Electronic Journal of Statistics
Vol. 19 (2025) 1–53
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2335

Efficient surrogate-assisted inference for patient-reported outcome measures with complex missing mechanism

Jaeyoung Park[1], **Muxuan Liang**[2], **Ying-Qi Zhao**[3] and **Xiang Zhong**[4]

[1]*School of Global Health Management & Informatics, University of Central Florida, e-mail:* jaeyoung.park@ucf.edu

[2]*Department of Biostatistics, University of Florida, e-mail:* muxuan.liang@ufl.edu

[3]*Public Health Sciences Divisions, Fred Hutchinson Cancer Center, e-mail:* yqzhao@fredhutch.org

[4]*Department of Industrial and Systems Engineering, University of Florida, e-mail:* xiang.zhong@ise.ufl.edu

Abstract: Patient-reported outcome (PRO) measures are increasingly collected as a means of measuring healthcare quality and value. The capability to predict such measures enables patient-provider shared decision-making and the delivery of patient-centered care. However, PRO measures often suffer from high missing rates, and the missingness may depend on many patient factors. Under such a complex missing mechanism, developing a predictive model for PRO measures with valid inference procedures is challenging, especially when flexible imputation models such as machine learning or nonparametric methods are used. Specifically, the slow convergence rate of the flexible imputation model may lead to non-negligible bias, and the traditional missing propensity, capable of removing such a bias, is hard to estimate due to the complex missing mechanism. To efficiently infer the parameters of interest, we propose to use an informative surrogate that enables a flexible imputation model lying in a low-dimensional subspace. To remove the bias due to the flexible imputation model, we identify a class of weighting functions as alternatives to the traditional propensity score and estimate the low-dimensional one within the identified function class. Based on the estimated low-dimensional weighting function, we construct a one-step debiased estimator without using any information on the true missing propensity. We establish the asymptotic normality of the one-step debiased estimator. Simulation and an application to real-world data demonstrate the superiority of the proposed method.

MSC2020 subject classifications: Primary 62-08; secondary 62D10.
Keywords and phrases: Missing data, dimension reduction, semiparametric inference, semi-supervised learning, double machine learning.

Received February 2023.

Contents

arXiv: 2210.09362v2

1. Introduction

Patient-reported outcome (PRO) measures are increasingly collected before and after an intervention or a treatment as a means of measuring healthcare quality and value, which is an important step toward patient-centered care. Knowing whether the measure goes up or down alone might not be sufficient to determine the effectiveness of the intervention. More importantly, whether the measure has changed with a sufficiently large margin, known as the minimally clinically important difference (MCID), needs to be evaluated. If the intervention is an elective surgery, identifying patients at risk of not achieving an MCID, particularly before the surgery, is important for pre-surgical decisions. There is a growing interest in applying machine learning techniques to predict whether a patient is likely to achieve an MCID before surgery and identify predictive factors associated with post-surgical PRO measures.

The increasing adoption of electronic health record (EHR) systems has provided unprecedented opportunities to learn an interpretable model for predicting PRO measures using massive observational data. Although the volume of observational data is large, the quality of such observational data may be uncertain. One of the major difficulties is missing data, especially missing the outcome data. In our motivating example, the MCIDs can only be observed from the participants who take both pre- and post-surgical surveys. The participants who completed both surveys may only account for a small portion (e.g., 1/3) of the participants whose EHR data is available, according to the response rate reported in literature [13, 26] and from our own data. Unfortunately, low survey

response rates are not uncommon in healthcare and other service industries. In this work, our objective is to develop an interpretable predictive model for the outcome subject to missing. Specifically, we aim to develop a linear prediction model by minimizing the deviance of a generalized linear model (GLM) with a valid inference procedure for the coefficients under possible model misspecification.

Many approaches have been developed to deal with missing outcomes under the assumption of missing at random (MAR) [16]. One seminal work is the propensity inverse weighting approach [30, 14]. For this approach, one first estimates the probability of missing w.r.t the covariate (also called the propensity) and then uses the inverse of the estimated propensity to adjust for the selection bias. When the propensity is poorly estimated, the propensity inverse weighting methods may not perform well. Another major type of approaches is known as imputation. This approach first learns an imputation model using the fully observed part of the data; then, imputes the missing outcomes with the predicted values; and finally, refits the predictive model based on the imputed outcomes [32]. When the estimated imputation model is misspecified, the refitted predictive model may be biased. To maintain robustness against the possible misspecification in the propensity and the imputation models, one possible solution is to use the doubly robust methods [29]. The doubly robust methods that incorporate both the propensity score and the imputation models can lead to a consistent estimate for the outcome as long as either model is correctly specified [35, 36, 27, 28, 33, 2, 9, 31, 10].

Statistical inference for the parameters in predictive modeling with outcome missingness is also challenging. In particular, when the missing mechanism depends on multiple covariates through a nonlinear relationship, a consistent estimator for the missing propensity with a fast convergence rate may be infeasible. For the inverse weighting approaches and the doubly robust methods, a parametric model for the propensity may not capture the potential non-linearity. To ensure a consistent propensity estimate, nonparametric regressions and machine learning methods have been adopted. These methods may lead to a slower convergence rate and hinder the inference of the parameters in the predictive model, especially when the number of covariates is large. When the number of covariates is small, to address the slow convergence rate, a double machine learning approach was proposed in [4]. They adopted a cross-fitting algorithm using a doubly robust formulation and proposed to estimate both the propensity and the imputation model using nonparametric or machine learning methods. They proved that, as long as the product of the convergence rates of the propensity and imputation estimates is smaller than $n^{-1/2}$, a valid inference procedure for the parameters in the predictive model is possible, where n represents the sample size. However, their required rate condition may be negated due to a large number of covariates or failure to meet the smoothness condition on the true propensity under a complex missing mechanism.

To address the above statistical inference challenge due to the presence of a large number of covariates, one possible strategy is to leverage a surrogate outcome. The surrogate outcomes herein are defined as alternative clinical out-

comes that are likely to predict the clinical benefit of primary interest. In our motivating example, the MCID of the global physical health T-score in the Patient-Reported Outcomes Measurement Information System (PROMIS) survey is a well-acknowledged measurement for evaluating surgery benefits. There are other PRO measures collected that represent different but related mental or physical health performances that can be considered surrogate outcomes. In many applications, a surrogate outcome can help improve efficiency or overcome the difficulties due to complex missing mechanisms. In causal inference, a surrogate can be used to improve the efficiency of estimating the average treatment effect (ATE) [25, 8, 6, 3, 1]. In the application of semi-supervised inference, under the assumption of missing completely at random (MCAR), [15] showed that a surrogate can help infer the predicted risk derived from a high-dimensional working model even when the true risk prediction model depends on multiple covariates. However, their approach cannot be applied under the assumption of MAR, which is the setting we need to deal with.

In this work, we focus on using surrogate outcomes to develop interpretable predictive models with outcome missingness. The parameter of interest herein is defined as the minimizer of the deviance under a GLM with possible model misspecification. We propose a concept of an informative surrogate, defined as a surrogate outcome that enables a low-dimensional imputation model conditional on the surrogate and the covariates (i.e., the imputation model lies in a low-dimensional subspace generated by the surrogate and the covariates). Under the MAR assumption, we exploit the role of this informative surrogate to 1) allow for a low-dimensional imputation model under a large number of covariates and 2) avoid estimating the complex missing propensity. To harvest the potential benefit brought by informative surrogate outcomes, we propose the following procedure. First, we estimate a flexible imputation model (e.g., using kernel regression or basis expansion) in a reduced subspace that is constructed by leveraging the information from informative surrogate outcomes. Subsequently, we can impute the missing outcomes and obtain an initial estimator for the parameters of interest. Then, we bypass the estimation of the complex missing propensity and instead estimate a low-dimensional weighting function based on the reduced subspace to adjust for the possible bias due to the estimated imputation model. Finally, a one-step debiased estimator for the parameters in the predictive model is constructed. Both the point and interval estimates of the parameters can be obtained. We further show that the proposed method can provide a valid inference procedure for the parameters of interest without requiring a consistent propensity estimation. In addition, when the true propensity lies in the same subspace as the imputation model does, the proposed method leads to a semiparametric efficient estimator for the parameters in the predictive model. Extensive simulation and a study of real-world data are provided to demonstrate the superior performance of the proposed method.

The remainder of the paper is organized as follows. In Section 2, we define the parameter of interest and introduce our proposed method. In Section 3, we demonstrate the theoretical validity of the proposed method. In Section 4, we provide numerical studies to bolster the superiority of the proposed methods

over other existing methods and methods without information on the surrogate. In Section 5, we apply the proposed method to derive a predictive rule to infer post-surgery improvement for joint replacement surgery patients. In Section 6, we discuss possible future works.

2. Method

Let $\boldsymbol{X}$ be a p-dimensional covariate and Y be a binary, categorical, or continuous outcome of interest. Without loss of generality, we choose a GLM as a working model for $E[Y \mid \boldsymbol{X}]$. Following the notation of exponential family distributions [34], a GLM assumes that $E[Y \mid \boldsymbol{X}] = b'(\boldsymbol{X}^\top \boldsymbol{\beta})$, where $b'(\cdot)$, the derivative of function $b(\cdot)$, is a known link function. The parameter of interest, $\boldsymbol{\beta}$, is often defined as the minimizer of the deviance (or equivalently, the negative log-likelihood) under the working model, i.e., $\boldsymbol{\beta}^* = \arg\min E[\ell(\boldsymbol{\beta})]$, where $\ell(\boldsymbol{\beta}) = b(\boldsymbol{X}^\top \boldsymbol{\beta}) - Y\boldsymbol{X}^\top \boldsymbol{\beta}$. If the working model is misspecified, i.e., $E[Y \mid \boldsymbol{X}] \neq b'(\boldsymbol{X}^\top \boldsymbol{\beta}^*)$, $\boldsymbol{\beta}^*$ that minimizes the deviance, a goodness-of-fit statistic, is still meaningful. For a linear working model, the link function $b'(t)$ is the identity function, and the function $b(t) = t^2/2$; the objective is equivalent to the ordinary least square (OLS). Notice that the parameter of interest, $\boldsymbol{\beta}^*$, is defined under the full distribution where Y and $\boldsymbol{X}$ are always observed. To ensure that $\boldsymbol{\beta}^*$ can be identified under the full distribution, we assume that $b''(\cdot)$, the second order derivative of function $b(\cdot)$, is always positive and $E[\boldsymbol{X}\boldsymbol{X}^\top]$ is positive definite.

For actual data, the outcome Y can be missing. We collect the covariate $\boldsymbol{X}$, the outcome Y, the informative surrogate outcome Z, and the missing indicator R from all samples. The missing indicator R indicates whether Y is observed ($R = 1$) or not ($R = 0$). We also assume that the surrogate Z can be fully observed. Collectively, the observed data can be denoted as $(\boldsymbol{X}, Z, R, RY)$. To ensure the identifiability of $\boldsymbol{\beta}^*$ using the actual data, we assume that $Y \perp R \mid \boldsymbol{X}, Z$.

2.1. First step: dimension reduction through informative surrogate

In this section, we propose a two-step procedure under the assumption of $Y \perp R \mid \boldsymbol{X}, Z$. To start with, we formally define the concept of informative surrogate outcomes and introduce additional conditions for the identifiability of $\boldsymbol{\beta}^*$.

An surrogate outcome Z is informative if 1) there exists a $(p+1) \times d$ matrix, $\boldsymbol{\Gamma}$, with orthogonal columns satisfying $Y \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$ and $d < p$, where $\widetilde{\boldsymbol{X}}^\top = (Z, \boldsymbol{X}^\top)$; 2) the coefficients in $\boldsymbol{\Gamma}$ corresponding to Z are not all zeros. The first requirement of the definition implies that, conditioning on the surrogate outcome, the dimension of the space constructed by the covariates and the surrogate can be reduced to d, which is expected to be much smaller than p. The columns of $\boldsymbol{\Gamma}$ represent the reduced subspace. Thus, if the surrogate is informative, $Q(Z, \boldsymbol{X}) := E[Y \mid Z, \boldsymbol{X}]$ is a function lying in a low-dimensional subspace, i.e., there exists an unknown link function g such that $Q(Z, \boldsymbol{X}) = g(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$. The

second requirement of the definition implies that $E[Y \mid Z, \boldsymbol{X}] \neq E[Y \mid \boldsymbol{X}]$, and thus an informative surrogate enables a more accurate imputation model. Consequently, an efficient estimator for this low-dimensional imputation model may have a faster convergence rate and more accurate imputations than directly using the kernel regression to estimate $E[Y \mid \boldsymbol{X}]$.

Various existing methods can be employed to estimate the reduced subspace when the actual data is fully observable. The assumption $Y \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$ is closely related to the (sufficient) dimension reduction. In dimension reduction literature [18, 5, 40, 39, 21, 22], the smallest space generated by the columns of $\boldsymbol{\Gamma}$ that satisfies $Y \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$ is referred to as the central subspace. When the data are fully observed, the dimension reduction methods, such as the minimum average variance estimation (MAVE) [40], the sliced-inverse regression (SIR) [18], or semiparametric approaches in [21, 22], can be directly applied to estimate the central subspace $\boldsymbol{\Gamma}$, and the estimator is asymptotically normal. When there are multiple surrogate outcomes in the observed data, the above-mentioned methods (e.g., the MAVE) can be used to select candidate informative surrogate outcomes. For example, we can select the surrogate outcome (or a combination of multiple surrogate outcomes) that leads to the lowest reduced dimension.

Remark 2.1. Our proposed method and theory can accommodate multiple surrogate outcomes. For ease of exposition, we focus on only one surrogate in the main text. In Appendix A.4, we provide the theory and its proof when multiple surrogate outcomes exist.

With incomplete outcome data, to ensure the identifiability of $\boldsymbol{\beta}^*$ and $\boldsymbol{\Gamma}$, the traditional positivity assumption requires that $P(R = 1 \mid \widetilde{\boldsymbol{X}}) > 0$. In this work, instead of imposing the traditional positivity assumption, we consider a relaxed positivity assumption, $P(R = 1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) > 0$. Under this assumption, if $Y \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$, we can show that $Y \perp R \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$, and $Q(Z, \boldsymbol{X}) = E[Y \mid Z, \boldsymbol{X}] = E\left[Y \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}\right] = E\left[Y \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}, R = 1\right]$. This implies that the conditional mean of Y restricted to $R = 1$ shares the same subspace with the unrestricted conditional mean. Thus, to estimate $\boldsymbol{\Gamma}$, we only need to apply these dimension reduction methods to the fully observed part of the data. Since $\boldsymbol{\Gamma}$ and subsequently $Q(Z, \boldsymbol{X})$ are identifiable, $\boldsymbol{\beta}^*$ is also identifiable under the relaxed positivity assumption.

Based on the identifiability under the relaxed positivity assumption and an estimator for $\boldsymbol{\Gamma}$, we can construct an initial estimator for $\boldsymbol{\beta}^*$. Specifically, after obtaining $\widehat{\boldsymbol{\Gamma}}$, we can use nonparametric regressions or machine learning methods to fit Y w.r.t $\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}$ to derive the unknown link function g and estimate $\widehat{Q}(Z, \boldsymbol{X}) = \widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})$. Then, we obtain an initial estimator for $\boldsymbol{\beta}^*$ by minimizing $\widehat{E}_n\left[b(\boldsymbol{X}^\top \boldsymbol{\beta}) - \widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \boldsymbol{X}^\top \boldsymbol{\beta}\right]$, or equivalently, solving the estimating equation

$$\widehat{E}_n\left[\left\{b'(\boldsymbol{X}^\top \boldsymbol{\beta}) - \widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})\right\} \boldsymbol{X}\right] = 0.$$

Denote the solution as $\widehat{\boldsymbol{\beta}}$. Due to the slow convergence rate of nonparametric regressions or machine learning methods, the convergence rate of $\widehat{Q}$ is dominated

by that of $\widehat{g}$. Subsequently, $\widehat{\boldsymbol{\beta}}$ may suffer from the slow convergence rate of $\widehat{g}$. Thus, to obtain an estimator with a faster convergence rate, we need to remove the bias due to the estimation error of $\widehat{g}$, which will be discussed in the following section.

Remark 2.2. The term $\widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})$ is expected to form a good prediction for Y. As an imputation model, it should be predictive of Y; however, there are at least two reasons that it may not be satisfactory. First, $\widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})$ depends on the surrogate outcome Z, which may not be available at the time of making the prediction and thus is not appropriate to be treated as a covariate. Second, when the reduced dimension, i.e., the dimension of $\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}$, is greater than 2, it could be hard to interpret the model.

2.2. Second step: debias using a low-dimensional weighting function

In this section, we attempt to remove the bias due to the estimation error of $\widehat{g}$ using a low-dimensional weighting function. For ease of exposition, we focus on how to construct an improved estimator for β_1^*, which is the first coefficient in $\boldsymbol{\beta}^*$. The proposed method can be extended to infer $\boldsymbol{u}^\top \boldsymbol{\beta}^*$ for any $\boldsymbol{u}$. To get an improved estimator for $\boldsymbol{\beta}^*$, we can implement the proposed method for each coordinate of $\boldsymbol{\beta}^*$ and ensemble these estimates to construct an estimator for $\boldsymbol{\beta}^*$.

To start with, we consider a class of estimating equations for β_1^*. We first derive the efficient influence function (See Appendix A.1) of β_1^* without assuming any relationship between Y and Z given covariates $\boldsymbol{X}$. Motivated by the efficient influence function, we consider the following class of estimating equations for β_1^*,

$$E\left[\{S(\boldsymbol{\beta};\pi,Q)\}^\top \boldsymbol{v}\right] = 0, \tag{2.1}$$

where

$$S(\boldsymbol{\beta};\pi,Q) \quad = \quad \left[\left\{b'(\boldsymbol{X}^\top\boldsymbol{\beta}) - Q(Z,\boldsymbol{X})\right\} + \pi^{-1}(\boldsymbol{X},Z)R\left\{Q(Z,\boldsymbol{X}) - Y\right\}\right]\boldsymbol{X},$$

$Q(Z,\boldsymbol{X})$ is the true imputation model for Y, and $\pi(\boldsymbol{X},Z)$ is an arbitrary weighting function. In this class of estimating equations, the first term

$$\left\{b'(\boldsymbol{X}^\top\boldsymbol{\beta}) - Q(Z,\boldsymbol{X})\right\}\boldsymbol{X}^\top\boldsymbol{v}$$

corresponds to the estimating equation using $Q(Z,\boldsymbol{X})$ as the imputation for all the outcomes. The second term

$$\pi^{-1}(\boldsymbol{X},Z)R\left\{Q(Z,\boldsymbol{X}) - Y\right\}\boldsymbol{X}^\top\boldsymbol{v}$$

can be interpreted as an efficiency augmentation term using the weighting function $\pi^{-1}(\boldsymbol{X},Z)$. In the first step, we have obtained an imputation model $\widehat{Q}(Z,\boldsymbol{X})$ and the initial estimator $\widehat{\boldsymbol{\beta}}$, which can be plugged into estimating

equations (2.1). That is, we will consider $E\left[\left\{S(\boldsymbol{\beta};\pi,\widehat{Q})\right\}^{\top}\boldsymbol{v}\right]=0$ with $\boldsymbol{\beta}$ being replaced by $(\beta_1,\widehat{\boldsymbol{\beta}}_{-1})$, where $\widehat{\boldsymbol{\beta}}_{-1}$ is the sub-vector of $\widehat{\boldsymbol{\beta}}$, excluding the first coordinate.

However, directly solving this estimating equation for an arbitrary choice of $\boldsymbol{v}$ and the weighting function may not lead to an improved estimator due to the estimation error of $\widehat{g}$. In the following, we discuss how to choose $\boldsymbol{v}$ and $\pi(\boldsymbol{X},Z)$ such that the first-order bias of the estimating equation induced by $\widehat{g}$ can be removed. This estimation error affects the estimating equation via two paths. First, the estimating equation depends on $\widehat{Q}$, which is directly affected by the estimation error of $\widehat{g}$; second, the estimating equation depends on $\widehat{\boldsymbol{\beta}}_{-1}$, which is also affected by the estimation error of $\widehat{g}$.

To obviate the impact of the estimation error of $\widehat{g}$ on $\widehat{\boldsymbol{\beta}}_{-1}$, we resort to the de-correlated score [24]. The de-correlated score projects the score in a chosen direction such that the projected estimating equation is not affected by the estimation error of $\widehat{\boldsymbol{\beta}}_{-1}$. Assimilating this idea, we choose the following $\boldsymbol{v}$ to achieve this goal. Let $\boldsymbol{w}^*$ be the minimizer of

$$E\left[b''(\boldsymbol{X}^{\top}\boldsymbol{\beta}^*)(X_1-\boldsymbol{X}_{-1}^{\top}\boldsymbol{w})^2\right],$$

where X_1 is the first covariate in $\boldsymbol{X}$ and $\boldsymbol{X}_{-1}$ is the covariate vector of $\boldsymbol{X}$ excluding X_1. Consider the following estimating equation for β_1^*,

$$E\left[\left\{S(\boldsymbol{\beta};\pi,\widehat{Q})\right\}^{\top}\boldsymbol{v}\right]=0, \tag{2.2}$$

where $\boldsymbol{v}^{\top}=(1,-\boldsymbol{w}^{*\top})$ and $\boldsymbol{\beta}=(\beta_1,\widehat{\boldsymbol{\beta}}_{-1})$. Using this estimating equation, the estimation error of $\widehat{\boldsymbol{\beta}}_{-1}$ will not affect the estimation of β_1^*.

To obviate the impact of the estimation error of $\widehat{g}$ on $\widehat{Q}$, we choose a tailored weighting function. Under the proposed estimating equation (2.2), for any choice of π, the first-order bias of the proposed estimating equation (2.2) with Q being replaced by $\widehat{Q}$ is

$$\begin{aligned} & E\left[\left\{R/\pi(\widetilde{\boldsymbol{X}})-1\right\}\left\{\widehat{g}(\widehat{\boldsymbol{\Gamma}}^{\top}\widetilde{\boldsymbol{X}})-g(\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{X}})\right\}\boldsymbol{X}^{\top}\boldsymbol{v}\right] \\ \approx\ & E\left[\left\{R/\pi(\widetilde{\boldsymbol{X}})-1\right\}\left\{\widehat{g}(\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{X}})-g(\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{X}})\right\}\boldsymbol{X}^{\top}\boldsymbol{v}\right]. \end{aligned}$$

In order to remove the estimation error of $\widehat{Q}$, one possible strategy is to choose $\pi(\widetilde{\boldsymbol{X}})$ such that

$$E\left[\left\{R/\pi(\widetilde{\boldsymbol{X}})-1\right\}f(\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{X}})\boldsymbol{X}^{\top}\boldsymbol{v}\right]=0,\ \text{for any } f\in L_2(\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{X}}). \tag{2.3}$$

Let $P(R=1)=\rho>0$ and $\eta(\cdot\mid R)$ be the conditional density function of $\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{X}}$ given R. Define

$$J_r(\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{x}})=E\left[\boldsymbol{X}^{\top}\boldsymbol{v}\mid\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{X}}=\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{x}},R=r\right]\eta(\boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{x}}\mid R=r),$$

for $r=0$ and 1. Theorem 2.1 characterizes the solution to Equation (2.3).

Theorem 2.1. *We define two regularity conditions:*

1. $P(J_1(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \neq 0) = 1$;
2. $\{\widetilde{\boldsymbol{x}} : J_1(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{x}}) = 0\} \subset \{\widetilde{\boldsymbol{x}} : J_0(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{x}}) = 0\}$.

Under any of the two regularity conditions, we can point-wisely define $\pi_*^{-1}(\widetilde{\boldsymbol{x}})$ *as*

$$\pi_*^{-1}(\widetilde{\boldsymbol{x}}) = E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} = \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{x}}\right] / E\left[R\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} = \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{x}}\right]$$

on $\{\widetilde{\boldsymbol{x}} : J_1(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{x}}) \neq 0\}$. *In addition, such a* $\pi_*(\widetilde{\boldsymbol{x}})$ *is a solution to Equation (2.3). Further, the solution set of Equation (2.3) can be characterized as all the functions of the form* $\pi_*^{-1}(\widetilde{\boldsymbol{x}}) + \mathcal{T}h(\widetilde{\boldsymbol{x}})$ *on* $\{\widetilde{\boldsymbol{x}} : J_1(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{x}}) \neq 0\}$, *where* $\mathcal{T}$ *is a linear operator defined as*

$$\mathcal{T}h = h - E\left[h\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}, R = 1\right] / E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}, R = 1\right]$$

and h *is an arbitrary function in* $L_2(\widetilde{\boldsymbol{X}})$.

The proof of Theorem 2.1 can be found in Appendix A.2.

Remark 2.3. The term $\pi_*^{-1}(\widetilde{\boldsymbol{x}})$ can also be written as

$$1 + \left\{\rho J_1(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{x}})\right\}^{-1} J_0(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{x}})(1-\rho).$$

Thus, to estimate $\pi_*^{-1}(\widetilde{\boldsymbol{x}})$, we can estimate J_1, J_0, and ρ, and then plug-in these estimates and $\widehat{\boldsymbol{\Gamma}}$ to construct an estimator for $\pi_*^{-1}(\widetilde{\boldsymbol{x}})$.

Remark 2.4. The regularity conditions can be easily satisfied. For example, assume the density of $\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \mid R = 1$ is bounded above 0. If

$$E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}, R = 1\right]$$

has a continuous distribution, then we have $P\left(J_1(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \neq 0\right) = 1$. In addition, when $P(R = 1 \mid Z, \boldsymbol{X}) = P(R = 1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$, the regularity condition is also satisfied because

$$E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}, R = 1\right] = E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}, R = 0\right].$$

Remark 2.5. The term $\pi_*^{-1}(\widetilde{\boldsymbol{x}})$ is a function of $\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{x}}$. This can be interpreted as the consequence of $Y \perp R \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$. However, the term $\pi_*^{-1}(\widetilde{\boldsymbol{x}}) + \mathcal{T}h(\widetilde{\boldsymbol{x}})$ may not include $P(R = 1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$. This implies that the true propensity based on $\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$ may not be sufficient to remove the bias.

We then use $\pi_*^{-1}(\widetilde{\boldsymbol{X}})$ to replace the $\pi^{-1}(\widetilde{\boldsymbol{X}})$ in estimating equation (2.2). The weighting function $\pi_*^{-1}(\widetilde{\boldsymbol{X}})$ only depends on $\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$ and thus is a low-dimensional function. To estimate the weighting function $\pi_*^{-1}(\widetilde{\boldsymbol{X}})$, we consider the trimmed kernel estimates. First, we estimate $\boldsymbol{w}$ by minimizing

$$\widehat{E}_n\left[b''(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}})(X_1 - \boldsymbol{X}_{-1}^\top \boldsymbol{w})^2\right],$$

and construct $\widehat{\boldsymbol{v}}^\top = (1, -\widehat{\boldsymbol{w}}^\top)$. Then, using kernel regressions, we consider the following estimator for π_*^{-1}:

$$\widehat{\pi}^{-1}(\widetilde{\boldsymbol{x}}; \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{v}}) = \begin{cases} 1 + \left\{ \widehat{J}_1(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \widehat{\rho} \right\}^{-1} \widehat{J}_0(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}})(1 - \widehat{\rho}) & \left| \widehat{J}_1(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \right| > c_n, \\ \widehat{\rho}^{-1}, & \left| \widehat{J}_1(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \right| \leqslant c_n, \end{cases}$$

where $\widehat{J}_r(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) = \widehat{E}_n \left[\boldsymbol{X}^\top \widehat{\boldsymbol{v}} K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \mid R = r \right]$, $\widehat{\rho} = \widehat{E}_n[R]$, $K_\hbar(\cdot) = K(\cdot)/\hbar^d$, and $\widehat{E}_n[\cdot \mid R = r]$ is the empirical mean over the samples with $R = r$. The function $K(\cdot)$ is a kernel function with the order of $\nu - 1$, and the bandwidth parameter $\hbar$ is selected according to Theorem 3.1. The proposed estimator equals to the kernel regression when $\left| \widehat{J}_1(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \right|$ is far from 0, and equals to $\widehat{\rho}^{-1}$, when $\left| \widehat{J}_1(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \right|$ is close to 0. The term c_n is used to trim possible extremities of the kernel regression estimates.

After obtaining the estimator $\widehat{\pi}^{-1}$ for $\pi_*^{-1}(\widetilde{\boldsymbol{X}})$, we construct the estimating equation by incorporating $\widehat{\pi}^{-1}$, i.e., $\widehat{E}_n \left[\left\{ S(\boldsymbol{\beta}; \widehat{Q}, \widehat{\pi}) \right\}^\top \widehat{\boldsymbol{v}} \right]$ with a constraint $\boldsymbol{\beta}_{-1} = \widehat{\boldsymbol{\beta}}_{-1}$. To avoid possible computational issues in solving this estimating equation (Chapter 5 in Van der Vaart [37]), we use its first-order expansion and construct a one-step debiased estimator $\widetilde{\beta}_1 = \widehat{\beta}_1 - \bar{I}^{-1} S$, where $S = \widehat{E}_n \left[\left\{ S(\widehat{\boldsymbol{\beta}}; \widehat{Q}, \widehat{\pi}) \right\}^\top \widehat{\boldsymbol{v}} \right]$, and $\bar{I} = \widehat{E}_n \left[b''(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}) X_1 \boldsymbol{X}^\top \widehat{\boldsymbol{v}} \right]$. Another challenge in constructing the debiased estimator is that the estimation errors of $\widehat{Q}$, $\widehat{\pi}$ and the samples used to construct the estimator are correlated. We adopt the cross-fitting procedure proposed in [4] in the implementation.

2.3. Implementation

The entire procedure can be separated into two steps. In the first step, using all fully observed data, we obtain $\widehat{\boldsymbol{\Gamma}}$ and then regress Y on $\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}$ using kernel regressions and denote the estimated link function as $\widehat{g}$. Using the estimated imputation model, $\widehat{Q}(Z, \boldsymbol{X}) = \widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})$, we obtain an initial estimate $\widehat{\boldsymbol{\beta}}$ by solving

$$\widehat{E}_n \left[\left\{ b'(\boldsymbol{X}^\top \boldsymbol{\beta}) - \widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \right\} \boldsymbol{X} \right] = 0,$$

which is equivalent to fitting a working generalized linear model for $\widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})$ using $\boldsymbol{X}$. Using the initial estimate, we solve

$$\min_{\boldsymbol{w}} \widehat{E}_n \left[b''(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}})(X_1 - \boldsymbol{X}_{-1}^\top \boldsymbol{w})^2 \right],$$

and denote its minimizer as $\widehat{\boldsymbol{w}}$. Then we can construct $\widehat{\boldsymbol{v}}^\top = (1, -\widehat{\boldsymbol{w}}^\top)$. In the second step, we estimate the identified weighting function and use it to

form a one-step debiased estimator for β_1^*. First, we split the data into K subsets $(I_1, \ldots, I_K)$ with equal sample sizes. The choice of K does not affect the theoretical results; in our simulation and real data analysis, for simplicity of the computation, we set $K = 2$. For a specific index k, the estimated link function, denoted as $\widehat{g}_{(-k)}(\cdot)$, is obtained through kernel regression of Y w.r.t. $\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}$ using the data excluding I_k. The estimated weighting function denoted as $\widehat{\pi}^{-1}_{(-k)}(\widetilde{\boldsymbol{X}}; \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{v}})$ is obtained through the truncated kernel regression using the data excluding I_k. Specifically,

$$\widehat{\pi}^{-1}_{(-k)}(\widetilde{\boldsymbol{x}}; \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{v}}) = \begin{cases} 1 + \left\{ \widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \widehat{\rho} \right\}^{-1} \widehat{J}_0^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}})(1 - \widehat{\rho}) & \left| \widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \right| > c_n, \\ \widehat{\rho}^{-1}, & \left| \widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \right| \leqslant c_n, \end{cases}$$

where $\widehat{J}_r^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) = \widehat{E}_n^{(-k)} \left[\boldsymbol{X}^\top \widehat{\boldsymbol{v}} K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \mid R = r \right]$ and $\widehat{E}_n^{(-k)}[\cdot \mid R = r]$ is the empirical average over the samples with $R = r$ and excluding those in I_k. In theory, the value of c_n can be chosen following Theorem 3.1; in our implementation, we choose $c_n = 0.01$ to trim extreme estimates. Then, the one-step debiased estimator is $\widetilde{\beta}_1 = \widehat{\beta}_1 - \bar{I}^{-1} \bar{S}$, where

$$\bar{S} = \sum_{k=1}^{K} S^{(k)}/K, \quad S^{(k)} = \widehat{E}_n^{(k)} \left[\left\{ S(\widehat{\boldsymbol{\beta}}; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)}) \right\}^\top \widehat{\boldsymbol{v}} \right].$$

A summary of the entire algorithm can be found in the Appendix A.3. In our simulation and real data analysis, for kernel regressions obtaining $\widehat{g}(\cdot)$ and $\widehat{J}_r^{(-k)}$, we adopt the Gaussian kernel function, which is a symmetric kernel function of order 1 (i.e., $\nu = 2$); to determine the bandwidth, we specify the bandwidth $\hbar$ as the theoretical optimal bandwidth to minimize the l_2-estimation errors, i.e., $\hbar_{\text{opt}} = (n\rho)^{-1/(2\nu+d)}$. In practice, we use $(n\widehat{\rho})^{-1/(2\nu+d)}$, where $n\widehat{\rho}$ is the sample size of fully observed data.

To estimate the asymptotic variance of $\widetilde{\beta}_1$, following [20], we bootstrap based on the entire sample for B times; for the bth bootstrapped dataset, we implement the algorithm and obtain $\widetilde{\beta}_1^{(b)}$, where $b = 1, \cdots, B$. We use the sample variance of $\left\{ \widetilde{\beta}_1^{(b)} \right\}_{b=1}^{B}$ as the estimate for the asymptotic variance to construct interval estimations. The confidence interval is constructed as follows: $\widetilde{\beta}_1 \pm z_{1-\alpha/2} \times$ standard deviation of $\left\{ \widetilde{\beta}_1^{(b)} \right\}_{b=1}^{B}$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution, and $1 - \alpha$ is the pre-specified nominal coverage. In our simulation and real data analysis, we choose $B = 500$.

Remark 2.6. To generalize the proposed procedure to infer $\boldsymbol{u}^\top \boldsymbol{\beta}^*$ for any $\boldsymbol{u}$, we modify the construction of $\widehat{\boldsymbol{v}}$. To construct $\widehat{\boldsymbol{v}}$, we calculate

$$\widehat{\boldsymbol{v}} = \left\{ \widehat{E}_n \left[b''(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}) \boldsymbol{X} \boldsymbol{X}^\top \right] \right\}^{-1} \boldsymbol{u}.$$

When constructing the one-step debiased estimator, we will use $\boldsymbol{u}^\top\widehat{\boldsymbol{\beta}} - \bar{I}^{-1}\bar{S}$, where

$$\bar{I} = \widehat{E}_n\left[\left(\left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q},\widehat{\pi})\right\}^\top \widehat{\boldsymbol{v}}\right)^2\right].$$

This modification is based on the proof of Theorem A.1. In the proof, we shown that when $\boldsymbol{u} = (1,0,\cdots,0)^\top$, $\boldsymbol{v}^\top$ is proportional to $(1,-(\boldsymbol{w}^*)^\top)$, where

$$\boldsymbol{v} = \left\{E\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^\top\right]\right\}^{-1}\boldsymbol{u}.$$

In general cases, it de-correlates the estimation error of $\boldsymbol{u}^\top\widehat{\boldsymbol{\beta}}$ w.r.t. other orthogonal complements. Thus, this procedure is an extension of the de-correlated score for an arbitrary $\boldsymbol{u}$.

3. Theoretical properties

In this section, we provide the asymptotic property of the proposed estimator. To accommodate the situation where the marginal missing rate may be close to 1, we assume that the distribution of $(\boldsymbol{X}, Z)$ and the conditional distribution $Y \mid Z, \boldsymbol{X}$ do not depend on n; the missing propensity $P(R=1 \mid Z, \boldsymbol{X})$ may depend on n. Specifically, we consider two scenarios: 1) the missing propensity $P(R=1 \mid Z, \boldsymbol{X})$ does not change with n; 2) $P(R=1 \mid Z, \boldsymbol{X}) = \rho_n w(\widetilde{\boldsymbol{X}})$ with $\rho_n \to 0$, where $w(\widetilde{\boldsymbol{X}})$ does not depend on n, $w(\widetilde{\boldsymbol{X}})$ is always bounded away from ∞, and $E\left[w(\widetilde{\boldsymbol{X}})\right] = 1$. For both scenarios, we assume the relaxed positivity assumption that $P(R=1 \mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) > 0$ rather than the traditional positivity assumption that $P(R=1 \mid Z, \boldsymbol{X}) > 0$, which is a benefit of not using the inverse of the true propensity $P(R=1 \mid Z, \boldsymbol{X})$ in the estimation. For brevity, we focus on the theoretical results for Scenario 1) in the main text and leave those for Scenario 2) in Appendix A.4. All the proofs of the theorems can also be found in Appendix A.4.

For Scenario 1), the following assumptions are required.

Assumption 3.1. The covariate $\boldsymbol{X}$'s and the surrogate outcome Z are bounded, and the function $b''(\cdot)$ is continuously differentiable; $\max\{\|\boldsymbol{\beta}^*\|_2, \|\boldsymbol{v}\|_2\}$ is bounded.

Assumption 3.2. There is a positive constant $\gamma_d > 1/4$ such that

$$\|\widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - Q(Z,\boldsymbol{X})\|_\infty = O_p(n^{-\gamma_d}),$$

and

$$\left\{\mathrm{vec}(\widehat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma})\right\} = n^{-1}\sum_{i=1}^{n} 1\{R_i=1\}\psi(\boldsymbol{X}_i, Z_i, Y_i) + o_p(n^{-1/2}),$$

where $\psi(\boldsymbol{X}_i, Z_i, Y_i)$ is bounded with mean 0 and $\mathrm{vec}(\cdot)$ represents the vectorization of the matrix. In addition, we assume that

$$\sup_{\widetilde{\boldsymbol{x}}}\left|(\widehat{g}-g)(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{x}}) - (\widehat{g}-g)(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}})\right| = o_p(n^{-1/2}).$$

Assumption 3.3. The function $K(\cdot)$ is a kernel function with the order of $\nu-1$. Function $J_r(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}})$'s are at least νth order differentiable w.r.t $\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}}$ with bounded derivatives. Define $G(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}}) = J_1^{-1}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}})J_0(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}})$. We assume that $G(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}})$ is bounded away from $+\infty$ on the open set $\left\{\widetilde{\boldsymbol{x}} : J_1(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}}) \neq 0\right\}$. We also assume that for any matrix $\boldsymbol{A}$ of full rank, the density function of $\boldsymbol{A}^\top\widetilde{\boldsymbol{X}}$, $\eta(\boldsymbol{A}^\top\widetilde{\boldsymbol{X}})$, is bounded away from 0 and $+\infty$, and at least νth order differentiable with bounded derivatives.

Assumption 3.4. When $t > 0$ is small enough, there exist positive constants A_0 and γ_m such that $P\left\{0 \neq \left|E[\boldsymbol{X}^\top\boldsymbol{v} \mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}, R=1]\right| \leqslant t\right\} \leqslant A_0 t^{\gamma_m}$.

Assumption 3.5. Set $c_n = \widetilde{\delta}_n^{2/(2+\gamma_m)}$, where $\widetilde{\delta}_n = (n\hbar^d/\log n)^{-1/2} + \hbar^\nu + n^{-\gamma_d}$. We assume that $n^{-\gamma_d}\widetilde{\delta}_n^{\gamma_m/(2+\gamma_m)} = o(n^{-1/2})$.

In Assumption 3.1, for ease of exposition, we assume a bounded design for each $\boldsymbol{X}$ and Z. Assumption 3.2 includes the requirement for the chosen dimension reduction method to estimate $\boldsymbol{\Gamma}$ and the chosen method to estimate g. Specifically, we assume that the uniform convergence rate of $\widehat{Q}$ is $O_p(n^{-\gamma_d})$, and the estimated subspace $\widehat{\boldsymbol{\Gamma}}$ is asymptotically linear. In addition, we assume that $\sup_{\widetilde{\boldsymbol{x}}}\left|(\widehat{g}-g)(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{x}}) - (\widehat{g}-g)(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}})\right| = o_p(n^{-1/2})$. Many dimension reduction methods and nonparametric methods satisfy Assumption 3.2. An example is given in Appendix A.5. Assumption 3.3 requires $J_r(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}})$ and $\eta(\boldsymbol{A}^\top\widetilde{\boldsymbol{X}})$ to be at least ν-th order differentiable. These requirements guarantee the convergence rates of the kernel regressions using a kernel function of order $\nu-1$ to estimate $J_r(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{x}})$ and $\eta(\boldsymbol{A}^\top\widetilde{\boldsymbol{X}})$. Assumption 3.4 restricts the concentration near $E[\boldsymbol{X}^\top\boldsymbol{v} \mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}, R=1] = 0$ by the parameter γ_m. When $E[\boldsymbol{X}^\top\boldsymbol{v} \mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}, R=1]$ is bounded away from 0, we have $\gamma_m = +\infty$; when $E[\boldsymbol{X}^\top\boldsymbol{v} \mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}, R=1]$ has a continuous distribution, we have $\gamma_m \geqslant 1$. Assumption 3.5 specifies the condition on γ_d and γ_m. In the following, we provide an example on the values of γ_d and γ_m. When $\boldsymbol{X} \mid R=1$ follows a multi-variate Gaussian distribution, $E[\boldsymbol{X}^\top\boldsymbol{v} \mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}, R=1]$ is continuous, and thus we have $\gamma_m = 1$. Further, if we choose the optimal $\hbar = n^{-1/(2\nu+d)}$, we have that $\widetilde{\delta}_n = n^{-\nu/(2\nu+d)} + n^{-\gamma_d}$, and $\gamma_d = \nu/(2\nu+d)$. Assumption 3.5 can be satisfied if $d < 2\nu/3$.

Under these assumptions, we show that the one-step debiased estimator is asymptotically normal.

Theorem 3.1. *Under Assumptions 3.1–3.5, we have* $\sqrt{n}(\widetilde{\beta}_1 - \beta_1^*) \to N(0, \sigma^2)$, *where the formula of the asymptotic variance* σ^2 *can be found in the Appendix A.4.*

In addition, Corollary 3.1 shows a sufficient condition that the semiparametric lower bound can be achieved.

Corollary 3.1. *When the true propensity* $P(R=1 \mid Z, \boldsymbol{X}) = P(R=1 \mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})$, *the one-step debiased estimator* $\widetilde{\beta}_1$ *obtains the semiparametric lower bound.*

When $P(R=1 \mid Z, \boldsymbol{X}) = P(R=1 \mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})$, we can show that $\pi_*^{-1}(\widetilde{X}) = P^{-1}(R=1 \mid Z, \boldsymbol{X})$, which is the inverse of the true propensity score. Thus,

the proposed method will lead to an estimator obtaining the semiparametric lower bound. In general cases where $P(R = 1 \mid Z, \boldsymbol{X}) \neq P(R = 1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$, the proposed estimator cannot obtain the semiparametric lower bound. Although the asymptotic variance of the proposed method may be higher than the semiparametric lower bound when $P(R = 1 \mid Z, \boldsymbol{X}) \neq P(R = 1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$, the proposed method may still show better performance in finite sample cases since low-dimensional imputation model and weighting function are easier to estimate.

4. Simulations

In this section, we conduct simulations and compare the proposed method with other methods to demonstrate 1) the advantage of avoiding complex propensity estimation and 2) the possible efficiency gain from incorporating the surrogate outcome. To show the advantage of avoiding modeling the complex propensity, we compare our proposed method with two baseline approaches. Baseline 1 follows the double machine learning procedure proposed in [4], which estimates both the propensity and the imputation model using kernel regressions. When using the kernel regressions to estimate the propensity and the imputation model, we first implement dimension reduction and then conduct the kernel regression. Another baseline approach (Baseline 2) executes the same procedure as Baseline 1 but employs a logistic regression to estimate the missing propensity. For both Baselines 1 and 2, we implement a threshold of 0.01 for the estimated propensities to trim extreme values. To show possible efficiency gain from incorporating the surrogate outcome, besides the proposed procedure using the surrogate outcome (denoted as "Proposed with Z"), we implement another approach (denoted as "Proposed w/o Z") following the same procedure but only using $\boldsymbol{X}$ (no surrogate outcome Z) in the dimension reduction, imputation model estimation, and weighting function estimation. For the dimension reduction step implemented in all these approaches, we use the kernel sliced regression method and choose the reduced dimension using cross-validation.

We consider 8 simulation scenarios with different missing rates, sample sizes, and outcome formats, i.e., continuous and binary outcomes. For each type of outcome, we consider a moderate marginal missing rate of 50% and a high marginal missing rate of 90%. For both scenarios, we choose the sample size to be 500 or 1000. To generate data under each design, we first generate the missing indicator R following a Bernoulli distribution with the success probability of 0.5 (moderate marginal missing rate) or 0.1 (high marginal missing rate). Then, we generate the covariates based on $R = 1$ or $R = 0$. When the outcome is missing ($R = 0$), the covariates, i.e., $\boldsymbol{X} \mid R = 0$, follow a standard multivariate Gaussian distribution with zero means and the identity covariance matrix; when the outcome is observed ($R = 1$), the covariates follow a mixture of two multivariate Gaussian distributions. With a probability of 0.7, the covariates are generated following a standard multivariate Gaussian distribution; otherwise, the covariates are generated following a multivariate Gaussian distribution $N(1, 1.5\boldsymbol{I}_p)$, i.e., $\boldsymbol{X} \mid R = 1 \sim \xi N(0, I_p) + (1 - \xi)N(1, 1.5\boldsymbol{I}_p)$, where ξ follows a Bernoulli

distribution with a success probability of 0.7. The surrogate outcome is generated from $Z = \delta \boldsymbol{X}^\top \boldsymbol{\beta}_0 + \sum_{j=5}^{8} |X_j|/4 + \epsilon_z$ where $\epsilon_z \sim N(0,1)$. To generate the outcome Y given the covariates, for the scenario with continuous outcomes, we consider $Y = \boldsymbol{X}^\top \boldsymbol{\beta}_0/2 + Z + \epsilon$, where $\boldsymbol{\beta}_0 = (1,1,-1,-1,0,0,0,0)^\top$ and $\epsilon \sim N(0,1)$. For binary outcomes, we consider $Y = 1\{\boldsymbol{X}^\top \boldsymbol{\beta}_0/4 + Z + \epsilon > 0\}$. In addition, δ is fixed at 0.5 for continuous outcomes and 0.25 for binary outcomes.

To evaluate the proposed methods, we use metrics including coverage, bias, standard deviation, and deviance of the estimates. For each scenario, we run 500 replicates. For each replicate, we estimate the coefficients and use bootstrapping (500 bootstraps) to construct a 95%-confidence interval using the training samples. We obtain the mean for each coefficient over 500 replicates to calculate the bias and obtain the standard deviation for each coefficient over 500 replicates. We also calculate the deviance using the coefficients estimated by each approach on an independently generated testing dataset with a sample size of 10^4.

Tables 1, 2, and 3 exhibit the coverage, bias, and standard deviation of the first four coefficients, respectively. From Table 1, overall, the proposed methods (Proposed w/o and with Z) achieve the nominal coverage in most scenarios even in the high missing rate settings, whereas Baseline 1 and Baseline 2 do not. This suggests both Baseline 1 and Baseline 2 are incapable of handling the complex missing mechanism. In the scenario with binary outcomes, for both proposed methods, there might be over-coverage issues, especially when the number of fully observed samples is below 250. When the number of fully observed samples increases, the coverage of both proposed methods tends to converge to the nominal coverage as the finite sample bias diminishes. From Table 2, our proposed method with Z yields lower bias compared with Baseline 1 and the proposed method without Z. From Table 3, our proposed method with Z provides smaller standard deviations than the proposed method without Z. Baseline 2 may achieve a relatively low bias/standard deviation due to a more stable estimation of the weights using logistic regressions, especially in the scenario $n = 500$ and with moderate missing rates. Figure 1 summarizes the deviance of the estimates for all scenarios. With the moderate missing rate, our proposed method with Z performs comparably to Baseline 2 but outperforms the other methods. In the high missing rate scenarios, Proposed with Z achieves the minimum deviance. Our proposed method with Z surpasses the method without Z in all the scenarios due to the reduced finite sample bias of the low-dimensional imputation model and weighting estimation induced by the surrogate Z. In summary, considering all the metrics, the proposed method with Z outperforms others.

As discussed in Section 2.1, machine learning methods, instead of kernel regression, can be used for the imputation model. The simulation results using random forests to fit the imputation model can be found in Appendix A.7.

5. Application to PROMIS global physical health T-score

In this section, we applied our proposed method to predict whether the improvement of the PROMIS global physical health T-score will exceed the MCID after

TABLE 1
Coverage of the 95% confidence interval for the coefficients.

	$n = 500$				$n = 1000$			
	Continuous, Missing rate of 50%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	0.872	0.840	0.886	0.864	0.847	0.769	0.861	0.817
Baseline 2	0.896	0.912	0.906	0.928	0.885	0.873	0.893	0.875
Proposed w/o Z	0.920	0.938	0.910	0.922	0.901	0.901	0.891	0.893
Proposed with Z	0.952	0.964	0.958	0.968	0.954	0.954	0.956	0.954
	Continuous, Missing rate of 90%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	0.866	0.840	0.844	0.854	0.740	0.798	0.752	0.762
Baseline 2	0.904	0.876	0.922	0.920	0.910	0.898	0.894	0.902
Proposed w/o Z	0.954	0.966	0.968	0.946	0.942	0.944	0.938	0.952
Proposed with Z	0.946	0.968	0.966	0.962	0.946	0.964	0.954	0.956
	Binary, Missing rate of 50%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	0.908	0.918	0.916	0.932	0.892	0.882	0.918	0.900
Baseline 2	0.928	0.932	0.916	0.930	0.920	0.898	0.894	0.912
Proposed w/o Z	0.968	0.970	0.956	0.950	0.962	0.964	0.964	0.964
Proposed with Z	0.970	0.972	0.948	0.962	0.958	0.968	0.956	0.946
	Binary, Missing rate of 90%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	0.950	0.940	0.948	0.962	0.892	0.942	0.964	0.940
Baseline 2	0.932	0.930	0.920	0.956	0.924	0.912	0.896	0.926
Proposed w/o Z	0.956	0.950	0.948	0.964	0.954	0.962	0.944	0.950
Proposed with Z	0.966	0.966	0.976	0.948	0.942	0.962	0.966	0.974

TABLE 2
Bias of coefficient estimates.

	n = 500				n= 1000			
	Continuous, Missing rate of 50%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	1.635	1.543	-1.404	-1.451	1.064	1.151	-0.997	-1.078
Baseline 2	0.070	0.054	-0.049	-0.067	0.057	0.055	-0.040	-0.054
Proposed w/o Z	0.694	0.693	-0.883	-0.926	0.554	0.635	-0.689	-0.656
Proposed with Z	0.097	0.072	-0.129	-0.087	0.054	0.072	-0.076	-0.068
	Continuous, Missing rate of 90%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	2.394	2.262	-2.374	-2.234	1.989	1.845	-1.915	-1.881
Baseline 2	0.074	0.141	-0.079	-0.108	0.171	0.193	-0.189	-0.131
Proposed w/o Z	-0.082	-0.103	0.125	0.105	-0.069	-0.049	0.028	0.041
Proposed with Z	0.056	0.050	-0.096	0.006	0.046	0.036	-0.084	-0.060
	Binary, Missing rate of 50%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	1.176	1.099	-1.123	-1.008	0.857	0.864	-0.809	-0.766
Baseline 2	0.048	0.031	-0.039	-0.037	0.023	0.027	-0.016	-0.035
Proposed w/o Z	0.288	0.318	-0.208	-0.299	0.261	0.191	-0.199	-0.226
Proposed with Z	0.007	0.000	-0.020	-0.016	-0.005	0.000	0.020	-0.006
	Binary, Missing rate of 90%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	9.457	9.361	-2.557	-2.440	7.875	6.110	-4.685	-4.458
Baseline 2	-0.034	-0.096	-0.218	-0.213	0.054	0.021	-0.243	-0.135
Proposed w/o Z	0.199	0.431	0.290	0.483	-0.054	0.009	0.029	0.084
Proposed with Z	-0.244	-0.218	0.214	0.191	-0.076	-0.155	0.108	0.113

TABLE 3
Standard deviations of coefficient estimates.

	n = 500				n= 1000			
	Continuous, Missing rate of 50%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	1.535	1.694	1.626	1.698	0.949	1.000	0.954	0.987
Baseline 2	0.176	0.173	0.179	0.171	0.118	0.120	0.120	0.123
Proposed w/o Z	1.312	1.306	1.324	1.263	0.759	0.761	0.790	0.772
Proposed with Z	0.423	0.414	0.452	0.390	0.226	0.257	0.255	0.254
	Continuous, Missing rate of 90%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	2.368	2.441	2.523	2.447	1.534	1.543	1.588	1.423
Baseline 2	0.698	0.724	0.692	0.644	0.522	0.553	0.513	0.521
Proposed w/o Z	1.038	0.968	0.869	0.965	0.617	0.566	0.653	0.597
Proposed with Z	0.726	0.640	0.709	0.623	0.404	0.338	0.418	0.378
	Binary, Missing rate of 50%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	2.022	1.961	1.981	2.009	1.316	1.355	1.219	1.200
Baseline 2	0.237	0.255	0.228	0.219	0.180	0.186	0.173	0.159
Proposed w/o Z	1.368	1.443	1.397	1.441	0.872	0.813	0.785	0.772
Proposed with Z	0.683	0.653	0.540	0.507	0.474	0.480	0.386	0.365
	Binary, Missing rate of 90%							
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Baseline 1	17.967	18.218	19.511	17.679	12.282	12.271	12.390	12.162
Baseline 2	1.489	1.479	1.348	1.272	1.340	1.271	1.067	1.003
Proposed w/o Z	3.233	3.307	3.142	2.397	1.669	1.661	1.688	1.540
Proposed with Z	0.704	0.869	0.478	0.460	0.672	0.640	0.385	0.312

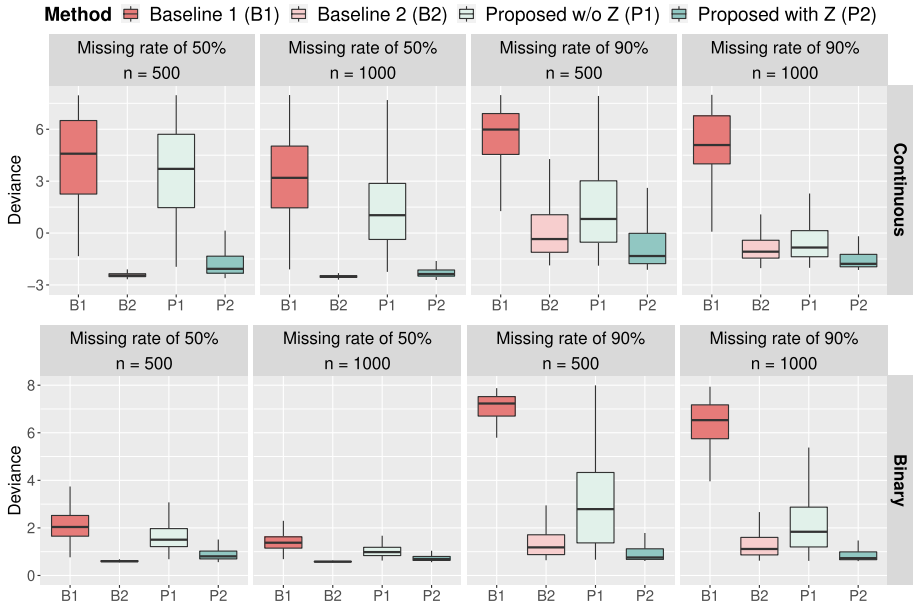

FIG 1. *Deviance under different missing rates, sample sizes, and outcome types.*

receiving total joint replacement using the information obtained before scheduling the surgery. In addition to making predictions, we also aimed to identify the driving factors of not achieving the MCID in the presence of highly incomplete outcomes. The dataset includes 1044 University of Florida Health patients who participated in the pre-surgical survey and underwent total joint replacement surgery. In the analysis, we incorporated many baseline covariates, including demographics, socioeconomic characteristics, medical history, and care characteristics before surgery (e.g., 30 days before admission for surgery). According to the convention of constructing MCID, the MCID is obtained based on the one-half standard deviation of the difference between the pre-and post-surgical PROMIS global health T-scores [7, 17].

For our data, the difference between the pre-and post-surgical scores has an average of 9.5 and a standard deviation of 8.1, and consequently, the MCID is 4.1. Thus, the outcome $Y = 1$ if the difference is less than 4.1, and $Y = 0$, otherwise. In terms of the missing proportion, all the 1044 identified patients took the pre-surgical survey, but only 261 patients (25%) responded to the post-surgical survey (the missing rate is 75%). Of the patients who took both surveys, 67 (25.7%) patients did not meet the MCID. Since the PROMIS global physical T-score is derived from the ten survey questionnaire items, individual items can be considered as candidates for the surrogate outcome. To construct an informative surrogate, we relied on actual data and fitted the outcome Y using the pre-surgical survey responses on the fully observed data. Then, we predicted the outcome Y using the pre-surgical survey responses and used the predicted values as a single informative surrogate outcome.

We conducted two analyses to investigate the performance of the proposed method. In the first analysis, we compared the proposed methods with the baseline methods in terms of the deviance $E\big[\ell(\widehat{\boldsymbol{\beta}})\big]$, where $\ell(\widehat{\boldsymbol{\beta}}) = b(\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}) - \widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}$. Specifically, we randomly split the entire dataset into a training dataset and a testing dataset with equal sample sizes. We estimated coefficients using the training dataset, and then calculated the deviance on the testing dataset. This procedure was repeated 1000 times. In the second analysis, we fitted the model on the entire dataset and compared the variables selected by different methods.

Table 4 shows that the proposed method with Z achieves the lowest deviance. In terms of the selected variables, compared with other methods, the proposed method with Z uniquely revealed that geriatric patients were less likely to achieve the MCID (estimated coefficient is 0.367; 95%-CI is $[0.020, 0.714]$). This is in accordance with the existing research finding that elderly patients were more likely to have post-operative adverse clinical outcomes than younger patients in total joint replacement [11, 23]. The coefficients with confidence intervals for other covariates are presented in the Appendix A.8.

TABLE 4
Comparison of averaged deviances (standard deviations) in real data example.

Baseline1	Baseline2	Proposed w/o Z	Proposed w/ Z
0.714 (0.105)	0.696 (0.094)	0.685 (0.091)	0.670 (0.094)

6. Discussion

In this work, we propose a debias approach to estimating the parameters of interest under a possibly misspecified GLM. This approach uses an informative surrogate outcome that leads to a low-dimensional flexible imputation model and estimates a low-dimensional weighting function instead of the complex propensity score. When the true propensity happens to enjoy the same low-dimensional structure, the proposed method achieves the semi-parametric efficiency lower bound. Compared with the double machine learning method, the proposed approach relaxes the requirement on the propensity estimation and maintains almost the same flexibility or requirement on the imputation model estimation. In addition, we only require a relaxed positivity assumption.

There are multiple future directions to extend the proposed approach. First, we can consider extending the proposed approach to high-dimensional settings where $p/n \to +\infty$. In a high-dimensional setting, $\widehat{\mathbf{\Gamma}}$ may not be asymptotic normal due to the possible penalization. In this case, in order to achieve an asymptotic normal estimator, an additional debias procedure is needed to adjust for the bias due to the estimation error of $\widehat{\mathbf{\Gamma}}$. Second, we can investigate more choices of the function h. In this work, to pursue a low-dimensional weighting function, we choose h such that $\mathcal{T}h = 0$ when constructing the weighting function. However, we can choose other potential alternatives to mitigate certain deficiencies, such as minimizing the asymptotic variance of the debiased estimator or avoiding possible negative weights. Third, we can combine the proposed approach with the augmented minimax linear estimation [12] to avoid the computation of the Riesz representer. In the high-dimensional setting, an explicit form of the weights to debias $\widehat{\mathbf{\Gamma}}$ might be intractable.

Appendix A: Appendix

Appendix A contains the derivation of all the theorems and the variables selected in the real data study.

A.1. Semiparametric lower bound with and without using surrogate outcome

In this section, we derive the semiparametric lower bound (i.e., the efficient influence function) for estimating β_1^* with or without surrogate outcomes under two scenarios: 1) the distribution of $(\boldsymbol{X}, Z)$ is known; 2) the distribution of $(\boldsymbol{X}, Z)$ is unknown.

Theorem A.1. *Assuming $(Y, Z) \perp R \mid \boldsymbol{X}$, regardless of the distribution of $(\boldsymbol{X}, Z)$ is known or not, a semiparametric efficient estimator for β_1^* obtained using the surrogate outcome Z is more efficient than that using only $\boldsymbol{X}$ if $E[Y \mid \boldsymbol{X}] \neq E[Y \mid Z, \boldsymbol{X}]$, where β_1^* is the first coordinate of β^*.*

Proof. To show this result, we derive the semiparametric lower bounds with surrogate outcome Z assuming the distribution of $(\boldsymbol{X}, Z)$ is 1) known; 2) unknown. To start with, we present the likelihood of the full data.

$$\eta_{\widetilde{\boldsymbol{X}}}(\widetilde{\boldsymbol{X}})\left\{1-\pi(\widetilde{\boldsymbol{X}})\right\}^{1-R}\left\{\pi(\widetilde{\boldsymbol{X}})\eta_Y(Y,\widetilde{\boldsymbol{X}})\right\}^{R},$$

where $\eta_{\widetilde{\boldsymbol{X}}}$ is the density of $\widetilde{\boldsymbol{X}}$, and $\eta_{\widetilde{Y}}$ is the density of $Y \mid \widetilde{\boldsymbol{X}}$. Under the assumption $Y \perp R \mid (\boldsymbol{X}, Z)$, the definition of β_1^* does not depend on $\pi_R(\boldsymbol{X})$, and thus the form of the semiparametric lower bounds does not depend on whether $\pi_R(\boldsymbol{X})$ is known or not. Thus, we will derive the semiparametric lower bounds assuming $\pi_R(\boldsymbol{X})$ is known.

First, we characterize the nuisance tangent space generated by $\eta_{\widetilde{\boldsymbol{X}}}$ and $\eta_{\widetilde{Y}}$.

$$\begin{aligned}\Lambda_{\widetilde{\boldsymbol{X}}} &= \left\{f(\widetilde{\boldsymbol{X}}) : E\left[f(\widetilde{\boldsymbol{X}})\right]=0\right\},\\ \Lambda_Y &= \left\{Rf(Y,\widetilde{\boldsymbol{X}}) : E\left[f(Y,\widetilde{\boldsymbol{X}}) \mid \widetilde{\boldsymbol{X}}\right]=0\right\}.\end{aligned}$$

Let $\Lambda_\pi = \left\{\left[R-\pi(\widetilde{\boldsymbol{X}})\right]h(\widetilde{\boldsymbol{X}}) : \forall h \in L_2(\widetilde{\boldsymbol{X}})\right\}$. By the decomposition of the likelihood, we have

$$\mathcal{H} = \Lambda_{\widetilde{\boldsymbol{X}}} \oplus \Lambda_\pi \oplus \Lambda_Y.$$

Second, we find an influence function of β_1^*. [15] verified that

$$\phi = \pi^{-1}(\widetilde{\boldsymbol{X}})R\{Y - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\}\boldsymbol{X}^\top\tilde{\boldsymbol{u}}_0$$

is an influence function of β_1^*, where $\tilde{\boldsymbol{u}}_0$ is defined as

$$\left\{E\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^\top\right]\right\}^{-1}(1,0,\cdots,0)^\top.$$

Finally, denote the nuisance tangent space as Λ. We derive the efficient influence function by subtracting its projection onto $\Lambda^\perp$ from ϕ. When the distribution of $(\boldsymbol{X}, Z)$ is unknown, we have $\Lambda = \Lambda_{\widetilde{\boldsymbol{X}}} \oplus \Lambda_Y$. Thus, $\Lambda^\perp = \Lambda_\pi$. To derive the projection of ϕ onto $\Lambda^\perp$, we solve the following equation for h,

$$E\left[\left\{\phi - \left[R-\pi(\widetilde{\boldsymbol{X}})\right]h(\widetilde{\boldsymbol{X}})\right\}\left[R-\pi(\widetilde{\boldsymbol{X}})\right]\tilde{h}(\widetilde{\boldsymbol{X}})\right]=0, \forall \tilde{h} \in L_2(\widetilde{\boldsymbol{X}}).$$

This equation is equivalent to

$$E\left[\left\{\phi - \left[R-\pi(\widetilde{\boldsymbol{X}})\right]h(\widetilde{\boldsymbol{X}})\right\}\left[R-\pi(\widetilde{\boldsymbol{X}})\right] \mid \widetilde{\boldsymbol{X}}\right]=0.$$

Solving this equation, we have $h(\widetilde{\boldsymbol{X}}) = \pi^{-1}(\widetilde{\boldsymbol{X}})\left\{E\left[Y \mid \widetilde{\boldsymbol{X}}\right] - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\right\}\boldsymbol{X}^\top\tilde{\boldsymbol{u}}_0$. Thus, when the distribution of $(\boldsymbol{X}, Z)$ is unknown and the surrogate outcome Z is available, the efficient influence function is

$$\begin{aligned}\phi_{1,Z} &= \pi^{-1}(\widetilde{\boldsymbol{X}})R\{Y - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\}\boldsymbol{X}^\top\tilde{\boldsymbol{u}}_0\\ &\quad -\pi^{-1}(\widetilde{\boldsymbol{X}})\left\{R-\pi(\widetilde{\boldsymbol{X}})\right\}\left\{E\left[Y \mid \widetilde{\boldsymbol{X}}\right] - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\right\}\boldsymbol{X}^\top\tilde{\boldsymbol{u}}_0.\end{aligned}$$

The semiparametric lower bound of estimating β_1^* is given by $\mathrm{var}(\widetilde{\phi}_{1,Z})$.

Now, we compare the semiparametric lower bound with and w/o the surrogate outcome Z when the distribution of $(\boldsymbol{X}, Z)$ is unknown. Under the assumption that $R \perp Z \mid \boldsymbol{X}$, we have

$$\begin{aligned}\widetilde{\phi}_{1,Z} &= \pi^{-1}(\boldsymbol{X})R\{Y - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\}\boldsymbol{X}^\top\widetilde{\boldsymbol{u}}_0 \\ &\quad -\pi^{-1}(\boldsymbol{X})\left\{R - \pi(\boldsymbol{X})\right\}\left\{E\left[Y \mid \widetilde{\boldsymbol{X}}\right] - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\right\}\boldsymbol{X}^\top\widetilde{\boldsymbol{u}}_0.\end{aligned}$$

Likewise, we can derive the efficient influence function without the surrogate outcome Z, i.e.,

$$\begin{aligned}\phi_1 &= \pi^{-1}(\widetilde{\boldsymbol{X}})R\{Y - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\}\boldsymbol{X}^\top\widetilde{\boldsymbol{u}}_0 \\ &\quad -\pi^{-1}(\widetilde{\boldsymbol{X}})\left\{R - \pi(\widetilde{\boldsymbol{X}})\right\}\left\{E\left[Y \mid \boldsymbol{X}\right] - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\right\}\boldsymbol{X}^\top\widetilde{\boldsymbol{u}}_0.\end{aligned}$$

The difference $\phi_1 - \widetilde{\phi}_{1,Z} = \pi^{-1}(\boldsymbol{X})\left\{R - \pi(\boldsymbol{X})\right\}\left\{E\left[Y \mid \widetilde{\boldsymbol{X}}\right] - E\left[Y \mid \boldsymbol{X}\right]\right\}\boldsymbol{X}^\top\widetilde{\boldsymbol{u}}_0 \in \Lambda_\pi$ and $\widetilde{\phi}_{1,Z} \in \Lambda_\pi^\perp$. Thus, we have $\mathrm{var}(\phi_1) \geqslant \mathrm{var}(\widetilde{\phi}_{1,Z})$ and the equality holds if and only if $E\left[Y \mid \widetilde{\boldsymbol{X}}\right] = E\left[Y \mid \boldsymbol{X}\right]$.

When the distribution of $(\boldsymbol{X}, Z)$ is known, the nuisance tangent space is $\Lambda = \Lambda_Y$. Thus, we have $\Lambda^\perp = \Lambda_\pi \oplus \Lambda_{\widetilde{\boldsymbol{X}}}$. To derive the projection of ϕ onto $\Lambda^\perp$, we solve the following equations for h_1 and h_2,

$$\begin{aligned}E\left[\left\{\phi - \left[R - \pi(\widetilde{\boldsymbol{X}})\right]h_1(\widetilde{\boldsymbol{X}}) - h_2(\widetilde{\boldsymbol{X}})\right\}\left\{R - \pi(\widetilde{\boldsymbol{X}})\right\}\widetilde{h}_3(\widetilde{\boldsymbol{X}})\right] &= 0, \\ E\left[\left\{\phi - \left[R - \pi(\widetilde{\boldsymbol{X}})\right]h_1(\widetilde{\boldsymbol{X}}) - h_2(\widetilde{\boldsymbol{X}})\right\}\widetilde{h}_4(\widetilde{\boldsymbol{X}})\right] &= 0,\end{aligned}$$

for any $\widetilde{h}_3 \in L_2(\widetilde{\boldsymbol{X}})$ and $\widetilde{h}_4 \in \Lambda_{\widetilde{\boldsymbol{X}}}$.

These equations are equivalent to

$$\begin{aligned}E\left[\left\{\phi - \left[R - \pi(\widetilde{\boldsymbol{X}})\right]h_1(\widetilde{\boldsymbol{X}}) - h_2(\widetilde{\boldsymbol{X}})\right\}\left\{R - \pi(\widetilde{\boldsymbol{X}})\right\} \mid \widetilde{\boldsymbol{X}}\right] &= 0, \\ E\left[\left\{\phi - \left[R - \pi(\widetilde{\boldsymbol{X}})\right]h_1(\widetilde{\boldsymbol{X}}) - h_2(\widetilde{\boldsymbol{X}})\right\} \mid \widetilde{\boldsymbol{X}}\right] &= 0.\end{aligned}$$

By solving these equations, we have that

$$h_1(\widetilde{\boldsymbol{X}}) = \pi^{-1}(\widetilde{\boldsymbol{X}})\left\{E\left[Y \mid \widetilde{\boldsymbol{X}}\right] - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\right\}\boldsymbol{X}^\top\widetilde{\boldsymbol{u}}_0$$

and

$$h_2(\widetilde{\boldsymbol{X}}) = \left\{E\left[Y \mid \widetilde{\boldsymbol{X}}\right] - b'(\boldsymbol{X}^\top\boldsymbol{\beta})\right\}\boldsymbol{X}^\top\widetilde{\boldsymbol{u}}_0.$$

Thus, when the distribution of $(\boldsymbol{X}, Z)$ is known, and the surrogate outcome Z is available, the efficient influence function is

$$\phi_{2,Z} = \pi^{-1}(\widetilde{\boldsymbol{X}})R\left\{Y - E\left[Y \mid \widetilde{\boldsymbol{X}}\right]\right\}\boldsymbol{X}^\top\widetilde{\boldsymbol{u}}_0.$$

The semiparametric lower bound of estimating β_1^* is given by $\mathrm{var}(\widetilde{\phi}_{2,Z})$.

Now, we compare the semiparametric lower bound with and w/o the surrogate outcome Z when the distribution of $(\boldsymbol{X}, Z)$ is known. Under the assumption that $R \perp Z \mid \boldsymbol{X}$, we have

$$\widetilde{\phi}_{2,Z} = \pi^{-1}(\boldsymbol{X}) R \left\{ Y - E\left[Y \mid \widetilde{\boldsymbol{X}} \right] \right\} \boldsymbol{X}^\top \widetilde{\boldsymbol{u}}_0.$$

Likewise, we can derive the efficient influence function without the surrogate outcome Z, i.e.,

$$\phi_2 = \pi^{-1}(\boldsymbol{X}) R \left\{ Y - E\left[Y \mid \boldsymbol{X} \right] \right\} \boldsymbol{X}^\top \widetilde{\boldsymbol{u}}_0.$$

The difference

$$\begin{aligned} & \phi_2 - \widetilde{\phi}_{2,Z} \\ = \;& \pi^{-1}(\boldsymbol{X}) R \left\{ E\left[Y \mid \widetilde{\boldsymbol{X}} \right] - E\left[Y \mid \boldsymbol{X} \right] \right\} \boldsymbol{X}^\top \widetilde{\boldsymbol{u}}_0 \\ = \;& \pi^{-1}(\boldsymbol{X}) \left\{ R - \pi(\boldsymbol{X}) \right\} \left\{ E\left[Y \mid \widetilde{\boldsymbol{X}} \right] - E\left[Y \mid \boldsymbol{X} \right] \right\} \boldsymbol{X}^\top \widetilde{\boldsymbol{u}}_0 \\ & + \left\{ E\left[Y \mid \widetilde{\boldsymbol{X}} \right] - E\left[Y \mid \boldsymbol{X} \right] \right\} \boldsymbol{X}^\top \widetilde{\boldsymbol{u}}_0 \\ \in \;& \Lambda_\pi \oplus \Lambda_{\widetilde{\boldsymbol{X}}} = \Lambda^\perp. \end{aligned}$$

Since $\widetilde{\phi}_{2,Z} \in \Lambda$, we have $\mathrm{var}(\phi_2) \geqslant \mathrm{var}(\widetilde{\phi}_{2,Z})$, and the equality holds if and only if $E\left[Y \mid \widetilde{\boldsymbol{X}} \right] = E\left[Y \mid \boldsymbol{X} \right]$. □

A.2. Proof of Theorem 2.1

In this section, we show the characterization of the solution to

$$E\left[\left\{ R - \pi(\widetilde{\boldsymbol{X}}) \right\} \pi^{-1}(\widetilde{\boldsymbol{X}}) f(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \boldsymbol{X}^\top \boldsymbol{v} \right] = 0, \forall f \in L_2(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}).$$

Proof. First, we assume that $\pi(\widetilde{\boldsymbol{X}})$ is a $\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$-measurable function and show $\pi_*(\widetilde{\boldsymbol{X}})$ is the unique solution. Equation (2.3) is equivalent to

$$E\left[\left\{ R - \pi(\widetilde{\boldsymbol{X}}) \right\} \pi^{-1}(\widetilde{\boldsymbol{X}}) \boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right] = 0.$$

Since $\pi(\widetilde{\boldsymbol{X}})$ is assumed to be a $\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$-measurable function, we can obtain

$$\pi^{-1}(\widetilde{\boldsymbol{X}}) E\left[R \boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right] = E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right].$$

Thus, on $E\left[R \boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right] \neq 0$, we have

$$\begin{aligned} \pi^{-1}(\widetilde{\boldsymbol{X}}) &= E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right] / E\left[R \boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right] \\ &= 1 + E\left[(1 - R) \boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right] / E\left[R \boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right]. \end{aligned}$$

Since

$$E\left[(1-R)\,\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}\right] / E\left[R\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}\right] = \left\{\rho J_1(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})\right\}^{-1} J_0(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})(1-\rho),$$

we have that $\pi^{-1}(\widetilde{\boldsymbol{X}}) = \pi_*^{-1}(\widetilde{\boldsymbol{X}})$ on $J_1(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \neq 0$ if $\pi(\widetilde{\boldsymbol{X}})$ is a $\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$-measurable function.

Next, we show that any solution to Equation (2.3) bears such form, i.e., $\pi^{-1}(\widetilde{\boldsymbol{X}}) = \pi_*^{-1}(\widetilde{\boldsymbol{X}}) + \mathcal{T}h(\widetilde{\boldsymbol{X}})$. Suppose that $\pi^{-1}(\widetilde{\boldsymbol{X}})$ is a solution to Equation (2.3). Define $g(\widetilde{\boldsymbol{X}}) := \pi^{-1}(\widetilde{\boldsymbol{X}}) - \pi_*^{-1}(\widetilde{\boldsymbol{X}})$. Suppose we have

$$E\left[R\left\{g(\widetilde{\boldsymbol{X}}) + \pi_*^{-1}(\widetilde{\boldsymbol{X}})\right\}\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}\right] = E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}\right].$$

Since

$$\pi_*^{-1}(\widetilde{\boldsymbol{X}}) E\left[R\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}\right] = E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}\right],$$

then we must have

$$E\left[Rg(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}\right] = 0.$$

In fact, for any g satisfies that

$$E\left[Rg(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top \boldsymbol{v} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}\right] = 0,$$

we have $g = \mathcal{T}(g)$. Thus, $\pi^{-1}(\widetilde{\boldsymbol{X}})$ must have the form $\pi_*^{-1}(\widetilde{\boldsymbol{X}}) + \mathcal{T}h(\widetilde{\boldsymbol{X}})$ for some h. □

A.3. Outline of the proposed debias algorithm

An outline of the proposed debias algorithm is exhibited in Algorithm 1.

A.4. Proof of Theorem 3.1

In this section, we provide proofs of Theorem 3.1, and Corollary 3.1. In the proofs, we also accommodate the case where there are multiple surrogates. We use $\boldsymbol{Z}$ to denote the multi-dimensional surrogate vector. To derive the main theorem, we assume that $(\boldsymbol{X}, \boldsymbol{Z})$ and $Y \mid (\boldsymbol{Z}, \boldsymbol{X})$ do not depend on n; in contrast, the missing propensity $P(R = 1 \mid \boldsymbol{Z}, \boldsymbol{X})$ may depend on n. In this work, we consider two scenarios: 1) the missing propensity $P(R = 1 \mid \boldsymbol{Z}, \boldsymbol{X})$ does not change with n; 2) $P(R = 1 \mid \boldsymbol{Z}, \boldsymbol{X}) = \rho_n w(\widetilde{\boldsymbol{X}})$ with $\rho_n \to 0$, where $w(\widetilde{\boldsymbol{X}})$ does not depend on n, $w(\widetilde{\boldsymbol{X}})$ is always non-negative and bounded away from ∞, and $E\left[w(\widetilde{\boldsymbol{X}})\right] = 1$. Notice that for the first scenario, we can treat it as a special case of the second scenario with ρ_n, which does not change with n, and $P(R = 1 \mid \boldsymbol{Z}, \boldsymbol{X})/P(R = 1) = w(\widetilde{\boldsymbol{X}})$. In both scenarios, we have that $P(R = 1) = \rho_n$. In addition to these assumptions, we also assume the following conditions.

Algorithm 1: A debias procedure for individual coefficient estimation.

Input: A random seed; n samples; a positive integer K.
Output: Estimator $\widetilde{\beta}_1$.
Use the entire observed dataset and selected dimension reduction approach to get $\widehat{\mathbf{\Gamma}}$;
Use the entire dataset to fit the imputation model $\widehat{g}(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}})$; then we fit an initial coefficients $\widehat{\boldsymbol{\beta}}$ by solving

$$\min_{\beta} \widehat{E}_n \left[b(\boldsymbol{X}^\top \boldsymbol{\beta}) - \widehat{g}(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \boldsymbol{X}^\top \boldsymbol{\beta} \right]. \tag{A.1}$$

;
On the entire dataset, using the estimated initial coefficients $\widehat{\boldsymbol{\beta}}$, we estimate v by optimizing

$$\min_{\boldsymbol{w}} \widehat{E}_n \left[b''(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}})(X_1 - \boldsymbol{X}_{-1}^\top \boldsymbol{w})^2 \right];$$

denote the minimizer as $\widehat{\boldsymbol{w}}$;
Randomly split data into K subsets with equal sample sizes, denote the index sets as $\{I_k\}_{k=1}^K$, and set $k = 1$;
Estimate the imputation model using the fully observed data in I_k^c by kernel regression based on the reduced dimension, and estimate the $\widehat{\pi}_{(-k)}^{-1}$ using the data in I_k^c by the truncated kernel regression based on the reduced dimension;
Similarly, we repeat Step (1) for $k = 2, \cdots, K$;
Obtain the one-step debiased estimator $\widehat{\beta}_1$ by $\widetilde{\beta}_1 = \widehat{\beta}_1 - \bar{I}^{-1}\bar{S}$, where $\bar{S} = \sum_{k=1}^K S^{(k)}/K$, $\quad S^{(k)} = \widehat{E}_n^{(k)} \left[\left\{ S(\widehat{\boldsymbol{\beta}}; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)}) \right\}^\top \widehat{\boldsymbol{v}} \right]$.

Assumption A.1. The covariate $\boldsymbol{X}$'s and the surrogate outcome $\boldsymbol{Z}$ are bounded, and the function $b''(\cdot)$ is continuously differentiable; $\max\{\|\boldsymbol{\beta}^*\|_2, \|\boldsymbol{v}\|_2\}$ is bounded. In addition, the smallest eigenvalue of $E\left[\boldsymbol{X}\boldsymbol{X}^\top\right]$ is bounded away from 0.

Assumption A.2. There is a positive constant $\gamma_d > 1/4$ such that

$$\|\widehat{g}(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - Q(\boldsymbol{Z}, \boldsymbol{X})\|_\infty = O_p\left\{(n\rho_n)^{-\gamma_d}\right\},$$

and

$$\left\{\mathrm{vec}(\widehat{\mathbf{\Gamma}}) - \mathrm{vec}(\mathbf{\Gamma})\right\} = n^{-1}\sum_{i=1}^n 1\{R_i = 1\}\psi(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i) + o_p\left\{(n\rho_n)^{-1/2}\right\},$$

where $\psi(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i)$ is bounded with mean 0 and $\mathrm{vec}(\cdot)$ represents the vectorization of the matrix. In addition, we assume that

$$\sup_{\widetilde{\boldsymbol{X}}} \left| (\widehat{g} - g)(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - (\widehat{g} - g)(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \right| = o_p\left\{(n\rho_n)^{-1/2}\right\}.$$

Assumption A.3. The function $K(\cdot)$ is a kernel function with the order of $\nu - 1$. The $J_w(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}})$ and $J(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}})$ are νth order differentiable w.r.t $\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}$ with bounded derivatives, where

$$J_w(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) = E\left[w(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top \boldsymbol{v} \mid \mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}\right], \quad J(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) = E\left[\boldsymbol{X}^\top \boldsymbol{v} \mid \mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}\right].$$

Define $G(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$ as $J_w^{-1}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) J(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$. We assume that $G(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$ is bounded away from $+\infty$ on the open set $\{J_w(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \neq 0\}$. We also assume that the density function of $A^\top \widetilde{\boldsymbol{X}}$, $\eta(A^\top \widetilde{\boldsymbol{X}})$, is bounded away from 0 and $+\infty$, and νth order differentiable with bounded derivatives. Define

$$H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v}) = E\left[\left\{J_w^{-1}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) w(\widetilde{\boldsymbol{X}}) J(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) - 1\right\} \nabla g(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \boldsymbol{X}^\top \boldsymbol{v}\right],$$

where $\otimes$ is the Kronecker product. The $H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})$ is bounded for any $\|v\|_2 = 1$.

Assumption A.4. When $t > 0$ is small enough, there exist positive constants A_0 and γ_m such that $P\left\{0 \neq \left|J_w(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})\right| \leqslant t\right\} \leqslant A_0 t^{\gamma_m}$.

Assumption A.5. Take $c_n = \widetilde{\delta}_n^{2/(2+\gamma_m)}$, where $\widetilde{\delta}_n = \rho_n^{-1/2}(n\hbar^d/\log n)^{-1/2} + \hbar^\nu + (n\rho_n)^{-\gamma_d}$. We assume that $n\rho_n \to +\infty$ and $(n\rho_n)^{-\gamma_d} \widetilde{\delta}_n^{\gamma_m/(2+\gamma_m)} = o\left\{(n\rho_n)^{-1/2}\right\}$.

Assumptions A.1–A.5 are different from those in the main text. In the main text, we focus on Scenario 1) where ρ_n is a constant. When ρ_n is a constant, Assumptions A.1–A.5 are equivalent to Assumptions 3.1–3.5 in the main text.

To start with the proof, Lemma A.2 provides the convergence rate of the initial estimator $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{w}}$.

Lemma A.2. *Under Assumptions A.1-A.2, we have that*

$$\begin{aligned} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &= O_p\left\{(p/n)^{1/2} + (n\rho_n)^{-\gamma_d}\right\}, \\ \|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2 &= O_p\left\{(p/n)^{1/2} + (n\rho_n)^{-\gamma_d}\right\}. \end{aligned}$$

Proof. First, we show the consistency and convergence rate of $\widehat{\boldsymbol{\beta}}$. To show consistency, notice that the optimization problem (A.1) is strictly convex. Let $l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \widehat{Q}) = b(\boldsymbol{X}^\top \boldsymbol{\beta}) - \widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \boldsymbol{X}^\top \boldsymbol{\beta}$. Using the strong convexity of $E\left[l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; Q)\right]$, there is a positive constant $\lambda_{\min}$ such that

$$\begin{aligned} \lambda_{\min}\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 &\leqslant E\left[l_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}; Q)\right] - E\left[l_{\boldsymbol{\beta}}(\boldsymbol{\beta}^*; Q)\right] \\ &\leqslant \sup_{\boldsymbol{\beta} \in \mathcal{B}}\left|(\widehat{E}_n - E)\left[l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; Q)\right]\right| \\ &\quad + \sup_{\boldsymbol{\beta} \in \mathcal{B}}\left|\widehat{E}_n\left[l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; Q) - l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \widehat{Q})\right]\right| \\ &= I_1 + I_2. \end{aligned}$$

For I_2, we have

$$\begin{aligned} &\sup_{\boldsymbol{\beta} \in \mathcal{B}}\left|\widehat{E}_n\left[l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; Q) - l_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \widehat{Q})\right]\right| \\ = \ &\|\widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - Q(Z, \boldsymbol{X})\|_\infty \sup_{\boldsymbol{\beta} \in \mathcal{B}} \widehat{E}_n\left(|\boldsymbol{X}^\top \boldsymbol{\beta}|\right) = O_p\left\{(n\rho_n)^{-\gamma_d}\right\}, \end{aligned}$$

where $\mathcal{B}$ is a centered l_2-ball with a radius sufficiently large (see Condition A.1). For I_1, Theorem 2.1 in [38], we have

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left|(\widehat{E}_n - E)\left[l_{\boldsymbol{\beta}}(\boldsymbol{\beta};Q)\right]\right| = O_p\left\{(p/n)^{1/2}\right\}.$$

Thus, we have

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left\{(p/n)^{1/2} + (n\rho_n)^{-\gamma_d}\right\}.$$

Similarly we can derive that

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2 = O_p\left\{(p/n)^{1/2} + (n\rho_n)^{-\gamma_d}\right\}.$$

Specifically, let $l_{\boldsymbol{w}}(\boldsymbol{w};\widehat{\boldsymbol{\beta}}) = b''(\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}})(X_1 - \boldsymbol{X}_{-1}^\top\boldsymbol{w})^2$. Using the strong convexity of $E\left[l_{\boldsymbol{w}}(\boldsymbol{w};\widehat{\boldsymbol{\beta}})\right]$, there is a positive constant $\lambda'_{\min}$ such that

$$\begin{aligned}
\lambda_{\min}\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2 &\leqslant E\left[l_w(\widehat{\boldsymbol{w}};\boldsymbol{\beta}^*)\right] - E\left[l_{\boldsymbol{\beta}}(\boldsymbol{w}^*;\boldsymbol{\beta}^*)\right] \\
&\leqslant \sup_{\boldsymbol{w}\in\mathcal{B}}\left|(\widehat{E}_n - E)\left[l_{\boldsymbol{w}}(\boldsymbol{w};\boldsymbol{\beta}^*)\right]\right| \\
&\quad + \sup_{\boldsymbol{w}\in\mathcal{B}}\left|\widehat{E}_n\left[l_{\boldsymbol{w}}(\boldsymbol{w};\boldsymbol{\beta}^*) - l_{\boldsymbol{w}}(\boldsymbol{w};\widehat{\boldsymbol{\beta}})\right]\right| \\
&= I_1 + I_2.
\end{aligned}$$

For I_1, we have

$$\sup_{\boldsymbol{w}\in\mathcal{B}}\left|(\widehat{E}_n - E)\left[l_{\boldsymbol{w}}(\boldsymbol{w};\boldsymbol{\beta}^*)\right]\right| = O_p\left\{(p/n)^{1/2}\right\}.$$

For I_2, we have

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{B}}\left|\widehat{E}_n\left\{l_{\boldsymbol{w}}(\boldsymbol{w};\boldsymbol{\beta}^*) - l_{\boldsymbol{w}}(\boldsymbol{w};\widehat{\boldsymbol{\beta}})\right\}\right| \\
= &\sup_{\boldsymbol{w}\in\mathcal{B}}\left|\widehat{E}_n\left[\left\{b''(\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}) - b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\right\}(X_1 - \boldsymbol{X}_{-1}^\top\boldsymbol{w})^2\right]\right| \\
\leqslant &\, C\widehat{E}_n\left\{\left|\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right|\right\} \\
\lesssim &\, O_p\left\{(p/n)^{1/2} + (n\rho_n)^{-\gamma_d}\right\}.
\end{aligned}$$

□

Lemma A.3. *When $n\rho_n \to +\infty$, for any r, we have*

$$\widehat{\rho}^r/\rho_n^r - 1 = O_p\left\{(n\rho_n)^{-1/2}\right\},$$

where $\widehat{\rho} = \widehat{E}_n[R]$.

Proof. To show this result, we use the Bernstein's inequality. By Bernstein's inequality, we have

$$P\left\{\left|\sum_{i=1}^{n}(R_i-\rho_n)\right|\geqslant t\right\}\leqslant \exp\left[-t^2/\{2n\rho_n(1-\rho_n)+t/3\}\right].$$

Notice that when $n\rho_n\to+\infty$, we have that

$$P\left(|\widehat{\rho}-\rho_n|\geqslant t\right)\leqslant \exp\left[-nt^2/\{3\rho_n(1-\rho_n)\}\right].$$

Set $t=(\rho_n/n)^{1/2}$, we have

$$\widehat{\rho}-\rho_n=O_p\left\{(\rho_n/n)^{1/2}\right\},$$

for n large enough.

By $n\rho_n\to+\infty$, we further have $\rho_n^{-1}(\widehat{\rho}-\rho_n)=O_p\left\{(n\rho_n)^{-1/2}\right\}=o_p(1)$. Thus, we have

$$\widehat{\rho}^r-\rho_n^r=r\rho_n^{r-1}(\widehat{\rho}-\rho_n)+o_p\left\{\rho_n^{r-1}(\widehat{\rho}-\rho_n)\right\}=O_p\left\{\rho_n^r(n\rho_n)^{-1/2}\right\}. \qquad \square$$

Lemma A.4. *Under Assumptions A.1–A.5, we have*

$$\left\|\left\{\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right\}R\right\|_{P,2}=O_p\left\{\rho_n^{-1/2}c_n^{\gamma_m/2}+\rho_n^{-1/2}(n\rho_n)^{-1/2}\right\},$$

where

$$c_n^{1+\gamma_m/2}=\rho_n^{-1/2}(n\hbar^d/\log n)^{-1/2}+\hbar^{-\nu}+(n\rho_n)^{-\gamma_d},$$

and $\hbar$ is the kernel bandwidth in estimating $\widehat{\pi}_{(-k)}^{-1}$; in addition, we have

$$\left\|\left\{\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right\}R\right\|_{P,1}=O_p(c_n^{\gamma_m/2}+(n\rho_n)^{-1/2}).$$

Proof. Conditional on I_k^c, taking the expectation over $\widetilde{\boldsymbol{X}}$ and R over I_k, we have

$$\begin{aligned}
& E\left[\left\{\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right\}R\right]^2\\
= \;& E\left[\pi\left\{\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right\}^2\right]\\
= \;& E\left[\pi\left\{\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right\}^2 1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\neq 0,\left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right|>c_n\right\}\right]\\
& +E\left[\pi\left\{\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right\}^2 1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})=0,\left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right|>c_n\right\}\right]\\
& +E\left[\pi\left(\widehat{\rho}^{-1}-\pi_*^{-1}\right)^2 1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\neq 0,\left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right|\leqslant c_n\right\}\right]\\
& +E\left[\pi\left(\widehat{\rho}^{-1}-\rho_n^{-1}\right)^2 1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})=0,\left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right|\leqslant c_n\right\}\right]\\
= \;& I_{21}+I_{22}+I_{23}+I_{24}.
\end{aligned}$$

Notice we can rewrite $\widehat{\pi}_{(-k)}^{-1}$ and π_*^{-1} as

$$\begin{aligned}
\widehat{\pi}_{(-k)}^{-1}(\widetilde{\boldsymbol{X}}) &= \widehat{\rho}^{-1}\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})/\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}),\\
\pi_*^{-1}(\widetilde{\boldsymbol{X}}) &= \rho_n^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})/J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}),\\
\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) &= h^{-d}\widehat{E}_n^{(k)}\left[X^\top\hat{\boldsymbol{v}}K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}-\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right].
\end{aligned}$$

Notice that

$$\widehat{\rho}\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) = h^{-d}\widehat{E}_n^{(k)}\left[R\boldsymbol{X}^\top\hat{\boldsymbol{v}}K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}-\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right].$$

To bound each term, we first consider the convergence rate of $\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})$ and $\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})$, i.e.,

$$\begin{aligned}
&\sup_x\left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - \rho_n J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|,\\
&\sup_x\left|\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|,
\end{aligned}$$

where $\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})$ is the density function of $\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}$.

By Lemma 6.6 in [39], for any positive α_1, we can choose constants α_2 and α_3 such that

$$\begin{aligned}
&P\left\{\sup_{|A|,\widetilde{\boldsymbol{X}}}\left|\left(\widehat{E}_n^{(k)}-E\right)\left[R\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(A^\top\widetilde{\boldsymbol{X}}-A^\top\widetilde{\boldsymbol{X}})\right]\right|\right.\\
&\left.\geqslant \alpha_2\rho_n^{1/2}(n\hbar^d/\log n)^{-1/2}\right\}\\
=\;& o(1),\\
&P\left\{\sup_{|A|,\widetilde{\boldsymbol{X}}}\left|\left(\widehat{E}_n^{(k)}-E\right)\left[\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(A^\top\widetilde{\boldsymbol{X}}-A^\top\widetilde{\boldsymbol{X}})\right]\right|\geqslant \alpha_2(n\hbar^d/\log n)^{-1/2}\right\}\\
=\;& o(1).
\end{aligned}$$

To bound $E\left[R\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(A^\top\widetilde{\boldsymbol{X}}-A^\top\widetilde{\boldsymbol{X}})\right] = \rho_n E\left[w(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(A^\top\widetilde{\boldsymbol{X}}-A^\top\widetilde{\boldsymbol{X}})\right]$, by Condition A.3, we directly calculate

$$\begin{aligned}
&\left|E\left[w(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(A^\top\widetilde{\boldsymbol{X}}-A^\top\widetilde{\boldsymbol{X}})\right]\right.\\
&\left.-E\left[w(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top\boldsymbol{v}\mid A^\top\widetilde{\boldsymbol{X}}=A^\top\widetilde{\boldsymbol{X}}\right]\eta(A^\top\widetilde{\boldsymbol{X}})\right|\\
=\;&\left|E\left[E\left[w(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top\boldsymbol{v}\mid A^\top\widetilde{\boldsymbol{X}}\right]K_\hbar(A^\top\widetilde{\boldsymbol{X}}-A^\top\widetilde{\boldsymbol{X}})\right]\right.\\
&\left.-E\left[w(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top\boldsymbol{v}\mid A^\top\widetilde{\boldsymbol{X}}=A^\top\widetilde{\boldsymbol{X}}\right]\eta(A^\top\widetilde{\boldsymbol{X}})\right|\\
=\;&\left|\int E\left[w(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top\boldsymbol{v}\mid A^\top\widetilde{\boldsymbol{X}}=A^\top\widetilde{\boldsymbol{X}}+\hbar s\right]K(s)\eta(A^\top\widetilde{\boldsymbol{X}}+\hbar s)\mathrm{d}s\right.
\end{aligned}$$

$$
\begin{aligned}
& -E\left[w(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top\boldsymbol{v} \mid A^\top\widetilde{\boldsymbol{X}} = A^\top\widetilde{\boldsymbol{X}}\right]\eta(A^\top\widetilde{\boldsymbol{X}})\Big| \\
\leqslant\ & \alpha_4\hbar^{-\nu},
\end{aligned}
$$

uniformly holds for any A and $\widetilde{\boldsymbol{X}}$ with a constant α_4, where $\eta(\cdot)$ is the density function of $A\widetilde{\boldsymbol{X}}$. Similarly, we can choose a sufficiently large α_4 such that

$$
\begin{aligned}
& \sup_{A,\widetilde{\boldsymbol{X}}}\left|E\left[\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(A^\top\widetilde{\boldsymbol{X}} - A^\top\widetilde{\boldsymbol{X}})\right] - E\left[\boldsymbol{X}^\top\boldsymbol{v} \mid A^\top\widetilde{\boldsymbol{X}} = A^\top\widetilde{\boldsymbol{X}}\right]\eta(A^\top\widetilde{\boldsymbol{X}})\right| \\
\leqslant\ & \alpha_4\hbar^{-\nu}.
\end{aligned}
$$

Since

$$
\|\mathrm{vec}(\widehat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma})\|_2 = O_p\left\{(n\rho_n)^{-1/2}\right\},
$$

we have

$$
\begin{aligned}
& \sup_{\widetilde{\boldsymbol{X}}}\left|E\left[w(\widetilde{\boldsymbol{X}})\boldsymbol{X}^\top\boldsymbol{v} \mid \widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}} = \widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}\right]\eta(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| \\
=\ & O_p\left\{(n\rho_n)^{-1/2}\right\}, \\
& \sup_{\widetilde{\boldsymbol{X}}}\left|E\left[\boldsymbol{X}^\top\boldsymbol{v} \mid \widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}\right]\eta(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| \\
=\ & O_p\left\{(n\rho_n)^{-1/2}\right\}.
\end{aligned}
$$

Combining these inequalities, we have

$$
\begin{aligned}
& \sup_{\widetilde{\boldsymbol{X}}}\left|\widehat{E}_n^{(k)}\left[R\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right] - \rho_n J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| \\
=\ & O_p\left[\rho_n\left\{(n\rho_n\hbar^d/\log n)^{-1/2} + \hbar^{-\nu} + (n\rho_n)^{-1/2}\right\}\right]; \\
& \sup_{\widetilde{\boldsymbol{X}}}\left|\widehat{E}_n^{(k)}\left[\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right] - J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| \\
=\ & O_p\left\{(n\rho_n\hbar^d/\log n)^{-1/2} + \hbar^{-\nu} + (n\rho_n)^{-\gamma_d}\right\}.
\end{aligned}
$$

To obtain the convergence rate of $\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})$ and $\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})$, we further consider

$$
\begin{aligned}
& \sup_{\widetilde{\boldsymbol{X}}}\left|\widehat{E}_n^{(k)}\left[R\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right] - \widehat{\rho}\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| \\
=\ & \sup_{\widetilde{\boldsymbol{X}}}\left|\widehat{E}_n^{(k)}\left[R\boldsymbol{X}^\top(\boldsymbol{v} - \widehat{\boldsymbol{v}})K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right]\right|,
\end{aligned}
$$

and

$$
\begin{aligned}
& \sup_{\widetilde{\boldsymbol{X}}}\left|\widehat{E}_n^{(k)}\left[\boldsymbol{X}^\top\boldsymbol{v}K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right] - \widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| \\
=\ & \sup_{\widetilde{\boldsymbol{X}}}\left|\widehat{E}_n^{(k)}\left[\boldsymbol{X}^\top(\boldsymbol{v} - \widehat{\boldsymbol{v}})K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right]\right|.
\end{aligned}
$$

To bound this term, following the proof above, we can show that

$$\begin{aligned}
&\sup_{\widetilde{\boldsymbol{X}}} \left\| \widehat{E}_n^{(k)} \left[R\boldsymbol{X} K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \right] - \right. \\
&\left. \rho_n E \left[w(\widetilde{\boldsymbol{X}}) \boldsymbol{X} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} = \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right] \eta(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \right\|_\infty \\
= \;& O_p \left[\rho_n \left\{ (n\rho_n \hbar^d / \log n)^{-1/2} + \hbar^{-\nu} + (n\rho_n)^{-1/2} \right\} \right]; \\
&\sup_{\widetilde{\boldsymbol{X}}} \left\| \widehat{E}_n^{(k)} \left[\boldsymbol{X} K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \right] - E \left[\boldsymbol{X} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} = \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}} \right] \eta(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \right\|_\infty \\
= \;& O_p \left\{ (n\hbar^d / \log n)^{-1/2} + \hbar^{-\nu} + (n\rho_n)^{-1/2} \right\}.
\end{aligned}$$

Notice that

$$\begin{aligned}
&\sup_{\widetilde{\boldsymbol{X}}} \left| \widehat{E}_n^{(k)} \left[R\boldsymbol{X}^\top (\boldsymbol{v} - \widehat{\boldsymbol{v}}) K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \right] \right| \\
\leqslant \;& \sup_{\widetilde{\boldsymbol{X}}} \left\| \widehat{E}_n^{(k)} \left[R\boldsymbol{X} K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \right] \right\|_\infty \| \boldsymbol{v} - \widehat{\boldsymbol{v}} \|_1 \\
= \;& O_p \left\{ \rho_n \left[(p/n)^{1/2} + (n\rho_n)^{-\gamma_d} \right] \right\},
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{\widetilde{\boldsymbol{X}}} \left| \widehat{E}_n^{(k)} \left[\boldsymbol{X}^\top (\boldsymbol{v} - \widehat{\boldsymbol{v}}) K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \right] \right| \\
\leqslant \;& \sup_{\widetilde{\boldsymbol{X}}} \left\| \widehat{E}_n^{(k)} \left[\boldsymbol{X} K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) \right] \right\|_\infty \| \boldsymbol{v} - \widehat{\boldsymbol{v}} \|_1 \\
= \;& O_p \left\{ (p/n)^{1/2} + (n\rho_n)^{-\gamma_d} \right\}.
\end{aligned}$$

Thus, because p is fixed and $\gamma_d \leqslant 1/2$, for the convergence rate of $\widehat{\rho}\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})$ and $\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})$, we have

$$\begin{aligned}
&\sup_x \left| \widehat{\rho}\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - \rho_n J_w(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \eta(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \right| \\
= \;& O_p \left\{ \rho_n \left[(n\rho_n \hbar^d / \log n)^{-1/2} + \hbar^{-\nu} + (n\rho_n)^{-\gamma_d} \right] \right\}, \\
&\sup_x \left| \widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - J(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \eta(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \right| \\
= \;& O_p \left\{ (n\rho_n \hbar^d / \log n)^{-1/2} + \hbar^{-\nu} + (n\rho_n)^{-\gamma_d} \right\}.
\end{aligned}$$

By Lemma A.3, these imply that

$$\begin{aligned}
&\sup_x \left| \widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - J_w(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \eta(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \right| \\
= \;& O_p \left\{ (n\rho_n \hbar^d / \log n)^{-1/2} + \hbar^{-\nu} + (n\rho_n)^{-\gamma_d} \right\}.
\end{aligned}$$

Let $\delta_n = \hbar^{-\nu} + (n\rho_n)^{-\gamma_d}$. Take $c_n^r = (n\rho_n\hbar^d/\log n)^{-1/2} + \delta_n$, where $r > 1$ is a constant to be chosen. Define the event Ω as

$$\begin{aligned}
\sup_{\widetilde{\boldsymbol{X}}}\left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| &\leqslant C(n\rho_n\hbar^d/\log n)^{-1/2} + C\delta_n,\\
\sup_{\widetilde{\boldsymbol{X}}}\left|\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| &\leqslant C(n\hbar^d/\log n)^{-1/2} + C\delta_n,\\
\widehat{\rho}^{-1}\rho_n &\leqslant 1 + C(n\rho_n)^{-1/2},
\end{aligned}$$

where ϵ is any positive constant. The derivation above and Lemma A.3 shows that $\lim_{C\to+\infty}\lim_n P(\Omega) = 1$.

Next, on Ω, we bound I_{21}, I_{22}, I_{23}, and I_{24}. For I_{22}, on Ω, when n is large enough, we have

$$1\left\{\left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| > c_n, J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) = 0\right\} = 0,$$

for all $\widetilde{\boldsymbol{X}}$. Thus, $I_{22} = 0$.

For I_{23}, since $P(R = 1 \mid Z, \boldsymbol{X}) = \rho_n w(\widetilde{\boldsymbol{X}})$, and $w(\widetilde{\boldsymbol{X}})$ is bounded away from 0, on Ω, we have

$$\begin{aligned}
&\pi\left(\widehat{\rho}^{-1} - \pi_*^{-1}\right)^2 1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) \neq 0, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| \leqslant c_n\right\}\\
\leqslant\ & c_{23}\rho_n^{-1}1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) \neq 0, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| \leqslant c_n\right\},
\end{aligned}$$

for some constant c_{23} large enough. By Lemma A.3, on Ω, we have

$$I_{23} \leqslant c_{23}\rho_n^{-1}P\left\{0 \neq \left|J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| \leqslant c_n + Cc_n^r\right\}.$$

For n large enough, we have $Cc_n^r \leqslant c_n$. Thus, by Condition A.4, we have

$$I_{23} \leqslant c_{23}\rho_n^{-1}(2c_n)^{\gamma_m}.$$

For I_{21}, we separate the discussion based on the value of $J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})$. When $|J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})| \leqslant c_n/2$, the upper bound for I_{23} can be applied, since we apply truncation to impose that $\widehat{\pi}_{(-k)}^{-1}(\widetilde{\boldsymbol{X}}) \leqslant \widehat{\rho}^{-1}M$, where M is a large constant. Specifically, in this case

$$I_{23} \leqslant 2\rho_n E\left[w(\widetilde{\boldsymbol{X}})(\widehat{\rho}^{-2}M^2 + \rho_n^{-2}M^2)1\left\{0 \neq |J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})| \leqslant c_n/2\right\}\right] \leqslant 2c_{211}c_n^{\gamma_m},$$

for a sufficiently large constant c_{211}.

When $|J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})| > c_n/2$, to derive an upper bound for I_{21}, we first derive an upper bound for $\sup_{\widetilde{\boldsymbol{X}}}\left|\widehat{\pi}_{(-k)}^{-1} - \pi_*^{-1}\right|$. When $\left|J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| > c_n/2, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| > c_n$, we write

$$\begin{aligned}
&\left|\widehat{\pi}_{(-k)}^{-1} - \pi_*^{-1}\right|\\
=\ &\left|\left\{\widehat{\rho}\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - \left\{\rho_n J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|
\end{aligned}$$

$$\leqslant \left|\left\{\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - \left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|\widehat{\rho}^{-1}$$
$$+\left|\widehat{\rho}^{-1}-\rho_n^{-1}\right|\left|\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|.$$

By Condition A.3, we have that $|\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})|$ is bounded. By Lemma A.3 and Condition A.3, we have $\left|\widehat{\rho}^{-1}-\rho_n^{-1}\right|\left|\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| \leqslant C\rho_n^{-1}(n\rho_n)^{-1/2}$ on Ω. Further, on Ω, when $\left|J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| > c_n/2, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| > c_n$, we have

$$\left|\left\{\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - \left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| \leqslant Cc_n^{\boldsymbol{\Gamma}-1},$$

for a large enough C. Thus, we have

$$I_{21} \lesssim \rho_n^{-1}(c_n^{2r-2}+(n\rho_n)^{-1}).$$

Similarly, we can obtain $I_{24} \leqslant \rho_n\left|\widehat{\rho}^{-1}-\rho_n^{-1}\right|^2 \leqslant c_{24}\rho_n^{-1}(n\rho_n)^{-1}$, for some constant c_{24}.

Therefore, we have

$$\left\|\left\{\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right\}R\right\|_{P,2} = O_p\left\{\rho_n^{-1/2}c_n^{r-1}+\rho_n^{-1/2}c_n^{\gamma_m/2}+\rho_n^{-1/2}(n\rho_n)^{-1/2}\right\}.$$

Take $r = 1+\gamma_m/2$, we have

$$\left\|\left\{\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right\}R\right\|_{P,2} = O_p\left\{\rho_n^{-1/2}c_n^{\gamma_m/2}+\rho_n^{-1/2}(n\rho_n)^{-1/2}\right\},$$

where $c_n^{1+\gamma_m/2} = \rho_n^{-1/2}(n\hbar^d/\log n)^{-1/2}+\delta_n$.

Similarly, we can show that

$$\begin{aligned}\left\|\left\{\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right\}R\right\|_{P,1} &= O_p\left\{c_n^{\gamma_m}+c_n^{r-1}+(n\rho_n)^{-1/2}\right\}\\ &= O_p\left\{c_n^{\gamma_m/2}+(n\rho_n)^{-1/2}\right\}.\end{aligned}$$

To show this, we write

$$\begin{aligned}&E\left[\left|\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right|R\right]\\ =\;& E\left[\pi\left|\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right|\right]\\ =\;& E\left[\pi\left|\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right|1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\neq 0, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| > c_n\right\}\right]\\ &+E\left[\pi\left|\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right|1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})= 0, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| > c_n\right\}\right]\\ &+E\left[\pi\left|\widehat{\rho}^{-1}-\pi_*^{-1}\right|1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\neq 0, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right| \leqslant c_n\right\}\right]\end{aligned}$$

$$
\begin{aligned}
&+E\left[\pi\left|\widehat{\rho}^{-1}-\rho_n^{-1}\right| 1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})=0, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right|\leqslant c_n\right\}\right]\\
=\;& \widetilde{I}_{21}+\widetilde{I}_{22}+\widetilde{I}_{23}+\widetilde{I}_{24}.
\end{aligned}
$$

Similar to I_{22}, on Ω, when n is large enough, we have $\widetilde{I}_{22}=0$. For $\widetilde{I}_{23}$, since $P(R=1\mid Z,\boldsymbol{X})=\rho_n w(\widetilde{\boldsymbol{X}})$, and $w(\widetilde{\boldsymbol{X}})$ is bounded away from 0, on Ω, we have

$$
\begin{aligned}
&\pi\left(\widehat{\rho}^{-1}-\pi_*^{-1}\right) 1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\neq 0, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right|\leqslant c_n\right\}\\
\leqslant\;& \tilde{c}_{23} 1\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\neq 0, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right|\leqslant c_n\right\},
\end{aligned}
$$

for some constant $\tilde{c}_{23}$ large enough. By Lemma A.3, on Ω, we have

$$
I_{23} \leqslant \tilde{c}_{23} P\left\{0\neq\left|J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\eta(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|\leqslant c_n+Cc_n^r\right\}.
$$

For n large enough, we have $Cc_n^r\leqslant c_n$. Thus, we have

$$
\widetilde{I}_{23} \lesssim c_n^{\gamma_m}.
$$

For $\widetilde{I}_{24}$, we have

$$
\widetilde{I}_{24}\leqslant \rho_n\left|\widehat{\rho}^{-1}-\rho_n^{-1}\right| = O_p\left\{(n\rho_n)^{-1/2}\right\}.
$$

For $\widetilde{I}_{21}$, when $|J_1(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})|\leqslant c_n/2$, the upper bound for $\widetilde{I}_{23}$ can be applied; when $\left|J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|>c_n/2, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right|>c_n$, we write

$$
\begin{aligned}
&\rho_n\left|\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right|\\
\leqslant\;& \left|\left\{\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})-\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|\widehat{\rho}^{-1}\rho_n\\
&+\rho_n\left|\widehat{\rho}^{-1}-\rho_n^{-1}\right|\left|\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|.
\end{aligned}
$$

By Condition A.3, we have that

$$
\left|\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|
$$

is bounded. By Lemma A.3 and Condition A.3, we have

$$
\rho_n\left|\widehat{\rho}^{-1}-\rho_n^{-1}\right|\left|\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right| = O_p\left\{(n\rho_n)^{-1/2}\right\}.
$$

Further, on Ω, when $\left|J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|>c_n/2, \left|\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right|>c_n$, we have

$$
\left|\left\{\widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}\widehat{J}^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})-\left\{J_w(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}^{-1}J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|\leqslant Cc_n^{\boldsymbol{\Gamma}-1},
$$

for a large enough C. Thus, we have

$$\widetilde{I}_{21} = O_p\left\{c_n^{r-1} + (n\rho_n)^{-1/2}\right\}.$$

Combining these results, we can obtain

$$\left\|\left\{\widehat{\pi}_{(-k)}^{-1} - \pi_*^{-1}\right\} R\right\|_{P,1} = O_p\left\{c_n^{\gamma_m} + c_n^{r-1} + (n\rho_n)^{-1/2}\right\} = O_p\left\{c_n^{\gamma_m/2} + (n\rho_n)^{-1/2}\right\},$$

where $c_n^{1+\gamma_m/2} = \rho_n^{-1/2}(n\hbar^d/\log n)^{-1/2} + \delta_n$. □

Lemma A.5.

$$\begin{aligned} & K^{-1}\sum_{k=1}^{K} E\left[\left\{S(\widehat{\boldsymbol{\beta}}; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)}) - S(\boldsymbol{\beta}^*; Q, \pi_*)\right\}^\top \boldsymbol{u}\right] \\ = \; & E\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right] \\ & + K^{-1}\sum_{k=1}^{K} E\left[\left(R\pi_*^{-1} - 1\right)\left\{\widehat{g}_{(-k)}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) - g(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}\boldsymbol{X}^\top\boldsymbol{u}\right] \\ & + H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{u})\left\{\mathrm{vec}(\widehat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma})\right\} + o_p\left\{(n\rho_n)^{-1/2}\right\} \end{aligned}$$

uniformly holds for any vector $\boldsymbol{u}$ *with* $\|\boldsymbol{u}\|_2 = 1$*, where*

$$H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{u}) = E\left[\left\{J_w^{-1}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})w(\widetilde{\boldsymbol{X}})J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) - 1\right\}(\boldsymbol{X}^\top\boldsymbol{u})\left\{\nabla g(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) \otimes X\right\}^\top\right],$$

$\otimes$ *is the Kronecker product, and* $\mathrm{vec}(\boldsymbol{\Gamma})$ *is the vectorization of* $\boldsymbol{\Gamma}$*. When* $\boldsymbol{u} = \boldsymbol{v}/\|\boldsymbol{v}\|_2$*, we have*

$$\begin{aligned} & K^{-1}\sum_{k=1}^{K} E\left[\left\{S(\widehat{\boldsymbol{\beta}}; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)}) - S(\boldsymbol{\beta}^*; Q, \pi_*)\right\}^\top \boldsymbol{v}\right] \\ = \; & E\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{v}\right] \\ & + H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\left\{\mathrm{vec}(\widehat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma})\right\} + o_p\left\{(n\rho_n)^{-1/2}\right\}. \end{aligned}$$

Proof. We write

$$\begin{aligned} & E\left[\left\{S(\widehat{\boldsymbol{\beta}}; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)}) - S(\boldsymbol{\beta}^*; Q, \pi_*)\right\}^\top \boldsymbol{u}\right] \\ = \; & E\left[\left\{S(\widehat{\boldsymbol{\beta}}; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)}) - S(\boldsymbol{\beta}^*; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)})\right\}^\top \boldsymbol{u}\right] \\ & + E\left[\left\{S(\boldsymbol{\beta}^*; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)}) - S(\boldsymbol{\beta}^*; Q, \pi_*)\right\}^\top \boldsymbol{u}\right] \\ = \; & I_{31} + I_{32}. \end{aligned}$$

For I_{31}, we have

$$I_{31} = E\left[\left\{b'(\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}) - b'(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\right\}\boldsymbol{X}^\top\boldsymbol{u}\right].$$

We use the Taylor expansion, and we have

$$I_{31} = E\left[\left\{b''(\boldsymbol{X}^\top\boldsymbol{b})\right\}\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right],$$

where $\boldsymbol{b} = \lambda\widehat{\boldsymbol{\beta}} + (1-\lambda)\boldsymbol{\beta}^*$. Thus, we have that

$$\begin{aligned}
&\left|I_{31} - E\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right]\right| \\
= &\left|E\left[\left\{b''(\boldsymbol{X}^\top\boldsymbol{b}) - b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\right\}\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right]\right| \\
\leqslant &\, CE\left[\left\{\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\}^2\right] \\
= &\, O_p\left\{n^{-1} + (n\rho_n)^{-2\gamma_d}\right\}
\end{aligned}$$

uniformly holds for all $\|\boldsymbol{u}\|_2 = 1$. By $\gamma_d > 1/4$, we can conclude that

$$I_{31} = E\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right] + o_p\left\{(n\rho_n)^{-1/2}\right\}.$$

For I_{32}, we write

$$\begin{aligned}
I_{32} = &\, E\left[\left(\widehat{\pi}_{(-k)}^{-1} - \pi_*^{-1}\right)R\left(Q - Y\right)\boldsymbol{X}^\top\boldsymbol{u}\right] \\
&+E\left[\left(R\pi_*^{-1} - 1\right)\left(\widehat{Q}_{(-k)} - Q\right)\boldsymbol{X}^\top\boldsymbol{u}\right] \\
&+E\left[\left(\widehat{\pi}_{(-k)}^{-1} - \pi_*^{-1}\right)R\left(\widehat{Q}_{(-k)} - Q\right)\boldsymbol{X}^\top\boldsymbol{u}\right] \\
= &\, I_{321} + I_{322} + I_{323}.
\end{aligned}$$

For I_{321}, because $E\left[Q - Y \mid R, \widetilde{\boldsymbol{X}}\right] = 0$, we have $I_{321} = 0$. By Lemma A.4 and Condition A.2, we have

$$\begin{aligned}
|I_{323}| \leqslant &\sup_X\left|\left(\widehat{Q}_{(-k)} - Q\right)\boldsymbol{X}^\top\boldsymbol{v}\right|\left\|\left(\widehat{\pi}_{(-k)}^{-1} - \pi_*^{-1}\right)R\right\|_{P,1} \\
= &\, O_p((n\rho_n)^{-\gamma_d}(c_n^{\gamma_m/2} + (n\rho_n)^{-1/2})).
\end{aligned}$$

By Condition A.5, we have $(n\rho_n)^{-\gamma_d}(c_n^{\gamma_m/2} + (n\rho_n)^{-1/2}) = o\left\{(n\rho_n)^{-1/2}\right\}$, and $I_{323} = o_p\left\{(n\rho_n)^{-1/2}\right\}$. For I_{322}, we write

$$\begin{aligned}
\widehat{Q}_{(-k)} - Q = &\, \widehat{g}_{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - g(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) \\
= &\, \widehat{g}_{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - \widehat{g}_{(-k)}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) + \widehat{g}_{(-k)}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) - g(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
I_{322} = &\, E\left[\left(R\pi_*^{-1} - 1\right)\left\{\widehat{g}_{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}}) - \widehat{g}_{(-k)}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}\boldsymbol{X}^\top\boldsymbol{u}\right] \\
&+E\left[\left(R\pi_*^{-1} - 1\right)\left\{\widehat{g}_{(-k)}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}) - g(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}\boldsymbol{X}^\top\boldsymbol{u}\right].
\end{aligned}$$

We discuss these two terms separately. The second term depends on the value of u. For example, by the definition of π_*, when $u = v/\|v\|_2$, the second term is 0. In the following, we focus on the first term.

$$\begin{aligned} & E\left[\left(R\pi_*^{-1}-1\right)\left\{\hat{g}_{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})-\hat{g}_{(-k)}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}\boldsymbol{X}^\top\boldsymbol{u}\right] \\ = \; & E\left[\left(R\pi_*^{-1}-1\right)\left\{(\hat{g}_{(-k)}-g)(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})-(\hat{g}_{(-k)}-g)(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}\boldsymbol{X}^\top\boldsymbol{u}\right] \\ & +E\left[\left(R\pi_*^{-1}-1\right)\left\{g(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})-g(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right\}\boldsymbol{X}^\top\boldsymbol{u}\right] \\ = \; & I_{3221}+I_{3222}. \end{aligned}$$

Notice that

$$E\left(R\pi_*^{-1}\mid\widetilde{\boldsymbol{X}}\right)=J_w^{-1}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})w(\widetilde{\boldsymbol{X}})J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}}).$$

By Taylor's expansion, we have

$$I_{3222} \;=\; H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\left\{\mathrm{vec}(\widehat{\boldsymbol{\Gamma}})-\mathrm{vec}(\boldsymbol{\Gamma})\right\}+O_p(\|\mathrm{vec}(\widehat{\boldsymbol{\Gamma}})-\mathrm{vec}(\boldsymbol{\Gamma})\|_2^2).$$

By $\|\mathrm{vec}(\widehat{\boldsymbol{\Gamma}})-\mathrm{vec}(\boldsymbol{\Gamma})\|_2^2=(n\rho_n)^{-1}=o_p\left\{(n\rho_n)^{-1/2}\right\}$, we have

$$I_{3222} \;=\; H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})(\mathrm{vec}(\widehat{\boldsymbol{\Gamma}})-\mathrm{vec}(\boldsymbol{\Gamma}))+o_p((n\rho_n)^{1/2}).$$

By Condition A.2, we have

$$|I_{3221}|\lesssim\sup_{\widetilde{\boldsymbol{X}}}\left|\left(\hat{g}_{(-k)}-g\right)(\widehat{\boldsymbol{\Gamma}}^\top\widetilde{\boldsymbol{X}})-\left(\hat{g}_{(-k)}-g\right)(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})\right|=o_p\left\{(n\rho_n)^{-1/2}\right\}.$$

Combining these equations, we can conclude the proof. □

Lemma A.6. *Under Assumptions A.1–A.5, we have that*

$$\begin{aligned} & K^{-1}\sum_{k=1}^{K}\widehat{E}_n^{(k)}\left[\left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{u}\right] \\ & -E\left[\left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{u}\right]=o_p\left\{(n\rho_n)^{-1/2}\right\} \end{aligned}$$

uniformly holds for $\boldsymbol{u}$ *with* $\|\boldsymbol{u}\|_2=1$.

Proof.

$$\begin{aligned} & K^{-1}\sum_{k=1}^{K}\widehat{E}_n^{(k)}\left[\left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{u}\right] \\ & -E\left[\left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{u}\right] \\ = \; & K^{-1}\sum_{k=1}^{K}\widehat{E}_n^{(k)}\left[\left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})\right\}^\top\boldsymbol{u}\right] \end{aligned}$$

$$
\begin{aligned}
& -E\left[\left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})\right\}^\top \boldsymbol{u}\right] \\
& +K^{-1}\sum_{k=1}^{K}\widehat{E}_n^{(k)}\left[\left\{S(\boldsymbol{\beta}^*;\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top \boldsymbol{u}\right] \\
& -E\left[\left\{S(\boldsymbol{\beta}^*;\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top \boldsymbol{u}\right] \\
= \;& I_{11}+I_{12}.
\end{aligned}
$$

For I_{11}, notice that

$$
\begin{aligned}
& \left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})\right\}^\top \boldsymbol{u} \\
= \;& \left\{b'(\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}})-b'(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\right\}\boldsymbol{X}^\top\boldsymbol{u}.
\end{aligned}
$$

Thus, by Taylor's expansion, we have

$$
\begin{aligned}
& K^{-1}\sum_{k=1}^{K}\widehat{E}_n^{(k)}\left[\left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})\right\}^\top \boldsymbol{u}\right] \\
& -E\left[\left\{S(\widehat{\boldsymbol{\beta}};\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})-S(\boldsymbol{\beta}^*;\widehat{Q}_{(-k)},\widehat{\pi}_{(-k)})\right\}^\top \boldsymbol{u}\right] \\
= \;& K^{-1}\sum_{k=1}^{K}\widehat{E}_n^{(k)}\left[b''(\boldsymbol{X}^\top\boldsymbol{b})\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right] \\
& -E\left[b''(\boldsymbol{X}^\top\boldsymbol{b})\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right],
\end{aligned}
$$

where $\boldsymbol{b}=\lambda\widehat{\boldsymbol{\beta}}+(1-\lambda)\boldsymbol{\beta}^*$. Notice that

$$
|b''(\boldsymbol{X}^\top\boldsymbol{b})-b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)|\leqslant C\lambda\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2.
$$

We have

$$
\begin{aligned}
& K^{-1}\sum_{k=1}^{K}\widehat{E}_n^{(k)}\left[b''(\boldsymbol{X}^\top\boldsymbol{b})\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right] \\
& -E\left[b''(\boldsymbol{X}^\top\boldsymbol{b})\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right] \\
= \;& K^{-1}\sum_{k=1}^{K}\widehat{E}_n^{(k)}\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right] \\
& -E\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right]+O_p\left\{\left[(p/n)^{1/2}+(n\rho_n)^{-\gamma_d}\right]^2\right\}.
\end{aligned}
$$

Write

$$
\sup_{\|\boldsymbol{u}\|_2=1}\left|K^{-1}\sum_{k=1}^{K}\widehat{E}_n^{(k)}\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right]\right.
$$

$$
\begin{aligned}
& -E\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{u}\right]\Big| \\
= & \left\| K^{-1}\sum_{k=1}^{K}(\widehat{E}_n^{(k)}-E)[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^\top](\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right\|_2
\end{aligned}
$$

Thus, we have

$$
I_{11} = O_p\left\{n^{-1/2}\left[(p/n)^{1/2}+(n\rho_n)^{-\gamma_d}\right]\right\}.
$$

For I_{12}, we write

$$
\begin{aligned}
I_{12} = & K^{-1}\sum_{k=1}^{K}\left(\widehat{E}_n^{(k)}-E\right)\left[\left(\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right)R\,(Q-Y)\,\boldsymbol{X}^\top\boldsymbol{u}\right] \\
& +K^{-1}\sum_{k=1}^{K}\left(\widehat{E}_n^{(k)}-E\right)\left[\left(R\pi_*^{-1}-1\right)\left(\widehat{Q}_{(-k)}-Q\right)\boldsymbol{X}^\top\boldsymbol{u}\right] \\
& +K^{-1}\sum_{k=1}^{K}\left(\widehat{E}_n^{(k)}-E\right)\left[\left(\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right)R\left(\widehat{Q}_{(-k)}-Q\right)\boldsymbol{X}^\top\boldsymbol{u}\right] \\
= & I_{121}+I_{122}+I_{123}.
\end{aligned}
$$

To bound each term, we use the maximal inequality in Lemma 19.38 in [37]. For the first term, the envelope function is

$$
F_{v,1} = C\left|\left(\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right)R\,(Q-Y)\right|.
$$

By calculation, $\|F_{v,1}\|_{P,2} \leqslant C'\left\|\left(\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right)R\right\|_{P,2}$ for some constant C'. By Lemma A.4, we have $\left\|\left(\widehat{\pi}_{(-k)}^{-1}-\pi_*^{-1}\right)R\right\|_{P,2} = o_p(\rho_n^{-1/2})$. Therefore, $I_{121} = o_p\left\{(n\rho_n)^{-1/2}\right\}$. Similarly, we have $I_{123} = o_p\left\{(n\rho_n)^{-1/2}\right\}$.

For the second term, the envelope function is

$$
F_{v,2} = C\left|\left(R\pi_*^{-1}-1\right)\left(\widehat{Q}_{(-k)}-Q\right)\right|.
$$

By the L_∞-convergence rate of $\widehat{Q}_{(-k)}$, $\|F_{v,2}\|_{P,2} = C\left\{E\left[\left(R\pi_*^{-1}-1\right)^2\right]\right\}^{1/2}(n\rho_n)^{-\gamma_d}$. Under Condition A.3, we have $E\left[\left(R\pi_*^{-1}-1\right)^2\right] \leqslant 4E\left[\pi_*^{-2}\rho_n w(\widetilde{\boldsymbol{X}})\right]+4 = O(\rho_n^{-1})$, we have

$$
I_{122} \lesssim O_p\left\{(n\rho_n)^{-1/2-\gamma_d}\right\}.
$$

Combining I_{121}, I_{122}, and I_{123}, we have

$$
I_{12} = o_p\left\{(n\rho_n)^{-1/2}\right\}.
$$

This concludes the proof. □

Theorem 3.1 shows that the one-step debiased estimator is asymptotically normal.

Proof. Denote

$$S_* = \widehat{E}_n \left[\{S(\boldsymbol{\beta}^*; Q, \pi_*)\}^\top \boldsymbol{v} \right].$$

First, we will show that

$$(n\rho_n)^{1/2} \bar{I}(\widetilde{\beta}_1 - \beta_1^* + S^*) = (n\rho_n)^{1/2} \left\{ \bar{I} \left(\widehat{\beta}_1 - \beta_1^* \right) - \bar{S} + S^* \right\}$$

is asymptotically normal when $P(R = 1 \mid \boldsymbol{Z}, \boldsymbol{X}) \neq P(R = 1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$; and is negligible when $P(R = 1 \mid \boldsymbol{Z}, \boldsymbol{X}) = P(R = 1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$.

We decompose the right-hand side by

$$\begin{aligned}
& (n\rho_n)^{1/2} \left\{ \bar{I} \left(\widehat{\beta}_1 - \beta_1^* \right) - \bar{S} + S^* \right\} \\
= \; & (n\rho_n)^{1/2} \left\{ I^* \left(\widehat{\beta}_1 - \beta_1^* \right) \right. \\
& \left. - K^{-1} \sum_{k=1}^{K} \widehat{E}_n^{(k)} \left[\left\{ S(\widehat{\boldsymbol{\beta}}; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)}) - S(\boldsymbol{\beta}^*; Q, \pi_*) \right\}^\top \boldsymbol{v} \right] \right\} \\
& + (n\rho_n)^{1/2} \left\{ K^{-1} \sum_{k=1}^{K} \widehat{E}_n^{(k)} \left[\left\{ S(\widehat{\boldsymbol{\beta}}; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)}) - S(\boldsymbol{\beta}^*; Q, \pi_*) \right\}^\top (\boldsymbol{v} - \widehat{\boldsymbol{v}}) \right] \right\} \\
& + (n\rho_n)^{1/2} \left\{ K^{-1} \sum_{k=1}^{K} \widehat{E}_n^{(k)} \left[\{ S(\boldsymbol{\beta}^*; Q, \pi_*) \}^\top (\boldsymbol{v} - \widehat{\boldsymbol{v}}) \right] \right\} \\
& + (n\rho_n)^{1/2} \left\{ (\bar{I} - I^*) \left(\widehat{\beta}_1 - \beta_1^* \right) \right\} \\
= \; & I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

By (the proof of) Lemma A.6 and A.5, we have

$$\begin{aligned}
I_1 \;=\; & (n\rho_n)^{1/2} \left[I^* \left(\widehat{\beta}_1 - \beta_1^* \right) - E \left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \boldsymbol{X}^\top \boldsymbol{v} \right] \right] \\
& - (n\rho_n)^{1/2} H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v}) \left\{ \mathrm{vec}(\widehat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma}) \right\} + o_p(1), \\
I_2 \;=\; & (n\rho_n)^{1/2} \left[K^{-1} \sum_{k=1}^{K} E \left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \boldsymbol{X}^\top (\boldsymbol{v} - \widehat{\boldsymbol{v}}) \right] \right] \\
& + (n\rho_n)^{1/2} E \left[(R\pi_*^{-1} - 1) \left\{ \widehat{g}_{(-k)}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) - g(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \right\} \boldsymbol{X}^\top (\boldsymbol{v} - \widehat{\boldsymbol{v}}) \right] \\
& + (n\rho_n)^{1/2} \left\{ H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \widehat{\boldsymbol{v}}) - H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v}) \right\} \left\{ \mathrm{vec}(\widehat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma}) \right\}.
\end{aligned}$$

For I_1, by the definition of v, we have

$$K^{-1} \sum_{k=1}^{K} E \left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \boldsymbol{X}^\top \boldsymbol{v} \right]$$

$$\begin{aligned} &= K^{-1}\sum_{k=1}^{K} E\left[b''(\boldsymbol{X}^\top\boldsymbol{\beta}^*)X_1(\widehat{\beta}_1-\beta_1^*)\boldsymbol{X}^\top\boldsymbol{v}\right] \\ &= I^*\left(\widehat{\beta}_1-\beta_1^*\right). \end{aligned}$$

Thus, when $P(R=1\mid \boldsymbol{Z},\boldsymbol{X}) \neq P(R=1\mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})$, we have that

$$I_1 = (n\rho_n)^{1/2}H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\left\{\mathrm{vec}(\widehat{\boldsymbol{\Gamma}})-\mathrm{vec}(\boldsymbol{\Gamma})\right\},$$

is asymptotic normal by Condition A.2; when $P(R=1\mid \boldsymbol{Z},\boldsymbol{X}) = P(R=1\mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})$, we can verify that $H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})=0$, and thus, I_1 is negligible.

For I_2, the first term is bounded by

$$(n\rho_n)^{1/2}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2\|\widehat{\boldsymbol{v}}-\boldsymbol{v}\|_2 = O_p\left\{(n\rho_n)^{1/2}\left[n^{-1/2}+(n\rho_n)^{-\gamma_d}\right]^2\right\} = o_p(1).$$

The second term is bounded by

$$(n\rho_n)^{1/2}\|\widehat{g}-g\|_\infty\|\widehat{\boldsymbol{v}}-\boldsymbol{v}\|_2 = O_p\left\{(n\rho_n)^{1/2}(n\rho_n)^{-\gamma_d}\left[n^{-1/2}+(n\rho_n)^{-\gamma_d}\right]\right\} = o_p(1).$$

The third term is bounded by

$$\begin{aligned} &(n\rho_n)^{1/2}\|\widehat{\boldsymbol{v}}-\boldsymbol{v}\|_2\|\mathrm{vec}(\widehat{\boldsymbol{\Gamma}})-\mathrm{vec}(\boldsymbol{\Gamma})\|_2 \\ =\ & O_p\left\{(n\rho_n)^{1/2}\left[n^{-1/2}+(n\rho_n)^{-\gamma_d}\right](n\rho_n)^{-1/2}\right\} \\ =\ & o_p(1). \end{aligned}$$

Thus, we have $I_2 = o_p(1)$.

For I_3, by Bernstein's inequality, we have

$$\left\|\widehat{E}_n^{(k)}\left[S(\boldsymbol{\beta}^*;Q,\pi_*)\right]\right\|_\infty = O_p\left\{(n\rho_n)^{-1/2}\right\}.$$

Thus, we have

$$I_3 = O_p\left\{(n\rho_n)^{1/2}(n\rho_n)^{-1/2}\left[n^{-1/2}+(n\rho_n)^{-\gamma_d}\right]\right\} = o_p(1).$$

For I_4, it is easy to see that

$$\left|\bar{I}-I^*\right| = O_p(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2+\|\widehat{\boldsymbol{v}}-\boldsymbol{v}\|_2).$$

Thus, we have

$$I_4 = O_p\left\{(n\rho_n)^{1/2}\left[n^{-1/2}+(n\rho_n)^{-\gamma_d}\right]^2\right\} = o_p(1).$$

Therefore,

$$\begin{aligned} &(n\rho_n)^{1/2}\left\{\bar{I}\left(\widehat{\beta}_1-\beta_1^*\right)-\bar{S}+S^*\right\} \\ =\ & -(n\rho_n)^{1/2}H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\left\{\mathrm{vec}(\widehat{\boldsymbol{\Gamma}})-\mathrm{vec}(\boldsymbol{\Gamma})\right\}+o_p(1); \end{aligned}$$

and $(n\rho_n)^{1/2}H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\left\{\text{vec}(\widehat{\mathbf{\Gamma}})-\text{vec}(\mathbf{\Gamma})\right\}=o_p(1)$ when $P(R=1\mid \boldsymbol{Z},\boldsymbol{X})=P(R=1\mid \mathbf{\Gamma}^\top\widetilde{\boldsymbol{X}})$.

Next, we will show that

$$(n\rho_n)^{1/2}\left[-S_*-H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\left\{\text{vec}(\widehat{\mathbf{\Gamma}})-\text{vec}(\mathbf{\Gamma})\right\}\right]\to N(0,\sigma_S^2),$$

where σ_S^2 is some positive constant. Since the asymptotic variance σ_S^2 depends on whether $\rho_n\to 0$ or $\rho_n=\rho>0$. Thus, we separate the discussion based on the value of ρ_n.

Notice that with

$$\begin{aligned}&H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})(\text{vec}(\widehat{\mathbf{\Gamma}})-\text{vec}(\mathbf{\Gamma})\\ =\;&\widehat{E}_n\left[R\rho_n^{-1}H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right]+o_p\left\{(n\rho_n)^{-1/2}\right\},\end{aligned}$$

we have

$$\begin{aligned}&-S_*-H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\left\{\text{vec}(\widehat{\mathbf{\Gamma}})-\text{vec}(\mathbf{\Gamma})\right\}\\ =\;&\widehat{E}_n\left[-\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{v}-R\rho_n^{-1}H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right].\end{aligned}$$

We will verify that

$$-\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{v}-R\rho_n^{-1}H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)$$

satisfies the Lindeberg condition.

First, we calculate the mean

$$E\left[-\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{v}-R\rho_n^{-1}H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right]\;=\;0;$$

and the variance

$$\begin{aligned}&E\left[-\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{v}-R\rho_n^{-1}H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right]^2\\ =\;&E\left[\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{v}\right]^2\\ &+E\left[R\rho_n^{-2}\left\{H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right\}^2\right]\\ &+2E\left[R\rho_n^{-1}\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{v}H_1(\text{vec}(\mathbf{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right].\end{aligned}$$

We calculate $E\left[\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{v}\right]^2$ as

$$\begin{aligned}&E\left[\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top\boldsymbol{v}\right]^2\\ =\;&E\left[\pi\pi_*^{-2}(Q-Y)^2\left(\boldsymbol{X}^\top\boldsymbol{v}\right)^2\right]+E\left[\left\{b'(\boldsymbol{X}^\top\boldsymbol{\beta}^*)-Q\right\}^2\left(\boldsymbol{X}^\top\boldsymbol{v}\right)^2\right]\end{aligned}$$

$$
\begin{aligned}
&= \rho_n^{-1} E\left[w(\widetilde{\boldsymbol{X}}) J_w^{-2}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) J^2(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})(Q-Y)^2 \left(\boldsymbol{X}^\top \boldsymbol{v}\right)^2\right] \\
&\quad + E\left[\left\{b'(\boldsymbol{X}^\top \boldsymbol{\beta}^*) - Q\right\}^2 \left(\boldsymbol{X}^\top \boldsymbol{v}\right)^2\right].
\end{aligned}
$$

The second term is

$$
\begin{aligned}
& E\left[R\rho_n^{-2}\left\{H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right\}^2\right] \\
= \; & \rho_n^{-1} E\left[w(\widetilde{\boldsymbol{X}})\left\{H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right\}^2\right].
\end{aligned}
$$

The third term is

$$
\begin{aligned}
& E\left[R\rho_n^{-1}\left\{S(\boldsymbol{\beta}^*; Q, \pi_*)\right\}^\top \boldsymbol{v} H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right] \\
= \; & \rho_n^{-1} E\left[w(\widetilde{\boldsymbol{X}}) J_w^{-1}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) J(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})(Q-Y)\left(\boldsymbol{X}^\top \boldsymbol{v}\right) H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right] \\
& + E\left[w(\widetilde{\boldsymbol{X}})\left\{b'(\boldsymbol{X}^\top \boldsymbol{\beta}^*) - Q\right\}\left(\boldsymbol{X}^\top \boldsymbol{v}\right) H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right].
\end{aligned}
$$

When $\rho_n \equiv \rho > 0$, define

$$
\begin{aligned}
& \sigma_S^2 \\
= \; & E\left[w(\widetilde{\boldsymbol{X}}) J_w^{-2}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) J^2(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})(Q-Y)^2 \left(\boldsymbol{X}^\top \boldsymbol{v}\right)^2\right] \\
& + \rho E\left[\left\{b'(\boldsymbol{X}^\top \boldsymbol{\beta}^*) - Q\right\}^2 \left(\boldsymbol{X}^\top \boldsymbol{v}\right)^2\right] \\
& + E\left[w(\widetilde{\boldsymbol{X}})\left\{H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right\}^2\right] \\
& + 2E\left[w(\widetilde{\boldsymbol{X}}) J_w^{-1}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) J(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})(Q-Y)\left(\boldsymbol{X}^\top \boldsymbol{v}\right) H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right] \\
& + 2\rho E\left[w(\widetilde{\boldsymbol{X}})\left\{b'(\boldsymbol{X}^\top \boldsymbol{\beta}^*) - Q\right\}\left(\boldsymbol{X}^\top \boldsymbol{v}\right) H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right].
\end{aligned}
$$

We have

$$
\begin{aligned}
& E\left[-\left\{S(\boldsymbol{\beta}^*; Q, \pi_*)\right\}^\top \boldsymbol{v} - R\rho_n^{-1} H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right]^2 \\
= \; & \rho^{-1}\sigma_S^2.
\end{aligned}
$$

When $P(R=1 \mid \boldsymbol{Z}, \boldsymbol{X}) = P(R=1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$, we have

$$
\begin{aligned}
& \sigma_S^2 \\
= \; & E\left[w(\widetilde{\boldsymbol{X}}) J_w^{-2}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) J^2(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})(Q-Y)^2 \left(\boldsymbol{X}^\top \boldsymbol{v}\right)^2\right] \\
& + \rho E\left[\left\{b'(\boldsymbol{X}^\top \boldsymbol{\beta}^*) - Q\right\}^2 \left(\boldsymbol{X}^\top \boldsymbol{v}\right)^2\right].
\end{aligned}
$$

Notice that

$$
\left|-\left\{S(\boldsymbol{\beta}^*; Q, \pi_*)\right\}^\top \boldsymbol{v} - R\rho_n^{-1} H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\psi(\boldsymbol{X}, \boldsymbol{Z}, Y)\right| \leqslant C\rho^{-1},
$$

for a large enough C. Thus, the Lindeberg condition satisfies because for any $\epsilon > 0$, we have

$$\begin{aligned} & 1\left\{\left|-\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top \boldsymbol{v} - R\rho_n^{-1}H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right| > \epsilon n^{1/2}\sigma_S\right\} \\ \leqslant\ & 1\left\{C\rho^{-1} \geqslant \epsilon n^{1/2}\sigma_S\right\} \\ =\ & 0, \end{aligned}$$

for a large enough n. By the Proposition 2.27 (Lindeberg-Feller CLT) of [37], we have

$$(n\rho)^{1/2}\left[-S_* - H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\left\{\mathrm{vec}(\widehat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma})\right\}\right] \to N(0,\sigma_S^2).$$

Thus, we have

$$(n\rho)^{1/2}(\widetilde{\beta}_1 - \beta_1^*) \to N(0,\sigma_1^2),$$

where $\sigma_1^2 = \sigma_S^2/\left(I^*\right)^2$. Especially, when $P(R=1 \mid \boldsymbol{Z},\boldsymbol{X}) = P(R=1 \mid \boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})$, we have

$$\rho^{-1}\sigma_S^2 = E\left[\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top \boldsymbol{v}\right]^2,$$

and $\pi_* = P(R=1 \mid \boldsymbol{Z},\boldsymbol{X})$.

When $\rho_n \to 0$, we define

$$\begin{aligned} & \sigma_S^2 \\ =\ & E\left[w(\widetilde{\boldsymbol{X}})J_w^{-2}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})J^2(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})(Q-Y)^2\left(\boldsymbol{X}^\top\boldsymbol{v}\right)^2\right] \\ & + E\left[w(\widetilde{\boldsymbol{X}})\left\{H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right\}^2\right] \\ & + 2E\left[w(\widetilde{\boldsymbol{X}})J_w^{-1}(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})J(\boldsymbol{\Gamma}^\top\widetilde{\boldsymbol{X}})(Q-Y)\left(\boldsymbol{X}^\top\boldsymbol{v}\right)H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right]. \end{aligned}$$

Notice that

$$E\left[-\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top \boldsymbol{v} - R\rho_n^{-1}H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right]^2 / (\rho_n^{-1}\sigma_S^2) \to 1.$$

In addition, we have

$$\left|-\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top \boldsymbol{v} - R\rho_n^{-1}H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right| \leqslant C\rho_n^{-1},$$

for a large enough C. We can also verify that the Lederberg condition satisfies because for any $\epsilon > 0$, we have

$$\begin{aligned} & 1\left\{\left|-\left\{S(\boldsymbol{\beta}^*;Q,\pi_*)\right\}^\top \boldsymbol{v} + R\rho_n^{-1}H_1(\mathrm{vec}(\boldsymbol{\Gamma});\boldsymbol{v})\psi(\boldsymbol{X},\boldsymbol{Z},Y)\right| > \epsilon n^{1/2}\rho_n^{-1/2}\sigma_S\right\} \\ \leqslant\ & 1\left\{C\rho_n^{-1} \geqslant \epsilon n^{1/2}\rho_n^{-1/2}\sigma_S\right\} \\ =\ & 0, \end{aligned}$$

for a large enough n given $n\rho_n \to +\infty$. By the Proposition 2.27 (Lindeberg-Feller CLT) of [37], we have

$$(n\rho)^{1/2}\left[-S_* - H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v})\left\{\mathrm{vec}(\widehat{\boldsymbol{\Gamma}}) - \mathrm{vec}(\boldsymbol{\Gamma})\right\}\right] \to N(0, \sigma_S^2).$$

Thus, we have

$$(n\rho_n)^{1/2}(\widetilde{\beta}_1 - \beta_1^*) \to N(0, \sigma_1^2),$$

where $\sigma_1^2 = \sigma_S^2/\left(I^*\right)^2$. Especially, when $P(R=1 \mid \boldsymbol{Z}, \boldsymbol{X}) = P(R=1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$, we have

$$\begin{aligned}\rho_n^{-1}\sigma_S^2 &= \rho_n^{-1} E\left[w(\widetilde{\boldsymbol{X}}) J_w^{-2}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) J^2(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})(Q-Y)^2 \left(\boldsymbol{X}^\top \boldsymbol{v}\right)^2\right] \\ &= E\left[R\pi_*^{-2}(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})(Q-Y)^2 \left(\boldsymbol{X}^\top \boldsymbol{v}\right)^2\right]\end{aligned}$$

and $\pi_* = P(R=1 \mid \boldsymbol{Z}, \boldsymbol{X})$. □

In Corollary 3.1, we show that the debiased estimator is locally efficient.

Proof. Notice that when $P(R=1 \mid \boldsymbol{Z}, \boldsymbol{X}) = P(R=1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$, we have $J_w(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$ bounded away from 0 and $\pi_*(\widetilde{\boldsymbol{X}}) = P(R=1 \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}})$. Thus, $H_1(\mathrm{vec}(\boldsymbol{\Gamma}); \boldsymbol{v}) = 0$.

We separate the discussions for the cases where $\rho_n = \rho > 0$ and $\rho_n \to 0$. When $\rho_n = \rho > 0$, we can verify that the asymptotic variance of $\widetilde{\beta}_1$ is

$$\boldsymbol{v}^\top E\left[\Omega^2 \boldsymbol{X}\boldsymbol{X}^\top\right] \boldsymbol{v}/(I^*)^2,$$

where

$$\begin{aligned}\Omega &= P^{-1}(R=1 \mid \boldsymbol{Z}, \boldsymbol{X}) R\{Y - b'(\boldsymbol{X}^\top \boldsymbol{\beta})\} \\ &\quad - P^{-1}(R=1 \mid \boldsymbol{Z}, \boldsymbol{X})\left\{R - P(R=1 \mid \boldsymbol{Z}, \boldsymbol{X})\right\}\left\{E\left[Y \mid \widetilde{\boldsymbol{X}}\right] - b'(\boldsymbol{X}^\top \boldsymbol{\beta})\right\}.\end{aligned}$$

By Theorem A.1, the semiparametric lower bound is given by $\mathrm{var}(\widetilde{\phi}_{1,\boldsymbol{Z}})$, i.e.,

$$\widetilde{\boldsymbol{u}}_0^\top E\left[\Omega^2 \boldsymbol{X}\boldsymbol{X}^\top\right] \widetilde{\boldsymbol{u}}_0,$$

where

$$\widetilde{\boldsymbol{u}}_0 = E^{-1}\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}\boldsymbol{X}^\top\right](1, 0, \cdots, 0)^\top.$$

Thus, to show that $\widetilde{\beta}_1$ achieves the semiparametric lower bound, we just need to verify $\widetilde{\boldsymbol{u}}_0 = \boldsymbol{v}/I^*$.

When $\rho_n \to 0$, we can verify that the asymptotic variance of $\widetilde{\beta}_1$ is of the form

$$\boldsymbol{v}^\top E\left[\widetilde{\Omega}^2 \boldsymbol{X}\boldsymbol{X}^\top\right] \boldsymbol{v}/(I^*)^2,$$

where

$$\widetilde{\Omega} \quad = \quad P^{-1}(R=1 \mid \boldsymbol{Z}, \boldsymbol{X}) R \left\{ Y - E\left[Y \mid \widetilde{\boldsymbol{X}} \right] \right\}.$$

By Theorem A.1, the semiparametric lower bound is given by $\mathrm{var}(\widetilde{\phi}_{2,\boldsymbol{Z}})$, i.e.,

$$\widetilde{\boldsymbol{u}}_0^\top E\left[\widetilde{\Omega}^2 \boldsymbol{X}\boldsymbol{X}^\top \right]) \widetilde{\boldsymbol{u}}_0.$$

Thus, to show that $\widetilde{\beta}_1$ achieves the semiparametric lower bound, we just need to verify $\widetilde{\boldsymbol{u}}_0 = \boldsymbol{v}/I^*$.

To show $\widetilde{\boldsymbol{u}}_0 = \boldsymbol{v}/I^*$, we compute each term directly. First, we can use the formula of the inverse of a block symmetric matrix to derive

$$E^{-1}\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}\boldsymbol{X}^\top \right] = (a, \boldsymbol{b}^\top; \boldsymbol{b}, \boldsymbol{c})/\mu,$$

where

$$\begin{aligned}
a \quad &= \quad 1, \\
\boldsymbol{b} \quad &= \quad -E^{-1}\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}_{-1}\boldsymbol{X}_{-1}^\top \right] E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1 \boldsymbol{X}_{-1} \right] \\
\boldsymbol{c} \quad &= \quad \mu E^{-1}\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}_{-1}\boldsymbol{X}_{-1}^\top \right] + \boldsymbol{b}\boldsymbol{b}^\top, \\
\mu \quad &= \quad E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1^2 \right] - \\
&\qquad E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1 \boldsymbol{X}_{-1}^\top \right] E^{-1}\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}_{-1}\boldsymbol{X}_{-1}^\top \right] E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1 \boldsymbol{X}_{-1} \right].
\end{aligned}$$

Thus, we have

$$\widetilde{\boldsymbol{u}}_0^\top = \left(1, -E^{-1}\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}_{-1}\boldsymbol{X}_{-1}^\top \right] E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1 \boldsymbol{X}_{-1} \right] \right) / \mu.$$

By the definition of v, we have

$$\boldsymbol{v}^\top = \left(1, -E^{-1}\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}_{-1}\boldsymbol{X}_{-1}^\top \right] E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1 \boldsymbol{X}_{-1} \right] \right).$$

Thus, to show $\widetilde{\boldsymbol{u}}_0 = \boldsymbol{v}/I^*$, it is sufficient to verify that $\mu = I^*$. Notice that

$$\begin{aligned}
I^* \quad &= \quad E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1 \boldsymbol{X}^\top \boldsymbol{v} \right] \\
&= \quad E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1^2 \right] - \\
&\qquad E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1 \boldsymbol{X}_{-1}^\top \right] E^{-1}\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) \boldsymbol{X}_{-1}\boldsymbol{X}_{-1}^\top \right] E\left[b''(\boldsymbol{X}^\top \boldsymbol{\beta}^*) X_1 \boldsymbol{X}_{-1} \right] \\
&= \quad \mu.
\end{aligned}$$

This concludes the proof. □

A.5. An example for Assumption A.2

In Assumption A.2, we require that there is a positive constant $\gamma_d > 1/4$ such that $\|\widehat{g}(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - Q(Z, \boldsymbol{X})\|_\infty = O_p\left\{(n\rho_n)^{-\gamma_d}\right\}$, and $\left\{\text{vec}(\widehat{\mathbf{\Gamma}}) - \text{vec}(\mathbf{\Gamma})\right\} = n^{-1}\sum_{i=1}^n 1\{R_i = 1\}\psi(\boldsymbol{X}_i, Z_i, Y_i) + o_p\left\{(n\rho_n)^{-1/2}\right\}$, where $\psi(\boldsymbol{X}_i, Z_i, Y_i)$ is bounded and vec$(\cdot)$ represents the vectorization of the matrix. In addition, we assume that $\sup_{\widetilde{\boldsymbol{X}}}\left|(\widehat{g} - g)(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - (\widehat{g} - g)(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}})\right| = o_p\left\{(n\rho_n)^{-1/2}\right\}$.

This assumption has two parts. The first part is the assumption on dimension reduction method, i.e., $\left\{\text{vec}(\widehat{\mathbf{\Gamma}}) - \text{vec}(\mathbf{\Gamma})\right\} = n^{-1}\sum_{i=1}^n 1\{R_i = 1\}\psi(\boldsymbol{X}_i, Z_i, Y_i) + o_p\left\{(n\rho_n)^{-1/2}\right\}$, where $\psi(\boldsymbol{X}_i, Z_i, Y_i)$ is bounded and vec$(\cdot)$ represents the vectorization of the matrix. This assumption is satisfied if the dimension reduction methods lead to asymptotically normal estimates (e.g., Sliced inverse regression [18], or minimum average variance estimation [40] and its variant [39]). The second part of the assumption on $\widehat{g}(\cdot)$ involve a common condition adopted by double machine learning literature [4], i.e., there is a positive constant $\gamma_d > 1/4$ such that $\|\widehat{g}(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - Q(Z, \boldsymbol{X})\|_\infty = O_p\left\{(n\rho_n)^{-\gamma_d}\right\}$.

In this section, we provide one of the examples to show that

$$\sup_{\widetilde{\boldsymbol{X}}}\left|(\widehat{g} - g)(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - (\widehat{g} - g)(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}})\right| = o_p\left\{(n\rho_n)^{-1/2}\right\}.$$

Specifically, we assume that $g(\cdot)$ can be written using basis functions, i.e.,

$$g(t) = \sum_{l=\infty}^{\infty} \alpha_l^* \phi_l(t),$$

where $\{\phi_l(\cdot)\}_{l=1}^\infty$ are basis functions. We assume that $\sum_{l=\infty}^\infty |\alpha_l^*| > +\infty$, and $\phi_l(t)$'s and their gradients are uniformly bounded. The gradient of $g(\cdot)$ is

$$\nabla g(t) = \sum_{l=\infty}^{\infty} \alpha_l^* \nabla\phi_l(t).$$

We consider to approximate $g(\cdot)$ and its gradient $\nabla g(\cdot)$ using L terms of basis functions, i.e.,

$$\begin{aligned} \widetilde{g}(t) &= \sum_{l=\infty}^{L} \alpha_l^* \phi_l(t), \\ \nabla\widetilde{g}(t) &= \sum_{l=\infty}^{L} \alpha_l^* \nabla\phi_l(t). \end{aligned}$$

We will show that if the approximation error of $\nabla\widetilde{g}(t)$ vanishes as L increases ($L \to +\infty$ as $n \to +\infty$), then any method satisfying that

$$\sum_{l=1}^{L} |\widehat{\alpha}_l - \alpha_l^*| = o_p(1).$$

will satisfy the required condition, where $\widehat{\alpha}_l$'s are estimators for α_l^*'s. To show this, notice that

$$\widehat{g}(\widehat{\mathbf{\Gamma}}^\top \tilde{X}) = \sum_{l=1}^{L} \widehat{\alpha}_l \phi_l(\widehat{\mathbf{\Gamma}}^\top \tilde{X}).$$

By direct calculation, we have

$$\begin{aligned}
& \left|(\widehat{g}-g)\,(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - (\widehat{g}-g)\,(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}})\right| \\
= & \left|\widehat{g}(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - \widehat{g}(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) - \left\{\nabla g(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\}^\top \left\{vec(\widehat{\mathbf{\Gamma}}) - vec(\mathbf{\Gamma})\right\}\right| \\
& + o_p\left\{(n\rho_n)^{-1/2}\right\} \\
= & \left|\sum_{l=1}^{L} \widehat{\alpha}_l \left\{\phi_l(\widehat{\mathbf{\Gamma}}^\top \widetilde{\boldsymbol{X}}) - \phi_l(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}})\right\} - \left\{\nabla g(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\}^\top \left\{vec(\widehat{\mathbf{\Gamma}}) - vec(\mathbf{\Gamma})\right\}\right| \\
& + o_p\left\{(n\rho_n)^{-1/2}\right\} \\
= & \left|\left\{\sum_{l=1}^{L} \widehat{\alpha}_l \nabla \phi_l(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\}^\top \left\{vec(\widehat{\mathbf{\Gamma}}) - vec(\mathbf{\Gamma})\right\}\right. \\
& \left. - \left\{\nabla g(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\}^\top \left\{vec(\widehat{\mathbf{\Gamma}}) - vec(\mathbf{\Gamma})\right\}\right| + o_p\left\{(n\rho_n)^{-1/2}\right\}.
\end{aligned}$$

Notice that

$$\begin{aligned}
& \left|\left\{\sum_{l=1}^{L} \widehat{\alpha}_l \nabla \phi_l(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\}^\top \left\{vec(\widehat{\mathbf{\Gamma}}) - vec(\mathbf{\Gamma})\right\}\right. \\
& \left. - \left\{\nabla g(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\}^\top \left\{vec(\widehat{\mathbf{\Gamma}}) - vec(\mathbf{\Gamma})\right\}\right| \\
\leqslant & \left|\left\{\sum_{l=1}^{L} \widehat{\alpha}_l \nabla \phi_l(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\} - \left\{\nabla g(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\}\right|_\infty \left\|vec(\widehat{\mathbf{\Gamma}}) - vec(\mathbf{\Gamma})\right\|_1 \\
\leqslant & \left|\left\{\sum_{l=1}^{L} \widehat{\alpha}_l \nabla \phi_l(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\} - \left\{\nabla g(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\}\right|_\infty O_p\left\{(n\rho_n)^{-1/2}\right\} \\
\leqslant & \left|\sum_{l=1}^{L} \widehat{\alpha}_l \nabla \phi_l(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}} - \sum_{l=1}^{L} \alpha_l^* \nabla \phi_l(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\|_\infty O_p\left\{(n\rho_n)^{-1/2}\right\} \\
& + \left\|\left\{\sum_{l=1}^{L} \alpha_l^* \nabla \phi_l(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\} - \left\{\nabla g(\mathbf{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}}\right\}\right\|_\infty O_p\left\{(n\rho_n)^{-1/2}\right\}.
\end{aligned}$$

The second term is $o_p\left\{(n\rho_n)^{-1/2}\right\}$ if the approximation error of $\nabla \widetilde{g}(t)$ vanishes

as L increases; the first term is $o_p\left\{(n\rho_n)^{-1/2}\right\}$ if

$$\begin{aligned} & \left\| \sum_{l=1}^{L} \widehat{\alpha}_l \nabla \phi_l(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}} - \sum_{l=1}^{L} \alpha_l^* \nabla \phi_l(\boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}) \otimes \widetilde{\boldsymbol{X}} \right\|_\infty \\ \leqslant \quad & C \sum_{l=1}^{L} |\widehat{\alpha}_l - \alpha_l^*| = o_p(1). \end{aligned}$$

A.6. An extension to a doubly robust procedure

The proposed method in the main text assumes that $Q(Z, \boldsymbol{X})$ is correctly specified. In this section, we can follow the idea in our main text to extend our proposed method by incorporating a working model for the true propensity, which enables a doubly robust estimation. However, since the extended algorithm is complicated (needs to fit nuisance parameters twice), we focus on only the case where we assume that $Q(Z, \boldsymbol{X})$ is correctly specified, i.e., $Y \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$ in the main text. In this extension, instead of assuming that $Y \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$ only, we assume that either $Y \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$ or $R \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^\top \widetilde{\boldsymbol{X}}$. To estimate such $\boldsymbol{\Gamma}$, we can conduct sufficient dimension reduction for Y and R, separately. Suppose that $\widehat{\boldsymbol{\Gamma}}_Y$ is the reduced dimension for Y and $\widehat{\boldsymbol{\Gamma}}_R$ is the reduced dimension for R, then we can set $\widehat{\boldsymbol{\Gamma}} = \left(\widehat{\boldsymbol{\Gamma}}_Y, \widehat{\boldsymbol{\Gamma}}_R\right)$. Based on the $\widehat{\boldsymbol{\Gamma}}$, we further modify how to construct the initial estimator (to ensure doubly robustness) and how to debias the initial estimator.

To obtain the initial estimator of $\boldsymbol{\beta}^*$, we regress Y on $\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}$ using kernel regressions and denote the estimated link function as $\widehat{g}$. Similarly, we regress R on $\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}}$ using kernel regressions and denote the estimated link function as $\widehat{m}$. Using the estimated imputation model and estimated propensity score, $\widehat{Q}(Z, \boldsymbol{X}) = \widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})$ and $\widehat{\pi}(Z, \boldsymbol{X}) = \widehat{m}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}})$, we obtain an initial estimate $\widehat{\boldsymbol{\beta}}$ by solving

$$\widehat{E}_n \left[S(\boldsymbol{\beta}; \widehat{\pi}, \widehat{Q}) \right] = 0.$$

Using the initial estimate, we can construct $\widehat{\boldsymbol{v}}^\top = (1, -\widehat{\boldsymbol{w}}^\top)$ following the procedure in the main text.

To remove the bias of the initial estimator due to the estimation of $\widehat{Q}$ and $\widehat{\pi}$ in the cross-fitting procedure, using the data excluding I_k, we will first obtain $\widehat{\pi}^{-1}_{(-k)}(\widetilde{\boldsymbol{x}}; \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{v}})$ following

$$\widehat{\pi}^{-1}_{(-k)}(\widetilde{\boldsymbol{x}}; \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{v}}) = \begin{cases} 1 + \left\{ \widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \widehat{\rho} \right\}^{-1} \widehat{J}_0^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}})(1 - \widehat{\rho}) & \left| \widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \right| > c_n, \\ \widehat{\rho}^{-1}, & \left| \widehat{J}_1^{(-k)}(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \right| \leqslant c_n. \end{cases}$$

Then we will fit $\widehat{Q}_{(-k)}(\widetilde{\boldsymbol{x}}; \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{v}})$ following

$$\widehat{Q}_{(-k)}(\widetilde{\boldsymbol{x}}; \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{v}}) = \frac{\widehat{E}_n^{(-k)}[Y \boldsymbol{X}^\top \widehat{\boldsymbol{\nu}} K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \mid R = 1]}{\widehat{E}_n^{(-k)}[\boldsymbol{X}^\top \widehat{\boldsymbol{\nu}} K_\hbar(\widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{\Gamma}}^\top \widetilde{\boldsymbol{x}}) \mid R = 1]}$$

Then, the one-step debiased estimator is $\widetilde{\beta}_1 = \widehat{\beta}_1 - \bar{I}^{-1}\bar{S}$, where

$$\bar{S} = \sum_{k=1}^{K} S^{(k)}/K, \quad S^{(k)} = \widehat{E}_n^{(k)}\left[\left\{S(\widehat{\boldsymbol{\beta}}; \widehat{Q}_{(-k)}, \widehat{\pi}_{(-k)})\right\}^{\top} \widehat{\boldsymbol{v}}\right].$$

It can be shown that the limit of $\widehat{\boldsymbol{\Gamma}}$, $\boldsymbol{\Gamma}$, satisfies either $Y \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{X}}$ or $R \perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{\Gamma}^{\top}\widetilde{\boldsymbol{X}}$, we will have that either $\widehat{Q}(Z, \boldsymbol{X})$ converges to $E[Y \mid \widetilde{\boldsymbol{X}}]$ or $\widehat{\pi}(Z, \boldsymbol{X})$ converges to $P(R = 1 \mid \widetilde{\boldsymbol{X}})$, and thus the initial estimator $\widehat{\boldsymbol{\beta}}$ is consistent. Further, we will also have that either $\widehat{Q}_{(-k)}(\widetilde{\boldsymbol{X}}; \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{v}})$ converges to $E[Y \mid \widetilde{\boldsymbol{X}}]$ or $\widehat{\pi}_{(-k)}(\widetilde{\boldsymbol{X}}; \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{v}})$ converges to $P(R = 1 \mid \widetilde{\boldsymbol{X}})$, and thus the debiased estimator $\widetilde{\beta}_1$ is also consistent. The asymptotic normality of $\widetilde{\beta}_1$ will be left for future investigation.

A.7. Simulations with the random forest

We use the random forest with the default parameter setting in the *randomForest* R package [19] and evaluate the four methods using the random forest as the method to fit the imputation model. We choose the simulation scenario where we have a binary outcome with a high missing rate and $n = 1000$. Table 5 summarizes the coverage, bias, and standard deviations of the coefficient estimates, and Figure 2 exhibits the deviance. Our proposed method achieves nominal coverage and has the smallest deviance among other methods. In terms of the bias and standard deviations, our proposed method has a smaller or comparable bias and standard deviation compared with Baseline 1 and Baseline 2.

TABLE 5
Coverage, Bias, and standard deviation when the imputation model is fitted using the random forest.

Binary, Missing rate of 90%, $n = 1000$				
	Coverage			
	β_1	β_2	β_3	β_4
Baseline 1	0.816	0.816	0.848	0.868
Baseline 2	0.854	0.852	0.850	0.866
Proposed w/o Z	0.958	0.950	0.974	0.978
Proposed with Z	0.940	0.948	0.948	0.962
	Bias			
	β_1	β_2	β_3	β_4
Baseline 1	9.862	8.470	-9.641	-8.353
Baseline 2	0.104	0.101	-0.239	-0.204
Proposed w/o Z	-0.071	-0.072	0.069	0.115
Proposed with Z	-0.192	-0.123	0.133	0.128
	Standard Deviation			
	β_1	β_2	β_3	β_4
Baseline 1	9.137	9.648	9.580	9.327
Baseline 2	1.054	1.017	0.938	0.826
Proposed w/o Z	1.374	1.357	1.198	1.210
Proposed with Z	0.815	0.768	0.383	0.415

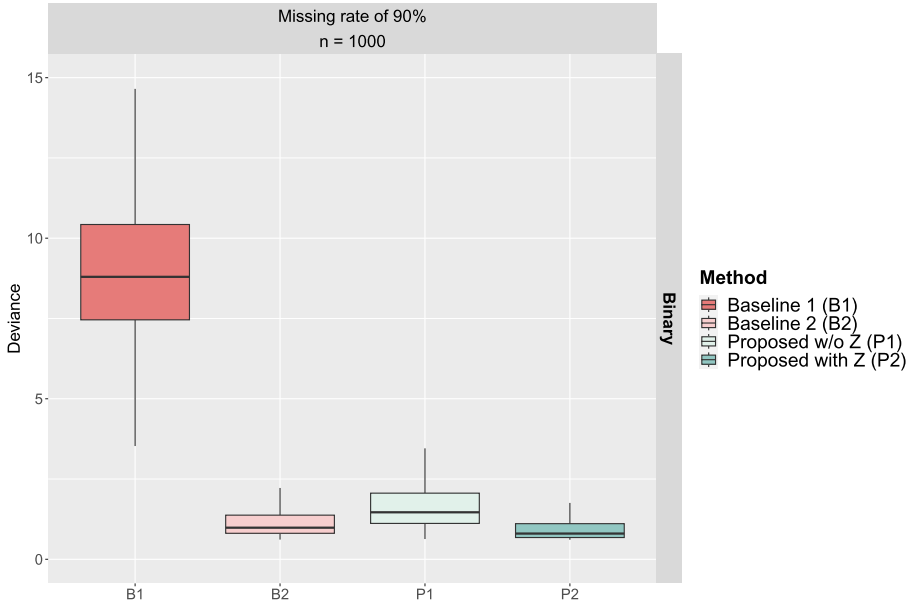

FIG 2. *Deviance when the imputation model is fitted using the random forest.*

A.8. Selected risk factors in the real data example

Table 6 reports the coefficients estimates and their 95%-confidence intervals for the predictive model developed using the proposed approach. The model provides some interesting observations. Elderly patients were significantly more likely to achieve the MCID, and non-white patients may be more likely to achieve the MCID. These findings can potentially be used to support patient-provider shared decision-making.

TABLE 6
Risk factors of the failure of the MCID identified from Proposed w/ Z.

Covariates	Coefficients (95% Confidence Interval)
Intercept	-1.449 (-1.880, -1.018)
Age	0.367 (0.020, 0.714)
Race - Not White	0.867 (-0.008, 1.743)
Income (ref: <40,000)	
40,000 - 60,000	-0.071 (-0.792, 0.650)
>60,000	0.332 (-0.344, 1.008)
Cardiovascular	-0.203 (-1.161, 0.755)
Respiratory	0.378 (-0.821, 1.577)
Weight	-0.419 (-1.289, 0.451)
BMI	-0.133 (-0.412, 0.146)
No. of orthopedics visits	-0.128 (-0.504, 0.247)

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved the quality of this paper. We also thank Dr. Jiwei Zhao and Dr. Yanyao Yi for their helpful suggestions on earlier versions of this work. Dr. Jaeyoung Park conducted this research during his PhD studies at the University of Florida under the supervision of Dr. Muxuan Liang and Dr. Xiang Zhong. Dr. Muxuan Liang is the co-first and corresponding author of this paper.

References

[1] Anderer, A., Bastani, H. and Silberholz, J. (2022). Adaptive clinical trial designs with surrogates: When should we bother? *Management Science* **68** 1982–2002.

[2] Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96** 723–734. MR2538768

[3] Cheng, D., Ananthakrishnan, A. N. and Cai, T. (2021). Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics* **77** 413–423. MR4307644

[4] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21** C1–C68. https://doi.org/10.1111/ectj.12097 MR3769544

[5] Cook, R. D. (2007). Fisher lecture: dimension reduction in regression. *Statistical Science* **22** 1–26. https://doi.org/10.1214/088342306000000682 MR2408655

[6] Fleming, T. R., Prentice, R. L., Pepe, M. S. and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* **13** 955–968.

[7] Fontana, M. A., Lyman, S., Sarker, G. K., Padgett, D. E. and MacLean, C. H. (2019). Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clinical Orthopaedics and Related Research* **477** 1267.

[8] Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

[9] Han, P. (2012). A note on improving the efficiency of inverse probability weighted estimator using the augmentation term. *Statistics & Probability Letters* **82** 2221–2228. MR2979759

[10] Han, P., Wang, L. and Song, P. X.-K. (2016). Doubly robust and locally efficient estimation with missing outcomes. *Statistica Sinica* 691–719. MR3497767

[11] Higuera, C. A., Elsharkawy, K., Klika, A. K., Brocone, M. and Barsoum, W. K. (2011). 2010 mid-america orthopaedic association physi-

cian in training award: predictors of early adverse outcomes after knee and hip arthroplasty in geriatric patients. *Clinical Orthopaedics and Related Research®* **469** 1391–1400.

[12] Hirshberg, D. A. and Wager, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics* **49** 3206–3227. MR4352528

[13] Ho, A., Purdie, C., Tirosh, O. and Tran, P. (2019). Improving the response rate of patient-reported outcome measures in an Australian tertiary metropolitan hospital. *Patient Related Outcome Measures* **10** 217.

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

[15] Hou, J., Guo, Z. and Cai, T. (2023). Surrogate assisted semi-supervised inference for high dimensional risk prediction. *Journal of Machine Learning Research* **24** 1–58. MR4664702

[16] Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22** 523–539. MR2420458

[17] Katakam, A., Karhade, A. V., Collins, A., Shin, D., Bragdon, C., Chen, A. F., Melnic, C. M., Schwab, J. H. and Bedair, H. S. (2022). Development of machine learning algorithms to predict achievement of minimal clinically important difference for the KOOS-PS following total knee arthroplasty. *Journal of Orthopaedic Research®* **40** 808–815.

[18] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327. https://doi.org/10.1080/01621459.1991.10475035 MR1137117

[19] Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News* **2** 18–22.

[20] Liu, M., Zhang, Y., Liao, K. P. and Cai, T. (2023). Augmented transfer regression learning with semi-non-parametric nuisance models. *Journal of Machine Learning Research* **24** 1–50. MR4664730

[21] Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107** 168–179. https://doi.org/10.1080/01621459.2011.646925 MR2949349

[22] Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics* **41** 250–268. https://doi.org/10.1214/12-AOS1072 MR3059417

[23] Malkani, A. L., Dilworth, B., Ong, K., Baykal, D., Lau, E., Mackin, T. N. and Lee, G.-C. (2017). High risk of readmission in octogenarians undergoing primary hip arthroplasty. *Clinical Orthopaedics and Related Research®* **475** 2878–2888.

[24] Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45** 158–195. https://doi.org/10.1214/16-AOS1448 MR3611489

[25] Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8** 431–440.

[26] Pronk, Y., Pilot, P., Brinkman, J. M., van Heerwaarden, R. J.

and VAN DER WEEGEN, W. (2019). Response rate and costs for automated patient-reported outcomes collection alone compared to combined automated and manual collection. *Journal of Patient-reported Outcomes* **3** 1–8.

[27] QIN, J., SHAO, J. and ZHANG, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association* **103** 797–810. MR2524011

[28] QIN, J. and ZHANG, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 101–122. MR2301502

[29] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89** 846–866. MR1294730

[30] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

[31] ROTNITZKY, A., LEI, Q., SUED, M. and ROBINS, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99** 439–456. MR2931264

[32] RUBIN, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. MR2117498

[33] RUBIN, D. B. and VAN DER LAAN, M. J. (2008). Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics* **4**. MR2399288

[34] SHAO, J. (2003). *Mathematical statistics*. Springer Science & Business Media.

[35] TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101** 1619–1637. MR2279484

[36] TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97** 661–682. MR2672490

[37] VAN DER VAART, A. W. (2000). *Asymptotic statistics*. Cambridge university press. MR1652247

[38] VAN DER VAART, A. and WELLNER, J. A. (2011). A local maximal inequality under uniform entropy. *Electronic Journal of Statistics* **5** 192. MR2792551

[39] XIA, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics* **35** 2654–2690. https://doi.org/10.1214/009053607000000352 MR2382662

[40] XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 363–410. https://doi.org/10.1111/1467-9868.03411 MR1924297