Electronic Journal of Statistics
Vol. 14 (2020) 3766–3820
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1752

# Kernel machines with missing responses

**Tiantian Liu**[1,2] **and Yair Goldberg**[1]

[1] *Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Israel*

[2] *School of Statistics, East China Normal University, China*
*e-mail:* tiantian.liu@campus.technion.ac.il; yairgo@technion.ac.il

**Abstract:** Missing responses is a common type of data where the interested outcomes are not always observed. In this paper, we develop two new kernel machines to handle such a case, which can be used for both regression and classification. The first proposed kernel machine uses only the complete cases where both response and covariates are observed. It is, however, subject to some assumption limitations. Our second proposed doubly-robust kernel machine overcomes such limitations regardless of the misspecification of either the missing mechanism or the conditional distribution of the response. Theoretical properties, including the oracle inequalities for the excess risk, universal consistency, and learning rates are established. We demonstrate the superiority of the proposed methods to some existing methods by simulation and illustrate their application to a real data set concerning a survey about homeless people.

**MSC2020 subject classifications:** Primary 60K35.
**Keywords and phrases:** Kernel machines, missing responses, inverse probability weighted estimator, doubly-robust estimator, oracle inequality, consistency, learning rate.

Received October 2019.

## Contents

1 Introduction . . . 3767
2 Preliminaries . . . 3770
3 Kernel machines with missing responses . . . 3772
3.1 Weighted-complete-case kernel machines . . . 3772
3.2 Doubly-robust kernel machines . . . 3773
3.3 Estimation of the augmentation term . . . 3774
3.3.1 Regression . . . 3775
3.3.2 Classification . . . 3776
3.4 Least-squares kernel machines with missing responses . . . 3776
4 Theoretical results . . . 3777
4.1 Assumptions and conditions . . . 3777
4.2 Theoretical results of the weighted-complete-case kernel machines 3778
4.3 Theoretical results of the doubly-robust kernel machines . . . 3779
5 Simulation . . . 3781
5.1 Setup . . . 3782
5.2 Results . . . 3784

6 Application to Los Angeles homeless population data . . . . . 3792
7 Conclusion and discussion . . . . . 3793
A Computation details in Subsection 3.4 . . . . . 3794
A.1 Weighted-complete-case kernel machines . . . . . 3794
A.2 Doubly-robust kernel machines . . . . . 3794
B Tables of simulations . . . . . 3796
C Proofs . . . . . 3798
C.1 Proof of Lemma 3.2 . . . . . 3798
C.2 Proof of Lemma 3.3 . . . . . 3799
C.3 Oracle inequality for the weighted-complete-case kernel machines 3800
C.4 Proof of Theorem 4.1 . . . . . 3803
C.5 Proof of Corollary 4.1 . . . . . 3805
C.6 Oracle inequality for the doubly-robust kernel machines . . . . . 3806
C.7 Proof of Lemma 4.1 . . . . . 3810
C.8 Proof of Lemma 4.2 . . . . . 3811
C.9 Proof of Theorem 4.2 . . . . . 3812
C.10 Proof of Corollary 4.2 . . . . . 3816
Acknowledgements . . . . . 3819
Supplementary material . . . . . 3819
References . . . . . 3819

## 1. Introduction

We consider the problem of statistical learning in the presence of missing responses. Missing response is a type of data in which the response variable cannot always be observed. Missing responses are common in market surveys, medical research, and opinion polls. Our first motivating example is from the Los Angeles County homeless survey directed by the Los Angeles Homeless Services Authority (LAHSA) (Kriegler and Berk, 2010). The LAHSA was interested in the number of homeless counts in the different survey tracts, among a total of 2,054. For each tract, information about the median household income, the percentage of unoccupied housing units, etc., were collected as covariates. It was known that there were 244 tracts having a large homeless population. All of these tracts, called "hot tracts", were included in the survey. Out of the remaining 1,810 tracts, 265 were randomly sampled. The sampling probability of such a tract depends on the Service Provision Areas (SPAs). Different areas have a different probability of being visited. Consequently, missing responses (number of homeless count) occurred in those tracts (not included in the survey) while having their covariates still available. (More details can be found in Kriegler and Berk (2010).) Another example incurring missing responses concerns a biomedical study where genetic information (treated as covariates) is collected on all participants, but the level of a biomarker is collected only on a subset of them based on the corresponding genetic information.

The missing mechanism involved in these two examples can be cataloged as missing at random (MAR), where the missingness depends only on the components that are observed, but not on the components that are missing (Little and Rubin, 2002, Chapter 1). In particular, in the first example, the sampled tracts depend on the areas but are independent of the actual counts in these tracts. In the second example, the observation of the responses of the biomarker levels depends on the genetic profile but not on their actual values.

For the general missing data problem, there are four common approaches. The first approach uses only the observations without any missing data, which are referred to as complete cases. This approach is simple but subject to obvious information loss and severe estimation bias. The second approach first imputes the missing data and then conducts an analysis based on the imputed data (Rubin, 2004). The imputation methods are typically more efficient. However, they can involve extrapolation, which is difficult to diagnose and can lead to bias. The third approach first imposes some distributional assumptions for both response variable and covariates; then makes inferences using the likelihood principle. In that case the distributional assumptions need to be justified in practice. Finally, the inverse probability weighting (IPW) approach constructs estimators by weighting complete cases (Horvitz and Thompson, 1952). We refer to Pelckmans et al. (2005) and Tsiatis (2006, Chapter 6) for excellent summaries of these approaches.

For the missing responses problem, various methods have been developed, including the augmented inverse probability weighting (AIPW) methods, semiparametric methods, and kernel machine methods, among others.

The AIPW method was first introduced by Robins et al. (1994) to estimate regression coefficients. Rotnitzky et al. (1998) proposed a semiparametic estimator for repeated outcomes with nonignorable response. Scharfstein et al. (1999) noted the double robustness of the AIPW estimator, which later attracted development of many new estimation methods. Wang and Rao (2002) first imputed the missing response values under the MAR assumption by kernel regression imputation and then constructed a complete data empirical likelihood. Wang et al. (2004) extended a semiparametric regression analysis method to include missing responses and built a doubly-robust estimator for the population mean. Tan (2010) built on a nonparametric likelihood approach and proposed a doubly-robust estimator which is locally and intrinsically efficient and sample-bounded. Vermeulen and Vansteelandt (2015) proposed an approach that locally minimized the squared first-order asymptotic bias of the doubly-robust estimator against the misspecification of both working models. Seaman and Vansteelandt (2018) proposed to use an estimation equation with inverse probability weighting and imputation to obtain a doubly-robust estimator. Wang et al. (2010) proposed a class of augmented inverse probability weighted kernel estimating equations for the nonparametric regression where the weights in the estimating equations contain kernel functions. The aim of aforementioned methods, except Wang et al. (2010), was to estimate the population mean but not to predict the response.

It is worthwhile to mention the difference between the doubly-robust kernel machines and the AIPW method. In statistical analysis, under certain conditions, one can construct different AIPW estimators of the population mean (Scharfstein et al., 1999; Tsiatis, 2006, Chapter 6.5) and different estimating equations (Wang et al., 2010; Seaman and Vansteelandt, 2018). The AIPW estimators can possess the doubly-robust property. While for the kernel machines approach, the augmented inverse probability weighted loss should not only ensure the doubly-robust property but also guarantee the convexity of the augmented loss function. The convexity is essential to warrant the theoretical properties of the kernel machine method. Under certain conditions, our proposed augmented loss function fulfills this requirement.

For the semiparametric methods, Liang et al. (2007) proposed a partially linear model for missing responses with measurement errors on the covariates. Azriel et al. (2016) studied a regression problem with missing responses. They showed that when the conditional expectation of the response is not linear in the predictors, additional observations can provide more information. In their work, they constructed the best linear predictor which depends also on the incomplete data.

The kernel machines, which include support vector machines (SVM) as a special case, are known for the advantages of easy computation and weak assumption about the distributions (Steinwart and Christmann, 2008; Hofmann et al., 2008; James et al., 2013, Chapter 9). In recent years, kernel methods have been developed for many types of data including some missing data settings (Goldberg and Kosorok, 2017; Stewart et al., 2018). For the kernel machine methods, Smola et al. (2005) developed a framework where the kernel methods are treated as an estimation problem in an exponential family, which can handle missing responses as well as missing covariates. They extended the concave convex procedure of Yuille and Rangarajan (2003) to find a local optimum of the estimator. However, the convergence of the estimator is not clear, the computation can be demanding, and the exponential family condition is hardly satisfied in real situations.

In this paper, we develop kernel machines with missing responses. We first propose a family of kernel machines that use the estimated inverse probabilities of the observed cases to weight the loss function of the complete cases. We call it 'inverse-weighted-probability complete-case estimator' (Tsiatis, 2006). More specifically, we first estimate the missing mechanism by a model and then show that if the model is correctly specified, the empirical risk is consistent.

When the model for the missing mechanism is misspecified in the aforementioned kernel machine method, the resulting estimator can be biased. Therefore, secondly, we propose a doubly-robust kernel machine estimator to overcome this problem. The new doubly-robust kernel machine estimator is derived with respect to an augmented loss, which is a kind of augmented inverse-probability-weighted-complete-case estimator, introduced by Scharfstein et al. (1999). See also Bang and Robins (2005) and Seaman and Vansteelandt (2018) for an overview. There are two key challenging issues that have to be dealt with when constructing a doubly-robust kernel-machine estimator. The first is-

sue is that the augmented loss function needs to be both doubly-robust and convex. As pointed out earlier, the convexity is not always satisfied by the usual AIPW method. The second issue is that the augmented estimator has to converge uniformly over a set of functions that grows with the sample size. We are not aware of any such doubly-robust estimator in the context of kernel machines.

Our construction of the proposed doubly-robust kernel machine estimator can be summarized in the following two steps. First, we estimate the missing mechanism and the conditional distribution of the response given the covariates. The latter is used to compute the conditional risk. Second, we augment a weighted conditional risk to the weighted loss function. We show that the proposed empirical risk is doubly-robust consistent against misspecification of either the missing mechanism or the conditional distributions.

The second contribution of our paper is to establish the theoretical properties for the proposed kernel machines including the oracle inequalities, universal consistency and learning rates. Here we emphasize that the techniques that we use to derive these properties can be applied to obtain doubly-robustness in the general context of minimization problems.

The rest of the paper is organized as follows. Section 2 introduces some notation and assumptions. Section 3 presents the main methods of the proposed kernel machines. Section 4 provides the theoretical results, including the oracle inequalities, consistency, and convergence rate. Section 5 contains the simulation study of comparison with some existing methods and demonstrates the superiority of the proposed methods in terms of the empirical risk of predicting a new response, as well as the overall expectation. Section 6 illustrates the application of the proposed methods to the Los Angeles homeless data. Section 7 concludes the paper with some discussion about future directions. All technical proofs are deferred to Appendix C. An R package called `KM4ICD` that integrates easily to the package `mlr` (Machine Learning in R) for the kernel machine estimators is given in the Supplementary Materials.

## 2. Preliminaries

Let $Y$ denote the response random variable taking values in $\mathcal{Y} \subset \mathbb{R}$, where $\mathcal{Y}$ can be $\{-1, 1\}$ for (binary) classification and a bounded compact set of $\mathbb{R}$ for regression. Let $R$ denote the missingness indicator with $R = 1$ if $Y$ is observed, called the 'complete case', and $R = 0$ otherwise, called the 'incomplete case'. Let $X$ denote the associated covariates taking values in a compact set $\mathcal{X} \subset \mathbb{R}^\ell$.

We make some assumptions as in Tsiatis (2006, Chapter 6).

**Assumption 2.1.** *The missingness indicator $R$ and the response $Y$ are independent given the covariate vector $X$, i.e., the missing mechanism is missing at random (MAR).*

Under Assumption 2.1, the propensity score, defined by the conditional probability of observing $Y$ given $X$, is

$$\pi^0(X) = \mathrm{P}(R = 1 \mid X) = \mathrm{P}(R = 1 \mid X, Y).$$

**Assumption 2.2.** *There exists a positive constant* $c < 1/2$ *such that* $\inf_{x \in \mathcal{X}} \pi^0(x) \geq 2c$.

Assumption 2.2 implies that a certain portion of responses is always missing.

Let $\mathcal{P}$ be the set of all joint distributions $(R, X, Y)$ for which Assumptions 2.1 and 2.2 hold. In what follows, we will focus our analysis on the probability measures in $\mathcal{P}$.

Let $f : \mathcal{X} \mapsto \mathcal{Y}$ be a function. Let $L : \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty)$ be a loss function where $L(Y, f(X))$ can be interpreted as the cost of predicting $Y$ by $f(X)$.

**Assumption 2.3.** *Assume that* $L$ *is convex and locally Lipschitz continuous. The latter is in the sense that for all* $u > 0$ *there exists a constant* $C_L(u) \geq 0$ *such that* $\sup_{y \in \mathcal{Y}} |L(y, t) - L(y, t')| \leq C_L(u)|t - t'|$ *for* $t, t' \in [-u, u]$.

Without loss of generality, assume that $L(y, 0)$ is bounded by 1. Denote the risk with respect to the loss function $L$ by $\mathcal{R}_{L,\mathrm{P}}(f) \equiv \mathrm{E}[L(Y, f(X))]$ and the Bayes risk by $\mathcal{R}^*_{L,\mathrm{P}} \equiv \inf_{f \text{ is measurable}} \mathcal{R}_{L,\mathrm{P}}(f)$.

Let $\mathcal{H}$ be a separable reproducing kernel Hilbert space (RKHS) of a bounded measurable kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Denote its norm by $\|\cdot\|_{\mathcal{H}}$. Assume that $k$ is a universal kernel which means $\mathcal{H}$ is dense in the space of bounded continuous functions with respect to the supremum norm. Throughout the paper we assume that $\|k\|_\infty \leq 1$. (Steinwart and Christmann, 2008; Hofmann et al., 2008). Denote $\mathcal{H}_n$ as a subspace of $\mathcal{H}$ that grows with the sample size $n$. Let $\lambda$ denote a positive tuning parameter. A kernel machine $f_{P,\lambda}$ is the minimizer of the regularized risk,

$$f_{P,\lambda} \equiv \min_{f \in \mathcal{H}} \lambda \|f\|^2_{\mathcal{H}} + \mathcal{R}_{L,P}(f).$$

Given a simple random sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$ (without any missing data), denote $\mathcal{R}_{L,D}(f) = n^{-1} \sum_{i=1}^n L(Y_i, f(X_i))$ as the empirical risk of $\mathcal{R}_{L,P}(f)$. A kernel machine estimator of $f_{P,\lambda}$ is

$$f_{D,\lambda} \equiv \min_{f \in \mathcal{H}} \lambda \|f\|^2_{\mathcal{H}} + \mathcal{R}_{L,D}(f), \tag{2.1}$$

where $\lambda$ can be obtained by cross validation in practice. We denote the associated $\lambda$ by $\lambda_n$ for later use.

Since $L$ is a convex loss function, the representer theorem (Steinwart and Christmann, 2008, Theorem 5.5) implies that there is a unique solution of $f_{D,\lambda}$ in the form of

$$f_{D,\lambda}(x) = \sum_{i=1}^n \alpha_i k(x, X_i), \tag{2.2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\intercal \in \mathbb{R}^n$ is a vector of coefficients.

## 3. Kernel machines with missing responses

In this section, we propose two types of kernel machines to estimate $f$ with missing responses. Suppose $(R_1, X_1, R_1Y_1), \ldots, (R_n, X_n, R_nY_n)$ are independent and identically distributed samples of the triplet $(R, X, RY)$ where $Y_i$ is observed only when $R_i = 1$.

### 3.1. Weighted-complete-case kernel machines

With missing responses, a naive estimator of $\mathcal{R}_{L,P}(f)$ using the complete cases is

$$\mathcal{R}_{L,D}(f) = \frac{\sum_{i=1}^n R_i L(Y_i, f(X_i))}{\sum_{i=1}^n R_i}. \tag{3.1}$$

When all $R_i = 1$, i.e., there is no missing response, it reduces to the usual empirical risk. By the law of large numbers, $\mathcal{R}_{L,D}(f)$ converges to $\mathrm{E}\{RL(Y, f(X))\}/\mathrm{E}(R)$ in probability, which equals $\mathrm{E}\{L(Y, f(X))\}$ if and only if $\mathrm{E}\{RL(Y, f(X))\} = \mathrm{E}(R)\mathrm{E}\{L(Y, f(X))\}$. This equation holds when $R$ is independent of $(X, Y)$. Thus, the consistency of $\mathcal{R}_{L,D}(f)$ to $\mathcal{R}_{L,P}(f)$ cannot be guaranteed in general.

On the other hand, observe that by Assumption 2.1,

$$\begin{aligned}
\mathrm{E}\left\{\frac{RL(Y, f(X))}{\pi^0(X)}\right\} &= \mathrm{E}\left[\mathrm{E}\left\{\frac{RL(Y, f(X))}{\pi^0(X)} \mid X\right\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left\{L(Y, f(X)) \mid X\right\} \mathrm{E}\left\{\frac{R}{\pi^0(X)} \mid X\right\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\{L(Y, f(X)) \mid X\}\right] \\
&= \mathrm{E}\{L(Y, f(X))\}.
\end{aligned} \tag{3.2}$$

Therefore, an unbiased estimator of $\mathcal{R}_{L,P}(f)$ can be achieved by weighting the complete cases appropriately (Tsiatis, 2006, Chapter 6).

Let $\pi(X) \in (0, 1]$ denote a generic conditional probability $R$ given $X$. Define the weighted loss function for the data with missing responses by

$$L_W(\pi, R, X, Y, f(X)) \equiv \frac{RL(Y, f(X))}{\pi(X)} = \begin{cases} \frac{L(Y, f(X))}{\pi(X)} & R = 1, \\ 0 & R = 0, \end{cases} \tag{3.3}$$

where the weight function is $W(X) = R/\pi(X)$. Note that for any $\pi(X) \in (0, 1]$, $L_W(\pi, R, X, Y, f(X))$ is a convex function, since $L(Y, f(X))$ is a convex function.

Let $0 < \widehat{\pi}(X) \leq 1$ be an estimator of $\pi^0(X)$. Write $L_{\widehat{W}} = L_W(\widehat{\pi}, R, X, Y, f(X))$.

$$\mathcal{R}_{L_{\widehat{W}},D}(f) \equiv \frac{1}{n}\sum_{i=1}^n L_W(\widehat{\pi}, R_i, X_i, Y_i, f(X_i)) = \frac{1}{n}\sum_{i=1}^n \frac{R_i L(Y_i, f(X_i))}{\widehat{\pi}(X_i)}. \tag{3.4}$$

We now propose the weighted-complete-case kernel machine as

$$f_{D,\lambda}^{\widehat{W}} \equiv \arg\min_{f\in\mathcal{H}} \lambda\|f\|_{\mathcal{H}}^2 + \mathcal{R}_{L_{\widehat{W}},D}(f), \tag{3.5}$$

where the first term is the same as defined in (2.1).

**Lemma 3.1.** *Assume that $\widehat{\pi}(X)$ converges to $\pi^0(X)$ in probability uniformly and that Assumption 2.1 holds. Then, for any $f \in \mathcal{H}$, $\mathcal{R}_{L_{\widehat{W}},D}(f) \xrightarrow{\mathrm{P}} \mathcal{R}_{L,\mathrm{P}}(f)$.*

*Proof.* Since $\widehat{\pi}(X)$ is consistent,

$$\begin{aligned}\mathcal{R}_{L_{\widehat{W}},D}(f) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i L(Y_i, f(X_i))}{\widehat{\pi}(X_i)}\right\} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i L(Y_i, f(X_i))}{\pi^0(X)}\right\} + \mathrm{o}_p(1)\\ &\xrightarrow{\mathrm{P}} \mathrm{E}\left\{\frac{RL(Y, f(X))}{\pi^0(X)}\right\} = \mathcal{R}_{L,\mathrm{P}}(f).\end{aligned}$$

□

The consistency assumption of $\widehat{\pi}(X)$ in Lemma 3.1 may not be easy to verify in practice. We next propose a family of doubly-robust kernel machines with this assumption relaxed.

### *3.2. Doubly-robust kernel machines*

Denote the conditional distribution of $Y$ given $X$ by $F_{Y|X}(y \mid x, \beta^0)$, where $\beta^0 \in \mathbb{B}$ is an unknown parameter vector and $\mathbb{B}$ is the parameter space of finite dimension. For any $\beta \in \mathbb{B}$, define

$$H(x, \beta, f(x)) = \int_{y\in\mathcal{Y}} L(y, f(x)) dF_{Y|X}\left(y \mid x, \beta\right).$$

Then, the conditional expectation of the loss function $L(Y, f(X))$ given $X$, or the conditional risk, is expressed as

$$H(x, \beta^0, f(x)) = \int_{y\in\mathcal{Y}} L(y, f(x)) dF_{Y|X}\left(y \mid x, \beta^0\right) = \mathrm{E}\left\{L(Y, f(X)) \mid X = x\right\}. \tag{3.6}$$

By the law of total expectation, $\mathrm{E}\left\{H(X, \beta^0, f(X))\right\} = \mathrm{E}\left\{L(Y, f(X))\right\} = R_{L,\mathrm{P}}(f)$.

**Assumption 3.1.** *For any $\beta \in \mathbb{B}$, assume that $F_{Y|X}(y \mid x, \beta)$ is continuously differentiable with respect to $\beta$ for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and that $H(x, \beta, f(x))$ is a continuous function of $\beta$ for $x \in \mathcal{X}$.*

Let $\widehat{\beta} \in \mathbb{B}$ be some estimator of $\beta^0$. We make the following assumption about $\widehat{\beta}$ and $\widehat{\pi}(X)$.

**Assumption 3.2.** *Assume that $\widehat{\pi}(x) \xrightarrow{\mathrm{P}} \pi^*(x)$ uniformly for $x \in \mathcal{X}$, where $\pi^*(x)$ does not necessarily have to be $\pi^0(x)$; and that $\widehat{\beta} \xrightarrow{\mathrm{P}} \beta^* \in \mathbb{B}$, where $\beta^*$ does not necessarily equal $\beta^0$.*

Define the augmented loss function by

$$L_{W,H}(\pi, H, R, X, Y, f(X)) \equiv \frac{RL(Y, f(X))}{\pi(X)} - \frac{R - \pi(X)}{\pi(X)} H(X, \beta, f(X)). \quad (3.7)$$

The corresponding empirical risk is

$$\begin{aligned}\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f) &\equiv \frac{1}{n}\sum_{i=1}^{n} L_{W,H}\left(\widehat{\pi}, \widehat{H}, R_i, X_i, Y_i, f(X_i)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i L(Y_i, f(X_i))}{\widehat{\pi}(X_i)} - \frac{R_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)} H\left(X_i, \widehat{\beta}, f(X_i)\right)\right\}.\end{aligned}$$

Denote $H^0 = H(x, \beta^0, f(x))$, $\widehat{H} = H\left(x, \widehat{\beta}, f(x)\right)$, $L_{W^0,H^0} = L_{W,H}(\pi^0, H^0, R, X, Y, f(X))$ and $L_{\widehat{W},\widehat{H}} = L_{W,H}\left(\widehat{\pi}, \widehat{H}, R, X, Y, f(X)\right)$.

Unlike $L(Y, f(X))$ and $L_W(\pi, R, X, Y, f(X))$, $L_{W,H}(\pi, H, R, X, Y, f(X))$ is not necessarily convex by the construction. In order to obtain the doubly-robust kernel machine, we need both $L_{W^0,H^0}$ and $L_{\widehat{W},\widehat{H}}$ to be convex functions. The following lemma gives a sufficient condition.

**Lemma 3.2.** *Suppose that $L(Y, f(X))$ is the quadratic loss, i.e., $L(Y, f(X)) = \{Y - f(X)\}^2$. Then, $L_{W^0,H^0}$ and $L_{\widehat{W},\widehat{H}}$ are both convex functions.*

We propose the doubly-robust kernel machine to be

$$f_{D,\lambda}^{\widehat{W},\widehat{H}} \equiv \arg\min_{f\in\mathcal{H}} \lambda\|f\|_{\mathcal{H}}^2 + \mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f). \quad (3.8)$$

**Remark 3.1.** *(i) The kernel machine $f_{D,\lambda}$ in (2.1) is the decision function with respect to the loss function $L(Y, f(X))$ when the data are fully observed. (ii) The weighted-complete-case kernel machine $f_{D,\lambda}^{\widehat{W}}$ in (3.5) is the decision function with respect to the weighted loss function $L_{\widehat{W}}$ when some responses are missing. (iii) The doubly-robust kernel machine $f_{D,\lambda}^{\widehat{W},\widehat{H}}$ in (3.8) is the decision function with respect to the augmented loss function $L_{\widehat{W},\widehat{H}}$ when some responses are missing. Additionally, we choose $L(Y, f(X))$ to be the quadratic loss to warrant the convexity of $L_{\widehat{W},\widehat{H}}$.*

At last, we claim the consistency of the proposed doubly-robust kernel machine.

**Lemma 3.3.** *Under Assumptions 2.1, 3.1, and 3.2, for any $f \in \mathcal{H}$, if either $\pi^*(X) = \pi^0(X)$ or $\beta^* = \beta^0$, then $\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f) \xrightarrow{\mathrm{P}} \mathcal{R}_{L,\mathrm{P}}(f)$.*

### *3.3. Estimation of the augmentation term*

We present two explicit examples for estimating the augmentation term of (3.7). We limit the loss function to the quadratic loss.

*3.3.1. Regression*

Consider the following location-shift regression model (Tsiatis, 2006, Chapter 5)

$$Y = \mu(X, \beta^0) + \varepsilon,$$

where $\varepsilon$ is the random error with mean zero and variance $\sigma_\varepsilon^2$. Assume that $\varepsilon$ is independent with $X$. The form of $\mu(X, \beta^0)$ can be arbitrary, for example, linear $\mu(X, \beta^0)$, or log-linear, $\log(X^\intercal \beta^0)$. By the assumptions of $L$ and $\varepsilon$,

$$\begin{aligned} H(X, \beta^0, f(X)) &= \mathrm{E}\left[\left\{\mu(X, \beta^0) + \varepsilon - f(X)\right\}^2 \mid X\right] \\ &= \left\{\mu(X, \beta^0) - f(X)\right\}^2 + \sigma_\varepsilon^2. \end{aligned} \tag{3.9}$$

We next estimate $\beta^0$ and $\sigma_\varepsilon^2$ by maximizing the likelihood. Let $F_{R,X}(r, x)$ and $F(r, x, y, \beta^0)$ denote the joint distributions of $(R, X)$ and $(R, X, Y)$ respectively. By Assumption 2.1,

$$F(r, x, y, \beta^0) = F_{Y|R,X}(y \mid r, x, \beta^0) F_{R,X}(r, x) = F_{Y|X}(y \mid r, \beta^0) F_{R,X}(r, x).$$

Without loss of generality, suppose that the first $n_1$ triples of $(R_i, X_i, Y_i)$ are the complete cases (with $R_i = 1$), and that the last $n - n_1$ triples have missing responses (with $R_i = 0$). The likelihood of $\beta^0$ is

$$\begin{aligned} &\prod_{i=1}^{n_1} F(r_i, x_i, y_i, \beta^0) \prod_{i=n_1+1}^{n} F_{R,X}(r_i, x_i) \\ =&\prod_{i=1}^{n_1} F_{Y|X}(y_i \mid x_i, \beta^0) F_{R,X}(r_i, x_i) \prod_{i=n_1+1}^{n} F_{R,X}(r_i, x_i) \\ =&\prod_{i=1}^{n_1} F_{Y|X}(y_i \mid x_i, \beta^0) \prod_{i=1}^{n} F_{R,X}(r_i, x_i). \end{aligned} \tag{3.10}$$

It suffices to maximize the first factor of (3.10) involving $\beta^0$ (without knowing the joint distributions). Denote the resulting maximum likelihood estimator (MLE) of $\beta^0$ by $\widehat{\beta}$. Then, we estimate $\sigma_\varepsilon^2$ by

$$\widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n_1} R_i \left\{Y_i - \mu(X_i, \widehat{\beta})\right\}^2}{n_1},$$

which is a consistent estimator whenever $\mu(X, \widehat{\beta})$ is a consistent estimator of $\mu(X, \beta^0)$ since $X$ is independent of $\varepsilon$.

Replacing $\beta^0$ and $\sigma_\varepsilon^2$ by $\widehat{\beta}$ and $\widehat{\sigma}_\varepsilon^2$ respectively in (3.9), we obtain the estimate of the augmented term $H(X, \widehat{\beta}, f(X))$ and consequently $\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f)$.

*3.3.2. Classification*

Consider $Y \in \{-1, 1\}$ in a classification problem. Suppose we use the logistic model for the conditional probability of $Y$ given $X$ through

$$\mathrm{P}\left(Y = 1 \mid X, \beta^0\right) = \frac{\exp\left(X^\intercal \beta^0\right)}{1 + \exp\left(X^\intercal \beta^0\right)} \equiv \mathrm{logit}\left(X^\intercal \beta^0\right).$$

In this case, we have

$$\begin{aligned} H\left(X, \beta^0, f(X)\right) &= \mathrm{E}\left[\{Y - f(X)\}^2 \mid X\right] \\ &= 1 + f(X)^2 - 2f(X)\mathrm{E}\left(Y \mid X\right) \\ &= 1 + f(X)^2 - 2f(X)\left\{2\mathrm{P}\left(Y = 1 \mid X, \beta^0\right) - 1\right\}. \end{aligned} \tag{3.11}$$

Under the same assumption of the data as in Section 3.3.1, we obtain the MLE of $\beta^0$, $\widehat{\beta}$, by maximizing

$$\prod_{i=1}^{n_1} F_{Y|X}\left(y_i \mid x_i, \beta^0\right) = \prod_{i=1}^{n_1} \mathrm{P}\left(Y = 1 \mid x_i, \beta^0\right)^{\frac{y_i+1}{2}} \left\{1 - \mathrm{P}(Y = 1 \mid x_i, \beta^0)\right\}^{\frac{1-y_i}{2}}.$$

Substituting $H(X, \widehat{\beta}, f(X))$ in $\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f)$ and minimizing (3.8) lead to the doubly-robust kernel machine for classification.

### *3.4. Least-squares kernel machines with missing responses*

In fact, under the quadratic loss function, the proposed kernel machines can be obtained explicitly.

Let $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\intercal$, $\mathbf{Y} = (Y_1, \cdots, Y_n)^\intercal$, $W = \mathrm{diag}(R_1/\widehat{\pi}(X_1), \cdots, R_n/\widehat{\pi}(X_n))$, and $A = W^{1/2}$. Denote the kernel matrix $K = (K_{ij})_{n\times n}$ with $K_{ij} = k(X_i, X_j)$. Denote $\boldsymbol{\mu}(X, \widehat{\beta}) = (\mu(X_1, \widehat{\beta}), \ldots, \mu(X_n, \widehat{\beta}))^\intercal$. For example, in the considered classification problem, $\mu(X, \widehat{\beta}) = 2\mathrm{logit}(X^\intercal\widehat{\beta}) - 1$. Let $I$ denote the $n \times n$ identity matrix.

For the weighted-complete-case kernel machine, we have

$$\boldsymbol{\alpha}^{\widehat{W}} = (n\lambda I + WK)^{-1} W\mathbf{Y}. \tag{3.12}$$

For the doubly-robust kernel machine, we have

$$\boldsymbol{\alpha}^{\widehat{W},\widehat{H}} = (n\lambda I + K)^{-1}\left\{WY + (I - W)\boldsymbol{\mu}\left(X, \widehat{\beta}\right)\right\}. \tag{3.13}$$

The details are given in Appendix A.

In the end, the proposed kernel machine estimators with respect to the weighted loss in (3.3) and the augmented loss in (3.7) are respectively

$$f_{D,\lambda}^{\widehat{W}}(x) = \sum_{i=1}^{n} \alpha_i^{\widehat{W}} k(x, X_i) \quad \text{and} \quad f_{D,\lambda}^{\widehat{W},\widehat{H}}(x) = \sum_{i=1}^{n} \alpha_i^{\widehat{W},\widehat{H}} k(x, X_i).$$

## 4. Theoretical results

### *4.1. Assumptions and conditions*

In Section 3, we have shown that for any given $f \in \mathcal{H}$ the empirical risk based on the two proposed kernel machines are consistent estimators of the risk function $\mathcal{R}_{L,\mathrm{P}}(f)$. In this section, we prove the universal consistency and derive the learning rates of the proposed kernel machines. Here, the universal consistency means when the training set is sufficiently large, the learning methods produce nearly optimal decision functions with large probability for all $P \in \mathcal{P}$. The learning rates provide a framework that is more closely related to the practical needs. It answers the question of how fast $\mathcal{R}_{L,\mathrm{P}}(f_{D,\lambda})$ converges to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$. The learning rate of the learning method is defined in Steinwart and Christmann (2008, Lemma 6.5). For universal consistency, we shall prove the oracle inequalities which bound the finite-sample distance between the empirical risk and the true risk of the omniscient oracle.

First, we make one additional assumption about the estimator $\widehat{\pi}(X)$.

**Assumption 4.1.** *Assume that* $\widehat{\pi}(X)$ *satisfies*

$$0 < c_{n,L} \leq \widehat{\pi}(x) \leq c_{n,U} < 1, \text{ for all } x \in \mathcal{X}$$

*where* $1/c_{n,L}$ *and* $1/(1 - c_{n,U})$ *are* $O(n^d)$*, for some* $0 \leq d < 1/2$.

The assumption is satisfied when $\widehat{\pi}(X) \equiv \min\{\max(c_{n,\mathrm{L}}, \widetilde{\pi}(X)), c_{n,\mathrm{U}}\}$, where $\widetilde{\pi}(X)$ is some estimator. Moreover, if a lower bound of the constant $c$ in Assumption 2.2 is known, then we can choose $d = 0$.

Second, we introduce two conditions about the kernel space and the loss function, which we need in order to show the universal consistency of the two proposed kernel machines. Let $B_{\mathcal{H}} \equiv \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$ denote the unit ball in the RKHS $\mathcal{H}$. Define $\mathcal{N}(B_{\mathcal{H}}, \|\cdot\|_\infty, \varepsilon)$ as the $\varepsilon$-covering number of $B_{\mathcal{H}}$ with respect to the supremum norm $\|\cdot\|_\infty$, defined by $\|f\|_\infty = \operatorname{ess\,sup}\{|f(x)|, x \in \mathcal{X}\}$.

**Condition 4.1.** *There exist constants* $a > 1$ *and* $p > 0$ *such that for every* $\epsilon > 0$*, the entropy of* $B_{\mathcal{H}}$ *is bounded by* $\log \mathcal{N}(B_{\mathcal{H}}, \|\cdot\|_\infty, \epsilon) \leq a\epsilon^{-2p}$.

**Condition 4.2.** *There exist constants* $q > 0$ *and* $m \geq 1$ *such that the locally Lipschitz constant* $C_L(u)$ *defined in Assumption 2.3 is bounded by* $mu^q$.

**Remark 4.1.** *(i) Condition 4.1 is used to bound the entropy of the function space* $\mathcal{H}$*. The linear, Taylor, and Gaussian radial basis function (RBF) kernels satisfy Condition 4.1 for all* $p > 0$ *(Steinwart and Christmann, 2008, Section 6.4). (ii) For the hinge loss, Condition 4.2 holds with* $q = 0$*. For the quadratic loss, Condition 4.2 holds with* $q = 1$ *(Steinwart and Christmann, 2008, Section 2.2). (iii) Both conditions can be verified.*

Define

$$\mathrm{Err}_{1,n} = \sup_{x \in \mathcal{X}} \left|\widehat{\pi}(x) - \pi^0(x)\right|,$$

$$\mathrm{Err}_{2,n} \quad = \quad \sup_{f\in\mathcal{H}_n}\sup_{x\in\mathcal{X}} |H(x,\widehat{\beta},f(x)) - H(x,\beta^0,f(x))|, \tag{4.1}$$

as the estimation errors regarding the missing mechanism and the estimation error regarding the conditional risk, respectively. $\mathrm{Err}_{1,n}$ is used later in Appendix C to derive the oracle inequalities. Denote the approximation error function of $f_{\mathrm{P},\lambda}$ with respect to the Bayes risk by

$$A_2(\lambda) \equiv \lambda\|f_{\mathrm{P},\lambda}\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}}.$$

**Assumption 4.2.** *There exist constants $b > 0$ and $\gamma \in (0,1]$ such that $A_2(\lambda) \le b\lambda^\gamma$ for all $\lambda \ge 0$.*

### *4.2. Theoretical results of the weighted-complete-case kernel machines*

Recall that $\mathcal{P}$ is the set of all probability distributions for which Assumptions 2.1 and 2.2 hold.

**Theorem 4.1.** *Let Assumptions 2.1, 2.2, and 4.1 hold. Suppose the universal kernel $k$ and the loss function $L$ satisfy Conditions 4.1 and 4.2 respectively. Assume that $\left|\widehat{\pi}(X) - \pi^0(X)\right| = \mathrm{O}_p(n^{-1/2})$. Suppose $\lambda_n$ is a sequence in $(0,1)$ satisfying $\lambda_n \to 0$ and*

$$\lambda_n^{\frac{q+1}{2}} n^{\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)-\frac{(q+1)d}{2}} \longrightarrow \infty, \tag{4.2}$$

*where $d$, $p$, and $q$ are defined in Assumption 2.2, Conditions 4.1, and 4.2 respectively. Then, the weighted-complete-case kernel machine $f^{\widehat{W}}_{D,\lambda}$ in (3.5) is $\mathcal{P}$-universally consistent, i.e., $\mathcal{R}_{L,\mathrm{P}}\left(f^{\widehat{W}}_{D,\lambda}\right) \xrightarrow{\mathrm{P}} \mathcal{R}^*_{L,\mathrm{P}}$ for all $\mathrm{P} \in \mathcal{P}$.*

The proof of Theorem 4.1 is based on the oracle inequality derived for the weighted-complete-case kernel machines in Theorem C.1 of Appendix C. When $L$ is the quadratic loss with $q = 1$, the kernel $k$ is Gaussian, and $d = 0$; (4.2) reduces to $\lambda_n n^{\frac{1}{2(p+1)}}$ which by (i) of Remark 4.1 holds for all $p > 0$ and therefore is equivalent to $\lambda_n n^{\frac{1}{2}-\epsilon} \to \infty$ for arbitrarily small $\epsilon > 0$.

**Corollary 4.1.** *Let Assumptions 2.1, 2.2, 4.1, and 4.2 hold. Suppose the universal kernel $k$ and the loss function $L$ satisfy Conditions 4.1 and 4.2 respectively. Assume that $\left|\widehat{\pi}(X) - \pi^0(X)\right| = \mathrm{O}_p(n^{-1/2})$. Then, the learning rate of the weighted-complete-case kernel machine $f^{\widehat{W}}_{D,\lambda}$ in (3.5) is*

$$n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2\gamma}{2\gamma+q+1}},$$

*where $d$ and $\gamma$ are defined in Assumptions 2.2 and 4.2, $p$ and $q$ are defined in Conditions 4.1, and 4.2, respectively.*

When $L$ is the quadratic loss with $q = 1$, the kernel $k$ is Gaussian, and $d = 0$; the learning rate reduces to $n^{-\frac{\gamma}{2(\gamma+1)(p+1)}}$, which by (i) of Remark 4.1 holds for all $p > 0$ and therefore is equivalent to $n^{-\frac{\gamma}{2(\gamma+1)}+\epsilon}$ for arbitrarily small $\epsilon > 0$.

### *4.3. Theoretical results of the doubly-robust kernel machines*

First, we present some convergence orders related to this learning method.

Let $\mathbb{P}_n g = \frac{1}{n}\sum_{i=1}^n g(X_i)$ denote an operator taking the sample mean of $g(X)$ over $X_1, \ldots, X_n$. Denote

$$a_n \equiv \mathbb{P}_n \left[\frac{R - \pi^0(X)}{\pi^0(X)}\right],$$
$$h_n \equiv \sup_{f \in \mathcal{H}_n} \left|\mathbb{P}_n \left\{L(Y, f(X)) - H(X, \beta^0, f(X))\right\}\right|.$$

By the central limit theorem, $a_n = \mathrm{O}_p(n^{-1/2})$. The term $h_n$ is a supremum of a random process over $\mathcal{H}_n$. The following lemma gives the order of $h_n$.

**Lemma 4.1.** *Let Assumption 4.1 hold. Suppose the loss function $L$ satisfies Condition 4.2. Then,*

$$h_n = \mathrm{O}_p\left(n^{-\left\{\frac{1}{2} - \frac{qd}{2}\right\}} \lambda^{-\frac{q}{2}}\right).$$

We will discuss more about the functional space $\mathcal{H}_n$ in the proof of Lemma 4.1.

**Lemma 4.2.** *Let Assumptions 3.1 and 4.1 hold. Suppose the loss function $L$ satisfies Condition 4.2. Assume $|\widehat{\beta} - \beta^0| = \mathrm{O}_p(n^{-1/2})$. Then, the estimation error $\mathrm{Err}_{2,n}$ defined in (4.1) is of $\mathrm{O}_p(n^{-\frac{1}{2}+qd}\lambda^{-q})$.*

Under the quadratic loss with $q = 1$, Lemmas 4.1 and 4.2 imply that

$$h_n = \mathrm{O}_p\left(n^{-\frac{1}{2}+\frac{d}{2}}\lambda^{-\frac{1}{2}}\right) \quad \text{and} \quad \mathrm{Err}_{2,n} = \mathrm{O}_p\left(n^{-\frac{1}{2}+d}\lambda^{-1}\right).$$

With the convergence orders of $a_n$, $h_n$, and $\mathrm{Err}_{2,n}$ in position, we are ready to show the universal consistency and derive the learning rate of the doubly-robust kernel machine $f_{D,\lambda}^{\widehat{W},\widehat{H}}$.

**Theorem 4.2.** *Let Assumptions 2.1, 2.2, 3.1, 3.2, and 4.1 hold. Suppose that the universal kernel $k$ satisfies Condition 4.1 and the loss function $L$ is the quadratic loss. Assume that either $|\hat{\pi}(X) - \pi^0(X)| = \mathrm{O}_p(n^{-1/2})$ or $|\widehat{\beta} - \beta^0| = \mathrm{O}_p(n^{-1/2})$. Suppose $\lambda_n$ is a sequence in $(0, 1)$ satisfying $\lambda_n \to 0$ and*

$$\lambda_n n^{\min\left(\frac{1}{2p+2}, \frac{1}{2}-d\right)-d} \longrightarrow \infty.$$

*Then, the doubly-robust kernel machine $f_{D,\lambda}^{\widehat{W},\widehat{H}}$ in (3.8) is $\mathcal{P}$-universally consistent, i.e., $\mathcal{R}_{L,\mathrm{P}}(f_{D,\lambda}^{\widehat{W},\widehat{H}}) \xrightarrow{\mathrm{P}} \mathcal{R}_{L,\mathrm{P}}^*$ for all $P \in \mathcal{P}$.*

The proof of Theorem 4.2 is based on the oracle inequality derived for the doubly-robust kernel machines in Theorem C.2 of Appendix C.

**Corollary 4.2.** *Let Assumptions 2.1, 2.2, 3.1, 3.2, 4.1, and 4.2 hold. Suppose that the universal kernel $k$ satisfies Condition 4.1 and the loss function $L$ is the quadratic loss. Assume that either $|\hat{\pi}(X) - \pi^0(X)| = \mathrm{O}_p(n^{-1/2})$ or $|\widehat{\beta} - \beta^0| = \mathrm{O}_p(n^{-1/2})$. Then, the learning rate of the doubly-robust kernel machine $f_{D,\lambda}^{\widehat{W},\widehat{H}}$ in (3.8) is $n^{\left\{-\min\left(\frac{1}{2p+2}, \frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}$.*

When the kernel $k$ is Gaussian and $d = 0$, the learning rate is $n^{-\frac{\gamma}{2(\gamma+1)(p+1)}}$, which by (i) of Remark 4.1 holds for all $p > 0$ and therefore is equivalent to $n^{-\frac{\gamma}{2\gamma+2}+\epsilon}$ for arbitrarily small $\epsilon > 0$.

**Remark 4.2.** *When $L$ is the quadratic loss, for the doubly-robust kernel machine estimator, using the similar argument as in Appendix A.2, we can show that one can estimate $\mathrm{E}\{L(Y, f(X)) \mid X = x\}$ by estimating $\mathrm{E}(Y \mid X) \equiv \mu^0(X)$ directly.*

*Let $\widehat{\mu}(X)$ be the estimator of $\mu^0(X)$. In this case, the requirement $\widehat{\beta} \xrightarrow{\mathrm{P}} \beta^*$ in Assumption 3.2 can be replaced with $\widehat{\mu}(X) \xrightarrow{\mathrm{P}} \mu^*(X)$ where $\mu^*(X)$ does not necessarily equal $\mu^0(X)$. The requirement $|\widehat{\beta} - \beta^0| = \mathrm{O}_p(n^{-1/2})$ that in Lemma 4.2, Theorem 4.2, and Corollary 4.2 can be replaced with $\left|\widehat{\mu}(X) - \mu^0(X)\right| = \mathrm{O}_p(n^{-1/2})$. We have the following theoretical results for the doubly-robust kernel machines.*

1. *The estimation error of the conditional risk $\mathrm{Err}_{2,n}$ in Lemma 4.2 is of $\mathrm{O}_p(n^{-\frac{1}{2}+\frac{d}{2}}\lambda^{-\frac{1}{2}})$. (The proof is given in Remark C.1 of Appendix C.8.)*
2. *When $|\widehat{\pi}(X) - \pi^0(X)| = \mathrm{O}_p(n^{-1/2})$, the $\mathcal{P}$-universally consistency in Theorem 4.2 and the learning rate in Corollary 4.2 remain to hold.*
3. *When $\left|\widehat{\mu}(X) - \mu^0(X)\right| = \mathrm{O}_p(n^{-1/2})$, the $\mathcal{P}$-universally consistency in Theorem 4.2 holds if*

$$\lambda_n n^{\min\left(\frac{1}{2p+2}, \frac{1}{2}-\frac{d}{2}\right)-d} \longrightarrow \infty.$$

*The learning rate in Corollary 4.2 is $n^{\left\{-\min\left(\frac{1}{2p+2}, \frac{1}{2}-\frac{d}{2}\right)+d\right\}\frac{\gamma}{\gamma+1}}$. The proofs are given in Remark C.2 of Appendix C.9 and Remark C.3 of Appendix C.10.*

**Remark 4.3.** *Let $L$ be the quadratic loss and $\widehat{\mu}(X)$ be the estimator of $\mu^0(X)$. Assume that the conditions in Theorem 4.1 hold for the weighted-complete-case kernel machine and the conditions in Theorem 4.2 hold for the doubly-robust kernel machine.*

*(i) First, when $\lambda_n$ satisfies*

$$\lambda_n n^{\min\left(\frac{1}{2p+2}, \frac{1}{2}-d\right)-d} \longrightarrow \infty,$$

*the weighted-complete-case kernel machine is $\mathcal{P}$-universally consistent. When $|\widehat{\pi}(X) - \pi^0(X)| = \mathrm{O}_p(n^{-1/2})$ and $\lambda_n$ satisfies*

$$\lambda_n n^{\min\left(\frac{1}{2p+2}, \frac{1}{2}-d\right)-d} \longrightarrow \infty,$$

*or $\left|\widehat{\mu}(X) - \mu^0(X)\right| = \mathrm{O}_p(n^{-1/2})$ and $\lambda_n$ satisfies*

$$\lambda_n n^{\min\left(\frac{1}{2p+2}, \frac{1}{2}-\frac{d}{2}\right)-d} \longrightarrow \infty$$

*the doubly-robust kernel machine is $\mathcal{P}$-universally consistent. This shows that the doubly-robust kernel machine requires no stronger condition than the weighted-complete-case kernel machine to warrant $\mathcal{P}$-universally consistency.*

*(ii) Second, when Assumption 4.2 holds, the learning rate of the weighted-complete-case kernel machine is*

$$n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}.$$

*When* $|\widehat{\pi}(X)-\pi^0(X)| = \mathrm{O}_p(n^{-1/2})$*, the learning rate of the doubly-robust kernel machine is*

$$n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}.$$

*When* $\left|\widehat{\mu}(X)-\mu^0(X)\right| = \mathrm{O}_p(n^{-1/2})$ *the learning rate of the doubly-robust kernel machine is*

$$n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-\frac{d}{2}\right)+d\right\}\frac{\gamma}{\gamma+1}}.$$

*This indicates that the doubly-robust kernel machine has equal or faster learning rate than the weighted-complete-case kernel machine.*

## 5. Simulation

We conduct simulation studies to compare the finite-sample performance of the proposed kernel methods with some existing methods in terms of predicting a new response of either a regression or classification model and estimating the mean of a regression model.

We denote the competing methods as follows.

**Reg:** the linear regression method which uses only complete cases.
**SSLR:** the semi-supervised linear regression method by Azriel et al. (2016) which takes into account the missing responses.
**BEDR:** the bounded, efficient and doubly-robust estimation by Tan (2010).
**BRDR:** the bias-reduced doubly-robust estimation by Vermeulen and Vansteelandt (2015).
**CC:** the naive kernel machines which use only the complete observations.
**WCC:** the proposed weighted-complete-case kernel machines.
**DR:** the proposed doubly-robust kernel machines.

Note that the BEDR and BRDR are only used to compare the population mean estimation since they are not designed for predicting new responses. For all kernel machine methods, we use Gaussian RBF kernel function, where the kernel width parameter and the tuning parameter $\lambda$ are chosen by (five-fold) cross validation.

For the WCC, the missing mechanism needs to be estimated. For the BEDR, BRDR, both the missing mechanism and the condition mean $E(Y \mid X)$ need to be estimated. For the DR, both the missing mechanism and the conditional distribution $F_{Y|X}\left(y \mid X, \beta^0\right)$ need to be estimated. Here, we consider the following four scenarios regarding misspecification.

**S-1** Both the missing mechanism and the conditional distribution working model are correctly specified.
**S-2** The missing mechanism is misspecified while the conditional distribution working model is correctly specified.

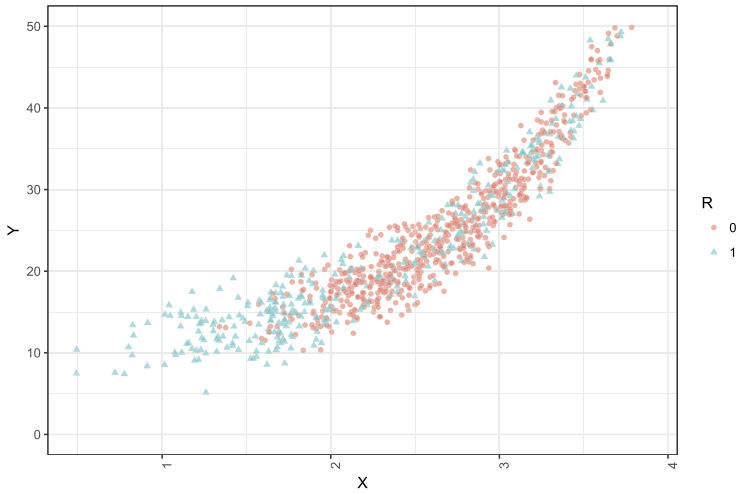


FIG 1. *Plot of Setting 1 with sample size 1000: The blue triangle points are observed responses and the red circle points are missing responses. Observations with X on [0,2] have lower probability to be generated than X on (2,4] and is more easily observed.*

**S-3** The conditional distribution working model is misspecified while the missing mechanism is correctly specified.

**S-4** Both the missing mechanism and the conditional distribution working model are misspecified.

### *5.1. Setup*

We consider the four settings to generate data with missing responses.

In the first setting, the response is generated by

$$Y = \exp(X) + U_2 + U_3 + U_4 + U_5 + \varepsilon,$$

where $X \sim 4 \times \text{Beta}(5,3)$, $U_2, \ldots U_5 \sim \text{Unif}(0,4)$, $\varepsilon \sim N(0,1)$, and $X$, $U_2 \ldots U_5$, and $\varepsilon$ are mutually independent. The missing mechanism is given by

$$\text{P}(R = 1 \mid X) = \begin{cases} \left[1 + \exp\left\{\frac{9}{2}(X-2)\right\}\right]^{-1} & X \in (0,2], \\ \left[1 + \exp\left\{-(X-4)\right\}\right]^{-1} & X \in (2,4). \end{cases}$$

The missing rates for $X$ in $(0,2]$ and $(2,4)$ are about 22% and 77%, respectively. The overall missing rate is about 64%. In this example, both the probability of the appearance of the large value of $Y$ and the missing rate increase in $X$. That means larger responses are more likely to be missed, as illustrated in Figure 1. Thus, ignoring the missingness will yield a biased estimator of the nonlinear predictor of $Y$, as the estimation is based on the smaller values of $X$.

Setting 2 is an example of the classification used by Laber and Murphy (2011). The response is generated as

$$Y = \text{sign}\left(X_2 - 0.16X_1^2 - 1 + \varepsilon\right),$$

where $X_1, X_2 \sim \text{Unif}(0,5)$, $\varepsilon \sim N(0, 0.25)$. The missing mechanism is

$$\text{P}\left(R = 1 \mid X\right) = \text{logit}\{1.5(X_2 - X_1)\}.$$

The missing rate when $Y = 1$ and $-1$ are about 20% and 84%, respectively. This implies that the positive labels of $Y$ are more easily observed. The overall missing rate is about 50%.

Settings 3 and 4 are taken from Liu et al. (2007), where they studied the generic pathway effect of the prostate-specific antigen (PSA), a biomarker for prostate cancer screening. Consider generating the response by a generic regression model through

$$Y = 3\cos(X_1) + 2U + h(X_1, \ldots, X_p) + \varepsilon,$$

where $U, X_1, \ldots X_p \sim \text{Unif}(0,1)$, $\varepsilon \sim N(0,1)$, which are mutually independent, and $h(\cdot)$ is a centered smooth function. In Setting 3, $p = 5$ and

$$\begin{aligned} h(X_1, \ldots, X_5) = {} & 10\cos(X_1) - 15X_2^2 + 10\exp(-X_3)X_4 - 8\sin(X_5)\cos(X_3) \\ & + 20X_1X_5. \end{aligned}$$

The missing mechanism is given by

$$\text{P}\left(R = 1 \mid X\right) = \text{logit}\left(-\frac{4\log 3}{3} + \frac{2\log 3}{3}\sum_{i=1}^{5}\frac{X_i}{5}\right).$$

In Setting 4, $p = 10$ and

$$\begin{aligned} & h\left(X_1, \ldots, X_{10}\right) \\ = {} & 10\cos(X_1) - 15X_2^2 + 10\exp(-X_3)X_4 - 8\sin(X_5)\cos(X_3) \\ & + 20X_1X_5 + 9X_6\sin(X_7) - 8\cos(X_6)X_7 + 20X_8\sin(X_9)\sin(X_{10}) \\ & - 15X_8^3 - 10X_8X_9 - \exp(X_{10})\cos(X_{10}). \end{aligned}$$

The missing mechanism is given by

$$\text{P}\left(R = 1 \mid X\right) = \text{logit}\left(-\frac{4\log 3}{3} + \frac{2\log 3}{3}\sum_{i=1}^{10}\frac{X_i}{10}\right).$$

In these two settings, the missing rates are both about 75%.

For the misspecification of the missing mechanism, we use the generalized linear model with the probit link to estimate the missing mechanism. For the misspecification of the conditional distribution working model, we used the simple linear model.

Throughout, we set the (training) sample size $n$ to be 100, 200, 400, and 800, respectively, for each data set.

We examine the performance of the competing estimation methods by the following two quantities. The first quantity is the empirical risk of predicting a new response, given by

$$Q_1 = \frac{1}{N}\sum_{i=1}^{N}\{Y_i - \widehat{f}(X_i)\}^2,$$

where $\widehat{f}$ denotes a generic estimator of the decision function obtained by various methods, $\{(X_i, Y_i) : i = 1, \ldots, N\}$ is the testing sample of size $N = 100{,}000$. The second quantity is the absolute bias of estimating the overall expectation of the response, given by

$$Q_2 = |\widehat{\mathrm{E}}(Y) - \mathrm{E}(Y)|,$$

where $\widehat{\mathrm{E}}(Y) = N^{-1}\sum_{i=1}^{N}\widehat{f}(X_i)$ and $\mathrm{E}(Y)$ is approximated by $N^{-1}\sum_{i=1}^{N} Y_i$, $\{(X_i, Y_i) : i = 1, \ldots, N\}$ and $N$ are as defined for $Q_1$. Throughout, we set the number of replications (simulation size) to be 100.

### 5.2. Results

Table 2 presents the sample mean, median, and standard deviation of $Q_1$ (over 100 replications) obtained by the five competing methods (Reg, SSL, CC, WCC, DR) under four settings, four different sample sizes, and four different scenarios of misspecification. The corresponding distributions of $Q_1$ are displayed by the boxplots in Figures 2 to 5.

It is seen that (i) when little information about the conditional distribution or regression model is given, e.g., there is no information about the correct covariates, the proposed weighted-complete-case kernel machine performs better than Reg, SSL, and CC, especially for large sample-size datasets; (ii) when either the conditional distribution model or the missing mechanism is correctly specified, the proposed doubly-robust kernel machine performs the best; (iii) for Settings 2 and 3, even if both the missing mechanism and the condition distribution model are misspecified, the proposed doubly-robust kernel machine performs the best.

Table 3 presents the sample mean, median, and standard deviation of $Q_2$ (over 100 replications) obtained by the seven competing methods (Reg, SSL, BEDR, BRDR, CC, WCC, DR) under Settings 1, 3, and 4, four different sample sizes, and four different scenarios of misspecification. The corresponding distributions of $\widehat{\mathrm{E}}(Y)$ are displayed by the boxplots in Figures 6 to 8.

It is seen that (i) the proposed weighted-completed-case kernel machine outperforms the considered five existing methods; (ii) the proposed doubly-robust kernel machine yields the most precise estimation in terms of absolute bias, even when the missing mechanism and the conditional distribution working model are both misspecified.

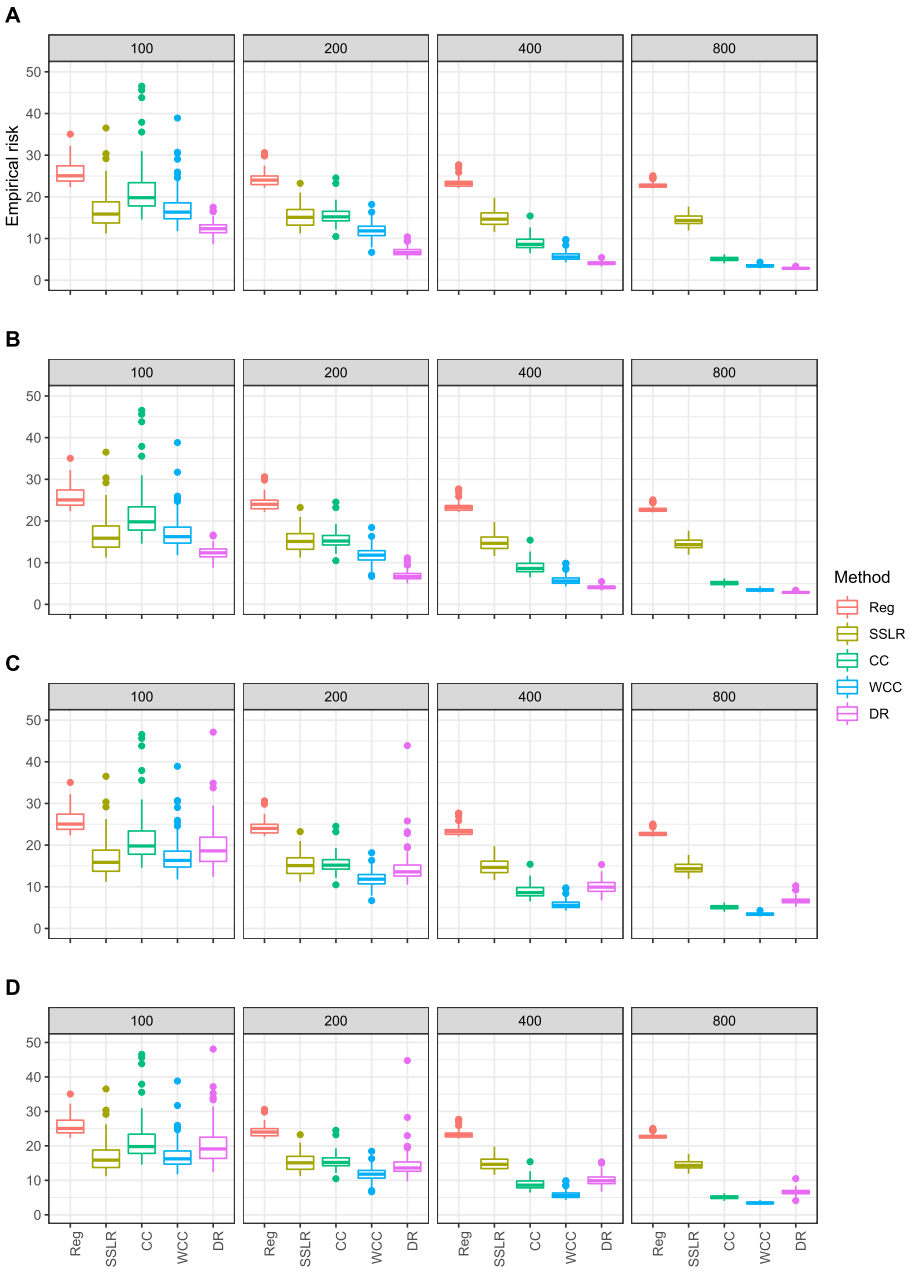


FIG 2. *Boxplots of the empirical risk over 100 replications obtained by the five competing methods under various sample size of n in Setting 1. Rows A-D present the results under misspecification scenarios S-1 to S-4, respectively. The results of Reg, SSLR, and CC are the same across all four scenarios/rows.*

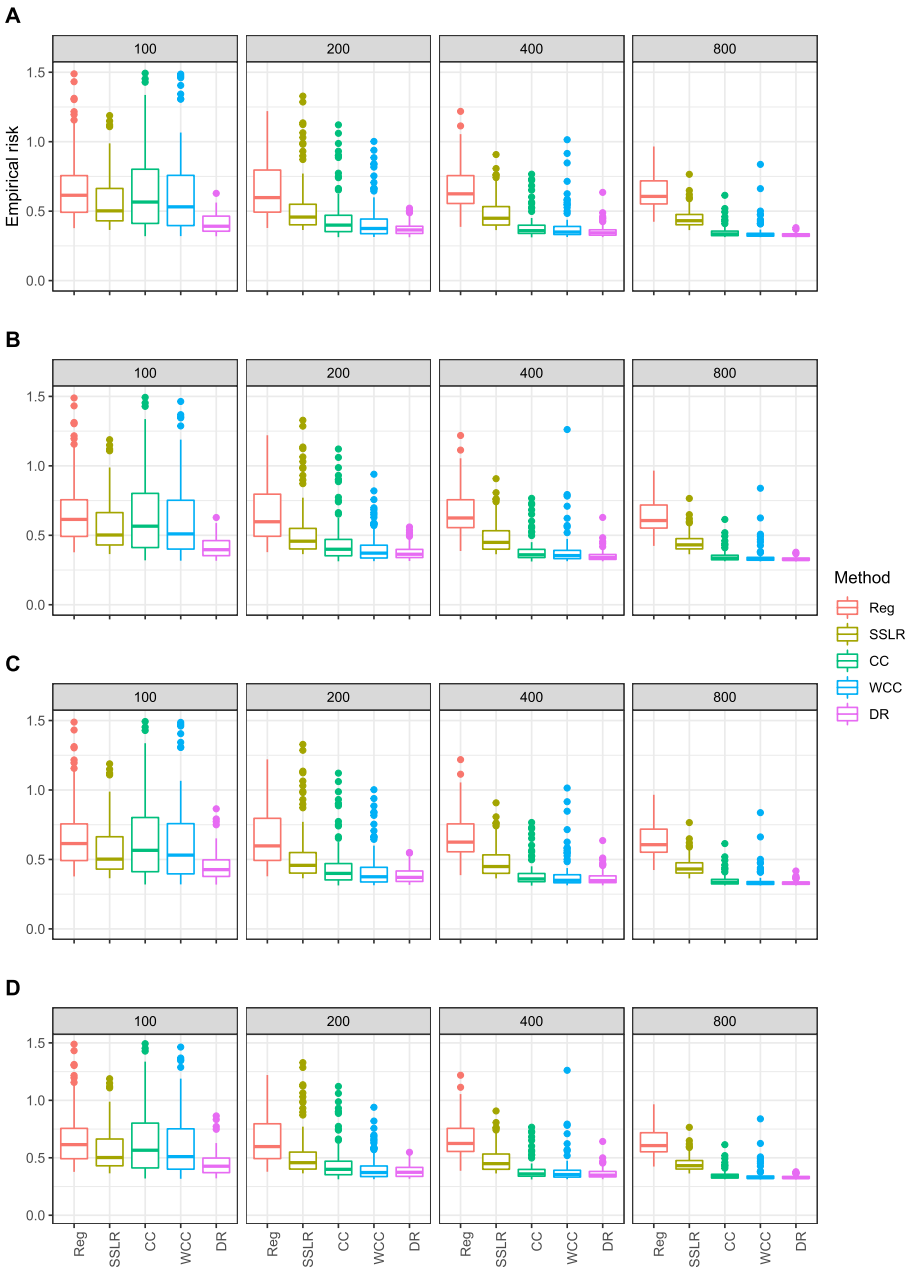


FIG 3. *Boxplots of the empirical risk over 100 replications obtained by the five competing methods under various sample size of n in Setting 2. Rows A-D present the results under misspecification scenarios S-1 to S-4, respectively. The results of Reg, SSLR, and CC are the same across all four scenarios/rows.*

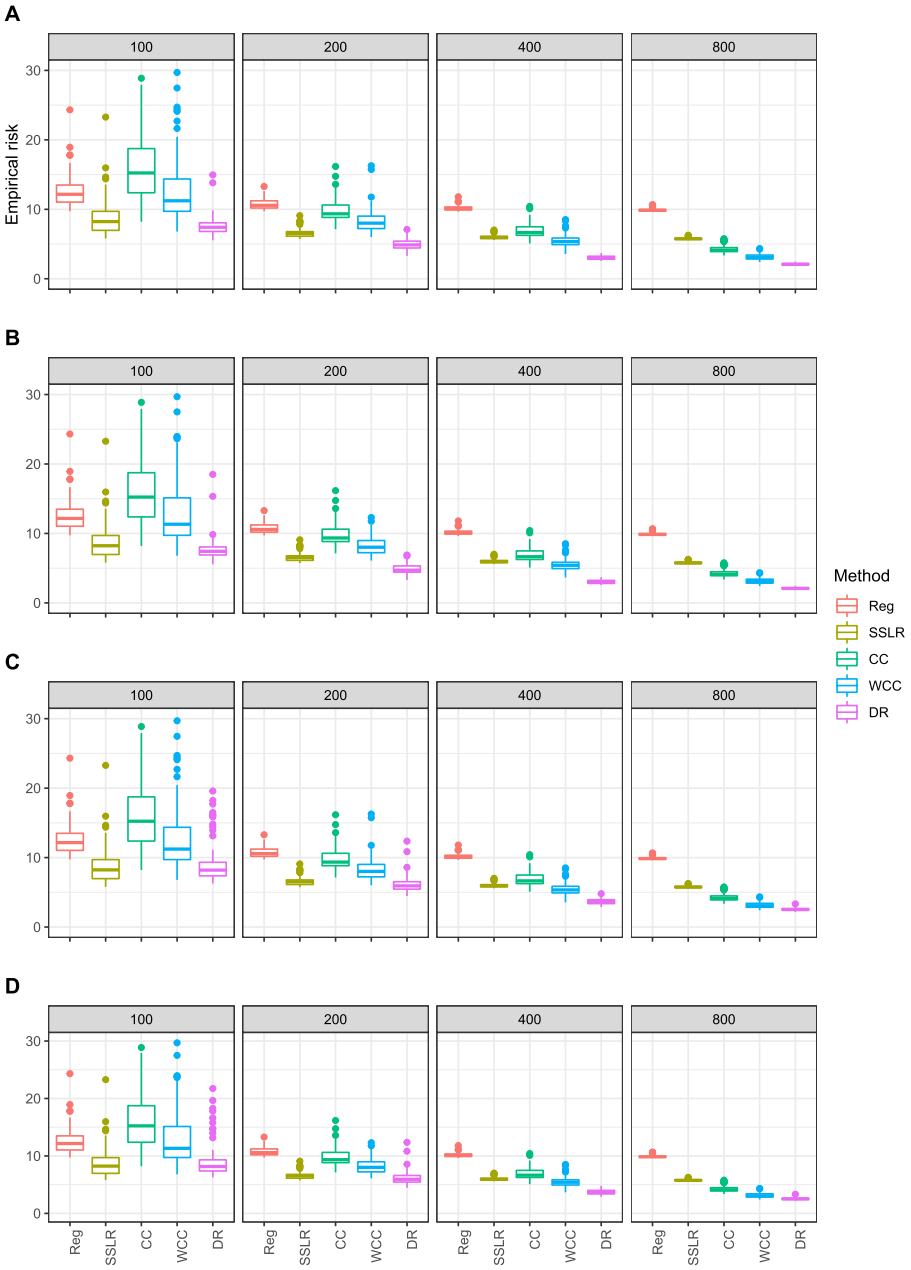


FIG 4. *Boxplots of the empirical risk over 100 replications obtained by the five competing methods under various sample size of n in Setting 3. Rows A-D present the results under misspecification scenarios S-1 to S-4, respectively. The results of Reg, SSLR, and CC are the same across all four scenarios/rows.*

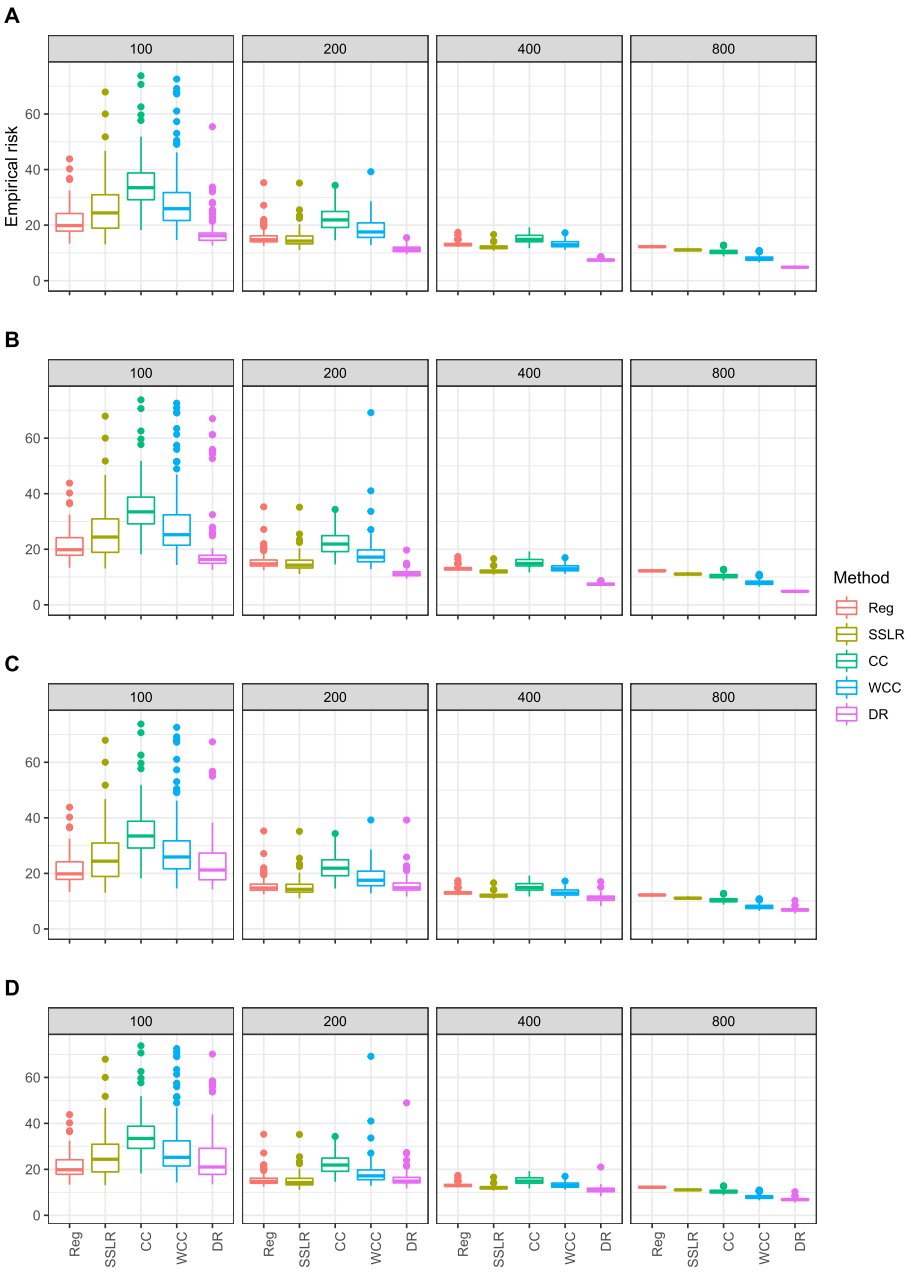


FIG 5. *Boxplots of the empirical risk over 100 replications obtained by the five competing methods under sample size of n in Setting 4. Rows A-D present the results under misspecification scenarios S-1 to S-4, respectively. The results of Reg, SSLR, and CC are the same across all four scenarios/rows.*

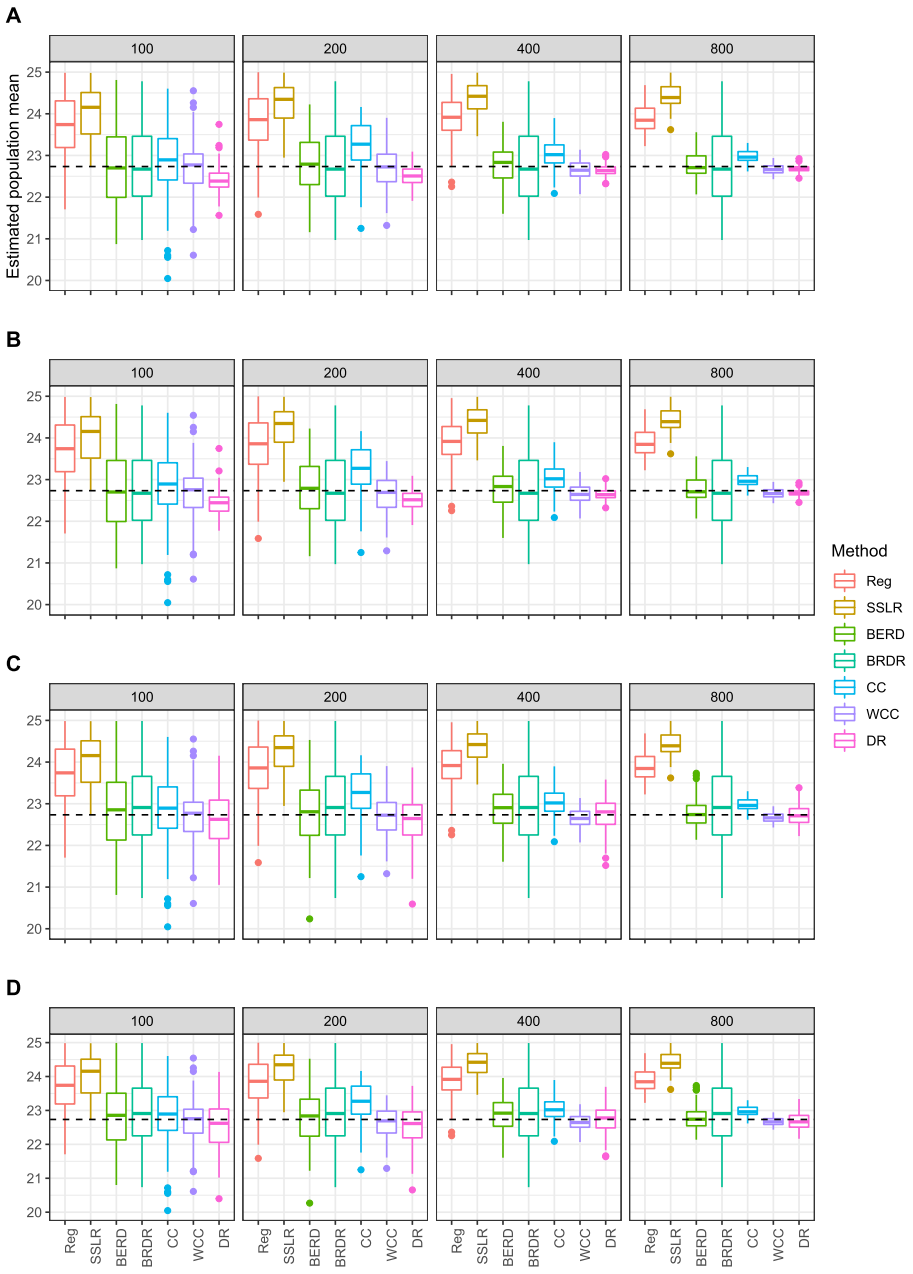


FIG 6. *Boxplots of the estimated population mean over 100 replications obtained by the seven competing methods under various sample size of n in Setting 1. Rows A-D present the results under misspecification scenarios S-1 to S-4, respectively. The dashed line is the approximated mean. The results of Reg, SSLR, and CC are the same across all four scenarios/rows.*

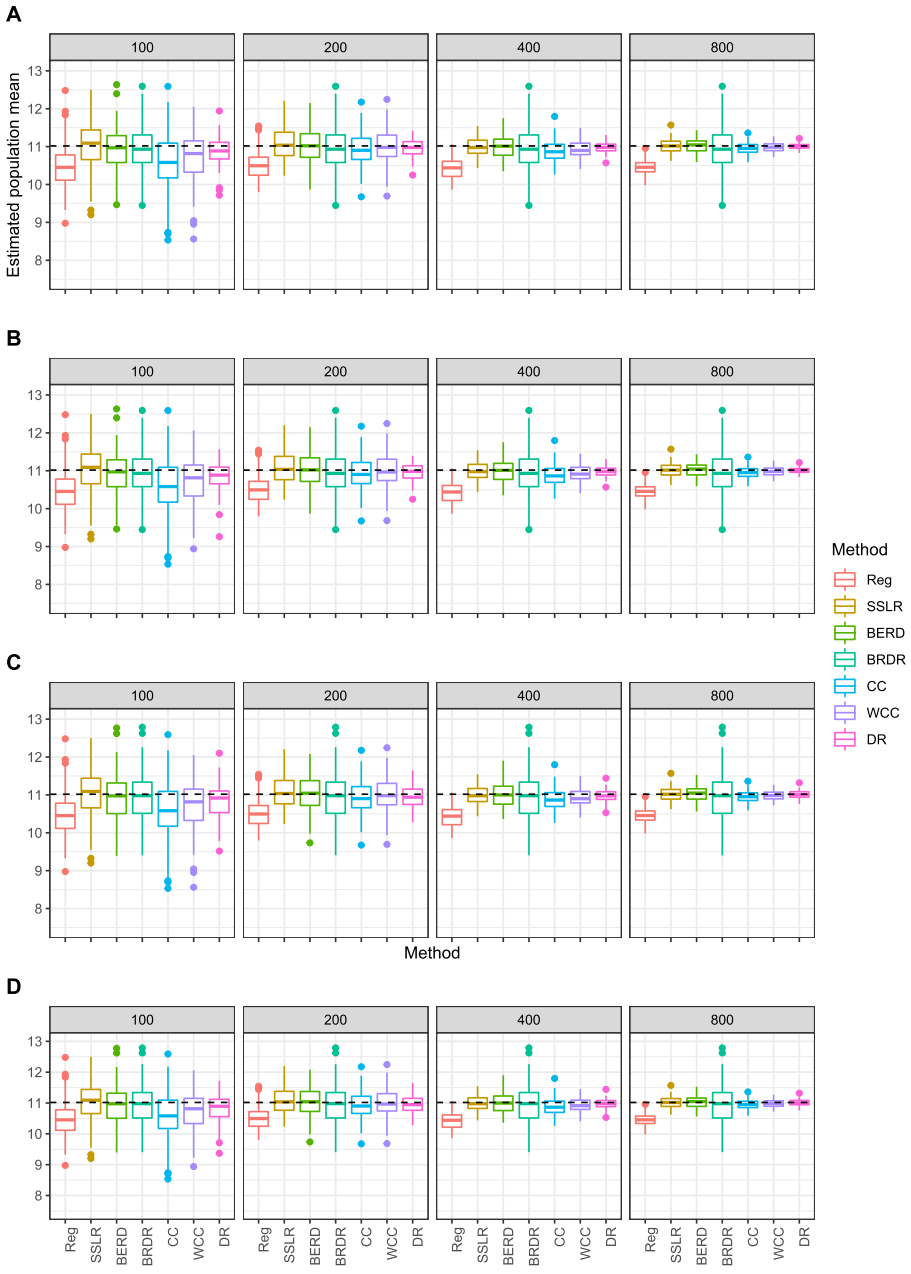


FIG 7. *Boxplots of the estimated population mean over 100 replications obtained by the seven competing methods under various sample size of n in Setting 3. Rows A-D present the results under misspecification scenarios S-1 to S-4, respectively. The dashed line is the approximated mean. The results of Reg, SSLR, and CC methods are the same across all four scenarios/rows.*

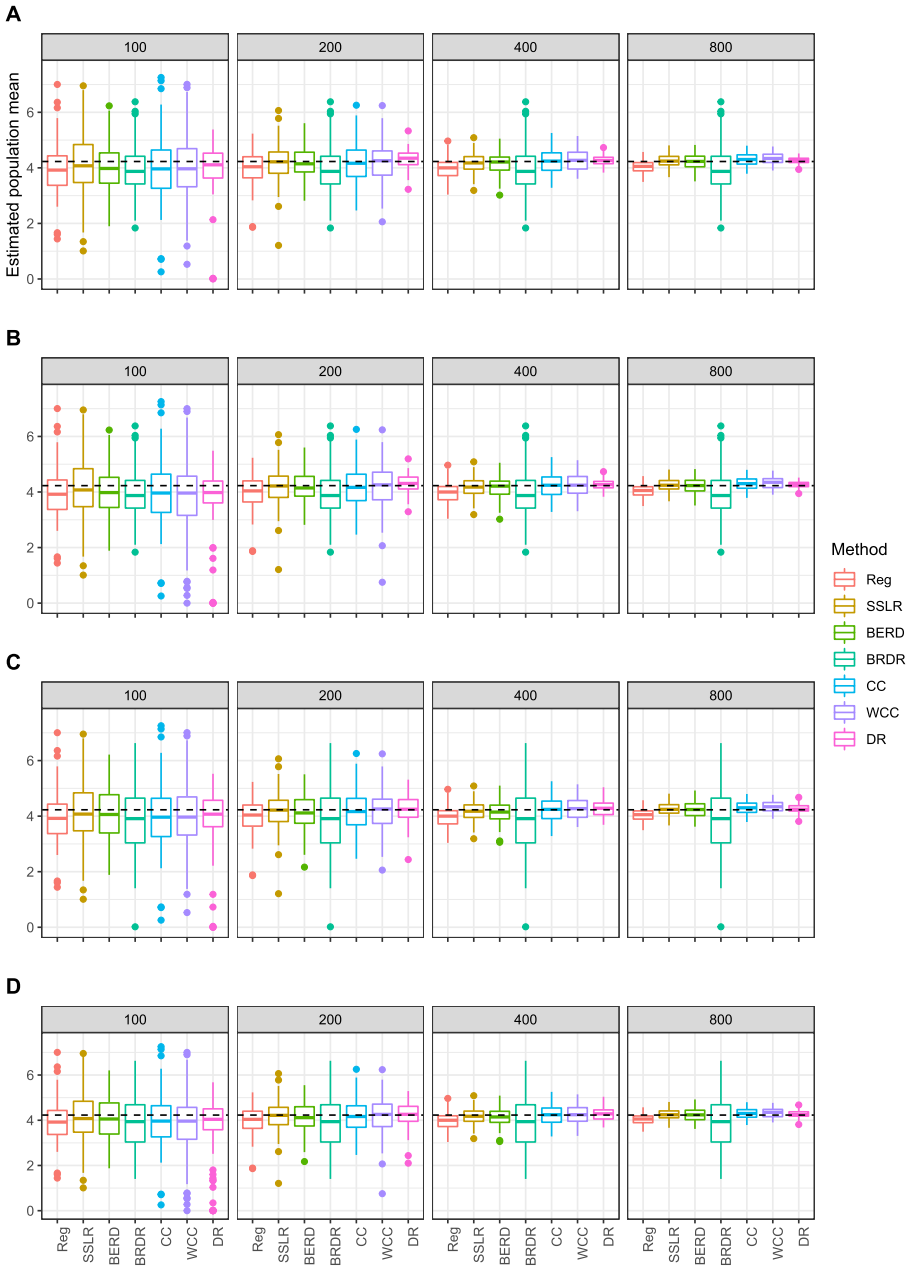


FIG 8. *Boxplots of the estimated population mean over 100 replications obtained by the seven competing methods under various sample size of n in Setting 4. Rows A-D present the results under misspecification scenarios S-1 to S-4, respectively. The dashed line is the approximated mean. The results of Reg, SSLR, and CC methods are the same across all four scenarios/rows.*

## 6. Application to Los Angeles homeless population data

We applied the proposed kernel machine methods to the Los Angeles homeless data, where the goal is to estimate the number of homeless in each tract (Kriegler and Berk, 2010; Azriel et al., 2016). The data set contains information about 2,054 census tracts in the Los Angeles county. Due to the budget limitation, some tracts were not visited, and consequently, the numbers of homeless in these tracts are missing. The missing mechanism depends on the service provision areas (SPA) to which the tract belongs. We used the same covariates of "Industrial" (percentage of land for industrial purposes), "PctVacant" (percentage of unoccupied housing units), "Commercial" (percentage of land used for commercial purposes), "MedianHouseIncome", "Residential" (percentage of land used for residential purposes) and "PctMinority" (percentage population that is non-Caucasian) as in Azriel et al. (2016).

We pre-process the data in the following three steps. First, similar to Azriel et al. (2016), we delete all tracts with zero median household income and the 244 highly-populated "hot tracts", which results in 1,797 tracts for subsequent analysis. The missing rate in this data set is about 85%. Second, to correct the skewness of variables (as a common preprocess for the SVM), we applied log transformation to the response variable of the number of homeless, the covariates "Industrial", "PctVacant", "Commercial", and "MedianHouseIncome". Third, all variables are standardized to have zero mean and unit standard deviation.

The boxplots of the data after transformation and normalization are shown in Figure 9.

We applied the same five competing methods (Reg, SSLR, CC, WCC, and DR) as in Table 2 to the dataset. We used the semi-supervised linear regression method to estimate the conditional expectation $\mu(X, \widehat{\beta})$ in the doubly-robust kernel machine. In this data, the missing mechanism is known; the weights are the inverse probability of the number of tracts assigned to each SPA. To evaluate the performance of these methods, we randomly sample 1,597 tracts as the training set and use the remaining 200 tracts as the testing set. Since some of the responses (of the number of homeless) of these 200 tracts are missing, we consider the empirical risks with respect to two different loss functions. The first loss function is the complete-case quadratic loss in (3.1). The second loss function is the weighted complete-case quadratic loss in (3.4), where the weights are the inverse probability that tract would have been included in the survey. Tracts for which the responses are observed are used to compute the empirical risk.

Table 1 shows the sample mean, median, and standard deviation of the empirical risk over 100 replications obtained by the five competing methods. It is seen that overall these five methods produce similar results. The proposed weighted-complete-case kernel machines performed the best with the minimum mean of empirical risk together with the smallest variation in terms of the standard deviation. The proposed doubly-robust kernel machine did not show its superiority as in the simulation studies. It is possibly due to the inferior performance of the semi-supervised linear regression used for the augmentation term.

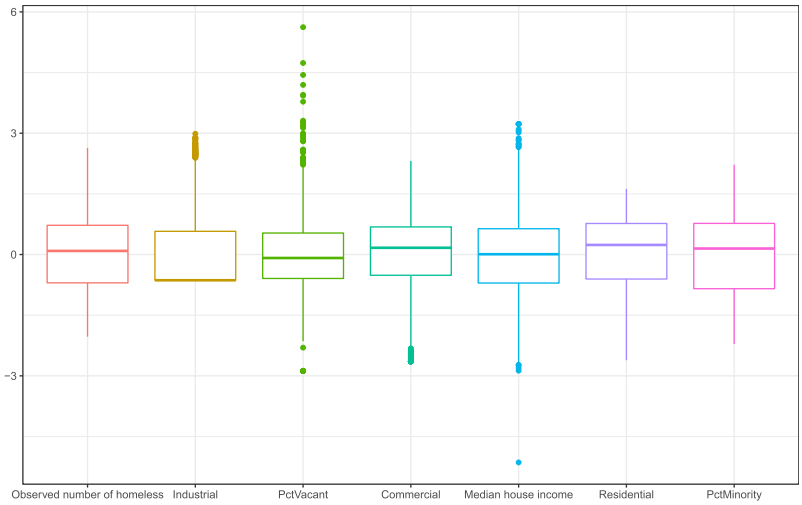


FIG 9. *From left to right, boxplots of the observed number of homeless, and the covariates Industrial, PctVacant, Commercial, and Median house income, after log transformation and normalization. The last two boxplots are of the covariates Residential and PctMinority after normalization.*

## 7. Conclusion and discussion

We proposed two kernel machines with missing responses. In particular, the proposed inverse-probability complete-case estimator can be applied under any convex loss function. The proposed doubly-robust estimator has the feature of being doubly-robust against misspecification of either the missing mechanism or the conditional distribution working model. The empirical risks of these new data-dependent loss functions are shown to be consistent under mild conditions. We also establish the universal consistency via the oracle inequalities for both kernel machines.

In the present work, we posited two parametric models for estimating the missing mechanism and the conditional distribution of $Y$ given the covatiates $X$. These two working models can be replaced by either semi-parametric models or nonparametric models such as smoothing kernel estimators and kernel machine estimators. The theoretical results can be adapted depending on the convergence rates of these working models. As recommended by Chernozhukov et al. (2018), one can split the dataset into two subsets and use the first one to estimate the two working models and the second one to obtain the kernel machine decision function. This strategy is worth a new investigation for our method. Other possible directions are some generalizations and extensions, for example, the extension of the doubly-robust estimator to include other convex loss functions. and the generalization of the data-dependent loss function to handle the situation of missing covariates. We shall report these investigations in separate works.

TABLE 1
*Sample median, mean, and standard deviation of the empirical risk under two loss functions over 100 replications obtained by the five competing methods.*

| | | Reg | SSLR | CC | WCC | DR |
|---|---|---|---|---|---|---|
| mean | weighted quadratic loss | 9488 | 9735 | 9488 | 9100 | 9585 |
| | quadratic loss | 1097 | 1116 | 1096 | 1054 | 1093 |
| median | weighted | 9399 | 9868 | 8441 | 8293 | 9320 |
| | not weighted | 943 | 1007 | 973 | 902 | 952 |
| std | weighted | 6811 | 6691 | 7173 | 6825 | 7141 |
| | not weighted | 872 | 842 | 921 | 884 | 914 |

## Appendix A: Computation details in Subsection 3.4

### A.1. Weighted-complete-case kernel machines

On replacing $f(x)$ by (2.2), the regularized empirical risk in (3.5) is expressed as

$$\begin{aligned} g(\boldsymbol{\alpha}) =& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{R_i\left\{Y_i-\sum_{j=1}^{n}\alpha_j k(X_i,X_j)\right\}^2}{\widehat{\pi}(X_i)}+\lambda\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(X_i,X_j)\\ =&(AY-AK\boldsymbol{\alpha})^{\intercal}(A\mathbf{Y}-AK\boldsymbol{\alpha})/n+\lambda\boldsymbol{\alpha}^{\intercal}K\alpha\\ =&(Y^{\intercal}W\mathbf{Y}-2\boldsymbol{\alpha}^{\intercal}K^{\intercal}W\mathbf{Y}+\boldsymbol{\alpha}^{\intercal}K^{\intercal}WK\boldsymbol{\alpha})/n+\lambda\boldsymbol{\alpha}^{\intercal}K\boldsymbol{\alpha}, \end{aligned}$$

where $\boldsymbol{\alpha}$, $A$ and $W$ are defined as in Section 3.4.

Setting $\partial g(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}=0$ leads to (3.12).

### A.2. Doubly-robust kernel machines

(i) For regression, by (2.2) and (3.9),

$$\begin{aligned} H(X_i,\widehat{\beta},f(X_i)) =& \left\{\mu(X_i,\widehat{\beta})-\sum_{j=1}^{n}\alpha_j k(X_i,X_j)\right\}^2\\ &+\frac{\sum_{j=1}^{n}R_j\left\{Y_j-\mu(X_j,\widehat{\beta})\right\}^2}{\sum_{j=1}^{n}R_j}, \end{aligned}$$

where the second term is free of $\boldsymbol{\alpha}$. Denote the first term by $H_1(X_i,\widehat{\beta},f(X_i))$.

The regularized empirical risk in (3.8) with $H(X_i, \widehat{\beta}, f(X_i))$ replaced by $H_1(X_i, \widehat{\beta}, f(X_i))$ is expressed as

$$\begin{aligned} g_1(\boldsymbol{\alpha}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i\left\{Y_i - \sum_{j=1}^{n}\alpha_j k(X_i, X_j)\right\}^2}{\widehat{\pi}(X_i)} - \frac{R_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)} H_1\left(X_i, \widehat{\beta}, f(X_i)\right)\right] \\ &\quad + \lambda\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(X_i, X_j) \\ &= \frac{1}{n}\{\mathbf{Y}^\intercal W\mathbf{Y} - 2\boldsymbol{\alpha}^\intercal K^\intercal W\mathbf{Y} + \boldsymbol{\alpha}^\intercal K^\intercal W K\boldsymbol{\alpha} + \boldsymbol{\mu}(X, \widehat{\beta})^\intercal (I - W)\boldsymbol{\mu}(X, \widehat{\beta}) \\ &\quad - 2\boldsymbol{\alpha}^\intercal K^\intercal (I - W)\boldsymbol{\mu}(X, \widehat{\beta}) + \boldsymbol{\alpha}^\intercal K^\intercal (I - W)K\boldsymbol{\alpha}\} + \lambda\boldsymbol{\alpha}^\intercal K^\intercal\boldsymbol{\alpha}, \end{aligned}$$

where $\boldsymbol{\mu}(X, \widehat{\beta})$ and $W$ are defined as in Subsection 3.4,

Setting $\partial g_1(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = 0$ solves (3.13).

(ii) For classification, by (3.11),

$$\begin{aligned} H(X_i, \widehat{\beta}, f(X_i)) &= 1 + f(X_i)^2 + 2f(X_i) - 4f(X_i)\text{logit}(X_i^\intercal\widehat{\beta}) \\ &= \left[f(X_i) - \left\{2\text{logit}(X_i^\intercal\widehat{\beta}) - 1\right\}\right]^2 + 1 - \left\{2\text{logit}(X_i^\intercal\widehat{\beta}) - 1\right\}^2, \end{aligned}$$

where the second term is free of $\boldsymbol{\alpha}$. Denote the first term by $H_2(X_i, \widehat{\beta}, f(X_i))$.

The regularized empirical risk in (3.8) with $H(X_i, \widehat{\beta}, f(X_i))$ replaced by $H_2(X_i, \widehat{\beta}, f(X_i))$ is expressed as

$$\begin{aligned} g_2(\boldsymbol{\alpha}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i\left\{Y_i - \sum_{j=1}^{n}\alpha_j k(X_i, X_j)\right\}^2}{\widehat{\pi}(X_i)} - \frac{R_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)} H_2(X_i, \widehat{\beta}, f(X_i))\right] \\ &\quad + \lambda\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(X_i, X_j) \\ &= \frac{1}{n}[\mathbf{Y}^\intercal W\mathbf{Y} - 2\boldsymbol{\alpha}^\intercal K^\intercal W\mathbf{Y} + \boldsymbol{\alpha}^\intercal K^\intercal W K\boldsymbol{\alpha} \\ &\quad + \left\{2\text{logit}(X^\intercal\widehat{\beta}) - 1\right\}^\intercal (I - W)\left\{2\text{logit}(X^\intercal\widehat{\beta}) - 1\right\} \\ &\quad - 2\boldsymbol{\alpha}^\intercal K^\intercal (I - W)\left\{2\text{logit}(X^\intercal\widehat{\beta}) - 1\right\} + \boldsymbol{\alpha}^\intercal K^\intercal (I - W)K\boldsymbol{\alpha}] \\ &\quad + \lambda\boldsymbol{\alpha}^\intercal K^\intercal\boldsymbol{\alpha}. \end{aligned}$$

Setting $\partial g_2(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = 0$ solves (3.13).

## Appendix B: Tables of simulations

TABLE 2

*The median, mean, and standard deviation of the empirical risk for Settings 1-4. A superscript j means that the method was evaluated with respect to scenario **S-j**. For example, $WCC^1$ means the WCC method with correctly specified the missing mechanism; $DR^1$ means the DR method with correctly specified both the missing mechanism and the conditional distribution model.*

| Setting | n | | Reg | SSL | CC | $WCC^1$ | $WCC^2$ | $DR^1$ | $DR^2$ | $DR^3$ | $DR^4$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | median | 25.06 | 15.87 | 19.80 | 16.34 | 16.24 | 12.37 | 12.36 | 18.80 | 19.15 |
| | 100 | mean | 25.84 | 16.92 | 21.43 | 22.54 | 17.22 | 14.30 | 12.30 | 26.88 | 22.34 |
| | | std | 2.64 | 4.49 | 5.98 | 51.82 | 4.12 | 19.90 | 1.55 | 58.32 | 13.97 |
| | | median | 24.01 | 15.09 | 15.19 | 11.83 | 11.80 | 6.63 | 6.67 | 13.59 | 13.65 |
| | 200 | mean | 24.27 | 15.23 | 15.57 | 11.94 | 11.86 | 6.81 | 6.85 | 14.59 | 15.12 |
| | | std | 1.72 | 2.47 | 2.06 | 1.85 | 1.91 | 0.97 | 1.07 | 3.98 | 6.19 |
| | | median | 23.21 | 14.65 | 8.61 | 5.56 | 5.60 | 4.09 | 4.09 | 9.92 | 9.89 |
| | 400 | mean | 23.35 | 14.86 | 8.99 | 5.77 | 5.83 | 4.10 | 4.11 | 10.11 | 10.08 |
| | | std | 1.06 | 1.92 | 1.54 | 1.04 | 1.11 | 0.42 | 0.42 | 1.60 | 1.56 |
| | | median | 22.64 | 14.34 | 5.10 | 3.39 | 3.40 | 2.85 | 2.85 | 6.68 | 6.69 |
| | 800 | mean | 22.79 | 14.55 | 5.14 | 3.46 | 3.47 | 2.86 | 2.86 | 6.66 | 6.65 |
| | | std | 0.61 | 1.26 | 0.48 | 0.31 | 0.31 | 0.18 | 0.18 | 0.84 | 0.85 |
| 2 | | median | 0.64 | 0.51 | 0.57 | 0.54 | 0.51 | 0.39 | 0.40 | 0.43 | 0.43 |
| | 100 | mean | 0.73 | 0.63 | 0.67 | 0.65 | 0.63 | 0.41 | 0.41 | 0.45 | 0.45 |
| | | std | 0.33 | 0.32 | 0.32 | 0.32 | 0.31 | 0.07 | 0.07 | 0.11 | 0.11 |
| | | median | 0.60 | 0.46 | 0.40 | 0.38 | 0.37 | 0.36 | 0.36 | 0.37 | 0.37 |
| | 200 | mean | 0.67 | 0.53 | 0.46 | 0.43 | 0.41 | 0.37 | 0.38 | 0.39 | 0.38 |
| | | std | 0.23 | 0.21 | 0.17 | 0.15 | 0.12 | 0.05 | 0.05 | 0.05 | 0.05 |
| | | median | 0.63 | 0.45 | 0.36 | 0.35 | 0.35 | 0.34 | 0.34 | 0.35 | 0.35 |
| | 400 | mean | 0.67 | 0.48 | 0.39 | 0.39 | 0.39 | 0.36 | 0.35 | 0.37 | 0.36 |
| | | std | 0.19 | 0.11 | 0.09 | 0.12 | 0.12 | 0.04 | 0.04 | 0.05 | 0.05 |
| | | median | 0.61 | 0.43 | 0.34 | 0.33 | 0.33 | 0.33 | 0.32 | 0.33 | 0.33 |
| | 800 | mean | 0.63 | 0.45 | 0.35 | 0.35 | 0.35 | 0.33 | 0.33 | 0.33 | 0.33 |
| | | std | 0.11 | 0.07 | 0.04 | 0.07 | 0.07 | 0.01 | 0.01 | 0.02 | 0.01 |
| 3 | | median | 12.16 | 8.23 | 15.48 | 11.24 | 11.33 | 7.41 | 7.43 | 8.20 | 8.16 |
| | 100 | mean | 12.69 | 8.75 | 16.65 | 14.08 | 13.34 | 7.64 | 7.70 | 9.05 | 9.02 |
| | | std | 2.31 | 2.63 | 5.75 | 10.24 | 5.41 | 1.32 | 1.59 | 2.75 | 2.88 |
| | | median | 10.58 | 6.51 | 9.35 | 8.01 | 8.02 | 4.88 | 4.71 | 5.93 | 5.93 |
| | 200 | mean | 10.77 | 6.56 | 9.85 | 8.32 | 8.22 | 4.99 | 4.93 | 6.10 | 6.13 |
| | | std | 0.75 | 0.62 | 1.64 | 1.66 | 1.31 | 0.75 | 0.72 | 1.13 | 1.14 |
| | | median | 10.09 | 5.95 | 6.65 | 5.34 | 5.42 | 2.97 | 2.97 | 3.65 | 3.68 |
| | 400 | mean | 10.18 | 5.99 | 6.91 | 5.47 | 5.51 | 3.02 | 3.02 | 3.69 | 3.72 |
| | | std | 0.37 | 0.29 | 0.98 | 0.85 | 0.85 | 0.27 | 0.27 | 0.45 | 0.45 |
| | | median | 22.64 | 14.34 | 5.10 | 3.39 | 3.40 | 2.85 | 2.85 | 6.68 | 6.69 |
| | 800 | mean | 22.79 | 14.55 | 5.14 | 3.46 | 3.47 | 2.86 | 2.86 | 6.66 | 6.65 |
| | | std | 0.61 | 1.26 | 0.48 | 0.31 | 0.31 | 0.18 | 0.18 | 0.84 | 0.85 |
| 4 | | median | 19.83 | 24.40 | 33.46 | 25.91 | 25.59 | 16.24 | 16.36 | 21.22 | 23.12 |
| | 100 | mean | 21.55 | 26.36 | 35.15 | 29.47 | 30.77 | 19.97 | 23.26 | 26.70 | 31.65 |
| | | std | 5.67 | 10.22 | 9.68 | 12.23 | 14.92 | 13.35 | 17.42 | 14.90 | 19.74 |
| | | median | 14.76 | 14.27 | 21.88 | 17.54 | 17.17 | 11.18 | 11.18 | 14.83 | 14.90 |
| | 200 | mean | 15.59 | 15.17 | 22.61 | 18.39 | 18.71 | 11.30 | 11.37 | 15.68 | 15.90 |
| | | std | 3.01 | 3.26 | 4.30 | 3.93 | 6.58 | 1.06 | 1.32 | 3.40 | 4.31 |
| | | median | 12.97 | 11.99 | 14.89 | 12.94 | 12.99 | 7.45 | 7.45 | 11.12 | 11.17 |
| | 400 | mean | 13.15 | 12.07 | 15.15 | 13.14 | 13.17 | 7.45 | 7.45 | 11.13 | 11.12 |
| | | std | 0.84 | 0.86 | 1.62 | 1.30 | 1.29 | 0.42 | 0.42 | 1.28 | 1.42 |
| | | median | 22.64 | 14.34 | 5.10 | 3.39 | 3.40 | 2.85 | 2.85 | 6.68 | 6.69 |
| | 800 | mean | 22.79 | 14.55 | 5.14 | 3.46 | 3.47 | 2.86 | 2.86 | 6.66 | 6.65 |
| | | std | 0.61 | 1.26 | 0.48 | 0.31 | 0.31 | 0.18 | 0.18 | 0.84 | 0.85 |

TABLE 3. *The median, mean, and standard deviation of the estimated population mean absolute bias for Settings 1, 3, and 4. A superscript j means that the method was evaluated with respect to scenario* ***S-j****. For example, BEDR*$^1$ *means the BEDR method with correctly specified both the missing mechanism and the conditional distribution model; WCC*$^1$ *means the WCC method with correctly specified missing mechanism.*

| Setting | n | | Reg | SSLR | BEDR$^1$ | BEDR$^2$ | BEDR$^3$ | BEDR$^4$ | BRDR$^1$ | BRDR$^2$ | BRDR$^3$ | BRDR$^4$ | CC | WCC$^1$ | WCC$^2$ | DR$^1$ | DR$^2$ | DR$^3$ | DR$^4$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | median | 1.08 | 1.58 | 0.73 | 0.74 | 0.78 | 0.77 | 0.73 | 0.73 | 0.75 | 0.75 | 0.53 | 0.36 | 0.36 | 0.37 | 0.33 | 0.46 | 0.50 |
| | 100 | mean | 1.18 | 1.59 | 0.76 | 0.76 | 0.91 | 0.91 | 0.77 | 0.77 | 0.91 | 0.91 | 0.68 | 0.70 | 0.49 | 0.39 | 0.36 | 0.80 | 0.61 |
| | | std | 0.73 | 0.84 | 0.53 | 0.53 | 0.64 | 0.64 | 0.52 | 0.52 | 0.68 | 0.68 | 0.56 | 2.12 | 0.44 | 0.26 | 0.24 | 2.24 | 0.56 |
| | | median | 1.15 | 1.70 | 0.51 | 0.51 | 0.53 | 0.53 | 0.73 | 0.73 | 0.75 | 0.75 | 0.67 | 0.33 | 0.33 | 0.25 | 0.25 | 0.37 | 0.40 |
| | 200 | mean | 1.19 | 1.66 | 0.56 | 0.56 | 0.64 | 0.64 | 0.77 | 0.77 | 0.91 | 0.91 | 0.68 | 0.39 | 0.38 | 0.26 | 0.26 | 0.45 | 0.46 |
| | | std | 0.66 | 0.59 | 0.37 | 0.37 | 0.46 | 0.46 | 0.52 | 0.52 | 0.68 | 0.68 | 0.40 | 0.31 | 0.30 | 0.18 | 0.18 | 0.39 | 0.39 |
| | | median | 1.19 | 1.74 | 0.33 | 0.33 | 0.38 | 0.37 | 0.73 | 0.73 | 0.75 | 0.75 | 0.30 | 0.16 | 0.16 | 0.11 | 0.11 | 0.26 | 0.26 |
| | 400 | mean | 1.21 | 1.72 | 0.36 | 0.36 | 0.45 | 0.45 | 0.77 | 0.77 | 0.91 | 0.91 | 0.36 | 0.19 | 0.19 | 0.13 | 0.13 | 0.33 | 0.33 |
| | | std | 0.51 | 0.44 | 0.26 | 0.26 | 0.33 | 0.33 | 0.52 | 0.52 | 0.68 | 0.68 | 0.25 | 0.13 | 0.14 | 0.09 | 0.09 | 0.25 | 0.25 |
| | | median | 1.11 | 1.66 | 0.20 | 0.20 | 0.21 | 0.22 | 0.73 | 0.73 | 0.75 | 0.75 | 0.23 | 0.10 | 0.10 | 0.06 | 0.06 | 0.17 | 0.15 |
| | 800 | mean | 1.16 | 1.71 | 0.25 | 0.25 | 0.28 | 0.28 | 0.77 | 0.77 | 0.91 | 0.91 | 0.25 | 0.11 | 0.11 | 0.07 | 0.08 | 0.20 | 0.21 |
| | | std | 0.33 | 0.29 | 0.19 | 0.19 | 0.23 | 0.23 | 0.52 | 0.52 | 0.68 | 0.68 | 0.14 | 0.08 | 0.07 | 0.06 | 0.06 | 0.16 | 0.16 |
| 3 | | median | 0.63 | 0.40 | 0.37 | 0.37 | 0.41 | 0.40 | 0.40 | 0.40 | 0.36 | 0.36 | 0.54 | 0.46 | 0.44 | 0.22 | 0.23 | 0.34 | 0.31 |
| | 100 | mean | 0.69 | 0.48 | 0.49 | 0.49 | 0.52 | 0.52 | 0.49 | 0.49 | 0.51 | 0.51 | 0.69 | 0.62 | 0.55 | 0.29 | 0.29 | 0.35 | 0.36 |
| | | std | 0.44 | 0.37 | 0.39 | 0.39 | 0.43 | 0.43 | 0.38 | 0.38 | 0.43 | 0.43 | 0.56 | 0.88 | 0.52 | 0.25 | 0.27 | 0.28 | 0.30 |
| | | median | 0.52 | 0.28 | 0.32 | 0.32 | 0.33 | 0.33 | 0.40 | 0.40 | 0.36 | 0.36 | 0.33 | 0.29 | 0.28 | 0.15 | 0.15 | 0.21 | 0.21 |
| | 200 | mean | 0.56 | 0.33 | 0.36 | 0.36 | 0.39 | 0.39 | 0.49 | 0.49 | 0.51 | 0.51 | 0.37 | 0.34 | 0.34 | 0.18 | 0.18 | 0.23 | 0.22 |
| | | std | 0.30 | 0.24 | 0.28 | 0.28 | 0.29 | 0.29 | 0.38 | 0.38 | 0.43 | 0.43 | 0.28 | 0.28 | 0.27 | 0.14 | 0.14 | 0.16 | 0.16 |
| | | median | 0.58 | 0.18 | 0.20 | 0.20 | 0.24 | 0.24 | 0.40 | 0.40 | 0.36 | 0.36 | 0.20 | 0.16 | 0.17 | 0.08 | 0.08 | 0.11 | 0.11 |
| | 400 | mean | 0.58 | 0.20 | 0.24 | 0.24 | 0.25 | 0.25 | 0.49 | 0.49 | 0.51 | 0.51 | 0.24 | 0.19 | 0.19 | 0.10 | 0.10 | 0.13 | 0.13 |
| | | std | 0.26 | 0.14 | 0.17 | 0.17 | 0.17 | 0.17 | 0.38 | 0.38 | 0.43 | 0.43 | 0.17 | 0.13 | 0.13 | 0.08 | 0.08 | 0.10 | 0.10 |
| | | median | 0.56 | 0.13 | 0.13 | 0.13 | 0.14 | 0.14 | 0.40 | 0.40 | 0.36 | 0.36 | 0.11 | 0.09 | 0.09 | 0.05 | 0.05 | 0.07 | 0.07 |
| | 800 | mean | 0.57 | 0.14 | 0.16 | 0.16 | 0.17 | 0.17 | 0.49 | 0.49 | 0.51 | 0.51 | 0.13 | 0.11 | 0.11 | 0.06 | 0.06 | 0.08 | 0.08 |
| | | std | 0.19 | 0.11 | 0.11 | 0.11 | 0.13 | 0.13 | 0.38 | 0.38 | 0.43 | 0.43 | 0.10 | 0.08 | 0.08 | 0.05 | 0.05 | 0.06 | 0.06 |
| 4 | | median | 0.61 | 0.72 | 0.57 | 0.57 | 0.69 | 0.71 | 0.68 | 0.68 | 0.69 | 0.69 | 0.74 | 0.68 | 0.73 | 0.42 | 0.45 | 0.51 | 0.61 |
| | 100 | mean | 0.79 | 0.92 | 0.69 | 0.69 | 0.78 | 0.79 | 0.75 | 0.75 | 0.95 | 0.92 | 0.96 | 0.95 | 1.02 | 0.59 | 0.71 | 0.74 | 0.99 |
| | | std | 0.64 | 0.79 | 0.54 | 0.54 | 0.56 | 0.56 | 0.57 | 0.57 | 0.77 | 0.67 | 0.82 | 0.80 | 0.96 | 0.81 | 0.95 | 0.90 | 1.21 |
| | | median | 0.45 | 0.41 | 0.36 | 0.36 | 0.45 | 0.46 | 0.68 | 0.68 | 0.69 | 0.69 | 0.48 | 0.48 | 0.49 | 0.21 | 0.21 | 0.31 | 0.32 |
| | 200 | mean | 0.51 | 0.48 | 0.46 | 0.46 | 0.54 | 0.54 | 0.75 | 0.75 | 0.95 | 0.92 | 0.59 | 0.54 | 0.57 | 0.25 | 0.25 | 0.38 | 0.40 |
| | | std | 0.42 | 0.46 | 0.35 | 0.35 | 0.40 | 0.40 | 0.57 | 0.57 | 0.77 | 0.67 | 0.46 | 0.44 | 0.53 | 0.19 | 0.18 | 0.30 | 0.35 |
| | | median | 0.32 | 0.23 | 0.24 | 0.24 | 0.26 | 0.26 | 0.68 | 0.68 | 0.69 | 0.69 | 0.31 | 0.30 | 0.30 | 0.11 | 0.11 | 0.19 | 0.21 |
| | 400 | mean | 0.37 | 0.28 | 0.28 | 0.28 | 0.30 | 0.30 | 0.75 | 0.75 | 0.95 | 0.92 | 0.32 | 0.31 | 0.31 | 0.14 | 0.14 | 0.22 | 0.22 |
| | | std | 0.25 | 0.22 | 0.23 | 0.23 | 0.24 | 0.24 | 0.57 | 0.57 | 0.77 | 0.67 | 0.23 | 0.21 | 0.21 | 0.11 | 0.11 | 0.14 | 0.15 |
| | | median | 0.19 | 0.16 | 0.19 | 0.19 | 0.21 | 0.21 | 0.68 | 0.68 | 0.69 | 0.69 | 0.17 | 0.17 | 0.17 | 0.09 | 0.09 | 0.10 | 0.10 |
| | 800 | mean | 0.24 | 0.19 | 0.22 | 0.22 | 0.24 | 0.23 | 0.75 | 0.75 | 0.95 | 0.92 | 0.20 | 0.19 | 0.19 | 0.10 | 0.10 | 0.12 | 0.12 |
| | | std | 0.17 | 0.15 | 0.16 | 0.16 | 0.17 | 0.17 | 0.57 | 0.57 | 0.77 | 0.67 | 0.14 | 0.12 | 0.12 | 0.07 | 0.07 | 0.10 | 0.10 |

## Appendix C: Proofs

### C.1. Proof of Lemma 3.2

*Proof.* Let $L(y,t) = (y-t)^2$. We now prove that $L_{\widehat{W},\widehat{H}}$ is convex. The same argument can be used for $L_{W^0,H^0}$.

Recall that

$$H\left(X,\widehat{\beta},t\right) = \int_{y\in\mathcal{Y}} L(y,t)dF_{Y|X}\left(y \mid X,\widehat{\beta}\right).$$

We first show that for every convex loss $L$, $H\left(X,\widehat{\beta},t\right)$ is convex. For any $\alpha \in (0,1)$, by the convexity of $L(y,t)$ with respect to $t$,

$$\begin{aligned}
&H\left(X,\widehat{\beta},\alpha t+(1-\alpha)t'\right)\\
=&\int_{y\in\mathcal{Y}} L\{y,\alpha t+(1-\alpha)t'\}dF_{Y|X}\left(y \mid X,\widehat{\beta}\right)\\
\leq&\int_{y\in\mathcal{Y}} \{\alpha L(y,t)+(1-\alpha)L(y,t')\}\, dF_{Y|X}\left(y \mid X,\widehat{\beta}\right)\\
=&\alpha\int_{y\in\mathcal{Y}} L(y,t)dF_{Y|X}\left(y \mid X,\widehat{\beta}\right)+(1-\alpha)\int_{y\in\mathcal{Y}} L(y,t')dF_{Y|X}\left(y \mid X,\widehat{\beta}\right)\\
=&\alpha H\left(X,\widehat{\beta},t\right)+(1-\alpha)H\left(X,\widehat{\beta},t'\right),
\end{aligned}$$

which indicates that $H\left(X,\widehat{\beta},t\right)$ is a convex function with respect to $t$.

Recall that

$$\begin{aligned}
L_{\widehat{W},\widehat{H}} &= L_{W,H}\left(\widehat{\pi},\widehat{H},R,X,Y,f(X)\right)\\
&= \frac{RL(Y,f(X))}{\widehat{\pi}(X)} - \frac{R-\widehat{\pi}(X)}{\widehat{\pi}(X)}H\left(X,\widehat{\beta},f(X)\right).
\end{aligned}$$

Therefore, when $R=0$, $L_{W,H}\left(\widehat{\pi},\widehat{H},0,X,Y,t\right) = H\left(X,\widehat{\beta},t\right)$ which is a convex function for any loss $L$. When $R=1$ and $L$ is the quadratic loss,

$$H\left(X,\widehat{\beta},t\right) = \int_{y\in\mathcal{Y}} (y-t)^2\, dF_{Y|X}\left(y \mid X,\widehat{\beta}\right) = t^2 - 2tU\left(X,\widehat{\beta}\right) + V\left(X,\widehat{\beta}\right),$$

where $U(X,\widehat{\beta}) = \int_{y\in\mathcal{Y}} ydF_{Y|X}(y|X,\widehat{\beta})$ and $V(X,\widehat{\beta}) = \int_{y\in\mathcal{Y}} y^2dF_{Y|X}(y|X,\widehat{\beta})$. Note that $U$ and $V$ are not functions of $t$. Hence, for $R=1$,

$$\begin{aligned}
&L_{W,H}\left(\widehat{\pi},\widehat{H},R,X,Y,t\right)\\
=&t^2 - 2t\left\{\frac{Y}{\widehat{\pi}(X)} - \frac{1-\widehat{\pi}(X)}{\widehat{\pi}(X)}U\left(X,\widehat{\beta}\right)\right\} + \frac{Y^2 - \{1-\widehat{\pi}(X)\}\, V\left(X,\widehat{\beta}\right)}{\widehat{\pi}(X)}.
\end{aligned}$$

Since the second derivative with respect to $t$ is positive, $L_{W,H}\left(\widehat{\pi}, \widehat{H}, R, X, Y, t\right)$ is convex with respect to $t$. □

### *C.2. Proof of Lemma 3.3*

*Proof.* Case 1: The missing mechanism is correctly specified, that is $\widehat{\pi}(X) \xrightarrow{\mathrm{P}} \pi^*(X) = \pi^0(X)$, but $\widehat{\beta} \xrightarrow{\mathrm{P}} \beta^*$, where $\beta^*$ does not necessarily equal $\beta^0$.

$$\begin{aligned}
&L_{W,H}\left(\widehat{\pi}, \widehat{H}, R, X, Y, f(X)\right)\\
=&\frac{RL(Y, f(X))}{\pi^0(X)} - \frac{R-\pi^0(X)}{\pi^0(X)} H(X, \beta^*, f(X)) + \mathrm{o}_p(1)\\
=&L(Y, f(X)) + \frac{R-\pi^0(X)}{\pi^0(X)} \{L(Y, f(X)) - H(X, \beta^*, f(X))\} + \mathrm{o}_p(1).
\end{aligned}$$

By the Law of Large Numbers (LLN), we have

$$\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f) \xrightarrow{\mathrm{P}} \mathcal{R}_{L,\mathrm{P}} + \mathrm{E}\left[\frac{R-\pi^0(X)}{\pi^0(X)} \{L(Y, f(X)) - H(X, \beta^*, f(X))\}\right].$$

Note that

$$\begin{aligned}
&\mathrm{E}\left[\frac{R-\pi^0(X)}{\pi^0(X)} \{L(Y, f(X)) - H(X, \beta^*, f(X))\}\right]\\
=&\mathrm{E}\left(\mathrm{E}\left[\frac{R-\pi^0(X)}{\pi^0(X)} \{L(Y, f(X)) - H(X, \beta^*, f(X))\} \middle| X, Y\right]\right)\\
=&\mathrm{E}\left(\{L(Y, f(X)) - H(X, \beta^*, f(X))\} \mathrm{E}\left[\frac{R-\pi^0(X)}{\pi^0(X)} \middle| X, Y\right]\right)\\
=&\mathrm{E}\left(\{L(Y, f(X)) - H(X, \beta^*, f(X))\} \mathrm{E}\left[\frac{R-\pi^0(X)}{\pi^0(X)} \middle| X\right]\right) = 0.
\end{aligned}$$

The third equality holds because by Assumption 2.1, $R$ and $Y$ are independent given $X$. As a conclusion, we have, $\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f) \xrightarrow{\mathrm{P}} \mathcal{R}_{L,\mathrm{P}}(f)$.

Case 2: When $\widehat{\beta} \xrightarrow{\mathrm{P}} \beta^* = \beta^0$, but $\widehat{\pi}(X) \xrightarrow{\mathrm{P}} \pi^*(X)$ which is not necessarily $\pi^0(X)$,

$$\begin{aligned}
&L_{W,H}\left(\widehat{\pi}, \widehat{H}, R, X, Y, f(X)\right)\\
=&H\left(X, \beta^0, f(X)\right) + \frac{R\left\{L(Y, f(X)) - H\left(X, \beta^0, f(X)\right)\right\}}{\pi^*(X)} + \mathrm{o}_p(1).
\end{aligned}$$

Then

$$\begin{aligned}
\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f) \xrightarrow{\mathrm{P}} &\mathrm{E}\{H\left(X, \beta^0, f(X)\right)\}\\
&+ \mathrm{E}\left[\frac{R\left\{L(Y, f(X)) - H\left(X, \beta^0, f(X)\right)\right\}}{\pi^*(X)}\right].
\end{aligned}$$

The second expression can be shown equal to 0. Indeed,

$$\begin{aligned}
&\mathrm{E}\left[\frac{R\left\{L(Y,f(X))-H\left(X,\beta^0,f(X)\right)\right\}}{\pi^*(X)}\right]\\
=&\mathrm{E}\left(\mathrm{E}\left[\frac{R\left\{L(Y,f(X))-H\left(X,\beta^0,f(X)\right)\right\}}{\pi^*(X)}\middle| R,X\right]\right)\\
=&\mathrm{E}\left(\frac{R}{\pi^*(X)}\mathrm{E}\left[\left\{L(Y,f(X))-H\left(X,\beta^0,f(X)\right)\right\}\mid X\right]\right)\\
=&\mathrm{E}\left[\frac{R}{\pi^*(X)}\mathrm{E}\left\{L(Y,f(X))\mid X\right\}-H\left(X,\beta^0,f(X)\right)\right]=0,
\end{aligned}$$

where the last equation holds since by (3.6), $H\left(X,\beta^0,f(X)\right)$ is defined as $\mathrm{E}\left\{L(Y,f(X))\mid X\right\}$. Note that $\mathrm{E}\left\{H\left(X,\beta^0,f(X)\right)\right\}=\mathcal{R}_{L,\mathrm{P}}(f)$ and the result follows. □

### C.3. Oracle inequality for the weighted-complete-case kernel machines

**Theorem C.1.** *Let Assumptions 2.1, 2.2, and 4.1 hold. Then, for fixed $\lambda>0$, $n\geq 1$, $\varepsilon>0$, and $\eta>0$, with probability not less than $1-e^{-\eta}$,*

$$\begin{aligned}
&\lambda\left\|f_{D,\lambda}^{\widehat{W}}\right\|_{\mathcal{H}}^2+\mathcal{R}_{L,\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W}}\right)-\inf_{f\in\mathcal{H}}\mathcal{R}_{L,\mathrm{P}}(f)\\
<&A_2(\lambda)+d_{2n}(\lambda)\varepsilon\\
&+d_{3n}(\lambda)\left[\sqrt{\frac{2\eta+2\log\left\{2\mathcal{N}\left(B_{\mathcal{H}},\|\cdot\|_\infty,d_{1n}(\lambda)\varepsilon\right)\right\}}{n}}+\frac{\mathrm{Err}_{1,n}}{c_{n,L}}\right],
\end{aligned}$$

*where* $d_{1n}=(c_{n,L}\lambda)^{\frac{1}{2}}$, $d_{2n}=\frac{2C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)}{c}$, $d_{3n}=\frac{C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)(c_{n,L}\lambda)^{-\frac{1}{2}}+1}{c}$.

*Proof.* Recall that $L_W\left(\pi^0,R,X,Y,f(X)\right)\equiv\frac{RL(Y,f(X))}{\pi^0(X)}$, similarly, write $L_{W^0}$ for short. By the definition of $f_{D,\lambda}^{\widehat{W}}$,

$$\lambda\left\|f_{D,\lambda}^{\widehat{W}}\right\|_{\mathcal{H}}^2+\mathcal{R}_{L_{\widehat{W}},D}\left(f_{D,\lambda}^{\widehat{W}}\right)\leq\lambda\left\|f_{\mathrm{P},\lambda}\right\|_{\mathcal{H}}^2+\mathcal{R}_{L_{\widehat{W}},D}\left(f_{\mathrm{P},\lambda}\right). \tag{C.1}$$

Recall that

$$A_2(\lambda)=\lambda\|f_{\mathrm{P},\lambda}\|_{\mathcal{H}}^2+\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})-\mathcal{R}_{L,\mathrm{P}}^*.$$

Let

$$A^{\widehat{W}}(\lambda)=\lambda\left\|f_{D,\lambda}^{\widehat{W}}\right\|_{\mathcal{H}}^2+\mathcal{R}_{L,\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W}}\right)-\mathcal{R}_{L,\mathrm{P}}^*.$$

Hence,

$$\begin{aligned}
&A^{\widehat{W}}(\lambda)-A_2(\lambda)\\
=&\lambda\left\|f_{D,\lambda}^{\widehat{W}}\right\|_{\mathcal{H}}^2+\mathcal{R}_{L,\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W}}\right)-\lambda\|f_{\mathrm{P},\lambda}\|_{\mathcal{H}}^2-\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})
\end{aligned}$$

$$
\begin{aligned}
&+\mathcal{R}_{L_{\widehat{W}},D}\left(f_{D,\lambda}^{\widehat{W}}\right)-\mathcal{R}_{L_{\widehat{W}},D}\left(f_{D,\lambda}^{\widehat{W}}\right)\\
\leq&\lambda\|f_{\mathrm{P},\lambda}\|_{\mathcal{H}}^2+\mathcal{R}_{L_{\widehat{W}},D}(f_{\mathrm{P},\lambda})+\mathcal{R}_{L,\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W}}\right)-\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})\\
&-\mathcal{R}_{L_{\widehat{W}},D}\left(f_{D,\lambda}^{\widehat{W}}\right)-\lambda\|f_{\mathrm{P},\lambda}\|_{\mathcal{H}}^2\\
=&\mathcal{R}_{L_{\widehat{W}},D}(f_{\mathrm{P},\lambda})-\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})+\mathcal{R}_{L,\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W}}\right)-\mathcal{R}_{L_{\widehat{W}},D}\left(f_{D,\lambda}^{\widehat{W}}\right)\\
\equiv&B^{\widehat{W}}(\lambda),
\end{aligned}
$$

where the inequality follows from (C.1).

Note that by (3.2),

$$
\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(f)=\mathrm{E}\left\{L\left(Y,f(X)\right)\right\}&=\mathrm{E}\left[\mathrm{E}\left\{\frac{RL\left(Y,f(X)\right)}{\pi^0(X)}\middle|X,Y\right\}\right]\\
&=\mathrm{E}\left\{\frac{RL\left(Y,f(X)\right)}{\pi^0(X)}\right\}=\mathcal{R}_{L_{W^0},\mathrm{P}}(f),
\end{aligned}
$$

where the second equality holds for Assumption 2.1 and the third equality holds by conditional expectation. Hence,

$$
B^{\widehat{W}}(\lambda)=\mathcal{R}_{L_{\widehat{W}},D}(f_{\mathrm{P},\lambda})-\mathcal{R}_{L_{W^0},\mathrm{P}}(f_{\mathrm{P},\lambda})+\mathcal{R}_{L_{W^0},\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W}}\right)-\mathcal{R}_{L_{\widehat{W}},D}\left(f_{D,\lambda}^{\widehat{W}}\right).
$$

Therefore,

$$
\begin{aligned}
B^{\widehat{W}}(\lambda)=&\mathcal{R}_{L_{\widehat{W}},D}(f_{\mathrm{P},\lambda})-\mathcal{R}_{L_{W^0},D}(f_{\mathrm{P},\lambda})+\mathcal{R}_{L_{W^0},D}(f_{\mathrm{P},\lambda})-\mathcal{R}_{L_{W^0},\mathrm{P}}(f_{\mathrm{P},\lambda})\\
&+\mathcal{R}_{L_{W^0},\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W}}\right)-\mathcal{R}_{L_{W^0},D}\left(f_{D,\lambda}^{\widehat{W}}\right)\\
&+\mathcal{R}_{L_{W^0},D}\left(f_{D,\lambda}^{\widehat{W}}\right)-\mathcal{R}_{L_{\widehat{W}},D}\left(f_{D,\lambda}^{\widehat{W}}\right)\\
\leq&\left|\mathcal{R}_{L_{W^0},\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W}}\right)-\mathcal{R}_{L_{W^0},D}\left(f_{D,\lambda}^{\widehat{W}}\right)\right|\\
&+\left|\mathcal{R}_{L_{W^0},D}(f_{\mathrm{P},\lambda})-\mathcal{R}_{L_{W^0},\mathrm{P}}(f_{\mathrm{P},\lambda})\right|\\
&+\left|\mathcal{R}_{L_{\widehat{W}},D}(f_{\mathrm{P},\lambda})-\mathcal{R}_{L_{W^0},D}\left(f_{\mathrm{P},\lambda}\right)\right|\\
&+\left|\mathcal{R}_{L_{W^0},D}\left(f_{D,\lambda}^{\widehat{W}}\right)-\mathcal{R}_{L_{\widehat{W}},D}\left(f_{D,\lambda}^{\widehat{W}}\right)\right|\\
\equiv&A_n+B_n+C_n+D_n.
\end{aligned}
$$

We first bound expressions $A_n$ and $B_n$. Note that $L(y,0)\leq 1$ for all $y\in\mathcal{Y}$. By Assumption 4.1, $L_W\left(\widehat{\pi},R,X,Y,0\right)=\frac{RL(Y,0)}{\widehat{\pi}(X)}\leq\frac{1}{c_{n,L}}$. Thus,

$$
\lambda\left\|f_{D,\lambda}^{\widehat{W}}\right\|_{\mathcal{H}}^2\leq\mathcal{R}_{L_{\widehat{W}},D}(f_0)\leq\frac{1}{c_{n,L}},
$$

for $f_0(X)\equiv 0$ for all $X$.

By Assumption 2.2, for every $f \in (c_{n,L}\lambda)^{-\frac{1}{2}} B_{\mathcal{H}}$, where $B_{\mathcal{H}}$ is the unit ball of $\mathcal{H}$,

$$\begin{aligned} L_{W^0} &= L_W\left(\pi^0, R, X, Y, f(X)\right) \\ &\le \left|L_W\left(\pi^0, R, X, Y, f(X)\right) - L_W\left(\pi^0, R, X, Y, 0\right)\right| + \left|L_W\left(\pi^0, R, X, Y, 0\right)\right| \\ &\le \left|\frac{R\left\{L(Y, f(X)) - L(Y, 0)\right\}}{\pi^0(X)}\right| + \frac{1}{2c} \\ &\le \left|\frac{L(Y, f(X)) - L(Y, 0)}{\pi^0(X)}\right| + \frac{1}{2c} \\ &\le \frac{1}{2c}\left\{L(Y, f(X)) - L(Y, 0)\right\} + \frac{1}{2c} \\ &\le \frac{1}{2c}\left\{C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)(c_{n,L}\lambda)^{-\frac{1}{2}} + 1\right\} :\equiv Q_n; \end{aligned} \tag{C.2}$$

$c$ is defined in Assumption 2.2 and $C_L(\cdot)$ is a Lipschiz constant defined in Assumption 2.3.

Let $\mathcal{F}_\varepsilon$ be an $\varepsilon$-net with cardinality $|\mathcal{F}_\varepsilon| = \mathcal{N}\left((c_{n,L}\lambda)^{-\frac{1}{2}} B_{\mathcal{H}}, \|\cdot\|_\infty, \varepsilon\right) = \mathcal{N}\left(B_{\mathcal{H}}, \|\cdot\|_\infty, (c_{n,L}\lambda)^{\frac{1}{2}}\varepsilon\right)$. For every function $f \in (c_{n,L}\lambda)^{-\frac{1}{2}} B_{\mathcal{H}}$, there exists a function $g \in \mathcal{F}_\varepsilon$ such that $\|f - g\|_\infty \le \varepsilon$. Since,

$$\begin{aligned} \left|\mathcal{R}_{L_{W^0},\mathrm{P}}(f) - \mathcal{R}_{L_{W^0},\mathrm{P}}(g)\right| &= \left|\mathrm{E}\left\{\frac{L(Y, f(X)) - L(Y, g(X))}{\pi^0(X)}\right\}\right| \\ &\le \frac{1}{2c} C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)\varepsilon. \end{aligned}$$

This inequality also holds for $\left|\mathcal{R}_{L_{W^0},D}(f) - \mathcal{R}_{L_{W^0},D}(g)\right|$. Thus,

$$\begin{aligned} &\left|\mathcal{R}_{L_{W^0},\mathrm{P}}(f) - \mathcal{R}_{L_{W^0},D}(f)\right| \\ \le &\left|\mathcal{R}_{L_{W^0},\mathrm{P}}(f) - \mathcal{R}_{L_{W^0},\mathrm{P}}(g)\right| + \left|\mathcal{R}_{L_{W^0},D}(f) - \mathcal{R}_{L_{W^0},D}(g)\right| \\ &+ \left|\mathcal{R}_{L_{W^0},\mathrm{P}}(g) - \mathcal{R}_{L_{W^0},D}(g)\right| \\ \le &\frac{1}{c} C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)\varepsilon + \left|\mathcal{R}_{L_{W^0},\mathrm{P}}(g) - \mathcal{R}_{L_{W^0},D}(g)\right| \text{ for some } g \in \mathcal{F}_\varepsilon. \end{aligned}$$

Using Hoeffding's inequality (Steinwart and Christmann, 2008, Theorem 6.10), and similarly to the proof of Theorem 6.25 therein, for any $\eta > 0$, we have

$$\begin{aligned} &\mathrm{P}\left(A_n + B_n \ge Q_n\sqrt{\frac{2\eta}{n}} + \frac{2}{c} C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)\varepsilon\right) \\ \le &\mathrm{P}\left(2\sup_{g\in\mathcal{F}_\varepsilon}\left|\mathcal{R}_{L_{W^0},\mathrm{P}}(g) - \mathcal{R}_{L_{W^0},D}(g)\right| \ge Q_n\sqrt{\frac{2\eta}{n}}\right) \\ \le &\sum_{g\in\mathcal{F}_\varepsilon}\mathrm{P}\left(\left|\mathcal{R}_{L_{W^0},\mathrm{P}}(g) - \mathcal{R}_{L_{W^0},D}(g)\right| \ge Q_n\sqrt{\frac{\eta}{2n}}\right) \\ \le &2\mathcal{N}\left(B_{\mathcal{H}}, \|\cdot\|_\infty, (c_{n,L}\lambda)^{\frac{1}{2}}\varepsilon\right) e^{-\eta}. \end{aligned}$$

Elementary algebraic transformation shows that for fixed $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\eta > 0$, with probability not less than $1 - e^{-\eta}$,

$$\begin{aligned} &A_n + B_n \\ &\leq \frac{C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)(c_{n,L}\lambda)^{-\frac{1}{2}} + 1}{2c}\left[\sqrt{\frac{2\eta + 2\log\left\{2\mathcal{N}\left(B_{\mathcal{H}}, \|\cdot\|_\infty, (c_{n,L}\lambda)^{\frac{1}{2}}\varepsilon\right)\right\}}{n}}\right] \\ &\quad + \frac{2C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)\varepsilon}{c}. \end{aligned} \tag{C.3}$$

Next, we bound $C_n + D_n$.

$$\begin{aligned} \left|\mathcal{R}_{L_{W^0},D}(f) - \mathcal{R}_{L_{\widehat{W}},D}(f)\right| &= \left|\mathbb{P}_n\left\{\frac{RL(Y,f(X))}{\pi^0(X)} - \frac{RL(Y,f(X))}{\widehat{\pi}(X)}\right\}\right| \\ &= \left|\mathbb{P}_n\left[\frac{RL(Y,f(X))}{\pi^0(X)\widehat{\pi}(X)}\left\{\widehat{\pi}(X) - \pi^0(X)\right\}\right]\right| \\ &\leq \frac{C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)(c_{n,L}\lambda)^{-\frac{1}{2}} + 1}{2c \cdot c_{n,L}}\mathrm{Err}_{1,n}. \end{aligned}$$

Then

$$C_n + D_n \leq \frac{C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right)(c_{n,L}\lambda)^{-\frac{1}{2}} + 1}{c \cdot c_{n,L}}\mathrm{Err}_{1,n}. \tag{C.4}$$

By the definition of $A^{\widehat{W}}(\lambda)$ and $B^{\widehat{W}}(\lambda)$,

$$A^{\widehat{W}}(\lambda) - A_2(\lambda) \leq B^{\widehat{W}}(\lambda) \leq A_n + B_n + C_n + D_n,$$

and the result thus follows from (C.3) and (C.4). □

### *C.4. Proof of Theorem 4.1*

*Proof.* By Condition 4.2,

$$C_L\left((c_{n,L}\lambda)^{-\frac{1}{2}}\right) \leq m(c_{n,L})^{-\frac{q}{2}}\lambda^{-\frac{q}{2}}.$$

Then, together with Condition 4.1, we have

$$\begin{aligned} &\sqrt{\frac{2\log\left\{2\mathcal{N}\left(B_{\mathcal{H}}, \|\cdot\|_\infty, (c_{n,L}\lambda)^{\frac{1}{2}}\varepsilon\right)\right\}}{n}} \\ \leq &\sqrt{\frac{2\ln 2 + 2a\left\{(c_{n,L}\lambda)^{\frac{1}{2}}\varepsilon\right\}^{-2p}}{n}} \leq \sqrt{\frac{2a}{n}}\left\{(c_{n,L}\lambda)^{\frac{1}{2}}\varepsilon\right\}^{-p}. \end{aligned}$$

Therefore, combined with the oracle inequality in Theorem C.1, for fixed $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\eta > 0$, with probability not less than $1 - e^{-\eta}$,

$$\begin{aligned} & A^{\widehat{W}}(\lambda) - A_2(\lambda) \\ \leq & \frac{m(c_{n,L})^{-\frac{q}{2}} \lambda^{-\frac{q}{2}} (c_{n,L}\lambda)^{-\frac{1}{2}} + 1}{c} \\ & \times \left[ 2(c_{n,L}\lambda)^{\frac{1}{2}} \varepsilon + \sqrt{\frac{2a}{n}} \left\{ (c_{n,L}\lambda)^{\frac{1}{2}} \varepsilon \right\}^{-p} + \left( \frac{2\eta}{n} \right)^{\frac{1}{2}} + \frac{\mathrm{Err}_{1,n}}{c_{n,L}} \right]. \end{aligned}$$

Let

$$\varepsilon = (c_{n,L}\lambda)^{-\frac{1}{2}} \left( \frac{p}{2} \right)^{\frac{1}{p+1}} \left( \frac{2a}{n} \right)^{\frac{1}{2p+2}}.$$

Then using some algebra,

$$2(c_{n,L}\lambda)^{\frac{1}{2}} \varepsilon + \sqrt{\frac{2a}{n}} \left\{ (c_{n,L}\lambda)^{\frac{1}{2}} \varepsilon \right\}^{-p} = (p+1) \left( \frac{2}{p} \right)^{\frac{p}{p+1}} \left( \frac{2a}{n} \right)^{\frac{1}{2p+2}} \leq 3 \left( \frac{2a}{n} \right)^{\frac{1}{2p+2}},$$

where the last inequality can be verified by Steinwart and Christmann (2008, Lemma A.1.5).

Since $\left| \widehat{\pi}(X) - \pi^0(X) \right| = \mathrm{O}_p\left( n^{-\frac{1}{2}} \right)$, we have $\mathrm{Err}_{1,n} = \mathrm{O}_p\left( n^{-\frac{1}{2}} \right)$, and there exists a constant $b_1(\eta)$ such that for all $n \geq 1$

$$\mathrm{P}\left( \mathrm{Err}_{1,n} \geq b_1(\eta) n^{-\frac{1}{2}} \right) < e^{-\eta}.$$

For fixed $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\eta > 0$, with probability not less than $1 - 2e^{-\eta}$,

$$\begin{aligned} & A^{\widehat{W}}(\lambda) - A_2(\lambda) \\ \leq & \frac{m(c_{n,L})^{-\frac{q+1}{2}} \lambda^{-\frac{q+1}{2}} + 1}{c} \left[ 3 \left( \frac{2a}{n} \right)^{\frac{1}{2p+2}} + \left( \frac{2\eta}{n} \right)^{\frac{1}{2}} + \frac{b_1(\eta) n^{-\frac{1}{2}}}{c_{n,L}} \right]. \end{aligned} \tag{C.5}$$

Note that by Assumption 4.1, $c_{n,L} = \mathrm{O}\left( n^{-d} \right)$,

$$m(c_{n,L})^{-\frac{q+1}{2}} = \mathrm{O}\left( n^{\frac{(q+1)d}{2}} \right). \tag{C.6}$$

Recall that

$$A^{\widehat{W}}(\lambda) = \lambda \left\| f_{D,\lambda}^{\widehat{W}} \right\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{P}}\left( f_{D,\lambda}^{\widehat{W}} \right) - \mathcal{R}_{L,\mathrm{P}}^*.$$

Taking $\lambda = \lambda_n$, then $\lambda_n \left\| f_{D,\lambda}^{\widehat{W}} \right\|_{\mathcal{H}}^2 \xrightarrow[n\to\infty]{} 0$, and by Steinwart and Christmann (2008, Lemma 5.15) $A_2(\lambda_n) \xrightarrow[n\to\infty]{} 0$. When $\lambda_n^{\frac{q+1}{2}} n^{\min\left(\frac{1}{2}-d, \frac{1}{2p+2}\right) - \frac{(q+1)d}{2}} \longrightarrow \infty$, the right-hand side of (C.5) converges to zero. Therefore, for all $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathrm{P}\left( \mathcal{R}_{L,\mathrm{P}}\left( f_{D,\lambda}^{\widehat{W}} \right) \leq \mathcal{R}_{L,\mathrm{P}}^* + \varepsilon \right) = 1.$$

Since it is true for all $\mathrm{P} \in \mathcal{P}$, the $\mathcal{P}$-universal consistency holds. □

### *C.5. Proof of Corollary 4.1*

*Proof.* By (C.5) and Assumption 4.2, for fixed $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\eta > 0$, with probability not less than $1 - 2e^{-\eta}$

$$\begin{aligned}
&A^{\widehat{W}}(\lambda) \\
\leq &A_2(\lambda) + \leq \frac{m(c_{n,L})^{-\frac{q+1}{2}}\lambda^{-\frac{q+1}{2}} + 1}{c}\left[3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}} + \left(\frac{2\eta}{n}\right)^{\frac{1}{2}} + \frac{b_1(\eta)n^{-\frac{1}{2}}}{c_{n,L}}\right] \\
\leq &b\lambda^{\gamma} + \left(\frac{m(c_{n,L})^{-\frac{q+1}{2}}}{c}\lambda^{-\frac{q+1}{2}} + \frac{1}{c}\right)\left[3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}} + \left(\frac{2\eta}{n}\right)^{\frac{1}{2}} + \frac{b_1(\eta)n^{-\frac{1}{2}}}{c_{n,L}}\right].
\end{aligned}$$

Let

$$\begin{aligned}
&G_1(\lambda) \\
= &b\lambda^{\gamma} + \left(\frac{m(c_{n,L})^{-\frac{q+1}{2}}}{c}\lambda^{-\frac{q+1}{2}} + \frac{1}{c}\right)\left[3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}} + \left(\frac{2\eta}{n}\right)^{\frac{1}{2}} + \frac{b_1(\eta)n^{-\frac{1}{2}}}{c_{n,L}}\right].
\end{aligned}$$

Taking the derivative with respect to $\lambda$ and setting it equal to 0,

$$b\gamma\lambda^{\gamma-1} = \frac{m(c_{n,L})^{-\frac{q+1}{2}}}{c}\frac{q+1}{2}\lambda^{-\frac{q+3}{2}}\left[3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}} + \left(\frac{2\eta}{n}\right)^{\frac{1}{2}} + \frac{b_1(\eta)n^{-\frac{1}{2}}}{c_{n,L}}\right].$$

By (C.6) and Assumption 4.1

$$\begin{aligned}
\lambda^{\gamma+\frac{q+1}{2}} &\propto \left(\frac{1}{n}\right)^{\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)} n^{\frac{(q+1)d}{2}}, \\
\Rightarrow \lambda &\propto n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2}{2\gamma+q+1}}.
\end{aligned}$$

Note that by choosing large $m$ where $m$ is defined in Condition 4.2, $G_1''\left(n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2}{2\gamma+q+1}}\right)$ can be positive. Then, for

$$\begin{aligned}
&\lambda_n = n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2}{2\gamma+q+1}}, \\
&G_1(\lambda_n) \\
= &bn^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2\gamma}{2\gamma+q+1}} \\
&+ \left\{\frac{m(c_{n,L})^{-\frac{q+1}{2}}}{c}n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2}{2\gamma+q+1}\left(-\frac{q+1}{2}\right)} + \frac{1}{c}\right\} \\
&\times \left[3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}} + \left(\frac{2\eta}{n}\right)^{\frac{1}{2}} + \frac{b_1(\eta)n^{-\frac{1}{2}}}{c_{n,L}}\right] \\
\leq &bn^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2\gamma}{2\gamma+q+1}}
\end{aligned}$$

$$
\begin{aligned}
&+ n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2}{2\gamma+q+1}\left(-\frac{q+1}{2}\right)-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}} \\
&\times c_{\mathrm{P}}\left(c_a+\sqrt{\eta}+b_1(\eta)\right) \\
\leq & Q\left(\sqrt{\eta}+b_1(\eta)+c_{a,b}\right) n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2\gamma}{2\gamma+q+1}},
\end{aligned}
$$

where $c_a$ is a constant related to $a$, $c_{a,b}$ is a constant related to $a$, $b$, and $c$, $c_{\mathrm{P}}$ and $Q$ are constants related to P. None of them is related to $\eta$.

Consequently, for fixed $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\eta > 0$, with probability not less than $1 - 2e^{-\eta}$,

$$
A^{\widehat{W}}(\lambda) \leq G_1(\lambda) \leq Q\left(\sqrt{\eta}+b_1(\eta)+c_{a,b}\right) n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2\gamma}{2\gamma+q+1}}.
$$

Therefore, the learning rate is $n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+\frac{(q+1)d}{2}\right\}\frac{2\gamma}{2\gamma+q+1}}$. □

### C.6. Oracle inequality for the doubly-robust kernel machines

**Theorem C.2.** *Let Assumptions 2.1, 2.2, and 4.1 hold. When $L$ is the quadratic loss, for fixed $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\eta > 0$, with probability not less than $1 - e^{-\eta}$*

$$
\begin{aligned}
&\lambda \left\| f_{D,\lambda}^{\widehat{W},\widehat{H}} \right\|_{\mathcal{H}}^2 + \mathcal{R}_{L,\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W},\widehat{H}}\right) - \inf_{f\in\mathcal{H}} \mathcal{R}_{L,\mathrm{P}}(f) \\
<& A_2(\lambda) + u_{2n}(\lambda)\varepsilon + 3u_{3n}(\lambda)\left[\sqrt{\frac{2\eta + 2\log\left\{2\mathcal{N}(B_{\mathcal{H}}, \|\cdot\|_\infty, u_{1n}(\lambda)\varepsilon)\right\}}{n}}\right] \\
&+ \frac{2u_{3n}(\lambda)}{c_{n,L}}\mathrm{Err}_{1,n} + 2\left(\frac{1}{c_{n,L}}+1\right)\mathrm{Err}_{2,n},
\end{aligned}
$$

*where* $u_{1n} = (c_{2,n}\lambda)^{\frac{1}{2}}$, $u_{2n} = \frac{6r(c_{2,n}\lambda)^{-\frac{1}{2}}}{c}$, $u_{3n} = \frac{r(c_{2,n}\lambda)^{-1}+1}{c}$, $c_{2,n} = \frac{c_{n,L}}{2+c_{n,U}}$, *$c_{n,L}$ and $c_{n,U}$ are defined as in Assumption 4.1.*

*Proof.* Let $c_1 = \frac{3}{2c}$, where $c$ is defined as in Assumption 2.2. Since $L(Y,0) \leq 1$, we also have $H\left(X,\beta^0,0\right) = \mathrm{E}\left\{L(Y,0) \mid X\right\} \leq 1$. Recall that

$$
\begin{aligned}
L_{W^0,H^0} &= L_{W,H}\left(\pi^0, H^0, R, X, Y, f(X)\right) \\
&\equiv \frac{RL(Y,f(X))}{\pi^0(X)} - \frac{R-\pi^0(X)}{\pi^0(X)} H\left(X,\beta^0,0\right).
\end{aligned}
$$

We have,

$$
\begin{aligned}
\left|L_{W,H}\left(\pi^0,H^0,R,X,Y,0\right)\right| &= \left|\frac{ML(Y,0)}{\pi^0(X)} - \frac{R-\pi^0(X)}{\pi^0(X)} H(X,\beta^0,0)\right| \\
&\leq \frac{RL(Y,0)}{\pi^0(X)} + \frac{R+\pi^0(X)}{\pi^0(X)} H(X,\beta^0,0) \\
&\leq \frac{2R+\pi^0(X)}{\pi^0(X)} \leq \frac{3}{2c} \equiv c_1,
\end{aligned}
$$

where the last inequality follows from Assumption 2.2.

Since $H\left(X,\widehat{\beta},0\right)=\int_{y\in\mathcal{Y}}L(y,0)dF_{Y|X}(y\mid X,\widehat{\beta})\leq 1$,

$$\begin{aligned}\left|L_{W,H}\left(\widehat{\pi},\widehat{H},R,X,Y,0\right)\right| &= \left|\frac{RL(Y,0)}{\widehat{\pi}(X)}-\frac{R-\widehat{\pi}(X)}{\widehat{\pi}(X)}H\left(X,\widehat{\beta},0\right)\right| \\ &\leq \frac{RL(Y,0)}{\widehat{\pi}(X)}+\frac{R+\widehat{\pi}(X)}{\widehat{\pi}(X)}H\left(X,\widehat{\beta},0\right) \\ &\leq \frac{2R+\widehat{\pi}(X)}{\widehat{\pi}(X)}\leq\frac{2+c_{n,U}}{c_{n,L}}\equiv\frac{1}{c_{2,n}}.\end{aligned}$$

Note that

$$\lambda\left\|f_{D,\lambda}^{\widehat{W},\widehat{H}}\right\|_{\mathcal{H}}^{2}\leq\lambda\left\|f_{D,\lambda}^{\widehat{W},\widehat{H}}\right\|_{\mathcal{H}}^{2}+\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}\left(f_{D,\lambda}^{\widehat{W},\widehat{H}}\right)\leq\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(0)\leq\frac{1}{c_{2,n}}. \quad \text{(C.7)}$$

For every $f\in(c_{2,n}\lambda)^{-\frac{1}{2}}B_{\mathcal{H}}$,

$$\begin{aligned}|L(Y,f(X))| &\leq |L(Y,f(X))-L(Y,0)|+L(Y,0) \\ &\leq C_L\left((c_{2,n}\lambda)^{-\frac{1}{2}}\right)(c_{2,n}\lambda)^{-\frac{1}{2}}+1\leq m(c_{2,n}\lambda)^{-1}+1. \quad \text{(C.8)}\end{aligned}$$

The last inequality holds since by Condition 4.2, for the quadratic loss $C_L\left((c_{2,n}\lambda)^{-\frac{1}{2}}\right)\leq m(c_{2,n}\lambda)^{-\frac{1}{2}}$.

Using the argument as in (C.2) in Theorem C.1,

$$L_{W,H}\left(\pi^0,H^0,R,X,Y,f(X)\right)\leq\frac{3\left\{m(c_{2,n}\lambda)^{-1}+1\right\}}{2c}\equiv T_n.$$

Let

$$A^{\widehat{W},\widehat{H}}(\lambda)=\lambda\left\|f_{D,\lambda}^{\widehat{W},\widehat{H}}\right\|_{\mathcal{H}}^{2}+\mathcal{R}_{L,\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W},\widehat{H}}\right)-\mathcal{R}_{L,\mathrm{P}}^{*}.$$

Since $\mathcal{R}_{L,\mathrm{P}}(f)=\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(f)$, using the same technique as in the proof of Theorem C.1,

$$\begin{aligned}&A^{\widehat{W},\widehat{H}}(\lambda)-A_2(\lambda) \\ \leq&\left|\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(f_{\mathrm{P},\lambda})-\mathcal{R}_{L_{W^0,H^0},D}(f_{\mathrm{P},\lambda})\right| \\ &+\left|\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}\left(f_{D,\lambda}^{\widehat{W},\widehat{H}}\right)-\mathcal{R}_{L_{W^0,H^0},D}\left(f_{D,\lambda}^{\widehat{W},\widehat{H}}\right)\right| \\ &+\left|\mathcal{R}_{L_{W^0,H^0},D}(f_{\mathrm{P},\lambda})-\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f_{\mathrm{P},\lambda})\right| \\ &+\left|\mathcal{R}_{L_{W^0,H^0},D}\left(f_{D,\lambda}^{\widehat{W},\widehat{H}}\right)-\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}\left(f_{D,\lambda}^{\widehat{W},\widehat{H}}\right)\right| \\ \equiv&A_n+B_n+C_n+D_n. \quad \text{(C.9)}\end{aligned}$$

Let $\mathcal{F}_\varepsilon$ be an $\varepsilon$-net of $(c_{2,n}\lambda)^{-\frac{1}{2}}B_{\mathcal{H}}$ with cardinality $|\mathcal{F}_\varepsilon|=\mathcal{N}\left((c_{2,n}\lambda)^{-\frac{1}{2}}B_{\mathcal{H}},\|\cdot\|_\infty,\varepsilon\right)=\mathcal{N}\left(B_{\mathcal{H}},\|\cdot\|_\infty,(c_{2,n}\lambda)^{\frac{1}{2}}\varepsilon\right)$. For every function $f\in(c_{2,n}\lambda)^{-\frac{1}{2}}B_{\mathcal{H}}$, there exists a function $g\in\mathcal{F}_\varepsilon$, such that $\|f-g\|_\infty\leq\varepsilon$. Thus,

$$|L(Y,f(X))-L(Y,g(X))|\leq C_L\left((c_{2,n}\lambda)^{-\frac{1}{2}}\right)\|f-g\|_\infty\leq m(c_{2,n}\lambda)^{-\frac{1}{2}}\varepsilon,$$

and

$$\begin{aligned}
&\left|H\left(X,\beta^0,f(X)\right)-H\left(X,\beta^0,g(X)\right)\right| \\
=&\left|\int_{y\in\mathcal{Y}}\{L(y,f(X))-L(y,g(X))\}dF_{Y|X}\left(y\mid X,\beta^0\right)\right| \\
\leq&\int_{y\in\mathcal{Y}}\left|L(y,f(X))-L(y,g(X))\right|dF_{Y|X}\left(y\mid X,\beta^0\right) \\
\leq&m(c_{2,n}\lambda)^{-\frac{1}{2}}\varepsilon.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\left|\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(f)-\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(g)\right| \\
=&\left|\mathrm{E}\left[\frac{R\{L(Y,f(X))-L(Y,g(X))\}}{\pi^0(X)}\right.\right. \\
&\left.\left.-\frac{\{R-\pi^0(X)\}\{H\left(X,\beta^0,f(X)\right)-H\left(X,\beta^0,g(X)\right)\}}{\pi^0(X)}\right]\right| \\
\leq&\mathrm{E}\left[\left|\frac{R\{L(Y,f(X))-L(Y,g(X))\}}{\pi^0(X)}\right.\right. \\
&\left.\left.-\frac{\{R-\pi(X)\}\{H\left(X,\beta^0,f(X)\right)-H\left(X,\beta^0,g(X)\right)\}}{\pi^0(X)}\right|\right] \\
\leq&\mathrm{E}\left[\frac{R\left|L(Y,f(X))-L(Y,g(X))\right|}{\pi^0(X)}\right. \\
&\left.+\frac{\{R+\pi^0(X)\}\left|H\left(X,\beta^0,f(X)\right)-H\left(X,\beta^0,g(X)\right)\right|}{\pi^0(X)}\right] \\
\leq&\frac{3m(c_{2,n}\lambda)^{-\frac{1}{2}}\varepsilon}{2c}. \qquad \text{(C.10)}
\end{aligned}$$

Similarly,

$$\left|\mathcal{R}_{L_{W^0,H^0},D}(f)-\mathcal{R}_{L_{W^0,H^0},D}(g)\right|\leq\frac{3m(c_{2,n}\lambda)^{-\frac{1}{2}}\varepsilon}{2c}.$$

Using (C.10) we can bound $A_n$ and $B_n$ of (C.9)

$$\begin{aligned}
&\left|\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(f)-\mathcal{R}_{L_{W^0,H^0},D}(f)\right| \\
\leq&\left|\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(f)-\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(g)\right|+\left|\mathcal{R}_{L_{W^0,H^0},D}(f)-\mathcal{R}_{L_{W^0,H^0},D}(g)\right| \\
&+\left|\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(g)-\mathcal{R}_{L_{W^0,H^0},D}(g)\right| \\
\leq&\frac{3m(c_{2,n}\lambda)^{-\frac{1}{2}}\varepsilon}{c}+\left|\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(g)-\mathcal{R}_{L_{W^0,H^0},D}(g)\right|.
\end{aligned}$$

Using the similar argument as Steinwart and Christmann (2008, Theorem 6.25) for any $\eta > 0$, we have

$$\begin{aligned}
&\mathrm{P}\left(A_n + B_n \geq T_n\sqrt{\frac{2\eta}{n}} + \frac{6m(c_{2,n}\lambda)^{-\frac{1}{2}}\varepsilon}{c}\right)\\
\leq&\mathrm{P}\left(2\sup_{g\in\mathcal{F}_\varepsilon}\left|\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(g) - \mathcal{R}_{L_{W^0,H^0},D}(g)\right| \geq T_n\sqrt{\frac{2\eta}{n}}\right)\\
\leq&\sum_{g\in\mathcal{F}_\varepsilon}\mathrm{P}\left(\left|\mathcal{R}_{L_{W^0,H^0},\mathrm{P}}(g) - \mathcal{R}_{L_{W^0,H^0},D}(g)\right| \geq T_n\sqrt{\frac{\eta}{2n}}\right)\\
\leq&2\mathcal{N}\left(B_{\mathcal{H}}, \|\cdot\|_\infty, (c_{2,n}\lambda)^{\frac{1}{2}}\varepsilon\right)e^{-\eta},
\end{aligned}$$

where the last inequality is from Hoeffding's inequality (Steinwart and Christmann, 2008, Theorem 6.10).

Elementary algebraic transformation shows that for fixed $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\eta > 0$, with probability not less than $1 - e^{-\eta}$,

$$\begin{aligned}
&A_n + B_n\\
\leq&\frac{3\left\{m(c_{2,n}\lambda)^{-1} + 1\right\}}{c}\left[\sqrt{\frac{2\eta + 2\log\left\{2\mathcal{N}\left(B_{\mathcal{H}}, \|\cdot\|_\infty, (c_{2,n}\lambda)^{\frac{1}{2}}\varepsilon\right)\right\}}{n}}\right]\\
&+\frac{6m(c_{2,n}\lambda)^{-\frac{1}{2}}\varepsilon}{c}.
\end{aligned} \tag{C.11}$$

Next bound $C_n$ and $D_n$,

$$\begin{aligned}
&\left|\mathcal{R}_{L_{W^0,H^0},D}(f) - \mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f)\right|\\
=&\mathbb{P}_n\left|\frac{RL(Y,f(X))}{\pi^0(X)} - \frac{RL(Y,f(X))}{\widehat{\pi}(X)} + \frac{R-\widehat{\pi}(X)}{\widehat{\pi}(X)}H\left(X,\widehat{\beta},f(X)\right)\right.\\
&\left.-\frac{R-\pi^0(X)}{\pi^0(X)}H\left(X,\beta^0,f(X)\right)\right|\\
\leq&\mathbb{P}_n\left[\frac{RL(Y,f(X))}{\pi^0(X)\widehat{\pi}(X)}\left|\pi^0(X) - \widehat{\pi}(X)\right| + \left|H\left(X,\widehat{\beta},f(X)\right) - H\left(X,\beta^0,f(X)\right)\right|\right.\\
&\left.+\left|\frac{RH\left(X,\widehat{\beta},f(X)\right)}{\widehat{\pi}(X)} - \frac{RH\left(X,\beta^0,f(X)\right)}{\pi^0(X)}\right|\right].
\end{aligned} \tag{C.12}$$

Then,

$$\left|\frac{RH\left(X,\widehat{\beta},f(X)\right)}{\widehat{\pi}(X)} - \frac{RH\left(X,\beta^0,f(X)\right)}{\pi^0(X)}\right|$$

$$
\begin{aligned}
&= \left| \frac{RH\left(X, \widehat{\beta}, f(X)\right)}{\widehat{\pi}(X)} - \frac{RH\left(X, \beta^0, f(X)\right)}{\widehat{\pi}(X)} \right. \\
&\quad \left. + \frac{RH\left(X, \beta^0, f(X)\right)}{\widehat{\pi}(X)} - \frac{RH\left(X, \beta^0, f(X)\right)}{\pi^0(X)} \right| \\
&\leq \frac{R}{\widehat{\pi}(X)} \left| H\left(X, \widehat{\beta}, f(X)\right) - H\left(X, \beta^0, f(X)\right) \right| \\
&\quad + \frac{RH\left(X, \beta^0, f(X)\right)}{\pi^0(X)\widehat{\pi}(X)} \left| \pi^0(X) - \widehat{\pi}(X) \right|. \qquad \text{(C.13)}
\end{aligned}
$$

Hence, by inequality (C.12) and (C.13), and definition of $\mathrm{Err}_{1,n}$ and $\mathrm{Err}_{2,n}$ in Subsection 4.1,

$$
\begin{aligned}
&\left| \mathcal{R}_{L_{W^0,H^0},D}(f) - \mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f) \right| \\
\leq &\frac{L(Y, f(X))}{2c \cdot c_{n,L}} \mathrm{Err}_{1,n} + \mathrm{Err}_{2,n} + \frac{1}{c_{n,L}} \mathrm{Err}_{2,n} + \frac{H\left(X, \beta^0, f(X)\right)}{2c \cdot c_{n,L}} \mathrm{Err}_{1,n}.
\end{aligned}
$$

Similarly to inequality (C.8), $\left| H\left(X, \beta^0, f(X)\right) \right| \leq m(c_{2,n}\lambda)^{-1} + 1$. Then we have

$$
\left| \mathcal{R}_{L_{W^0,H^0},D}(f) - \mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f) \right| \leq \frac{m(c_{2,n}\lambda)^{-1} + 1}{c \cdot c_{n,L}} \mathrm{Err}_{1,n} + \left( \frac{1}{c_{n,L}} + 1 \right) \mathrm{Err}_{2,n}. \qquad \text{(C.14)}
$$

By (C.11) and (C.14), and using

$$
A^{\widehat{W},\widehat{H}}(\lambda) - A_2(\lambda) \leq A_n + B_n + C_n + D_n,
$$

the result follows. □

### C.7. *Proof of Lemma 4.1*

*Proof.* Define $X_i(f) = L(X_i, Y_i, f(X_i)) - H\left(X_i, \beta^0, f(X_i)\right)$ and let

$$
\tilde{h}_n(f) = \frac{1}{n} \sum_{i=1}^{n} L(X_i, Y_i, f(X_i)) - H\left(X_i, \beta^0, f(X_i)\right) = \frac{1}{n} \sum_{i=1}^{n} X_i(f).
$$

By inequality (C.7) we have $\|f\|_{\mathcal{H}_n} \leq (c_{2,n}\lambda)^{-\frac{1}{2}}$, where $c_{2,n} = \frac{c_{n,L}}{2 + c_{n,U}}$. By the Cauchy-Schwarz inequality and the reproducing property (see Steinwart and Christmann (2008, Lemma 4.23)),

$$
|f(x)| = \langle f, k(\cdot, x) \rangle \leq \|f(x)\|_{\mathcal{H}_n} \sqrt{k(x, x)} \leq \|f(x)\|_{\mathcal{H}_n} \|k\|_\infty.
$$

Consequently, $\|f\|_\infty \leq \|f\|_{\mathcal{H}_n} \|k\|_\infty \leq \|f\|_{\mathcal{H}_n}$, where the inequality follows since we assume that $\|k\|_\infty \leq 1$.

Since $\|f\|_\infty \leq \|f\|_{\mathcal{H}_n}$, the space $\mathcal{H}_n$ over which the supremum $h_n$ is taken is contained in $(c_{2,n}\lambda)^{-\frac{q}{2}} B_{\mathcal{H}}$.

By (C.8), $\|X_i(f)\|_\infty \leq 2\left\{m\left(c_{2,n}\lambda\right)^{-q}+1\right\}$. Using the functional Hoeffding's inequality (Berestycki et al., 2009, Section 6.5),

$$\mathrm{P}\left[\frac{1}{\sqrt{2\left\{m\left(c_{2,n}\lambda\right)^{-1}+1\right\}}}\left\|\sum_{i=1}^{n} X_i(f)\right\|_\infty \geq C\right] \leq \frac{1}{K_u}\exp\left(-\frac{C^2}{K_u n}\right),$$

where $K_u$ is a universal constant and $C$ is any constant.

Let $\widetilde{C} = \frac{C}{\sqrt{n}}$, so $C = \sqrt{n}\widetilde{C}$. Then,

$$\mathrm{P}\left[\frac{\sqrt{n}}{\sqrt{2\left\{m\left(c_{2,n}\lambda\right)^{-1}+1\right\}}}\left\|\frac{1}{n}\sum_{i=1}^{n} X_i(f)\right\|_\infty \geq \widetilde{C}\right] \leq \frac{1}{K_u}\exp\left(-\text{const }\widetilde{C}^2\right). \tag{C.15}$$

Since $c_{2,n} = \frac{c_{n,L}}{2+c_{n,U}}$ for $0 < c_{n,U} < 1$, and $\frac{1}{c_{n,L}} = \mathrm{O}\left(n^d\right)$, then $\frac{1}{c_{2,n}} = \mathrm{O}\left(n^d\right)$. Thus,

$$m\left(c_{2,n}\lambda\right)^{-q} = \mathrm{O}\left(n^{qd}\lambda^{-q}\right).$$

Consequently,

$$\frac{\sqrt{n}}{\sqrt{2\left\{m\left(c_{2,n}\lambda\right)^{-1}+1\right\}}} = \mathrm{O}\left(n^{\frac{1}{2}-\frac{qd}{2}}\lambda^{\frac{q}{2}}\right). \tag{C.16}$$

We have, from (C.15) that $h_n = \mathrm{O}_p\left(n^{-\left(\frac{1}{2}-\frac{qd}{2}\right)}\lambda^{-\frac{q}{2}}\right)$. □

### *C.8. Proof of Lemma 4.2*

*Proof.* Note that for every $f$,

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|H\left(x,\widehat{\beta},f(x)\right)-H\left(x,\beta^0,f(x)\right)\right| \\
&=\sup_{x\in\mathcal{X}}\left|\int_{y\in\mathcal{Y}} L(y,f(x))dF_{Y|X}\left(y\mid x,\widehat{\beta}\right)-\int_{y\in\mathcal{Y}} L(y,f(x))dF_{Y|X}\left(y\mid x,\beta^0\right)\right| \\
&=\sup_{x\in\mathcal{X}}\left|\int_{y\in\mathcal{Y}} L(y,f(x))d\left\{F_{Y|X}\left(y\mid x,\widehat{\beta}\right)-F_{Y|X}\left(y\mid x,\beta^0\right)\right\}\right|.
\end{aligned}$$

By (C.8)

$$|L(y,f(x))| \leq m(c_{2,n}\lambda)^{-q}+1 = \mathrm{O}\left(n^{qd}\lambda^{-q}\right).$$

We have

$$\mathrm{Err}_{2,n} = \sup_{f\in\mathcal{H}_n}\sup_{x\in\mathcal{X}}\left|H\left(x,\widehat{\beta},f(x)\right)-H\left(x,\beta^0,f(x)\right)\right|$$

$$\leq \left\{m(c_{2,n}\lambda)^{-q}+1\right\} \sup_{x\in\mathcal{X}} \left| \int_{y\in\mathcal{Y}} d\left\{F_{Y|X}\left(y \mid x, \widehat{\beta}\right) - F_{Y|X}\left(y \mid x, \beta^0\right)\right\}\right|.$$

Define the function $\phi : \mathcal{B} \mapsto \mathcal{L}_\infty(X)$ by $\phi(\beta) = \int_y dF_{Y|X}(y \mid \cdot, \beta)$. Note that $\phi$ is Hadamard differentiable as a composite of $\beta \mapsto F_{Y|X}(\cdot \mid \cdot, \beta) \mapsto \int_y dF_{Y|X}(y \mid \cdot, \beta)$. The first mapping is Hadamard differentiable by the assumption of continuous differentiability with respect to $\beta$ and the definition of Hadamard differentiability (Kosorok, 2008, Section 2.2.4), and the second by Kosorok (2008, Lemma 12.3). Thus, by the function delta method (Kosorok, 2008, Theorem 2.8),

$$\int_{y\in\mathcal{Y}} d\left\{F_{Y|X}\left(y \mid x, \widehat{\beta}\right) - F_{Y|X}\left(y \mid x, \beta^0\right)\right\} = \mathrm{O}_p\left(n^{-\frac{1}{2}}\right). \tag{C.17}$$

Consequently, by the definition of convergence in probability (Kosorok, 2008, Section 2.2.1), we conclude that $\mathrm{Err}_{2,n} = \mathrm{O}_p\left(n^{-\frac{1}{2}+qd}\lambda^{-q}\right)$.

**Remark C.1.** *When $L$ is the quadratic loss, as mentioned in Remark 4.2, one can directly estimate the conditional expectation $\mu^0(X)$. When $\left|\widehat{\mu}(X) - \mu^0(X)\right| = \mathrm{O}_p(n^{-1/2})$,*

$$\begin{aligned}
\mathrm{Err}_{2,n} &= \sup_{f\in\mathcal{H}_n}\sup_{x\in\mathcal{X}} \left|H\left(x, \widehat{\beta}, f(x)\right) - H\left(x, \beta^0, f(x)\right)\right| \\
&= \sup_{f\in\mathcal{H}_n}\sup_{x\in\mathcal{X}} \left|\mathrm{E}(Y^2 \mid X) + f^2(X) - 2\widehat{\mu}(X)f(X)\right. \\
&\qquad\qquad \left. -\mathrm{E}(Y^2 \mid X) - f^2(X) + 2\mu^0(X)f(X)\right|. \\
&= \sup_{f\in\mathcal{H}_n}\sup_{x\in\mathcal{X}} \left|2\mu^0(X)f(X) - 2\widehat{\mu}(X)f(X)\right|
\end{aligned}$$

*By (C.8), $(Y - f(X))^2 \leq \mathrm{O}\left(n^d\lambda^{-1}\right)$ and $Y$ is bounded, thus,*

$$|f(x)| \leq \mathrm{O}\left(n^{\frac{d}{2}}\lambda^{-\frac{1}{2}}\right).$$

*Consequently,*

$$\mathrm{Err}_{2,n} = \mathrm{O}_p\left(n^{-\frac{1}{2}+\frac{d}{2}}\lambda^{-\frac{1}{2}}\right) \leq \mathrm{O}_p\left(n^{-\frac{1}{2}+d}\lambda^{-1}\right). \qquad \square$$

### C.9. Proof of Theorem 4.2

*Proof.* In the proof of Theorem C.2,

$$A^{\widehat{W},\widehat{H}}(\lambda) - A_2(\lambda) \leq A_n + B_n + C_n + D_n,$$

where $A_n$, $B_n$, $C_n$, and $D_n$ are the same as defined in (C.9). For $A_n + B_n$, we have the same result as (C.11). Next we bound $C_n$ and $D_n$ in the two different situations of (i) and (ii).

Recall that

$$\begin{aligned}\mathcal{R}_{L_{W^0,H^0},D} &= \mathbb{P}_n\left\{L_{W,H}\left(\pi^0, H^0, R, X, Y, f(X)\right)\right\}\\ &= \mathbb{P}_n\left\{\frac{RL(Y,f(X))}{\pi^0(X)} - \frac{R-\pi^0(X)}{\pi^0(X)}H(X,\beta^0,f(X))\right\}.\end{aligned}$$

By Assumption 3.2, we have $\widehat{\pi} \xrightarrow{\mathrm{P}} \pi^*$ and $\widehat{\beta} \xrightarrow{\mathrm{P}} \beta^*$.

Case 1: $\left|\hat{\pi}(X) - \pi^0(X)\right| = \mathrm{O}_p\left(n^{-\frac{1}{2}}\right)$ which means $\mathrm{Err}_{1,n} = \mathrm{O}_p\left(n^{-\frac{1}{2}}\right)$, and $\widehat{\beta} \xrightarrow{\mathrm{P}} \beta^*$ where $\beta^*$ is not necessarily $\beta^0$. Since

$$\begin{aligned}&\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f)\\ =&\mathbb{P}_n\left[\frac{RL(Y,f(X))}{\widehat{\pi}(X)} - \frac{R-\widehat{\pi}(X)}{\widehat{\pi}(X)}H\left(X,\widehat{\beta},f(X)\right)\right]\\ =&\mathbb{P}_n\left[\frac{RL(Y,f(X))}{\pi^0(X)} - \frac{R-\pi^0(X)}{\pi^0(X)}H\left(X,\widehat{\beta},f(X)\right)\right.\\ &\left.+\left\{\frac{RL(Y,f(X)) - RH\left(X,\widehat{\beta},f(X)\right)}{\widehat{\pi}(X)\pi^0(X)}\right\}\left(\pi^0(X)-\widehat{\pi}(X)\right)\right].\end{aligned}$$

Then,

$$\begin{aligned}&\left|\mathcal{R}_{L_{\widehat{W},\widehat{H}},D}(f) - \mathcal{R}_{L_{W^0,H^0},D}(f)\right|\\ =&\left|\mathbb{P}_n\left[\frac{R-\pi^0(X)}{\pi^0(X)}\left\{H(X,\beta^0(X),f(X)) - H\left(X,\widehat{\beta},f(X)\right)\right\}\right]\right.\\ &\left.+\mathbb{P}_n\left[\frac{RL(Y,f(X)) - RH\left(X,\widehat{\beta},f(X)\right)}{\widehat{\pi}(X)\pi^0(X)}\left(\pi^0(X)-\widehat{\pi}(X)\right)\right]\right|\\ \le&\left|\mathbb{P}_n\left[\frac{R-\pi^0(X)}{\pi^0(X)}\right]\right|\mathrm{Err}_{2,n} + \frac{m(c_{2,n}\lambda)^{-1}+1}{c\cdot c_{n,L}}\mathrm{Err}_{1,n}\\ =&|a_n|\mathrm{Err}_{2,n} + \frac{m(c_{2,n}\lambda)^{-1}+1}{c\cdot c_{n,L}}\mathrm{Err}_{1,n}.\end{aligned}$$

Since both $a_n$ and $\mathrm{Err}_{1,n}$ are $\mathrm{O}_p\left(n^{-\frac{1}{2}}\right)$, for every given $\eta > 0$, there exists a constant $b_3(\eta)$ such that for all $n \ge 1$,

$$\mathrm{P}\left(\max\{|a_n|, \mathrm{Err}_{1,n}\} > b_3(\eta)n^{-\frac{1}{2}}\right) < e^{-\eta}.$$

Note that

$$\mathrm{Err}_{2,n} = \sup_{f\in\mathcal{H}_n}\sup_{x\in\mathcal{X}}\left|H\left(x,\widehat{\beta},f(x)\right) - H\left(x,\beta^0,f(x)\right)\right| \le 2\left\{m(c_{2,n}\lambda)^{-1}+1\right\}.$$

Therefore, with probability not less than $1-e^{-\eta}$,

$$C_n + D_n \leq b_3(\eta) n^{-\frac{1}{2}} \left[ 4\left\{ m(c_{2,n}\lambda)^{-1} + 1 \right\} + \frac{2\left\{ m(c_{2,n}\lambda)^{-1} + 1 \right\}}{c \cdot c_{n,L}} \right].$$

Combining this bound with (C.11), for every fixed $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\eta > 0$, with probability not less than $1 - 2e^{-\eta}$,

$$\begin{aligned} & A^{\widehat{W},\widehat{H}}(\lambda) \\ \leq & A_2(\lambda) + \frac{3\left\{ m(c_{2,n}\lambda)^{-1} + 1 \right\}}{c} \left[ 2(c_{2,n}\lambda)^{\frac{1}{2}}\varepsilon \right. \\ & \left. + \sqrt{\frac{2\eta + 2\log\left\{ 2\mathcal{N}(B_{\mathcal{H}}, \|\cdot\|_\infty, (c_{2,n}\lambda)^{\frac{1}{2}}\varepsilon) \right\}}{n}} + \frac{4cb_3(\eta)n^{-\frac{1}{2}}}{3} + \frac{b_3(\eta)n^{-\frac{1}{2}}}{c_{n,L}} \right]. \end{aligned}$$

Together with Condition 4.1 and letting

$$\varepsilon = (c_{2,n}\lambda)^{-\frac{1}{2}} \left(\frac{p}{2}\right)^{\frac{1}{p+1}} \left(\frac{2a}{n}\right)^{\frac{1}{2p+2}},$$

with probability not less than $1 - 2e^{\eta}$,

$$\begin{aligned} & A^{\widehat{W},\widehat{H}}(\lambda) \\ & \leq \frac{3\left\{ m(c_{2,n}\lambda)^{-1} + 1 \right\}}{c} \left[ 3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}} + \left(\frac{2\eta}{n}\right)^{\frac{1}{2}} + \frac{b_3(\eta)n^{-\frac{1}{2}}}{c_{n,L}} + \frac{4cb_3(\eta)n^{-\frac{1}{2}}}{3} \right] \\ & \quad + A_2(\lambda). \end{aligned} \tag{C.18}$$

Since $\frac{1}{c_{2,n}} = \mathrm{O}(n^d)$, using the similar argument as in the proof of Theorem 4.1, when $\lambda_n n^{\min\left(\frac{1}{2p+2}, \frac{1}{2}-d\right)-d} \longrightarrow \infty$, the $\mathcal{P}$-universal consistency holds.

Case 2: $\left|\widehat{\beta} - \beta^0\right| = \mathrm{O}_p\left(n^{-\frac{1}{2}}\right)$, whereas $\widehat{\pi}(X) \xrightarrow{\mathrm{P}} \pi^*(X)$ which is not necessarily equal to $\pi^0(X)$.

$$\begin{aligned} \mathcal{R}_{L_{\widehat{W},\widehat{H},D}}(f) = & \mathbb{P}_n \left[ \frac{RL(Y, f(X))}{\widehat{\pi}(X)} - \frac{R - \widehat{\pi}(X)}{\widehat{\pi}(X)} H\left(X, \beta^0, f(X)\right) \right] \\ & - \mathbb{P}_n \left[ \frac{R - \widehat{\pi}(X)}{\widehat{\pi}(X)} \left\{ H\left(X, \widehat{\beta}, f(X)\right) - H\left(X, \beta^0, f(X)\right) \right\} \right] \end{aligned}$$

Then,

$$\begin{aligned} & \left| \mathcal{R}_{L_{\widehat{W},\widehat{H},D}}(f) - \mathcal{R}_{L_{W^0,H^0,D}}(f) \right| \\ = & \left| \mathbb{P}_n \left[ \frac{RL(Y, f(X)) - RH\left(X, \beta^0, f(X)\right)}{\widehat{\pi}(X)\pi^0(X)} \left\{ \pi^0(X) - \widehat{\pi}(X) \right\} \right] \right. \end{aligned}$$

$$
\begin{aligned}
&- \mathbb{P}_n \left[ \frac{R - \widehat{\pi}(X)}{\widehat{\pi}(X)} \left\{ H\left(X, \widehat{\beta}, f(X)\right) - H\left(X, \beta^0, f(X)\right) \right\} \right] \Bigg| \\
\leq & \frac{\mathrm{Err}_{1,n}}{2c \cdot c_{n,L}} \left| \mathbb{P}_n \left\{ L(Y, f(X)) - H\left(X, \beta^0, f(X)\right) \right\} \right| + \frac{1 + c_{n,U}}{c_{n,L}} \mathrm{Err}_{2,n} \\
\leq & h_n \frac{\mathrm{Err}_{1,n}}{2c \cdot c_{n,L}} + \frac{1 + c_{n,U}}{c_{n,L}} \mathrm{Err}_{2,n}.
\end{aligned}
$$

Note that $\mathrm{Err}_{1,n} \leq 2$. By Lemma 4.1, $h_n = \mathrm{O}_p\left(n^{-\left(\frac{1}{2}-d\right)} \lambda^{-\frac{1}{2}}\right)$, and thus there exists a constant $b_4(\eta)$ such that for all $n \geq 1$,

$$
\mathrm{P}\left\{ |h_n| > b_4(\eta) n^{-\left(\frac{1}{2}-\frac{d}{2}\right)} \lambda^{-\frac{1}{2}} \right\} < e^{-\eta}.
$$

By Lemma 4.2, $\mathrm{Err}_{2,n} = \mathrm{O}_p\left(n^{-\left(\frac{1}{2}-d\right)} \lambda^{-1}\right)$

$$
\mathrm{P}\left\{ \mathrm{Err}_{2,n} \geq b_2(\eta) n^{-\left(\frac{1}{2}-d\right)} \lambda^{-1} \right\} < e^{-\eta}.
$$

For a fixed $\eta > 0$, with probability not less than $1 - 2e^{-\eta}$,

$$
C_n + D_n \leq \frac{2(1 + c_{n,U})}{c_{n,L}} b_2(\eta) n^{-\left(\frac{1}{2}-d\right)} \lambda^{-1} + \frac{1}{c \cdot c_{n,L}} b_4(\eta) n^{-\left(\frac{1}{2}-\frac{d}{2}\right)} \lambda^{-\frac{1}{2}}.
$$

Combining this bound with (C.11), using Condition 4.1, and letting

$$
\varepsilon = (c_{2,n}\lambda)^{-\frac{1}{2}} \left(\frac{p}{2}\right)^{\frac{1}{p+1}} \left(\frac{2a}{n}\right)^{\frac{1}{2p+2}},
$$

for every fixed $\lambda > 0$, $n \geq 1$, $\varepsilon > 0$, and $\eta > 0$, with probability not less than $1 - 3e^{-\eta}$,

$$
\begin{aligned}
& A^{\widehat{W},\widehat{H}}(\lambda) - A_2(\lambda) \\
\leq & \frac{3\left\{ m(c_{2,n}\lambda)^{-1} + 1 \right\}}{c} \left\{ 3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}} + \left(\frac{2\eta}{n}\right)^{\frac{1}{2}} \right\} + \frac{2(1 + c_{n,U})}{c_{n,L}} b_2(\eta) n^{-\left(\frac{1}{2}-d\right)} \lambda^{-1} \\
& + \frac{1}{c \cdot c_{n,L}} b_4(\eta) n^{-\left(\frac{1}{2}-\frac{d}{2}\right)} \lambda^{-\frac{1}{2}}. \qquad \text{(C.19)}
\end{aligned}
$$

Note that $\frac{1+c_{n,U}}{c_{n,L}} = \mathrm{O}\left(n^d\right)$, and that $1 - 2d > \frac{1}{2} - d$. Hence, using the similar argument as in the proof of Theorem 4.1, when $\lambda_n n^{\min\left(\frac{1}{2p+2}, \frac{1}{2}-d\right)-d} \longrightarrow \infty$, the $\mathcal{P}$-universal consistency holds.

**Remark C.2.** *Let $L$ be the quadratic loss. Consider estimating the conditional expectation $\mu^0(X)$. If $\left|\widehat{\mu}(X) - \mu^0(X)\right| = \mathrm{O}_p(n^{-1/2})$, by Remark C.1, $\mathrm{Err}_{2,n} = \mathrm{O}_p\left(n^{-\frac{1}{2}+\frac{d}{2}} \lambda^{-\frac{1}{2}}\right)$. Then, the RHS of (C.19) becomes*

$$
\frac{3\left\{ m(c_{2,n}\lambda)^{-1} + 1 \right\}}{c} \left\{ 3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}} + \left(\frac{2\eta}{n}\right)^{\frac{1}{2}} \right\} + \frac{2(1 + c_{n,U})}{c_{n,L}} b_2(\eta) n^{-\left(\frac{1}{2}-\frac{d}{2}\right)} \lambda^{-\frac{1}{2}}
$$

$$+\frac{1}{c\cdot c_{n,L}}b_4(\eta)n^{-\left(\frac{1}{2}-\frac{d}{2}\right)}\lambda^{-\frac{1}{2}}.$$

*Hence, when*

$$\lambda_n n^{\min\left(\frac{1}{2p+2},\frac{1}{2}-\frac{d}{2}\right)-d}\longrightarrow\infty,$$

*the doubly-robust kernel machine in (3.8) is $\mathcal{P}$-universally consistent.* □

### C.10. Proof of Corollary 4.2

*Proof.* Case 1: $\left|\hat{\pi}(X)-\pi^0(X)\right|=\mathrm{O}_p\left(n^{-\frac{1}{2}}\right)$. By (C.18) and Assumption 4.2, for every fixed $\lambda>0$, $n\geq 1$, $\varepsilon>0$, and $\eta>0$, with probability not less $1-2e^{-\eta}$

$$\begin{aligned}&A^{\widehat{W},\widehat{H}}(\lambda)\\ \leq&\frac{3\left\{m(c_{2,n}\lambda)^{-1}+1\right\}}{c}\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}+\frac{b_3(\eta)n^{-\frac{1}{2}}}{c_{n,L}}+\frac{4cb_3(\eta)n^{-\frac{1}{2}}}{3}\right\}\\ &+b\lambda^{\gamma}.\end{aligned}$$

Let

$$\begin{aligned}&G_2(\lambda)\\ =&\frac{3\left\{m(c_{2,n}\lambda)^{-1}+1\right\}}{c}\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}+\frac{b_3(\eta)n^{-\frac{1}{2}}}{c_{n,L}}+\frac{4c}{3}b_3(\eta)n^{-\frac{1}{2}}\right\}\\ &+b\lambda^{\gamma}.\end{aligned}$$

Taking the derivative with respect to $\lambda$ and setting it equal to 0,

$$b\gamma\lambda^{\gamma-1}=\frac{3m(c_{2,n})^{-1}}{c}\lambda^{-2}\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}+\frac{b_3(\eta)n^{-\frac{1}{2}}}{c_{n,L}}+\frac{4c}{3}b_1(\eta)n^{-\frac{1}{2}}\right\}.$$

Note that $\frac{1}{c_{2,n}}=\mathrm{O}\left(n^d\right)$. Thus,

$$\begin{aligned}\lambda^{\gamma+1}&\propto\left(\frac{1}{n}\right)^{\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)-d}\\ \Rightarrow\lambda&\propto n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{1}{\gamma+1}}.\end{aligned}$$

Note that by choosing large $m$, we have $G_2''\left(n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{1}{\gamma+1}}\right)>0$. Then for $\lambda_n=n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{1}{\gamma+1}}$,

$$\begin{aligned}&G_2(\lambda_n)\\ =&bn^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}\end{aligned}$$

$$
\begin{aligned}
&+\left\{\frac{3m(c_{2,n})^{-1}}{c}n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{-1}{\gamma+1}}+\frac{1}{c}\right\}\\
&\quad\times\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}+\frac{b_3(\eta)n^{-\frac{1}{2}}}{c_{n,L}}+\frac{4c}{3}b_3(\eta)n^{-\frac{1}{2}}\right\}\\
\le& bn^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}\\
&+c_{\mathrm{P}}^{*}\left(c_a^{*}+\sqrt{\eta}+2b_3(\eta)\right)n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{-1}{\gamma+1}-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d}\\
\le& Q^{*}\left\{c_{a,b}^{*}+\sqrt{\eta}+2b_3(\eta)\right\}n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}},
\end{aligned}
$$

where $c_a^{*}$ is a constant related to $a$, $c_{a,b}^{*}$ is a constant related to $a$, $b$, and $c$, $c_{\mathrm{P}}^{*}$ and $Q^{*}$ are constants related to P. None of them is related to $\eta$.

Therefore, for fixed $\lambda>0$, $n\ge 1$, $\varepsilon>0$, and $\eta>0$, with probability not less than $1-3e^{\eta}$,

$$
A^{\widehat{W},\widehat{H}}(\lambda)\le G_2(\lambda_n)\le Q^{*}\left\{c_{a,b}^{*}+\sqrt{\eta}+b_1(\eta)+b_3(\eta)\right\}n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}.
$$

The obtained learning rate is $n^{\left\{-\min\left(\frac{1}{2p+2},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}$.

Case 2: $\left|\widehat{\beta}-\beta^{0}\right|=\mathrm{O}_p\left(n^{-\frac{1}{2}}\right)$. By (C.19) and Assumption 4.2, for every fixed $\lambda>0$, $n\ge 1$, $\varepsilon>0$, and $\eta>0$, with probability not less than $1-3e^{-\eta}$,

$$
\begin{aligned}
A^{\widehat{W},\widehat{H}}(\lambda)\le& b\lambda^{\gamma}+\frac{3\left\{m(c_{2,n}\lambda)^{-1}+1\right\}}{c}\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}\right\}\\
&+\frac{2(1+c_{n,U})}{c_{n,L}}b_2(\eta)n^{-\left(\frac{1}{2}-d\right)}\lambda^{-1}+\frac{1}{c\cdot c_{n,L}}b_4(\eta)n^{-\left(\frac{1}{2}-\frac{d}{2}\right)}\lambda^{-\frac{1}{2}}.
\end{aligned}
$$

Choosing $0<\lambda<1$, we have

$$
\begin{aligned}
A^{\widehat{W},\widehat{H}}(\lambda)\le& b\lambda^{\gamma}+\frac{3\left\{m(c_{2,n}\lambda)^{-1}+1\right\}}{c}\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}\right\}\\
&+\frac{2(1+c_{n,U})}{c_{n,L}}b_2(\eta)n^{-\left(\frac{1}{2}-d\right)}\lambda^{-1}+\frac{1}{c\cdot c_{n,L}}b_4(\eta)n^{-\left(\frac{1}{2}-\frac{d}{2}\right)}\lambda^{-1}.
\end{aligned}
$$

Let

$$
\begin{aligned}
G_3(\lambda)=& b\lambda^{\gamma}+\frac{3\left\{m(c_{2,n}\lambda)^{-1}+1\right\}}{c}\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}\right\}\\
&+\frac{2(1+c_{n,U})}{c_{n,L}}b_2(\eta)n^{-\left(\frac{1}{2}-d\right)}\lambda^{-1}+\frac{1}{c\cdot c_{n,L}}b_4(\eta)n^{-\left(\frac{1}{2}-\frac{d}{2}\right)}\lambda^{-1}.
\end{aligned}
$$

Taking the derivative with respect to $\lambda$ and setting it equal to 0,

$$
b\gamma\lambda^{\gamma-1}=\frac{3m(c_{2,n})^{-1}}{c}\lambda^{-2}\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}\right\}
$$

$$+\frac{2(1+c_{n,U})}{c_{n,L}}b_2(\eta)n^{-\left(\frac{1}{2}-d\right)}\lambda^{-2}+\frac{1}{c\cdot c_{n,L}}b_4(\eta)n^{-\left(\frac{1}{2}-\frac{d}{2}\right)}\lambda^{-2}.$$

Then

$$\lambda^{\gamma+1}\propto\left(\frac{1}{n}\right)^{\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)-d}\Rightarrow\lambda\propto n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{1}{\gamma+1}}.$$

Note that by choosing large $m$, we have $G_3''\left(n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{1}{\gamma+1}}\right)>0$.

Then for $\lambda_n=n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{1}{\gamma+1}}$,

$$\begin{aligned}
&G_3(\lambda_n)\\
=&bn^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}\\
&+\left\{\frac{3m(c_{2,n})^{-1}}{c}n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{-1}{\gamma+1}}+\frac{1}{c}\right\}\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}\right\}\\
&+\frac{1+c_{n,U}}{c_{n,L}}b_2(\eta)n^{-\left(\frac{1}{2}-d\right)}n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{-1}{\gamma+1}}\\
&+\frac{2}{c\cdot c_{n,L}}b_4(\eta)n^{-\left(\frac{1}{2}-\frac{d}{2}\right)}n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{-1}{\gamma+1}}\\
\le&bn^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}\\
&+c_{\mathrm{P}}^{\star}\left(c_a^{\star}+\sqrt{\eta}+b_2(\eta)+b_4(\eta)\right)n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{-1}{\gamma+1}-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d}\\
\le&Q^{\star}\left(c_{a,b}^{\star}+\sqrt{\eta}+b_2(\eta)+b_4(\eta)\right)n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}},
\end{aligned}$$

where $c_a^{\star}$ is a constant related to $a$, $c_{a,b}^{\star}$ is a constant related to $a$, $b$, and $c$, $c_{\mathrm{P}}^{\star}$ and $Q^{\star}$ are constants related to P. None of them is related to $\eta$.

Therefore, for fixed $0<\lambda<1$, $n\ge1$, $\varepsilon>0$, and $\eta>0$, with probability not less than $1-3e^{\eta}$,

$$A^{\widehat{W},\widehat{H}}(\lambda)\le G_3(\lambda_n)\le Q^{\star}\left(c_{a,b}^{\star}+\sqrt{\eta}+b_2(\eta)+b_4(\eta)\right)n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}.$$

The obtained learning rate is $n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-d\right)+d\right\}\frac{\gamma}{\gamma+1}}$.

**Remark C.3.** *When $L$ is the quadratic loss and $\left|\widehat{\mu}(X)-\mu^0(X)\right|=\mathrm{O}_p(n^{-1/2})$. By Remark C.2, we have*

$$\begin{aligned}
b\gamma\lambda^{\gamma-1}=&\frac{3m(c_{2,n})^{-1}}{c}\lambda^{-2}\left\{3\left(\frac{2a}{n}\right)^{\frac{1}{2p+2}}+\left(\frac{2\eta}{n}\right)^{\frac{1}{2}}\right\}\\
&+\frac{2(1+c_{n,U})}{c_{n,L}}b_2(\eta)n^{-\left(\frac{1}{2}-\frac{d}{2}\right)}\lambda^{-2}+\frac{1}{c\cdot c_{n,L}}b_4(\eta)n^{-\left(\frac{1}{2}-\frac{d}{2}\right)}\lambda^{-2}.
\end{aligned}$$

*and*

$$\lambda\propto n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-\frac{d}{2}\right)+d\right\}\frac{1}{\gamma+1}}.$$

*By the similar argument as in case 2 of the proof, the learning rate is*

$$n^{\left\{-\min\left(\frac{1}{2p+1},\frac{1}{2}-\frac{d}{2}\right)+d\right\}\frac{\gamma}{\gamma+1}}.$$

□

**Acknowledgements**

The authors thank the editor, the associate editor and the two referees for carefully reading our manuscript and their helpful comments and constructive suggestion to improve the paper. Yair Goldberg was partially supported by the Israeli Science Foundation (grant No. 849/17). Tiantian Liu was partially supported by China Scholarship Council.

**Supplementary material**

**Supplement: R code**
(`https://tinyurl.com/KM4ICDcode`). R package `KM4ICD` for the weighted-complete-case kernel machine estimator and the doubly-robust kernel machine estimator, and R code for Section 5 and Section 6.

**References**

D. Azriel, L. D. Brown, M. Sklar, R. Berk, A. Buja, and L. Zhao. Semi-supervised linear regression. `https://arxiv.org/abs/1612.02391v1`, 2016.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973, 2005. MR2216189

N. Berestycki, R. Nickl, and B. Schlein. Concentration of measure. 12, 2009.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68, 2018. MR3769544

Y. Goldberg and M. R. Kosorok. Support vector regression for right censored data. *Electronic Journal of Statistics*, 11:532–569, 2017. MR3619316

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

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47:663–685, 1952. MR0053460

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, New York, 2013. MR3100153

M. R. Kosorok. *Introduction to Empirical Inference Processes and Semiparametric Inference*. Springer, New York, 2008. MR2724368

B. Kriegler and R. Berk. Small area estimation of the homeless in Los Angeles: An application of cost-sensitive gradient boosting. *Annals of Applied Statistics*, 4:1234–1255, 2010. MR2751340

E. B. Laber and S. A. Murphy. Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association*, 106:904–913, 2011. MR2894746

H. Liang, S. J. Wang, and R. J. Carroll. Partially linear models with missing response variables and error-prone covariates. *Biometrika*, 94:185, 2007. MR2307903

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 2nd edition, 2002. MR1925014

D. W. Liu, X. H. Lin, and D. Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63:1079–1088, 2007. MR2414585

K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Handling missing values in support vector machine classifiers. *Neural Networks: The Official Journal of the International Neural Network Society*, 18:684–692, 2005.

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

A. Rotnitzky, J. M. Robins, and D. O. Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93:1321–1339, 1998. MR1666631

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 2004. MR2117498

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*, 33:184–197, 2018. MR3797709

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

T. G. Stewart, D. L. Zeng, and M. C. Wu. Constructing support vector machines with missing data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10:e1430, 2018. MR3826095

Z. Q. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97:661–682, 2010. MR2672490

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

K. Vermeulen and S. Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110:1024–1036, 2015. MR3420681

L. Wang, A. Rotnitzky, and X. H. Lin. Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association*, 105:1135–1146, 2010. MR2752609

Q. H. Wang and J. N. K. Rao. Empirical likelihood-based inference under imputation for missing response data. *Annals of Statistics*, 30:896–924, 2002. MR1922545

Q. H. Wang, O. Linton, and W. Härdle. Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, 99:334–345, 2004. MR2062820

A. L. Yuille and A. Rangarajan. The concave convex procedure. *Neural Computation*, 15:915–936, 2003.