Can we tell if the justification of the validity of Wald confidence intervals of doubly robust functionals may be incorrect, without assumptions?

LIN LIU[1], RAJARSHI MUKHERJEE[2], AND JAMES M. ROBINS[3]

Abstract. In this article we study and generalize the assumption-lean tests in Liu et al. (2020a) that can *refute an analyst's justification* for the validity of a reported nominal $(1-\alpha)$ Wald confidence interval (CI) centered at a double machine learning (DML) estimator for doubly robust functionals (DRFs) studied by Rotnitzky et al. (2021). The class of DRFs is broad and of central importance in economics and biostatistics. It strictly includes both (i) the class of mean-square continuous functionals that can be written as an expectation of an affine functional of a conditional expectation studied by Chernozhukov et al. (2018c) and the class of functionals studied by Robins et al. (2008). The state-of-the-art estimators for DRFs ψ are DML estimators $\widehat{\psi}$. The bias of DR estimators depends on the product of the rates at which two nuisance functions b and p are estimated. Most commonly an analyst justifies the validity of their Wald CIs by proving that, under their complexity-reducing assumptions, the Cauchy-Schwarz (CS) upper bound $\mathsf{E}[(\widehat{b}(X)-b(X))^2]^{1/2}\mathsf{E}[(\widehat{p}(X)-p(X))^2]^{1/2}$ for the bias of $\widehat{\psi}$ is $o(n^{-1/2})$. Thus if we can falsify the hypothesis H_0: the CS upper bound is $o(n^{-1/2})$, we refute the analyst's justification for the validity of her Wald CIs. In this work, we exhibit a valid test of H_0, without relying on complexity-reducing assumptions on (b, p) and their estimates $(\widehat{b}, \widehat{p})$. Furthermore, we also develop a test of a different hypothesis from H_0 if refined DML estimators are used, for which the CS upper bound is not used to justify the theoretical properties. To the best of our knowledge, these proposed tests are the first of their kind. Simulation experiments are performed to demonstrate how the proposed assumption-lean test can be used in practice.

Key words: Econometrics and Machine Learning, Assumption-lean Statistics, Doubly Robust Functionals, Higher-Order Influence Functions, U-Statistics

[1] Institute of Natural Sciences, School of Mathematical Sciences, CMA-Shanghai, MOE-LSC, SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University linliu@sjtu.edu.cn.

[2] Department of Biostatistics, Harvard University, ram521@mail.harvard.edu.

[3] Department of Epidemiology and Biostatistics, Harvard University, robins@hsph.harvard.edu.

This work is based on a talk given by the last author in the conference of "Whitney Newey's Contributions to Econometrics" held at MIT in 2019.

1. Introduction

In this paper, we consider the following problem: What can we tell about the coverage of Wald confidence intervals (CIs) for a class of low dimensional functionals of high dimensional multivariate distributions that are of great interest in economics and statistics, using empirical data alone, without any essential assumptions?

Specifically, we consider the general class of mixed bias functionals or doubly robust functionals (DRFs) studied in Rotnitzky et al. (2021). This class strictly includes both (i) the class of mean-square continuous functionals that can be written as an expectation of an affine functional of a conditional expectation studied by Chernozhukov et al. (2018c) and the class of functionals studied by Robins et al. (2008). Many DRFs are of substantive scientific and economic interests. One obvious example is $\psi(\theta) = -\mathsf{E}_\theta[Y(a=1)]$, the (minus[1]) counterfactual mean of an outcome Y when the binary treatment A is set to 1 under strong ignorability. Here θ denotes the nuisance parameters: in this case, θ contains the outcome regression function $b(x) = \mathsf{E}_\theta[Y|X=x, A=1]$, the inverse propensity score $p(x) = \{\mathsf{E}_\theta[A|X=x]\}^{-1}$ and the conditional density $g(x|A=1)$ of X in the treatment group, where X is the (potentially high-dimensional) covariates with dimension d.

There is now a fast-growing literature in statistics and econometrics on estimating and drawing statistical inference for (certain members of) DRFs, by establishing sufficient (and sometimes necessary) conditions to guarantee the $\sqrt{n}$-consistency and/or semiparametric efficiency of $\psi(\theta)$. Those conditions mainly concern with the smoothness (Bradic et al., 2019a; Liu et al., 2017; Robins et al., 2009), sparsity (Bradic et al., 2019b; Chernozhukov et al., 2018a; Farrell, 2015; Smucler et al., 2019), or other (Chen et al., 2020; Farrell et al., 2021) related complexity reducing assumptions of the nuisance functions. One of the most popular ones is the standard Double Machine Learning (DML) estimator (Chernozhukov et al., 2018a; Smucler et al., 2019) (abbreviated as the "standard DML estimator"), denoted as $\widehat{\psi}_{1,\mathsf{cf}}$.More specifically, the data is randomly divided into two (or more) samples – the estimation sample of size n and the training or nuisance sample of size $n_{\mathsf{tr}} = N - n$ with $1 - c > n/N > c$ for some $c \in (0,1)$. To simplify the exposition we take c to be $1/2$. Machine learning estimators $\widehat{b}(x)$ of the outcome regression $b(x)$ and $\widehat{p}(x)$ of the inverse propensity score $p(x)$ are fit using the training sample data. The one step estimator of $\psi(\theta)$ is $\widehat{\psi}_1 = \psi(\widehat{\theta}) + \mathsf{P}_n\left[\mathsf{IF}_{1,\psi}(\widehat{\theta})\right] = \mathsf{P}_n\left[-\widehat{b}(X) - A\widehat{p}(X)(Y - \widehat{b}(X))\right]$ of $\psi(\theta)$ is computed from the estimation sample. Then the cross-fitting DML estimator $\widehat{\psi}_{1,\mathsf{cf}}$ is simply an arithmetic average between $\widehat{\psi}_1$ and $\widehat{\psi}_1'$, which is the same as $\widehat{\psi}_1$ except the roles of training and estimation samples are reversed. In general a data analyst who reports the Wald CI $\widehat{\psi}_{1,\mathsf{cf}} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_{1,\mathsf{cf}})$ for an DRF justifies its validity by (i) imposing complexity-reducing assumptions on the nuisance functions b and p and then

[1] The minus sign plays no essential role and is added purely for notational convenience that will be made clear later.

(ii) appealing to theorems wherein the asymptotic validity of the Wald CI has been proved under these assumptions. In this paper, we focus more on the non-cross-fitting version $\widehat{\psi}_1$ for simplicity.

Most notably, standard DML estimators $\widehat{\psi}_1$ can have bias $o(n^{-1/2})$, which are needed for the validity of the Wald CI, under quite weak complexity reducing assumptions on the nuisance functions. This is because the bias of $\widehat{\psi}_1$ is a product of estimation errors $\widehat{b} - b$ and $\widehat{p} - p$, where $\widehat{b}$ and $\widehat{p}$ are b and p's estimates from the nuisance sample. To be concrete, when b and p are s_b- and s_p-Hölder smooth with $s_b = s_p = s > d/2$, then up to constant, there exists estimators $\widehat{b}$ and $\widehat{p}$ (e.g. wavelet projections), that converge to the true b and p only at the nonparametric $O(n^{-\frac{s}{d+2s}}) = o(n^{-\frac{1}{4}})$ rate in L_2-norm, yet the absolute bias of $\widehat{\psi}_1$ conditioning on the training sample is bounded by

(1.1)

$$\begin{aligned}|\mathsf{Bias}_\theta(\widehat{\psi}_1)| = \left|\mathsf{E}_\theta[A(\widehat{b}(X) - b(X))(\widehat{p}(X) - p(X))]\right| &\overset{(*)}{\leqslant} \left\{\mathsf{E}_\theta[A(\widehat{b}(X) - b(X))^2]\right\}^{1/2} \left\{\mathsf{E}_\theta[A(\widehat{p}(X) - p(X))^2]\right\}^{1/2} \\ &= O\left(n^{-\frac{2s}{d+2s}}\right) = o(n^{-1/2}),\end{aligned}$$

where $(*)$ follows from Cauchy-Schwarz (CS) inequality. So after cross-fitting, $\widehat{\psi}_{1,\mathsf{cf}}$ is semiparametric efficient and the $(1-\alpha)$ Wald CI $\widehat{\mathsf{CI}}_\alpha$ centered around $\widehat{\psi}_{1,\mathsf{cf}}$ has nominal coverage asymptotically. Therefore, the key to **justify** the asymptotic validity of $\widehat{\mathsf{CI}}_\alpha$ is via the CS inequality and the $L_2(\mathsf{P}_\theta)$ convergence rates of the nuisance estimators.

Some recent works also consider further refinement of standard DML estimators using techniques including but not limited to higher order influence function (HOIF) (Robins et al., 2008) based estimators (Liu et al., 2017), multiple sample splitting and/or undersmoothing (Cattaneo and Jansson, 2018; Cattaneo et al., 2019; Kennedy, 2020; Kline et al., 2020; Newey and Robins, 2018), and convex programming (Bradic et al., 2019a). These refined estimators are able to achieve semiparametric efficiency under weaker assumptions than standard DML estimators, or sometimes even under minimal assumptions (Bradic et al., 2019a; Liu et al., 2017; Robins et al., 2009). For convenience, we call these refined estimators the "refined DML estimators", to be distinguished from the standard DML estimators.

Given this state of affairs, estimation and statistical inference of DRFs appear to be relatively well understood. However, the devil is often in the detail. First, the complexity reducing assumptions on the underlying true nuisance parameters b and p in the above works are generally not empirically checkable without making further complexity-reducing assumptions. Second, in view of the increasing popularity of deep neural networks (DNNs) in statistics and econometrics (Chen et al., 2020; Chernozhukov et al., 2021; Farrell et al., 2021), despite the tremendous theoretical progress made in recent years, their statistical properties are yet well-understood[2]. Even when the complexity reducing assumptions on b and p actually hold, we cannot guarantee the desired convergence rates of the DNN or other complex machine

[2] For example, the implicit regularization due to the training dynamics by SGD is often not incorporated when deriving the convergence rates of DNNs under nonparametric regression frameworks, an issue also brought up in Farrell et al. (2021).

learning estimators $\widehat{b}$ and $\widehat{p}$. Third, except for Liu et al. (2017), the more refined DML estimators often rely on carefully crafted nuisance estimators $\widehat{b}$ and $\widehat{p}$ with nice algebraic structures for the bias analysis to bypass the CS inequality step as in (1.1). When complex machine learning methods are used, the desired structures of $\widehat{b}$ and $\widehat{p}$ are rarely preserved, and hence it is unclear if these refined estimators still work well in practice.

Is it at all possible to make valid inference on $\psi(\theta)$ without complexity reducing assumptions on the nuisance parameters and their estimates? Unfortunately, Robins and Ritov (1997), Robins et al. (2009) and Ritov et al. (2014) proved that, if the parameter space Θ does not entail any restriction on the complexity of b and p [beyond the assumption that the propensity score $1/p(X)$ is bounded away from zero, which is needed to insure a positive semiparametric variance bound], then (i) the bias of any estimator of $\psi(\theta) = -\mathsf{E}_\theta[b(X)] = -\mathsf{E}_\theta[Y(a = 1)]$ may be order 1 [i.e., there does not exist an estimator of $\psi(\theta)$ such that the bias of estimator converges to zero as $N \to \infty$ for all $\theta \in \Theta$]; and (ii) there does not exist a uniform (in $\theta \in \Theta$) consistent test of the composite null hypothesis that b and p satisfy given smoothness or sparsity assumptions. From (ii) it follows that no matter the sample size N, for any $\alpha^\dagger$-level test of this composite null, there exists $\theta = (b, p, \theta \setminus \{b, p\})$ in the complement to the null such that the power under P_θ is less than or equal to $\alpha^\dagger$. From (i) it follows that there exists no valid nominal $(1 - \alpha)$ CI whose median length converges to 0 for all $\theta \in \Theta$ as the sample size $N \to \infty$. In this sense, without imposing possibly incorrect restrictive assumptions on b and p, no meaningful inference concerning $\psi(\theta)$ is possible.

Given the above concerns, we take a radically different perspective from all the above works. Specifically, we adopt a *Skeptic's Approach to Statistics and Empirical Sciences* and ask the following question:

> From the empirical data alone, can we tell if the justification of the validity of a Wald CI centered around a DML estimator of a DRF may be incorrect, without making complexity-reducing assumptions on the nuisance parameters?

To be concrete, suppose, as is often the case, the analyst justifies the validity of the Wald CI $\widehat{\mathsf{CI}}_\alpha(\widehat{\psi}_1) := \widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_1)$ and thus its cross-fit version $\widehat{\mathsf{CI}}_\alpha(\widehat{\psi}_{1,\mathsf{cf}}) := \widehat{\psi}_{1,\mathsf{cf}} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_{\mathsf{cf},1})$ by following (1.1) and proving that, under her complexity reducing assumptions, the Cauchy-Schwarz (CS) bias functional

$$\mathsf{CSBias}_\theta(\widehat{\psi}_1) = \left\{\mathsf{E}_\theta[A(\widehat{b}(X) - b(X))^2]\right\}^{1/2} \left\{\mathsf{E}_\theta[A(\widehat{p}(X) - p(X))^2]\right\}^{1/2} \tag{1.2}$$

is $o(n^{-1/2})$ conditional on the training sample data[3] (and thus also on the functions $\widehat{b}, \widehat{p}$ computed from the training sample). It then follows if we could empirically show that $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$ exceeds some given

[3] In this paper, essentially all expectations and probabilities are to be understood as being conditional on the training sample (except when explicitly noted otherwise). Hence we can and do omit this conditioning event in our notation. The randomness in the training sample will also be ignored.

multiple $\delta > 0$, say $\delta = 3/4$, times $\mathsf{s.e.}_\theta[\widehat{\psi}_1]$, the conditional standard error of $\widehat{\psi}_1$ of order $n^{-1/2}$, then we have falsified the justification for the analyst's claim that her Wald CI is valid, or equivalently has the nominal coverage probability.

In this paper, we show that the answer to the question posed above is a firmly "yes" when the Wald CI is centered around standard DML estimators. In a nutshell, instead of directly targeting $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$ in (1.2), which depends on the unknown b and p, our approach targets the following lower bound of $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$:

$$
\begin{aligned}
&|\underbrace{\mathsf{E}_\theta[\Pi_\theta[A(\widehat{b}-b)|A\bar{\mathsf{z}}_k](X)\Pi_\theta[A(\widehat{p}-p)|A\bar{\mathsf{z}}_k](X)]}_{=:\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}| \\
&\leqslant \underbrace{\{\mathsf{E}_\theta[\Pi_\theta[A(\widehat{b}-b)|A\bar{\mathsf{z}}_k](X)^2]\}^{1/2}\{\mathsf{E}_\theta[\Pi_\theta[A(\widehat{p}-p)|A\bar{\mathsf{z}}_k](X)^2]\}^{1/2}}_{=:\mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)} \leqslant \mathsf{CSBias}_\theta(\widehat{\psi}_1)
\end{aligned}
\tag{1.3}
$$

where the first inequality again follows from CS inequality and the second inequality follows from the fact that projection contracts norms. Here for any $h \in L_2(\mathsf{P}_{\theta,X})$, $\Pi_\theta[h|A\bar{\mathsf{z}}_k]$ denotes the projection of h onto the linear span of a k-dimensional basis functions $A\bar{\mathsf{z}}_k = A(\mathsf{v}_1, \cdots, \mathsf{v}_k)^\top$. Later we will show that the above lower bound $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ can be consistently estimated by a third-order U-statistic with sufficiently fast rate, which forms the basis of our proposed approach to empirically refute the justification (1.1). More formally, we will construct a test of the following null hypothesis

$$
\mathsf{H}_{0,\mathsf{CS}}(\delta) : \mathsf{CSBias}_\theta(\widehat{\psi}_1) \leqslant \delta \mathsf{s.e.}_\theta(\widehat{\psi}_1)
\tag{1.4}
$$

by targeting a tempered null hypothesis[4]

$$
\mathsf{H}_{0,k}(\delta) : |\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)| \leqslant \delta \mathsf{s.e.}_\theta(\widehat{\psi}_1).
\tag{1.6}
$$

Note that $\mathsf{H}_{0,k}(\delta)$ is a tempered version of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ in the following sense: if $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ is true, then $\mathsf{H}_{0,k}(\delta)$ is true; if $\mathsf{H}_{0,k}(\delta)$ is false, then $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ is false; but the converse of the above two clauses do not necessarily hold.

Main contribution. A preliminary version of the above idea has appeared in Liu et al. (2020a), but the results therein are not as sharp. The main contributions of this paper are four-fold. First, we extend the results in Liu et al. (2020a) to the more realistic case where the distribution of the potentially high-dimensional covariates X is unknown. Liu et al. (2020a) restricted attention to the case where this distribution was known, except for an abbreviated discussion of the unknown case in an online supplement. In particular, we will show that a test statistic based on a third-order U-statistic (which is the third-order

[4] The reason why we do not focus on a "less-tempered" null hypothesis

$$
\mathsf{H}_{0,k,\mathsf{CS}}(\delta) : \mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1) \leqslant \delta \mathsf{s.e.}_\theta(\widehat{\psi}_1).
\tag{1.5}
$$

will be explained in Remark 3.5 of Section 3. In principle, though, one can also consider testing $\mathsf{H}_{0,k,\mathsf{CS}}(\delta)$.

influence function of $\mathsf{Bias}_\theta(\widehat{\psi}_1)$) can be used to refute the justification (1.1) with desired statistical guarantees without knowing the distribution of X. Second, the results in this paper require much weaker assumptions[5] on the basis functions $\bar{\mathsf{z}}_k$ than those in Liu et al. (2020a), allowing the proposed methodology to be used more broadly in substantive studies. Third, we consider the same problem not only for the standard DML estimators, but also for the refined DML estimators in this paper, whereas Liu et al. (2020a) did not make such a distinction. Fourth, we extend the results of Liu et al. (2020a) for expected conditional (co)variance to the entire DRF class. This also requires we derive the form of higher order influence function (HOIF) (Robins et al., 2008; van der Vaart, 2014) estimators for functionals in the DRF class. The HOIFs for the DRF class may be of independent interest, considering the role of HOIFs in constructing minimax rate-optimal estimators for smooth/differentiable functionals (Liu et al., 2017; Robins et al., 2017).

The remainder of the paper is arranged as follows. In Section 2, we describe the mathematical setup formally and review the definition and properties of DRFs recently characterized in Rotnitzky et al. (2021). In Section 2.1, we review the statistical properties of their standard DML estimators, based upon which we motivate and formally define our approach.

In Section 3, we first derive the second- and third-order influence functions of DRFs, denoted as $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$, following the notation used in Robins et al. (2008). Here k is the dimension of the basis functions $\bar{\mathsf{z}}_k$ and $\widehat{\Sigma}_k$ is the empirical estimate of the population λ-weighted Gram matrix $\Sigma_k \coloneqq \mathsf{E}_\theta[\lambda(X)\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]$, where the weight function $\lambda(x)$ appears in the definition of DRFs and will be made clear in Section 2. These statistics are estimators of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$, and hence are the building blocks of the test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ (1.4). The statistical properties of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ will be studied in this section, including their bias, variance, and distributional limits as estimators. These preparatory results will culminate at a valid $\alpha^\dagger$-level test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ based on $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$, which we denote as $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$, a notation that will be defined later in this paper. This is the test that we recommend analysts use to refute the justification for the validity of the Wald CI $\widehat{\mathsf{CI}}_\alpha(\widehat{\psi}_1)$, centered around the standard DML estimator $\widehat{\psi}_{1,\mathsf{cf}}$.

In Section 4, we consider whether our framework can be useful even for refined DML estimators, which do not rely on the justification in (1.1) by bypassing the CS inequality. Since different refined DML estimators are developed under different complexity-reducing assumptions on the nuisance parameters, and are geared towards different algebraic structures of the nuisance estimators, it is unclear (at least to us) if it is at all possible to treat these estimators in a unified manner as we have done for standard DML estimators. As a result, we only consider the estimator with multiple sample splitting plus undersmoothing, first investigated by Newey and Robins (2018), as a case study. In this case, we argue that

[5] Note that both the assumptions in this paper and in Liu et al. (2020a) are easily checkable by the analysts; see Conditions W and SW.

the tempered null hypothesis $\mathsf{H}_{0,k}(\delta)$ should be a more natural null hypothesis of actual interest and construct a valid $\alpha^{\dagger}$-level test of $\mathsf{H}_{0,k}(\delta)$ based on higher order U-statistics. Since higher order U-statistics are computational costly, we also propose an early-stopping strategy that takes the analyst's computational budget into account.

In Section 5 we present results of simulation studies to evaluate the finite sample performance of our methods. Section 6 concludes with a discussion of some open problems. Many of the technical details are deferred to the online supplementary materials.

2. Doubly robust functionals, their influence functions and standard DML estimators

The formal setup is as follows. We observe N i.i.d. copies of the data vector $O = (W, X)$ drawn from some unknown probability distribution P_θ belonging to a (locally) nonparametric model

$$\mathcal{M} = \{\mathsf{P}_\theta; \theta = (b, p, \theta \setminus \{b, p\}) \in \Theta = \mathcal{B} \times \mathcal{P} \times \Theta \setminus \{\mathcal{B}, \mathcal{P}\}\}$$

parameterized by the parameter $\theta = (b, p, \theta \setminus \{b, p\})$ with $b, p, \theta \setminus \{b, p\}$ variation independent. X is a d-dimensional random vector with compact support whose density is bounded away from 0 and ∞ on its support, and d is allowed to increase with N. Here the maps $x \mapsto b(x) \in \mathcal{B}$ and $x \mapsto p(x) \in \mathcal{P}$ have range contained in $\mathbb{R}$. $\mathcal{M}$ is a non-parametric model in the sense that we do not impose conditions on $\mathcal{B}$ and $\mathcal{P}$ except that $\mathcal{B}, \mathcal{P} \subseteq L_2(\mathsf{P}_{\theta,X})$ where $L_2(\mathsf{P}_{\theta,X})$ denotes the space of functions over X that are square integrable with respect to the marginal distribution $\mathsf{P}_{\theta,X}$ of X.

To avoid extraneous technical issues, we assume that observed data O is bounded with probability 1 (see Remark 3.2 for further discussion). We consider functionals (i.e. parameters) $\psi : \Theta \to \mathbb{R}$ that possess a (first order) influence function[6] $\mathsf{IF}_{1,\psi}(\theta) = \mathsf{if}_{1,\psi}(O; \theta)$ (and thus a positive semiparametric variance bound (Ichimura and Newey, 2021; Newey, 1990)) and are contained in the mixed bias or doubly robust class of functionals of Rotnitzky et al. (2021) defined as follows.

Definition 2.1 (Definition of a mixed bias or doubly robust functional (DRF) (Definition 1 of Rotnitzky et al. (2021))). *$\psi(\theta)$ is a doubly robust functional if, for each $\theta \in \Theta$ there exists $b : x \mapsto b(x) \in \mathcal{B}$ and $p : x \mapsto p(x) \in \mathcal{P}$ such that (i) $\theta = (b, p, \theta \setminus \{b, p\})$ and $\Theta = \mathcal{B} \times \mathcal{P} \times \Theta \setminus \{\mathcal{B}, \mathcal{P}\}$ and (ii) for any θ, θ'*

$$\psi(\theta) - \psi(\theta') + \mathsf{E}_\theta\left[\mathsf{IF}_{1,\psi}(\theta')\right] = \mathsf{E}_\theta\left[S_{bp}(b(X) - b'(X))(p(X) - p'(X))\right] \tag{2.1}$$

where $S_{bp} \equiv s_{bp}(O)$ with $o \mapsto s_{bp}(o)$ a known function that does not depend on θ or θ' satisfying either $\mathsf{P}_\theta(S_{bp} \geqslant 0) = 1$ or $\mathsf{P}_\theta(S_{bp} \leqslant 0) = 1$. b and p are called nuisance parameters in the semiparametric statistics literature, a terminology we also adopt in this paper. We also denote $\lambda(X) := \mathsf{E}_\theta[S_{bp}|X]$.

[6] The term "influence function" when used without further qualification is to be understood to be the first order influence function.

To understand the importance of DRFs, let P_N be the empirical expectation operator and suppose θ' were an estimate of θ from a separate sample (which we regard as fixed, i.e. non-random). It then follows that the one step estimator $\psi(\theta') + \mathsf{P}_N\left[\mathsf{IF}_{1,\psi}(\theta')\right]$ is doubly robust (Bang and Robins, 2005; Chernozhukov et al., 2016; Robins and Rotnitzky, 2001; Scharfstein et al., 1999a,b). That is, by (2.1), it is unbiased for $\psi(\theta)$ under P_θ if either $b = b'$ or $p = p'$. Because of this fact, we will use the term DRF in this paper, as is done in much of the current literature, instead of the "mixed bias" terminology employed in Rotnitzky et al. (2021). To ease notation, we will restrict consideration to DRFs for which $\mathsf{P}_\theta(S_{bp} \geqslant 0) = 1$. For a DRF $\psi^\dagger(\theta)$ of substantive interest for which $\mathsf{P}_\theta(S_{bp} \leqslant 0) = 1$, we will instead analyze $\psi(\theta) = -\psi^\dagger(\theta)$.

Let $W = (Y, A)$. Below are some examples of DRFs that are of substantive interest in economics and statistics.

(1) The counterfactual mean of Y when $\{0, 1\}$-valued A is set to 1 (under strong ignorability), $\mathsf{E}_\theta[Y(a = 1)]$ is identified from the distribution of $W = (A, AY, X)$ by the DRF $\psi^\dagger(\theta) = \mathsf{E}_\theta[b(X)]$ with $b(x) = \mathsf{E}_\theta[Y|X = x, A = 1]$, $p(x) = 1/\mathsf{E}_\theta[A|X = x]$, $S_{bp} = -A$, and $\lambda(x) = -p(x)^{-1}$. Here $p(x)$ is the inverse of the propensity score $\pi(x) = \mathsf{E}_\theta[A|X = x]$. Because $S_{bp} = -A$, we instead analyze the DRF

$$\psi(\theta) = -\psi^\dagger(\theta) = -\mathsf{E}_\theta[b(X)] = -\mathsf{E}_\theta[Y(a = 1)]$$

for which $S_{bp} = A$. We will use $\psi(\theta) = -\mathsf{E}_\theta[b(X)] = -\mathsf{E}_\theta[Y(a = 1)]$ as a running example below[7]. Average treatment effect, one of the most intensively studied causal parameters, is thus a sum of two DRFs.

(2) The expected conditional covariance

$$\psi(\theta) = \mathsf{E}_\theta[(Y - b(X))(A - p(X))]$$

with $b(x) = \mathsf{E}_\theta[Y|X = x], p(x) = \mathsf{E}_\theta[A|X = x]$ is a DRF (equivalently, a MB functional) with $S_{bp} = \lambda(X) \equiv 1$. When Y and $b(X)$ are replaced by A and $p(X)$, $\psi(\theta)$ reduces to the expected conditional variance $\phi(\theta) = \mathsf{E}_\theta[(A - p(X))^2]$. When A is binary, the variance-weighted average treatment effect, another commonly encountered causal parameter, is thus the ratio $\frac{\psi(\theta)}{\phi(\theta)}$ of two DRFs (Robins et al., 2008).

Before proceeding further, we collect some frequently used notation. We use $\mathsf{E}_\theta[\cdot], \mathsf{var}_\theta[\cdot]$ and etc. to denote the expectation, variance, and etc. with respect to P_θ. The data is randomly divided into an estimation sample and a training (equivalently, nuisance) sample of size $n = N/2$. To avoid notational clutter, all expectations, variances, and probabilities are conditional on the training sample unless otherwise stated. For a (random) vector V, $\|V\|_\theta \equiv \mathsf{E}_\theta[V^{\otimes 2}]^{1/2} = \mathsf{E}_\theta[V^\top V]^{1/2}$ denotes its $L_2(\mathsf{P}_\theta)$ norm conditioning on the training sample, $\|V\| \equiv (V^{\otimes 2})^{1/2} = (V^\top V)^{1/2}$ its ℓ_2 norm and $\|V\|_\infty$ its $L_\infty(\mathsf{P}_\theta)$

[7] Chernozhukov et al. (2018c) used an alternative decomposition under which they included our A, X in their X. See Example 1 of Rotnitzky et al. (2021).

norm. For any matrix A, $\|A\|$ will be reserved for its operator norm. Given an integer k, and a random vector $\bar{\mathsf{z}}_k = \bar{\mathsf{z}}_k(X)$, $\Pi_\theta[\cdot|\bar{\mathsf{z}}_k]$ denotes the population linear projection operator onto the linear space spanned by $\bar{\mathsf{z}}_k$ conditioning on the training sample, and $\Pi_\theta^\perp[\cdot|\bar{\mathsf{z}}_k] = (\mathsf{I} - \Pi_\theta)[\cdot|\bar{\mathsf{z}}_k]$ is the projection onto the ortho-complement of $\bar{\mathsf{z}}_k$ in the Hilbert space $L_2(\mathsf{P}_{\theta,X})$ where $\mathsf{P}_{\theta,X}$ denotes the marginal law of X under θ. That is, for a random variable V,

$$\Pi_\theta[V|\bar{\mathsf{z}}_k](x) = \bar{\mathsf{z}}_k^\top(x)\{\mathsf{E}_\theta[\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]\}^{-1}\mathsf{E}_\theta[\bar{\mathsf{z}}_k(X)V], \Pi_\theta^\perp[V|\bar{\mathsf{z}}_k](x) = V - \Pi_\theta[V|\bar{\mathsf{z}}_k](x). \tag{2.2}$$

The following common asymptotic notations are used throughout the paper: $x \lesssim y$ (equivalently $x = O(y)$ or $y = \Omega(x)$) denotes that there exists some constant $C > 0$ such that $x \leqslant Cy$, $x \asymp y$ (equivalently $x = \Theta(y)$) means there exist some constants $c_1 > c_2 > 0$ such that $c_2|y| \leqslant |x| \leqslant c_1|y|$. $x = o(y)$ or $y = \omega(x)$ or $y \gg x$ or $x \ll y$ is equivalent to $\lim_{x,y\to\infty} \frac{x}{y} = 0$. For a random variable X_n with law P possibly depending on the sample size n, $X_n = O_{\mathsf{P}}(a_n)$ denotes that X_n/a_n is bounded in P-probability, and $X_n = o_{\mathsf{P}}(a_n)$ means that $\lim_{n\to\infty} \mathsf{P}(|X_n/a_n| \geqslant \epsilon) = 0$ for every positive real number ϵ.

2.1. Influence functions of DRFs and natural estimators of nuisance functions.

In this section, we first review the influence functions of DRFs (see Definition 2.1) established in Rotnitzky et al. (2021). Although the definition of DRF is quite abstract, Rotnitzky et al. (2021) derived the (nonparametric) influence functions of DRFs by directly leveraging the form of the product bias appeared in Definition 2.1, which we summarize their results below for the sake of completeness.

Theorem 2.1 (Theorems 1 and 2 of Rotnitzky et al. (2021))**.** *Suppose (1) $\psi(\theta)$ for $\theta = (b, p, \theta \setminus \{b, p\}) \in \Theta \equiv \mathcal{B} \times \mathcal{P} \times [\Theta \setminus \{\mathcal{B}, \mathcal{P}\}]$ is a DRF (equivalently MB functional) according to Definition 2.1; (2) the regularity Condition A.1 in Appendix A.1 (Condition 1 of Rotnitzky et al. (2021)) holds; (3) the functions b, p, λb and λp belong to $L_2(\mathsf{P}_{\theta,X})$ for all $\theta \in \Theta$.*

Then there exists a statistic S_0 and linear maps $h \mapsto m_1(O, h)$ for $h \in \mathcal{B}$ and $h \mapsto m_2(O, h)$ for $h \in \mathcal{P}$ independent of θ satisfying the following:

- *The influence function of $\psi(\theta)$ is given by:*

$$\mathsf{IF}_{1,\psi}(\theta) = \mathcal{H}(b, p) - \psi(\theta) \tag{2.3}$$

where $\mathcal{H}(b, p) \coloneqq S_{bp}b(X)p(X) + m_1(O, b) + m_2(O, p) + S_0$;

- *Furthermore, the following first-order doubly robust moment conditions hold:*

$$\begin{cases} \mathsf{E}_\theta[S_{bp}h(X)p(X) + m_1(O, h)] = 0 \ \text{ for all } h \in \mathcal{B}, \\ \mathsf{E}_\theta[S_{bp}b(X)h(X) + m_2(O, h)] = 0 \ \text{ for all } h \in \mathcal{P}. \end{cases} \tag{2.4}$$

And $\psi(\theta) = \mathsf{E}_\theta[m_1(O, b) + S_0] = \mathsf{E}_\theta[m_2(O, p) + S_0] = \mathsf{E}_\theta[-S_{bp}b(X)p(X) + S_0]$.

- *Further suppose the maps $h \mapsto \mathsf{E}_\theta[m_1(O,h)]$ and $h \mapsto \mathsf{E}_\theta[m_2(O,h)]$ for $h \in L_2(\mathsf{P}_{\theta,X})$ are continuous and linear with Riesz representers $\mathcal{R}_1(X) \equiv \mathcal{R}_1(X;\theta)$ and $\mathcal{R}_2(X) \equiv \mathcal{R}_2(X;\theta)$[8] respectively. Then $b(X) = -\mathcal{R}_2(X)/\lambda(X)$ and $p(X) = -\mathcal{R}_1(X)/\lambda(X)$; moreover, b and p are the (global) minimizers of the following minimization problems*

$$
\begin{aligned}
b(\cdot) &= \mathsf{arg\ min}_{h \in L_2(\mathsf{P}_{\theta,X})} \mathsf{E}_\theta \left[S_{bp} \frac{h(X)^2}{2} + m_2(O,h) \right], \\
p(\cdot) &= \mathsf{arg\ min}_{h \in L_2(\mathsf{P}_{\theta,X})} \mathsf{E}_\theta \left[S_{bp} \frac{h(X)^2}{2} + m_1(O,h) \right].
\end{aligned}
\tag{2.5}
$$

We first give some examples of the influence functions $\mathsf{IF}_{1,\psi}(\theta)$ and Riesz representers of DRFs.

Example 1 (Some examples of Riesz representers for DRFs)**.**

(1) $\psi(\theta) = -\mathsf{E}_\theta[b(X)] = -\mathsf{E}_\theta[Y(a=1)]$. Then $\mathsf{IF}_{1,\psi}(\theta) = \mathcal{H}(b,p) - \psi(\theta)$ with

$$\mathcal{H}(b,p) = -\left(-Ab(X)p(X) + b(X) + AYp(X)\right).$$

Here $S_{bp} = A$ and $\lambda(X) = p(X)^{-1}$, $m_1(O,b) = -b(X)$, $m_2(O,p) = -AYp(X)$ and $S_0 = 0$. Then $\mathcal{R}_1(X) = -1$ and $\mathcal{R}_2(X) = -b(X)/p(X)$ since

$$
\begin{cases}
\mathsf{E}_\theta\left[m_1(O,h)\right] = \mathsf{E}_\theta[-h(X)], \\
\mathsf{E}_\theta\left[m_2(O,h)\right] = \mathsf{E}_\theta[-AYh(X)] = \mathsf{E}_\theta\left[-\left\{p(X)\right\}^{-1} b(X)h(X)\right].
\end{cases}
$$

(2) $\psi(\theta) = \mathsf{E}_\theta\left[(Y - b(X))(A - p(X))\right]$ – expected conditional covariance between A and Y given X with $b(X) = \mathsf{E}_\theta[Y|X], p(X) = \mathsf{E}_\theta[A|X]$: $\mathsf{IF}_{1,\psi}(\theta) = \mathcal{H}(b,p) - \psi(\theta)$ with $\mathcal{H}(b,p) = b(X)p(X) - Ab(X) - Yp(X) + AY$, $S_{bp} = 1$ and $\lambda(X) = 1$, $m_1(O,b) = -Ab(X)$, $m_2(O,p) = -Yp(X)$, and $S_0 = AY$. Then $\mathcal{R}_1(X) = -p(X)$ and $\mathcal{R}_2(X) = -b(X)$ since

$$
\begin{cases}
\mathsf{E}_\theta\left[m_1(O,h)\right] = \mathsf{E}_\theta[-Ah(X)] = \mathsf{E}_\theta[-p(X)h(X)], \\
\mathsf{E}_\theta\left[m_2(O,h)\right] = \mathsf{E}_\theta[-Yh(X)] = \mathsf{E}_\theta[-b(X)h(X)].
\end{cases}
$$

We now briefly comment on the three parts of Theorem 2.1 as they are all important for future development of the paper.

Equation (2.3) exhibits the general formula of influence functions of DRFs, which are the basic building blocks for standard DML estimators and most refined DML estimators. Our framework heavily relies on higher order influence functions, which can be derived from the (first-order) influence functions.

Equations (2.4) and (2.5) together suggest natural loss functions that could be used to fit the nuisance functions b and p from data. To see why, we first make the following important observation:

[8] The Riesz representer $\mathcal{R}(X)$ of a continuous linear functional $h \mapsto \mathsf{E}[m(O,h)]$ for $h \in L_2(\mathsf{P}_X)$ is, by definition, the function of X satisfying $\mathsf{E}[m(O,h)] = \mathsf{E}\left[\mathcal{R}(X)h(X)\right]$; see Chernozhukov et al. (2018b,c); Rotnitzky et al. (2021).

Lemma 2.1. *The minimization problems given in* (2.5)*, with the function class* $L_2(\mathsf{P}_{\theta,X})$ *replaced by some* $\mathcal{F} \subset L_2(\mathsf{P}_{\theta,X})$

$$\begin{aligned} \widetilde{b}(\cdot) &= \text{arg min}_{h \in \mathcal{F}} \mathsf{E}_\theta \left[S_{bp} \frac{h(X)^2}{2} + m_2(O,h) \right], \\ \widetilde{p}(\cdot) &= \text{arg min}_{h \in \mathcal{F}} \mathsf{E}_\theta \left[S_{bp} \frac{h(X)^2}{2} + m_1(O,h) \right] \end{aligned} \tag{2.6}$$

are equivalent to:

$$\begin{aligned} \widetilde{b}(\cdot) &= \text{arg min}_{h \in \mathcal{F}} \mathsf{E}_\theta \left[\lambda(X)(b(X) - h(X))^2 \right], \\ \widetilde{p}(\cdot) &= \text{arg min}_{h \in \mathcal{F}} \mathsf{E}_\theta \left[\lambda(X)(p(X) - h(X))^2 \right]. \end{aligned} \tag{2.7}$$

Proof. The following algebra makes this statement explicit:

$$\begin{aligned} \frac{1}{2}\mathsf{E}_\theta[\lambda(X)(b(X) - h(X))^2] &= \frac{1}{2}\mathsf{E}_\theta[S_{bp}(b(X) - h(X))^2] \\ &= \mathsf{E}_\theta[S_{bp}\frac{h(X)^2}{2} - S_{bp}b(X)h(X)] + \frac{1}{2}\mathsf{E}_\theta[S_{bp}b(X)^2] \\ &= \mathsf{E}_\theta[S_{bp}\frac{h(X)^2}{2} + m_2(O,h)] + \frac{1}{2}\mathsf{E}_\theta[S_{bp}b(X)^2] \end{aligned}$$

where in the third line we use equation (2.4) and the extra term $\frac{1}{2}\mathsf{E}_\theta[S_{bp}b(X)^2]$ is unrelated to minimizing $\frac{1}{2}\mathsf{E}_\theta[\lambda(X)(b(X) - h(X))^2]$. By symmetry,

$$\frac{1}{2}\mathsf{E}_\theta[\lambda(X)(p(X) - h(X))^2] = \mathsf{E}_\theta[S_{bp}\frac{h(X)^2}{2} + m_1(O,h)] + \frac{1}{2}\mathsf{E}_\theta[S_{bp}p(X)^2].$$

■

Lemma 2.1 thus motivates the **Squared Error Loss Minimization Strategy** defined below:

Squared Error Loss Minimization Strategy is any strategy in which $\widehat{b}(x)$ and $\widehat{p}(x)$ are near minimizers of the expected λ-weighted ($\mathsf{E}_\theta[\lambda(X)\{\widehat{b}(X)-b(X)\}^2]$ and $\mathsf{E}_\theta[\lambda(X)\{\widehat{p}(X)-p(X)\}^2]$) squared error losses over the class of functions chosen by the analysts. Hence to construct $\widehat{b}$ and $\widehat{p}$, it is then natural to use an ML algorithm to try to obtain the global minimizers:

$$\begin{aligned} \widehat{b} &= \underset{h \in \mathcal{F}}{\text{arg min}} \mathsf{P}_{n_{\mathsf{tr}}} \left[S_{bp}\frac{h^2}{2} + m_2(O,h) \right], \\ \widehat{p} &= \underset{h \in \mathcal{F}}{\text{arg min}} \mathsf{P}_{n_{\mathsf{tr}}} \left[S_{bp}\frac{h^2}{2} + m_1(O,h) \right]. \end{aligned} \tag{2.8}$$

where $\mathcal{F}$ is the set of functions computable by the ML algorithm.

Researchers who use the ML algorithm with the objective functions in (2.8) are explicitly acknowledging that they have adopted the **Squared Error Loss Minimization Strategy**.

Remark 2.1. *The* ***Squared Error Loss Minimization Strategy*** *is not restrictive.*

(1) *As discussed in detail in Appendix A.2, many commonly used loss functions are equivalent to using the above* ***Squared Error Loss Minimization Strategy****: e.g. (1) when $p(X)$ is the inverse propensity score of a binary treatment A, the commonly used cross-entropy loss is asymptotically equivalent to the above λ-weighted squared error loss; (2) the unweighted squared error losses $\mathsf{E}_\theta[\{\widehat{b}(X) - b(X)\}^2]$ and $\mathsf{E}_\theta[\{\widehat{p}(X) - p(X)\}^2]$ are equivalent to the λ-weighted squared error losses up to constant.*

(2) *We also want to point out the connection to recent works on using adversarial losses (but with unconstrained discriminator $h \in L_2(\mathsf{P}_{\theta,X})$) (Chernozhukov et al., 2020; Ghassami et al., 2021):*

$$
\begin{aligned}
&\mathsf{min}_{b' \in \mathcal{F}} \mathsf{max}_{h \in L_2(\mathsf{P}_{\theta,X})} \mathsf{E}_\theta \left[S_{bp} b(X) h(X) + m_2(O, h) \right] - \|h\|^2 \\
&\mathsf{min}_{p' \in \mathcal{F}} \mathsf{max}_{h \in L_2(\mathsf{P}_{\theta,X})} \mathsf{E}_\theta \left[S_{bp} p(X) h(X) + m_1(O, h) \right] - \|h\|^2
\end{aligned}
\tag{2.9}
$$

to train the nuisance estimators, which essentially leverages the moment conditions in (2.4). Denote the minimizers of the above adversarial losses as $\widetilde{b}_{\text{min-max}}, \widetilde{p}_{\text{min-max}}$. As shown in Theorem 1 of Ghassami et al. (2021), $\widetilde{b}_{\text{min-max}}$ and $\widetilde{p}_{\text{min-max}}$ are also the minimizers to the losses in (2.5) when $L_2(\mathsf{P}_{\theta,X})$ is replaced by $\mathcal{F}$. Therefore, minimizing adversarial losses also belongs to the ***Squared Error Loss Minimization Strategy****.*

2.2. Standard DML estimators.

The following algorithm defines the standard DML estimators $\widehat{\psi}_1$ and $\widehat{\psi}_{\mathsf{cf},1}$ of a DRF $\psi(\theta)$ satisfying the suppositions of Theorem 2.1.

(i) The N study subjects are randomly split into two parts: an estimation sample of size n and a training (nuisance) sample of size $n_{tr} = N - n$ with $n/N \approx 1/2$. Without loss of generality we shall assume that $i = 1, \ldots, n$ corresponds to the estimation sample.

(ii) Nuisance estimators $\widehat{b}$ and $\widehat{p}$ are constructed from the training sample data using ML methods and define $\widehat{\theta} := (\widehat{b}, \widehat{p}, \theta \setminus \{b, p\})$[9].

(iii) Denote $\mathbb{IF}_1 \equiv \mathbb{IF}_1(\theta) := \frac{1}{n}\sum_{i=1}^{n} \mathsf{IF}_{1,\psi}(\theta)$ and $\widehat{\mathbb{IF}}_1 \equiv \mathbb{IF}_1(\widehat{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \mathsf{IF}_{1,\psi}(\widehat{\theta})$. Then

$$
\widehat{\psi}_1 = \widehat{\mathbb{IF}}_1 + \psi(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{H}_i(\widehat{b}, \widehat{p}) = \frac{1}{n}\sum_{i=1}^{n} \left(S_{bp,i} \widehat{b}(X_i) \widehat{p}(X_i) + m_1(O_i, \widehat{b}) + m_2(O_i, \widehat{p}) + S_{0,i} \right)
$$

from n subjects in the estimation sample and $\widehat{\psi}_{\mathsf{cf},1} = \dfrac{1}{2}\left(\widehat{\psi}_1 + \overline{\widehat{\psi}}_1\right)$ where $\overline{\widehat{\psi}}_1$ is $\widehat{\psi}_1$ but with the training and estimation samples reversed.

The following theorem provides asymptotic properties of standard DML estimators $\widehat{\psi}_1$ (and $\widehat{\psi}_{1,\mathsf{cf}}$) of DRFs.

[9] Note that unlike $\{\widehat{b}, \widehat{p}\}$, $\theta \setminus \{b, p\}$ in the definition of $\widehat{\theta}$ is not estimated from data in the standard DML estimator. In fact, in this paper, we only need $\widehat{\theta}$ to depend on estimated b and p because none of the estimators/tests in this paper will depend on estimates of $\theta \setminus \{b, p\}$.

Theorem 2.2. *Under the conditions of Theorem 2.1, conditional on the training sample, if a) $\widehat{\psi}_1$ has conditional bias*

(2.10)
$$\mathsf{Bias}_\theta(\widehat{\psi}_1) := \mathsf{E}_\theta[S_{bp}(b(X) - \widehat{b}(X))(p(X) - \widehat{p}(X))] \equiv \mathsf{E}_\theta[\lambda(X)(b(X) - \widehat{b}(X))(p(X) - \widehat{p}(X))] = o(n^{-1/2})$$

and b) $\widehat{b}(x)$ and $\widehat{p}(x)$ converge to $b(x)$ and $p(x)$ in $L_2(\mathsf{P}_\theta)$, then:

(1) *$\widehat{\psi}_1 - \psi(\theta) = n^{-1}\sum_{i=1}^{n} \mathsf{IF}_{1,\psi,i}(\theta) + o(n^{-1/2})$ and $\widehat{\psi}_{\mathsf{cf},1} - \psi(\theta) = N^{-1}\sum_{i=1}^{N} \mathsf{IF}_{1,\psi,i}(\theta) + o(N^{-1/2})$. Further $n^{1/2}(\widehat{\psi}_1 - \psi(\theta))$ converges conditionally and unconditionally to a normal distribution with mean zero; $\widehat{\psi}_{\mathsf{cf},1}$ is a regular, asymptotically linear estimator; i.e., $N^{1/2}(\widehat{\psi}_{\mathsf{cf},1} - \psi(\theta))$ converges unconditionally to a normal distribution with mean zero and variance equal to the semiparametric variance bound $\mathsf{var}_\theta\left[\mathsf{IF}_{1,\psi}(\theta)\right]$.*

(2) *The nominal $(1-\alpha)$ Wald CIs*

$$\mathsf{CI}_\alpha(\widehat{\psi}_1) := \widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\widehat{\psi}_1], \ \mathsf{CI}_\alpha(\widehat{\psi}_{\mathsf{cf},1}) := \widehat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\widehat{\psi}_{\mathsf{cf},1}] \tag{2.11}$$

are asymptotically valid nominal $(1-\alpha)$ (unconditional) CI for $\psi(\theta)$. Here $\widehat{\mathsf{s.e.}}[\widehat{\psi}_1] = \{\widehat{\mathsf{var}}[\widehat{\psi}_1]\}^{1/2}$ with $\widehat{\mathsf{var}}[\widehat{\psi}_1] = \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathsf{IF}_{1,\psi,i}(\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\mathsf{IF}_{1,\psi,i}(\widehat{\theta})\right)^2$ and $\widehat{\mathsf{s.e.}}[\widehat{\psi}_{\mathsf{cf},1}] = \frac{1}{2}\{\widehat{\mathsf{var}}[\widehat{\psi}_1] + \widehat{\mathsf{var}}[\overline{\widehat{\psi}}_1]\}^{1/2}$.

The proof of Theorem 2.2 is straightforward and can be found in Chernozhukov et al. (2018a), Smucler et al. (2019) or Farrell et al. (2021, Theorem 3). One often invokes complexity reducing assumptions on b and p. Further, under these complexity reducing assumptions, if one adopts the **Squared Error Loss Minimization Strategy**, rates at which $\mathsf{E}_\theta[\lambda(X)(\widehat{b}(X) - b(X))^2]$ and $\mathsf{E}_\theta[\lambda(X)(\widehat{p}(X) - p(X))^2]$ converge to zero can often be established (e.g. see many papers (Chernozhukov et al., 2018a; Farrell, 2015; Farrell et al., 2021; Smucler et al., 2019) on standard DML estimators) under a variety of complexity reducing assumptions on b and p. Then as in (1.1), CS inequality and these two convergence rates are used to justify if

$$|\mathsf{Bias}_\theta(\widehat{\psi}_1)| \leqslant \mathsf{CSBias}_\theta(\widehat{\psi}_1) \equiv \left\{\mathsf{E}_\theta[\lambda(X)(\widehat{b}(X) - b(X))^2]\mathsf{E}_\theta[\lambda(X)(\widehat{p}(X) - p(X))^2]\right\}^{1/2} = o(n^{-1/2}). \tag{2.12}$$

2.3. **A prelude to our approach.**

As briefly described in Section 1, our goal is to develop assumption-lean empirical methods to refute the justification as in (2.12). However, (2.12) involves an asymptotic statement which cannot be operationalized in practice. Therefore, we shall instead construct $\alpha^\dagger$-level tests of the null hypothesis $\mathsf{H}_{0,\mathsf{CS}}(\delta) : \mathsf{CSBias}_\theta(\widehat{\psi}_1) < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$, where $\delta > 0$ is chosen by the analysts.

Remark 2.2. *The null hypothesis* $\mathsf{H}_0(\delta)$*, if true, guarantees that the nominal* $(1-\alpha)$ *Wald CI covers* $\psi(\theta)$ *with probability at least*

$$\rho = \Phi(z_{\alpha/2} - \delta) - \Phi(-z_{\alpha/2} - \delta) \tag{2.13}$$[10]

in large sample. As a rule of thumb, the analyst could first decide the lowest coverage probability ρ *that can still be tolerated for a nominal* $(1-\alpha)$ *Wald CI, possibly depending on the application context, and then calculate the corresponding* δ *by inverting equation* (2.13)*. For example, when* $\alpha = 0.10$*, if* ρ *is required to be at least 80%, then the corresponding* $\delta = 3/4$.

However, since $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$ involves unknown functions b and p, we need to instead test a tempered null hypothesis $\mathsf{H}_{0,k}(\delta) : |\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)| < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$. For general DRFs,

$$\begin{aligned} \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) &\coloneqq \mathsf{E}_\theta[\Pi_\theta[\lambda^{1/2}(\widehat{b}-b)|\lambda^{1/2}\bar{\mathsf{z}}_k](X)\Pi_\theta[\lambda^{1/2}(\widehat{p}-p)|\lambda^{1/2}\bar{\mathsf{z}}_k]] \\ &\equiv \mathsf{E}_\theta[\lambda(X)(\widehat{b}(X)-b(X))\bar{\mathsf{z}}_k(X)]^\top \Sigma_k^{-1}\mathsf{E}_\theta[\bar{\mathsf{z}}_k(X)\lambda(X)(\widehat{p}(X)-p(X))] \end{aligned} \tag{2.14}$$

where $\bar{\mathsf{z}}_k$ is a vector of k-dimensional basis functions chosen by the analyst satisfying mild regularity conditions (see Condition SW in Section 3) and $\Sigma_k \coloneqq \mathsf{E}_\theta[\lambda(X)\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top] \equiv \mathsf{E}_\theta[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]$[11] is the population Gram matrix of $\lambda^{1/2}\bar{\mathsf{z}}_k$.

The test statistic of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, or more directly of $\mathsf{H}_{0,k}(\delta)$, that we will reveal in Section 3 is via higher-order influence function (HOIF)-based estimators of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ introduced in Liu et al. (2017); Robins et al. (2008, 2017). Due to space limitation, we refer the interested readers to works Robins et al. (2008, 2016) or van der Vaart (2014) for a more comprehensive review of HOIF-related theory at different technical levels. The HOIF estimators of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ are higher order U-statistics; in particular, an m-th order influence function is an m-th order U-statistic. *En route* to construct these HOIF estimators and associated tests, we require access to the study data and the nuisance estimators $\widehat{b}$ and $\widehat{p}$ obtained from the training sample. However, our tests are constructed without: i) refitting, modifying, or even having knowledge of the ML algorithms that have been employed to compute $\widehat{b}, \widehat{p}$ from the training sample, as long as they were trained based on the **Squared Error Loss Minimization Strategy** and ii) requiring any assumptions at all (aside from a few standard, quite weak assumptions given later) – in particular, without making any assumptions about the smoothness or sparsity of the true outcome regression b or propensity score $1/p$.

However, there is an unavoidable limitation to what can be achieved with our or any other method. No test, including ours, of the null hypothesis $\mathsf{CSBias}_\theta(\widehat{\psi}_1) < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$ can be uniformly (in θ) consistent [without making additional complexity reducing assumptions]. Thus, when our $\alpha^\dagger$-level test rejects this null hypothesis for $\alpha^\dagger$ small, we have strong evidence that the null hypothesis is false; nonetheless when

[10] We use Φ to denote standard normal cdf.

[11] To relate to the discussion in Section 1, $S_{bp} = A$ for the functional $\psi(\theta) = -\mathsf{E}_\theta[Y(1)]$.

the test does not reject, we cannot conclude that there is good evidence that $\mathsf{CSBias}_\theta(\widehat{\psi}_1) < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$, no matter how large the sample size. This reflects the fact that, because we make (essentially) no assumptions, $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$ may be order 1 and thus $n^{1/2}$ times greater than $\mathsf{s.e.}_\theta(\widehat{\psi}_1)$ under the law P_θ generating the data and yet our $\alpha^\dagger$-level test is really trying to test a tempered null hypothesis $\mathsf{H}_{0,k}(\delta)$.

Formally, our proposed approach begins by noticing that $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ can be unbiasedly estimated by the following infeasible second order U-statistic

$$
\begin{aligned}
&\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) := \frac{1}{n(n-1)} \sum_{1\leqslant i_1 \neq i_2 \leqslant n} \left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^\top \Sigma_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}, \text{ where} \\
&\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O) := S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k), \mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O) := S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)
\end{aligned}
\tag{2.15}
$$

with standard error of order $\frac{\sqrt{k}}{n} \vee \frac{1}{\sqrt{n}}$. In fact, $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is the second-order influence function of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$[12]; see Appendix A.6 for more details. $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is infeasible because in general Σ_k^{-1} is unknown. The unbiasedness follows directly from equation (2.4) in Theorem 2.1; the proof of the standard error bound of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is deferred to Appendix A.3, and can also be found in Liu et al. (2020a), which mostly focused on the infeasible case by assuming Σ_k^{-1} to be known. Since Σ_k^{-1} is generally unknown, we propose to estimate Σ_k^{-1} by $\widehat{\Sigma}_k^{-1}$ from the training sample data, where $\widehat{\Sigma}_k := \mathsf{P}_{n_{\mathsf{tr}}}\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]$ is simply the empirical Gram matrix estimator. In particular, throughout this paper, we choose $k \ll n$ for the following reasons.

Remark 2.3 (On the choice of k)**.** *We shall always choose k to be much less than the sample size $n = N/2$ of the estimation sample, for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$'s standard error of order to be smaller than or equal to the order $n^{-1/2}$ of the standard error of $\widehat{\psi}_1$, thereby creating the possibility of detecting, for any given $\delta > 0$, that the (absolute value) of the ratio of the estimable part of the bias of $\widehat{\psi}_1$ to its standard error exceeds δ, when the sample size n is sufficiently large.*

Moreover, we also need $k \ll n_{\mathsf{tr}} = N/2$, now the sample size of the training sample, to ensure $\widehat{\Sigma}_k^{-1}$ to be a operator-norm consistent estimator of Σ_k^{-1}, without imposing, possibly incorrect, additional smoothness assumption on the distribution of (S_{bp}, X) or sparsity assumptions on Σ_k or Σ_k^{-1} required for accurate estimation when $k \geqslant n$.

One might view the choice of $k \ll n$ a disadvantage. However, we argue that this is not the case for two reasons: (1) the nuisance parameters are allowed to be estimated by any procedure, be it high-dimensional or not, from the training samples; (2) one of the main motivation of this paper is to bypass any complexity reducing assumptions on the nuisance parameters, including the marginal density of X and this requires $k \ll n$.

Since Σ_k^{-1} is replaced by $\widehat{\Sigma}_k^{-1}$, the bias of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ as an estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ is no longer 0. In fact, we cannot provably show that simply replacing $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ by $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ in the infeasible test developed in Liu et al. (2020a) directly gives us a feasible valid test of the null hypothesis $\mathsf{H}_{0,\mathsf{CS}}(\delta)$. But

[12] This is why we adopt the notation $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ by following the convention in Robins et al. (2008)

fortunately, in Section 3, we will show how to use even higher order U-statistic estimators of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$, which are in fact higher order influence functions of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ (Robins et al., 2008), to construct a valid test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$.

3. Main results: A valid test for null hypothesis $\mathsf{H}_{0,\mathsf{CS}}(\delta)$

In this section, we show how to construct a valid nominal level-$\alpha^\dagger$ test statistic of the null hypothesis

$$\mathsf{H}_{0,\mathsf{CS}}(\delta) : \frac{\mathsf{CSBias}_\theta(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta(\widehat{\psi}_1)} \leqslant \delta.$$

This is the main theoretical result of the paper.

3.1. What constitutes a valid test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$?

Before proceeding to the main result, we first recall the infeasible valid test statistic of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ proposed in Liu et al. (2020a) based on $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$[13]:

$$\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta) = \mathbb{1}\left\{\frac{|\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} \geqslant \delta\right\} \tag{3.1}$$

where $\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]$ and $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ are estimators of the standard errors $\mathsf{s.e.}_\theta[\widehat{\psi}_1]$ and $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ (see Theorem 2.2 and Appendix A.8). Here one can choose the cut-off $\varsigma_k = z_{\alpha^\dagger/2}$ by asymptotic normal approximation. The asymptotic validity of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ relies on the unbiasedness of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, but the unbiasedness can be relaxed as in the proposition below.

Proposition 3.1. *Consider any estimator* $\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1)$ *of* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ *that satisfies:*

(1) Under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$*:* $\mathsf{E}_\theta[\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)] = o\left(\frac{\sqrt{k}}{n}\right)$*;*

(2) $\mathsf{s.e.}_\theta[\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1)] \asymp \mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$*;*

(3) Under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$*:* $\frac{\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1)]} \xrightarrow{d} N(0,1)$

Then the test

$$\widehat{\chi}_k(\varsigma_k, \delta) = \mathbb{1}\left\{\frac{|\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1)|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1)]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} \geqslant \delta\right\} \tag{3.2}$$

is an asymptotically valid test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ *with* $\varsigma_k = z_{\alpha^\dagger/2}$*, where* $\widehat{\mathsf{s.e.}}[\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1)]$ *is a consistent estimator of* $\mathsf{s.e.}_\theta[\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1)]$*.*

The statement of this proposition also holds if one replaces $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ *by the tempered null hypothesis* $\mathsf{H}_{0,k}(\delta)$*.*

The intuition for Proposition 3.1 is quite simple. Since $\mathsf{s.e.}_\theta[\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1)]$ can be as small as of order $\frac{\sqrt{k}}{n}$, one needs the bias $\mathsf{E}_\theta[\widehat{\mathsf{Bias}}_{\theta,k}(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)]$ to be dominated by this order under the null $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ to protect the level of the test. The final clause of Proposition 3.1 will be most relevant to the development of Section 4. One can read the current Section 3 without paying much attention to it.

[13] We also provide relevant results in Appendix A.3.

3.2. **Bias-reduced estimator of** $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$**.**

As mentioned above, the feasible second-order U-statistic estimator $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ is biased. Indeed, we *cannot*[14] prove the test below

$$\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) = \mathbb{1}\left\{ \frac{|\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - z_{\alpha^\dagger/2} \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} \geqslant \delta \right\} \tag{3.3}$$

is asymptotically valid for $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, because Theorem 3.1 below will show the bias

$$\mathsf{EB}_{\theta,k,2} \coloneqq \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)] = O\left(\frac{\sqrt{k}}{n} \sqrt{\mathsf{log} k} \right) \tag{3.4}$$

under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$. The upper bound $o\left(\frac{\sqrt{k \mathsf{log} k}}{n} \right)$ dominates the required upper bound for the bias given in Condition (1) of Proposition 3.1, failing to guarantee the asymptotic validity of $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$.

Fortunately, the same Theorem 3.1 below will show that the following third-order U-statistic estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$

(3.5)

$$\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) \coloneqq \widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) + \widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1}), \text{ where}$$

$$\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1}) \coloneqq \frac{(n-2)!}{n!} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O) \right]_{i_1}^\top \widehat{\Sigma}_k^{-1} \left[S_{bp} \bar{\mathsf{z}}_k(X) \bar{\mathsf{z}}_k(X)^\top - \widehat{\Sigma}_k \right]_{i_3} \widehat{\Sigma}_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O) \right]_{i_2}$$

of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ can provably reduce the bias upper bound in (3.4) to

$$\mathsf{EB}_{\theta,k,3} \coloneqq \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)] = O\left(\frac{\sqrt{k}}{n} \sqrt{\frac{k \mathsf{log} k}{n}} \right) = o\left(\frac{\sqrt{k}}{n} \right) \tag{3.6}$$

whence $k \mathsf{log} k = o(n)$.

Remark 3.1. *$\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ can be viewed as a de-biased version of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, where part of the bias is corrected by $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$. As will be shown in Appendix A.6, $\widehat{\mathbb{IF}}_{22\to33,k}(\Sigma_k^{-1})$ is the third-order influence function of* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ *and $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ is simply its feasible version.*

Before stating Theorem 3.1, we need to introduce the following additional notation for various norms:

$$\mathbb{L}_{\theta,2,\widehat{b}} \coloneqq \left\{ \mathsf{E}_\theta[\lambda(X)(\widehat{b}(X) - b(X))^2] \right\}^{1/2}, \mathbb{L}_{\theta,2,\widehat{b},k} \coloneqq \left\{ \mathsf{E}_\theta[\Pi_\theta[\lambda^{1/2}(\widehat{b} - b) | \lambda^{1/2} \bar{\mathsf{z}}_k](X)^2] \right\}^{1/2},$$
$$\mathbb{L}_{\theta,2,\widehat{p}} \coloneqq \left\{ \mathsf{E}_\theta[\lambda(X)(\widehat{p}(X) - p(X))^2] \right\}^{1/2}, \mathbb{L}_{\theta,2,\widehat{p},k} \coloneqq \left\{ \mathsf{E}_\theta[\Pi_\theta[\lambda^{1/2}(\widehat{p} - p) | \lambda^{1/2} \bar{\mathsf{z}}_k](X)^2] \right\}^{1/2},$$
$$\mathbb{L}_{\theta,\infty,\widehat{b},k} \coloneqq \| \Pi_\theta[\lambda^{1/2}(\widehat{b} - b) | \lambda^{1/2} \bar{\mathsf{z}}_k] \|_\infty, \mathbb{L}_{\theta,\infty,\widehat{p},k} \coloneqq \| \Pi_\theta[\lambda^{1/2}(\widehat{p} - p) | \lambda^{1/2} \bar{\mathsf{z}}_k] \|_\infty, \mathbb{L}_{\theta,2,\widehat{\Sigma},k} \coloneqq \| \widehat{\Sigma}_k - \Sigma_k \|.$$

which will also be useful for the later development of this paper.

We also need to further impose the following weak regularity conditions (Condition SW).

[14] We cannot ruled out the possibility that $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is asymptotically valid yet.

Condition SW.

(1) *All the eigenvalues of Σ_k are bounded away from 0 and ∞.*

(2) *The observed data $O = (W, X)$, the true nuisance functions $b(X)$ and $p(X)$, the estimated nuisance functions $\widehat{b}(X)$ and $\widehat{p}(X)$ are bounded with P_θ-probability 1, and $\lambda(X)$ are bounded away from 0 and ∞ with P_θ-probability 1.*

(3) $\|\bar{\mathsf{z}}_k^\top \bar{\mathsf{z}}_k\|_\infty \leqslant Bk$ *for some constant* $B > 0$.

Note that Condition SW should also be compared to the following slightly stronger Condition W in Liu et al. (2020a), which is the same as Condition SW except (3) shall be replaced by the following (3′):

Condition W.

(3′) $\|\bar{\mathsf{z}}_k^\top \bar{\mathsf{z}}_k\|_\infty \leqslant Bk$ *for some constant* $B > 0$*; in addition,* $\mathbb{L}_{\theta,\infty,\widehat{b},k} < \infty$ *and* $\mathbb{L}_{\theta,\infty,\widehat{p},k} < \infty$.

Remark 3.2. *Condition W(3′) is stronger than Condition SW(3) and it holds for Cohen-Vial-Daubechies wavelets, local polynomial partition and B-spline series (Belloni et al., 2015). But it does not hold in general for Fourier series or monomial transformation of the covariates X when X is compactly supported (Belloni et al., 2015). We also refer interested readers to Kennedy et al. (2020, Section 6) and Liu et al. (2020b, Section 2.1) for further motivation on why we decide to relax Condition W by Condition SW. However, as we will see, when Σ_k is unknown and need to be estimated from the training sample, violation of Condition W(3′) but not Condition SW(3), will incur a loss in the power of the test, but the validity of the test (i.e. level) is still guaranteed. Nevertheless, such loss in power happens (or equivalently, Condition SW holds but Condition W fails to hold), only if the L_∞-norms of the projections $\Pi_\theta[\lambda^{1/2}(\widehat{b} - b)|\lambda^{1/2}\bar{\mathsf{z}}_k]$ and $\Pi_\theta[\lambda^{1/2}(\widehat{p} - p)|\lambda^{1/2}\bar{\mathsf{z}}_k]$ are not bounded even though those of $\lambda^{1/2}(\widehat{b} - b)$ and $\lambda^{1/2}(\widehat{p} - p)$ are bounded.*

Finally, we summarize the statistical properties of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ in Theorem 3.1 below:

Theorem 3.1. *Under the conditions of Theorem 2.1 and Condition SW, with $k, n \to \infty$ but $k \log k = o(n)$, conditioning on the training sample, the followings hold on the event that $\widehat{\Sigma}_k$ is invertible:*

(1) *The bias and variance of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ as an estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ are of the following order:*

$$|\mathsf{EB}_{\theta,2,k}| \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\mathbb{L}_{\theta,2,\widehat{\Sigma},k} \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\sqrt{\frac{k\log(k)}{n}},$$

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}^2_{\theta,2,\widehat{b},k} + \mathbb{L}^2_{\theta,2,\widehat{p},k}\right\}$$

where $\mathsf{EB}_{\theta,2,k}$ is defined in equation (3.4).

(2) *The bias and variance of $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) \coloneqq \widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) + \widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ are of the following order:*

$$|\mathsf{EB}_{\theta,3,k}| \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\mathbb{L}^2_{\theta,2,\widehat{\Sigma},k} \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\frac{k\log(k)}{n},$$

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2 + k\mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,2,\widehat{p},k}^2\right\}$$

where $\mathsf{EB}_{\theta,3,k}$ is defined in equation (3.4).

If, however, Condition SW is strengthened to Condition W

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2\right\}.$$

(3) *Under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$:*

$$\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})\right]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]} \text{ and } \frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})\right]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]} \xrightarrow{d} N(0,1).$$

When we estimate $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ respectively by their consistent bootstrap estimators $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ described in detail in Appendix A.8, we also have

$$\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})\right]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]} \text{ and } \frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})\right]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]} \xrightarrow{d} N(0,1).$$

Remark 3.3.

- *The bias bounds have been proved in Liu et al. (2017). The variance bound proof can be found in Appendix A.7. The normal approximation under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ follows easily from the asymptotic normality of U-statistic proved in Bhattacharya and Ghosh (1992).*
- *The generic bias bounds in Theorem 3.1 imply those stated in (3.4) and (3.6) under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, as*

$$\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k} \leqslant \mathbb{L}_{\theta,2,\widehat{b}}\mathbb{L}_{\theta,2,\widehat{p}} = \mathsf{CSBias}_\theta(\widehat{\psi}_1) \lesssim n^{-1/2}$$

when $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ is true.

3.3. The proposed assumption-lean test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$.

All the previous discussions in this section culminate in (1) the following test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$:

$$\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) \coloneqq \mathbb{1}\left\{\frac{|\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - z_{\alpha^\dagger/2}\frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} \geqslant \delta\right\}, \tag{3.7}$$

and (2) the following theorem provably showing its asymptotically validity:

Theorem 3.2. *Assume all the conditions in Theorem 3.1. Under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$: $\frac{\mathsf{CSBias}_\theta(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta(\widehat{\psi}_1)} \leqslant \delta$,*

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) = 1\right) \leqslant \alpha^\dagger. \tag{3.8}$$

That is, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is an asymptotically level-$\alpha^\dagger$ test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$.

The proof of Theorem 3.2 is a direct consequence of Theorem 3.1 and deferred to Appendix A.9. As discussed previously, under assumptions on b and p that can guarantee $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ to be true, one can justify a nominal $(1-\alpha)$ Wald CI $\widehat{\mathsf{CI}}_{\alpha}$ centered around the usual DML estimator has the correct coverage, or equivalently, is valid. Based on Theorem 3.2, if $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^{\dagger}/2}, \delta)$ rejects $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, then the analyst should not report $\widehat{\mathsf{CI}}_{\alpha}$ in a research paper as evidence of a causal effect or a lack thereof.

Remark 3.4. *One might wonder if the basis functions $\bar{\mathsf{z}}_k$ can be made data-adaptive, so as to improve the power of our approach. To avoid unnecessary complications, we consider a new independent sample $\mathcal{D}_{sel}$ independent of the estimation and the nuisance samples. With $\mathcal{D}_{sel}$, there are at least two heuristics that one could exploit for constructing data-adaptive basis functions that we could envision practitioners may use in applied settings:*

(i) If given a huge dictionary of basis functions $\bar{\mathsf{z}}_K$ with $K \gg n$, one could regress between and $\bar{\mathsf{z}}_K$ with sparsity-inducing regression techniques such as Lasso or greedy algorithms.

(ii) One could also consider directly learning a representation (i.e. data-adaptive basis) $\widetilde{\bar{\mathsf{z}}}_k$ via modern deep neural networks, the last hidden layer or a distillation of which could be viewed as a data-adaptive representation (Ansuini et al., 2019; Dou and Liang, 2021; Ha et al., 2021).

The current framework allows for data-adaptive basis functions, especially when the relaxed Condition SW is assumed instead of Condition W.

Remark 3.5. *Now we address one question we posed in footnote 4 in Section 1. In principle, there is no reason not to consider another tempered null hypothesis*

$$\mathsf{H}_{0,k,\mathsf{CS}}(\delta) : \mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)^2 = \mathsf{E}_{\theta}\left[\Pi_{\theta}[\lambda^{1/2}(\widehat{b}-b)|\lambda^{1/2}\bar{\mathsf{z}}_k](X)^2\right]\mathsf{E}_{\theta}\left[\Pi_{\theta}[\lambda^{1/2}(\widehat{p}-p)|\lambda^{1/2}\bar{\mathsf{z}}_k](X)^2\right] \leqslant \delta^2 \mathsf{var}_{\theta}(\widehat{\psi}_1).$$

$\mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)^2$ *can be unbiasedly estimated by the following infeasible fourth-order U-statistic:*

$$\widehat{\mathbb{IF}}_{44,\mathsf{CS},k}(\Sigma_k^{-1}) := \frac{(n-4)!}{n!} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} [\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)]_{i_1}^{\top} \Sigma_k^{-1} [\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)]_{i_2} [\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)]_{i_3}^{\top} \Sigma_k^{-1} [\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)]_{i_4}.$$

Using $\widehat{\mathbb{IF}}_{44,\mathsf{CS},k}(\Sigma_k^{-1})$ as a starting point, one can indeed follow the previous development in Section 3 to construct a feasible test of $\mathsf{H}_{0,k,\mathsf{CS}}(\delta)$. We decide not to cover this alternative strategy in the paper because the statistical properties of a fourth-order U-statistic are more tedious than a second-order U-statistic[15]*, yet it does not add more new insights to our understanding of the problem.*

4. What if refined DML estimators are used in practice?

4.1. An overview on some refined DML estimators and motivation for testing $\mathsf{H}_{0,k}(\delta)$.

Standard DML estimators are agnostic to how the nuisance estimators $\widehat{b}$ and $\widehat{p}$ are fit, but recent works (Bradic et al., 2019a; Cattaneo and Jansson, 2018; Kennedy, 2020; Kline et al., 2020; Newey and Robins, 2018) have shown standard DML estimators can be suboptimal compared to certain refined estimators that leverage specific algebraic structures of particular nuisance estimators $\widehat{b}$ and $\widehat{p}$, together with certain

[15] To get a feasible estimator one might need to use even higher-order U-statistics.

complexity-reducing assumptions on the true nuisance functions b and p. Since this section focuses on the refined DML estimators, we use the same notation $\widehat{b}$, $\widehat{p}$ and $\widehat{\psi}_1$ without causing any confusion. But recall that $\widehat{b}$ and $\widehat{p}$ are now carefully crafted: e.g. in Newey and Robins (2018) and Kennedy (2020), $\widehat{b}$ and $\widehat{p}$ are wavelets projection estimators with the basis dimension k chosen to be close to n.

A common theme of all the refined DML estimators is to control their bias $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ through a direct analysis bypassing the CS inequality upper bound as in (1.1) or (2.12). But the detail might differ[16]. As a result, each refined DML estimator needs to be treated in a case-specific manner. In this paper, we consider the refined DML estimator of Newey and Robins (2018) as an example, which fits the nuisance estimators $\widehat{b}$ and $\widehat{p}$ separately from two independent subsets of training sample data with wavelet bases with dimension k' of order $n/\log^2(n)$ but otherwise estimates $\widehat{\psi}_1$ in the same way as the standard DML estimator. We denote the basis functions used by nuisance estimators as $\bar{\mathsf{v}}_{k'}$. Then the bias of $\widehat{\psi}_1$ is simply

$$\mathsf{Bias}_\theta(\widehat{\psi}_1) = \mathsf{Bias}_{\theta,\bar{\mathsf{v}}_{k'}}(\widehat{\psi}_1) + \mathsf{TB}_{\theta,\bar{\mathsf{v}}_{k'}}(\widehat{\psi}_1)$$

where $\mathsf{Bias}_{\theta,\bar{\mathsf{v}}_{k'}}(\widehat{\psi}_1) = \mathsf{E}_\theta\left[\Pi_\theta[\lambda^{1/2}(\widehat{b}-b)|\lambda^{1/2}\bar{\mathsf{v}}_{k'}](X)\Pi_\theta[\lambda^{1/2}(\widehat{p}-p)|\lambda^{1/2}\bar{\mathsf{v}}_{k'}](X)\right]$. We then immediately have (e.g. see Lemma A.2 in Appendix A.3)

$$\mathsf{TB}_{\theta,\bar{\mathsf{v}}_{k'}} \coloneqq \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,\bar{\mathsf{v}}_{k'}}(\widehat{\psi}_1) = \mathsf{E}_\theta\left[\Pi_\theta^\perp[\lambda^{1/2}(\widehat{b}-b)|\lambda^{1/2}\bar{\mathsf{v}}_{k'}](X)\Pi_\theta^\perp[\lambda^{1/2}(\widehat{p}-p)|\lambda^{1/2}\bar{\mathsf{v}}_{k'}](X)\right]. \tag{4.1}$$

In Newey and Robins (2018), $\mathsf{Bias}_\theta(\widehat{\psi}_1) = O(n^{-1/2})$ is justified by the following conditions:

Condition 4.1 (Justification for refined DML estimators)**.** *(1) b and p are Hölder-continuous functions with smoothness s_b and s_p and (2) the average smoothness $\frac{s_b+s_p}{2}$ of b and p is at least[17] above $d/4$, and (3) $\bar{\mathsf{v}}_{k'}$ are optimal approximation basis functions for Hölder-continuous functions with dimension $k' = O(n/\log^2(n))$, satisfying Condition SW.*

Under these three assumptions, both $\mathsf{Bias}_{\theta,\bar{\mathsf{v}}_{k'}}(\widehat{\psi}_1)$ and $\mathsf{TB}_{\theta,\bar{\mathsf{v}}_{k'}}$ are $O(n^{-1/2})$.

Condition 4.2. *Importantly, none of the refined DML estimators justify their bias to be $O(n^{-1/2})$ when $\mathsf{Bias}_{\theta,\bar{\mathsf{v}}_{k'}}(\widehat{\psi}_1)$ and $\mathsf{TB}_{\theta,\bar{\mathsf{v}}_{k'}}$ are of greater order than $n^{-1/2}$, yet $\mathsf{Bias}_\theta(\widehat{\psi}_1) = O(n^{-1/2})$ by some "miracle cancellation". In Liu et al. (2020a), we termed this assumption "the faithfulness condition"; if a data generating distribution allows such "miracle cancellation", then it is "unfaithful".*

[16] The estimators considered in Kline et al. (2020) are similar to those in Newey and Robins (2018) and Liu et al. (2017), but they only focused on quadratic functionals rather than the more general class of DRFs. Finally, Cattaneo and Jansson (2018) considered under-smoothing kernel-based leave-out estimators (which are also U-statistics) with bootstrap distributional approximation.

[17] The actual sufficient condition may be stronger than $\frac{s_b+s_p}{2} > \frac{d}{4}$, depending on the parameter of interest. But we decide to ignore such subtleties to focus on the most important issues.

Since the refined DML estimators do not rely on the CS inequality but the above Condition 4.2 to justify $\mathsf{Bias}_\theta(\widehat{\psi}_1)$, one shall instead be interested in testing

$$\text{(4.2)} \qquad \mathsf{H}_0 : \frac{\mathsf{Bias}_\theta(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta(\widehat{\psi}_1)} \leqslant \delta', \text{ for every } \delta' > 0.$$

Similarly, $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ depends on unknowns like b and p, we need to consider the following tempered null hypothesis,

$$\text{(4.3)} \qquad \mathsf{H}_{0,k}(\delta) : \frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta(\widehat{\psi}_1)} \leqslant \delta$$

by choosing a set of basis functions $\bar{\mathsf{z}}_k$. Under H_0 and Condition 4.2, $\mathsf{H}_{0,k}(\delta)$ holds for any given $\delta > 0$.

The following remark explains why $\mathsf{H}_{0,k}(\delta)$ is a reasonable null hypothesis of interest in the current context.

Remark 4.1.

(i) *If* $\bar{\mathsf{z}}_k \equiv \bar{\mathsf{v}}_{k'}$ *(and of course* $k = k'$*),* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \equiv \mathsf{Bias}_{\theta,\bar{\mathsf{v}}_{k'}}(\widehat{\psi}_1)$*. Hence under* H_0 *and Condition 4.2, any valid* $\alpha^\dagger$*-level test of* $\mathsf{H}_{0,k}(\delta)$ *is automatically a valid test of* H_0*. The tempered null hypothesis* $\mathsf{H}_{0,k}(\delta)$ *can be interpreted as testing if the justification on* $\mathsf{Bias}_{\theta,\bar{\mathsf{v}}_{k'}}(\widehat{\psi}_1) = o(n^{-1/2})$ *is warranted.*

(ii) *If* $\bar{\mathsf{z}}_k \neq \bar{\mathsf{v}}_{k'}$ *but span*$[\bar{\mathsf{z}}_k] \perp\!\!\!\perp$ *span*$[\bar{\mathsf{v}}_{k'}]$*, then* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{TB}_{\theta,\bar{\mathsf{z}}_k \cup \bar{\mathsf{v}}_{k'}}$*. Again, under* H_0 *and the faithfulness Condition 4.2, any valid* $\alpha^\dagger$*-level test of* $\mathsf{H}_{0,k}(\delta)$ *is automatically a valid test of* H_0*. Here the tempered null hypothesis* $\mathsf{H}_{0,k}(\delta)$ *can be interpreted as testing if the justification on* $\mathsf{TB}_{\theta,\bar{\mathsf{v}}_{k'}} = o(n^{-1/2})$ *is warranted. The basis functions that satisfy Condition W rather than Condition SW are mostly local basis functions (e.g. wavelets, B-splines, ...) with small local support. When using such local basis functions, it is easy to empirically check if* $\bar{\mathsf{z}}_k$ *and* $\bar{\mathsf{v}}_{k'}$ *have non-overlapping support, which, if true, is sufficient for span*$[\bar{\mathsf{z}}_k] \perp\!\!\!\perp$ *span*$[\bar{\mathsf{v}}_{k'}]$*.*

4.2. **Issues with $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ as a test of $\mathsf{H}_{0,k}(\delta)$ and a rescue by HOIFs.**

One may wonder if the test $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ based on $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ can be directly applied to test $\mathsf{H}_{0,k}(\delta)$. Let us recall the bias upper bound of $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ appeared in Theorem 3.1:

$$\mathsf{EB}_{3,\theta,k} = \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)] \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\sqrt{\frac{k\mathsf{log}k}{n}}.$$

Under $\mathsf{H}_{0,k}(\theta)$, we have $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \lesssim n^{-1/2}$; yet since $\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k} \geqslant \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$, one cannot guarantee $\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k} \lesssim n^{-1/2}$ and $\mathsf{EB}_{3,\theta,k} = o\left(\frac{\sqrt{k}}{n}\right)$, per required in Proposition 3.1, without making further unverifiable assumptions on $b, p, \widehat{b}, \widehat{p}$ and $\bar{\mathsf{z}}_k$. A natural solution would be to further de-bias $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$. This is indeed the main motivation of the following higher-order U-statistic estimator of

$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$:

(4.4)

$$\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) \coloneqq \sum_{j=2}^{m} \widehat{\mathbb{IF}}_{jj,k}(\widehat{\Sigma}_k^{-1}), \text{ where}$$

$$\widehat{\mathbb{IF}}_{jj,k}(\widehat{\Sigma}_k^{-1}) \coloneqq (-1)^j \frac{(n-j)!}{n!} \left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^{\top} \left\{\prod_{s=3}^{j} \widehat{\Sigma}_k^{-1} \left(\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^{\top}\right]_{i_s} - \widehat{\Sigma}_k\right)\right\} \widehat{\Sigma}_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}$$

where m can be chosen to slowly increase with n, e.g. $m = \sqrt{\mathsf{log} n}$. $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ is actually the m-th order influence function of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ (see HOIF-related theory in Appendix A.6) and can also be viewed as a further de-biased version of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, to reduce the bias induced by estimating Σ_k^{-1} using $\widehat{\Sigma}_k^{-1}$. In particular, we have the following theorem on the statistical properties of $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$, similar to Theorem 3.1.

Theorem 4.1. *Under the same conditions as in Theorem 3.1: with $m = O(\sqrt{\mathsf{log} n})$,*

(1) *The bias and variance of $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ as an estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ are of the following order:*

$$|\mathsf{EB}_{\theta,m,k}| \coloneqq \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)\right] \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{\frac{m-1}{2}} \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\left(\frac{k\mathsf{log}k}{n}\right)^{\frac{m-1}{2}}$$

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k} + \mathbb{L}_{\theta,2,\widehat{p},k} + k\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\right\}.$$

If, however, Condition SW is strengthened to Condition W,

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k} + \mathbb{L}_{\theta,2,\widehat{p},k}\right\}.$$

(2) *Under $\mathsf{H}_{0,k}(\delta)$:*

$$\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})\right]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \xrightarrow{d} N(0,1).$$

When we estimate $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ by its consistent bootstrap estimator $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ (e.g. derived by generalizing the strategy in Appendix A.8), we also have

$$\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})\right]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \xrightarrow{d} N(0,1).$$

Remark 4.2. *Note that:*

(1) *As m increases, the bias of $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ converges to 0 at faster rates.*

(2) *But the greater m is, the higher the computational cost incurred in computing $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$.*

4.3. The proposed assumption-lean test of $\mathsf{H}_{0,k}(\delta)$, early-stopping and computational-assumption trade-off.

Based on the above discussion, we propose the following nominal $\alpha^{\dagger}$-level test of $\mathsf{H}_{0,k}(\delta)$: for an fixed $m \geqslant 3$,

$$\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^{\dagger}/2}, \delta) := \mathbb{1}\left\{ \frac{|\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - z_{\alpha^{\dagger}/2} \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta \right\}. \tag{4.5}$$

Theorem 4.1 immediately implies the following:

Proposition 4.1. *Under the conditions in Theorem 4.1 with Condition W, $k\mathsf{log}(k) = o(n)$ together with the additional restriction*

$$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o\left(\mathbb{L}_{\theta,2,\widehat{b},k} \mathbb{L}_{\theta,2,\widehat{p},k} \left(\frac{k\mathsf{log}(k)}{n} \right)^{\frac{m-1}{2}} \right) \tag{4.6}$$

for any given $\delta > 0$, suppose that $\frac{|\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)|}{\mathsf{s.e.}_{\theta}[\widehat{\psi}_1]} = \gamma$ for some (sequence) $\gamma = \gamma(n)$ (where $\gamma(n)$ can diverge with n), then the rejection probability of $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^{\dagger}/2}, \delta)$ converges to

(4.7)
$$2 - \Phi\left(z_{\alpha^{\dagger}/2} - \lim_{n\to\infty} (\gamma - \delta) \frac{\mathsf{s.e.}_{\theta}[\widehat{\psi}_1]}{\mathsf{s.e.}_{\theta}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \right) - \Phi\left(z_{\alpha^{\dagger}/2} + \lim_{n\to\infty} (\gamma + \delta) \frac{\mathsf{s.e.}_{\theta}[\widehat{\psi}_1]}{\mathsf{s.e.}_{\theta}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \right)$$

as $n \to \infty$. In particular,

(1) *under $\mathsf{H}_{0,k}(\delta) : \gamma \leqslant \delta$, $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^{\dagger}/2}, \delta)$ rejects the null with probability less than or equal to $\alpha^{\dagger}$, as $n \to \infty$;*

(2) *under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any diverging sequence $c = c(n) \to \infty$, $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^{\dagger}/2}, \delta)$ rejects the null with probability converging to 1, as $n \to \infty$.*

(2′) *If $\mathbb{L}_{\theta,2,\widehat{b},k}$ and $\mathbb{L}_{\theta,2,\widehat{p},k}$ converge to 0, under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any fixed $c > 0$ or any diverging sequence $c = c(n) \to \infty$, $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^{\dagger}/2}, \delta)$ has rejection probability converging to 1, as $n \to \infty$.*

If Condition W is strengthened to Condition SW, we have the following instead of (1), (2) and (2′):

(i) *under $\mathsf{H}_{0,k}(\delta) : \gamma \leqslant \delta$, $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^{\dagger}/2}, \delta)$ rejects the null with probability equal to $\alpha^{\dagger}$, as $n \to \infty$;*

(ii) *under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any diverging sequence $c = c(n) \gg k\mathbb{L}_{\theta,2,\widehat{b},k}^2 \mathbb{L}_{\theta,2,\widehat{p},k}^2 = k\mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)$, $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^{\dagger}/2}, \delta)$ rejects the null $\mathsf{H}_{0,k}(\delta)$ with probability converging to 1, as $n \to \infty$.*

Remark 4.3. *Note that the greater m is, the less restrictive (4.6) becomes. If we let $m = O(\sqrt{\mathsf{log} n})$ as $n \to \infty$, (4.6) converges to a null requirement.*

However, as stated in Remark 4.2, although increasing m relaxes the restriction (4.6), it incurs higher computational cost. In practice, one might hope m to be not too large. In view of such a computational concern in practice, we propose the following "early-stopping" test that stops increasing m if $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^{\dagger}/2}, \delta)$ fails to reject $\mathsf{H}_{0,k}(\delta)$ at a smaller m.

Formally, for a fixed integer $M > 0$ that is determined by the analyst's computational budget, define

(4.8)
$$\widehat{\chi}^{es}_{M,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) := \mathbb{1}\left\{\frac{|\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - z_{\alpha^\dagger/2}\frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta : \forall\, m = 2, \ldots, M\right\}.$$

If $\widehat{\chi}^{es}_{M,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ fails to reject $\mathsf{H}_{0,k}(\delta)$ at some $m \leqslant M$, we stop the test at m and claim that we fail to reject $\mathsf{H}_{0,k}(\delta)$. This "early stopping" procedure is again an asymptotically valid $\alpha^\dagger$-level test:

Proposition 4.2. *Assume all the conditions of Proposition 4.1, but with (4.6) replaced by*

$$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o\left(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\left(\frac{k\log(k)}{n}\right)^{\frac{M-1}{2}}\right).$$

Under $\mathsf{H}_{0,k}(\delta)$:

$$\mathsf{P}_\theta\left(\widehat{\chi}^{es}_{M,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) = 1\right) \leqslant \alpha^\dagger$$

that is, $\widehat{\chi}^{es}_{M,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid level-$\alpha^\dagger$ test of $\mathsf{H}_{0,k}(\delta)$.

Apparently, there is a trade-off between computational cost and the level of assumptions that one can relax to: whether (4.6) is assumed for a pre-fixed M or an increasing sequence $m \equiv m(n)$ that eventually leads to exponential-time computation). We conjecture that this trade-off cannot be avoided. Finally, the power of the above "early-stopping" procedure for testing $\mathsf{H}_{0,k}(\delta)$ is more challenging to analyze, which we leave for future work.

5. Monte Carlo experiments

In the Monte Carlo (MC) experiments, we focus on parameter the counterfactual mean of Y when $\{0,1\}$-valued A is set to 1: $\psi^\dagger(\theta) \equiv \mathsf{E}_\theta[Y(a=1)] \equiv \mathsf{E}_\theta[b(X)]$ under ignorability. Here $p(X) = 1/\mathsf{E}_\theta[A|X]$ is the inverse propensity score and $b(X) = \mathsf{E}_\theta[Y|A=1,X]$ is the conditional mean of the outcome in the treatment group. For this parameter, $\Sigma_k = \mathsf{E}_\theta[A\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]$ and $\widehat{\Sigma}_k = n^{-1}\sum_{i\in\mathsf{tr}} A_i\bar{\mathsf{z}}_k(X_i)\bar{\mathsf{z}}_k(X_i)^\top$. We choose $\psi^\dagger(\theta) \equiv 0$.

We consider two simulation setups. In simulation setup I, we draw $N = 100,000$ i.i.d. X_j for $j = 1, \ldots, 4$ (so $d = 4$). The marginal density f of each X_j is supported on $[0,1]$ with $f \in$ Hölder$(0.1 + c)$ for some small $c > 0$, as defined in Appendix B. The correlation between each pair of X_j and X_k, with $j \neq k$, is introduced based on the algorithm described in Appendix B.1. We then simulate Y and A according to the following data generating mechanism:

$$Y \sim b(X) + N(0,1) = \sum_{j=1}^{4} \tau_{b,j} h_b(X_j; 0.25) + N(0,1)$$

and

$$A \sim \mathsf{Bernoulli}\left(1/p(X) \equiv \mathsf{expit}\left\{\sum_{j=1}^{4} \tau_{p,j} h_p(X_j; 0.25)\right\}\right)$$

where $h_b(\cdot; 0.25)$ and $h_p(\cdot; 0.25)$ have the forms as defined in Appendix B and hence both belong to Hölder$(0.25 + c)$ for some very small $c > 0$. The numerical values for $(\tau_{b,j}, \tau_{p,j})_{j=1}^4$ are provided in Table 5. *We fix half of the $N = 100,000$ samples as the training sample so $n_{\mathsf{tr}} = n = 50,000$ and only consider the randomness from the estimation sample in the simulation.* In simulation setup II, we consider the same data generating mechanism as in setup I except that we choose $b(X) = \sum_{j=1}^4 \tau_{b,j} h_b(X_j; 0.6)$ and $1/p(X) \equiv \mathsf{expit}\left\{\sum_{j=1}^4 \tau_{p,j} h_p(X_j; 0.6)\right\}$ where $h_b(\cdot; 0.6)$ and $h_p(\cdot; 0.6)$ have the forms as defined in Appendix B and hence both belong to Hölder$(0.6 + c)$ for some very small $c > 0$.

We choose D12 (or equivalently db6) Daubechies wavelets at resolutions $\ell \in (6, 7, 8)$ to form the basis functions

$$\bar{\mathsf{z}}_k(X) = (\bar{\mathsf{z}}_{k'}(X_1)^\top, \bar{\mathsf{z}}_{k'}(X_2)^\top, \bar{\mathsf{z}}_{k'}(X_3)^\top, \bar{\mathsf{z}}_{k'}(X_4)^\top)^\top,$$

with the corresponding $k' \in \{2^6 = 64, 2^7 = 128, 2^8 = 256\}$ and $k \in \{64 \cdot 4 = 256, 128 \cdot 4 = 512, 256 \cdot 4 = 1024\}$. To compute the oracle statistics and tests, we evaluate Σ_k through Monte Carlo integration by simulating $L = 10^7$ independent (A, X) from the true data generating law. To investigate the finite sample performance of the statistical procedures developed in this article, all the summary statistics of the Monte Carlo experiments are calculated based on 100 replicates. We estimate the nuisance functions $1/p(x)$ and $b(x)$ using generalized additive models (GAMs) (Hastie and Tibshirani, 1986, 1987). In particular, the smoothing parameters were selected by generalized cross validation, the default setup in gam function from R package mgcv (Wood et al., 2016). We choose db6 father wavelets to construct HOIF estimators when analyzing the simulated data.

5.1. **Finite sample performance of tests for $\mathsf{H}_{0,k}(\delta)$.** In this section, we consider testing the null hypothesis $\mathsf{H}_{0,k}(\delta)$. Henceforth we investigate the finite sample performance of the oracle statistics and tests $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$, and $\widehat{\chi}_{2,k}(\Sigma_k^{-1})$ where $\widehat{\psi}_{2,k}(\Sigma_k^{-1}) = \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, together with the statistics and tests relying on $\widehat{\Sigma}_k^{-1}$: $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$, and $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1})$, where $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1}) = \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1}) = \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$

First, we check the asymptotic normalities of $\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}$, $\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]}$, $\frac{\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]}$, and $\frac{\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]}$ through normal qq-plots displayed in Figures 1 (simulation setup I) and 2 (simulation setup II). We observe that the distributions of most of these statistics are close to normal at different k's ($k = 256$: left panels; $k = 512$: middle panels; $k = 1024$: right panels). In the third row of Figures 1 and 2), we observe some deviation from normality in the tails of $\frac{\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]}$, but the asymptotic distribution of $\frac{\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]}$ is not much affected and quite close to normal based on the qqplot.

In the simulation, we use nonparametric bootstrap to estimate the standard errors of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\hat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\hat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\hat{\Sigma}_k^{-1})$, as described in Appendix A.8. Hence we study if the estimated standard errors of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\hat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\hat{\Sigma}_k^{-1})$, and $\widehat{\mathbb{IF}}_{22\to 33,k}(\hat{\Sigma}_k^{-1})$ by nonparametric bootstrap are close to their true standard errors (calibrated by the MC variance from 100 replicates in the simulation). We use $B = 100$ bootstrap samples to compute the bootstrapped standard errors as the estimated standard errors for all four statistics. In Tables 1 (simulation setup I) and 3 (simulation setup II), we display the MC standard deviations (the upper numerical values in each cell), accompanied with the MC averages of the estimated standard errors (the lower numerical values outside the parenthesis in each cell) and *MC standard deviations* of the estimated standard errors (the lower numerical values inside the parenthesis in each cell) of all three statistics at $k = 256$ (left panel), $k = 512$ (middle panel) and $k = 1024$ (right panel). From Table 1 and Table 3, we observe that the estimated standard errors only slightly differ from the MC standard deviations.

Then we investigate (1) the finite sample performance of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\hat{\Sigma}_k^{-1})$, and $\widehat{\mathbb{IF}}_{22\to 33,k}(\hat{\Sigma}_k^{-1})$ and evaluate how close they are to $\mathsf{Bias}_\theta(\hat{\psi}_1)$, which is evaluated based on the MC bias of 100 replicates, and (2) the rejection rate of the tests $\hat{\chi}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta)$, $\hat{\chi}_{2,k}(\hat{\Sigma}_k^{-1}; z_{0.10/2}, \delta)$ and $\hat{\chi}_{3,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta)$ for the null hypothesis $\mathsf{H}_{0,k}(\delta)$. The numerical results are shown in Tables 2 (simulation setup I) and 4 (simulation setup II).

- In the upper panel of Table 2 (simulation setup I), the first row and the third column, we display the MC bias of the DML estimator $\hat{\psi}_1$ and the MC average of $\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$, which are -34.26×10^{-3} and 8.77×10^{-3}. Since the ratio between the bias and standard error is around 4, we expect the associated 90% Wald CI $\hat{\psi}_1 \pm z_{0.05}\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$ does not have the nominal coverage. This is indeed the case by reading from the first row, the second column of the upper panel of Table 2, showing the MC coverage probability for the 90% Wald CI of $\hat{\psi}_1$ is 0%.

  Similarly, in the upper panel of Table 4 (simulation setup II), the first row and the third column, we display the MC bias of the DML estimator $\hat{\psi}_1$ and the MC average of $\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$, which are -8.88×10^{-3} and 8.10×10^{-3}. This is as expected because the true nuisance functions b and p belong to Hölder$(0.6 + c)$ for some $c > 0$ and DML estimator (with b and p estimated in optimal rate in L_2 norm) is expected to have bias of $o(n^{-1/2})$ when the average smoothness between b and p is above 0.5 (Robins et al., 2009). Here the ratio between the bias and standard error is around 1 and we expect the associated 90% Wald CI $\hat{\psi}_1 \pm z_{0.05}\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$ has near nominal coverage. This is indeed the case by reading from the first row, the second column of the upper panel of Table 4, showing the MC coverage probability for the 90% Wald CI of $\hat{\psi}_1$ is 83%.
- In the second column of the upper panel of Table 2, we display the MC averages of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and the MC averages of its estimated standard errors; in the middle column, we display those of $\widehat{\mathbb{IF}}_{22,k}(\hat{\Sigma}_k^{-1})$; and in the lower panel, we display those of $\widehat{\mathbb{IF}}_{22\to 33,k}(\hat{\Sigma}_k^{-1})$ after third-order bias

correction. In the upper panel, we observe that increasing k from 256 to 512 does improve the amount of bias recovered by $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ (from -18.76×10^{-3} to -25.34×10^{-3}), but there is no obvious difference in the MC averages of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ between $k = 512$ and $k = 1024$ (-25.34×10^{-3} and -25.43×10^{-3}, about 75% of the total bias). The MC average of the estimated standard error increases with k, from 2.45×10^{-3} at $k = 256$ to 3.43×10^{-3} at $k = 1024$. This is consistent with the theoretical prediction that the variability of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ should grow with k. In the middle, we observe very similar numerical results between $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ for all the statistics that we are interested in. Interestingly, we did not see much improvement of using $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ instead of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, when compared to $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, at least in this example.

In the second column of the upper panel of Table 4, in the upper panel, we also observe that increasing k from 256 to 512 also slightly improve the amount of bias recovered by $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ (from -4.54×10^{-3} to -4.94×10^{-3}). The MC average of the estimated standard error increases with k, from 1.51×10^{-3} at $k = 256$ to 2.43×10^{-3} at $k = 1024$. In the middle panel, we observe very similar numerical results between $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ for all the statistics that we are interested in. Interestingly, we again did not see much improvement of using $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ instead of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, when compared to $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, at least in this example.

- In the third columns of Table 2, we display the MC coverage probabilities of the 90% two-sided CIs of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel). Comparing the oracle bias-corrected estimator $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ vs. $\widehat{\psi}_1$, the coverage probability improves from 0% to 33% (82%) at $k = 256$ (at $k = 1024$). Similar observations can be made for $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ as well when Σ_k is unknown.

  In the third columns of Table 4, we display the MC coverage probabilities of the 90% two-sided CIs of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel). Comparing the oracle bias-corrected estimator $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ vs. $\widehat{\psi}_1$, the coverage probability improves from 83% to 100%) at $k = 256$ and $k = 1024$. Similar observations can be made for $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ as well when Σ_k is unknown.
- In the fourth columns of Tables 2 and 4, we display the MC biases of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), together with their MC standard deviations (in the parentheses). As expected, the standard deviations of the estimators after bias correction are very similar to those of $\widehat{\psi}_1$. This indeed confirms that we are able to correct bias without drastically inflating the variance.
- In the fifth column of Table 2, we display the MC rejection rates of the test statistic: the upper panel shows the MC rejection rates of the oracle test $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ and the lower panel shows the MC rejection rates of the test $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$. These simulation results demonstrate that when a "good" basis (db6 father wavelets) is used, our proposed test does have

k	256	512	1024
$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$	2.437 2.268 (0.218)	3.011 2.716 (0.268)	3.247 2.841 (0.322)
$\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$	2.456 2.271 (0.218)	3.075 2.778 (0.278)	3.546 3.032 (0.378)
$\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$	0.495 0.603 (0.0598)	0.814 1.126 (0.126)	1.518 1.998 (0.285)
$\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$	2.282 2.251 (0.227)	2.745 2.738 (0.290)	2.743 3.009 (0.427)

Table 1. For data generating mechanism in simulation setup I: We reported the MC standard deviations $\times 10^{-3}$ (upper values in each cell), the MC averages of the estimated standard errors $\times 10^{-3}$ (lower values in each cell outside the parenthesis) and the MC standard deviations of the estimated standard errors $\times 10^{-3}$ (lower values in each cell inside the parenthesis) through nonparametric bootstrap resampling for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$.

		MC Coverage		
k	$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \times 10^{-3}$	($\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ 90% Wald CI)	Bias($\widehat{\psi}_{2,k}(\Sigma_k^{-1})$) $\times 10^{-3}$	$\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$
0	NA (NA)	0%	-34.26 (8.77)	NA
256	-18.76 (2.27)	31%	-15.50 (8.60)	100%
512	-25.34 (2.72)	83%	-8.92 (8.61)	100%
1024	-25.43 (2.84)	82%	-8.83 (8.79)	100%

		MC Coverage		
k	$\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) \times 10^{-3}$	($\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ 90% Wald CI)	Bias($\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$) $\times 10^{-3}$	$\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$
256	-18.54 (2.25)	30%	-15.72 (8.60)	100%
512	-24.65 (2.74)	81%	-9.61 (8.60)	100%
1024	-23.82 (3.01)	76%	-10.44 (8.76)	100%

Table 2. For data generating mechanism in simulation setup I: We reported the MC averages of point estimates and standard errors (the first column in each panel) of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), together with the coverage probabilities of two-sided 90% Wald CIs (the second column in each panel) of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), the MC biases and MC standard deviations (the third column in each panel) of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel) and the empirical rejection rates (the fourth column in each panel) of $\widehat{\chi}_{2,k}^{(1)}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (upper panel) and $\widehat{\chi}_{3,k}^{(1)}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (lower panel). In the upper panel, for $k = 0$, $\widehat{\psi}_{2,k=0} \equiv \widehat{\psi}_1$.

power to reject the null hypothesis of actual interest $\mathsf{H}_0(\delta)$ that the ratio between the bias of $\widehat{\psi}_1$ and its standard error is lower than δ. All the rejection rates in Table 2 are 100% when $\delta = 3/4$, regardless of whether we are using the oracle test or not. When $\delta = 2$, we obtain nontrivial rejection rates. This is as expected because the ratio between $\widehat{\mathbb{IF}}_{22,k}$ and $\widehat{\mathsf{s.e.}}(\widehat{\psi}_1)$ is around 2 to 3. In the fifth column of Table 4, all the rejection rates are 0% when $\delta = 3/4$. This is also as expected because the ratio between $\widehat{\mathbb{IF}}_{22,k}$ and $\widehat{\mathsf{s.e.}}(\widehat{\psi}_1)$ is close to $1/2$.

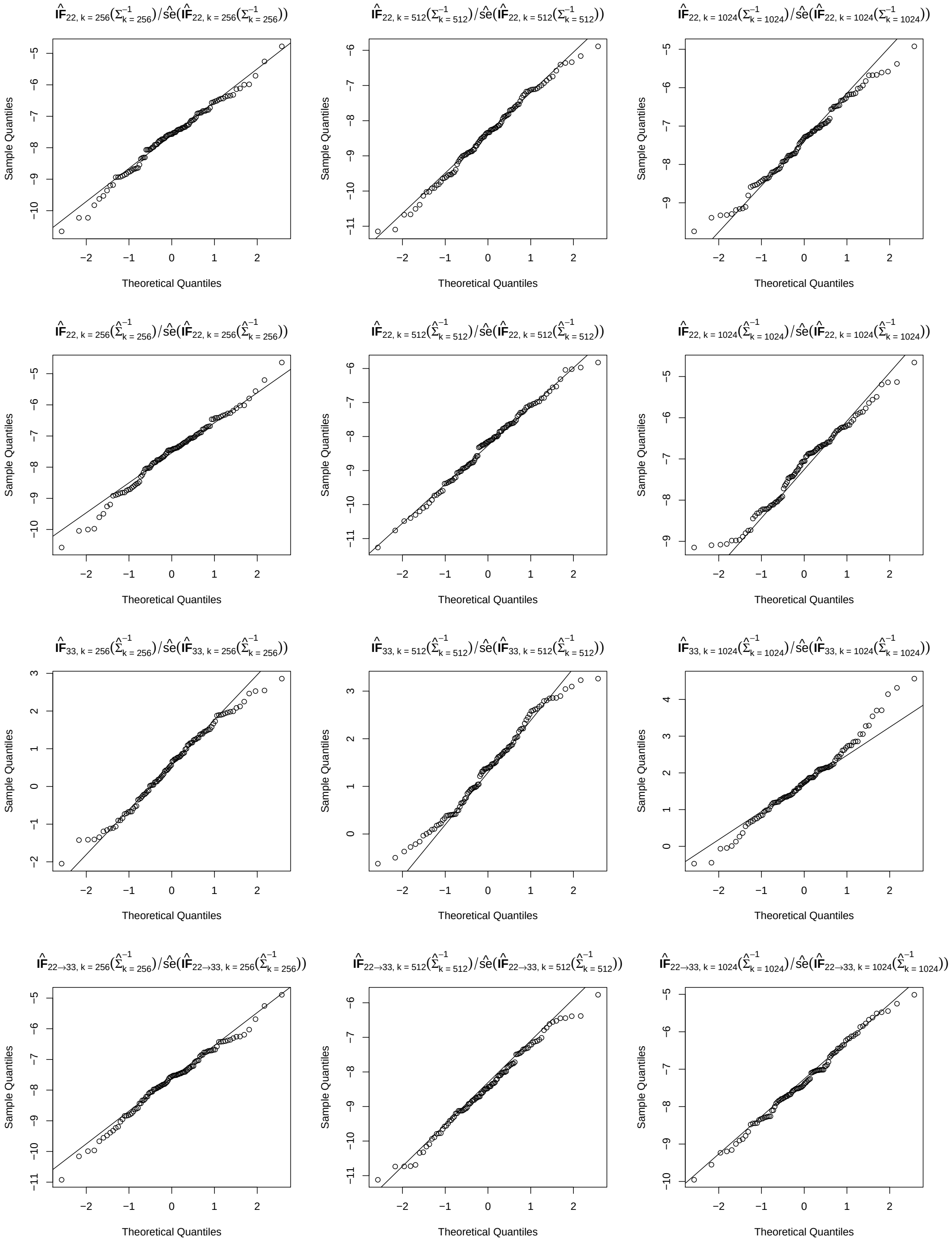

FIGURE 1. For data generating mechanism in simulation setup I: normal qq-plots for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ (the first row), $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]$ (the second row), $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]$ (the third row) and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]$ (the fourth row) for $k = 256$ (left panels), $k = 512$ (middle panels) and $k = 1024$ (right panels).

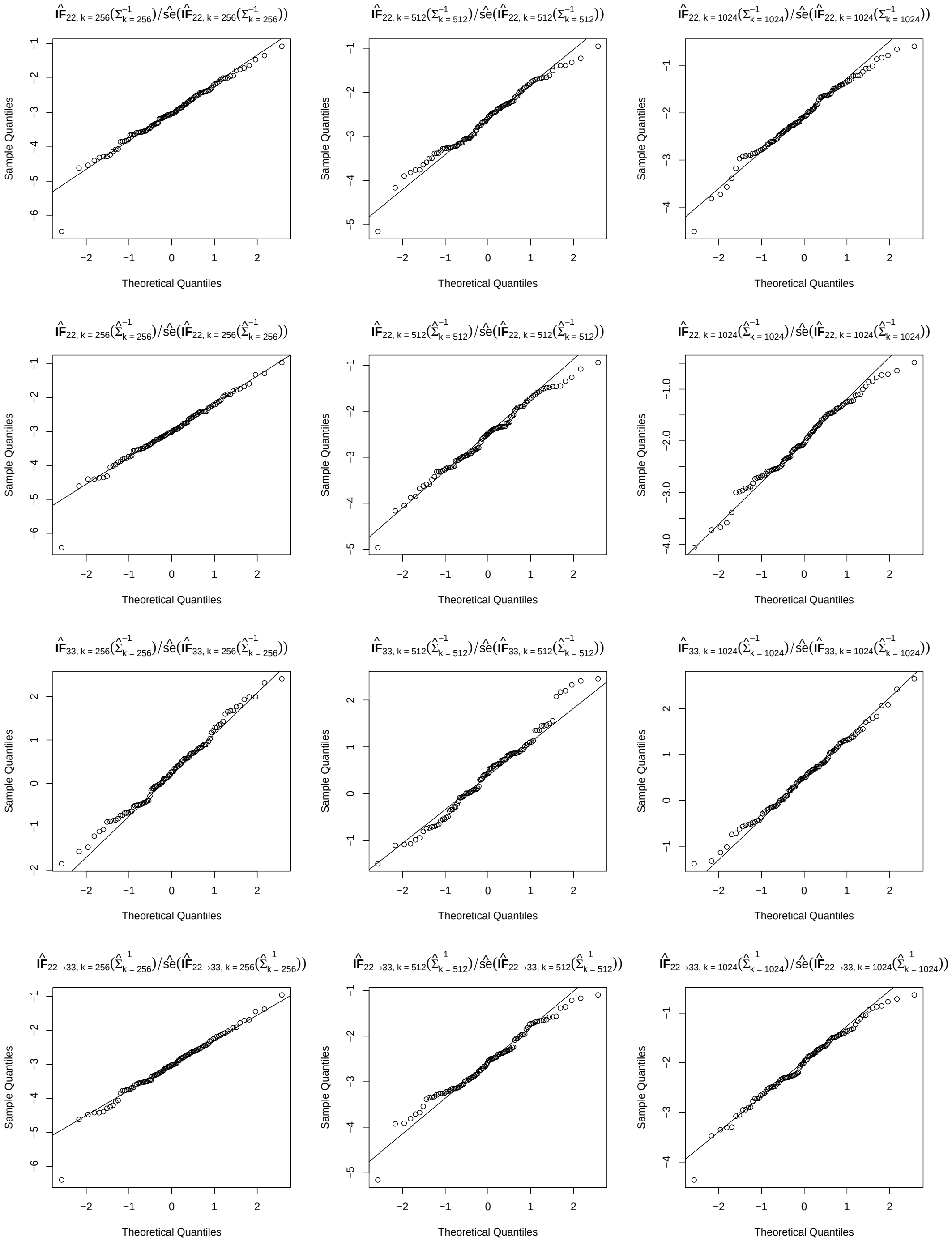

FIGURE 2. For data generating mechanism in simulation setup II: normal qq-plots for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ (the first row), $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]$ (the second row), $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]$ (the third row) and $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]$ (the fourth row) for $k = 256$ (left panels), $k = 512$ (middle panels) and $k = 1024$ (right panels).

k	256	512	1024
$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$	1.166 1.209 (0.156)	1.367 1.408 (0.191)	1.673 1.584 (0.267)
$\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$	1.187 1.219 (0.159)	1.433 1.439 (0.198)	1.828 1.697 (0.285)
$\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$	0.251 0.331 (0.0451)	0.417 0.629 (0.0997)	0.854 1.251 (0.247)
$\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$	1.109 1.186 (0.156)	1.256 1.346 (0.191)	1.488 1.560 (0.383)

Table 3. For data generating mechanism in simulation setup II: We reported the MC standard deviations $\times 10^{-3}$ (upper values in each cell), the MC averages of the estimated standard errors $\times 10^{-3}$ (lower values in each cell outside the parenthesis) and the MC standard deviations of the estimated standard errors $\times 10^{-3}$ (lower values in each cell inside the parenthesis) through nonparametric bootstrap resampling for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$.

k	$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \times 10^{-3}$	MC Coverage ($\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ 90% Wald CI)	Bias($\widehat{\psi}_{2,k}(\Sigma_k^{-1})$) $\times 10^{-3}$	$\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$
0	NA (NA)	83%	-8.88 (8.10)	NA
256	-4.54 (1.21)	99%	-4.34 (8.13)	0%
512	-4.89 (1.41)	99%	-3.99 (8.21)	0%
1024	-4.94 (1.58)	100%	-3.94 (8.35)	0%
k	$\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) \times 10^{-3}$	MC Coverage ($\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ 90% Wald CI)	Bias($\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$) $\times 10^{-3}$	$\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$
256	-4.49 (1.19)	99%	-4.39 (8.13)	0%
512	-4.76 (1.35)	99%	-4.12 (8.20)	0%
1024	-4.62 (1.56)	100%	-4.26 (8.32)	0%

Table 4. For data generating mechanism in simulation setup II: We reported the MC averages of point estimates and standard errors (the first column in each panel) of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), together with the coverage probabilities of two-sided 90% Wald CIs (the second column in each panel) of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), the MC biases and MC standard deviations (the third column in each panel) of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel) and the empirical rejection rates (the fourth column in each panel) of $\widehat{\chi}_{2,k}^{(1)}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (upper panel) and $\widehat{\chi}_{3,k}^{(1)}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (lower panel). In the upper panel, for $k = 0$, $\widehat{\psi}_{2,k=0} \equiv \widehat{\psi}_1$.

6. Concluding remarks

In this paper, we developed a valid assumption-lean third-order U-statistic based test that can empirically refute the justification for a Wald CI $\mathsf{CI}_\alpha(\widehat{\psi}_1)$ centered around a standard DML estimator $\widehat{\psi}_1$ covering the underlying DRF $\psi(\theta)$ at nominal rate (see Section 3). When more refined DML estimators are used, we also develop a valid test that is based on higher-order U-statistics, which are in fact HOIFs of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$.

An interesting future direction is to extend our framework to endogeneity settings, but some auxiliary variables are available for point identifying the causal effects. In Appendix C, we document the HOIFs for the running example $\psi(\theta) = -\mathsf{E}_{\theta}[Y(a=1)]$ in Section 1 under the so-call "proximal causal inference" or "double negative control" framework (Cui et al., 2020; Tchetgen Tchetgen et al., 2020). This framework is also closely related to other approaches dealing with endogeneity, e.g. synthetic controls (Shi et al., 2021). Based on the form of these HOIFs, it will be easy to apply the methods in this paper to the setting with endogeneity.

Another direction worth studying is how to extend our framework to the scenario where wider confidence intervals are used, e.g. confidence intervals centered around non-$\sqrt{n}$ higher-order influence function-based estimators. In Liu et al. (2020a), we provide some partial results. But it is currently beyond our capability to obtain guarantees as sharp as those in this paper.

Finally, evaluating the performance in real data applications will be important for the eventual adoption of our proposal in practice and more works are needed towards such a goal.

Acknowledgement

The authors would like to thank three anonymous referees and the editor Xiaohong Chen for comments that significantly improve the structure of the paper. Lin Liu's research was partially supported by NSFC Grant No. 12090024 and 12101397, Shanghai NSF Grant No.21ZR1431000 and No.21JC1402900, Shanghai Municipal Science and Technology Major Project No.2021SHZDZX0102, and Institute of Modern Analysis – A Shanghai Frontier Research Center Grant No. ZXWH2071101/002. Rajarshi Mukherjee's research was partially supported by NSF Grant EAGER-1941419. James M. Robins' research was partially supported by the U.S. Office of Naval Research grant N000141912446, and National Institutes of Health (NIH) awards R01 AG057869 and R01 AI127271.

References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 6111–6122, 2019.

Miguel A Arcones and Evarist Gine. On the bootstrap of U and V statistics. *The Annals of Statistics*, 20 (2):655–674, 1992.

Rose Baker. An order-statistics-based method for constructing multivariate distributions with fixed marginals. *Journal of Multivariate Analysis*, 99(10):2312–2327, 2008.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

Rabi N Bhattacharya and Jayanta K Ghosh. A class of U-statistics and asymptotic normality of the number of k-clusters. *Journal of Multivariate Analysis*, 43(2):300–330, 1992.

Jelena Bradic, Victor Chernozhukov, Whitney K Newey, and Yinchu Zhu. Minimax semiparametric learning with approximate sparsity. *arXiv preprint arXiv:1912.12213*, 2019a.

Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019b.

Matias D Cattaneo and Michael Jansson. Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency. *Econometrica*, 86(3):955–995, 2018.

Matias D Cattaneo, Michael Jansson, and Xinwei Ma. Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies*, 86(3):1095–1122, 2019.

Xiaohong Chen, Ying Liu, Shujie Ma, and Zheng Zhang. Efficient estimation of general treatment effects using neural networks with a diverging number of confounders. *arXiv preprint arXiv:2009.07055*, 2020.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018a.

Victor Chernozhukov, Whitney Newey, and James Robins. Double/de-biased machine learning using regularized Riesz representers. *arXiv preprint arXiv:1802.08667*, 2018b.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Learning L2 continuous regression functionals via regularized Riesz representers. *arXiv preprint arXiv:1809.05224*, 2018c.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of Riesz representers. *arXiv preprint arXiv:2101.00009*, 2020.

Victor Chernozhukov, Whitney K Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. *arXiv preprint arXiv:2110.03031*, 2021.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411*, 2020.

Xialiang Dou and Tengyuan Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *Journal of the American Statistical Association*, pages 1–14, 2021.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

AmirEmad Ghassami, Andrew Ying, Ilya Shpitser, and Eric Tchetgen Tchetgen. Minimax kernel machine learning for a class of doubly robust functionals. *arXiv preprint arXiv:2104.02929*, 2021.

Wooseok Ha, Chandan Singh, Francois Lanusse, Eli Song, Song Dang, Kangmin He, Srigokul Upadhyayula, and Bin Yu. Adaptive wavelet distillation from neural networks through interpretations. *arXiv preprint arXiv:2107.09145*, 2021.

Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 1998.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Marie Huskova and Paul Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, 21(4):1811–1823, 1993.

Hidehiko Ichimura and Whitney K Newey. The influence function of semiparametric estimators. *Quantitative Economics (To Appear)*, 2021.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Edward H Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Discussion of "On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning". *Statistical Science*, 35(3):540–544, 2020.

Patrick Kline, Raffaele Saggio, and Mikkel Sølvsten. Leave-out estimation of variance components. *Econometrica*, 88(5):1859–1898, 2020.

Lin Liu, Rajarshi Mukherjee, Whitney K Newey, and James M Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017.

Lin Liu, Rajarshi Mukherjee, and James M Robins. On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 35(3):518–539, 2020a.

Lin Liu, Rajarshi Mukherjee, and James M Robins. Rejoinder: On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 35(3):545–554, 2020b.

Wang Miao, Xu Shi, and Eric Tchetgen Tchetgen. A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945*, 2018.

James E Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1-2):65–82, 1962.

Justin T Newcomer, Nagaraj K Neerchal, and Jorge G Morel. Computation of higher order moments from two multinomial overdispersion likelihood models. Technical report, Department of Mathematics and Statistics, University of Maryland, 2008.

Whitney K Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Whitney K Newey and James M Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Ya'acov Ritov, Peter J Bickel, Anthony C Gamst, and Bastiaan Jan Korneel Kleijn. The Bayesian analysis of complex, high-dimensional models: Can it be CODA? *Statistical Science*, 29(4):619–639, 2014.

James Robins, Lingling Li, Eric Tchetgen Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

James Robins, Eric Tchetgen Tchetgen, Lingling Li, and Aad van der Vaart. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321, 2009.

James Robins, Lingling Li, Eric Tchetgen Tchetgen, and Aad van der Vaart. Technical report: Higher order influence functions and minimax estimation of nonlinear functionals. *arXiv preprint arXiv:1601.05820*, 2016.

James M Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997.

James M Robins and Andrea Rotnitzky. Comments on "Inference for semiparametric models: some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5): 1951–1987, 2017.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999a.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Rejoinder. *Journal of the American Statistical Association*, 94(448):1135–1146, 1999b.

Xu Shi, Wang Miao, Mengtong Hu, and Eric Tchetgen Tchetgen. On proximal causal inference with synthetic controls. *arXiv preprint arXiv:2108.13935*, 2021.

Ilya Shpitser, Zach Wood-Doughty, and Eric J Tchetgen Tchetgen. The proximal ID algorithm. *arXiv preprint arXiv:2108.06818*, 2021.

Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Aad van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, 29(4):679–686, 2014.

Simon N Wood, Natalya Pya, and Benjamin Säfken. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563, 2016.

Appendix A. Derivations and proofs

A.1. **Regularity condition for doubly robust functionals (DRFs).** All the theoretical results in this paper rely on the following (standard) regularity condition (Condition 1 in Rotnitzky et al. (2021)):

Condition A.1. *There exists a dense set $\mathcal{H}_b$ of $L_2(\mathsf{P}_{\theta,X})$ such that $\mathcal{F}_b \cap \mathcal{B} \neq \emptyset$, and for each $\theta \in \Theta$ and $h \in \mathcal{H}_b$, there exists $\varepsilon(\theta, h) > 0$ such that $b + th \in \mathcal{B}$ if $|t| < \varepsilon(\theta, h)$. The same holds replacing b with p and $\mathcal{B}$ with $\mathcal{P}$. Furthermore $\mathsf{E}_\theta[|S_{bp}p(X)h(X)|] < \infty$ for any $h \in \mathcal{H}_b$ and $\mathsf{E}_\theta[|S_{bp}b(X)h(X)|] < \infty$ for any $h \in \mathcal{H}_p$. Moreover for all $\theta \in \Theta$, $\mathsf{E}_\theta[|S_{bp}b'(X)p'(X)|] < \infty$ for any $b' \in \mathcal{B}$ and $p' \in \mathcal{P}$.*

A.2. **Examples of losses that are equivalent to the squared error losses.** The following Lemma shows $\mathsf{CSBias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$ if and only if

$$\mathsf{CSBias}_\theta^{uw}(\widehat{\psi}_1) \coloneqq \left\{\mathsf{E}_\theta[(b(X) - \widehat{b}(X))^2]\mathsf{E}_\theta[(p(X) - \widehat{p}(X))^2]\right\}^{1/2} = o(n^{-1/2})$$

so rejection of either of these two hypotheses implies rejection of the other.

Lemma A.1.

$$\mathsf{CSBias}_\theta(\widehat{\psi}_1) \lesssim \mathsf{CSBias}_\theta^{uw}(\widehat{\psi}_1) \lesssim \mathsf{CSBias}_\theta(\widehat{\psi}_1).$$

That is, there exists $1 > c > 0$ such that $c > \mathsf{CSBias}_\theta(\widehat{\psi}_1)/\mathsf{CSBias}_\theta^{uw}(\widehat{\psi}_1) < 1/c$.

Proof. By Condition SW, we have

$$\mathsf{E}_\theta[(b(X) - \widehat{b}(X))^2] = \mathsf{E}_\theta\left[\frac{1}{\lambda(X)}\lambda(X)(b(X) - \widehat{b}(X))^2\right]$$
$$\Rightarrow \inf_x |\lambda^{-1}(x)|\mathsf{E}_\theta[\lambda(X)(b(X) - \widehat{b}(X))^2] \leqslant \mathsf{E}_\theta[(b(X) - \widehat{b}(X))^2] \leqslant \sup_x |\lambda^{-1}(x)|\mathsf{E}_\theta[\lambda(X)(b(X) - \widehat{b}(X))^2]$$

and similarly

$$\inf_x |\lambda^{-1}(x)|\mathsf{E}_\theta[\lambda(X)(p(X) - \widehat{p}(X))^2] \leqslant \mathsf{E}_\theta[(p(X) - \widehat{p}(X))^2] \leqslant \sup_x |\lambda^{-1}(x)|\mathsf{E}_\theta[\lambda(X)(p(X) - \widehat{p}(X))^2].$$

Thus combining the above calculations, we have

$$\mathsf{CSBias}_\theta(\widehat{\psi}_1) \lesssim \mathsf{CSBias}_\theta^{uw}(\widehat{\psi}_1) \lesssim \mathsf{CSBias}_\theta(\widehat{\psi}_1).$$

■

Thus the λ-weighted and unweighted squared error loss functions are equivalent up to constants by Condition SW. Henceforth we restrict to consideration of $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$.

For when $p(X) = \mathsf{P}_\theta(A = 1|X)^{-1}$ is the inverse propensity score of a binary treatment A, using the cross-entropy loss is also asymptotically equivalent to the squared error loss with a simple Taylor expansion:

$$\mathsf{E}_\theta\left[\frac{1}{p(X)}\mathsf{log}\left(\frac{1}{\widehat{p}(X)}\right) + \left(\frac{p(X) - 1}{p(X)}\right)\mathsf{log}\left(\frac{\widehat{p}(X) - 1}{\widehat{p}(X)}\right)\right]$$
$$\approx \mathsf{E}_\theta\left[\frac{1}{p(X)}\left\{\mathsf{log}\left(\frac{1}{p(X)}\right) - \frac{p(X) - \widehat{p}(X)}{\widehat{p}(X)} + \frac{(\widehat{p}(X) - p(X))^2}{\widehat{p}(X)^2}\right\}\right]$$

$$+ \mathsf{E}_\theta \left[\frac{p(X) - 1}{p(X)} \left\{ \log \left(\frac{p(X) - 1}{p(X)} \right) - \frac{p(X)}{p(X) - 1} \frac{\widehat{p}(X) - p(X)}{p(X)\widehat{p}(X)} + \left(\frac{p(X)}{p(X) - 1} \right)^2 \frac{(\widehat{p}(X) - p(X))^2}{p(X)^2 \widehat{p}(X)^2} \right\} \right]$$

$$= \mathsf{const} + \mathsf{E}_\theta \left[\frac{p(X)}{p(X) - 1} \frac{(\widehat{p}(X) - p(X))^2}{p(X)\widehat{p}(X)^2} \right]$$

A.3. **The infeasible procedure when $\Sigma_k^{-1} := \{\mathsf{E}_\theta[\lambda(X)\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]\}^{-1}$ is known.** It is pedagogically useful to first consider an infeasible procedure that would only be implementable if Σ_k^{-1} were known. We will focus on such situation in this section.

A.3.1. *Truncated parameters and their unbiased estimators.* Though there does not exist uniformly (in θ) consistent test of the null hypothesis that the bias $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ of $\widehat{\psi}_1$ for estimating $\psi(\theta)$ is smaller than a fraction δ of its standard error $\mathsf{s.e.}_\theta(\widehat{\psi}_1)$, this is not the case for the bias of $\widehat{\psi}_1$ as an estimator of the truncated parameter $\widetilde{\psi}_k(\theta)$ of $\psi(\theta)$ (Robins et al. (2008, Section 3.2)), defined next. Recall from Section 1 that the bias of $\widehat{\psi}_1$ for $\widetilde{\psi}_k(\theta)$ is the estimable part of the bias of $\widehat{\psi}_1$ for $\psi(\theta)$.

Definition A.1 (The truncated parameter $\widetilde{\psi}_k(\theta)$, truncation bias $\mathsf{TB}_{\theta,k}$, and the projected bias $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$)**.** *Given estimators $\widehat{b} \in \mathcal{B}$ and $\widehat{p} \in \mathcal{P}$ of b and p computed from the training sample and some k dimensional basis functions $\bar{\mathsf{z}}_k(x)$ in $L_2(\mathsf{P}_{\theta,X})$:*

- *Define the (conditional) truncated parameter of a DRF $\psi(\theta)$ as*

$$\widetilde{\psi}_k(\theta) := \mathsf{E}_\theta \left[\mathcal{H}\left(\widetilde{b}_{k,\theta}, \widetilde{p}_{k,\theta} \right) \right],$$

where $\widetilde{b}_{k,\theta}$ and $\widetilde{p}_{k,\theta}$ are defined as follows. Consider the working linear models

$$\begin{cases} b_k^*(X; \bar{\zeta}_{b,k}) = \widehat{b}(X) + \bar{\zeta}_{b,k}^\top \bar{\mathsf{z}}_k(X), \\ p_k^*(X; \bar{\zeta}_{p,k}) = \widehat{p}(X) + \bar{\zeta}_{p,k}^\top \bar{\mathsf{z}}_k(X). \end{cases}$$

Define $\widetilde{\bar{\zeta}}_b(\theta)$ and $\widetilde{\bar{\zeta}}_p(\theta)$ as the solutions to the following system of equations

$$\begin{cases} 0 = \mathsf{E}_\theta \left[\dfrac{\partial \mathcal{H}\left(b_k^*(X; \bar{\zeta}_{b,k}), p^*(X; \bar{\zeta}_{p,k}) \right)}{\partial \bar{\zeta}_{b,k}} \right] = \mathsf{E}_\theta \left[S_{bp} p_k^*(X; \bar{\zeta}_{p,k}) \bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k) \right] \\ 0 = \mathsf{E}_\theta \left[\dfrac{\partial \mathcal{H}\left(b_k^*(X; \bar{\zeta}_{b,k}), p^*(X; \bar{\zeta}_{p,k}) \right)}{\partial \bar{\zeta}_{p,k}} \right] = \mathsf{E}_\theta \left[S_{bp} b_k^*(X; \bar{\zeta}_{b,k}) \bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k) \right]. \end{cases}$$

where, for $j = 1, 2$, $m_j(O, \bar{\mathsf{z}}_k) = (m_j(O, z_1), \ldots, m_j(O, z_k))$. Hence

$$\begin{cases} \widetilde{\bar{\zeta}}_{b,k}(\theta) = -\Sigma_k^{-1} \mathsf{E}_\theta \left[\left(S_{bp} \widehat{b}(X) \bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k) \right) \right], \\ \widetilde{\bar{\zeta}}_{p,k}(\theta) = -\Sigma_k^{-1} \mathsf{E}_\theta \left[(S_{bp} \widehat{p}(X) \bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)) \right] \end{cases} \tag{A.1}$$

where

$$\Sigma_k := \mathsf{E}_\theta \left[S_{bp} \bar{\mathsf{z}}_k(X) \bar{\mathsf{z}}_k(X)^\top \right] \equiv \mathsf{E}_\theta \left[\lambda(X) \bar{\mathsf{z}}_k(X) \bar{\mathsf{z}}_k(X)^\top \right].$$

Now define

$$\begin{cases} \widetilde{b}_{k,\theta}(X) := b_k^*(X; \widetilde{\bar{\zeta}}_{b,k}(\theta)) \equiv \widehat{b}(X) + \widetilde{\bar{\zeta}}_{b,k}(\theta)^\top \bar{\mathsf{z}}_k(X), \\ \widetilde{p}_{k,\theta}(X) := p_k^*(X; \widetilde{\bar{\zeta}}_{p,k}(\theta)) \equiv \widehat{p}(X) + \widetilde{\bar{\zeta}}_{p,k}(\theta)^\top \bar{\mathsf{z}}_k(X). \end{cases}$$

- *Define the difference between $\widetilde{\psi}_k(\theta)$ and $\psi(\theta)$ as the truncation bias $\mathsf{TB}_{\theta,k}$:*

$$\mathsf{TB}_{\theta,k} \coloneqq \widetilde{\psi}_k(\theta) - \psi(\theta) = \mathsf{E}_\theta \left[S_{bp} \left(\widetilde{b}_{k,\theta}(X) - b(X) \right) \left(\widetilde{p}_{k,\theta}(X) - p(X) \right) \right]. \tag{A.2}$$

- *Define the bias of $\widehat{\psi}_1$ as an estimator of $\widetilde{\psi}_k(\theta)$ as*

$$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \coloneqq \mathsf{E}_\theta \left[\widehat{\psi}_1 - \widetilde{\psi}_k(\theta) \right]. \tag{A.3}$$

By definition, $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{TB}_{\theta,k}$. Combining equation (A.2), we have $\widetilde{\psi}_k(\theta) - \psi(\theta) \equiv \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$. As discussed further below, $\mathsf{TB}_{\theta,k}$ is not consistently estimable without imposing smoothness/sparsity assumptions on the nuisance function spaces $\mathcal{B}$ and $\mathcal{P}$. But $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ can be unbiasedly estimated by the following oracle (i.e. Σ_k^{-1} known) second-order U-statistic:

$$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \coloneqq \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \widehat{\mathsf{IF}}_{22,k,\bar{\imath}_2}(\Sigma_k^{-1})$$

where

$$\widehat{\mathsf{IF}}_{22,k,\bar{\imath}_2}(\Sigma_k^{-1}) \coloneqq \left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O) \right]_{i_1}^{\top} \Sigma_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O) \right]_{i_2}.$$

Here $\mathcal{E}_{1,b'}(h)$ and $\mathcal{E}_{2,p'}(h)$ are linear functionals defined as: given any $h \in L_2(\mathsf{P}_{\theta,X})$, $b' \in \mathcal{B}$ and $p' \in \mathcal{P}$

$$\mathcal{E}_{b',m_2}(h)(O) \coloneqq S_{bp} b'(X) h(X) + m_2(O, h) \tag{A.4}$$

and

$$\mathcal{E}_{p',m_1}(h)(O) \coloneqq S_{bp} p'(X) h(X) + m_1(O, h). \tag{A.5}$$

When the range of h is multidimensional (e.g. $h = \bar{\mathsf{z}}_k$), the linear functionals $\mathcal{E}_{1,b'}$ and $\mathcal{E}_{2,p'}$ act on h componentwise.

For the examples covered in Example 1, $\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)$ and $\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)$ are

(1) Example 1(1):

$$\begin{cases} \mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O) = A\widehat{b}(X)\bar{\mathsf{z}}_k(X) - AY\bar{\mathsf{z}}_k(X), \\ \mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O) = A\widehat{p}(X)\bar{\mathsf{z}}_k(X) - \bar{\mathsf{z}}_k(X). \end{cases}$$

(2) Example 1(2):

$$\begin{cases} \mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O) = \widehat{b}(X)\bar{\mathsf{z}}_k(X) - Y\bar{\mathsf{z}}_k(X), \\ \mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O) = \widehat{p}(X)\bar{\mathsf{z}}_k(X) - A\bar{\mathsf{z}}_k(X). \end{cases}$$

The following lemma proves that (i) $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is an unbiased estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ and (ii) $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ is the expected product of the $L_2(\mathsf{P}_\theta)$ projections of the weighted residuals

$$\Delta_{1,\widehat{b}} \coloneqq \lambda(X)^{1/2} \left(\widehat{b}(X) - b(X) \right),$$
$$\Delta_{2,\widehat{p}} \coloneqq \lambda(X)^{1/2} \left(\widehat{p}(X) - p(X) \right)$$

onto the subspace spanned by the $\lambda^{1/2}$-weighted basis functions $\bar{v}_k \coloneqq \lambda^{1/2}\bar{\mathsf{z}}_k$:

Lemma A.2. *Define* $\beta_{\widehat{b},k} \coloneqq \mathsf{E}_\theta \left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]^\top \Sigma_k^{-1}$ *and* $\beta_{\widehat{p},k} \coloneqq \mathsf{E}_\theta \left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]^\top \Sigma_k^{-1}$*. Under the conditions of Theorem 2.1, we have*

$$\begin{aligned} \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) &= \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})\right] \\ &= \mathsf{E}_\theta\left[\Pi_\theta\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)\Pi_\theta\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right] = \beta_{\widehat{b},k}^\top \Sigma_k \beta_{\widehat{p},k} \end{aligned} \tag{A.6}$$

and

$$\mathsf{TB}_{\theta,k} \coloneqq \widetilde{\psi}_k(\theta) - \psi(\theta) = \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{E}_\theta\left[\Pi_\theta^\perp\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)\Pi_\theta^\perp\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right]. \tag{A.7}$$

where $\Pi_\theta^\perp[\cdot|\bar{v}_k]$ *is the projection operator onto the ortho-complement of the linear span of* $\bar{v}_k$*.*

Furthermore, define the bias corrected estimator $\widehat{\psi}_{2,k}(\Sigma_k^{-1}) \coloneqq \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$*. Then*

$$\begin{aligned} \mathsf{Bias}_\theta(\widehat{\psi}_{2,k}(\Sigma_k^{-1})) \equiv \mathsf{E}_\theta\left[\widehat{\psi}_{2,k}(\Sigma_k^{-1}) - \psi(\theta)\right] = \mathsf{TB}_{\theta,k}, \\ \textit{and thus } \mathsf{E}_\theta\left[\widehat{\psi}_{2,k}(\Sigma_k^{-1})\right] = \psi(\theta) + \mathsf{TB}_{\theta,k} = \widetilde{\psi}_k(\theta). \end{aligned} \tag{A.8}$$

Remark A.1. *The proof of the above lemma is deferred to Appendix A.4.1. By* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})\right] = \beta_{\widehat{b},k}^\top \Sigma_k^{-1} \beta_{\widehat{p},k}$*, we can view* $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ *as an unbiased estimator of the bilinear functional* $\beta_{\widehat{b},k}^\top \Sigma_k^{-1} \beta_{\widehat{p},k}$*, where* $\beta_{\widehat{b},k}$ *and* $\beta_{\widehat{p},k}$ *are the population coefficients of the following linear regressions:*

$$\textit{regress } \lambda^{1/2}(\widehat{b} - b) \textit{ against } \lambda^{1/2}\bar{\mathsf{z}}_k,$$
$$\textit{regress } \lambda^{1/2}(\widehat{p} - p) \textit{ against } \lambda^{1/2}\bar{\mathsf{z}}_k.$$

A.4. **Projection, truncation bias, and the motivation of higher-order influence functions.**

A.4.1. *Proof of Lemma A.2.*

Proof. First, we show the first part of equation (A.6).

$$\begin{aligned} \mathsf{TB}_k(\theta) &= \mathsf{E}_\theta\left[S_{bp}\left(\widetilde{b}_{k,\theta}(X) - b(X)\right)(\widetilde{p}_{k,\theta}(X) - p(X))\right] \\ &= \mathsf{E}_\theta\left[\begin{array}{c} S_{bp}\left\{\widehat{b}(X) - b(X) - \mathsf{E}_\theta\left[(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k))\right]^\top \Sigma_k^{-1}\bar{\mathsf{z}}_k(X)\right\} \\ \times\left\{\widehat{p}(X) - p(X) - \mathsf{E}_\theta\left[(S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k))\right]^\top \Sigma_k^{-1}\bar{\mathsf{z}}_k(X)\right\} \end{array}\right] \\ &= \mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))(\widehat{p}(X) - p(X))\right] \\ &\quad - \mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right] \\ &\quad - \mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right] \\ &\quad + \mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right] \\ &= \underbrace{\mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))(\widehat{p}(X) - p(X))\right]}_{\equiv \mathsf{Bias}_\theta(\widehat{\psi}_1)} \end{aligned}$$

$$-\,\mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X)-b(X))\bar{\mathsf{z}}_k(X)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X)-p(X))\bar{\mathsf{z}}_k(X)\right]$$

where in the third equality we apply Theorem 2.1(3). It follows that

$$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)=\mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X)-b(X))\bar{\mathsf{z}}_k(X)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X)-p(X))\bar{\mathsf{z}}_k(X)\right].$$

Again by Theorem 2.1(3),

$$\begin{aligned}\mathsf{E}_\theta\left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]&=\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)\right]=\mathsf{E}_\theta\left[S_{bp}\left(\widehat{b}(X)-b(X)\right)\bar{\mathsf{z}}_k(X)\right]\\ \mathsf{E}_\theta\left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]&=\mathsf{E}_\theta\left[S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X)+m_1(O,\bar{\mathsf{z}}_k)\right]=\mathsf{E}_\theta\left[S_{bp}\left(\widehat{p}(X)-p(X)\right)\bar{\mathsf{z}}_k(X)\right]\end{aligned}$$

and thus

$$\mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})\right]=\mathsf{E}_\theta\left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]\equiv\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1).$$

In terms of the second part of equation (A.6):

$$\begin{aligned}\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)&=\mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X)-b(X))\bar{\mathsf{z}}_k(X)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X)-p(X))\bar{\mathsf{z}}_k(X)\right]\\&=\mathsf{E}_\theta\left[\lambda(X)(\widehat{b}(X)-b(X))\bar{\mathsf{z}}_k(X)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[\lambda(X)(\widehat{p}(X)-p(X))\bar{\mathsf{z}}_k(X)\right]\\&=\mathsf{E}_\theta\left[\lambda(X)^{1/2}(\widehat{b}(X)-b(X))\bar{v}_k(X)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[\lambda(X)^{1/2}(\widehat{p}(X)-p(X))\bar{v}_k(X)\right]\\&=\mathsf{E}_\theta\left[\Delta_{1,\widehat{b}}\bar{v}_k(X)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[\Delta_{2,\widehat{p}}\bar{v}_k(X)\right]\\&=\mathsf{E}_\theta\left[\Pi_\theta\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)\Pi_\theta\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right].\end{aligned}$$

where in the last line we use the definition of the projection operator $\Pi_\theta\left[\cdot|\bar{v}_k\right]$ (see equation (2.2)).

Then equation (A.7) immediately follows:

$$\begin{aligned}\mathsf{TB}_{\theta,k}&=\mathsf{Bias}_\theta(\widehat{\psi}_1)-\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)\\&=\mathsf{E}_\theta\left[\Delta_{1,\widehat{b}}\Delta_{2,\widehat{p}}\right]-\mathsf{E}_\theta\left[\Pi_\theta\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)^\top\Pi_\theta\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right]\\&=\mathsf{E}_\theta\left[\Pi_\theta^\perp\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)\right]^\top\mathsf{E}_\theta\left[\Pi_\theta^\perp\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right].\end{aligned}$$

Finally, we prove equation (A.8):

$$\begin{aligned}\mathsf{Bias}_\theta(\widehat{\psi}_{2,k}(\Sigma_k^{-1}))&=\mathsf{E}_\theta\left[\widehat{\psi}_{2,k}(\Sigma_k^{-1})-\psi(\theta)\right]\\&=\mathsf{E}_\theta\left[\widehat{\psi}_1-\psi(\theta)-\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})\right]\\&=\mathsf{Bias}_\theta(\widehat{\psi}_1)-\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)=\mathsf{TB}_{\theta,k}\end{aligned}$$

where the third line follows from the definition of $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ and equation (A.6). ■

The statistical properties of the oracle estimator $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and the oracle biased-corrected estimator $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ are given by the following theorem.

Theorem A.1. *Under the conditions of Theorem 2.1 and Condition SW, with $k, n \to \infty$, and $k = o(n^2)$, conditional on the training sample, we have*

(1) *$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is unbiased for $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ with variance of order*

$$\frac{1}{n}\left(\frac{k}{n} + \mathbb{L}^2_{\theta,2,\widehat{b},k} + \mathbb{L}^2_{\theta,2,\widehat{p},k}\right).$$

(2) *$\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} \xrightarrow{d} N(0,1)$, where $\xrightarrow{d}$ stands for convergence in distribution. Further, $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})] \coloneqq \mathsf{var}_\theta^{1/2}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ can be estimated by the bootstrap estimator $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})] \coloneqq \widehat{\mathsf{var}}^{1/2}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ defined in Appendix A.8.*

(3) *$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \pm z_{\alpha^\dagger/2}\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ is a $(1-\alpha^\dagger)$ asymptotic two-sided Wald CI for $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ with length of order*

$$\frac{1}{\sqrt{n}}\left(\sqrt{\frac{k}{n}} + \mathbb{L}_{\theta,2,\widehat{b},k} + \mathbb{L}_{\theta,2,\widehat{p},k}\right).$$

Proof. The variance order of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is proved in Appendix A.7. When $k = o(n^2)$ and $k \to \infty$ as $n \to \infty$, the conditional asymptotic normality of $\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}$ follows directly from Hoeffding decomposition, with the conditional asymptotic normality of the degenerate second-order U-statistic part implied by Bhattacharya and Ghosh (1992, Corollary 1.2). Appendix A.8 proves that the bootstrap standard error estimators $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ satisfy $\frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} = 1 + o_{\mathsf{P}_\theta}(1)$. ∎

Remark A.2. *When $k \gtrsim n^2$, the Gaussian limit of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ does not hold. Further if $k \gg n^2$, $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})] = O(\frac{k}{n^2})$ is of order greater than 1, and therefore $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ cannot consistently estimate $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ regardless of its magnitude.*

A.5. **Statistical properties of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$.** As mentioned above, based on the statistical properties of $\widehat{\psi}_1$ and $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ summarized in Theorems 2.2 and A.1, for a DRF $\psi(\theta)$, we study the statistical property of the oracle two-sided test $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ as a test of the surrogate null hypothesis $\mathsf{H}_{0,k}(\delta)$.

The following theorem characterizes the asymptotic level and power of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ for $\mathsf{H}_{0,k}(\delta)$ and the asymptotic level under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$.

Theorem A.2. *For a DRF $\psi(\theta)$, under the conditions in Theorem 2.1, Theorem A.1 and Condition SW, when $k \to \infty$ but $k = o(n)$, for any given $\varsigma_k, \delta > 0$, suppose that $\frac{|\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)|}{\mathsf{s.e.}_\theta[\widehat{\psi}_1]} = \gamma$ for some (sequence) $\gamma = \gamma(n)$ (where $\gamma(n)$ can diverge with n), then the rejection probability of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ converges to*

$$2 - \Phi\left(\varsigma_k - \lim_{n\to\infty}(\gamma - \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) - \Phi\left(\varsigma_k + \lim_{n\to\infty}(\gamma + \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \tag{A.9}$$

as $n \to \infty$. In particular,

(1) *under $\mathsf{H}_{0,k}(\delta) : \gamma \leqslant \delta$, $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ rejects the null with probability less than or equal to $2(1 - \Phi(\varsigma_k))$, as $n \to \infty$;*

(2) *under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any diverging sequence $c = c(n) \to \infty$, $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ rejects the null with probability converging to 1, as $n \to \infty$.*

(2′) *If $\mathbb{L}_{\theta,2,\widehat{b},k}$ and $\mathbb{L}_{\theta,2,\widehat{p},k}$ converge to 0, under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any fixed $c > 0$ or any diverging sequence $c = c(n) \to \infty$, $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ has rejection probability converging to 1, as $n \to \infty$.*

As an immediate consequence, under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, the rejection probability of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is asymptotically no greater than $\alpha^\dagger$.

Remark A.3. *In Appendix A.5.1, we prove equation (A.9). We now prove that equation (A.9) implies Theorem A.2(1)-(2) and (2′).*

- *Regarding (1), under $\mathsf{H}_{0,k}(\delta) : \gamma \leqslant \delta$,*
$$-(\gamma - \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}]} \geqslant 0 \text{ and } (\gamma + \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}]} \geqslant 0,$$
which implies that the rejection probability is less than or equal to $2 - 2\Phi(\varsigma_k)$. Choose $\varsigma_k = z_{\alpha^\dagger/2}$, $2(1 - \Phi(\varsigma_k)) = 2\alpha^\dagger/2 = \alpha^\dagger$ and conclude that the test is a valid level $\alpha^\dagger$ test of the null.
- *Under the alternative to $\mathsf{H}_{0,k}(\delta)$ with $\gamma = \delta + c$ for some $c > 0$, it follows from Theorem A.1 and equation (A.9) that the rejection probability of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$, as $n \to \infty$, is no smaller than*
$$2 - \Phi\left(\varsigma_k - c\Theta(b, p, \widehat{b}, \widehat{p}, f_X, \bar{\mathsf{z}}_k)\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k} + \mathbb{L}_{\theta,2,\widehat{p},k}\right\}^{-1}\right) - \Phi(\infty)$$
where $\Theta(b, p, \widehat{b}, \widehat{p}, f_X, \bar{\mathsf{z}}_k)$ is some positive constant depending on the true regression functions b and p, the estimated functions $\widehat{b}, \widehat{p}$ from the training sample, the density f_X of X and the chosen basis functions $\bar{\mathsf{z}}_k$. To have power approaching 1 to reject $\mathsf{H}_{0,k}(\delta)$, we need one of the following:
  - *If one of $\mathbb{L}_{\theta,2,\widehat{b},k}$ and $\mathbb{L}_{\theta,2,\widehat{p},k}$ is $O(1)$, we need $c \to \infty$ to guarantee that the rejection probability of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ converges to $1 - \Phi(-\infty) = 1$. Hence we have Theorem A.2(2).*
  - *If c is fixed, we need both $\mathbb{L}_{\theta,2,\widehat{b},k}$ and $\mathbb{L}_{\theta,2,\widehat{p},k}$ to be $o(1)$ to guarantee that the rejection probability of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ converges to $1 - \Phi(-\infty) = 1$. Hence we have Theorem A.2(2′).*

A.5.1. *Proof of Theorem A.2.*

Proof. The rejection probability of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ is computed as follows:

$$\begin{aligned}
&\lim_{n\to\infty} \mathsf{P}_\theta\left(\frac{|\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta\right) \\
= &\lim_{n\to\infty} \left\{\begin{array}{c} \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} > \varsigma_k + \delta\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \\ + \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} < -\varsigma_k - \delta\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \end{array}\right\} \\
= &\lim_{n\to\infty} \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} > \varsigma_k - \frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} + \delta\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \\
&+ \lim_{n\to\infty} \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} < -\varsigma_k - \frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} - \delta\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \mathsf{P}_\theta \left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} (1 + o_{\mathsf{P}_\theta}(1)) > \varsigma_k - (\gamma - \delta) \frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} (1 + o_{\mathsf{P}_\theta}(1)) \right) \\
&\quad + \lim_{n\to\infty} \mathsf{P}_\theta \left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})} (1 + o_{\mathsf{P}_\theta}(1)) < -\varsigma_k - (\gamma + \delta) \frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} (1 + o_{\mathsf{P}_\theta}(1)) \right) \\
&= 1 - \Phi \left(\varsigma_k - (\gamma - \delta) \frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} \right) + \Phi \left(-\varsigma_k - (\gamma + \delta) \frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} \right) \\
&= 2 - \Phi \left(\varsigma_k - (\gamma - \delta) \frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} \right) - \Phi \left(\varsigma_k + (\gamma + \delta) \frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} \right).
\end{aligned}
$$

■

A.6. **Derivation of the higher order influence functions of $\widetilde{\psi}_k(\theta)$ in Theorem A.4.** In this section, we give an almost self-contained derivation of higher order influence functions (HOIFs) of $\widetilde{\psi}_k(\theta)$, except for Robins et al. (2008, Theorem 2.3), which provides the "calculus" of deriving m-th order influence functions from $(m-1)$-th order influence functions.

A.6.1. *A review of the theory of higher order influence functions.* As mentioned above, in this section, we present the higher order influence functions (HOIFs) for the DRFs (Rotnitzky et al., 2021) under a (locally) nonparametric model of every order with the derivation deferred to Appendix A.6. First, we need to formally define HOIFs of a generic functional $\psi(\theta)$, not necessarily in the DRF/DR class. In particular, we use the characterization of HOIFs under the (locally) nonparametric models given in Theorem 2.3 of Robins et al. (2008). We note that in the following $\mathbb{IF}_m(\theta)$ is a m-th order U-statistic and $\mathbb{IF}_{mm}(\theta)$ is a degenerate m-th order U-statistic under P_θ.

Definition A.2. *For a generic functional $\psi(\theta)$, under a (locally) nonparametric model for which the first order influence function $\mathsf{IF}_1(\theta)$ exists. For $m \geqslant 2$, if the $(m-1)$-th order influence function $\mathbb{IF}_{m-1}(\theta)$ exists, then there exists a m-th order influence function $\mathbb{IF}_m(\theta)$ of $\psi(\theta)$, defined recursively via the following rule:*

$$
\mathbb{IF}_1(\theta) \equiv \mathbb{IF}_{11}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{IF}_{1,i}(\theta),
$$

$$
\mathbb{IF}_m(\theta) \equiv \mathbb{IF}_{m-1}(\theta) + \mathbb{IF}_{mm}(\theta)
$$

where $m\mathbb{IF}_{mm}(\theta)$ is the m-th order degenerate U-statistic component of the Hoeffding decomposition of

$$
\frac{(n-m)!}{n!} \sum_{1 \leqslant i_1 \neq \cdots \neq i_m \leqslant n} \mathsf{if}_{1,\mathsf{if}_{m-1,m-1}(O_{i_1},\cdots,O_{i_{m-1}};\cdot)}(O_{i_m};\theta),
$$

if, for each $(o_{i_1}, \cdots, o_{i_{m-1}})$ in the support of $(O_{i_1}, \cdots, O_{i_{m-1}})$, the first order influence function

$$
\mathsf{IF}_{1,\mathsf{if}_{m-1,m-1}(o_{i_1},\cdots,o_{i_{m-1}};\cdot)}(O_{i_m};\theta)
$$

exists for the functional

$$
\mathsf{if}_{m-1,m-1}(o_{i_1}, \cdots, o_{i_{m-1}}; \theta))
$$

given by the kernel of the $(m-1)$-th order U-statistic $\mathbb{IF}_{m-1,m-1}(\theta)$ evaluated at fixed data $o_{i_1}, \cdots, o_{i_{m-1}}$. Otherwise $\mathbb{IF}_m(\theta)$ does not exist.

The general use of HOIF estimators is to decrease bias as given in the following Theorem 2.2 of Robins et al. (2008) .

Theorem A.3 (Theorem 2.2 of Robins et al. (2008))**.** *Given an m-th order influence function $\mathbb{IF}_m(\theta)$ for $\psi(\theta)$, if both $\mathsf{E}_\theta[\mathbb{IF}_m(\widehat{\theta})]$ and $\psi(\widehat{\theta}) - \psi(\theta) \equiv \widehat{\psi} - \psi(\theta)$ have bounded Fréchet derivatives with respect to $\widehat{\theta}$ to order m for a norm $\|\cdot\|$, then the estimator $\psi(\widehat{\theta}) + \mathbb{IF}_m(\widehat{\theta})$ has bias of order $\|\widehat{\theta} - \theta\|^{m+1}$ as an estimator of $\psi(\theta)$ where $\widehat{\theta}$ and $\mathbb{IF}_m(\cdot)$ are computed from independent samples.*

A.6.2. *Higher order influence functions of the DRFs.* With the above preparation, we are ready to state the following theorem characterizing the HOIFs of the DRFs:

Theorem A.4. *Under the conditions of Theorem 2.1, in a locally nonparametric model, under $\mathsf{P}_{\widehat{\theta}}$, the m-th order influence function of $\widetilde{\psi}_k(\theta)$ is $\widehat{\mathbb{IF}}_{m,k}(\Sigma_k^{-1}) = \widehat{\mathbb{IF}}_1 - \widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1})$, where $\widehat{\mathbb{IF}}_{\ell\ell\to mm,k}(\Sigma_k^{-1}) \coloneqq \sum_{j=\ell}^m \widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$,*

$$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \coloneqq \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1})$$

and for $j > 2$,

$$\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1}) \coloneqq \frac{(n-j)!}{n!} \sum_{1 \leqslant i_1 \neq \ldots \neq i_j \leqslant n} \widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}).$$

Here

$$\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \equiv \widehat{\mathsf{IF}}_{22,\widetilde{\psi}_k,\bar{i}_2}(\theta) = \left[\mathcal{E}_{1,\widehat{b}}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^\top \Sigma_k^{-1} \left[\mathcal{E}_{2,\widehat{p}}(\bar{\mathsf{z}}_k)(O)\right]_{i_2},$$

and

$$\begin{aligned} &\widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}) \equiv \widehat{\mathsf{IF}}_{jj,\widetilde{\psi}_k,\bar{i}_j}(\theta) \\ &= (-1)^j \left[\mathcal{E}_{1,\widehat{b}}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^\top \left\{\prod_{s=3}^j \Sigma_k^{-1} \left(\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]_{i_s} - \Sigma_k\right)\right\} \Sigma_k^{-1} \left[\mathcal{E}_{2,\widehat{p}}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}. \end{aligned} \tag{A.10}$$

are the kernels of second order influence function $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and j-th order influence function $\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$ respectively, where $\bar{i}_j \coloneqq \{i_1, \ldots, i_j\}$.

Furthermore, define

$$\widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1}) \coloneqq \sum_{j=2}^m \widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1}).$$

As a consequence, $\widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1})$ is the m-th order influence function of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ under $\mathsf{P}_{\widehat{\theta}}$.

Proof. Without loss of generality, we assume $\mathsf{P}_\theta(S_{bp} \geqslant 0) = 1$. First, let's recall the results from Theorem 2.1 (or Rotnitzky et al. (2021, Theorems 2 (ii))), the influence function of $\psi(\theta)$ has the following form

$$\mathsf{IF}_{1,\psi(\theta)} = S_{bp} b(X) p(X) + m_1(O, b) + m_2(O, p) + S_0 - \psi(\theta),$$

where the linear functionals $m_1(O, h)$ and $m_2(O, h)$ have the Riesz representers $\mathcal{R}_1(X)$ and $\mathcal{R}_2(X)$ respectively. By the definition of Riesz representers, they must obey the following identities:

$$\begin{cases} \mathsf{E}_\theta\left[m_1(O, h)\right] = \mathsf{E}_\theta\left[h(X)\mathcal{R}_1(X)\right] \\ \mathsf{E}_\theta\left[m_2(O, h)\right] = \mathsf{E}_\theta\left[h(X)\mathcal{R}_2(X)\right] \end{cases}$$

for any $h \in L_2(g)$.

To derive the HOIFs for $\widetilde{\psi}_k(\theta)$, we need its first-order influence function. Recall from Definition A.1 that $\widetilde{\psi}_k(\theta)$ is of the following form:

$$\begin{aligned}
&\widetilde{\psi}_k(\theta) \\
&\equiv \mathsf{E}_\theta\left[\mathcal{H}\left(\widetilde{b}_{k,\theta}, \widetilde{p}_{k,\theta}\right)\right] \\
&= \mathsf{E}_\theta\left[S_{bp}\widetilde{b}_{k,\theta}(X)\widetilde{p}_{k,\theta}(X) + m_1(O, \widetilde{b}_{k,\theta}) + m_2(O, \widetilde{p}_{k,\theta}) + S_0\right] \\
&= \mathsf{E}_\theta\left[\begin{array}{c} S_{bp}\left(\widehat{b}(X) + \widetilde{\bar{\zeta}}_{b,k}^\top(\theta)\bar{\mathsf{z}}_k(X)\right)\left(\widehat{p}(X) + \widetilde{\bar{\zeta}}_{p,k}^\top(\theta)\bar{\mathsf{z}}_k(X)^\top\right) \\ + m_1(O, \widehat{b} + \widetilde{\bar{\zeta}}_{b,k}^\top(\theta)\bar{\mathsf{z}}_k) + m_2(O, \widehat{p} + \widetilde{\bar{\zeta}}_{p,k}^\top(\theta)\bar{\mathsf{z}}_k) + S_0 \end{array}\right] \\
&= \mathsf{E}_\theta\left[\begin{array}{c} S_{bp}\left(\widehat{b}(X) - \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)^\top + m_2(O, \bar{\mathsf{z}}_k(X))^\top\right]\bar{\mathsf{z}}_k(X)\right) \\ \times\left(\widehat{p}(X) - \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X)^\top + m_1(O, \bar{\mathsf{z}}_k(X))^\top\right]\bar{\mathsf{z}}_k(X)\right) \\ + m_1(O, \widehat{b} - \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)^\top + m_2(O, \bar{\mathsf{z}}_k(X))^\top\right]\bar{\mathsf{z}}_k) \\ + m_2(O, \widehat{p} - \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X)^\top + m_1(O, \bar{\mathsf{z}}_k(X))^\top\right]\bar{\mathsf{z}}_k) + S_0 \end{array}\right].
\end{aligned}$$

Then

$$\begin{aligned}
\mathsf{IF}_{1,\widetilde{\psi}_k(\theta),i}(\theta) &= \mathcal{H}(\widetilde{b}_k(X,\theta), \widetilde{p}_k(X,\theta)) - \widetilde{\psi}_k(\theta) \\
&\quad + \ \mathsf{E}_\theta\left[\partial\mathcal{H}\left(b_k^*(X, \widetilde{\bar{\zeta}}_{b,k}(\theta)), p_k^*(X, \widetilde{\bar{\zeta}}_{b,k}(\theta))\right)/\partial\bar{\zeta}_{b,k}^\top\right]\mathsf{IF}_{1,\widetilde{\bar{\zeta}}_{b,k}}(\theta) \\
&\quad + \ \mathsf{E}_\theta\left[\partial\mathcal{H}\left(b_k^*(X, \widetilde{\bar{\zeta}}_{b,k}(\theta)), p_k^*(X, \widetilde{\bar{\zeta}}_{b,k}(\theta))\right)/\partial\bar{\zeta}_{p,k}^\top\right]\mathsf{IF}_{1,\widetilde{\bar{\zeta}}_{p,k}}(\theta) \\
&= \mathcal{H}(\widetilde{b}_{k,\theta}, \widetilde{p}_{k,\theta}) - \widetilde{\psi}_k(\theta) \\
&= S_{bp,i}\widetilde{b}_{k,\theta}(X_i)\widetilde{p}_{k,\theta}(X_i) + m_1(O_i, \widetilde{b}_{k,\theta}) + m_2(O_i, \widetilde{p}_{k,\theta}) + S_{0,i} - \widetilde{\psi}_k(\theta).
\end{aligned}$$

where the second equality follows from the definitions of $\widetilde{\bar{\zeta}}_{b,k}(\theta)$ and $\widetilde{\bar{\zeta}}_{p,k}(\theta)$ in equation (A.1). Then the influence function of $-\mathsf{IF}_{1,\widetilde{\psi}_k(\theta),i}(\theta)$, denoted as $\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)$, can be calculated as follows:

$$\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta) = -\frac{\partial\mathcal{H}_{i_1}\left(\widetilde{b}_{k,\theta}(X_{i_1}), \widetilde{p}_{k,\theta}(X_{i_1})\right)}{\partial\bar{\zeta}_{b,k}^\top}\mathsf{IF}_{1,\widetilde{\bar{\zeta}}_{b,k},i_2}(\theta) - \frac{\partial\mathcal{H}_{i_1}\left(\widetilde{b}_{k,\theta}(X_{i_1}), \widetilde{p}_{k,\theta}(X_{i_1})\right)}{\partial\bar{\zeta}_{p,k}^\top}\mathsf{IF}_{1,\widetilde{\bar{\zeta}}_{p,k},i_2}(\theta)$$

where

$$\begin{cases} \frac{\partial\mathcal{H}_{i_1}\left(\widetilde{b}_{k,\theta}(X_{i_1}),\widetilde{p}_{k,\theta}(X_{i_1})\right)}{\partial\bar{\zeta}_{b,k}} = S_{bp,i_1}\widetilde{p}_{k,\theta}(X_{i_1})\bar{\mathsf{z}}_k(X_{i_1}) + m_1(O_{i_1}, \bar{\mathsf{z}}_k), \\ \frac{\partial\mathcal{H}_{i_1}\left(\widetilde{b}_{k,\theta}(X_{i_1}),\widetilde{p}_{k,\theta}(X_{i_1})\right)}{\partial\bar{\zeta}_{p,k}} = S_{bp,i_1}\widetilde{b}_{k,\theta}(X_{i_1})\bar{\mathsf{z}}_k(X_{i_1}) + m_2(O_{i_1}, \bar{\mathsf{z}}_k) \end{cases}$$

and by Lemma A.3

$$\begin{cases} \mathsf{IF}_{1,\widetilde{\bar{\zeta}}_{b,k},i_2}(\theta) = -\Sigma_k^{-1}\left(S_{bp,i_2}\widetilde{b}_{k,\theta}(X_{i_2})\bar{\mathsf{z}}_k(X_{i_2}) + m_2(O_{i_2},\bar{\mathsf{z}}_k)\right), \\ \mathsf{IF}_{1,\widetilde{\bar{\zeta}}_{p,k},i_2}(\theta) = -\Sigma_k^{-1}\left(S_{bp,i_2}\widetilde{p}_{k,\theta}(X_{i_2})\bar{\mathsf{z}}_k(X_{i_2}) + m_1(O_{i_2},\bar{\mathsf{z}}_k)\right). \end{cases}$$

Then

$$\begin{aligned} & \mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta) \\ = & \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^\top \Sigma_k^{-1}\left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2} \\ & + \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_1}^\top \Sigma_k^{-1}\left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_2}. \end{aligned}$$

Then by Robins et al. (2008, part 5.c of Theorem 2.3),

$$\mathbb{IF}_{22,k}(\Sigma_k^{-1}) = \frac{1}{2}\Pi_\theta\left[\mathbb{V}_2\left[\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)\right]|\mathcal{U}_1^{\perp,2}(\theta)\right].$$

Here $\mathcal{U}_m(\theta)$ denotes the Hilbert space of all m-th order U-statistics with mean zero and finite variance with inner product defined by covariances w.r.t. the n-fold product measure $\mathsf{P}_\theta(\cdot)^n$ and $\mathcal{U}_m^{\perp,m+1}(\theta)$ denotes the orthocomplement of $\mathcal{U}_m(\theta)$ in $\mathcal{U}_{m+1}(\theta)$. The notation $\mathbb{V}_m[U_{i_1,\ldots,i_m}]$ maps an m-th order U-statistic kernel $U_{i_1,\ldots,i_m}$ to an m-th order U-statistic:

$$\mathbb{V}_m[U_{i_1,\ldots,i_m}] = \frac{(n-m)!}{n!}\sum_{1\leqslant i_1\neq\cdots\neq i_m\leqslant n} U_{i_1,\ldots,i_m}.$$

For the 2nd order U-statistic kernel $\mathsf{IF}_{1,\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)$, we thus have

$$\begin{aligned} & \mathbb{V}_2\left[\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)\right] \\ = & \mathbb{V}_2\left[\begin{array}{c} \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^\top \Sigma_k^{-1}\left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2} \\ + \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_1}^\top \Sigma_k^{-1}\left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_2} \end{array}\right] \\ = & \mathbb{V}_2\left[2\left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^\top \Sigma_k^{-1}\left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2}\right] \end{aligned}$$

where the last equality is due to the symmetrization of $\mathbb{V}_2[\cdot]$. In addition, we have

$$\Pi_\theta\left[\mathbb{V}_2\left[\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)\right]|\mathcal{U}_1(\theta)\right] = 0$$

since

$$\begin{cases} \mathsf{E}_\theta\left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right] = 0, \\ \mathsf{E}_\theta\left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right] = 0. \end{cases}$$

Thus

$$\mathbb{IF}_{22,k}(\Sigma_k^{-1}) = \frac{1}{2}\mathbb{V}_2\left[\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)\right]$$

$$= \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \left[S_{bp} \widetilde{p}_{k,\theta}(X) \bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp} \widetilde{b}_{k,\theta}(X) \bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right]_{i_2} .$$

Hence formally we have

$$\mathsf{IF}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) = \left[S_{bp} \widetilde{p}_{k,\theta}(X) \bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp} \widetilde{b}_{k,\theta}(X) \bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right]_{i_2}$$

and

$$\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) = \left[S_{bp} \widehat{p}(X) \bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp} \widehat{b}(X) \bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right]_{i_2} .$$

We now complete the proof by induction. We assume that the form for $\widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1})$ is true and thus by the induction hypothesis

$$\begin{aligned} & \mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}) \\ & = (-1)^j \left[S_{bp} \widetilde{p}_{k,\theta}(X) \bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \prod_{s=3}^{j} \Sigma_k^{-1} \left(\left[S_{bp} \bar{\mathsf{z}}_k(X) \bar{\mathsf{z}}_k(X)^{\top}\right]_{i_s} - \Sigma_k \right) \\ & \times \ \Sigma_k^{-1} \left[S_{bp} \widetilde{b}_{k,\theta}(X) \bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right]_{i_2} \end{aligned}$$

and

$$\mathbb{IF}_{jj,k}(\Sigma_k^{-1}) = \mathbb{V}_j \left[\mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1})\right].$$

Then

$$\begin{aligned} & \mathbb{V}_{j+1} \left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right] \\ & = \frac{1}{j+1} \mathbb{V}_{j+1} \left[\Pi_\theta \left[\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j}, i_{j+1}}(\theta) | \mathcal{U}_j^{\perp, j+1}(\theta)\right]\right]. \end{aligned}$$

Then the derivatives w.r.t. θ's in $\widetilde{\widetilde{\zeta}}_{b,k}(\theta)$, $\widetilde{\widetilde{\zeta}}_{p,k}(\theta)$ and $(j-1)$ terms of Σ_k^{-1} will be contributing terms to $\mathbb{V}_{j+1} \left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right]$. But differentiating w.r.t. $(j-2)$ terms of Σ_k will not contribute terms to $\mathbb{V}_{j+1} \left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right]$ as the contribution of these $(j-2)$ terms to $\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j}, i_{j+1}}(\theta)$ is a function of j units' data and is thus an element of $\mathcal{U}_j(\theta)$. Then as in the proof of Lemma A.3,

$$\mathsf{IF}_{1,\Sigma_k^{-1}}(\theta) = -\Sigma_k^{-1} \left(S_{bp} \bar{\mathsf{z}}_k(X) \bar{\mathsf{z}}_k(X)^{\top} - \Sigma_k\right) \Sigma_k^{-1}$$

and upon permuting the indices, the contribution of each of these $(j-1)$ terms to $\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j}, i_{j+1}}(\theta)$ is

$$\begin{aligned} & -(-1)^j \left[S_{bp} \widetilde{p}_{k,\theta}(X) \bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \\ & \times \sum_{s=3}^{j+1} \Sigma_k^{-1} \left\{ \left(S_{bp} \bar{\mathsf{z}}_k(X) \bar{\mathsf{z}}_k(X)^{\top}\right)_{i_s} - \Sigma_k \right\} \\ & \times \Sigma_k^{-1} \left[S_{bp} \widetilde{b}_{k,\theta}(X) \bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right]_{i_2} \end{aligned} \tag{A.11}$$

which is degenerate and thus orthogonal to $\mathcal{U}_j(\theta)$. Notice that equation (A.11) is exactly the kernel of $\mathbb{IF}_{(j+1),(j+1),k}(\Sigma_k^{-1})$. Then differentiating w.r.t. θ's in $\widetilde{p}_{k,\theta}$ and $\widetilde{b}_{k,\theta}$, we further obtain terms of the following form contributing to

$\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)$:

$$\begin{aligned}&(-1)^j \mathsf{IF}_{1,\tilde{\tilde{\zeta}}_{p,k},i_{j+1}}(\theta)^\top \left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]_{i_1} \left[\sum_{s=3}^{j} \Sigma_k^{-1} \left\{\left(S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right)_{i_s} - \Sigma_k\right\}\right] \\ &\times \Sigma_k^{-1}\left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2} \\ &+ (-1)^j \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^\top \left[\sum_{s=3}^{j} \Sigma_k^{-1} \left\{\left(S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right)_{i_s} - \Sigma_k\right\}\right] \\ &\times \Sigma_k^{-1}\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]_{i_2} \mathsf{IF}_{1,\tilde{\tilde{\zeta}}_{b,k},i_{j+1}}(\theta).\end{aligned} \tag{A.12}$$

Applying Lemma A.3 again and projecting the above display onto $\mathcal{U}_j^{\perp,j+1}(\theta)$, we obtain that

$$\Pi_\theta\left[\text{equation (A.12)}|\mathcal{U}_j^{\perp,j+1}(\theta)\right] \equiv 2 \times \text{equation (A.11)}.$$

Recall that there are $(j-1)$ terms of equation (A.11) in total contributing to $\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)$, thus eventually there are $(j+1)$ terms of equation (A.11) in total contributing to $\mathbb{V}_{j+1}\left[\Pi_\theta\left[\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)|\mathcal{U}_j^{\perp,j+1}(\theta)\right]\right]$. Further recall that

$$\begin{aligned}&\mathbb{V}_{j+1}\left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right] \\ &= \frac{1}{j+1}\mathbb{V}_{j+1}\left[\Pi_\theta\left[\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)|\mathcal{U}_j^{\perp,j+1}(\theta)\right]\right].\end{aligned}$$

Therefore we have

$$\begin{aligned}\mathbb{IF}_{(j+1),(j+1),k}(\Sigma_k^{-1}) &\equiv \mathbb{V}_{j+1}\left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right] \\ &= \frac{1}{j+1}\mathbb{V}_{j+1}\left[(j+1)\times \text{equation (A.11)}\right] \\ &= \mathbb{V}_{j+1}\left[\text{equation (A.11)}\right].\end{aligned}$$

We thus prove the results for general j by induction. ■

Lemma A.3. *As defined in equation (A.1), the influence functions of $\tilde{\tilde{\zeta}}_{b,k}(\theta)$ and $\tilde{\tilde{\zeta}}_{p,k}(\theta)$ are*

$$\begin{cases}\mathsf{IF}_{1,\tilde{\tilde{\zeta}}_{b,k}}(\theta) = -\Sigma_k^{-1}\left(S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right), \\ \mathsf{IF}_{1,\tilde{\tilde{\zeta}}_{p,k}}(\theta) = -\Sigma_k^{-1}\left(S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right).\end{cases} \tag{A.13}$$

Proof. We only need to show $\widehat{\mathsf{IF}}_{1,\tilde{\tilde{\zeta}}_{b,k}}(\theta)$ and $\widehat{\mathsf{IF}}_{1,\tilde{\tilde{\zeta}}_{p,k}}(\theta)$ will follow by symmetry. Recall that

$$\tilde{\tilde{\zeta}}_{b,k}(\theta) = -\left\{\mathsf{E}_\theta\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]\right\}^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right].$$

Then

$$\begin{aligned}&\mathsf{IF}_{1,\tilde{\tilde{\zeta}}_{b,k}}(\theta) \\ &= -\left\{\mathsf{E}_\theta\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]\right\}^{-1}\mathsf{IF}_{1,\mathsf{E}.\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)\right]}(\theta)\end{aligned}$$

$$
\begin{aligned}
& - \mathsf{IF}_{1,\left\{\mathsf{E}_{\cdot}\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]\right\}^{-1}}(\theta)\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right] \\
= & - \Sigma_k^{-1}\left(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k) - \mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right]\right) \\
& + \Sigma_k^{-1}\left(S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top - \Sigma_k\right)\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right] \\
= & - \Sigma_k^{-1}\left(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k) - S_{bp}\bar{\mathsf{z}}_k(X)^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right]\bar{\mathsf{z}}_k(X)\right) \\
= & - \Sigma_k^{-1}\left(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k) + S_{bp}\bar{\mathsf{z}}_k(X)^\top\widetilde{\overline{\zeta}}_{b,k}(\theta)\bar{\mathsf{z}}_k(X)\right) \\
= & - \Sigma_k^{-1}\left\{S_{bp}\left(\widehat{b}(X) + \bar{\mathsf{z}}_k(X)^\top\widetilde{\overline{\zeta}}_{b,k}(\theta)\right)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right\} \\
= & - \Sigma_k^{-1}\left(S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)\right)
\end{aligned}
$$

where the fourth equality follows from the definition of $\widetilde{\overline{\zeta}}_{b,k}(\theta)$ and the sixth equality follows from the definition of $\widetilde{b}_{k,\theta}(X)$. ∎

Under conditions in Theorem 2.1, $\widetilde{b}_{k,\widehat{\theta}} = \widehat{b}$ and $\widetilde{p}_{k,\widehat{\theta}} = \widehat{p}$ or equivalently $\widetilde{\overline{\zeta}}_{b,k}(\widehat{\theta}) = \widetilde{\overline{\zeta}}_{p,k}(\widehat{\theta}) = 0$. This follows from results in Theorem 2.1 with θ evaluated at $\widehat{\theta}$:

$$\mathsf{E}_{\widehat{\theta}}\left[S_{bp}\widehat{b}(X)h(X) + m_2(O, h)\right] = 0 \text{ for all } h \in \mathcal{B}, \mathsf{E}_{\widehat{\theta}}\left[S_{bp}\widehat{p}(X)h(X) + m_1(O, h)\right] = 0 \text{ for all } h \in \mathcal{P}.$$

Thus under $\widehat{\theta}$, the HOIFs of $\widetilde{\psi}_k(\widehat{\theta})$ are: $\widehat{\mathbb{IF}}_{m,k}(\Sigma_k^{-1}) = \widehat{\mathbb{IF}}_1 - \widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1})$, where $\widehat{\mathbb{IF}}_{\ell\ell\to mm,k}(\Sigma_k^{-1}) := \sum_{j=\ell}^m \widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$,

$$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) := \frac{1}{n(n-1)} \sum_{1\leqslant i_1 \neq i_2 \leqslant n} \widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \tag{A.14}$$

and for $j > 2$,

$$\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1}) := \frac{(n-j)!}{n!} \sum_{1\leqslant i_1 \neq \ldots \neq i_j \leqslant n} \widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}). \tag{A.15}$$

Here

$$\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \equiv \mathsf{IF}_{22,\widetilde{\psi}_k,\bar{i}_2}(\widehat{\theta}) = \left[\mathcal{E}_{\widehat{b},m_2}\left(\bar{\mathsf{z}}_k\right)(O)\right]_{i_1}^\top \Sigma_k^{-1}\left[\mathcal{E}_{\widehat{p},m_1}\left(\bar{\mathsf{z}}_k\right)(O)\right]_{i_2}, \tag{A.16}$$

and

$$
\begin{aligned}
& \widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}) \equiv \mathsf{IF}_{jj,\widetilde{\psi}_k,\bar{i}_j}(\widehat{\theta}) \\
& = (-1)^j\left[\mathcal{E}_{\widehat{b},m_2}\left(\bar{\mathsf{z}}_k\right)(O)\right]_{i_1}^\top\left\{\prod_{s=3}^j \Sigma_k^{-1}\left(\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]_{i_s} - \Sigma_k\right)\right\}\Sigma_k^{-1}\left[\mathcal{E}_{\widehat{p},m_1}\left(\bar{\mathsf{z}}_k\right)(O)\right]_{i_2}.
\end{aligned}
\tag{A.17}
$$

are the kernels of second order influence function $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and j-th order influence function $\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$ respectively, where $\mathcal{E}_{\widehat{b},m_2}$ and $\mathcal{E}_{\widehat{p},m_1}$ are defined in equations (A.4) and (A.5).

As a consequence of Theorem A.4, the m-th order influence function for the truncated parameter $\widetilde{\psi}_k(\theta)$ under $\mathsf{P}_{\widehat{\theta}}$ is simply $\widehat{\psi}_{m,k}(\Sigma_k^{-1}) - \widetilde{\psi}_k(\theta) \equiv \widehat{\mathbb{IF}}_{m,k}(\Sigma_k^{-1})$ and the m-th order influence function for the bias $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ is $\widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1})$, with the b and p in $\mathbb{IF}_{m,k}(\Sigma_k^{-1})$ and $\mathbb{IF}_{22\to mm,k}(\Sigma_k^{-1})$ replaced by $\widehat{b}$ and $\widehat{p}$.

A.7. **General formula for the variance order of $\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})$.** In this section we derive the following formula for the order of $\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})\right]$ for general m. Then we have

Lemma A.4. *When $m \geqslant 3$,*

(A.18)

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})\right] \overset{\textit{Condition } W}{\lesssim} \frac{1}{n}\left\{\begin{array}{c}\left(\frac{k}{n}\right)^{m-1} + \sum_{j=2}^{m-1}\binom{m}{j}^2\left(\frac{k}{n}\right)^{j-1}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-j-2)\vee 0}\left\{\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^4 + \mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + k\mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,2,\widehat{p},k}^2\right\} \\ + m^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-3)}\left\{\mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + k\mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,2,\widehat{p},k}^2\right\}\end{array}\right\}$$

$$\overset{\textit{Condition } SW}{\lesssim} \frac{1}{n}\left\{\begin{array}{c}\left(\frac{k}{n}\right)^{m-1} + \sum_{j=2}^{m-1}\binom{m}{j}^2\left(\frac{k}{n}\right)^{j-1}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-j-2)\vee 0}\left\{\begin{array}{c}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^4 + \mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 \\ + \left\{\mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,\infty,\widehat{p},k}^2 \wedge \mathbb{L}_{\theta,2,\widehat{p},k}^2\mathbb{L}_{\theta,\infty,\widehat{b},k}^2\right\}\end{array}\right\} \\ + m^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-3)}\left\{\mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \left\{\mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,\infty,\widehat{p},k}^2 \wedge \mathbb{L}_{\theta,2,\widehat{p},k}^2\mathbb{L}_{\theta,\infty,\widehat{b},k}^2\right\}\right\}\end{array}\right\}.$$

Proof. Equation A.18 follows from Hoeffding decomposition (Hoeffding, 1948) of U-statistics $\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})\right]$. We simply bound each term in the decomposition. The detailed calculations can be found in the forthcoming updated version of Liu et al. (2017). ∎

A.8. **Bootstrapping higher order influence functions.** In this section, we will use higher order moments of multinomial distributions frequently. Two good references are Mosimann (1962); Newcomer et al. (2008). For notational convenience, we suppress the dependence on Σ_k^{-1} or $\widehat{\Sigma}_k^{-1}$ in $\widehat{\mathbb{IF}}_{mm,k}(\Sigma_k^{-1})$ and $\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})$ as the proposed bootstrap resampling procedure will not depend on Σ_k^{-1} or the training sample. We propose the following bootstrap estimator of $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]$: choose some large integer M as the number of bootstrap resamples, and define

$$\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}] \coloneqq \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{22,k}^{(m)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{22,k}^{(m)}\right)^2}_{T_1} - \underbrace{\frac{2}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{22,k}^{(m),c} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{22,k}^{(m),c}\right)^2}_{T_2} \tag{A.19}$$

where

$$\widehat{\mathbb{IF}}_{22,k}^{(m)} = \frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n} W_{i_1}^{(m)}W_{i_2}^{(m)}\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}$$

$$\widehat{\mathbb{IF}}_{22,k}^{(m),c} = \frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n} (W_{i_1}^{(m)}-1)(W_{i_2}^{(m)}-1)\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}$$

and $\boldsymbol{W}^{(m)} = (W_1^{(m)}, \ldots, W_n^{(m)}) \overset{i.i.d.}{\sim} \mathsf{Multinom}(n, \underbrace{n^{-1}, \ldots, n^{-1}}_{(n-1)'\mathrm{s}})$ for $m = 1, \ldots, M$ are M multinomial weights independent of the observed data $\{O_i, i = 1, \ldots, N\}$.

Remark A.4. *We now explain why $\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]$ is proposed as an estimator of $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]$. Conditional on the observed data (including both the training and estimation samples), the expectation over the random multinomial weights $\boldsymbol{W}$ of the term T_1 is:*

$$\mathsf{E}_{\boldsymbol{W}}[T_1|\{O_i\}_{i=1}^N]$$

$$
\begin{aligned}
&= \mathsf{E}_{\boldsymbol{W}}\left[\left(\frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n} W_{i_1}^{(m)}W_{i_2}^{(m)}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\right)^2 \lvert\{O_i\}_{i=1}^N\right] \\
&\quad - \left(\mathsf{E}_{\boldsymbol{W}}\left[\frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n} W_{i_1}^{(m)}W_{i_2}^{(m)}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\lvert\{O_i\}_{i=1}^N\right]\right)^2 \\
&= \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\leqslant n}\{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}^2W_{i_2}^2]-(\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}])^2\} \\
&\quad + \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}\left\{\begin{array}{c}
\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}^2W_{i_2}W_{i_3}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_3}]\} \\
+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}^2W_{i_2}W_{i_3}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_3}W_{i_1}]\} \\
+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}^2W_{i_3}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_2}W_{i_3}]\} \\
+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}^2W_{i_3}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_3}W_{i_2}]\}
\end{array}\right\} \\
&\quad + \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\neq i_4\leqslant n}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}W_{i_3}W_{i_4}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_3}W_{i_4}]\}.
\end{aligned}
$$

Plugging in the following higher order moments of $\mathsf{Multinom}(n, \underbrace{n^{-1},\ldots,n^{-1}}_{(n-1)'s})$ (Newcomer et al., 2008)*:*

$$
\begin{aligned}
\mathsf{E}_{\boldsymbol{W}}[W_1W_2] &= \frac{n-1}{n}, \mathsf{E}_{\boldsymbol{W}}[W_1W_2W_3] = \frac{(n-1)(n-2)}{n^2}, \mathsf{E}_{\boldsymbol{W}}[W_1W_2W_3W_4] = \frac{(n-1)(n-2)(n-3)}{n^3}, \\
\mathsf{E}_{\boldsymbol{W}}[W_1^2W_2] &= \frac{2(n-1)^2}{n^2}, \mathsf{E}_{\boldsymbol{W}}[W_1^2] = \frac{2n-1}{n}, \\
\mathsf{E}_{\boldsymbol{W}}[W_1^2W_2W_3] &= \frac{(n-1)(n-2)(n-3)+n(n-1)(n-2)}{n^3} \\
&= \frac{(n-1)(2n^2-7n+6)}{n^3}, \\
\mathsf{E}_{\boldsymbol{W}}[W_1^2W_2^2] &= \frac{(n-1)(n-2)(n-3)+2n(n-1)(n-2)+n^2(n-1)}{n^3} \\
&= \frac{(n-1)(4n^2-9n+6)}{n^3},
\end{aligned}
$$

we have

$$
\begin{aligned}
&\mathsf{E}_{\boldsymbol{W}}[T_1\lvert\{O_i\}_{i=1}^N] \\
&= \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\leqslant n}\{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\left(3-\frac{11}{n}+\frac{14}{n^2}-\frac{6}{n^3}\right) \\
&\quad + \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}\left\{\begin{array}{c}
\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\
+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}
\end{array}\right\}\left(1-\frac{7}{n}+\frac{12}{n^2}-\frac{6}{n^3}\right) \\
&\quad + \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\neq i_4\leqslant n}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\left(-\frac{4}{n}+\frac{10}{n^2}-\frac{6}{n^3}\right)
\end{aligned}
\tag{A.20}
$$

The term T_2 is the bootstrap variance estimator for second order degenerate U-statistics and it has been proposed previously in Arcones and Gine (1992); Huskova and Janssen (1993)*. Similarly, conditional on the observed data (including both the* *training and estimation samples), the expectation over the random multinomial weights $\boldsymbol{W}$ of the term T_2 is:*

$$
\begin{aligned}
&\mathsf{E}_{\boldsymbol{W}}[T_2 | \{O_i\}_{i=1}^N] \\
= &\frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \{\widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_1}\} \{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)^2 (W_{i_2}-1)^2] - (\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)])^2\} \\
+ &\frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left\{ \begin{array}{c} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_3} \{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)^2 (W_{i_2}-1)(W_{i_3}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)] \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_3}-1)]\} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_1} \{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)^2 (W_{i_2}-1)(W_{i_3}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)] \mathsf{E}_{\boldsymbol{W}}[(W_{i_3}-1)(W_{i_1}-1)]\} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_3} \{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)^2 (W_{i_3}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)] \mathsf{E}_{\boldsymbol{W}}[(W_{i_2}-1)(W_{i_3}-1)]\} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_2} \{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)^2 (W_{i_3}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)] \mathsf{E}_{\boldsymbol{W}}[(W_{i_3}-1)(W_{i_2}-1)]\} \end{array} \right\} \\
+ &\frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_4} \{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)(W_{i_3}-1)(W_{i_4}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)] \mathsf{E}_{\boldsymbol{W}}[(W_{i_3}-1)(W_{i_4}-1)]\} \\
= &\frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \{\widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_1}\} \{\mathsf{E}_{\boldsymbol{W}}[(W_1^2 - 2W_1 + 1)(W_2^2 - 2W_2 + 1)] - (\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)])^2\} \\
+ &\frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left\{ \begin{array}{c} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_2} \end{array} \right\} \{\mathsf{E}_{\boldsymbol{W}}[(W_1-1)^2 (W_2-1)(W_3-1)] - (\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)])^2\} \\
+ &\frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_4} \{\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)(W_3-1)(W_4-1)] - (\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)])^2\} \\
= &\frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \{\widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_1}\} \left(1 - \frac{3}{n} + \frac{7}{n^2} - \frac{6}{n^3} - \frac{1}{n^2}\right) \\
+ &\frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left\{ \begin{array}{c} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_2} \end{array} \right\} \left(-\frac{1}{n} + \frac{5}{n^2} - \frac{6}{n^3} - \frac{1}{n^2}\right) \\
+ &\frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_4} \left(\frac{3}{n^2} - \frac{6}{n^3} - \frac{1}{n^2}\right).
\end{aligned}
$$

Hence T_2 serves as a correction term of the inflated factor 3 appeared in equation (A.20).

Combining the above computations, we have

$$
\begin{aligned}
&\mathsf{E}_{\boldsymbol{W}}[\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}] | \{O_i\}_{i=1}^N] = \mathsf{E}_{\boldsymbol{W}}[T_1 | \{O_i\}_{i=1}^N] - \mathsf{E}_{\boldsymbol{W}}[T_2 | \{O_i\}_{i=1}^N] \\
= &\frac{1}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \{\widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_1}\} \left(1 - \frac{5}{n} + \frac{2}{n^2} + \frac{6}{n^3}\right) \\
+ &\frac{1}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left\{ \begin{array}{c} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_2} \end{array} \right\} \left(1 - \frac{5}{n} + \frac{6}{n^3}\right) \\
- &\frac{4}{n} \frac{1}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_4} \left(1 - \frac{1.5}{n} - \frac{1.5}{n^2}\right).
\end{aligned}
$$

Finally, it is not difficult to see that

$$
\mathsf{E}_{\theta}[\mathsf{E}_{\boldsymbol{W}}[\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}] | \{O_i\}_{i=1}^N]] = \mathsf{var}_{\theta}[\widehat{\mathbb{IF}}_{22,k}] \left(1 + O(n^{-1})\right).
$$

Proposition A.1. *Under Condition SW, as $n \to \infty$ and $M \to \infty$,*

$$\frac{\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]} = 1 + o_{\mathsf{P}_\theta}(1).$$

Proof. Let the $\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]$ be the limit (in $\mathsf{P}_{\boldsymbol{W}}$-probability) of $\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]$ as $M \to \infty$ given all the observed data. Using the moments of multinomial distributions (Newcomer et al., 2008), together with the explanation in Comment A.4, it is easy to show that $\mathsf{E}_\theta\left[\frac{\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]}\right] = 1 + o(1)$ and $\mathsf{var}_\theta\left[\frac{\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]}\right] = o(1)$. Combining the above two claims, we have $\frac{\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]} = 1 + o_{\mathsf{P}_\theta}(1)$. ■

Following computations similar in spirit to but more tedious than the computations above, we propose the following estimator of $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{33,k}]$:

$$\begin{aligned}
\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{33,k}] \coloneqq\ & \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{33,k}^{(m)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m)}\right)^2}_{S_1} \\
& - \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,p)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,p)}\right)^2}_{S_2} \\
& - \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,\Sigma)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,\Sigma)}\right)^2}_{S_3} \\
& - \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{33,k}^{(m),c(p,\Sigma)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m),c(p,\Sigma)}\right)^2}_{S_4}
\end{aligned} \tag{A.21}$$

where

$$\begin{aligned}
\widehat{\mathbb{IF}}_{33,k}^{(m)} &= \frac{1}{n(n-1)(n-2)}\sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} W_{i_1}^{(m)}W_{i_2}^{(m)}W_{i_3}^{(m)}\widehat{\mathsf{IF}}_{33,k,\bar{i}_3} \\
\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,p)} &= \frac{1}{n(n-1)(n-2)}\sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} (W_{i_1}^{(m)}-1)W_{i_2}^{(m)}(W_{i_3}^{(m)}-1)\widehat{\mathsf{IF}}_{33,k,\bar{i}_3} \\
\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,\Sigma)} &= \frac{1}{n(n-1)(n-2)}\sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} (W_{i_1}^{(m)}-1)(W_{i_2}^{(m)}-1)W_{i_3}^{(m)}\widehat{\mathsf{IF}}_{33,k,\bar{i}_3} \\
\widehat{\mathbb{IF}}_{33,k}^{(m),c(p,\Sigma)} &= \frac{1}{n(n-1)(n-2)}\sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} W_{i_1}^{(m)}(W_{i_2}^{(m)}-1)(W_{i_3}^{(m)}-1)\widehat{\mathsf{IF}}_{33,k,\bar{i}_3}.
\end{aligned}$$

We then have:

Proposition A.2. *Under Condition SW, as $n \to \infty$ and $M \to \infty$,*

$$\frac{\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{33,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{33,k}]} = 1 + o_{\mathsf{P}_\theta}(1).$$

Finally, we consider estimating $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}]$ through nonparametric bootstrap. It is not difficult to see that the following estimator has the desired property $\frac{\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22\to 33,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}]} = 1 + o_{\mathsf{P}_\theta}(1)$ as $M \to \infty$:

$$\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22\to 33,k}] = \widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}] + \widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{33,k}] + 2\widehat{\mathsf{cov}}[\widehat{\mathbb{IF}}_{22,k}, \widehat{\mathbb{IF}}_{33,k}]$$

where

$$\begin{aligned}\widehat{\mathsf{cov}}[\widehat{\mathbb{IF}}_{22,k}, \widehat{\mathbb{IF}}_{33,k}] = {} & \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{22,k}^{(m)} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{22,k}^{(m')}\right)\left(\widehat{\mathbb{IF}}_{33,k}^{(m)} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m')}\right) \\ & - \frac{2}{M-1}\sum_{m=1}^{M-1}\left(\widehat{\mathbb{IF}}_{22,k}^{(m),c} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{22,k}^{(m'),c}\right)\left(\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,p)} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m'),c(b,p)}\right).\end{aligned}$$

It is not difficult to generalize the above arguments to construct bootstrapped estimators of $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{mm,k}]$ for general $m \geqslant 2$. We plan to report the general construction elsewhere.

A.9. **Proof of Theorem 3.2 and Proposition 4.1.** As discussed in Sections 3 and 4.3, since Σ_k^{-1} is generally unknown, $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ needs to be replaced by $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$. We consider the following statistic used in the test $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ after standardization:

(A.22)

$$\begin{aligned}
& \frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \delta\right) \\
= {} & \left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right) + \frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \delta\right) \\
= {} & \left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right) + \frac{\mathsf{EB}_{\theta,m,k}}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \\
& + \frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \delta\right) \\
= {} & \left\{\left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right) + \left(\frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) + \mathsf{EB}_{\theta,m,k}}{\mathsf{s.e.}_\theta[\widehat{\psi}_1]} - \delta\right)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right\}(1 + o_{\mathsf{P}_\theta}(1)) \\
= {} & \left\{\left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right) + \left(\gamma + \frac{\mathsf{EB}_{\theta,m,k}}{\mathsf{s.e.}_\theta[\widehat{\psi}_1]} - \delta\right)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right\}(1 + o_{\mathsf{P}_\theta}(1)) \\
= {} & \left\{\underbrace{\left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right) - (\delta - \gamma)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}}_{A} + \underbrace{\frac{\mathsf{EB}_{\theta,m,k}}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}}_{B}\right\}(1 + o_{\mathsf{P}_\theta}(1)).
\end{aligned}$$

The effect of estimating Σ_k^{-1} on the asymptotic validity of the test $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ of $\mathsf{H}_{0,k}(\delta)$ thus depends on the orders of terms A and B. A has variance 1 and mean $-(\delta - \gamma)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}$. B depends on the estimation bias due to estimating Σ_k^{-1} by $\widehat{\Sigma}_k^{-1}$. Hence if we have:

(1) $\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}$ is asymptotically $N(0,1)$ conditional on the training sample;

(2) $\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \Big/ \frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \to 1$;

(3) $B = o(1)$ under $\mathsf{H}_{0,k}(\delta)$ or fixed alternatives to $\mathsf{H}_{0,k}(\delta)$;

(3′) $B \ll -(\delta-\gamma)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}$ under diverging alternatives to $\mathsf{H}_{0,k}(\delta)$ i.e. $\delta - \gamma = c$ for some $c \to \infty$ (at any rate).

$\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is an asymptotically level $\alpha^\dagger$ two-sided test for $\mathsf{H}_{0,k}(\delta)$ and rejects the null with probability approaching 1 under diverging alternatives to $\mathsf{H}_{0,k}(\delta)$.

According to Theorem 4.1, both (1) and (2) hold for $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$. In terms of (3) and (3′), under the conditions of Proposition 4.1:

- Under $\mathsf{H}_{0,k}(\delta)$ or fixed alternatives to $\mathsf{H}_{0,k}(\delta)$, $\mathsf{EB}_{\theta,m,k}(\widehat{\Sigma}_k^{-1}) \lesssim \frac{(k\log k)^{(m-1)/2}}{n^{m/2}} \ll \frac{\sqrt{k}}{n} + \frac{1}{\sqrt{n}}$ and hence $B = o(1)$. Thus (3) is satisfied.
- Under diverging alternatives to $\mathsf{H}_{0,k}(\delta)$ i.e. $\gamma - \delta = c \to \infty$, $(\gamma-\delta)\mathsf{s.e.}_\theta(\widehat{\psi}_1) \asymp \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}$,
$$\begin{aligned} B &\lesssim \frac{\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{k\log k}{n}\right)^{(m-1)/2} \ll \frac{\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \\ &\asymp -(\delta-\gamma)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}. \end{aligned}$$
Thus (3′) is satisfied.

In summary, $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid level $\alpha^\dagger$ two-sided test for $\mathsf{H}_{0,k}(\delta)$ and rejects the null with probability approaching 1 under diverging alternatives to $\mathsf{H}_{0,k}(\delta)$.

A.10. **A brief review of refined DML estimators.**

Appendix B. Supplementary information for simulation studies

The functions h_f, h_b and h_p appeared in Section 5 are of the following forms:

$$h_f(x; s_f) \propto 1 + \exp\left\{\frac{1}{2}\sum_{j\in\mathcal{J},\ell\in\mathbb{Z}} 2^{-j(s_f+0.5)}\omega_{j,\ell}(x)\right\}, \tag{B.1}$$

$$h_b(x; s_b) = \sum_{j\in\mathcal{J},\ell\in\mathbb{Z}} 2^{-j(s_b+0.5)}\omega_{j,\ell}(x), \tag{B.2}$$

$$h_p(x; s_p) = -2\sum_{j\in\mathcal{J},\ell\in\mathbb{Z}} 2^{-j(s_p+0.5)}\omega_{j,\ell}(x) \tag{B.3}$$

where $\mathcal{J} = \{0, 3, 6, 9, 10, 16\}$ and $\omega_{j,\ell}(\cdot)$ is the D12 (or equivalently db6) father wavelets function dilated at resolution j, shifted by ℓ. Härdle et al. (1998)[Theorem 9.6] indeed implies that $h_f(\cdot; s_f) \in$ Hölder(s_f), $h_b(\cdot; s_b) \in$ Hölder(s_b) and $h_p(\cdot; s_p) \in$ Hölder(s_p). We fix $s_f = 0.1$. In simulation setup I we choose $s_b = s_p = 0.25$ whereas in simulation setup II we choose $s_b = s_p = 0.6$.

In Table 5, we provide the numerical values for $(\tau_{b,j}, \tau_{p,j})_{j=1}^{8}$ used in generating the simulation experiments in Section 5.

j	$\tau_{b,j}$	$\tau_{p,j}$
1	-0.2819	0.09789
2	0.4876	0.08800
3	-0.1515	-0.4823
4	-0.1190	0.4588

Table 5. Coefficients used in constructing b and π in Section 5.

B.1. **Generating correlated multidimensional covariates X with fixed non-smooth marginal densities.** In the simulation study conducted in Section 5, one key step of generating the simulated datasets is to draw correlated multidimensional covariates $X \in [0,1]^d$ with fixed non-smooth marginal densities. First, we fix the marginal densities of X in each dimension proportional to $h_f(\cdot)$ (equation (B.1)). Then we make $2K$ independent draws of $\widetilde{X}_{i,j}$, $i = 1, \ldots, 2K$, from h_f for every $j = 1, \ldots, d$ so $\widetilde{X} = (\widetilde{X}_{1,\cdot}, \ldots, \widetilde{X}_{2K,\cdot})^\top \in [0,1]^{2K \times d}$. Next, to create correlations between different dimensions, we follow the strategy proposed in Baker (2008). First we group every two consecutive draws:

$$(\widetilde{X}_{1,\cdot}, \widetilde{X}_{2,\cdot})^\top, (\widetilde{X}_{3,\cdot}, \widetilde{X}_{4,\cdot})^\top, \ldots, (\widetilde{X}_{2K-1,\cdot}, \widetilde{X}_{2K,\cdot})^\top.$$

Then for each pair $(\widetilde{X}_{2i-1,\cdot}, \widetilde{X}_{2i,\cdot})^\top$ for $i = 1, \ldots, K$, we form the following d-dimensional random vectors

$$U_i \coloneqq (\mathsf{max}(\widetilde{X}_{2i-1,1}, \widetilde{X}_{2i,1}), \ldots, \mathsf{max}(\widetilde{X}_{2i-1,d}, \widetilde{X}_{2i,d}))^\top,$$
$$V_i \coloneqq (\mathsf{min}(\widetilde{X}_{2i-1,1}, \widetilde{X}_{2i,1}), \ldots, \mathsf{min}(\widetilde{X}_{2i-1,d}, \widetilde{X}_{2i,d}))^\top.$$

Lastly, we construct K independent d-dimensional vectors X by the following rule: for each $i = 1, \ldots, K$, we draw a Bernoulli random variable B_i with probability 1 / 2, and if $B_i = 0$, $X_{i,\cdot} = U_i$, otherwise $X_{i,\cdot} = V_i$. Following the above strategy, we conserve the marginal density of $X_{\cdot,j}$ as that of $\widetilde{X}_{\cdot,j}$ but create dependence between different dimensions.

Appendix C. Proximal higher-order influence functions

In this section, we further generalize the derivations in Appendix A.6 for $\psi(\theta) = -\mathsf{E}_\theta[Y(a)]$ to the case under the proximal causal inference framework considered in Cui et al. (2020); Miao et al. (2018); Tchetgen Tchetgen et al. (2020). The purpose of this section is to show that our framework can be extended to the endogeneity settings which economists are extremely interested in; e.g. Newey and Powell (2003). We only consider $\psi(\theta) = -\mathsf{E}_\theta[Y(a)]$, since it is yet unclear what is the "right" generalization under endogeneity for the general DRFs considered in Rotnitzky et al. (2021). We leave such generalization to future work.

We immediately begin with the formal setup, followed with a theorem on the forms of HOIFs of $\psi(\theta)$. Instead of observing i.i.d. $O_i = (X_i, A_i, Y_i)_{i=1}^n$ under strong ignorability, there might exist latent confounding variable U, so $\psi(\theta)$ is not point identified in general. But suppose we observe $O_i = (X_i, Z_i, A_i, W_i, Y_i)_{i=1}^n$, where W and Z are "proxy variables" of Y and A respectively, satisfying the following proximal causal identification conditions given in Tchetgen Tchetgen et al. (2020):

Condition C.1.

(1) *Consistency:* $Y = Y_A$ *almost surely;*

(2) *No direct effect from the treatment* A *and treatment proxy* Z *to the outcome proxy* W*:* $W_{a,z} = W$ *for all* a *and* z *almost surely;*

(3) *No direct effect from the treatment proxy* Z *to the outcome* Y*:* $Y_{a,z} = Y_a$ *for all* a *and* z *almost surely;*

(4) $(Z, A) \perp\!\!\!\perp (Y_a, W) | (U, X)$ *for* $a = 0, 1$*;*

(5) $0 < \Pr(A = a | U, X) < 1$ *almost surely for* $a = 0, 1$*;*

(6) *For any* a, x*, if* $\mathsf{E}_\theta[g(U) | W, A = a, X = x] = 0$ *almost surely, then* $g(U) = 0$ *almost surely;*

(7) *For any* a, x*, if* $\mathsf{E}_\theta[g(W) | Z, A = a, X = x] = 0$ *almost surely, then* $g(W) = 0$ *almost surely;*

For more details of proximal causal inference and the intuition behind the above regularity conditions, we refer interested readers to many recent papers on this topic (Cui et al., 2020; Kallus et al., 2021; Miao et al., 2018; Shpitser et al., 2021; Tchetgen Tchetgen et al., 2020).

Then we have

Theorem C.1. *Under the above Condition C.1 (Assumptions 1, and 4-9 of Cui et al. (2020)), together with the following conditions on the so-called confounding bridge functions* r *and* q *(Assumption 10 of Cui et al. (2020)): Denote* $\Psi(z, a, x) \equiv \mathsf{E}_\theta[(Y - r(W, A = a, X))^2 | Z = z, A = a, X = x]$ *and* $\Gamma(w, a, x) \equiv \mathsf{E}_\theta\left[\left(q(Z, A, X) - \frac{1}{f(A|W,X)}\right)^2 | W = w, A = a, X = x\right]$*:*

(1) $0 < \inf\limits_{z,a,x} \Psi(z, a, x) \leqslant \sup\limits_{z,a,x} \Psi(z, a, x) < \infty$;

(2) $0 < \inf\limits_{w,a,x} \Gamma(w, a, x) \leqslant \sup\limits_{w,a,x} \Gamma(w, a, x) < \infty$;

(3) *Let* $T : L_2(W, A, X) \to L_2(Z, A, X)$ *be the conditional expectation operator given by* $T(g) \equiv \mathsf{E}_\theta[g(W, A, X) | Z, A, X]$ *and the adjoint* $T' : L_2(Z, A, X) \to L_2(W, A, X)$ *be* $T'(g) \equiv \mathsf{E}_\theta[g(Z, A, X) | W, A, X]$*.*

Suppose that the inverse mapping $\{T' \Psi^{-1} T\}^{-1}$ *exists and is uniquely defined and that*

$T' \Psi^{-1} \mathsf{E}_\theta[(Y - r(W, A, X))(r(W, A = a, X) - \psi(\theta)) | Z, A, X] \in$ *Domain*$(\{T' \Psi^{-1} T\}^{-1})$

$\mathbb{1}\{A = a\} / f(A | W, X) \in$ *Domain*$(\{T' \Psi^{-1} T\}^{-1})$.

Then the confounding bridge functions r *and* q *that solve integral equations*

$$
\text{(C.1)} \qquad \begin{aligned}
\mathsf{E}_\theta[Y | X, A = a, Z] &= \int r(w, A = a, X) \mathrm{d}F_\theta(w | X, Z), \\
\frac{1}{\Pr_\theta(A = a | X, W)} &= \int q(z, A = a, X) \mathrm{d}F_\theta(z | X, W, A = a).
\end{aligned}
$$

are uniquely identified. The nonparametric first order influence function of $\psi(\theta) = -\mathsf{E}_\theta[Y(a)]$ *is given by*

$$
\text{(C.2)} \quad \mathsf{IF}_{1,\psi}(\theta) = \mathbb{1}\{A = a\} q(Z, A = a, X) r(W, A = a, X) - \mathbb{1}\{A = a\} Y q(Z, A = a, X) - r(W, A = a, X) - \psi(\theta).
$$

In the main text, all the notations related to b *and* p *are extended to* h *and* q*. For the truncated nuisance parameters* $\widetilde{r}_{k,\theta}$ *and* $\widetilde{q}_{k,\theta}$*, we use* k*-dimensional basis* $\bar{\mathfrak{z}}_k(W, A, X)$ *and* $\bar{\mathsf{z}}_k(Z, A, X)$*, respectively.*

Furthermore, the HOIFs of order m *of* $\widetilde{\psi}_k(\theta) = \mathsf{E}_\theta\left[\mathcal{H}(\widetilde{r}_{k,\theta}, \widetilde{q}_{k,\theta})\right]$ *are* $\mathbb{IF}_{m,k}(\Sigma_k^{-1}) = \mathbb{IF}_1 - \mathbb{IF}_{22\to mm,k}(\Sigma_k^{-1})$*, where* $\mathbb{IF}_{\ell\ell\to mm,k}(\Sigma_k^{-1}) \coloneqq \sum_{j=\ell}^{m} \mathbb{IF}_{jj,k}(\Sigma_k^{-1})$*,* $\Sigma_k \coloneqq \mathsf{E}_\theta[\mathbb{1}\{A=a\}\bar{\mathfrak{z}}_k(W, A=a, X)\bar{\mathsf{z}}_k(Z, A=a, X)^\top]$ *which is assumed to be invertible,*

$$\mathbb{IF}_{22,k}(\Sigma_k^{-1}) \coloneqq \frac{1}{n(n-1)} \sum_{1\leqslant i_1 \neq i_2 \leqslant n} \mathsf{IF}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \tag{C.3}$$

and for $j > 2$*,*

$$\mathbb{IF}_{jj,k}(\Sigma_k^{-1}) \coloneqq \frac{(n-j)!}{n!} \sum_{1\leqslant i_1 \neq \ldots \neq i_j \leqslant n} \mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}). \tag{C.4}$$

Here

(C.5)

$$\mathsf{IF}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \equiv \mathsf{IF}_{22,\widetilde{\psi}_k,\bar{i}_2}(\theta) = \left[\mathbb{1}\{A=a\}(Y - \widetilde{r}_{k,\theta}(W, A=a, X))\bar{\mathsf{z}}_k(Z, A=a, X)\right]_{i_1}^\top \Sigma_k^{-1} \left[\bar{\mathfrak{z}}_k(W, A=a, X)(\mathbb{1}\{A=a\}\widetilde{q}_{k,\theta}(Z, A=a, X) - 1)\right]_{i_2},$$

and

$$\begin{aligned}
&\mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}) \equiv \mathsf{IF}_{jj,\widetilde{\psi}_k,\bar{i}_j}(\theta) \\
= {}& (-1)^j \left[\mathbb{1}\{A=a\}(Y - \widetilde{r}_{k,\theta}(W, A=a, X))\bar{\mathsf{z}}_k(Z, A=a, X)\right]_{i_1}^\top \left\{\prod_{s=3}^{j} \Sigma_k^{-1}\left(\left[\mathbb{1}\{A=a\}\bar{\mathfrak{z}}_k(W, A=a, X)\bar{\mathsf{z}}_k(Z, A=a, X)^\top\right]_{i_s} - \Sigma_k\right)\right\} \\
&\times \Sigma_k^{-1}\left[\bar{\mathfrak{z}}_k(W, A=a, X)(\mathbb{1}\{A=a\}\widetilde{q}_{k,\theta}(Z, A=a, X) - 1)\right]_{i_2}.
\end{aligned}$$

are the kernels of second order influence function $\mathbb{IF}_{22,k}(\Sigma_k^{-1})$ *and* j*-th order influence function* $\mathbb{IF}_{jj,k}(\Sigma_k^{-1})$ *respectively, where* $\bar{i}_j \coloneqq \{i_1, \ldots, i_j\}$*.*

Note that the first order influence function in the above theorem has appeared in Cui et al. (2020). We keep it in the above theorem for the sake of completeness. The forms of HOIFs are new, but are straightforward applications of the same derivation given in Appendix A.6 for DRFs.