Submitted to Statistical Science

A Geometric Perspective on Double Robustness by Semiparametric Theory and Information Geometry

Andrew Ying

Abstract. Double robustness (DR) is a widely-used property of estimators that provides protection against model misspecification and slow convergence of nuisance functions. While DR is a global property on the probability distribution manifold, it often coincides with influence curves, which only ensure orthogonality to nuisance directions locally. This apparent discrepancy raises fundamental questions about the theoretical underpinnings of DR.

In this short communication, we address two key questions: (1) Why do influence curves frequently imply DR "for free"? (2) Under what conditions do DR estimators exist for a given statistical model and parameterization? Using tools from semiparametric theory, we show that convexity is the crucial property that enables influence curves to imply DR. We then derive necessary and sufficient conditions for the existence of DR estimators under a mean squared differentiable path-connected parameterization.

Our main contribution also lies in the novel geometric interpretation of DR using information geometry. By leveraging concepts such as parallel transport, m-flatness, and m-curvature freeness, we characterize DR in terms of invariance along submanifolds. This geometric perspective deepens the understanding of when and why DR estimators exist.

The results not only resolve apparent mysteries surrounding DR but also have practical implications for the construction and analysis of DR estimators. The geometric insights open up new connections and directions for future research. Our findings aim to solidify the theoretical foundations of a fundamental concept and contribute to the broader understanding of robust estimation in statistics.

MSC2020 subject classifications: Primary 62D20; secondary 62M99.
Key words and phrases: Double Robustness, Multiple Robustness, Estimating Function, Semiparametric Theory, Information Geometry.

1. INTRODUCTION

Double robustness (DR) has emerged as a critical property of estimators in various fields, offering protection against model misspecification and slow convergence of nuisance functions. Despite its widespread use, the theoretical underpinnings of DR remain elusive, particularly in understanding the connection between its local and global properties. This raises two fundamental questions: (1) Why do influence curves, which ensure orthogonality to nuisance directions locally, often coincide with DR, a global property? (2) Under what conditions do DR estimators exist for a given statistical model and parameterization?

Addressing these questions is crucial for deepening our understanding of DR and its application in practice. Existing work has explored DR estimators in various contexts (Robins and Rotnitzky, 2001; Robins et al., 2008) but has not fully resolved the apparent discrepancy between local and global properties or provided a complete characterization of the existence of DR estimators.

In this paper, we tackle these questions head-on by leveraging tools from semiparametric theory and information geometry. Our main contributions are two-fold. First, we show that convexity is the key property that enables influence curves to imply DR, resolving the apparent discrepancy between local and global properties. Second, we provide necessary and sufficient conditions for the existence of DR estimators under a mean squared differentiable path-connected parameterization.

To strengthen the geometric interpretation, we introduce novel concepts from information geometry, including parallel transport, m-flatness, and m-curvature freeness. These tools allow us to characterize DR in terms of invariance along submanifolds, providing new insights into when and why DR estimators exist. This geometric perspective deepens our understanding of DR and opens up new connections and directions for future research.

Our findings have both theoretical and practical implications. Understanding the conditions for existence can guide the construction of DR estimators and the choice of parameterizations. The geometric interpretation offers a new lens through which to view DR and its relationship to other concepts in statistics. By solidifying the theoretical foundations of DR, our work contributes to the broader understanding of robust estimation and its application in various fields.

The remainder of the paper is organized as follows. Section 2 introduces the necessary background on DR and semiparametric theory, setting the stage for our main results. Sections 3 and 4 present our findings on the role of convexity and the existence of DR estimators, respectively. Section 5 develops the geometric interpretation using information geometry, providing new insights into the nature of DR. Finally, Section 6 discusses the implications of our work and potential directions for future research.

2. BACKGROUND AND PROBLEM SETUP

Double robustness (DR) is a desirable property of estimators that provides protection against model misspecification and slow convergence of nuisance functions when estimating scientific parameters. Formally, given n i.i.d. samples, an estimator $\hat{\theta}(\gamma_1, \gamma_2)$ of θ that relies on two nuisance functions (γ_1, γ_2) is called doubly robust if $\hat{\theta}(\hat{\gamma}_1, \hat{\gamma}_2)$ is consistent for θ provided that either $\hat{\gamma}_1 - \gamma_1$ or $\hat{\gamma}_2 - \gamma_2$ converges to zero in some metric, where $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are some sample-based functions.

For θ that can be estimated at a root-n rate, DR can be further categorized into "model double robustness" and "rate double robustness" (Chernozhukov et al., 2018; Smucler, Rotnitzky and Robins, 2019; Rotnitzky, Smucler and Robins, 2021). Model double robustness refers to estimators that are asymptotically normal when either of the parametric models for the nuisance functions is correctly specified. Rate double robustness, on the other hand, refers to estimators that are asymptotically normal when the product of the error rates of the nuisance estimators converges faster than root-n, even if they are estimated nonparametrically.

Define a sample space Ω, an event space $\mathscr{F}$, a state space $\mathcal{E}$, and some random mapping X. A random mapping can be a random variable, a random vector, a random process, or a random field. We define $\mathcal{P} = \{\mathbb{P}\}$ as the collection of all possible distributions of X, that is, the model. We use $\mathbb{P}$ and $\mathbb{E}$ as a probability law and its expectation. We are interested in inferring a differentiable parameter $\theta = \theta(\mathbb{P}) : \mathcal{P} \to \mathbb{R}^k$. A parameterization $\gamma(\mathbb{P})$ is a mapping from the probability space $\mathcal{P}$ to a (semi-)metric space Γ. Since in this paper, we focus on double robustness, all parameterizations considered are "two-dimensional" $\gamma(\mathbb{P}) = (\gamma_1(\mathbb{P}), \gamma_2(\mathbb{P}))$. Therefore intuitively a parameterization acts as longitude and latitude over the statistical manifold.

For a pathwise differentiable estimand, every regular and asymptotically linear estimator is equivalent to an estimating function, up to some regularity conditions. To sharp the focus, we hence concentrate on population-level double robustness below. We formally define an adaptive estimating function and its double robustness property as follows:

Definition 1 (Adaptive estimating function). A possibly vector function $D(\theta, \gamma)$ of random mapping X, parameter of interest θ and possibly some nuisance function γ is called an estimating function of θ when it satisfies, for any $\mathbb{P} \in \mathcal{P}$

$$\mathbb{E}\{D(\theta(\mathbb{P}), \gamma(\mathbb{P}))\} = 0,$$

$$\mathbb{E}\{D(\theta', \gamma(\mathbb{P}))\} \neq 0,$$

when $\theta' \neq \theta(\mathbb{P})$ in a neighborhood of $\theta(\mathbb{P})$, and

$$\mathbb{E}\{D(\theta, \gamma(\mathbb{P}'))^2\} < \infty.$$

We call γ a "nuisance function" and $\gamma(\mathbb{P})$ over statistical model a "parameterization", though sometimes we use these two names interchangeably. Formally, we can define a population-level double robustness over an estimating function:

Definition 2 (Doubly robust estimating function). An adaptive estimating function with variation independent tuples $(\theta(\mathbb{P}), \gamma_1(\mathbb{P}), \gamma_2(\mathbb{P}))$ is doubly robust if it remains unbiased provided that either one of the nuisance functions is at the truth, that is,

$$\mathbb{E}[D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}), \gamma_2\}] = \mathbb{E}[D\{\theta(\mathbb{P}), \gamma_1, \gamma_2(\mathbb{P})\}] = 0,$$

for any γ_1 and γ_2.

Note that this is global property along $\mathcal{P}$. The following examples illustrate the concept of DR in various settings:

Example 1 (Partially linear model). Consider the following partially linear model (Härdle, Liang and Gao, 2000)

$$Y = \theta^\top A + \omega(L) + \varepsilon,$$

where $X=(Y,A,L)$, ω is unknown and $\mathbb{E}(Y-\theta^\top A|A,L)=\mathbb{E}(Y-\theta^\top A|L)$. A and L are both exploratory variables whilst A has linear effect and L is nonlinear. The model is semiparametric since it contains both parametric and nonparametric components. With appropriate causal conditions (Robins and Rotnitzky, 2001; Vansteelandt and Joffe, 2014), this model is also known as the structural mean model and β can also be understood causally. The estimating function

$$D\{\theta,\gamma_1,\gamma_2\}=\{d(A,L)-\gamma_1(X)\}\{Y-\theta^\top A-\gamma_2(X)\}$$

is doubly robust, where

$$\gamma_1(\mathbb{P})(X)=\mathbb{E}\{d(A,L)|L\},$$

and

$$\gamma_2(\mathbb{P})(X)=\omega(L).$$

We then proceed to the semiparametric odds ratio model to show how double robustness can be absent under one parameterization but present under a different parameterization.

EXAMPLE 2 (Odds ratio model, canonical parameterization). Suppose we observe $X=(Y,A,L)$, a binary Y, binary A, and some covariates L. Consider the semiparametric conditional odds ratio model:

$$\psi(Y,A,L;\theta)=\frac{f(Y|A,L)f(y_0|a_0,L)}{f(Y|a_0,L)f(y_0|A,L)}.$$

for some baseline point y_0, a_0, where ψ is a known function. We are interested in inferring θ. Suppose the parameterization is given by the canonical density decomposition

$$\gamma_1(\mathbb{P})(X)=f(A|L),$$

and

$$\gamma_2(\mathbb{P})(X)=f(Y|A,L),$$

Robins and Rotnitzky (2001) has shown that there does not exist DR estimating function.

EXAMPLE 3 (Odds ratio model, another parameterization). Continuing Example 2, however, consider a different parameterization by

$$\gamma_1(\mathbb{P})(X)=f(Y|a_0,L),$$

and

$$\gamma_2(\mathbb{P})(X)=f(A|y_0,L),$$

Chen (2007) has shown that all influence curves are doubly robust in this case. See more discussion in Tchetgen Tchetgen, Robins and Rotnitzky (2010) as well.

EXAMPLE 4 (Average treatment effect). Suppose we observe $X=(Y,A,L)$ where Y is an outcome of interest, A is the treatment, and L are baseline covariates that ensure there is no unmeasured confounding. We are interested in the average treatment mean on one treatment arm

$$\theta=\mathbb{E}\{\mathbb{E}(Y|A=a,L)\}.$$

The estimating function

$$\begin{aligned}&\gamma_2(X)(Y-\theta)\\&-\gamma_2(X)\{\gamma_1(X)-\theta\}\\&+\int\gamma_1(X)d\mathbb{1}(A=a)-\theta,\end{aligned}$$

is doubly robust, where

$$\gamma_1(\mathbb{P})(X)=\mathbb{E}(Y|A,L),$$

and

$$\gamma_2(\mathbb{P})(t,X)=\frac{\mathbb{1}(A=a)}{\mathbb{P}(A|L)}.$$

See Hernán and Robins (2020) for details.

2.1 Semiparametric theory and influence curves

Semiparametric theory (Newey, 1990; Bickel et al., 1993; Van der Vaart, 2000; Bickel and Kwon, 2001; Tsiatis, 2006; Kosorok, 2008) focuses on first order behavior of the estimand geometrically. By informally treating the statistical model $\mathcal{P}$ as a differential manifold and the parameter of interest $\theta(\mathbb{P})$ as a differentiable mapping to some Euclidean space, the semiparametric theory operates by computing the tangent space $T_{\mathbb{P}}(\mathcal{P})$ at each point $\mathbb{P}$. For any point $\mathbb{P}\in\mathcal{P}$, this is computed by exhausting all possible regular parametric submodels, that is, parametric submodels containing $\mathbb{P}$. In fact, one may define the tangent space for any submodel $\mathcal{M}\subset\mathcal{P}$ as

$$T_{\mathbb{P}}(\mathcal{M})=\text{clspan}\left(\left\{S(X)\in L_0^2(\mathbb{P}):\mathbb{P}_t\in\mathcal{M}\right\}\right).$$

A important concept is convexity, which gives an explicit form of the tangent space:

LEMMA 1. *When a model $\mathcal{M}$ is convex, the associated nuisance tangent space at each law $\mathbb{P}\in\mathcal{M}$ is*

$$T_{\mathbb{P}}(\mathcal{M})=clspan\left\{\frac{d\mathbb{P}'}{d\mathbb{P}}(X)-1:\mathbb{P}'\in\mathcal{M}\right\}.$$

With $\theta(\mathbb{P})$ defined, one may define the nuisance tangent space $\mathcal{N}_{\mathbb{P}}(\mathcal{M})$ at each point $\mathbb{P}$ by

$$\mathcal{N}_{\mathbb{P}}(\mathcal{M})=\text{clspan}\left(\left\{S\in T_{\mathbb{P}}(\mathcal{M}):\left.\frac{\mathrm{d}\theta(\mathbb{P}_t)}{\mathrm{d}t}\right|_{t=0}=0\right\}\right).$$

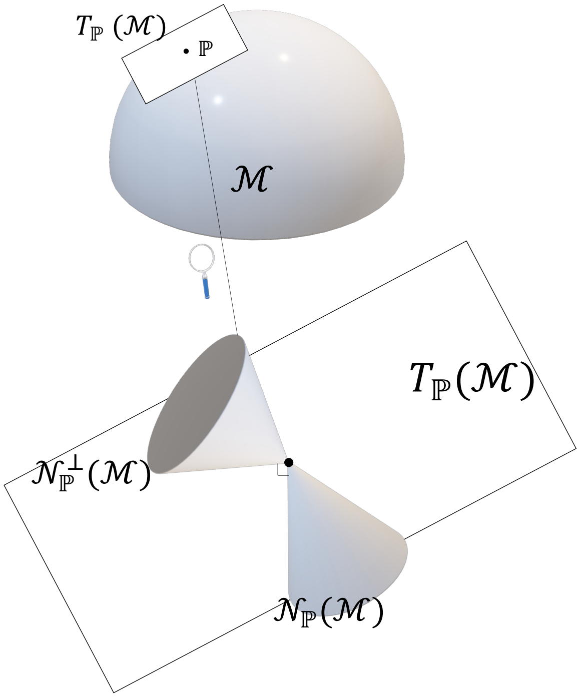
$T_{\mathbb{P}}(\mathcal{M})$
$\mathbb{P}$
$\mathcal{M}$
$T_{\mathbb{P}}(\mathcal{M})$
$\mathcal{N}_{\mathbb{P}}^{\perp}(\mathcal{M})$
$\mathcal{N}_{\mathbb{P}}(\mathcal{M})$

FIG 1. *The tangent space $T_{\mathbb{P}}(\mathcal{M})$, the nuisance tangent space $\mathcal{N}_{\mathbb{P}}(\mathcal{M})$, and the orthogonal complement $\mathcal{N}_{\mathbb{P}}^{\mathbb{P},\perp}(\mathcal{M})$ for a model $\mathcal{M}$.*

The tangent space together with the inner product structure upon it makes the statistical model a Riemann manifold unofficially. The understanding of the tangent space enables one to compute the first-order derivative of a parameter of interest. The set of pathwise derivatives of the parameter can be identified by the orthogonal complement of the nuisance tangent space by Riesz's lemma. The estimating function suggested by the pathwise derivatives is also called the influence curve, which informally is equal to zero first-order Taylor expansions against nuisance functions.

DEFINITION 3 (Influence curve). An influence curve $\mathrm{IC}(\mathbb{P}) \in L_0^2(\mathbb{P})$ is a Riesz representer for the pathwise derivative of the parameter of interest, that is, for any parametric submodel $\{\mathbb{P}_t\}$ and its corresponding score $S(X)$

$$\left.\frac{d\theta(\mathbb{P}_t)}{dt}\right|_{t=0} = \mathbb{E}\{\mathrm{IC}(\mathbb{P})S(X)\}.$$

Any IC lies in the orthogonal complement of the nuisance tangent space, that is $\mathrm{IC} \in \mathcal{N}_{\mathbb{P}}^{\mathbb{P},\perp}(\mathcal{P})$. See Figure 1 for a geometric view of the tangent space $T_{\mathbb{P}}(\mathcal{M})$, the nuisance tangent space $\mathcal{N}_{\mathbb{P}}(\mathcal{M})$, and the orthogonal complement $\mathcal{N}_{\mathbb{P}}^{\mathbb{P},\perp}(\mathcal{M})$ for a model $\mathcal{M}$. The efficient influence curve (EIC) is the unique IC with the minimum L^2 length. As we see in the examples, there are lots of cases when influence curves are indeed doubly robust or at least inspires DR estimating functions.

2.2 Problem formulation

We now formally state the two key questions we aim to address in this paper:

1. Why do influence curves frequently imply DR "for free"?
2. Under what conditions do DR estimators exist for a given statistical model and parameterization?

The first question arises from the observation that influence curves, which ensure orthogonality to nuisance directions locally, often coincide with DR estimators, which have a global robustness property. The second question seeks to characterize the necessary and sufficient conditions for the existence of DR estimators, given a model and parameterization.

In the following sections, we tackle these questions using tools from semiparametric theory and information geometry, providing new insights into the theoretical underpinnings of DR.

3. CONVEXITY AND DOUBLE ROBUSTNESS

In this section, we investigate the role of convexity in enabling influence curves to imply double robustness (DR) without additional conditions. We first introduce the concept of a section, which captures the subset of the model where one nuisance function remains fixed.

DEFINITION 4 (Section). The section of $\gamma_1()$ at $\mathbb{P}$ is defined as

$$\mathcal{M}_1(\mathbb{P}) = \{\mathbb{P}' \in \mathcal{M} : \theta(\mathbb{P}') = \theta(\mathbb{P}), \gamma_2(\mathbb{P}') = \gamma_2(\mathbb{P})\}.$$

Likewise, one can define $\mathcal{M}_2(\mathbb{P})$. Quick observations:

1. $\mathbb{P} \in \mathcal{M}_1(\mathbb{P})$;
2. For any $\mathbb{P}' \in \mathcal{M}_1(\mathbb{P})$ and any doubly robust estimating function $D(\theta, \gamma_1, \gamma_2)$,
$$\begin{aligned}&\mathbb{E}[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}]\\ &= \mathbb{E}[D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P})\}] = 0,\end{aligned}$$
also
$$\begin{aligned}&\mathbb{E}'[D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P})\}]\\ &= \mathbb{E}'[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}] = 0;\end{aligned}$$
3. For any $\mathbb{P}' \in \mathcal{M}_1(\mathbb{P})$, $\mathcal{M}_1(\mathbb{P}') = \mathcal{M}_1(\mathbb{P})$;
4. for any parametric submodel passing through $\mathcal{M}_1(\mathbb{P})$, the corresponding score lies in the nuisance tangent space $\mathcal{N}_{\mathbb{P}}(\mathcal{M})$ because $\theta(\mathbb{P})$ remains unchanged within $\mathcal{M}_1(\mathbb{P})$.

We now state a key proposition that characterizes the orthogonality of a DR estimating function to the nuisance tangent space of a section.

PROPOSITION 1. *For any $\mathbb{P}' \in \mathcal{M}_1(\mathbb{P})$ and any doubly robust estimating function $D(\theta, \gamma_1, \gamma_2)$,*

$$D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}), \gamma_2(\mathbb{P})\} \perp S(X),$$

at law $\mathbb{P}'$, for any $S(X) \in T_{\mathbb{P}'}(\mathcal{M}_1(\mathbb{P}')) = T_{\mathbb{P}'}(\mathcal{M}_1(\mathbb{P}))$. Therefore,

$$D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}), \gamma_2(\mathbb{P})\} \in T^{\mathbb{P},\perp}_{\mathbb{P}'}(\mathcal{M}_1(\mathbb{P})),$$

where $T^{\mathbb{P},\perp}_{\mathbb{P}'}(\mathcal{M}_1(\mathbb{P}))$ is the orthogonal complement of $T_{\mathbb{P}'}(\mathcal{M}_1(\mathbb{P}))$ at law $\mathbb{P}$. Moreover,

$$D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}), \gamma_2(\mathbb{P})\} \in \left\{\bigcap_{\mathbb{P}' \in \mathcal{M}_1(\mathbb{P})} T^{\mathbb{P}',\perp}_{\mathbb{P}'}(\mathcal{M}_1(\mathbb{P}))\right\} \bigcap \left\{\bigcap_{\mathbb{P}' \in \mathcal{M}_2(\mathbb{P})} T^{\mathbb{P}',\perp}_{\mathbb{P}'}(\mathcal{M}_2(\mathbb{P}))\right\}.$$

The proposition can be seen as a generalization of the necessary theorem outlined by Robins and Rotnitzky (2001, Lemma 1). This proposition establishes that a DR estimating function is orthogonal to the nuisance tangent space, not just at the true distribution P, but at any distribution $\mathbb{P}'$ within the section. This orthogonality property holds for both sections $\mathcal{M}_1(\mathbb{P})$ and $\mathcal{M}_1(\mathbb{P})$. We now present our main result, which shows that convexity of the sections is sufficient for an influence curve to be doubly robust.

THEOREM 1. *When the sections $\mathcal{M}_1(\mathbb{P})$ and $\mathcal{M}_2(\mathbb{P})$ are convex, any influence curve $D(\theta, \gamma_1, \gamma_2)$ is doubly robust.*

The proof of this theorem relies on the orthogonality property established in Proposition 1 and the fact that convexity allows us to construct a parametric submodel connecting any two distributions within a section. Along this submodel, the pathwise derivative of the expected estimating function vanishes, implying double robustness.

The geometric intuition behind this result is that convexity ensures that the nuisance tangent space remains unchanged as we move along the section. This invariance, combined with the orthogonality of the influence curve to the nuisance tangent space, enables the global robustness property of double robustness.

Theorem 1 answers to our first question, showing that convexity is a sufficient condition for influence curves to imply double robustness "for free." In the next section, we will explore necessary and sufficient conditions for the existence of DR estimators under more general settings.

4. EXISTENCE OF DOUBLY ROBUST ESTIMATING FUNCTIONS

In this section, we investigate the necessary and sufficient conditions for the existence of doubly robust (DR) estimators under a given parameterization. We introduce the concept of mean squared differentiable path-connectedness, which generalizes the convexity condition from the previous section.

DEFINITION 5 (Mean squared differentiable path-connected model). A model $\mathcal{M}$ is called mean squared differentiable path-connected if, for any pair $\mathbb{P}$, $\mathbb{P}' \in \mathcal{M}$, there exists a curve $\mathbb{P}_t \subset \mathcal{M}$ such that $\mathbb{P}_0 = \mathbb{P}$ and $\mathbb{P}_1 = \mathbb{P}'$ and mean squared differentiable at each point $\mathbb{P}_t$.

DEFINITION 6 (θ-connected parameterization). A parameterization $\gamma(\mathbb{P})$ is called θ-connected if the sections $\mathcal{M}_1(\mathbb{P})$ and $\mathcal{M}_2(\mathbb{P})$ are mean squared differentiable path-connected.

We now state our main result, which provides necessary and sufficient conditions for the existence of DR estimators.

THEOREM 2. *An adaptive estimating function $D(X, \theta, \gamma)$ is doubly robust under a mean squared differentiable path-connected parameterization γ if and only if for any $\mathbb{P} \in \mathcal{M}$,*

$$D(X, \theta, \gamma) \in \left\{\bigcap_{\mathbb{P}' \in \mathcal{M}_1(\mathbb{P})} T^{\mathbb{P}',\perp}_{\mathbb{P}'}(\mathcal{M}_1(\mathbb{P}'))\right\} \bigcap \left\{\bigcap_{\mathbb{P}' \in \mathcal{M}_2(\mathbb{P})} T^{\mathbb{P}',\perp}_{\mathbb{P}'}(\mathcal{M}_2(\mathbb{P}'))\right\}.$$

Theorem 2 provides a complete characterization of the existence of DR estimators under a given parameterization. The necessary and sufficient condition requires the estimating function to be orthogonal to the nuisance tangent spaces of both sections, not just at the true distribution, but at all distributions within the sections.

The geometric intuition behind this result is that the path-connectedness of the sections allows us to “move” between any two distributions within a section while preserving the orthogonality of the estimating function to the nuisance tangent spaces. This global orthogonality property is equivalent to double robustness.

Theorem 2 answers our second question and provides a powerful tool for determining the existence of DR estimators in practice. By checking the orthogonality condition, we can easily verify whether a given estimating function is doubly robust under a specific parameterization.

In the next section, we will explore the geometric aspects of double robustness in more depth, using tools from information geometry to gain further insights into the structure of the statistical model and the properties of DR estimators.

5. GEOMETRIC UNDERSTANDING BY INFORMATION GEOMETRY

Semiparametric theory only concerns the local property of the statistical model manifold and estimand, for instance, it only operationalizes over one tangent space at one point at a time. On the other hand, double robustness is a global property along certain curves over the statistical manifold. Therefore, intuitively we need some tools to operate globally to investigate relation among tangent spaces at different points. In this section, we deepen our geometric understanding of double robustness (DR) by leveraging tools from information geometry.

Information geometry (Amari and Kawanabe, 1997; Amari and Nagaoka, 2000; Amari, 2016; Ay et al., 2017) leverage global differential geometry theory onto the statistical manifold $\mathcal{P}$. For more elaborate introduction on information geometry, see appendix.

We begin by introducing the concepts of e-parallel transport and m-parallel transport, which are two types of parallel transport that are commonly used in information geometry, allowing one to “move” along curves on the statistical manifold.

DEFINITION 7 (E-parallel transport). The e-parallel transports of a vector $D(x)$ from $T_{\mathbb{P}}(\mathcal{M})$ to $T_{\mathbb{P}'}(\mathcal{M})$ is defined as

$$\prod_{\mathbb{P}\to\mathbb{P}'}^{e} D(x) = D(x) - \mathbb{E}'\{D(X)\}.$$

DEFINITION 8 (M-parallel transport). The m-parallel transports of a vector $D(x)$ from $T_{\mathbb{P}}(\mathcal{M})$ to $T_{\mathbb{P}'}(\mathcal{M})$ is defined as

$$\prod_{\mathbb{P}\to\mathbb{P}'}^{m} D(x) = \frac{d\mathbb{P}}{d\mathbb{P}'}(x) D(x).$$

These two parallel transports are dual with respect to the Riemann metric (or inner product), that is,

$$\langle D_1, D_2\rangle_{\mathbb{P}} = \left\langle \prod_{\mathbb{P}\to\mathbb{P}'}^{e} D_1, \prod_{\mathbb{P}\to\mathbb{P}'}^{m} D_2 \right\rangle_{\mathbb{P}'}.$$

These parallel transports are dual with respect to the Fisher information metric and preserve the inner product between vectors. Figure 2 provides a visual illustration of these concepts.

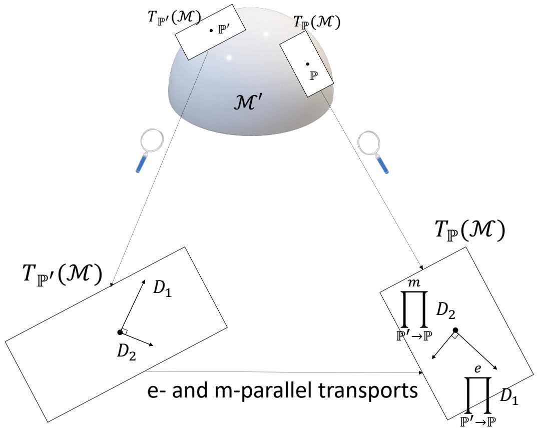
$T_{\mathbb{P}'}(\mathcal{M})$
$\mathbb{P}'$
$T_{\mathbb{P}}(\mathcal{M})$
$\mathbb{P}$
$\mathcal{M}'$
$T_{\mathbb{P}}(\mathcal{M})$
$T_{\mathbb{P}'}(\mathcal{M})$
D_1
D_2
$\prod_{\mathbb{P}'\to\mathbb{P}}^{m}$
D_2
$\prod_{\mathbb{P}'\to\mathbb{P}}^{e}$
D_1
e- and m-parallel transports

FIG 2. *A graphical illustration of e- and m-parallel transports. These two dual transports keep the inner product of a pair of vectors. In particular, as shown in the graph, two perpendicular vectors (D_1, D_2) will remain to be perpendicular after transportation.*

We now establish a key result connecting e-parallel transport and double robustness.

PROPOSITION 2. *An estimating function $D(\theta, \gamma_1, \gamma_2)$ is doubly robust if and only if it remains e-parallel transport invariant along sections $\mathcal{M}_1(\mathbb{P})$ and $\mathcal{M}_2(\mathbb{P})$.*

The proof of this proposition relies on the properties of e-parallel transport and the variation independence of the nuisance functions. Proposition 2 provides a geometric characterization of DR estimators in terms of their invariance under e-parallel transport.

with m-parallel transport, Theorem 2 leads to the following corollary.

COROLLARY 1. *An adaptive estimating function $D(X, \theta, \gamma)$ is doubly robust under a mean squared differentiable path-connected parameterization γ if and only if for any $\mathbb{P} \in \mathcal{M}$,*

$$D(X,\theta,\gamma) \in \left\{ \bigcap_{\mathbb{P}'\in\mathcal{M}_1(\mathbb{P})} \left\{ \prod_{\mathbb{P}'\to\mathbb{P}}^{m} T_{\mathbb{P}'}(\mathcal{M}_1(\mathbb{P})) \right\}^{\mathbb{P},\perp} \right\} \bigcap \left\{ \bigcap_{\mathbb{P}'\in\mathcal{M}_2(\mathbb{P})} \left\{ \prod_{\mathbb{P}'\to\mathbb{P}}^{m} T_{\mathbb{P}'}(\mathcal{M}_2(\mathbb{P})) \right\}^{\mathbb{P},\perp} \right\}.$$

See Figure 3 for a geometric view of the above statements.

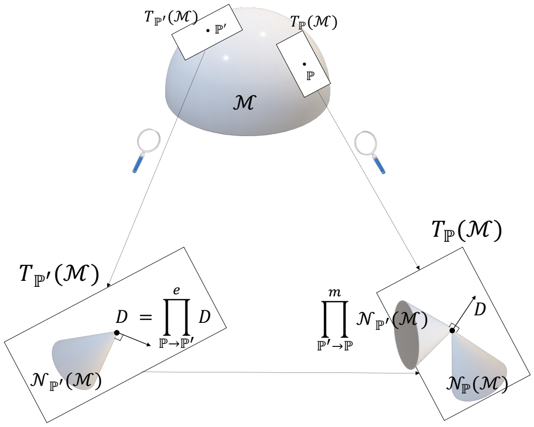
$T_{\mathbb{P}'}(\mathcal{M})$
$\mathbb{P}'$
$T_{\mathbb{P}}(\mathcal{M})$
$\mathbb{P}$
$\mathcal{M}$
$T_{\mathbb{P}}(\mathcal{M})$
$T_{\mathbb{P}'}(\mathcal{M})$
$D = \prod^{e}_{\mathbb{P}\to\mathbb{P}'} D$
$\prod^{m}_{\mathbb{P}'\to\mathbb{P}} \mathcal{N}_{\mathbb{P}'}(\mathcal{M})$
D
$\mathcal{N}_{\mathbb{P}'}(\mathcal{M})$
$\mathcal{N}_{\mathbb{P}}(\mathcal{M})$

FIG 3. *E-parallel transport of an estimating function and m-parallel transport of the nuisance tangent space.*

This corollary provides an alternative characterization of DR estimators in terms of the orthogonality of the estimating function to the m-parallel transported nuisance tangent spaces.

Amari and Kawanabe (1997) investigated the existence of an estimating function (without nuisance function) under a semiparametric model. They introduced the concepts of m-flatness and m-curvature freeness, which capture the behavior of the nuisance tangent spaces under m-parallel transport, to enrich geometric understanding of their topic. Here we generalize these concepts to understand double robustness.

DEFINITION 9 (M-flat parameterization). A model $\mathcal{P}$ with a variation independent parameterization $(\theta, \gamma_1, \gamma_2)$ is called m-flat if

$$\prod^{m}_{\mathbb{P}'\to\mathbb{P}} T_{\mathbb{P}'}(\mathcal{M}_j) \subset T_{\mathbb{P}}(\mathcal{M}_j),$$

for any $\mathbb{P}'$ and $j = 1, 2$.

Intuitively, m-flatness requires that $T_{\mathbb{P}'}(\mathcal{M}_j)$ remains barely changed (within $T_{\mathbb{P}}(\mathcal{M}_j)$) under a m-parallel transport. The convexity of $\mathcal{M}_j$ indeed implies m-flatness.

DEFINITION 10 (M-curvature free parameterization). A model $\mathcal{P}$ with the a variation independent parameterization $(\theta, \gamma_1, \gamma_2)$ is called m-curvature free if

$$\mathrm{EIC} \in \left\{ \prod^{m}_{\mathbb{P}'\to\mathbb{P}} T_{\mathbb{P}'}(\mathcal{M}_j) \right\}^{\mathbb{P},\perp},$$

$\mathbb{P}'$ and $j = 1, 2$.

M-curvature freeness is weaker than m-flatness as $\mathrm{EIC}(\mathbb{P}) \in \mathcal{N}_{\mathbb{P}}(\mathcal{P})^{\mathbb{P},\perp} \subset \{\prod^{m}_{\mathbb{P}'\to\mathbb{P}} \mathcal{N}_{\mathbb{P}'}(\mathcal{P})\}^{\mathbb{P},\perp}$ when m-flatness holds. Intuitively m-flatness says that $\{\prod^{m}_{\mathbb{P}'\to\mathbb{P}} T_{\mathbb{P}'}(\mathcal{M}_j)\}^{\perp}$ contains $T_{\mathbb{P}}(\mathcal{M}_j)$ whilst m-curvature freeness weakens this argument but still requires the most efficient direction EIC is still contained in $\{\prod^{m}_{\mathbb{P}'\to\mathbb{P}} T_{\mathbb{P}'}(\mathcal{M}_j)\}^{\perp}$. We show that m-flatness and m-curvature freeness are sufficient conditions for the existence of DR estimators.

COROLLARY 2. *When a mean squared differentiable path-connected parameterization $(\theta, \gamma_1, \gamma_2)$ is m-flat, any influence functions are doubly robust.*

COROLLARY 3. *When a mean squared differentiable path-connected parameterization $(\theta, \gamma_1, \gamma_2)$ is m-curvature free, the efficient influence curve is a doubly robust estimating function.*

These corollaries highlight the role of the geometry of the nuisance tangent spaces in determining the existence of DR estimators. In particular, they show that certain "nice" geometric properties, such as m-flatness and m-curvature freeness, are sufficient for the existence of DR estimators.

The information geometric perspective developed in this section provides a deeper understanding of the global structure of the statistical model and the properties of DR estimators. By studying the behavior of estimating functions and nuisance tangent spaces under parallel transport, we gain new insights into the geometric nature of double robustness.

6. DISCUSSION AND FUTURE DIRECTIONS

In this paper, we have deepened the geometric understanding of double robustness (DR) by leveraging tools from semiparametric theory and information geometry. Our results provide new insights into the theoretical underpinnings of DR and the conditions for the existence of DR estimators.

We have shown that convexity of the sections of the statistical model is a sufficient condition for influence curves to imply DR "for free," resolving the apparent discrepancy between the local and global robustness properties of DR estimators. We have also provided necessary and sufficient conditions for the existence of DR estimators under a θ-connected parameterization, characterizing the global orthogonality properties of DR estimating functions.

Furthermore, we have introduced novel geometric concepts, such as m-flatness and m-curvature freeness, which capture the behavior of the nuisance tangent spaces under parallel transport. These concepts provide a deeper understanding of the global structure of the statistical model and the geometric properties that enable the existence of DR estimators.

Our findings have both theoretical and practical implications. From a theoretical perspective, our results contribute to the foundational understanding of DR and highlight the importance of geometric considerations in the study of robust estimators. The geometric characterizations we provide can serve as a basis for further investigations into the properties and behavior of DR estimators in different settings.

From a practical perspective, our results can guide the construction of DR estimators and the choice of parameterizations in applied settings. The necessary and sufficient conditions we establish can be used to check the existence of DR estimators for a given statistical model and parameterization, informing the development of robust inference procedures. The geometric insights we provide can also aid in the design of more efficient and stable estimators by taking into account the structure of the statistical model.

One promising direction for future research is to explore the implications of our geometric characterizations for the design of adaptive estimators that can achieve DR without explicit knowledge of the nuisance functions. Developing data-adaptive methods that can exploit the geometric structure of the statistical model to achieve DR could have significant practical impact in settings where the nuisance functions are difficult to estimate or specify.

In conclusion, this paper provides a deepened understanding of double robustness by leveraging tools from semiparametric theory and information geometry. We have resolved the apparent discrepancy between the local properties of influence curves and the global properties of doubly robust estimators, showing that convexity is the key condition that enables this connection. We have also characterized the necessary and sufficient conditions for the existence of doubly robust estimators under a mean squared differentiable path-connected parameterization. Furthermore, our novel geometric perspective, using concepts such as parallel transport, m-flatness, and m-curvature freeness, sheds new light on the structure of the statistical model and the properties of doubly robust estimators. These insights not only advance the theoretical understanding of double robustness but also have practical implications for the construction and analysis of robust estimators in various fields. We hope that our work will inspire further research at the intersection of semiparametric theory, information geometry, and robust estimation, ultimately contributing to the development of more reliable and efficient statistical methods.

ACKNOWLEDGMENTS

The author would like to thank Jiangang Ying, Cong Ding for countless discussions. The author would also like to thank Eric J. Tchetgen Tchetgen for inspiration of this idea. The author would like to acknowledge the helpful discussions and insights gained through interactions with the Claude AI assistant, which contributed to refining the presentation of this work.

APPENDIX A: AN INTRODUCTION TO INFORMATION GEOMETRY

There is a long history of studies on geometry of manifolds of probability distributions. C.R. Rao is believed to have been the first who introduced a Riemannian metric by using the Fisher information matrix (Rao, 1945), which is a monumental work from which information geometry has emerged. Later, Efron (1975) investigated old unpublished calculations by R.A. Fisher and elucidated the results by defining the statistical curvature of a statistical model. This work was commented on by A.P. Dawid in discussions of Efron's paper, where the e- and m-connections were suggested. Following Efron's and Dawid's works, Amari (1982) further developed the differential geometry of statistical models and elucidated its dualistic nature. It was applied to statistical inference to establish a higher-order statistical theory (Amari, 1982, 1985; Kumon and Amari, 1983). Since then, information geometry has become widely known and a number of competent researchers have joined from the fields of statistics, vision, optimization, machine learning, etc. Many international conferences have been organized on this subject.

In differential geometry, affine connection and the corresponding parallel transport are tools to move tangent vector between tangent spaces. Note that there is a unique natural affine connection on a Riemann manifold that is torsion-free and respects the inner product structure, called the Levi-Civita connection (or Riemann connection), which usually is the only connection geometricians are concerned about. However, researchers of information geometry are typically interested in other torsion-free connections that do not respect the inner product structure. We adopt the e-connection and its dual, the m-connection because it turns out that the corresponding e-parallel transport has a natural connection with double robustness. For estimating function related studies, it takes a further step investigating global property by constructing parallel transport from affine connection on fibre bundles. Geometrically, parameterization is simply a way of expressing curves along manifold and doubly robust estimating functions are orthogonal to the nuisance tangent space parallelly transported along two kinds of curves. Indeed, many statistical models commonly considered are flat manifolds, meaning that the nuisance tangent space remains unchanged along parallel transport, which equalizes nuisance tangent space and nuisance fibre space and therefore an influence function becomes doubly (or doubly) robust.

Though general infinite-dimensional statistical models do not admit a formal manifold structure (unless with map into complex Banach spaces), certain concepts can still be carried over from differential geometry. The Hilbert space $L_0^2(\mathbb{P})$ associated at every point $\mathbb{P}$ has informally given the statistical model the Riemann manifold structure. $(\mathcal{P}, L_0^2(\mathcal{P}))$ together is called a fibre bundle. We may compute the tangent space associated with $\mathcal{M}$ called $T_{\mathbb{P}}(\mathcal{M})$ at each point $\mathbb{P}$, $(\mathcal{M}, T(\mathcal{M}))$ is called a tangent bundle. This has given the nuisance tangent sub-bundle $(\mathcal{M}, \mathcal{N})$. $\mathcal{N}(\mathbb{P})^{\perp}$ is the orthogonal complement of the nuisance tangent space. A section of a vector bundle is called a vector field. An adaptive estimating equation is a vector field that has non-trivial directions along θ, that is, $\mathbb{E}\{D(\mathbb{P})|\mathcal{N}^{\perp}(\mathbb{P})\} \neq 0$, for every $\mathbb{P}$.

APPENDIX B: PROOFS

Proof of Lemma 1

For any $\mathbb{P}' \in \mathcal{M}$, by convexity of $\mathcal{M}$, the parametric submodel $\{t\,\mathbb{P}' + (1-t)\,\mathbb{P}\}_{t \in [0,1]} \subset \mathcal{M}$, then

$$\frac{\mathrm{d}\,\mathbb{P}'}{\mathrm{d}\,\mathbb{P}}(X) - 1 = \lim_{t \to 0} \frac{t\,\mathbb{P}' + (1-t)\,\mathbb{P} - \mathbb{P}}{t\,\mathbb{P}}$$

is in $T_{\mathbb{P}}(\mathcal{M})$. Now since $\mathbb{P}' \in \mathcal{M}$ is arbitrary, we have proved the proposition.

Proof of Theorem 1

First, an immediate implication for the influence curve is

$$D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}), \gamma_2(\mathbb{P})\} \in T_{\mathbb{P}}^{\mathbb{P},\perp}(\mathcal{M}_1(\mathbb{P})) \bigcap T_{\mathbb{P}}^{\mathbb{P},\perp}(\mathcal{M}_2(\mathbb{P})).$$

That is,

$$\mathbb{E}\left[D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}), \gamma_2(\mathbb{P})\}\left\{\frac{d\,\mathbb{P}'}{d\,\mathbb{P}}(X) - 1\right\}\right]$$
$$= \mathbb{E}'\left[D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}), \gamma_2(\mathbb{P})\}\left\{\frac{d\,\mathbb{P}}{d\,\mathbb{P}'}(X) - 1\right\}\right] = 0.$$

Therefore,

$$D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}), \gamma_2(\mathbb{P})\} \in T_{\mathbb{P}'}^{\mathbb{P}',\perp}(\mathcal{M}_1(\mathbb{P})).$$

Define $g(t) = \mathbb{E}_t[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}]$, where $\mathbb{P}_t := (1-t)\,\mathbb{P}' + t\,\mathbb{P}$. Therefore

$$\frac{dg(t)}{dt} = \mathbb{E}_t\left[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}S(X)\right] = 0,$$

since

$$S(X) \in T_{\mathbb{P}_t}(\mathcal{M}_1(\mathbb{P})).$$

By ODE theory, we have $g(1) = \mathbb{E}[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}] = \mathbb{E}[D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P})\}] = 0$ and hence doubly robust.

Proof of Theorem 2

"If" part: for any $\mathbb{P}' \in \mathcal{M}_1(\mathbb{P})$, define $g(t) = \mathbb{E}_t[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'),$ where $\mathbb{P}_t := (1-t)\,\mathbb{P}' + t\,\mathbb{P}$. Therefore

$$\frac{dg(t)}{dt} = \mathbb{E}_t\left[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}S(X)\right] = 0,$$

by the condition and the fact that

$$S(X) \in T_{\mathbb{P}_t}(\mathcal{M}_1(\mathbb{P})),$$

By ODE theory, we have $g(1) = \mathbb{E}[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}] = \mathbb{E}[D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P})\}] = 0$ and hence doubly robust.

"Only if" part follows from Proposition 1.

Proof of Proposition 2

We only need to show for γ_1. γ_2 is symmetric. "If" part: for any γ_1, because of variation independence, we can find $\mathbb{P}' \in \mathcal{M}_1(\mathbb{P})$ such that $\gamma_1(\mathbb{P}') = \gamma_1$. Also since $D(\theta, \gamma_1, \gamma_2)$ is e-parallel transport invariant along the section $\mathcal{M}_1(\mathbb{P})$, that is,

$$D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}$$
$$= \prod_{\mathbb{P}' \to \mathbb{P}}^{e} D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}$$
$$= D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\} - \mathbb{E}[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}],$$

implying $\mathbb{E}[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}] = 0$. We have

$$\begin{aligned} &\mathbb{E}[D\{\theta(\mathbb{P}), \gamma_1, \gamma_2(\mathbb{P})\}] \\ &= \mathbb{E}[D\{\theta(\mathbb{P}), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P})\}] \\ &= \mathbb{E}[D\{\theta(\mathbb{P}'), \gamma_1(\mathbb{P}'), \gamma_2(\mathbb{P}')\}] = 0. \end{aligned}$$

"Only if" part is trivial.

REFERENCES

AMARI, S.-I. (1982). Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics* **10** 357–385.

AMARI, S.-I. (1985). Differential-geometrical methods in statistics. *Lecture Notes on Statistics* **28** 1.

AMARI, S.-I. (2016). *Information geometry and its applications* **194**. Springer.

AMARI, S.-I. and KAWANABE, M. (1997). Information geometry of estimating functions in semi-parametric statistical models. *Bernoulli* 29–54.

AMARI, S.-I. and NAGAOKA, H. (2000). *Methods of information geometry* **191**. American Mathematical Soc.

AY, N., JOST, J., VÂN LÊ, H. and SCHWACHHÖFER, L. (2017). *Information geometry* **64**. Springer.

BICKEL, P. J. and KWON, J. (2001). Inference for semiparametric models: some questions and an answer. *Statistica Sinica* 863–886.

BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and adaptive estimation for semiparametric models* **4**. Johns Hopkins University Press Baltimore.

CHEN, H.-Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics* **63** 413–421.

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.

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics* 1189–1242.

HÄRDLE, W., LIANG, H. and GAO, J. (2000). *Partially linear models*. Springer Science & Business Media.

HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

KOSOROK, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer.

KUMON, M. and AMARI, S. (1983). Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* **387** 429–458.

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

RAO, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. Kotz S & Johnson NL (eds.), Breakthroughs in Statistics Volume I: Foundations and Basic Theory, 235–248.

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** 920–936.

ROBINS, J., LI, L., TCHETGEN, E., VAN DER VAART, A. et al. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and statistics: essays in honor of David A. Freedman* **2** 335–421.

ROTNITZKY, A., SMUCLER, E. and ROBINS, J. M. (2021). Characterization of parameters with a mixed bias property. *Biometrika* **108** 231–238.

SMUCLER, E., ROTNITZKY, A. and ROBINS, J. (2019). A unifying approach for doublyrobust l1 regularized estimation of causal contrasts. arXiv e-prints. *arXiv preprint arXiv:1904.03737*.

TCHETGEN TCHETGEN, E. J., ROBINS, J. M. and ROTNITZKY, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika* **97** 171–180.

TSIATIS, A. A. (2006). *Semiparametric theory and missing data*. Springer.

VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.

VANSTEELANDT, S. and JOFFE, M. (2014). Structural nested models and G-estimation: the partially realized promise. *Statistical Science* **29** 707–731.