Electronic Journal of Statistics
Vol. 17 (2023) 2485–2538
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2158

Kernel machines with missing covariates

Tiantian Liu and Yair Goldberg

Faculty of Data and Decision Science
Technion–Israel Institute of Technology
Haifa, 3200003, Israel e-mail: tiantian.liu@campus.technion.ac.il*;* yairgo@technion.ac.il

Abstract: We develop a family of doubly robust kernel machines for classification in the presence of missing covariates. We assume that the missingness is missing at random and the missing pattern is homogeneous over a subset of covariates. First, we construct a novel convex augmented loss function using inverse probability weighting, multiple imputation, and surrogacy. It features (i) the double robustness against misspecification of the missing mechanism or the imputation model, and (ii) computation feasibility via a constrained quadratic optimization. Second, we obtain theoretical results for the proposed kernel machine, which include Fisher consistency, an upper bound of the excess risk, and the rate of convergence. We demonstrate the finite sample performance of the proposed kernel machine through simulation and real data analysis.

MSC2020 subject classifications: Primary 60K35.
Keywords and phrases: Classification, missing covariates, kernel machines, multiple imputation, doubly robust estimators.

Received November 2022.

Contents

1. Introduction

Classification is the problem of identifying to which category an observation belongs. A classifier is an algorithm that maps input data to a category. In recent years, kernel machine methods, such as support vector machines, have became popular for their flexibility and computational ease. Typically, the algorithms assume that the training data is fully observed. However, there are many situations where only a fraction of the covariates are available for some subjects. Such cases of missing covariates arise either by chance or by design. For example, in a two-stage clinical trial, information of some covariates is collected on all patients in the first stage. While, in the second stage, information of additional covariates is collected only on a subgroup of the patients. This could happen because of patient dropouts which is referred to as missing by chance. It could also happen because the information of the second stage is expensive or time consuming (w.r.t. some patients). Hence, the collection or missing information depends on the information of the first stage. This type of missingness is referred to as missing by design. Another example of missing covariates often occurs in a market surveys, where companies are interested in whether customers are willing to purchase a new electric product (a binary response). Sensitive questions such as personal income are often skipped in surveys when other covariates, such as age, gender, and occupation, are collected.

Throughout the paper, we assume that the missing mechanism is missing at random (MAR), which means that the missingness does not depend on the missing components given the observed data. This assumption holds true for missingness by design, although in general, the MAR assumption is non-identifiable. The missing mechanism of the aforementioned two examples can be categorized as MAR. For the two-stage clinical trial example, the missingness of the information in the second stage only depends on the information collected from the first stage which is observed and independent of the information in the second stage. For the market survey example, whether a customer answers the income question depends on the observed occupation and is assumed to be independent of the actual value of income given the observed occupation. Under the MAR assumption, three major approaches have been developed to handle the missing covariates, namely, the maximum likelihood method, the imputation method and the inverse probability weighting method Little and Rubin (2002); Pelckmans et al. (2005); Tsiatis (2006).

In the kernel machine literature, various methods have been proposed for the classification problem in the presence of missing covariates. Smola et al. (2005) constructed a framework that can handle missing covariates and missing response under the condition that the kernel machines can be written as an estimator in an exponential family. Pelckmans et al. (2005) proposed a classifier by imposing a distribution for the covariates under the assumption of missing completely at random (MCAR). Shivaswamy et al. (2006) proposed the support vector machines (SVM) using (conditional) probabilistic constraints in the presence of missing covariates. Anderson and Gupta (2011) assumed a specific distribution for the covariates and proposed a classifier using the expectation of the kernel matrix to circumvent the problem of missing covariates. Luengo et al. (2012) compared different imputation techniques for three groups of classification methods which include the SVM based on a working set selection using second-order information (Fan et al., 2005). Hazan et al. (2015) presented a kernel-gradient-based online algorithm under the low rank assumption of the joint distribution of the covariates, observed attributes, and the response variable. Stewart et al. (2018) provided an overview of some SVM-specific strategies where the methods of imputation, multiple imputations, and probability constraints are employed first to handle the missing covariates. Śmieja et al. (2019) constructed a generalized Gaussian radial bias (RBF) kernel for incomplete data under the normality assumption.

Other approaches in the domain of machine learning include K-nearest neighbours (García-Laencina et al., 2009; Choudhury and Kosorok, 2020), the neural network ensemble models (Sharpe and Solly, 1995), classification trees (Saar-Tsechansky and Provost, 2007; Ding and Simonoff, 2010), pattern classification (García-Laencina et al., 2010), learning with limited attribute observation (Bullins et al., 2016), adjusted weight voting random forests (Xia et al., 2017), neural network with generalized neuron's response (Śmieja et al., 2018), deep learning (Qiu et al., 2018), and graph representation learning (You et al., 2020). As far as we know, the consistency of the classifier has not been established by any of these methods. In a different domain, Wang et al. (2019) proposed doubly robust joint learning for recommendation when ratings (i.e., the response variables) are potentially missing and the missing mechanism is assumed to be not missing at random. Their method combines error-imputation based approach and inverse probability weighting to build a doubly robust estimator for the prediction inaccuracy, not for the loss function as is done in this work.

In this paper, we first introduce a kernel machine with an inverse-probability-weighted-complete-case loss function, where the weight of a complete case is chosen to be the inverse of the probability of observing this case. This kernel machine is inefficient since only the complete cases are used. Additionally, the consistency of the corresponding kernel machine is guaranteed only when the estimator of the missing mechanism is consistent. Secondly, we propose a novel kernel machine with a convex augmented loss to overcome the aforementioned drawbacks. The construction of the proposed loss consists of two steps. First, through multiple imputations, the information of the incomplete cases is used to construct an augmented loss function to achieve double robustness. Second, a surrogate loss and nonnegative weights are used together to modify the augmented loss to achieve the convexity. The latter is essential for a computationally tractable solution and a simple representation of kernel machines (Steinwart and Christmann, 2008, Theorem 5.5). The surrogate losses we consider here are the classification calibrated losses which possess the property that minimizing the risk with respect to the surrogate losses implies minimizing the risk with respect to the classification loss as proposed by Bartlett et al. (2006). See more discussion of the classification calibrated loss in Steinwart and Christmann (2008, Chap. 3) and Bao et al. (2020). We show that the proposed convex augmented loss is indeed a calibrated loss for the classification loss in the presence of missing covariates.

We establish some desired theoretical properties of the proposed kernel machine. Specifically, we show that the optimal classifier is Fisher consistent and doubly robust against the misspecification of either the missing mechanism or the imputation model but not necessarily both. To demonstrate the closeness to the Bayes risk, we derive an upper bound of the excess risk with respect to the classification loss (Steinwart and Christmann, 2008, Sect. 2). This upper bound holds if either the imputation model or the missingness model is correctly specified. In addition, we obtain an upper bound for the excess risk of the proposed kernel machine classifier and establish the convergence rate of the risk. We show that the proposed kernel machine classifier can be implemented using constrained quadratic programming. The R package `drkm4mc` is provided for implementation of the proposed methods. We demonstrate the performance of the proposed kernel machine through simulation and real data analysis. Lastly, we extend the proposed kernel machine to accommodate a more complicated type of missing pattern.

Here, we would like to emphasize the difference between the proposed convex augmented loss and the usual augmented inverse probability weighting (AIPW) methods (Robins et al., 1994; An and Fuller, 1998; Fuller, 2011). The AIPW technique is used to construct an augmented term to an estimator (Scharfstein et al., 1999; Tsiatis, 2006, Chap. 6.5) or an estimation equation (Carpenter et al., 2006; Han et al., 2019) in the presence of missing covariates. Our approach is different from the usual AIPW method since we do not add an augmented term to an estimator or an estimating equation. Specifically, in our proposed kernel machine, we construct an augmented term to an inverse probability weighted (IPW) loss function to achieve double robustness. However, this raises mathematical challenges regarding the convexity of the augmented loss function and further the ability to use the kernel trick (Hofmann et al., 2008). The proposed kernel machine classifier is obtained by minimizing a regularized empirical risk over a reproducing kernel Hilbert space (RKHS), in contrast to an optimization program with respect to the IPW loss function.

So far, the only similar work we are aware of using the convex augmented loss function is in Liu and Goldberg (2020) who developed kernel machines with a different missingness type, namely, missing responses. However, their result is limited to the quadratic loss. The loss functions considered in the present work are much broader, which include the hinge loss, the quadratic loss, the logistic loss, and the exponential loss, among others.

The rest of the paper is organized as follows. Section 2 introduces some notation and preliminary about kernel machine classification, the assumption of missing at random, and a naive method to handle classification in the presence of missing covariates. Section 3 presents the construction of a novel convex augmented loss function which serves as the basis for the proposed kernel machine. Section 4 provides the theoretical results including Fisher consistency, an upper bound of the excess risk, and rate of convergence. Sections 5 and 6 compare the proposed method with some existing methods through simulation and real data, respectively. Section 7 concludes the paper with some discussion. All technical details are deferred to Appendix.

2. Notation and preliminaries

Let $Y \in \mathcal{Y} = \{1, -1\}$ denote a binary response variable which represents two different categories. Let $X \in \mathcal{X} \subset \mathbb{R}^d$ denote a d-dimensional vector of covariates, which is partitioned as $X = (X_1^\intercal, X_2^\intercal)^\intercal$ with $X_1 \in \mathcal{X}_1 \subset \mathbb{R}^{d_1}$ and $X_2 \in \mathcal{X}_2 \subset \mathbb{R}^{d_2}$, respectively. Assume that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ is a compact set. Assume that (X, Y) jointly follows an unknown probability distribution P. Denote the marginal distribution of X by P_X.

Let $I\{\cdot\}$ denote the indicator function. Define $\mathrm{sign}(t) = 2I(t \geq 0) - 1 \in \mathcal{Y}$. For a measurable function $f : X \mapsto \mathbb{R}$, denote the classification 0-1 loss as $L(X, Y, f) = I[Y \neq \mathrm{sign}\{f(X)\}] = I[Y\mathrm{sign}\{f(X)\} \leq 0]$. Denote the risk by $\mathcal{R}(f) = \mathrm{E}\{L(X, Y, f)\}$. The Bayes risk is defined by $\mathcal{R}^* = \inf_f \mathcal{R}(f)$, where the infimum is taken over all measurable functions f. It is attained at $f_{I,\mathrm{opt}}$ such that $\mathrm{sign}\,(f_{I,\mathrm{opt}}) = \mathrm{sign}\,\{2\mathrm{P}(Y = 1 \mid X) - 1\}$ with the value $\mathcal{R}^* = \int_{\mathcal{X}} \min\{\mathrm{P}(Y = 1 \mid x), 1 - \mathrm{P}(Y = 1 \mid x)\}d\mathrm{P}_X$ (Steinwart and Christmann, 2008, Sect. 2.1).

Let $\mathcal{H}$ be a separable reproducing kernel Hilbert space (RKHS) of a bounded measurable kernel on $\mathcal{X}$. Denote its norm by $\| \cdot \|_{\mathcal{H}}$. Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be the kernel function satisfying that $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$ for all x, $x' \in \mathcal{X}$, where Φ is called the feature map, and $\langle \cdot, \cdot \rangle$ denotes the inner product. For $f \in \mathcal{H}$, $\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle$. RKHS possesses the property that the norm convergence implies the point-wise convergence. We further assume that k is universal in the sense that $\mathcal{H}$ is dense in the space of bounded continuous functions with respect to the supremum norm, denoted by $\| \cdot \|_\infty$. It is used to facilitate the derivation of rate of convergence in Sect. 4. Without loss of generality, assume that $\|k\|_\infty \leq 1$ (Hofmann et al., 2008; Steinwart and Christmann, 2008, Chap. 4). A kernel machine is a function $f \in \mathcal{H}$ which minimizes the regularized empirical risk.

Now, assume that X_1 is fully observed and X_2 is potentially missing. Let R denote the missingness indicator with $R = 1$ if X_2 is observed, called the 'complete case', and $R = 0$ if X_2 is missing, called the 'incomplete case'. We elaborate the missing at random assumption as follows.

Assumption 1. *The missingness indicator R and the potentially missing covariates X_2 are independent given the fully observed covariates X_1 and the response Y, i.e., the missing mechanism is missing at random (MAR) (Little and Rubin, 2002).*

Under Assumption 1, define the probability of observing X_2 given X_1 and Y, i.e., the propensity score, as $\pi^0(X_1, Y) = \mathrm{P}(R = 1|X_1, Y)$.

Suppose that $(R_i, X_{i1}, R_i X_{i2}, Y_i)$, $i = 1, \ldots, n$, are independent and identically distributed samples of (R, X_1, RX_2, Y) where X_{i2} is observed only when $R_i = 1$. Let $\mathbb{P}_n(F(U)) = n^{-1}\sum_{i=1}^n F(U_i)$ denote the empirical process for an arbitrary function F and a simple random sample of U in $U_1, \ldots, U_n$.

When there is no missing covariates, $\mathcal{R}(f)$ is simply estimated by $\mathcal{R}_D(f) = \mathbb{P}_n[I\{Y\mathrm{sign}(f(X)) \le 0\}]$. When some covariates are missing, $\mathcal{R}_D(f)$ is not available since $\mathrm{sign}(f(X))$ is not defined for the missing observations. A naive method is to just use the complete cases and estimate $\mathcal{R}(f)$ by

$$\mathcal{R}_D^{\mathrm{C}}(f) = \frac{\sum_{i=1}^n R_i I[Y_i \mathrm{sign}\{f(X_i)\} \le 0]}{\sum_{i=1}^n R_i}.$$

Here, $X_i = (X_{i1}, X_{i2})$. However, this complete-case-based estimator is biased unless the missing mechanism is missing completely at random (MCAR) (Tsiatis, 2006, Sect. 6.1).

A common method to correct such bias is to use the inverse probability weighted (IPW) loss defined by

$$L_{\mathrm{IPW}}(R, X, Y, \widehat{\pi}, f) = \frac{RI[Y\mathrm{sign}(f(X)\} \le 0]}{\widehat{\pi}(X_1, Y)}, \tag{2.1}$$

where $\widehat{\pi}(X_1, Y)$ is an estimator of $\pi^0(X_1, Y)$ (Tsiatis, 2006). Denote the empirical risk with respect to (2.1) as $\mathcal{R}_{L_{\mathrm{IPW}},D}(f) = \mathbb{P}_n\{L_{\mathrm{IPW}}(R, X, Y, \widehat{\pi}, f)\}$. The minimizer of the regularized version of $\mathcal{R}_{L_{\mathrm{IPW}},D}(f)$, i.e.,

$$\widehat{f}_{\mathrm{WCC}} = \arg\min_{f\in\mathcal{H}} \lambda\|f\|_{\mathcal{H}}^2 + \mathcal{R}_{L_{\mathrm{IPW}},D}(f), \tag{2.2}$$

is defined as the weighted-complete-case kernel machine estimator. It can be shown (Liu and Goldberg, 2020) that under MAR, $\mathcal{R}_{L_{\mathrm{IPW}},D}(f)$ is consistent whenever $\widehat{\pi}(X_1, Y)$ is a consistent estimator of $\pi^0(X_1, Y)$. However, the consistency of $\widehat{\pi}(X_1, Y)$ is not guaranteed in general. Secondly, the minimizer of $\mathcal{R}_{L_{\mathrm{IPW}},D}(f)$ is subject to inefficiency since only complete cases are used directly.

In the next section, we will propose a convex augmented surrogate loss to overcome these two problems.

3. Convex augmented loss

3.1. Augmented loss

We first introduce an augmented loss which serves as a stepping stone for the later convex augmented loss. Our construction of the augmented term is based on a conditional expectation with multiple imputations.

Let $F^0_{2|1,Y}(x_2)$ denote the conditional distribution function of X_2 given X_1 and Y; for example, the conditional normal distribution as used in the simulation study. Let $\widehat{F}_{2|1,Y}(x_2)$ be an estimator of $F^0_{2|1,Y}(x_2)$ based on the sample.

Assumption 2. *$\widehat{F}_{2|1,Y}(x_2)$ converges in probability to $F^*_{2|1,Y}(x_2)$, i.e., for any fixed x_1, x_2, y, and $\varepsilon > 0$,*

$$\mathrm{P}\{|\widehat{F}_{2|1,Y}(x_2) - F^*_{2|1,Y}(x_2)| \geq \varepsilon\} \longrightarrow 0;$$

and $\widehat{\pi}(x_1, y)$ converges in probability to $\pi^*(x_1, y)$, i.e., for any fixed x_1, y, and $\varepsilon > 0$,

$$\mathrm{P}\{|\widehat{\pi}(x_1, y) - \pi^*(x_1, y)| \geq \varepsilon\} \longrightarrow 0.$$

Remark 3.1. *Assumption 2 requires that both $\widehat{F}_{2|1,Y}(x_2)$ and $\widehat{\pi}(x_1, y)$ converge to some deterministic functions $F^*_{2|1,Y}(x_2)$ and $\pi^*(x_1, y)$ which are not necessarily the true functions $F^0_{2|1,Y}(x_2)$ and $\pi^0(x_1, y)$ due to possible model misspecification.*

Let $X_{2|1,Y}$, $X^{\mathrm{imp}}_{2|1,Y}$, and $X^*_{2|1,Y}$ denote the random variables whose distribution functions are $F^0_{2|1,Y}(x_2)$, $\widehat{F}_{2|1,Y}(x_2)$, and $F^*_{2|1,Y}(x_2)$, respectively.

Remark 3.2. *Assumption 1 implies that the missingness R is independent of both $X^{\mathrm{imp}}_{2|1,Y}$ and $X^*_{2|1,Y}$, given X_1 and Y.*

Denote $X^0 = (X_1^\intercal, X_{2|1,Y}^\intercal)^\intercal$, $X^{\mathrm{imp}} = (X_1^\intercal, X^{\mathrm{imp}\intercal}_{2|1,Y})^\intercal$ and $X^* = (X_1^\intercal, X^{*\intercal}_{2|1,Y})^\intercal$ for later use. Denote $X^{\mathrm{g}} = (X_1^\intercal, Z^\intercal)^\intercal$ as a generic random vector (of dimension $d_1 + d_2$) where Z is some d_2 dimensional random vector. We shall use X^{g} to represent X^0, X^{imp} and X^* in a generic form for later development. Let $\pi^{\mathrm{g}}(X_1, Y)$ denote a generic conditional probability R given X_1 and Y.

Define the augmented loss by

$$\begin{aligned}
&L_{\mathrm{aug}}(\pi^*, X^*, f)\\
=&\frac{R}{\pi^*(X_1, Y)} I[Y \mathrm{sign}\{f(X)\} \leq 0]\\
&+ \frac{\pi^*(X_1, Y) - R}{\pi^*(X_1, Y)} \mathrm{E}_{X^*_{2|1,Y}}(I[Y \mathrm{sign}\{f(X_1, X^*_{2|1,Y})\} \leq 0] \mid X_1, Y), \qquad (3.1)\\
=&I[Y \mathrm{sign}\{f(X)\} \leq 0] + \frac{R - \pi^*(X_1, Y)}{\pi^*(X_1, Y)}\\
&\quad \times \{I[Y \mathrm{sign}\{f(X)\} \leq 0] - \mathrm{E}_{X^*_{2|1,Y}}(I[Y \mathrm{sign}\{f(X_1, X^*_{2|1,Y})\} \leq 0] \mid X_1, Y)\},\\
&\qquad (3.2)
\end{aligned}$$

where the expectation is taken with respect to $X^*_{2|1,Y}$. Note that the augmented loss function $L_{\mathrm{aug}}(\pi^*, X^*, f)$ depends on X, Y, and R. To simplify the notation, we omit these three terms in the argument. Similar simplification is used to denote loss functions in Sect. 3.2. For fixed f, the empirical risk of (3.2) is used to estimate $\mathrm{E}[I\{Y \mathrm{sign}(f(X)) \leq 0\}]$. A similar form has been used to estimate the population mean (Tsiatis, 2006, Sect. 6) or the regression coefficients (Seaman and Vansteelandt, 2018) in the presence of missing responses. However, we consider $L_{\text{aug}}(\pi^*, X^*, f)$ as a loss function over different f which is unlike the aforementioned forms. We call the risk with respect to $L_{\text{aug}}(\pi^*, X^*, f)$, i.e., $\mathcal{R}_{L^*_{\text{aug}}}(f) = \mathrm{E}\{L_{\text{aug}}(\pi^*, X^*, f)\}$, the auxiliary risk.

We introduce the following two conditions regarding model specification.

Condition 1 (CD)**.** *The conditional distributional model is correctly specified if $F^*_{2|1,Y}(x_2) = F^0_{2|1,Y}(x_2)$ while $\pi^*(x_1, y)$ is not necessarily $\pi^0(x_1, y)$.*

Condition 2 (PS)**.** *The propensity score model is correctly specified if $\pi^*(x_1, y) = \pi^0(x_1, y)$ while $F^*_{2|1,Y}(x_2)$ is not necessarily $F^0_{2|1,Y}(x_2)$.*

Theorem 3.1. *Under Assumption 1, $\mathcal{R}_{L^*_{\text{aug}}}(f) = \mathcal{R}(f)$ whenever Condition 1 or Condition 2 holds.*

The proof of Theorem 3.1 basically shows that under either one of the two conditions the expectation of the second term of (3.2) vanishes, thus establishing the doubly robust property of $\mathcal{R}_{L^*_{\text{aug}}}(f)$.

By replacing π^* and X^* as $\widehat{\pi}$ and X^{imp} respectively, we obtain the sample version of $L_{\text{aug}}(\pi^*, X^*, f)$ as

$$\begin{aligned} & L_{\text{aug}}(\widehat{\pi}, X^{\text{imp}}, f) \\ =& \frac{R}{\widehat{\pi}(X_1, Y)} I[Y \operatorname{sign}\{f(X)\} \leq 0] \\ & + \frac{\widehat{\pi}(X_1, Y) - R}{\widehat{\pi}(X_1, Y)} \mathrm{E}_{X^{\text{imp}}_{2|1,Y}}(I[Y \operatorname{sign}\{f(X_1, X^{\text{imp}}_{2|1,Y})\} \leq 0] \mid X_1, Y), \end{aligned} \tag{3.3}$$

where the expectation is taken with respect to the conditional distribution $X^{\text{imp}}_{2|1,Y}$.

To estimate the conditional expectation in (3.3), we multiply impute $X^{\text{imp}}_{2|1,Y}$ m times, denoted by $X^{\text{imp}}_{2j|1,Y}$, $j = 1, \ldots, m$, based on the distribution $\widehat{F}_{2|1,Y}(x_2)$. Denote $X^{\text{imp}}_j = (X_1^{\intercal}, X^{\text{imp}\intercal}_{2j|1,Y})^{\intercal}$, $j = 1, \ldots, m$, and $\mathbf{X}^{\text{imp}} = (X^{\text{imp}}_1, \ldots, X^{\text{imp}}_m)^{\intercal}_{m \times d}$. Then, we modify $L_{\text{aug}}(\widehat{\pi}, X^{\text{imp}}, f)$ in (3.3) as

$$\begin{aligned} & L_{\text{aug}}(\widehat{\pi}, \mathbf{X}^{\text{imp}}, f) \\ =& \frac{R}{\widehat{\pi}(X_1, Y)} I[Y \operatorname{sign}\{f(X)\} \leq 0] \\ & + \frac{\widehat{\pi}(X_1, Y) - R}{\widehat{\pi}(X_1, Y)} \left(\frac{1}{m} \sum_{j=1}^{m} I[Y \operatorname{sign}\{f(X^{\text{imp}}_j)\} \leq 0] \right), \end{aligned} \tag{3.4}$$

where the augmented second term is the weighted empirical risk with respect to the imputed data. By the weak law of large numbers, the term in the brackets of (3.4) converges in probability to $\mathrm{E}_{X^{\text{imp}}_{2|1,Y}}(I[Y \operatorname{sign}\{f(X_1, X^{\text{imp}}_{2|1,Y})\} \leq 0] \mid X_1, Y)$ as $m \to \infty$. This conditional expectation further converges to the conditional expectation in (3.1) as $n \to \infty$. In practice, we find that a moderate number of m is adequate as illustrated in the simulation.

At last, denote the empirical risk with respect to $L_{\text{aug}}(\widehat{\pi}, \mathbf{X}^{\text{imp}}, f)$ by $\mathcal{R}_{L_{\text{aug}}^{\text{imp}},D}(f) = \mathbb{P}_n L_{\text{aug}}(\widehat{\pi}, \mathbf{X}^{\text{imp}}, f)$. The corresponding kernel machine estimator is given by

$$f_{D,\lambda} = \arg\min_{f \in \mathcal{H}} \lambda \|f\|_{\mathcal{H}}^2 + \mathcal{R}_{L_{\text{aug}}^{\text{imp}},D}(f), \tag{3.5}$$

where λ is the tuning parameter governing the penalty term $\|f\|_{\mathcal{H}}^2$.

3.2. *Convex surrogate*

The computation of (3.5) involves non-convex optimization because of the 0-1 loss. It is common practice to replace the 0-1 loss by some convex surrogate. To be more precise, by introducing

$$W_j(\pi^{\text{g}}) = \frac{R}{\pi^{\text{g}}(X_1, Y)} I(Y = j), \quad V_j(\pi^{\text{g}}) = \frac{\pi^{\text{g}}(X_1, Y) - R}{\pi^{\text{g}}(X_1, Y)} I(Y = j), \quad j = 1, -1,$$

we write a generic loss for (3.4) by

$$\begin{aligned} & L_{\text{aug}}(\pi^{\text{g}}, X^{\text{g}}, f) \\ = & W_1(\pi^{\text{g}}) I[\text{sign}\{f(X)\} \le 0] + W_{-1}(\pi^{\text{g}}) I[-\text{sign}\{f(X)\} \le 0] \\ & + \frac{1}{m} \sum_{j=1}^{m} V_1(\pi^{\text{g}}) I[\text{sign}\{f(X_j^{\text{g}})\} \le 0] + \frac{1}{m} \sum_{j=1}^{m} V_{-1}(\pi^{\text{g}}) I[-\text{sign}\{f(X_j^{\text{g}})\} \le 0]. \end{aligned} \tag{3.6}$$

The formula (3.6) is a general formula for any π^{g} and X^{g}. While (3.4) is a special case of (3.6) by substituting $\widehat{\pi}$ and $\mathbf{X}^{\text{imp}}$ into (3.6). For simplicity, we fix $m = 1$. The results can be easily derived for general m.

Notice that the negative value of V_j when $R = 1$ causes $L_{\text{aug}}(\pi^{\text{g}}, X^{\text{g}}, f)$ to remain non-convex even after replacing the 0-1 loss by some convex loss. So our first step toward the construction of a convex loss is to find a way of using only nonnegative weights. It is achieved by the following proposed loss function,

$$\begin{aligned} & L_{\text{abs}}(\pi^{\text{g}}, X^{\text{g}}, f) \\ = & W_1(\pi^{\text{g}}) I[\text{sign}\{f(X)\} \le 0] + W_{-1}(\pi^{\text{g}}) I[-\text{sign}\{f(X)\} \le 0] \\ & + |V_1(\pi^{\text{g}})| I[\text{sign}(V_1)\{f(X^{\text{g}})\} \le 0] + |V_{-1}(\pi^{\text{g}})| I[-\text{sign}(V_{-1})\{f(X^{\text{g}})\} \le 0]. \end{aligned} \tag{3.7}$$

For general $m > 1$, we shall replace the last two terms of (3.7) by

$$\begin{aligned} & \frac{1}{m} \sum_{j=1}^{m} |V_1(\pi^{\text{g}})| I[\text{sign}(V_1) \, \text{sign}\{f(X_j^{\text{g}})\} \le 0] \\ + & \frac{1}{m} \sum_{j=1}^{m} |V_{-1}(\pi^{\text{g}})| I[-\text{sign}(V_{-1}) \, \text{sign}\{f(X_j^{\text{g}})\} \le 0]. \end{aligned} \tag{3.8}$$

Lemma 3.1. $L_{\text{abs}}(\pi^{\text{g}}, X^{\text{g}}, f) - L_{\text{aug}}(\pi^{\text{g}}, X^{\text{g}}, f) = -V_1 I(V_1 < 0) - V_{-1} I(V_{-1} < 0)$.

Remark 3.3. *Lemma 3.1 shows that the difference between the two losses (3.7) and (3.6) is free of f, which implies that the minimizer of the risk w.r.t. $L_{\text{aug}}(\pi^{\text{g}}, X^{\text{g}}, f)$ is the same as the minimizer of the risk w.r.t. $L_{\text{abs}}(\pi^{\text{g}}, X^{\text{g}}, f)$, i.e., $\arg\min_f \mathrm{E}\{L_{\text{aug}}(\pi^{\text{g}}, X^{\text{g}}, f)\} = \arg\min_f \mathrm{E}\{L_{\text{abs}}(\pi^{\text{g}}, X^{\text{g}}, f)\}$, where $(\pi^{\text{g}}, X^{\text{g}})$ can be either (π^0, X^0), (π^*, X^*), or $(\hat{\pi}, X^{\text{imp}})$. The loss function $L_{\text{abs}}(\pi^{\text{g}}, X^{\text{g}}, f)$ is a nonnegative loss function with all positive weights. After replacing the 0-1 loss by some convex surrogate, a convex loss can be built. The proof of Lemma 3.1 is not trivial. The main technique enabled to guarantee that the following equation holds for the classification loss*

$$I[-\text{sign}\{f(X^{\text{g}})\} \leq 0] = 1 - I[\text{sign}\{f(X^{\text{g}})\} \leq 0].$$

Then the term $|V_1(\pi^{\text{g}})| I[\text{sign}(V_1)\text{sign}\{f(X^{\text{g}})\} \leq 0]$ in the loss function L_{abs} can be related to the term $V_1 I[\text{sign}\{f(X^{\text{g}})\} \leq 0]$ in the loss function L_{aug} (free of f). Similarly, the difference between the terms $|V_{-1}(\pi^{\text{g}})| I[-\text{sign}(V_{-1})$ $\text{sign}\{f(X^{\text{g}})\} \leq 0]$ and $V_{-1} I[-\text{sign}\{f(X^{\text{g}})\} \leq 0]$ is free of f.

Next, we introduce a surrogate loss to deal with the non-convex classification loss in (3.7). Let $\phi(t)$ be a convex surrogate loss for the classification loss. Define $G(t) = u\phi(t) + v\phi(-t)$, where u and v are two positive constants. Let $t_{\min} = \arg\min_{t\in\mathbb{R}} G(t)$.

Assumption 3. *$\phi(t)$ is differentiable at 0 and $\phi'(0) < 0$. $\phi(-t)$ is convex with $\phi(0) = 1$ and ϕ satisfies $\text{sign}(t_{\min}) = \text{sign}(u - v)$.*

By Theorem 2 of Bartlett et al. (2006), Assumption 3 ensures that $\phi(t)$ is a classification-calibrated loss, which is used to facilitate the derivation for the excess risk in Sect. 4.

Lemma 3.2. *Assumption 3 holds for the hinge loss, $\phi(t) = \max\{0, 1 - t\}$, the quadratic loss, $\phi(t) = (1 - t)^2$, the logistic loss, $\phi(t) = \log(1 + e^{-t})$, and the exponential loss, $\phi(t) = e^{-t}$.*

Now, on replacing the 0-1 loss by ϕ, we obtain a convex augmented loss by

$$\begin{aligned} & L_\phi(\pi^{\text{g}}, X^{\text{g}}, f) \\ =& W_1(\pi^{\text{g}})\phi\{f(X)\} + W_{-1}(\pi^{\text{g}})\phi\{-f(X)\} \\ & + |V_1(\pi^{\text{g}})|\phi\{\text{sign}(V_1) f(X^{\text{g}})\} + |V_{-1}(\pi^{\text{g}})|\phi\{-\text{sign}(V_{-1}) f(X^{\text{g}})\}. \end{aligned} \tag{3.9}$$

Define the corresponding risk as $\mathcal{R}_{L_\phi^{\text{g}}}(f) = \mathrm{E}\{L_\phi(\pi^{\text{g}}, X^{\text{g}}, f)\}$.

Remark 3.4. *The risk $\mathcal{R}_{L_\phi^{\text{g}}}(f)$, consequently its corresponding kernel machine, depends on the conditional distribution of X_2 given X_1 and Y and the propensity score model $\pi^{\text{g}}(X_1, Y)$. For example, when $(\pi^{\text{g}}, X^{\text{g}}) = (\pi^*, X^*)$, by Assumption 1 and Remark 3.2,*

$$\mathcal{R}_{L_\phi^*}(f)$$

$$= \int_{\mathcal{X}_1,\mathcal{Y}} \int_{\{0,1\}} \int_{\mathcal{X}_2} [W_1(\pi^*)\phi\{f(x_1,x_2)\} \\ + W_{-1}(\pi^*)\phi\{-f(x_1,x_2)\}] dF^0_{2|1,Y}(x_2)\, dF_{R|1,Y}(r) dF_{X_1,Y}(x_1,y) \\ + \int_{\mathcal{X}_1,\mathcal{Y}} \int_{\{0,1\}} \int_{\mathcal{X}_2} [|V_1(\pi^*)|\phi\{\text{sign}(V_1)f(x_1,x_2)\} \\ + |V_{-1}(\pi^*)|\phi\{-\text{sign}(V_{-1})f(x_1,x_2)\}] dF^*_{2|1,Y}(x_2)\, dF_{R|1,Y}(r) dF_{X_1,Y}(x_1,y)$$

where $F_{R|1,Y}(r)$ is the conditional distribution of the missing mechanism R given X_1 and Y.

Denote the Bayes risk with respect to $L_\phi(\pi^*, X^*, f)$ by $\mathcal{R}^*_{L^*_\phi} = \inf_f \mathcal{R}_{L^*_\phi}(f)$, which by convexity is attained at $f_{\phi,\text{opt}} = \arg\min_f \mathcal{R}_{L^*_\phi}(f)$. Further denote the kernel machine in $\mathcal{H}$ by

$$f_{\phi,\text{opt},\lambda} = \arg\min_{f\in\mathcal{H}} \lambda\|f\|^2_{\mathcal{H}} + \mathcal{R}_{L^*_\phi}(f). \tag{3.10}$$

Finally, denote the empirical risk of $\mathcal{R}_{L^*_\phi}(f)$ by $\mathcal{R}_{L^{\text{imp}}_\phi,D}(f) = \mathbb{P}_n L_\phi(\widehat{\pi}, \mathbf{X}^{\text{imp}}, f)$. The corresponding kernel machine estimator is

$$\widehat{f}_\phi = \arg\min_{f\in\mathcal{H}} \lambda\|f\|^2_{\mathcal{H}} + \mathcal{R}_{L^{\text{imp}}_\phi,D}(f). \tag{3.11}$$

We refer to $\widehat{f}_\phi$ in (3.11) as a doubly robust kernel machine estimator; this property is shown in the next section.

The computation of $\widehat{f}_\phi$ through a constrained quadratic optimization under the hinge loss is provided in Appendix A.

Remark 3.5. *For the augmented loss $L_{\text{aug}}(\widehat{\pi}, X^{\text{imp}}, f)$ in (3.3), though it is a common approach to construct the augmented term for the general AIPW estimator, directly estimating the conditional expectation in (3.3) does not work because the signs of V_1 and V_{-1} and the convexity of the conditional expectation are unknown. We provide more details in Appendix B.*

4. Theoretical results

4.1. Fisher consistency

Recall that $f_{I,\text{opt}}$ is the minimizer of the risk function $\mathcal{R}(f)$ in Sect. 2 and $f_{\phi,\text{opt}}$ is the minimizer of the auxiliary risk function $\mathcal{R}_{L^*_\phi}$ in Sect. 3.2.

Theorem 4.1. *Under Assumptions 1 and 3,*

$$\mathcal{R}(f_{\phi,\text{opt}}) = \mathcal{R}(f_{L^*_{\text{aug}},\text{opt}}) = \mathcal{R}(f_{I,\text{opt}}) = \mathcal{R}^*,$$

whenever either Condition 1 or Condition 2 holds, where

$$f_{L^*_{\text{aug}},\text{opt}} = \arg\min_f \mathcal{R}_{L^*_{\text{aug}}}(f).$$

Theorem 4.1 states that whenever $F^*_{2|1,Y}(x_2)$ or π^* is correctly specified, the minimizer of $\mathcal{R}_{L^*_\phi}(f)$ achieves the Bayes risk.

Remark 4.1. *Replacing the 0-1 loss by a surrogate loss cannot guarantee the same minimizer due to different loss functions. However, for the classification problem, the sign of the minimizer determines the classification rule. By the Fisher consistency result (Theorem 4.1) and Remark C.1 (in the Appendix C.5), we have shown that* $\operatorname{sign}\{f_{\phi,\mathrm{opt}}(x)\} = \operatorname{sign}\{f_{I,\mathrm{opt}}(x)\}$*, i.e., the two classifiers* $f_{\phi,\mathrm{opt}}(x)$ *and* $f_{I,\mathrm{opt}}$ *determine the same classification rule.*

4.2. Excess risk

Let $L_{\phi,1}(f) = L_\phi\left(\pi^*, X^0, f\right)$ and $L_{\phi,2}(f) = L_\phi\left(\pi^0, X^*, f\right)$ denote the augmented convex loss with $F^*_{2|1,Y}(x_2)$ correctly specified and π^* correctly specified, respectively. For $j = 1, 2$, define the risk with respect to $L_{\phi,j}$ as $\mathcal{R}_{L_{\phi,j}}(f) = \mathrm{E}\left\{L_{\phi,j}(f)\right\}$. Denote the corresponding Bayes risk by $\mathcal{R}^*_{L_{\phi,j}} = \inf_f \mathcal{R}_{L_{\phi,j}}(f)$, which is attained at $f_{\phi,j,\mathrm{opt}}$.

Following Bartlett et al. (2006), we define the optimal conditional ϕ-risk as

$$H(\eta) = \inf_{t\in\mathbb{R}}\{\eta\phi(t) + (1-\eta)\phi(-t)\}, \quad \eta \in [0,1],$$

where $\phi(t)$ is a classification-calibrated loss. Let

$$\psi(t) = \phi(0) - H\left(\frac{1+t}{2}\right). \tag{4.1}$$

Similar to Bartlett et al. (2006, Theorem 1), we use $\psi(t)$ to relate the excess risk with respect to the classification loss to the excess risk with respect to the proposed convex augmented loss. Recall that $\phi(t)$ is a classification-calibrated loss. By Bartlett et al. (2006, Lemma 2), $\psi(t)$ is invertible. In particular, $\psi(t) = |t|$ for the hinge loss, $\psi(t) = 2t^2 - 1$ for the quadratic loss, $\psi(t) = \frac{1+t}{2}\log(1+t) + \frac{1-t}{2}\log(1-t)$ for the logistic loss, and $\psi(t) = 1 - \sqrt{1-t^2}$ for the exponential loss. We will use ψ to derive the bound of the excess risk $\mathcal{R}(f) - \mathcal{R}^*$ in Theorems 4.2, 4.3, and 4.4.

Assumption 4. *There exist constants* c_ℓ *and* c_u *such that*

$$0 < c_\ell \le \pi^0(x_1,y), \widehat{\pi}(x_1,y), \pi^*(x_1,y) \le c_u < 1$$

for all $x_1 \in \mathcal{X}_1$ *and* $y \in \mathcal{Y}$*, where* $\pi^0(x_1,y)$*,* $\widehat{\pi}(x_1,y)$ *and* $\pi^*(x_1,y)$ *are defined in (2.1) and Assumption 2.*

Remark 4.2. *Assumption 4 is used to bound the propensity score and its estimator as in Tsiatis (2006, Chap. 6).*

Theorem 4.2. *Under Assumptions 1, 2, 3, and 4, (i) when Condition 1 holds,*

$$\psi\left\{\frac{\mathcal{R}(f) - \mathcal{R}^*}{\sup_{x\in\mathcal{X}} c_1(x)}\right\} \le \frac{\mathcal{R}_{L_{\phi,1}}(f) - \mathcal{R}^*_{L_{\phi,1}}}{\inf_{x\in\mathcal{X}} c_1(x)},$$

where $c_1(x) \geq 1$ is some function (defined in the proof) taking value in $[2c_\ell c_u^{-1} + 1 - c_\ell - c_u, 2c_u c_\ell^{-1} - 2c_\ell + 1]$; (ii) when Condition 2 holds and the Radon-Nikodym derivative of $F^*_{2|1,Y}(x_2)$ with respect to $F^0_{2|1,Y}(x_2)$ is bounded by M_h,

$$\psi\left\{\frac{\mathcal{R}(f) - \mathcal{R}^*}{\sup_{x\in\mathcal{X}} c_2(x)}\right\} \leq \frac{\mathcal{R}_{L_{\phi,2}}(f) - \mathcal{R}^*_{L_{\phi,2}}}{\inf_{x\in\mathcal{X}} c_2(x)},$$

where $c_2(x)$ is some function (defined in the proof) taking value in $[1, 1+M_h(1-c_\ell)]$.

Remark 4.3. *(i) When the derivative of $F^*_{2|1,Y}(x_2)$ is bounded and the derivative of $F^0_{2|1,Y}(x_2)$ is greater than a positive constant, the Radon-Nikodym derivative of $F^*_{2|1,Y}(x_2)$ with respect to $F^0_{2|1,Y}(x_2)$ is bounded. (ii) The properties of double robustness (Theorem 3.1), Fisher consistency (Theorem 4.1), and excess risk (Theorem 4.2) hold for general regularized empirical risk minimization, such as neural network, in the presence of missing covariates with MAR mechanism.*

For general f, Theorem 4.2 provides a bound of the excess risk $\mathcal{R}(f) - \mathcal{R}^*$ in terms of the excess risks $\mathcal{R}_{L_{\phi,1}}(f) - \mathcal{R}^*_{L_{\phi,1}}$ and $\mathcal{R}_{L_{\phi,2}}(f) - \mathcal{R}^*_{L_{\phi,2}}$, respectively. Next, we show the upper bound of the excess risk when the kernel machine $\widehat{f}_\phi$ in (3.11) is used.

To this end, we need some additional notations and assumptions. Denote the excess risk of the population-version kernel machine in (3.10) with respect to the Bayes risk by

$$a(\lambda) = \inf_{f\in\mathcal{H}}\left\{\lambda\|f\|^2_{\mathcal{H}} + \mathcal{R}_{L^*_\phi}(f)\right\} - \inf_{f\in\mathcal{H}} \mathcal{R}_{L^*_\phi}(f).$$

Clearly, $a(\lambda) \to 0$ as $\lambda \to 0$.

Let B denote the closed unit ball of $\mathcal{H}$. Let $M = \{(4 + c_u)c_\ell^{-1}\}^{1/2}$. Let $B_{\mathcal{H}}(\lambda) = \lambda^{-1/2}MB$ denote the ball with the radius $\lambda^{-1/2}M$. Define the empirical L_2 norm of $B_{\mathcal{H}}$ by

$$\|f - g\|_{L_2(\mathbb{P}_n)} = \{\mathbb{P}_n|f(X) - g(X)|^2\}^{1/2} \quad \text{for } f, g \in B_{\mathcal{H}}. \tag{4.2}$$

By the density of universal RKHS, define the $L_2(\mathbb{P}_n)$ ϵ-balls around $g \in B_{\mathcal{H}}$ as the set $\left\{f \in B_{\mathcal{H}} : \|f - g\|_{L_2(\mathbb{P}_n)} < \epsilon\right\}$. For $\epsilon > 0$, the covering number of $B_{\mathcal{H}}$ with respect to $L_2(\mathbb{P}_n)$, denoted by $N(B_{\mathcal{H}}, \epsilon, L_2(\mathbb{P}_n))$, is the smallest number of $L_2(\mathbb{P}_n)$ ϵ-balls needed to cover $B_{\mathcal{H}}$ (Zhao et al., 2015).

Assumption 5. *For $p \in (0, 2]$, there exists a constant $C_{2,p}$ depending on p such that $\sup_{\{x_1,\ldots,x_n\}\in\mathcal{X}^n} \log N(B, \epsilon, L_2(\mathbb{P}_n)) \leq C_{2,p}\epsilon^{-p}$, where $\mathbb{P}_n$ is the empirical measure obtained by observing $x_1, \ldots, x_n$ and $L_2(\mathbb{P}_n)$ is the empirical L_2 norm defined as in (4.2).*

Remark 4.4. *By Corollary 9.5 of Kosorok (2008), Assumption 5 holds for any Vapnik-Červonenkis classes (Kosorok, 2008, Sect. 9.1.1) of measurable functions. Assumption 5 is also satisfied by the Gaussian RBF kernel with $k(x, x') =$*

$\exp(-\sigma_n^2\|x-x'\|^2)$ where $\sigma_n > 0$ is the bandwidth parameter. By Theorem 2.1 of Steinwart and Scovel (2007), for any $\epsilon > 0$, $\sup_{\mathbb{P}_n} \log N\left(B_{\mathcal{H}}, \epsilon, L_2(\mathbb{P}_n)\right) \leq c_{p,\delta,d}\sigma_n^{(1-p/2)(1+\delta)d}\epsilon^{-p}$, where $0 < p \leq 2$, $\delta > 0$, and $c_{p,\delta,d}$ is a constant depending on p, δ, and d.

Assumption 6. *There exist two positive constants ρ_1 and ρ_2 such that $|\widehat{F}_{2|1,Y}\left(x_2\right) - F^*_{2|1,Y}\left(x_2\right)| = \mathrm{O}_p(n^{-\rho_1})$ uniformly, i.e.,*

$$\sup_{x_1\in\mathcal{X}_1, x_2\in\mathcal{X}_2, y\in\mathcal{Y}} |\widehat{F}_{2|1,Y}\left(x_2\right) - F^*_{2|1,Y}\left(x_2\right)| = \mathrm{O}_p(n^{-\rho_1}),$$

and $|\widehat{\pi}(x_1,y) - \pi^*(x_1,y)| = \mathrm{O}_p(n^{-\rho_2})$ uniformly, i.e.,

$$\sup_{x_1\in\mathcal{X}_1, y\in\mathcal{Y}} |\widehat{\pi}(x_1,y) - \pi^*(x_1,y)| = \mathrm{O}_p(n^{-\rho_2}).$$

Assumption 6 is stronger than Assumption 2. It is used to derive the upper bound of the excess risk and the universal consistency. The existence of the constants can be warranted by using an order-preserving nonparametric estimator and appropriate choices of smooth kernel function and bandwidths after Hall and Müller (2003, Theorem 3.4).

Assumption 7. *Suppose that $\phi(t)$ in Assumption 3 is locally Lipschitz continuous, i.e., for any $\beta \geq 0$, there exist constants $C_\phi(\beta)$ such that $|\phi(t) - \phi(t')| \leq C_\phi(\beta)|t-t'|$.*

Remark 4.5. *Assumption 7 holds for the hinge loss, the quadratic loss, the logistic loss, and the exponential loss.*

Define

$$C_{L_\phi}(\beta) = \frac{(2c_u+4)C_\phi(\beta)}{c_l}. \tag{4.3}$$

Theorem 4.3. *Under Assumptions 1–7, when either Condition 1 or Condition 2 holds, for any $b > 0$, with probability no less than $1 - e^{-2b}$,*

$$\frac{1}{\inf_{x\in\mathcal{X}} c_j(x)}\psi\left\{\frac{\mathcal{R}(\widehat{f}_\phi) - \mathcal{R}^*}{\sup_{x\in\mathcal{X}} c_j(x)}\right\}$$
$$\leq a(\lambda) + \mathrm{O}_p\{C_\phi(\lambda^{-1/2})\lambda^{-1/2}n^{-\min(\rho_1,\rho_2)}\} + \epsilon_{n,\lambda,b}$$

where $j = 1, 2$, and

$$\begin{aligned}
&\epsilon_{n,\lambda,b}\\
&= \max\left[12C_{1,\lambda}c_p \max\left\{(C_{1,\lambda}{}^{-2}c_\lambda\varepsilon)^{1/2-p/4}\left(\frac{C_{2,p}}{n}\right)^{1/2}, \left(\frac{C_{2,p}}{n}\right)^{2/(2+p)}\right\},\right.\\
&\qquad\left.\frac{36cQb}{n}, \frac{8QC_1b}{n}\right],
\end{aligned}$$

$C_{1,\lambda} = C_{L_\phi}(\lambda^{-1/2})M\lambda^{-1/2} + 2M^2$, c_p is a constant depending p, $C_{2,p}$ is the constant defined in Assumption 5, $c_\lambda = \frac{2}{\lambda}\{C_{L_\phi}(\lambda^{-1/2}) + 2M\lambda^{1/2}\}^2$; ε is an arbitrarily small enough positive constant, C_ϕ is Lipschitz constants defined in Assumption 7, C_{L_ϕ} is defined in (4.3), and Q is an absolute constant.

The universal consistency of the doubly robust kernel machine $\widehat{f}_\phi$ is discussed in the next theorem.

Theorem 4.4. *Suppose Assumptions 1–7 hold. Further suppose the constant $C_{L_\phi}(\beta)$ in (4.3) is bounded by $\delta\beta^q$ for some $q > 0$ and $\delta \geq 1$ and the tuning parameter λ in (3.11) satisfies $\lambda \to 0$ and $\lambda^{(q+2)/2}n^{\min(\rho_1,\rho_2)} \to \infty$. Let ψ, defined in (4.1), be increasing in $[0, \infty)$. Then, whenever either Condition 1 or Condition 2 holds, for any $b > 0$, with probability no less than $1 - e^{-2b}$,*

$$\mathcal{R}(\widehat{f}_\phi) - \mathcal{R}^* \longrightarrow 0.$$

Remark 4.6. *(i) When ϕ is either the hinge loss, the quadratic loss, the logistic loss, or the exponential loss, ψ increases in $[0, \infty)$. (ii) By Theorems 4.2 and 4.3, we conclude that the proposed convex augmented loss is indeed a calibrated loss in the presence of missing covariates. (iii) In Theorem 4.4, we claim the consistency of $\widehat{f}$ through excess risk instead of deriving the bound of $\|\widehat{f} - f^*\|_2$ directly is because even if by bounding $\|\widehat{f} - f^*\|_2$ under appropriate assumptions, we can conclude $\widehat{f}$ converges to f^*. However, for the classification loss $I[Y\mathrm{sign}\{f(X)\} \leq 0]$, the convergence of $\widehat{f}$ cannot guarantee $\mathcal{R}_I(\widehat{f})$ converges to $\mathcal{R}_I(f^*) = \mathcal{R}^*$ (due to the discontinuity of classification loss).*

5. Simulation

We conduct simulation studies to compare the finite-sample performance of the proposed kernel-machine methods with some existing methods in terms of classification error.

We denote the seven competing methods as follows.

CC: The kernel machine that use only the complete observations without any partially observed subjects.

Full: The kernel machine that use the full observations with the missing covariates assumed to be known from the generating model, i.e., the oracle method.

$\mathrm{IPT_a}$: The kernel machine with the missing covariates imputed by the sample mean.

$\mathrm{IPT_k}$: The kernel machine with the missing covariates imputed by the sample mean from the k nearest neighbors.

$\mathrm{IPT_m}$: The kernel machine with the missing covariates imputed by multiple imputations.

WCC: The proposed weighted-complete-case kernel machine presented in (2.2).

DR: The proposed doubly robust kernel machine presented in (3.11).

For all kernel machines, we use the Gaussian RBF kernel. The tuning parameter λ and the Gaussian RBF kernel width parameter are chosen by (the five-fold) cross validation where λ is from $\{1/1000, 1/100, 1/10, 1, 10\}$ and the kernel width parameter is from $\{1/100, 1/10, 1, 10\}$. The hinge loss is used as the surrogate loss in our proposed methods. The last three methods (IPT_m, WCC, DR) involve estimation of the missing mechanism $\mathrm{P}(R = 1|X_1, Y)$ and/or multiple imputations of the missing covariates X_2.

The probability that an observation is fully observed is generated using logistic regression of R on X_1 and Y. We estimate the propensity score using either logistic regression (under a correctly-specified model) or probit regression (under a misspecified model). The missing values X_2 are imputed either from the multivariate normal distribution or from the regression of X_2 on X_1 and Y. For the former imputation, the procedure works as follows. We assume that given X_1 and Y, the conditional distribution of X_2 is multivariate normal. Then, we obtain the estimated distribution $\widehat{F}_{2|1,Y}(x_2)$ by replacing the mean vector and covariance matrix by their MLEs. At last, we use $\widehat{F}_{2|1,Y}(x_2)$ to generate m independent samples $\{X_{2j|1,Y}^{\mathrm{imp}} : j = 1, \dots, m\}$ to form $\{X_j^{\mathrm{imp}} = (X_1^\intercal, X_{2j|1,Y}^{\mathrm{imp}^\intercal})^\intercal : j = 1, \dots, m\}$. An alternative way to estimate the imputation model is discussed in Appendix D.

We consider three classification models with missing covariates in Table 1, where the actual generating model $\mathrm{P}(Y = 1|X)$ and the missing mechanism $\mathrm{P}(R = 1|X_1, Y)$ are specified in detail.

Under the MLE imputation, denote the two methods of estimating the propensity score by logistic regression and probit regression by E1 and E2, respectively. Under the regression imputation, denote the two methods of estimating the propensity score by logistic regression and probit regression by E3 and E4, respectively. The missingness indicator is generated through the logistic model. Thus, Condition 2 holds when we apply methods E1 and E3.

We set the training sample size n to be 100, 200, and 400, respectively, and the number of imputations m in (3.4) to be 5. Let $\{(X_i, Y_i) : i = 1, \dots, N\}$ denote a generic testing sample of size $N = 10,000$. In the simulation, the test instances are fully observed. In reality, when the covariates X_2 are missing in some instances, we can impute them by X_2^{imp} based on the estimated conditional distribution $\widehat{F}_{2|1,Y}(x_2)$. Let $\widehat{Y}_i$ denote the generic estimate of Y_i obtained by the competing method based on common training data and X_i. Denote the corresponding empirical classification error by $\widetilde{\mathcal{R}} = N^{-1}\sum_{i=1}^N (Y_i - \widehat{Y}_i)^2$. Finally, we set the number of replications to be 100. The implementation of both the WCC and the DR algorithms, as well as the dataset generation can be found in the R package `drkm4mc`.

Table 2 reports the sample mean, median, and standard deviation of $\widetilde{\mathcal{R}}$ over 100 replications obtained by the competing methods for the three data generating models under various sample sizes and different estimation methods of the missing mechanism and imputation (E1–E4). The corresponding distributions of $\widetilde{\mathcal{R}}$ are displayed by boxplots in Figs. 1–3. The boxplots of IPT_a and IPT_k are omitted since they show nearly uniform inferiority to that of IPT_m

Table 1. *Specification of three data generating models with missing covariates X_2 in terms of the actual generating model $\mathrm{P}(Y = 1|X)$ and missing mechanism $\mathrm{P}(R = 1|X_1, Y)$, where $\mathrm{logit}(x) = \log\{x/(1-x)\}$, $\mathbf{0}$ and $\mathbf{1}$ denote vectors of zeros and ones respectively, $\Omega_p = (\omega_{ij})$ with $\omega_{ij} = I(i = j) + |i-j|^{-1} I(i \neq j)$, $U^* = \{U - \mathrm{E}(U)\}/\mathrm{sd}(U)$. For model 2, $U = z_1 + z_2$, $\mathrm{E}(U) = 1.3273$, $\mathrm{sd(U)} = 0.7840$; for model 3, $U = z_1 + \cdots + z_6$, $\mathrm{E}(U) = 1.4659$, $\mathrm{sd(U)} = 0.5306$.*

model	X	X_2	$\mathrm{logit}(\mathrm{P}(Y = 1\|X))$	$\mathrm{logit}(\mathrm{P}(R = 1\|X_1, Y))$
1	$\mathbf{z}_2 \sim N(\mathbf{0}, I_2)$	z_1	$\mathrm{logit}\{\Phi(5(z_1^2 - z_2^2) + 1)\}$	$-\frac{2}{3} + 4z_1Y$
2	$\mathbf{z}_{10} \sim N(\Omega_{10}\mathbf{1}_{10}, \Omega_{10}^2)$	z_9, z_{10}	$\frac{1}{8}\{6.3 + z_{10} - \sum_{i=1}^{9}(z_i - 1)^2\}$	$-\frac{1}{2} + 3(Y+1)U^*$
3	$\mathbf{z}_{17} \sim N(\Omega_{17}\mathbf{1}_{17}, \Omega_{17}^2)$ $z_{18} = z_6 + z_7 + z_8 + z_9^2$ $z_{19} = z_{10} + z_{11} + z_{12} + z_{13}^2$ $z_{20} = z_{14} + z_{15} + z_{16} + z_{17}^2$	z_9, z_{13}, z_{17} z_{18}, z_{19}, z_{20}	$-\frac{1}{10}\{20 + \frac{z_{18}+z_{19}+z_{20}}{3} - \sum_{i=1}^{17}(z_i - 1)^2\}$	$-\frac{1}{2} + YU^*$

Table 2. *Sample median, mean, and standard deviation of the empirical risk* $\widetilde{\mathcal{R}}$ *over 100 replications obtained by the seven competing methods under the three data generating models, where the superscript indicates the method of estimating missing mechanism and imputation, e.g., WCC*^{1} *stands for the WCC method under E1. Note that by design,* $\mathrm{IPT}^1_{\mathrm{m}}$ *coincides with* IPT^2_m*;* IPT^3_m *coincides with* IPT^4_m*; WCC*^{1} *coincides with WCC*^{3} *and WCC*^{2} *coincides with WCC*^{4}*.*

model	n		CC	$\mathrm{IPT}_{\mathrm{a}}$	$\mathrm{IPT}_{\mathrm{k}}$	$\mathrm{IPT}^1_{\mathrm{m}}$	$\mathrm{IPT}^3_{\mathrm{m}}$	WCC^1	WCC^2	DR^1	DR^2	DR^3	DR^4	Full
1	100	median	1.66	1.30	1.29	1.09	1.09	1.66	1.66	0.94	0.96	0.94	0.89	0.39
		mean	1.67	1.30	1.31	1.10	1.09	1.53	1.52	0.96	1.03	1.02	0.97	0.40
		std	0.14	0.22	0.25	0.18	0.18	0.29	0.30	0.31	0.37	0.32	0.33	0.06
	200	median	1.66	1.30	1.29	0.92	0.91	1.50	1.48	0.72	0.76	0.71	0.78	0.34
		mean	1.65	1.32	1.30	0.90	0.90	1.41	1.39	0.77	0.79	0.77	0.82	0.35
		std	0.11	0.18	0.21	0.17	0.16	0.30	0.31	0.24	0.27	0.25	0.27	0.04
	400	median	1.66	1.26	1.34	0.80	0.80	1.29	1.34	0.62	0.64	0.68	0.64	0.31
		mean	1.64	1.26	1.32	0.79	0.80	1.26	1.28	0.65	0.68	0.70	0.69	0.31
		std	0.11	0.16	0.18	0.13	0.13	0.35	0.36	0.18	0.19	0.22	0.21	0.02
2	100	median	1.90	1.64	1.79	1.70	1.69	1.73	1.75	1.64	1.64	1.64	1.64	1.64
		mean	1.96	1.69	1.79	1.72	1.71	1.78	1.78	1.72	1.70	1.71	1.71	1.68
		std	0.29	0.09	0.12	0.09	0.07	0.17	0.17	0.13	0.11	0.13	0.12	0.08
	200	median	1.88	1.64	1.76	1.65	1.66	1.68	1.65	1.64	1.64	1.64	1.64	1.64
		mean	1.90	1.69	1.76	1.67	1.67	1.73	1.72	1.67	1.66	1.67	1.67	1.67
		std	0.21	0.09	0.11	0.08	0.08	0.12	0.11	0.07	0.07	0.08	0.06	0.09
	400	median	1.90	1.64	1.79	1.64	1.64	1.64	1.65	1.64	1.64	1.64	1.64	1.63
		mean	1.92	1.68	1.77	1.64	1.64	1.72	1.73	1.65	1.65	1.66	1.66	1.62
		std	0.18	0.09	0.10	0.04	0.04	0.14	0.16	0.04	0.04	0.06	0.05	0.05
3	100	median	1.83	1.44	1.62	1.51	1.51	1.75	1.77	1.44	1.44	1.44	1.44	1.44
		mean	1.88	1.48	1.63	1.53	1.53	1.76	1.77	1.46	1.45	1.46	1.45	1.47
		std	0.28	0.09	0.14	0.09	0.10	0.29	0.30	0.07	0.05	0.08	0.06	0.08
	200	median	1.82	1.44	1.66	1.44	1.48	1.71	1.70	1.44	1.44	1.44	1.44	1.44
		mean	1.86	1.46	1.65	1.46	1.49	1.71	1.70	1.44	1.44	1.44	1.44	1.44
		std	0.19	0.07	0.11	0.05	0.07	0.23	0.21	0.00	0.00	0.02	0.01	0.06
	400	median	1.77	1.41	1.62	1.41	1.43	1.66	1.64	1.44	1.44	1.44	1.44	1.39
		mean	1.78	1.42	1.61	1.41	1.43	1.66	1.65	1.44	1.44	1.44	1.44	1.40
		std	0.09	0.05	0.08	0.03	0.05	0.15	0.16	0.00	0.00	0.00	0.00	0.05

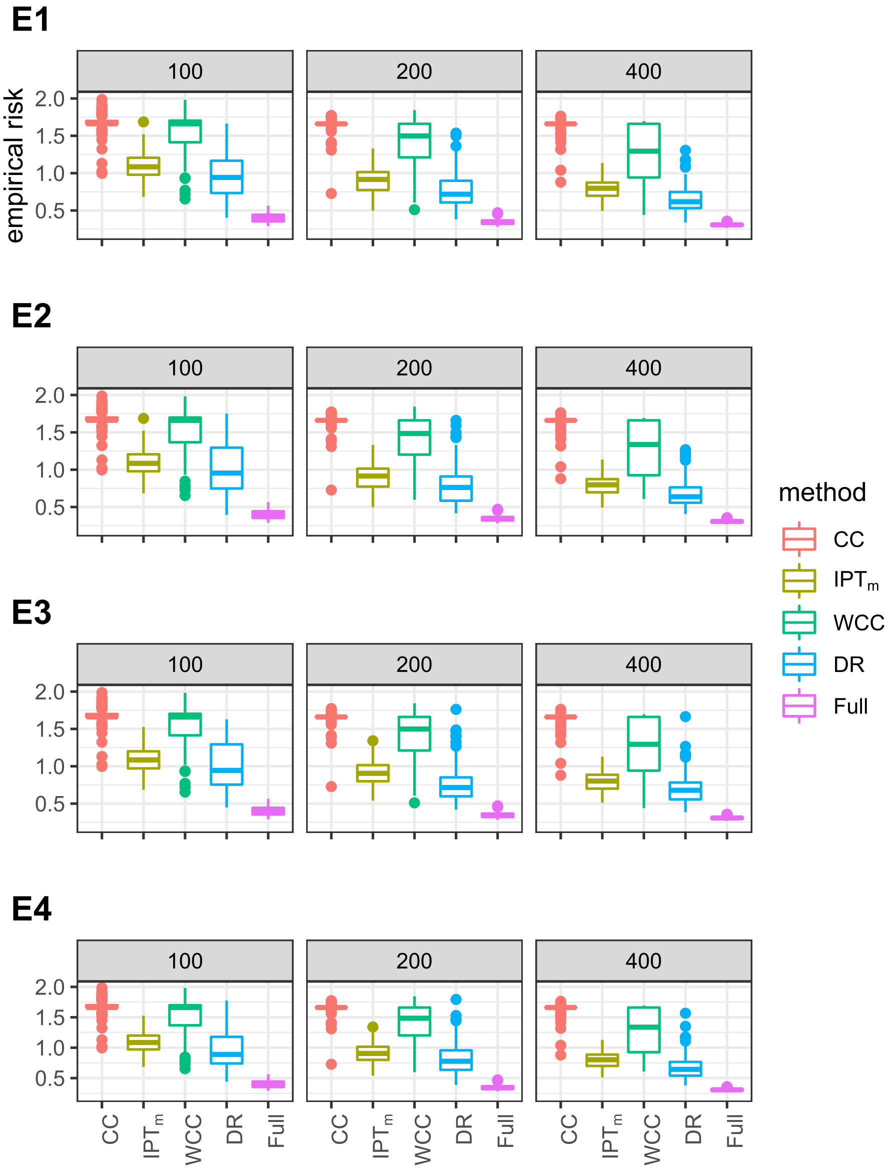

Fig 1. *Box plots of $\widetilde{\mathcal{R}}$ obtained by the five competing methods (CC, IPT$_{\rm m}$, WCC, DR and Full) under model 1.*

as reflected in Table 2. The results of the kernel machines that use only the complete observation (CC) method and the oracle (Full) method remain the same across the four estimation methods since they do not involve missing values and imputation.

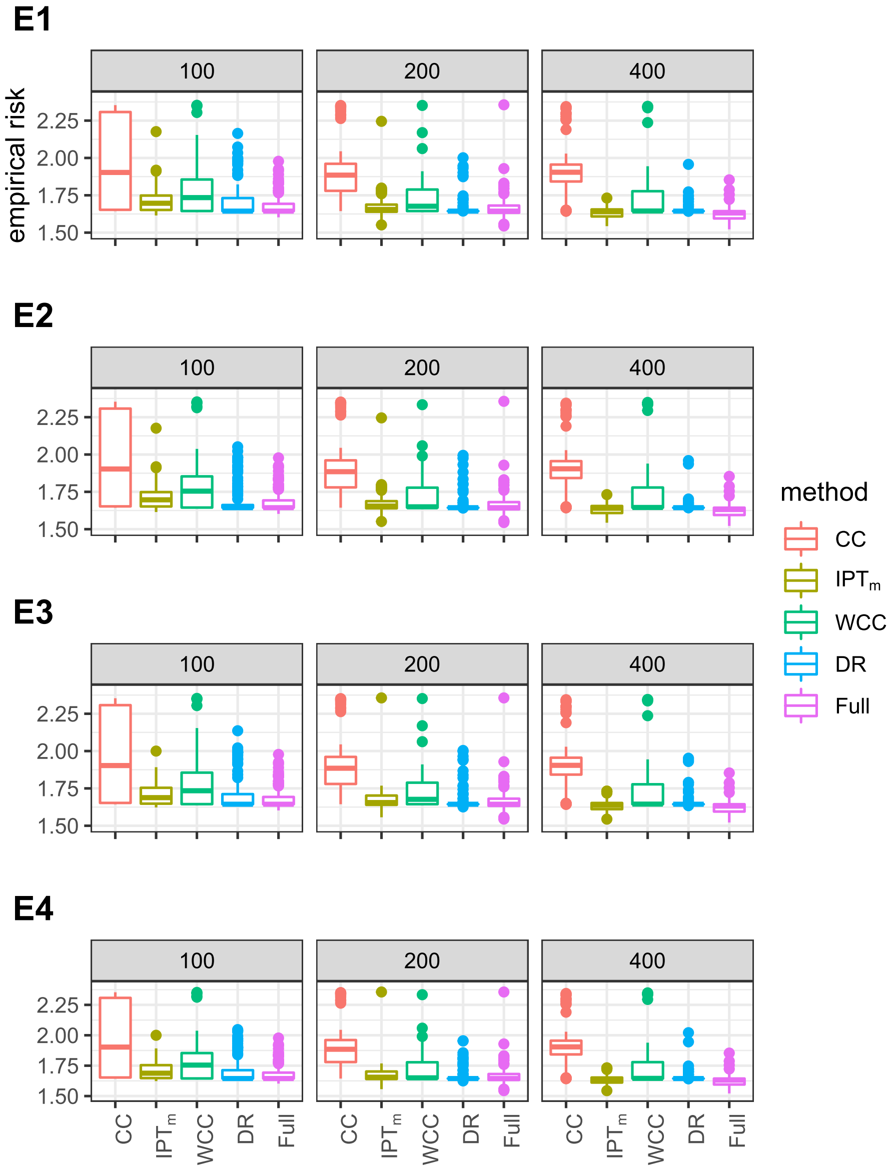

FIG 2. *Box plots of $\widetilde{\mathcal{R}}$ obtained by the five competing methods (CC, IPT$_{\mathrm{m}}$, WCC, DR and Full) under model 2.*

It is seen that the doubly robust (DR) kernel machine method performs the best in almost all cases regardless of the modeling of the missing mechanism and imputation. The multiple imputation method (IPT$_{\mathrm{m}}$) performs the second best. The only exception that IPT$_{\mathrm{m}}$ performs better than DR occurs in model 2 when

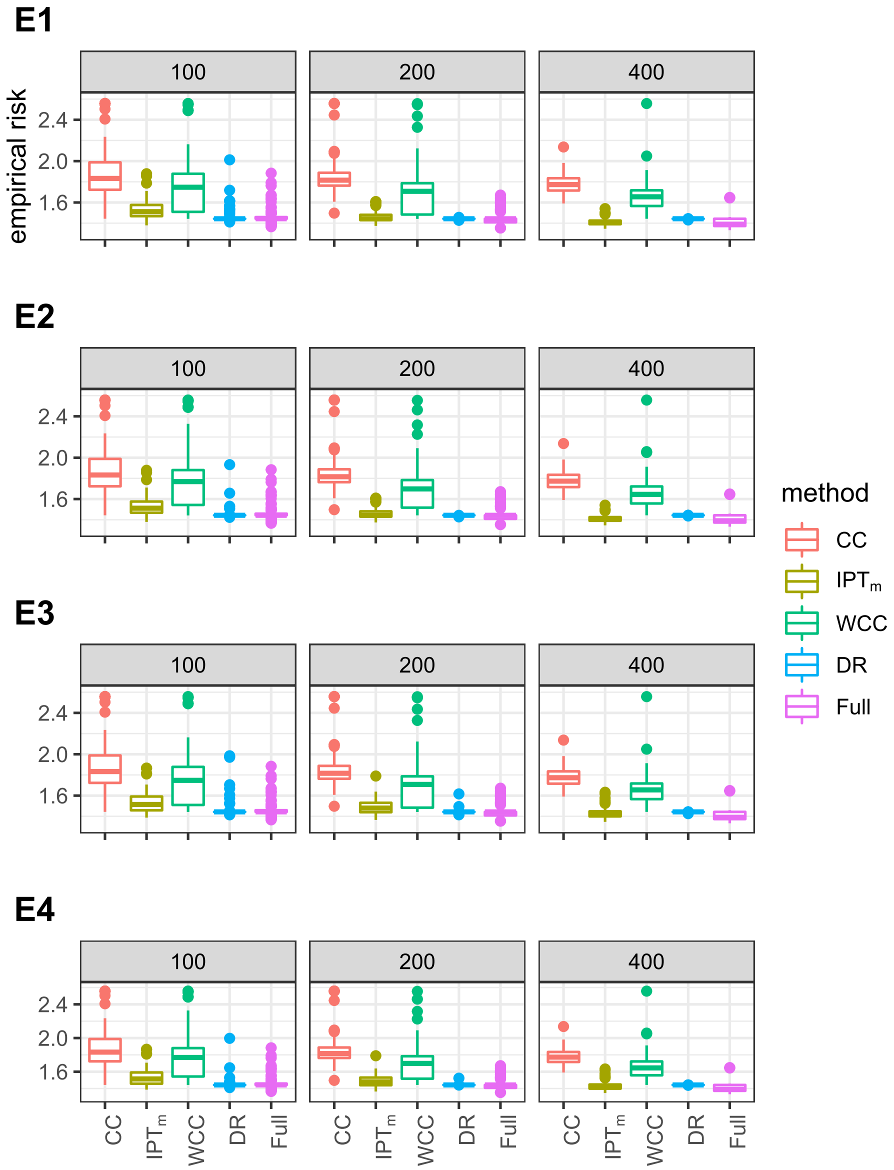

Fig 3. *Box plots of $\widetilde{\mathcal{R}}$ obtained by the five competing methods (CC, IPT$_{\rm m}$, WCC, DR and Full) under model 3.*

$n = 400$. Overall, the proposed DR method exhibits the desired robustness in regard to the choice of methods used in estimating the missing mechanism and imputing missing covariates.

Table 3

Description of the responses and covariates of three real data sets and specification of the missing mechanism of X_2 of the training sets, where z_1^*, z_2^* and z_3^* are the standardized variables of z_1, z_2 and z_3, respectively.

data	$(n,\ n_0)$	response and covatiates	$\text{logit}\{\text{P}(R=1\|X_1,Y)\}$
1	(748, 500)	$Y=1$: donating blood	
		z_1: months since last donation	
		z_2: months since first donation	
		z_3: total number of donations	
		$X_2 = z_3$	$-1 - z_2^* Y$
2	(345, 200)	$Y=1$: in training set	
		z_1: mean corpuscular volume	
		z_2: alkaline phosphotase	
		z_3: alanine aminotransferase	
		z_4: aspartate aminotransferase	
		z_5: glutamyl transpeptidase	
		z_6: alcohol drank per day	
		$X_2 = (z_1, z_6)$	$-\frac{1}{6} - 3z_3^* Y$
3	(310, 200)	$Y=1$: abnormal	
		z_1: pelvic incidence	
		z_2: pelvic tilt	
		z_3: lumbar lordosis angle	
		z_4: sacral slope	
		z_5: lumbar lordosis angle	
		z_6: grade of spondylolisthesis	
		$X_2 = (z_4, z_5)$	$2 - \frac{11}{2}(z_1^* + z_2^*)Y$

6. Real data application

We apply the proposed methods to three classification data sets regarding blood transfusion, liver disorders, and the vertebral column, respectively, which are publicly available from the UCI machine learning repository (www.ics.uci.edu/~mlearn/MLRepository.html). The three datasets and the code that generate the missing data appear in the package `drkm4mc`. The descriptions of the data are given in Table 3, where the proportions of positive response ($Y = 1$) are about 24%, 41%, and 67%, respectively.

We randomly divide each data into a training set of size n_0 (of about two thirds of n) and a testing set of size $n - n_0$. Further, we specify the missing mechanism for the covariates X_2 in the last column of Table 3 through some logit models under which the missing rates are about 30%, 45% and 55%, respectively.

We then apply the proposed methods to these data sets. For all three datasets, we use the linear regression of X_1 to impute the missing values of X_2. For the third data, since the partially observed $X_2 = \{z_4, z_5\}$ appears to follow a bivariate normal distribution, we also use the bivariate normal variables to impute the missing covariates. Thus, the estimation methods E1 and E2 are also adopted for this data.

Table 4 reports the sample mean, median, and standard deviation of $\widetilde{\mathcal{R}}$ (over 100 replications) obtained by the different methods. Figure 4 shows the

Table 4. *Sample median, mean, and standard deviation of the empirical risk* $\widetilde{\mathcal{R}}$ *over 100 replications obtained by the seven competing methods for the three real data sets. The results under* $\mathrm{IPT}^1_{\mathrm{m}}$ *and* DR^1 *and* DR^2 *are not available for the first two data sets since the estimation methods E1 and E2 are not considered.*

data		CC	$\mathrm{IPT}_{\mathrm{a}}$	$\mathrm{IPT}_{\mathrm{k}}$	$\mathrm{IPT}^1_{\mathrm{m}}$	$\mathrm{IPT}^3_{\mathrm{m}}$	WCC^1	WCC^2	DR^1	DR^2	DR^3	DR^4	Full
1	median	1.26	1.28	1.23	–	1.05	0.96	0.97	–	–	0.95	0.95	0.87
	mean	1.24	1.26	1.22	–	1.06	0.96	0.97	–	–	0.95	0.95	0.88
	std	0.15	0.18	0.20	–	0.14	0.12	0.12	–	–	0.09	0.08	0.09
2	median	1.64	1.47	1.54	–	1.46	1.64	1.64	–	–	1.44	1.44	1.31
	mean	1.63	1.46	1.53	–	1.47	1.63	1.64	–	–	1.46	1.46	1.30
	std	0.16	0.16	0.17	–	0.14	0.15	0.15	–	–	0.19	0.19	0.14
3	median	1.05	0.95	0.87	0.63	0.75	1.09	1.07	0.74	0.75	0.73	0.73	0.65
	mean	1.06	0.96	0.87	0.63	0.75	1.08	1.07	0.75	0.76	0.73	0.74	0.65
	std	0.14	0.16	0.17	0.10	0.16	0.22	0.21	0.17	0.18	0.17	0.17	0.12

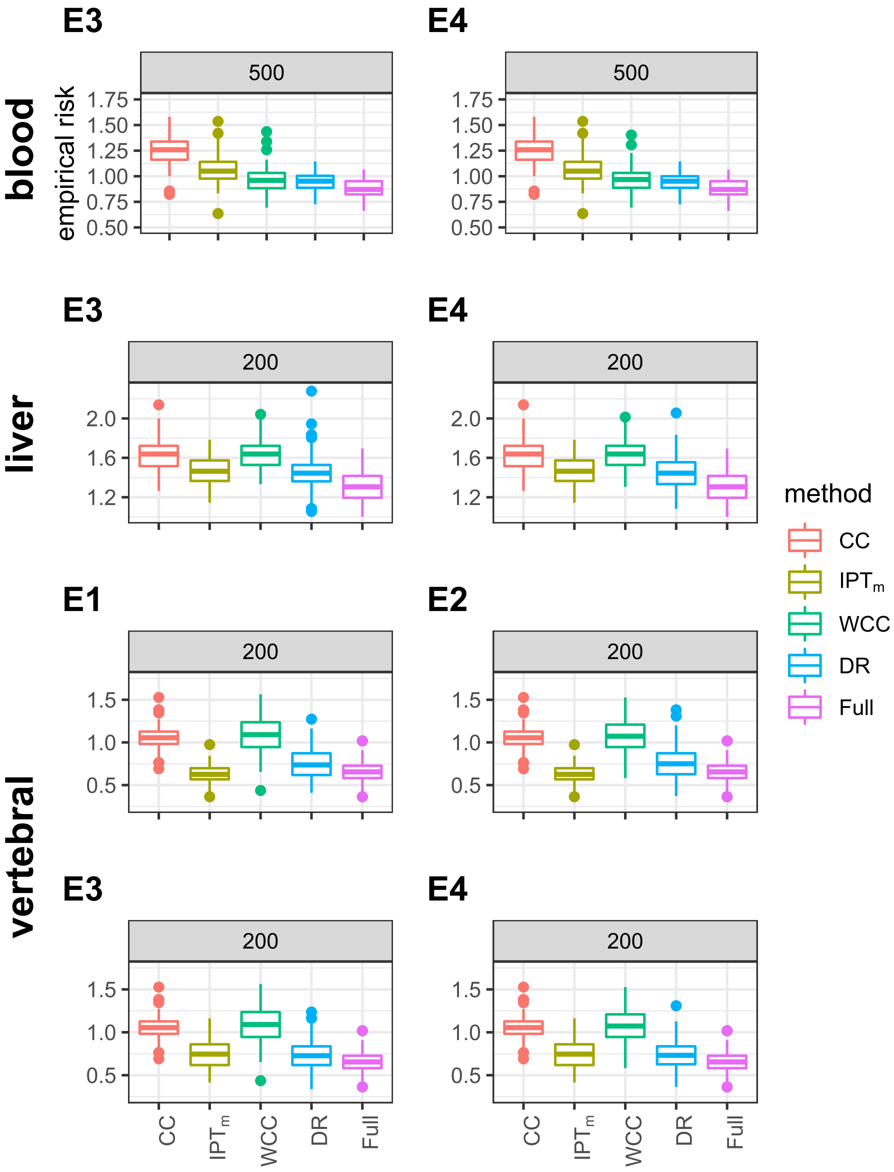

FIG 4. *Box plots of $\widetilde{\mathcal{R}}$ obtained by five competing methods for the three real datasets under various estimations of the missing mechanism.*

corresponding boxplots of $\widetilde{\mathcal{R}}$. The results of $\mathrm{IPT_a}$ and $\mathrm{IPT_k}$ are again omitted since they are uniformly inferior to that of $\mathrm{IPT_m}$. It is seen that (i) for the first data, the proposed DR method performs best under both E3 and E4. The WCC method performs better than IPT_m. (ii) For the second data, the DR method is comparable with IPT_m under both E3 and E4. (iii) For the third data, the DR method performs best under both E3 and E4. While the IPT_m performs best under both E1 and E2 when the missing data are imputed from a multivariate normal distribution.

The findings are consistent with the results in the simulation section. The proposed doubly robust kernel machine method is recommended when the missing covarites are imputed by a regression model. It can serve as a good alternative to the multiple imputation method when the missing covariates follow from a multivariate normal distribution.

7. Concluding remarks

We developed two kernel machines for classification in the presence of missing covariates. A novel convex augmented loss function was proposed to obtain the doubly robust kernel machine. Its construction combines the techniques of inverse probability weighting and multiple imputations. Theoretical results regarding Fisher consistency, excess risk, and convergence are established. The proposed doubly robust kernel machine is recommended in general after simulation comparison with some existing methods.

We would like to note that the algorithm we develop for calculating the doubly robust estimator involves multiple imputations within the loss and, therefore, is time consuming. It is of interest to pursue a faster version. Other directions for future work include extensions to the regression model with the continuous response and to handle missing covariates with the non-monotonic pattern.

Appendix A: Computation details of the doubly robust kernel machine $\widehat{f}_\phi$ (3.11) in Sect. 3.2

Write the original data as $\{(R_i, X_i, Y_i) : i = 1, \ldots, n\}$. When $R_i = 0$, X_{2i} is missing. Write the data with multiple imputation as $\{(R_i, X_{i,j}^{\text{imp}}, Y_i) : i = 1, \ldots, n,\ j = 1, \ldots, m\}$, where $X_{i,j}^{\text{imp}} = (X_{1i}, X_{2j|1i,Y_i}^{\text{imp}})$.

The empirical risk of $\mathcal{R}_{L_\phi^*}(f)$ is

$$\begin{aligned}
&\mathcal{R}_{L_\phi^{\text{imp}},D}(f)\\
=&\frac{1}{n}\sum_{i=1}^{n}\left[\left|W_1\left(\widehat{\pi}\right)\right|\phi\left\{\text{sign}(W_1)f(x_i)\right\}+\left|W_{-1}\left(\widehat{\pi}\right)\right|\phi\left\{-\text{sign}(W_{-1})f(x_i)\right\}\right]\\
&+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}[\left|V_1\left(\widehat{\pi}\right)\right|\phi\{\text{sign}(V_1)f(x_{i,j}^{\text{imp}})\}+\left|V_{-1}\left(\widehat{\pi}\right)\right|\phi\{-\text{sign}(V_{-1})f(x_{i,j}^{\text{imp}})\}].
\end{aligned}$$

Let $N = n(m+1)$. For computational convenience, we will re-order the nm records of the triplet data with multiple imputation by a single index with $i = n+1, \ldots, n(m+1)$.

Then, $\widehat{f}_\phi = \arg\min_{f\in\mathcal{H}} \lambda\|f\|^2_{\mathcal{H}} + \mathcal{R}_{L^{\mathrm{imp}}_\phi, D}(f)$ in (3.11) can be expressed as

$$\begin{aligned}
&\widehat{f}_\phi \\
&= \arg\min_{f\in\mathcal{H}} \Bigg\{ \frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i}{\widehat{\pi}} I\left(Y_i = 1\right)\phi\{f(X_i)\} + \frac{R_i}{\widehat{\pi}} I\left(Y_i = -1\right)\phi\{-f(X_i)\}\right] \\
&+ \frac{1}{nm}\sum_{i=n+1}^{n(m+1)} \Bigg(R_i \left[\frac{1-\widehat{\pi}}{\widehat{\pi}} I\left(Y_i = 1\right)\phi\{-f\left(X_i\right)\} + \frac{1-\widehat{\pi}}{\widehat{\pi}} I\left(Y_i = -1\right)\phi\{f\left(X_i\right)\}\right] \\
&\quad + (1-R_i)\left[I\left(Y_i = 1\right)\phi\{f\left(X_i\right)\} + I\left(Y_i = -1\right)\phi\{-f\left(X_i\right)\}\right]\Bigg)\Bigg\} + \lambda\|f\|^2_{\mathcal{H}} \\
&= \arg\min_{f\in\mathcal{H}} \Bigg\{ \frac{1}{N}\sum_{i=1}^{n}(m+1)\left[\frac{R_i}{\widehat{\pi}} I\left(Y_i = 1\right)\phi\{f(X_i)\} + \frac{R_i}{\widehat{\pi}} I\left(Y_i = -1\right)\phi\{-f(X_i)\}\right] \\
&+ \frac{1}{N}\sum_{i=n+1}^{n(m+1)} \frac{N}{nm}\Bigg(R_i \left[\frac{1-\widehat{\pi}}{\widehat{\pi}} I\left(Y_i = 1\right)\phi\{-f(X_i)\} + \frac{1-\widehat{\pi}}{\widehat{\pi}} I\left(Y_i = -1\right)\phi\{f(x_i)\}\right] \\
&+ (1-R_i)\left[I\left(Y_i = 1\right)\phi\{f(X_i)\} + I\left(Y_i = -1\right)\phi\{-f(X_i)\}\right]\Bigg)\Bigg\} + \lambda\|f\|^2_{\mathcal{H}}.
\end{aligned}$$

Let

$$\mu_i = \begin{cases} (m+1)\frac{R_i}{\widehat{\pi}} I\left(Y_i = 1\right), & i = 1,\dots,n, \\ \frac{N}{nm}\left\{R_i \frac{1-\widehat{\pi}}{\widehat{\pi}} I\left(Y_i = -1\right) + (1-R_i) I\left(Y_i = 1\right)\right\}, & i = n+1,\dots,n(m+1), \end{cases}$$

$$\nu_i = \begin{cases} (m+1)\frac{R_i}{\widehat{\pi}} I\left(Y_i = -1\right), & i = 1,\dots,n; \\ \frac{N}{nm}\left\{R_i \frac{1-\widehat{\pi}}{\widehat{\pi}} I\left(Y_i = 1\right) + (1-R_i) I\left(Y_i = -1\right)\right\}, & i = n+1,\dots,n(m+1). \end{cases}$$

The objective function can be expressed as

$$\frac{1}{N}\sum_{i=1}^{N}[\mu_i\phi\{f(X_i)\} + \nu_i\phi\{-f(X_i)\}] + \lambda\|f\|^2_{\mathcal{H}}. \tag{A.1}$$

(i) We first consider the hinge loss. Then, $\phi(f) = \max(0, 1-f)$ and $\phi(-f) = \max(0, 1+f)$. For the logistic loss and the exponential loss, the derivation is similar (Steinwart and Christmann, 2008, Sect. 11.1).

Since μ_i and ν_i are both nonnegative, (A.1) is a convex function of f. By the representer theorem of Steinwart and Christmann (2008, Theorem 5.5), there exists $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_N)^\intercal$ such that $\widehat{f}_\phi(\cdot) = \sum_{j=1}^N \alpha_j k(\cdot, X_j)$.

Let $C = (2N\lambda)^{-1}$. Minimizing (A.1) is equivalent to minimizing

$$C\sum_{i=1}^{N}(\mu_i\xi_i + \nu_i\eta_i) + \frac{1}{2}\|f\|^2_{\mathcal{H}},$$

subject to

$$\xi_i \ge 0, \quad \xi_i \ge 1 - f(X_i), \quad \eta_i \ge 0, \quad \eta_i \ge 1 + f(X_i), \quad i = 1,\dots,N.$$

Next, introduce the objective function with Lagrange multiplier

$$\begin{aligned}\text{Lagr} = C\sum_{i=1}^{N}(\mu_i\xi_i+\nu_i\eta_i)+\frac{1}{2}\|f\|_{\mathcal{H}}^2+\sum_{i=1}^{N}\gamma_{1i}\{1-f(X_i)-\xi_i\}-\sum_{i=1}^{N}\gamma_{2i}\xi_i \\ +\sum_{i=1}^{N}\gamma_{3i}\{1+f(X_i)-\eta_i\}-\sum_{i=1}^{N}\gamma_{4i}\eta_i, \qquad \text{(A.2)}\end{aligned}$$

where γ_{1i}, γ_{2i}, γ_{3i} and γ_{4i} are all nonnegative for $i=1,\dots,N$.

Write $U=\operatorname{diag}(\mu_1,\dots,\mu_N)$, $V=\operatorname{diag}(\nu_1,\dots,\nu_N)$, $\boldsymbol{\xi}=(\xi_1,\dots,\xi_N)^\intercal$, $\boldsymbol{\eta}=(\eta_1,\dots,\eta_N)^\intercal$, $\Gamma_a=(\gamma_{a1},\dots,\gamma_{aN})^\intercal$ for $a=1,\dots,4$, $K=(k(x_i,x_j))_{N\times N}$. Denote $\mathbf{1}$ and $\mathbf{0}$ as the vectors of ones and zeros, respectively.

We have

$$\begin{aligned}\text{Lagr} =& C\left(U\boldsymbol{\xi}+V\boldsymbol{\eta}\right)^\intercal\mathbf{1}+\frac{1}{2}\boldsymbol{\alpha}^\intercal K\boldsymbol{\alpha}+\Gamma_1^\intercal\left(\mathbf{1}-K\boldsymbol{\alpha}-\boldsymbol{\xi}\right) \\ &-\Gamma_2^\intercal\boldsymbol{\xi}+\Gamma_3^\intercal\left(\mathbf{1}+K\boldsymbol{\alpha}-\boldsymbol{\eta}\right)-\Gamma_4^\intercal\boldsymbol{\eta}. \qquad \text{(A.3)}\end{aligned}$$

Setting the partial derivatives of Lagr with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ to zeros, respectively, we get the following equations

$$\begin{aligned}\frac{\partial\text{Lagr}}{\partial\boldsymbol{\alpha}} &= K\boldsymbol{\alpha}-K\Gamma_1+K\Gamma_3=\mathbf{0},\\ \frac{\partial\text{Lagr}}{\partial\boldsymbol{\xi}} &= CU\mathbf{1}-\Gamma_1-\Gamma_2=\mathbf{0},\\ \frac{\partial\text{Lagr}}{\partial\boldsymbol{\eta}} &= CV\mathbf{1}-\Gamma_3-\Gamma_4=\mathbf{0}.\end{aligned}$$

Solving $\boldsymbol{\alpha}$, U and V by $\Gamma_1,\dots,\Gamma_4$ and substituting them in the primal problem (A.3), we obtain the dual program

$$\begin{aligned}\text{Larg} &= -\frac{1}{2}\left(\Gamma_1-\Gamma_3\right)^\intercal K\left(\Gamma_1-\Gamma_3\right)+\Gamma_1^\intercal\mathbf{1}+\Gamma_3^\intercal\mathbf{1}\\ &= -\frac{1}{2}\begin{pmatrix}\Gamma_1\\ \Gamma_3\end{pmatrix}^\intercal\begin{pmatrix}K & -K\\ -K & K\end{pmatrix}\begin{pmatrix}\Gamma_1\\ \Gamma_3\end{pmatrix}+\begin{pmatrix}\Gamma_1\\ \Gamma_3\end{pmatrix}^\intercal\begin{pmatrix}\mathbf{1}\\ \mathbf{1}\end{pmatrix},\end{aligned}$$

where $\mathbf{0}\le_e\Gamma_1\le_e CU\mathbf{1}$ and $\mathbf{0}\le_e\Gamma_3\le_e CV\mathbf{1}$; the subscript 'e' stands for element-wise. It then reduces to a quadratic optimization with box constraints.

(ii) Second, consider the quadratic loss, i.e., $\phi(f)=(1-f)^2$. Substituting $\phi(f)$ in (A.1), we get

$$\frac{1}{N}\sum_{i=1}^{N}[\mu_i\{1-f(X_i)\}^2+\nu_i\{1+f(X_i)\}^2]+\lambda\|f\|_{\mathcal{H}}^2. \qquad \text{(A.4)}$$

Since $\phi(f)$ and $\phi(-f)$ are both convex, there exists $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_N)^\intercal$ such that $\widehat{f}_\phi(\cdot)=\sum_{j=1}^{N}\alpha_j k(\cdot,X_j)$. After some simple algebra, we express

$$\text{(A.4)}=\frac{1}{N}\{\boldsymbol{\alpha}^\intercal K(U+V)K\boldsymbol{\alpha}-2\boldsymbol{\alpha}^\intercal K(U-V)+\mathbf{1}^\intercal(U^2+V^2)\mathbf{1}\}+\lambda\boldsymbol{\alpha}^\intercal K\boldsymbol{\alpha}.$$

Setting its derivative with respect to $\boldsymbol{\alpha}$ to be the zero vector, we solve

$$\boldsymbol{\alpha} = (U + V + \lambda N I_N)^{-1}(U - V),$$

where I_N is the $N \times N$ identity matrix.

Appendix B: Discussion of the infeasibility of construction of a convex augmented loss with the augmented term obtained by the conditional expectation estimation

The conditional expectation estimation approach is a common method to construct the augment term in the AIPW literature. Here, we reason the infeasibility of this approach for the classification problem in the presence of missing covariates.

First, we recap our approach. The augmented loss function in (3.6) is

$$\begin{aligned} L_{\text{aug}}\left(\pi^{\text{g}}, X^{\text{g}}, f\right) =& W_1(\pi^{\text{g}}) I[\text{sign}\{f(X)\} \le 0] + W_{-1}(\pi^{\text{g}}) I[-\text{sign}\{f(X)\} \le 0] \\ &+ V_1(\pi^{\text{g}}) I[\{\text{sign}\{f(X^{\text{g}})\} \le 0] + V_{-1}(\pi^{\text{g}}) I[\{-\text{sign}\{f(X^{\text{g}})\} \le 0]. \end{aligned}$$

Based on (3.6), we introduce a new loss (3.7) given by

$$\begin{aligned} & L_{\text{abs}}(\pi^{\text{g}}, X^{\text{g}}, f) \\ =& W_1(\pi^{\text{g}}) I[\text{sign}\{f(X)\} \le 0] + W_{-1}(\pi^{\text{g}}) I[-\text{sign}\{f(X)\} \le 0] \\ +& |V_1(\pi^{\text{g}})| I[\text{sign}(V_1)\text{sign}\{f(X^{\text{g}})\} \le 0] + |V_{-1}(\pi^{\text{g}})| I[-\text{sign}(V_{-1})\text{sign}\{f(X^{\text{g}})\} \le 0]. \end{aligned}$$

Then, minimization with respect to L_{aug} is equivalent to minimization with respect to L_{abs}. Replacing the 0-1 loss by the convex surrogate loss ϕ, we obtain a convex augmented loss (3.9) given by

$$\begin{aligned} L_\phi(\pi^{\text{g}}, X^{\text{g}}, f) =& W_1(\pi^{\text{g}})\phi\{f(X)\} + W_{-1}(\pi^{\text{g}})\phi\{-f(X)\} \\ &+ |V_1(\pi^{\text{g}})|\phi\{\text{sign}(V_1) f(X^{\text{g}})\} + |V_{-1}(\pi^{\text{g}})|\phi\{-\text{sign}(V_{-1}) f(X^{\text{g}})\}. \end{aligned}$$

The loss function L_{abs} serves as a connection between the nonconvex augmented loss L_{aug} and the convex augmented loss function L_ϕ in (3.9).

Second, we explain the problems in building a convex augmented loss as (3.9) by directly estimating the required conditional expectations.

Denote the conditional expectation given X_1 and Y as

$$Q^0(X_1, Y, f) = \mathrm{E}_{2|1,Y}\left(I[Y \text{sign}\{f(X_1, X_2)\} \le 0] \mid X_1, Y\right) \tag{B.1}$$

Based on (B.1), define the augmented loss by

$$L_{\text{aug1}}(\widehat{\pi}, \widehat{Q}, f) = \frac{R}{\widehat{\pi}(X_1, Y)} I[Y \text{sign}\{f(X)\} \le 0] + \frac{\widehat{\pi}(X_1, Y) - R}{\widehat{\pi}(X_1, Y)} \widehat{Q}(X_1, Y, f)$$

where $\widehat{Q}$ is an estimator of $Q^0(X_1, Y)$. Then, in the same way as L_{aug} we can express

$$L_{\text{aug1}}\left(\pi^{\text{g}}, Q^{\text{g}}, f\right) = W_1(\pi^{\text{g}}) I[\text{sign}\{f(X)\} \le 0] + W_{-1}(\pi^{\text{g}}) I[-\text{sign}\{f(X)\} \le 0]$$

$$+V_1(\pi^{\mathrm{g}})Q^{\mathrm{g}}(X_1,1,f)+V_{-1}(\pi^{\mathrm{g}})Q^{\mathrm{g}}(X_1,-1,f),$$

where Q^{g} denotes a generic conditional expectation.

However,all positive weights as L_{abs} because the signs of $V_1(\pi^{\mathrm{g}})$ and $V_{-1}(\pi^{\mathrm{g}})$ are unknown. Even if they are positive, since $Q^{\mathrm{g}}(X_1,1,f)$ and $Q^{\mathrm{g}}(X_1,1,f)$ are not necessarily convex, it is not clear that the obtained loss function

$$\begin{aligned}&W_1(\pi^{\mathrm{g}})\phi\{f(X)\}+W_{-1}(\pi^{\mathrm{g}})\phi\{-f(X)\}+V_1(\pi^{\mathrm{g}})Q^{\mathrm{g}}(X_1,1,f)\\&+V_{-1}(\pi^{\mathrm{g}})Q^{\mathrm{g}}(X_1,-1,f)\end{aligned}$$

is convex after replacing the 0-1 loss by the convex surrogate loss ϕ as L_{abs}.

Appendix C: Proofs

C.1. Proof of Theorem 3.1

Proof. Case I: Under Condition 1, $F^*_{2|1,Y}(x_2)=F^0_{2|1,Y}(x_2)$. By Assumption 1,

$$\begin{aligned}&\mathrm{E}\left(\frac{\pi^*(X_1,Y)-R}{\pi^*(X_1,Y)}I[Y\mathrm{sign}\{f(X)\}\le 0]\right)\\
=&\mathrm{E}\left\{\mathrm{E}\left(\frac{\pi^*(X_1,Y)-R}{\pi^*(X_1,Y)}I[Y\mathrm{sign}\{f(X)\}\le 0]\mid X_1,Y\right)\right\}\\
=&\mathrm{E}\left\{\mathrm{E}\left(\frac{\pi^*(X_1,Y)-R}{\pi^*(X_1,Y)}\mid X_1,Y\right)\mathrm{E}\left(I[Y\mathrm{sign}\{f(X_1,X_2)\}\le 0]\mid X_1,Y\right)\right\}\\
=&\mathrm{E}\left\{\mathrm{E}\left(\frac{\pi^*(X_1,Y)-R}{\pi^*(X_1,Y)}\mid X_1,Y\right)\mathrm{E}\left(I[Y\mathrm{sign}\{f(X_1,X_{2|1,Y})\}\le 0]\mid X_1,Y\right)\right\}\\
=&E\left(\frac{\pi^*(X_1,Y)-R}{\pi^*(X_1,Y)}I[Y\mathrm{sign}\{f(X^0)\}\le 0]\right).\end{aligned}$$

Then,

$$\begin{aligned}&\mathcal{R}_{L_{\mathrm{aug}}}(\pi^*,X^*,f)\\
=&\mathrm{E}\left(\frac{R}{\pi^*(X_1,Y)}I[Y\mathrm{sign}\{f(X)\}\le 0]+\frac{\pi^*(X_1,Y)-R}{\pi^*(X_1,Y)}I[Y\mathrm{sign}\{f(X^0)\}\le 0]\right)\\
=&\mathrm{E}(I[Y\mathrm{sign}\{f(X)\}\le 0])+\mathrm{E}\left(\frac{R-\pi^*(X_1,Y)}{\pi^*(X_1,Y)}I[Y\mathrm{sign}\{f(X)\}\le 0]\right.\\
&\left.\qquad+\frac{\pi^*(X_1,Y)-R}{\pi^*(X_1,Y)}I[Y\mathrm{sign}\{f(X^0)\}\le 0]\right)\\
=&\mathrm{E}\left[I\{Y\mathrm{sign}(f(X))\le 0\}\right].\end{aligned}$$

Case II: Under Condition 2, $\pi^*(X_1,Y)=\pi^0(X_1,Y)$. Then,

$$\mathcal{R}_{L_{\mathrm{aug}}}(\pi^*,X^*,f)$$

$$
\begin{aligned}
=&\mathrm{E}\left(I[Y\operatorname{sign}\{f(X)\}\leq 0]\right)+\mathrm{E}\left(\frac{\pi^0(X_1,Y)-R}{\pi^0(X_1,Y)}I[Y\operatorname{sign}\{f(X^*)\}\leq 0]\right)\\
&-\mathrm{E}\left(\frac{\pi^0(X_1,Y)-R}{\pi^0(X_1,Y)}I[Y\operatorname{sign}\{f(X)\}\leq 0]\right). \qquad \text{(C.1)}
\end{aligned}
$$

Observe that the second term of (C.1) equals

$$
\begin{aligned}
&\mathrm{E}\left\{\mathrm{E}\left(\frac{\pi^0(X_1,Y)-R}{\pi^0(X_1,Y)}I[Y\operatorname{sign}\{f(X^*)\}\leq 0]\mid X^*,Y\right)\right\}\\
=&\mathrm{E}\left(I[Y\operatorname{sign}\{f(X^*)\}\leq 0]\mathrm{E}\left(\frac{\pi^0(X_1,Y)-R}{\pi^0(X_1,Y)}\mid X^*,Y\right)\right)\\
=&\mathrm{E}\left(I[Y\operatorname{sign}\{f(X^*)\}\leq 0]\mathrm{E}\left(\frac{\pi^0(X_1,Y)-R}{\pi^0(X_1,Y)}\mid X_1,Y\right)\right),
\end{aligned}
$$

which is zero after Remark 3.2. Similarly, the third term of (C.1) is zero. This completes the proof of Case II. □

C.2. *Proof of Lemma 3.1*

Proof. Recall that

$$
\begin{aligned}
&L_{\mathrm{abs}}(\pi^{\mathrm{g}},X^{\mathrm{g}},f)\\
=&W_1(\pi^{\mathrm{g}})I[\operatorname{sign}\{f(X)\}\leq 0]+W_{-1}(\pi^{\mathrm{g}})I[-\operatorname{sign}\{f(X)\}\leq 0]\\
&+|V_1(\pi^{\mathrm{g}})|I[\operatorname{sign}(V_1)\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0]\\
&+|V_{-1}(\pi^{\mathrm{g}})|I[-\operatorname{sign}(V_{-1})\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0]. \qquad \text{(C.2)}
\end{aligned}
$$

We now show the relation between $L_{\mathrm{aug}}(\pi^{\mathrm{g}},X^{\mathrm{g}},f)$ and $L_{\mathrm{abs}}(\pi^{\mathrm{g}},X^{\mathrm{g}},f)$.

Recall that $\operatorname{sign}(t)=2I(t\geq 0)-1\in\mathcal{Y}$; thus, $I[\operatorname{sign}\{f(X^{\mathrm{g}})\}=0]=0$. Consequently,

$$
I[-\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0]=I[\operatorname{sign}\{f(X^{\mathrm{g}})\}\geq 0]=1-I[\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0].
$$

Then, the third term of (C.2) is

$$
\begin{aligned}
&V_1I(V_1\geq 0)I[\operatorname{sign}\{f(X^{\mathrm{g}})\leq 0]-V_1I(V_1<0)I[-\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0]\\
=&V_1I(V_1\geq 0)I[\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0]-V_1I(V_1<0)(1-I[\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0])\\
=&V_1I[\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0]-V_1I(V_1<0). \qquad \text{(C.3)}
\end{aligned}
$$

Similarly, the fourth term of (C.2) can be expressed as

$$
V_{-1}I[-\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0]-V_{-1}I(V_{-1}<0). \qquad \text{(C.4)}
$$

Combining (C.2), (C.3), and (C.4),

$$
\begin{aligned}
L_{\mathrm{abs}}\left(\pi^{\mathrm{g}},X^{\mathrm{g}},f\right)=&W_1(\pi^{\mathrm{g}})I[\operatorname{sign}\{f(X)\}\leq 0]+W_{-1}(\pi^{\mathrm{g}})I[-\operatorname{sign}\{f(X)\}\leq 0]\\
&+V_1I[\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0]+V_{-1}I[-\operatorname{sign}\{f(X^{\mathrm{g}})\}\leq 0]\\
&-V_1I(V_1<0)-V_{-1}I(V_{-1}<0)\\
=&L_{\mathrm{aug}}(\pi^0,X^0,f)-V_1I(V_1<0)-V_{-1}I(V_{-1}<0).
\end{aligned}
$$

□

C.3. Proof of Lemma 3.2

Proof. Recall that $G(t) = u\phi(t)+v\phi(-t)$ and $t_{\min} = \arg\min_{t\in\mathbb{R}} G(t)$. When $\phi(t)$ is the hinge loss, the quadratic loss, the logistic loss, or the exponential loss, $\phi(t)$ is differentiable at 0 and $\phi'(0) < 1$. Also the convexity of $\phi(-t)$ holds with $\phi(0) = 1$.

Next we will show that $\operatorname{sign}(t_{\min}) = \operatorname{sign}(u - v)$.

(i) Suppose $\phi(t)$ is the hinge loss. Write $\phi(t) = \max(0, 1 - t) = \frac{1-t+|1-t|}{2}$. Then,

$$G(t) = \frac{t}{2}(v - u) + \frac{|1 - t|}{2}u + \frac{|1 + t|}{2}v + \frac{u + v}{2}.$$

Case 1: $u > v$. We show that for every $\alpha < 0$, there exists $\beta \geq 0$ such that $G(\alpha) > G(\beta)$.

If $-1 < \alpha < 0$, choose $0 < \beta < 1$. Then, $G(\alpha) - G(\beta) = (v - u)(\alpha - \beta) > 0$. If $\alpha \leq -1$, choose $\beta = 1$. Then, $G(\alpha) - G(\beta) = -u(\alpha - 1) - 2v \geq 2(u - v) > 0$. Therefore, the minimizer of $G(t)$ is non-negative and the sign of the minimizer is the same as $u - v$.

Case 2: $u < v$. We show that for every $\alpha \geq 0$, there exists $\beta < 0$ such that $G(\alpha) > G(\beta)$.

If $0 \leq \alpha < 1$, choose $-1 < \beta < 0$. Then, $G(\alpha) - G(\beta) = (v - u)(\alpha - \beta) > 0$. If $\alpha \geq 1$, choose $\beta = -1$. Then, $G(\alpha)-G(\beta) = v(\alpha+1)+u(-1-1) \geq 2(v-u) > 0$. Therefore, the minimizer of $G(t)$ is negative and its sign is the same as $u - v$.

Case 3: $u = v$. We show that $\operatorname{sign}(t_{\min}) = 1 = \operatorname{sign}(u - v)$.

Since for every $\alpha > 1$, $G(\alpha) = \alpha u + u > 2u = G(1)$; thus, $t_{\min} \leq 1$. Second, since for every $\alpha < -1$, $G(\alpha) = -\alpha u + u > 2u = G(-1)$; thus, $t_{\min} \geq -1$. Observe that $G(\alpha) = 2u$ when $\alpha \in [-1, 1]$. Then, $t_{\min} = 1/2$ is also the minimizer of $G(t)$.

Combining all three cases, we have $t_{\min} = \operatorname{sign}(u - v)$.

(ii) Suppose $\phi(t)$ is the quadratic loss, i.e., $\phi(t) = (1 - t)^2$. Then,

$$G'(t) = 2(u + v)t - 2(u - v).$$

Thus, $t_{\min} = \frac{u-v}{u+v}$. Consequently, $\operatorname{sign}(t_{\min}) = \operatorname{sign}(u - v)$.

(iii) Suppose $\phi(t)$ is the logistic loss, i.e., $\phi(t) = \ln\{1 + \exp(-t)\}$.

The derivative of $G(t)$ is

$$G'(t) = \frac{-\exp(-t)}{1 + \exp(-t)}u + \frac{\exp(t)}{1 + \exp(t)}v.$$

If $u > 0$ and $v > 0$, by solving $G'(t) = 0$, we have $t_{\min} = \log\left(\frac{u}{v}\right)$. Then, $\operatorname{sign}(t_{\min}) = \operatorname{sign}(u - v)$.

If $u = 0$ and $v > 0$, then $G'(t) > 0$. Thus, $t_{\min} = -\infty$ and $\operatorname{sign}(t_{\min}) = \operatorname{sign}(-v) = -1$.

If $u > 0$ and $v = 0$, then $G'(t) < 0$. Thus, $t_{\min} = \infty$, and $\operatorname{sign}(t_{\min}) = \operatorname{sign}(u) = 1$.

(iv) Suppose $\phi(t)$ is the exponential loss. The proof is similar to (ii). □

C.4. Proof of Theorem 4.1

Proof. Recall the decision function $f_{\phi,\text{opt}}$ in Sect. 3.2 with respect to $F^*_{2|1,Y}(x_2)$ and π^*.

Case I: Suppose Condition 1 holds.

Recall that

$$\begin{aligned}\mathcal{R}_{L^*_\phi}(f) = & \,\mathrm{E}\left[W_1(\pi^*)\phi\left\{f(X)\right\} + W_{-1}(\pi^*)\phi\left\{-f(X)\right\}\right] \\ & + \mathrm{E}\left[\left|V_1(\pi^*)\right|\phi\left\{\operatorname{sign}(V_1)f(X^*)\right\} + \left|V_{-1}(\pi^*)\right|\phi\left\{-\operatorname{sign}(V_{-1})f(X^*)\right\}\right]\end{aligned}$$

and $X^* = (X_1, X^*_{2|1,Y})$.

On replacing $F^*_{2|1,Y}(x_2)$ by $F^0_{2|1,Y}(x_2)$,

$$\begin{aligned}& \mathrm{E}\left[\left|V_1(\pi^*)\right|\phi\left\{\operatorname{sign}(V_1)f(X^*)\right\}\right] \\ = & \int_{\mathcal{X}_1,\mathcal{Y}}\int_{\{0,1\}}\int_{\mathcal{X}_2}\left|V_1(\pi^*)\right|\phi\left\{\operatorname{sign}\left(V_1\right)f\left(x_1,x_2\right)\right\}dF^*_{2|1,Y}\left(x_2\right)dF_{R|1,Y}(r)dF_{X_1,Y}(x_1,y) \\ = & \int_{\mathcal{X}_1,\mathcal{Y}}\int_{\{0,1\}}\int_{\mathcal{X}_2}\left|V_1(\pi^*)\right|\phi\left\{\operatorname{sign}\left(V_1\right)f\left(x_1,x_2\right)\right\}dF^0_{2|1,Y}\left(x_2\right)dF_{R|1,Y}(r)dF_{X_1,Y}(x_1,y) \\ = & \mathrm{E}\left[\left|V_1(\pi^*)\right|\phi\left\{\operatorname{sign}(V_1)f(X)\right\}\right].\end{aligned}$$

Similarly, we get

$$\mathrm{E}\left[\left|V_{-1}(\pi^*)\right|\phi\left\{-\operatorname{sign}(V_{-1})f(X^*)\right\}\right] = \mathrm{E}\left[\left|V_{-1}(\pi^*)\right|\phi\left\{-\operatorname{sign}(V_{-1})f(X)\right\}\right].$$

Then,

$$\begin{aligned}\mathcal{R}_{L^*_\phi}(f) &= \mathrm{E}\left\{L_\phi\left(\pi^*,X,f\right)\right\} = \mathrm{E}\left[\mathrm{E}\left\{L_\phi\left(\pi^*,X,f\right) \mid X\right\}\right] \\ &= \mathrm{E}\left(\mathrm{E}\left[W_1(\pi^*)\phi\left\{f(X)\right\} + W_{-1}(\pi^*)\phi\left\{-f(X)\right\}\right.\right. \\ &\quad\left.\left. + \left|V_1(\pi^*)\right|\phi\left\{\operatorname{sign}(V_1)f(X)\right\} + \left|V_{-1}(\pi^*)\right|\phi\left\{-\operatorname{sign}(V_{-1})f(X)\right\} \mid X\right]\right).\end{aligned}$$

Let

$$\begin{aligned}u_1(x) &= \mathrm{E}\left\{W_1(\pi^*) + V_1(\pi^*)I(V_1 \geq 0) - V_{-1}(\pi^*)I(V_{-1} < 0) \mid X = x\right\}, \\ v_1(x) &= \mathrm{E}\left\{W_{-1}(\pi^*) + V_{-1}(\pi^*)I(V_{-1} \geq 0) - V_1(\pi^*)I(V_1 < 0) \mid X = x\right\}.\end{aligned} \tag{C.5}$$

Then,

$$\mathcal{R}_{L^*_\phi}(f) = \mathrm{E}\left[u_1(X)\phi\{f(X)\} + v_1(X)\phi\{-f(X)\}\right].$$

Let

$$G_1(x,f) = u_1(x)\phi\{f(x)\} + v_1(x)\phi\{-f(x)\}.$$

Given x, $u_1(x)$ and $v_1(x)$ are known. Then, minimizing $G_1(x,f)$ is equivalent to minimizing $\mathcal{R}_{L^*_\phi}(f)$. Further, since both $u_1(x)$ and $v_1(x)$ are nonnegative, by Assumption 3, we have $\operatorname{sign}(f_{\phi,\text{opt}}(x)) = \operatorname{sign}(u_1(x) - v_1(x))$.

Next, we show that

$$\operatorname{sign}\{u_1(x) - v_1(x)\} = \operatorname{sign}\{2\mathrm{P}(Y = 1 \mid X = x) - 1\}.$$

Recall that for $j = -1, 1$,

$$W_j\left(\pi^*\right) = \frac{R}{\pi^*} I(Y = j);\ V_j\left(\pi^*\right) = \frac{\pi^* - R}{\pi^*} I(Y = j).$$

Then,

$$\begin{aligned}
&\mathrm{E}\{W_1\left(\pi^*\right) \mid X = x\} \\
=&\mathrm{E}\left\{\frac{R}{\pi^*(X_1, Y)} I(Y = 1) \mid X = x\right\} \\
=&\frac{1}{\pi^*(x_1, 1)} \mathrm{P}(Y = 1, R = 1 \mid X = x) \\
=&\frac{1}{\pi^*(x_1, 1)} \mathrm{P}(R = 1 \mid X = x, Y = 1)\mathrm{P}(Y = 1 \mid X = x) \\
=&\frac{\pi^0(x_1, 1)}{\pi^*(x_1, 1)} \mathrm{P}(Y = 1 \mid X = x),
\end{aligned}$$

where $\pi^0(x_1, y) = \mathrm{P}(R = 1|X_1 = x_1, Y = y)$ is the propensity score.

On the other hand,

$$\begin{aligned}
&\mathrm{E}\left\{V_1(\pi^*) I(V_1 \geq 0) \mid X = x\right\} \\
=&\mathrm{E}\left\{\frac{\pi^*(X_1, Y) - R}{\pi^*(X_1, Y)} I(Y = 1) I(V_1 \geq 0) \mid X = x\right\} \\
=&\frac{\pi^*(x_1, 1) - 0}{\pi^*(x_1, 1)} \mathrm{P}(Y = 1, R = 0 \mid X = x) \\
=&\mathrm{P}(R = 0 \mid X = x, Y = 1)\mathrm{P}(Y = 1 \mid X = x) \\
=&\{1 - \pi^0(x_1, 1)\}\mathrm{P}(Y = 1 \mid X = x),
\end{aligned}$$

where the second equality holds because $\frac{\pi^*(X_1,Y)-R}{\pi^*(X_1,Y)} I(Y = 1)I(V_1 \geq 0) \neq 0$ if and only if $Y = 1$ and $R = 0$. Also,

$$\begin{aligned}
&\mathrm{E}\left\{V_{-1}\left(\pi^*\right) I(V_{-1} < 0) \mid X = x\right\} \\
=&\mathrm{E}\left\{\frac{\pi^*(X_1, Y) - R}{\pi^*(X_1, Y)} I(Y = -1) I(V_{-1} < 0) \mid X = x\right\} \\
=&\frac{\pi^*(x_1, -1) - 1}{\pi^*(x_1, -1)} \mathrm{P}(Y = -1, R = 1 \mid X = x) \\
=&\frac{\pi^*(x_1, -1) - 1}{\pi^*(x_1, -1)} \mathrm{P}(R = 1 \mid X = x, Y = -1)\mathrm{P}(Y = -1 \mid X = x) \\
=&\frac{\pi^*(x_1, -1) - 1}{\pi^*(x_1, -1)} \pi^0(x_1, -1)\mathrm{P}(Y = -1 \mid X = x),
\end{aligned}$$

where the second equality holds because $\frac{\pi^*(X_1,Y)-R}{\pi^*(X_1,Y)} I(Y=-1)I(V_{-1}<0) \neq 0$ if and only if $Y=-1$ and $R=1$.

Similarly, we have

$$\begin{aligned}
\mathrm{E}\{W_{-1}(\pi^*) \mid X=x\} &= \frac{\pi^0(x_1,-1)}{\pi^*(x_1,-1)} \mathrm{P}(Y=-1 \mid X=x) \\
\mathrm{E}\{V_{-1}(\pi^*) I(V_{-1} \geq 0) \mid X=x\} &= \{1-\pi^0(x_1,-1)\} \mathrm{P}(Y=-1 \mid X=x) \\
\mathrm{E}\{V_1(\pi^*) I(V_1 < 0) \mid X=x\} &= \frac{\pi^*(x_1,1)-1}{\pi^*(x_1,1)} \pi^0(x_1,1) \mathrm{P}(Y=1 \mid X=x).
\end{aligned}$$

Thus,

$$\begin{aligned}
u_1(x) =& \frac{\pi^0(x_1,1)}{\pi^*(x_1,1)} \mathrm{P}(Y=1 \mid X=x) + \{1-\pi^0(x_1,1)\} \mathrm{P}(Y=1 \mid X=x) \\
& - \frac{\pi^*(x_1,-1)-1}{\pi^*(x_1,-1)} \pi^0(x_1,-1) \mathrm{P}(Y=-1 \mid X=x), \\
v_1(x) =& \frac{\pi^0(x_1,-1)}{\pi^*(x_1,-1)} \mathrm{P}(Y=-1 \mid X=x) + \{1-\pi^0(x_1,-1)\} \mathrm{P}(Y=-1 \mid X=x) \\
& - \frac{\pi^*(x_1,1)-1}{\pi^*(x_1,1)} \pi^0(x_1,1) \mathrm{P}(Y=1 \mid X=x).
\end{aligned}$$

Observe now

$$\begin{aligned}
& \pi^*(x_1,1)\pi^*(x_1,-1)\{u_1(x)-v_1(x)\} \\
=& \pi^*(x_1,-1)\pi^0(x_1,1)\mathrm{P}(Y=1 \mid X=x) \\
& + \pi^*(x_1,1)\pi^*(x_1,-1)\{1-\pi^0(x_1,1)\}\mathrm{P}(Y=1 \mid X=x) \\
& + \pi^*(x_1,1)\{1-\pi^*(x_1,-1)\}\pi^0(x_1,-1)\mathrm{P}(Y=-1 \mid X=x) \\
& - \pi^*(x_1,1)\pi^0(x_1,-1)\mathrm{P}(Y=-1 \mid X=x) \\
& - \pi^*(x_1,1)\pi^*(x_1,-1)\{1-\pi^0(x_1,-1)\}\mathrm{P}(Y=-1 \mid X=x) \\
& - \pi^*(x_1,-1)\{1-\pi^*(x_1,1)\}\pi^0(x_1,1)\mathrm{P}(Y=1 \mid X=x) \\
=& \pi^*(x_1,1)\pi^*(x_1,-1)\left\{2\mathrm{P}(Y=1 \mid X=x)-1\right\}.
\end{aligned}$$

This implies that

$$\begin{aligned}
& \operatorname{sign}\{f_{\phi,\mathrm{opt}}(x)\} \\
=& \operatorname{sign}\{u_1(x)-v_1(x)\} = \operatorname{sign}\{2\mathrm{P}(Y=1 \mid X=x)-1\} = \operatorname{sign}\{f_{I,\mathrm{opt}}(x)\}.
\end{aligned}$$

Case II: Suppose Condition 2 holds.

In this case,

$$\begin{aligned}
\mathcal{R}_{L_\phi^*}(f) = & \mathrm{E}[W_1(\pi^0)\phi\{f(X)\} + W_{-1}(\pi^0)\phi\{-f(X)\}] \\
& + \mathrm{E}[|V_1(\pi^0)|\phi\{\operatorname{sign}(V_1)f(X^*)\} + |V_{-1}(\pi^0)|\phi\{-\operatorname{sign}(V_{-1})f(X^*)\}].
\end{aligned} \tag{C.6}$$

The first expectation of (C.6) is

$$\int_{\mathcal{X}_1,\mathcal{Y}}\int_{\{0,1\}}\int_{\mathcal{X}_2}|W_1(\pi^0)|\phi\{\operatorname{sign}(W_1)f(x_1,x_2)\} \\ +|W_{-1}(\pi^0)|\phi\{-\operatorname{sign}(W_{-1})f(x_1,x_2)\}dF^0_{2|1,Y}(x_2)\,dF_{R|1,Y}(r)dF_{X_1,Y}(x_1,y), \tag{C.7}$$

where the third integration follows from the independence in Assumption 1 and Remark 3.2. Note that when $R=1$, (C.7) is nonzero. Then, given X_1 and Y, the conditional distribution of X_2 is $F^0_{2|1,Y}(x_2)$.

Given $R=r$, $Y=y$, define

$$h(x_1,x_2,y)=\frac{dF^*_{2|1,Y}(x_2)}{dF^0_{2|1,Y}(x_2)}, \tag{C.8}$$

which is the Radon-Nikodym derivative of $F^*_{2|1,Y}(x_2)$ with respect to $F^0_{2|1,Y}(x_2)$.

Then, by Remark 3.2, the second expectation of (C.6) is

$$\begin{aligned}&\int_{\mathcal{X}_1,\mathcal{Y}}\int_{\{0,1\}}\int_{\mathcal{X}_2}|V_1(\pi^0)|\phi\{\operatorname{sign}(V_1)f(x_1,x_2)\}\\ &\qquad+|V_{-1}(\pi^0)|\phi\{-\operatorname{sign}(V_{-1})f(x_1,x_2)\}dF^*_{2|1,Y}(x_2)\,dF_{R|Y,1}(r)dF_{X_1,Y}(x_1,y)\\ &=\int_{\mathcal{X}_1,\mathcal{Y}}\int_{\{0,1\}}\int_{\mathcal{X}_2}[|V_1(\pi^0)|\phi\,\{\operatorname{sign}(V_1)f(x_1,x_2)\}\\ &\qquad+|V_{-1}(\pi^0)|\phi\{-\operatorname{sign}(V_{-1})f(x_1,x_2)\}]\\ &\qquad\frac{dF^*_{2|1,Y}(x_2)}{dF^0_{2|1,Y}(x_2)}dF^0_{2|1,Y}(x_2)\,dF_{R|Y,1}(r)dF_{X_1,Y}(x_1,y)\\ &=\int_{\mathcal{X}_1,\mathcal{Y}}\int_{\{0,1\}}\int_{\mathcal{X}_2}[|V_1(\pi^0)|\phi\{\operatorname{sign}(V_1)f(x_1,x_2)\}\\ &\qquad+|V_{-1}(\pi^0)|\phi\{-\operatorname{sign}(V_{-1})f(x_1,x_2)\}]\\ &\qquad h(x_1,x_2,y)dF^0_{2|1,Y}(x_2)\,dF_{R|Y,1}(r)dF_{X_1,Y}(x_1,y),\end{aligned} \tag{C.9}$$

where the first equality holds because of the change of measure and the second equality follows after (C.8).

Consequently, combining (C.7) and (C.9),

$$\begin{aligned}\mathcal{R}_{L^*_\phi}(f)&=\mathrm{E}[W_1(\pi^0)\phi\{f(X)\}+W_{-1}(\pi^0)\phi\{-f(X)\}]\\ &\quad+\mathrm{E}(h(X,Y)[|V_1(\pi^0)|\phi\{\operatorname{sign}(V_1)f(X)\}\\ &\quad+|V_{-1}(\pi^0)|\phi\{-\operatorname{sign}(V_{-1})f(X)\}]).\end{aligned}$$

Define

$$\begin{aligned}u_2(x)&=\mathrm{E}\{W_1(\pi^0)\mid X=x\}+\mathrm{E}[h(X,Y)\{V_1(\pi^0)I(V_1\geq 0)\}\mid X=x]\\ &\quad-\mathrm{E}[h(X,Y)\{V_{-1}(\pi^0)I(V_{-1}<0)\}\mid X=x]\end{aligned}$$

$$\begin{aligned}
&= u_{21}(x) + u_{22}(x) + u_{23}(x), \\
v_2(x) &= \mathrm{E}\{W_{-1}(\pi^0) \mid X = x\} + \mathrm{E}[h(X,Y)\{V_{-1}(\pi^0)I(V_{-1} \geq 0)\} \mid X = x] \\
&\quad - \mathrm{E}[h(X,Y)\{V_1(\pi^0)I(V_1 < 0)\} \mid X = x] \\
&= v_{21}(x) + v_{22}(x) + v_{23}(x). \qquad \text{(C.10)}
\end{aligned}$$

Clearly, both $u_2(x)$ and $v_2(x)$ are nonnegative.

Then,

$$\mathcal{R}_{L_\phi}(f) = \mathrm{E}[u_2(X)\phi\{f(X)\} + v_2(X)\phi\{-f(X)\}].$$

Define

$$G_2(x, f) = u_2(x)\phi\{f(x)\} + v_2(x)\phi\{-f(x)\}.$$

By the similar argument in Case I, we focus on minimizing $G_2(x, f)$.

By Assumption 3, we have $\mathrm{sign}\{f_{\phi,\mathrm{opt}}(x)\} = \mathrm{sign}\{u_2(x) - v_2(x)\}$.

In what follows, we show that

$$\mathrm{sign}\{u_2(x) - v_2(x)\} = \mathrm{sign}\{2\mathrm{P}(Y = 1 \mid X = x) - 1\}.$$

Recall that $W_j(\pi^0) = \frac{R}{\pi^0}I(Y = j)$, $V_j(\pi^0) = \frac{\pi^0 - R}{\pi^0}I(Y = j)$ for $j = -1, 1$. By (C.10),

$$\begin{aligned}
u_{21}(x) &= \mathrm{E}\left\{\frac{R}{\pi^0(X_1, Y)}I(Y = 1) \mid X = x\right\} \\
&= \frac{1}{\pi^0(x_1, 1)}\mathrm{P}(R = 1, Y = 1 \mid X = x) \\
&= \frac{1}{\pi^0(x_1, 1)}\mathrm{P}(R = 1 \mid X = x, Y = 1)\mathrm{P}(Y = 1 \mid X = x) \\
&= \frac{1}{\pi^0(x_1, 1)}\mathrm{P}(R = 1 \mid X_1 = x_1, Y = 1)\mathrm{P}(Y = 1 \mid X = x) \\
&= \mathrm{P}(Y = 1 \mid X), \qquad \text{(C.11)}
\end{aligned}$$

where the fourth equality is obtained by Assumption 1 and the definition of $\pi^0(X_1, Y)$. Similarly,

$$v_{21}(x) = \mathrm{P}(Y = -1 \mid X = x). \qquad \text{(C.12)}$$

By the definition of $u_{22}(x)$ in (C.10),

$$\begin{aligned}
u_{22}(x) &= \mathrm{E}\left\{h(X,Y)\frac{\pi^0(X_1, Y) - R}{\pi^0(X_1, Y)}I(Y = 1)I(V_1 \geq 0) \mid X = x\right\} \\
&= h(x, 1)\frac{\pi^0(x_1, 1)}{\pi^0(x_1, 1)}\mathrm{P}(R = 0, Y = 1 \mid X = x) \\
&= h(x, 1)\{1 - \pi^0(x_1, 1)\}\mathrm{P}(Y = 1 \mid X = x), \qquad \text{(C.13)}
\end{aligned}$$

where the second equality holds by the fact that $h(X,Y)\frac{\pi^0(X_1,Y)-R}{\pi^0(X_1,Y)}I(Y=1)I\left(V_1\geq 0\right)\neq 0$ if and only if $Y=1$ and $R=0$.

Similarly, we get

$$\begin{aligned} v_{23}(x) &= -h(x,1)\frac{\pi^0(x_1,1)-1}{\pi^0(x_1,1)}\mathrm{P}(R=1,Y=1\mid x) \\ &= -h(x,1)\{\pi^0(x_1,1)-1\}\mathrm{P}(Y=1\mid X=x), \end{aligned} \tag{C.14}$$

and

$$\begin{aligned} v_{22}(x) &= \mathrm{E}\left\{h(X,Y)\frac{\pi^0(X_1,Y)-R}{\pi^0(X_1,Y)}I(Y=-1)I(V_{-1}\geq 0)\mid X=x\right\} \\ &= h(x,-1)\frac{\pi^0(x_1,-1)}{\pi^0(x_1,-1)}\mathrm{P}(R=0,Y=-1\mid X=x) \\ &= h(x,-1)\{1-\pi^0(x_1,-1)\}\mathrm{P}(Y=-1\mid X=x), \end{aligned} \tag{C.15}$$

where the second equality holds by the fact that $h(X,Y)V_{-1}\left(\pi^0\right)I\left(V_{-1}\geq 0\right)\neq 0$ if and only if $Y=-1$ and $R=0$.

By the definition of $u_{23}(x)$ in (C.10) and the fact that the integrated term $h(X,Y)\{V_{-1}(\pi^0)I\left(V_{-1}<0\right)\}\neq 0$ if and only if $Y=-1$ and $R=1$. Hence,

$$\begin{aligned} u_{23}(x) &= -h(x,-1)\frac{\pi^0(x_1,-1)-1}{\pi^0(x_1,-1)}\mathrm{P}\left(R=1,Y=-1\mid X=x\right) \\ &= -h(x,-1)\{\pi^0(x_1,-1)-1\}\mathrm{P}\left(Y=-1\mid X=x\right). \end{aligned} \tag{C.16}$$

Substituting (C.11)–(C.16) in (C.10),

$$\begin{aligned} & u_2(x)-v_2(x) \\ =& \mathrm{P}(Y=1\mid X=x)+h(x,1)\{1-\pi^0(x_1,1)\}\mathrm{P}(Y=1\mid X=x) \\ & +h(x,-1)\{1-\pi^0(x_1,-1)\}\mathrm{P}(Y=-1\mid X=x) \\ & -\big[\mathrm{P}(Y=-1\mid X=x)+h(x,-1)\{1-\pi^0(x_1,-1)\}\mathrm{P}(Y=-1\mid X=x) \\ & +h(x,1)(1-\pi^0(x_1,1))\mathrm{P}(Y=1\mid X=x)\big] \\ =& 2\mathrm{P}(Y=1\mid X=x)-1. \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{sign}\{u_2(x)-v_2(x)\} &= \mathrm{sign}\{f_{\phi,\mathrm{opt}}(x)\}=\mathrm{sign}\{2\mathrm{P}(Y=1\mid X=x)-1\} \\ &= \mathrm{sign}\{f_{I,\mathrm{opt}}(x)\}. \end{aligned}$$

By Theorem 3.1, $\mathcal{R}_{L^*_{\mathrm{aug}}}(f)=\mathcal{R}(f)$. Then,

$$\mathcal{R}(f_{I,\mathrm{opt}})=\mathcal{R}_{L^*_{\mathrm{aug}}}(f_{I,\mathrm{opt}})\geq\mathcal{R}_{L^*_{\mathrm{aug}}}(f_{L^*_{\mathrm{aug}},\mathrm{opt}})=\mathcal{R}(f_{L^*_{\mathrm{aug}},\mathrm{opt}})\geq\mathcal{R}(f_{I,\mathrm{opt}}). \qquad \square$$

C.5. Proof of Theorem 4.2

Proof. Recall that

$$\begin{aligned}&L_{\text{aug}}(\pi^0, X^0, f)\\ =&W_1(\pi^0)I[\text{sign}\{f(X)\} \leq 0] + W_{-1}(\pi^0)I[-\text{sign}\{f(X)\} \leq 0]\\ &+ V_1(\pi^0)I[\text{sign}\{f(X^0)\} \leq 0] + V_{-1}(\pi^0)I[-\text{sign}\{f(X^0)\} \leq 0]\\ =&\frac{R}{\pi^0}I[Y\text{sign}\{f(X)\} \leq 0] + \frac{\pi^0 - R}{\pi^0}I[Y\text{sign}\{f(X^0)\} \leq 0],\end{aligned}$$

where $X^0 = (X_1, X_{2|1,Y})$. By Theorem 3.1,

$$\mathcal{R}_{L_{\text{aug}}}(\pi^0, X^0, f) = \text{E}\{L_{\text{aug}}(\pi^0, X^0, f)\} = \mathcal{R}(f) = \text{E}(I[Y\,\text{sign}\{f(X) \leq 0\}]).$$

Define $L_{\text{aug},1}(f) = L_{\text{aug}}(\pi^*, X^0, f)$, $L_{\text{aug},2}(f) = L_{\text{aug}}(\pi^0, X^*, f)$. Then, for $j = 1, 2$, the risk function and the Bayes risk are given by $\mathcal{R}_{L_{\text{aug},j}}(f) = \text{E}\{L_{\text{aug},j}(f)\}$ and $\mathcal{R}^*_{L_{\text{aug},j}} = \inf_f \mathcal{R}_{L_{\text{aug},j}}(f)$, respectively.

By Theorem 3.1, we have

$$\mathcal{R}_{L_{\text{aug},1}}(f) = \mathcal{R}_{L_{\text{aug},2}}(f) = \mathcal{R}(f).$$

Thus, the excess risk of $\mathcal{R}(f) - \mathcal{R}^*$ is equivalent to $\mathcal{R}_{L_{\text{aug},j}}(f) - \mathcal{R}^*_{L_{\text{aug},j}}$ for both $j = 1$ and 2.

Define $L_{\text{abs},1}(f) = L_{\text{abs}}(\pi^*, X^0, f)$, $L_{\text{abs},2}(f) = L_{\text{abs}}(\pi^0, X^*, f)$ where L_{abs} is the form of (3.7). For $j = 1, 2$, the risk, the Bayes risk, and the Bayes decision function are

$$\mathcal{R}_{L_{\text{abs},j}}(f) = \text{E}\left\{L_{\text{abs},j}(f)\right\},\ \mathcal{R}^*_{L_{\text{abs},j}} = \inf_f \mathcal{R}_{L_{\text{abs},j}}(f),$$

and

$$f_{\text{abs},j,\text{opt}} = \arg\min_f \mathcal{R}_{L_{\text{abs},j}}(f).$$

Then,

$$\mathcal{R}(f) - \mathcal{R}^* = \mathcal{R}_{L_{\text{aug},j}}(f) - \mathcal{R}^*_{L_{\text{aug},j}} = \mathcal{R}_{L_{\text{abs},j}}(f) - \mathcal{R}^*_{L_{\text{abs},j}}, \tag{C.17}$$

where the first equality holds after Theorem 3.1 and the second equality holds after Lemma 3.1. Thus, the excess risk $\mathcal{R}(f) - \mathcal{R}^*$ equals $\mathcal{R}_{L_{\text{abs},1}}(f) - \mathcal{R}^*_{L_{\text{abs},1}}$ when Condition 1 holds and equals $\mathcal{R}_{L_{\text{abs},2}}(f) - \mathcal{R}^*_{L_{\text{abs},2}}$ when Condition 2 holds. Subsequently, we will focus our analysis on these two excess risks.

Recall that

$$\begin{aligned}&\mathcal{R}_{L_{\text{abs},1}}(f)\\ =&\text{E}(W_1(\pi^*)I[\text{sign}\{f(X)\} \leq 0] + W_{-1}(\pi^*)I[-\text{sign}\{f(X)\} \leq 0]\\ &+ |V_1(\pi^*)|I[\text{sign}(V_1)\text{sign}\{f(X^0)\} \leq 0]\\ &+ |V_{-1}(\pi^*)|I[-\,\text{sign}(V_{-1})\text{sign}\{f(X^0)\} \leq 0])\end{aligned}$$

and

$$\begin{aligned}
&\mathcal{R}_{L_{\text{abs},2}}(f)\\
=&\text{E}(W_1(\pi^0)I[\text{sign}\{f(X)\} \le 0] + W_{-1}(\pi^0)I[-\text{sign}\{f(X)\} \le 0]\\
&+ |V_1(\pi^0)|I[\text{sign}(V_1)\text{sign}\{f(X^*)\} \le 0]\\
&+ |V_{-1}(\pi^0)|I[-\,\text{sign}(V_{-1})\text{sign}\{f(X^*)\} \le 0]).
\end{aligned}$$

For $j = 1, 2$, let

$$c_j(x) = u_j(x) + v_j(x), \tag{C.18}$$

where $u_1(x)$ and $v_1(x)$ are defined in (C.5), and $u_2(x)$ and $v_2(x)$ are defined in (C.10).

Using the integration argument in the proof of Theorem 4.1,

$$\begin{aligned}
&\mathcal{R}_{L_{\text{abs},j}}(f)\\
=&\text{E}(u_j(X)I[\text{sign}\{f(X)\} \le 0] + v_1(X)I[-\text{sign}\{f(X)\} \le 0])\\
=&\text{E}\{c_j(X)(\eta_j(X)I[\text{sign}\{f(X)\} \le 0] + \{1-\eta_j(X)\}I[-\text{sign}\{f(X)\} \le 0])\},
\end{aligned}$$

where $\eta_j(X) = \frac{u_j(X)}{c_j(X)}$. Since $c_j(X)$ is positive and by Steinwart and Christmann (2008, Sect. 2.1), minimizing $\mathcal{R}_{L_{\text{abs},j}}(f)$ yields

$$\text{sign}\{f_{\text{abs},j,\text{opt}}(x)\} = \text{sign}\{2\eta_j(x) - 1\}.$$

For the classification loss function,

$$\mathcal{R}^*_{L_{\text{abs},j}} = \mathcal{R}_{L_{\text{abs},j}}(f_{\text{abs},j,\text{opt}}) = \text{E}\left[c_j(X)\min\left\{\eta_j(X), 1-\eta_j(X)\right\}\right].$$

Thus,

$$\begin{aligned}
&\mathcal{R}_{L_{\text{abs},j}}(f) - \mathcal{R}^*_{L_{\text{abs},j}}\\
=&\text{E}\left\{c_j(X)\left(\eta_j(X)I[\text{sign}\{f(X)\} \le 0] + \{1-\eta_j(X)\}I[-\text{sign}\{f(X)\} \le 0]\right.\right.\\
&\qquad\qquad \left.\left.-\min\left\{\eta_j(X), 1-\eta_j(X)\right\}\right)\right\}\\
=&\text{E}(c_j(X)\left|2\eta_j(X)-1\right|I[\{2\eta_j(X)-1\}\text{sign}\{f(X)\} \le 0]).
\end{aligned} \tag{C.19}$$

Similarly,

$$\mathcal{R}_{L_{\phi,j}}(f) = \text{E}(c_j(X)[\eta_j(X)\phi\{f(X)\} + \{1-\eta_j(X)\}\phi\{-f(X)\}]).$$

Define $U(\eta, t) = \eta\phi(t) + (1-\eta)\phi(-t)$. Then, $\mathcal{R}^*_{L_{\phi,j}} = \text{E}[c_j(X)\inf_{t\in\mathbb{R}} U\{\eta_j(X), t\}] = \text{E}[c_j(X)H\{\eta_j(X)\}]$. Hence,

$$\mathcal{R}_{L_{\phi,j}}(f) - \mathcal{R}^*_{L_{\phi,j}} = \text{E}(c_j(X)[U\{\eta_j(X), f(X)\} - \inf_{t\in\mathbb{R}} U\{\eta_j(X), t\}]).$$

Following Definition 2 of Bartlett et al. (2006), the ψ-transform of ϕ can be written as

$$\psi\{2\eta_j(X) - 1\} = \inf_{f:\ f(2\eta_j - 1)\le 0} U\{\eta_j(X), f(X)\} - \inf_{t\in\mathbb{R}} U\{\eta_j(X), t\}. \tag{C.20}$$

Specially, ψ is convex when $\phi(\alpha)$ is the hinge loss, the logistic loss, or the exponential loss.

Thus,

$$\begin{aligned}
&\psi\left\{\frac{\mathcal{R}(f)-\mathcal{R}^*}{\sup_{x\in\mathcal{X}} c_j(x)}\right\}\\
=&\psi\left(\frac{\mathcal{R}_{L_{\mathrm{abs},j}}(f)-\mathcal{R}^*_{L_{\mathrm{abs},j}}}{\sup_{x\in\mathcal{X}} c_j(x)}\right)\\
=&\psi\left\{\frac{1}{\sup_{x\in\mathcal{X}} c_j(x)}\mathrm{E}(c_j(x)|2\eta_j(X)-1|I[\{2\eta_j(X)-1\}\mathrm{sign}\{f(X)\}\le 0])\right\}\\
\le&\psi\left\{\mathrm{E}(|2\eta_j(X)-1|I[\{2\eta_j(X)-1\}\mathrm{sign}\{f(X)\}\le 0])\right\}\\
\le&\mathrm{E}\{\psi(|2\eta_j(X)-1|I[\{2\eta_j(X)-1\}\mathrm{sign}\{f(X)\}\le 0])\}\\
=&\mathrm{E}(I[\{2\eta_j(X)-1\}\mathrm{sign}\{f(X)\}\le 0]\psi\{|2\eta_j(X)-1|\})\\
=&\mathrm{E}(I[\{2\eta_j(X)-1\}\mathrm{sign}\{f(X)\}\le 0]\psi\{2\eta_j(X)-1\})\\
=&\mathrm{E}(I[\{2\eta_j(X)-1\}\mathrm{sign}\{f(X)\}\le 0]\\
&\qquad[\inf_{f:\ f(2\eta_j-1)\le 0} U\{\eta_j(X),f(X)\}-\inf_{t\in\mathbb{R}} U\{\eta_j(X),t\}])\\
\le&\mathrm{E}[U\{\eta_j(X),f(X)\}-\inf_{t\in\mathbb{R}} U\{\eta_j(X),t\}]\\
\le&\frac{1}{\inf_{x\in\mathcal{X}} c_j(x)}\mathrm{E}(c_j(x)[U\{\eta_j(X),f(X)\}-\inf_{t\in\mathbb{R}} U\{\eta_j(X),t\}])\\
=&\frac{\mathcal{R}_{L_{\phi,j}}(f)-\mathcal{R}^*_{L_{\phi,j}}}{\inf_{x\in\mathcal{X}} c_j(x)},
\end{aligned}$$

where the first equality follows from (C.17), the second equality follows from (C.19), the fourth equality follows from the convexity of ψ and Jenssen inequality, the fifth equality holds after applying Lemma 2 of Bartlett et al. (2006) to the nonnegative loss function $\phi(\alpha)$ and $\psi(0)=0$, the sixth equality holds after Lemma 2 of Bartlett et al. (2006) and $\psi(t)=\psi(-t)$, and the seventh equality holds after the definition of ψ in (C.20).

Finally, we derive the bounds for $c_j(x)$ for $j=1,2$. By the definition of $u_1(x)$ and $v_1(x)$ in (C.5),

$$\begin{aligned}
c_1(x)&=u_1(x)+v_1(x)\\
&=\mathrm{E}\left\{\frac{R}{\pi^*(X_1,Y)}+I(R=0)+\frac{1-\pi^*(X_1,Y)}{\pi^*(X_1,Y)}I(R=1)\mid X=x\right\}\\
&=\frac{\pi^0(X_1,Y)}{\pi^*(X_1,Y)}+1-\pi^0(X_1,Y)+\frac{1-\pi^*(X_1,Y)}{\pi^*(X_1,Y)}\pi^0(X_1,Y)\\
&=2\frac{\pi^0(X_1,Y)}{\pi^*(X_1,Y)}-2\pi^0(X_1,Y)+1.
\end{aligned}$$

By Assumption 4, $0<c_\ell\le\pi^*(X_1,Y)$, $\pi^0(X_1,Y)\le c_u<1$. Therefore, $2c_\ell c_u^{-1}+1-c_\ell-c_u\le c_1(x)\le 2c_u c_\ell^{-1}-2c_\ell+1$.

Recall that $u_2(x)$ and $v_2(x)$ are defined in (C.10), and $h(x_1, x_2, y)$ is defined in (C.8). Then,

$$\begin{aligned}c_2(x) &= u_2(x) + v_2(x)\\ &= 1 + 2h(x,1)\{1-\pi^0(x_1,1)\}\mathrm{P}(Y=1 \mid X=x)\\ &\quad + 2h(x,-1)\{1-\pi^0(x_1,-1)\}\mathrm{P}(Y=-1 \mid X=x).\end{aligned}$$

Note that $h(x_1, x_2, y)$ is nonnegative and when $|h(x_1, x_2, y)| \leq M_h$, then $1 \leq c_2(x) \leq 1 + M_h(1-c_\ell)$. □

Remark C.1. *By the definition of $\eta_j(X)$, i.e., $\eta_j(X) = \frac{u_j(X)}{c_j(X)}$,*

$$\begin{aligned}\mathrm{sign}\{f_{\phi,\mathrm{opt}}(x)\} = \mathrm{sign}\{f_{I,\mathrm{opt}}(x)\} &= \mathrm{sign}\{f_{abs,j,\mathrm{opt}}(x)\}\\ &= \mathrm{sign}\{2\mathrm{P}(Y=1 \mid X=x) - 1\},\end{aligned}$$

where $j = 1, 2$. This implies the Fisher consistency of the decision function $\mathrm{sign}\{f_{\phi,\mathrm{opt}}(x)\}$.

C.6. Proof of Theorem 4.3

Proof. From Theorem 3.1, if either $\pi^* = \pi^0$ or $F^*_{2|1,Y}(x_2) = F^0_{2|1,Y}(x_2)$, $\mathcal{R}(f) = \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, f)$. Then, $\mathcal{R}(f) = \mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, f)$, where $X^0 = (X_1, X_{2|1,Y})$.

Now we show the upper bound of

$$\mathcal{R}(\widehat{f}_\phi) - \mathcal{R}^* = \mathcal{R}(\widehat{f}_\phi) - \mathcal{R}(f_{I,\mathrm{opt}}).$$

Observe that

$$\begin{aligned}&\mathcal{R}(\widehat{f}_\phi) - \mathcal{R}(f_{I,\mathrm{opt}})\\ =&\mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, \widehat{f}_\phi) - \mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, f_{I,\mathrm{opt}})\\ =&\mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, \widehat{f}_\phi) - \inf_{f\in\mathcal{H}} \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, f) + \inf_{f\in\mathcal{H}} \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, f)\\ &- \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, \widehat{f}_\phi) + \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, \widehat{f}_\phi) - \mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, f_{I,\mathrm{opt}})\\ \leq&\mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, \widehat{f}_\phi) - \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, \widehat{f}_\phi)\\ &+ \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, f_{I,\mathrm{opt}}) - \mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, f_{I,\mathrm{opt}})\\ &+ \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, \widehat{f}_\phi) - \inf_{f\in\mathcal{H}} \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, f)\\ \leq&|\mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, \widehat{f}_\phi) - \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, \widehat{f}_\phi)|\\ &+ |\mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, f_{I,\mathrm{opt}}) - \mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, f_{I,\mathrm{opt}})|\\ &+ \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, \widehat{f}_\phi) - \inf_{f\in\mathcal{H}} \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, f)\\ \leq&2 \sup_{f\in\mathcal{H}} |\mathcal{R}_{L_{\mathrm{aug}}}(\pi^0, X^0, f) - \mathcal{R}_{L_{\mathrm{aug}}}(\pi^*, X^*, f)|\end{aligned}$$

$$+ \mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, \widehat{f}_\phi) - \inf_{f\in\mathcal{H}} \mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, f).$$

If either Condition 1 or Condition 2 is correctly specified, both $\mathcal{R}_{L_{\text{aug}}}(\pi^*,X^*,f) = \mathcal{R}(f)$ and $\mathcal{R}_{L_{\text{aug}}}(\pi^0, X^0, f) = \mathcal{R}(f)$. Thus,

$$2 \sup_{f\in\mathcal{H}} |\mathcal{R}_{L_{\text{aug}}}(\pi^0, X^0, f) - \mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, f)| = 0.$$

Hence, it suffices to derive the upper bound of

$$\mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, \widehat{f}_\phi) - \inf_{f\in\mathcal{H}} \mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, f).$$

By $\mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, f) = \mathcal{R}(f)$,

$$\mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, \widehat{f}_\phi) = \mathcal{R}(\widehat{f}_\phi).$$

Combining

$$\inf_{f\in\mathcal{H}} \mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, f) = \inf_{f\in\mathcal{H}} \mathcal{R}(f) \geq \mathcal{R}(f_{I,\text{opt}})$$

and Theorem 4.1, we get

$$\mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, \widehat{f}_\phi) - \inf_{f\in\mathcal{H}} \mathcal{R}_{L_{\text{aug}}}(\pi^*, X^*, f) \leq \mathcal{R}(\widehat{f}_\phi) - \mathcal{R}(f_{\phi,\text{opt}}).$$

By Theorem 4.2,

$$\{\inf_{x\in\mathcal{X}} c_j(x)\}\psi\left\{\frac{\mathcal{R}(\widehat{f}_\phi) - \mathcal{R}(f_{\phi,\text{opt}})}{\sup_{x\in\mathcal{X}} c_j(x)}\right\} \leq \mathcal{R}_{L_{\phi,j}}(\widehat{f}_\phi) - \mathcal{R}^*_{L_{\phi,j}}. \tag{C.21}$$

Then, we focus on the upper bound of the RHS of (C.21).

$$\begin{aligned}
&\text{RHS of (C.21)}\\
=&\inf_{f\in\mathcal{H}} \mathrm{E}\{L_\phi(\pi^*, X^*, f)\} + \lambda\|f\|^2_{\mathcal{H}} - \mathrm{E}\{L_\phi(\pi^*, X^*, f_{\phi,\text{opt}})\}\\
&- \inf_{f\in\mathcal{H}} \mathrm{E}\{L_\phi(\pi^*, X^*, f)\} - \lambda\|f\|^2_{\mathcal{H}} + \mathrm{E}\{L_\phi(\pi^*, X^*, \widehat{f}_\phi)\}\\
=&a_n(\lambda) + \mathrm{E}\{L_\phi(\pi^*, X^*, \widehat{f}_\phi)\} - \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, \widehat{f}_\phi)\}\\
&+ \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, \widehat{f}_\phi)\} - \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, f_{\phi,\text{opt},\lambda})\}\\
&+ \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, f_{\phi,\text{opt},\lambda})\} - \inf_{f\in\mathcal{H}} \mathrm{E}\{L_\phi(\pi^*, X^*, f)\} - \lambda\|f\|^2_{\mathcal{H}}\\
=&a_n(\lambda) + \mathrm{E}\{L_\phi(\pi^*, X^*, \widehat{f}_\phi)\} - \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, \widehat{f}_\phi)\}\\
&+ \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, \widehat{f}_\phi)\} - \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, f_{\phi,\text{opt},\lambda})\}\\
&+ \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, f_{\phi,\text{opt},\lambda})\} - \mathrm{E}\{L_\phi(\pi^*, X^*, f_{\phi,\text{opt},\lambda})\} - \lambda\|f_{\phi,\text{opt},\lambda}\|^2_{\mathcal{H}}\\
\leq& a_n(\lambda) + \lambda\|\widehat{f}_\phi\|^2_{\mathcal{H}} + \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, \widehat{f}_\phi)\}\\
&- \lambda\|f_{\phi,\text{opt},\lambda}\|^2_{\mathcal{H}} - \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\text{imp}}, f_{\phi,\text{opt},\lambda})\}
\end{aligned}$$

$$
\begin{aligned}
&+ \mathrm{E}\{L_\phi(\pi^*, X^*, \widehat{f}_\phi)\} - \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, \widehat{f}_\phi)\} \\
&+ \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f_{\phi,\mathrm{opt},\lambda})\} - \mathrm{E}\{L_\phi(\pi^*, X^*, f_{\phi,\mathrm{opt},\lambda})\}.
\end{aligned}
$$

Define

$$
\begin{aligned}
(J_1) &\equiv \lambda\|\widehat{f}_\phi\|_{\mathcal{H}}^2 + \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, \widehat{f}_\phi)\} - \lambda\|f_{\phi,\mathrm{opt},\lambda}\|_{\mathcal{H}}^2 - \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f_{\phi,\mathrm{opt},\lambda})\}; \\
(J_2) &\equiv \mathrm{E}\{L_\phi(\pi^*, X^*, \widehat{f}_\phi)\} - \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, \widehat{f}_\phi)\}; \\
(J_3) &\equiv \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f_{\phi,\mathrm{opt},\lambda})\} - \mathrm{E}\{L_\phi(\pi^*, X^*, f_{\phi,\mathrm{opt},\lambda})\}.
\end{aligned}
$$

We begin with the upper bound of J_2 and J_3.

By the fact that $\mathbb{P}_n L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, \widehat{f}_\phi)$ is nonnegative and by the definition of $\widehat{f}_\phi$ in (3.11),

$$
\lambda\|\widehat{f}_\phi\|_{\mathcal{H}}^2 \le \lambda\|\widehat{f}_\phi\|_{\mathcal{H}}^2 + \mathbb{P}_n L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, \widehat{f}_\phi) \le \mathbb{P}_n L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, 0).
$$

For all the hinge loss, the quadratic loss, the exponential loss, and the logistic loss, $\phi(0) = 1$. Then,

$$
\begin{aligned}
&L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, 0) \\
=&\frac{R}{\widehat{\pi}} I(Y=1) + \frac{R}{\widehat{\pi}} I(Y=-1) + \left|\frac{\widehat{\pi}-R}{\widehat{\pi}} I(Y=1)\right| + \left|\frac{\widehat{\pi}-R}{\widehat{\pi}} I(Y=-1)\right| \\
\le&\frac{2}{c_L} + 2\frac{c_U+1}{c_L} \equiv M^2.
\end{aligned}
$$

Thus, $\|\widehat{f}_\phi\|_{\mathcal{H}} \le M\lambda^{-1/2}$. Similarly, $\lambda\|f_{\phi,\mathrm{opt},\lambda}\|_{\mathcal{H}}^2 \le M^2$.

Next we examine the difference between $L_\phi(\pi^*, X^*, f)$ and $L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f)$.

Observe that

$$
\begin{aligned}
&L_\phi(\pi^*, X^*, f) - L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f) \\
=&\Bigg[\frac{R}{\pi^*} I(Y=1)\phi\{f(X)\} + \frac{R}{\pi^*} I(Y=-1)\phi\{-f(X)\} \\
&+ I(R=0)\frac{\pi^*-0}{\pi^*} I(Y=1)\phi\{f(X^*)\} - I(R=1\}\frac{\pi^*-1}{\pi^*} I(Y=1)\phi(-f(X^*)\} \\
&+ I(R=0)\frac{\pi^*-0}{\pi^*} I\{Y=-1)\phi\{-f(X^*)\} \\
&-I(R=1)\frac{\pi^*-1}{\pi^*} I(Y=-1)\phi\{f(X^*)\}\Bigg] \\
&- \Bigg[\frac{R}{\widehat{\pi}} I(Y=1)\phi\{f(X)\} + \frac{R}{\widehat{\pi}} I(Y=-1)\phi\{-f(X)) \\
&+ I(R=0)\frac{\widehat{\pi}-0}{\widehat{\pi}} I(Y=1)\phi\{f(X^{\mathrm{imp}})\} - I(R=1)\frac{\widehat{\pi}-1}{\widehat{\pi}} I(Y=1)\phi\{-f(X^{\mathrm{imp}})\} \\
&+ I(R=0)\frac{\widehat{\pi}-0}{\widehat{\pi}} I(Y=-1)\phi\{-f(X^{\mathrm{imp}})\} \\
&-I(R=1)\frac{\widehat{\pi}-1}{\widehat{\pi}} I(Y=-1)\phi\{f(X^{\mathrm{imp}})\}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{R}{\pi^*} - \frac{R}{\widehat{\pi}}\right)[I(Y=1)\phi\{f(X)\} + I(Y=-1)\phi\{-f(X)\}] \\
&\quad + (I(R=0)I(Y=1)[\phi\{f(X^*)\} - \phi\{f(X^{\text{imp}})\}]) \\
&\quad - (I(R=1)I(Y=1)[\phi\{-f(X^*)\} - \phi\{-f(X^{\text{imp}})\}]) \\
&\quad + (I(R=0)I(Y=-1)[\phi\{-f(X^*)\} - \phi\{-f(X^{\text{imp}})\}]) \\
&\quad - (I(R=1)I(Y=-1)[\phi\{f(X^*)\} - \phi\{f(X^{\text{imp}})\}]) \\
&\quad + \left(I(R=1)I(Y=1)\left[\frac{1}{\pi^*}\phi\{-f(X^*)\} - \frac{1}{\widehat{\pi}}\phi\{-f(X^{\text{imp}})\}\right]\right) \\
&\quad + \left(I(R=1)I(Y=-1)\left[\frac{1}{\pi^*}\phi\{f(X^*)\} - \frac{1}{\widehat{\pi}}\phi\{f(X^{\text{imp}})\}\right]\right). \qquad \text{(C.22)}
\end{aligned}
$$

Observe that

$$
\begin{aligned}
&\frac{1}{\pi^*}\phi\{-f(X^*)\} - \frac{1}{\widehat{\pi}}\phi\{-f(X^{\text{imp}})\} \\
=&\frac{1}{\pi^*}\phi\{-f(X^*)) - \frac{1}{\widehat{\pi}}\phi\{-f(X^*)\} + \frac{1}{\widehat{\pi}}\phi\{-f(X^*)\} - \frac{1}{\widehat{\pi}}\phi\{-f(X^{\text{imp}})\}. \qquad \text{(C.23)}
\end{aligned}
$$

Combining (C.22) and (C.23),

$$
\begin{aligned}
&L_\phi(\pi^*, X^*, f) - L_\phi(\widehat{\pi}, X^{\text{imp}}, f) \\
=&\sum_{j=\pm 1}\left\{I(R=0)I(Y=j) + I(R=1)I(Y=-j)\left(\frac{1}{\widehat{\pi}} - 1\right)\right\} \\
&\qquad [\phi\{jf(X^*)\} - \phi\{jf(X^{\text{imp}})\}] \\
&+ \left(\frac{R}{\pi^*} - \frac{R}{\widehat{\pi}}\right)\sum_{j=\pm 1} I(Y=j)\phi\{jf(X)\} \\
&+ \left(\frac{1}{\pi^*} - \frac{1}{\widehat{\pi}}\right)\sum_{j=\pm 1} I(Y=j)I(R=1)\phi\{-jf(X^*)\}. \qquad \text{(C.24)}
\end{aligned}
$$

Recall that $B_{\mathcal{H}}(\lambda) = M\lambda^{-1/2}B$, and that $B_{\mathcal{H}}$ is the ball covering $\mathcal{H}$ with radius $M\lambda^{-1/2}$. Then, for $j \in \{-1, 1\}$

$$
\begin{aligned}
|\phi\{jf(x)\}| &\leq |\phi\{jf(x)\} - \phi(0)| + |\phi(0)| \leq C_\phi(\lambda^{-1/2})\|f\|_\infty + \phi(0) \\
&\leq C_\phi(\lambda^{-1/2})M\lambda^{-1/2} + 1,
\end{aligned}
$$

where the second inequality follows after Assumption 7 with $\phi(0) = 1$ and C_ϕ is the Lipschitz constant in Assumption 7.

By Assumption 6, $|\widehat{F}_{2|1,Y}(x_2) - F^*_{2|1,Y}(x_2)| = \text{O}_p(n^{-\rho_1})$ and Lemma 4.2 of Liu and Goldberg (2020),

$$
\begin{aligned}
&\sup_{f \in B_{\mathcal{H}}} |\text{E}[\phi\{jf(X^{\text{imp}})\}] - \text{E}[\phi\{jf(X^*)\}]| \\
=&\{C_\phi(\lambda^{-1/2})M\lambda^{-1/2} + 1\}\text{O}_p(n^{-\rho_1})
\end{aligned}
$$

$$=\mathrm{O}_p\{C_\phi(\lambda^{-1/2})\lambda^{-1/2}n^{-\rho_1}\} \tag{C.25}$$

for $j \in \{-1, 1\}$.

Then, combining (C.24), (C.25) and Assumption 6 that $|\widehat{\pi} - \pi^*| = \mathrm{O}_p(n^{-\rho_2})$, we obtain

$$\sup_{f\in B_{\mathcal{H}}} |\mathrm{E}\{L_\phi(\pi^*, X^*, f)\} - \mathrm{E}\{L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f)\}| = \mathrm{O}_p\{C_\phi(\lambda^{-1/2})\lambda^{-1/2}n^{-\min(\rho_1,\rho_2)}\}.$$

This leads to

$$(J_2) + (J_3) = \mathrm{O}_p\{C_\phi(\lambda^{-1/2})\lambda^{-1/2}n^{-\min(\rho_1,\rho_2)}\}. \tag{C.26}$$

Next, we derive the upper bound of (J_1) using Lemma 6 of Bartlett et al. (2006).

Define the functional class

$$\mathcal{L}_{\phi,\lambda} = \{L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f) + \lambda\|f\|_{\mathcal{H}}^2 - L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f_{\phi,\mathrm{opt},\lambda}) - \lambda\|f_{\phi,\mathrm{opt},\lambda}\|_{\mathcal{H}}^2 : f \in B_{\mathcal{H}}\}.$$

Define

$$\begin{aligned} \mathcal{G}_{\phi,\lambda_n} &= \{\mathrm{E}(\ell) - \ell : \mathrm{E}(\ell) = \varepsilon, \ell \in \mathcal{L}_{\phi,\lambda_n}\}, \\ Z &= \sup_{g\in\mathcal{G}_{\phi,\lambda_n}} \mathbb{P}_n g, \text{ where } g \in \mathcal{G}_{\phi,\lambda_n}. \end{aligned} \tag{C.27}$$

We now bound the function in $\mathcal{L}_{\phi,\lambda}$. Lemma 6 of Bartlett et al. (2006) requires the following three conditions.

1. $\sup_{\ell\in\mathcal{L}_{\phi,\lambda_n}} \|\ell\|_\infty \le C_{1,\lambda}$.
2. There exists $c \ge 1$ and $0 < \beta \le 1$, for any $\ell \in \mathcal{L}_{\phi,\lambda_n}$, $\mathrm{E}(\ell^2) \le c\{\mathrm{E}(\ell)\}^\beta$.
3. Fixed $0 < \alpha, \varepsilon_1 < 1$, suppose if some $\ell \in \mathcal{L}_{\phi,\lambda_n}$ has $\mathbb{P}_n\ell \le \alpha\varepsilon_1$ and $\mathrm{E}(\ell) \ge \varepsilon_1$, then some $\ell' \in \mathcal{L}_{\phi,\lambda_n}$ has $\mathbb{P}_n\ell' \le \alpha\varepsilon_1$ and $\mathrm{E}(\ell') = \varepsilon_1$.

Under these three conditions, for any $\ell \in \mathcal{L}_{\phi,\lambda_n}$ that satisfies $\mathbb{P}_n\ell \le \alpha\varepsilon_1$, we have

$$\mathrm{P}\{\mathrm{E}(\ell) \ge \varepsilon_1\} \ge 1 - e^{-b},$$

provided that

$$\varepsilon_1 \ge \max\left\{\varepsilon^*, \frac{9cQb}{(1-\alpha)^2 n}, \frac{4QC_1 b}{(1-\alpha)n}\right\},$$

where Q is an absolute constant and $\varepsilon^* \ge \frac{6}{1-\alpha}\mathrm{E}(Z)$.

Then, it suffices to verify Conditions 1, 2, 3 and bound $\mathrm{E}(Z)$.

(i) To verify Condition 1, for all $\ell \in \mathcal{L}_{\phi,\lambda_n}$, since L_ϕ is a convex function, by Assumption 7 and Lemma 4.23 of Steinwart and Christmann (2008)

$$\begin{aligned} |\ell| &\le |L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f) - L_\phi(\widehat{\pi}, X^{\mathrm{imp}}, f_{\phi,\mathrm{opt},\lambda})| + \lambda\left|\|f\|_{\mathcal{H}}^2 - \|f_{\phi,\mathrm{opt},\lambda}\|_{\mathcal{H}}^2\right| \\ &\le C_{L_\phi}(\lambda^{-1/2})\|f - f_{\phi,\mathrm{opt},\lambda}\|_\infty + \lambda\|f\|_{\mathcal{H}}^2 + \lambda\|f_{\phi,\mathrm{opt},\lambda}\|_{\mathcal{H}}^2 \\ &\le 2C_{L_\phi}(\lambda^{-1/2})M\lambda^{-1/2} + 2M^2 \equiv C_{1,\lambda}. \end{aligned} \tag{C.28}$$

Thus, $\sup_{\ell \in \mathcal{L}_{\phi,\lambda}} \|\ell\|_\infty \leq C_{1,\lambda}$ which implies Condition 1. By (C.28), we also have $\|\ell\|_{\mathcal{H}} \leq C_{1,\lambda}$ by Lemma 4.23 of Steinwart and Christmann (2008) and $\|f\|_\infty \leq \|f\|_{\mathcal{H}}$.

(ii) To verify Condition 2, observe that

$$\begin{aligned}
&|L_\phi(\widehat{\pi}, X^{\text{imp}}, f) - L_\phi(\widehat{\pi}, X^{\text{imp}}, f_{\phi,\text{opt},\lambda})| \\
\leq &\left|W_1(\widehat{\pi})\phi\{f(X^{\text{imp}})\} - W_1(\widehat{\pi})\phi\{f_{\phi,\text{opt},\lambda}(X^{\text{imp}})\}\right| \\
&+ \left|W_{-1}(\widehat{\pi})\phi\{-f(X^{\text{imp}})\} - W_{-1}(\widehat{\pi})\phi\{-f_{\phi,\text{opt},\lambda}(X^{\text{imp}})\}\right| \\
&+ |V_1(\widehat{\pi})\phi\{\text{sign}(V_1)f(X^{\text{imp}})\} - V_1(\widehat{\pi})\phi\{\text{sign}(V_1)f_{\phi,\text{opt},\lambda}(X^{\text{imp}})\}| \\
&+ |V_{-1}(\widehat{\pi})\phi\{-\text{sign}(V_1)f(X^{\text{imp}})\} - V_{-1}(\widehat{\pi})\phi\{-\text{sign}(V_1)f_{\phi,\text{opt},\lambda}(X^{\text{imp}})\}| \\
= &W_1(\widehat{\pi})|\phi\{f(X^{\text{imp}})\} - \phi\{f_{\phi,\text{opt},\lambda}(X^{\text{imp}})\}| \\
&+ W_{-1}(\widehat{\pi})|\phi\{-f(X^{\text{imp}})\} - \phi\{-f_{\phi,\text{opt},\lambda}(X^{\text{imp}})\}| \\
&+ V_1(\widehat{\pi})|\phi\{\text{sign}(V_1)f(X^{\text{imp}})\} - \phi\{\text{sign}(V_1)f_{\phi,\text{opt},\lambda}(X^{\text{imp}})\}| \\
&+ V_{-1}(\widehat{\pi})|\phi\{-\text{sign}(V_1)f(X^{\text{imp}})\} - \phi\{-\text{sign}(V_1)f_{\phi,\text{opt},\lambda}(X^{\text{imp}})\}| \\
\leq &W_1(\widehat{\pi})C_\phi(\lambda^{-1/2})\|f - f_{\phi,\text{opt},\lambda}\|_\infty + W_{-1}(\widehat{\pi})C_\phi(\lambda^{-1/2})\|f - f_{\phi,\text{opt},\lambda}\|_\infty \\
&+ V_1(\widehat{\pi})C_\phi(\lambda^{-1/2})\|f - f_{\phi,\text{opt},\lambda}\|_\infty + V_{-1}(\widehat{\pi})C_\phi(\lambda^{-1/2})\|f - f_{\phi,\text{opt},\lambda}\|_\infty \\
\leq &\frac{(2c_u + 4)C_\phi(\lambda^{-1/2})}{c_l}\|f - f_{\phi,\text{opt},\lambda}\|_\infty \\
= &C_{L_\phi}(\lambda^{-1/2})\|f - f_{\phi,\text{opt},\lambda}\|_\infty,
\end{aligned}$$

where the second inequality holds because of the locally Lipschitz continuity of ϕ in Assumption 7; the third inequality holds after Assumption 4.

Note that

$$\begin{aligned}
&\ell|L_\phi(\widehat{\pi}, X^{\text{imp}}, f) - L_\phi(\widehat{\pi}, X^{\text{imp}}, f_{\phi,\text{opt},\lambda})| + \lambda|\|f\|_{\mathcal{H}}^2 - \|f_{\phi,\text{opt},\lambda}\|_{\mathcal{H}}^2| \\
\leq &C_{L_\phi}(\lambda^{-1/2})\|f - f_{\phi,\text{opt},\lambda}\|_\infty + \lambda\|f + f_{\phi,\text{opt},\lambda}\|_{\mathcal{H}}\|f - f_{\phi,\text{opt},\lambda}\|_{\mathcal{H}} \\
\leq &\{C_{L_\phi}(\lambda^{-1/2}) + \lambda\|f + f_{\phi,\text{opt},\lambda}\|_{\mathcal{H}}\}\|f - f_{\phi,\text{opt},\lambda}\|_{\mathcal{H}} \\
\leq &\{C_{L_\phi}(\lambda^{-1/2}) + 2M\lambda^{1/2}\}\|f - f_{\phi,\text{opt},\lambda}\|_{\mathcal{H}}.
\end{aligned}$$

Then,

$$\mathrm{E}(\ell^2) \leq \{C_{L_\phi}(\lambda^{-1/2}) + 2M\lambda^{1/2}\}^2\|f - f_{\phi,\text{opt},\lambda}\|_{\mathcal{H}}^2.$$

Using the same argument as in the proof of Theorem 3.4 of Zhao et al. (2012), we can show that

$$\mathrm{E}(\ell) \geq \frac{\lambda\|f - f_{\phi,\text{opt},\lambda}\|_{\mathcal{H}}^2}{2}.$$

Thus,

$$\mathrm{E}(\ell^2) \leq \frac{2}{\lambda}\{C_{L_\phi}(\lambda^{-1/2}) + 2M\lambda^{1/2}\}^2\mathrm{E}(\ell). \tag{C.29}$$

Then Condition 2 is satisfied for $c_\lambda = \frac{2}{\lambda}\{C_{L_\phi}(\lambda^{-1/2}) + 2M\lambda^{1/2}\}^2$ and $\beta = 1$.

(iii) To verify Condition 3, fixed $0 < \alpha, \varepsilon_1 < 1$, recall that for a function $f_1 \in B_{\mathcal{H}}(\lambda)$, $\ell(f_1) \in \mathcal{L}_{\phi,\lambda}$, which is defined as

$$\ell(f_1) = L_\phi(\widehat{\pi}, X^{\text{imp}}, f_1) + \lambda \|f_1\|^2_{\mathcal{H}} - L_\phi(\widehat{\pi}, X^{\text{imp}}, f_{\phi,\text{opt},\lambda}) - \lambda \|f_{\phi,\text{opt},\lambda}\|^2_{\mathcal{H}}.$$

Assume that $\mathbb{P}_n \ell(f_1) \leq \alpha\varepsilon_1$ and $\mathrm{E}\{\ell(f_1)\} \geq \varepsilon_1$. Since $\ell(f_{\phi,\text{opt},\lambda}) \equiv 0$, $\mathbb{P}_n \ell(f_{\phi,\text{opt},\lambda}) = 0$ and $\mathrm{E}\{\ell(f_{\phi,\text{opt},\lambda})\} = 0$.

Also $\mathbb{P}_n \ell(f)$ and $\mathrm{E}\{\ell(f)\}$ are both convex functions of f. There exists f' between f_1 and $f_{\phi,\text{opt},\lambda}$ such that $E\{\ell(f')\} = \varepsilon_1$ and $\mathbb{P}_n \ell(f') \leq \alpha\varepsilon_1$. This implies Condition 3.

Consider the difference

$$L_\phi(\widehat{\pi}, X^{\text{imp}}, f) + \lambda \|f\|^2_{\mathcal{H}} - L_\phi(\widehat{\pi}, X^{\text{imp}}, f_{\phi,\text{opt},\lambda}) - \lambda \|f_{\phi,\text{opt},\lambda}\|^2_{\mathcal{H}}. \tag{C.30}$$

By (3.11),

$$\mathbb{P}_n L_\phi(\widehat{\pi}_i, X^{\text{imp}}, \widehat{f}_\phi) + \lambda \|\widehat{f}_\phi\|^2_{\mathcal{H}} \leq \mathbb{P}_n L_\phi(\widehat{\pi}_i, X^{\text{imp}}, f_{\phi,\text{opt},\lambda}) + \lambda \|f_{\phi,\text{opt},\lambda}\|^2_{\mathcal{H}},$$

and

$$\mathbb{P}_n\{L_\phi(\widehat{\pi}_i, X^{\text{imp}}, \widehat{f}_\phi) + \lambda \|\widehat{f}_\phi\|^2_{\mathcal{H}} - L_\phi(\widehat{\pi}_i, X^{\text{imp}}, f_{\phi,\text{opt},\lambda}) - \lambda \|f_{\phi,\text{opt},\lambda}\|^2_{\mathcal{H}}\} \leq 0 < \frac{\varepsilon_1}{2}.$$

Since Conditions 1, 2, and 3 hold, applying Lemma 6 of Bartlett et al. (2006) to (C.30),

$$\begin{aligned} &\mathrm{P}[\mathrm{E}\{L_\phi(\widehat{\pi}_i, X^{\text{imp}}, \widehat{f}_\phi) + \lambda \|\widehat{f}_\phi\|^2_{\mathcal{H}} - L_\phi(\widehat{\pi}_i, X^{\text{imp}}, f_{\phi,\text{opt},\lambda}) - \lambda \|f_{\phi,\text{opt},\lambda}\|^2_{\mathcal{H}}\} \leq \varepsilon_1] \\ &> 1 - e^{-b}, \end{aligned} \tag{C.31}$$

where

$$\varepsilon_1 \geq \max\left(\varepsilon^*, \frac{36 c_\lambda Q b}{n}, \frac{8 Q C_{1,\lambda} b}{n}\right),$$

with $c_\lambda = \frac{2}{\lambda}\{C_{L_\phi}(\lambda^{-1/2}) + 2M\lambda^{1/2}\}^2$, $C_{1,\lambda} = 2C_{L_\phi}(\lambda^{-1/2})M\lambda^{-1/2} + 2M^2$, $\varepsilon^* \geq 12\mathrm{E}(Z)$. Note that both c_λ and $C_{1,\lambda}$ are functions of λ.

By Assumption 5 and Lemma C.1,

$$\mathrm{E}(Z) \leq C_{1,\lambda} c_p \max\left\{(C_{1,\lambda}^{-2} c_\lambda \varepsilon)^{1/2-p/4}\left(\frac{C_{2,p}}{n}\right)^{1/2}, \left(\frac{C_{2,p}}{n}\right)^{2/(2+p)}\right\}, \tag{C.32}$$

where $c_p > 0$ is a constant depending on p and $\varepsilon > 0$ is an arbitrarily small enough positive constant. Let

$$\begin{aligned} &\epsilon_{n,\lambda,b} \\ &= \varepsilon_1 \\ &\geq \max\left[12 C_{1,\lambda} c_p \max\left\{(C_{1,\lambda}^{-2} c_\lambda \varepsilon)^{1/2-p/4}\left(\frac{C_{2,p}}{n}\right)^{1/2}, \left(\frac{C_{2,p}}{n}\right)^{2/(2+p)}\right\}, \frac{36 c Q b}{n}, \right. \end{aligned}$$

$$\left.\frac{8QC_1b}{n}\right].$$

By (C.26) and (C.31), with probability no less than $1-e^{-2b}$,

$$(J_1)+(J_2)+(J_3)\le \mathrm{O}_p\{C_\phi(\lambda^{-1/2})n^{-\min(\rho_1,\rho_2)}\}+\epsilon_{n,\lambda,b}.$$

Therefore, with probability no less than $1-e^{-2b}$,

$$\begin{aligned}&\frac{1}{\inf_{x\in\mathcal{X}}c_j(x)}\psi\left\{\frac{\mathcal{R}(\widehat{f}_\phi)-\mathcal{R}(f_{\phi,\mathrm{opt}})}{\sup_{x\in\mathcal{X}}c_j(x)}\right\}\\ \le & a_n(\lambda)+\mathrm{O}_p\{C_\phi(\lambda^{-1/2})n^{-\min(\rho_1,\rho_2)}\}+\epsilon_{n,\lambda,b},\end{aligned}\tag{C.33}$$

for $j=1,2$. □

C.7. Lemma C.1

The following lemma is to bound $\mathrm{E}(Z)$ in (C.27).

Lemma C.1. *Under Assumption 5, for any* $\varepsilon>0$,

$$\mathrm{E}(Z)\le C_{1,\lambda}c_p\max\left\{\left(C_{1,\lambda}^{-2}c_\lambda\varepsilon\right)^{1/2-p/4}\left(\frac{C_{2,p}}{n}\right)^{1/2},\left(\frac{C_{2,p}}{n}\right)^{2/(2+p)}\right\},$$

where c_p *is a positive constant depending only on* p,

$$C_{1,\lambda}=C_{L_\phi}(\lambda^{-1/2})M\lambda^{-1/2}+2M^2,\quad c_\lambda=\frac{2}{\lambda}\{C_{L_\phi}(\lambda^{-1/2})+2M\lambda^{1/2}\}^2.$$

Proof. Recall $\mathcal{G}_{\phi,\lambda_n}=\{\mathrm{E}(\ell)-\ell:\mathrm{E}(\ell)=\varepsilon,\ell\in\mathcal{L}_{\phi,\lambda_n}\}$. For any $\ell\in\mathcal{G}_{\phi,\lambda_n}$, by (C.29)

$$\mathrm{E}(\ell^2)\le c_\lambda\mathrm{E}(\ell)=c_\lambda\varepsilon.$$

Then,

$$\begin{aligned}\mathrm{E}(Z)=\mathrm{E}(\sup_{g\in\mathcal{G}_{\phi,\lambda_n}}\mathbb{P}_ng)&=\mathrm{E}\left(\sup_{\substack{\ell\in\mathcal{L}_{\phi,\lambda_n},\\ \mathrm{E}(\ell)=\varepsilon}}[\mathrm{E}(\ell)-\frac{1}{n}\sum_{i=1}^n\ell\{f(X_i)\}]\right)\\ &\le\mathrm{E}\left[\sup_{\ell\in\mathcal{L}_{\phi,\lambda_n},\mathrm{E}(\ell^2)\le c_\lambda\varepsilon}|\mathrm{E}(\ell)-\frac{1}{n}\sum_{i=1}^n\ell\{f(X_i)\}|\right]=\mathrm{Rad}(\mathcal{L}_{\phi,\lambda_n},n,c_\lambda\varepsilon),\end{aligned}$$

where $\mathrm{Rad}(\mathcal{L}_{\phi,\lambda_n},n,c_\lambda\varepsilon)$ is the local Rademacher average of $\mathcal{L}_{\phi,\lambda_n}$ for $c_\lambda\varepsilon$. (See Sect. 5.2 of Steinwart and Scovel, 2007 for more details.)

Recall that

$$\mathcal{L}_{\phi,\lambda_n}=\{L_\phi(\widehat{\pi},X^{\mathrm{imp}},f)-\lambda\|f\|_{\mathcal{H}}^2-L_\phi(\widehat{\pi},X^{\mathrm{imp}},f_{\phi,\mathrm{opt},\lambda_n})-\lambda\|f_{\phi,\mathrm{opt},\lambda_n}\|_{\mathcal{H}}^2,$$

$$f \in B_{\mathcal{H}}\}$$

and $\|\ell\|_{\mathcal{H}} \leq C_{1,\lambda}$. Then, the ball $C_{1,\lambda}B$ covers $\mathcal{L}_{\phi,\lambda_n}$. It is sufficient to consider the local Rademacher average, defined by

$$\begin{aligned}\mathrm{Rad}(C_{1,\lambda}B, n, c_\lambda\varepsilon) &= \mathrm{E}\left[\sup_{\{\ell:\ell\in C_{1,\lambda}B \text{ and } \mathrm{E}(\ell^2)\leq c_\lambda\varepsilon\}} |\mathrm{E}(\ell) - \frac{1}{n}\sum_{i=1}^{n} \ell\{f(X_i)\}|\right] \\ &= C_{1,\lambda}\mathrm{Rad}(B, n, C_{1,\lambda}^{-2}c_\lambda\varepsilon),\end{aligned}$$

where the second equality holds after (37) of Steinwart and Scovel (2007, Sect. 5.2).

By Assumption 5,

$$\sup_{\mathbb{P}_n} \log N(B, C_{1,\lambda}^{-2}c_\lambda\varepsilon, L_2(\mathbb{P}_n)) \leq C_{2,p}(C_{1,\lambda}^{-2}c_\lambda\varepsilon)^{-p}.$$

Since B is a closed unit ball, by Proposition 5.5 of Steinwart and Scovel (2007),

$$\mathrm{Rad}(B, n, C_{1,\lambda}^{-2}c_\lambda\varepsilon) \leq c_p \max\left\{(C_{1,\lambda}^{-2}c_\lambda\varepsilon)^{1/2-p/4}\left(\frac{C_{2,p}}{n}\right)^{1/2}, \left(\frac{C_{2,p}}{n}\right)^{2/(2+p)}\right\},$$

where $c_p > 0$ depends only on p. Therefore,

$$\begin{aligned}\mathrm{E}(Z) &\leq \mathrm{Rad}(\mathcal{L}_{\phi,\lambda_n}, n, c_\lambda\varepsilon) \leq \mathrm{Rad}(C_{1,\lambda}B, n, c_\lambda\varepsilon) \\ &\leq C_{1,\lambda}c_p \max\left\{(C_{1,\lambda}^{-2}c_\lambda\varepsilon)^{1/2-p/4}\left(\frac{C_{2,p}}{n}\right)^{1/2}, \left(\frac{C_{2,p}}{n}\right)^{2/(2+p)}\right\}.\end{aligned}$$

□

C.8. Proof of Theorem 4.4

Proof. We first consider $\epsilon_{n,\lambda,b}$. When $p \in (0,2)$, $\frac{1}{2} - \frac{p}{4} > 0$. Since $C_{L_\phi}(\beta) \leq \delta\beta^q$, then

$$C_{1,\lambda} = 2C_{L_\phi}(\lambda^{-1/2})M\lambda^{-1/2} + 2M^2 \leq 2\delta\lambda^{-q/2}M\lambda^{-1/2} + 2M^2 \leq \delta_{1,M}\lambda^{-(q+1)/2}, \tag{C.34}$$

where $\delta_{1,M}$ is some constant depending on δ and M. Also for $\lambda \in (0,1)$ and $t < 0 < s$, we have $\lambda^s < \lambda^t$. Then,

$$c_\lambda = \frac{2}{\lambda}\{C_{L_\phi}(\lambda^{-1/2}) + 2M\lambda^{1/2}\}^2 \leq \frac{2}{\lambda}(\delta\lambda^{-(q+1)/2} + 2M\lambda^{1/2})^2 \leq \delta_{2,M}\lambda^{-(q+2)}, \tag{C.35}$$

where $\delta_{2,M}$ is some constant depending on δ and M. Thus,

$$\begin{aligned}C_{1,\lambda}c_p(C_{1,\lambda}^{-2}c_\lambda\varepsilon)^{1/2-p/4} &= c_p C_{1,\lambda}{}^{p/2} c_\lambda{}^{1/2-p/4}\varepsilon^{1/2-p/4} \\ &\leq c_p(\delta_{1,M})^{p/2}\lambda^{-(q+1)p/4}(\delta_{2,M}\lambda^{-(q+2)})^{1/2-p/4}\varepsilon^{1/2-p/4}\end{aligned}$$

$$\begin{aligned}&\le c_p\delta_{M,\varepsilon}\lambda^{-(q+1)p/4}\lambda^{-(q+2)(1/2-p/4)}\\&= c_p\delta_{M,\varepsilon}\lambda^{-(2q+4-p)/4},\end{aligned}$$

where $\delta_{M,\varepsilon}$ is some constant related to $\delta_{1,M}$, $\delta_{2,M}$, and ε. Hence,

$$\begin{aligned}&C_{1,\lambda}c_p\left(C_{1,\lambda}^{-2}c\varepsilon\right)^{1/2-p/4}\left(\frac{C_{2,p}}{n}\right)^{1/2}\\&\le c_p\delta_{M,\varepsilon}C_{2,p}^{1/2}\lambda^{-(q+1)p/4}\lambda^{-(q+2)(1/2-p/4)}n^{-1/2}\end{aligned}\tag{C.36}$$

and,

$$\begin{aligned}&C_{1,\lambda}c_p\left(\frac{C_{2,p}}{n}\right)^{2/(2+p)}\\&\le c_p\delta_{1,M}\lambda^{-(q+1)/2}C_{2,p}^{2/(2+p)}n^{-2/(2+p)}.\end{aligned}\tag{C.37}$$

Substituting (C.36) and (C.37) in (C.32),

$$\begin{aligned}&\mathrm{E}(Z)\\&\le\max\left\{c_p\delta_{M,\varepsilon}C_{2,p}^{1/2}\lambda^{-(2q+4-p)/4}n^{-1/2},c_p\delta_{1M}C_{2,p}^{2/(2+p)}\lambda^{-(q+1)/2}n^{-2/(2+p)}\right\}.\end{aligned}$$

Consequently, if $\min\left\{\lambda^{(2q+4-p)/4}n^{1/2},\lambda^{(q+1)/2}n^{2/(2+p)}\right\}\to\infty$, $\mathrm{E}(Z)\to 0$.

By (C.34) and (C.35), if $\lambda^{(q+1)/2}n\longrightarrow\infty$ and $\lambda^{q+2}n\to\infty$, then,

$$\min\{\lambda^{(2q+4-p)/4}n^{1/2},\lambda^{(q+1)/2}n^{2/(2+p)}\}\to\infty,\quad \frac{8QC_{1,\lambda}b}{n}\to 0,\quad \frac{36c_\lambda Qb}{n}\to 0.$$

Hence, if $\lambda^{q+2}n\to\infty$, then $\epsilon_{n,\lambda,b}\to 0$.

Next we consider $\mathrm{O}_p\{C_\phi(\lambda^{-1/2})\lambda^{-1/2}n^{-\min(\rho_1,\rho_2)}\}$. Since $C_\phi(\beta)\le\delta\beta^q$, the following equality holds,

$$\begin{aligned}C_\phi(\lambda^{-1/2})\lambda^{-1/2}n^{-\min(\rho_1,\rho_2)}&\le\delta\lambda^{-q/2}\lambda^{-1/2}n^{-\min(\rho_1,\rho_2)}\\&=\delta\lambda^{-(q+1)/2}n^{-\min(\rho_1,\rho_2)}.\end{aligned}$$

Hence, if $\lambda^{(q+1)/2}n^{\min(\rho_1,\rho_2)}\to\infty$, then $\mathrm{O}_p\{C_\phi(\lambda^{-1/2})\lambda^{-1/2}n^{-\min(\rho_1,\rho_2)}\}\to 0$. Since $\lambda\to 0$, $a(\lambda)\to 0$. Therefore, for $p\in(0,2]$, if $\lambda\to 0$ and $\lambda^{(q+2)/2}n^{\min(\rho_1,\rho_2)}$ $\to\infty$, then the RHS of (C.33) converges to zero, that is,for any $b>0$, with probability no less than $1-e^{-2b}$,

$$\frac{1}{\inf_{x\in\mathcal{X}}c_j(x)}\psi\left\{\frac{\mathcal{R}(\widehat{f}_\phi)-\mathcal{R}^*}{\sup_{x\in\mathcal{X}}c_j(x)}\right\}\longrightarrow 0.$$

Since $\mathcal{R}^*$ is the Bayes risk, then $\mathcal{R}(\widehat{f}_\phi)-\mathcal{R}^*$ is nonnegative. Note that $\inf_{x\in\mathcal{X}}c_j(x)$ and $\sup_{x\in\mathcal{X}}c_j(x)$ are positive and finite and ψ is increasing in $[0,\infty)$. Thus, for any $b>0$, with probability no less than $1-e^{-2b}$,

$$\mathcal{R}(\widehat{f}_\phi)-\mathcal{R}^*\longrightarrow 0.$$

This completes the proof. □

Appendix D: An alternative way to estimate the imputation model

An alternative way to estimate the imputation model is to employ the EM algorithm. When all the models were fully parametric, one can use the following EM algorithm

- E-step: Compute the prodictive distribution for X_2 by

$$f(x_2|x_1, y; \theta^{(t)}) = \frac{\mathrm{P}(Y = y|x_1, x_2; \theta_1^{(t)}) f(x_2|x_1; \theta_2^{(t)})}{\int \mathrm{P}(Y = y|x_1, x_2, y; \theta_1^{(t)}) f(x_2|x_1; \theta_2^{(t)}) dx_2}$$

- M-step: Update the parameters by solving the score equations for θ_1 and θ_2:

$$\sum_{i=1}^{n} [R_i S_1(\theta_1; x_{1i}, x_{2i}, y_i) + (1 - R_i)\mathrm{E}\{S_1(\theta_1; x_{1i}, X_2, y_i)|x_{1i}, y_i; \theta^{(t)}\}] = 0$$

and

$$\sum_{i=1}^{n} [R_i S_2(\theta_2; x_{1i}, x_{2i}, y_i) + (1 - R_i)\mathrm{E}\{S_2(\theta_2; x_{1i}, X_2, y_i)|x_{1i}, y_i; \theta^{(t)}\}] = 0,$$

where S_1 and S_2 are the score functions of of θ_1 and θ_2 are respectively.

Noticed that the normalization constant $\int \mathrm{P}(Y = y|x_1, x_2, y; \theta_1^{(t)})$ $f(x_2|x_1;$ $\theta_2^{(t)}) dx_2$ is known to be difficult to calculate as it involves the multivariate integral. See discussion in Tsiatis (2006, Remark 2, Sect. 6.2). One possible strategy is to apply Metropolis-Hastings algorithm to construct a Markov chain with a stationary distribution is $f(x_2|x_1; \theta_2^{(t)})$ for every given x_1 and at every iteration t of the EM algorithm.

Funding

Yair Goldberg was partially supported by the Israeli Science Foundation (grant No. 849/17).

Supplementary Material

R code

(doi: 10.1214/23-EJS2158SUPP; .zip). The R package `drkm4mc` for the weighted-complete-case kernel machine estimator and the doubly robust kernel machine estimator, and R code for Sects. 5 and 6 are provided at `https://github.com/LTTGH`.

References

A. B. An and W. A. Fuller. Regression adjustments for nonresponse. *Journal of the Indian Society of Agricultural Statistics*, 1998. MR1776585

H. S. Anderson and M. R. Gupta. Expected kernel for missing features in support vector machines. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 285–288. IEEE, 2011.

H. Bao, C. Scott, and M. Sugiyama. Calibrated surrogate losses for adversarially robust classification. volume 125 of *Proceedings of Machine Learning Research*, pages 408–451, 2020.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006. MR2268032

B. Bullins, E. Hazan, and T. Koren. The limits of learning with missing data. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3503–3511, 2016.

J. R. Carpenter, M. G. Kenward, and S. Vansteelandt. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A*, 169:571–584, 2006. MR2236921

A. Choudhury and M. R. Kosorok. Missing data imputation for classification problems. arXiv:2002.10709, 2020.

Y. F. Ding and J. S. Simonoff. An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, 11:131–170, 2010. MR2591624

R. E. Fan, P. H. Chen, C. J. Lin, and T. Joachims. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, 6:1889–1918, 2005. MR2249875

W. A. Fuller. *Sampling statistics*. John Wiley & Sons, 2011.

P. J. García-Laencina, J. L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, 72:1483–1493, 2009.

P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19:263–282, 2010.

P. Hall and H. G . Müller. Order-preserving nonparametric regression, with applications to conditional distribution and quantile function estimation. *Journal of the American Statistical Association*, 98:598–608, 2003. MR2011674

P. S. Han, L. L. Kong, J. W. Zhao, and X. C. Zhou. A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society: Series B*, 81:305–333, 2019. MR3928144

E. Hazan, R. Livni, and Y. Mansour. Classification with low rank and missing data. In *International Conference on Machine Learning*, pages 257–266, 2015.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36:1171–1220, 2008. MR2418654

M. R. Kosorok. *Introduction to Empirical Inference Processes and Semipara-*

metric Inference. Springer, New York, 2008. MR2724368

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, second edition, 2002. MR1925014

T. Liu and Y. Goldberg. Kernel machines with missing responses. *Electronic Journal of Statistics*, 14:3766–3820, 2020. MR4164463

J. Luengo, S. García, and F. Herrera. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, 32:77–108, 2012.

K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18:684–692, 2005.

Y. L. Qiu, H. Zheng, and O. Gevaert. A deep learning framework for imputing missing values in genomic data. *bioRxiv*, 2018.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89:846–866, 1994. MR1294730

M. Saar-Tsechansky and F. Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8:1623–1657, 2007.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1120, 1999. MR1731478

S. R. Seaman and S. Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical Science*, 33:184, 2018. MR3797709

P. K. Sharpe and R.J. Solly. Dealing with missing values in neural network-based diagnostic systems. *Neural Computing and Applications*, 3:73–77, 1995.

P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, 2006. MR2274406

M. Śmieja, L. Struski, J. Tabor, B. Zieliński, and P. Spurek. Processing of missing data by neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2724–2734, 2018.

M. Śmieja, L. Struski, J. Tabor, and M. Marzec. Generalized RBF kernel for incomplete data. *Knowledge-Based Systems*, 173:150–162, 2019.

A. J. Smola, S. V. N. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 325–332, 2005.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008. MR2796580

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35:575–607, 2007. MR2336860

T. G. Stewart, D. L. Zeng, and M. C. Wu. Constructing support vector machines with missing data. *WIREs Computational Statistics*, page e1430, 2018. MR3826095

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2006. MR2233926

X. J. Wang, R. Zhang, Y. Sun, and J. Z. Qi. Doubly robust joint learning for recommendation on data missing not at random. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6638–6647, 2019.

J. Xia, S. Y. Zhang, G. L. Cai, L. Li, Q. Pan, J. Yan, and G. M. Ning. Adjusted weight voting algorithm for random forests in handling missing values. *Pattern Recognition*, 69:52–60, 2017.

J. X. You, X. B. Ma, D. Y. Ding, M. Kochenderfer, and J. Leskovec. Handling missing data with graph representation learning. In *34th Conference on Neural Information Processing Systems*, 2020.

Y. Q. Zhao, D. L. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107:1106–1118, 2012. MR3010898

Y. Q. Zhao, D. l. Zeng, E. B. Laber, R. Song, M. Yuan, and M. R. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102:151–168, 2015. MR3335102