\|f\|_\infty\}\\
&\leqslant 2E\left[\mathscr{N}(\varepsilon'', \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{X})})\max_{g\in\mathscr{G}_{\ell,\varepsilon''}} P\left\{\frac{1}{\ell}\left|\sum_{i=1}^{\ell}\sigma_i\delta_i g(\mathbf{Z}_i)\mathscr{K}\left(\frac{\mathbf{X}_i-\mathbf{X}}{h_\ell}\right)\right| > \frac{\gamma h_\ell^d}{16}\,\middle|\,\mathbf{X}, \mathscr{D}_\ell\right\}\,\middle|\,\mathbf{X}\right]\\
&\quad + 2P\{\widehat{f}_\ell(\mathbf{X}) \geqslant 2\|f\|_\infty\}\\
&:= \mathbf{I}_{n,1} + \mathbf{I}_{n,2} \quad \text{(say)}.
\end{aligned} \tag{73}$$

On the other hand, conditional on $\mathscr{D}_\ell$ and $\mathbf{X}$, the terms $\sigma_i\delta_i g(\mathbf{Z}_i)\mathscr{K}((\mathbf{X}_i-\mathbf{X})/h_\ell)$, $i = 1, \ldots, \ell$ are independent, zero-mean random variables bounded by $-B\mathscr{K}((\mathbf{X}_i-\mathbf{X})/h_\ell)$ and $+B\mathscr{K}((\mathbf{X}_i-\mathbf{X})/h_\ell)$. Therefore, bounding the above inner conditional probability (in the definition of $\mathbf{I}_{n,1}$) via Hoeffding's inequality, one obtains

$$\mathbf{I}_{n,1} \leqslant 4E\left[\mathscr{N}\left(\frac{\gamma}{32\|f\|_\infty}, \mathscr{G}, \|\|_{\mathscr{W}_\ell(\mathbf{x})}\right)\exp\left\{-\frac{\gamma^2\ell^2 h_\ell^{2d}}{512B^2\sum_{i=1}^{\ell}\mathscr{K}^2((\mathbf{X}_i-\mathbf{X})/h_\ell)}\right\}\,\middle|\,\mathbf{X}\right]. \tag{74}$$

Put $V_i = \mathscr{K}((\mathbf{X}_i - \mathbf{X})/h_\ell) - E[\mathscr{K}((\mathbf{X}_i - \mathbf{X})/h_\ell)|\mathbf{X}]$ and observe that since $E[\mathscr{K}((\mathbf{X}_i - \mathbf{X})/h_\ell)|\mathbf{X}] \leqslant \|f\|_\infty h_\ell^d$, one may write

$$
\begin{aligned}
&E\left[\exp\left\{-\frac{\gamma^2\ell^2 h_\ell^{2d}}{256B^2\sum_{i=1}^{\ell}\mathscr{K}^2((\mathbf{X}_i-\mathbf{X})/h_\ell)}\right\}\middle|\mathbf{X}\right]\\
&\quad\leqslant E\left[\exp\left\{-\frac{\gamma^2\ell^2 h_\ell^{2d}}{256B^2\|\mathscr{K}\|_\infty(|\sum_{i=1}^{\ell}V_i| + \|f\|_\infty \ell h_\ell^d)}\right\}\right.\\
&\qquad\left.\times\left(I\left\{\left|\sum_{i=1}^{\ell}V_i\right| < \ell h_\ell^d\right\} + I\left\{\left|\sum_{i=1}^{\ell}V_i\right| \geqslant \ell h_\ell^d\right\}\right)\middle|\mathbf{X}\right]\\
&\quad\leqslant \exp\left\{-\frac{\gamma^2\ell h_\ell^d}{256B^2\|\mathscr{K}\|_\infty(1+\|f\|_\infty)}\right\} + P\left\{\left|\sum_{i=1}^{\ell}V_i\right| \geqslant \ell h_\ell^d\middle|\mathbf{X}\right\}.
\end{aligned}
$$

At the same time, since $E(V_i|\mathbf{X}) = 0$ and $\mathrm{Var}(V_i|\mathbf{X}) \leqslant h_\ell^d\|\mathscr{K}\|_\infty\|f\|_\infty$, one may once again invoke Bennett's inequality to conclude that $P\{|\sum_{i=1}^{\ell}V_i| \geqslant \ell h_\ell^d|\mathbf{X}\} \leqslant 2\exp\{-\ell h_\ell^d/[2\|\mathscr{K}\|_\infty(1+\|f\|_\infty)]\}$. Now the first term on the r.h.s. of (45) follows from (71)–(74) in conjunction with Cauchy–Schwarz inequality and the elementary fact that $|x+y|^r \leqslant |x|^r + |y|^r$, which holds for all $0 < r \leqslant 1$. As for the term $\mathbf{I}_{n,2}$ in (73), it is straightforward to show that $\mathbf{I}_{n,2} \leqslant 2P\{|\widehat{f}_\ell(\mathbf{X}) - f(\mathbf{X})| > \|f\|_\infty\} \leqslant 4\exp\{-\ell h_\ell^d\|f\|_\infty/(12\|\mathscr{K}\|_\infty)\}$. This completes the proof of (45). □

References

Cheng, P.E., 1994. Nonparametric estimation of mean functionals with data missing at random. J. Amer. Statist. Assoc. 89, 81–87.

Cheng, P.E., Chu, C.K., 1996. Kernel estimation of distribution functions and quantiles with missing data. Statist. Sinica 6, 63–78.

Hazelton, M.L., 2000. Marginal density estimation from incomplete bivariate data. Statist. Probab. Lett. 47, 75–84.

Hirano, K.I., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161–1189.

Kolmogorov, A.N., Tikhomirov, V.M., 1959. ε-entropy and ε-capacity of sets in function spaces. Uspehi Mat. Nauk 14, 3–86.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis With Missing Data. Wiley, New York.

Pollard, D., 1984. Convergence of Stochastic Processes. Springer, New York.

Polonik, W., Yao, Q., 2002. Set-indexed conditional empirical and quantile processes based on dependent data. J. Multivariate Anal. 80, 234–255.

Robins, J.M., Rotnitzky, A., Zhao, L., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89, 846–866.

Rubin, D., 1987. Applied Probability and Statistics. Wiley, New York.

van de Geer, S.A., 2000. Applications of Empirical Process Theory. Cambridge University Press, Cambridge.

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer, New York.

Wang, Q., Linton, O., Härdle, W., 2004. Semiparametric regression analysis with missing response at random. J. Amer. Statist. Assoc. 99, 334–345.