Stat Papers (2019) 60:1487–1513
https://doi.org/10.1007/s00362-017-0883-y

REGULAR ARTICLE

Kernel classification with missing data and the choice of smoothing parameters

Levon Demirdjian[1] · **Majid Mojirsheibani**[2]

Received: 29 August 2015 / Revised: 17 October 2016 / Published online: 2 February 2017

Abstract Methods are proposed for selecting smoothing parameters of kernel classifiers in the presence of missing covariates. Here the missing covariates can appear in both the data and in the unclassified observation that has to be classified. The proposed methods are quite straightforward to implement. Exponential performance bounds will be derived for the resulting classifiers. Such bounds, in conjunction with the Borel–Cantelli lemma, provide various strong consistency results. Several numerical examples are presented to illustrate the effectiveness of the proposed procedures.

Keywords Classification · Kernel · Missing covariate · Consistency · Shatter coefficient

1 Introduction

Consider the following standard two-class classification problem. Let $(\mathbf{Z}, Y)$ be a random pair with an underlying distribution $F_{\mathbf{Z},Y}$, where $\mathbf{Z} \in \mathbb{R}^s$, $s \geq 1$, is a vector of observed covariates, and $Y \in \{0, 1\}$ is the unobserved class membership of $\mathbf{Z}$. The problem is then to predict Y based on $\mathbf{Z}$. More specifically, in classification, one seeks to find a function (classifier) $\psi : \mathbb{R}^s \to \{0, 1\}$ for which the misclassification error probability, $P\{\psi(\mathbf{Z}) \neq Y\}$, is as small as possible (Devroye et al. 1996). The optimal or best classifier, denoted by ψ_B, is given by

✉ Majid Mojirsheibani
majid.mojirsheibani@csun.edu

[1] Department of Statistics, University of California, Los Angeles, CA 90095, USA

[2] Department of Mathematics, California State University, Northridge, CA 91330, USA

$$\psi_B(\mathbf{z}) = \begin{cases} 1 & \text{if } E(Y|\mathbf{Z}=\mathbf{z}) > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

and is sometimes called the Bayes classifier or the Bayes decision or rule in the literature on statistical classification [see, for example, (Devroye 1981; Devroye et al. 1996; Györfi et al. 2002)].

When the distribution $F_{\mathbf{Z},Y}$ is completely known, finding ψ_B does not pose a challenge. In practice however, $F_{\mathbf{Z},Y}$ is usually unknown (fully or partially), thus finding ψ_B is virtually impossible. One only has access to a random sample (data), denoted by $D_n = \{(\mathbf{Z}_i, Y_i); i = 1, \ldots, n\}$, where $(\mathbf{Z}_i, Y_i)$ are independently and identically distributed (iid) according to $F_{\mathbf{Z},Y}$. The goal is then to construct a sample-based classifier ψ_n whose error rate is in some sense small. A data-based rule ψ_n is said to be strongly consistent if

$$P\{\psi_n(\mathbf{Z}) \neq Y \mid D_n\} \xrightarrow{a.s.} P\{\psi_B(\mathbf{Z}) \neq Y\}.$$

If the convergence holds in probability, we say ψ_n is weakly consistent. Several nonparametric classifiers such as general partitioning methods (Glick 1973; Gordon and Olshen 1978, 1980), nearest neighbor rules (Devroye and Wagner 1982), and kernel rules (Devroye and Krzyzak 1989; Devroye and Wagner 1980; Krzyzak 1986) have been proposed in the literature with strong consistency properties.

In this paper we focus on the important problem where there may be missing covariates in $\mathbf{Z}_i$, $i = 1, \ldots, n$, and in the new unclassified covariate vector $\mathbf{Z}$. We consider kernel classifiers and propose methods to estimate the smoothing parameters of the kernels when missing covariates are present.

Much of the research on statistical estimation in the presence of missing covariates has relied on assumptions such as missingness at random (MAR); see, for example, Hu and Zhang (2012), Hirano and Ridder (2003) ,Wang et al. (2004), and Hazelton (2000). The MAR assumption means that the probability that a covariate is missing does not depend on that covariate itself. In classification, most of the existing results only deal with the case where covariates can be missing in the data (i.e., in $\mathbf{Z}_i$, $i = 1, \ldots, n$), but not in the new observation $\mathbf{Z}$, which has to be classified. See, for example, Chung and Han (2000) for the parametric case, and Pawlak (1993) for the nonparametric case. One source of difficulty is the fact that the optimal classifier in (1) corresponding to the case with no missing covariates is not necessarily the best when there are missing covariates in the new unclassified observation $\mathbf{Z}$. In fact, Mojirsheibani and Montazeri (2007) derive the optimal classifier in the presence of missing covariates, which is in general different from (1) and which works without imposing any MAR-type assumptions. Another representation of this classifier is given in Mojirsheibani (2012). Furthermore, Mojirsheibani and Montazeri (2007) propose nonparametric kernel estimators of the optimal classifier for this setup. Although these authors establish the strong consistency of their proposed kernel classifiers, they do not provide any directions as to how to estimate the unknown smoothing parameter of the kernels used. Finding good data-driven estimates of the kernel smoothing parameter is not just an important theoretical

consideration; from an applied point of view, a carefully selected bandwidth can help to minimize the error probability.

In this paper, we propose methods for selecting kernel smoothing parameters in the complicated case where covariates can be missing in both the data and in the new unclassified observation. In Sect. 2.1, we revisit kernel-based estimators of the optimal classifier corresponding to this setup. The proposed classifiers do not require any MAR-type assumptions on the missingness probability mechanism. In Sects. 2.2 and 2.3, we turn to the question of bandwidth selection by considering the methods of data-splitting and resubstitution. To evaluate the performance of the corresponding classification rules, exponential performance bounds are established on the deviations of their error probabilities from those of the optimal classifier. In Sect. 3, we present several numerical examples highlighting the effectiveness of the proposed methods. Proofs are postponed to the end of the paper.

2 Main results

2.1 Kernel classifier with missing covariates

Our discussion and results for classification with missing covariates are based on the following setup. Let $\mathbf{Z} = (\mathbf{X}', \mathbf{V}')' \in \mathbb{R}^{d+p}$ be the vector of covariates to be used to predict the class membership $Y \in \{0, 1\}$, where $\mathbf{X} \in \mathbb{R}^d$, $d \geq 1$ is always observable but $\mathbf{V} \in \mathbb{R}^p$, $p \geq 1$ can be missing. Also, let δ be a $\{0, 1\}$-valued random variable defined by

$$\delta = \begin{cases} 0 & \text{if } \mathbf{V} \text{ is missing} \\ 1 & \text{otherwise.} \end{cases}$$

Mojirsheibani (2012) and Mojirsheibani and Montazeri (2007) show that in this case, the optimal classifier is given by

$$\phi_B(\mathbf{z}, \delta) := \begin{cases} 1 & \text{if } \delta \dfrac{E(\delta Y|\mathbf{Z} = \mathbf{z})}{E(\delta|\mathbf{Z} = \mathbf{z})} + (1 - \delta) \dfrac{E[(1 - \delta)Y|\mathbf{X} = \mathbf{x}]}{E[(1 - \delta)|\mathbf{X} = \mathbf{x}]} > \dfrac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

with the convention $0/0 = 0$. The corresponding probability of misclassification is given by

$$L(\phi_B) = P\{\phi_B(\mathbf{Z}, \delta) \neq Y\}. \tag{3}$$

Observe that the classifier $\psi_B(\mathbf{z})$ in (1) is a special case of $\phi_B(\mathbf{z}, \delta)$ in (2); to see this, simply note that $\phi_B(\mathbf{z}, \delta)$ reduces to $\psi_B(\mathbf{z})$ whenever $P\{\delta = 1\} = 1$ (i.e., whenever there are no missing covariates).

Since the joint distribution of $(\mathbf{Z}, Y)$ is almost always unknown, the classifier ϕ_B in (2) is not available in practice and must be estimated by some sample-based classifier.

One typically has access to only some iid data $D_n = \{(\mathbf{Z}_1, Y_1), \ldots, (\mathbf{Z}_n, Y_n)\}$, which can also be represented by

$$\{(\mathbf{X}_1, \mathbf{V}_1, Y_1, \delta_1), \ldots, (\mathbf{X}_n, \mathbf{V}_n, Y_n, \delta_n)\}.$$

To define the kernel classification rule corresponding to (2), we replace the quantities $E(\delta Y|\mathbf{Z} = \mathbf{z})/E(\delta|\mathbf{Z} = \mathbf{z})$ and $E[(1-\delta)Y|\mathbf{X} = \mathbf{x}]/E[(1-\delta)|\mathbf{X} = \mathbf{x}]$ that appear in (2) by their corresponding kernel estimates given by

$$\hat{\tau}_1(\mathbf{z}) := \frac{\sum_{i=1}^{n} \delta_i Y_i K_1\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h_1}\right)}{\sum_{i=1}^{n} \delta_i K_1\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h_1}\right)} \text{ and } \hat{\tau}_0(\mathbf{x}) := \frac{\sum_{i=1}^{n} (1-\delta_i) Y_i K_0\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_0}\right)}{\sum_{i=1}^{n} (1-\delta_i) K_0\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_0}\right)}, \tag{4}$$

respectively. Here, the kernels K_1 and K_0 are maps of the form $K_j : \mathbb{R}^{d+jp} \to \mathbb{R}^+$ for $j = 0, 1$, and $h_1 \in \mathbb{R}^+$ and $h_0 \in \mathbb{R}^+$ are called the smoothing parameters or bandwidths of the kernels K_1 and K_0 respectively. The idea of replacing the unknown regression functions with their kernel regression estimates is fairly common in the literature on regression function estimation with missing data. For example, Mojirsheibani and Reese (2015) prove the strong consistency of such kernel regression estimates when the response variables can be missing; see also Karimi and Mohammadzadeh (2012) and Toutenburg and Shalabh (2003) for more on the estimation of regression functions for correlated data in the presence of missing response variables.

The kernel classification rule corresponding to (2) is then given by

$$\phi_n(\mathbf{z}, \delta) = \begin{cases} 1 & \text{if } \delta\hat{\tau}_1(\mathbf{z}) + (1-\delta)\hat{\tau}_0(\mathbf{x}) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The classifier in (5) would be useful if the values of h_0 and h_1 were known. As for the choice of the unknown parameters h_1 and h_0, we propose a number of methods to construct the estimates $\hat{h}_1$ and $\hat{h}_0$; these include a *data-splitting* approach and the *resubstitution* method.

2.2 Data splitting

We start by randomly splitting D_n into a training sequence D_m of size $m \equiv m(n)$ and a testing sequence $T_\ell = D_n - D_m$ of size $\ell \equiv \ell(n)$, where $m + \ell = n$. Here, $D_n = D_m \cup T_\ell$ and $D_m \cap T_\ell = \emptyset$. The training sequence D_m is used to construct a class Φ_m of kernel classifiers of the form

$$\phi_m(\mathbf{z}, \delta) = \begin{cases} 1 & \text{if } \delta\hat{\tau}_{1,m}(\mathbf{z}) + (1-\delta)\hat{\tau}_{0,m}(\mathbf{x}) > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where

$$\hat{\tau}_{1,m}(\mathbf{z}) = \frac{\sum_{i:(\mathbf{Z}_i,Y_i,\delta_i)\in D_m} \delta_i Y_i K_1\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h_1}\right)}{\sum_{i:(\mathbf{Z}_i,Y_i,\delta_i)\in D_m} \delta_i K_1\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h_1}\right)}, \quad \hat{\tau}_{0,m}(\mathbf{x}) = \frac{\sum_{i:(\mathbf{Z}_i,Y_i,\delta_i)\in D_m} (1-\delta_i) Y_i K_0\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_0}\right)}{\sum_{i:(\mathbf{Z}_i,Y_i,\delta_i)\in D_m} (1-\delta_i) K_0\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_0}\right)}, \tag{7}$$

and where K_1, K_0, h_1, and h_0 are defined as before. Observe that the class Φ_m is indexed by the values of h_1 and h_0. In what follows, we consider two cases: the case where the cardinality of Φ_m is finite (this is the case where the free parameters h_1 and h_0 can take a finite number of values), and the case where Φ_m has an infinite number of classifiers (i.e., the case where h_1 and h_0 can take an infinite number of values).

Case (i): Φ_m is finite

In the case where Φ_m is a finite class of such kernel rules, the parameters h_1 and h_0 will be chosen from a finite set of possible values. Specifically, assume that $h_1 \in H_1 = \{h_{1,1}, h_{1,2}, \ldots, h_{1,N_1}\}$ and $h_0 \in H_0 = \{h_{0,1}, h_{0,2}, \ldots, h_{0,N_0}\}$. Observe that in this setup, the cardinality of Φ_m is $|\Phi_m| = N_1 N_0$. Next, we use the testing sequence T_ℓ to select a classifier from Φ_m that minimizes the following estimate of the error probability

$$\widehat{L}_{m,\ell}(\phi_m) = \frac{1}{\ell} \sum_{i:(\mathbf{Z}_i,Y_i,\delta_i)\in T_\ell} I\{\phi_m(\mathbf{Z}_i, \delta_i) \neq Y_i\}. \tag{8}$$

Let $\hat{\phi}_n$ denote the classifier selected from Φ_m that minimizes (8), i.e., $\widehat{L}_{m,\ell}(\hat{\phi}_n) \leq \widehat{L}_{m,\ell}(\phi_m)$ for all $\phi_m \in \Phi_m$. Equivalently, $\hat{\phi}_n$ is the classifier that minimizes (8) as h_1 and h_0 vary over H_1 and H_0 respectively. Here, the subscript n of $\hat{\phi}_n$ emphasizes the fact that it depends on the entire data set of size n.

How good is $\hat{\phi}_n$ for predicting Y? To answer this question, let

$$L(\phi_m) = P\{\phi_m(\mathbf{Z}, \delta) \neq Y \mid D_m\} \quad \text{and} \quad L(\hat{\phi}_n) = P\{\hat{\phi}_n(\mathbf{Z}, \delta) \neq Y | D_n\}$$

denote the error probabilities of ϕ_m and $\hat{\phi}_n$, respectively. The following result gives exponential performance bounds for $\hat{\phi}_n$:

Theorem 1 *Let ϕ_m and $\hat{\phi}_n$ be the data-based classifiers defined as above. Then for any distribution of $(\mathbf{Z}, Y)$ and every $\epsilon > 0$,*

$$P\left\{L(\hat{\phi}_n) - \inf_{\phi_m \in \Phi_m} L(\phi_m) > \epsilon\right\} \leq 2N_1 N_0 e^{-\ell\epsilon^2/2}.$$

Remark 1 Of course, as $\ell \to \infty$, the bound in Theorem 1 yields

$$L(\hat{\phi}_n) - \inf_{\phi_m \in \Phi_m} L(\phi_m) \to 0$$

with probability one. But this falls short of achieving strong consistency for $\hat{\phi}_n$. To appreciate this, consider the decomposition

$$L(\hat{\phi}_n) - L(\phi_B) = \left\{ L(\hat{\phi}_n) - \inf_{\phi_m \in \Phi_m} L(\phi_m) \right\} + \left\{ \inf_{\phi_m \in \Phi_m} L(\phi_m) - L(\phi_B) \right\}.$$

Then although the first bracketed term goes to zero, there is no guarantee that the term $\{\inf_{\phi_m \in \Phi_m} L(\phi_m) - L(\phi_B)\}$ can become small unless the class Φ_m can capture the classifier ϕ_B as $m \to \infty$. To overcome such difficulties, we next consider the case where Φ_m has an infinite cardinality.

Case (ii): Φ_m is infinite

Now consider the case where Φ_m is the class of all kernel rules of the form in (6), but with parameters h_1 and h_0 chosen from an infinite set of possible values. Once again, let $\hat{\phi}_n$ denote the classifier that minimizes the empirical misclassification error in (8). To study the performance of $\hat{\phi}_n$, we first need to state the definition of the *shatter coefficient* of a set. Let $\mathcal{A}$ be a class of measurable sets in $\mathbb{R}^s$, where $s \geq 1$. The n^{th} shatter coefficient of $\mathcal{A}$, denoted by $S(\mathcal{A}, n)$, is defined by

$$S(\mathcal{A}, n) = \max_{\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^s} \{\text{number of different sets in } \{\{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \cap A | A \in \mathcal{A}\}\}.$$

The shatter coefficient $S(\mathcal{A}, n)$ measures the richness of the class $\mathcal{A}$. Let $\mathcal{A}_\Phi$ be the class of all sets of the form

$$A = \{\{\mathbf{z} | \phi(\mathbf{z}) = 1\} \times \{0\}\} \cup \{\{\mathbf{z} | \phi(\mathbf{z}) = 0\} \times \{1\}\}, \quad \phi \in \Phi \tag{9}$$

and define the n^{th} shatter coefficient of the class of classifiers Φ to be $S(\Phi, n) = S(\mathcal{A}_\Phi, n)$. Note that the size of $S(\mathcal{A}_\Phi, n)$ depends on the class Φ. If, for example, we take Φ to be the class of all linear classifiers of the form

$$\phi(\mathbf{z}) = \begin{cases} 1 & \text{if } a_0 + a_1 z_1 + \cdots + a_{d+p} z_{d+p} > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $a_0, a_1, \ldots, a_{d+p} \in \mathbb{R}$, then $S(\mathcal{A}_\Phi, n) \leq n^{d+p+1}$; see, for example, Devroye et al. (1996) and Pollard (1984).

To state our main results, we shall assume that the chosen kernels are *regular:* a nonnegative kernel K is said to be regular if there are positive constants $b > 0$ and $r > 0$ for which $K(\mathbf{x}) \geq bI\{\mathbf{x} \in S_{0,r}\}$ and $\int \sup_{\mathbf{y} \in \mathbf{x} + S_{0,r}} K(\mathbf{y}) d\mathbf{x} < \infty$, where $S_{0,r}$ is the ball of radius r centered at the origin. For more on this see, for example, Györfi et al. (2002).

Theorem 2 *Let ϕ_B be as in (2) and let $\hat{\phi}_n$ be the classifier that minimizes (8) as h_0 and h_1 vary over sets of the form $H_0 = [0, A_0]$ and $H_1 = [0, A_1]$, where $A_0 > 0$ and $A_1 > 0$ are arbitrary real numbers. Also, let K_0 and K_1 be regular kernels. Then for any distribution of $(\mathbf{Z}, Y)$ and every $\epsilon > 0$, there is an integer $n_0 \equiv n_0(\epsilon) > 0$ such that for all $n > n_0$*

$$P\{L(\hat{\phi}_n) - L(\phi_B) > \epsilon\} \leq 4e^8 E\left[S(\Phi_m, \ell^2)\right] e^{-\ell\epsilon^2/8} + 2e^{-c_1 m\epsilon^2},$$

where c_1 is a positive constant that does not depend on n.

In passing we also note that the bound in Theorem 2, along with the Borel–Cantelli lemma, immediately provides the following strong consistency result:

Corollary 1 *If $\ell^{-1}\log\{E[S(\Phi_m, \ell^2)]\} \to 0$, as $n \to \infty$, then under the conditions of Theorem 2,*

$$L(\hat{\phi}_n) - L(\phi_B) \xrightarrow{a.s.} 0 \quad as \quad n \to \infty.$$

In other words, the error probability of the classifier $\hat{\phi}_n$ converges, with probability one, to that of the optimal classifier.

Some special kernels

In most situations, the term $S(\Phi_m, \ell^2)$, which appears in the bound of Theorem 2, may be difficult to compute, in which case an upper bound on the shatter coefficient may be convenient. Fortunately, we can bound $S(\Phi_m, \ell^2)$ based on the notion of *kernel complexity*. Doing so will allow us to find computable performance bounds for several widely used classes of kernels including the Gaussian kernel. To present such bounds, first note that when $\delta = 1$ (i.e., when the new observation $\mathbf{Z}$ has no missing components), the kernel estimator (6) of the optimal classifier ϕ_B can be written as

$$\phi_{m,1}(\mathbf{z}) := \begin{cases} 1 & \text{if} \sum\limits_{i:(\mathbf{Z}_i, Y_i, \delta_i)\in D_m} (2Y_i - 1)\delta_i K_1\left(\dfrac{\mathbf{Z}_i - \mathbf{z}}{h_1}\right) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

If $\delta = 0$, however, the kernel classifier in (6) becomes

$$\phi_{m,0}(\mathbf{x}) := \begin{cases} 1 & \text{if} \sum\limits_{i:(\mathbf{Z}_i, Y_i, \delta_i)\in D_m} (2Y_i - 1)(1 - \delta_i) K_0\left(\dfrac{\mathbf{X}_i - \mathbf{x}}{h_0}\right) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Next, we borrow the following definitions from Devroye et al. (1996, Chap. 25). Define the quantities $\kappa_m^{(1)}$ and $\kappa_m^{(0)}$ as follows:

$$\kappa_m^{(1)} = \sup_{\mathbf{z},(\mathbf{z}_1,y_1),\ldots,(\mathbf{z}_m,y_m)} \left\{ \text{Number of sign changes of} \right.$$
$$\left. \sum_{i:(\mathbf{z}_i,y_i,1)\in D_m} (2y_i - 1)\delta_i K_1\left(\frac{\mathbf{z}_i - \mathbf{z}}{h_1}\right) \text{ as } h_1 \text{ varies from 0 to infinity} \right\},$$

$$\kappa_m^{(0)} = \sup_{\mathbf{x},(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_m,y_m)} \left\{ \text{Number of sign changes of } \sum_{i:(\mathbf{z}_i,y_i,0)\in D_m} (2y_i-1)(1-\delta_i) K_0\left(\frac{\mathbf{x}_i-\mathbf{x}}{h_0}\right) \text{ as } h_0 \text{ varies from 0 to infinity} \right\}.$$

Finally, define the *kernel complexity* to be

$$\kappa_m^* = \max(\kappa_m^{(1)}, \kappa_m^{(0)}). \tag{12}$$

The kernel complexity of a classifier is closely related to the shatter coefficient of the class Φ_m. To appreciate this, suppose we have a rule ϕ_m with complexity κ_m^*. Then, as h_1 and h_0 vary from 0 to infinity, the binary ℓ-vectors

$$\left(\text{sign}\left[\sum_{i:(\mathbf{Z}_i,Y_i,\delta_i)\in D_m} (2Y_i-1)\delta_i K_1\left(\frac{\mathbf{Z}_i-\mathbf{Z}_j}{h_1}\right)\right]\right)_{j=m+1}^{m+\ell}$$

and

$$\left(\text{sign}\left[\sum_{i:(\mathbf{Z}_i,Y_i,\delta_i)\in D_m} (2Y_i-1)(1-\delta_i) K_0\left(\frac{\mathbf{X}_i-\mathbf{X}_j}{h_0}\right)\right]\right)_{j=m+1}^{m+\ell}$$

change at most $\ell\kappa_m^*$ times each. Therefore, they can take at most $\ell\kappa_m^* + 1$ different values, which implies that

$$S(\Phi_m, \ell) \le \ell\kappa_m^* + 1. \tag{13}$$

The following corollary is an immediate consequence of Theorem 2 and the bound in (13):

Corollary 2 *Suppose the kernel classifier ϕ_m in (6) has complexity κ_m^* as defined above. Then, under the conditions of Theorem 2, one has, for large n,*

$$P\{L(\hat{\phi}_n) - L(\phi_B) > \epsilon\} \le 4e^8(\ell^2\kappa_m^* + 1)e^{-\ell\epsilon^2/8} + 2e^{-c_3 m\epsilon^2}$$

where the positive constant c_3 depends only on the choice of kernels used.

Corollary 2 applies to a broad range of kernels. In what follows, we examine such bounds for two popular classes: the exponential kernels and the polynomial kernels [also, see (Devroye et al. 1996, Sect. 25)].

(i) Exponential kernels
Consider exponential kernels of the form

$$K(\mathbf{u}) = e^{-\|\mathbf{u}\|^\alpha}$$

where $\alpha > 0$, $\mathbf{u} \in \mathbb{R}^s$, and $\|\cdot\|$ is any norm on $\mathbb{R}^s$ (the popular Gaussian kernel falls into this category). If $\delta = 1$ then (10) becomes

$$\phi_{m,1}(\mathbf{z}) = \begin{cases} 1 & \text{if } \sum_{i:(\mathbf{Z}_i, Y_i, \delta_i) \in D_m} (2Y_i - 1)\delta_i e^{-(\|\mathbf{Z}_i - \mathbf{z}\|/h_1)^{\alpha_1}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

for some positive constant α_1. Similarly, when $\delta = 0$, the expression in (11) becomes

$$\phi_{m,0}(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i:(\mathbf{Z}_i, Y_i, \delta_i) \in D_m} (2Y_i - 1)(1 - \delta_i) e^{-(\|\mathbf{X}_i - \mathbf{x}\|/h_0)^{\alpha_0}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

for some positive constant α_0. We can now state the following version of Theorem 2 when K_0 and K_1 are exponential kernels.

Theorem 3 *Let Φ_m be the class of kernel classifiers ϕ_m in (6), where K_0 and K_1 are exponential kernels, as defined above. Then under the conditions of Theorem 2, and for any distribution of $(\mathbf{Z}, Y)$, one has, for large n,*

$$P\{L(\hat{\phi}_n) - L(\phi_B) > \epsilon\} \leq 4e^8(\ell^2 m + 1)e^{-\ell\epsilon^2/8} + 2e^{-c_4 m\epsilon^2}$$

where the positive constant c_4 depends only on the choice of kernels used.

Once again, the above bound (in conjunction with the Borel–Cantelli lemma) yields $L(\hat{\phi}_n) - L(\phi_B) \xrightarrow{a.s.} 0$ as $n \to \infty$, provided that $\ell^{-1}\log(\ell^2 m + 1) \to 0$, as n (and thus $m, \ell) \to \infty$.

(ii) Polynomial kernels

Consider kernels of the form

$$K(\mathbf{z}) = \left(\sum_{i=1}^{t} a_i \|\mathbf{z}\|^{b_i}\right) I\{\|\mathbf{z}\| \leq 1\},$$

where $a_i \in \mathbb{R}$, $b_i \geq 1$, and t is the number of terms in the above sum. When $\delta = 1$, the expression in (10) becomes

$$\phi_{m,1}(\mathbf{z}) = \begin{cases} 1 & \text{if } \sum_{i:(\mathbf{Z}_i, Y_i, \delta_i) \in D_m} (2Y_i - 1)\delta_i \left(\sum_{i=1}^{r_1} a_i \frac{\|\mathbf{Z}_i - \mathbf{z}\|^{b_i}}{{h_1}^{b_i}}\right) I\left\{\frac{\|\mathbf{Z}_i - \mathbf{z}\|}{h_1} \leq 1\right\} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, when $\delta = 0$, we find

$$\phi_{m,0}(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i:(\mathbf{Z}_i, Y_i, \delta_i) \in D_m} (2Y_i - 1)(1 - \delta_i) \left(\sum_{i=1}^{r_0} c_i \frac{\|\mathbf{X}_i - \mathbf{x}\|^{d_i}}{{h_0}^{d_i}}\right) I\left\{\frac{\|\mathbf{X}_i - \mathbf{x}\|}{h_0} \leq 1\right\} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The following performance bound holds for the class of polynomial kernels:

Theorem 4 *Let Φ_m be the class of kernel classifiers ϕ_m in (6), where K_0 and K_1 are polynomial kernels, as defined above. Then under the conditions of Theorem 2, and for any distribution of $(\mathbf{Z}, Y)$, one has, for large n,*

$$P\{L(\hat{\phi}_n) - L(\phi_B) > \epsilon\} \leq 4e^8(\ell^2 rm + 1)e^{-\ell\epsilon^2/8} + 2e^{c_5 m\epsilon^2},$$

where $r = \max(r_0, r_1)$ and where the positive constant c_5 depends only on the choice of kernels used.

Once again, the strong consistency of $\hat{\phi}_n$ follows from the above theorem, provided that $\ell^{-1}\log(\ell^2 rm + 1) \to 0$, as $n \to \infty$.

2.3 The resubstitution method

In the previous section, we considered methods based on data-splitting to find a data-driven value of the kernel bandwidth that would yield strongly consistent classifiers. One problem with this approach, however, is that it is not always clear as to how one should choose the splitting ratio m/n (i.e., what m should be). In this section, we propose to consider the alternative approach (for choosing the bandwidth) based on the *resubstitution method*. Let Φ_n be the collection of classifiers $\phi_n \equiv \phi_{n,h_1,h_0}$ defined via (4) and (5), as h_1 and h_0 vary over sets of positive real numbers (possibly infinite sets). Then for any $\phi_n \in \Phi_n$, the resubstitution estimate of the error of ϕ_n is simply

$$\begin{aligned}\widehat{L}_n^{(R)}(\phi_n) &:= \frac{1}{n}\sum_{i=1}^{n} I\{\phi_n(\mathbf{Z}_i, \delta_i) \neq Y_i\} \\ &= \frac{1}{n}\sum_{i=1}^{n} I\left\{I\left\{\delta_i \widehat{\tau}_{h_1}(\mathbf{Z}_i) + (1-\delta_i)\widehat{\tau}_{h_0}(\mathbf{X}_i) > \frac{1}{2}\right\} \neq Y_i\right\}, \end{aligned} \tag{14}$$

where $\widehat{\tau}_{h_1} := \widehat{\tau}_1$ and $\widehat{\tau}_{h_0} := \widehat{\tau}_0$ are as in (4). In other words, the resubstitution estimates of (h_1, h_0), which we shall denote by $(\tilde{h}_1, \tilde{h}_0)$, satisfy

$$(\tilde{h}_1, \tilde{h}_0) = \text{argmin}_{\phi_{n,h_1,h_0}\in\Phi_n} \widehat{L}_n^{(R)}(\phi_{n,h_1,h_0}). \tag{15}$$

Note that the data has been used twice here: once to construct ϕ_n and a second time to estimate the error of ϕ_n (error committed on the same data). Let $\tilde{\phi}_n \in \Phi_n$ be the classifier corresponding to $(\tilde{h}_1, \tilde{h}_0)$, i.e., $\widehat{L}_n^{(R)}(\tilde{\phi}_n) \leq \widehat{L}_n^{(R)}(\phi_n)$ for all $\phi_n \in \Phi_n$. Also, let

$$L(\phi_n) = P\{\phi_n(\mathbf{Z}, \delta) \neq Y | D_n\} \quad \text{and} \quad L(\tilde{\phi}_n) = P\{\tilde{\phi}_n(\mathbf{Z}, \delta) \neq Y | D_n\}$$

denote the error probabilities of ϕ_n and $\tilde{\phi}_n$, respectively. The following theorem establishes strong consistency of the resubstitution method in the case where the components of the random vector $\mathbf{Z}$ are discrete:

Theorem 5 *Suppose that the kernels $K_1 \geq 0$ and $K_0 \geq 0$ satisfy the conditions $K_1(z) \to 0$ as $\|z\| \to \infty$ and $K_0(x) \to 0$ as $\|x\| \to \infty$. Let $\tilde{\phi}_n$ be the classifier defined by (4) and (5) with $\tilde{h}_1$ and $\tilde{h}_0$ selected via (15). Then $L(\tilde{\phi}_n) \to L(\phi_B)$, with probability one, whenever the random vector $\mathbf{Z}$ has a discrete distribution.*

Remark 2 While consistency is guaranteed when the random vector $\mathbf{Z}$ is discrete, Theorem 5 makes no conclusions about the performance of the procedure when $\mathbf{Z}$ is continuous. It is true that the resubstitution procedure can be inconsistent when the random vector $\mathbf{Z}$ is not discrete [see (Devroye et al. 1996, Sect. 25.6) for a counter-example], yet there may be cases where consistency can be achieved.

3 Numerical examples

In this section, we carry out several numerical studies to assess the performance of our proposed classification methods.

Example 1 Consider the class membership $Y = 0$ or $Y = 1$, of an entity based on the covariates $\mathbf{Z} = (X, V)'$, where $\mathbf{Z} \sim N_2(\boldsymbol{\mu}_0, \Sigma_0)$ if $Y = 0$ and $\mathbf{Z} \sim N_2(\boldsymbol{\mu}_1, \Sigma_1)$ otherwise. The unconditional class probabilities were taken to be $P\{Y = 1\} = P\{Y = 0\} = 0.5$. The parameters were chosen as follows:

$$\boldsymbol{\mu}_0 = (0.7, 0)', \quad \Sigma_0 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 2 \end{pmatrix},$$

$$\boldsymbol{\mu}_1 = (1, 1)', \quad \Sigma_1 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}.$$

Here X is observable but V could be missing. The missingness probability mechanism was taken to be

$$\begin{aligned} p(\mathbf{z}, y) = p(x, v, y) &:= P\{\delta = 1 | X = x, V = v, Y = y\} \\ &= \exp\{-a(1 - 0.6y)(x - 1)^2 - b(1 + 0.6y)v^2 - cy\}, \end{aligned} \tag{16}$$

where $a, b, c > 0$ are constants. We considered three different choices for (a, b, c): $(0.5, 0.5, 0)$, $(0.45, 0, 1.3)$, and $(0, 0, 0)$. The choice $(0, 0, 0)$ corresponds to the case of no missing data. We considered two sample sizes: $n = 100$ and $n = 300$. As for the choice of the kernels, $K_1 : \mathbb{R}^2 \to \mathbb{R}$ and $K_0 : \mathbb{R} \to \mathbb{R}$ were taken to be standard Gaussian and the bandwidths h_1 and h_0 were chosen from grids of 10 equally spaced values in $[0.05, 1.2]$ and $[0.08, 1.2]$, respectively. Here we employed four different methods to estimate the smoothing parameters h_1 and h_0 of the kernel classifiers: (i) The data-splitting procedure based on (6) and (7) with a splitting ratio of 65%, i.e., $m = 0.65n$ and $\ell = 0.35n$. (ii) Breiman's *out-of-bag* procedure (Breiman 1996). (iii) The resubstitution method which was explained in Sect. 2. (iv) The method that selects h_1 and h_0 as the minimizers of the asymptotic mean integrated squared error (AMISE) of the corresponding kernel density estimator, as given in Wand and Jones (1995). Further properties of such procedures are discussed in Hall and Marron (1987).

Table 1 Error rates for $\hat{\phi}_n$ (data-splitting), $\tilde{\phi}_n$ (resubstitution), ϕ_n^G (out-of-bag), and ϕ_n^D (density) when $n = 100$ for Example 1

a	b	c	Error($\hat{\phi}_n$)	Error($\tilde{\phi}_n$)	Error(ϕ_n^G)	Error(ϕ_n^D)
0	0	0	.374	.361	.382	.384
			(.0017)	(.0015)	(.0017)	(.0013)
0.5	0.5	0	.456	.450	.458	.448
			(.0013)	(.0013)	(.0013)	(.0012)
0.45	0	1.3	.279	.272	.281	.278
			(.0012)	(.0011)	(.0012)	(.0010)

See also Bontemps et al. (2009) for results pertaining to the selection of bandwidths for kernel based conditional density estimates. In passing, we also note that the optimal choice of the smoothing parameters in kernel density estimation is not necessarily optimal in the problem of kernel classification. For the out-of-bag method, we first generated a bootstrap training sample of size n from the original sample (that is, a sample of size n drawn with replacement from the original sample). The remaining values (i.e. observations not appearing in the bootstrap sample) were then used as our testing sequence. The bootstrap sample is employed to construct the family Φ_m of classifiers of the form (6), whereas the testing sequence is used to choose the empirically best classifier from Φ_m.

Finally, to assess the error rates of these four classifiers, we also generated an additional 1000 observations, (the same way we generated the data), to be used as our test sample. The entire above process was repeated a total of 500 times and the average misclassification error probability estimates, committed on the testing sequences over 500 such training and testing samples, were calculated. The results appear in Table 1 for the case of $n = 100$. The error rates that are reported are averages over the 500 runs and the numbers in parentheses are the standard errors of those averages.

As Table 1 shows, the proposed resubstitution method appears to outperform the other three procedures for most values of the constants a, b, and c. These results further validate our earlier remarks about the performance of this method when the random vector $\mathbf{Z}$ is continuous. It is also worth noting that the approach based on density estimation performs nearly as well as the other procedures, at least in the case of missing data. These observations are consistent with the results in Hall and Kang (2005), which demonstrate that in the multivariate setting and under suitable conditions, the bandwidths selected to minimize the mean square error between the kernel density estimators and the true densities are on the same order of magnitude as the optimal bandwidths for classification.

The results for the case where $n = 300$ are given in Table 2. Once again, it is clear that the resubstitution method is superior (followed by the data-splitting classifier $\hat{\phi}_n$).

Tables 1 and 2 also show that, for some choices of the constants a, b, and c, the misclassification error of $\hat{\phi}_n$ can be less than that of the case with no missing covariates. See, for example, the third row of Table 2, where $a = 0.45, b = 0$ and $c = 1.3$, in which case the error of $\hat{\phi}_n$ is $0.263 < 0.344$. This illustrates, somewhat counter-

Table 2 Error rates for $\hat{\phi}_n$ (data-splitting), $\tilde{\phi}_n$ (resubstitution), ϕ_n^G (out-of-bag), and ϕ_n^D (density) when $n = 300$ for Example 1

a	b	c	Error($\hat{\phi}_n$)	Error($\tilde{\phi}_n$)	Error(ϕ_n^G)	Error(ϕ_n^D)
0	0	0	.344	.338	.348	.371
			(.0010)	(.0009)	(.0012)	(.0009)
0.5	0.5	0	.434	.429	.438	.433
			(.0011)	(.0011)	(.0012)	(.0009)
0.45	0	1.3	.263	.257	.265	.269
			(.0009)	(.0008)	(.0009)	(.0007)

intuitively, that classification with missing covariates can sometimes have a lower misclassification error than the case with no missing covariates. This phenomenon, which is also noted in Mojirsheibani and Montazeri (2007), is partially explained by the relationship between the correlation of Y and V, and that of Y and δ. To appreciate this, note that when $a = b = c = 0$, the correlation between V and Y is 0.378. When $a = 0.45, b = 0$, and $c = 1.3$, however, the correlation between δ and Y is -0.476. The fact that $|-0.476| > 0.378$ implies that the random variable δ, which is always observable, can sometimes do better (than the covariate V) at predicting Y.

Example 2 In this example, we consider both continuous and discrete covariates for predicting the class Y. More specifically, we consider covariate vectors of the form $\mathbf{Z} = (X_1, X_2, V)'$, where $X_1 = Z_1$ and $X_2 = I\{|Z_2| < 2\}$, i.e., X_2 is a discrete covariate. When $Y = 0$, $(Z_1, Z_2, Z_3)' \sim N_3(\boldsymbol{\mu}_0, \Sigma_0)$, where

$$\boldsymbol{\mu}_0 = (0.7, 0.7, 0.7)', \quad \Sigma_0 = \begin{pmatrix} 1 & 0.4 & 0.16 \\ 0.4 & 1 & 0.4 \\ 0.16 & 0.4 & 1 \end{pmatrix}.$$

When $Y = 1$, the vector $(Z_1, Z_2, Z_3)'$ has a standard Cauchy distribution with independent components, i.e., Z_j, $j = 1, 2, 3$ are independent standard Cauchy random variables. The missingness probability mechanism was taken to be

$$\begin{aligned} p(\mathbf{z}, y) = p(x_1, x_2, v, y) &:= P\{\delta = 1 | X_1 = x_1, X_2 = x_2, V = v, Y = y\} \\ &= \exp\{-a(1 - 0.6y)(x_1 - 0.7)^2 - b(1 - 0.4y)(v - 0.5)^2 \\ &\quad - c(1 + 0.6y)^2(x_2 - 0.5)^2 - dy\}, \end{aligned} \tag{17}$$

where $a, b, c, d > 0$ are constants. We considered three different choices for (a, b, c, d), namely $(0.15, 0, 0.5, 1.2)$, $(0.25, 0.5, 0.25, 0.6)$, and $(0, 0, 0, 0)$. Note that the choice $(0, 0, 0, 0)$ corresponds to no missing data. The kernels, bandwidths, and sample sizes are as in Example 1. Once again, we considered data-splitting, resubstitution, the out-of-bag, and the procedure based on density estimation to select a kernel classifier from a class indexed by values of the bandwidths h_1 and h_0. The

Table 3 Error rates for $\hat{\phi}_n$ (data-splitting), $\tilde{\phi}_n$ (resubstitution), ϕ_n^G (out-of-bag), and ϕ_n^D (density) when $n = 100$ for Example 2

a	b	c	d	Error($\hat{\phi}_n$)	Error($\tilde{\phi}_n$)	Error(ϕ_n^G)	Error(ϕ_n^D)
0	0	0	0	.301	.284	.305	.301
				(.0017)	(.0014)	(.0016)	(.0020)
0.15	0	0.5	1.2	.243	.229	.246	.260
				(.0013)	(.0011)	(.0013)	(.0013)
0.25	0.5	0.25	0.6	.357	.345	.364	.365
				(.0014)	(.0013)	(.0016)	(.0016)

Table 4 Error rates for $\hat{\phi}_n$ (data-splitting), $\tilde{\phi}_n$ (resubstitution), ϕ_n^G (out-of-bag), and ϕ_n^D (density) when $n = 300$ for Example 2

a	b	c	d	Error($\hat{\phi}_n$)	Error($\tilde{\phi}_n$)	Error(ϕ_n^G)	Error(ϕ_n^D)
0	0	0	0	.264	.254	.269	.278
				(.0011)	(.0010)	(.0011)	(.0019)
0.15	0	0.5	1.2	.213	.205	.214	.233
				(.0009)	(.0008)	(.0008)	(.0010)
0.25	0.5	0.25	0.6	.325	.317	.329	.344
				(.0010)	(.0009)	(.0010)	(.0011)

results appear in Table 3 for the case of $n = 100$ and Table 4 for $n = 300$ (see Example 1 for more details on how these values were calculated).

The results show that the proposed resubstitution method outperforms all of the other procedures for every choice of the constants a, b, c, and d appearing in (17), and for both samples of size $n = 100$ and $n = 300$. Once again, as noted in Mojirsheibani and Montazeri (2007), we see that classification with missing covariates can sometimes perform better than the case with no missing covariates; see, for example, row 2 of Table 4 (corresponding to $a = 0.15, b = 0, c = 0.5, d = 1.2$). This is explained in part by the fact that the correlation between δ and Y is -0.37, whereas the correlation between V and Y in row 1 (corresponding to no missing data) is only 0.04.

Example 3 (Mammogram data)

We now turn to a real data example involving the classification of mammographic masses in the screening for breast cancer; there are many discrete covariates in this data set. The data set consists of 961 patients, 516 of whom have mammographic masses which are benign (class 0), and the remaining 445 patients have mammographic masses which are malignant (class 1). The covariates used to predict the class a patient belongs to are $x_1 =$ 'patient's age' (in years), $x_2 =$ 'mass shape' (nominal value $\in \{1, 2, 3, 4\}$), $x_3 =$ 'mass margin' (nominal value $\in \{1, 2, 3, 4, 5\}$), and $v =$ 'mass density' (nominal value $\in \{1, 2, 3, 4\}$). A full description of this data set is available from the University of California, Irvine, repository of machine learning database at http://archive.ics.uci.edu/ml/. In this example, we focus on one dominant missingness pattern: for 56

Table 5 Error rates for $\hat{\phi}_n$ (data-splitting), $\tilde{\phi}_n$ (resubstitution), ϕ_n^G (out-of-bag), and ϕ_n^D (density) for Example 3

Error($\hat{\phi}_n$)	Error($\tilde{\phi}_n$)	Error(ϕ_n^G)	Error(ϕ_n^D)
0.148	0.0800	0.156	0.0821
(0.0013)	(0.0008)	(0.0012)	(0.0008)

patients, the value of v = 'mass density' was missing. To better present our results, we consider only this particular missingness pattern.

For each of the four procedures mentioned in the previous examples, a kernel classifier was constructed using two-thirds of the data (randomly selected), and the performance of the chosen classifiers was tested on the remaining portion (the kernels were taken to be as in Example 1). This entire process was repeated 500 times, producing the results in Table 5. It is interesting to note that the resubstitution method and the procedure based on density estimation perform quite similarly, with estimated error rates that are roughly half of those of data-splitting and the out-of-bag method. In this example, all of the covariates used to predict Y were discrete, which might partly explain the superior performance of the resubstitution procedure.

Remark 3 Although we have stated our results for a two-class classification problem, our results can readily be extended to the more general M-class setup, where $M \geq 2$. More specifically, if we put

$$\phi_B(\mathbf{z}, \delta) = \delta\psi_1(\mathbf{z}) + (1-\delta)\psi_0(\mathbf{x}),$$

where for $1 \leq j \leq M$

$$\psi_1(\mathbf{z}) = j \ \text{ if } \ E[\delta I\{Y=j\}|\mathbf{Z}=\mathbf{z}] = \max_{1\leq j\leq M} E[\delta I\{Y=j\}|\mathbf{Z}=\mathbf{z}]$$
$$\psi_0(\mathbf{x}) = j \ \text{ if } \ E[(1-\delta) I\{Y=j\}|\mathbf{X}=\mathbf{x}] = \max_{1\leq j\leq M} E[(1-\delta) I\{Y=j\}|\mathbf{X}=\mathbf{x}],$$

then it follows from Mojirsheibani and Montazeri (2007) that ϕ_B is indeed the optimal classifier. Now, for $j = 1 \ldots M$, define $\hat{\psi}_{1,m}(\mathbf{z}) = j$ if

$$\sum_{i:(\mathbf{Z}_i, Y_i, \delta_i)\in D_m} \delta_i I\{Y_i = j\} K_1\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h_1}\right)$$
$$= \max_{1\leq j\leq M} \sum_{i:(\mathbf{Z}_i, Y_i, \delta_i)\in D_m} \delta_i I\{Y_i = j\} K_1\left(\frac{\mathbf{Z}_i - \mathbf{z}}{h_1}\right).$$

Also, for $j = 1 \ldots M$, define $\hat{\psi}_{0,m}(\mathbf{x}) = j$ if

$$\sum_{i:(\mathbf{X}_i, Y_i, \delta_i)\in D_m} (1-\delta_i) I\{Y_i = j\} K_0\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_1}\right)$$
$$= \max_{1\leq j\leq M} \sum_{i:(\mathbf{X}_i, Y_i, \delta_i)\in D_m} (1-\delta_i) I\{Y_i = j\} K_0\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_1}\right),$$

where K_1, K_0, h_1, and h_0 are as in Sect. 2, and put

$$\psi_m(\mathbf{z}, \delta) = \delta \hat{\psi}_{1,m}(\mathbf{z}) + (1-\delta)\hat{\psi}_{0,m}(\mathbf{x}).$$

Finally, letting h_1 and h_0 vary over a set of prescribed values, the optimal classifier $\hat{\psi}_n$ is the classifier that minimizes the empirical error committed on the testing sequence T_ℓ, i.e., minimizing $\ell^{-1} \sum_{i:(\mathbf{Z}_i, Y_i, \delta_i) \in T_\ell} I\{\psi_m(\mathbf{Z}_i, \delta_i) \neq Y_i\}$.

Acknowledgements This research was supported by a grant from the Interdisciplinary Research Institute of the Sciences at California State University Northridge of Levon Demirdjian and the NSF Grant DMS-1407400 of Majid Mojirsheibani.

Appendix: Proofs

In order to prove our main results, we first state a number of technical lemmas.

Lemma 1 *Define* $T_1(z) = E[\delta(2Y-1)|\mathbf{Z}=z]$ *and* $T_0(\boldsymbol{x}) = E[(1-\delta)(2Y-1)|\boldsymbol{X}=\boldsymbol{x}]$ *and let* $\widehat{T}_1$ *and* $\widehat{T}_0$ *be any approximations to* T_1 *and* T_0*, based on the sample* D_m*. Also, put* $\hat{\phi}_1(z) = I\{\widehat{T}_1(z) > 0\}$ *and* $\hat{\phi}_0(\boldsymbol{x}) = I\{\widehat{T}_0(\boldsymbol{x}) > 0\}$*. Then the classifier*

$$\hat{\phi}(\mathbf{Z}, \delta) = \delta\hat{\phi}_1(\mathbf{Z}) + (1-\delta)\hat{\phi}_0(\boldsymbol{X})$$

satisfies

$$L(\hat{\phi}) - L(\phi_B) \leq E\left(\left|E\left[\delta(2Y-1)|\mathbf{Z}\right] - \widehat{T}_1(\mathbf{Z})\right| \middle| D_m\right) + E\left(\left|E\left[(1-\delta)(2Y-1)|\boldsymbol{X}\right] - \widehat{T}_0(\boldsymbol{X})\right| \middle| D_m\right), \text{ a.s.}$$

The proof of Lemma 1 will be deferred to the end of this section.

The next lemma provides exponential performance bounds for kernel regression estimates.

Lemma 2 (Devroye et al. (1996) and Györfi et al. (2002)). *Let* $D_n = \{(U_1, \boldsymbol{V}_1), \ldots, (U_n, \boldsymbol{V}_n)\}$ *be iid* $[-L, L] \times \mathbb{R}^d$*-valued random vectors. Put* $m(\boldsymbol{v}) = E(U|\boldsymbol{V}=\boldsymbol{v})$ *and define*

$$m_n(\boldsymbol{v}) = \sum_{i=1}^{n} U_i K\left(\frac{\boldsymbol{V}_i - \boldsymbol{v}}{h_n}\right) \Big/ nE\left\{K\left(\frac{\boldsymbol{V}-\boldsymbol{v}}{h_n}\right)\right\},$$

where the kernel $K(\cdot)$ *is regular. If* $h_n \to 0$ *and* $nh_n^d \to \infty$ *as* $n \to \infty$*, then for every* $\epsilon > 0$ *and any distribution of* $(U, \boldsymbol{V})$*, and* n *large enough,*

$$P\{E[|m_n(\boldsymbol{V}) - m(\boldsymbol{V})||D_n] > \epsilon\} \leq \exp\left(-\frac{n\epsilon^2}{64\rho^2 L}\right),$$

where $\rho = \rho(K)$ is a positive constant that depends on K only.

Next, we state a result from the empirical process theory. Given an iid sample $D_n = \{(\mathbf{Z}_1, Y_1), \ldots, (\mathbf{Z}_n, Y_n)\}$, let ν_n be the empirical measure of the set A, i.e.,

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I\{(\mathbf{Z}_i, Y_i) \in A\}.$$

Lemma 3 (Devroye (1982) and Massart (1990)). *Let ν be the probability measure of $(\mathbf{Z}, Y)$ on $\mathbb{R}^{d+p} \times \{0, 1\}$ and let ν_n be the empirical measure based on D_n. If $\mathcal{A}$ is a collection of measurable sets, then $\forall \epsilon \leq 1$,*

$$P\left\{\sup_{A\in\mathcal{A}} \left|\nu_n(A) - \nu(A)\right| > \epsilon\right\} \leq c_2 S(\mathcal{A}, n^2) e^{-2n\epsilon^2}$$

where the constant c_2 is positive, does not depend on n, and does not exceed $4e^{4\epsilon+4\epsilon^2} \leq 4e^8$, and $S(\mathcal{A}, n^2)$ is the $(n^2)^{th}$ shatter coefficient of the class $\mathcal{A}$.

Lemma 4 (Devroye et al. (1996)). *Let $a_1, \ldots, a_m$ be fixed real numbers, and let $b_1, \ldots, b_m$ be different nonnegative reals. If $\alpha \neq 0$, then the function*

$$g(x) = \sum_{i=1}^{m} a_i e^{-b_i x^{\alpha}}, \quad x \geq 0$$

is either identically zero, or takes the value 0 at most m times.

Proof of Theorem 1

The proof is based on standard arguments. First, note that

$$\begin{aligned} L(\hat{\phi}_n) - \inf_{\phi_m\in\Phi_m} L(\phi_m) &= \left(L(\hat{\phi}_n) - \widehat{L}_{m,\ell}(\hat{\phi}_n)\right) + \left(\widehat{L}_{m,\ell}(\hat{\phi}_n) - \inf_{\phi_m\in\Phi_m} L(\phi_m)\right) \\ &\leq \sup_{\phi_m\in\Phi_m} |\widehat{L}_{m,\ell}(\phi_m) - L(\phi_m)| + \left(\widehat{L}_{m,\ell}(\hat{\phi}_n) - \inf_{\phi_m\in\Phi_m} L(\phi_m)\right) \\ &\leq 2 \sup_{\phi_m\in\Phi_m} |\widehat{L}_{m,\ell}(\phi_m) - L(\phi_m)|. \end{aligned}$$

Therefore,

$$\begin{aligned} P\left\{L(\hat{\phi}_n) - \inf_{\phi_m\in\Phi_m} L(\phi_m) > \epsilon \middle| D_m\right\} &\leq P\left\{2 \sup_{\phi_m\in\Phi_m} \left|\widehat{L}_{m,\ell}(\phi_m) - L(\phi_m)\right| > \epsilon \middle| D_m\right\} \\ &= P\left\{\sup_{\phi_m\in\Phi_m} \left|\widehat{L}_{m,\ell}(\phi_m) - L(\phi_m)\right| > \epsilon/2 \middle| D_m\right\} \end{aligned}$$

$$
\begin{aligned}
&\leq |\Phi_m| \max_{\phi_m \in \Phi_m} P\{|\widehat{L}_{m,\ell}(\phi_m) - L(\phi_m)| > \epsilon/2 | D_m\} \qquad \text{(by the union bound)} \\
&\leq 2|\Phi_m| e^{-2\ell\epsilon^2/2^2}, \ \text{(by Hoeffding's inequality)} \\
&= 2N_1 N_0 e^{-\ell\epsilon^2/2}
\end{aligned}
$$

The result follows by taking expectation of both sides with respect to the distribution of D_m. □

Proof of Theorem 2

Start by writing

$$
\begin{aligned}
L(\hat{\phi}_n) - L(\phi_B) &= \left[L(\hat{\phi}_n) - \inf_{\phi_m \in \Phi_m} L(\phi_m)\right] + \left[\inf_{\phi_m \in \Phi_m} L(\phi_m) - L(\phi_B)\right] \\
&:= \mathrm{I}_n + \mathrm{II}_n, \ \text{(say)}.
\end{aligned}
$$

Employing the arguments used in the proof of Theorem 1, we find

$$
\mathrm{I}_n \leq 2 \sup_{\phi_m \in \Phi_m} \left|\widehat{L}_{m,\ell}(\phi_m) - L(\phi_m)\right|. \tag{18}
$$

To deal with the term II_n, let $h_0 \in H_0$ and $h_1 \in H_1$ be given and define

$$
\phi_{m,h_1}(\mathbf{z}) = \begin{cases} 1 & \text{if } \hat{\tau}_{1,m}(\mathbf{z}) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \phi_{m,h_0}(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{\tau}_{0,m}(\mathbf{x}) > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases}
$$

where $\hat{\tau}_{1,m}(\mathbf{z})$ and $\hat{\tau}_{0,m}(\mathbf{x})$ are as in (7). Now, observe that the kernel classifier in (6) can be written as

$$
\phi_m(\mathbf{Z}, \delta) = \delta \phi_{m,h_1}(\mathbf{Z}) + (1 - \delta)\phi_{m,h_0}(\mathbf{X}). \tag{19}
$$

Furthermore, it is a simple exercise to show that the functions $\phi_{m,h_1}(\mathbf{z})$ and $\phi_{m,h_0}(\mathbf{x})$ can equivalently be written as

$$
\phi_{m,h_1}(\mathbf{z}) = \begin{cases} 1 & \text{if } \widehat{T}_1(\mathbf{z}) > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \phi_{m,h_0}(\mathbf{x}) = \begin{cases} 1 & \text{if } \widehat{T}_0(\mathbf{x}) > 0 \\ 0 & \text{otherwise} \end{cases}
$$

respectively, where

$$
\widehat{T}_1(\mathbf{z}) = \frac{\sum_{i:(\mathbf{Z}_i, Y_i, \delta_i) \in D_m} \delta_i (2Y_i - 1) K_1\left(\dfrac{\mathbf{Z}_i - \mathbf{z}}{h_1}\right)}{m E\left[K_1\left(\dfrac{\mathbf{Z} - \mathbf{z}}{h_1}\right)\right]} \tag{20}
$$

$$\text{and} \quad \widehat{T}_0(\mathbf{x}) = \frac{\sum_{i:(\mathbf{Z}_i, Y_i, \delta_i) \in D_m} (1-\delta_i)(2Y_i - 1) K_0\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_0}\right)}{m E\left[K_0\left(\frac{\mathbf{X} - \mathbf{x}}{h_0}\right)\right]}. \tag{21}$$

Now let $\tilde{h}_1 \equiv \tilde{h}_1(m)$ and $\tilde{h}_0 \equiv \tilde{h}_0(m)$ be sequences in H_1 and H_0, respectively, such that, as $m \to \infty$, one has $\tilde{h}_1 \to 0$, $\tilde{h}_0 \to 0$, $m\tilde{h}_1^{p+d} \to \infty$, and $m\tilde{h}_0^d \to \infty$. [For example, one can take $\tilde{h}_0 = \min(A_0, m^{-\frac{1}{d+c_0}})$, $c_0 > 0$, and $\tilde{h}_1 = \min(A_1, m^{-\frac{1}{d+p+c_1}})$, $c_1 > 0$.] Also, let $\tilde{\phi}_m \in \Phi_m$ be the classifier corresponding to $\tilde{h}_1$ and $\tilde{h}_0$. In view of (19), (20), and (21), one has

$$\tilde{\phi}_m(\mathbf{Z}, \delta) = \delta \phi_{m,\tilde{h}_1}(\mathbf{Z}) + (1-\delta)\phi_{m,\tilde{h}_0}(\mathbf{X}),$$

with

$$\phi_{m,\tilde{h}_1}(\mathbf{z}) = I\{\widehat{T}_1(\mathbf{z}) > 0\}, \qquad \phi_{m,\tilde{h}_0}(\mathbf{x}) = I\{\widehat{T}_0(\mathbf{x}) > 0\},$$

and where $\widehat{T}_1$ and $\widehat{T}_0$ are as in (20) and (21) with h_1 and h_0 replaced by $\tilde{h}_1$ and $\tilde{h}_0$. Then, by Lemma 1,

$$\begin{aligned}
\text{II}_n &= \inf_{\phi_m \in \Phi_m} L(\phi_m) - L(\phi_B) \le L(\tilde{\phi}_m) - L(\phi_B) \\
&\le E\left(\left|E\left[\delta(2Y-1)|\mathbf{Z}\right] - \widehat{T}_1(\mathbf{Z})\right| \middle| D_m\right) \\
&\quad + E\left(\left|E\left[(1-\delta)(2Y-1)|\mathbf{X}\right] - \widehat{T}_0(\mathbf{X})\right| \middle| D_m\right), \text{ a.s.},
\end{aligned}$$

where $\widehat{T}_1$ and $\widehat{T}_0$ are as in (20) and (21). Putting all the above together, we have

$$\begin{aligned}
&P\{L(\hat{\phi}_n) - L(\phi_B) > \epsilon\} \\
&= P\left\{L(\hat{\phi}_n) - \inf_{\phi_m \in \Phi_m} L(\phi_m) + \inf_{\phi_m \in \Phi_m} L(\phi_m) - L(\phi_B) > \epsilon\right\} \\
&\le P\left\{2 \sup_{\phi_m \in \Phi_m} \left|\widehat{L}_{m,\ell}(\phi_m) - L(\phi_m)\right| > \frac{\epsilon}{2}\right\} \\
&\quad + P\left\{E\left(\left|E\left[\delta(2Y-1)|\mathbf{Z}\right] - \widehat{T}_1(\mathbf{Z})\right| \middle| D_m\right) > \frac{\epsilon}{4}\right\} \\
&\quad + P\left\{E\left(\left|E\left[(1-\delta)(2Y-1)|\mathbf{X}\right] - \widehat{T}_0(\mathbf{X})\right| \middle| D_m\right) > \frac{\epsilon}{4}\right\} \\
&:= P_{n,1} + P_{n,2} + P_{n,3}.
\end{aligned}$$

Now,

$$P_{n,1} = E\left[P\left\{\sup_{\phi_m\in\Phi_m}\left|\widehat{L}_{m,\ell}(\phi_m) - L(\phi_m)\right| > \frac{\epsilon}{4}\middle| D_m\right\}\right]$$
$$\leq 4e^8 E\left[S(\Phi_m, \ell^2)\right] e^{-\ell\epsilon^2/8} \quad \text{(by Lemma 3)}\,.$$

Furthermore, by Lemma 2

$$P_{n,2} + P_{n,3} \leq \exp\left\{-\frac{m\epsilon^2}{(64)(16)\rho_1^2}\right\} + \exp\left\{-\frac{m\epsilon^2}{(64)(16)\rho_0^2}\right\}$$
$$\leq 2\exp\left\{-\frac{m\epsilon^2}{(64)(16)(\rho_1^2\vee\rho_0^2)}\right\},$$

where $\rho_j \equiv \rho(K_j)$, $j = 0, 1$, is a positive constant that depends on K_j only. This completes the proof of Theorem 2. □

Proof of Theorem 3

Theorem 3 follows immediately from Corollary 2 and Lemma 4.

Proof of Theorem 4

The arguments in Devroye et al. (1996, Sec 25.3) can be used to show that $\kappa_m^{(1)} \leq r_1 m$ and $\kappa_m^{(0)} \leq r_0 m$, where $\kappa_m^{(1)}$ and $\kappa_m^{(0)}$ are as in (12). Therefore, $\kappa_m^* \leq \max(r_1 m, r_0 m) = rm$. The result now follows from an application of Corollary 2.

Proof of Theorem 5

The following proof employs some of the arguments used in Devroye et al. (1996, Chap. 25). Fix $D_n = \{(\mathbf{Z}_1, \delta_1, Y_1), \ldots, (\mathbf{Z}_n, \delta_n, Y_n)\}$ and let Φ_n be the class of all kernel rules in (5). Denote a typical rule in Φ_n as $\phi_n := \phi_{n,h_1,h_0}$ and let $\mathcal{A}_n$ be the class of all sets of the form

$$A_n = \left\{(\mathbf{z}, \delta, y)\middle| I\left\{\delta\widehat{\tau}_1(\mathbf{z}) + (1-\delta)\widehat{\tau}_0(\mathbf{x}) > \frac{1}{2}\right\} \neq y\right\}$$
$$:= \left\{(\mathbf{z}, \delta, y)\middle| I\left\{G(\mathbf{z}, \delta, D_n) > \frac{1}{2}\right\} \neq y\right\},$$

where $G(\mathbf{z}, \delta, D_n) := \delta\widehat{\tau}_1(\mathbf{z}) + (1-\delta)\widehat{\tau}_0(\mathbf{x})$, and $\widehat{\tau}_1$ and $\widehat{\tau}_0$ are defined as in (4). Similarly, define $\mathcal{A}_{n,1}$ and $\mathcal{A}_{n,0}$ to be classes of sets of the form

$$A_{n,1} = \left\{(\mathbf{z}, \delta, y)\middle| I\left\{G(\mathbf{z}, \delta, D_n) > \frac{1}{2}\right\} = 1, y = 0\right\}$$
$$\text{and} \quad A_{n,0} = \left\{(\mathbf{z}, \delta, y)\middle| I\left\{G(\mathbf{z}, \delta, D_n) > \frac{1}{2}\right\} = 0, y = 1\right\}$$

respectively. Note that any set $A_n \in \mathcal{A}_n$ can be written as

$$\begin{aligned} A_n &= \left\{(\mathbf{z}, \delta, y) \middle| I\left\{G(\mathbf{z}, \delta, D_n) > \frac{1}{2}\right\} = 1, y = 0\right\} \\ &\quad \cup \left\{(\mathbf{z}, \delta, y) \middle| I\left\{G(\mathbf{z}, \delta, D_n) > \frac{1}{2}\right\} = 0, y = 1\right\} \\ &= A_{n,1} \cup A_{n,0}, \end{aligned}$$

for some $A_{n,1} \in \mathcal{A}_{n,1}$ and $A_{n,0} \in \mathcal{A}_{n,0}$. Define

$$\begin{aligned} \nu_n(A_n) &= P\{(\mathbf{Z}, \delta, Y) \in A_n | D_n\} \ \ (\text{conditioned on } D_n \text{ since } A_n \text{ depends on } D_n) \\ &= E\left[I\{(\mathbf{Z}, \delta, Y) \in A_n\} \middle| D_n\right] \\ &= E\left[I\left\{(\mathbf{Z}, \delta, Y) \in A_{n,1} \cup A_{n,0}\right\} \middle| D_n\right] \\ &= E\left[I\left\{(\mathbf{Z}, \delta, Y) \in A_{n,1}\right\} \middle| D_n\right] + E\left[I\left\{(\mathbf{Z}, \delta, Y) \in A_{n,0}\right\} \middle| D_n\right] \\ &= P\{(\mathbf{Z}, \delta, Y) \in A_{n,1} | D_n\} + P\{(\mathbf{Z}, \delta, Y) \in A_{n,0} | D_n\} \\ &:= \nu_{n,1}(A_{n,1}) + \nu_{n,0}(A_{n,0}). \end{aligned} \tag{22}$$

Similarly, for any $A_n \in \mathcal{A}_n$, define the empirical measure of A_n as

$$\begin{aligned} \widehat{\nu}_n(A_n) &= \frac{1}{n}\sum_{i=1}^{n} I\{(\mathbf{Z}_i, \delta_i, Y_i) \in A_n\} = \frac{1}{n}\sum_{i=1}^{n} I\left\{I\left\{G(\mathbf{Z}_i, \delta_i, D_n) > \frac{1}{2}\right\} \neq Y_i\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} I\left\{I\left\{G(\mathbf{Z}_i, \delta_i, D_n) > \frac{1}{2}\right\} = 1, Y_i = 0\right\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} I\left\{I\left\{G(\mathbf{Z}_i, \delta_i, D_n) > \frac{1}{2}\right\} = 0, Y_i = 1\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} I\{(\mathbf{Z}_i, \delta_i, Y_i) \in A_{n,1}\} + \frac{1}{n}\sum_{i=1}^{n} I\{(\mathbf{Z}_i, \delta_i, Y_i) \in A_{n,0}\} \\ &:= \widehat{\nu}_{n,1}(A_{n,1}) + \widehat{\nu}_{n,0}(A_{n,0}). \end{aligned} \tag{23}$$

Next, observe that

$$\begin{aligned} L(\tilde{\phi}_n) - \inf_{\phi_n \in \Phi_n} L(\phi_n) &= \left(L(\tilde{\phi}_n) - \widehat{L}_n^{(R)}(\tilde{\phi}_n)\right) + \left(\widehat{L}_n^{(R)}(\tilde{\phi}_n) - \inf_{\phi_n \in \Phi_n} L(\phi_n)\right) \\ &\leq \sup_{\phi_n \in \Phi_n} \left|\widehat{L}_n^{(R)}(\phi_n) - L(\phi_n)\right| + \left(\widehat{L}_n^{(R)}(\tilde{\phi}_n) - \inf_{\phi_n \in \Phi_n} L(\phi_n)\right) \\ &\leq 2 \sup_{\phi_n \in \Phi_n} \left|\widehat{L}_n^{(R)}(\phi_n) - L(\phi_n)\right| \end{aligned}$$

$$= 2 \sup_{A_n \in \mathcal{A}_n} \left| \widehat{\nu}_n(A_n) - \nu_n(A_n) \right|$$
$$\leq 2 \sup_{A_{n,1} \in \mathcal{A}_{n,1}} \left| \widehat{\nu}_{n,1}(A_{n,1}) - \nu_{n,1}(A_{n,1}) \right| + 2 \sup_{A_{n,0} \in \mathcal{A}_{n,0}} \left| \widehat{\nu}_{n,0}(A_{n,0}) - \nu_{n,0}(A_{n,0}) \right| \quad (24)$$

where the last expression follows from (22) and (23). Let

$$\mathcal{C}_n = \left\{ (\mathbf{z}, \delta) \middle| I\left\{ G(\mathbf{z}, \delta, D_n) > \frac{1}{2} \right\} = 1 \right\}$$

and for every $C_n \in \mathcal{C}_n$, let $\widehat{\mu}_n(C_n) = \frac{1}{n} \sum_{i=1}^n I\{(\mathbf{Z}_i, \delta_i) \in C_n\}$ be the empirical measure of C_n. Also, let $\mu_n(C_n) = P\{(\mathbf{Z}, \delta) \in C_n | D_n\}$. Then we have

$$\text{R.H.S of (24)} \leq 2 \sup_{C_n \in \mathcal{C}_n} \left| \widehat{\mu}_n(C_n) - \mu_n(C_n) \right| + 2 \sup_{C_n \in \mathcal{C}_n^c} \left| \widehat{\mu}_n(C_n) - \mu_n(C_n) \right|$$
$$\leq 4 \sup_{B \in \mathcal{B}} \left| \widehat{\mu}_n(B) - \mu(B) \right| \quad (25)$$

where $\mathcal{B}$ is the collection of all Borel sets in $\mathbb{R}^{d+p} \times \{0, 1\}$, $\mu(B) = P\{(\mathbf{Z}, \delta) \in B\}$, and $\widehat{\mu}_n(B) = \frac{1}{n} \sum_{i=1}^n I\{(\mathbf{Z}_i, \delta_i) \in B\}$. To see that the bound in (25) goes to zero with probability one, let Ξ be the set of all possible values of $(\mathbf{Z}, \delta)$ and let S be an arbitrary finite subset of Ξ. It follows that

$$\sup_{B \in \mathcal{B}} \left| \widehat{\mu}_n(B) - \mu(B) \right| = \frac{1}{2} \sum_{(\mathbf{z}, \delta) \in \Xi} \left| \widehat{\mu}_n(\{\mathbf{z}, \delta\}) - \mu(\{\mathbf{z}, \delta\}) \right| \text{ (Scheffé's theorem)}$$
$$= \frac{1}{2} \sum_{(\mathbf{z}, \delta) \in S} \left| \widehat{\mu}_n(\{\mathbf{z}, \delta\}) - \mu(\{\mathbf{z}, \delta\}) \right| + \frac{1}{2} \sum_{(\mathbf{z}, \delta) \in S^c} \left| \widehat{\mu}_n(\{\mathbf{z}, \delta\}) - \mu(\{\mathbf{z}, \delta\}) \right|$$
$$\leq \frac{1}{2} \sum_{(\mathbf{z}, \delta) \in S} \left| \widehat{\mu}_n(\{\mathbf{z}, \delta\}) - \mu(\{\mathbf{z}, \delta\}) \right| + \widehat{\mu}_n(S^c) + \mu(S^c)$$
$$\leq \frac{1}{2} \sum_{(\mathbf{z}, \delta) \in S} \left| \widehat{\mu}_n(\{\mathbf{z}, \delta\}) - \mu(\{\mathbf{z}, \delta\}) \right| + \left| \widehat{\mu}_n(S^c) - \mu(S^c) \right| + 2\mu(S^c). \quad (26)$$

The first $|S|$ terms in (26) can be bounded via Hoeffding's inequality: that is, given $\epsilon > 0$,

$$P\left(\left| \widehat{\mu}_n(\{\mathbf{z}, \delta\}) - \mu(\{\mathbf{z}, \delta\}) \right| > \epsilon \middle| D_n \right) \leq 2e^{-2n\epsilon^2}, \text{ a.s.}$$

Similarly, the term $|\widehat{\mu}_n(S^c) - \mu(S^c)|$ can be bounded by one more application of Hoeffding's inequality. Finally, $\mu(S^c)$ can be made as small as desired by choosing S large enough. Thus, by the Borel–Cantelli lemma, $L(\tilde{\phi}_n) - \inf_{\phi_n \in \Phi_n} L(\phi_n) \to 0$, with probability one, as $n \to \infty$. Next, note that $L(\tilde{\phi}_n) - L(\phi_B)$ can be rewritten as

$$L(\tilde{\phi}_n) - L(\phi_B) = \left(L(\tilde{\phi}_n) - \inf_{\phi_n \in \Phi_n} L(\phi_n)\right) + \left(\inf_{\phi_n \in \Phi_n} L(\phi_n) - L(\phi_B)\right)$$
$$:= \mathrm{I}_n + \mathrm{II}_n.$$

To show that $\mathrm{II}_n \xrightarrow{a.s.} 0$, let ψ_n^{mult} be the *multinomial discrimination rule*, which is given by

$$\psi_n^{mult}(\mathbf{z}, \delta) = \begin{cases} 1 & \text{if } \delta\hat{\lambda}_1(\mathbf{z}) + (1-\delta)\hat{\lambda}_0(\mathbf{x}) > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\hat{\lambda}_1(\mathbf{z}) = \frac{\sum_{i=1}^n \delta_i Y_i I\{\mathbf{Z}_i = \mathbf{z}\}}{\sum_{i=1}^n \delta_i I\{\mathbf{Z}_i = \mathbf{z}\}} \quad \text{and} \quad \hat{\lambda}_0(\mathbf{x}) = \frac{\sum_{i=1}^n (1-\delta_i) Y_i I\{\mathbf{X}_i = \mathbf{x}\}}{\sum_{i=1}^n (1-\delta_i) I\{\mathbf{X}_i = \mathbf{x}\}}.$$

Note the similarity between the above classifier and the kernel classifier defined via (4) and (5). It turns out that the multinomial classifier is exactly equal to the kernel classifier in (5) provided h_1 and h_0 are taken to be 0 [see Devroye et al. (1996, p. 462)]. The strong consistency of the multinomial discrimination rule ψ_n^{mult} then follows from the consistency results for general partitioning estimates of regression functions along with expression (2). For more on partitioning estimates see, for example, Mojirsheibani and Montazeri (2007). Also, see Gyorfi et al. (2002, Sect. 23.1).

The discussion above implies that $\inf_{\phi_n \in \Phi_n} L(\phi_n) \leq L(\psi_n^{mult})$. But since $L(\psi_n^{mult}) \xrightarrow{a.s.} L(\phi_B)$, we see that $\inf_{\phi_n \in \Phi_n} L(\phi_n) \xrightarrow{a.s.} L(\phi_B)$ which yields the desired result (that $\mathrm{II}_n \xrightarrow{a.s.} 0$). Therefore, $L(\tilde{\phi}_n) - L(\phi_B) \xrightarrow{a.s.} 0$ as $n \to \infty$. □

Proof of Lemma 1

Let

$$\phi_1(\mathbf{z}) = \begin{cases} 1 & \text{if } \dfrac{E[\delta Y | \mathbf{Z} = \mathbf{z}]}{E[\delta | \mathbf{Z} = \mathbf{z}]} > \dfrac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \phi_0(\mathbf{x}) = \begin{cases} 1 & \text{if } \dfrac{E[(1-\delta) Y | \mathbf{X} = \mathbf{x}]}{E[(1-\delta) | \mathbf{X} = \mathbf{x}]} > \dfrac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{27}$$

with the convention $0/0 = 0$, and note that the optimal classifier in (2) can be written as

$$\phi_B(\mathbf{Z}, \delta) = \delta\phi_1(\mathbf{Z}) + (1-\delta)\phi_0(\mathbf{X}).$$

Next, write

$$\begin{aligned} P\{\phi_B(\mathbf{Z},\delta)=Y\} &= P\{\delta\phi_1(\mathbf{Z})+(1-\delta)\phi_0(\mathbf{X})=Y\} \\ &= P\{\delta=1,\phi_1(\mathbf{Z})=1,Y=1\}+P\{\delta=1,\phi_1(\mathbf{Z})=0,Y=0\} \\ &\quad +P\{\delta=0,\phi_0(\mathbf{X})=1,Y=1\}+P\{\delta=0,\phi_0(\mathbf{X})=0,Y=0\} \\ &:= A+B+C+D \text{ (say).} \end{aligned}$$

Put $p(\mathbf{Z},Y)=P(\delta=1|\mathbf{Z},Y)$, $q(\mathbf{X},Y)=P(\delta=1|\mathbf{X},Y)$, $\eta(\mathbf{Z})=P(Y=1|\mathbf{Z})$, and $\lambda(\mathbf{X})=P(Y=1|\mathbf{X})$, and observe that upon conditioning on $\mathbf{Z}$ and Y, one finds

$$\begin{aligned} A &= E\left[I\{\phi_1(\mathbf{Z})=1\}I\{Y=1\}p(\mathbf{Z},Y)\right] \\ &= E\left[I\{\phi_1(\mathbf{Z})=1\}I\{Y=1\}\big(I\{Y=1\}p(\mathbf{Z},1)+I\{Y=0\}p(\mathbf{Z},0)\big)\right] \\ &= E\left[I\{\phi_1(\mathbf{Z})=1\}I\{Y=1\}p(\mathbf{Z},1)\right]=E\left[E\left(I\{\phi_1(\mathbf{Z})=1\}I\{Y=1\}p(\mathbf{Z},1)|\mathbf{Z}\right)\right] \\ &= E\left[I\{\phi_1(\mathbf{Z})=1\}p(\mathbf{Z},1)\eta(\mathbf{Z})\right]. \end{aligned}$$

Similarly,

$$B=E\left[I\{\phi_1(\mathbf{Z})=0\}p(\mathbf{Z},0)(1-\eta(\mathbf{Z}))\right],\quad C=E\left[I\{\phi_0(\mathbf{X})=1\}q(\mathbf{X},1)\lambda(\mathbf{X})\right],$$
$$D=E\left[I\{\phi_0(\mathbf{X})=0\}q(\mathbf{X},0)(1-\lambda(\mathbf{X}))\right].$$

Next, let $\hat{\phi}_1$, $\hat{\phi}_0$, and $\hat{\phi}$ be as in the statement of the lemma and note that

$$P\{\hat{\phi}(\mathbf{Z},\delta)=Y|D_m\}=A'+B'+C'+D', \text{ a.s.,}$$

where

$$\begin{aligned} A' &= E\left[I\{\hat{\phi}_1(\mathbf{Z})=1\}p(\mathbf{Z},1)\eta(\mathbf{Z})\Big|D_m\right], \\ B' &= E\left[I\{\hat{\phi}_1(\mathbf{Z})=0\}p(\mathbf{Z},0)(1-\eta(\mathbf{Z}))\Big|D_m\right], \\ C' &= E\left[I\{\hat{\phi}_0(\mathbf{X})=1\}q(\mathbf{X},1)\lambda(\mathbf{X})\Big|D_m\right], \\ D' &= E\left[I\{\hat{\phi}_0(\mathbf{X})=0\}q(\mathbf{X},0)(1-\lambda(\mathbf{X}))\Big|D_m\right]. \end{aligned}$$

Thus,

$$\begin{aligned} L(\hat{\phi})-L(\phi_B) &= P\{\phi_B(\mathbf{Z},\delta)=Y\}-P\{\hat{\phi}(\mathbf{Z},\delta)=Y|D_m\} \\ &= (A-A')+(B-B')+(C-C')+(D-D') \\ &= E\left[p(\mathbf{Z},1)\eta(\mathbf{Z})\left(I\{\phi_1(\mathbf{Z})=1\}-I\{\hat{\phi}_1(\mathbf{Z})=1\}\right)\Big|D_m\right] \\ &\quad +E\left[p(\mathbf{Z},0)\,(1-\eta(\mathbf{Z}))\left(I\{\phi_1(\mathbf{Z})=0\}-I\{\hat{\phi}_1(\mathbf{Z})=0\}\right)\Big|D_m\right] \\ &\quad +E\left[q(\mathbf{X},1)\lambda(\mathbf{X})\left(I\{\phi_0(\mathbf{X})=1\}-I\{\hat{\phi}_0(\mathbf{X})=1\}\right)\Big|D_m\right] \\ &\quad +E\left[q(\mathbf{X},0)\,(1-\lambda(\mathbf{X}))\left(I\{\phi_0(\mathbf{X})=0\}-I\{\hat{\phi}_0(\mathbf{X})=0\}\right)\Big|D_m\right], \text{ a.s.} \end{aligned}$$

$$= E\left[\{p(\mathbf{Z},1)\eta(\mathbf{Z}) - p(\mathbf{Z},0)(1-\eta(\mathbf{Z}))\}\left(I\{\phi_1(\mathbf{Z})=1\} - I\{\hat{\phi}_1(\mathbf{Z})=1\}\right)\Big|D_m\right]$$
$$+ E\left[\{q(\mathbf{X},1)\lambda(\mathbf{X}) - q(\mathbf{X},0)(1-\lambda(\mathbf{X}))\}\left(I\{\phi_0(\mathbf{X})=1\} - I\{\hat{\phi}_0(\mathbf{X})=1\}\right)\Big|D_m\right], \text{ a.s.} \tag{28}$$

On the other hand,

$$\begin{aligned} E(\delta Y|\mathbf{Z}) &= E[E(\delta Y|\mathbf{Z},Y)|\mathbf{Z}] = E[Yp(\mathbf{Z},Y)|\mathbf{Z}] = E[Y\{Yp(\mathbf{Z},1) \\ &\quad +(1-Y)p(\mathbf{Z},0)\}|\mathbf{Z}] \\ &= E[Yp(\mathbf{Z},1)|\mathbf{Z}] = p(\mathbf{Z},1)\eta(\mathbf{Z}), \text{ a.s.} \end{aligned}$$

We also have

$$\begin{aligned} E(\delta|\mathbf{Z}) &= E[E(\delta|\mathbf{Z},Y)|\mathbf{Z}] = E[p(\mathbf{Z},Y)|\mathbf{Z}] = E[Yp(\mathbf{Z},1) + (1-Y)p(\mathbf{Z},0)|\mathbf{Z}] \\ &= p(\mathbf{Z},1)\eta(\mathbf{Z}) + p(\mathbf{Z},0)(1-\eta(\mathbf{Z})), \text{ a.s.} \end{aligned}$$

Therefore,

$$\begin{aligned} E(\delta(2Y-1)|\mathbf{Z}) &= 2E(\delta Y|\mathbf{Z}) - E(\delta|\mathbf{Z}) \\ &= 2p(\mathbf{Z},1)\eta(\mathbf{Z}) - [p(\mathbf{Z},1)\eta(\mathbf{Z}) + p(\mathbf{Z},0)(1-\eta(\mathbf{Z}))] \\ &= p(\mathbf{Z},1)\eta(\mathbf{Z}) - p(\mathbf{Z},0)(1-\eta(\mathbf{Z})), \text{ a.s.} \end{aligned} \tag{29}$$

Similarly, one can show that

$$E[(1-\delta)(2Y-1)|\mathbf{X}] = q(\mathbf{X},1)\lambda(\mathbf{X}) - q(\mathbf{X},0)(1-\lambda(\mathbf{X})), \text{ a.s.} \tag{30}$$

Substituting (29) and (30) in (28), one has

$$\begin{aligned} &L(\hat{\phi}) - L(\phi_B) \\ &= E\left\{E[\delta(2Y-1)|\mathbf{Z}]\left(I\{\phi_1(\mathbf{Z})=1\} - I\{\hat{\phi}_1(\mathbf{Z})=1\}\right)\Big|D_m\right\} \\ &\quad + E\left\{E[(1-\delta)(2Y-1)|\mathbf{X}]\left(I\{\phi_0(\mathbf{X})=1\} - I\{\hat{\phi}_0(\mathbf{X})=1\}\right)\Big|D_m\right\}, \text{ a.s.} \end{aligned} \tag{31}$$

Now, in view of the definition of $\hat{\phi}_1$ and the fact that the function ϕ_1 [see (27)] can alternatively be written in the form

$$\phi_1(\mathbf{z}) = \begin{cases} 1 & \text{if } E[\delta(2Y-1)|\mathbf{Z}=\mathbf{z}] > 0 \\ 0 & \text{otherwise,} \end{cases}$$

one finds

$$\begin{aligned} &E[\delta(2Y-1)|\mathbf{Z}=\mathbf{z}]\left(I\{\phi_1(\mathbf{z})=1\} - I\{\hat{\phi}_1(\mathbf{z})=1\}\right) \\ &\le \left|E[\delta(2Y-1)|\mathbf{Z}=\mathbf{z}] - \widehat{T}_1(\mathbf{z})\right|, \end{aligned} \tag{32}$$

which follows by considering the two cases $\{\phi_1(\mathbf{z}) = 1, \hat{\phi}_1(\mathbf{z}) = 0\}$ and $\{\phi_1(\mathbf{z}) = 0, \hat{\phi}_1(\mathbf{z}) = 1\}$ separately. Similarly, since ϕ_0 in (27) can alternatively be written as

$$\phi_0(\mathbf{x}) = \begin{cases} 1 & \text{if } E[(1-\delta)(2Y-1)|\mathbf{X} = \mathbf{x}] > 0 \\ 0 & \text{otherwise,} \end{cases}$$

by considering the two cases $\{\phi_0(\mathbf{x}) = 1, \hat{\phi}_0(\mathbf{x}) = 0\}$ and $\{\phi_0(\mathbf{x}) = 0, \hat{\phi}_0(\mathbf{x}) = 1\}$ separately we find that

$$\begin{aligned} E[(1-\delta)(2Y-1)|\mathbf{X} = \mathbf{x}] \left(I\{\phi_0(\mathbf{x}) = 1\} - I\{\hat{\phi}_0(\mathbf{x}) = 1\} \right) \\ \leq \left| E[(1-\delta)(2Y-1)|\mathbf{X} = \mathbf{x}] - \widehat{T}_0(\mathbf{x}) \right|. \end{aligned} \tag{33}$$

Integrating both sides of (32) with respect to the probability measure of $\mathbf{Z}$, and both sides of (33) with respect to the probability measure of $\mathbf{X}$, gives the desired result [in conjunction with (31)]. □

References

Breiman L (1996) Out-of-bag estimation. Unpublished Technical Report, UC Berkeley

Bontemps C, Racine JS, Simioni M (2009) Nonparametric vs parametric binary choice models: an empirical investigation. Toulouse School of Economics TSE Working Papers with number 09-126

Chung H, Han C (2000) Discriminant analysis when a block of observations is missing. Ann Inst Stat Math 52:544–556

Devroye L (1981) On the almost everywhere convergence of nonparametric regression function estimates. Ann Stat 9:1310–1319

Devroye L (1982) Bounds for the uniform deviation of empirical measures. J Multivar Anal 12:72–79

Devroye L, Krzyzak A (1989) An equivalence theorem for L_1 convergence of the kernel regression estimate. J Stat Plann Inference 23:71–82

Devroye L, Wagner T (1980) Distribution-free consistency results in nonparametric discrimination and regression function estimation. Ann Stat 8:231–239

Devroye L, Wagner T (1982) Nearest neighbor methods in discrimination. Handb Stat 2:193–197

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

Glick N (1973) Sample-based multinomial classification. Biometrics 29:241–256

Gordon L, Olshen R (1978) Asymptotically efficient solutions to the classification problem. Ann Stat 6:515–533

Gordon L, Olshen R (1980) Consistent nonparametric regression from recursive partitioning schemes. J Multivar Anal 10:611–627

Györfi L, Kohler M, Krzyzak A, Walk H (2002) A distribution-free theory of non-parametric regression. Springer, New York

Hall P, Kang KH (2005) Bandwidth choice for nonparametric classification. Ann Stat 33:284–306

Hall P, Marron JS (1987) On the amount of noise inherent in bandwidth selection for a kernel density estimator. Ann Stat 15:163–181

Hazelton ML (2000) Marginal density estimation from incomplete bivariate data. Stat Probab Lett 47:75–84

Hirano KI, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71:1161–1189

Hu XJ, Zhang B (2012) Pseudolikelihood ratio test with biased observations. Stat Papers 53:387–400

Karimi O, Mohammadzadeh M (2012) Bayesian spatial regression models with closed skew normal correlated errors and missing observations. Stat Pap 53:205–218

Krzyzak A (1986) The rates of convergence of kernel regression estimates and classification rules. IEEE Trans Inform Theory 32:668–679

Massart P (1990) The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. Ann Probab 18:1269–1283
Mojirsheibani M (2012) On the correct regression function (in L_2) and its applications when the dimension of the covariate vector is random. J Stat Plann Inference 142:2586–2598
Mojirsheibani M, Montazeri Z (2007) Statistical classification with missing covariates. J R Stat Soc Ser B 69:839–857
Mojirsheibani M, Reese T (2015) Kernel regression estimation for incomplete data with applications. Stat Pap. doi:10.1007/s00362-015-0693-z
Pawlak M (1993) Kernel classification rules from missing data. IEEE Trans Inform Theory 39:979–988
Pollard D (1984) Convergence of stochastic processes. Springer, New York
Toutenburg H, Shalabh, (2003) Estimation of regression models with equicorrelated responses when some observations on the response variable are missing. Stat Pap 44:217–232
Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall/CRC, Boca Raton
Wang Q, Linton O, Härdle W (2004) Semiparametric regression analysis with missing response at random. J Am Stat Ass 99:334–345