# New $\sqrt{n}$-consistent, numerically stable higher-order influence function estimators

Chang Li[1] and Lin Liu*[2]

[1]Department of Statistics, University of Virginia, Charlottesville, VA, USA
[2]Institute of Natural Sciences, MOE-LSC, School of Mathematical Sciences, CMA-Shanghai, SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University and Shanghai Artificial Intelligence Laboratory, Shanghai, China

February 3, 2023

**Abstract**

Higher-Order Influence Functions (HOIFs) provide a unified theory for constructing rate-optimal estimators for a large class of low-dimensional (smooth) statistical functionals/parameters (and sometimes even infinite-dimensional functions) that arise in substantive fields including epidemiology, economics and the social sciences. Since the introduction of HOIFs in Robins et al. (2008) or Robins et al. (2016)[1], they have been viewed mostly as a theoretical benchmark rather than a useful tool for statistical practice. Works aimed to flip the script are scant, but a few recent papers Liu et al. (2017, 2021b) make some partial progress. However, based on strong empirical evidence from large-scale synthetic experiments (Liu et al., 2020a), all these efforts still fell short. In this paper, we take a fresh attempt at achieving this goal by constructing new, numerically stable HOIF estimators (or sHOIF estimators for short with "s" standing for "stable") with provable statistical, numerical and computational guarantees. This new class of sHOIF estimators (up to the 3rd order) was foreshadowed in synthetic experiments conducted in Liu et al. (2020a).

**Keywords:** Causal Inference, Functional Estimation, Higher-Order Influence Functions, Semiparametric Theory, Combinatorics

*Correspondence: linliu@sjtu.edu.cn. The authors would like to thank Fengnan Gao, Zhenyu Liao, Rajarshi Mukherjee, Jamie Robins, and Zheng Zhang for invaluable discussions on this paper. The authors gratefully acknowledges funding support by NSFC Grant No.12101397 and No.12090024, Shanghai Municipal Science and Technology Grant No.2021SHZDZX0102, Shanghai Science and Technology Commission Grant No.21JC1402900, Shanghai Natural Science Foundation Grant No.21ZR1431000.

[1]Robins et al. (2016) is the complete version of Robins et al. (2008), including more results and proofs. We therefore only refer to Robins et al. (2016) in the sequel.

# 1 Introduction

*Higher-Order Influence Functions* (HOIFs) (Robins et al., 2016) are higher-order generalizations of the first-order influence functions (IFs), a staple in semiparametric statistical theory (Newey, 1990; Bickel et al., 1998; van der Vaart, 2002). HOIFs are a powerful and unified approach to constructing minimax rate-optimal estimators for a class of statistical functionals/parameters (and sometimes even functions; see Kennedy et al. (2022)) that arise in (bio)statistics, epidemiology, economics, and the social sciences. HOIF estimators originally proposed in Robins et al. (2016, 2017)[2] remain the only known minimax rate-optimal estimators for statistical functionals/parameters with substantive interests in the above disciplines, including the Average Treatment Effect (ATE) under the strong ignorability assumption[3] and the expected conditional covariance of two random variables $A$ and $Y$ given a third random variable $X$, even after highly active research by the statistics and econometrics communities in recent years (Newey and Robins, 2018; Kennedy, 2020; Hirshberg and Wager, 2021; Yu and Wang, 2020). More recent works (Kennedy et al., 2022; Bonvini and Kennedy, 2022) also initiated the application of the HOIF machinery to the minimax optimal estimation of Conditional Average Treatment Effect (CATE) function or dose response curves. Their results lay important theoretical foundation for individualized decision making problems, e.g. personalized medicine. This is the first instance when HOIF estimators are also shown to be effective, at least in theory, for function estimation problems, or more precisely, "hybrid function and functional estimation problems". Similar idea has also been applied to dose-response curve estimation (Bonvini and Kennedy, 2022). For an introductory level review of HOIFs, we refer the interested readers to van der Vaart (2014) and Section 1 of Liu et al. (2020b). A relatively more technical review of HOIFs is delegated to Section 1.3.

Over the past decade, Robins and colleagues initiated the research program of establishing theoretical foundations for HOIFs and estimators based on HOIFs (Robins, 2004; van der Vaart, 2014; Robins et al., 2016, 2017; Liu et al., 2017) for a class of statistical functionals/parameters recently characterized in Rotnitzky et al. (2021), which are heretofore termed as *Doubly Robust Functionals* (DRF) in this paper. We adopt this terminology to reflect the fact that their nonparametric first-order IFs give rise to doubly robust estimators (Scharfstein et al., 1999; Robins and Rotnitzky, 2001; Chernozhukov et al., 2018a). This class of DRFs subsumes the class of functionals studied in Robins et al. (2016) and Chernozhukov et al. (2018b). Under the standard Hölder-regularity assumptions on the nuisance parameters (abbreviated as Hölder nuisance models), Robins et al. (2017) constructed minimax optimal but non-adaptive HOIF estimators for a sub-class of DRFs. Liu et al. (2021a) constructed adaptive second-order IF estimators for DRFs using the celebrated Lepskiĭ's adaptation scheme (Lepskii, 1991), within a strict submodel of the Hölder nuisance models. But both estimators require estimating the density of the potentially high-dimensional covariates $X$, even in $\sqrt{n}$-estimable regimes. When the dimension $d$ of the covariates is only moderately

[2]See Robins et al. (2022) for corrections of the proofs in Robins et al. (2017).

[3]In Liu et al. (2021b), we derived the HOIFs for the ATE functional even when the strong ignorability assumption fails to hold, provided that we have access to valid proxies for both the treatment and outcome, following a series of works on proximal causal learning (Tchetgen Tchetgen et al., 2020).

large (e.g. $d = 10$), nonparametric density estimation is already a daunting computational and statistical task, a manifestation of the so-called "curse-of-dimensionality".

To overcome the above issues, Liu et al. (2017) introduced *empirical* HOIF (eHOIF for short) estimators that obviate multi-dimensional density estimation by inverting the sample/empirical Gram matrix of vector-valued basis transformation of the covariates computed using a separate sample independent of the sample used to construct the estimator of the DRF. This sample-splitting strategy is adopted mainly for simplifying the mathematical analysis, leading to rather straightforward analysis of the statistical properties of the eHOIF estimators. In particular, the eHOIF estimators are still the only class of estimators that achieves $\sqrt{n}$-consistency and semiparametric efficiency for DRFs under the minimal Hölder-regularity assumptions (Robins et al., 2009). These nice statistical properties of the eHOIF estimators also motivate the development of a class of assumption-lean hypothesis tests statistic that is designed to falsify if the standard $(1 - \alpha) \times 100\%$ Wald confidence interval of the DRF has the claimed coverage probability (Liu et al., 2020a, 2021b). At this point, astute readers must wonder why we need a new class of empirical HOIF estimators at all, which is what this article is all about.

## 1.1 Motivation and main contributions

Despite the effort in Liu et al. (2017), from our past experience of using eHOIF estimators in practice (Liu et al., 2017, 2020a, 2021b; Wanis et al., 2023), several singular issues of their finite-sample performance were unveiled by large-scale simulation experiments[4]:

(i) Numerical instability: In Liu et al. (2017), although eHOIF estimators exhibit better finite-sample performance than the original HOIF estimators in Robins et al. (2017), the simulations were restricted to very low condition number $k/n$: e.g. $k \approx 1,000$ and $n \approx 10,000$. In the simulation studies of Liu et al. (2020a), when $k$ gets near $n$, eHOIF estimators blow up numerically (see Section S3.1 of Liu et al. (2020a)) already at order two. What is more striking is that the eHOIF estimators at higher orders, though supposed to be correcting the bias, can only exacerbate the numeric blow-up.

(ii) Non-monotone bias reduction: Theoretical results in Liu et al. (2017) hint that increasing the orders of the estimator should *in principle* reduce the bias. However, we found that this is not usually the case for eHOIF estimators in practice (e.g. see Section 5 of (Liu et al., 2021b)). Interestingly, sHOIF estimators do not seem to suffer from this problem in simulations, elevating the theoretical results from *mere principles* closer to *empirical facts*; see Liu et al. (2020a) or Wanis et al. (2023) for simulations at orders 2 or 3.

Our contributions are three-fold.

[4] For interested readers, these simulation experiments have also been used to expose the gap between the (nonparametric) statistical theory deep neural networks (DNNs) and their practice in Xu et al. (2022). One can access computer codes of generating such simulations here.

- Methodology and practical relevance: This article proposes a new class of numerically stable sHOIF estimators for DRFs, that overcomes the above two major limitations of eHOIF estimators. The stable Second-Order IF (SOIF) estimators first appeared in the simulation studies of Liu et al. (2020a), but their statistical properties remain elusive.

- Theory and the proof strategy: Obtaining a deeper theoretical underpinning of this phenomenon mandates meticulous calculations rather than crude upper bounds. This is the critical technical innovation vis-à-vis other HOIF-related works. In particular, we intensively use the following proof techniques: leave-out analysis, matrix-valued Taylor expansion, and combinatorial calculations (i.e. corollaries of the binomial identity). The proof strategy developed in this paper may be of independent interest.

- Extensions of sHOIFs beyond standard settings: We also generalize sHOIF estimators to all the DRFs and ATEs under the proximal causal learning setting, allowing us to handle the tantalizing issue of unmeasured confounding (or endogeneity) in the current causal inference (or econometrics) literature.

## 1.2 Notation

Before proceeding, we gather some frequently used notation throughout the paper. We denote the observed data random vector as $O \in \mathcal{O}$, where $\mathcal{O}$ is its corresponding sample space. Let $\bar{\mathsf{z}}_k \coloneqq (z_1, \cdots, z_k)^\top$ denote a collection of $k$ different functions, each of which has input domain $\mathcal{X}$. Fix some $\theta' \in \Theta$. $\mathbb{E}_{\theta'}$, $\mathsf{var}_{\theta'}$, and $\mathsf{cov}_{\theta'}$ are, respectively, the expectation, variance, and covariance operators under the probability law $\mathbb{P}_{\theta'}$. For any measurable function $h : \mathcal{X} \to \mathbb{R}$, let $\|h\|_\infty \coloneqq \operatorname{ess\,sup}_{x \in \mathcal{X}} h(x)$ and $\|h\|_{\theta',p} \coloneqq \{\mathbb{E}_{\theta'}[h(X)^p]\}^{1/p}$ for any $p \geq 1$. We adopt standard (stochastic) asymptotic notation $\lesssim$, $\gtrsim$, $\asymp$, $\gg$, $\ll$, $o(\cdot)$, $\omega(\cdot)$, $O(\cdot)$, $\Omega(\cdot)$, $o_{\mathbb{P}_{\theta'}}(\cdot)$, $O_{\mathbb{P}_{\theta'}}(\cdot)$. For any real-valued vector $\bar{v}$ and any $q \in \mathbb{R}$, let $v^q$ be the element-wise $q$-th power of $v$.

Furthermore, define $\Sigma_{\theta'} \coloneqq \mathbb{E}_{\theta'}[Q]$, where $Q \coloneqq A\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top$, as the ($A$-weighted) population Gram matrix of $\bar{\mathsf{z}}_k(X)$, till Section 5.2, in which we generalize all our results from $\psi(\theta) \coloneqq \mathbb{E}_\theta[Y(a=1)]$[5], the mean of outcome $Y$ in the treated group under strong ignorability, to all members of the DRFs. Similarly, define $\widehat{\Sigma} \coloneqq \mathbb{P}_n[Q_k] \equiv n^{-1}\sum_{i=1}^n Q_i$ as the ($A$-weighted) sample Gram matrix of $\bar{\mathsf{z}}_k(X)$, again till Section 5.2. Here $\mathbb{P}_n[\cdot]$ denotes the sample mean operator. To further lighten the notation, we let $Q_I \coloneqq \sum_{i \in I} Q_i$ for any multi-index set $I \subseteq [n]$. For convenience, we also denote multi-index set $\{i_1, i_2, \cdots, i_j\} \subseteq [n]$ as $\bar{i}_j$ for $j \leq n$. $\Omega_{\theta'} \equiv \Sigma_{\theta'}^{-1}$ and $\widehat{\Omega} \equiv \widehat{\Sigma}^{-1}$, when they exist, are respectively the inverse of the population and sample Gram matrices. The kernels constructed from $\bar{\mathsf{z}}_k$ are denoted as $K_{\theta',k}(x, x') \coloneqq \bar{\mathsf{z}}_k(x)^\top \Omega_{\theta'} \bar{\mathsf{z}}_k(x')$ and $\widehat{K}_k(x, x') = \bar{\mathsf{z}}_k(x)^\top \widehat{\Omega} \bar{\mathsf{z}}_k(x')$. Given a set of functions $\bar{\mathsf{v}}_k : \mathcal{X} \to \mathbb{R}^k$ and any $L_2$ function $h : \mathcal{O} \to \mathbb{R}$, $\Pi[h|\bar{\mathsf{v}}_k]$ denotes the linear projection operator of projecting $h$ onto the linear span of $\bar{\mathsf{v}}_k$: formally,

$$\Pi[h|\bar{\mathsf{v}}_k](x) \coloneqq \bar{\mathsf{v}}_k(x)^\top \mathbb{E}[\bar{\mathsf{v}}_k(X)\bar{\mathsf{v}}_k(X)^\top]^{-1} \mathbb{E}[\bar{\mathsf{v}}_k(X)h(O)].$$

[5] We use the potential outcome notation without introducing it, which will not affect the understanding of the main theme of this work.

We use $\mathbb{P}_\theta$ to denote the true data generating law, unless stated otherwise. When the reference measure is the true law $\mathbb{P}_\theta$, we often drop the dependence on $\theta$: for example, we write $\|h\|_p \equiv \|h\|_{\theta,p}$, $\Sigma \equiv \Sigma_\theta$, $\Omega \equiv \Omega_\theta$ and $\mathbb{P}$, $\mathbb{E}$, $\mathsf{var}$, $\mathsf{cov}$ correspond to $\mathbb{P}_\theta$, $\mathbb{E}_\theta$, $\mathsf{var}_\theta$, $\mathsf{cov}_\theta$. Note that $\Omega$ should not be confused with the asymptotic notation $\Omega(\cdot)$ and this will be clear from the context. A statistic is said to be "oracle" whenever it depends on some part(s) of the unknown *true* data generating law $\mathbb{P}_\theta$ (such as $\Omega$); otherwise it is said to be "feasible". We also introduce $\mathsf{Diag}$ as the operator of extracting the diagonal elements of a matrix.

Finally, let $\mathbb{U}_{n,m}[\cdot]$ denote the $m$-th order $U$-statistic operator: for any function $h : \mathbb{R}^m \to \mathbb{R}$

$$\mathbb{U}_{n,m}[h(O_1, \cdots, O_m)] \coloneqq \frac{(n-m)!}{n!} \sum_{1 \le i_1 \ne \cdots \ne i_m \le n} h(O_{i_1}, \cdots, O_{i_m}).$$

When $m = 1$, $\mathbb{U}_{n,m}[\cdot]$ reduces to the sample mean operator $\mathbb{P}_n[\cdot]$. Similarly, let $\mathbb{V}_{n,m}$ be the corresponding $V$-statistic operator[6]:

$$\mathbb{V}_{n,m}[h(O_1, \cdots, O_m)] \coloneqq \frac{(n-m)!}{n!} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(O_{i_1}, \cdots, O_{i_m}).$$

Later in the paper, for $m \ge 2$, we will define "oracle" $m$-th order influence function estimators constructed using the dictionary $\bar{\mathsf{z}}_k$, denoted as $\widehat{\mathbb{IF}}_{m,m,k}(\Omega) \equiv \widehat{\mathbb{IF}}_{m,m,k} = \mathbb{U}_{n,m}[\widehat{\mathsf{IF}}_{m,m,k,\bar{i}_m}]$ with $U$-statistic kernel $\mathsf{IF}_{m,m,k,\bar{i}_m} \equiv \widehat{\mathsf{IF}}_{m,m,k,\bar{i}_m}(\Omega)$. Its stable feasible version is denoted by $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})$ with the corresponding kernel $\widehat{\mathsf{IF}}_{m,m,k,\bar{i}_m}(\widehat{\Omega})$.

## 1.3 The setup and a review of the theory of HOIFs

With the notation just introduced, we are poised to state the problem setup and briefly review the theory of HOIFs relevant for this paper, in particular the theory of eHOIFs. For other related works, see Section **??** for an extended discussion.

Suppose that we are given $N$ i.i.d. observations $\{O_i\}_{i=1}^N \sim \mathbb{P}_\theta$, where $\theta \in \Theta$ is the so-called nuisance parameter and $\Theta$ is its underlying parameter space. Let $\mathcal{P} \coloneqq \{\mathbb{P}_\theta : \theta \in \Theta\}$ be the space of data generating probability measures. Our primary interest is to estimate and draw statistical inference on a smooth statistical functional $\psi(\theta) :\to \mathbb{R}$, in the sense of van der Vaart (1991). We restrict $\psi(\theta)$ to be the DRFs defined in Rotnitzky et al. (2021). As mentioned, our running example is $\psi(\theta) = \mathbb{E}_\theta[Y(a = 1)]$ the mean of an outcome $Y$ in the treated group $A = 1$. Here the observed data specializes to $O = (X, A, Y)$: respectively the $d$-dimensional covariates belonging to a compact subset $\mathcal{X} \equiv [-B, B]^d$ of $\mathbb{R}^d$, the binary treatment assignment, and the bounded outcome variable. Under unconfoundedness assumption (relaxed in Section 5.3), $\psi(\theta)$ can be identified by either of the two statistical functionals of the observed data distribution:

$$\psi(\theta) \equiv \mathbb{E}\left[Aa(X)Y\right] \equiv \mathbb{E}[b(X)] \tag{1}$$

where $a(x) \coloneqq \{\mathbb{E}[A|X = x]\}^{-1}$ and $b(x) \coloneqq \mathbb{E}[Y|X = x, A = 1]$ except Section 5.2. For this functional $\psi(\theta)$, the nuisance parameter is $\theta \equiv (a, b, g)$ where $g(x)$ is the probability

[6] Here we use the scaling $\frac{(n-m)!}{n!}$ instead of the more conventional $\frac{1}{n^m}$ for notational convenience.

density/mass function of the covariates $X$ conditional on $A = 1$. Hence the nuisance parameter space $\Theta = \mathcal{A} \times \mathcal{B} \times \mathcal{G}$, where $\mathcal{A}, \mathcal{B}, \mathcal{G}$ are, respectively, the space where $a, b, g$ lie. We further divide the whole $N$ data points into two parts: one with sample size $n$, called the estimation sample, and the other with sample size $N - n$, called the nuisance sample used to estimate the nuisance parameter $\theta$. Throughout this paper, we condition on the nuisance sample data by treating it or any quantity computed from it as fixed.

For a smooth statistical functional $\psi(\theta) : \Theta \to \mathbb{R}$ in the sense of van der Vaart (1991), its first-order influence function $\mathbb{IF}_1(\theta)$ is a mean-zero first-order $U$-statistic satisfying the following functional equation

$$\left.\frac{\mathrm{d}\psi(\theta_t)}{\mathrm{d}t}\right|_{t=0} = \mathbb{E}\left[\mathbb{IF}_1(\theta) \cdot \mathbb{S}_1\right]$$

where $\mathbb{P}_{\theta_t}$ is any parametric submodels in $\{\mathbb{P}_\theta, \theta \in \Theta\}$, such that when $t = 0$, $\mathbb{P}_{\theta_t} \equiv \mathbb{P}_\theta$, the true data generating law, and $\mathbb{S}_1$ is its first-order score vector, as defined in Waterman and Lindsay (1996); also see Robins et al. (2016). Here $\mathbb{IF}_1(\theta)$ has the following form (Robins et al., 1994):

$$\mathbb{IF}_1(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathsf{IF}_{1,i}(\theta), \text{ where } \mathsf{IF}_1(\theta) = Aa(X)(Y - b(X)) + b(X) - \psi(\theta). \tag{2}$$

Typically, classical semiparametric theory (Newey, 1990; Bickel et al., 1998) constructs semiparametric efficient first-order estimators $\widehat{\psi}_1$ of $\psi(\theta)$ based on its first-order influence function follows:

$$\widehat{\psi}_1 = \frac{1}{n}\sum_{i=1}^{n} A_i \widehat{a}(X_i)(Y_i - \widehat{b}(X_i)) + \widehat{b}(X_i)$$

where $\widehat{a}, \widehat{b}$ are nuisance parameter estimates computed from the nuisance sample. In particular, $\widehat{\psi}_1$ has bias

$$\begin{aligned} \mathsf{bias}(\widehat{\psi}_1) &= \mathbb{E}[\widehat{\psi}_1 - \psi(\theta)] = \mathbb{E}[\mathsf{IF}_1(\widehat{\theta}) - \mathsf{IF}_1(\theta)] \\ &= \mathbb{E}\left[\left(\frac{\widehat{a}(X)}{a(X)} - 1\right)(b(X) - \widehat{b}(X))\right]. \end{aligned} \tag{3}$$

Formally, $\mathsf{bias}(\widehat{\psi}_1)$ is a *product of two nuisance estimation errors*[7], and hence *doubly-robust* (Scharfstein et al., 1999).

Despite being doubly-robust, the veracity of inference based on first-order estimators like $\widehat{\psi}_1$ may nonetheless be questionable when the nuisance parameter $\theta$ is of high complexity: e.g. functions with low smoothness or without sparsity. For example, when $a, b$ belong to Hölder functions with smoothness $s_a, s_b$ and $g$ arbitrarily complex, by far no first-order estimators are known to be $\sqrt{n}$-consistency for estimating $\psi(\theta)$ throughout the entire range

$$\{(s_a, s_b) : (s_a + s_b)/2 \geq d/4\} \tag{4}$$

[7] Rotnitzky et al. (2021) actually define the general class of statistical functionals that permit doubly-robust estimators based on this second-order bias property; see Section 5.2.

but the eHOIF estimators of Liu et al. (2017) or the original HOIF estimators of Robins et al. (2016) if additionally assuming $g$ to be Hölder with smoothness $s_g > 0$. In fact, Robins et al. (2009) also showed that (4) is the minimal condition for the existence of $\sqrt{n}$-consistent estimators of $\psi(\theta)$ under the Hölder nuisance modeling assumption. Outside (4), $\psi(\theta)$ is non $\sqrt{n}$-estimable and the only known estimator with the optimal rate of convergence in minimax sense is again the HOIF estimator (Robins et al., 2016, 2017, 2022). When restricting to highly smooth $g$, Liu et al. (2021a) construct minimax optimal and adaptive estimator of $\psi(\theta)$ by combining the HOIF estimators with the celebrated Lepskii's adaptation scheme (Lepskii, 1991).

This article is about the $\sqrt{n}$-estimable regime (4), so we will focus our attention on the eHOIF estimators. First, we choose a set of $k$-dimensional functions $\bar{\mathsf{z}}_k \equiv (z_1, \cdots, z_k)^\top : \mathcal{X} \to \mathbb{R}^k$ satisfying certain regularity conditions to be given later in Section 2. The Second-Order Influence Function (SOIF) estimator of $\psi(\theta)$ is the following second-order $U$-statistic:

$$\widehat{\psi}_{2,k}(\Omega) \coloneqq \widehat{\psi}_1 + \widehat{\mathbb{IF}}_{2,2,k} \text{ where } \widehat{\mathbb{IF}}_{2,2,k} \equiv \widehat{\mathbb{IF}}_{2,2,k}(\Omega) \coloneqq \mathbb{U}_{n,2}\left[\mathsf{IF}_{2,2,k;1,2}\right] \tag{5}$$

and

$$\begin{aligned}\mathsf{IF}_{2,2,k;1,2} \equiv \mathsf{IF}_{2,2,k;1,2}(\Omega) &\coloneqq \left(A_1 \widehat{a}(X_1) - 1\right) \bar{\mathsf{z}}_k(X_1)^\top \Omega \bar{\mathsf{z}}_k(X_2) A_2 (Y_2 - \widehat{b}(X_2)) \\ &\equiv \left(A_1 \widehat{a}(X_1) - 1\right) K_k(X_1, X_2) A_2 (Y_2 - \widehat{b}(X_2)).\end{aligned}$$

Based on the definition of HOIFs (Robins et al., 2016), $-\widehat{\mathbb{IF}}_{2,2,k}$ is in fact the SOIF of $\mathsf{bias}(\widehat{\psi}_1)$[8]. A more intuitively appealing explanation goes as follows: $-\widehat{\mathbb{IF}}_{2,2,k}$ is an unbiased estimator of the following quantity:

$$\begin{aligned}\mathsf{bias}_k(\widehat{\psi}_1) &= \mathbb{E}\left[\left(\frac{\widehat{a}(X)}{a(X)} - 1\right) \bar{\mathsf{z}}_k(X)^\top\right] \Omega \mathbb{E}\left[A \bar{\mathsf{z}}_k(X)(b(X) - \widehat{b}(X))\right] \\ &= \mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)} - 1\right) K_k(X_1, X_2) A_2 (b(X_2) - \widehat{b}(X_2))\right]\end{aligned} \tag{6}$$

which is simply replacing the estimation errors $\widehat{a}/a - 1$ and $b - \widehat{b}$ in (3) by

$$\Pi\left[\frac{\widehat{a}}{a} - 1 \middle| \bar{\mathsf{z}}_k\right] \text{ and } \Pi\left[b - \widehat{b} \middle| A \bar{\mathsf{z}}_k\right].$$

Hence $\widehat{\mathbb{IF}}_{2,2,k}$ can be interpreted as a bias correction term that *partially* debiases $\mathsf{bias}(\widehat{\psi}_1)$.

However, evaluating $\Omega$ in practice relies on the knowledge of $g$, which is generally unknown to the analyst. The initial attempt by Robins et al. (2016) and Robins et al. (2017) was to estimate $g$ from the nuisance sample by $\widehat{g}$, leading to statistical properties affected by $g - \widehat{g}$ and thus complexity-reducing assumptions on $\mathcal{G} \ni g$. To completely resolve this reliance, Liu et al. (2017) choose to estimate $\Omega$ by its empirical analogue using the *nuisance sample*, denoted as $\widehat{\Omega}_{\text{nuis}} = \widehat{\Sigma}_{\text{nuis}}^{-1}$. The resulting estimated kernel is denoted as $\widehat{K}_k^{\text{nuis}}(x, x')$, similar

[8] The difference in the signs in $\widehat{\mathbb{IF}}_{2,2,k}$ between here and Robins et al. (2016) is non-essential.

to $\widehat{K}_k$ defined in Section 1.2. Then the empirical SOIF (eSOIF) estimator $\widehat{\mathbb{IF}}_{2,2,k}(\hat{\Omega}_{\text{nuis}})$ of $\mathsf{bias}_k(\widehat{\psi}_1)$ is

$$\widehat{\mathbb{IF}}_{2,2,k}(\hat{\Omega}_{\text{nuis}}) \equiv \mathbb{U}_{n,2}\left[\mathsf{IF}_{2,2,k;1,2}(\hat{\Omega}_{\text{nuis}})\right],$$

which, unlike $\widehat{\mathbb{IF}}_{2,2,k}$, incurs a kernel estimation bias

$$\begin{aligned}\mathbb{E}[\widehat{\mathbb{IF}}_{2,2,k}(\hat{\Omega}_{\text{nuis}}) - \widehat{\mathbb{IF}}_{2,2,k}] &= \mathbb{E}[(A_1\widehat{a}(X_1)-1)\bar{\mathsf{z}}_k(X_1)^\top](\hat{\Omega}_{\text{nuis}} - \mathbb{I})\mathbb{E}[\bar{\mathsf{z}}_k(X_2)A_2(Y_2 - \widehat{b}(X_2))] \\ &= \mathbb{E}[(A_1\widehat{a}(X_1)-1)(\widehat{K}_k^{\text{nuis}}(X_1,X_2) - K_k(X_1,X_2))A_2(Y_2-\widehat{b}(X_2))],\end{aligned}$$

shown to be of order at most $\sqrt{k\log k/n}$ in Liu et al. (2017). To further reduce the kernel estimation bias, one can consider the following $m$-th order eHOIF estimator, which is an $m$-th order $U$-statistic:

$$\begin{aligned}\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\hat{\Omega}_{\text{nuis}}) &\coloneqq \sum_{j=2}^{m}\widehat{\mathbb{IF}}_{j,j,k}(\hat{\Omega}_{\text{nuis}}) \\ \text{where } \widehat{\mathbb{IF}}_{j,j,k}(\hat{\Omega}_{\text{nuis}}) &\coloneqq \mathbb{U}_{n,j}\left[\hat{\mathsf{IF}}_{j,j,k;1,\cdots,j}(\hat{\Omega}_{\text{nuis}})\right]\end{aligned}$$

and

$$\hat{\mathsf{IF}}_{j,j,k;1,\cdots,j}(\hat{\Omega}_{\text{nuis}}) = (-1)^j(A_1\widehat{a}(X_1)-1)\bar{\mathsf{z}}_k(X_1)^\top\hat{\Omega}_{\text{nuis}}\prod_{s=3}^{j}\left\{(Q_s - \hat{\Sigma}_{\text{nuis}})\hat{\Omega}_{\text{nuis}}\right\}\bar{\mathsf{z}}_k(X_2)A_2(Y_2-\widehat{b}(X_2)).$$

Liu et al. (2017) showed that the kernel estimation bias of $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\hat{\Omega}_{\text{nuis}})$ is of order at most $(k\log k/n)^{m/2}$ and variance of order at most $1/n \vee k/n^2$. Hence by taking $m \asymp \sqrt{\log n}$ and $k \asymp n/\log^c n$ for some absolute constant $c > 0$, we could estimate $\mathsf{bias}_k(\widehat{\psi}_1)$ with essentially no bias without inflating the order of the variance of $\widehat{\psi}_1$. Furthermore, under Hölder nuisance models on $\mathcal{A}\times\mathcal{B}$, Liu et al. (2017) demonstrate that the sHOIF estimator $\widehat{\psi}_1 + \widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\hat{\Omega}_{\text{nuis}})$, with said choices of $m$ and $k$, is $\sqrt{n}$-consistent in (4) and semiparametric efficient in the interior of (4) under some additional mild assumptions. In this paper, the sHOIF estimators to be introduced in Section 3 simply replace $\hat{\Sigma}_{\text{nuis}}$ and $\hat{\Omega}_{\text{nuis}}$ in the eHOIF estimators by $\hat{\Sigma}$ and $\hat{\Omega}$, the empirical analogues of $\Sigma$ and $\Omega$ computed from the estimation sample. One can easily see that, due to the correlation between $\hat{\Omega}$ and the estimation sample, the analysis of the statistical properties of sHOIF estimators becomes significantly more challenging.

## 1.4 Plan

The rest of the paper is organized as follows. Section 2 defines the stable Second-Order IF (sSOIF) estimators and studies their statistical and numerical properties as a warm-up. Section 3 presents the full version of sHOIF estimators, together with their statistical, numerical, and computational properties. We then apply sHOIF estimators and their statistical properties to two concrete problems Section 4: one is to show that sHOIF estimators for $\psi(\theta)$ achieve semiparametric efficiency under the minimal conditions within the classical Hölder nuisance models; the other is to use sHOIF estimators to test if the nominal $(1-\alpha)\times 100\%$ Wald confidence interval centered at the first-order DML estimator has the claimed coverage, a novel assumption-lean statistical procedure recently proposed in Liu et al. (2020a), and further developed in Liu et al. (2021b). To demonstrate the generality of sHOIF estimators, Section 5 extends results heretofore in several directions. Finally, Section 6 concludes the paper and discusses several open problems and possible future directions. Appendix contains technical details that provide insights on the proof strategy. The remaining technical details are deferred to Supplementary Materials (Li and Liu, 2023).

# 2 Assumptions and warm-up: Stable second-order influence function estimators

In this section, we disclose the main assumptions, accompanied with an illustration of the main results using the stable second-order influence function (sSOIF) estimator $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ as a warm-up of what follows.

The assumptions below are imposed throughout the paper unless stated otherwise.

**Assumption 1** (Conditions on initial first-step nuisance parameter estimates.)**.** Nuisance parameter estimators $\widehat{a}$ and $\widehat{b}$ are attained from a separate independent frozen nuisance sample. For simplicity, we assume this sample to also have size $n$. $\widehat{a}$ and $\widehat{b}$ further satisfy the following properties until otherwise noticed:

(i) $\|\widehat{a}-a\|_2 = o(1)$ and $\|\widehat{b}-b\|_2 = o(1)$, i.e. both nuisance parameter estimators are $L_2$-consistent;

(ii) $\|a\|_\infty$, $\|b\|_\infty$, $\|\widehat{a}\|_\infty$ and $\|\widehat{b}\|_\infty$ are bounded by some absolute constant $B > 0$.

(iii) In the case of $\psi(\theta) = \mathbb{E}[Y(1)]$ under strong ignorability, we additionally need $1/a$ and $1/\widehat{a}$ to be bounded between $(c, 1-c)$ for some absolute constant $0 < c < 0.5$.

**Assumption 2** (Conditions on $\bar{\mathsf{z}}_k$ related quantities.)**.** The following are assumed on the basis functions $\bar{\mathsf{z}}_k$ and the corresponding (inverse) Gram matrices $\Sigma, \widehat{\Sigma}, \Omega, \widehat{\Omega}$ and projection kernels $K_k$ and $\widehat{K}_k$:

(i) There exists an absolute constant $B > 0$ such that $\sup_{x\in\mathcal{X}} K_k(x,x) \leq Bk$ and $\sup_{x\in\mathcal{X}} \widehat{K}_k(x,x) \leq Bk$;

(ii) Both $\Sigma$ and $\widehat{\Sigma}$ have bounded spectra;

(iii) The projection kernel satisfies the following $L_\infty$-stability condition: for any measurable function $h : \mathcal{X} \to \mathbb{R}$,

$$\|\Pi\left[h|\bar{\mathsf{z}}_k\right](\cdot)\|_\infty \lesssim \|h\|_\infty. \tag{7}$$

**Remark 1** (Comments on Assumptions 1 and 2)**.**

(i) Given Assumption 2(ii), there is no loss of generality by assuming $\Sigma \equiv \Omega \equiv \mathbb{I}$, the identity matrix of the same size as $\Sigma$ or $\Omega$. We make such a simplification throughout the paper, unless stated otherwise.

(ii) The assumptions on the nuisance parameters and their estimators in Assumption 1 are quite mild. In particular, we do not assume $\widehat{a}$, $\widehat{b}$ converge to $a$, $b$ at any algebraic rate in $L_2$-norm. In fact, if content with $\sqrt{n}$-consistency instead of semiparametric efficiency, $\|\widehat{a} - a\|_2 = o(1)$ and $\|\widehat{b} - b\|_2 = o(1)$ can be even relaxed to $\|\widehat{a} - a\|_2 = O(1)$ and $\|\widehat{b} - b\|_2 = O(1)$; see Liu et al. (2017).

(iii) Assumption 2 on the dictionary $\bar{\mathsf{z}}_k$ also appeared in Robins et al. (2017); Liu et al. (2017, 2020a, 2021b); also see comments in Liu et al. (2020b). The $L_\infty$-stability condition (iii) have been established for Cohen-Daubechies-Vial wavelets, B-splines, and local polynomial partition series (Belloni et al., 2015). It is possible to relax such a condition to a high-probability version, which we decide not to further pursue in this paper.

■

The following result on the sSOIF estimator $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ is a special case of Theorem 1 to be revealed in Section 3.

**Proposition 1** (Bias and variance bounds of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$.)**.** *Under Assumptions 1 – 2, with $k = o(n)$, one has the following:*

(i) *The kernel estimation bias of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ satisfies*

$$
\begin{aligned}
&\mathsf{kern\text{-}bias}_{2,k}(\widehat{\psi}_1) \coloneqq \mathbb{E}\left[\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})\right] - \mathsf{bias}_{\theta,k}(\widehat{\psi}_1) \equiv \mathbb{E}\left[\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega}) - \widehat{\mathbb{IF}}_{2,2,k}\right] \\
&\lesssim \frac{k}{n}\left\{\left\|\frac{\widehat{a}-1}{a}\right\|_2 \|\widehat{b}-b\|_2 + \left\|\frac{\widehat{a}-a}{a}\right\|_2 \|\widehat{b}-b\|_2 + \left(\left\|\frac{\widehat{a}-1}{a}\right\|_2 \left\|\widehat{b}-b\right\|_\infty \wedge \left\|\frac{\widehat{a}-1}{a}\right\|_\infty \left\|\widehat{b}-b\right\|_2\right)\right\}.
\end{aligned} \tag{8}
$$

(ii) *The variance of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ satisfies*

$$
\mathsf{var}\left[\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \left(\left\|\frac{\widehat{a}-1}{a}\right\|_2 \left\|\widehat{b}-b\right\|_\infty \wedge \left\|\frac{\widehat{a}-1}{a}\right\|_\infty \left\|\widehat{b}-b\right\|_2\right)\right\}. \tag{9}
$$

**Remark 2.** The dependence on the condition number $k/n$ in the kernel estimation bias upper bound of the eSOIF estimator $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega}_{\text{nuis}})$ in Liu et al. (2017) ($\sqrt{k \log k/n}$) is worse than that of the sSOIF estimator $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ reported here ($k/n$). ■

## 2.1 Proof sketch of Proposition 1

### 2.1.1 Kernel estimation bias bound

$\mathsf{kern\text{-}bias}_{2,k}(\widehat{\psi}_1)$ can be controlled by repeatedly using the matrix identity $(A-B)^{-1} - A^{-1} = -A^{-1}B(A-B)^{-1}$ with $A \equiv \widehat{\Sigma}^{\dagger} \coloneqq n^{-1}\sum_{i=3}^{n} Q_i$ and $B = n^{-1}Q_{1,2}$:

$$
\begin{aligned}
&\mathsf{kern\text{-}bias}_{2,k}(\widehat{\psi}_1) \\
&= \mathbb{E}\left[(A_1\widehat{a}(X_1)-1)\bar{\mathsf{z}}_k(X_1)^{\top}(\widehat{\Omega}^{-1}-\mathbb{I})\bar{\mathsf{z}}_k(X_2)A_2(Y_2-\widehat{b}(X_2))\right] \\
&= \sum_{j=1}^{J-1}\mathbb{E}\left[(A_1\widehat{a}(X_1)-1)\bar{\mathsf{z}}_k(X_1)^{\top}\left(\mathbb{I}-\widehat{\Sigma}^{\dagger}-\frac{Q_{1,2}}{n}\right)^{j}\bar{\mathsf{z}}_k(X_2)A_2(Y_2-\widehat{b}(X_2))\right] \\
&\quad + \mathbb{E}\left[(A_1\widehat{a}(X_1)-1)\bar{\mathsf{z}}_k(X_1)^{\top}\left(\mathbb{I}-\widehat{\Sigma}^{\dagger}-\frac{Q_{1,2}}{n}\right)^{J}\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)A_2(Y_2-\widehat{b}(X_2))\right].
\end{aligned}
$$

By choosing $J \asymp \log n$, the second term of the above display can be shown to be $o(n^{-1/2})$.

For the first term, we only look at $j = 1, 2$ in the main text and the remaining analysis is a special case of the proof of Theorem 1 in Appendix C.

For $j = 1$, we have

$$
\begin{aligned}
&\mathbb{E}\left[(A_1\widehat{a}(X_1)-1)\bar{\mathsf{z}}_k(X_1)^{\top}\left(\mathbb{I}-\widehat{\Sigma}^{\dagger}-\frac{Q_{1,2}}{n}\right)\bar{\mathsf{z}}_k(X_2)A_2(Y_2-\widehat{b}(X_2))\right] \\
&= \frac{2}{n}\mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)}-1\right)\bar{\mathsf{z}}_k(X_1)^{\top}\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right] \\
&\quad - \frac{1}{n}\mathbb{E}\left[(A_1\widehat{a}(X_1)-1)\,\bar{\mathsf{z}}_k(X_1)^{\top}Q_{1,2}\bar{\mathsf{z}}_k(X_2)A_2\left(Y_2-\widehat{b}(X_2)\right)\right] \qquad (10) \\
&= O\left(\frac{1}{n}\right) - \frac{1}{n}\mathbb{E}\left[\left(\frac{\widehat{a}(X_1)-1}{a(X_1)}\right)\bar{\mathsf{z}}_k(X_1)^{\top}\bar{\mathsf{z}}_k(X_1)\bar{\mathsf{z}}_k(X_1)^{\top}\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right] \\
&\quad - \frac{1}{n}\mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)}-1\right)\bar{\mathsf{z}}_k(X_1)^{\top}\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)\bar{\mathsf{z}}_k(X_2)^{\top}\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right] \\
&\lesssim \frac{1}{n} + \frac{k}{n}\left\|\frac{\widehat{a}-1}{a}\right\|_2\|b-\widehat{b}\|_2 + \frac{k}{n}\left\|\frac{\widehat{a}}{a}-1\right\|_2\|b-\widehat{b}\|_2
\end{aligned}
$$

where the last line follows from triangle inequality, Cauchy-Schwarz inequality and Assumptions 1, 2(i) and 2(ii).

For $j = 2$, we have

$$
\begin{aligned}
&\mathbb{E}\left[(A_1\widehat{a}(X_1)-1)\bar{\mathsf{z}}_k(X_1)^{\top}\left(\mathbb{I}-\widehat{\Sigma}^{\dagger}-\frac{Q_{1,2}}{n}\right)^{2}\bar{\mathsf{z}}_k(X_2)A_2(Y_2-\widehat{b}(X_2))\right] \\
&= \mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)}-1\right)\bar{\mathsf{z}}_k(X_1)^{\top}\right]\mathbb{E}\left[\left(\mathbb{I}-\widehat{\Sigma}^{\dagger}\right)^{2}\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right] \\
&\quad - \frac{2}{n^2}\mathbb{E}\left[(A_1\widehat{a}(X_1)-1)\,\bar{\mathsf{z}}_k(X_1)^{\top}Q_{1,2}\bar{\mathsf{z}}_k(X_2)A_2\left(Y_2-\widehat{b}(X_2)\right)\right]
\end{aligned}
$$

$$
+ \frac{1}{n^2}\mathbb{E}\left[\left(A_1\widehat{a}(X_1)-1\right)\bar{\mathsf{z}}_k(X_1)^\top Q_{1,2}^2\bar{\mathsf{z}}_k(X_2)A_2\left(Y_2-\widehat{b}(X_2)\right)\right] \tag{11}
$$
$$
=: \text{(I)} + \text{(II)} + \text{(III)}.
$$

Since (II) is dominated by the term for $j = 1$, we only need to further analyze (I) and (III). (III) can be bounded by

$$
\text{(III)} \lesssim \left(\frac{k}{n}\right)^2 \left\{\left\|\frac{\widehat{a}-1}{a}\right\|_2 \|b-\widehat{b}\|_2 + \left\|\frac{\widehat{a}}{a}-1\right\|_2 \|b-\widehat{b}\|_2 + \left(\left\|\frac{\widehat{a}-1}{a}\right\|_\infty \|b-\widehat{b}\|_2\right) \wedge \left(\left\|\frac{\widehat{a}-1}{a}\right\|_2 \|b-\widehat{b}\|_\infty\right)\right\} \tag{12}
$$

where the first two terms are due to the first three terms in the (non-commutative) expansion of

$$
Q_{1,2}^2 = Q_1^2 + Q_2^2 + Q_1 Q_2 + Q_2 Q_1, \tag{13}
$$

and the third term comes from the last term in the above expansion. The appearance of the estimation error in $L_\infty$-norm is due to the opposite order of sample points indexed by 1 and 2 between the "meat" $Q_2 Q_1$ and the "bread slices" of the "sandwich" structure $\bar{\mathsf{z}}_k(X_1)^\top[\cdots]\bar{\mathsf{z}}_k(X_2)$. We formally prove (12) in Appendix **??**.

For (I), we need to expand $(\mathbb{I}-\widehat{\Sigma}^\dagger)^2$.

$$
\begin{aligned}
\text{(I)} &= \mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)}-1\right)\bar{\mathsf{z}}_k(X_1)^\top\right]\mathbb{E}\left[\mathbb{I}-2\widehat{\Sigma}^\dagger+\widehat{\Sigma}^{\dagger 2}\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right] \\
&= \left(\frac{4}{n}-1\right)\mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)}-1\right)\bar{\mathsf{z}}_k(X_1)^\top\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right] \\
&\quad + \frac{(n-2)(n-3)}{n^2}\mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)}-1\right)\bar{\mathsf{z}}_k(X_1)^\top\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right] \\
&\quad + \frac{n-2}{n^2}\mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)}-1\right)\bar{\mathsf{z}}_k(X_1)^\top\right]\mathbb{E}\left[Q_3^2\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right] \\
&= \left(\frac{6}{n^2}-\frac{1}{n}\right)\mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)}-1\right)\bar{\mathsf{z}}_k(X_1)^\top\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right] \\
&\quad + \left(\frac{1}{n}-\frac{2}{n^2}\right)\mathbb{E}\left[\left(\frac{\widehat{a}(X_1)}{a(X_1)}-1\right)\bar{\mathsf{z}}_k(X_1)^\top\right]\mathbb{E}\left[Q_3^2\right]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)(b(X_2)-\widehat{b}(X_2))\right].
\end{aligned} \tag{14}
$$

It is straightforward to see the first term in the last equality of the above display is dominated by the term for $j = 1$, whereas the second term can be shown to be bounded by

$$
\frac{k}{n}\left\|\frac{\widehat{a}}{a}-1\right\|_2 \|b-\widehat{b}\|_2.
$$

Taken together, the terms for $j = 1$ and $j = 2$ give the desired bound for $\mathsf{kern\text{-}bias}_{2,k}(\widehat{\psi}_1)$ in (8). It remains to prove the terms for $j \geq 3$ are of smaller order, which is deferred to Appendix **??**. For $j \geq 3$, the corresponding term is of order

$$
\left(\frac{k}{n}\right)^{j-1}\left\|\frac{\widehat{a}}{a}-1\right\|_2 \|b-\widehat{b}\|_2.
$$

**Remark 3.** Now is a perfect time to compare how the analysis of the kernel estimation bias of sSOIF differs from that of eSOIF of Liu et al. (2017). The only difference between the eSOIF and sSOIF estimators are the samples used to estimate $\Omega$. Using the nuisance sample instead, the (conditional) kernel estimation bias of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega}_{\text{nuis}})$ conditioning on the nuisance sample data is

$$\begin{aligned}&\mathbb{E}\left[\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega}_{\text{nuis}}) - \widehat{\mathbb{IF}}_{2,2,k}\right]\\ = {}&\mathbb{E}\left[(A_1\widehat{a}(X_1) - 1)\bar{\mathsf{z}}_k(X_1)^\top\right]\left(\widehat{\Omega}_{\text{nuis}} - \mathbb{I}\right)\mathbb{E}\left[A_2\bar{\mathsf{z}}_k(X_2)(Y_2 - \widehat{b}(X_2))\right].\end{aligned}$$

From this, we can conclude

$$\left|\mathbb{E}\left[\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega}_{\text{nuis}}) - \widehat{\mathbb{IF}}_{2,2,k}\right]\right| \lesssim \left(\frac{k\log k}{n}\right)^{1/2}\left\|\frac{\widehat{a}}{a} - 1\right\|_2 \|\widehat{b} - b\|_2$$

by using matrix Bernstein or Khintchine inequality (Rudelson, 1999; Bandeira et al., 2021); also see Couillet and Liao (2022). Liu et al. (2017) further show that

$$\left|\mathbb{E}\left[\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega}_{\text{nuis}}) - \widehat{\mathbb{IF}}_{2,2,k}\right]\right| \lesssim \left(\frac{k\log k}{n}\right)^{m/2}\left\|\frac{\widehat{a}}{a} - 1\right\|_2 \|\widehat{b} - b\|_2.$$

However, as pointed out in Section 1.1, the finite-sample performance of eHOIF estimators is not well-reflected by these upper bounds, prompting the need of developing sHOIF estimators. ■

### 2.1.2 Variance bound

The variance bound is technically involved. The missing steps can be found in Appendix A. The key step is to show

$$\mathbb{E}\left[A_1\bar{\mathsf{z}}_k(X_1)^\top\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)A_3\bar{\mathsf{z}}_k(X_3)^\top\widehat{\Omega}\bar{\mathsf{z}}_k(X_4)\right] - \left(\mathbb{E}\left[A_1\bar{\mathsf{z}}_k(X_1)^\top\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)\right]\right)^2 = O\left(\frac{1}{n}\right). \tag{15}$$

To prove (15), it is sufficient to exhibit

$$\begin{aligned}&\mathbb{E}\left[A_1\bar{\mathsf{z}}_k(X_1)^\top\bar{\mathsf{z}}_k(X_2)A_3\bar{\mathsf{z}}_k(X_3)^\top\left(\widehat{\Omega} - \mathbb{I}\right)\bar{\mathsf{z}}_k(X_4)\right]\\ &- \mathbb{E}\left[A_1\bar{\mathsf{z}}_k(X_1)^\top\bar{\mathsf{z}}_k(X_2)\right]\mathbb{E}\left[A_1\bar{\mathsf{z}}_k(X_1)^\top\left(\widehat{\Omega} - \mathbb{I}\right)\bar{\mathsf{z}}_k(X_2)\right] = O\left(\frac{1}{n}\right)\end{aligned} \tag{16}$$

and

$$\begin{aligned}&\mathbb{E}\left[A_1\bar{\mathsf{z}}_k(X_1)^\top\left(\widehat{\Omega} - \mathbb{I}\right)\bar{\mathsf{z}}_k(X_2)A_3\bar{\mathsf{z}}_k(X_3)^\top\left(\widehat{\Omega} - \mathbb{I}\right)\bar{\mathsf{z}}_k(X_4)\right]\\ &- \mathbb{E}\left[A_1\bar{\mathsf{z}}_k(X_1)^\top\left(\widehat{\Omega} - \mathbb{I}\right)\bar{\mathsf{z}}_k(X_2)\right]\mathbb{E}\left[A_1\bar{\mathsf{z}}_k(X_1)^\top\left(\widehat{\Omega} - \mathbb{I}\right)\bar{\mathsf{z}}_k(X_2)\right] = O\left(\frac{1}{n}\right).\end{aligned} \tag{17}$$

Recall $\widehat{\Omega} = \widehat{\Sigma}^{-1}$ and $\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n} Q_i$. We introduce independent "ghost copies" $Q_1', Q_2'$ of $Q_1, Q_2$ and denote $\widehat{\Omega}' = (\widehat{\Sigma}')^{-1}$ and $\widehat{\Sigma}' = n^{-1}\sum_{i=3}^{n} Q_i + n^{-1}(Q_1' + Q_2')$. Then (16) is equivalent to

$$\mathbb{E}\left[A_1 \bar{\mathsf{z}}_k(X_1)^\top \bar{\mathsf{z}}_k(X_2) A_3 \bar{\mathsf{z}}_k(X_3)^\top \left(\widehat{\Omega} - \widehat{\Omega}'\right) \bar{\mathsf{z}}_k(X_4)\right] = O\left(\frac{1}{n}\right). \tag{18}$$

Let $\bar{\Omega} \equiv \bar{\Sigma}$ and $\bar{\Sigma} \equiv n^{-1}\sum_{i=5}^{n} Q_i$. Repeating the matrix identity $(A+B)^{-1} - A^{-1} = -A^{-1}B(A+B)^{-1}$ on (18) by setting $A = \bar{\Sigma}$ and $B = n^{-1}(Q_{1,2} + Q_{3,4})$ or $B = n^{-1}(Q_{1,2}' + Q_{3,4})$, we have

$$\begin{aligned}
&\mathbb{E}\left[A_1 \bar{\mathsf{z}}_k(X_1)^\top \bar{\mathsf{z}}_k(X_2) A_3 \bar{\mathsf{z}}_k(X_3)^\top \left(\widehat{\Omega} - \widehat{\Omega}'\right) \bar{\mathsf{z}}_k(X_4)\right] \\
= &\sum_{j=1}^{J-1} (-1)^j \mathbb{E}\left[A_1 \bar{\mathsf{z}}_k(X_1)^\top \bar{\mathsf{z}}_k(X_2) A_3 \bar{\mathsf{z}}_k(X_3)^\top \left\{\left(\bar{\Omega}\frac{Q_{1,2}+Q_{3,4}}{n}\right)^j - \left(\bar{\Omega}\frac{Q_{1,2}'+Q_{3,4}}{n}\right)^j\right\} \bar{\Omega}\bar{\mathsf{z}}_k(X_4)\right] \\
&+ (-1)^J \mathbb{E}\left[A_1 \bar{\mathsf{z}}_k(X_1)^\top \bar{\mathsf{z}}_k(X_2) A_3 \bar{\mathsf{z}}_k(X_3)^\top \left\{\left(\bar{\Omega}\frac{Q_{1,2}+Q_{3,4}}{n}\right)^J - \left(\bar{\Omega}\frac{Q_{1,2}'+Q_{3,4}}{n}\right)^J\right\} \widehat{\Omega}\bar{\mathsf{z}}_k(X_4)\right].
\end{aligned}$$

Let $J \asymp \log n$. The second term of the above display can be shown to be $o(1/n)$. Proceeding to the first term, it is easy to see from Assumption 2(iii) that the term corresponding to $j = 1$:

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}\left[A_1 \bar{\mathsf{z}}_k(X_1)^\top \bar{\mathsf{z}}_k(X_2) A_3 \bar{\mathsf{z}}_k(X_3)^\top \bar{\Omega}(Q_{1,2} - Q_{1,2}')\bar{\Omega}\bar{\mathsf{z}}_k(X_4)\right] \\
= &\frac{1}{n}\left(\begin{array}{c}
\mathbb{E}\left\{A_1 \bar{\mathsf{z}}_k(X_1)^\top \mathbb{E}[\bar{\mathsf{z}}_k(X_2)] \mathbb{E}[A_3 \bar{\mathsf{z}}_k(X_3)]^\top \bar{\Omega}\bar{\mathsf{z}}_k(X_1)\bar{\mathsf{z}}_k(X_1)^\top \bar{\Omega}\mathbb{E}[\bar{\mathsf{z}}_k(X_4)]\right\} \\
+ \mathbb{E}\left\{\mathbb{E}[A_1 \bar{\mathsf{z}}_k(X_1)]^\top \bar{\mathsf{z}}_k(X_2) \mathbb{E}[A_3 \bar{\mathsf{z}}_k(X_3)]^\top \bar{\Omega} A_2 \bar{\mathsf{z}}_k(X_2)\bar{\mathsf{z}}_k(X_2)^\top \bar{\Omega}\mathbb{E}[\bar{\mathsf{z}}_k(X_4)]\right\} \\
- \mathbb{E}\left\{\mathbb{E}[A_1 \bar{\mathsf{z}}_k(X_1)]^\top \bar{\mathsf{z}}_k(X_2) \mathbb{E}[A_3 \bar{\mathsf{z}}_k(X_3)]^\top \bar{\Omega}\mathbb{E}[Q_{1,2}']\bar{\Omega}\mathbb{E}[\bar{\mathsf{z}}_k(X_4)]\right\}
\end{array}\right) = O\left(\frac{1}{n}\right).
\end{aligned}$$

Similarly, the terms corresponding to $j \geq 2$ can be shown to be $O\left(\frac{1}{n}\left(\frac{2k}{n}\right)^{j-1}\right)$, a consequence of Lemma 1 below. Note that the extra factor 2 appears because there are $O(2^j)$ terms in total by expanding out $\left(\frac{Q_{1,2}+Q_{3,4}}{n}\right)^j$ and $\left(\frac{Q_{1,2}'+Q_{3,4}}{n}\right)^j$.

**Lemma 1.** *Given a positive integer $j$. Given any pair of integers $j_1 \geq 0, j_2 > 0$ such that $j_1 + j_2 = j$, for any subset $\mathfrak{j} \subseteq \{0,1\}^j$ of the $j$-dimensional Boolean hypercube with $\|\mathfrak{j}\|_1 = j_1$, we have*

$$\mathbb{E}\left[\bar{\mathsf{z}}_k(X_1)^\top \bar{\mathsf{z}}_k(X_2) \bar{\mathsf{z}}_k(X_3)^\top \left(\prod_{\ell=1}^{j} Q_{1,2}^{\mathfrak{j}_\ell} Q_{3,4}^{(1-\mathfrak{j}_\ell)}\right) \bar{\mathsf{z}}_k(X_4)\right] \lesssim k^{j-1}. \tag{19}$$

*However, if $j_1 = 0$ and $j_2 = j$, we have*

$$\mathbb{E}\left[\bar{\mathsf{z}}_k(X_1)^\top \bar{\mathsf{z}}_k(X_2) \bar{\mathsf{z}}_k(X_3)^\top Q_{3,4}^j \bar{\mathsf{z}}_k(X_4)\right] \lesssim k^j. \tag{20}$$

The proof of Lemma 1 can be found in Appendix A.1. Finally, we defer the proof of (17) to Section S3.3, which can be proved in a similar fashion.

## 2.2 Numerical stability and time complexity of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$

As discussed in Section 1, the key motivation for proposing sHOIF estimators is the numerical instability observed for eHOIF estimators. As a warm-up, we rigorously prove the numerical stability and calculate the time complexity of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ in this section. The reason why $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ can be numerically unstable is that when $k$ is near $n$, it is highly likely $\lambda_{\min}(\widehat{\Sigma}) \approx 0$ and hence $\lambda_{\max}(\widehat{\Omega})$ is close to infinity. But:

**Proposition 2.** *$\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ does not depend on the eigenvalues of $\widehat{\Omega}$.*

For ease of exposition, in what follows we let

- $\bar{\mathsf{Z}}_{n,k} \coloneqq (\bar{\mathsf{z}}_k(X_1), \cdots, \bar{\mathsf{z}}_k(X_n))^\top$ as the $n \times k$-matrix of the dictionary vectors for all $n$ samples;
- $\bar{\mathsf{Z}}^A_{n,k} \coloneqq (A_1\bar{\mathsf{z}}_k(X_1), \cdots, A_n\bar{\mathsf{z}}_k(X_n))^\top$ as the $n \times k$-matrix of the $A$-weighted dictionary vectors for all $n$ samples;
- $\boldsymbol{\mathcal{E}}_{n,a}(\widehat{a}) \coloneqq (\mathcal{E}_a(\widehat{a}, O_1), \cdots, \mathcal{E}_a(\widehat{a}, O_n))^\top$ and $\boldsymbol{\mathcal{E}}_{n,b}(\widehat{b}) \coloneqq \left(\mathcal{E}_b(\widehat{b}, O_1), \cdots, \mathcal{E}_b(\widehat{b}, O_n)\right)^\top$.

*Proof.* We can rewrite $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ as

$$\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega}) = \frac{1}{n-1}\boldsymbol{\mathcal{E}}_{n,a}(\widehat{a})^\top \left[\mathbb{I} - \mathsf{Diag}\right] \left\{\bar{\mathsf{Z}}_{n,k}\left(\bar{\mathsf{Z}}^\top_{n,k}\bar{\mathsf{Z}}^A_{n,k}\right)^{-1}\bar{\mathsf{Z}}^{A\top}_{n,k}\right\}\boldsymbol{\mathcal{E}}_{n,b}(\widehat{b}). \tag{21}$$

Now apply Singular Value Decomposition (SVD) on the matrices $\bar{\mathsf{Z}}^A_{n,k}$:

$$\bar{\mathsf{Z}}^A_{n,k} = U^A \mathsf{Diag}(D^A) V^{A\top}.$$

Then

$$\bar{\mathsf{Z}}_{n,k}\left(\bar{\mathsf{Z}}^\top_{n,k}\bar{\mathsf{Z}}^A_{n,k}\right)^{-1}\bar{\mathsf{Z}}^{A\top}_{n,k} = \bar{\mathsf{Z}}^A_{n,k}\left(\bar{\mathsf{Z}}^{A\top}_{n,k}\bar{\mathsf{Z}}^A_{n,k}\right)^{-1}\bar{\mathsf{Z}}^{A\top}_{n,k} = U^A U^{A\top}.$$

So

$$(21) = \frac{1}{n-1}\boldsymbol{\mathcal{E}}_{n,a}(\widehat{a})^\top \left[\mathbb{I} - \mathsf{Diag}\right]\left\{U^A U^{A\top}\right\}\boldsymbol{\mathcal{E}}_{n,b}(\widehat{b}),$$

which is completely independent of the eigenvalues of $\widehat{\Omega}$ ($(D^A)^2$ up to constant). □

Hence it is not surprising that $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ is numerically stable even when $k \to n$.

**Remark 4.** In a sense, $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ can be viewed as a *self-normalized* version of $\widehat{\mathbb{IF}}_{2,2,k}$. It is generally expected that self-normalized statistics could have better statistical properties than the non-self-normalized ones (Peña et al., 2008). However, whether the perspective of self-normalization is useful for establishing statistical properties of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ is still unclear to us and is worth pursuing as a research problem. ■

Furthermore, not only does the alternative formula (21) of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ directly imply its numerical stability, but also it hints at the complexity of computing $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$. Barring the time complexity of SVD ($O(nk^2)$), the time complexity of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ scales with $kn$ at a linear instead of a quadratic rate. This can be seen from (21), in which only two vector-matrix products are involved, each taking $O(nk)$ operations. Thus we have

**Proposition 3.** *The time complexity of computing $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ is $O(nk^2)$, dominated by that of SVD.*

**Remark 5.** Another alternative way of arriving at the above conclusion is to observe that the $U$-statistic kernel of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$, denoted as $\widehat{\mathsf{IF}}_{2,2,k,\bar{i}_2}(\widehat{\Omega})$, is separable, in the following sense: there exists a pair (but not necessarily a unique pair) of functions $h_1, h_2$ such that

$$\widehat{\mathsf{IF}}_{2,2,k,\bar{i}_2}(\widehat{\Omega}) \equiv h_1(O_{i_1}, \widehat{\Omega}) \cdot h_2(O_{i_2}, \widehat{\Omega}).$$

■

# 3 The hierarchy of sHOIF estimators

As indicated in Section 1.3, the $m$-th order sHOIF estimator $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})$ takes the same form as the $m$-th order eHOIF estimator $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega}_{\text{nuis}})$, with the sole difference that $\widehat{\Omega}_{\text{nuis}}$ is replaced by $\widehat{\Omega}$. Formally, the $m$-th order sHOIF and the corresponding $m$-th order estimator of $\psi(\theta)$ read as follows:

$$\begin{aligned} \widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega}) &\coloneqq (-1)^m \mathbb{U}_{n,m}\left[\mathcal{E}_a(\widehat{a}; O_1)\bar{\mathsf{z}}_k(X_1)^\top \widehat{\Omega} \prod_{s=3}^{m}\left\{\left(Q_s - \widehat{\Sigma}\right)\widehat{\Omega}\right\}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b}; O_2)\right], \\ \widehat{\psi}_{m,k}(\widehat{\Omega}) &\coloneqq \widehat{\psi}_1 + \sum_{j=2}^{m}\widehat{\mathbb{IF}}_{j,j,k}(\widehat{\Omega}) \equiv \widehat{\psi}_1 + \widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega}). \end{aligned} \tag{22}$$

**Remark 6.** The above sHOIF statistics are the same as the eHOIF statistics except that $\widehat{\Omega}$ is constructed from the estimation sample instead of the nuisance sample. ■

In this section, we first explain heuristically why $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})$ is enough to correct for the kernel estimation bias (see Section 3.1), after which the statistical, numerical and computational properties of $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})$ are stated formally.

## 3.1 Heuristic explanation

In what follows we explain heuristically why the kernel estimation bias $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ can be further corrected by adding:

$$\widehat{\mathbb{IF}}_{3,3,k}(\widehat{\Omega}) \coloneqq \frac{n-2}{n}\mathbb{U}_{n,3}\left[- \mathcal{E}_a(\widehat{a}; O_1)\bar{\mathsf{z}}_k(X_1)^\top \widehat{\Omega}\left(Q_3 - \widehat{\Sigma}\right)\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b}; O_2)\right]$$

and

$$\widehat{\mathbb{IF}}_{4,4,k}(\widehat{\Omega}) := \frac{n-3}{n}\mathbb{U}_{n,4}\left[(A_1\widehat{a}(X_1)-1)\bar{\mathsf{z}}_k(X_1)^\top\widehat{\Omega}\left(Q_3-\widehat{\Sigma}\right)\widehat{\Omega}\left(Q_4-\widehat{\Sigma}\right)\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)A_2(Y_2-\widehat{b}(X_2))\right].$$

For short, we define $\widehat{\mathbb{IF}}_{(2,2)\to(3,3),k}(\widehat{\Omega}) := \widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega}) + \widehat{\mathbb{IF}}_{3,3,k}(\widehat{\Omega})$ and $\widehat{\mathbb{IF}}_{(2,2)\to(4,4),k}(\widehat{\Omega}) := \sum_{j=2}^{4}\widehat{\mathbb{IF}}_{j,j,k}(\widehat{\Omega})$. Also note that the majority of this section is written for mathematical rigor.

Simple algebra gives

$$\begin{aligned}
&\widehat{\mathbb{IF}}_{3,3,k}(\widehat{\Omega}) \\
&\equiv -\frac{n-2}{n}\frac{(n-3)!}{n!}\sum_{1\le i_1\neq i_2\neq i_3\le n}\mathcal{E}_a(\widehat{a};O_{i_1})\bar{\mathsf{z}}_k(X_{i_1})^\top\widehat{\Omega}\left(Q_{i_3}-\widehat{\Sigma}\right)\widehat{\Omega}\bar{\mathsf{z}}_k(X_{i_2})\mathcal{E}_b(\widehat{b};O_{i_2}) \\
&= \frac{1}{n}\mathbb{U}_{n,2}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\widehat{\Omega}Q_{1,2}\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] - \frac{2}{n}\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega}) \\
&\approx \frac{1}{n}\mathbb{U}_{n,2}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\widehat{\Omega}Q_{1,2}\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] =: \widetilde{\widehat{\mathbb{IF}}}_{3,3,k}(\widehat{\Omega})
\end{aligned}$$

and

$$\begin{aligned}
&\widehat{\mathbb{IF}}_{4,4,k}(\widehat{\Omega}) \\
&\equiv \frac{n-3}{n}\frac{(n-4)!}{n!}\sum_{1\le i_1\neq i_2\neq i_3\neq i_4\le n}\mathcal{E}_a(\widehat{a};O_{i_1})\bar{\mathsf{z}}_k(X_{i_1})^\top\widehat{\Omega}\prod_{s=3}^{4}\left[\left(Q_{i_s}-\widehat{\Sigma}\right)\widehat{\Omega}\right]\bar{\mathsf{z}}_k(X_{i_2})\mathcal{E}_b(\widehat{b};O_{i_2}) \\
&= \frac{1}{n(n-2)}\mathbb{U}_{n,2}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\widehat{\Omega}Q_{1,2}\widehat{\Omega}Q_{1,2}\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] \\
&\quad - \frac{1}{n}\mathbb{U}_{n,3}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\widehat{\Omega}Q_3\widehat{\Omega}Q_3\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] \\
&\quad - \frac{1}{n-2}\left(6\widehat{\mathbb{IF}}_{3,3,k}(\widehat{\Omega}) - \left(1-\frac{6}{n}\right)\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})\right) \\
&\approx \frac{1}{n^2}\mathbb{U}_{n,2}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\widehat{\Omega}Q_{1,2}\widehat{\Omega}Q_{1,2}\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] \\
&\quad - \frac{1}{n}\mathbb{U}_{n,3}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\widehat{\Omega}Q_3\widehat{\Omega}Q_3\widehat{\Omega}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] \\
&=: \widetilde{\widehat{\mathbb{IF}}}_{4,4,k}(\widehat{\Omega}).
\end{aligned}$$

First, observe that the expectation of the oracle version of $\widetilde{\widehat{\mathbb{IF}}}_{3,3,k}(\widehat{\Omega})$

$$\mathbb{E}\left[\widetilde{\widehat{\mathbb{IF}}}_{3,3,k}\right] = \frac{1}{n}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top Q_{1,2}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right]$$

exactly cancels (10), the leading-order part of the kernel estimation bias of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ corresponding to $j = 1$.

Next, observe that the expectation of the oracle version of $\widetilde{\widehat{\mathbb{IF}}}_{4,4,k}(\widehat{\Omega})$ is

$$\mathbb{E}\left[\widetilde{\widehat{\mathbb{IF}}}_{4,4,k}\right] = \frac{1}{n^2}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top Q_{1,2}^2 \bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] - \frac{1}{n}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\right]\mathbb{E}[Q_3^2]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right]$$

which again cancels the kernel estimation bias of $\widehat{\mathbb{IF}}_{(2,2)\to(3,3),k}(\widehat{\Omega})$ truncated at $j=2$, dominated by

$$-\frac{1}{n^2}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top Q_{1,2}^2 \bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] + \frac{1}{n}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\right]\mathbb{E}[Q_3^2]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] \tag{23}$$

which can be derived from (11), (14) and the kernel estimation bias of $\widehat{\mathbb{IF}}_{3,3,k}(\widehat{\Omega})$ truncated at level $j=1$; see Appendix B for a more detailed calculation. Hence $\widehat{\mathbb{IF}}_{(3,3)\to(4,4),k}(\widehat{\Omega})$ further reduces the kernel estimation bias of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$.

### 3.2 Characterization of the bias and variance of the sHOIF estimators

We now state the main theoretical result of this paper.

**Theorem 1.** *Under Assumptions 1 – 2, with $k \lesssim \frac{n}{\log^2 n}$ and $m \gtrsim \sqrt{\log n}$, one has the following:*

(i) *The kernel estimation bias of $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$ satisfies*

$$\begin{aligned}&\mathsf{kern\text{-}bias}_{m,k}(\widehat{\psi}_1) := \mathbb{E}\left[\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})\right] - \mathsf{bias}_{\theta,k}(\widehat{\psi}_1) \equiv \mathbb{E}\left[\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega}) - \widehat{\mathbb{IF}}_{2,2,k}\right]\\ &\lesssim \left(\frac{km}{n}\right)^{\lceil\frac{\lceil\frac{m-1}{2}\rceil-1}{2}\rceil\vee 1}\left\{\begin{array}{c}\left\|\frac{\widehat{a}-1}{a}\right\|_2\|\widehat{b}-b\|_2 + \left\|\frac{\widehat{a}-a}{a}\right\|_2\|\widehat{b}-b\|_2\\ +\left(\left\|\frac{\widehat{a}-1}{a}\right\|_2\left\|\widehat{b}-b\right\|_\infty \wedge \left\|\frac{\widehat{a}-1}{a}\right\|_\infty\left\|\widehat{b}-b\right\|_2\right)\end{array}\right\}.\end{aligned} \tag{24}$$

(ii) *For $m \geq 2$, the variance of $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})$ satisfies*

$$\mathsf{var}\left[\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \left(\left\|\frac{\widehat{a}-1}{a}\right\|_2\left\|\widehat{b}-b\right\|_\infty \wedge \left\|\frac{\widehat{a}-1}{a}\right\|_2\left\|\widehat{b}-b\right\|_\infty\right)\right\}. \tag{25}$$

*And thus the variance of $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$ satisfies*

$$\mathsf{var}\left[\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \left(\left\|\frac{\widehat{a}-1}{a}\right\|_2\left\|\widehat{b}-b\right\|_\infty \wedge \left\|\frac{\widehat{a}-1}{a}\right\|_2\left\|\widehat{b}-b\right\|_\infty\right)\right\}. \tag{26}$$

The proof of the above theorem can be found in Appendix C (for kernel estimation bias bound) and S3 (for variance bound).

**Remark 7** (Asymptotic normality and the bootstrap approximation)**.** As shown in Liu et al. (2020a), the asymptotic normality of the oracle statistic $\frac{\widehat{\mathbb{IF}}_{2,2,k} - \mathsf{bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{se}_\theta(\widehat{\mathbb{IF}}_{2,2,k})}$ follows from Theorem 1 of Bhattacharya and Ghosh (1992) whence $1 \ll k \ll n^2$. Thus to show CLT of $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$ for any $m \geq 2$, it is sufficient to demonstrate under what conditions $\mathsf{kern\text{-}bias}_{m,k}(\widehat{\Omega}) \ll \mathsf{se}_\theta(\widehat{\mathbb{IF}}_{2,2,k})$. Bootstrap approximation (and its rate) of the distribution of $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$, or even of $\widehat{\mathbb{IF}}_{2,2,k}$ is still an important open problem, though Liu et al. (2021b) have made some partial progress. A more thorough study of the conditions under which central limit theorem (CLT) or bootstrap approximation holds is beyond the scope of this paper. ■

## 3.3 Numerical stability and time complexity of sHOIF estimators

In what follows we consider the numerical and computational properties of sHOIF estimators, which extends the results in Section 2.2 to higher-order. The first result in this section, Theorem 2, earmarks the "stability" of sHOIF estimators in terms of their independence of the eigenvalues of $\widehat{\Omega}$, the root cause of the instability of eHOIF estimators.

**Theorem 2.** $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})$ *does not depend on the eigenvalues of* $\widehat{\Omega}$*.*

*Proof.* The proof resembles the proof of Proposition 2 closely by realizing that, for any $i, j \in [n]$,

$$A_i \bar{\mathsf{z}}_k(X_i)^\top \widehat{\Omega} \bar{\mathsf{z}}_k(X_j) = U^A_{i,\bullet} U^{A\top} U U_{j,\bullet},$$

which is completely independent of the eigenvalues of $\widehat{\Omega}$. □

Hence sHOIF estimators do not suffer from any numerical instability resulted from the large condition number of the sample Gram matrix when we let $k$ near $n$ in practice.

**Remark 8.** Theorem 2 also suggests a better way to compute sHOIF estimators. Instead of computing the sample Gram matrix $\widehat{\Sigma}$ and its inverse $\widehat{\Omega}$ using numerical methods, we should instead perform SVD on the basis matrices $\bar{\mathsf{Z}}_{n,k}$ and $\bar{\mathsf{Z}}^A_{n,k}$ and then compute $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})$. In fact, the upcoming R package (Wanis et al., 2023) for computing HOIF related statistics exactly uses this strategy. ■

Since sHOIF estimators are numerically stable and thus are potentially useful tools for statistical practice (Liu et al., 2020a; Wanis et al., 2023), it is worth discussing the computational complexity of sHOIF estimators for general order $m$ as well.

**Theorem 3.** *The time complexity of computing* $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})$ *is* $O(\max\{(nk)^{\lceil (m-1)/2 \rceil}, nk^2\})$*.*

*Proof.* Similar to the proof of Proposition 2, we need to rewrite $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$ in the form of a linear combination of $V$-statistics. Without loss of generality, we take $\mathcal{E}_a(\widehat{a}; O) \equiv \mathcal{E}_b(\widehat{b}; O) \equiv 1$. But let us first represent $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$ as the following series:

$$\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega}) \equiv \sum_{j=1}^{m} (-1)^j \binom{m-1}{j-1} \mathbb{U}_{n,j} \left[ \bar{\mathsf{z}}_k(X_1)^\top \widehat{\Omega} \cdot \prod_{s=3}^{m} \left( Q_s \widehat{\Omega} \right) \cdot \bar{\mathsf{z}}_k(X_2) \right].$$

Figure 1: The amount of computational resource (a-axis) vs. the amount of kernel estimation bias reduction (y-axis)

Note that the number of summations in $\widehat{\mathbb{IF}}_{m,m,k}(\widehat{\Omega})$ is

$$n(n-1)\cdots(n-m+1) = \sum_{j=1}^{m}(-1)^{m-j}s(m,j)n^j$$

where $s(m,j)$ are unsigned Stirling numbers of the first kind, or the number of permutations on $m$ elements with $j$ cycles. Accordingly one can write an $m$-th order $U$-statistic into a linear combination of $V$-statistics from order 1 to order $m$, with the number of $j$-th order $V$-statistics, for $j = 1, \cdots, m$, equal to $s(m,j)$.

We further leverage the special structure of the $U$-statistic kernel for sHOIF estimators. □

**Remark 9.** Considering Theorem 1 and Theorem 3 in tandem, we attain the statistical-computational trade-off diagram depicted in Figure 1:

However, whether or not such statistical-computational trade-off is an emanation of possibly intrinsic computational hardness of estimating certain smooth statistical functionals is still an open problem[9].

Finally, we briefly comment on our philosophical stance on the usefulness of sHOIF estimators. sHOIF estimators are effectively infinite-order $U$-statistics, so given the current computing devices, there is no doubt that practitioners are not using sHOIF estimators in practice in near term. This is "conditional" on the availability of hardware. The numerical stability or lack thereof, however, is independent from such hardware constraint. ■

**Remark 10.** Theorem 3 also applies to eHOIF estimators (Liu et al., 2017) and the original HOIF estimators of Robins et al. (2016), that needs an estimate of the density of the covariates $X$, if the time for density estimation is not counted. ■

# 4 Applications of the statistical properties of sHOIF estimators

## 4.1 Semiparametric efficiency under minimal Hölder assumptions on the nuisance functions

In nonparametric statistics, the optimality of a statistical procedure is often evaluated under the Hölder nuisance models.

The above calculations culminate into the following theorem, which is the second main result of this paper.

[9]It seems not to have been explicitly recognized in the literature before us, except being alluded to during a discussion by James Robins (see https://www.youtube.com/watch?v=AUOnAfUjDVE) on Kennedy (2020) in the online causal inference seminar.

**Theorem 4.** *If $\mathcal{A} \times \mathcal{B} \subseteq \mathcal{H}(s_a, \mathcal{X}) \times \mathcal{H}(s_b, \mathcal{X})$ with $(s_a + s_b)/2 \geq d/4$, and choosing $m \asymp \sqrt{\log n}$ and $k \lesssim n / \log(n)^2$,*

$$\sqrt{n} \left( \widehat{\psi}_{m,k} - \psi(\theta) \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{E}[\mathsf{IF}_1(\theta)^2]) \tag{27}$$

*where $\mathbb{E}[\mathsf{IF}_1(\theta)^2]$ is the semiparametric efficiency bound of $\psi(\theta)$.*

**Remark 11.** According to the lower bound of Robins et al. (2009) under the Hölder nuisance model, $(s_a + s_b)/2 \geq d/4$ is the minimal condition for the existence of a semiparametric efficient estimator of $\psi(\theta)$. It is not unreasonable to expect that this minimal condition also holds for most, if not all, of the DRFs. ■

## 4.2 Implications on the assumption-free bias testing procedure of Liu et al. (2020a) and Liu et al. (2021b)

In light of the growing interest in understanding the performance of deep-learning-based causal inference (Farrell et al., 2021; Chen et al., 2020) and the gap between these theoretical results and empirical performance (Xu et al., 2022), Liu et al. (2020a) proposed the following oracle assumption-free valid nominal $\alpha$-level test statistic:

$$\widehat{\chi} \coloneqq \mathbb{1} \left\{ \frac{\widehat{\mathbb{IF}}_{2,2,k}}{\widehat{\mathsf{se}}[\widehat{\psi}_1]} - z_{\alpha/2} \frac{\widehat{\mathsf{se}}[\widehat{\mathbb{IF}}_{2,2,k}]}{\widehat{\mathsf{se}}[\widehat{\psi}_1]} > \delta \right\} \tag{28}$$

for the following null hypothesis:

$$\mathsf{H}_0(\delta) : \frac{\mathsf{cs\text{-}bias}(\widehat{\psi}_1)}{\mathsf{se}[\widehat{\psi}_1]} \leq \delta \tag{29}$$

where

$$\mathsf{cs\text{-}bias}(\widehat{\psi}_1) \coloneqq \left\{ \mathbb{E}\left[ \lambda(X)(\widehat{a}(X) - a(X))^2 \right] \mathbb{E}\left[ \lambda(X)(\widehat{b}(X) - b(X))^2 \right] \right\}^{1/2}. \tag{30}$$

Liu et al. (2021b) in turn constructed a feasible assumption-lean valid nominal $\alpha$-level test statistic

$$\widehat{\chi}_{3,k}(\widehat{\Omega}_{\mathrm{nuis}}) \coloneqq \mathbb{1} \left\{ \frac{\widehat{\mathbb{IF}}_{(2,2)\to(3,3),k}(\widehat{\Omega}_{\mathrm{nuis}})}{\widehat{\mathsf{se}}[\widehat{\psi}_1]} - z_{\alpha/2} \frac{\widehat{\mathsf{se}}[\widehat{\mathbb{IF}}_{(2,2)\to(3,3),k}(\widehat{\Omega}_{\mathrm{nuis}})]}{\widehat{\mathsf{se}}[\widehat{\psi}_1]} > \delta \right\} \tag{31}$$

and the following higher-order test statistic based on eHOIF estimators:

$$\widehat{\chi}_{m,k}(\widehat{\Omega}_{\mathrm{nuis}}) \coloneqq \mathbb{1} \left\{ \frac{\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega}_{\mathrm{nuis}})}{\widehat{\mathsf{se}}[\widehat{\psi}_1]} - z_{\alpha/2} \frac{\widehat{\mathsf{se}}[\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega}_{\mathrm{nuis}})]}{\widehat{\mathsf{se}}[\widehat{\psi}_1]} > \delta \right\}. \tag{32}$$

Liu et al. (2021b) showed that all the standard errors in the above test statistics can be estimated consistently by certain bootstrapping procedure. More importantly, they proved the following.

**Proposition 4.** *Let*

$$\mathsf{cs\text{-}bias}_k(\widehat{\psi}_1) \coloneqq \left\{ \mathbb{E}\left[ \Pi[(A\widehat{a} - 1)|\bar{\mathsf{z}}_k](X)^2 \right] \mathbb{E}\left[ \Pi[A(\widehat{b} - b)|\bar{\mathsf{z}}_k](X)^2 \right] \right\}^{1/2}.$$

*Under the assumptions of Theorem 4, with $k \lesssim n/(\log n)^2$ and the following extra condition:*

$$|\mathsf{kern\text{-}bias}_{3,k}(\widehat{\psi}_1)| \not\ll \mathsf{cs\text{-}bias}_k(\widehat{\psi}_1) \tag{33}$$

*then $\widehat{\chi}_{3,k}(\widehat{\Omega}_{tr})$ is a valid nominal $\alpha$-level test of $\mathsf{H}_0(\delta)$ (29). The extra condition (33) can be relaxed to*

$$|\mathsf{kern\text{-}bias}_{m,k}(\widehat{\psi}_1)| \not\ll \mathsf{cs\text{-}bias}_k(\widehat{\psi}_1) \left( \frac{k \log k}{n} \right)^{\frac{m-1}{2}} \tag{34}$$

*if one uses $\widehat{\chi}_{m,k}(\widehat{\Omega}_{tr})$ instead of $\widehat{\chi}_{3,k}(\widehat{\Omega}_{tr})$.*

We can similarly define the following sHOIF-based test statistics: for $m \geq 2$,

$$\widehat{\chi}_{m,k}(\widehat{\Omega}) \coloneqq \mathbb{1}\left\{ \frac{\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})}{\widehat{\mathsf{se}}[\widehat{\psi}_1]} - z_{\alpha/2} \frac{\widehat{\mathsf{se}}[\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})]}{\widehat{\mathsf{se}}[\widehat{\psi}_1]} > \delta \right\}.$$

Then as an immediate corollary of Theorem 1, we have

**Theorem 5.** *Under the assumptions of Theorem 4, with $k \lesssim n/(\log n)^2$ and a different relaxed extra condition from (34):*

$$|\mathsf{kern\text{-}bias}_{m,k}(\widehat{\psi}_1)| \not\ll \mathsf{cs\text{-}bias}_k(\widehat{\psi}_1) \left( \frac{k}{n} \right)^{m-1} \tag{35}$$

*then $\widehat{\chi}_{m,k}(\widehat{\Omega})$ is a valid nominal $\alpha$-level test of $\mathsf{H}_0(\delta)$ (29).*

Given the above theoretical guarantees, and further considering that the sHOIF estimators and tests have better finite-sample performance than the corresponding eHOIF estimators and tests, we recommend using $\widehat{\chi}_{m,k}(\widehat{\Omega})$ in practice. For more examples of its application, see Wanis et al. (2023).

# 5 Further extensions of sHOIF estimators

## 5.1 HOIFs for fixed-design settings

In Kennedy et al. (2020), the authors asked the following question:

> "$\cdots$ it would be useful to know how higher-order estimators and/or the bias test proposed by Liu et al. (2020a) perform in a fixed design setup, insofar as this is the right context for conditional-on-covariate inference $\cdots$"

In their rejoinder, Liu et al. (2020b) conjectured that sHOIFs may be *the* HOIFs under the fixed-design setting. In this section, we affirmatively prove this conjecture and hence make some partial progress towards answering the above question.

In the fixed-design setting, we need to redefine the parameter of interest $\psi(\theta)$ as

$$\psi^{\text{fixed}}(\theta) \coloneqq \mathbb{P}_n\left[b(X)\right] \equiv \mathbb{P}_n\left[\mathbb{E}[AY|X]a(X)\right]$$

with the same first-order IF as in the random-design setting

$$\mathbb{IF}_1^{\text{fixed}}(\theta) \equiv \mathbb{P}_n\left\{\mathbb{E}[A(Y-b(X))|X=X_i]a(X_i)+b(X_i)-\psi^{\text{fixed}}(\theta)\right\}. \tag{36}$$

**Remark 12.** The form of $\mathbb{IF}_1^{\text{fixed}}(\theta)$ in (36) can be derived by following van der Vaart (1991) and Ichimura and Newey (2022). ■

Following Section 3.2.1 of Robins et al. (2016), to derive HOIFs we need to consider a finite-dimensional sieve approximation $\widetilde{\psi}_{\theta,k}^{\text{fixed}}(\widehat{\theta})$ of $\psi^{\text{fixed}}(\theta)$ given initial nuisance estimates $\widehat{\theta}=(\widehat{a},\widehat{b})$ by the following DR estimating functions:

$$\begin{aligned}
\mathbb{P}_n\left[\bar{\mathsf{z}}_k(X)\{1-A(\widehat{a}(X)+\bar{\mathsf{z}}_k(X)^\top\widetilde{\alpha}_k)\}\right]&=0,\\
\mathbb{P}_n\left[\bar{\mathsf{z}}_k(X)A\{Y-(\widehat{b}(X)+\bar{\mathsf{z}}_k(X)^\top\widetilde{\beta}_k)\}\right]&=0
\end{aligned}$$

which gives

$$\begin{aligned}
\widetilde{\alpha}_k&=\mathbb{P}_n[\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]^{-1}\mathbb{P}_n[\bar{\mathsf{z}}_k(X)(1-A\widehat{a}(X))],\\
\widetilde{\beta}_k&=\mathbb{P}_n[A\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]^{-1}\mathbb{P}_n[\bar{\mathsf{z}}_k(X)A(Y-\widehat{b}(X))].
\end{aligned}$$

In turn, we have

$$\widetilde{\psi}_{\theta,k}^{\text{fixed}}(\widehat{\theta})=\widetilde{\alpha}_k^\top\mathbb{P}_n[\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]\widetilde{\beta}_k\equiv\mathbb{P}_n[\mathcal{E}_a(\widehat{a};O)\bar{\mathsf{z}}_k(X)^\top]\widehat{\Omega}\mathbb{P}_n[\bar{\mathsf{z}}_k(X)\mathcal{E}_b(\widehat{b};O)].$$

Then Theorem 3 of Robins et al. (2016), with the population expectations $\mathbb{E}[\cdot]$ replaced by empirical averages $\mathbb{P}_n[\cdot]$, immediately yields that the 2nd order and $m$-th order influence functions of $\widetilde{\psi}_{\theta,k}^{\text{fixed}}(\widehat{\theta})$ are $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ and $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$.

## 5.2 Generalization to the entire class of DRFs

In this subsection, we briefly comment on how our results can be generalized to the entire class of DRFs characterized in Rotnitzky et al. (2021). The class of DRFs includes many other functionals that arise in substantive studies in (bio)statistics, epidemiology, economics, and social sciences, including:

- the expected conditional variance, which is useful for constructing confidence/predictive sets (Robins and van der Vaart, 2006);
- the expected conditional covariance, which is useful for both causal inference and conditional independence testing (Shah and Peters, 2020);

- average causal effect of continuous treatment, which is important for treatment allocations (Ai et al., 2021; Bonvini and Kennedy, 2022).

Rotnitzky et al. (2021) defined the class of DRFs as follows:

**Definition 1** (Doubly Robust Functionals; Definition 1 of Rotnitzky et al. (2021))**.** $\psi(\theta)$ is a doubly robust functional if, for each $\theta \in \Theta$ there exists $a : x \mapsto a(x) \in \mathcal{A}$ and $b : x \mapsto b(x) \in \mathcal{B}$ such that (i) $\theta = (a, b, \theta \setminus \{b, p\})$ and $\Theta = \mathcal{A} \times \mathcal{B} \times \Theta \setminus \{\mathcal{A}, \mathcal{B}\}$ and (ii) for any $\theta, \theta' \in \Theta$

$$\psi(\theta) - \psi(\theta') + \mathbb{E}\left[\mathsf{IF}_1(\theta')\right] = \mathbb{E}\left[S(a(X) - a'(X))(b(X) - b'(X))\right] \tag{37}$$

where $S \equiv s(O)$ with $o \mapsto s(o)$ a known function that does not depend on $\theta$ or $\theta'$ satisfying either $\mathbb{P}_\theta(S \geq 0) = 1$ or $\mathbb{P}_\theta(S \leq 0) = 1$. We also denote $\lambda(x) \coloneqq \mathbb{E}[S|X = x]$. Then the first-order influence function of $\psi(\theta)$ has the following form: given $\theta' \equiv (a', b', \theta' \setminus \{a', b'\})^\top \in \Theta$,

$$\mathsf{IF}_1(\theta') \equiv Sa'(X)b'(X) + m_a(O, a') + m_b(O, b') + S_0 \tag{38}$$

where $S_0$ is some known statistic that does not depend on $a'$ and $b'$, and $h \mapsto m_a(o, h)$ for $h \in \mathcal{A}$ and $h \mapsto m_b(o, h)$ for $h \in \mathcal{B}$ are two known linear maps satisfying

$$\begin{aligned} &\mathbb{E}\left[Sh(X)b(X) + m_a(O, h)\right] \equiv 0, \ \forall\ h \in \mathcal{A}, \\ &\mathbb{E}\left[Sa(X)h(X) + m_b(O, h)\right] \equiv 0, \ \forall\ h \in \mathcal{B}. \end{aligned}$$

As a result, $\psi(\theta) \equiv \mathbb{E}[m_a(O, a) + S_0] \equiv \mathbb{E}[m_b(O, b) + S_0]$.

**Remark 13.** For $\psi(\theta) \equiv \mathbb{E}[Y(1)]$ under strong ignorability, $S$, $a$, $b$, $m_a(O, a)$, $m_b(O, b)$, and $S_0$ correspond to $-A$, $\{\mathbb{E}[A|X]\}^{-1}$, $\mathbb{E}[Y|X, A = 1]$, $AYa(X)$, $b(X)$, and 0, respectively. Thus $\lambda(X) = \mathbb{E}[-A|X] = -\frac{1}{a(X)}$. We also have

$$\begin{aligned} \mathbb{E}[\mathcal{E}_a(\widehat{a}; O)\bar{\mathsf{z}}_k(X)] &= \mathbb{E}\left[\frac{1}{a(X)}(\widehat{a}(X) - a(X))\bar{\mathsf{z}}_k(X)\right], \\ \mathbb{E}[\mathcal{E}_b(\widehat{b}; O)\bar{\mathsf{z}}_k(X)] &= \mathbb{E}\left[\frac{1}{a(X)}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right]. \end{aligned}$$

■

We have the following notation correspondence that maps the results for $\psi(\theta) \equiv \mathbb{E}[Y(1)]$ under strong ignorability to any DRF $\psi(\theta)$:

- $\mathcal{E}_a(\widehat{a}; O)\bar{\mathsf{z}}_k(X) \Rightarrow \mathcal{E}_a(\widehat{a}, \bar{\mathsf{z}}_k; O)$ and $\mathcal{E}_b(\widehat{b}; O)\bar{\mathsf{z}}_k(X) \Rightarrow \mathcal{E}_b(\widehat{b}, \bar{\mathsf{z}}_k; O)$ where $\mathcal{E}_a(\widehat{a}, \bar{\mathsf{z}}_k; O)$ and $\mathcal{E}_b(\widehat{b}, \bar{\mathsf{z}}_k; O)$ satisfy

$$\begin{aligned} \mathbb{E}[\mathcal{E}_a(\widehat{a}, \bar{\mathsf{z}}_k; O)] &= \mathbb{E}\left[\lambda(X)(\widehat{a}(X) - a(X))\bar{\mathsf{z}}_k(X)\right], \\ \mathbb{E}[\mathcal{E}_b(\widehat{b}, \bar{\mathsf{z}}_k; O)] &= \mathbb{E}\left[\lambda(X)(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right]. \end{aligned}$$

- $\Sigma = \mathbb{E}[A\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top] \Rightarrow \Sigma = \mathbb{E}[S\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]$ and $\widehat{\Sigma} = \mathbb{P}_n[A\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top] \Rightarrow \widehat{\Sigma} = \mathbb{P}_n[S\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]$.

With the above mappings, all the theoretical results for $\psi(\theta) \equiv \mathbb{E}[Y(1)]$ developed herein can be applied to those for an arbitrary DRF $\psi(\theta)$ *mutatis mutandis*.

### 5.2.1 A special case: the expected conditional covariance

Before concluding our paper, we further study the implications of the sHOIF theory developed so far for a special cases of DRFs: the expected conditional covariance between two random variables $A$ and $Y$ given a third random variable $X$, $\psi \coloneqq \mathbb{E}[\mathsf{cov}(A, Y|X)]$. When $A = Y$ almost surely, $\psi$ reduces to the expected conditional variance of $A$ given $X$, $\psi \coloneqq \mathbb{E}[\mathsf{cov}(A|X)]$. For differences between these two parameters, see an extended discussion in Liu et al. (2020a).

The main feature that distinguishes $\psi$ from many other DRFs is $S = 1$, which leads to its SOIF:

$$\widehat{\mathbb{IF}}_{2,2,k} = \frac{1}{n(n-1)} \sum_{1 \leq i_1 \neq i_2 \leq n} (A_{i_1} - \widehat{a}(X_{i_1}))\bar{\mathsf{z}}_k(X_{i_1})^\top \bar{\Omega} \bar{\mathsf{z}}_k(X_{i_2})(Y_{i_2} - \widehat{b}(X_{i_2})),$$

in which $\bar{\Omega} \equiv \{\mathbb{E}[S\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]\}^{-1} \equiv \{\mathbb{E}[\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]\}^{-1}$ only depends on the distribution of $X$. Also, $\widehat{a}$ and $\widehat{b}$ are nuisance estimates of $a(x) = \mathbb{E}[A|X = x]$ and $b(x) = \mathbb{E}[Y|X = x]$ in this context. This leads to the following improved kernel estimation bias bound:

**Corollary 1.** *Under Assumptions 1 – 2, with* $k \lesssim \frac{n}{\log^2 n}$ *and* $m \gtrsim \sqrt{\log n}$*, one has the following: The kernel estimation bias of* $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$ *satisfies*

$$\begin{aligned} &\mathsf{kern\text{-}bias}_{m,k}(\widehat{\psi}_1) \coloneqq \mathbb{E}\left[\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})\right] - \mathsf{bias}_{\theta,k}(\widehat{\psi}_1) \equiv \mathbb{E}\left[\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega}) - \widehat{\mathbb{IF}}_{2,2,k}\right] \\ &\lesssim \left(\frac{km}{n}\right)^{\lceil \frac{\lceil \frac{m-1}{2} \rceil - 1}{2} \rceil \vee 1} \left\{ \|\widehat{a} - a\|_2 \|\widehat{b} - b\|_2 + \left( \|\widehat{a} - a\|_2 \left\|\widehat{b} - b\right\|_\infty \wedge \|\widehat{a} - a\|_\infty \left\|\widehat{b} - b\right\|_2 \right) \right\}. \end{aligned} \tag{39}$$

Note that the variance bound is improved in a similar manner and is omitted here.

## 5.3 sHOIF estimators for proximal causal learning

In this section, we generalize the previous results to the estimation of $\psi(\theta)$ under the proximal causal learning framework (Tchetgen Tchetgen et al., 2020), allowing to handle endogeneity issues by leveraging auxiliary variables called "proxies".

To state sHOIFs of proximal $\psi(\theta)$, we need to introduce some extra notation and assumptions. Instead of observing i.i.d. copies of $(X, A, Y) \sim \mathbb{P}_\theta$, we observe i.i.d. copies of $(X, Z, A, W, Y) \sim \mathbb{P}_\theta$ where $Z$ and $W$ are treatment and outcome proxies (abbreviated as $A$-proxy and $Y$-proxy in the sequel) that satisfy

**Assumption 3.**

(1) Validity: There exists *baseline* measured factor $X$ and unmeasured factor $U$ such that ignorability holds conditional on both $X, U$; $Y \perp\!\!\!\perp Z|U, X, A$ and $(A, Z) \perp\!\!\!\perp W|U, X$.

(2) Completeness:

$$\begin{aligned}
&\mathbb{E}[g(U)|X=x,Z,A=\mathfrak{a}]=0 \text{ almost surely for all } \mathfrak{a}\in\mathcal{A} \text{ and } x\in\mathcal{X}\\
\Leftrightarrow\ & g(U)\equiv 0 \text{ almost surely,}\\
&\mathbb{E}[g(U)|X=x,A=\mathfrak{a},W]=0 \text{ almost surely for all } \mathfrak{a}\in\mathcal{A} \text{ and } x\in\mathcal{X}\\
\Leftrightarrow\ & g(U)\equiv 0 \text{ almost surely.}
\end{aligned} \tag{40}$$

(3) Existence and uniqueness of confounding bridge functions: There exists uniquely identified $A$- and $Y$-confounding bridge functions that satisfy the following (generally ill-posed) Fredholm integral equations of the first kind

$$\begin{aligned}
&\mathbb{E}[a(X,Z,A=\mathfrak{a})|X,A=\mathfrak{a},W]=\frac{1}{\mathbb{P}_\theta(A=\mathfrak{a}|X,W)},\\
&\mathbb{E}[b(X,A,W)|X,Z,A]=\mathbb{E}[Y|X,Z,A].
\end{aligned} \tag{41}$$

Then $\psi(\theta)=\mathbb{E}[Y(1)]$ can be identified as the following statistical functional of the observed data distribution:

$$\psi(\theta)\equiv\mathbb{E}\left[\mathbb{E}[b(X,A,W)|X,A=1]\right]\equiv\mathbb{E}\left[Aa(X,Z,A)Y\right].$$

Cui et al. (2020) derived the first-order IF of $\psi(\theta)$ under Condition 3:

$$\mathsf{IF}_{1,\theta}=Aa(X,Z,A)(Y-b(X,A,W))+b(X,A,W)-\psi(\theta) \tag{42}$$

which gives rise to first-order doubly robust estimators of $\psi(\theta)$:

$$\widehat{\psi}_1=\frac{1}{n}\sum_{i=1}^{n}A_i\widehat{a}(X_i,Z_i,A_i)(Y_i-\widehat{b}(X_i,A_i,W_i))+\widehat{b}(X_i,A_i,W_i). \tag{43}$$

Liu et al. (2021b) derived the HOIFs of $\psi(\theta)$ under the proximal causal learning framework:

$$\widehat{\mathbb{IF}}_{m,m,k}=\mathbb{U}_{n,m}\left\{\begin{array}{c}(-1)^m\left[(A\widehat{a}(X,Z,A)-1)\,\bar{\mathsf{v}}_k(X,A,W)\right]_{i_{m-1}}^\top\Omega_{zv}\\ \times\prod_{s=1}^{m-2}\left[\left(A\bar{\mathsf{z}}_k(X,A,Z)\bar{\mathsf{v}}_k(X,A,W)^\top-\Sigma_{zv}\right)\Omega_{zv}\right]_{i_s}\\ \times\left[\bar{\mathsf{z}}_k(X,Z,A)A\left(Y-\widehat{b}(X,A,W)\right)\right]_{i_m}\end{array}\right\} \tag{44}$$

where $\Sigma_{zv}:=\mathbb{E}\left[\bar{\mathsf{z}}_k(X,A,Z)\bar{\mathsf{v}}_k(X,A,W)^\top\right]$ and $\Omega_{zv}:=\Sigma_{zv}^{-1}$. The corresponding sHOIF estimators simply replace $\Omega_{zv}$ by $\widehat{\Omega}_{zv}:=n^{-1}\sum_{i=1}^n\bar{\mathsf{z}}_k(X_i,A_i,Z_i)\bar{\mathsf{v}}_k(X_i,A_i,Z_i)^\top$. Hence the results in this paper can be straightforwardly extended to the proximal causal inference setting.

# 6 Discussion

In this paper, we propose a novel class of HOIF estimators, stable HOIF (sHOIF) estimators, for the doubly robust functionals (DRFs) characterized in Rotnitzky et al. (2021). They are semiparametric efficient under the minimal Hölder-smoothness condition $\frac{s_a+s_b}{2} > \frac{d}{4}$ of Robins et al. (2009), allowing the dimension $k$ of the basis function diverging at a rate just slower than the sample size $n$. As can be seen from Theorem 1, sHOIF estimators have improved rate of convergence than eHOIF developed in Liu et al. (2017). More importantly, as well documented in the simulation studies of Liu et al. (2020a) and Wanis et al. (2023), the sHOIF estimators also have significantly better finite-sample performance over existing higher-order estimators in practice, making them more amenable for tasks such as testing if the bias of a first-order DML estimator $\widehat{\psi}_1$ of a causal effect is dominated by its standard error (Liu et al., 2020a, 2021b; Wanis et al., 2023). Finally, we end our paper by mentioning several future research directions:

(1) It will be interesting to study if one can extend the idea of sHOIF estimators to the non-$\sqrt{n}$-estimable regimes by, for instance, estimating $\Omega$ via some shrinkage or regularized algorithms. As conjectured in Robins et al. (2016), the minimax convergence rate of the functionals studied in this paper may depend on the regularity of the density of the covariates $X$. Hence it is expected that the shrinkage or regularization also depends on the density of $X$. Simulation studies in Liu et al. (2020a) suggest the nonlinear shrinkage covariance matrix estimators of Ledoit and Wolf (2012) could be a viable option. Preliminary simulation studies in Liu et al. (2020a) and Wanis et al. (2023) suggest that the performance of these shrinkage covariance matrix estimators does degrade with the smoothness of the design density.

(2) As pointed out in Kennedy et al. (2022), their Second-Order R-Learner (SORL) for CATE also involves inverting large Gram matrices of certain basis functions (in which they use the Legendre polynomials) under additional complexity-reducing assumptions on the covariates $X$. It will be interesting to investigate if the sHOIF estimators can be generalized to the CATE estimation problems and stabilize their SORL or even HORL estimators.

(3) Another important open problem was also mentioned in van der Vaart (2014); Liu et al. (2020a, 2021b). To define HOIFs for DRFs, one needs to choose a set of $k$-dimensional basis functions $\bar{\mathsf{z}}_k$ or an approximation kernel $K$ of the Kronecker delta function, ideally in prior to the data analysis. However, such a strategy seems to go against the current data analytic paradigm, which strongly advocates learning representations (e.g. in the form of bases or kernels) adaptively from data rather than choosing some fixed bases/frames *a priori*. Prominent examples include DNNs, autoencoders, and GANs. It is thus interesting to construct HOIF estimators along different basis directions and then select one or aggregate all, guided by certain optimality criterion. We leave this important and difficult problem to future endeavor.

(4) It will be interesting to also derive HOIFs and sHOIFs for identifiable causal effect functionals in graphical models with latent variables (Bhattacharya et al., 2022) and implicitly defined functionals (Robins et al., 2016; Ai et al., 2021) in general semiparametric regression problems for improved quality of estimation and statistical inference, which however requires extension of the current work to $U$-processes, a much more difficult research problem that we are working on in a separate paper.

(5) Finally, as pointed out in Section 1.1, deeper connections between sHOIF estimators and Kikuchi hierarchy from statistical physics (Yedidia et al., 2003; Wein et al., 2019) except for the similarity in their mathematical structure and analysis strategy could be a fruitful future research direction.

# References


Chunrong Ai, Oliver Linton, Kaiji Motegi, and Zheng Zhang. A unified framework for efficient estimation of general treatment models. Quantitative Economics, 12(3):779–816, 2021.

Afonso S Bandeira, March T Boedihardjo, and Ramon van Handel. Matrix concentration inequalities and free probability. arXiv preprint arXiv:2108.06312, 2021.

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.

Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. Journal of Machine Learning Research, 23:1–76, 2022.

Peter J Bickel, Chris A J Klaassen, Ya'acov Ritov, and Jon A Wellner. Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences. Springer New York, 1998. ISBN 9780387984735.

Matteo Bonvini and Edward H Kennedy. Fast convergence rates for dose-response estimation. arXiv preprint arXiv:2207.11825, 2022.

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

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.

Romain Couillet and Zhenyu Liao. Random Matrix Methods for Machine Learning. Cambridge University Press, 2022.

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

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

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. The Annals of Statistics, 49(6):3206–3227, 2021.

Hidehiko Ichimura and Whitney K Newey. The influence function of semiparametric estimators. Quantitative Economics, 13(1):29–61, 2022.

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.

Edward H Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Minimax rates for heterogeneous causal effect estimation. arXiv preprint arXiv:2203.00837, 2022.

Tonu Kollo and Dietrich von Rosen. Advanced multivariate statistics with matrices. Springer, 2005.

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics, 40(2):1024–1060, 2012.

Oleg V Lepskii. On a problem of adaptive estimation in Gaussian white noise. Theory of Probability & Its Applications, 35(3):454–466, 1991.

Chang Li and Lin Liu. Supplement to "New numerically stable, semiparametric efficient empirical higher-order influence functions". 2023.

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.

Lin Liu, Rajarsh Mukherjee, James M Robins, and Eric Tchetgen Tchetgen. Adaptive estimation of nonparametric functionals. Journal of Machine Learning Research, 22(99): 1–66, 2021a.

Lin Liu, Rajarshi Mukherjee, and James M Robins. Assumption-lean tests of the validity of published double-machine-learning Wald confidence intervals for doubly-robust functionals. Technical report, Shanghai Jiao Tong University, 2021b. URL https://linliu-stats.github.io/files/Dec16-Assumption-freetestecon.pdf.

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

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

Alexandru Nica and Roland Speicher. Lectures on the combinatorics of free probability, volume 13. Cambridge University Press, 2006.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. Self-normalized processes: Limit theory and statistical applications. Springer Science & Business Media, 2008.

James Robins and Aad van der Vaart. Adaptive nonparametric confidence sets. The Annals of Statistics, 34(1):229–253, 2006.

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. Optimal structural nested models for optimal sequential decisions. In Proceedings of the Second Seattle Symposium in Biostatistics, pages 189–326. Springer, 2004.

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, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89(427):846–866, 1994.

James M Robins, Lingling Li, Lin Liu, 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.

James M Robins, Lingling Li, Lin Liu, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Corrigenda to "Minimax estimation of a functional on a structured high-dimensional model". 2022.

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

Mark Rudelson. Random vectors in the isotropic position. Journal of Functional Analysis, 164(1):60–72, 1999.

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, 1999.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. The Annals of Statistics, 48(3):1514–1538, 2020.

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. On differentiable functionals. The Annals of Statistics, 19(1):178–204, 1991.

Aad van der Vaart. Part III: Semiparameric statistics. Lectures on Probability Theory and Statistics, pages 331–457, 2002.

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

Kerollos Wanis, Chang Li, Lin Liu, Nelya Melnitchoukc, and James M Robins. Machine learning and causal inference: Quantifying bias using higher order influence functions. Technical report, Harvard University, 2023.

Richard P Waterman and Bruce G Lindsay. Projected score methods for approximating conditional scores. Biometrika, 83(1):1–13, 1996.

Alexander S Wein, Ahmed El Alaoui, and Cristopher Moore. The Kikuchi hierarchy and tensor PCA. In 2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS), pages 1446–1468. IEEE, 2019.

Siqi Xu, Lin Liu, and Zhonghua Liu. DeepMed: Semiparametric causal mediation analysis with debiased deep learning. Advances in Neural Information Processing Systems, 36, 2022.

Jonathan S Yedidia, William Freeman, and Yair Weiss. Understanding Belief Propagation and Its Generalizations, page 239–269. Morgan Kaufmann Publishers Inc., 2003.

Ruoqi Yu and Shulei Wang. Treatment effects estimation by uniform transformer. arXiv preprint arXiv:2008.03738, 2020.

# Appendix A Proof of the variance part of Proposition 1

## A.1 Proof of Lemma 1

We prove the second statement (20) first.

$$\begin{aligned}
&\left|\mathbb{E}\left[\bar{\mathsf{z}}_k(X_1)^\top \bar{\mathsf{z}}_k(X_2)\bar{\mathsf{z}}_k(X_3)^\top Q_{3,4}^j \bar{\mathsf{z}}_k(X_4)\right]\right| \\
= &\left|\mathbb{E}[\bar{\mathsf{z}}_k(X_1)]^\top \mathbb{E}[\bar{\mathsf{z}}_k(X_2)]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_3)^\top Q_{3,4}^j \bar{\mathsf{z}}_k(X_4)\right]\right| \\
\lesssim &\left|\mathbb{E}\left[\bar{\mathsf{z}}_k(X_3)^\top Q_{3,4}^j \bar{\mathsf{z}}_k(X_4)\right]\right| \\
\leq &\left\{\mathbb{E}\left[\bar{\mathsf{z}}_k(X_4)^\top Q_{3,4}^j \bar{\mathsf{z}}_k(X_3)\bar{\mathsf{z}}_k(X_3)^\top Q_{3,4}^j \bar{\mathsf{z}}_k(X_4)\right]\right\}^{1/2} \lesssim k^j.
\end{aligned}$$

For the first statement (19), since $j_1 > 0$, there must exist at least one $Q_{1,2}$ between $\bar{\mathsf{z}}_k(X_3)$ and $\bar{\mathsf{z}}_k(X_4)$ in (19). We conduct induction on $j$. $j = 1$ has been proved in the main text. Suppose (19) holds for $j - 1$. Then by applying Cauchy-Schwarz inequality, one can easily exhibit (19) for $j$.

# Appendix B Derivation of (23)

Recall that the kernel estimation bias $\mathsf{kern\text{-}bias}_{2,k}(\widehat{\psi}_1)$ of $\widehat{\mathbb{IF}}_{2,2,k}(\widehat{\Omega})$ truncated at level $j = 2$ is dominated by

$$\frac{1}{n^2}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top Q_{1,2}^2 \bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] + \frac{1}{n}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\right]\mathbb{E}[Q_3^2]\mathbb{E}\left[\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] \tag{45}$$

which equals the sum of (11) and (14).

We further consider the kernel estimation bias of $\widehat{\mathbb{IF}}_{3,3,k}(\widehat{\Omega})$ as an estimator of (10), the dominating term of $\mathsf{kern\text{-}bias}_{2,k}(\widehat{\psi}_1)$ truncated at level $j = 1$:

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mathbb{IF}}_{3,3,k}(\widehat{\Omega}) - (10)\right] \\
= &\,\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top \left(Q_{1,2}^2 - \widehat{\Omega}Q_{1,2}^2\widehat{\Omega}\right)\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] + O(n^{-1}) \\
= &-\frac{2}{n^2}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top Q_{1,2}^2 \bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] + \mathsf{Rem} + O(n^{-1})
\end{aligned}$$

where the remainder term $\mathsf{Rem}$ can be shown to dominated by the first term. Adding this term to (45), we conclude that the dominating part of the kernel estimation bias of $\widehat{\mathbb{IF}}_{(2,2)\to(3,3),k}(\widehat{\Omega})$ cancels with $\mathbb{E}[\widetilde{\widehat{\mathbb{IF}}}_{4,4,k}]$.

# Appendix C Proof of the kernel estimation bias bound in Theorem 3.2

We divide the proof of Theorem 3.2 into several steps. First, in Section C.1, we provide alternative characterization of $m$-th order sHOIFs $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$ to facilitate the bias control.

## C.1 Alternative characterization of sHOIFs

We have the following alternative characterization of $m$-th order sHOIFs, which can be shown by induction:

$$\begin{aligned}
&\widehat{\mathbb{IF}}_{2,2,k} - \widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega}) \\
&\equiv \mathbb{U}_{n,2}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\left\{\sum_{j=0}^{m-1}(-1)^j\binom{m-1}{j}\mathbb{U}_{n-2,j-1}\left(\widehat{\Omega}\prod_{s=3}^{j+1}Q_s\widehat{\Omega}\right)\right\}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right] \\
&\equiv \mathbb{U}_{n,2}\left[\mathcal{E}_a(\widehat{a};O_1)\bar{\mathsf{z}}_k(X_1)^\top\left\{\sum_{j=1}^{m-1}(-1)^j\binom{m-1}{j}\mathbb{U}_{n-2,j}\left(\widehat{\Omega}\prod_{s=3}^{j+1}Q_s\widehat{\Omega}-\mathbb{I}\right)\right\}\bar{\mathsf{z}}_k(X_2)\mathcal{E}_b(\widehat{b};O_2)\right]
\end{aligned} \tag{46}$$

where $\widehat{\Omega}\prod_{s=3}^{2}Q_s\widehat{\Omega}$ and $\widehat{\Omega}\prod_{s=3}^{1}Q_s\widehat{\Omega}$ are understood to be $\widehat{\Omega}$ and $\mathbb{I}$, respectively.

Armed with (46), we can characterize the kernel estimation bias of $\widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})$ as follows.

**Lemma 2.**

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mathbb{IF}}_{2,2,k} - \widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})\right] \\
&= \sum_{j=1}^{m-1}(-1)^j\binom{m-1}{j}\sum_{\ell=1}^{j-1}\sum_{\substack{S\subseteq[j-1],\\|S|=\ell}}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_{m-1})\bar{\mathsf{z}}_k(X_{m-1})^\top\prod_{s=0}^{j-1}\left\{Q_s(\widehat{\Omega}-\mathbb{I})^{\mathbb{1}\{s\in S\}}\right\}\bar{\mathsf{z}}_k(X_m)\mathcal{E}_b(\widehat{b};O_m)\right]
\end{aligned} \tag{47}$$

$$= \sum_{\mathfrak{c}=1}^{m-1}\sum_{j=\mathfrak{c}}^{m-1}(-1)^j\binom{m-1}{j}\sum_{\substack{S\subseteq[j-1],\\|S|=\mathfrak{c}}}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_{m-1})\bar{\mathsf{z}}_k(X_{m-1})^\top\prod_{s=0}^{j-1}\left\{Q_s(\widehat{\Omega}-\mathbb{I})^{\mathbb{1}\{s\in S\}}\right\}\bar{\mathsf{z}}_k(X_m)\mathcal{E}_b(\widehat{b};O_m)\right]. \tag{48}$$

*Proof.* To avoid notation clutter, we introduce the alias term $Q_0 \coloneqq \mathbb{I}$. We also overload the notation $[l] \coloneqq \{0, 1, \cdots, l\}$ for any non-negative integer $l$. We first prove (47):

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mathbb{IF}}_{2,2,k} - \widehat{\mathbb{IF}}_{(2,2)\to(m,m),k}(\widehat{\Omega})\right] \\
&= \sum_{j=1}^{m-1}(-1)^j\binom{m-1}{j}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_{m-1})\bar{\mathsf{z}}_k(X_{m-1})^\top\left(\prod_{s=0}^{j-1}(Q_s\widehat{\Omega})-\mathbb{I}\right)\bar{\mathsf{z}}_k(X_m)\mathcal{E}_b(\widehat{b};O_m)\right]
\end{aligned}$$

$$= \sum_{j=1}^{m-1}(-1)^j\binom{m-1}{j}\sum_{\ell=1}^{j-1}\sum_{\substack{S\subseteq[j-1],\\|S|=\ell}}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_{m-1})\bar{\mathsf{z}}_k(X_{m-1})^\top\prod_{s=0}^{j-1}\left\{Q_s(\widehat{\Omega}-\mathbb{I})^{\mathbb{1}\{s\in S\}}\right\}\bar{\mathsf{z}}_k(X_m)\mathcal{E}_b(\widehat{b};O_m)\right]$$

where the second equality follows from decomposing each $\widehat{\Omega}$ into $\mathbb{I}-(\mathbb{I}-\widehat{\Omega})$ for $j=2,\cdots,m-1$.

For (48), we proceed by reorganizing all the summands in (47) according to the copy number $\mathfrak{c}$ of $\mathbb{I}-\widehat{\Omega}$ in the product:

$$(47) = \sum_{\mathfrak{c}=1}^{m-1}\sum_{j=\mathfrak{c}}^{m-1}(-1)^j\binom{m-1}{j}\sum_{\substack{S\subseteq[j-1],\\|S|=\mathfrak{c}}}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_{m-1})\bar{\mathsf{z}}_k(X_{m-1})^\top\prod_{s=0}^{j-1}\left\{Q_s(\widehat{\Omega}-\mathbb{I})^{\mathbb{1}\{s\in S\}}\right\}\bar{\mathsf{z}}_k(X_m)\mathcal{E}_b(\widehat{b};O_m)\right].$$

□

## C.2 Analysis by matrix expansion and combinatorics

After different terms in the kernel estimation bias are reorganized as in (48), we perform the following expansion of $\mathbb{I}-\widehat{\Omega}$:

$$\widehat{\Omega}-\mathbb{I} = \sum_{j=1}^{J}(\mathbb{I}-\widehat{\Sigma})^j + (\mathbb{I}-\widehat{\Sigma})^{J+1}\widehat{\Omega}. \tag{49}$$

We denote the above expansion up to $J$-th order as

$$[\widehat{\Omega}-\mathbb{I}]_J \coloneqq \sum_{j=1}^{J}\left(\mathbb{I}-\widehat{\Sigma}\right)^j.$$

We then proceed by collecting different terms together by the copy number $\mathfrak{c}$ of $\widehat{\Sigma}-\mathbb{I}$, and obtain the following lemma.

**Lemma 3.** *With $\mathbb{I}-\widehat{\Omega}$ replaced by $[\mathbb{I}-\widehat{\Omega}]_J$, (48) can be rewritten as $\sum_{\mathfrak{c}=1}^{(m-1)J}\mathfrak{M}_{\mathfrak{c}}$, where*

$$\mathfrak{M}_{\mathfrak{c}} \coloneqq \sum_{j=1}^{m-1}(-1)^j\binom{m-1}{j}\sum_{\mathfrak{c}'=1}^{\mathfrak{c}\wedge j}\sum_{\substack{S\subseteq[j-1],\\|S|=\mathfrak{c}'}}\sum_{\substack{\{\ell_s,s\in S\}\subseteq\{1,\cdots,J\}^{\mathfrak{c}'},\\ \sum_{s\in S}\ell_s=\mathfrak{c}}}\mathbb{E}\left[\begin{array}{c}\mathcal{E}_a(\widehat{a};O_{m-1})\bar{\mathsf{z}}_k(X_{m-1})^\top\\ \times\prod_{s=0}^{j-1}\left\{Q_s(\mathbb{I}-\widehat{\Sigma})^{\ell_s\mathbb{1}\{s\in S\}}\right\}\\ \times\,\bar{\mathsf{z}}_k(X_m)\mathcal{E}_b(\widehat{b};O_m)\end{array}\right]. \tag{50}$$

*Proof.* The proof follows from a few lines of algebra.

$$\sum_{\mathfrak{c}'=1}^{m-1}\sum_{j=\mathfrak{c}'}^{m-1}(-1)^j\binom{m-1}{j}\sum_{\substack{S\subseteq[j-1],\\|S|=\mathfrak{c}'}}\mathbb{E}\left[\mathcal{E}_a(\widehat{a};O_{m-1})\bar{\mathsf{z}}_k(X_{m-1})^\top\prod_{s=0}^{j-1}\left\{Q_s[\widehat{\Omega}-\mathbb{I}]_J^{\mathbb{1}\{s\in S\}}\right\}\bar{\mathsf{z}}_k(X_m)\mathcal{E}_b(\widehat{b};O_m)\right]$$

$$= \sum_{\mathfrak{c}'=1}^{m-1} \sum_{j=\mathfrak{c}'}^{m-1} (-1)^j \binom{m-1}{j} \sum_{\substack{S \subseteq [j-1], \\ |S| = \mathfrak{c}'}} \mathbb{E}\left[ \begin{array}{c} \mathcal{E}_a(\widehat{a}; O_{m-1}) \bar{\mathsf{z}}_k(X_{m-1})^\top \\ \times \prod_{s=0}^{j-1} \left\{ Q_s \left( \sum_{\ell_s=1}^{J} (\mathbb{I} - \widehat{\Sigma})^{\ell_s} \right)^{\mathbb{1}\{s \in S\}} \right\} \\ \times\ \bar{\mathsf{z}}_k(X_m) \mathcal{E}_b(\widehat{b}; O_m) \end{array} \right]$$

$$= \sum_{\mathfrak{c}=1}^{(m-1)J} \sum_{\mathfrak{c}'=1}^{(m-1)\wedge \mathfrak{c}} \sum_{j=\mathfrak{c}'}^{m-1} (-1)^j \binom{m-1}{j} \sum_{\substack{S \subseteq [j-1], \\ |S| = \mathfrak{c}'}} \sum_{\substack{\{\ell_s, s \in S\} \subseteq \{1, \cdots, J\}^{\mathfrak{c}'}, \\ \sum_{s \in S} \ell_s = \mathfrak{c}}} \mathbb{E}\left[ \begin{array}{c} \mathcal{E}_a(\widehat{a}; O_{m-1}) \bar{\mathsf{z}}_k(X_{m-1})^\top \\ \times \prod_{s=0}^{j-1} \left\{ Q_s (\mathbb{I} - \widehat{\Sigma})^{\ell_s \mathbb{1}\{s \in S\}} \right\} \\ \times\ \bar{\mathsf{z}}_k(X_m) \mathcal{E}_b(\widehat{b}; O_m) \end{array} \right]$$

$$= \sum_{\mathfrak{c}=1}^{(m-1)J} \sum_{j=1}^{m-1} (-1)^j \binom{m-1}{j} \sum_{\mathfrak{c}'=1}^{\mathfrak{c} \wedge j} \sum_{\substack{S \subseteq [j-1], \\ |S| = \mathfrak{c}'}} \sum_{\substack{\{\ell_s, s \in S\} \subseteq \{1, \cdots, J\}^{\mathfrak{c}'}, \\ \sum_{s \in S} \ell_s = \mathfrak{c}}} \mathbb{E}\left[ \begin{array}{c} \mathcal{E}_a(\widehat{a}; O_{m-1}) \bar{\mathsf{z}}_k(X_{m-1})^\top \\ \times \prod_{s=0}^{j-1} \left\{ Q_s (\mathbb{I} - \widehat{\Sigma})^{\ell_s \mathbb{1}\{s \in S\}} \right\} \\ \times\ \bar{\mathsf{z}}_k(X_m) \mathcal{E}_b(\widehat{b}; O_m) \end{array} \right].$$

□

To proceed further, we also need the following elementary lemma:

**Lemma 4.** *Given $j$ nonnegative integers $\ell_0, \cdots, \ell_{j-1} \in \mathbb{Z}_{\geq 0}$ and let $\bar{\ell}_h := \sum_{s=0}^{h} \ell_s$ for $h = 0, \cdots, j-1$ with $\bar{\ell}_{-1} \equiv 0$,*

$$\begin{aligned} &\mathbb{E}\left[ \mathcal{E}_a(\widehat{a}; O_{m-1}) \bar{\mathsf{z}}_k(X_{m-1})^\top \prod_{s=0}^{j-1} \left\{ Q_s (\mathbb{I} - \widehat{\Sigma})^{\ell_s} \right\} \bar{\mathsf{z}}_k(X_m) \mathcal{E}_b(\widehat{b}; O_m) \right] \\ &= n^{-\bar{\ell}_{j-1}} \sum_{i_1=1}^{n} \cdots \sum_{i_{\bar{\ell}_{j-1}}=1}^{n} \mathbb{E}\left[ \mathcal{E}_a(\widehat{a}; O_{m-1}) \bar{\mathsf{z}}_k(X_{m-1})^\top \left\{ \prod_{s=0}^{j-1} Q_s \prod_{h=\bar{\ell}_{s-1}+1}^{\bar{\ell}_s} (\mathbb{I} - Q_{i_h}) \right\} \bar{\mathsf{z}}_k(X_m) \mathcal{E}_b(\widehat{b}; O_m) \right]. \end{aligned} \tag{51}$$

*Proof.* Without essential loss of generality, we take $\mathcal{E}_a(\widehat{a}; O) \equiv A, \mathcal{E}_b(\widehat{b}; O) \equiv 1$ to simplify the exposition. Repeatedly invoking the identity $\mathbb{I} - \widehat{\Sigma} \equiv n^{-1} \sum_{i=1}^{n} (\mathbb{I} - Q_i)$, together with the convention $\prod_{h=\ell+1}^{\ell} (\cdot)_h \equiv \mathbb{I}$ for any nonnegative integer $\ell$, we have

$$\begin{aligned} &\mathbb{E}\left[ A_{m-1} \bar{\mathsf{z}}_k(X_{m-1})^\top \prod_{s=0}^{j-1} \left\{ Q_s (\mathbb{I} - \widehat{\Sigma})^{\ell_s} \right\} \bar{\mathsf{z}}_k(X_m) \right] \\ &= n^{-\bar{\ell}_{j-1}} \mathbb{E}\left[ A_{m-1} \bar{\mathsf{z}}_k(X_{m-1})^\top \left\{ \prod_{s=0}^{j-1} Q_s \prod_{h=1}^{\ell_s} \left[ \sum_{i_{s,h}=0}^{n} (\mathbb{I} - Q_{i_{s,\ell_s}}) \right] \right\} \bar{\mathsf{z}}_k(X_m) \right] \\ &= n^{-\bar{\ell}_{j-1}} \sum_{i_1=1}^{n} \cdots \sum_{i_{\bar{\ell}_{j-1}}=1}^{n} \mathbb{E}\left[ A_{m-1} \bar{\mathsf{z}}_k(X_{m-1})^\top \left\{ \prod_{s=0}^{j-1} Q_s \prod_{h=\bar{\ell}_{s-1}+1}^{\bar{\ell}_s} (\mathbb{I} - Q_{i_h}) \right\} \bar{\mathsf{z}}_k(X_m) \right]. \end{aligned}$$

□

With the above preparatory steps, the following "cancellation lemma" is the first key milestone towards completing the proof.

**Lemma 5.** *For copy numbers satisfying* $\mathfrak{c} < \lceil (m-1)/2 \rceil$,

$$\text{Equation (50)} \equiv 0.$$

*Proof.* Again, without loss of generality, we take $\mathcal{E}_a(\widehat{a}; O) \equiv A, \mathcal{E}_b(\widehat{b}; O) \equiv 1$. Aided by Lemma 4, the summand in Equation (50) at any given $\mathfrak{c}$ can be rewritten as

$$\sum_{j=1}^{m-1} (-1)^j \binom{m-1}{j} \sum_{\mathfrak{c}'=1}^{\mathfrak{c} \wedge j} \sum_{\substack{S \subseteq [j-1], \\ |S| = \mathfrak{c}'}} \sum_{\substack{\{\ell_s, s \in S\} \subseteq \{1, \cdots, J\}^{\mathfrak{c}'}, \\ \sum_{s \in S} \ell_s = \mathfrak{c}}} \mathbb{E}\left[A_{m-1} \bar{\mathsf{z}}_k(X_{m-1})^\top \prod_{s=0}^{j-1} \left\{Q_s (\mathbb{I} - \widehat{\Sigma})^{\ell_s \mathbb{1}\{s \in S\}}\right\} \bar{\mathsf{z}}_k(X_m)\right] \tag{52}$$

$$= n^{-\mathfrak{c}} \sum_{i_1, \cdots, i_{\mathfrak{c}} = 1}^{n} \sum_{j=1}^{m-1} \binom{m-1}{j} \sum_{\mathfrak{c}'=1}^{\mathfrak{c} \wedge j} \sum_{\substack{S \subseteq [j-1], \\ |S| = \mathfrak{c}'}} \sum_{\substack{\{\ell_s, s \in S\} \subseteq \{1, \cdots, J\}^{\mathfrak{c}'}, \\ \sum_{s \in S} \ell_s = \mathfrak{c}}} \mathbb{E}\left[\begin{array}{c} A_{m-1} \bar{\mathsf{z}}_k(X_{m-1})^\top \\ \times \prod_{s=0}^{j-1} (-\,Q_s) \prod_{h=\bar{\ell}_{s-1}+1}^{\bar{\ell}_s} (\mathbb{I} - Q_{i_h}) \\ \times\ \bar{\mathsf{z}}_k(X_m) \end{array}\right]$$

$$= n^{-\mathfrak{c}} \sum_{i_1, \cdots, i_{\mathfrak{c}} = 1}^{n} \sum_{j=1}^{m-1} \binom{m-1}{j} \sum_{\mathfrak{c}'=1}^{\mathfrak{c} \wedge j} \sum_{\substack{S \subseteq [j-1], \\ |S| = \mathfrak{c}'}} \sum_{\substack{\{\ell_s, s \in S\} \subseteq \{1, \cdots, J\}^{\mathfrak{c}'}, \\ \sum_{s \in S} \ell_s = \mathfrak{c}}} \mathbb{E}\left[\begin{array}{c} A_{m-1} \bar{\mathsf{z}}_k(X_{m-1})^\top \\ \times \prod_{s=0}^{j-1} \left\{\begin{array}{c} (\mathbb{I} - Q_s) \prod_{h=\bar{\ell}_{s-1}+1}^{\bar{\ell}_s} (\mathbb{I} - Q_{i_h}) \\ - \prod_{h=\bar{\ell}_{s-1}+1}^{\bar{\ell}_s} (\mathbb{I} - Q_{i_h}) \end{array}\right\} \\ \times\ \bar{\mathsf{z}}_k(X_m) \end{array}\right]. \tag{53}$$

Now we introduce another auxiliary copy number $\mathfrak{c}^\dagger$, collecting all the terms in the above display with $\mathfrak{c}^\dagger$ many $\mathbb{I} - \mathbb{Q}_s$'s after expanding the following product

$$\prod_{s=0}^{j-1} \left\{(\mathbb{I} - Q_s) \prod_{h=\bar{\ell}_{s-1}+1}^{\bar{\ell}_s} (\mathbb{I} - Q_{i_h}) - \prod_{h=\bar{\ell}_{s-1}+1}^{\bar{\ell}_s} (\mathbb{I} - Q_{i_h})\right\} \tag{54}$$

within the expectation of (53). Let $\mathsf{p}(n, k)$ be the number of all possible partitions of $n$ into $k$ positive integers. Upon expanding, for any given $\mathfrak{c}^\dagger$, the expectations are all of the following form:

$$\mathbb{E}\left[A_{m-1} \bar{\mathsf{z}}_k(X_{m-1})^\top \prod_{s \in S \subseteq [j-1], |S| = \mathfrak{c}^\dagger} (\mathbb{I} - Q_s) \prod_{l=1}^{\mathfrak{c}} (\mathbb{I} - Q_{i_l}) \bar{\mathsf{z}}_k(X_m)\right] \tag{55}$$

up to permuting the orders of different $(\mathbb{I} - Q_s)$ and $(\mathbb{I} - Q_{i_l})$. Hence the coefficient constant of the corresponding expectation (again, up to permutations) is

$$\sum_{j=1}^{m-1} \binom{m-1}{j} \left\{\sum_{\ell=1}^{\mathfrak{c} \wedge j} \binom{j}{\ell} \mathsf{p}(\mathfrak{c}, \ell)\right\} (-1)^{j - \mathfrak{c}^\dagger} \binom{j}{\mathfrak{c}^\dagger}$$

$$
\begin{aligned}
&= (-1)^{-\mathfrak{c}^\dagger} \sum_{j=0}^{m-1} (-1)^j \binom{m-1}{j} \binom{j+\mathfrak{c}-1}{j-1} \binom{j}{\mathfrak{c}^\dagger} \\
&= \frac{(-1)^{-\mathfrak{c}^\dagger}}{\mathfrak{c}^\dagger! \mathfrak{c}!} \sum_{j=0}^{m-1} (-1)^j \binom{m-1}{j} j \prod_{\ell=-\mathfrak{c}^\dagger+1}^{\mathfrak{c}-1} (j+\ell)
\end{aligned} \tag{56}
$$

where the first equality follows from Lemma S7. The coefficient constant is simply counting the number of terms after expanding (53): in the first line of the above display, $\binom{m-1}{j}$ comes from $\binom{m-1}{j}$ of (53), $\left\{\sum_{\mathfrak{c}'=1}^{\mathfrak{c}\wedge j} \binom{j}{\mathfrak{c}'} \mathsf{p}(\mathfrak{c}, \mathfrak{c}')\right\}$ arises from the three summations after $\binom{m-1}{j}$ of (53), and $(-1)^{j-\mathfrak{c}^\dagger} \binom{j}{\mathfrak{c}^\dagger}$ counts the number of terms with $\mathfrak{c}^\dagger$ many $(\mathbb{I} - Q_s)$'s, for $s \in \{0, 1, \cdots, j-1\}$, upon expanding (54).

Another key observation is that after expansion, the expectation of (53) is identically zero when $\mathfrak{c}^\dagger > \mathfrak{c}$, leading to zero summands regardless of its coefficient constant. By virtue of this observation, we only need to consider the case when $\mathfrak{c}^\dagger \leq \mathfrak{c}$. It takes elementary calculations to show there exists integers $\gamma_1, \cdots, \gamma_{\mathfrak{c}+\mathfrak{c}^\dagger-1}$ such that

$$
(56) = \frac{(-1)^{-\mathfrak{c}^\dagger}}{\mathfrak{c}^\dagger! \mathfrak{c}!} \sum_{j=0}^{m-1} (-1)^j \binom{m-1}{j} \left( j^{\mathfrak{c}+\mathfrak{c}^\dagger} + \sum_{\ell=1}^{\mathfrak{c}+\mathfrak{c}^\dagger-1} \gamma_\ell j^\ell \right).
$$

Hence when

$$
m - 1 - (\mathfrak{c} + \mathfrak{c}^\dagger) > 0, \tag{57}
$$

(56) $\equiv 0$ by differentiating the binomial identity as in Lemma S8. Since we have assumed that $\mathfrak{c} \geq \mathfrak{c}^\dagger$, $\mathfrak{c} < \lceil (m-1)/2 \rceil$ suffices for (57) to hold. This concludes the proof. $\square$

Following Lemma 5, the next important observation wraps up the proof.

**Lemma 6.** *For copy numbers satisfying $\mathfrak{c} \geq \lceil (m-1)/2 \rceil$,*

$$
|\text{Equation (50)}| \leq \left(\frac{km}{n}\right)^{\lceil \frac{\mathfrak{c}-1}{2} \rceil \vee 1}.
$$

*Proof.* This proof inherits the notations defined in the proof of Lemma 5. We consider the case $\mathfrak{c} \geq \lceil (m-1)/2 \rceil$, which, as shown in the previous lemma, is not identically zero. We need to count the number of non-zero expectations, which is easier to work out using the representation (52).

For any given copy number $\mathfrak{c}$, the number of non-zero expectations of the form (55) is

$$
\sum_{j=1}^{m-1} (-1)^j \binom{m-1}{j} \binom{j+\mathfrak{c}-1}{\mathfrak{c}} = (-1)^{m-1} \binom{\mathfrak{c}-1}{(\mathfrak{c}-m+1) \vee 0} \tag{58}
$$

by employing Lemma S9 in Appendix S5. When $\mathfrak{c} < m-1$, (58) $= (-1)^{m-1}$; whereas when $\mathfrak{c} \geq m-1$:

- if $m \geq 4$ and $\mathfrak{c} \geq 5$

$$|(58)| = \binom{\mathfrak{c}-1}{m-2} \leq (\mathfrak{c}-1)^{m-2} \leq m^{\mathfrak{c}};$$

- if $m = 4$ and $\mathfrak{c} = 4$,

$$|(58)| = \binom{3}{2} = 3 < m^{\mathfrak{c}};$$

- finally, if $m = 3$,

$$|(58)| = \mathfrak{c} - 1 < m^{\mathfrak{c}}.$$

The proof is completed by bounding the absolute value of these non-zero expectations of the form (55) by $k^{\mathfrak{c}}$ using Lemma S1, leading to the claim

$$|(50)| \leq \left(\frac{km}{n}\right)^{\lceil \frac{\mathfrak{c}-1}{2} \rceil \vee 1}.$$

□

**Remark 14.** It is possible to improve the upper bound $m^{\mathfrak{c}}$ of $|(58)|$. First, the upper bound for the binomial coefficient used here is not sharp. Second, not all the terms counted in (58) (i.e. the term $\binom{j+\mathfrak{c}-1}{\mathfrak{c}}$) are nonzero. We decide not to pursue an improvement over $m^{\mathfrak{c}}$ for aesthetic purpose. ■

Finally, combining the above results, we have the desired kernel estimation bias bound given in Theorem 1.