# Minimax Kernel Machine Learning for a Class of Doubly Robust Functionals with Application to Proximal Causal Inference

**AmirEmad Ghassami**[1] **Andrew Ying**[2] **Ilya Shpitser**[1] **Eric Tchetgen Tchetgen**[2]

[1]Department of Computer Science, Johns Hopkins University
[2]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania

## Abstract

Robins et al. (2008) introduced a class of influence functions (IFs) which could be used to obtain doubly robust moment functions for the corresponding parameters. However, that class does not include the IF of parameters for which the nuisance functions are solutions to integral equations. Such parameters are particularly important in the field of causal inference, specifically in the recently proposed proximal causal inference framework of Tchetgen Tchetgen et al. (2020), which allows for estimating the causal effect in the presence of latent confounders. In this paper, we first extend the class of Robins et al. to include doubly robust IFs in which the nuisance functions are solutions to integral equations. Then we demonstrate that the double robustness property of these IFs can be leveraged to construct estimating equations for the nuisance functions, which enables us to solve the integral equations without resorting to parametric models. We frame the estimation of the nuisance functions as a minimax optimization problem. We provide convergence rates for the nuisance functions and conditions required for asymptotic linearity of the estimator of the parameter of interest. The experiment results demonstrate that our proposed methodology leads to robust and high-performance estimators for average causal effect in the proximal causal inference framework.

## 1 INTRODUCTION

Suppose independent and identically distributed data from a distribution $P$ over variables $V$ are given and we are interested in estimating a finite-dimensional functional of the distribution $\psi_0 := \psi(P)$. A common way to do so is to use moment functions to construct estimating equations for the parameter of interest. Such moment functions usually depend on other (possibly infinite-dimensional) functions indexing the distribution called nuisance functions. That is, despite the fact that these functions are not of primary interest, they may be needed to obtain an estimator of the target parameter. The concern is that bias in estimating the required nuisance functions can induce excessive bias for the parameter of interest, thus compromising one's ability for accurate inference about $\psi_0$.

For many applications, one can construct an estimating equation in which the moment function depends on two variation independent nuisance functions. An estimator based on such a moment function is called doubly robust if it is consistent even if one of the nuisance functions is not estimated consistently, provided that the other nuisance function is. That is, double robustness gives the user two chances to estimate the parameter of interest correctly. In addition, in cases where one of the nuisance functions has a slow convergence rate, a parametric ($\sqrt{n}$) convergence rate for the functional of interest can still be obtained if the other nuisance function can be estimated at a fast enough rate (Tsiatis, 2007). The average causal effect is perhaps the most well-studied functional for which under certain conditions a doubly robust moment function exists. For this functional, under the assumption of no unobserved confounders, the nuisance functions are the outcome regression function and the propensity score (Hernán and Robins, 2020).

Robins and Rotnitzky (2001) and Chernozhukov et al. (2016a) gave conditions for the existence and constructing doubly robust moment conditions in semi-parametric models. A common approach for obtaining such a moment function is based on using the

influence function (IF) of the parameter of interest. Robins et al. (2008) introduced a large class of doubly robust IFs in which nuisance functions are always in the form of a regression function. However, for some functionals of interest, specially in the field of causal inference, the nuisance functions are solutions to complicated integral equations. A prominent example of this case appears in the recently proposed proximal causal inference framework (Tchetgen Tchetgen et al., 2020; Miao et al., 2018; Cui et al., 2020). This framework enables estimating the average causal effect when unobserved confounders are present in the system, yet requires the existence of two conditionally independent sets of proxies of the latent confounders that are sufficiently rich to fulfill certain completeness conditions, also referred to as proximal relevance assumption (e.g., using measurements of biomarkers as proxies of patients underlying biological confounding variables). The proximal causal inference framework unifies and connects several existing causal identification and inference frameworks which leverage various types of proxies, such as the instrumental variables, negative controls, synthetic controls, and difference-in-differences. As an example of a nuisance function which is a solution to an integral equation, as we will explain in detail in Section 4, under the assumptions of the proximal causal inference framework, a nuisance function $h$ can be used for identifying the average causal effect of a treatment variable on an outcome variable, where $h$ is a solution to the integral equation in display (8).

Motivated by proximal causal inference framework, in this paper, we first extend the IF class of Robins et al. (2008) to include doubly robust IFs in which the nuisance functions are solutions to integral equations (Section 3). In this case, one cannot simply fit a flexible model for the nuisance functions by solving a regression problem. We show that in this case, the doubly robust moment functions, besides their aforementioned desired properties, can also be used for constructing estimating equations for the nuisance functions (Section 3.1). This enables us to solve the integral equations without resorting to parametric models. The main idea in our nuisance function estimation approach is to estimate each function such that it keeps the expected value of the moment function fixed for the perturbations in the other nuisance function. To implement this idea, we first define a perturbation function which captures the change in the mean of the moment function for a deviation in the nuisance functions. Then we solve a regularized minimax optimization problem which estimates one nuisance function as the function minimizing the perturbation for the worst-case deviation in the other nuisance function. The proposed approach yields unbiased estimat-

ing equations for each nuisance function that are free of any nuisance function. Moreover, the use of perturbation function elaborates the connection between nuisance function estimation and reduces the sensitivity/bias in the estimation of the functional of interest.

Framing the problem of estimating the nuisance parameters as an optimization problem enables us to use high-performance non-parametric machine learning tools to design our learners. Here we use reproducing kernel Hilbert spaces (RKHS) (Section 5) and present the closed-form solution for the minimax optimization problem for estimating the nuisance functions. We characterize the convergence rate of the nuisance functions based on the recently proposed method of Dikkala et al. (2020) that leverages the localized Rademacher complexity of the class of nuisance functions.

Equipped with high-performance estimators for the nuisance functions, we use the cross-fitting approach (Bickel and Ritov, 1988; Van der Vaart, 2000; Robins et al., 2008; Zheng and Van Der Laan, 2010; Chernozhukov et al., 2018) to design an estimator for the parameter of interest (Section 3.2). We characterize the measure of ill-posedness of the integral equations that we need to solve, and investigate the complexities caused by ill-posedness of the integral equations and their effect on convergence rate of the parameter of interest. We present the requirements on the convergence rate of the nuisance functions and ill-posedness of the system which guarantees the estimator to be asymptotically linear and root $n$ consistent, i.e., attains parametric convergence rate (Section 3.3). Hence, one can use the influence function of the proposed estimator to obtain confidence intervals for the parameter of interest. In Section 4, we demonstrate how proximal causal inference framework fits in our estimation setup and evaluate our proposed method by estimating average causal effect in the proximal causal inference framework on synthetic data as well as real-data in Section 6.

## 2 RELATED WORK

There are few other works on using a doubly robust moment function for estimating the nuisance functions. One approach is to estimate the nuisance functions by minimizing the variance of the doubly robust estimator (Cao et al., 2009; Tsiatis et al., 2011; van der Laan and Gruber, 2010). Another perspective is focusing on bias reduction rather than variance reduction and our proposed method also falls into this category (Van der Laan, 2014; Vermeulen and Vansteelandt, 2015; Avagyan and Vansteelandt, 2017; Cui and Tchetgen Tchetgen, 2019). Especially, Vermeulen and Vansteelandt (2015) proposed the bias reduced doubly robust estimation approach, which locally minimizes

the squared first-order asymptotic bias of the doubly robust estimator in the direction of the nuisance parameters under misspecification of both working models. Vermeulen and Vansteelandt (2015) only consider parametric working models for the nuisance functions; a restriction that we avoid in the present work similar to Robins et al. (2008) and Cui and Tchetgen Tchetgen (2019). In fact, our method can be viewed as a proposal for generalization of that work to the nonparametric setup.

There is a growing attention in the literature to the use of machine learning approaches for estimating causal quantities Athey et al. (2016); Farbmacher et al. (2020); Kallus and Zhou (2018); Nie and Wager (2017); Shalit et al. (2017); Wager and Athey (2018); Chernozhukov et al. (2019); Oprescu et al. (2019); Kallus et al. (2019); Dikkala et al. (2020); Bennett et al. (2019); Hartford et al. (2017), with a particular recent interest in minimax machine learning methods (Bennett et al., 2019; Dikkala et al., 2020; Muandet et al., 2019; Liao et al., 2020; Chernozhukov et al., 2020; Kallus et al., 2021). However, it is important to note that in several of these works such as (Dikkala et al., 2020) and (Bennett et al., 2019), the target is a dose-response curve. Therefore, the target of estimation is unique. In our work, we use the machine learning tools for estimating functions which are nuisance functions for our parameter of interest, and in general, they do not need to be uniquely identified for the parameter of interest to be uniquely identified.

Finally, regarding the proximal causal inference framework, existing results mostly rely on parametric assumptions for the working models for the nuisance functions, with the exception of (Singh, 2020), (Mastouri et al., 2021), and the independent, concurrent work (Kallus et al., 2021). However, the functional classes, assumptions, and convergence analysis for the parameter of interest in those works are different from ours. Specifically, Singh (2020) does not use the influence function for estimating the causal effect. Mastouri et al. (2021) only considers the so-called proximal g-formula approach for estimation, which is the result mentioned in Theorem 3 of our paper. Because the influence function is not used in that work, the bias will be first order and therefore, larger than the bias in our work which is not only second order but also product bias. That is, the bias in our method is guaranteed to be of smaller order. In fact, while our estimator is root-$n$ consistent (under the assumptions of Theorem 2), the estimator in that work will generally fail to be root-$n$ consistent due to slow convergence rate of their nonparametric estimator of the bridge function. Kallus et al. (2021) consider the use of a minimax learning method in proximal causal inference frame-

work. However, the functional class, assumptions, and convergence analysis for the parameter of interest in that work is different from ours. That work is primarily focused on average causal effect in proximal causal inference, yet we work with a broad functional class and we focus on the double robustness property to derive the estimators and convergence analysis. The functional class considered in (Kallus et al., 2021) is a special case of our class of functionals. This can be seen by including the function $\pi$ in that paper in functions $g_1$ and $g_2$ in our functional. Also, we make explicit use of double-robustness or product-bias property in establishing that our estimator is root-$n$ consistent, regular and asymptotically linear. Please see the Supplementary Materials for a detailed description of the related work.

## 3 PROPOSED METHOD

Let $V$ denote the variables from which independent and identically distributed data is collected, and let $\psi_0$ be the finite-dimensional parameter of interest. We consider the class of regular parameters $\psi_0$ for which the influence function (IF) is of the form[1]

$$
\begin{aligned}
IF_{\psi_0}(V) = {} & q_0(V_q)h_0(V_h)g_1(V) + q_0(V_q)g_2(V) \\
& + h_0(V_h)g_3(V) + g_4(V) - \psi_0,
\end{aligned} \tag{1}
$$

where $h_0(\cdot)$ and $q_0(\cdot)$ are nuisance functions, $V_q$ and $V_h$ are (not necessarily disjoint) subsets of $V$, and $g_1, g_2, g_3$, and $g_4$ are known measurable functions. The IF can be used as a moment function for estimating $\psi_0$, and with slight abuse of notation, we define the moment function

$$
\begin{aligned}
IF(V; \psi, q, h) \coloneqq {} & q(V_q)h(V_h)g_1(V) + q(V_q)g_2(V) \\
& + h(V_h)g_3(V) + g_4(V) - \psi.
\end{aligned}
$$

Parameters with IF of the form (1) appear in many settings, especially in causal inference and missing data problems. It includes functionals such as average causal effect, expected product of conditional expectations, expected conditional covariance, semiparametric regression, marginal structural mean models, marginal mean under missing at random assumption and certain cases of data missing not-at-random problems, etc. This also applies to the nonparametric instrumental variable problems. For instance, the moment function in (Chernozhukov et al., 2016b, p. 20) is in the proposed class with the choice of $\gamma$ as $h$ and $\lambda$ as $q$. This class is an extension of the class introduced in (Robins et al., 2008) in that we allow the nuisance functions to have different arguments. That is, in the class of Robins et al. (2008), $V_q = V_h$. However, as mentioned in Section 1, requiring that $V_q = V_h$ will exclude functionals for which the nuisance functions

---

[1]See (Van der Vaart, 2000) for the definition of regular parameters and their influence functions.

are not solutions to regression problems. Specially, it will exclude functionals in the proximal causal inference framework such as the proximal extensions of the aforementioned examples. Also it will exclude functionals arising in shadow variable approaches to missing data (Miao et al., 2015). In (Li et al., 2021), noting that $\mathbb{E}[h(X, Z) \mid R = 1, X, Y] = \gamma(X, Y)$, the influence function given in Equation (16) is in the proposed class with the choice of $\delta$ as $h$ and $\gamma$ as $q$. Our framework provides a unified nonparametric adversarial estimation framework for all these problems. An important example we will use in this paper is the proximal average causal effect, with the IF given in (10).

We require the following assumption on the IF, which is satisfied for a large class of doubly robust IFs, including the one for average causal effect, discussed in Section 4.

**Assumption 1.** *The nuisance functions $q_0$ and $h_0$ are variation independent, that is, they vary over a Cartesian product set $\mathcal{Q} \times \mathcal{H}$, where we assume $\mathcal{Q}$ and $\mathcal{H}$ are dense in $\mathcal{L}_2(P_{V_q})$ and $\mathcal{L}_2(P_{V_h})$, respectively. In addition, the (possibly infinite dimensional) parameters governing the marginal distribution of $V_q \cup V_h$ are variation independent of $q_0$ and $h_0$.*

As an example, in the proximal causal inference framework that we focus on in Section 4, one can always choose parameters such that Assumption 1 is satisfied. In the integral equation for the proximal setup for estimating the average causal effect, two out of three parameters of the function $h$, the conditional distribution of $W$ given $Z, A, X$, and the conditional distribution of $Y$ given $Z, A, X$ can always be chosen to be variational independent. One can choose to pick the first two, which is the requirement of Assumption 1.

The form of the IF (1) provides us with the double robustness property, which states that an estimator based on the moment function $IF(V; \psi, q, h)$ is consistent even if we misspecify one of the nuisance functions (but not both). This property is especially crucial in setups with high-dimensional or non-parametric nuisance functions.

**Proposition 1** (Double-robustness). *For all choices of nuisance functions $h$ and $q$,*

$$\mathbb{E}[IF(V; \psi_0, q_0, h_0)] = \mathbb{E}[IF(V; \psi_0, q, h_0)]$$
$$= \mathbb{E}[IF(V; \psi_0, q_0, h)] = 0. \quad (2)$$

## 3.1 Leveraging Double-Robustness to Estimate Nuisance Functions

We note that the nuisance functions in IF (1) are not functions of the parameter of interest. Therefore, we can first focus on estimating the nuisance functions. In this subsection, we establish that the doubly robust moment function, besides its desired properties

for estimating the parameter of interest, can also be leveraged for constructing estimating equations for the nuisance functions.

We define the perturbation at function pair $(q, h)$ towards the pair $(\dot{q}, \dot{h})$ as

$$prt(q, h; \dot{q}, \dot{h}) := IF(V; \psi_0, q + \dot{q}, h + \dot{h}) - IF(V; \psi_0, q, h).$$

Double robustness property in equation (2) indicates that at the pair $(q_0, h)$ for true function $q_0$ and any function $h$, the expected value of the perturbation function towards any other function $h + \dot{h}$ should be zero. This motivates the following approach for estimating the nuisance functions: We choose $(\hat{q}, \hat{h})$ as the true nuisance function pair if it is the solution to the following estimating equations.

$$\mathbb{E}[prt(\hat{q}, h; 0, \dot{h})] = 0 \qquad \forall h, \dot{h},$$
$$\mathbb{E}[prt(q, \hat{h}; \dot{q}, 0)] = 0 \qquad \forall q, \dot{q}.$$

Using these equations, and leveraging the fact that the IF is linear in each nuisance function, we propose the following optimization-based estimation approach for estimating the nuisance functions.

$$(\hat{q}, \hat{h}) = \arg\min_{(q,h)} \Big\{ \max_{\dot{h}} \mathbb{E}[prt(q, h; 0, \dot{h})] + \max_{\dot{q}} \mathbb{E}[prt(q, h; \dot{q}, 0)] \Big\} \quad (3)$$

Optimization (3) demonstrates the connection between estimating the nuisance functions and the parameter of interest. Specifically, it shows that finding the nuisance functions indeed requires minimizing the perturbation function, and hence the sensitivity of the parameter of interest to the nuisance functions.

We note that due to Assumption 1, the optimum value of (3) can be obtained by solving two separate optimization problems.

**Proposition 2.** *Under Assumption 1, the solution $(\hat{q}, \hat{h})$ to the optimization (3) is obtained by solving*

$$\hat{q} = \arg\min_q \max_{\dot{h}} \mathbb{E}\Big[\dot{h}(V_h)[q(V_q)g_1(V) + g_3(V)]\Big],$$
$$\hat{h} = \arg\min_h \max_{\dot{q}} \mathbb{E}\Big[\dot{q}(V_q)[h(V_h)g_1(V) + g_2(V)]\Big],$$

*That is, $\hat{q}$ and $\hat{h}$ can be estimated using separate optimizations. The solutions $(\hat{q}, \hat{h})$ satisfy*

$$\mathbb{E}[\hat{h}(V_h)g_1(V) + g_2(V)|V_q] = 0,$$
$$\mathbb{E}[\hat{q}(V_q)g_1(V) + g_3(V)|V_h] = 0. \quad (4)$$

In the case that $V_q = V_h$, conditional moment equations (4) suggest that the nuisance functions can be estimated by solving standard regression problems. However when $V_q \neq V_h$, those moment equations will be inverse problems known as Fredholm integral equation of the first kind.

Unfortunately the optimization problems in Proposition 2 are not stable. For instance, for the first optimization, if for a choice of $q$, $\mathbb{E}[q(V_q)g_1(V)+g_3(V) \mid V_h]$ is not zero, there exists a choice of $\dot{h}$ that can make $\mathbb{E}[\dot{h}(V_h)[q(V_q)g_1(V) + g_3(V)]]$ arbitrary large, regardless of how close $q$ is to the ground-truth. That is, the objective function can be made arbitrarily large for sufficiently rich function space for $\dot{h}$ unless $q$ is evaluated at the truth, and this can happen even in the neighborhood of the truth. Since in reality we work with finite data, for any choice of $q$ we get a large value for the objective function depending on the function space and data, and we cannot converge to the correct parameter; leading to miss-specification. Therefore, we add a squared regularization term for robustness against misspecification and stability of the optimization, as well as improving the convergence rates. Hence, we will instead focus on the following regularized optimization problems.

$$\hat{q} = \arg\min_{q}\max_{\dot{h}} \mathbb{E}\Big[\dot{h}(V_h)[q(V_q)g_1(V)+g_3(V)]-\dot{h}^2(V_h)\Big],$$
(5)

$$\hat{h} = \arg\min_{h}\max_{\dot{q}} \mathbb{E}\Big[\dot{q}(V_q)[h(V_h)g_1(V)+g_2(V)]-\dot{q}^2(V_q)\Big].$$
(6)

In the following result, we show that the proposed penalized optimizations solve the integral equations (4). That is, the stability term does not introduce bias.

**Theorem 1.** *The solutions to the optimization problems (5) and (6) satisfy the conditional moment equations in (4).*

Proposition 2 and Theorem 1 show the connection between perturbing doubly robust functionals and conditional moment function; an observation which to the best of our knowledge is new to this work.

### 3.2 Estimation Procedure

We use cross-fitting estimation approach of Chernozhukov et al. (2018) for separating the estimation of the nuisance functions from the parameter of interest. This approach provides us with the benefit that weaker smoothness requirements are needed for the nuisance functions. In the cross-fitting approach, we partition the samples into $L$ equal size parts $\{I_1, ..., I_L\}$. Consider the finite data version of the estimators from the optimization problems (5) and (6), obtained by replacing the population expectation operator $\mathbb{E}[\cdot]$ with the sample average operator $\hat{\mathbb{E}}[\cdot]$. For $\ell \in \{1, ..., L\}$, we estimate the nuisance functions $(\hat{q}_\ell, \hat{h}_\ell)$ on data from all parts but $I_\ell$. For all $\ell$, let $\hat{\psi}_\ell$ be the estimation of $\psi_0$ obtained by solving $\frac{1}{|I_\ell|}\sum_{i\in I_\ell} IF(V_i; \hat{\psi}_\ell, \hat{q}_\ell, \hat{h}_\ell) = 0$. Our final estimator of $\psi_0$ is obtained by

$$\hat{\psi} = \frac{1}{L}\sum_{\ell=1}^{L} \hat{\psi}_\ell.$$
(7)

### 3.3 Asymptotic Analysis

In this subsection, we study the asymptotic properties of the cross-fitting estimator in display (7). We require the following regularity conditions for the results.

**Assumption 2.** *(i) Functions $g_1$, $g_2$, $g_3$, and $g_4$ are bounded.*

*(ii) There exists a constant $\sigma_1$ such that $\big|\mathbb{E}[g_1(V)|V_q, V_h]\big| > \sigma_1 > 0$.*

*(iii) Functions $q_0$ and $h_0$ are square-integrable with respect to the underlying measure $P$.*

*(iv)* $\min\Big\{ \sup_{v_q} |\hat{q}_\ell(v_q)| + \sup_{v_h} |h_0(v_h)|,$

$\qquad \sup_{v_q} |q_0(v_q)| + \sup_{v_h} |\hat{h}_\ell(v_h)| \Big\} < \infty.$

In addition to the regularity conditions, we require conditions on the convergence of the nuisance functions. In our results, we show that the projected space is the right space to impose the convergence constraints on. Below, we use the conventional notations $\mathbb{P}_n[\cdot]$ and $P[\cdot]$ to denote empirical and population expectations with respect to the variable $V$. Note that for a function $\hat{f}$, unlike the operator $\mathbb{E}[\hat{f}(V)]$, in $P[\hat{f}(V)]$, the operator does not take the expectation with respect to the possible randomness in the function $\hat{f}$. For any function $f$, we use the norm notation $\|f\|_2 := \sqrt{P[f^2]}$. We need the following definition for our convergence assumption.

**Definition 1.** *For any given value $\delta > 0$, for given function spaces $\mathcal{Q}$ and $\mathcal{H}$, let $\mathcal{Q}^{|\delta} := \big\{q \in \mathcal{Q} : \big\|P[q(V_q)|V_h]\big\|_2 \leq \delta\big\}$, and $\mathcal{H}^{|\delta} := \big\{h \in \mathcal{H} : \big\|P[h(V_h)|V_q]\big\|_2 \leq \delta\big\}$. We define the local measure of ill-posedness as $\tau_q(\delta) := \sup_{q\in\mathcal{Q}^{|\delta}} \|q\|_2$, and $\tau_h(\delta) := \sup_{h\in\mathcal{H}^{|\delta}} \|h\|_2$.*

$\tau_q(\delta)$ and $\tau_h(\delta)$ are defined as measures of ill-posedness of the conditional expectation operator. This measure was originally proposed in (Chen and Pouzo, 2012) and is commonly used in the econometrics literature. The intuition behind this measure is as follows: Projecting functions on a space makes them closer (and hence hard to distinguish) in terms of the $L_2$-norm. The measure of ill-posedness relates a certain $L_2$ distance post-projection, to the corresponding largest $L_2$ distance of the functions before projection (i.e. the extent to which projection shrinks the distance). In other words, how much resolution is lost due to the projection. Therefore, a large ill-posedness measure $\tau_h(\delta)$ implies that $V_q$ does not capture much of the information in $V_h$.

**Assumption 3.** *(i) $\|\hat{q}_\ell - q_0\|_2 = o_p(1)$, $\|\hat{h}_\ell - h_0\|_2 = o_p(1)$, i.e., they converge to zero in probability.*

*(ii) $\big\|P[g_1(V)\{\hat{q}_\ell(V_q) - q_0(V_q)\}|V_h]\big\|_2 = O(r_q(n))$, and $\big\|P[g_1(V)\{\hat{h}_\ell(V_h) - h_0(V_h)\}|V_q]\big\|_2 = O(r_h(n))$, where $(r_q(n), r_h(n))$ satisfy*

$$\min\left\{r_h(n)\tau_q\big(r_q(n)\big), r_q(n)\tau_h\big(r_h(n)\big)\right\} = o(n^{-\frac{1}{2}}).$$

Part $(i)$ of Assumption 3 requires the consistency of the estimators of the nuisance functions. In Section 5, we provide an RKHS-based estimation framework, under which, for certain choices of the kernel (such as the Gaussian or Sobolev kernels), this assumption is satisfied. Part $(ii)$ of Assumption 3 requires a certain rate of convergence for the estimated nuisance functions. However, instead of requiring a convergence rate for each nuisance function, we require a rate for their product. Hence, in case that one of them is not converging fast enough the other nuisance function can compensate. This is one of the main attractions of double robust estimators. In Section 5, we will study the local measure of ill-posedness and convergence rates related to part $(ii)$ for the case that a reproducing kernel Hilbert space is used as the hypothesis class.

We have the following result regarding the convergence of the cross-fitting estimator.

**Theorem 2.** *Under Assumptions 2 and 3, the estimator $\hat{\psi}$ in (7) is asymptotically linear and satisfies*

$$\sqrt{n}\{\hat{\psi} - \psi_0\} = \sqrt{n}\,\hat{\mathbb{E}}[IF(V; \psi_0, q_0, h_0)] + o_p(1),$$

*where $o_p(1)$ demonstrates a sequence of variables which converges to zero in probability, and $\sqrt{n}\{\hat{\psi} - \psi_0\}$ converges in distribution to the Gaussian distribution $\mathcal{N}\big(0, var\,(IF(V; \psi_0, q_0, h_0))\big)$.*

As a corollary of Theorem 2, one can use the influence function of the proposed estimator to obtain confidence intervals for the parameter of interest.

**Remark 1.** *In our approach, we use an IF, which always leads to a second order bias. This fundamental property was recently popularized by Chernozhukov et al. (2018) in terms of Neyman Orthogonality property. However, note that our result is even stronger because not only does our estimator have second order bias, but also it has the so-called mixed bias property, which is the basis of double-robustness and is not guaranteed by Neyman Orthogonality. Please see the Supplementary Materials for a detailed discussion.*

## 4 AVERAGE CAUSAL EFFECT UNDER PROXIMAL CAUSAL INFERENCE FRAMEWORK

In this section, we apply the proposed minimax estimation approach to estimating average causal effect from observational data. Let $A$ be a binary treatment variable and $Y$ be the outcome variable. For $a \in \{0, 1\}$, let $Y^{(a)}$ be the counterfactual outcome variable representing the outcome if (contrary to the fact) the treatment is set to value $a$. The Average causal
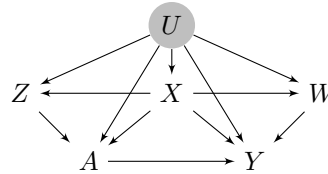


Figure 1: A proximal DAG.

effect (ACE) captures the difference in the expected value of the counterfactual outcome variables, that is $ACE = \mathbb{E}[Y^{(1)} - Y^{(0)}]$. Therefore, it suffices to focus on the problem of estimating the counterfactual mean of form $\psi_0 = \mathbb{E}[Y^{(a)}]$, for $a \in \{0, 1\}$. The most popular assumption for studying average causal effect is the so-called *conditional exchangeability* assumption (Hernán and Robins, 2020). Intuitively, this assumption requires that we have collected enough covariates of the units in the study that conditional on those covariates, the observational data is effectively as good as a conditionally randomized experiment. That is, conditional on the covariates, the units in the treated and untreated groups are exchangeable.

Despite its popularity, the exchangeability assumption is often violated even in laboratory settings as there may be confounders of the treatment and the outcome which are latent. Hence, causal inference approaches are needed that are capable of handling unobserved confounders. Proximal causal inference is one such approach in which, although latent confounders are allowed, presence of certain proxy variables for the confounders is required. Formally, the proximal causal inference framework allows for unobserved confounder variable (possibly vector-valued) $U$, observed confounder $X$, and requires access to proxy variables $Z$ and $W$ which satisfy the following conditions.

**Assumption 4.**

- $Y^{(a,z)} = Y^{(a)}$ almost surely, for all $a$ and $z$.
- $W^{(a,z)} = W$ almost surely, for all $a$ and $z$.
- $(Y^{(a)}, W) \perp\!\!\!\perp (A, Z)|(U, X)$, for $a \in \{0, 1\}$.

Note that Assumption 4 implies $Y \perp\!\!\!\perp Z|A, U, X$ and $W \perp\!\!\!\perp (A, Z)|U, X$, which can be taken as primitive identification conditions if one does not wish to entertain an intervention on $Z$. Figure 1 depicts an example graphical representation that satisfies these assumptions (the gray variable $U$ is unobserved).

In order to obtain identifiability, Miao et al. (2018) considered the following assumptions.

**Assumption 5** (Consistency and Positivity). $(i)$ $Y^{(A,Z)} = Y$, $(ii)$ $0 < Pr(A = 1|U, X) < 1$ a.s.

**Assumption 6.** $(i)$ For any $a$, $x$, if $\mathbb{E}[g(U)|Z, A = a, X = x] = 0$ almost surely, then $g(U) = 0$ almost surely. $(ii)$ There exists an outcome confounding

bridge function $h_0(w, a, x)$ that solves the following integral equation

$$\mathbb{E}[Y|Z, A, X] = \mathbb{E}[h_0(W, A, X)|Z, A, X]. \quad (8)$$

Miao et al. (2018) established the following nonparametric identification result.

**Theorem 3** (Miao et al. (2018)). *Under Assumptions 4, 5, and 6, the counterfactual mean $\mathbb{E}[Y^{(a)}]$ is nonparametrically identified by $\mathbb{E}[Y^{(a)}] = \mathbb{E}[h_0(W, a, X)]$.*

Cui et al. (2020) proposed an alternative proximal identification result based on the following counterpart of Assumption 6.

**Assumption 7.** $(i)$ *For any $a$, $x$, if $\mathbb{E}[g(U)|W, A = a, X = x] = 0$ almost surely, then $g(U) = 0$ almost surely. $(ii)$ There exists a treatment confounding bridge function $q_0(z, a, x)$ that solves the following integral equation*

$$\mathbb{E}[q_0(Z, a, X)|W, A = a, X] = \frac{1}{P(A = a|W, X)}. \quad (9)$$

Cui et al. (2020) established the following nonparametric identification result.

**Theorem 4** (Cui et al. (2020)). *Under Assumptions 4, 5, and 7, the counterfactual mean $\mathbb{E}[Y^{(a)}]$ is nonparametrically identified by $\mathbb{E}[Y^{(a)}] = \mathbb{E}[I(A = a)q_0(Z, a, X)Y]$.*

**Remark 2.** *Part $(i)$ in Assumption 6 and 7 is known as completeness condition; a technical condition central to the study of sufficiency in statistical inference. Regarding part $(ii)$ of these assumptions, note that equations (8) and (9) define inverse problems known as Fredholm integral equations of the first kind. A sufficient condition for part $(ii)$ of Assumption 6 is the following completeness assumption together with certain mild regularity conditions (Miao and Tchetgen Tchetgen, 2018): For any $a$, $x$, if $\mathbb{E}[g(Z)|W, A = a, X = x] = 0$ almost surely, then $g(Z) = 0$ a.s. Similar result holds for part $(ii)$ of Assumption 7.*

Let $V = (X, Z, W, A, Y)$. Cui et al. (2020) derived the locally semiparametric efficient influence function for the parameter of interest $\psi_0$ in a nonparametric model in which both confounding bridge functions are unrestricted, at the intersection submodel where both are uniquely identified as

$$IF_{\psi_0}(V) = -I(A = a)q_0(Z, A, X)h_0(W, A, X) \\ + I(A = a)Yq_0(Z, A, X) + h_0(W, a, X) - \psi_0, \quad (10)$$

This influence function satisfies our generic doubly robust functional form in expression (1) with the choice of $g_1(V) = -I(A = a)$, $g_2(V) = I(A = a)Y$, $g_3(V) = 1$, and $g_4(V) = 0$. Therefore, we can apply Theorem 1 to obtain the following result.

**Proposition 3.** *For arm $A=a$, let $(\tilde{q}_a, \tilde{h}_a)$ be a solution to the population-level minimax optimizations*

$$\tilde{h}_a = \arg\min_h \max_q \mathbb{E}[\{-I(A = a)h(W, X) \\ + I(A = a)Y\}q(Z, X) - q^2(Z, X)],$$
$$\tilde{q}_a = \arg\min_q \max_h \mathbb{E}[\{-I(A = a)q(Z, X) \\ + 1\}h(W, X) - h^2(W, X)].$$

*Then functions $\tilde{q}(z, a, x) = I(a = 0)\tilde{q}_0(z, x) + I(a = 1)\tilde{q}_1(z, x)$, and $\tilde{h}(w, a, x) = I(a = 0)\tilde{h}_0(w, x) + I(a = 1)\tilde{h}_1(w, x)$ solve the integral equations (8) and (9), respectively.*

Note that surprisingly, the propensity score does not appear in the moment equations in Proposition 3; a fact previously noted in (Cui et al., 2020), although motivated from a different perspective.

# 5 DOUBLY ROBUST KERNEL ESTIMATOR

In this section, we describe solving the minimax problems (5) and (6) when an RKHS is used as the hypothesis class. We demonstrate that in this case, fast convergence rates can be obtained and we present a closed-form solution for the optimization problem which renders an easy implementation of the method possible. We further provide results regarding estimating the ill-posedness measure for RKHSes.

Consider the finite data version of the estimators by replacing the population expectation operator $\mathbb{E}[\cdot]$ with sample average operator $\hat{\mathbb{E}}[\cdot]$. We propose the following Tikhonov regularization-based optimizations:

$$\hat{q} = \arg\min_{q \in \mathcal{Q}} \sup_{h \in \mathcal{H}} \hat{\mathbb{E}}[h(V_h)[q(V_q)g_1(V) + g_3(V)] \\ - h^2(V_h)] - \lambda_{\mathcal{H}}^q \|h\|_{\mathcal{H}}^2 + \lambda_{\mathcal{Q}}^q \|q\|_{\mathcal{Q}}^2, \quad (11)$$

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \sup_{q \in \mathcal{Q}} \hat{\mathbb{E}}[q(V_q)[h(V_h)g_1(V) + g_2(V)] \\ - q^2(V_q)] - \lambda_{\mathcal{Q}}^h \|q\|_{\mathcal{Q}}^2 + \lambda_{\mathcal{H}}^h \|h\|_{\mathcal{H}}^2, \quad (12)$$

where we assume $\mathcal{H}$ and $\mathcal{Q}$ are RKHSes with kernels $K_{\mathcal{H}}$ and $K_{\mathcal{Q}}$, respectively, equipped with the RKHS norms $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{Q}}$. Since the form of the two optimization problems are the same, in the following, we only discuss optimization problem (12). We drop the superscripts $h$ from $\lambda_{\mathcal{Q}}^h$ and $\lambda_{\mathcal{H}}^h$ to make the notations less cluttered, and denote sample size by $n$.

## 5.1 Closed-Form Solution for the Minimax Optimization

We start by showing that the optimization problem in Equation (12) has a closed-form solution. Define $K_{\mathcal{Q},n} = (K_{\mathcal{Q}}(V_{q_i}, V_{q_j}))_{i,j=1}^n$ and $K_{\mathcal{H},n} = (K_{\mathcal{H}}(V_{h_i}, V_{h_j}))_{i,j=1}^n$ as the empirical kernel matrices corresponding to spaces $\mathcal{Q}$ and $\mathcal{H}$, respectively.

**Proposition 4.** *Equation (12) achieves its optimum*

at $\hat{h} = \sum_{i=1}^n \alpha_i K_{\mathcal{H}}(V_{h_i}, \cdot)$, with $\alpha$ defined as

$$\alpha = - \big( K_{\mathcal{H},n} diag(g_{1,n}) \Gamma diag(g_{1,n}) K_{\mathcal{H},n} + n^2 \lambda_{\mathcal{H}} K_{\mathcal{H},n} \big)^\dagger$$
$$K_{\mathcal{H},n} diag(g_{1,n}) \Gamma g_{2,n},$$

where $\Gamma = \frac{1}{4} K_{\mathcal{Q},n} (\frac{1}{n} K_{\mathcal{Q},n} + \lambda_{\mathcal{Q}} I_n)^{-1}$, $diag(g_{1,n})$ is a diagonal matrix with $g_1(V_i)$ as the $i$-th diagonal entry, and $g_{2,n} := [g_2(V_1) \cdots g_2(V_n)]^\top$, and $\dagger$ denotes Moore-Penrose pseudoinverse.

## 5.2 Convergence Analysis of the Nuisance Functions

In Section 3, we observed that the solution to the population level optimization problem (6) satisfies the conditional moment equation $\mathbb{E}[\hat{h}(V_h) g_1(V) + g_2(V)|V_q] = 0$, which is equivalent to $\mathbb{E}[g_1(V)\{\hat{h}(V_h) - h_0(V_h)\}|V_q] = 0$. Hence, to quantify the performance of an estimator $\hat{h}$, we define the *projected risk* of the estimator as

$$R(\hat{h}) := \big\| P[g_1(V)\{\hat{h}(V_h) - h_0(V_h)\}|V_q] \big\|_2.$$

We will use the recently proposed approach of Dikkala et al. (2020) for bounding the projected risk of the regularized minimax estimator in (12). The bound that we provide is based on the critical radii of the involved function classes, which are defined by upper bounding the localized Rademacher complexity of the class. See the Supplementary Materials for the definition of localized Rademacher complexity, critical radius, as well as a computation method when an RKHS is used as the function class.

For any function class $\mathcal{F}$, let $\mathcal{F}_B := \{f \in \mathcal{F} : \|f\|_{\mathcal{F}}^2 \leq B\}$. We require the following conditions on the function classes $\mathcal{H}$ and $\mathcal{Q}$ for the convergence rate results.

**Assumption 8.** *(i) $\mathcal{H}$ and $\mathcal{Q}$ are classes of $b$-uniformly bounded functions, where without loss of generality, we assume $b = 1$. The functions $g_1$ and $g_2$ are also bounded. (ii) $\mathcal{Q}$ is a symmetric class, i.e., if $q \in \mathcal{Q}$, then $-q \in \mathcal{Q}$. Also, $\mathcal{Q}$ is a star-convex class, i.e., if $q \in \mathcal{Q}$, then $\alpha q \in \mathcal{Q}$, for all $\alpha \in [0,1]$. (iii) $h_0 \in \mathcal{H}$ with $B$ an upper bound on $\|h_0\|_{\mathcal{H}}^2$. (iv) For all $h \in \mathcal{H}$, $P[g_1(V)\{h(V_h) - h_0(V_h)\}|V_q] \in \mathcal{Q}_{L^2 \|h - h_0\|_{\mathcal{H}}^2}$. Specifically, with $U = 2L^2 B$, for all $h \in \mathcal{H}_B$, we have $P[g_1(V)\{h(V_h) - h_0(V_h)\}|V_q] \in \mathcal{Q}_U$.*

**Theorem 5.** *Define the function class $\mathcal{G}_B := \{V \to \alpha g_1(V)\{h(V_h) - h_0(V_h)\} P[g_1(V)\{h(V_h) - h_0(V_h)\}|V_q], where \alpha \in [0,1], h \in \mathcal{H}, h - h_0 \in \mathcal{H}_B\}$. Let $\delta_n$ be an upper bound on the critical radii of $\mathcal{Q}_U$ and $\mathcal{G}_B$, and define $\delta := \delta_n + c_0 \sqrt{\frac{\log(c_1/\zeta)}{n}}$, for some constants $c_0$ and $c_1$. If $\lambda_{\mathcal{Q}} \geq \frac{\delta^2}{U}$ and $\lambda_{\mathcal{H}} \geq \lambda_{\mathcal{Q}} L^2 c_2$, for some constant $c_2$, then under Assumption 8, with probability $1 - 3\zeta$, the estimator $\hat{h}$ in (12) satisfies*

$$R(\hat{h}) = O\Big(\delta + \frac{\lambda_{\mathcal{H}}}{\delta} \|h_0\|_{\mathcal{H}}^2\Big).$$

## 5.3 Characterization of Ill-posedness Measure

In Section 3, for any given value $\delta > 0$, we defined the local measure of ill-posedness. Due to the Tikhonov regularization used in our minimax estimator, assume that there exists a constant $C$, such that the output of the estimator satisfies $\hat{h}^* - h_0 \in \mathcal{H}_C$. Therefore, we define $\mathcal{H}_C^{|\delta} := \{h \in \mathcal{H}_C : \|P[h(V_h)|V_q]\|_2 \leq \delta\}$, and focus on the measure of ill-posedness $\tau_h(\delta) := \sup_{h \in \mathcal{H}_C^{|\delta}} \|h\|_2$.

If $\mathcal{H}$ is an RKHS, $\tau_h(\delta)$ can be bounded using the eigenvalues of the RKHS and a measure of dependency between $V_q$ and $V_h$. This also implies that convergence of the projected risk to zero leads to vanishing of the RMSE, i.e., as required in Assumption 3, the estimator is consistent. Let $\{\mu_j\}_{j=1}^\infty$ and $\{\phi_j\}_{j=1}^\infty$ be the eigenvalues and eigenfunctions of the RKHS $\mathcal{H}$. For any $m \in \mathbb{N}_+$, let $V_m$ be the $m \times m$ matrix with entry $(i,j)$ defines as $[V_m]_{i,j} = \mathbb{E}\big[\mathbb{E}[\phi_i(V_h)|V_q]\mathbb{E}[\phi_j(V_h)|V_q]\big]$. Note that the quantity on the right hand side of the equality measures how much the eigenfucntions are smoothened by the conditional expectation operator. For instance, if $V_h = V_q$, as is in the case of ACE under no unmeasured confounders, then $V_m$ will be the identity matrix of size $m$. Similarly, for the proximal setup, if the proxy variables are very informative about the unobserved confounder, $V_h$ and $V_q$ will be highly correlated, and the matrix $V_m$ will be close to identity matrix.

**Lemma 1** (Dikkala et al. (2020)). *Let $\lambda_m$ be the minimum eigenvalue of $V_m$, and suppose that for all $i \leq m < j$, $\big|\mathbb{E}\big[\mathbb{E}[\phi_i(V_h)|V_q]\mathbb{E}[\phi_j(V_h)|V_q]\big]\big| \leq c\lambda_m$, for some constant $c$. Then*
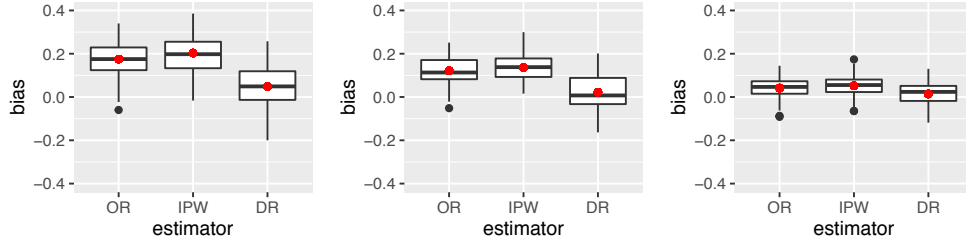
$$\tau_h(\delta)^2 \leq \min_{m \in \mathbb{N}_+} \Big\{ \frac{4\delta^2}{\lambda_m} + (4c^2 + 1)C\mu_{m+1} \Big\}.$$

As an example of the application of Lemma 1, suppose we use an RKHS, such as a Sobolev RKHS, with polynomial eigendecay, i.e., $\mu_m \sim m^{-a}$, and the smoothing effect of the conditional expectation operator is also polynomial, i.e., $\lambda_m \sim m^{-b}$, which implies that $\tau_h(\delta) = O(\delta^{\frac{a}{a+b}})$. By Theorem 2, we need $\sqrt{n} r_h(n) r_q(n)^{\frac{a}{a+b}} = o(1)$. Therefore, if $r_h(n) = n^{-r_h}$ and $r_q(n) = n^{-r_q}$, we require $\frac{1}{2} < r_h + \frac{a r_q}{a+b}$ for $\sqrt{n}$-convergence of the functional of interest.

# 6 EXPERIMENTS

**Synthetic-Data Experiments.** We investigate operational characteristics of our framework by applying doubly robust kernel estimators developed in Section 5 to estimate $h_0$ and $q_0$ in Section 4 in a range of simulation settings.[2] We generate the data such that

---

[2]The implementation is publicly available at https://github.com/andrewyyp/Kernel-Doubly-Robust.

Figure 2: Boxplots of simulation results when $n = 800, 1600, 3200$, respectively.

the assumptions in Section 4 hold. We consider sample size $n \in \{800, 1600, 3200\}$ and repeat each simulation for 100 times. The details of the data generating mechanisms as well as additional simulation results considering different data generating mechanisms are provided in the Supplementary Materials. We estimate the average causal effect using 5 folds cross fitting, with 5 folds cross validation to tune the hyperparameters and kernel bandwidths when learning $h_0$ and $q_0$. We then compare the performance of estimation of ACE obtained from the following three approaches: (i) Proximal outcome regression (POR) estimator, which is based on the result in Theorem 3, (ii) Proximal inverse probability weighting (PIPW) estimator, which is based on the result in Theorem 4, and (iii) Proximal doubly robust (PDR) estimator, which is based on the IF in display (10). The results of the comparisons are shown in Figure 2. All three estimators attain smaller bias as sample size becomes larger. As expected, the PDR estimator outperforms POR and PIPW estimators in all cases.

**Real-Data Analysis.** We reanalyze the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) with the aim of evaluating the causal effect of right heart catheterization (RHC) during the initial care of critically ill patients in the intensive care unit (ICU) on survival time up to 30 days (Connors et al., 1996). The same dataset has been analyzed using proximal framework in (Tchetgen Tchetgen et al., 2020; Cui et al., 2020) with parametric estimation on nuisance parameters. Data are available on 5735 individuals, 2184 treated and 3551 controls. In total, 3817 patients survived and 1918 died within 30 days. The outcome Y is the number of days between admission and death or censoring at day 30. We include 11 baseline covariates to adjust for potential confounding. We kept choices of $Z$ and $W$ the same as in (Tchetgen Tchetgen et al., 2020; Cui et al., 2020). Details on baseline covariates, hyperparameters and bandwidths are given in the Supplementary Materials. The results are summarized in Table 1.

Concordance between the three proximal estimators offers confidence in modeling assumptions, indicating

Table 1: Causal effect estimates (standard deviations) and 95% confidence intervals.

|  | ACE (SDs) | 95% CIs |
|---|---|---|
| $\hat{\psi}_{\text{POR}}$ | -1.76 (0.24) | (-2.23, -1.29) |
| $\hat{\psi}_{\text{PIPW}}$ | -1.59 (0.24) | (-2.07, -1.11) |
| $\hat{\psi}_{\text{PDR}}$ | -1.70 (0.24) | (-2.17, -1.22) |

that RHC may have a more harmful effect on 30 day survival among critically ill patients admitted into an intensive care unit than previously reported. In closing, our proximal estimates are well aligned with results obtained in (Tchetgen Tchetgen et al., 2020; Cui et al., 2020).

## 7 CONCLUSION

We have demonstrated that the double-robustness property of a moment function can be used for constructing estimating equations for the nuisance components of the moment function. We framed the idea in terms of a minimax optimization approach, in which we choose a nuisance function such that it minimizes the perturbation of the expected value of the moment function for the worst-case deviation in the other nuisance function. Framing the problem in terms of an optimization, enabled us to use tools from machine learning for implementing the idea and techniques from statistical learning theory to analyze the convergence rates. Specifically, we used kernel methods for designing the learners and Rademacher complexity analysis for obtaining the convergence rates. As is the case with most of nonparametric learners, the approach introduces challenges such as choosing the hyper parameters and having larger time complexity. However, it gives us robustness with respect to model misspecification compared to parametric learners. We provided conditions on the product error of the nuisance functions, as well as the ill-posedness of a conditional expectation operator to obtain the property of asymptotic linearity for the estimator of the parameter of interest. We investigated the application of the developed methodology in estimating the average causal effect in the proximal causal inference framework. The experiment results confirmed the superiority of the double robust learner compared to methods that only use one part of the distribution.

## Acknowledgements

## References

Athey, S., Imbens, G. W., and Wager, S. (2016). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*.

Avagyan, V. and Vansteelandt, S. (2017). Honest data-adaptive inference for the average treatment effect under model misspecification using penalised bias-reduced double-robust estimation. *arXiv preprint arXiv:1708.03787*.

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

Bennett, A., Kallus, N., and Schnabel, T. (2019). Deep generalized method of moments for instrumental variable analysis. *arXiv preprint arXiv:1905.12495*.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

Chen, X. and Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84.

Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., Demirer, M., Lewis, G., and Syrgkanis, V. (2019). Semi-parametric efficient policy learning with continuous actions.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016a). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016b). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033v2*.

Chernozhukov, V., Newey, W., Singh, R., and Syrgkanis, V. (2020). Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, 276(11):889–897.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2020). Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411*.

Cui, Y. and Tchetgen Tchetgen, E. (2019). Selective machine learning of doubly robust functionals. *arXiv preprint arXiv:1911.02029*.

Deaner, B. (2021). Proxy controls and panel data. `https://bdeaner.mit.edu/sites/default/files/images/Ben_Deaner_JMP_Proxy_Controls_and_Panel_Data_updated.pdf`.

Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. (2020). Minimax estimation of conditional moment models. *arXiv preprint arXiv:2006.07201*.

Dukes, O., Shpitser, I., and Tchetgen Tchetgen, E. (2021). Proximal mediation analysis. *arXiv preprint arXiv:2109.11904*.

Farbmacher, H., Huber, M., Lafférs, L., Langen, H., and Spindler, M. (2020). Causal mediation analysis with double machine learning. *arXiv preprint arXiv:2002.12710*.

Foster, D. J. and Syrgkanis, V. (2019). Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*.

Ghassami, A., Shpitser, I., and Tchetgen Tchetgen, E. (2021). Proximal causal inference with hidden mediators: Front-door and related mediation problems. *arXiv preprint arXiv:2111.02927*.

Ghassami, A., Shpitser, I., and Tchetgen Tchetgen, E. (2022). Combining experimental and observational data for identification of long-term causal effects. *arXiv preprint arXiv:2201.10743*.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR.

Hernán, M. A. and Robins, J. M. (2020). Causal inference: what if.

Imbens, G., Kallus, N., and Mao, X. (2021). Controlling for unmeasured confounding in panel data using minimal bridge functions: From two-way fixed effects to factor models. *arXiv preprint arXiv:2108.03849*.

Imbens, G., Kallus, N., Mao, X., and Wang, Y. (2022). Long-term causal inference under persistent confounding via data combination. *arXiv preprint arXiv:2202.07234*.

Kallus, N., Mao, X., and Uehara, M. (2019). Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *arXiv preprint arXiv:1912.12945*.

Kallus, N., Mao, X., and Uehara, M. (2021). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*.

Kallus, N. and Zhou, A. (2018). Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251. PMLR.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Li, W., Miao, W., and Tchetgen Tchetgen, E. (2021). Identification and estimation of nonignorable missing outcome mean without identifying the full data distribution. *arXiv preprint arXiv:2110.05776*.

Liao, L., Chen, Y.-L., Yang, Z., Dai, B., Wang, Z., and Kolar, M. (2020). Provably efficient neural estimation of structural equation model: An adversarial approach. *arXiv preprint arXiv:2007.01290*.

Lipsitch, M., Tchetgen Tchetgen, E., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*, pages 7512–7523. PMLR.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Miao, W., Liu, L., Tchetgen Tchetgen, E., and Geng, Z. (2015). Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556*.

Miao, W. and Tchetgen Tchetgen, E. (2018). A confounding bridge approach for double negative control inference on causal effects (supplement and sample codes are included). *arXiv preprint arXiv:1808.04945*.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. (2019). Dual instrumental variable regression. *arXiv preprint arXiv:1910.12358*.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135.

Nie, X. and Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.

Okui, R., Small, D. S., Tan, Z., and Robins, J. M. (2012). Doubly robust instrumental variable regression. *Statistica Sinica*, pages 173–205.

Oprescu, M., Syrgkanis, V., and Wu, Z. S. (2019). Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941. PMLR.

Robins, J., Li, L., Tchetgen Tchetgen, E., van der Vaart, A., et al. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics.

Robins, J. M., Li, L., Mukherjee, R., Tchetgen Tchetgen, E., and van der Vaart, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987.

Robins, J. M. and Rotnitzky, A. (2001). Comment on the bickel and kwon article,"inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of*

the *American Statistical Association*, 94(448):1096–1120.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Shi, X., Miao, W., Hu, M., and Tchetgen Tchetgen, E. (2021). Theory for identification and inference with synthetic controls: A proximal causal inference framework. *arXiv preprint arXiv:2108.13935*.

Shi, X., Miao, W., Nelson, J. C., and Tchetgen Tchetgen, E. J. (2020a). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):521–540.

Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2020b). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*, pages 1–13.

Singh, R. (2020). Kernel methods for unobserved confounding: Negative controls, proxies, and instruments. *arXiv preprint arXiv:2012.10315*.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen Tchetgen, E. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(3):348.

Tchetgen Tchetgen, E. J., Robins, J. M., and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97(1):171–180.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Tsiatis, A. A., Davidian, M., and Cao, W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics*, 67(2):536–545.

Van der Laan, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *The international journal of biostatistics*, 10(1):29–57.

van der Laan, M. J. and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, 6(1).

Van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vansteelandt, S., VanderWeele, T. J., Tchetgen Tchetgen, E. J., and Robins, J. M. (2008). Multiply robust inference for statistical interactions. *Journal of the American Statistical Association*, 103(484):1693–1704.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.

Vlassis, N., Hebda, P., McBride, S., and Noulas, A. (2020). On proximal causal learning with many hidden confounders. *arXiv preprint arXiv:2012.06725*.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Ying, A., Miao, W., Shi, X., and Tchetgen Tchetgen, E. (2021). Proximal causal inference for complex longitudinal studies. *arXiv preprint arXiv:2109.07030*.

Zheng, W. and Van Der Laan, M. J. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation.

# Supplementary Material:
# Minimax Kernel Machine Learning for a Class of Doubly Robust Functionals with Application to Proximal Causal Inference

## A   RELATED WORK

In recent years, doubly robust moment functions have been proposed for several different parameters of interest (Scharfstein et al., 1999; Robins and Rotnitzky, 2001; Bang and Robins, 2005; Kang et al., 2007; Vansteelandt et al., 2008; Rotnitzky et al., 2012; Van der Laan and Rose, 2011; Okui et al., 2012; Tchetgen Tchetgen and Shpitser, 2012; Chernozhukov et al., 2016a, 2018). The predominant approach for constructing doubly robust moment functions is by using the influence function of the parameter of interest. The seminal work of Scharfstein et al. (1999) took this approach for the study of estimating the average outcome in the presence of missing data. Chernozhukov et al. (2016a) derived several novel classes of doubly robust moment conditions by adding to identifying moment functions the nonparametric influence functions, and discussed sufficient conditions for the existence of doubly robust estimating functions. In our work, we study a general class of doubly robust influence functions as our moment function, which is an extension of a class introduced in (Robins et al., 2008).

There are few other works on using a doubly robust moment function for estimating the nuisance functions. One approach is to estimate the nuisance functions by minimizing the variance of the doubly robust estimator (Cao et al., 2009; Tsiatis et al., 2011; van der Laan and Gruber, 2010). Another perspective is focusing on bias reduction rather than variance reduction and our proposed method also falls into this category (Van der Laan, 2014; Vermeulen and Vansteelandt, 2015; Avagyan and Vansteelandt, 2017; Cui and Tchetgen Tchetgen, 2019). Especially, Vermeulen and Vansteelandt (2015) proposed the bias reduced doubly robust estimation approach, which locally minimizes the squared first-order asymptotic bias of the doubly robust estimator in the direction of the nuisance parameters under misspecification of both working models. Vermeulen and Vansteelandt (2015) only considers parametric working models and the estimating equations are based on the derivative of the functions. There is no straight forward generalization of their approach in the literature, and stability concerns in terms of framing that approach as optimization problem can arise. In fact, our method can be viewed as a proposal for generalization of that work to the non-parametric setup.

Our approach is related to the series of work (Blundell et al., 2007; Chen and Pouzo, 2012; Chen and Christensen, 2018) in which functionals of non-parametric instrumental variable response curve are considered. Specifically, Chen and Pouzo (2012) proposed a class of penalized sieve minimum distance estimators for this task. Our work provides analogous techniques for a large class of doubly robust functionals using minimax estimation of nuisance functions with the hypothesis class chosen to be an RKHS.

There is a growing attention in the literature to the use of machine learning approaches for estimating causal quantities (Athey et al., 2016; Farbmacher et al., 2020; Kallus and Zhou, 2018; Nie and Wager, 2017; Shalit et al., 2017; Wager and Athey, 2018; Chernozhukov et al., 2019; Oprescu et al., 2019; Kallus et al., 2019; Dikkala et al., 2020; Bennett et al., 2019; Hartford et al., 2017), with a particular recent interest in minimax machine learning methods (Bennett et al., 2019; Dikkala et al., 2020; Muandet et al., 2019; Liao et al., 2020; Chernozhukov et al., 2020; Kallus et al., 2021). It is important to note that in works such as (Dikkala et al., 2020) and (Bennett et al., 2019), the target is a dose-response curve. Therefore, the target of estimation is unique. In our work, functions $h$ and $q$ are just nuisance functions for our parameter of interest, and in general, they do not need to be uniquely identified for the parameter of interest to be uniquely identified. For the convergence analysis of the kernel estimator, we used the results of Dikkala et al. (2020). However, their loss function is a special case of the loss functions resulted from our function class. Also, we utilized the universality property of RKHSes and tailored and streamlined the proof and arguments to account for this property. Also, we modified the presentation form of the hyper parameters. Also, the starting point of our approach is a doubly robust functional not a conditional moment function as in (Dikkala et al., 2020), and the connection between the former and the latter was not stated in the literature before and is one of the contributions of our work. In our work, we start from perturbing

a doubly robust functional and prove that this can lead to a conditional moment function.

The authors in the concurrent work (Chernozhukov et al., 2020) also use results of Dikkala et al. (2020) for function estimation, yet they do not use the double-robustness property for estimating nuisance function and their approach is for a different functional class. The functional class considered in that work does not includes the influence function of the proximal framework. Therefore, we believe that class is not suited for the goal of our paper. Also, that work does not leverage the mixed bias property which is our main tool for estimation. In general, their functional class is restricted to those that do not require ill-posedness concerns regarding projection; both hypothesis spaces are on functions with same arguments.

Finally, regarding the proximal causal inference framework, there is a growing literature on using proxy variables for causal inference from observational data. Extensive discussion of proxy variables encountered in health and social sciences can be found in (Lipsitch et al., 2010; Kuroki and Pearl, 2014; Miao et al., 2018; Shi et al., 2020a,b; Sofer et al., 2016). The proximal causal inference framework has been recently also extended to other setups such as mediation analysis (Dukes et al., 2021; Ghassami et al., 2021), data fusion (Ghassami et al., 2022; Imbens et al., 2022), and longitudinal data analysis (Ying et al., 2021; Shi et al., 2021; Imbens et al., 2021). The specific model that we use in our paper is developed in (Miao et al., 2018; Miao and Tchetgen Tchetgen, 2018; Tchetgen Tchetgen et al., 2020). Specifically, Miao et al. (2018) established sufficient conditions for nonparametric identification of causal effects using a pair of proxies. Their setup was further studied in (Vlassis et al., 2020) and (Deaner, 2021). Also, Cui et al. (2020) developed an efficient semiparametric estimator for the proximal framework. Cui et al. (2020) also proposed estimating equations for the nuisance functions. That work considers a parametric working model for the nuisance functions. Then considers the parameters of the nuisance functions as the parameter of interest and derives the influence functions for these parameters in order to generate estimating equations for the latter. This leads to unconditional moment equations that may be used to estimate parametric models for nuisance bridge functions, yet under certain considerations, as we establish in this paper, these moments equations can generate conditional moment equations allowing for nonparametric estimation of bridge functions. We emphasize that our nonparametric conditional moment equations are generated using a completely different, and more general argument than that of Cui et al. as it applies to a broader class of functionals.

Existing results for proximal causal inference mostly rely on parametric assumptions for the working models for the nuisance functions, with the following exceptions: (Singh, 2020) which uses RKHSes, however, does not use the influence function for estimating the causal effect. Mastouri et al. (2021) only considers the so-called proximal g-formula approach for estimation, which is the result mentioned in Theorem 3 of our paper. Because the influence function is not used in that work, the bias will be first order and therefore, larger than the bias in our work which is not only second order but also product bias. That is, the bias in our method is guaranteed to be of smaller order. In fact, while our estimator is root-$n$ consistent (under the assumptions of Theorem 2), the estimator in that work will generally fail to be root-$n$ consistent due to slow convergence rate of their nonparametric estimator of the bridge function. Also, in an independent, concurrent work, Kallus et al. (2021) consider the use of a minimax learning method in proximal causal inference framework. However, the functional class, assumptions, and convergence analysis for the parameter of interest in that work is different from ours. The convergence analysis for the stabilized minimax optimization in that work is also based on the results of Dikkala et al. (2020). However, that work is primarily focused on ACE in proximal causal inference, yet we work with a broad functional class and we focus on the double robustness property to derive the estimators and convergence analysis. The functional class considered in (Kallus et al., 2021) is a special case of our class of functionals. This can be seen by including the function $\pi$ in that paper in functions $g_1$ and $g_2$ in our functional.

## A.1 Connection to the Double/Debiased Machine Learning Framework and Neyman Orthogonality

Our results are formally grounded in semiparametric theory. We use a first order influence function, which always leads to a second order bias (Bickel et al., 1993; Newey, 1990; Robins et al., 2017). This fundamental property was recently popularized by Chernozhukov et al. (2018) in terms of Neyman orthogonality property. However, note that our result is even stronger because not only our estimator has second order bias, but also it has the so-called mixed bias property, which is the basis of double-robustness. That is, simply having Neyman orthogonality property does not necessarily lead to double-robustness, which is the main feature of our function class, and hence the results of Chernozhokov et al. are not enough to derive ours. Also, due to the same fact, our proof techniques are different. Neyman orthogonality is not the starting assumption in our theorem even

though, again, by virtue of the fact that we are using an influence function, our functional does satisfy the Neyman orthogonality property.

Neyman Orthogonality can be interpreted as the moment equation incurring second order bias in terms of its nuisance functions. Influence functions in semiparametric/nonparametric models have been known to have this property since the 1980s, going back to the works of Newey, Bickel et al., and Van der Vaart. Therefore, all influence functions are guaranteed to satisfy Neyman Orthogonality. Nevertheless, Neyman Orthogonality can be considerably weaker than double robustness, as although influence functions are guaranteed to have second order bias, double robustness of an influence function requires evaluating the influence function under an appropriate choice of parametrization of the observed data distribution. See for example (Tchetgen Tchetgen et al., 2010) for an example of an influence function for the semiparametric logistic regression model which is not doubly robust if one of the nuisance parameters is the standard propensity score (i.e., Pr(Treatment = 1 | Covariates)), however, the same influence function becomes doubly robust if an alternative parametrization replacing the propensity score with the retrospective propensity score which further conditions on the outcome not having occurred (i.e., Pr(Treatment = 1 | Covariates, Outcome = 0)), a somewhat surprising result given the central role the standard propensity plays in causal inference.

Therefore, one could use arguments in Chernozhukov et al. (2018) to establish that the bias of our estimator must be second order, however those arguments are insufficient to establish the exact form of the second order bias, potentially leading to conservative statements about the order the bias. In fact, note that in (Chernozhukov et al., 2018), all nuisance functions are lumped together as one nuisance function, and the work requires a convergence rate for that nuisance function. On the other hand, in our work, we have two nuisance functions, and we do not require convergence rates on either of them, but we require a convergence rate on the product of the biases. That is, one of the nuisance functions can converge arbitrarily slow (contrary to (Chernozhukov et al., 2018)), as long as the other one converges fast enough so that their product is of order smaller than root $n$. The approach of Chernozhukov et al. (2018) generally requires both nuisance functions to converge at rate faster than quarter root $n$. It is the special form of our functional of interest which enabled us to obtain such important property.

Note that besides the aforementioned difference between the two works, due to the fact that we have different arguments for the nuisance functions, throughout the proof, we have to also deal with projection of one function space to the other for our convergence analysis, which is a feature new in our work.

## B  CRITICAL RADIUS

For the bounding approach in Section 5, we need the following definitions from the statistical learning theory literature (Wainwright, 2019).

**Definition 2** (Localized Rademacher Complexity). *For a given $\delta > 0$, and function class $\mathcal{F}$, the localized Rademacher complexity of $\mathcal{F}$ is defined as*

$$\mathcal{R}(\delta, \mathcal{F}) := \mathbb{E}_{\epsilon, X} \left[ \sup_{\substack{f \in \mathcal{F} \\ \|f\|_2 \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right],$$

*where $\{X_i\}_{i=1}^{n}$ are i.i.d. samples from the underlying distribution and $\{\epsilon_i\}_{i=1}^{n}$ are i.i.d. Rademacher variables taking values in $\{-1, +1\}$ with equal probability, independent of $\{X_i\}_{i=1}^{n}$.*

**Definition 3** (Critical Radius). *The critical radius of a function class $\mathcal{F}$ is the smallest solution $\delta_n^*$ to the inequality $\mathcal{R}(\delta, \mathcal{F}) \leq \delta^2$.*

Wainwright (2019) also provided the empirical counterparts of the localized Rademacher complexity and critical radius, which can be used to estimate the corresponding true values.

One of the attractive computational properties of RKHSes is that their localized Rademacher complexity can be determined by their eigendecay. For any function class $\mathcal{F}$, let $\mathcal{F}_B := \{f \in \mathcal{F} : \|f\|_{\mathcal{F}}^2 \leq B\}$. If $\mathcal{F}$ is an RKHS with eigenvalues $\{\mu_j\}_{j=1}^{\infty}$, then the localized Rademacher complexity of $\mathcal{F}_B$ can be upper bounded as (Wainwright, 2019)

$$\mathcal{R}(\delta, \mathcal{F}_B) \leq B \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\}}.$$

Therefore, we can find an upper bound for the critical radius of $\mathcal{F}_B$ by finding a solution to the inequality

$$B\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{\infty}\min\{\mu_j, \delta^2\}} \leq \delta^2.$$

## C   A REMARK ON THE MEASURE OF ILL-POSEDNESS

Chen and Pouzo (2012) also proposed the following measure of ill-posedness in their work

$$\tau = \sup_{h \in \mathcal{H}} \frac{\|h\|_2}{\left\|\mathrm{P}[h(V_h)|V_q]\right\|_2}.$$

This measure can also be used for bounding error $\|\hat{h} - h_0\|_2$ simply by

$$\|\hat{h} - h_0\|_2 \leq \tau \left\|\mathrm{P}[\hat{h}(V_h) - h_0(V_h)|V_q]\right\|_2.$$

However, we note that

$$\begin{aligned}
\tau_h(\delta) &= \delta \sup_{h \in \mathcal{H}^{|\delta}} \frac{\|h\|_2}{\delta} \\
&= \delta \sup_{h \in \mathcal{H}^{|\delta}} \frac{\|h\|_2}{\left\|\mathrm{P}[h(V_h)|V_q]\right\|_2} \cdot \frac{\left\|\mathrm{P}[h(V_h)|V_q]\right\|_2}{\delta} \\
&\leq \delta \sup_{h \in \mathcal{H}^{|\delta}} \frac{\|h\|_2}{\left\|\mathrm{P}[h(V_h)|V_q]\right\|_2} \\
&\leq \delta\tau.
\end{aligned}$$

Hence, working with $\tau_h(\delta)$, as we do in our approach, leads to tighter bounds.

## D   PROOFS

*Proof of Proposition 1.* Due to Assumption 1, the model governing the distribution of the data contains parametric submodels $\mathcal{M}^l = \{P^{l,t} : t \geq 0\}$ for $l = 1, 2, \ldots$ (the parameter of each submodel is denoted by $t$), such that

- $P^{l,0} = P$

- $P_X^{l,t} = P_X$

- $q^{l,t}(v_q) = q_0(v_q) + tf^l(v_q)$

- $h^{l,t}(v_h) = h_0(v_h)$

where the sequence $\{f^l\}$ is dense in $\mathcal{L}_2(P_{V_q})$.

Let $S^l(P)$ be the score of the submodel $P^{l,t}$ at $t = 0$, that is,

$$S^l(P) = \frac{d}{dt} \log P^{l,t}\Big|_{t=0}.$$

We note that

$$\begin{aligned}
\psi(P^{l,t}) &= \mathbb{E}_{P^{l,t}}[q^{l,t}(V_q)h^{l,t}(V_h)g_1(V) + q^{l,t}(V_q)g_2(V) + h^{l,t}(V_h)g_3(V) + g_4(V)] \\
&= \mathbb{E}_{P^{l,t}}[\{q_0(V_q) + tf^l(V_q)\}h_0(V_h)g_1(V) \\
&\quad + \{q_0(V_q) + tf^l(V_q)\}g_2(V) + h_0(V_h)g_3(V) + g_4(V)].
\end{aligned}$$

Therefore,

$$\frac{d}{dt}\psi(P^{l,t})\big|_{t=0} = \mathbb{E}_P[f^l(V_q)h_0(V_h)g_1(V) + f^l(V_q)g_2(V)]$$

$$+ \mathbb{E}_P[\{q_0(V_q)h_0(V_h)g_1(V) + q_0(V_q)g_2(V) + h_0(V_h)g_3(V) + g_4(V)\}S^l(P)].$$

But we notice that since $q_0(V_q)h_0(V_h)g_1(V) + q_0(V_q)g_2(V) + h_0(V_h)g_3(V) + g_4(V) - \psi_0$ is the influence function of $\psi_0$, we have

$$\frac{d}{dt}\psi(P^{l,t})\big|_{t=0} = \mathbb{E}_P[\{q_0(V_q)h_0(V_h)g_1(V) + q_0(V_q)g_2(V) + h_0(V_h)g_3(V) + g_4(V)\}S^l(P)],$$

which implies that

$$\mathbb{E}_P[f^l(V_q)h_0(V_h)g_1(V) + f^l(V_q)g_2(V)] = \mathbb{E}_P[\{h_0(V_h)g_1(V) + g_2(V)\}f^l(V_q)] = 0. \tag{13}$$

Equation (13) is true for every submodel $l$. Therefore, since the sequence $\{f^l\}$ is dense in $\mathcal{L}_2(P_{V_q})$, we conclude that $\mathbb{E}[h_0(V_h)g_1(V) + g_2(V)|V_q] = 0$. Similarly, we can show that $\mathbb{E}[q_0(V_q)g_1(V) + g_3(V)|V_h] = 0$.

For any choice of $q$, we have

$$\mathbb{E}[IF(V;\psi_0,q,h_0)] - \mathbb{E}[IF(V;\psi_0,q_0,h_0)]$$
$$= \mathbb{E}[g_1(V)h_0(V_h)\{q(V_q) - q_0(V_q)\} + g_2(V)\{q(V_q) - q_0(V_q)\}]$$
$$= \mathbb{E}[\{g_1(V)h_0(V_h) + g_2(V)\}\{q(V_q) - q_0(V_q)\}]$$
$$= \mathbb{E}[\mathbb{E}[h_0(V_h)g_1(V) + g_2(V)|V_q]\{q(V_q) - q_0(V_q)\}] = 0,$$

For any choice of $h$, we have

$$\mathbb{E}[IF(V;\psi_0,q_0,h)] - \mathbb{E}[IF(V;\psi_0,q_0,h_0)]$$
$$= \mathbb{E}[g_1(V)q_0(V_q)\{h(V_h) - h_0(V_h)\} + g_3(V)\{h(V_h) - h_0(V_h)\}]$$
$$= \mathbb{E}[\{g_1(V)q_0(V_q) + g_3(V)\}\{h(V_h) - h_0(V_h)\}]$$
$$= \mathbb{E}[\mathbb{E}[q_0(V_q)g_1(V) + g_3(V)|V_h]\{h(V_h) - h_0(V_h)\}] = 0,$$

$\square$

*Proof of Proposition 2.* We note that the function $q$ does not appear in

$$\mathbb{E}[prt(q,h;\dot{q},0)] = \mathbb{E}\big[\dot{q}(V_q)[h(V_h)g_1(V) + g_2(V)]\big].$$

Similarly, the function $h$ does not appear in

$$\mathbb{E}[prt(q,h;0,\dot{h})] = \mathbb{E}\big[\dot{h}(V_h)[q(V_q)g_1(V) + g_3(V)]\big].$$

Also, recall that the nuisance functions are assumed to be variation independent. Therefore, $\hat{q}$ and $\hat{h}$ can be estimated using separate optimizations.

Since $\mathbb{E}[\dot{q}(V_q)[\hat{h}(V_h)g_1(V) + g_2(V)]] = 0$ for all $\dot{q}$, the choice of $\dot{q}(V_q) = \mathbb{E}[\hat{h}(V_h)g_1(V) + g_2(V) \mid V_q]$ and law of iterative expectations imply that $\mathbb{E}[\hat{h}(V_h)g_1(V) + g_2(V)|V_q] = 0$. Similarly, we conclude that $\mathbb{E}[\hat{q}(V_q)g_1(V) + g_3(V)|V_h] = 0$.

$\square$

*Proof of Theorem 1.* We note that

$$\min_h \max_{\dot{q}} \mathbb{E}\big[\dot{q}(V_q)[h(V_h)g_1(V) + g_2(V)] - \dot{q}^2(V_q)\big]$$

$$= \min_h -\Big(\min_{\dot{q}} \mathbb{E}\big[(\dot{q}(V_q) - \frac{1}{2}[h(V_h)g_1(V) + g_2(V)])^2\big] - \frac{1}{4}\mathbb{E}\big[[h(V_h)g_1(V) + g_2(V)]^2\big]\Big)$$

Similarly,

$$\min_q \max_{\dot{h}} \mathbb{E}\Big[\dot{h}(V_h)[q(V_q)g_1(V) + g_3(V)] - \dot{h}^2(V_h)\Big]$$

$$= \min_q -\Big( \min_{\dot{h}} \mathbb{E}\Big[(\dot{h}(V_h) - \frac{1}{2}[q(V_q)g_1(V) + g_3(V)])^2\Big] - \frac{1}{4}\mathbb{E}\Big[[q(V_q)g_1(V) + g_3(V)]^2\Big]\Big),$$

The inner optimizations achieve their optimum at

$$\dot{q}^*(V_q) = \frac{1}{2}\mathbb{E}[h(V_h)g_1(V) + g_2(V)|V_q],$$

$$\dot{h}^*(V_h) = \frac{1}{2}\mathbb{E}[q(V_q)g_1(V) + g_3(V)|V_h],$$

Therefore, the outer minimization problems are

$$\min_h \frac{1}{4}\mathbb{E}\Big[\mathbb{E}[h(V_h)g_1(V) + g_2(V)|V_q]^2\Big] \geq 0,$$

$$\min_q \frac{1}{4}\mathbb{E}\Big[\mathbb{E}[q(V_q)g_1(V) + g_3(V)|V_h]^2\Big] \geq 0.$$

But we note that zero is achievable. This is due to Proposition 2, where zero is achieved at the ground-truth nuisance functions $q_0$ and $h_0$. Therefore, the optimizers satisfy

$$\mathbb{E}[h(V_h)g_1(V) + g_2(V)|V_q] = 0,$$

$$\mathbb{E}[q(V_q)g_1(V) + g_3(V)|V_h] = 0.$$

$\square$

*Proof of Theorem 2.* To simplify the notations, assume the size of each partition of the data is $m$, that is, $n = mL$. Let $(\hat{q}_\ell, \hat{h}_\ell)$ be the estimators of $(q_0, h_0)$ obtained on subsamples $I_\ell^c$. For parameters $q$, $h$, and $\psi$ define

$$\phi(V; q, h) := IF(V; \psi, q, h) - \psi,$$

which implies that $\hat{\psi}_\ell$ can be written as

$$\hat{\psi}_\ell = \mathbb{P}_m^\ell[\phi(V; \hat{q}_\ell, \hat{h}_\ell)].$$

Therefore, we have

$$\sqrt{n}\{\hat{\psi} - \psi_0\} = \sqrt{n}\Big\{\frac{1}{L}\sum_{\ell=1}^L \mathbb{P}_m^\ell[\phi(V; \hat{q}_\ell, \hat{h}_\ell)] - \psi_0]\Big\}$$

$$= \frac{\sqrt{n}}{L\sqrt{m}}\sum_{\ell=1}^L \Big\{\Big\{\mathbb{G}_m^\ell[\phi(V; \hat{q}_\ell, \hat{h}_\ell)] - \mathbb{G}_m^\ell[\phi(V; q_0, h_0)]\Big\}$$

$$+ \mathbb{G}_m^\ell[\phi(V; q_0, h_0)]$$

$$+ \sqrt{m}\Big\{P[\phi(V; \hat{q}_\ell, \hat{h}_\ell)] - \psi_0\Big\}\Big\}$$

where for any function $f(V)$,

$$\mathbb{G}_n[f] = \sqrt{n}\{\mathbb{P}_n[f] - P[f]\}$$

represents the empirical process. Recalling that $n = mL$, we have

$$\sqrt{n}\{\hat{\psi} - \psi_0\} = \frac{1}{\sqrt{L}}\sum_{\ell=1}^L \Big\{\mathbb{G}_m^\ell[\phi(V; \hat{q}_\ell, \hat{h}_\ell)] - \mathbb{G}_m^\ell[\phi(V; q_0, h_0)]\Big\} \qquad (T_1)$$

$$+ \frac{1}{\sqrt{L}}\sum_{\ell=1}^L \mathbb{G}_m^\ell[\phi(V; q_0, h_0)] \qquad (T_2)$$

$$+ \frac{1}{\sqrt{L}}\sum_{\ell=1}^L \sqrt{m}\Big\{P[\phi(V; \hat{q}_\ell, \hat{h}_\ell)] - \psi_0\Big\} \qquad (T_3)$$

We will show that $T_2$ provides us with the term $\sqrt{n}\ \hat{\mathbb{E}}[IF(V;\psi_0,q_0,h_0)]$, and under the assumptions of the theorem, $T_1$ and $T_3$ are $o_p(1)$.

## Analysis of $T_1$

Let $V^{\ell,c}$ be the data in all but the $\ell$-th fold, and define

$$A_m^\ell := \left\{\mathbb{G}_m^\ell[\phi(V;\hat{q}_\ell,\hat{h}_\ell)] - \mathbb{G}_m^\ell[\phi(V;q_0,h_0)]\right\}.$$

We note that

$$
\begin{aligned}
var(A_m^\ell|V^{\ell,c}) &= m\ var\big(\mathbb{P}_m^\ell[\phi(V;\hat{q}_\ell,\hat{h}_\ell)] - \mathbb{P}_m^\ell[\phi(V;q_0,h_0)]\big|V^{\ell,c}\big) \\
&= var\big(\phi(V;\hat{q}_\ell,\hat{h}_\ell) - \phi(V;q_0,h_0)\big|V^{\ell,c}\big) \\
&\leq \mathrm{P}\big[\{\phi(V;\hat{q}_\ell,\hat{h}_\ell) - \phi(V;q_0,h_0)\}^2\big|V^{\ell,c}\big] \\
&= \left\|\phi(V;\hat{q}_\ell,\hat{h}_\ell) - \phi(V;q_0,h_0)\right\|_2^2 \\
&= \Big\|g_1(V)\hat{q}_\ell(V_q)\hat{h}_\ell(V_h) + g_2(V)\hat{q}_\ell(V_q) + g_3(V)\hat{h}_\ell(V_h) \\
&\qquad - g_1(V)q_0(V_q)h_0(V_h) - g_2(V)q_0(V_q) - g_3(V)h_0(V_h)\Big\|_2^2.
\end{aligned}
$$

Note that the last expression is equal to both

$$\left\|(\hat{h}_\ell(V_h) - h_0(V_h))(g_3(V) + g_1(V)\hat{q}_\ell(V_q)) + (\hat{q}_\ell(V_q) - q_0(V_q))(g_2(V) + g_1(V)h_0(V_h))\right\|_2^2$$

and

$$\left\|(\hat{h}_\ell(V_h) - h_0(V_h))(g_3(V) + g_1(V)q_0(V_q)) + (\hat{q}_\ell(V_q) - q_0(V_q))(g_2(V) + g_1(V)\hat{h}_\ell(V_h))\right\|_2^2.$$

Therefore,

$$
\begin{aligned}
var(A_m^\ell|V^{\ell,c}) &\leq 2\min\Bigg\{\left\|(\hat{h}_\ell(V_h) - h_0(V_h))(g_3(V) + g_1(V)\hat{q}_\ell(V_q))\right\|_2^2 \\
&\qquad + \left\|(\hat{q}_\ell(V_q) - q_0(V_q))(g_2(V) + g_1(V)h_0(V_h))\right\|_2^2, \\
&\qquad \left\|(\hat{h}_\ell(V_h) - h_0(V_h))(g_3(V) + g_1(V)q_0(V_q))\right\|_2^2 \\
&\qquad + \left\|(\hat{q}_\ell(V_q) - q_0(V_q))(g_2(V) + g_1(V)\hat{h}_\ell(V_h))\right\|_2^2\Bigg\} \\
&\leq 2\min\Bigg\{\sup_v\{(g_3(v) + g_1(v)\hat{q}_\ell(v_q))^2\}\|\hat{h}_\ell - h_0\|_2^2 \\
&\qquad + \sup_v\{(g_2(v) + g_1(v)h_0(v_h))^2\}\|\hat{q}_\ell - q_0\|_2^2, \\
&\qquad \sup_v\{(g_3(v) + g_1(v)q_0(v_q))^2\}\|\hat{h}_\ell - h_0\|_2^2 \\
&\qquad + \sup_v\{(g_2(v) + g_1(v)\hat{h}_\ell(v_h))^2\}\|\hat{q}_\ell - q_0\|_2^2\Bigg\}
\end{aligned}
$$

By Assumptions 2 and 3, $\|\hat{q}_\ell - q_0\|_2$ and $\|\hat{h}_\ell - h_0\|_2$ converge to zero in probability, and either the first two, or the second two supremums are finite, which implies that $var(A_m^\ell|V^{\ell,c})$ converges to zero in probability as $m \to \infty$. Also note that $\mathrm{P}[A_m^\ell|V^{\ell,c}] = 0$. Therefore,

$$\mathrm{P}[(A_m^\ell)^2|V^{\ell,c}] = var(A_m^\ell|V^{\ell,c}) \xrightarrow{p} 0 \qquad \text{as } m \to \infty.$$

Then, by conditional Chebyshev inequality, for all $\delta > 0$, we have

$$P(|A_m^\ell| > \delta \mid V^{\ell,c}) \xrightarrow{p} 0 \qquad \text{as } m \to \infty.$$

That is, $P(|A_m^\ell| > \delta \mid V^{\ell,c})$ is a bounded sequence that converges to zero in probability. Therefore,

$$\mathbb{E}\big[P(|A_m^\ell| > \delta \mid V^{\ell,c})\big] = P(|A_m^\ell| > \delta) \xrightarrow{p} 0 \qquad \text{as } m \to \infty.$$

That is, $A_m^\ell \xrightarrow{p} 0$ as $m \to \infty$.

The conclusion holds for all $\ell \in \{1, ..., L\}$ and hence, we conclude that $T_1 = o_p(1)$.

## Analysis of $T_2$

$$\begin{aligned}
T_2 &= \frac{1}{\sqrt{L}} \sum_{\ell=1}^{L} \mathbb{G}_m^\ell[\phi(V; q_0, h_0)] \\
&= \frac{\sqrt{m}}{\sqrt{L}} \sum_{\ell=1}^{L} \left\{ \mathbb{P}_m^\ell[\phi(V; q_0, h_0)] - \mathrm{P}[\phi(V; q_0, h_0)] \right\} \\
&= \frac{\sqrt{m}}{\sqrt{L}} \frac{1}{m} \sum_{\ell=1}^{L} \sum_{i=1}^{m} \left\{ \phi(V_i; q_0, h_0) - \mathrm{P}[\phi(V; q_0, h_0)] \right\} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \phi(V_i; q_0, h_0) - \mathrm{P}[\phi(V; q_0, h_0)] \right\}.
\end{aligned}$$

By central limit theorem, the last expression converges to the distribution $\mathcal{N}(0, var\ (\phi(V; q_0, h_0)))$ if $\mathrm{P}[\phi(V; q_0, h_0)^2] < \infty$. Note that $\mathrm{P}[\phi(V; q_0, h_0)^2]$ can be upper bounded as

$$\begin{aligned}
&\mathrm{P}[\phi(V; q_0, h_0)^2] \\
&= \|\phi(V; q_0, h_0)\|_2^2 \\
&= \|g_1(V)q_0(V_q)h_0(V_h) + g_2(V)q_0(V_q) + g_3(V)h_0(V_h) + g_4(V)\|_2^2 \\
&\leq 2\|\{g_2(V) + g_1(V)h_0(V_h)\}q_0(V_q)\|_2^2 + 2\|g_3(V)h_0(V_h)\|_2^2 + 2\|g_4(V)\|_2^2 \\
&\leq 2\sup_v\{(g_2(v) + g_1(v)h_0(v_h))^2\}\|q_0(V_q)\|_2^2 + 2\sup_v\{g_3(V)^2\}\|h_0(V_h)\|_2^2 + 2\sup_v\{g_4(V)^2\},
\end{aligned}$$

or upper bounded as

$$\begin{aligned}
&\mathrm{P}[\phi(V; q_0, h_0)^2] \\
&= \|\phi(V; q_0, h_0)\|_2^2 \\
&= \|g_1(V)q_0(V_q)h_0(V_h) + g_2(V)q_0(V_q) + g_3(V)h_0(V_h) + g_4(V)\|_2^2 \\
&\leq 2\|\{g_3(V) + g_1(V)q_0(V_q)\}h_0(V_h)\|_2^2 + 2\|g_2(V)q_0(V_q)\|_2^2 + 2\|g_4(V)\|_2^2 \\
&\leq 2\sup_v\{(g_3(v) + g_1(v)q_0(v_q))^2\}\|h_0(V_h)\|_2^2 + 2\sup_v\{g_2(V)^2\}\|q_0(V_q)\|_2^2 + 2\sup_v\{g_4(V)^2\}.
\end{aligned}$$

Therefore by Assumption 2, we have $\mathrm{P}[\phi(V; q_0, h_0)^2] < \infty$.

## Analysis of $T_3$

Using the double robustness property of the estimating equation, we have

$$T_3 = \frac{1}{\sqrt{L}} \sum_{\ell=1}^{L} \sqrt{m} \Big\{ \mathrm{P}[\phi(V; \hat{q}_\ell, \hat{h}_\ell)] - \psi_0 \Big\}$$

$$= \frac{1}{\sqrt{L}} \sum_{\ell=1}^{L} \sqrt{m} \Big\{ \mathrm{P}[\phi(V; \hat{q}_\ell, \hat{h}_\ell)] - \mathrm{P}[\phi(V; q_0, \hat{h}_\ell)] \Big\}$$

$$= \sqrt{n} \Big\{ \mathrm{P}[\phi(V; \hat{q}_\ell, \hat{h}_\ell)] - \mathrm{P}[\phi(V; q_0, \hat{h}_\ell)] \Big\}$$

(14)

Therefore (dropping the subscript $\ell$ for readability), we have

$$T_3 = \sqrt{n}\mathrm{P}\Big[ \hat{q}(V_q)\hat{h}(V_h)g_1(V) + \hat{q}(V_q)g_2(V) + \hat{h}(V_h)g_3(V) + g_4(V)$$

$$- q_0(V_q)\hat{h}(V_h)g_1(V) - q_0(V_q)g_2(V) - \hat{h}(V_h)g_3(V) - g_4(V) \Big]$$

$$= \sqrt{n}\mathrm{P}\Big[ (\hat{h}(V_h)g_1(V) + g_2(V))(\hat{q}(V_q) - q_0(V_q)) \Big]$$

$$= \sqrt{n}\mathrm{P}\Big[ \mathrm{P}[\hat{h}(V_h)g_1(V) + g_2(V)|V_q](\hat{q}(V_q) - q_0(V_q)) \Big]$$

As seen in the proof of Proposition 1, we have $\mathbb{E}\big[h_0(V_h)g_1(V) + g_2(V)|V_q\big] = 0$. Therefore, for all functions $h$, we have

$$\mathbb{E}\big[h(V_h)g_1(V) + g_2(V)|V_q\big] = \mathbb{E}\big[g_1(V)\{h_0(V_h) - h(V_h)\}|V_q\big].$$

Applying this equality to the last expression for $T_3$ and using Cauchy–Schwarz inequality, we have

$$T_3 = \sqrt{n}\mathrm{P}\Big[ \mathrm{P}\big[g_1(V)\{\hat{h}(V_h) - h_0(V_h)\}|V_q\big](\hat{q}(V_q) - q_0(V_q)) \Big]$$

$$\leq \sqrt{n}\Big\| \mathrm{P}\big[g_1(V)\{\hat{h}(V_h) - h_0(V_h)\}|V_q\big] \Big\|_2 \times \big\| \hat{q}(V_q) - q_0(V_q) \big\|_2.$$

Note that by Assumption 2,

$$\Big\| \mathrm{P}\big[g_1(V)\{\hat{q}(V_q) - q_0(V_q)\}|V_h\big] \Big\|_2$$

$$= \Big\| \mathrm{P}\big[ \mathrm{P}[g_1(V)|V_q, V_h]\{\hat{q}(V_q) - q_0(V_q)\}|V_h\big] \Big\|_2$$

$$\geq \min_{v_q, v_h} \big| \mathrm{P}[g_1(V)|V_q = v_q, V_h = v_h] \big| \Big\| \mathrm{P}\big[\hat{q}(V_q) - q_0(V_q)|V_h\big] \Big\|_2$$

$$\geq \sigma_1 \Big\| \mathrm{P}\big[\hat{q}(V_q) - q_0(V_q)|V_h\big] \Big\|_2.$$

Therefore, there exists constant $C_1$ such that

$$\Big\| \mathrm{P}\big[\hat{q}(V_q) - q_0(V_q)|V_h\big] \Big\|_2 \leq C_1 r_q(n).$$

This implies that

$$\big\| \hat{q}(V_q) - q_0(V_q) \big\|_2 \leq C_2 \tau_q\big(r_q(n)\big),$$

for some constant $C_2$. Therefore,

$$T_3 \leq \sqrt{n}\Big\| \mathrm{P}\big[g_1(V)\{\hat{h}(V_h) - h_0(V_h)\}|V_q\big] \Big\|_2 \times \big\| \hat{q}(V_q) - q_0(V_q) \big\|_2$$

$$= O\Big( \sqrt{n}\ r_h(n)\tau_q\big(r_q(n)\big) \Big).$$

We can also replace $\psi_0$ in equation (14) with $\mathrm{P}[\phi(V; \hat{q}_\ell, h_0)]$. Hence, we also have

$$T_3 = O\Big( \sqrt{n}\ r_q(n)\tau_h\big(r_h(n)\big) \Big),$$

which implies that

$$T_3 = O\Big( \sqrt{n}\ \min\Big\{ r_h(n)\tau_q\big(r_q(n)\big), r_q(n)\tau_h\big(r_h(n)\big) \Big\} \Big) = o(1),$$

where the last equality is due to Assumption 3.

$\square$

*Proof of proposition 3.* By Theorem 1, for arm $A = a$, for function $\tilde{h}_a$, we have

$$\mathbb{E}[-I(A = a)\tilde{h}_a(W, X) + I(A = a)Y|Z, X] = 0$$
$$\Leftrightarrow \mathbb{E}[\tilde{h}_a(W, X) - Y|Z, A = a, X] = 0$$
$$\Leftrightarrow \mathbb{E}[\tilde{h}(W, a, X) - Y|Z, A = a, X] = 0,$$

where the function $\tilde{h}$ is defined in the statement of the proposition.

By Theorem 1, for arm $A = a$, for function $\tilde{q}_a$, we have

$$\mathbb{E}[-I(A = a)\tilde{q}_a(Z, X) + 1|W, X] = 0$$
$$\Leftrightarrow \mathbb{E}[\mathbb{E}[I(A = a)\tilde{q}_a(Z, X)|W, A, X]|W, X] = 1$$
$$\Leftrightarrow \mathbb{E}[I(A = a)|W, X]\mathbb{E}[\tilde{q}_a(Z, X)|W, A = a, X] = 1$$
$$\Leftrightarrow \mathbb{E}[\tilde{q}_a(Z, X)|W, A = a, X] = \frac{1}{P(A = a|W, X)}$$
$$\Leftrightarrow \mathbb{E}[\tilde{q}(Z, a, X)|W, A = a, X] = \frac{1}{P(A = a|W, X)},$$

where the function $\tilde{q}$ is defined in the statement of the proposition.

$\square$

*Proof of Proposition 4.* We first show that for any function $h$,

$$\sup_{q \in \mathcal{Q}} \hat{\mathbb{E}}\Big[q(V_q)[h(V_h)g_1(V) + g_2(V)] - q^2(V_q)\Big] - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2 \tag{15}$$
$$= \frac{1}{4}\{\xi_n(h)\}^\top K_{\mathcal{Q},n}(\frac{1}{n}K_{\mathcal{Q},n} + \lambda_{\mathcal{Q}}I_n)^{-1}\{\xi_n(h)\},$$

where $\xi_n(h) = \frac{1}{n}(h(V_{h_i})g_1(V_i) + g_2(V_i))_{i=1}^n$.

To see this, we note that by the generalized representer theorem (Schölkopf et al., 2001), the solution to this maximization will be of the form

$$q(v_q) = \sum_{j=1}^n \alpha_j K_{\mathcal{Q}}(V_{q_j}, v_q).$$

Hence, only the coefficients $\alpha = (\alpha_j)_{j=1}^n$ are needed to be calculated.

Note that we have

$$\hat{\mathbb{E}}[q(V_q)[h(V_h)g_1(V) + g_2(V)]] = \sum_{i=1}^n q(V_{q_i})\{\xi_n(h)\}_i = \alpha^\top K_{\mathcal{Q},n}\{\xi_n(h)\},$$

$$\hat{\mathbb{E}}[q^2(V_q)] = \frac{1}{n}\alpha^\top K_{\mathcal{Q},n}^2\alpha,$$

$$\|q\|_{\mathcal{Q}}^2 = \alpha^\top K_{\mathcal{Q},n}\alpha.$$

Therefore, we have

$$\hat{\mathbb{E}}\Big[q(V_q)[h(V_h)g_1(V) + g_3(V)] - q^2(V_q)\Big] - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2 = \alpha^\top K_{\mathcal{Q},n}\{\xi_n(h)\} - \alpha^\top(\frac{1}{n}K_{\mathcal{Q},n}^2 + \lambda_{\mathcal{Q}}K_{\mathcal{Q},n})\alpha.$$

Taking the derivative and setting it to zero, the optimal coefficients can be obtained as

$$\alpha^* = \frac{1}{2}(\frac{1}{n}K_{\mathcal{Q},n} + \lambda_{\mathcal{Q}}I_n)^{-1}\{\xi_n(h)\},$$

where $I_n$ is the identity matrix of size $n$.

Consequently we have

$$\sup_{q \in \mathcal{Q}} \hat{\mathbb{E}}\Big[q(V_q)[h(V_h)g_1(V) + g_2(V)] - q^2(V_q)\Big] - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2$$

$$= \frac{1}{4}\{\xi_n(h)\}^\top K_{\mathcal{Q},n}(\frac{1}{n}K_{\mathcal{Q},n} + \lambda_{\mathcal{Q}}I_n)^{-1}\{\xi_n(h)\}.$$

Therefore, the outer minimization problem in (12) is reduced to the following:

$$\hat{h} = \arg\min_{h \in \mathcal{H}}\{\xi_n(h)\}^\top \Gamma\{\xi_n(h)\} + \lambda_{\mathcal{H}}\|h\|_{\mathcal{H}}^2,$$

where $\Gamma = \frac{1}{4}K_{\mathcal{Q},n}(\frac{1}{n}K_{\mathcal{Q},n} + \lambda_{\mathcal{Q}}I_n)^{-1}$.

We note that by the generalized representer theorem (Schölkopf et al., 2001), the solution to this minimization problem will be of the form

$$h(v_h) = \sum_{j=1}^{n} \alpha_j K_{\mathcal{H}}(V_{h_j}, v_h).$$

Hence, only the coefficients $\alpha = (\alpha_j)_{j=1}^{n}$ are needed to be calculated.

Recall that $\xi_n(h) = \frac{1}{n}(h(V_{h_i})g_1(V_i) + g_2(V_i))_{i=1}^{n}$; that is,

$$\xi_n(h) = \frac{1}{n}(diag(g_{1,n})K_{\mathcal{H},n}\alpha + g_{2,n}),$$

where $diag(g_{1,n})$ is a diagonal matrix with $g_1(V_i)$ as the $i$-th diagonal entry, and $g_{2,n} := [g_2(V_i) \cdots g_2(V_i)]^\top$. Also note that

$$\|h\|_{\mathcal{H}}^2 = \alpha^\top K_{\mathcal{H},n}\alpha.$$

Therefore, we have

$$\min_{h \in \mathcal{H}}\{\xi_n(h)\}^\top \Gamma\{\xi_n(h)\} + \lambda_{\mathcal{H}}\|h\|_{\mathcal{H}}^2$$

$$= \min_{\alpha \in \mathbb{R}^n} \frac{1}{n^2}\alpha^\top K_{\mathcal{H},n}diag(g_{1,n})\Gamma diag(g_{1,n})K_{\mathcal{H},n}\alpha + \frac{2}{n^2}g_{2,n}^\top \Gamma diag(g_{1,n})K_{\mathcal{H},n}\alpha + \lambda_{\mathcal{H}}\alpha^\top K_{\mathcal{H},n}\alpha + c,$$

which is solved by

$$\alpha^* = -\Big(K_{\mathcal{H},n}diag(g_{1,n})\Gamma diag(g_{1,n})K_{\mathcal{H},n} + n^2\lambda_{\mathcal{H}}K_{\mathcal{H},n}\Big)^\dagger K_{\mathcal{H},n}diag(g_{1,n})\Gamma g_{2,n}.$$

$\square$

*Proof of Theorem 5.* The proof is for the most part the same as the proof of Theorem 1 in (Dikkala et al., 2020). For a function $h$, define its empirical $L_2$ norm as $\|h\|_{2,n} := \sqrt{\mathbb{P}_n[h^2]}$. Define

$$\Psi(h, q) := \mathbb{E}\big[\{g_1(V)h(V_h) + g_2(V)\}q(V_q)\big] = \mathbb{E}\big[g_1(V)\{h(V_h) - h_0(V_h)\}q(V_q)\big],$$
$$\Psi_n(h, q) := \hat{\mathbb{E}}\big[\{g_1(V)h(V_h) + g_2(V)\}q(V_q)\big],$$

and

$$\Psi^\lambda(h, q) := \Psi(h, q) - \frac{1}{2}\|q\|_2^2 - \frac{\lambda_{\mathcal{Q}}}{2}\|q\|_{\mathcal{Q}}^2,$$
$$\Psi_n^\lambda(h, q) := \Psi_n(h, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2.$$

Note that the minimax optimization (12) can be written as

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \sup_{q \in \mathcal{Q}} \hat{\mathbb{E}}\Big[q(V_q)[h(V_h)g_1(V) + g_2(V)] - q^2(V_q)\Big] - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2 + \lambda_{\mathcal{H}}\|h\|_{\mathcal{H}}^2$$

$$= \arg\min_{h \in \mathcal{H}} \sup_{q \in \mathcal{Q}} \Psi_n^\lambda(h, q) + \lambda_{\mathcal{H}}\|h\|_{\mathcal{H}}^2.$$

The proof proceeds in 3 steps:

**Step 1.** We will show that with probability $1 - 2\zeta$,

$$\sup_{q \in \mathcal{Q}} \left\{ \Psi_n(\hat{h}, q) - \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_\mathcal{Q} \|q\|_\mathcal{Q}^2 \right\} \leq \lambda_\mathcal{H} \left( \|h_0\|_\mathcal{H}^2 - \|\hat{h}\|_\mathcal{H}^2 \right) + O(\delta^2).$$

**Step 2.** For any function $h$, define $q_h := \mathrm{P}\left[ g_1(V)\{h(V_h) - h_0(V_h)\} \big| V_q \right]$. We will show that if $\|q_{\hat{h}}\|_2 \geq \delta$, then with probability $1 - \zeta$,

$$\sup_{q \in \mathcal{Q}} \left\{ \Psi_n(\hat{h}, q) - \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_\mathcal{Q} \|q\|_\mathcal{Q}^2 \right\} \geq \frac{\delta}{2} R(\hat{h}) - C\delta^2 - \lambda_\mathcal{H} \left( \|h_0\|_\mathcal{H}^2 + \|\hat{h}\|_\mathcal{H}^2 \right).$$

**Step 3.** Finally, using Steps 1 and 2, if $\|q_{\hat{h}}\|_2 \geq \delta$, with probability $1 - 3\zeta$, we have

$$\frac{\delta}{2} R(\hat{h}) \leq 2\lambda_\mathcal{H} \|h_0\|_\mathcal{H}^2 + O(\delta^2),$$

which concludes that

$$R(\hat{h}) = O\left( \delta + \frac{\lambda_\mathcal{H}}{\delta} \|h_0\|_\mathcal{H}^2 \right). \tag{16}$$

Note that $\|q_{\hat{h}}\|_2 = R(\hat{h})$. Hence, $\|q_{\hat{h}}\|_2 \leq \delta$ implies $R(\hat{h}) \leq \delta$. Therefore, (16) in either case holds.

We next prove the inequalities in Steps 1 and 2. The main ingredients in the proofs are the following two results

**Theorem 6** (Wainwright (2019), Theorem 14.1). *Let $\mathcal{F}$ be a star-convex, 1-uniformly bounded function class with critical radius $\delta_n$. Then for any $t \geq \delta_n$, we have*

$$\left| \|f\|_{2,n}^2 - \|f\|_2^2 \right| \leq \frac{1}{2} \|f\|_2^2 + \frac{t^2}{2}, \quad \textit{for all } f \in \mathcal{F},$$

*with probability at least $1 - c_1 e^{-c_2 n t^2}$.*

Theorem 6 implies that for the choice of $t$ equal to $\delta = \delta_n + c_0 \sqrt{\frac{\log(c_1/\zeta)}{n}}$, where $\delta_n$ is an upper bound on the critical radius of $\mathcal{Q}_U$, for some constants $c_0$ and $c_1$, we have

$$\left| \|q\|_{2,n}^2 - \|q\|_2^2 \right| \leq \frac{1}{2} \|q\|_2^2 + \frac{1}{2} \delta^2, \quad \text{for all } q \in \mathcal{Q}_U,$$

with probability at least $1 - c_1 e^{-c_2 n \delta^2}$, and hence, with probability larger than $1 - \zeta$.

Moreover, for any function $q \in \mathcal{Q}$ such that $\|q\|_\mathcal{Q}^2 \geq U$, since $\mathcal{Q}$ is assumed to be star-convex, the function $\frac{q\sqrt{U}}{\|q\|_\mathcal{Q}}$ is in $\mathcal{Q}_U$. Therefore, again by Theorem 6, we have

$$\left| \|q\|_{2,n}^2 - \|q\|_2^2 \right| \leq \frac{1}{2} \|q\|_2^2 + \frac{1}{2} \delta^2 \frac{\|q\|_\mathcal{Q}^2}{U},$$

with probability larger than $1 - \zeta$. The last two displays imply that with probability larger than $1 - \zeta$,

$$\left| \|q\|_{2,n}^2 - \|q\|_2^2 \right| \leq \frac{1}{2} \|q\|_2^2 + \frac{\delta^2}{2U} \|q\|_\mathcal{Q}^2 + \frac{1}{2} \delta^2, \quad \text{for all } q \in \mathcal{Q}. \tag{17}$$

Therefore, with probability at least $1 - \zeta$,

$$\lambda_\mathcal{Q} \|q\|_\mathcal{Q}^2 + \|q\|_{2,n}^2 \geq \left( \lambda_\mathcal{Q} - \frac{\delta^2}{2U} \right) \|q\|_\mathcal{Q}^2 + \frac{1}{2} \|q\|_2^2 - \frac{\delta^2}{2}.$$

Therefore, assuming that $\lambda_{\mathcal{Q}} \geq \frac{\delta^2}{U}$,

$$\lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2 + \|q\|_{2,n}^2 \geq \frac{\lambda_{\mathcal{Q}}}{2}\|q\|_{\mathcal{Q}}^2 + \frac{1}{2}\|q\|_2^2 - \frac{\delta^2}{2}. \qquad (18)$$

The second key result used in the proof is an extension of Theorem 6, proposed in (Foster and Syrgkanis, 2019). Since $g_1$ and $g_2$ are bounded and $\mathcal{H}$ is a class of bounded functions, the function $\{g_1(V)h(V_h) + g_2(V)\}q(V_q)$ is Lipschitz with respect to $q$. Therefore, by Lemma 11 in (Foster and Syrgkanis, 2019), with probability $1 - \zeta$,

$$\left|\Psi_n(h_0, q) - \Psi(h_0, q)\right| \leq O(\delta\|q\|_2 + \delta^2), \quad \text{for all } q \in \mathcal{Q}_U.$$

Moreover, for any function $q \in \mathcal{Q}$ such that $\|q\|_{\mathcal{Q}}^2 \geq U$, since $\mathcal{Q}$ is assumed to be star-convex, the function $\frac{q\sqrt{U}}{\|q\|_{\mathcal{Q}}}$ is in $\mathcal{Q}_U$. Therefore,

$$\left|\Psi_n(h_0, q) - \Psi(h_0, q)\right| \leq O(\delta\|q\|_2 + \frac{\delta^2}{\sqrt{U}}\|q\|_{\mathcal{Q}}),$$

with probability larger than $1 - \zeta$. The last two displays imply that with probability larger than $1 - \zeta$,

$$\left|\Psi_n(h_0, q) - \Psi(h_0, q)\right| \leq O(\delta\|q\|_2 + \delta^2 + \frac{\delta^2}{\sqrt{U}}\|q\|_{\mathcal{Q}}), \quad \text{for all } q \in \mathcal{Q}. \qquad (19)$$

For any function $h$, define $q_h \coloneqq P[g_1(V)\{h(V_h) - h_0(V_h)\}|V_q]$, which is a function in $\mathcal{Q}$. Again, by Lemma 11 in (Foster and Syrgkanis, 2019), since $\delta_n$ is an upper bound on the critical radius of $\mathcal{G}_B$, for all functions $h$ such that $h - h_0 \in \mathcal{H}_B$, with probability $1 - \zeta$,

$$\left|\left(\Psi_n(h, q_h) - \Psi_n(h_0, q_h)\right) - \left(\Psi(h, q_h) - \Psi(h_0, q_h)\right)\right| \leq O(\delta\|g_1(V)\{h(V_h) - h_0(V_h)\}q_h(V_q)\|_2 + \delta^2)$$
$$= O(\delta\|q_h(V_q)\|_2 + \delta^2).$$

Moreover, for any function $h \in \mathcal{H}$ such that $\|h - h_0\|_{\mathcal{H}}^2 \geq B$, since $\mathcal{H}$ is assumed to be star-convex, the function $\frac{(h-h_0)\sqrt{B}}{\|h-h_0\|_{\mathcal{H}}}$ is in $\mathcal{H}_B$. Therefore,

$$\left|\left(\Psi_n(h, q_h) - \Psi_n(h_0, q_h)\right) - \left(\Psi(h, q_h) - \Psi(h_0, q_h)\right)\right| \leq O(\delta\|q_h(V_q)\|_2\frac{\|h-h_0\|_{\mathcal{H}}^2}{B} + \delta^2\frac{\|h-h_0\|_{\mathcal{H}}^2}{B}),$$

with probability larger than $1 - \zeta$. The last two displays imply that with probability larger than $1 - \zeta$,

$$\left|\left(\Psi_n(h, q_h) - \Psi_n(h_0, q_h)\right) - \left(\Psi(h, q_h) - \Psi(h_0, q_h)\right)\right|$$
$$\leq O(\delta\|q_h(V_q)\|_2 + \delta^2 + \delta\|q_h(V_q)\|_2\frac{\|h-h_0\|_{\mathcal{H}}^2}{B} + \delta^2\frac{\|h-h_0\|_{\mathcal{H}}^2}{B}), \quad \text{for all } h \in \mathcal{H}. \qquad (20)$$

Equipped with inequalities (17), (18), (19), and (20), we now present the proofs of Steps 1 and 2.

*Proof of Step 1.*

We note that,

$$\sup_{q \in \mathcal{Q}} \Psi_n^\lambda(\hat{h}, q) = \sup_{q \in \mathcal{Q}} \left\{\Psi_n(\hat{h}, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2\right\}$$

$$= \sup_{q \in \mathcal{Q}} \left\{\Psi_n(\hat{h}, q) - \Psi_n(h_0, q) + \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2\right\}$$

$$\geq \sup_{q \in \mathcal{Q}} \left\{\Psi_n(\hat{h}, q) - \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2\right\} - \sup_{q \in \mathcal{Q}} \Psi_n^\lambda(h_0, q).$$

That is,

$$\sup_{q \in \mathcal{Q}} \left\{\Psi_n(\hat{h}, q) - \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2\right\} \leq \sup_{q \in \mathcal{Q}} \Psi_n^\lambda(h_0, q) + \sup_{q \in \mathcal{Q}} \Psi_n^\lambda(\hat{h}, q).$$

Moreover, by definition,

$$\sup_{q \in \mathcal{Q}} \Psi_n^\lambda(\hat{h}, q) + \lambda_{\mathcal{H}} \|\hat{h}\|_{\mathcal{H}}^2 \leq \sup_{q \in \mathcal{Q}} \Psi_n^\lambda(h_0, q) + \lambda_{\mathcal{H}} \|h_0\|_{\mathcal{H}}^2.$$

Hence,

$$\sup_{q \in \mathcal{Q}} \Psi_n^\lambda(\hat{h}, q) \leq \sup_{q \in \mathcal{Q}} \Psi_n^\lambda(h_0, q) + \lambda_{\mathcal{H}}(\|h_0\|_{\mathcal{H}}^2 - \|\hat{h}\|_{\mathcal{H}}^2).$$

Therefore, we have

$$\sup_{q \in \mathcal{Q}} \left\{ \Psi_n(\hat{h}, q) - \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}} \|q\|_{\mathcal{Q}}^2 \right\} \leq 2 \sup_{q \in \mathcal{Q}} \Psi_n^\lambda(h_0, q) + \lambda_{\mathcal{H}}(\|h_0\|_{\mathcal{H}}^2 - \|\hat{h}\|_{\mathcal{H}}^2). \tag{21}$$

By inequalities (18) and (19), with probability $1 - 2\zeta$, for some constant $c$ we have

$$\sup_{q \in \mathcal{Q}} \Psi_n^\lambda(h_0, q) = \sup_{q \in \mathcal{Q}} \left\{ \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}} \|q\|_{\mathcal{Q}}^2 \right\}$$

$$\leq \sup_{q \in \mathcal{Q}} \left\{ \Psi(h_0, q) + c(\delta \|q\|_2 + \delta^2 + \frac{\delta^2}{\sqrt{U}} \|q\|_{\mathcal{Q}}) - \frac{\lambda_{\mathcal{Q}}}{2} \|q\|_{\mathcal{Q}}^2 - \frac{1}{2} \|q\|_2^2 + \frac{\delta^2}{2} \right\} \tag{22}$$

$$= \sup_{q \in \mathcal{Q}} \left\{ c(\delta \|q\|_2 + \delta^2 + \frac{\delta^2}{\sqrt{U}} \|q\|_{\mathcal{Q}}) - \frac{\lambda_{\mathcal{Q}}}{2} \|q\|_{\mathcal{Q}}^2 - \frac{1}{2} \|q\|_2^2 + \frac{\delta^2}{2} \right\},$$

where the last equality is due to the fact that $\Psi(h_0, q) = 0$. We note that since we assumed that $\lambda_{\mathcal{Q}} \geq \frac{\delta^2}{U}$, we have

$$\sup_{q \in \mathcal{Q}} \left\{ c\delta \|q\|_2 - \frac{1}{2} \|q\|_2^2 + c \frac{\delta^2}{\sqrt{U}} \|q\|_{\mathcal{Q}} - \frac{\lambda_{\mathcal{Q}}}{2} \|q\|_{\mathcal{Q}}^2 \right\}$$

$$\leq \sup_{q \in \mathcal{Q}} \left\{ c\delta \|q\|_2 - \frac{1}{2} \|q\|_2^2 \right\} + \sup_{q \in \mathcal{Q}} \left\{ c \frac{\delta^2}{\sqrt{U}} \|q\|_{\mathcal{Q}} - \frac{\delta^2}{2U} \|q\|_{\mathcal{Q}}^2 \right\} \tag{23}$$

$$\leq \frac{c^2 \delta^2}{2} + \frac{c^2 \delta^2}{2} = O(\delta^2).$$

From inequalities (22), (21), and (23) we conclude that

$$\sup_{q \in \mathcal{Q}} \left\{ \Psi_n(\hat{h}, q) - \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}} \|q\|_{\mathcal{Q}}^2 \right\} \leq \lambda_{\mathcal{H}}(\|h_0\|_{\mathcal{H}}^2 - \|\hat{h}\|_{\mathcal{H}}^2) + O(\delta^2).$$

*Proof of Step 2.*

Recall that for any function $h$, $q_h := \mathrm{P}\big[g_1(V)\{h(V_h) - h_0(V_h)\}\big|V_q\big]$, which is a function in $\mathcal{Q}$. Suppose $\|q_{\hat{h}}\|_2 \geq \delta$ and $\alpha = \frac{\delta}{2\|q_{\hat{h}}\|_2} \in [0, \frac{1}{2}]$. Since $\mathcal{Q}$ is star-convex and $q_{\hat{h}} \in \mathcal{Q}$, then also $\alpha q_{\hat{h}} \in \mathcal{Q}$. Therefore,

$$\sup_{q \in \mathcal{Q}} \left\{ \Psi_n(\hat{h}, q) - \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}} \|q\|_{\mathcal{Q}}^2 \right\}$$

$$\geq \alpha \Big( \Psi_n(\hat{h}, q_{\hat{h}}) - \Psi_n(h_0, q_{\hat{h}}) \Big) - \alpha^2 \Big( \|q_{\hat{h}}\|_{2,n}^2 + \lambda_{\mathcal{Q}} \|q_{\hat{h}}\|_{\mathcal{Q}}^2 \Big). \tag{24}$$

By inequality (20), with probability $1 - \zeta$, we have

$$\alpha \Big( \Psi_n(\hat{h}, q_{\hat{h}}) - \Psi_n(h_0, q_{\hat{h}}) \Big)$$

$$\geq \alpha \Big( \Psi(\hat{h}, q_{\hat{h}}) - \Psi(h_0, q_{\hat{h}}) \Big) - \alpha O(\delta \|q_{\hat{h}}(V_q)\|_2 + \delta^2 + \delta \|q_{\hat{h}}(V_q)\|_2 \frac{\|\hat{h} - h_0\|_{\mathcal{H}}^2}{B} + \delta^2 \frac{\|\hat{h} - h_0\|_{\mathcal{H}}^2}{B}).$$

Hence, recalling that $\alpha = \frac{\delta}{2\|q_{\hat{h}}\|_2} \in [0, \frac{1}{2}]$, and since $\Psi(h_0, q_{\hat{h}}) = 0$, there exists constant $c$, such that

$$\alpha \Big( \Psi_n(\hat{h}, q_{\hat{h}}) - \Psi_n(h_0, q_{\hat{h}}) \Big) \geq \alpha \Psi(\hat{h}, q_{\hat{h}}) - c\delta^2 - c\delta^2 \frac{\|\hat{h} - h_0\|_{\mathcal{H}}^2}{B}.$$

Also, we note that

$$
\begin{aligned}
\alpha\Psi(\hat{h}, q_{\hat{h}}) &= \frac{\delta}{2\|q_{\hat{h}}\|_2} \mathrm{P}\big[g_1(V)\{\hat{h}(V_h) - h_0(V_h)\}q_{\hat{h}}(V_q)\big] \\
&= \frac{\delta}{2\|q_{\hat{h}}\|_2} \mathrm{P}\big[\mathrm{P}[g_1(V)\{\hat{h}(V_h) - h_0(V_h)\}|V_q]^2\big] \\
&= \frac{\delta\|q_{\hat{h}}\|_2^2}{2\|q_{\hat{h}}\|_2} = \frac{\delta}{2}R(\hat{h}).
\end{aligned}
$$

Therefore,

$$
\alpha\Big(\Psi_n(\hat{h}, q_{\hat{h}}) - \Psi_n(h_0, q_{\hat{h}})\Big) \geq \frac{\delta}{2}R(\hat{h}) - c\delta^2 - c\delta^2\frac{\|\hat{h} - h_0\|_{\mathcal{H}}^2}{B}. \tag{25}
$$

By inequality (17), the fact that $\lambda_{\mathcal{Q}} \geq \frac{\delta^2}{U}$, and the fact that $\|q_{\hat{h}}\|_{\mathcal{Q}}^2 \leq L^2\|\hat{h} - h_0\|_{\mathcal{H}}^2$, we have

$$
\begin{aligned}
-\alpha^2\Big(\|q_{\hat{h}}\|_{2,n}^2 + \lambda_{\mathcal{Q}}\|q_{\hat{h}}\|_{\mathcal{Q}}^2\Big) &\geq -\alpha^2\Big(\frac{3}{2}\|q_{\hat{h}}\|_2^2 + \frac{\delta^2}{2U}\|q_{\hat{h}}\|_{\mathcal{Q}}^2 + \frac{1}{2}\delta^2 + \lambda_{\mathcal{Q}}\|q_{\hat{h}}\|_{\mathcal{Q}}^2\Big) \\
&\geq -\alpha^2\frac{3}{2}\|q_{\hat{h}}\|_2^2 - \frac{\delta^2}{2U}\|q_{\hat{h}}\|_{\mathcal{Q}}^2 - \frac{1}{2}\delta^2 - \lambda_{\mathcal{Q}}\|q_{\hat{h}}\|_{\mathcal{Q}}^2 \\
&\geq -\delta^2 - (\frac{\delta^2}{2U} + \lambda_{\mathcal{Q}})\|q_{\hat{h}}\|_{\mathcal{Q}}^2 \\
&\geq -\delta^2 - \frac{3\lambda_{\mathcal{Q}}}{2}\|q_{\hat{h}}\|_{\mathcal{Q}}^2 \\
&\geq -\delta^2 - \frac{3\lambda_{\mathcal{Q}}}{2}L^2\|\hat{h} - h_0\|_{\mathcal{H}}^2.
\end{aligned} \tag{26}
$$

From inequalities (24), (25), and (26) we have

$$
\begin{aligned}
\sup_{q\in\mathcal{Q}} &\Big\{\Psi_n(\hat{h}, q) - \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2\Big\} \\
&\geq \frac{\delta}{2}R(\hat{h}) - c\delta^2 - c\delta^2\frac{\|\hat{h} - h_0\|_{\mathcal{H}}^2}{B} - \delta^2 - \frac{3\lambda_{\mathcal{Q}}}{2}L^2\|\hat{h} - h_0\|_{\mathcal{H}}^2 \\
&\geq \frac{\delta}{2}R(\hat{h}) - c\delta^2 - cU\lambda_{\mathcal{Q}}\frac{\|\hat{h} - h_0\|_{\mathcal{H}}^2}{B} - \delta^2 - \frac{3\lambda_{\mathcal{Q}}}{2}L^2\|\hat{h} - h_0\|_{\mathcal{H}}^2 \\
&= \frac{\delta}{2}R(\hat{h}) - c\delta^2 - \delta^2 - \lambda_{\mathcal{Q}}\Big(\frac{cU}{B} + \frac{3}{2}L^2\Big)\|\hat{h} - h_0\|_{\mathcal{H}}^2 \\
&\geq \frac{\delta}{2}R(\hat{h}) - c\delta^2 - \delta^2 - \lambda_{\mathcal{Q}}\Big(\frac{2cU}{B} + 3L^2\Big)\big(\|h_0\|_{\mathcal{H}}^2 + \|\hat{h}\|_{\mathcal{H}}^2\big) \\
&= \frac{\delta}{2}R(\hat{h}) - c\delta^2 - \delta^2 - \lambda_{\mathcal{Q}}\Big(4cL^2 + 3L^2\Big)\big(\|h_0\|_{\mathcal{H}}^2 + \|\hat{h}\|_{\mathcal{H}}^2\big).
\end{aligned}
$$

Assuming $\lambda_{\mathcal{Q}}\Big(4cL^2 + 3L^2\Big) \leq \lambda_{\mathcal{H}}$, for some constant $C$ we conclude that if $\|q_{\hat{h}}\|_2 \geq \delta$, then with probability $1 - \zeta$,

$$
\sup_{q\in\mathcal{Q}}\Big\{\Psi_n(\hat{h}, q) - \Psi_n(h_0, q) - \|q\|_{2,n}^2 - \lambda_{\mathcal{Q}}\|q\|_{\mathcal{Q}}^2\Big\} \geq \frac{\delta}{2}R(\hat{h}) - C\delta^2 - \lambda_{\mathcal{H}}\big(\|h_0\|_{\mathcal{H}}^2 + \|\hat{h}\|_{\mathcal{H}}^2\big).
$$

$\square$

# E  ADDITIONAL MATERIALS FOR SECTION 6

## E.1  Data Generating Process in the Synthetic-Data Experiments

We first describe the data generating mechanism of the scenario considered in the main text, which is the same as in (Cui et al., 2020). Covariates $X$ are generated from a multivariate normal distribution $N(\Gamma_x, \Sigma_x)$. We then generate $A$ conditional on $X$ from a Bernoulli distribution.

Next, we generate $Z, W, U$ from the following multivariate normal distribution,

$$(Z, W, U) \,|A, X \sim MVN \left( \left( \begin{array}{c} \alpha_0 + \alpha_a A + \alpha_x X \\ \mu_0 + \mu_a A + \mu_x X \\ \kappa_0 + \kappa_a A + \kappa_x X \end{array} \right), \Sigma = \left( \begin{array}{ccc} \sigma_z^2 & \sigma_{zw} & \sigma_{zu} \\ \sigma_{zw} & \sigma_w^2 & \sigma_{wu} \\ \sigma_{zu} & \sigma_{wu} & \sigma_u^2 \end{array} \right) \right).$$

Finally, $Y$ is generated from

$$
\begin{aligned}
\mathbb{E}(Y|W, U, A, Z, X) &= \mathbb{E}(Y|U, A, Z, X) + \omega \{W - \mathbb{E}(W|U, A, Z, X)\} \\
&= \mathbb{E}(Y|U, A, X) + \omega \{W - \mathbb{E}(W|U, X)\} \\
&= b_0 + b_a A + b_x X + b_w \mathbb{E}(W|U, X) + \omega \{W - \mathbb{E}(W|U, X)\} \\
&= b_0 + b_a A + b_x X + (b_w - \omega)\mathbb{E}(W|U, X) + \omega W,
\end{aligned}
$$

where

$$\mathbb{E}(W|U, X) = \mathbb{E}(W|U, A, Z, X) = \mu_0 + \mu_x X + \frac{\sigma_{wu}}{\sigma_u^2}(U - \kappa_0 - \kappa_x X).$$

The parameters are set as follows:

- $\Gamma_x = (0.25, 0.25)^T$, $\Sigma_x = \left( \begin{array}{cc} \sigma_x^2 & 0 \\ 0 & \sigma_x^2 \end{array} \right)$, $\sigma_x = 0.25$.

- $\Pr(A = 1|X) = \left[ 1 + \exp\{(0.125, 0.125)^T X\} \right]^{-1}$.

- $\alpha_0 = 0.25$, $\alpha_a = 0.125$, $\alpha_x = (0.25, 0.25)^T$.

- $\mu_0 = 0.25$, $\mu_a = 0.25$, $\mu_x = (0.25, 0.25)^T$.

- $\kappa_0 = 0.25$, $\kappa_a = 0.25$, $\kappa_x = (0.25, 0.25)^T$.

- $\Sigma = \left( \begin{array}{ccc} 1 & 0.25 & 0.5 \\ 0.25 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{array} \right)$.

- $b_0 = 2$, $b_a = 2$, $b_x = (0.25, 0.25)^T$, $b_w = 4$, $\omega = 2$.

The validity of relevant assumptions imposed in Section 4 was proved in (Cui et al., 2020). This scenario sets dimensions of $X$, $Z$, $W$ to 2, 1, 1. Hyper parameters $\lambda_Q^q$, $\lambda_Q^h$, $\lambda_H^q$, and $\lambda_H^h$, as well as the kernel bandwidth are chosen by cross validation with minimizing the projected risk.

### E.2 Multidimensional Proxies

We further consider a more complicated scenario where the dimensions of $X$, $Z$, $W$ increase to 5, 2, 2. The data are generated by

$$U \sim \text{Normal}(0, \sigma_U^2) \tag{27}$$

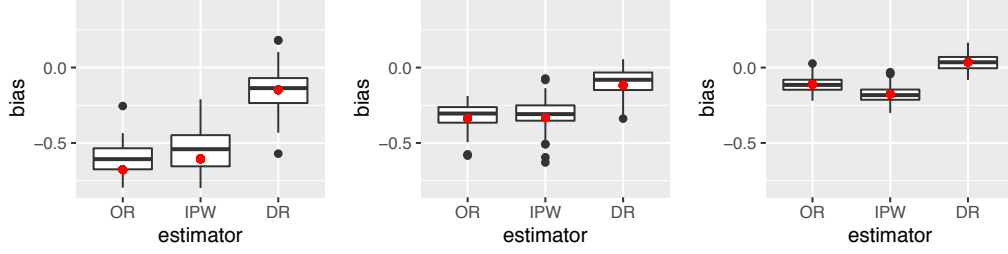$$X|U \sim \text{Normal}(\mu_{UX} U, \sigma_X^2) \tag{28}$$

$$A|X, U \sim \text{Bernoulli}(1/(1 + \exp(t_A + t_X X + t_Z(M_{AZ} + \mu_{XZ} X + \mu_{UZ} U) + 0.5 t_Z \sigma_Z^2 t_Z^\top))) \tag{29}$$

$$Z|A, X, U \sim \text{Normal}(\mu_{AZ} A + \mu_{XZ} X + \mu_{UZ} U, \sigma_Z^2) \tag{30}$$

$$W|X, U \sim \text{Normal}(\mu_{XW} X + \mu_{UW} U, \sigma_W^2) \tag{31}$$

$$Y|A, X, U \sim \text{Normal}(\mu_{AY} A + \mu_{WY} W + \mu_{XY} X + \mu_{UY} U, \sigma_Y^2) \tag{32}$$

The parameters are set as follows

Figure 3: Boxplots of simulation results when $N = 400, 800, 1600$, respectively.

- $\mu_U = 0$, $\sigma_U = 0.3$.

- $\mu_{UX} = \begin{pmatrix} 0.4 \\ 0.525 \\ 0.650 \\ 0.775 \\ 0.9 \end{pmatrix}$, $\sigma_X^2 = \begin{pmatrix} 0.3 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 0.3 \end{pmatrix}$.

- $t_Z = (0.9, 0.4)$, $t_X = (0.9, 0.775, 0.65, 0.525, 0.4)$, $t_A = -t_Z\mu_{AZ} - t_Z\sigma_Z^2 t_Z^\top = -0.981$.

- $\mu_{AZ} = (0.5, 0.6)$, $\mu_{XZ} = \begin{pmatrix} 0.4 & 0.511 & 0.622 & 0.733 & 0.844 \\ 0.456 & 0.567 & 0.678 & 0.789 & 0.9 \end{pmatrix}$, $\mu_{UZ} = (0.8, 0.9)$, $\sigma_Z^2 = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}$.

- $\mu_{XW} = \begin{pmatrix} 0.9 & 0.789 & 0.678 & 0.567 & 0.455 \\ 0.844 & 0.733 & 0.622 & 0.511 & 0.4 \end{pmatrix}$, $\mu_{UW} = (0.8, 0.9)$, $\sigma_W^2 = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}$.

- $\mu_{AY} = 2$, $\mu_{WY} = (0.4, 0.9)$, $\mu_{XY} = (0.4, 0.525, 0.65, 0.775, 0.9)$, $\mu_{UY} = (0.4, 0.9)$, $\sigma_Y^2 = 0.3$

The validity of relevant assumptions imposed in Section 4 follows immediately by the temporal order of data generating process and the underlying independence. The simulation results are presented in Figure 3. Once again, all three estimators attain smaller bias as sample size becomes larger and the PDR estimator beats POR and PIPW estimators in all cases.

### E.3 Nonlinear Confounding Bridge Functions

In this subsection we consider a simulating process such that the confounding bridge functions $h$ and $q$ are nonlinear. In particular, we add cubic terms of $(X, U)$ when generating $A$ and also interaction terms when generating $Y$. The data are generated by

$$U \sim \text{Normal}(0, 1), \tag{33}$$

$$X_1 \sim \text{Normal}(0, 1), \tag{34}$$

$$X_2 \sim \text{Normal}(0, 1), \tag{35}$$

$$A|X, U \sim \text{Bernoulli}(1/(1 + \exp(-0.25 - 0.2X_1 - 0.3X_2 - 0.1X_1^3 - 0.05X_2^3 - 0.25U + 0.1U^3)) \tag{36}$$

$$Z|A, X, U \sim \text{Normal}(0.5 + 0.5A + 0.2X_1 - 0.2X_2 + 0.75U, 1) \tag{37}$$

$$W|X, U \sim \text{Normal}(0.3 + 0.35X_1 + 0.25X_2 - 0.75U, 1) \tag{38}$$

$Y|A, X, U \sim$

$\text{Normal}(-0.5 + A + 0.25X_1 - 0.2X_2 - 0.5AX_1 + 0.3AX_2 - 0.025X_1^3 + 0.03X_2^3 - X - 0.3U + 0.25AU + 0.025U^3, 1)$ (39)

The validity of relevant assumptions imposed in Section 4 follows immediately by the temporal order of data generating process and the underlying independence. The simulation results are presented in Figure 4. Once again, all three estimators attain smaller bias as sample size becomes larger and the PDR estimator beats POR and PIPW estimators in all cases.
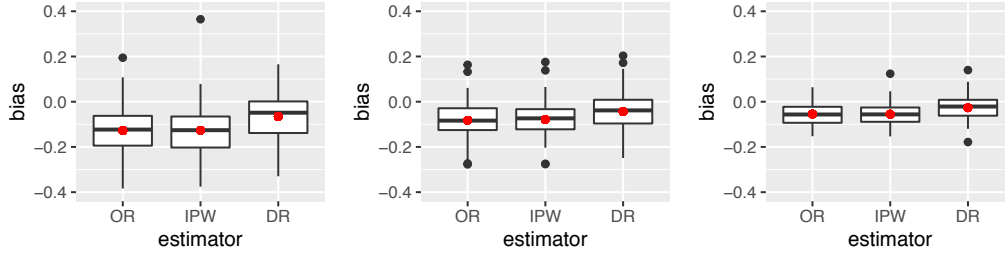
Figure 4: Boxplots of simulation results when $N = 400, 800, 1600$, respectively.

**Real-Data Analysis.** We allocated $Z = (\text{pafi1, paco21})$ and $W = (\text{ph1, hema1})$ as the same choices of treatment- and outcome-inducing confounding proxies in (Tchetgen Tchetgen et al., 2020; Cui et al., 2020). The rest baseline covariates are $X = (\text{age, sex, cat1\_coma, cat2\_coma, dnr1, surv2md1, aps1})$. We chose hyperparameters $\lambda_{\mathcal{Q}}^h = 0.01$, $\lambda_{\mathcal{H}}^q = 0.01$, $\lambda_{\mathcal{H}}^h = 0.001$ and $\lambda_{\mathcal{Q}}^q = 0.001$. We set the kernel bandwidths 35 and 20 in $K_{\mathcal{H},n}$ and $K_{\mathcal{Q},n}$.