Doubly-robust inference and optimality in structure-agnostic models with smoothness

Matteo Bonvini* Edward H. Kennedy† Oliver Dukes‡ Sivaraman Balakrishnan§

May 15, 2024

Preliminary. Comments welcome.

Abstract

We study the problem of constructing an estimator of the average treatment effect (ATE) that exhibits doubly-robust asymptotic linearity (DRAL). This is a stronger requirement than doubly-robust consistency. A DRAL estimator can yield asymptotically valid Wald-type confidence intervals even when the propensity score or the outcome model is inconsistently estimated. On the contrary, the celebrated doubly-robust, augmented-IPW (AIPW) estimator generally requires consistent estimation of both nuisance functions for standard root-n inference. We make three main contributions. First, we propose a new hybrid class of distributions that consists of the structure-agnostic class introduced in Balakrishnan et al (2023) with additional smoothness constraints. While DRAL is generally not possible in the pure structure-agnostic class, we show that it can be attained in the new hybrid one. Second, we calculate minimax lower bounds for estimating the ATE in the new class, as well as in the pure structure-agnostic one. Third, building upon the literature on doubly-robust inference (van der Laan, 2014, Benkeser et al, 2017, Dukes et al 2021), we propose a new estimator of the ATE that enjoys DRAL. Under certain conditions, we show that its rate of convergence in the new class can be much faster than that achieved by the AIPW estimator and, in particular, matches the minimax lower bound rate, thereby establishing its optimality. Finally, we clarify the connection between DRAL estimators and those based on higher-order influence functions (Robins et al, 2017) and complement our theoretical findings with simulations.

1 Introduction

The effect of a binary random variable A on an outcome Y is often measured by the average treatment effect (ATE). Letting Y^a denote the potential outcome that one would observe had treatment been set to $A = a$, the ATE is defined as $\mathbb{E}(Y^1 - Y^0)$. Suppose a sufficiently rich set of covariates $X \in \mathbb{R}^p$ is collected. Under consistency ($A = a \implies Y^a = Y$), positivity ($\mathbb{P}(A = 1 \mid X) \in (0, 1)$) and no-unmeasured-confounding ($A \perp\!\!\!\perp Y^a \mid X$), we have

$$\mathbb{E}(Y^1 - Y^0) = \mathbb{E}\{\mathbb{E}(Y \mid A = 1, X) - \mathbb{E}(Y \mid A = 0, X)\}.$$

In this work, we consider the problem of estimating $\mathbb{E}(Y^1 - Y^0)$ identified as above. For simplicity, however, we focus on $\psi \equiv \mathbb{E}\{\mathbb{E}(Y \mid A = 1, X)\}$, with the understanding that the same results apply

*Department of Statistics, Rutgers, the State University of New Jersey. Email: mb1662@stat.rutgers.edu.
†Department of Statistics & Data Science, Carnegie Mellon University. Email: edward@stat.cmu.edu.
‡Department of Applied Mathematics, Computer Science and Statistics, Ghent University. Email: Oliver.Dukes@ugent.be
§Department of Statistics & Data Science, Carnegie Mellon University. Email: siva@stat.cmu.edu.

to $\mathbb{E}\{\mathbb{E}(Y \mid A=0,X)\}$. The parameter ψ captures the average outcome if every unit in the population takes $A=1$; it can also be interpreted as the population mean outcome under missingness at random (Rubin 1976). Here, we consider estimation of ψ regardless of its interpretation.

One key feature of ψ is that it can be estimated at $n^{-1/2}$ rates even in nonparametric models where the best rate of convergence for estimating $b(X)=\mathbb{E}(Y \mid A=1,X)$ is slower than $n^{-1/2}$. One way to see this is to consider a randomized trial whereby the probability of receiving treatment given the measured covariates, $\mathbb{P}(A=1 \mid X)$, is known. Because ψ can be expressed as $\psi=\mathbb{E}[AYa(X)]$, for $a(X)=\{\mathbb{P}(A=1 \mid X)\}^{-1}$, by Chebyshev's inequality, it can be estimated at $n^{-1/2}$ rates simply as $n^{-1}\sum_{i=1}^{n} A_iY_ia(X_i)$. In non-randomized studies, however, both nuisance functions $a(X)$ and $b(X)$ are unknown. In this sense, the convergence rate of an estimator of ψ will typically depend on how accurately these nuisance parameters can be estimated. A vast literature has thus focused on weakening the dependence of the estimator on the nuisance functions' estimation error.

In the context of nonparametric modeling, many widely adopted estimators of ψ rely, in some form, on estimation of both $b(X)$ and $a(X)$. The augmented-inverse-probability-weighted (AIPW) estimator (Robins *et al.* 1994) and those based on Targeted-Maximum-Likelihood Estimation (TMLE) (Van der Laan & Rose 2011, 2018) are two prominent examples. In particular, these two approaches, based on the first-order influence function of ψ, are agnostic with respect to how the nuisances are estimated. Other approaches consider more specific, clever estimators of these nuisances, but can still be considered as variants of the two approaches above. Finally, the list of available estimators of ψ also includes other strategies, such as those based on matching (see, e.g., Imbens 2004).

In non-randomized studies, where both $b(X)$ and $a(X)$ are unknown, the state-of-the-art to conduct inference on ψ is to assume that both nuisances are estimated well enough. A standard requirement is that the product of the root-mean-square-errors is asymptotically negligible, i.e., it converges in probability to 0 when scaled by $\sqrt{n}$. In this case, an asymptotically valid confidence interval is simply the Wald interval.

When it is not possible to estimate the nuisances with the accuracy required for the validity of the Wald interval, one option is to attempt to estimate and take into account the bias of the original estimator. This idea is connected to the development of the general theory of Higher Order Influence Functions (HOIFs) (Robins *et al.* 2017a, 2008, 2009a,b; van der Laan *et al.* 2018), as well as the discovery of estimators that can be $\sqrt{n}$-consistent even if the model for $b(X)$ or $a(X)$ is misspecified (Benkeser *et al.* 2017; Dukes *et al.* 2021; Van der Laan 2014). In this work, we build upon these two streams of literature and propose a novel estimator of ψ, which remains $\sqrt{n}$-consistent even if one of the two nuisances is not consistently estimated. A more detailed list of our contributions, including minimax lower bounds, appears in Section 1.3, after introducing notation and the basic problem statement.

1.1 Notation

We assume that one observes n iid copies $O^n=O_1,\ldots,O_n$, where $O=(Y,A,X)\in\mathcal{Y}\times\{0,1\}\times\mathbb{R}^d$. Let $f(x)$ denote the density of X and

$$\pi(X)=\mathbb{P}(A=1 \mid X),\quad a(X)=1/\pi(X),\quad b(X)=\mathbb{E}(Y \mid A=1,X),\quad \text{and}\quad g(X)=\pi(X)f(X).$$

The parametrization of the density in terms of $a(x)$ instead of $\pi(x)$ is natural and convenient when deriving the lower bounds results in Section 2, but it is not essential for deriving the properties of

the estimators described in Section 3. With this notation, we can write

$$\psi = \mathbb{E}\{b(X)\} = \mathbb{E}\left\{\frac{AY}{\pi(X)}\right\} = \mathbb{E}\left\{a(X)AY\right\} = \int a(x)b(x)g(x)dx.$$

Let us use the notation $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(O_i)$, $\mathbb{P}f = \int f(o)d\mathbb{P}(o)$ and $\|f\|^2 = \mathbb{P}f^2$. We also let $a \wedge b$ denote $\min(a,b)$, $a \vee b$ denote $\max(a,b)$ and $a \lesssim b$ denote $a \leq Cb$ for some constant C that is independent of the sample size n. To lighten the notation, when it does not induce confusion, we will also use $a = a(X)$, $b = b(X)$, $a_i = a(X_i)$ and $b_i = b(X_i)$. A similar notation will be used for the estimators $\widehat{a}(x)$ and $\widehat{b}(x)$. We also use the notation $\mathbb{E}_{X_1|X_2}\{f(X_1)\} = \mathbb{E}\{f(X_1) \mid X_2\}$. Throughout, we assume that $\widehat{a}(x)$ and $\widehat{b}(x)$ are computed on a sample D^n that is independent of O^n. This can be accomplished by splitting O^n is subsamples and then swapping the roles of the samples for training the nuisance estimators. Further, we assume that all observations and nuisance functions are bounded; in particular, we assume that $\widehat{a}(x), a(x) \in [1, M_a]$, $|b(x)| \leq M_b$ and $|\widehat{b}(x)| \leq M_b$, for some constants M_a and M_b.

Finally, recall the definition of a Hölder smooth function. We consider this function class when we introduce our estimators in Sections 3.2 and 3.3.

Definition 1. Let $\alpha \in [0,1]$ and C be a positive constant. A function $f(x)$ is Hölder of order α if

$$|f(x) - f(y)| \leq C\|x - y\|^\alpha \text{ for every } x, y \text{ in the domain of } f.$$

Because our subsequent results only pertain to the vanishing rate of the mean-square-errors of our estimators as a function of the sample size, we will not keep track of constants. In this light, we will say that a function is smooth of order α if it satisfies Definition 1 for some constant C. Finally, letting $\lfloor\alpha\rfloor$ denote the greatest integer less than α, we say a function f is Hölder of order $\alpha > 1$ if f is $\lfloor\alpha\rfloor$-times differentiable with $\lfloor\alpha\rfloor^{\text{th}}$ derivative Hölder of order $\alpha - \lfloor\alpha\rfloor$ in the sense of Definition 1, and if f has all derivatives up to order $\lfloor\alpha\rfloor$ bounded above by some constant.

1.2 Problem statement

The AIPW estimator, also known as the doubly-robust (DR) estimator, is defined as

$$\widehat{\psi}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^n A_i\widehat{a}(X_i)\{Y_i - \widehat{b}(X_i)\} + \widehat{b}(X_i) \equiv \mathbb{P}_n\widehat{\varphi}, \tag{1}$$

where $\varphi(O) = Aa(X)\{Y - b(X)\} + b(X)$ is the (uncentered) influence function of ψ. The variance $\text{var}(\varphi)$ is the semiparametric efficiency bound for estimating ψ, i.e. the smallest variance any regular estimator of ψ can achieve (Kennedy 2022; Newey 1990; Tsiatis 2006). Under certain conditions, $\widehat{\psi}_{\text{DR}}$ achieves this semiparametric bound and it is thus efficient. This can be seen from the following decomposition. Let $\overline{a}$ and $\overline{b}$ denote the limits, as $n \to \infty$, of $\widehat{a}$ and $\widehat{b}$. In this sense, let us assume, without essential loss of generality, that $\|\widehat{a} - \overline{a}\| = o_{\mathbb{P}}(1)$ and $\|\widehat{b} - \overline{b}\| = o_{\mathbb{P}}(1)$. Letting $\overline{\varphi}(O) = A\overline{a}(X)\{Y - \overline{b}(X)\} + \overline{b}(X)$, we have

$$\widehat{\psi}_{\text{DR}} - \psi = (\mathbb{P}_n - \mathbb{P})(\widehat{\varphi} - \overline{\varphi}) + (\mathbb{P}_n - \mathbb{P})\overline{\varphi} + \mathbb{P}(\widehat{\varphi} - \varphi).$$

The central term, when scaled by $\sqrt{n}$, converges to $N(0, \text{var}(\overline{\varphi}))$ by the central limit theorem. Thus, by Slutsky's theorem, $\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N(0, \text{var}(\overline{\varphi}))$ as long as the first and last terms on the right-hand-side are $o_{\mathbb{P}}(n^{-1/2})$. Notice that $\text{var}(\overline{\varphi}) = \text{var}(\varphi)$ if $\overline{a} = a$ and $\overline{b} = b$. Next, in light of

Lemma 2 in Kennedy *et al.* 2020 (restated in the Appendix as Lemma 4) and the fact that $\widehat{a}$ and $\widehat{b}$ are trained on a separate sample, the first term is $o_{\mathbb{P}}(n^{-1/2})$. The most difficult term to control is the third one, which equals

$$R_n \equiv \mathbb{P}(\widehat{\varphi} - \varphi) = \int \{\widehat{a}(x) - a(x)\}\{b(x) - \widehat{b}(x)\}g(x)dx. \tag{2}$$

In this light, we rewrite the decomposition $\widehat{\psi}_{\text{DR}} - \psi$ as

$$\widehat{\psi}_{\text{DR}} - \psi = (\mathbb{P}_n - \mathbb{P})\overline{\varphi} + R_n + o_{\mathbb{P}}(n^{-1/2}). \tag{3}$$

For inference, we assume throughout that equation (3) holds. The quantity R_n is the conditional bias of $\widehat{\psi}_{\text{DR}}$ given the training sample D^n. By the Cauchy-Schwarz inequality,

$$|R_n| \lesssim \|\widehat{a} - a\|\|\widehat{b} - b\| \implies \widehat{\psi}_{\text{DR}} - \psi = O_{\mathbb{P}}\left(n^{-1/2} \vee \mathbb{E}\{\|\widehat{a} - a\|\|\widehat{b} - b\|\}\right).$$

Thus, in general, asymptotic negligibility of this term is guaranteed if

$$\mathbb{E}\{\|\widehat{a} - a\|\|\widehat{b} - b\|\} \leq \{\mathbb{E}(\|\widehat{a} - a\|^2)\}^{1/2} \cdot \{\mathbb{E}(\|\widehat{b} - b\|^2)\}^{1/2} = o(n^{-1/2}),$$

leading to the standard $n^{-1/4}$-rate requirement on the nuisance root-mean-square-errors. This is remarkable because $n^{-1/4}$-rates are achievable in nonparametric function classes under appropriate structural conditions, e.g. sufficient smoothness or sparsity. For example, if a and b are Hölder smooth of order s and are estimated by rate-optimal estimators (see, e.g., Section 1.6 in Tsybakov 2008 for a discussion on local polynomials), then $\mathbb{E}\{\|\widehat{a} - a\|\|\widehat{b} - b\|\} = o(n^{-1/2})$ follows if $s > d/2$. More generally, for some sequences of constants ϵ_n and δ_n, the Cauchy-Schwarz bound yields that $\mathbb{E}|\widehat{\psi}_{\text{DR}} - \psi| \lesssim n^{-1/2} \vee \delta_n\epsilon_n$ whenever $\|\widehat{a} - a\| \leq \epsilon_n$ and $\|\widehat{b} - b\| \leq \delta_n$.

When the condition ensuring $\epsilon_n\delta_n = o(n^{-1/2})$ fails, the Cauchy-Schwarz bound does not guarantee that the bias of $\widehat{\psi}_{\text{DR}}$ is vanishing at a rate faster than $n^{-1/2}$, the order of the standard error. The foundational work on Higher Order Influence Functions (HOIFs) by Robins and co-authors offers a general, principled way to carry out functional estimation optimally. This approach has led to new estimators of ψ based on higher-order U-statistics, which have been shown to be optimal in certain nonparametric models (Liu *et al.* 2017; Robins *et al.* 2017b, 2008, 2009a). This general theory considerably expands the use of U-statistics for optimal functional estimation, which has a long history in statistics; see, e.g., the literature on estimating integral functionals of a density (Bickel & Ritov 1988; Birgé & Massart 1995; Laurent 1996, 1997). With respect to the settings considered here, higher order corrections for estimating ψ can be viewed as effectively estimating R_n and subtract it off from $\widehat{\psi}_{\text{DR}}$, leading to better bias-variance trade-offs. Higher order corrections can also be done in the context of TMLE (Diaz *et al.* 2016; van der Laan *et al.* 2021; van der Laan *et al.* 2018). With a more direct focus on inference, Van der Laan 2014 and Benkeser *et al.* 2017 have discovered TMLE-based estimators of ψ that are $\sqrt{n}$-consistent and asymptotically normal even if one between $\widehat{a}$ or $\widehat{b}$ is not consistent. This represents an improvement upon $\widehat{\psi}_{\text{DR}}$ because, at best, the rate of convergence for $\widehat{a}$ and $\widehat{b}$ is of order $n^{-1/2}$ (corresponding to correctly specified parametric models), so that, if the Cauchy-Schwarz bound on $|R_n|$ is employed, R_n would *not* be asymptotically negligible if $\widehat{a}$ or $\widehat{b}$ are inconsistent. We remark that their estimators do not directly build on the theory of HOIFs and, in particular, do not employ U-statistics.

1.3 Main contributions

In this work, we make at least three main contributions. First, we propose and develop a new function class, which is data-dependent and can be viewed as a hybrid between the completely structure-agnostic ones considered in Balakrishnan *et al.* 2023 and more traditional ones based on smoothness conditions. The structure-agnostic class of distributions considered in Balakrishnan *et al.* 2023 is defined as the set of all distributions $\mathcal{P}(\epsilon_n, \delta_n)$ for which it holds that $\|\widehat{a} - a\| \leq \epsilon_n$ and $\|\widehat{b} - b\| \leq \delta_n$, for some rates ϵ_n and δ_n. We study a subset of this class that additionally impose some smoothness constraints on certain regression functions for which the estimators $\widehat{a}(X)$ and $\widehat{b}(X)$ enter as covariates. We find that the convergence rates admitted over this more regular class can be potentially much faster than those holding over the completely structure-agnostic ones studied in Balakrishnan *et al.* 2023. As our proposed function class does not directly impose regularity restrictions on a and b and yet can potentially admit fast rates of convergence, it can provide a nice middle ground between complete agnosticism at one extreme and much more structured smoothness, say encoded in Hölder smoothness restrictions on a and b, at the other.

Our second contribution is to provide minimax lower bounds for estimating ψ in the new hybrid class proposed, as well as in the pure structure agnostic one. We find that, over $\mathcal{P}(\epsilon_n, \delta_n)$, i.e., if the rate-condition for estimating a and b is the only information available (together with mild boundedness regularity conditions), the rate of convergence $n^{-1/2} \vee (\epsilon_n \delta_n)$ for ψ is not improvable in a minimax sense. This shows that, in this framework, $\widehat{\psi}_{\text{DR}}$ is not improvable without the introduction of additional assumptions. Our current construction only covers the case $\epsilon_n \leq \delta_n$. However, in a work concurrent to ours and developed independently, Jin & Syrgkanis 2024 show that the bound $\epsilon_n \delta_n$ also holds when $\epsilon_n > \delta_n$; their proof is conceptually similar to ours but employs a different parametrization of the data generating process.

On the contrary, the lower bound derived for estimating ψ in the proposed hybrid model is of order $n^{-1/2} \vee (\epsilon_n^2 \wedge \delta_n^2)$. In this sense, imposing additional smoothness constraints on $\mathcal{P}(\epsilon_n, \delta_n)$ allows for potentially much faster rates of convergence. The rate $\epsilon_n^2 \wedge \delta_n^2$ does not rule out the existence of valid confidence intervals shrinking at the rate $n^{-1/2}$ even if one of the two nuisance functions is not consistently estimated, i.e., even if one between ϵ_n or δ_n does not converge to zero. It thus allows for the possibility of conducting doubly-robust root-n inference. As mentioned above, in virtue of the lower bound rate $\epsilon_n \delta_n$, the purely structure-agnostic class of distributions studied in Balakrishnan *et al.* 2023 and Jin & Syrgkanis 2024 does not allow for doubly-robust root-n inference in nonparametric nuisance function classes.

Our third contribution is the construction and analysis of a new estimator that achieves the rate $n^{-1/2} \vee (\epsilon_n^2 \wedge \delta_n^2)$ under certain conditions. As this rate matches our minimax lower bound rate for estimating ψ in the proposed hybrid function class, this new estimator is a minimax optimal one. We view our new estimator as a hybrid between the estimator of the ATE based on the approximate second-order influence function (Diaz *et al.* 2016; Robins *et al.* 2009a) and the estimator(s) considered in Van der Laan 2014, Benkeser *et al.* 2017 and Dukes *et al.* 2021 that are specifically designed to conduct doubly-robust inference. Our estimator can be used for doubly-robust inference under arguably more transparent conditions than the ones previously considered in the literature, which did not posit a hybrid smoothness/structure-agnostic model like we do here. In addition, our constructions can be easily adjusted to conduct doubly-robust inference in settings where this is not readily feasible using currently available estimators, such as to estimate the parameters in a partially linear logistic model (see Appendix A). Finally, we evaluate the performance of the newly derived estimator against that of $\widehat{\psi}_{\text{DR}}$ and that of the estimator described in Benkeser *et al.* 2017

and implemented in the `R` package `drtmle` (Benkeser & Hejazi 2023). The code to replicate the simulations can be found at https://github.com/matteobonvini/dr_inference.

2 Structure-agnostic viewpoint

2.1 Purely structure-agnostic class of distributions

In this section, we describe the optimality viewpoint introduced in Balakrishnan *et al.* 2023. They consider the problem of functional estimation with nuisance parameters when all that is known are convergence rates for estimating the nuisance components. In our setting, we aim to derive the best possible rate for estimating ψ when the only information available are bounds of the form $\|\widehat{a} - a\| \leq \epsilon_n$ and $\|\widehat{b} - b\| \leq \delta_n$. This framework is particularly helpful for understanding how precisely one can estimate ψ without imposing structural assumptions on the data generating process; e.g., smoothness or sparsity on $a(x)$ and $b(x)$. In deriving the lower bounds in this section and the next one, we assume Y is binary and X is supported in $[0,1]^d$.

Given two arbitrary estimators $\widehat{a}(x)$ and $\widehat{b}(x)$, the class $\mathcal{P}(\epsilon_n, \delta_n)$ consists of all densities such that $\|\widehat{a} - a\| \leq \epsilon_n$ and $\|\widehat{b} - b\| \leq \delta_n$. Our first result is that the rate of convergence of any estimator of ψ over this class cannot be faster than $n^{-1/2} \vee \epsilon_n \cdot \delta_n$. Our current proof breaks for data generating processes where $\epsilon_n > \delta_n$. However, both cases are covered by the concurrent work by Jin & Syrgkanis 2024. The class $\mathcal{P}(\epsilon_n, \delta_n)$ aims to describe the setting where one constructs $\widehat{a}(\cdot)$ and $\widehat{b}(\cdot)$ on a separate independent training sample. The information available would then be encoded in the form of high-probability bounds for $\|\widehat{a} - a\| \leq \epsilon_n$ and $\|\widehat{b} - b\| \leq \delta_n$. Without essential loss of generality, in what follows, we assume these bounds hold exactly (not just with high probability) and refer to Section 3.1.1 in Balakrishnan *et al.* 2023 for further discussion.

Proposition 1. *Let $\mathcal{P}(\epsilon_n, \delta_n)$ denote the class of all densities such that $\sup_{p \in \mathcal{P}(\epsilon_n, \delta_n)} \|a_p - \widehat{a}\|_2 \leq \epsilon_n$ and $\sup_{p \in \mathcal{P}(\epsilon_n, \delta_n)} \|b_p - \widehat{b}\|_2 \leq \delta_n$. Then, provided that $\epsilon_n \leq \delta_n$,*

$$\inf_{T_n} \sup_{p \in \mathcal{P}(\epsilon_n, \delta_n)} \mathbb{E}|T_n - \psi_p| \gtrsim \epsilon_n \delta_n.$$

Proof. See Jin & Syrgkanis 2024 for a full proof that also covers the case when $\epsilon_n > \delta_n$. Our proof relying on $\epsilon_n \leq \delta_n$ is reported in Appendix B.2. □

Proposition 1 establishes that, if the only assumption on the data generating process consists of error bounds for estimating a and b, then ψ cannot be estimated at a rate faster than the product of these bounds. This means that, to improve upon the AIPW estimator $\widehat{\psi}_{\text{DR}}$, which achieves this rate under mild conditions, one needs to introduce other assumptions in addition to rate conditions on the nuisance components. Furthermore, if either a or b is inconsistently estimated, so that either ϵ_n or δ_n do not vanish as n goes to infinity, then ψ cannot be estimated at a rate faster than $\epsilon_n \wedge \delta_n$, without introducing other assumptions. This implies that nonparametric, doubly-robust root-n inference is possible only if one relies on additional conditions. Note that this is meant as a clarifying technical statement; we do not claim that relying on such additional conditions should necessarily be avoided in practice.

We conclude this section with a remark highlighting a connection between the result from Proposition 1 and the seminal results on optimal estimation of functionals indexed by Hölder nuisance components.

Remark 1. Suppose that a and b are Hölder smooth functions on a d-dimensional domain of orders α and β. Then, there exist rate-optimal estimators in these classes that satisfy, with high probability, $\|\widehat{a} - a\| \lesssim n^{-\alpha/(2\alpha+d)}$ and $\|\widehat{b} - b\| \lesssim n^{-\beta/(2\beta+d)}$, i.e., $\epsilon_n = C_a n^{-\alpha/(2\alpha+d)}$ and $\delta_n = C_b n^{-\beta/(2\beta+d)}$ for some constants C_a and C_b. Suppose that the good event $\|\widehat{a} - a\| \leq \epsilon_n$ and $\|\widehat{b} - b\| \leq \delta_n$ holds. Proposition 1 (in the case of $\alpha \geq \beta$) or more generally Theorem 2.1 in Jin & Syrgkanis 2024 yield that no estimator of ψ can achieve a rate faster than $n^{-1/2} \vee \epsilon_n \delta_n$. This rate is slower than the minimax rate for estimating ψ in this model, which is of the order $n^{-1/2} \vee n^{-4s/4s+d}$, where $s = (\alpha + \beta)/2$ (Robins *et al.* 2008, 2009b). This does not contradict Proposition 1: the class $\mathcal{P}(n^{-\alpha/(2\alpha+d)}, n^{-\beta/(2\beta+d)})$ is larger than the class of densities such that a and b are α- and β-Hölder smooth. In fact, the worst-case construction used to prove Proposition 1 relies on nuisance functions that are not necessarily Hölder smooth. In this sense, Proposition 1 suggests that imposing smoothness assumptions on a and b induces regularity on ψ that is in addition to the regularity induced on a and b. This then results in a model where $\widehat{\psi}_{\mathrm{DR}}$ is no longer optimal and one has to rely on additional corrections for optimal estimation, such as those based on HOIFs. This intuition aligns with the result that in a Hölder model where $\alpha + \beta > d/2$, ψ can be estimated at $n^{-1/2}$ rates without the need for consistent estimators of a and b. However, if the nuisance functions' estimators are consistent, then the HOIFs-based estimator is also semiparametric efficient in the sense of having the smallest asymptotic variance among all regular estimators. See Liu *et al.* 2017, particularly Corollary 5 and Remark 7.

2.2 Hybrid structure-agnostic class of distributions with smoothness

Next, we introduce our main hybrid class, as well as two other hybrid models that are tailored to settings where it known whether a or b is easier to estimate. In defining these classes, we depart from the structure agnostic framework introduced in Balakrishnan *et al.* 2023, encoded in $\mathcal{P}(\epsilon_n, \delta_n)$, to introduce certain smoothness conditions. In particular, we consider $\mathcal{P}_a(\epsilon_n)$ denoting the collection of all densities such that $\|\widehat{a} - a\| \leq \epsilon_n$ and $\mathbb{E}\{b(X) \mid A = 1, \widehat{a}(X) = t_1, a(X) = t_2, D^n\}$ is smooth. The reason why this class can be of interest in the context of estimating ψ is that one can write R_n as

$$\begin{aligned} R_n &= \mathbb{E}\left[A\{\widehat{a}(X) - a(X)\}\{b(X) - \widehat{b}(X)\} \mid D^n\right] \\ &= \mathbb{E}\left[A\{\widehat{a}(X) - a(X)\}\mathbb{E}\{b(X) \mid A = 1, \widehat{a}(X), a(X), D^n\} \mid D^n\right] - \mathbb{E}\left[A\{\widehat{a}(X) - a(X)\}\widehat{b}(X) \mid D^n\right] \\ &= \mathbb{E}\left[\{A\widehat{a}(X) - 1\}\mathbb{E}\{b(X) \mid A = 1, \widehat{a}(X), a(X), D^n\} \mid D^n\right] - \mathbb{E}\left[\{A\widehat{a}(X) - 1\}\widehat{b}(X) \mid D^n\right]. \end{aligned}$$

If $\mathbb{E}\{b(X) \mid A = 1, \widehat{a}(X), a(X), D^n\}$ was known, R_n could be estimated by a sample average with accuracy of order $n^{-1/2}$. The hope is that if this additional nuisance function is unknown, as it would be in practice, but smooth enough so that it can be estimated well, one may still estimate ψ with $n^{-1/2}$-accuracy. A more structure-agnostic way to define $\mathcal{P}_a(\epsilon_n)$ would be to simply impose a rate condition on the accuracy with which $\mathbb{E}\{b(X) \mid A = 1, \widehat{a}(X), a(X), D^n\}$ can be estimated, as opposed to assuming this function is smooth. We leave this refinement for future work.

Motivated by writing R_n as

$$R_n = \mathbb{E}\left[A\{Y - \widehat{b}(X)\}\widehat{a}(X) \mid D^n\right] - \mathbb{E}\left[A\{Y - \widehat{b}(X)\}\mathbb{E}\{a(X) \mid A = 1, \widehat{b}(X), b(X), D^n\} \mid D^n\right],$$

we also consider the class of densities $\mathcal{P}_b(\delta_n)$ such that $\|\widehat{b} - b\| \leq \delta_n$ and $\mathbb{E}\{a(X) \mid A = 1, \widehat{b}(X) = t_1, b(X) = t_2, D^n\}$ is smooth.

Finally, we consider $\mathcal{P}_{ab}(\epsilon_n, \delta_n)$, the main hybrid class that we propose. It restricts $\mathcal{P}(\epsilon_n, \delta_n)$ to include only densities for which $\mathbb{E}\{a(X) \mid A = 1, \widehat{b}(X) = t_1, b(X) = t_2, \widehat{a}(X) = t_3, D^n\}$ and $\mathbb{E}\{b(X) \mid A = 1, \widehat{a}(X) = t_1, a(X) = t_2, \widehat{b}(X) = t_3, D^n\}$ are smooth. In the following proposition, we derive a minimax lower bound for each of the three hybrid classes considered.

Proposition 2. *We consider three cases:*

1. *Let $\mathcal{P}_a(\epsilon_n)$ denote the class of all densities such that $\sup_{p \in \mathcal{P}_a(\epsilon_n)} \|a_p - \widehat{a}\| \leq \epsilon_n$ and $\mathbb{E}\{b(X) \mid A = 1, a(X) = t_1, \widehat{a}(X) = t_2, D^n\}$ is infinitely smooth. Then,*

$$\inf_{T_n} \sup_{p \in \mathcal{P}_a(\epsilon_n)} \mathbb{E}|T_n - \psi_p| \gtrsim \epsilon_n^2.$$

2. *Let $\mathcal{P}_b(\delta_n)$ denote the class of all densities such that $\sup_{p \in \mathcal{P}_b(\delta_n)} \|b_p - \widehat{b}\| \leq \delta_n$ and $\mathbb{E}\{a(X) \mid A = 1, \widehat{b}(X) = t_1, b(X) = t_2, D^n\}$ is infinitely smooth. Then,*

$$\inf_{T_n} \sup_{p \in \mathcal{P}_b(\delta_n)} \mathbb{E}|T_n - \psi_p| \gtrsim \delta_n^2.$$

3. *Let $\mathcal{P}_{ab}(\epsilon_n, \delta_n)$ be the class of all densities such that 1) $\sup_{p \in \mathcal{P}_{ab}} \|a_p - \widehat{a}\| \leq \epsilon_n$ and $\sup_{p \in \mathcal{P}_{ab}} \|b_p - \widehat{b}\| \leq \delta_n$, 2) $\mathbb{E}\{b(X) \mid A = 1, \widehat{a}(X) = t_1, a(X) = t_2, \widehat{b}(X) = t_3; D^n\}$ and $\mathbb{E}\{a(X) \mid A = 1, \widehat{b}(X) = t_1, b(X) = t_2, \widehat{a}(X) = t_3; D^n\}$ are infinitely smooth. Then,*

$$\inf_{T_n} \sup_{p \in \mathcal{P}_{ab}(\epsilon_n, \delta_n)} \mathbb{E}|T_n - \psi_p| \gtrsim \delta_n^2 \wedge \epsilon_n^2.$$

As exemplified in the third claim, the smaller class $\mathcal{P}_{ab}(\epsilon_n, \delta_n)$ is an example of a collection of densities for which our lower bound allows for the possibility of doubly-robust, root-n inference. In fact, $\delta_n^2 \wedge \epsilon_n^2 = o(n^{-1/2})$ if either $\|\widehat{a} - a\| = o(n^{-1/4})$ or $\|\widehat{b} - b\| = o(n^{-1/4})$. In Section 3.3, we construct an estimator achieving this rate, under certain conditions. This estimator can then be used for conducting doubly-robust inference via a standard Wald confidence interval.

Remark 2. All the lower bounds from Propositions 1 and 2 can be strengthened by taking the maximum between the rates shown and the parametric rate $n^{-1/2}$. Even if the nuisance functions $a(x)$ and $b(x)$ were known exactly, one would not typically be able to estimate ψ are a rate faster than $n^{-1/2}$. The resulting lower bounds can be derived by a standard argument; see, for example, Section B.3 (Case 1) in Balakrishnan *et al.* 2023.

3 New estimators of ψ

3.1 Preliminaries and overview

In this section, we consider three new estimators of ψ and derive upper bounds on their risk. Each estimator will be written as the doubly-robust estimator $\widehat{\psi}_{\text{DR}}$ (1) plus a term T_n taking a different form depending on the assumptions invoked. That is,

$$\widehat{\psi} = \widehat{\psi}_{\text{DR}} - T_n = \mathbb{P}_n \widehat{\varphi} - T_n, \text{ where } \varphi(O) = Aa(X)\{Y - b(X)\} + b(X).$$

We view T_n as an estimator of R_n (from (2)) that stems from merging ideas from the theory of HOIFs and some observations previously made in the doubly-robust inference literature. Relative

to the HOIF-based estimators of ψ, our constructions do not directly approximate $x \mapsto a(x) - \widehat{a}(x)$ and $x \mapsto b(x) - \widehat{b}(x)$ with a dictionary of basis functions; in particular, they do not require the use of computationally expensive tensor products bases to approximate these functions. Relative to existing estimators proposed in the doubly-robust inference literature, ours are "one-step" in the sense that they do not require iterative procedures. Further, the conditions under which they can be used to construct valid Wald-type confidence interval are arguably more transparent. For example, in contrast with the estimators studied in Benkeser *et al.* 2017 and Dukes *et al.* 2021, under mild assumptions, the limiting variance of the estimators studied in this work does not depend on which nuisance function is consistently estimated.

As noted, for instance, in Van der Laan 2014 and Benkeser *et al.* 2017, and briefly discussed in Section 2, R_n can be written in different ways other than as in equation (2). In this work, we consider a slight departure from their parametrizations of R_n, which is, however, still based on their idea of considering additional nuisance functions taking the form of regressions with both estimated outcomes and covariates. Define

$$
\begin{aligned}
s_a(t_1, t_2; D^n) &= \mathbb{E}\{b(X) - \widehat{b}(X) \mid A = 1, \widehat{a}(X) = t_1, a(X) = t_2, D^n\}, \\
s_b(t_1, t_2; D^n) &= \mathbb{E}\{\widehat{a}(X) - a(X) \mid A = 1, \widehat{b}(X) = t_1, b(X) = t_2, D^n\}, \\
f_a(t_1, t_2, t_3; D^n) &= \mathbb{E}\{b(X) - \widehat{b}(X) \mid A = 1, \widehat{a}(X) = t_1, a(X) = t_2, \widehat{b}(X) = t_3, D^n\}, \\
f_b(t_1, t_2, t_3; D^n) &= \mathbb{E}\{\widehat{a}(X) - a(X) \mid A = 1, \widehat{b}(X) = t_1, b(X) = t_2, \widehat{a}(X) = t_3, D^n\}.
\end{aligned}
$$

Notice that

$$
\begin{aligned}
s_a(t_1, t_2; D^n) &= \mathbb{E}\{Y - \widehat{b}(X) \mid A = 1, \widehat{a}(X) = t_1, a(X) = t_2, D^n\}, \\
f_a(t_1, t_2, t_3; D^n) &= \mathbb{E}\{Y - \widehat{b}(X) \mid A = 1, \widehat{a}(X) = t_1, a(X) = t_2, \widehat{b}(X) = t_3, D^n\},
\end{aligned}
$$

but s_b and f_b cannot be written as regressions of observed outcomes on partially observed covariates. The functions f_a and f_b enter the definition of the model considered in Claim 3 of Proposition 2, while s_a and s_b are similar but not quite the same functions as those defining the models considered in Claims 2 and 3 of Proposition 2.

The bias R_n can be written an expectation of an observed random variable times one of the four functions above:

$$
\begin{aligned}
R_n &= \mathbb{E}\{(A\widehat{a} - 1) s_a(\widehat{a}, a; D^n) \mid D^n\} = \mathbb{E}\{(A\widehat{a} - 1) f_a(\widehat{a}, a, \widehat{b}; D^n) \mid D^n\} \\
&= \mathbb{E}\{A(Y - \widehat{b}) s_b(\widehat{b}, b; D^n) \mid D^n\} = \mathbb{E}\{A(Y - \widehat{b}) f_b(\widehat{b}, b, \widehat{a}; D^n) \mid D^n\},
\end{aligned}
$$

where, for shorthand notation, $\widehat{a} = \widehat{a}(X)$ and $\widehat{b} = \widehat{b}(X)$. If either s_a, s_b, f_a or f_b were known, then R_n could be estimated efficiently by a sample average. In Section 3.2, we derive estimators tailored to models where s_a or s_b are Hölder functions. These estimators will improve upon $\widehat{\psi}_{\text{DR}}$, and match the lower bound rates from Claims 1 and 2 in Proposition 2, when it is known whether s_a or s_b is smoother and thus easier to estimate. Without such knowledge, they would perform, in terms of mean-square-error, as well as $\widehat{\psi}_{\text{DR}}$ but not necessarily better. To remedy this, in Section 3.3, we construct an estimator that, under certain conditions, can improve upon $\widehat{\psi}_{\text{DR}}$ without the knowledge of which nuisance function is easier to estimate. Further, it is shown to achieve the lower bound rate from Claim 3 in Proposition 2. We view this estimator as a possible one-step counterpart to the TMLE ones proposed in Benkeser *et al.* 2017, although the additional nuisance functions entering the parametrization of R_n that they use are not exactly the same as ours.

All these nuisance functions depend on estimated outcomes and covariates. The problem of non-parametrically estimating a regression function when some covariates need to be estimated in a first step has been considered, for example, by Mammen *et al.* 2012, Sperlich 2009 and Dukes *et al.* 2021. In non-randomized experiments, regression adjustments based on the propensity score, e.g. via matching (Abadie & Imbens 2016) or ordinary least squares (Robins *et al.* 1992; Vansteelandt & Joffe 2014), represent instances of such a general estimation problem. To understand one of the main challenges in this problem, due to its intrinsic non-smoothness, consider estimating $\mathbb{E}(Y \mid f(X) = t)$, for some function $f(X)$, estimated by $\widehat{f}(X)$, and observable random variables Y and X. Regressing Y onto $\widehat{f}(X)$ via some local method that depends on a vanishing bandwidth h will need to ensure that h does not shrink faster that the error $\widehat{f} - f$ or else the localization would be "misplaced." This, in turns, leads to difficulties in choosing the correct order of h as well as to potentially a dramatically slow rate of convergence since one may expect the error $\widehat{f} - f$ to be inflated by multiplication by h^{-1} (see, e.g., Theorem 1 in Mammen *et al.* 2012 and Theorem 3 in Dukes *et al.* 2021).

In the context of our problem, one approach is to assume that, say, $s_a(\widehat{a}, a; D^n)$ is sufficiently close to $\mathbb{E}\{b(X) - \widehat{b}(X) \mid A = 1, a(X), D^n\}$ and then proceed by bounding the error in estimating this latter regression function. This route was taken by Dukes *et al.* 2021 and Benkeser *et al.* 2017 and it relies on estimating a regression function with unknown but estimable covariates; in light of the discussion above, this can be challenging. Instead, we impose smoothness conditions directly on $s_a(t_1, t_2; D^n)$ and find that simply regressing $b(X) - \widehat{b}(X)$ on $\widehat{a}(X)$ among units with $A = 1$ would yield a good estimate of $s_a(t_1, t_2; D^n)$ up to an error depending on $\widehat{a} - a$. Which bounding approach, if any, is appropriate likely depends on the application considered. We conclude this section by providing some intuition for when one may expect that s_a, s_b, f_a and f_b possess some smoothness. However, a formal investigation that takes into account specific estimators of $a(X)$ and $b(X)$ is left for future work.

Remark 3. Our proofs of Propositions 3 and 4 below require control only for t_1 and t_2 of the form $t_1 = \widehat{a}(x_0)$ and $t_2 = a(x_0)$ for some arbitrary x_0 in the support of X. In this light, in this remark, we take $s_a(\widehat{a}(x_0), a(x_0); D^n)$ as the example and consider two cases where one may expect this function to be smooth. We leave the conditioning on $A = 1$ and D^n implicit.

Perhaps the easiest case to consider is when $s_a(\widehat{a}(X), a(X); D^n) = \mathbb{E}\{b(X) - \widehat{b}(X) \mid a(X), D^n\}$, i.e., $b(X) - \widehat{b}(X) = m(a(X)) + \epsilon$, where ϵ is mean-zero given both $a(X)$ and $\widehat{a}(X)$. If this is the case, then one needs to assume that $m(\cdot)$ possesses some smoothness, which is arguably a more standard requirement as it does not involve a generated regressor.

Next, we consider the case where, conditional on D^n, $(b(X) - \widehat{b}(X), a(X), \widehat{a}(X))$ is jointly Gaussian:

$$\begin{bmatrix} b(X) - \widehat{b}(X) \\ a(X) \\ \widehat{a}(X) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_b \\ \mu_a \\ \mu_{\widehat{a}} \end{bmatrix}, \begin{bmatrix} \sigma^2_{b-\widehat{b}} & \sigma_{a(b-\widehat{b})} & \sigma_{\widehat{a}(b-\widehat{b})} \\ \sigma_{a(b-\widehat{b})} & \sigma^2_a & \sigma_{\widehat{a}a} \\ \sigma_{\widehat{a}(b-\widehat{b})} & \sigma_{\widehat{a}a} & \sigma^2_{\widehat{a}} \end{bmatrix} \right), \text{ where } \sigma_{vw} = \text{cov}\{v(X), w(X)\}.$$

We have that

$$
\begin{aligned}
&s_a(\widehat{a}(x_0), a(x_0); D^n) - s_a(\widehat{a}(x_1), a(x_1); D^n) \\
&= \left(\frac{\sigma_{a(b-\widehat{b})}\sigma^2_{\widehat{a}} - \sigma_{\widehat{a}a}\sigma_{\widehat{a}(b-\widehat{b})}}{\sigma^2_a\sigma^2_{\widehat{a}} - \sigma^2_{\widehat{a}a}} \right) \{\widehat{a}(x_0) - \widehat{a}(x_1)\} + \left(\frac{\sigma_{\widehat{a}(b-\widehat{b})}\sigma^2_a - \sigma_{a(b-\widehat{b})}\sigma_{\widehat{a}a}}{\sigma^2_a\sigma^2_{\widehat{a}} - \sigma^2_{\widehat{a}a}} \right) \{a(x_0) - a(x_1)\} \\
&\equiv \Lambda_1\{\widehat{a}(x_0) - \widehat{a}(x_1)\} + \Lambda_2\{a(x_0) - a(x_1)\}.
\end{aligned}
$$

Therefore, $s(\widehat{a}(x_0), a(x_0); D^n)$ would be Lipschitz if $|\Lambda_1|$ and $|\Lambda_2|$ are bounded. Letting $\rho_{\widehat{a}a} = \text{cor}\{\widehat{a}(X), a(X)\}$, we can write

$$
\begin{aligned}
\Lambda_1 &= \frac{\sigma_{a(b-\widehat{b})}}{\sigma_{\widehat{a}}\sigma_a} \cdot \frac{1}{1+\rho_{\widehat{a}a}} + \frac{\sigma_{a(b-\widehat{b})}}{\sigma_{\widehat{a}}\sigma^2_a} \cdot \frac{\sigma_{\widehat{a}} - \sigma_a}{1-\rho^2_{\widehat{a}a}} + \frac{\rho_{\widehat{a}a}}{\sigma_{\widehat{a}}\sigma_a} \cdot \frac{\sigma_{(\widehat{a}-a)(b-\widehat{b})}}{(1-\rho^2_{\widehat{a}a})}, \\
\Lambda_2 &= \frac{\sigma_{a(b-\widehat{b})}}{\sigma_{\widehat{a}}\sigma_a} \cdot \frac{1}{1+\rho_{\widehat{a}a}} + \frac{\sigma_{\widehat{a}(b-\widehat{b})}}{\sigma^2_{\widehat{a}}\sigma_a} \cdot \frac{\sigma_a - \sigma_{\widehat{a}}}{1-\rho^2_{\widehat{a}a}} + \frac{1}{\sigma_{\widehat{a}}\sigma_a} \cdot \frac{\sigma_{(\widehat{a}-a)(b-\widehat{b})}}{(1-\rho^2_{\widehat{a}a})}.
\end{aligned}
$$

So, if $\sigma_a \gtrsim 1$, $\sigma_{\widehat{a}} \gtrsim 1$, $|\sigma_{\widehat{a}} - \sigma_a| \lesssim 1 - \rho^2_{\widehat{a}a}$ and $|\sigma_{(\widehat{a}-a)(b-\widehat{b})}| \lesssim 1 - \rho^2_{\widehat{a}a}$, then $|\Lambda_1|$ and $|\Lambda_2|$ are bounded and so $s_a(\widehat{a}(x_0), a(x_0); D^n)$ is Lipschitz. These straightforward calculations provide an example of more primitive conditions, albeit under idealized conditions, under which $s_a(t_1, t_2; D^n)$ may be expected to possess some smoothness.

3.2 Estimators exploiting the smoothness of either s_a or s_b but not both.

In this section, we present two estimators, $\widehat{\psi}_a$ and $\widehat{\psi}_b$, that are tailored to models where s_a and s_b have some smoothness, respectively. We will focus on $\widehat{\psi}_a$, but the same reasoning essentially applies to $\widehat{\psi}_b$ as well. Recall the notation $\widehat{a}_i = \widehat{a}(X_i)$ and $\widehat{a} = \widehat{a}(X)$. Consider the estimator $\widehat{\psi}_a = \widehat{\psi}_{\text{DR}} - T_{na}$ for

$$
T_{na} = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \sum (A_i\widehat{a}_i - 1) \frac{K_h(\widehat{a}_j - \widehat{a}_i)}{\widehat{Q}(\widehat{a}_i)} A_j(Y_j - \widehat{b}_j), \text{ where}
$$

$K_h(u) = h^{-1}K(u/h)$, for some vanishing bandwidth h, and $\widehat{Q}(\widehat{a}_i) = (n-1)^{-1}\sum_{j \ne i} A_j K_h(\widehat{a}_j - \widehat{a}_i)$. Throughout, we assume the following:

Assumption (Kernel). The kernel $K(u)$ is a non-negative, bounded, symmetric function (around zero) that is supported on $[-1, 1]$. One example is $K(u) = 0.5\mathbb{1}(|u| \le 1)$.

It can be seen that our approximation of R_n with T_{na} is based on a second-order U-statistic inspired by the theory of HOIFs. We expand on the similarities and differences between our contributions and the literature on HOIFs in Section 4.1. A more detailed comparison between our estimators and those proposed in Benkeser *et al.* 2017 can be found in Section 4.2. Interestingly, the term T_{na} is similar to the correction in the approximate second-order estimator of ψ proposed (but not analyzed in detail) in Section 3.2 in Diaz *et al.* 2016. However, they consider different nonparametric models and thus do not derive results comparable to ours. In fact, they conclude that their estimator, at least in terms of rates, does not improve upon estimators based on the first-order influence function, such as $\widehat{\psi}_{\text{DR}}$.

If $\mathbb{E}\{A_j K_h(\widehat{a}_j - \widehat{a}_i) \mid X_i, D^n\}$ and $\widehat{Q}^{-1}(\widehat{a}_i)$ are bounded, then by the Cauchy-Schwarz inequality:

$$
|\mathbb{E}(T_{na} \mid D^n)| \lesssim \|\widehat{a} - a\| \|\widehat{b} - b\|.
$$

In this light, this term has expectation of the same order as R_n, the conditional bias of $\widehat{\psi}_{\mathrm{DR}}$, and thus one may hope to not degrade the performance of $\widehat{\psi}_{\mathrm{DR}}$, at least asymptotically, by including T_{na} in the final estimator. A formal calculation, however, would need to consider the variance of T_{na} as well. Next, we outline the reasoning for why and when subtracting off T_{na} from $\widehat{\psi}_{\mathrm{DR}}$ may lead to a better estimator.

We write

$$A_j(Y_j - \widehat{b}_j) = A_j s_a(\widehat{a}_i, a_i; D^n) + A_j\{s_a(\widehat{a}_j, a_j; D^n) - s_a(\widehat{a}_i, a_i; D^n)\} + A_j\epsilon_j,$$

and notice that, by definition of $\widehat{Q}(\widehat{a}_i)$:

$$\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\sum\{A_i\widehat{a}_i - 1\}\frac{K_h(\widehat{a}_j - \widehat{a}_i)}{\widehat{Q}(\widehat{a}_i)}A_j s_a(\widehat{a}_i, a_i; D^n) = \frac{1}{n}\sum_{i=1}^{n}\{A_i\widehat{a}_i - 1\}s_a(\widehat{a}_i, a_i; D^n).$$

Conditioning on the training sample D^n, this term has mean exactly equal to R_n. Next, by definition:

$$\mathbb{E}(A_j\epsilon_j \mid \widehat{a}_j, a_j, D^n) = \mathbb{E}[A_j\{Y_j - \widehat{b}_j - s_a(\widehat{a}_j, a_j; D^n)\} \mid \widehat{a}_j, a_j, A_j, D^n] = 0.$$

This implies that $\mathbb{E}(A_j\epsilon_j \mid \widehat{a}_j, A_j, D^n) = 0$ and, by independence of O_i and O_j for $i \neq j$, also that $\mathbb{E}(A_j\epsilon_j \mid \widehat{a}_j, \widehat{a}_i, A_i, A_j, D^n) = 0$. In this respect, we have

$$\mathbb{E}(T_{na} \mid D^n) = R_n + \mathbb{E}\left[A_1(\widehat{a}_1 - a_1)\frac{K_h(\widehat{a}_2 - \widehat{a}_1)}{\widehat{Q}(\widehat{a}_1)}A_2\{s_a(\widehat{a}_2, a_2; D^n) - s_a(\widehat{a}_1, a_1; D^n)\} \mid D^n\right].$$

If $s_a(t_1, t_2; D^n)$ is Hölder of order $\alpha \in [0, 1]$ (Definition 1), we have

$$\begin{aligned}|s_a(\widehat{a}_2, a_2; D^n) - s_a(\widehat{a}_1, a_1; D^n)| &\lesssim \{(\widehat{a}_2 - \widehat{a}_1)^2 + (a_2 - a_1)^2\}^{\alpha/2} \\ &\leq \{4(\widehat{a}_2 - \widehat{a}_1)^2 + 3(a_2 - \widehat{a}_2)^2 + 3(\widehat{a}_1 - a_1)^2\}^{\alpha/2} \\ &\lesssim |\widehat{a}_2 - \widehat{a}_1|^\alpha + |\widehat{a}_2 - a_2|^\alpha + |\widehat{a}_1 - a_1|^\alpha.\end{aligned}$$

We therefore have

$$\begin{aligned}|\mathbb{E}(T_{na} \mid D^n) - R_n| &\leq \mathbb{E}\left[|\widehat{a}_1 - a_1|\frac{A_2K_h(\widehat{a}_2 - \widehat{a}_1)}{\widehat{Q}(\widehat{a}_1)}|s_a(\widehat{a}_2, a_2; D^n) - s_a(\widehat{a}_1, a_1; D^n)|| \; D^n\right] \\ &\lesssim \mathbb{E}\left[|\widehat{a}_1 - a_1|\frac{A_2K_h(\widehat{a}_2 - \widehat{a}_1)}{\widehat{Q}(\widehat{a}_1)}\{|\widehat{a}_2 - \widehat{a}_1|^\alpha + |\widehat{a}_2 - a_2|^\alpha + |\widehat{a}_1 - a_1|^\alpha\} \mid D^n\right].\end{aligned}$$

If $\widehat{Q}(\widehat{a}_1)$ is bounded away from zero and $\mathbb{E}\{A_2K_h(\widehat{a}_2 - \widehat{a}_1) \mid X_1, D^n\}$ is bounded, then

$$|\mathbb{E}(T_{na} \mid D^n) - R_n| \lesssim \left(\|\widehat{a} - a\|^{1+\alpha} + \|\widehat{a} - a\|h^\alpha\right) \wedge \|\widehat{a} - a\|\|\widehat{b} - b\|.$$

For $\alpha = 1$, i.e., $s_a(t_1, t_2; D^n)$ is Lipschitz, choosing $h \asymp n^{-1/2}$ yields

$$\left|\mathbb{E}(\widehat{\psi}_a \mid D^n) - \psi\right| \lesssim \|\widehat{a} - a\|^2 \wedge \|\widehat{a} - a\|\|\widehat{b} - b\|.$$

The bound on the right-hand-side of the display above improves upon the conditional bias of $\widehat{\psi}_{\mathrm{DR}}$

in the case when $\|\widehat{a}-a\|$ is of smaller order than $\|\widehat{b}-b\|$.

Next, we look for an estimator that achieves a bound on the bias of order $\|\widehat{b}-b\|^2 \wedge \|\widehat{a}-a\|\|\widehat{b}-b\|$. A natural candidate is $\widehat{\psi}_b = \widehat{\psi}_{\mathrm{DR}} - T_{nb}$, where:

$$T_{nb} = \frac{1}{n(n-1)} \sum_{1\le i\neq j\le n} \sum (A_i\widehat{a}_i - 1)\frac{K_h(\widehat{b}_j - \widehat{b}_i)}{\widehat{Q}_{-i}(\widehat{b}_j)} A_j(Y_j - \widehat{b}_j),$$

$$\widehat{Q}_{-i}(\widehat{b}_j) = (n-1)^{-1} \sum_{s=1, s\neq i}^{n} A_s K_h(\widehat{b}_s - \widehat{b}_j).$$

We remove the i^{th} observation in $\widehat{Q}_{-i}(\widehat{b}_j)$ because we need to ensure that

$$\mathbb{E}\left\{(A_i\widehat{a}_i - 1)\frac{K_h(\widehat{b}_j - \widehat{b}_i)}{\widehat{Q}_{-i}(\widehat{b}_j)} A_j(Y_j - \widehat{b}_j) \mid D^n\right\} = \mathbb{E}\left\{A_i(\widehat{a}_i - a_i)\frac{K_h(\widehat{b}_j - \widehat{b}_i)}{\widehat{Q}_{-i}(\widehat{b}_j)} A_j(Y_j - \widehat{b}_j) \mid D^n\right\}.$$

This would not be the case if the residual $A_i\widehat{a}_i - 1$ gets multiplied by a term $\widehat{Q}(\widehat{b}_j)$ involving A_i. With this modification in place, calculations analogous to the ones above for $\widehat{\psi}_a$ yield that

$$|\mathbb{E}(T_{nb} \mid D^n) - R_n| \lesssim \left(\|\widehat{b}-b\|^{1+\beta} \vee h^\beta \|\widehat{b}-b\| \vee \frac{\|\widehat{a}-a\|\|\widehat{b}-b\|}{\sqrt{nh}}\right) \wedge \|\widehat{a}-a\|\|\widehat{b}-b\|.$$

when $s_b(t_1, t_2; D^n)$ is Hölder $\beta \in [0,1]$. When $\beta = 1$ and $h \asymp n^{-1/2}$, the bound reduces to

$$\left|\mathbb{E}(\widehat{\psi}_b \mid D^n) - \psi\right| \lesssim (n^{-1/2} \vee \|\widehat{b}-b\|^2) \wedge \|\widehat{a}-a\|\|\widehat{b}-b\|.$$

Therefore, $\widehat{\psi}_b$ would have smaller conditional bias than $\widehat{\psi}_{\mathrm{DR}}$ if $b(x)$ is easier to estimate than $a(x)$ so that $\|\widehat{b}-b\|^2$ is the dominant term. In the following proposition, we collect more formal statements on the conditional bias and variance of $\widehat{\psi}_a$ and $\widehat{\psi}_b$.

Proposition 3. *Suppose that $\mathbb{E}\{A_j K_h(\widehat{a}_j - \widehat{a}_i) \mid X_i, D^n\} \lesssim 1$ and that $\widehat{Q}(\widehat{b}_j)$ is bounded away from zero for $1 \le i \neq j \le n$. Then, if $s_a(t_1, t_2; D^n)$ is Hölder of order $\alpha \in [0,1]$, it holds that*

$$\left|\mathbb{E}(\widehat{\psi}_a - \psi \mid D^n)\right| \lesssim \left(h^\alpha \|\widehat{a}-a\| \vee \|\widehat{a}-a\|^{1+\alpha}\right) \wedge \|\widehat{a}-a\|\|\widehat{b}-b\|$$

$$var(\widehat{\psi}_a \mid D^n) \lesssim n^{-1} \vee (n^2 h)^{-1} \vee \frac{\|\widehat{a}-a\|^2\|\widehat{b}-b\|^2}{nh}.$$

Suppose that $\mathbb{E}\{A_j K_h(\widehat{b}_j - \widehat{b}_i) \mid X_i, D^n\} \lesssim 1$ and that $\widehat{Q}_{-i}(\widehat{b}_j)$ is bounded away from zero for $1 \le i \neq j \le n$. Then, if $s_b(t_1, t_2; D^n)$ is Hölder of order $\beta \in [0,1]$, it holds that

$$\left|\mathbb{E}(\widehat{\psi}_b - \psi \mid D^n)\right| \lesssim \left(h^\beta \|\widehat{b}-b\| \vee \|\widehat{b}-b\|^{1+\beta} \vee \frac{\|\widehat{a}-a\|\|\widehat{b}-b\|}{\sqrt{nh}}\right) \wedge \|\widehat{a}-a\|\|\widehat{b}-b\|$$

$$var(\widehat{\psi}_b \mid D^n) \lesssim n^{-1} \vee (n^2 h)^{-1} \vee \frac{\|\widehat{a}-a\|^2\|\widehat{b}-b\|^2}{nh}.$$

The results from Proposition 3 yield that $\widehat{\psi}_a$ can improve upon the performance of $\widehat{\psi}_{\mathrm{DR}}$ in models where a is easier to estimate than b, while $\widehat{\psi}_b$ can improve upon $\widehat{\psi}_{\mathrm{DR}}$ when b is easier to estimate

than a. The main requirement for the proposition to hold is that s_a and s_b possess some minimal smoothness encoded in the Hölder condition. It can be seen from a standard change of variables argument that the conditions $\mathbb{E}\{A_j K_h(\widehat{a}_j - \widehat{a}_i) \mid X_i, D^n\}$ and $\mathbb{E}\{A_j K_h(\widehat{b}_j - \widehat{b}_i) \mid X_i, D^n\}$ are satisfied if $\widehat{a}(X)$ and $\widehat{b}(X)$ have densities with respect to the Lebesgue measure, respectively. We expect the assumption that $\widehat{Q}(\widehat{a}_j)$ and $\widehat{Q}_{-i}(\widehat{b}_j)$ are bounded away from zero to hold in practice as long as the sample size is large enough and $\widehat{a}(X)$ and $\widehat{b}(X)$ are evenly distributed on their support.

Remark 4. It is reasonable to consider applications where α and β could be greater than 1, i.e., s_a and s_b are potentially smoother than what considered in Proposition 3. This higher-order smoothness could then be exploited using higher-order kernels or kernels of local polynomials, for example. However, our analysis (not reported in the interest of space) suggests that the bound on the bias would still contain terms of order $\|\widehat{a} - a\|^{1+(\alpha \wedge 1)}$ and $\|\widehat{b} - b\|^{1+(\beta \wedge 1)}$. This is consistent with our lower bound analysis (Proposition 2), which establishes that the rate of convergence for $\widehat{\psi}_a$ and $\widehat{\psi}_b$ in these models cannot be faster than $\|\widehat{a} - a\|^2$ and $\|\widehat{b} - b\|^2$, respectively. Attempting to track higher order smoothness, however, may have benefits in terms of more flexibility in choosing the bandwidth h, particularly in finite samples. We leave the study of higher-smoothness regimes for future work.

Remark 5. Suppose that $\alpha = 1$ and set $h \asymp n^{-1/2}$. If $\|\widehat{a} - a\| = o_{\mathbb{P}}(n^{-1/4})$, then $|\widehat{\psi}_a - \psi| = O_{\mathbb{P}}(n^{-1/2})$. Further, by the decomposition (3), $\sqrt{n}(\widehat{\psi}_a - \psi) \rightsquigarrow N(0, \mathrm{var}(\overline{\varphi}))$. This means that $\widehat{\psi}_a$ can be used for constructing a Wald-type confidence interval even if $\widehat{b}$ is not consistent. However, if $\|\widehat{b} - b\| = o_{\mathbb{P}}(1)$, then $\widehat{\psi}_a$ will also achieve the semiparametric efficiency bound $\mathrm{var}(\varphi)$ because, in this case, $\overline{\varphi} = \varphi$ since $\overline{a} = a$ and $\overline{b} = b$. Similar considerations apply to $\widehat{\psi}_b$ with the roles of $\widehat{a}$ and $\widehat{b}$ reversed.

3.3 Main estimator

To be useful in practice, the estimators presented in Section 3.2 require knowledge of whether a or b is easier to estimate. In this section, we consider the estimator $\widehat{\psi} = \widehat{\psi}_{\mathrm{DR}} - T_n$, for

$$T_n = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \sum (A_i \widehat{a}_i - 1) \frac{K_h(\widehat{b}_j - \widehat{b}_i) K_h(\widehat{a}_j - \widehat{a}_i)}{\widehat{Q}(\widehat{a}_i, \widehat{b}_i)} A_j (Y_j - \widehat{b}_j)$$

where $\widehat{Q}(\widehat{a}_i, \widehat{b}_i) = (n-1)^{-1} \sum_{i \ne j} A_j K_h(\widehat{a}_j - \widehat{a}_i) K_h(\widehat{b}_j - \widehat{b}_i)$. We will show that, under certain conditions, this risk of this estimator is of the same order as the lower bound on the risk of any estimator derived in Proposition 2 (Claim 3), thereby establishing sufficient conditions under which this estimator is minimax optimal.

A key difference between $\widehat{\psi}$ and either $\widehat{\psi}_a$ or $\widehat{\psi}_b$ is that $\widehat{\psi}$ is tailored to models where both $f_a(t_1, t_2, t_3; D^n)$ and $f_b(t_1, t_2, t_3; D^n)$ have some smoothness. The central term multiplying the two residuals $A_i \widehat{a}_i - 1$ and $A_j(Y_j - \widehat{b}_j)$ is meant to act as a kernel of a local regression on $(\widehat{a}, \widehat{b})$ rather than on $\widehat{a}$ or $\widehat{b}$ alone. The reason for this change is that $\widehat{\psi}$ is designed to correct the bias of $\widehat{\psi}_{\mathrm{DR}}$ by subtracting off an estimate of R_n even when it is not known whether a or b is easier to estimate. In fact, one can see that estimating $f_a(\widehat{a}, a, \widehat{b}; D^n)$ and $f_b(\widehat{b}, b, \widehat{a}; D^n)$ can be carried out simply by modifying the outcome variable as both residuals would be regressed onto $(\widehat{a}, \widehat{b})$. This then allows for the construction of estimates of R_n that would be essentially the same whether one expresses R_n in terms of f_a or f_b. One caveat is that summing over i in the expression for T_n should return an estimate of $f_b(\widehat{b}_j, b_j, \widehat{a}_j; D^n)$, but this is not exactly the case because $\widehat{Q}(\widehat{a}_i, \widehat{b}_i)$ is localized at $(\widehat{a}_i, \widehat{b}_i)$ instead of at $(\widehat{a}_j, \widehat{b}_j)$. In Proposition 4, we deal with this issue by invoking a Lipschitz assumption on the density of $(\widehat{a}, \widehat{b})$ among units with $A = 1$. Whether this condition (or similar

ones) can be avoided remains an open question. The main advantage of using $\widehat{\psi}$ versus either $\widehat{\psi}_a$ or $\widehat{\psi}_b$ is that $\widehat{\psi}$ is able to correct for R_n when R_n is expressed as a function of f_a or f_b using the same term T_n. Relative to $\widehat{\psi}_a$ and $\widehat{\psi}_b$, the price is a moderate increase in the variance because of the product of two kernels, which is needed for estimating the regressions on a two-, as opposed to one-, dimensional domain. The following proposition summarizes our bounds on the bias and variance of $\widehat{\psi}$.

Proposition 4. *Suppose that the distribution of $(\widehat{a}, \widehat{b})$ among units with $A = 1$ has a density with respect to the Lebesgue measure that is Lipschitz. Suppose that $\widehat{Q}(\widehat{a}_i, \widehat{b}_i)$ is bounded away from zero for $1 \leq j \leq n$. Further, suppose that $f_a(t_1, t_2, t_3; D^n)$ and $f_b(t_1, t_2, t_3; D^n)$ are Hölder smooth order α and β, respectively, with $\alpha, \beta \in [0, 1]$. Finally, assume that $nh^2 \to \infty$. Then,*

$$\left|\mathbb{E}(\widehat{\psi} - \psi \mid D^n)\right| \lesssim \left(\|\widehat{a} - a\| h^\alpha \vee \|\widehat{a} - a\|^{1+\alpha}\right) \wedge \left(\|\widehat{b} - b\| h^\beta \vee \|\widehat{b} - b\|^{1+\beta} \vee \frac{\|\widehat{a} - a\| \|\widehat{b} - b\|}{\sqrt{nh^2}}\right)$$
$$\wedge \|\widehat{a} - a\| \|\widehat{b} - b\|$$
$$var(\widehat{\psi} \mid D^n) \lesssim \frac{1}{n} \vee \frac{1}{(nh)^2} \vee \frac{\|\widehat{a} - a\|^2 \|\widehat{b} - b\|^2}{nh^2}$$

It can be seen from Proposition 4 that $\widehat{\psi}$ has the potential to improve upon $\widehat{\psi}_a$ and $\widehat{\psi}_b$ in the sense that its bias is the minimum between their biases. This comes at the price of a smoothness assumption on both f_a and f_b as well as a Lipschitz condition on the density of $(\widehat{a}(X), \widehat{b}(X))$. Under these conditions, $\widehat{\psi}$ also improves upon $\widehat{\psi}_{\text{DR}}$ and can deliver doubly-robust root-n inference when either $\widehat{a}$ or $\widehat{b}$ converges at $n^{-1/4}$ rates, as long as $\alpha = \beta = 1$ and $h \asymp n^{-1/4}$. In practice, choosing h can be nontrivial. In Section 5, we select the cross-validated bandwidth that an estimator of the regression of $f_a(t_1, t_2, t_3; D^n)$ would choose. This choice can be implemented using off-the-shelf routines but we do not claim any optimality for it. How to choose the bandwidth in practice remains largely an open question. We note that this difficulty in selecting the tuning parameter also arises in Dukes *et al.* 2021.

Remark 6. The assumption that the density of $(\widehat{a}(X), \widehat{b}(X))$ among units with $A = 1$ is Lipschitz could potentially be relaxed to requiring that the density is Hölder of order $\gamma \in [0, 1]$. However, our bound on the bias would then be

$$\left|\mathbb{E}(\widehat{\psi} - \psi \mid D^n)\right| \lesssim \left(\|\widehat{a} - a\| h^\alpha \vee \|\widehat{a} - a\|^{1+\alpha}\right) \wedge \|\widehat{a} - a\| \|\widehat{b} - b\|$$
$$\wedge \left(\|\widehat{b} - b\| h^\beta \vee \|\widehat{b} - b\|^{1+\beta} \vee \frac{\|\widehat{a} - a\| \|\widehat{b} - b\|}{\sqrt{nh^2}} \vee h^\gamma \|\widehat{a} - a\| \|\widehat{b} - b\|\right).$$

Remark 7. Consider the case where $\alpha = \beta = 1$ and set $h \asymp n^{-1/4}$. Then, it holds that

$$\left|\mathbb{E}(\widehat{\psi} - \psi \mid D^n)\right| \lesssim \|\widehat{a} - a\|^2 \wedge \|\widehat{b} - b\|^2 + o_{\mathbb{P}}(n^{-1/2})$$
$$\text{var}(\widehat{\psi} \mid D^n) \lesssim n^{-1} \vee n^{-1/2} \|\widehat{a} - a\|^2 \|\widehat{b} - b\|^2.$$

In this light, from (3), $\sqrt{n}(\widehat{\psi} - \psi)$ is asymptotically normal as long as $\|\widehat{a} - a\| = o_{\mathbb{P}}(n^{-1/4})$ or $\|\widehat{b} - b\| = o_{\mathbb{P}}(n^{-1/4})$. Further, it is semiparametric efficient if, in addition, $\|\widehat{a} - a\| = o_{\mathbb{P}}(1)$ and $\|\widehat{b} - b\| = o_{\mathbb{P}}(1)$.

Remark 8. We believe our construction of $\widehat{\psi}$ sheds light on a point raised by Benkeser *et al.* 2017 about the putative superiority of TMLE over one-step corrections when it comes to performing doubly-robust inference (Section 4 in Benkeser *et al.* 2017). The authors note the difficulty in deriving doubly-robust, asymptotic linear estimators that take the form of $\widehat{\psi}_{\text{DR}}$ plus a correction term. This apparent difference in performance is surprising, as one-step estimators and TMLEs are grounded in the same semiparametric efficiency theory and typically, at least asymptotically and in regimes where ψ admits root-n-consistent estimators, share largely the same properties. Proposition 4 shows that, under certain conditions, there exists in fact an estimator that can estimate R_n "in one step" and could thus be used for doubly-robust inference.

Finally, in Appendix A, we briefly discuss how our approach can be applied to the partially linear logistic model. This is a type of parameter which the approach from Dukes *et al.* 2021 does not readily extend to. Instead, we show how a suitable modification of $\widehat{\psi}$ can deliver doubly-robust inference in this setting as well.

4 Connections to other literature

4.1 Higher order influence functions

In this section, we expand on the similarities and differences between $\widehat{\psi}$ and the estimator of ψ based on the (approximate) second-order influence function. The estimator is described in detail in a series of articles by Robins and co-authors (Liu *et al.* 2017; Robins *et al.* 2017b, 2008, 2009a). It takes the form:

$$\widehat{\psi}_{HOIF-2} = \widehat{\psi}_{\text{DR}} - \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \sum (A_i \widehat{a}_i - 1) p_i^T \widehat{\Omega}^{-1} p_j A_j (Y_j - \widehat{b}_j),$$

where $p(x) = (p_0(x), p_1(x), \ldots, p_k(x))$ is a vector of basis functions suitable for approximating the functions $a(x) - \widehat{a}(x)$ and $b(x) - \widehat{b}(x)$, and $\Omega = \mathbb{E}\{Ap(X)p^T(X)\}$. When X is multivariate, $p(x)$ can be taken to be a tensor product of univariate basis functions. The tuning parameter k is chosen to diverge with the sample size n to obtain progressively better approximations of R_n. A natural estimator of Ω is its empirical counterpart. However, in regimes of low-smoothness, k is selected to grow faster than n so that the empirical version is not invertible. In this case, an alternative estimator, which requires estimating the density of X, is $\int p(x)p^T(x)\widehat{g}(x)dx$, where $g(x)$ denotes the density of X among the units with $A = 1$ multiplied by $\mathbb{P}(A = 1)$. See Liu *et al.* 2017 and Liu & Li 2023 for additional details.

One can see that $\widehat{\psi}$ and $\widehat{\psi}_{HOIF-2}$ differ only in that $\widehat{\Pi}_k(x_i, x_j) \equiv p_i^T \widehat{\Omega}^{-1} p_j$ is replaced by

$$\widehat{\Pi}_h(\widehat{a}_i, \widehat{b}_i, \widehat{a}_j, \widehat{b}_j) = \widehat{Q}^{-1}(\widehat{a}_j, \widehat{b}_j) K_h(\widehat{a}_i - \widehat{a}_j) K_h(\widehat{b}_i - \widehat{b}_j)$$

in the former. The kernel $\widehat{\Pi}_k(x_i, x_j)$ is designed to approximate the kernel of an orthogonal projection in $L_2(g)$. That is,

$$\mathbb{E}_{Z_j | D^n} \left[\Pi_k(X_i, X_j) A_j \{Y_j - b_j(X)\} \right] = b(X_i) - \widehat{b}(X_i) + \text{approximation error}.$$

The larger the size of the image of the projection (i.e., the larger k), the smaller the approximation error is. However, the variance of $\widehat{\psi}_{HOIF-2}$ increases with k, so k needs to be carefully tuned to minimize the mean-square-error. In practice, one needs to use $\widehat{\Pi}_k$ in place of Π_k, so that the

approximation error depends also on the accuracy with which Ω can be estimated. Additional higher-order corrections are needed when $\|\widehat{\Omega}^{-1} - \Omega^{-1}\|_{\text{op}}$ or $\|\widehat{g} - g\|_\infty$ are not sufficiently small. One key feature of the population version $\Pi_k(x_i, x_j)$ is that it satisfies

$$\mathbb{E}\left\{(A_i\widehat{a}_i - 1)\Pi_k(X_i, X_j)A_j(Y_j - \widehat{b}_j) \mid D^n\right\} = \int \Pi_k(\widehat{a} - a)(x)\Pi_k(\widehat{b} - b)(x)g(x)dx,$$

where $\Pi_k(f)(x) = p(x)^T\Omega^{-1}\int p(x)f(x)g(x)dx$ denote the weighted orthogonal projection (with weight g) of f onto the space spanned by p. This is crucial to obtain that

$$R_n - \mathbb{E}\left\{(A_i\widehat{a}_i - 1)\Pi_k(X_i, X_j)A_j(Y_j - \widehat{b}_j) \mid D^n\right\} = \int (I - \Pi_k)(\widehat{a} - a)(x)(I - \Pi_k)(b - \widehat{b})(x)g(x)dx,$$

which is a remainder error term that is particularly small because it is a product of approximation errors. The kernel $\widehat{\Pi}_h(\widehat{a}_i, \widehat{a}_j, \widehat{b}_i, \widehat{b}_j)$ appearing in $\widehat{\psi}$ does not yield a product of approximation errors even if the true $Q(\widehat{a}_j, \widehat{b}_j)$ is used. However, it is designed to achieve the similar goal of approximating the kernel of a local regression in the sense that

$$\mathbb{E}_{Z_j|D^n}\left[\Pi_h(\widehat{a}_i, \widehat{b}_i, \widehat{a}_j, \widehat{b}_j)A_j\{Y_j - b_j(X)\}\right] = f_a(\widehat{a}_i, a_i, \widehat{b}_i; D^n) + \text{approximation error}.$$

Crucially, it retains a symmetry property from $\Pi_k(x_i, x_j)$, which is that taking the expectation with respect to Z_j yields an approximation of $f_b(\widehat{a}_i, a_i, \widehat{b}_i; D^n)$, while taking the expectation with respect to Z_i yields an approximation of $f_a(\widehat{b}_j, b_j, \widehat{a}_i; D^n)$:

$$\mathbb{E}_{Z_i|D^n}\left[\widehat{\Pi}_h(\widehat{a}_i, \widehat{b}_i, \widehat{a}_j, \widehat{b}_j)(A_i\widehat{a}(X_i) - 1)\right] = f_b(\widehat{b}_j, b_j, \widehat{a}_i; D^n) + \text{approximation error}.$$

This then allows for the estimation of R_n relying on f_a or f_b depending on which one is smoother. One advantage of $\widehat{\psi}$ over $\widehat{\psi}_{HOIF-2}$ (or higher order versions) is that the kernel Π_h in $\widehat{\psi}$ is low dimensional no matter how large the dimension of X is. The advantage of estimators based on HOIFs is that they can better exploit the regularity of a and b (as opposed to being agnostic with respect to their structure and instead exploiting the regularity of f_a and f_b) when the dictionary of basis functions is chosen appropriately. Further, they are grounded in a very general theoretical framework for functional estimation.

4.2 Doubly-robust inference

Van der Laan 2014 and Benkeser *et al.* 2017 derive estimators of ψ that remain $\sqrt{n}$-consistent and asymptotically normal even when either $\widehat{\pi} = 1/\widehat{a}$ or $\widehat{b}$ (but not both) is misspecified. More recently, Dukes *et al.* 2021 have also proposed estimators enjoying this property but focused on the expected conditional covariance functional, defined as $\mathbb{E}\{\text{cov}(Y, A \mid X)\}$. In this section, we briefly outline some similarities and differences between $\widehat{\psi}$ and the estimator proposed in Benkeser *et al.* 2017 and implemented in the `R` package `drtmle` (Benkeser & Hejazi 2023).

To start, the authors write $R_n = \mathbb{P}(\widehat{\psi}_{\text{DR}} - \psi)$ as

$$R_n = \mathbb{1}(\overline{\pi} = \pi)\mathbb{E}\left\{\frac{(\overline{b} - b)(\widehat{\pi} - \pi)}{\pi} \mid D^n\right\} + \mathbb{1}(\overline{b} = b)\mathbb{E}\left\{\frac{(\widehat{b} - b)(\overline{\pi} - \pi)}{\overline{\pi}} \mid D^n\right\}$$
$$+ \mathbb{E}\left\{\frac{(\widehat{b} - \overline{b})(\widehat{\pi} - \overline{\pi})}{\overline{\pi}} \mid D^n\right\} + \mathbb{E}\left\{\frac{(\widehat{b} - b)(\widehat{\pi} - \pi)(\widehat{\pi} - \overline{\pi})}{\overline{\pi}\widehat{\pi}} \mid D^n\right\}$$

where we recall the notation $\overline{\pi}$ and $\overline{b}$ to denote the limits as $n \to \infty$ of $\widehat{\pi}$ and $\widehat{b}$ respectively. The last two terms are asymptotically negligible as long as $\widehat{\pi}$ and $\widehat{b}$ converge to their corresponding limits at $n^{-1/4}$-rates and either $\overline{b} = b$ or $\overline{\pi} = \pi$. Next, they make the observation that the first term can be written as

$$\mathbb{E}\left\{\frac{(\overline{b} - b)(\widehat{\pi} - \pi)}{\pi} \mid D^n\right\} = -\mathbb{E}\left[\mathbb{E}\{Y - \overline{b}(X) \mid A = 1, \widehat{\pi}(X), \pi(X), D^n\}\left\{\frac{\widehat{\pi}(X) - \pi(X)}{\pi(X)}\right\} \mid D^n\right]$$

Notice that $\mathbb{E}\{Y - \overline{b}(X) \mid A = 1, \widehat{\pi}(X), \pi(X), D^n\}$ is essentially $s_a(t_1, t_2; D^n)$ up to the parametrization in terms of $\pi(x) = 1/a(x)$ as opposed to $a(x)$, and with $\overline{b}$ replacing $\widehat{b}$. Next, they show that

$$\mathbb{E}\left\{\frac{(\widehat{b} - b)(\overline{\pi} - \pi)}{\overline{\pi}} \mid D^n\right\} = \mathbb{E}\left[A\{Y - \widehat{b}(X)\}\frac{\mathbb{E}\left\{\frac{A - \overline{\pi}(X)}{\overline{\pi}(X)} \mid \widehat{b}(X), b(X), D^n\right\}}{\mathbb{E}\{A \mid \widehat{b}(X), b(X), D^n\}} \mid D^n\right].$$

Because one does not know whether $\overline{\pi} = \pi$ or $\overline{b} = b$, following the guiding principles of TMLE, they propose fluctuating $\widehat{\pi}$, $\widehat{b}$, as well as the estimators of the three other nuisance functions, $\widehat{\mathbb{E}}\left\{\frac{A - \widehat{\pi}(X)}{\widehat{\pi}(X)} \mid \widehat{b}(X), D^n\right\}$, $\widehat{\mathbb{E}}\{A \mid \widehat{b}(X), D^n\}$, and $\widehat{\mathbb{E}}\{Y - \widehat{b}(X) \mid A = 1, \widehat{\pi}(X), D^n\}$, so that they simultaneously satisfy:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\widehat{\pi}^*(X_i)}\{Y_i - \widehat{b}^*(X_i)\} \approx 0,$$
$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbb{E}}^*\{Y - \widehat{b}^*(X) \mid A = 1, \widehat{\pi}^*(X_i), D^n\}\left\{\frac{\widehat{\pi}^*(X_i) - A_i}{\widehat{\pi}^*(X_i)}\right\} \approx 0,$$
$$\frac{1}{n}\sum_{i=1}^{n}A_i\{Y_i - \widehat{b}^*(X_i)\}\frac{\widehat{\mathbb{E}}^*\left\{\frac{A - \widehat{\pi}^*(X)}{\widehat{\pi}^*(X)} \mid \widehat{b}^*(X), D^n\right\}}{\widehat{\mathbb{E}}^*\{A \mid \widehat{b}^*(X), D^n\}} \approx 0.$$

They propose an iterative procedure to solve these three moment conditions. However, their results are in terms of high-level conditions and the convergence properties of their algorithms are not fully analyzed. For example, two conditions are that

$$\int\left[\widehat{\mathbb{E}}^*\{Y - \overline{b}(X) \mid A = 1, \widehat{\pi}^*(x), D^n\} - \mathbb{E}\{Y - \overline{b}(X) \mid A = 1, \widehat{\pi}(x), \pi(x), D^n\}\right]^2 d\mathbb{P}(x) = o_{\mathbb{P}}(n^{-1/2}),$$
$$\int\left[\widehat{\mathbb{E}}^*\{Y - \overline{b}(X) \mid A = 1, \widehat{\pi}^*(x), D^n\} - \mathbb{E}\{Y - \overline{b}(X) \mid A = 1, \pi(x), D^n\}\right]^2 d\mathbb{P}(x) = o_{\mathbb{P}}(n^{-1/2}).$$

In certain applications, the last condition can be hard to justify because it pertains to the estimation of a regression function on unknown covariates for which the convergence rate can be slow; see Mammen *et al.* 2012 and Theorem 3 in Dukes *et al.* 2021.

Our construction relies on smoothness assumptions imposed directly on $f_a(\widehat{a}, a, \widehat{b}; D^n)$ and $f_b(\widehat{b}, b, \widehat{a}; D^n)$. This allows us to avoid devising an iterative procedure. However, we do not claim that our estimator should be preferred to that proposed in Benkeser *et al.* 2017 in applications, which in fact performs well in the small simulation study described in the next section. We view the introduction of a hybrid structure-agnostic class where doubly-robust inference is possible and the analysis of a one-step estimator in this context as our main contributions.

5 Simulation experiments

In this section, we consider simulation experiments to investigate the finite-sample properties of the estimators studied. The main goal is to verify that the estimator introduced in Section 3.3 yields a performance comparable to its Doubly-Robust TMLE counterpart implemented in the `R` package `drtmle` (Benkeser & Hejazi 2023). We consider the following data generating process:

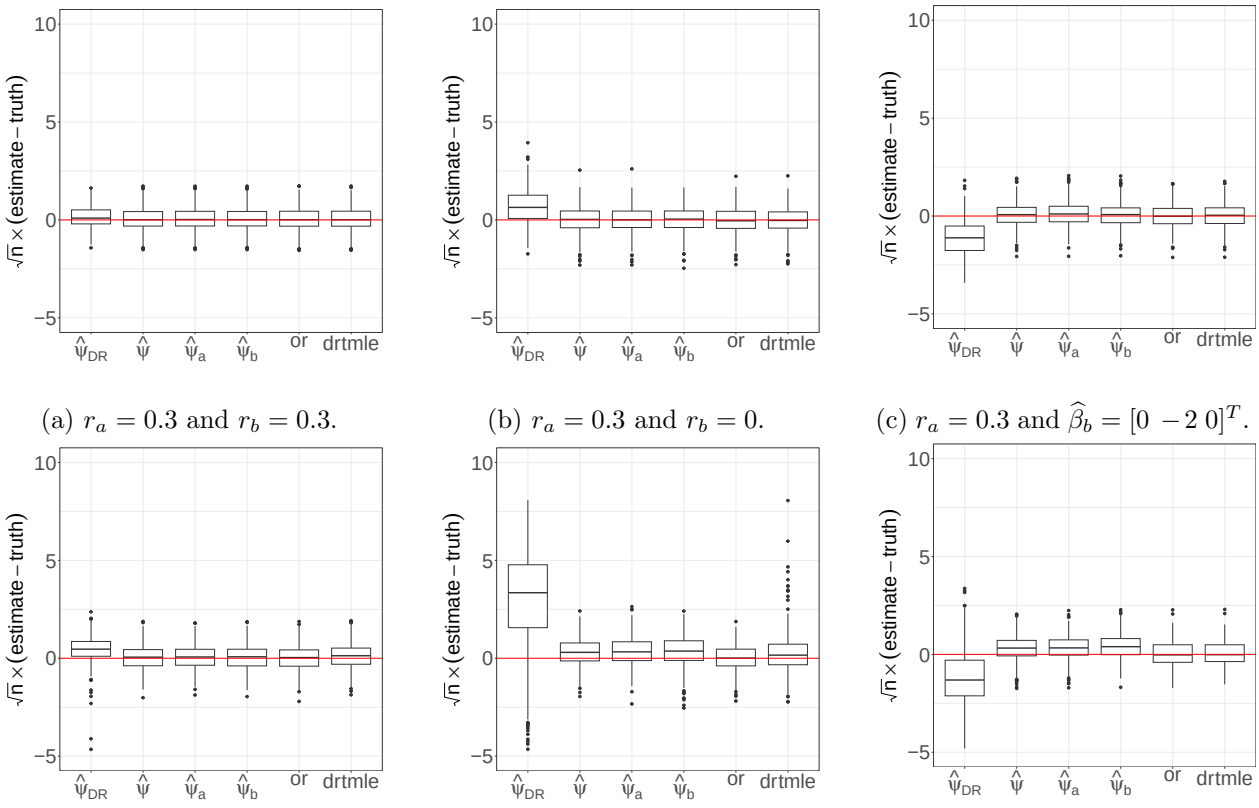
- $X = (X_1, X_2)$, where $X_i \sim \text{Unif}(-1, 1)$ for $i \in \{1, 2\}$, $X_1 \perp\!\!\!\perp X_2$,
- $A \sim \text{Binom}(\pi(x))$ and $Y = AY^1 + (1-A)Y^0$, where $Y^1 \sim \text{Binom}(b(x))$ and $Y^0 \sim \text{Binom}(0.5)$,
- $\pi(x) = \text{expit}(\begin{bmatrix}1 & x\end{bmatrix}^T \beta_a)$, where $\beta_a = \begin{bmatrix}-0.5 & 2 & 0.5\end{bmatrix}^T$,
- $b(x) = \text{expit}(\begin{bmatrix}1 & x\end{bmatrix}^T \beta_b)$, where $\beta_b = \begin{bmatrix}-2.5 & 5 & 2\end{bmatrix}^T$,
- $\widehat{\pi}(x) = \text{expit}(\begin{bmatrix}1 & x\end{bmatrix}^T \widehat{\beta}_a)$, where $\widehat{\beta}_a = \beta_a + N(n^{-r_a}, n^{-r_a})$,
- $\widehat{b}(x) = \text{expit}(\begin{bmatrix}1 & x\end{bmatrix}^T \widehat{\beta}_b)$, where $\widehat{\beta}_b = \beta_b + N(n^{-r_b}, n^{-r_b})$.

In this set-up, the target parameter is $\psi = \mathbb{E}\{b(X)\} \approx 0.66$ and $\widehat{a}$ and $\widehat{b}$ converge to a and b at rates n^{-r_a} and n^{-r_b} respectively. We vary r_a, r_b in $\{0, 0.3\}$ and the sample size n in $\{500, 1000, 2000, 4000\}$. We also consider a scenario where $\widehat{\beta}_a = \widehat{\beta}_b = \begin{bmatrix}0 & -2 & 0\end{bmatrix}^T$, i.e., both $a(x)$ and $b(x)$ are completely mis-specified but they are fixed across iterations. The reason why we investigate this scenario as well is that we noticed that for $\alpha = 0$, injecting $N(1, 1)$ noise in β_a often results in propensity scores that are quite extreme and potentially not very realistic. We run 500 simulations. We select the bandwidth for constructing $\widehat{\psi}_a$, $\widehat{\psi}_b$ and $\widehat{\psi}$ as the one that we would choose to estimate $\mathbb{E}(Y - \widehat{b}(X) \mid A = 1, \widehat{a}(X), \widehat{b}(X), D^n\}$. In particular, we use the cross-validation-based selector from the `R` package `sm` (`h.select()` function) (Bowman & Azzalini 2024). Our theoretical results do not justify this choice, but, in this simulation set-up, it yields reasonable results. As a benchmark, we also consider the performance of an oracle estimator that has access to the true nuisance functions $a(x)$ and $b(x)$.

Figure 1 reports the distribution of $\sqrt{n}(\overline{\psi} - \psi)$ for different choices of $\overline{\psi}$ and the 4 possible combinations of (α, β) plus the case when either $\widehat{\beta}_a$ or $\widehat{\beta}_b$ is inconsistent and set to $[0 \ -2 \ 0]^T$ while the other nuisance functions is consistently estimated at rate $n^{-0.3}$. The sample size is $n = 2000$. We leave out the three cases where either $r_a = 0$ or $r_b = 0$ and either $\widehat{\beta}_a$ or $\widehat{\beta}_b$ is set to $[0 \ -2 \ 0]^T$; in these settings, as expected, $\widehat{\psi}_{\text{DR}}$, $\widehat{\psi}$, $\widehat{\psi}_a$ and $\widehat{\psi}_b$ present very large errors. However, surprisingly, `drtmle` still performs well. We conjecture this is due to the the fact that, for `drtmle`, the highly mis-specified $\widehat{a}(x)$ and $\widehat{b}(x)$ are suitably fluctuated to solve the relevant moment conditions, whereas the other four estimators take these estimated nuisances as given. The main take-away from Figure 1 is that $\widehat{\psi}_{\text{DR}}$ is the only estimator suffering from the mis-specification of one nuisance function. Surprisingly, $\widehat{\psi}$, $\widehat{\psi}_a$ and $\widehat{\psi}_b$ perform similarly. We conjecture this could be due to the fact that the

kernel in their corrections give similar weights to the observations even when the kernel involves a mis-specified covariate. However, this warrants further investigation.

Figure 2 reports the coverage of a Wald confidence interval as a function of the sample size n. The most noticeable pattern is that an interval based on $\widehat{\psi}_{\text{DR}}$ fails to achieve nominal coverage in all scenarios considered except when both nuisance functions are correctly specified. When at least one nuisance function is correctly specified, the performance of `drtmle` is remarkable. This is also the case for the estimators $\widehat{\psi}$, $\widehat{\psi}_a$ and $\widehat{\psi}_b$ except that they exhibit under-coverage when the propensity score is mis-specified by injecting $N(1,1)$ noise to the true function. Our investigation of this issue reveals that this appears to be due to the fact that injecting $N(1,1)$ often leads to extreme propensity scores and an underestimation of the standard error. Possibly a better choice for the bandwidth could improve inference, but this is beyond the scope of this paper. Finally, as expected, when both nuisance functions are mis-specified all estimators exhibit under-coverage.



(a) $r_a = 0.3$ and $r_b = 0.3$.

(b) $r_a = 0.3$ and $r_b = 0$.

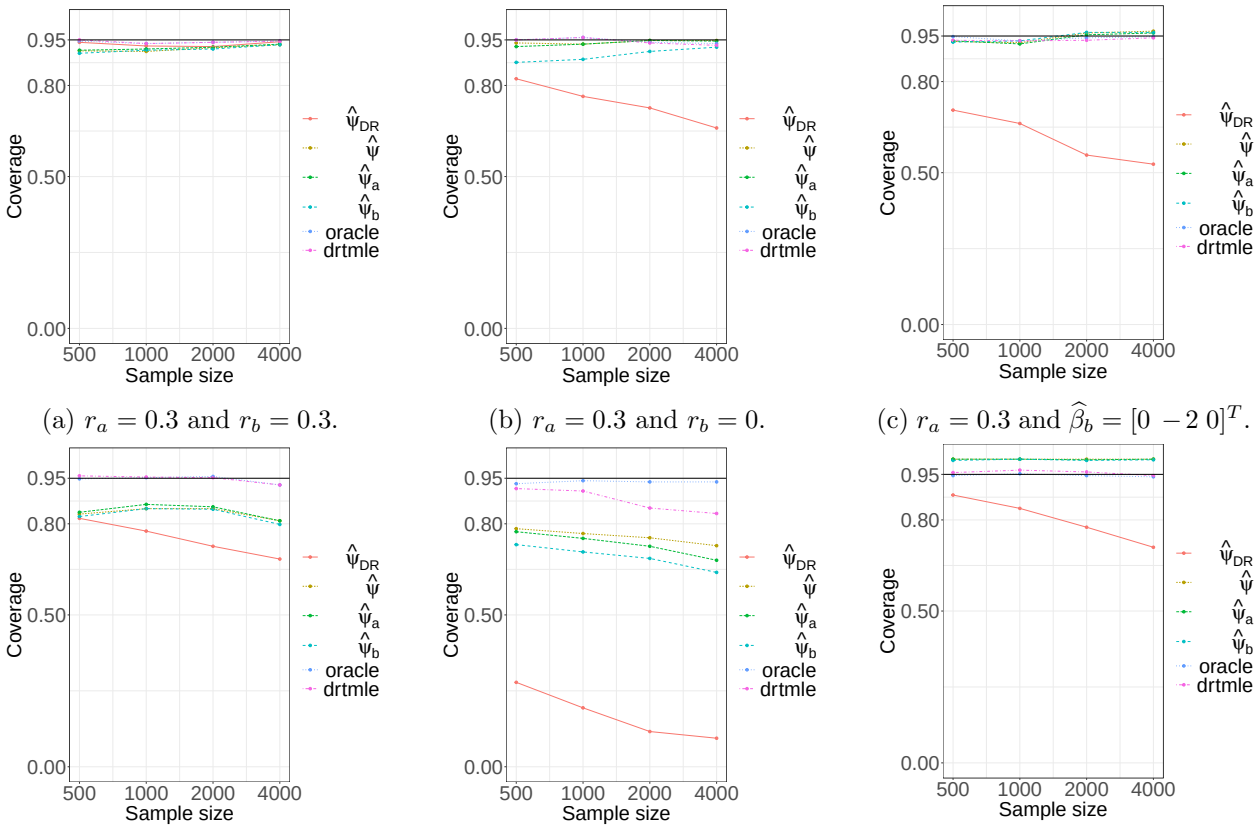
(c) $r_a = 0.3$ and $\widehat{\beta}_b = [0\ -2\ 0]^T$.

(d) $r_a = 0$ and $r_b = 0.3$.

(e) $r_a = 0$ and $r_b = 0$.

(f) $\widehat{\beta}_a = [0\ -2\ 0]^T$ and $r_b = 0.3$.

Figure 1: Distribution of the errors scaled by $\sqrt{n}$, where $n = 2000$. "Or" stands for "oracle," which refers to the estimator that has access to the true nuisance functions.



(a) $r_a = 0.3$ and $r_b = 0.3$. (b) $r_a = 0.3$ and $r_b = 0$. (c) $r_a = 0.3$ and $\widehat{\beta}_b = [0\ -2\ 0]^T$.

(d) $r_a = 0$ and $r_b = 0.3$. (e) $r_a = 0$ and $r_b = 0$. (f) $\widehat{\beta}_a = [0\ -2\ 0]^T$ and $r_b = 0.3$.

Figure 2: Coverage as a function of the sample size n. .

6 Conclusions

In this work, we have investigated the possibility of constructing estimators of the ATE functional that remain asymptotically linear even if one of the two nuisance functions is not consistently estimated. We have proposed a novel function class that is a hybrid between the purely structure-agnostic one proposed in Balakrishnan *et al.* 2023, and analyzed in detail in Jin & Syrgkanis 2024, and more traditional ones based on Hölder smoothness. In completely structure-agnostic models, where all that is known are rates of convergence for the two nuisance functions, our Proposition 1 (covering the case when the propensity score can be estimated at an equal or faster rate than the outcome model), and, more generally the concurrent work by Jin & Syrgkanis 2024, show that nonparametric, doubly-robust root-n inference is not attainable. On the contrary, we show that this is possible in the new hybrid class proposed. Further, merging ideas from the literature on doubly-robust inference and that on HOIFs, we have constructed an estimator that, under certain conditions, exhibits doubly-robust asymptotic linearity (DRAL). The sufficient conditions needed for DRAL that we have derived are relatively straightforward to describe, although they pertain to nonstandard regression models where both the outcome and the covariates depend on a training sample used to estimate the nuisance functions. In addition, we have shown that this estimator is minimax optimal in the new hybrid model proposed. In future work, it would be interesting to expand the results presented in Proposition 2 to cover cases where, for example, the additional nuisance functions involving generated regressors are Hölder smooth of orders less than

1. Furthermore, it would be also interesting to study other nonparametric models where these additional nuisance functions satisfy structure-agnostic rate conditions instead of the smoothness constrains that we have imposed.

7 Acknowledgements

EK gratefully acknowledges support from NIH R01 Grant LM013361-01A1. OD was supported by FWO grant 1222522N. MB thanks the participants at the Center for Causal Inference Seminar Series at the University of Pennsylvania for helpful discussions.

References

1. Abadie, A. & Imbens, G. W. Matching on the estimated propensity score. *Econometrica* **84,** 781–807 (2016).
2. Balakrishnan, S., Kennedy, E. H. & Wasserman, L. The Fundamental Limits of Structure-Agnostic Functional Estimation. *arXiv preprint arXiv:2305.04116* (2023).
3. Benkeser, D., Carone, M., Laan, M. V. D. & Gilbert, P. B. Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104,** 863–880 (2017).
4. Benkeser, D. & Hejazi, N. S. Doubly-robust inference in R using drtmle. *Observational Studies* **9,** 43–78 (2023).
5. Bickel, P. J. & Ritov, Y. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A,* 381–393 (1988).
6. Birgé, L. & Massart, P. Estimation of integral functionals of a density. *The Annals of Statistics* **23,** 11–29 (1995).
7. Bowman, A. W. & Azzalini, A. *R package* `sm`*: nonparametric smoothing methods (version 2.2-6.0)* (University of Glasgow, UK and Università di Padova, Italia, 2024). http://www.stats.gla.ac.uk/~adrian/sm/.
8. Diaz, I., Carone, M. & van der Laan, M. J. Second-order inference for the mean of a variable missing at random. *The international journal of biostatistics* **12,** 333–349 (2016).
9. Dukes, O., Vansteelandt, S. & Whitney, D. On doubly robust inference for double machine learning. *arXiv preprint arXiv:2107.06124* (2021).
10. Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics* **86,** 4–29 (2004).
11. Jin, J. & Syrgkanis, V. Structure-agnostic Optimality of Doubly Robust Learning for Treatment Effect Estimation. *arXiv preprint arXiv:2402.14264* (2024).
12. Kennedy, E. H. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469* (2022).
13. Kennedy, E. H., Balakrishnan, S. & G'Sell, M. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics* **48,** 2008–2030 (2020).
14. Laurent, B. Efficient estimation of integral functionals of a density. *The Annals of Statistics* **24,** 659–681 (1996).
15. Laurent, B. Estimation of integral functionals of a density and its derivatives. *Bernoulli,* 181–211 (1997).
16. Liu, L. & Li, C. New $\sqrt{n}$-consistent, numerically stable higher-order influence function estimators. *arXiv preprint arXiv:2302.08097* (2023).
17. Liu, L., Mukherjee, R., Newey, W. K. & Robins, J. M. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577* (2017).

18. Mammen, E., Rothe, C. & Schienle, M. Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics* **40,** 1132–1170 (2012).
19. Newey, W. K. Semiparametric efficiency bounds. *Journal of applied econometrics* **5,** 99–135 (1990).
20. Robins, J., Li, L., Mukherjee, R., Tchetgen, E. T. & van der Vaart, A. Higher order estimating equations for high-dimensional models. *Annals of statistics* **45,** 1951 (2017).
21. Robins, J., Li, L., Mukherjee, R., Tchetgen, E. T. & van der Vaart, A. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics* **45,** 1951–1987 (2017).
22. Robins, J., Li, L., Tchetgen, E. & van der Vaart, A. *Higher order influence functions and minimax estimation of nonlinear functionals* in *Probability and statistics: essays in honor of David A. Freedman* 335–421 (Institute of Mathematical Statistics, 2008).
23. Robins, J., Li, L., Tchetgen, E. & van der Vaart, A. W. Quadratic semiparametric von mises calculus. *Metrika* **69,** 227–247 (2009).
24. Robins, J., Tchetgen, E. T., Li, L. & van der Vaart, A. Semiparametric minimax rates. *Electronic journal of statistics* **3,** 1305 (2009).
25. Robins, J. M., Mark, S. D. & Newey, W. K. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics,* 479–495 (1992).
26. Robins, J. M., Rotnitzky, A. & Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* **89,** 846–866 (1994).
27. Rubin, D. B. Inference and missing data. *Biometrika* **63,** 581–592 (1976).
28. Sperlich, S. A note on non-parametric estimation with predicted variables. *The Econometrics Journal* **12,** 382–395 (2009).
29. Tan, Z. On doubly robust estimation for logistic partially linear models. *Statistics & Probability Letters* **155,** 108577 (2019).
30. Tsiatis, A. A. *Semiparametric theory and missing data* (Springer Science & Business Media, 2006).
31. Tsybakov, A. B. *Introduction to nonparametric estimation* (Springer Science & Business Media, 2008).
32. Van der Laan, M., Wang, Z. & van der Laan, L. Higher order targeted maximum likelihood estimation. *arXiv preprint arXiv:2101.06290* (2021).
33. Van der Laan, M. J. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The international journal of biostatistics* **10,** 29–57 (2014).
34. Van der Laan, M. J. & Rose, S. *Targeted learning: causal inference for observational and experimental data* (Springer Science & Business Media, 2011).
35. Van der Laan, M. J. & Rose, S. *Targeted learning in data science* (Springer, 2018).
36. Van der Laan, M. J., Rose, S., Carone, M., Díaz, I. & van der Laan, M. J. Higher-order targeted loss-based estimation. *Targeted learning in data science: causal inference for complex longitudinal studies,* 483–510 (2018).
37. Vansteelandt, S. & Joffe, M. Structural Nested Models and G-estimation: The Partially Realized Promise. *Statistical Science,* 707–731 (2014).

A Example: partially linear logistic model

We revisit Example 3 in Dukes *et al.* 2021. Here, the goal is to derive an estimator of θ under the model

$$\text{logit}\mathbb{P}(Y = 1 \mid A, X) = \theta_0 A + m_0(X),$$

where $A \in \mathbb{R}$ and $m_0(X) = \mathbb{E}(Y \mid A = 0, X)$. Tan 2019 showed that a doubly-robust estimator of θ can be found by solving the empirical version of the moment condition

$$\psi(\theta_0) = \mathbb{E}\left[\{A - v(X)\}\{Ye^{-\theta_0 A - m_0(X)} - (1 - Y)\}\right] = 0$$

where $v(X) = \mathbb{E}(A \mid Y = 0, X)$. By the usual decomposition for M-estimators, the conditional bias of $\widehat{\theta}$ solving the empirical version of the moment condition above can be derived from the conditional bias of the moment condition itself. Tan 2019 has shown that such conditional bias is

$$R_n = \mathbb{E}\left[(1 - Y)\left(e^{m_0(X) - \widehat{m}_0(X)} - 1\right)\{v(X) - \widehat{v}(X)\} \mid D^n\right],$$

which is not of the variety of remainder terms studied in Dukes *et al.* 2021. Our work shows that, if one is willing to assume smoothness of the functions

$$f_v(t_1, t_2, t_3; D^n) = \mathbb{E}\left\{e^{m_0(X) - \widehat{m}_0(X)} - 1 \mid Y = 0, \widehat{v}(X) = t_1, v(X) = t_2, \widehat{m}_0(X) = t_3, D^n\right\}$$
$$f_m(t_1, t_2, t_3; D^n) = \mathbb{E}\left\{v(X) - \widehat{v}(X) \mid Y = 0, \widehat{m}_0(X) = t_1, m_0(X) = t_2, \widehat{v}(X) = t_3, D^n\right\},$$

then it is possible to derive an estimator of θ that admits doubly-robust asymptotic linearity. In fact, such estimator would be based on an augmented moment condition of the same variety as the estimator considered in Section 3.3. In particular, let $\widehat{\theta}$ solve

$$\widehat{\psi}(\widehat{\theta}) = \mathbb{P}_n\left[\{A - \widehat{v}(X)\}\{Ye^{-\widehat{\theta}A - \widehat{m}_0(X)} - (1 - Y)\}\right] - T_n, \text{ where}$$
$$T_n = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \left\{e^{-\theta A_i - \widehat{m}_i} Y_i - (1 - Y_i)\right\} \frac{K_h(\widehat{v}_i - \widehat{v}_j) K_h(\widehat{m}_i - \widehat{m}_j)}{\widehat{Q}(\widehat{v}_i, \widehat{m}_i)} (1 - Y_j)(A_j - \widehat{v}_j),$$
$$\widehat{Q}(\widehat{v}_i, \widehat{m}_i) = \frac{1}{n-1} \sum_{s=1, s \neq i}^{n} (1 - Y_s) K_h(\widehat{v}_s - \widehat{v}_i) K_h(\widehat{m}_s - \widehat{m}_i),$$

where we recall the notation $\widehat{v}_i = \widehat{v}(X_i)$, $\widehat{m}_i = \widehat{m}_0(X_i)$.

A proposition analogous to Proposition 4 can be derived for $\psi(\theta_0) - \widehat{\psi}(\theta_0)$. To see this, notice that

$$R_n = \mathbb{E}\{(1 - Y) f_v(\widehat{v}, v, \widehat{m}_0; D^n)(v - \widehat{v}) \mid D^n\} = \mathbb{E}\left\{(1 - Y)\left(e^{m_0 - \widehat{m}} - 1\right) f_m(\widehat{m}_0, m_0, v; D^n) \mid D^n\right\}$$

Furthermore, for $Q(\widehat{v}_i, \widehat{m}_i) = \mathbb{E}\{\widehat{Q}(\widehat{v}_i, \widehat{m}_i) \mid X_i, D^n\}$:

$$\begin{aligned}
&\mathbb{E}_{Z_j|D^n}\left\{\frac{K_h(\widehat{v}_i - \widehat{v}_j)K_h(\widehat{m}_i - \widehat{m}_j)}{Q(\widehat{v}_j, \widehat{m}_j)}(1-Y_j)(A_j - \widehat{v}_j)\right\} \\
&= \mathbb{E}_{Z_j|D^n}\left\{\frac{K_h(\widehat{v}_i - \widehat{v}_j)K_h(\widehat{m}_i - \widehat{m}_j)}{Q(\widehat{v}_i, \widehat{m}_i)}(1-Y_j)(v_j - \widehat{v}_j)\right\} \\
&= \mathbb{E}_{Z_j|D^n}\left\{\frac{K_h(\widehat{v}_i - \widehat{v}_j)K_h(\widehat{m}_i - \widehat{m}_j)}{Q(\widehat{v}_i, \widehat{m}_i)}(1-Y_j)f_m(\widehat{m}_j, m_j, \widehat{v}_j; D^n)\right\} \\
&= f_m(\widehat{m}_i, m_i, \widehat{v}_j; D^n) + O(h^\alpha + \|\widehat{m} - m\|^\alpha)
\end{aligned}$$

where the last equality follows if $f_m(\widehat{m}, m, \widehat{v}; D^n)$ is Hölder of order α. Similarly, for $Q(\widehat{v}_j, \widehat{v}_j) = \mathbb{E}\{(1-Y)K_h(\widehat{v} - \widehat{v}_j)K_h(\widehat{m} - \widehat{m}_j) \mid D^n\}$:

$$\begin{aligned}
&\mathbb{E}_{Z_i|D^n}\left[\left\{e^{-\theta A_i - \widehat{m}_i}Y_i - (1-Y_i)\right\}\frac{K_h(\widehat{v}_i - \widehat{v}_j)K_h(\widehat{m}_i - \widehat{m}_j)}{Q(\widehat{v}_j, \widehat{m}_j)}\right] \\
&= \mathbb{E}_{Z_i|D^n}\left\{\frac{e^{m_i - \widehat{m}_i} - 1}{1 + e^{\theta A_i + m_i}}\frac{K_h(\widehat{v}_i - \widehat{v}_j)K_h(\widehat{m}_i - \widehat{m}_j)}{Q(\widehat{v}_j, \widehat{m}_j)}\right\} \\
&= \mathbb{E}_{Z_1|D^n}\left\{(1-Y_i)\left(e^{m_i - \widehat{m}_i} - 1\right)\frac{K_h(\widehat{v}_i - \widehat{v}_j)K_h(\widehat{m}_i - \widehat{m}_j)}{Q(\widehat{v}_j, \widehat{m}_j)}\right\} \\
&= \mathbb{E}_{Z_i|D^n}\left\{(1-Y_i)f_v(\widehat{v}_i, v_i, \widehat{m}_i; D^n)\frac{K_h(\widehat{v}_i - \widehat{v}_j)K_h(\widehat{m}_i - \widehat{m}_j)}{Q(\widehat{v}_j, \widehat{m}_j)}\right\} \\
&= f_v(\widehat{v}_j, v_j, \widehat{m}_j; D^n) + O(h^\beta + \|\widehat{v} - v\|^\beta)
\end{aligned}$$

where the last equality follows if $f_v(\widehat{v}, v, \widehat{m}; D^n)$ is Hölder of order β. The errors $\widehat{Q}(\widehat{v}_i, \widehat{m}_i) - Q(\widehat{v}_i.\widehat{m}_i)$ and $Q(\widehat{v}_i, \widehat{m}_i) - Q(\widehat{v}_j, \widehat{m}_j)$, as well as the variance calculations, can be carried out in a way analogous to that used to prove Proposition 4.

B Proofs regarding the lower bounds

B.1 Useful Lemmas

First, we recall the definition of Hellinger distance. Let P and Q be two probability measures with densities p and q relative to some σ-finite dominating measure ν.

Definition 2 (Hellinger distance). The Hellinger distance between P and Q is

$$H(P, Q) = \left\{\int (\sqrt{p} - \sqrt{q})^2 d\nu\right\}^{1/2}.$$

The following Lemma is a restatement of Theorem 2.15 in Tsybakov 2008 and Lemma 1 in Balakrishnan *et al.* 2023 with a few simplifications tailored to our settings.

Lemma 1 (Lemma 1 in Balakrishnan *et al.* 2023 / Theorem 2.15 in Tsybakov 2008)**.** *Let $F(\theta)$ be a functional and $\{P_\theta, \theta \in \Theta\}$ be a statistical model. Consider two priors distributions μ_0 and μ_1 on*

Θ *and define the posteriors:*

$$P_j(A) = \int P_\theta^n(A)\mu_j(d\theta), \text{ for all measurable } A \text{ and } j \in \{0,1\}.$$

Assume that

1. *There exists* c *and* $s > 0$ *such that* $\mu_0(\theta : F(\theta) \leq c) = 1$ *and* $\mu_1(\theta : F(\theta) \geq c + 2s) = 1$.
2. *The Hellinger distance satisfies* $H^2(P_1, P_0) \leq \alpha < 2$.

Then, it holds that

$$\inf_{T_n} \sup_{\theta\in\Theta} \mathbb{E}|T_n - F(\theta)| \geq s \cdot \frac{1 - \sqrt{\alpha(1-\alpha/4)}}{2}.$$

Lemma 2 (Theorem 2.1 in Robins *et al.* 2009b)**.** *Let* k *in* $\mathbb{N}$, *let* $\mathcal{X} = \cup_{j=1}^k \mathcal{X}_j$ *be a measurable partition of the sample space. Given a vector* $\lambda = (\lambda_1, \ldots, \lambda_k)$ *in some product measurable space* $\Lambda = \lambda_1 \times \cdots \lambda_k$ *let* P_λ *and* Q_λ *be probablity measures on* $\mathcal{X}$ *such that*

1. $P_\lambda(\mathcal{X}_j) = Q_\lambda(X_j) = p_j$ *for every* $\lambda \in \Lambda$, *for some probability vector* $(p_1, \ldots, p_k)$.
2. *The restrictions of* P_λ *and* Q_λ *to* $\mathcal{X}_j$ *depend on the* j*th coordinate* λ_j *of* $\lambda = (\lambda_1, \ldots, \lambda_k)$ *only.*

For p_λ *and* q_λ *densities of the measures* P_λ *and* Q_λ *that are jointly measurable in the parameter* λ *and the observation, and* π *a probability measure on* Λ, *define* $p = \int p_\lambda d\pi(\lambda)$ *and* $q = \int q_\lambda d\pi(\lambda)$, *and set*

$$\delta_1 = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(p_\lambda - p)^2}{p_\lambda}\frac{d\nu}{p_j}, \quad \delta_2 = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(q_\lambda - p_\lambda)^2}{p_\lambda}\frac{d\nu}{p_j}, \quad \delta_3 = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(q - p)^2}{p_\lambda}\frac{d\nu}{p_j}.$$

If $np_j(1 \vee \delta_1 \vee \delta_2) \leq A$ *for all* j *and* $\underline{B} \leq p_\lambda \leq \overline{B}$ *for positive constants* A, $\underline{B}$, $\overline{B}$, *then there exists a constant* C *that depends only on* A, $\underline{B}$, $\overline{B}$ *such that, for any product probability measure* $\pi = \pi_1 \otimes \cdots \otimes p_k$,

$$H^2(P_1, P_2) \leq Cn^2(\max_j p_j)(\delta_2^2 + \delta_1\delta_2) + Cn\delta_3.$$

We proceed essentially as in the proof of Theorem 3.1 in Robins *et al.* 2009b, but choose certain parameters and the fluctuations in their construction differently in accordance with the assumptions defining $\mathcal{P}(\epsilon_n, \delta_n)$, $\mathcal{P}_a(\epsilon_n)$ and $\mathcal{P}_b(\delta_n)$, and $\mathcal{P}_{ab}(\epsilon_n, \delta_n)$. The construction is conceptually very similar to the one used in Balakrishnan *et al.* 2023. The only changes pertain to the choice of the fluctuations for the nuisance parameters, which are tailored to the ATE functional considered here.

Let $B : \mathbb{R}^d \to \mathbb{R}$ be a function with support on $[0, 1/2]^d$ such that $\int B(u)du = 0$ and $\int B^2(u)du = 1$. Let k be an integer and $\mathcal{X}_1, \ldots, \mathcal{X}_k$ be traslates of the cube $k^{-1/d}[0, 1/2]^d$ that are disjoint and contained in $[0, 1]^d$. Let $m_1, \ldots, m_k$ be the bottom left corners of these cubes. Let $\lambda = (\lambda_1, \ldots, \lambda_k) \in \{-1, 1\}^k$. For shorthand notation, let $B_j(x) = B\{k^{1/d}(x - m_j)\}$.

B.2 Proof of Proposition 1

Suppose that $\epsilon_n \leq \delta_n$ and define:

$$a_{p_\lambda}(x) = a_{q_\lambda}(x) = \widehat{a}(x) + \epsilon_n \sum_{j=1}^{k} \lambda_j B_j(x), \quad b_{p_\lambda}(x) = \widehat{b}(x) - \epsilon_n \frac{\widehat{b}(x)}{\widehat{a}(x)} \sum_{j=1}^{k} \lambda_j B_j(x),$$

$$\text{and} \quad b_{q_\lambda}(x) = b_{1\lambda}(x) + \frac{\delta_n}{\widehat{a}(x)} \sum_{j=1}^{k} \lambda_j B_j(x).$$

We have $\|\widehat{a} - a_{p_\lambda}\|_2 = \|\widehat{a} - a_{q_\lambda}\| = \epsilon_n$, $\|\widehat{b} - \widehat{b}_{p_\lambda}\| \lesssim \epsilon_n \leq \delta_n$, and $\|\widehat{b} - b_{q_\lambda}\| \lesssim \delta_n$. Consider the densities $p_\lambda(x)$ and $q_\lambda(x)$ for $Y, A \in \{0,1\}$ and $X \in [0,1]^d$:

$$p_\lambda(Y,A,X) = g(X)\{a_{p_\lambda}(X) - 1\}^{1-A}[b_{p_\lambda}(X)^Y\{1 - b_{p_\lambda}(X)\}^{1-Y}]^A$$
$$q_\lambda(Y,A,X) = g(X)\{a_{q_\lambda}(X) - 1\}^{1-A}[b_{q_\lambda}(X)^Y\{1 - b_{q_\lambda}(X)\}^{1-Y}]^A$$

By construction, p_λ and q_λ are part of $\mathcal{P}(\epsilon_n, \delta_n)$. Set $g(X) = g = \{\int \widehat{a}(x)dx\}^{-1}$ so that the density of X, $f_{p_\lambda}(x) = f_{q_\lambda}(x) = a_{p_\lambda}(x)g(x) = a_{q_\lambda}(x)g(x)$, integrates to 1. We have

$$\psi_{p_\lambda} = g\int a_{p_\lambda}(x)b_{p_\lambda}(x)dx = g\int \widehat{a}(x)\widehat{b}(x)dx - g\epsilon_n^2 \sum_{j=1}^{k} \int B_j^2(x)\frac{\widehat{b}(x)}{\widehat{a}(x)}dx$$

$$\psi_{q_\lambda} = \int a_{q_\lambda}(x)b_{q_\lambda}(x)g(x)dx = \psi_{p_\lambda} + g\delta_n\epsilon_n \int \sum_{j=1}^{k} B_j^2(x)\{\widehat{a}(x)\}^{-1}dx$$

This means that $|\psi_{q_\lambda} - \psi_{p_\lambda}| \gtrsim \epsilon_n\delta_n$ for every λ. Let $\omega(\lambda)$ denote a product prior on $\lambda = (\lambda_1, \ldots, \lambda_k)$ such that $\omega_j(\lambda = -1) = \omega_j(\lambda = 1) = 1/2$. Let $O = (X, A, Y)$ and define $\mathcal{O} = \cup_{j=1}^{k}\mathcal{O}_j$, where $\mathcal{O}_j = \mathcal{X}_j \times \{0,1\}^2$. Notice that $p_j = \int_{\mathcal{O}_j} p_\lambda d\nu = \int_{\mathcal{O}_j} q_\lambda d\nu = k^{-1}$. Next, we have

$$p(O) := \int p_\lambda(Y,A,X)d\omega(\lambda) = g(X)\{\widehat{a}(X) - 1\}^{1-A}[\widehat{b}(X)^Y\{1 - \widehat{b}(X)\}^{1-Y}]^A,$$
$$p_\lambda(O) - p(O) = g(X)\{a_{p_\lambda}(X) - \widehat{a}(X)\}^{1-A}[\{b_{p_\lambda}(X) - \widehat{b}(X)\}^Y\{\widehat{b}(X) - b_{p_\lambda}(X)\}^{1-Y}]^A,$$
$$q_\lambda(O) - p_\lambda(O) = Ag(X)\{b_{q_\lambda}(X) - b_{p_\lambda}(X)\}^Y\{b_{p_\lambda}(X) - b_{q_\lambda}(X)\}^{1-Y},$$
$$q(O) - p(O) := \int \{q_\lambda(Y,A,X) - p_\lambda(Y,A,X)\}d\omega(\lambda).$$

Next, we apply Lemma 2. Notice that $\delta_1 \lesssim \epsilon_n^2$, $\delta_2 \lesssim \delta_n^2$ and $\delta_3 = 0$. Therefore,

$$H^2\left(\int p_\lambda^n d\omega(\lambda), \int q_\lambda^n d\omega(\lambda)\right) \lesssim n^2k^{-1}(\delta_n^4 + \delta_n^2\epsilon_n^2).$$

In this light, choosing k large enough yields that the Hellinger distance is bounded. The lower bound then follows by Lemma 1.

B.3 Proof of Proposition 2

Claim 1. To prove this claim, we modify the definitions of p_λ and q_λ. Let $a_{p_\lambda}(x) = \widehat{a}(x)$, $a_{q_\lambda}(x) = \widehat{a}(x) + \epsilon_n \sum_{j=1}^k \lambda_j B_j(x)$, $b_{p_\lambda}(x) = 1/\widehat{a}(x)$,

$$b_{q_\lambda}(x) = \frac{1}{\widehat{a}(x)} - \frac{a_{q_\lambda}(x) - \widehat{a}(x)}{\widehat{a}^2(x)} = \frac{1}{\widehat{a}(x)} - \frac{\epsilon_n \sum_{j=1}^k \lambda_j B_j(x)}{\widehat{a}^2(x)},$$

and $g(x) = g = \{\int \widehat{a}(x)dx\}^{-1}$. In this light, $\mathbb{E}_{p_\lambda}\{b_{p_\lambda}(X) \mid A = 1, a_{p_\lambda}(X) = t_1, \widehat{a}(X) = t_2, D^n\}$ and $\mathbb{E}_{q_\lambda}\{b_{q_\lambda}(X) \mid A = 1, a_{q_\lambda}(X) = t_1, \widehat{a}(X) = t_2, D^n\}$ are infinitely smooth functions in t_1 and t_2. Further, $\|\widehat{a} - a_{p_\lambda}\| = 0$ and $\|\widehat{a} - a_{q_\lambda}\| = \epsilon_n$, so p_λ and q_λ belong to $\mathcal{P}_a(\epsilon_n)$.

With these modifications in place, we have $\psi_{p_\lambda} = g$ and

$$\psi_{q_\lambda} = \psi_{p_\lambda} - g\epsilon_n^2 \sum_{j=1}^k \int \frac{B_j^2(x)}{\widehat{a}^2(x)} dx,$$

so that $|\psi_{p_\lambda} - \psi_{q_\lambda}| \gtrsim \epsilon_n^2$ for every λ. Following the same of reasoning to prove Proposition 1, we have $\delta_1 = 0$, $\delta_2 \lesssim \epsilon_n^2$ and $\delta_3 = 0$. Therefore, for k large enough, Claim 1 follows.

Claim 2. To prove Claim 3, we modify p_λ and q_λ. We set $a_{p_\lambda}(x) = \{\widehat{b}(x)\}^{-1}$, $b_{p_\lambda}(x) = \widehat{b}(x)$,

$$a_{q_\lambda}(x) = \frac{1}{\widehat{b}(x)} + \frac{b_{q_\lambda}(x) - \widehat{b}(x)}{\widehat{b}^2(x)} = \frac{1}{\widehat{b}(x)} + \delta_n \sum_{j=1}^k \lambda_j B_j(x), \quad \text{and}$$

$$b_{q_\lambda}(x) = \widehat{b}(x) - \widehat{b}^2(x)\delta_n \sum_{j=1}^k \lambda_j B_j(x)$$

Under this construction, both p_λ and q_λ are contained in $\mathcal{P}_b(\delta_n)$, because $\|\widehat{b} - b_{p_\lambda}\| = 0$, $\|\widehat{b} - b_{q_\lambda}\| \lesssim \delta_n$, as well as $\mathbb{E}_{p_\lambda}\{a_{p_\lambda}(X) \mid A = 1, \widehat{b}(X), b_{p_\lambda}(X), D^n\} = \{\widehat{b}(X)\}^{-1}$ and $\mathbb{E}_{q_\lambda}\{a_{q_\lambda}(X) \mid A = 1, \widehat{b}(X), b_{q_\lambda}(X), D^n\} = \{\widehat{b}(X)\}^{-1} + \{b_\lambda(X) - \widehat{b}(X)\}\{\widehat{b}^2(X)\}^{-1}$, which are both infinitely smooth functions of $\widehat{b}$ and b. We have

$$\psi_{p_\lambda} = g, \quad \text{and} \quad \psi_{q_\lambda} = \psi_{p_\lambda} - g\delta_n^2 \sum_{j=1}^k \int \widehat{b}^2(x) B_j^2(x) dx$$

where $g = \{\int 1/\widehat{b}(x)dx\}^{-1}$ under both q_λ and p_λ. Therefore, for any λ, $|\psi_{p_\lambda} - \psi_{q_\lambda}| \gtrsim \delta_n^2$. Because

$$\int a_{p_\lambda}(x)d\omega(\lambda) = \int a_{q_\lambda}(x)d\omega(\lambda) = \frac{1}{\widehat{b}(x)}$$
$$\int b_{p_\lambda}(x)d\omega(\lambda) = \int b_{q_\lambda}(x)d\omega(\lambda) = \widehat{b}(x),$$

we have $\delta_1 = 0$, $\delta_2 \lesssim \delta_n^2$ and $\delta_3 = 0$. In this light, for k large enough, by proceeding as in the proof of Proposition 1, Claim 2 follows.

Claim 3. Let $\gamma_n = \epsilon_n \wedge \delta_n$ and define

$$a_{p_\lambda}(x) = a_{q_\lambda}(x) = \widehat{a}(x) + \gamma_n \sum_{j=1}^{k} \lambda_j B_j(x), \quad b_{p_\lambda}(x) = \widehat{b}(x) - \gamma_n \frac{\widehat{b}(x)}{\widehat{a}(x)} \sum_{j=1}^{k} \lambda_j B_j(x)$$

$$\text{and} \quad b_{q_\lambda}(x) = b_{p_\lambda}(x) + \frac{\gamma_n}{\widehat{a}(x)} \sum_{j=1}^{k} \lambda_j B_j(x).$$

We have $\|\widehat{a} - a_{p_\lambda}\| = \|\widehat{a} - a_{q_\lambda}\| \leq \gamma_n \leq \epsilon_n$, $\|b_{p_\lambda} - \widehat{b}\| \lesssim \gamma_n \leq \delta_n$, and $\|b_{q_\lambda} - \widehat{b}\| \lesssim \gamma_n \leq \delta_n$. Furthermore,

$$a_{p_\lambda}(x) = a_{q_\lambda}(x) = \widehat{a}(x) + \frac{\{\widehat{b}(x) - b_{p_\lambda}(x)\}\widehat{a}(x)}{\widehat{b}(x)} = \widehat{a}(x) + \frac{\{b_{q_\lambda}(x) - \widehat{b}(x)\}\widehat{a}(x)}{1 - \widehat{b}(x)},$$

$$b_{p_\lambda}(x) = \widehat{b}(x) - \frac{\widehat{b}(x)\{a_{p_\lambda}(x) - \widehat{a}(x)\}}{\widehat{a}(x)}, \quad \text{and} \quad b_{q_\lambda}(x) = \widehat{b}(x) + \frac{\{1 - \widehat{b}(x)\}\{a_{q_\lambda}(x) - \widehat{a}(x)\}}{\widehat{a}(x)}.$$

Therefore, $\mathbb{E}\{a_{p_\lambda}(X) \mid A = 1, \widehat{b}(X) = t_1, b_{p_\lambda}(X) = t_2, \widehat{a}(X) = t_3, D^n\}$, $\mathbb{E}\{a_{q_\lambda}(X) \mid A = 1, \widehat{b}(X) = t_1, b_{q_\lambda}(X) = t_2, \widehat{a}(X) = t_3, D^n\}$, $\mathbb{E}\{b_{p_\lambda}(X) \mid A = 1, \widehat{a}(X) = t_1, a_{p_\lambda}(X) = t_2, \widehat{b}(X) = t_3, D^n\}$, as well as $\mathbb{E}\{b_{q_\lambda}(X) \mid A = 1, \widehat{a}(X) = t_1, a_{q_\lambda}(X) = t_2, \widehat{b}(X) = t_3, D^n\}$ are all smooth functions. In this respect, p_λ and q_λ belong to $\mathcal{P}_{ab}(\epsilon_n, \delta_n)$. Set $g(X) = g = \{\int \widehat{a}(x)dx\}^{-1}$ and define the densities:

$$p_\lambda(Y, A, X) = g(X)\{a_{p_\lambda}(X) - 1\}^{1-A}[b_{p_\lambda}(X)^Y\{1 - b_{p_\lambda}(X)\}^{1-Y}]^A,$$
$$q_\lambda(Y, A, X) = g(X)\{a_{q_\lambda}(X) - 1\}^{1-A}[b_{q_\lambda}(X)^Y\{1 - b_{q_\lambda}(X)\}^{1-Y}]^A.$$

We have

$$\psi_{p_\lambda} = g \int a_\lambda(x) b_{1\lambda}(x) dx = g \int \widehat{a}(x)\widehat{b}(x)dx - g\gamma_n^2 \sum_{j=1}^{k} \int \frac{\widehat{b}(x)}{\widehat{a}(x)} B_j^2(x)dx,$$

$$\psi_{q_\lambda} = \psi_{p_\lambda} + g\gamma_n^2 \sum_{j=1}^{k} \int \frac{B_j^2(x)}{\widehat{a}(x)} dx,$$

so that $|\psi_{p_\lambda} - \psi_{q_\lambda}| \gtrsim \gamma_n^2 = \epsilon_n^2 \wedge \delta_n^2$ for every λ. To conclude the proof, it is sufficient to note that $\delta_1^2 \lesssim \gamma_n^2$, $\delta_2 \lesssim \gamma_n^2$, and $\delta_3 = 0$ and then follow the arguments made to establish Proposition 1.

C Proofs regarding the upper bounds

C.1 Useful lemmas

Lemma 3 (Lemma 6 from Robins *et al.* 2009a)**.** *For any measurable function $f : \mathcal{X}^2 \to \mathbb{R}$, and $f_1(x_1) = \int f(x_1, x_2) d\mathbb{P}(x_2)$,*

$$var\left\{ \frac{1}{n(n-1)} \sum\sum_{1 \leq i \neq j \leq n} f(X_i, X_j) \right\} \leq \frac{4}{n}\mathbb{P}f_1^2 + \frac{2}{n(n-1)}\mathbb{P}f^2.$$

Lemma 4 (Lemma 2 from Kennedy *et al.* 2020)**.** *Let $\widehat{f}(o)$ be a function estimated from a sample $O^N = (O_{n+1}, \ldots, O_N)$ and let $\mathbb{P}_n$ denote the empirical measure over $(O_1, \ldots, O_n)$, which is independent of O^N. Then,*

$$(\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - f\|}{\sqrt{n}}\right).$$

C.2 Proof of Proposition 3: bias and variance of $\widehat{\psi}_a$

Recall the notation $\widehat{a} = \widehat{a}(X)$ and $\widehat{a}_i = \widehat{a}(X_i)$ and that

$$s_a(t_1, t_2; D^n) = \mathbb{E}(b - \widehat{b} \mid A = 1, \widehat{a} = t_1, a = t_2, D^n).$$

The estimator is $\widehat{\psi}_a = \mathbb{P}_n\widehat{\varphi} - T_n$, where $\varphi(O) = Aa(Y - b) + b$ and

$$T_n = \frac{1}{n(n-1)} \sum\sum_{1 \le i \ne j \le n} (A_i\widehat{a}_i - 1)\widehat{Q}^{-1}(\widehat{a}_i)K_h(\widehat{a}_j - \widehat{a}_i)A_j(Y_j - \widehat{b}_j)$$

for $\widehat{Q}(\widehat{a}_i) = (n-1)^{-1}\sum_{j=1, j\ne i}^n A_j K_h(\widehat{a}_j - \widehat{a}_i)$. From the decomposition (3), we have that $\mathbb{E}(\widehat{\psi} - \psi \mid D^n) = R_n - \mathbb{E}(T_n \mid D^n)$ and $\mathrm{var}(\widehat{\psi} \mid D^n) \lesssim n^{-1} \vee \mathrm{var}(T_n \mid D^n)$.

C.2.1 Bias of $\widehat{\psi}_a$

The bias bound follows exactly as described in the main text. We briefly report the calculation here for convenience. From the decomposition

$$A_j(Y_j - \widehat{b}_j) = A_j s_a(\widehat{a}_i, a_i; D^n) + A_j\{s_a(\widehat{a}_j, a_j; D^n) - s_a(\widehat{a}_i, a_i; D^n)\} + A_j\epsilon_j,$$

we obtain that

$$\begin{aligned} T_n = &\frac{1}{n}\sum_{i=1}^n (A_i\widehat{a}_i - 1)s_a(\widehat{a}_i, a_i; D^n) \\ &+ \frac{1}{n(n-1)} \sum\sum_{1 \le i \ne j \le n} (A_i\widehat{a}_i - 1)\widehat{Q}^{-1}(\widehat{a}_i)K_h(\widehat{a}_j - \widehat{a}_i)A_j\{s_a(\widehat{a}_j, a_j; D^n) - s_a(\widehat{a}_i, a_i; D^n) + \epsilon_j\}. \end{aligned}$$

Next, notice that

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n (A_i\widehat{a}_i - 1)s_a(\widehat{a}_i, a_i; D^n) \mid D^n\right\} = R_n$$

and that

$$\begin{aligned} \mathbb{E}(A_j\epsilon_j \mid A_i, A_j, \widehat{a}_j, \widehat{a}_i, D^n) &= \mathbb{E}(A_j\epsilon_j \mid A_j, \widehat{a}_j, D^n) \\ &= \mathbb{E}\{\mathbb{E}(A_j\epsilon_j \mid A_j, \widehat{a}_j, a_j, D^n) \mid A_j, \widehat{a}_j, D^n\} \\ &= \mathbb{E}(\mathbb{E}[A_j\{(Y_j - \widehat{b}_j) - s_a(\widehat{a}_j, a_j; D^n)\} \mid A_j, \widehat{a}_j, a_j, D^n] \mid A_j, \widehat{a}_j, D^n) \\ &= \mathbb{E}(\mathbb{E}[A_j\{(b_j - \widehat{b}_j) - s_a(\widehat{a}_j, a_j; D^n)\} \mid A_j, \widehat{a}_j, a_j, D^n] \mid A_j, \widehat{a}_j, D^n) \\ &= 0 \end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}(R_n - T_n \mid D^n)\\ &= \frac{1}{n(n-1)} \sum_{1\le i\neq j\le n}\sum \mathbb{E}\left[(A_i\widehat{a}_i - 1)\widehat{Q}^{-1}(\widehat{a}_i)K_h(\widehat{a}_j - \widehat{a}_i)A_j\{s_a(\widehat{a}_j, a_j; D^n) - s_a(\widehat{a}_i, a_i; D^n)\} \mid D^n\right]\end{aligned}$$

Under the assumption that $\widehat{Q}(\widehat{a}_i)$ is bounded away from zero, we have

$$\begin{aligned}&|\mathbb{E}(R_n - T_n \mid D^n)|\\ &\lesssim \frac{1}{n(n-1)} \sum_{1\le i\neq j\le n}\sum \mathbb{E}\left\{A_i|\widehat{a}_i - a_i|A_jK_h(\widehat{a}_j - \widehat{a}_i)|s_a(\widehat{a}_j, a_j; D^n) - s_a(\widehat{a}_i, a_i; D^n)|| \; D^n\right\}\end{aligned}$$

By the Hölder smoothness assumption,

$$\begin{aligned}|s_a(\widehat{a}_j, a_j; D^n) - s_a(\widehat{a}_i, a_i; D^n)| &\lesssim \{(\widehat{a}_j - \widehat{a}_i)^2 + (a_j - \widehat{a}_j)^2 + (\widehat{a}_i - a_i)^2\}^{\alpha/2}\\ &\lesssim |\widehat{a}_j - \widehat{a}_i|^\alpha + |\widehat{a}_j - a_j|^\alpha + |\widehat{a}_i - a_i|^\alpha\end{aligned}$$

We bound each of the 3 terms as follows. First, it holds that

$$\begin{aligned}\mathbb{E}\left\{A_i|\widehat{a}_i - a_i|A_jK_h(\widehat{a}_j - \widehat{a}_i)|\widehat{a}_j - \widehat{a}_i|^\alpha| \; D^n\right\} &\le h^\alpha \mathbb{E}\left[|\widehat{a}_i - a_i|\mathbb{E}\{A_jK_h(\widehat{a}_j - \widehat{a}_i) \mid X_i, D^n\} \mid D^n\right]\\ &\lesssim h^\alpha \|\widehat{a} - a\|,\end{aligned}$$

under the assumption that $\mathbb{E}\{A_jK_h(\widehat{a}_j - \widehat{a}_i) \mid X_i, D^n\} \lesssim 1$ and because $K(u)$ has support in $[-1, 1]$. The last equality follows because $\int |\widehat{a}(x) - a(x)|d\mathbb{P}(x) \le \|\widehat{a} - a\|$. Next, by the Cauchy-Schwarz inequality:

$$\begin{aligned}&\left[\mathbb{E}\left\{A_i|\widehat{a}_i - a_i|A_jK_h(\widehat{a}_j - \widehat{a}_i)|\widehat{a}_j - a_j|^\alpha| \; D^n\right\}\right]^2\\ &\le \mathbb{E}\left[(\widehat{a}_i - a_i)^2\mathbb{E}\{A_jK_h(\widehat{a}_j - \widehat{a}_i) \mid X_i, D^n\} \mid D^n\right] \mathbb{E}\left[(\widehat{a}_j - a_j)^{2\alpha}\mathbb{E}\{A_iK_h(\widehat{a}_j - \widehat{a}_i) \mid X_j, D^n\} \mid D^n\right]\\ &\lesssim \mathbb{E}\left\{(\widehat{a}_i - a_i)^2 \mid D^n\right\} \mathbb{E}\left\{(\widehat{a}_j - a_j)^{2\alpha} \mid D^n\right\}\\ &\lesssim \|\widehat{a} - a\|^{2+2\alpha}\end{aligned}$$

where the last inequality follows by Jensen's inequality because $|x|^\alpha$ is concave for $\alpha < 1$. Similarly,

$$\begin{aligned}\mathbb{E}\left\{(\widehat{a}_i - a_i)^{1+\alpha}A_jK_h(\widehat{a}_j - \widehat{a}_i) \mid D^n\right\} &= \mathbb{E}\left[(\widehat{a}_i - a_i)^{1+\alpha}\mathbb{E}\{A_jK_h(\widehat{a}_j - \widehat{a}_i) \mid D^n, X_i\} \mid D^n\right]\\ &\lesssim \|\widehat{a} - a\|^{1+\alpha}.\end{aligned}$$

Thus, we conclude that

$$|\mathbb{E}(R_n - T_n \mid D^n)| \lesssim \|\widehat{a} - a\|^{1+\alpha} + h^\alpha\|\widehat{a} - a\|.$$

C.2.2 Variance of $\widehat{\psi}_a$

To bound the variance, we use Lemma 6 from Robins *et al.* 2009a (Lemma 3). Let $\widetilde{T}_n$ be equal to T_n except that $\widehat{Q}(\widehat{a}_i)$ is replaced by $Q(\widehat{a}_i) = \mathbb{E}\{\widehat{Q}(\widehat{a}_i) \mid D^n, X_i\}$. Then,

$$\widehat{\psi} = \mathbb{P}_n\widehat{\varphi} - \widetilde{T}_n + (\widetilde{T}_n - T_n) \implies \operatorname{var}(\widehat{\psi} \mid D^n) \lesssim n^{-1} \vee \operatorname{var}(\widetilde{T}_n \mid D^n) \vee \mathbb{E}\{(\widetilde{T}_n - T_n)^2 \mid D^n\}.$$

Under the assumption that $\mathbb{E}\{A_j K_h(\widehat{a}_i - \widehat{a}_j) \mid X_i, D^n\} \lesssim 1$, we have

$$\mathbb{E}\left(\left[(A_i\widehat{a}_i - 1)^2 Q^{-2}(\widehat{a}_i)\mathbb{E}\left\{K_h(\widehat{a}_j - \widehat{a}_i)A_j(Y_j - \widehat{b}_j) \mid D^n, X_i, A_i\right\}\right]^2 \mid D^n\right) \lesssim 1,$$
$$\mathbb{E}\left\{(A_i\widehat{a}_i - 1)^2 Q^{-2}(\widehat{a}_i)A_j K_h^2(\widehat{a}_j - \widehat{a}_i)(Y_j - \widehat{b}_j)^2 \mid D^n\right\} \lesssim h^{-1}.$$

Therefore, $\mathrm{var}(\widetilde{T}_n \mid D^n) \lesssim n^{-1} + (n^2 h)^{-1}$ by Lemma 3. Finally, we have:

$$\begin{aligned}\widetilde{T}_n - T_n &= \frac{1}{n(n-1)} \sum\sum_{1\le i\neq j\le n} (A_i\widehat{a}_i - 1)Q^{-1}(\widehat{a}_i)\widehat{Q}^{-1}(\widehat{a}_i)\{\widehat{Q}(\widehat{a}_i) - Q(\widehat{a}_i)\}K_h(\widehat{a}_j - \widehat{a}_i)A_j(Y_j - \widehat{b}_j) \\ &\equiv \frac{1}{n(n-1)} \sum\sum_{1\le i\neq j\le n} T_{ij}\end{aligned}$$

We can break the square of the double sum as a sum of seven different terms:

$$\begin{aligned}\left(\sum\sum_{1\le i\neq j\le n} T_{ij}\right)^2 &= \sum\sum_{1\le i\neq j\le n}(T_{ij}^2 + T_{ij}T_{ji}) + \sum\sum\sum_{1\le i\neq j\neq l\le n}(T_{ij}T_{il} + T_{ij}T_{li} + T_{ij}T_{lj}) \\ &\quad + \sum\sum\sum\sum_{1\le i\neq j\neq l\neq k\le n} T_{ij}T_{lk}\end{aligned}$$

and bound the expectation of each term separately. In particular,

$$T_{ij}^2 \lesssim h^{-1}A_j K_h(\widehat{a}_j - \widehat{a}_i) \quad \text{and} \quad |T_{ij}T_{ji}| \lesssim h^{-1}A_i A_j K_h(\widehat{a}_j - \widehat{a}_i),$$

so that, in light of $\mathbb{E}\{A_j K_h(\widehat{a}_j - \widehat{a}_i) \mid X_i, D^n\} \lesssim 1$:

$$\mathbb{E}\left(\sum\sum_{1\le i\neq j\le n}(T_{ij}^2 + T_{ij}T_{ji}) \mid D^n\right) \lesssim n(n-1)h^{-1}.$$

Next, we have

$$\begin{aligned}&|T_{ij}T_{li}| \lesssim A_i A_j K_h(\widehat{a}_j - \widehat{a}_i)K_h(\widehat{a}_i - \widehat{a}_l), \quad |T_{ij}T_{il}| \lesssim A_j A_l K_h(\widehat{a}_j - \widehat{a}_i)K_h(\widehat{a}_l - \widehat{a}_i) \\ &|T_{ij}T_{jl}| \lesssim A_j A_l K_h(\widehat{a}_j - \widehat{a}_i)K_h(\widehat{a}_l - \widehat{a}_j), \quad \text{and} \quad |T_{ij}T_{lj}| \lesssim A_j K_h(\widehat{a}_j - \widehat{a}_i)K_h(\widehat{a}_j - \widehat{a}_l).\end{aligned}$$

Because

$$\begin{aligned}\mathbb{E}\{A_i A_j K_h(\widehat{a}_j - \widehat{a}_i)K_h(\widehat{a}_i - \widehat{a}_l) \mid D^n\} &= \mathbb{E}[A_i K_h(\widehat{a}_i - \widehat{a}_l)\mathbb{E}\{A_j K_h(\widehat{a}_j - \widehat{a}_i) \mid X_i, D^n\} \mid D^n] \\ &\lesssim \mathbb{E}[A_i K_h(\widehat{a}_i - \widehat{a}_l) \mid D^n] \\ &\lesssim 1,\end{aligned}$$

we have

$$\mathbb{E}\left\{\sum\sum\sum_{1\le i\neq j\neq l\le n}(T_{ij}T_{il} + T_{ij}T_{li} + T_{ij}T_{jl}) \mid D^n\right\} \lesssim n(n-1)(n-2).$$

Finally, to analyze the terms of the form $T_{ij}T_{lk}$ and $T_{ij}T_{lj}$, we need to address the fact that $\widehat{Q}(\widehat{a}_i)$ and $\widehat{Q}(\widehat{a}_l)$ contain the random variables A_l and A_i, respectively. This adds a minor complication to the bound because $\mathbb{E}\{A_i\widehat{a}_i - 1)\widehat{Q}^{-1}(\widehat{a}_l) \mid D^n\} \neq \mathbb{E}\{A_i(\widehat{a}_i - a_i)\widehat{Q}^{-1}(\widehat{a}_l) \mid D^n\}$, for example. In this light, we write

$$\begin{aligned}
\widehat{Q}(\widehat{a}_i) &= \frac{1}{n-1}\sum_{s=1,s\neq i}^{n} A_s K_h(\widehat{a}_s - \widehat{a}_i) \\
&= \frac{1}{n-1}\sum_{s=1,s\neq (i,l)}^{n} A_s K_h(\widehat{a}_s - \widehat{a}_i) + \frac{A_l K_h(\widehat{a}_l - \widehat{a}_i)}{n-1} \\
&\equiv \widehat{Q}_{-l}(\widehat{a}_i) + \frac{A_l K_h(\widehat{a}_l - \widehat{a}_i)}{n-1}.
\end{aligned}$$

This way, one has

$$\begin{aligned}
\left|\widehat{Q}^{-1}(\widehat{a}_i) - \widehat{Q}_{-l}^{-1}(\widehat{a}_i)\right| &= \left|\frac{A_l K_h(\widehat{a}_l - \widehat{a}_i)}{(n-1)\widehat{Q}(\widehat{a}_i)\widehat{Q}_{-l}(\widehat{a}_i)}\right| \lesssim \frac{K_h(\widehat{a}_l - \widehat{a}_i)}{n-1}, \\
\left|\widehat{Q}^{-1}(\widehat{a}_l) - \widehat{Q}_{-i}^{-1}(\widehat{a}_l)\right| &= \left|\frac{A_i K_h(\widehat{a}_i - \widehat{a}_l)}{(n-1)\widehat{Q}(\widehat{a}_l)\widehat{Q}_{-i}(\widehat{a}_l)}\right| \lesssim \frac{K_h(\widehat{a}_i - \widehat{a}_l)}{n-1}.
\end{aligned}$$

Next, we write

$$\begin{aligned}
T_{ij}T_{lk} = {} & (A_i\widehat{a}_i - 1)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i) - Q^{-1}(\widehat{a}_i)\}K_h(\widehat{a}_j - \widehat{a}_i)A_j(Y_j - \widehat{b}_j) \\
& \times (A_l\widehat{a}_l - 1)\{\widehat{Q}_{-i}^{-1}(\widehat{a}_l) - Q^{-1}(\widehat{a}_l)\}K_h(\widehat{a}_l - \widehat{a}_k)A_k(Y_k - \widehat{b}_k) \\
& - (A_i\widehat{a}_i - 1)\frac{A_l K_h(\widehat{a}_l - \widehat{a}_i)K_h(\widehat{a}_j - \widehat{a}_i)}{(n-1)\widehat{Q}(\widehat{a}_i)\widehat{Q}_{-l}(\widehat{a}_i)}A_j(Y_j - \widehat{b}_j) \\
& \times (A_l\widehat{a}_l - 1)\{\widehat{Q}^{-1}(\widehat{a}_l) - Q^{-1}(\widehat{a}_l)\}K_h(\widehat{a}_l - \widehat{a}_k)A_k(Y_k - \widehat{b}_k) \\
& - (A_i\widehat{a}_i - 1)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i) - Q^{-1}(\widehat{a}_i)\}K_h(\widehat{a}_j - \widehat{a}_i)A_j(Y_j - \widehat{b}_j) \\
& \times (A_l\widehat{a}_l - 1)\frac{A_i K_h(\widehat{a}_i - \widehat{a}_l)K_h(\widehat{a}_l - \widehat{a}_k)}{(n-1)\widehat{Q}(\widehat{a}_l)\widehat{Q}_{-i}(\widehat{a}_l)}A_k(Y_k - \widehat{b}_k)
\end{aligned}$$

The expectation of the last two terms can be upper bounded as

$$\begin{aligned}
& \mathbb{E}\left[\left|(-A_i\widehat{a}_i - 1)\frac{A_l K_h(\widehat{a}_l - \widehat{a}_i)K_h(\widehat{a}_j - \widehat{a}_i)}{(n-1)\widehat{Q}(\widehat{a}_i)\widehat{Q}_{-l}(\widehat{a}_i)}A_j(Y_j - \widehat{b}_j)\right.\right. \\
& \qquad \times (A_l\widehat{a}_l - 1)\{\widehat{Q}^{-1}(\widehat{a}_l) - Q^{-1}(\widehat{a}_l)\}K_h(\widehat{a}_l - \widehat{a}_k)A_k(Y_k - \widehat{b}_k) \\
& \quad - (A_i\widehat{a}_i - 1)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i) - Q^{-1}(\widehat{a}_i)\}K_h(\widehat{a}_j - \widehat{a}_i)A_j(Y_j - \widehat{b}_j) \\
& \qquad \left.\left. \times (A_l\widehat{a}_l - 1)\frac{A_i K_h(\widehat{a}_i - \widehat{a}_l)K_h(\widehat{a}_l - \widehat{a}_k)}{(n-1)\widehat{Q}(\widehat{a}_l)\widehat{Q}_{-i}(\widehat{a}_l)}A_k(Y_k - \widehat{b}_k)\right| \mid D^n\right] \\
& \lesssim \mathbb{E}\left\{\frac{A_j A_k A_l K_h(\widehat{a}_l - \widehat{a}_i)K_h(\widehat{a}_j - \widehat{a}_i)K_h(\widehat{a}_l - \widehat{a}_k)}{(n-1)} \mid D^n\right\} \\
& \lesssim n^{-1}
\end{aligned}$$

The last inequality follows because

$$\begin{aligned}
&\mathbb{E}\left\{A_jA_kA_lK_h(\widehat{a}_l-\widehat{a}_i)K_h(\widehat{a}_j-\widehat{a}_i)K_h(\widehat{a}_l-\widehat{a}_k)\mid D^n\right\}\\
&=\mathbb{E}\left[A_jA_lK_h(\widehat{a}_l-\widehat{a}_i)K_h(\widehat{a}_j-\widehat{a}_i)\mathbb{E}\{A_kK_h(\widehat{a}_l-\widehat{a}_k)\mid X_l,D^n\}\mid D^n\right]\\
&\lesssim\mathbb{E}\left\{A_jA_lK_h(\widehat{a}_l-\widehat{a}_i)K_h(\widehat{a}_j-\widehat{a}_i)\mid D^n\right\}\\
&=\mathbb{E}\left[A_lK_h(\widehat{a}_l-\widehat{a}_i)\mathbb{E}\{A_jK_h(\widehat{a}_j-\widehat{a}_i)\mid X_i,D^n\}\mid D^n\right]\\
&\lesssim\mathbb{E}\left\{A_lK_h(\widehat{a}_l-\widehat{a}_i)\mid D^n\right\}\\
&=\mathbb{E}\left[\mathbb{E}\left\{A_lK_h(\widehat{a}_l-\widehat{a}_i)\mid X_i,D^n\right\}\mid D^n\right]\\
&\lesssim 1.
\end{aligned}$$

Next, we proceed to bound the first term appearing in $T_{ij}T_{lk}$, which is the main term.

When $k=j$, we have the bound

$$|\mathbb{E}(T_{ij}T_{lj})\mid D^n|\lesssim\mathbb{E}\{A_iA_jA_lA_kK_h(\widehat{a}_j-\widehat{a}_i)K_h(\widehat{a}_l-\widehat{a}_k)\mid D^n\}+\frac{1}{n}\lesssim 1$$

In this light,

$$\mathbb{E}\left\{\sum\sum\sum_{1\le i\neq j\neq l\le n}T_{ij}T_{lj}\mid D^n\right\}\lesssim n(n-1)(n-2).$$

Next, we proceed for $k\neq j$. We have

$$\begin{aligned}
&\Big|\mathbb{E}\Big[(A_i\widehat{a}_i-1)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i)-Q^{-1}(\widehat{a}_i)\}K_h(\widehat{a}_j-\widehat{a}_i)A_j(Y_j-\widehat{b}_j)\\
&\qquad\times(A_l\widehat{a}_l-1)\{\widehat{Q}_{-i}^{-1}(\widehat{a}_l)-Q^{-1}(\widehat{a}_l)\}K_h(\widehat{a}_l-\widehat{a}_k)A_k(Y_k-\widehat{b}_k)\mid D^n\Big]\Big|\\
&=\Big|\mathbb{E}\Big[A_i(\widehat{a}_i-a_i)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i)-Q^{-1}(\widehat{a}_i)\}K_h(\widehat{a}_j-\widehat{a}_i)A_j(b_j-\widehat{b}_j)\\
&\qquad\times A_l(\widehat{a}_l-a_l)\{\widehat{Q}_{-i}^{-1}(\widehat{a}_l)-Q^{-1}(\widehat{a}_l)\}K_h(\widehat{a}_l-\widehat{a}_k)A_k(b_k-\widehat{b}_k)\mid D^n\Big]\Big|\\
&\lesssim\Big|\mathbb{E}\Big[|\widehat{a}_i-a_i||b_j-\widehat{b}_j||\widehat{a}_l-a_l||b_k-\widehat{b}_k|A_iA_jA_kA_lK_h(\widehat{a}_j-\widehat{a}_i)K_h(\widehat{a}_l-\widehat{a}_k)\\
&\qquad\times\mathbb{E}\left\{|\widehat{Q}_{-l}(\widehat{a}_i)-Q(\widehat{a}_i)||\widehat{Q}_{-i}(\widehat{a}_l)-Q(\widehat{a}_l)||\ A_i,A_j,A_k,A_l,X_i,X_j,X_l,X_k,D^n\right\}\Big]\mid D^n\Big|
\end{aligned}$$

Next, by the Cauchy-Schwarz inequality:

$$\begin{aligned}
&\left[\mathbb{E}\left\{|\widehat{Q}_{-l}(\widehat{a}_i)-Q(\widehat{a}_i)||\widehat{Q}_{-i}(\widehat{a}_l)-Q(\widehat{a}_l)||\ A_i,A_j,A_k,A_l,X_i,X_j,X_l,X_k,D^n\right\}\right]^2\\
&\le\mathbb{E}\left[\left\{\widehat{Q}_{-l}(\widehat{a}_i)-Q(\widehat{a}_i)\right\}^2\mid A_j,A_k,X_i,X_j,X_k,X_l,D^n\right]\\
&\qquad\times\mathbb{E}\left[\left\{\widehat{Q}_{-i}(\widehat{a}_l)-Q(\widehat{a}_l)\right\}^2\mid A_j,A_k,X_i,X_j,X_k,X_l,D^n\right].
\end{aligned}$$

We have

$$\left\{\widehat{Q}_{-l}(\widehat{a}_i) - Q(\widehat{a}_i)\right\}^2 = \left[\frac{1}{n-1}\sum_{1\leq s\leq n, s\neq(i,l)}\{A_s K_h(\widehat{a}_s - \widehat{a}_i) - Q(\widehat{a}_i)\}\right]^2$$
$$= \frac{1}{(n-1)^2}\sum_{1\leq s\neq m\leq n,(s,m)\neq(i,l)}\sum\{A_s K_h(\widehat{a}_s - \widehat{a}_i) - Q(\widehat{a}_i)\}\{A_m K_h(\widehat{a}_m - \widehat{a}_i) - Q(\widehat{a}_i)\}$$
$$+ \frac{1}{(n-1)^2}\sum_{1\leq s\leq n, s\neq(i,l)}\{A_s K_h(\widehat{a}_s - \widehat{a}_i) - Q(\widehat{a}_i)\}^2.$$

Furthermore,

$$\left|\mathbb{E}\left[\sum_{1\leq s\neq m\leq n,(s,m)\neq(i,l)}\sum\{A_s K_h(\widehat{a}_s - \widehat{a}_i) - Q(\widehat{a}_i)\}\{A_m K_h(\widehat{a}_m - \widehat{a}_i) - Q(\widehat{a}_i)\} \mid A_j, A_k, X_i, X_j, X_k, X_l, D^n\right]\right|$$
$$= |2\{A_j K_h(\widehat{a}_j - \widehat{a}_i) - Q(\widehat{a}_i)\}\{A_k K_h(\widehat{a}_k - \widehat{a}_i) - Q(\widehat{a}_i)\}|$$
$$\lesssim h^{-2}$$

and

$$\left|\mathbb{E}\left[\sum_{1\leq s\leq n, s\neq(i,l)}\{A_s K_h(\widehat{a}_s - \widehat{a}_i) - Q(\widehat{a}_i)\}^2 \mid D^n, A_j, A_k, X_i, X_j, X_k, X_l\right]\right|$$
$$= \Big|(n-4)\mathbb{E}\left[\{A_s K_h(\widehat{a}_s - \widehat{a}_i) - Q(\widehat{a}_i)\}^2 \mid D^n, X_i\right]$$
$$+\mathbb{E}\left[\{A_j K_h(\widehat{a}_j - \widehat{a}_i) - Q(\widehat{a}_i)\}^2 \mid D^n, A_j, X_i, X_j\right]$$
$$+\mathbb{E}\left[\{A_k K_h(\widehat{a}_k - \widehat{a}_i) - Q(\widehat{a}_i)\}^2 \mid D^n, A_k, X_i, X_k\right]\Big|$$
$$\lesssim (n-4)h^{-1} + h^{-2}$$

In this light, under the condition that $nh \to \infty$, we have reached

$$\mathbb{E}\left[\left\{\widehat{Q}_{-l}(\widehat{a}_i) - Q(\widehat{a}_i)\right\}^2 \mid D^n, A_j, A_k, X_i, X_j, X_k, X_l\right] \lesssim \frac{1}{nh}.$$

An identical reasoning yields

$$\mathbb{E}\left[\left\{\widehat{Q}_{-i}(\widehat{a}_l) - Q(\widehat{a}_l)\right\}^2 \mid D^n, A_j, A_k, X_i, X_j, X_k, X_l\right] \lesssim \frac{1}{nh}.$$

Therefore, we have reached

$$\mathbb{E}\left\{|\widehat{Q}_{-l}(\widehat{a}_i) - Q(\widehat{a}_i)||\widehat{Q}_{-i}(\widehat{a}_l) - Q(\widehat{a}_l)|| \, D^n, A_j, A_k, X_i, X_j, X_l, X_k\right\} \lesssim \frac{1}{nh},$$

which yields

$$
\begin{aligned}
&|\mathbb{E}(T_{ij}T_{lk} \mid D^n)| \\
&\lesssim \frac{1}{nh} \times \mathbb{E}\left[|\widehat{a}_i - a_i||\widehat{a}_l - a_l|A_jA_iA_kA_lK_h(\widehat{a}_j - \widehat{a}_i)K_h(\widehat{a}_k - \widehat{a}_l)|\widehat{b}_j - b_j||\widehat{b}_k - b_k||\ D^n\right] + \frac{1}{n} \\
&= \frac{\mathbb{E}\left[|\widehat{a}_i - a_i|A_iA_jK_h(\widehat{a}_j - \widehat{a}_i)|\widehat{b}_j - b_j||\ D^n\right]\mathbb{E}\left[A_kA_lK_h(\widehat{a}_k - \widehat{a}_l)|\widehat{a}_l - a_l||\widehat{b}_k - b_k||\ D^n\right]}{nh} + \frac{1}{n} \\
&\lesssim \frac{\|\widehat{a} - a\|^2\|\widehat{b} - b\|^2}{nh} + \frac{1}{n}.
\end{aligned}
$$

The last inequality follows by the Cauchy-Schwarz inequality:

$$
\begin{aligned}
&\left\{\mathbb{E}\left(|\widehat{a}_i - a_i|A_iA_jK_h(\widehat{a}_j - \widehat{a}_i)|\widehat{b}_j - b_j||\ D^n\right)\right\}^2 \\
&\leq \mathbb{E}\left[(\widehat{a}_i - a_i)^2\mathbb{E}\{A_jK_h(\widehat{a}_j - \widehat{a}_i) \mid X_i, D^n\} \mid D^n\right]\mathbb{E}\left[(\widehat{b}_j - b_j)^2\mathbb{E}\{A_iK_h(\widehat{a}_j - \widehat{a}_i) \mid X_j, D^n\} \mid D^n\right] \\
&\lesssim \|\widehat{a} - a\|^2\|\widehat{b} - b\|^2.
\end{aligned}
$$

Finally, this means that we have reached that

$$
\mathbb{E}\left(\sum\sum\sum\sum\sum_{1\leq i\neq j\neq l\neq k\leq n} T_{ij}T_{lk} \mid D^n\right) \lesssim n(n-1)(n-2)(n-3)\left(\frac{\|\widehat{a} - a\|^2\|\widehat{b} - b\|^2}{nh} + \frac{1}{n}\right)
$$

Putting everything together, we conclude that

$$
\mathbb{E}\{(\widetilde{T}_n - T_n)^2 \mid D^n\} \lesssim \frac{1}{n} \vee \frac{1}{n^2h} \vee \frac{\|\widehat{a} - a\|^2\|\widehat{b} - b\|^2}{nh}.
$$

This concludes our proof of the bound on $\text{var}(\widehat{\psi}_a \mid D^n)$.

C.3 Proof of Proposition 3: bias and variance of $\widehat{\psi}_b$

The proof essentially follows that for the bounds on the bias and variance of $\widehat{\psi}_a$, so we omit certain details. Recall that

$$
\widehat{Q}_{-i}(\widehat{b}_j) = \frac{1}{n-1}\sum_{s=1, s\neq i}^{n} A_sK_h(\widehat{b}_s - \widehat{b}_j),
$$

and define

$$
\begin{aligned}
Q(\widehat{b}_j) = \mathbb{E}\{\widehat{Q}_{-i}(\widehat{b}_j) \mid X_j, D^n\} &= \int \{a(x)\}^{-1}K_h(\widehat{b}(x) - \widehat{b}(X_j))d\mathbb{P}(x) \\
&= \mathbb{P}(A=1)\int K(u)d\mathbb{P}_{\widehat{b}|A=1,D^n}(uh + \widehat{b}(X_j)).
\end{aligned}
$$

Recall that $\widehat{\psi}_b = \mathbb{P}_n\widehat{\varphi} - T_n = \mathbb{P}_n\widehat{\varphi} - \widetilde{T}_n + (\widetilde{T}_n - T_n)$, where

$$T_n = \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n}\sum (A_i\widehat{a}_i - 1)\{\widehat{Q}_{-i}(\widehat{b}_j)\}^{-1} K_h(\widehat{b}_i - \widehat{b}_j) A_j(Y_j - \widehat{b}_j),$$
$$\widetilde{T}_n = \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n}\sum (A_i\widehat{a}_i - 1)\{Q(\widehat{b}_j)\}^{-1} K_h(\widehat{b}_i - \widehat{b}_j) A_j(Y_j - \widehat{b}_j).$$

For $R_n = \mathbb{P}(\widehat{\varphi} - \varphi)$, we have

$$\left|\mathbb{E}(\widehat{\psi}_b - \psi \mid D^n)\right| = \left|\mathbb{E}(R_n - \widetilde{T}_n \mid D^n)\right| + \left|\mathbb{E}(\widetilde{T}_n - T_n \mid D^n)\right|,$$
$$\mathrm{var}(\widehat{\psi}_b \mid D^n) \lesssim n^{-1} \vee (n^2h)^{-1} \vee \mathbb{E}\{(\widetilde{T}_n - T_n)^2 \mid D^n\}.$$

The inequality for the variance relies on Lemma 3, the boundedness of the observations, $\widehat{a}(X)$ and $\widehat{b}(X)$, as well as Lemma 4.

C.3.1 Bias of $\widehat{\psi}_b$

We start by bounding $\mathbb{E}(R_n - \widetilde{T}_n \mid D^n)$. We have

$$\begin{aligned}
\mathbb{E}(\widetilde{T}_n \mid D^n) &= \mathbb{E}\left\{ \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n}\sum (A_i\widehat{a}_i - 1)\{Q(\widehat{b}_j)\}^{-1} K_h(\widehat{b}_i - \widehat{b}_j) A_j(Y_j - \widehat{b}_j) \mid D^n \right\} \\
&= \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n}\sum \mathbb{E}\left\{ A_i(\widehat{a}_i - a_i)\{Q(\widehat{b}_j)\}^{-1} K_h(\widehat{b}_i - \widehat{b}_j) A_j(Y_j - \widehat{b}_j) \mid D^n \right\} \\
&= \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n}\sum \mathbb{E}\left\{ A_i s_b(\widehat{b}_j, b_j; D^n)\{Q(\widehat{b}_j)\}^{-1} K_h(\widehat{b}_i - \widehat{b}_j) A_j(Y_j - \widehat{b}_j) \mid D^n \right\} \\
&\quad + \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n}\sum \mathbb{E}\left[A_i\{s_b(\widehat{b}_i, b_i; D^n) - s_b(\widehat{b}_j, b_j; D^n)\} K_h(\widehat{b}_i - \widehat{b}_j)\{Q(\widehat{b}_j)\}^{-1} A_j(Y_j - \widehat{b}_j) \mid D^n \right] \\
&= R_n + \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n}\sum \mathbb{E}\left[A_i\{s_b(\widehat{b}_i, b_i; D^n) - s_b(\widehat{b}_j, b_j; D^n)\} K_h(\widehat{b}_i - \widehat{b}_j)\{Q(\widehat{b}_j)\}^{-1} A_j(b_j - \widehat{b}_j) \mid D^n \right].
\end{aligned}$$

The third equality follows because

$$\begin{aligned}
&\mathbb{E}\{A_i(\widehat{a}_i - a_i) - A_i s_b(\widehat{b}_i, b_i; D^n) \mid A_i, A_j, Y_j, \widehat{b}_i, \widehat{b}_j, D^n\} \\
&= \mathbb{E}\{A_i(\widehat{a}_i - a_i) - A_i s_b(\widehat{b}_i, b_i; D^n) \mid A_i, \widehat{b}_i, D^n\} \\
&= \mathbb{E}\left[\mathbb{E}\{A_i(\widehat{a}_i - a_i) \mid A_i, \widehat{b}_i, b_i, D^n\} - A_i s_b(\widehat{b}_i, b_i; D^n) \mid A_i, \widehat{b}_i, D^n\right] \\
&= \mathbb{E}\{A_i s_b(\widehat{b}_i, b_i; D^n) - A_i s_b(\widehat{b}_i, b_i; D^n) \mid A_i, \widehat{b}_i, D^n\} \\
&= 0.
\end{aligned}$$

The last equality follows because

$$\begin{aligned}
&\frac{1}{n(n-1)}\sum\sum_{1\le i\neq j\le n}\mathbb{E}\left\{A_i s_b(\widehat{b}_j,b_j;D^n)\{Q(\widehat{b}_j)\}^{-1}K_h(\widehat{b}_i-\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\mid D^n\right\}\\
&=\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left\{s_b(\widehat{b}_j,b_j;D^n)A_j(Y_j-\widehat{b}_j)\mid D^n\right\}\\
&=R_n.
\end{aligned}$$

Under the smoothness condition of $s_b(\widehat{b}_j,b_j;D^n)$, one also has

$$\begin{aligned}
&\left|\mathbb{E}\left[A_i\{s_b(\widehat{b}_i,b_i;D^n)-s_b(\widehat{b}_j,b_j;D^n)\}\frac{K_h(\widehat{b}_i-\widehat{b}_j)}{Q(\widehat{b}_j)}A_j(b_j-\widehat{b}_j)\mid D^n\right]\right|\\
&\lesssim\mathbb{E}\left[\{|\widehat{b}_j-\widehat{b}_i|^\beta+|\widehat{b}_j-b_j|^\beta+|\widehat{b}_i-b_i|^\beta\}A_iA_jK_h(\widehat{b}_i-\widehat{b}_j)(b_j-\widehat{b}_j)\mid D^n\right]\\
&\lesssim h^\beta\|\widehat{b}-b\|+\|\widehat{b}-b\|^{1+\beta}.
\end{aligned}$$

Next, we have

$$\begin{aligned}
&\left|\mathbb{E}(\widetilde{T}_n-T_n\mid D^n)\right|\\
&\le\frac{1}{n(n-1)}\left|\sum\sum_{1\le i\neq j\le n}\mathbb{E}\left[(A_i\widehat{a}_i-1)\{\widehat{Q}_{-i}^{-1}(\widehat{b}_j)-Q^{-1}(\widehat{b}_j)\}K_h(\widehat{b}_i-\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\right|.
\end{aligned}$$

Because $\widehat{Q}_{-i}(\widehat{b}_j)$ does not depend on A_i, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[(A_i\widehat{a}_i-1)\{\widehat{Q}_{-i}^{-1}(\widehat{b}_j)-Q^{-1}(\widehat{b}_j)\}K_h(\widehat{b}_i-\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\right|\\
&=\left|\mathbb{E}\left[A_i(\widehat{a}_i-a_i)\{\widehat{Q}_{-i}^{-1}(\widehat{b}_j)-Q^{-1}(\widehat{b}_j)\}K_h(\widehat{b}_i-\widehat{b}_j)A_j(b_j-\widehat{b}_j)\mid D^n\right]\right|\\
&\lesssim\left|\mathbb{E}\left[|\widehat{a}_i-a_i|\mathbb{E}\{|\widehat{Q}_{-i}(\widehat{b}_j)-Q(\widehat{b}_j)||\ A_j,X_i,X_j,D^n\}A_iA_jK_h(\widehat{b}_i-\widehat{b}_j)|b_j-\widehat{b}_j||\ D^n\right]\right|
\end{aligned}$$

By the same reasoning used to bound $\mathbb{E}[\{\widehat{Q}_{-l}(\widehat{a}_i)-Q(\widehat{a}_i)\}^2\mid A_j,A_k,X_i,X_j,X_l,X_k,D^n]$ in proving the variance bound for $\widehat{\psi}_a$, we have

$$\mathbb{E}\{|\widehat{Q}_{-i}(\widehat{b}_j)-Q(\widehat{b}_j)||\ A_j,X_i,X_j,D^n\}\lesssim\frac{1}{\sqrt{nh}}$$

under the condition that $nh\to\infty$. In this light, we have

$$\left|\mathbb{E}\left[(A_i\widehat{a}_i-1)\{\widehat{Q}_{-i}^{-1}(\widehat{b}_j)-Q^{-1}(\widehat{b}_j)\}K_h(\widehat{b}_i-\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\right|\lesssim\frac{\|\widehat{a}-a\|\|\widehat{b}-b\|}{\sqrt{nh}}.$$

Putting everything together, we have reached that

$$\begin{aligned}\left|\mathbb{E}(\widehat{\psi}_b - \psi \mid D^n)\right| &\lesssim |\mathbb{E}(R_n - T_n \mid D^n)| + \left|\mathbb{E}(\widetilde{T}_n - T_n \mid D^n)\right| \\ &\lesssim h^\beta \|\widehat{b} - b\| + \|\widehat{b} - b\|^{1+\beta} + \frac{\|\widehat{b} - b\|\|\widehat{a} - a\|}{\sqrt{nh}}.\end{aligned}$$

Finally, we also have the bound

$$\left|\mathbb{E}(\widehat{\psi}_b - \psi \mid D^n)\right| \lesssim |\mathbb{P}(\widehat{\varphi} - \varphi)| + |\mathbb{E}(T_n \mid D^n)| \lesssim \|\widehat{a} - a\|\|\widehat{b} - b\| + |\mathbb{E}(T_n \mid D^n)|$$

and

$$\begin{aligned}|\mathbb{E}(T_n \mid D^n)| &\leq \frac{1}{n(n-1)} \left| \sum\sum_{1 \leq i \neq j \leq n} \mathbb{E}\left\{ (A_i \widehat{a}_i - 1) \widehat{Q}_{-i}^{-1}(\widehat{b}_j) K_h(\widehat{b}_i - \widehat{b}_j) A_j (Y_j - \widehat{b}_j) \mid D^n \right\} \right| \\ &\lesssim \|\widehat{a} - a\|\|\widehat{b} - b\|.\end{aligned}$$

This concludes our derivation of the bound on the bias of $\widehat{\psi}_b$ given the training sample D^n.

C.3.2 Proof of Proposition 3: variance of $\widehat{\psi}_b$

Recall that $\mathrm{var}(\widehat{\psi}_b \mid D^n) \lesssim n^{-1} \vee (n^2 h)^{-1} \vee \mathbb{E}\{(\widetilde{T}_n - T_n)^2 \mid D^n\}$. Thus, we only need to bound $\mathbb{E}\{(\widetilde{T}_n - T_n)^2 \mid D^n\}$. Let

$$\begin{aligned}T_{ij} &= (A_i \widehat{a}_i - 1)\{\widehat{Q}_{-i}^{-1}(\widehat{b}_j) - Q^{-1}(\widehat{b}_j)\} K_h(\widehat{b}_i - \widehat{b}_j) A_j (Y_j - \widehat{b}_j) \\ &= (A_i \widehat{a}_i - 1) \widehat{Q}_{-i}^{-1}(\widehat{b}_j) Q^{-1}(\widehat{b}_j) \{Q(\widehat{b}_j) - \widehat{Q}_{-i}(\widehat{b}_j)\} K_h(\widehat{b}_i - \widehat{b}_j) A_j (Y_j - \widehat{b}_j)\end{aligned}$$

and notice that

$$\begin{aligned}(\widetilde{T}_n - T_n)^2 &= \left(\sum\sum_{1 \leq i \neq j \leq n} T_{ij} \right)^2 \\ &= \sum\sum_{1 \leq i \neq j \leq n} (T_{ij}^2 + T_{ij} T_{ji}) + \sum\sum\sum_{1 \leq i \neq j \neq l \leq n} (T_{ij} T_{il} + T_{ij} T_{li} + T_{ij} T_{jl} + T_{ij} T_{lj}) + \sum\sum\sum\sum_{1 \leq i \neq j \neq l \neq k \leq n} T_{ij} T_{lk}.\end{aligned}$$

Just like in the proof of the bound on the variance of $\widehat{\psi}_a$, we have

$$\begin{aligned}&T_{ij}^2 \lesssim A_j h^{-1} K_h(\widehat{b}_i - \widehat{b}_j), \quad |T_{ij} T_{ji}| \lesssim h^{-1} A_i A_j K_h(\widehat{b}_i - \widehat{b}_j) \\ &|T_{ij} T_{il}| \lesssim A_j A_l K_h(\widehat{b}_i - \widehat{b}_j) K_h(\widehat{b}_i - \widehat{b}_l), \quad |T_{ij} T_{li}| \lesssim A_i A_j K_h(\widehat{b}_i - \widehat{b}_j) K_h(\widehat{b}_l - \widehat{b}_i) \\ &|T_{ij} T_{jl}| \lesssim A_j A_l K_h(\widehat{b}_i - \widehat{b}_j) K_h(\widehat{b}_j - \widehat{b}_l), \quad |T_{ij} T_{lj}| \lesssim A_j K_h(\widehat{b}_i - \widehat{b}_j) K_h(\widehat{b}_l - \widehat{b}_j),\end{aligned}$$

which yields that

$$\left|\mathbb{E}\left\{\sum\sum_{1\le i\neq j\le n}(T_{ij}^2+T_{ij}T_{ji})\mid D^n\right\}\right|\lesssim n(n-1)h^{-1},$$
$$\left|\mathbb{E}\left\{\sum\sum\sum_{1\le i\neq j\neq l\le n}(T_{ij}T_{il}+T_{ij}T_{li}+T_{ij}T_{jl})\mid D^n\right\}\right|\lesssim n(n-1)(n-2).$$

Next, define

$$\widehat{Q}_{-il}(\widehat{b}_j)=\frac{1}{n-1}\sum_{s=1,s\neq(i,l)}^{n}A_sK_h(\widehat{b}_s-b_j),$$

so that

$$\widehat{Q}_{-il}(\widehat{b}_j)-\widehat{Q}_{-i}(\widehat{b}_j)=-\frac{A_lK_h(\widehat{b}_l-\widehat{b}_j)}{n-1}.$$

In this light, we have

$$\begin{aligned}
T_{ij}T_{lk}=&(A_i\widehat{a}_i-1)\{\widehat{Q}_{-il}^{-1}(\widehat{b}_j)-Q^{-1}(\widehat{b}_j)\}K_h(\widehat{b}_i-\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\\
&\times(A_l\widehat{a}_l-1)\{\widehat{Q}_{-il}^{-1}(\widehat{b}_k)-Q^{-1}(\widehat{b}_k)\}K_h(\widehat{b}_l-\widehat{b}_k)A_k(Y_k-\widehat{b}_k)\\
&-(A_i\widehat{a}_i-1)\frac{A_lK_h(\widehat{b}_l-\widehat{b}_j)K_h(\widehat{b}_i-\widehat{b}_j)}{(n-1)\widehat{Q}_{-il}(\widehat{b}_j)\widehat{Q}_{-i}(\widehat{b}_j)}A_j(Y_j-\widehat{b}_j)\\
&\times(A_l\widehat{a}_l-1)\{\widehat{Q}_{-l}^{-1}(\widehat{b}_k)-Q^{-1}(\widehat{b}_k)\}K_h(\widehat{b}_l-\widehat{b}_k)A_k(Y_k-\widehat{b}_k)\\
&-(A_i\widehat{a}_i-1)\{\widehat{Q}_{-il}^{-1}(\widehat{b}_j)-Q^{-1}(\widehat{b}_j)\}K_h(\widehat{b}_i-\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\\
&\times(A_l\widehat{a}_l-1)\frac{A_iK_h(\widehat{b}_i-\widehat{b}_k)K_h(\widehat{b}_l-\widehat{b}_k)}{(n-1)\widehat{Q}_{-il}(\widehat{b}_k)\widehat{Q}(\widehat{b}_k)}A_k(Y_k-\widehat{b}_k).
\end{aligned}$$

Notice that

$$\begin{aligned}
&\left|\mathbb{E}\left[(A_i\widehat{a}_i-1)\frac{A_lK_h(\widehat{b}_l-\widehat{b}_j)K_h(\widehat{b}_i-\widehat{b}_j)}{(n-1)\widehat{Q}_{-il}(\widehat{b}_j)\widehat{Q}(\widehat{b}_j)}A_j(Y_j-\widehat{b}_j)\right.\right.\\
&\quad\left.\left.\times(A_l\widehat{a}_l-1)\{\widehat{Q}_{-l}^{-1}(\widehat{b}_k)-Q^{-1}(\widehat{b}_k)\}K_h(\widehat{b}_l-\widehat{b}_k)A_k(Y_k-\widehat{b}_k)\mid D^n\right]\right|\\
&\lesssim\mathbb{E}\left\{\frac{A_lA_jA_kK_h(\widehat{b}_l-\widehat{b}_j)K_h(\widehat{b}_i-\widehat{b}_j)K_h(\widehat{b}_l-\widehat{b}_k)}{(n-1)}\mid D^n\right\}\\
&\lesssim n^{-1},
\end{aligned}$$

and, similarly,

$$\begin{aligned}
&\Big|\mathbb{E}\Big[(A_i\widehat{a}_i-1)\{\widehat{Q}_{-il}^{-1}(\widehat{b}_j)-Q^{-1}(\widehat{b}_j)\}K_h(\widehat{b}_i-\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\\
&\quad\times(A_l\widehat{a}_l-1)\frac{A_iK_h(\widehat{b}_i-\widehat{b}_k)K_h(\widehat{b}_l-\widehat{b}_k)}{(n-1)\widehat{Q}_{-il}(\widehat{b}_k)\widehat{Q}(\widehat{b}_k)}A_k(Y_k-\widehat{b}_k)\mid D^n\Big]\Big|\\
&\lesssim n^{-1}.
\end{aligned}$$

Therefore, when $k=j$, we have the bound

$$|\mathbb{E}(T_{ij}T_{lj}\mid D^n)|\lesssim \mathbb{E}\{A_iA_jA_lA_kK_h(\widehat{b}_i-\widehat{b}_j)K_h(\widehat{b}_l-\widehat{b}_k)\mid D^n\}+\frac{1}{n}\lesssim 1,$$

so that

$$\left|\mathbb{E}\left\{\sum\sum\sum_{1\le i\ne j\ne l\le n}T_{ij}T_{lj}\mid D^n\right\}\right|\lesssim n(n-1)(n-2).$$

Finally, we consider the case $k\ne j$:

$$\begin{aligned}
&\Big|\mathbb{E}\Big[(A_i\widehat{a}_i-1)\{\widehat{Q}_{-il}^{-1}(\widehat{b}_j)-Q^{-1}(\widehat{b}_j)\}K_h(\widehat{b}_i-\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\\
&\qquad\times(A_l\widehat{a}_l-1)\{\widehat{Q}_{-il}^{-1}(\widehat{b}_k)-Q^{-1}(\widehat{b}_k)\}K_h(\widehat{b}_l-\widehat{b}_k)A_k(Y_k-\widehat{b}_k)\mid D^n\Big]\Big|\\
&\lesssim\Big|\mathbb{E}\Big[|\widehat{a}_i-a_i|A_iA_jK_h(\widehat{b}_i-\widehat{b}_j)|b_j-\widehat{b}_j|A_lA_kK_h(\widehat{b}_l-\widehat{b}_k)|\widehat{a}_l-a_l||b_k-\widehat{b}_k|\\
&\qquad\times\mathbb{E}\{|\widehat{Q}_{-il}(\widehat{b}_j)-Q(\widehat{b}_j)||\widehat{Q}_{-il}(\widehat{b}_k)-Q(\widehat{b}_k)||\,A_j,A_k,X_i,X_j,X_l,X_k,D^n\}\mid D^n\Big]\Big|.
\end{aligned}$$

We have

$$\begin{aligned}
&\Big\{\widehat{Q}_{-il}(\widehat{b}_j)-Q(\widehat{b}_j)\Big\}^2\\
&=\frac{1}{(n-1)^2}\sum_{1\le s\ne m\le n,(s,m)\ne(i,l)}\{A_sK_h(\widehat{b}_s-\widehat{b}_j)-Q(\widehat{b}_j)\}\{A_mK_h(\widehat{b}_m-\widehat{b}_j)-Q(\widehat{b}_j)\}\\
&\quad+\frac{1}{(n-1)^2}\sum_{1\le s\le n,s\ne(i,l)}\{A_sK_h(\widehat{b}_s-\widehat{b}_j)-Q(\widehat{b}_j)\}^2,
\end{aligned}$$

Further,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{1\le s\ne m\le n,(s,m)\ne(i,l)}\{A_sK_h(\widehat{b}_s-\widehat{b}_j)-Q(\widehat{b}_j)\}\{A_mK_h(\widehat{b}_m-\widehat{b}_j)-Q(\widehat{b}_j)\}\mid A_j,A_k,X_i,X_j,X_l,X_k,D^n\right]\\
&=2|\{A_kK_h(\widehat{b}_l-\widehat{b}_j)-Q(\widehat{b}_j)\}\{A_jK_h(\widehat{b}_j-\widehat{b}_j)-Q(\widehat{b}_j)\}|\\
&\lesssim h^{-2}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\sum_{1\le s\le n, s\neq(i,l)} \{A_s K_h(\widehat{b}_s - \widehat{b}_j) - Q(\widehat{b}_j)\}^2 \mid A_j, A_k, X_i, X_j, X_l, X_k, D^n\right] \\
&= (n-4)\mathbb{E}\left[\{A_s K_h(\widehat{b}_s - \widehat{b}_j) - Q(\widehat{b}_j)\}^2 \mid X_j, D^n\right] + \{A_k K_h(\widehat{b}_k - \widehat{b}_j) - Q(\widehat{b}_j)\}^2 \\
&\quad + \{A_j K_h(\widehat{b}_j - \widehat{b}_j) - Q(\widehat{b}_j)\}^2 \\
&\lesssim (n-4)h^{-1} + h^{-2}
\end{aligned}$$

Thus, under the condition that $nh \to \infty$, it holds that

$$\mathbb{E}\left[\{\widehat{Q}_{-il}(\widehat{b}_j) - Q(\widehat{b}_j)\}^2 \mid A_j, A_k, X_i, X_j, X_k, X_l, D^n\right] \lesssim \frac{1}{nh}.$$

By the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\{|\widehat{Q}_{-il}(\widehat{b}_j) - Q(\widehat{b}_j)||\widehat{Q}_{-il}(\widehat{b}_k) - Q(\widehat{b}_k)|| \, A_j, A_k, X_i, X_j, X_l, X_k, D^n\} \lesssim \frac{1}{nh}.$$

In this respect, we conclude that

$$\begin{aligned}
&\Big|\mathbb{E}\Big[(A_i\widehat{a}_i - 1)\{\widehat{Q}_{-il}^{-1}(\widehat{b}_j) - Q^{-1}(\widehat{b}_j)\}K_h(\widehat{b}_i - \widehat{b}_j)A_j(Y_j - \widehat{b}_j) \\
&\qquad \times (A_l\widehat{a}_l - 1)\{\widehat{Q}_{-il}^{-1}(\widehat{b}_k) - Q^{-1}(\widehat{b}_k)\}K_h(\widehat{b}_l - \widehat{b}_k)A_k(Y_k - \widehat{b}_k) \mid D^n\Big]\Big| \\
&\lesssim \frac{\|\widehat{a} - a\|^2\|\widehat{b} - b\|^2}{nh}
\end{aligned}$$

Putting everything together, we have reached that

$$\mathbb{E}\{(\widetilde{T}_n - T_n)^2 \mid D^n\} \lesssim n^{-1} + \frac{\|\widehat{a} - a\|^2\|\widehat{b} - b\|^2}{nh}.$$

Thus, we conclude that

$$\mathrm{var}(\widehat{\psi}_b \mid D^n) \lesssim n^{-1} \vee (n^2h)^{-1} \vee \frac{\|\widehat{a} - a\|^2\|\widehat{b} - b\|^2}{nh}.$$

C.4 Proof of Proposition 4: bias and variance of $\widehat{\psi}$

For shorthand notation, let us use the notation $\widehat{a}(X_i) = \widehat{a}_i$, $\widehat{a} = \widehat{a}(X)$ and so on. Also let $K_{hi}(\widehat{a}_j) = h^{-1}K\left(\frac{\widehat{a}_i - \widehat{a}_j}{h}\right)$ and define $K_{hi}(\widehat{b}_j)$ similarly. Further, let

$$\begin{aligned}
\widehat{Q}(\widehat{a}_i, \widehat{b}_i) &= \frac{1}{n-1} \sum_{1 \le j \le n, j \ne i} A_j K_{hi}(\widehat{a}_j) K_{hi}(\widehat{b}_j), \\
Q(\widehat{a}_i, \widehat{b}_i) &= \mathbb{E}\left\{ \widehat{Q}(\widehat{a}_i, \widehat{b}_i) \mid X_i, D^n \right\} \\
&= \int \{a(x)\}^{-1} K_h(\widehat{a}(x) - a(X_i)) K_h(\widehat{b}(x) - \widehat{b}(X_i)) d\mathbb{P}(x) \\
&= \mathbb{P}(A=1) \int K(u)K(v) d\mathbb{P}_{\widehat{a},\widehat{b}|A=1,D^n}(hu + \widehat{a}(X_i), hv + \widehat{b}(X_i)), \\
Q(\widehat{a}_j, \widehat{b}_j) &= \int \{a(x)\}^{-1} K_h(\widehat{a}(x) - \widehat{a}(X_j)) K_h(\widehat{b}(x) - \widehat{b}(X_j)) d\mathbb{P}(x) \\
&= \mathbb{P}(A=1) \int K(u)K(v) d\mathbb{P}_{\widehat{a},\widehat{b}|A=1,D^n}(hu + \widehat{a}(X_j), hv + \widehat{b}(X_j)),
\end{aligned}$$

where $d\mathbb{P}_{\widehat{a},\widehat{b}|A=1,D^n}(u,v)$ is the density of $(\widehat{a}(X), \widehat{b}(X))$ among units with $A=1$ keeping $\widehat{a}(\cdot)$ and $\widehat{b}(\cdot)$ as fixed functions given the training sample D^n. Recall that $\widehat{\psi} = \mathbb{P}_n\widehat{\varphi} - T_n$, where $\widehat{\varphi}(O) = A\widehat{a}(Y - \widehat{b}) + \widehat{b}$ and

$$T_n = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \sum (A\widehat{a}_i - 1)\{\widehat{Q}(\widehat{a}_i, \widehat{b}_i)\}^{-1} K_{hi}(\widehat{a}_j) K_{hi}(\widehat{b}_j) A_j (Y_j - \widehat{b}_j) \equiv \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \sum T_{ij}.$$

C.4.1 Bias

Bound 1. We start with the following decomposition:

$$T_{ij} = (A_i\widehat{a}_i - 1)\{Q(\widehat{a}_j, \widehat{b}_j)\}^{-1} K_{hi}(\widehat{a}_j) K_{hi}(\widehat{b}_j) A_j (Y_j - \widehat{b}_j) + T_{2ij} + T_{3ij},$$

where

$$\begin{aligned}
T_{2ij} &= (A_i\widehat{a}_i - 1)\left[\{\widehat{Q}(\widehat{a}_i, \widehat{b}_i)\}^{-1} - \{Q(\widehat{a}_i, \widehat{b}_i)\}^{-1}\right] K_{hi}(\widehat{a}_j) K_{hi}(\widehat{b}_j) A_j (Y_j - \widehat{b}_j), \\
T_{3ij} &= (A_i\widehat{a}_i - 1)\left[\{Q(\widehat{a}_i, \widehat{b}_i)\}^{-1} - \{Q(\widehat{a}_j, \widehat{b}_j)\}^{-1}\right] K_{hi}(\widehat{a}_j) K_{hi}(\widehat{b}_j) A_j (Y_j - \widehat{b}_j).
\end{aligned}$$

We have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum(A_i\widehat{a}_i-1)\{Q(\widehat{a}_j,\widehat{b}_j)\}^{-1}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\\
&=\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum A_i(\widehat{a}_i-a_i)\{Q(\widehat{a}_j,\widehat{b}_j)\}^{-1}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\\
&=\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum A_i f_b(\widehat{b}_i,b_i,\widehat{a}_i;D^n)\{Q(\widehat{a}_j,\widehat{b}_j)\}^{-1}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\\
&=\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum A_i f_b(\widehat{b}_j,b_j,\widehat{a}_j;D^n)\{Q(\widehat{a}_j,\widehat{b}_j)\}^{-1}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\\
&\quad+\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum A_i\{f_b(\widehat{b}_i,b_i,\widehat{a}_i;D^n)-f_b(\widehat{b}_j,b_j,\widehat{a}_j;D^n)\}\frac{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)(b_j-\widehat{b}_j)}{\widehat{Q}(\widehat{a}_j,\widehat{b}_j)}\mid D^n\right]\\
&=R_n+\mathbb{E}\left(\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum T_{1ij}\mid D^n\right)
\end{aligned}
$$

The second equality follows because

$$
\begin{aligned}
&\mathbb{E}\{A_i(\widehat{a}_i-a_i)-A_if_b(\widehat{b}_i,b_i,\widehat{a}_i;D^n)\mid A_i,A_j,Y_j,\widehat{a}_i,\widehat{a}_j,\widehat{b}_i,\widehat{b}_j,D^n\}\\
&=\mathbb{E}\{A_i(\widehat{a}_i-a_i)-A_if_b(\widehat{b}_i,b_i,\widehat{a}_i;D^n)\mid A_i,\widehat{a}_i,\widehat{b}_i,D^n\}\\
&=\mathbb{E}\left[\mathbb{E}\{A_i(\widehat{a}_i-a_i)\mid A_i,\widehat{a}_i,\widehat{b}_i,b_i,D^n\}-A_if_b(\widehat{b}_i,b_i,\widehat{a}_i;D^n)\mid A_i,\widehat{a}_i,\widehat{b}_i,D^n\right]\\
&=\mathbb{E}\left\{A_i\mathbb{E}(\widehat{a}_i-a_i\mid A_i=1,\widehat{a}_i,\widehat{b}_i,b_i,D^n)-A_if_b(\widehat{b}_i,b_i,\widehat{a}_i;D^n)\mid A_i,\widehat{a}_i,\widehat{b}_i,D^n\right\}\\
&=0
\end{aligned}
$$

The last equality follows because

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum A_i f_b(\widehat{b}_j,b_j,\widehat{a}_j;D^n)\{Q(\widehat{a}_j,\widehat{b}_j)\}^{-1}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\\
&=\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum\frac{\mathbb{E}\{A_iK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\mid X_j,D^n\}}{Q(\widehat{a}_j,\widehat{b}_j)}f_b(\widehat{b}_j,b_j,\widehat{a}_j;D^n)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\\
&=\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum f_b(\widehat{b}_j,b_j,\widehat{a}_j;D^n)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\\
&=\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}f_b(\widehat{b}_j,b_j,\widehat{a}_j;D^n)A_j(Y_j-\widehat{b}_j)\mid D^n\right]\\
&=R_n.
\end{aligned}
$$

Thus, we have reached

$$\mathbb{E}\left\{\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum T_{ij} \mid D^n\right\} = R_n + \mathbb{E}\left(\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum T_{1ij}+T_{2ij}+T_{3ij} \mid D^n\right)$$

so that

$$\mathbb{E}(\widehat{\psi}-\psi \mid D^n) = \mathbb{E}(R_n - T_n \mid D^n) = \mathbb{E}\left(\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum T_{1ij}+T_{2ij}+T_{3ij} \mid D^n\right)$$

We bound each term separately.

Term T_1. By the smoothness assumption on f_b, the last term can be upper bounded as

$$\begin{aligned}
&\left|\mathbb{E}\left(\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum T_{1ij} \mid D^n\right)\right| \\
&= \left|\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum A_i\{f_b(\widehat{b}_i,b_i,\widehat{a}_i;D^n)-f_b(\widehat{b}_j,b_j,\widehat{a}_j;D^n)\}\frac{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)(b_j-\widehat{b}_j)}{Q(\widehat{a}_j,\widehat{b}_j)} \mid D^n\right]\right| \\
&\lesssim \mathbb{E}\left\{\left(A_i|\widehat{b}_i-\widehat{b}_j|^\beta + |b_i-b_j|^\beta + |\widehat{a}_i-\widehat{a}_j|^\beta\right)A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)|b_j-\widehat{b}_j||\ D^n\right\} \\
&\le \mathbb{E}\left\{A_i\left(h^\beta + |b_i-\widehat{b}_i|^\beta + |\widehat{b}_j-b_j|^\beta\right)A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)|b_j-\widehat{b}_j||\ D^n\right\} \\
&\lesssim h^\beta\mathbb{E}\left[\mathbb{E}\left\{A_iK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid X_j,D^n\right\}|b_j-\widehat{b}_j||\ D^n\right] \\
&\quad + \left(\mathbb{E}\left[\mathbb{E}\left\{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid X_i,D^n\right\}(b_i-\widehat{b}_i)^{2\beta} \mid D^n\right]\right)^{1/2} \\
&\qquad \times\left(\mathbb{E}\left[\mathbb{E}\left\{A_iK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid X_j,D^n\right\}(b_j-\widehat{b}_j)^2 \mid D^n\right]\right)^{1/2} \\
&\quad + \mathbb{E}\left[\mathbb{E}\left\{A_iK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid X_j,D^n\right\}|b_j-\widehat{b}_j|^{1+\beta}|\ D^n\right] \\
&\lesssim h^\beta\|\widehat{b}-b\| + \|\widehat{b}-b\|^{1+\beta}
\end{aligned}$$

where the last inequality follows because $\mathbb{E}\left\{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid X_i,D^n\right\} \lesssim 1$.

Term T_2. Let $A^n=(A_1,\ldots,A_n)$ and $A^n_{-i}=A^n\setminus A_i$. Because $\widehat{Q}(\widehat{a}_i,\widehat{b}_i)$ does not depend on A_i, we have $\mathbb{E}\{\widehat{Q}(\widehat{a}_i,\widehat{b}_i) \mid X^n,A^n_{-i},D^n\} = \widehat{Q}(\widehat{a}_i,\widehat{b}_i)$. Therefore:

$$\mathbb{E}(T_{2ij} \mid X^n,A^n_{-i},D^n) = \frac{(\widehat{a}_i-a_i)A_j(b_j-\widehat{b}_j)}{a_i\widehat{Q}(\widehat{a}_i,\widehat{b}_i)Q(\widehat{a}_i,\widehat{b}_i)}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\{Q(\widehat{a}_i,\widehat{b}_i)-\widehat{Q}(\widehat{a}_i,\widehat{b}_i)\}$$

Therefore, we have

$$
\begin{aligned}
|\mathbb{E}(T_{2ij} \mid D^n)| &\leq \mathbb{E}\left[\left|\frac{A_i(\widehat{a}_i - a_i)A_j(b_j - \widehat{b}_j)}{\widehat{Q}(\widehat{a}_i,\widehat{b}_i)Q(\widehat{a}_i,\widehat{b}_i)}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\{Q(\widehat{a}_i,\widehat{b}_i) - \widehat{Q}(\widehat{a}_i,\widehat{b}_i)\}\right| \mid D^n\right] \\
&\lesssim \mathbb{E}\left\{|\widehat{a}_i - a_i||b_j - \widehat{b}_j|A_iA_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\left|Q(\widehat{a}_i,\widehat{b}_i) - \widehat{Q}(\widehat{a}_i,\widehat{b}_i)\right| \mid D^n\right\} \\
&= \mathbb{E}\left[|\widehat{a}_i - a_i||b_j - \widehat{b}_j|A_iA_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\mathbb{E}\left\{\left|Q(\widehat{a}_i,\widehat{b}_i) - \widehat{Q}(\widehat{a}_i,\widehat{b}_i)\right| \mid A_j, X_i, X_j, D^n\right\} \mid D^n\right]
\end{aligned}
$$

Next, we bound

$$
\mathbb{E}\left\{\left|Q(\widehat{a}_i,\widehat{b}_i) - \widehat{Q}(\widehat{a}_i,\widehat{b}_i)\right| \mid A_j, X_i, X_j, D^n\right\} \leq \left(\mathbb{E}\left[\left\{Q(\widehat{a}_i,\widehat{b}_i) - \widehat{Q}(\widehat{a}_i,\widehat{b}_i)\right\}^2 \mid A_j, X_i, X_j, D^n\right]\right)^{1/2}
$$

We have

$$
\begin{aligned}
&\left\{\widehat{Q}(\widehat{a}_i,\widehat{b}_i) - Q(\widehat{a}_i,\widehat{b}_i)\right\}^2 = \left[\frac{1}{n-1}\sum_{1\leq s\leq n, s\neq i}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s) - Q(\widehat{a}_i,\widehat{b}_i)\}\right]^2 \\
&= \frac{1}{(n-1)^2}\sum\sum_{1\leq s\neq m\leq n,(s,m)\neq i}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s) - Q(\widehat{a}_i,\widehat{b}_i)\}\{A_mK_{hi}(\widehat{a}_m)K_{hi}(\widehat{b}_m) - Q(\widehat{a}_i,\widehat{b}_i)\} \\
&\quad + \frac{1}{(n-1)^2}\sum_{1\leq s\leq n, s\neq i}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s) - Q(\widehat{a}_i,\widehat{b}_i)\}^2
\end{aligned}
$$

Notice that

$$
\begin{aligned}
&\mathbb{E}\left[\sum\sum_{1\leq s\neq m\leq n,(s,m)\neq i}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s) - Q(\widehat{a}_i,\widehat{b}_i)\}\{A_mK_{hi}(\widehat{a}_m)K_{hi}(\widehat{b}_m) - Q(\widehat{a}_i,\widehat{b}_i)\} \mid A_j, X_i, X_j, D^n\right] \\
&= 0
\end{aligned}
$$

because, for any $s \neq (i,j)$, $\mathbb{E}\left\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s) - Q(\widehat{a}_i,\widehat{b}_i) \mid D^n, A_j, X_i, X_j\right\} = 0$ and either $s \neq j$ or $m \neq j$ in the double sum above (since $s \neq m$). Next, notice that:

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{(n-1)^2}\sum_{1\leq s\leq n, s\neq i}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s) - Q(\widehat{a}_i,\widehat{b}_i)\}^2 \mid A_j, X_i, X_j, D^n\right] \\
&= \mathbb{E}\left[\frac{1}{(n-1)^2}\sum_{1\leq s\leq n, s\neq (i,j)}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s) - Q(\widehat{a}_i,\widehat{b}_i)\}^2 \mid X_i, D^n\right] \\
&\quad + \frac{\{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) - Q(\widehat{a}_i,\widehat{b}_i)\}^2}{(n-1)^2}
\end{aligned}
$$

In this light, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{(n-1)^2}\sum_{1\le s\le n, s\neq i}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s)-Q(\widehat{a}_i,\widehat{b}_i)\}^2 \mid A_j,X_i,X_j,D^n\right]\right| \\
&\lesssim \frac{(n-2)}{(n-1)^2h^2}\mathbb{E}\{|AK_{hi}(\widehat{a})K_{hi}(\widehat{b})-Q(\widehat{a}_i,\widehat{b}_i)\mid X_i,D^n\}+\frac{1}{(n-1)^2h^4} \\
&\lesssim \frac{1}{nh^2}
\end{aligned}$$

under the condition that $nh^2\to\infty$. Putting everything together, we have reached

$$\mathbb{E}\left\{\left|Q(\widehat{a}_i,\widehat{b}_i)-\widehat{Q}(\widehat{a}_i,\widehat{b}_i)\right| \mid A_j,X_i,X_j,D^n\right\}\lesssim\frac{1}{\sqrt{nh^2}}.$$

Therefore, we have

$$|\mathbb{E}(T_{2ij}\mid D^n)|\lesssim\frac{\mathbb{E}\left\{|\widehat{a}_i-a_i||b_j-\widehat{b}_j|A_iA_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\mid D^n\right\}}{\sqrt{nh^2}}\lesssim\frac{\|\widehat{a}-a\|\|\widehat{b}-b\|}{\sqrt{nh^2}}$$

so that

$$\left|\mathbb{E}\left(\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum T_{2ij}\mid D^n\right)\right|\lesssim\frac{\|\widehat{a}-a\|\|\widehat{b}-b\|}{\sqrt{nh^2}}.$$

Term T_3. We have

$$\mathbb{E}(T_{3ij}\mid D^n)=\mathbb{E}\left[\frac{(\widehat{a}_i-a_i)(b_j-\widehat{b}_j)}{Q(\widehat{a}_i,\widehat{b}_i)Q(\widehat{a}_j,\widehat{b}_j)}A_iA_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\{Q(\widehat{a}_j,\widehat{b}_j)-Q(\widehat{a}_i,\widehat{b}_i)\}\mid D^n\right]$$

Under the assumption that the density of $(\widehat{a},\widehat{b})$ given $A=1$ is Lipschitz:

$$\begin{aligned}
&\{\mathbb{P}(A=1)\}^{-1}\left|Q(\widehat{a}_i,\widehat{b}_i)-Q(\widehat{a}_j,\widehat{b}_j)\right| \\
&=\left|\int K(u)K(v)\left\{d\mathbb{P}_{\widehat{a},\widehat{b}|A=1,D^n}(uh+\widehat{a}_i,vh+\widehat{b}_i)-d\mathbb{P}_{\widehat{a},\widehat{b}|A=1,D^n}(uh+\widehat{a}_j,vh+\widehat{b}_j)\right\}\right| \\
&\lesssim|\widehat{a}_i-\widehat{a}_j|+|\widehat{b}_i-\widehat{b}_j|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
|\mathbb{E}(T_{3ij}\mid D^n)| &\lesssim\mathbb{E}\left[|\widehat{a}_i-a_i||b_j-\widehat{b}_j|K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\{|\widehat{a}_i-\widehat{a}_j|+|\widehat{b}_i-\widehat{b}_j|\}\mid D^n\right] \\
&\lesssim h\mathbb{E}\left\{|\widehat{a}_i-a_i||b_j-\widehat{b}_j|K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\mid D^n\right\} \\
&\lesssim h\|\widehat{a}-a\|\|\widehat{b}-b\| \\
&\lesssim h\|\widehat{b}-b\|.
\end{aligned}$$

where the last inequality follows by the Cauchy-Schwarz inequality.

Putting everything together, we have reached that

$$\left|\mathbb{E}(\widehat{\psi}-\psi \mid D^n)\right| = \mathbb{E}\left(\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum T_{1ij}+T_{2ij}+T_{3ij} \mid D^n\right)$$
$$\lesssim h^{\beta}\|\widehat{b}-b\| + \|\widehat{b}-b\|^{1+\beta} + \frac{\|\widehat{a}-a\|\|\widehat{b}-b\|}{\sqrt{nh^2}}.$$

Next, we proceed with a different bound on the bias.

Bound 2. Recall that

$$T_n = \frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum (A\widehat{a}_i-1)\{\widehat{Q}(\widehat{a}_i,\widehat{b}_i)\}^{-1}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j) \equiv \frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum T_{ij}.$$

so that

$$\mathbb{E}(T_n \mid D^n) = \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum (A_i\widehat{a}_i-1)\{\widehat{Q}(\widehat{a}_i,\widehat{b}_i)\}^{-1}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j f_a(\widehat{a}_j,a_j,\widehat{b}_j;D^n) \mid D^n\right]$$
$$= \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum (A_i\widehat{a}_i-1)f_a(\widehat{a}_i,a_i,\widehat{b}_i;D^n) \mid D^n\right]$$
$$+ \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum (A_i\widehat{a}_i-1)\frac{K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)}{\widehat{Q}(\widehat{a}_i,\widehat{b}_i)}A_j\{f_a(\widehat{a}_j,a_j,\widehat{b}_j;D^n)-f_a(\widehat{a}_i,a_i,\widehat{b}_i;D^n)\} \mid D^n\right]$$
$$= R_n + \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum A_i(\widehat{a}_i-a_i)\frac{K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)}{\widehat{Q}(\widehat{a}_i,\widehat{b}_i)}A_j\{f_a(\widehat{a}_j,a_j,\widehat{b}_j;D^n)-f_a(\widehat{a}_i,a_i,\widehat{b}_i;D^n)\} \mid D^n\right]$$

The first equality follows because

$$\mathbb{E}\{A_j(Y_j-\widehat{b}_j)-A_jf_a(\widehat{a}_j,a_j,\widehat{b}_j;D^n) \mid A_j,\widehat{a}_j,\widehat{b}_j,A_i,\widehat{a}_i,\widehat{b}_i,D^n\}$$
$$= \mathbb{E}\{A_j(Y_j-\widehat{b}_j)-A_jf_a(\widehat{a}_j,a_j,\widehat{b}_j) \mid A_j,\widehat{a}_j,\widehat{b}_j,D^n\}$$
$$= \mathbb{E}\left[\mathbb{E}\{A_j(Y_j-\widehat{b}_j) \mid A_j,\widehat{a}_j,a_j,\widehat{b}_j,D^n\}-A_jf_a(\widehat{a}_j,a_j,\widehat{b}_j;D^n) \mid A_j,\widehat{a}_j,\widehat{b}_j,D^n\right]$$
$$= \mathbb{E}\left(\mathbb{E}[\mathbb{E}\{A_j(Y_j-\widehat{b}_j) \mid A_j,\widehat{a}_j,a_j,\widehat{b}_j,X_j,D^n\} \mid A_j,\widehat{a}_j,a_j,\widehat{b}_j,D^n]-A_jf_a(\widehat{a}_j,a_j,\widehat{b}_j;D^n) \mid A_j,\widehat{a}_j,\widehat{b}_j,D^n\right)$$
$$= \mathbb{E}\left[\mathbb{E}\{A_j(b_j-\widehat{b}_j) \mid A_j,\widehat{a}_j,a_j,\widehat{b}_j,D^n\}-A_jf_a(\widehat{a}_j,a_j,\widehat{b}_j;D^n) \mid A_j,\widehat{a}_j,\widehat{b}_j,D^n\right]$$
$$= \mathbb{E}\left[A_jf_a(\widehat{a}_j,a_j,\widehat{b}_j;D^n)-A_jf_a(\widehat{a}_j,a_j,\widehat{b}_j;D^n) \mid A_j,\widehat{a}_j,\widehat{b}_j,D^n\right]$$
$$= 0.$$

The last equality follows because $\widehat{Q}(\widehat{a}_i,\widehat{b}_i)$ does not depend on A_i so that

$$\mathbb{E}\{(A_i\widehat{a}_i-1)\widehat{Q}^{-1}(\widehat{a}_i,\widehat{b}_i) \mid X^n,A_{-i}^n,D^n\} = \left(\frac{\widehat{a}_i}{a_i}-1\right)\widehat{Q}^{-1}(\widehat{a}_i,\widehat{b}_i)$$
$$= \mathbb{E}\{(A_i(\widehat{a}_i-a_i)\widehat{Q}^{-1}(\widehat{a}_i,\widehat{b}_i) \mid X^n,A_{-i}^n,D^n\},$$

where $X^n = (X_1, \ldots, X_n)$ and $A^n_{-i} = A^n \setminus A_i$.

Next, by the smoothness assumption on f_a and the boundedness assumption on $\widehat{Q}^{-1}(\widehat{a}_i, \widehat{b}_i)$, we have

$$\begin{aligned}\left|f_a(\widehat{a}_j, a_j, \widehat{b}_j; D^n) - f_a(\widehat{a}_i, a_i, \widehat{b}_i; D^n)\right| &\lesssim |\widehat{a}_j - \widehat{a}_i|^\alpha + |a_j - a_i|^\alpha + |\widehat{b}_j - \widehat{b}_i|^\alpha \\ &\lesssim |\widehat{a}_j - \widehat{a}_i|^\alpha + |\widehat{a}_j - a_j|^\alpha + |\widehat{a}_i - a_i|^\alpha + |\widehat{b}_j - \widehat{b}_i|^\alpha\end{aligned}$$

so that

$$\begin{aligned}&\left|\mathbb{E}\left[\frac{1}{n(n-1)}\sum_{1_1 \neq j \leq n}\sum A_i(\widehat{a}_i - a_i)\frac{K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)}{\widehat{Q}(\widehat{a}_i, \widehat{b}_i)}A_j\{f_a(\widehat{a}_j, a_j, \widehat{b}_j; D^n) - f_a(\widehat{a}_i, a_i, \widehat{b}_i; D^n)\} \mid D^n\right]\right| \\ &\lesssim \mathbb{E}\left[|\widehat{a}_i - a_i|A_iA_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\{|\widehat{a}_j - \widehat{a}_i|^\alpha + |\widehat{a}_j - a_j|^\alpha + |\widehat{a}_i - a_i|^\alpha + |\widehat{b}_j - \widehat{b}_i|^\alpha\} \mid D^n\right] \\ &\lesssim h^\alpha \mathbb{E}\left[|\widehat{a}_i - a_i|\mathbb{E}\{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid X_i, D^n\} \mid D^n\right] \\ &\quad + \mathbb{E}\left\{|\widehat{a}_i - a_i|A_iA_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)|\widehat{a}_j - a_j|^\alpha| \, D^n\right\} \\ &\quad + \mathbb{E}\left[(\widehat{a}_i - a_i)^{1+\alpha}\mathbb{E}\{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid X_i, D^n\}\right] \\ &\lesssim h^\alpha\|\widehat{a} - a\| + \|\widehat{a} - a\|^{1+\alpha}\end{aligned}$$

under the condition that $\mathbb{E}\{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid X_i, D^n\}$. Thus, we also have that

$$|\mathbb{E}(\widehat{\psi} - \psi \mid D^n)| = |\mathbb{E}(T_n \mid D^n) - R_n| \lesssim h^\alpha\|\widehat{a} - a\| + \|\widehat{a} - a\|^{1+\alpha}.$$

Combining **Bound 1** and **Bound 2**, we conclude that

$$|\mathbb{E}(\widehat{\psi} - \psi \mid D^n)| \lesssim \left(h^\beta\|\widehat{b} - b\| + \|\widehat{b} - b\|^{1+\beta} + \frac{\|\widehat{a} - a\|\|\widehat{b} - b\|}{\sqrt{nh^2}}\right) \wedge \left(h^\alpha\|\widehat{a} - a\| + \|\widehat{a} - a\|^{1+\alpha}\right).$$

This concludes our proof of the bound on the bias.

C.4.2 Variance

The proof of the bound on the variance of $\widehat{\psi}$ conditional on the training sample D^n follows exactly the same logic as in the proof used to derive the bound on the conditional variance of ψ_a. Hence,

we omit some details. Recall the definitions

$$\begin{aligned}
\widehat{Q}(\widehat{a}_i, \widehat{b}_i) &= \frac{1}{n-1} \sum_{s=1,\ j\neq i}^{n} A_s K_{hi}(\widehat{a}_s) K_{hi}(\widehat{b}_s) \\
\widehat{Q}_{-l}(\widehat{a}_i, \widehat{b}_i) &= \frac{1}{n-1} \sum_{s=1,\ s\neq(i,l)}^{n} A_s K_{hi}(\widehat{a}_s) K_{hi}(\widehat{b}_s) \\
Q(\widehat{a}_i, \widehat{b}_i) &= \mathbb{E}\{\widehat{Q}(\widehat{a}_i, \widehat{b}_i) \mid X_i, D^n\} \\
&= \int \{a(x)\}^{-1} K_h(\widehat{a}(x) - a(X_i)) K_h(\widehat{b}(x) - b(X_i)) d\mathbb{P}(x) \\
&= \mathbb{P}(A=1) \int K(u)K(v) d\mathbb{P}_{\widehat{a},\widehat{b}|A=1,D^n}(uh + \widehat{a}(X_i), vh + \widehat{b}(X_i))
\end{aligned}$$

The estimator is $\widehat{\psi} = \mathbb{P}_n\widehat{\psi} - T_n = \mathbb{P}_n\widehat{\psi} - \widetilde{T}_n + (\widetilde{T}_n - T_n)$, where

$$\begin{aligned}
T_n &= \frac{1}{n(n-1)} \sum\sum_{1\leq i\neq j\leq n} (A_i\widehat{a}_i - 1)\{\widehat{Q}(\widehat{a}_i, \widehat{b}_i)\}^{-1} K_{hi}(\widehat{a}_j) K_{hi}(\widehat{b}_j) A_j(Y_j - \widehat{b}_j), \\
\widetilde{T}_n &= \frac{1}{n(n-1)} \sum\sum_{1\leq i\neq j\leq n} (A_i\widehat{a}_i - 1)\{Q(\widehat{a}_i, \widehat{b}_i)\}^{-1} K_{hi}(\widehat{a}_j) K_{hi}(\widehat{b}_j) A_j(Y_j - \widehat{b}_j).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathrm{var}(\widehat{\psi} \mid D^n) &\lesssim \mathrm{var}(\mathbb{P}_n\widehat{\varphi} \mid D^n) + \mathrm{var}(\widetilde{T}_n \mid D^n) + \mathbb{E}\{(\widetilde{T}_n - T_n)^2 \mid D^n\} \\
&\lesssim n^{-1} \vee (nh)^{-2} \vee \mathbb{E}\{(\widetilde{T}_n - T_n)^2 \mid D^n\}.
\end{aligned}$$

The last inequality follows by the independence and boundedness of the observations, as well as by Lemma 3 because

$$\begin{aligned}
&\mathbb{E}\left(\left[(A_i\widehat{a}_i - 1)^2 Q^{-2}(\widehat{a}_i, \widehat{b}_i) \mathbb{E}\left\{ K_{hi}(\widehat{a}_j) K_{hi}(\widehat{b}_j) A_j(Y_j - \widehat{b}_j) \mid D^n, X_i, A_i \right\} \right]^2 \mid D^n \right) \lesssim 1 \\
&\mathbb{E}\left\{ (A_i\widehat{a}_i - 1)^2 Q^{-2}(\widehat{a}_i, \widehat{b}_i) K_{hi}^2(\widehat{a}_j) K_{hi}^2(\widehat{b}_j) A_j(Y_j - \widehat{b}_j)^2 \mid D^n \right\} \lesssim h^{-2}.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\widetilde{T}_n - T_n &= \frac{1}{n(n-1)} \sum\sum_{1\leq i\neq j\leq n} (A_i\widehat{a}_i - 1)\{Q^{-1}(\widehat{a}_i, \widehat{b}_i) - \widehat{Q}^{-1}(\widehat{a}_i, \widehat{b}_i)\} K_{hi}(\widehat{a}_j) K_{hi}(\widehat{b}_j) A_j(Y_j - \widehat{b}_j) \\
&= \frac{1}{n(n-1)} \sum\sum_{1\leq i\neq j\leq n} T_{ij}
\end{aligned}$$

Just like in the proof for the bound on the variance of $\widehat{\psi}_a$, we break the square of the double sum above in seven terms:

$$\left(\sum_{1\le i\neq j\le n}\sum T_{ij}\right)^2 = \sum_{1\le i\neq j\le n}\sum (T_{ij}^2 + T_{ij}T_{ji}) + \sum_{1\le i\neq j\neq l\le n}\sum\sum (T_{ij}T_{il} + T_{ij}T_{li} + T_{ij}T_{jl} + T_{ij}T_{lj}) + \sum_{1\le i\neq j\neq l\neq k\le n}\sum\sum\sum T_{ij}T_{lk}.$$

We have

$$T_{ij}^2 \lesssim A_j h^{-2} K_{hi}(\widehat{a}_i)K_{hi}(\widehat{b}_i) \quad \text{and} \quad |T_{ij}T_{ji}| \lesssim A_i A_j h^{-2} K_{hi}(\widehat{a}_i)K_{hi}(\widehat{b}_i).$$

In this light,

$$\mathbb{E}\left(\sum_{1\le i\neq j\le n}\sum T_{ij}^2 + \sum_{1\le i\neq j\le n}\sum T_{ij}T_{ji} \mid D^n\right) \lesssim n(n-1)h^{-2},$$

under the condition that $\mathbb{E}\{A_j K_{hi}(\widehat{a}_i)K_{hi}(\widehat{b}_i) \mid X_i, D^n\} \lesssim 1$. Next, notice that

$$\begin{aligned}
|T_{ij}T_{il}| &\lesssim A_j A_l K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)K_{hi}(\widehat{a}_l)K_{hi}(\widehat{b}_l)\\
|T_{ij}T_{li}| &\lesssim A_i A_j K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)K_{hi}(\widehat{a}_l)K_{hi}(\widehat{b}_l)\\
|T_{ij}T_{jl}| &\lesssim A_j A_l K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)K_{hj}(\widehat{a}_l)K_{hj}(\widehat{b}_l)\\
|T_{ij}T_{lj}| &\lesssim A_j K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)K_{hj}(\widehat{a}_l)K_{hj}(\widehat{b}_l).
\end{aligned}$$

In this light, we have

$$\begin{aligned}
\mathbb{E}(|T_{ij}T_{il}|| \, D^n) &\lesssim \mathbb{E}\{A_j A_l K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)K_{hi}(\widehat{a}_l)K_{hi}(\widehat{b}_l) \mid D^n\}\\
&= \mathbb{E}\left[A_j K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)\mathbb{E}\{A_l K_{hi}(\widehat{a}_l)K_{hi}(\widehat{b}_l) \mid X_i, D^n\} \mid D^n\right]\\
&\lesssim \mathbb{E}\{A_j K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid D^n\}\\
&= \mathbb{E}\left[\mathbb{E}\{A_j K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j) \mid X_i, D^n\} \mid D^n\right]\\
&\lesssim 1.
\end{aligned}$$

The expectation of the terms $|T_{ij}T_{li}|$ and $|T_{ij}T_{jl}|$ can be similarly upper bounded by a constant. Therefore, we have

$$\mathbb{E}\left\{\sum_{1\le i\neq j\neq l\le n}\sum\sum (T_{ij}T_{il} + T_{ij}T_{li} + T_{ij}T_{jl}) \mid D^n\right\} \lesssim n(n-1)(n-2).$$

The remaining term to control is that involving the terms $T_{ij}T_{lk}$. We have

$$\begin{aligned}\widehat{Q}(\widehat{a}_i,\widehat{b}_i) &= \frac{1}{n-1}\sum_{s=1,s\neq i}^{n} A_s K_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_j)\\ &= \frac{1}{n-1}\sum_{s=1,s\neq (i,l)}^{n} A_s K_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s) + \frac{A_l K_{hi}(\widehat{a}_l)K_{hi}(\widehat{b}_l)}{n-1}\\ &\equiv \widehat{Q}_{-l}(\widehat{a}_i,\widehat{b}_i) + \frac{A_l K_{hi}(\widehat{a}_l)K_{hi}(\widehat{b}_l)}{n-1}.\end{aligned}$$

Next, we write

$$\begin{aligned}T_{ij}T_{lk} = {} & (A_i\widehat{a}_i - 1)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i,\widehat{b}_i) - Q^{-1}(\widehat{a}_i,\widehat{b}_i)\}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\\ & \times (A_l\widehat{a}_l - 1)\{\widehat{Q}_{-i}^{-1}(\widehat{a}_l,\widehat{b}_l) - Q^{-1}(\widehat{a}_l,\widehat{b}_l)\}K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)A_k(Y_k-\widehat{b}_k)\\ & - (A_i\widehat{a}_i - 1)\frac{A_l K_{hi}(\widehat{a}_l)K_{hi}(\widehat{b}_l)K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)}{(n-1)\widehat{Q}(\widehat{a}_i,\widehat{b}_i)\widehat{Q}_{-l}(\widehat{a}_i,\widehat{b}_l)}A_j(Y_j-\widehat{b}_j)\\ & \times (A_l\widehat{a}_l - 1)\{\widehat{Q}^{-1}(\widehat{a}_l,\widehat{b}_l) - Q^{-1}(\widehat{a}_l,\widehat{b}_l)\}K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)A_k(Y_k-\widehat{b}_k)\\ & - (A_i\widehat{a}_i - 1)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i,\widehat{b}_i) - Q^{-1}(\widehat{a}_i,\widehat{b}_i)\}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\\ & \times (A_l\widehat{a}_l - 1)\frac{A_i K_{hl}(\widehat{a}_i)K_{hl}(\widehat{b}_i)K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)}{(n-1)\widehat{Q}(\widehat{a}_l,\widehat{b}_l)\widehat{Q}_{-i}(\widehat{a}_l,\widehat{b}_l)}A_k(Y_k-\widehat{b}_k)\end{aligned}$$

The expectation of the last two terms can be upper bounded as

$$\begin{aligned}&\mathbb{E}\left[\left| -(A_i\widehat{a}_i - 1)\frac{A_l K_{hi}(\widehat{a}_l)K_{hi}(\widehat{b}_l)K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)}{(n-1)\widehat{Q}(\widehat{a}_i,\widehat{b}_i)\widehat{Q}_{-l}(\widehat{a}_i,\widehat{b}_l)}A_j(Y_j-\widehat{b}_j)\right.\right.\\ &\qquad \times (A_l\widehat{a}_l - 1)\{\widehat{Q}^{-1}(\widehat{a}_l,\widehat{b}_l) - Q^{-1}(\widehat{a}_l,\widehat{b}_l)\}K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)A_k(Y_k-\widehat{b}_k)\\ &\quad - (A_i\widehat{a}_i - 1)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i,\widehat{b}_i) - Q^{-1}(\widehat{a}_i,\widehat{b}_i)\}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\\ &\qquad \left.\left.\times (A_l\widehat{a}_l - 1)\frac{A_i K_{hl}(\widehat{a}_i)K_{hl}(\widehat{b}_i)K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)}{(n-1)\widehat{Q}(\widehat{a}_l,\widehat{b}_l)\widehat{Q}_{-i}(\widehat{a}_l,\widehat{b}_l)}A_k(Y_k-\widehat{b}_k)\right| \mid D^n\right]\\ &\lesssim \mathbb{E}\left\{\frac{A_l A_j A_k K_{hi}(\widehat{a}_l)K_{hi}(\widehat{b}_l)K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)}{(n-1)} \mid D^n\right\}\\ &\lesssim n^{-1}\end{aligned}$$

This means that, for $k=j$, we have the bound

$$|\mathbb{E}(T_{ij}T_{lj}\mid D^n)| \lesssim \mathbb{E}\{A_iA_jA_kA_lK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)K_{hl}(\widehat{a}_j)K_{hl}(\widehat{b}_j)\mid D^n\} + \frac{1}{n} \lesssim 1,$$

so that

$$\left|\mathbb{E}\left\{\sum\sum\sum_{1\leq i\neq j\neq l\leq n} T_{ij}T_{lj}\mid D^n\right\}\right| \lesssim n(n-1)(n-2).$$

Finally, for $k \neq j$, we have

$$
\begin{aligned}
&\Big|\mathbb{E}\Big[(A_i\widehat{a}_i-1)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i,\widehat{b}_i)-Q^{-1}(\widehat{a}_i,\widehat{b}_i)\}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(Y_j-\widehat{b}_j)\\
&\qquad\times(A_l\widehat{a}_l-1)\{\widehat{Q}_{-i}^{-1}(\widehat{a}_l,\widehat{b}_l)-Q^{-1}(\widehat{a}_l,\widehat{b}_l)\}K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)A_k(Y_k-\widehat{b}_k)\mid D^n\Big]\Big|\\
&=\Big|\mathbb{E}\Big[A_i(\widehat{a}_i-a_i)\{\widehat{Q}_{-l}^{-1}(\widehat{a}_i,\widehat{b}_i)-Q^{-1}(\widehat{a}_i,\widehat{b}_i)\}K_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)A_j(b_j-\widehat{b}_j)\\
&\qquad\times A_l(\widehat{a}_l-a_l)\{\widehat{Q}_{-i}^{-1}(\widehat{a}_l,\widehat{b}_l)-Q^{-1}(\widehat{a}_l,\widehat{b}_l)\}K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)A_k(b_k-\widehat{b}_k)\mid D^n\Big]\Big|\\
&\lesssim\Big|\mathbb{E}\Big[|\widehat{a}_i-a_i||b_j-\widehat{b}_j||\widehat{a}_l-a_l||b_k-\widehat{b}_k|A_iA_jA_kA_lK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)\\
&\qquad\times\mathbb{E}\Big\{|\widehat{Q}_{-l}(\widehat{a}_i,\widehat{b}_i)-Q(\widehat{a}_i,\widehat{b}_i)||\widehat{Q}_{-i}(\widehat{a}_l,\widehat{b}_l)-Q(\widehat{a}_l,\widehat{b}_l)||\,A_j,A_k,X_i,X_j,X_l,X_k,D^n\Big\}\Big]\mid D^n\Big|
\end{aligned}
$$

Further, we have

$$
\begin{aligned}
&\Big\{\widehat{Q}_{-l}(\widehat{a}_i,\widehat{b}_i)-Q(\widehat{a}_i,\widehat{b}_i)\Big\}^2=\left[\frac{1}{n-1}\sum_{1\le s\le n,s\neq(i,l)}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s)-Q(\widehat{a}_i,\widehat{b}_i)\}\right]^2\\
&=\frac{1}{(n-1)^2}\sum_{1\le s\neq m\le n,(s,m)\neq(i,l)}\sum\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s)-Q(\widehat{a}_i,\widehat{b}_i)\}\{A_mK_{hi}(\widehat{a}_m)K_{hi}(\widehat{b}_m)-Q(\widehat{a}_i,\widehat{b}_i)\}\\
&\quad+\frac{1}{(n-1)^2}\sum_{1\le s\le n,s\neq(i,l)}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s)-Q(\widehat{a}_i,\widehat{b}_s)\}^2
\end{aligned}
$$

In addition,

$$
\begin{aligned}
&\left|\mathbb{E}\left[\sum_{1\le s\neq m\le n,(s,m)\neq(i,l)}\sum\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s)-Q(\widehat{a}_i,\widehat{b}_i)\}\right.\right.\\
&\qquad\left.\left.\times\{A_mK_{hi}(\widehat{a}_m)K_{hi}(\widehat{b}_m)-Q(\widehat{a}_i,\widehat{b}_i)\}\mid A_j,A_k,X_i,X_j,X_k,X_l,D^n\right]\right|\\
&=\Big|2\{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)-Q(\widehat{a}_i,\widehat{b}_i)\}\{A_kK_{hi}(\widehat{a}_k)K_{hi}(\widehat{b}_k)-Q(\widehat{a}_i,\widehat{b}_i)\}\Big|\\
&\lesssim h^{-2}
\end{aligned}
$$

and

$$
\begin{aligned}
&\left|\mathbb{E}\left[\sum_{1\le s\le n,s\neq(i,l)}\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s)-Q(\widehat{a}_i,\widehat{b}_i)\}^2\mid D^n,A_j,A_k,X_i,X_j,X_k,X_l\right]\right|\\
&=\Big|(n-4)\mathbb{E}\Big[\{A_sK_{hi}(\widehat{a}_s)K_{hi}(\widehat{b}_s)-Q(\widehat{a}_i,\widehat{b}_i)\}^2\mid D^n,X_i\Big]\\
&\quad+\mathbb{E}\Big[\{A_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)-Q(\widehat{a}_i,\widehat{b}_i)\}^2\mid D^n,A_j,X_i,X_j\Big]\\
&\quad+\mathbb{E}\Big[\{A_kK_{hi}(\widehat{a}_k)K_{hi}(\widehat{b}_k)-Q(\widehat{a}_i,\widehat{b}_i)\}^2\mid D^n,A_k,X_i,X_k\Big]\Big|\\
&\lesssim(n-4)h^{-2}+h^{-4}
\end{aligned}
$$

In this light, under the condition that $nh^2 \to \infty$, we have reached

$$\mathbb{E}\left[\left\{\widehat{Q}_{-l}(\widehat{a}_i, \widehat{b}_i) - Q(\widehat{a}_i, \widehat{b}_i)\right\}^2 \mid D^n, A_j, A_k, X_i, X_j, X_k, X_l\right] \lesssim \frac{1}{nh^2}.$$

An identical reasoning yields

$$\mathbb{E}\left[\left\{\widehat{Q}_{-i}(\widehat{a}_l, \widehat{b}_l) - Q(\widehat{a}_l, \widehat{b}_l)\right\}^2 \mid D^n, A_j, A_k, X_i, X_j, X_k, X_l\right] \lesssim \frac{1}{nh^2}.$$

Therefore, by applying the Cauchy-Schwarz inequality, we have reached

$$\mathbb{E}\left\{|\widehat{Q}_{-l}(\widehat{a}_i, \widehat{b}_i) - Q(\widehat{a}_i, \widehat{b}_i)||\widehat{Q}_{-i}(\widehat{a}_l, \widehat{b}_l) - Q(\widehat{a}_l, \widehat{b}_l)|| \; D^n, A_j, A_k, X_i, X_j, X_l, X_k\right\} \lesssim \frac{1}{nh^2},$$

which yields

$$\begin{aligned}
&|\mathbb{E}(T_{ij}T_{lk} \mid D^n)| \\
&\lesssim \frac{1}{nh^2} \times \mathbb{E}\left[|\widehat{a}_i - a_i||\widehat{a}_l - a_l|K_{hi}(\widehat{a}_j)A_iA_jA_kA_lK_{hi}(\widehat{b}_j)K_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)|\widehat{b}_j - b_j||\widehat{b}_k - b_k|| \; D^n\right] + \frac{1}{n} \\
&= \frac{\mathbb{E}\left[|\widehat{a}_i - a_i|A_iA_jK_{hi}(\widehat{a}_j)K_{hi}(\widehat{b}_j)|\widehat{b}_j - b_j|| \; D^n\right]\mathbb{E}\left[A_kA_lK_{hl}(\widehat{a}_k)K_{hl}(\widehat{b}_k)|\widehat{a}_l - a_l||\widehat{b}_k - b_k|| \; D^n\right]}{nh^2} + \frac{1}{n} \\
&\lesssim \frac{\|\widehat{a} - a\|^2\|\widehat{b} - b\|^2}{nh^2} + \frac{1}{n}.
\end{aligned}$$

This means that we have reached that

$$\left|\mathbb{E}\left(\sum\sum\sum\sum\sum_{1\leq i\neq j\neq l\neq k\leq n} T_{ij}T_{lk} \mid D^n\right)\right| \lesssim n(n-1)(n-2)(n-3)\left(\frac{\|\widehat{a} - a\|^2\|\widehat{b} - b\|^2}{nh^2} + \frac{1}{n}\right)$$

Putting everything together, we have that

$$\mathbb{E}\{(\widetilde{T}_n - T_n)^2 \mid D^n\} \lesssim \frac{1}{n} \vee \frac{1}{(nh)^2} \vee \frac{\|\widehat{a} - a\|^2\|\widehat{b} - b\|^2}{nh^2}.$$

This concludes our derivation of the bound on $\text{var}(\widehat{\psi} \mid D^n)$.