University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2014 *Paper* 332

Statistical Inference for the Mean Outcome Under a Possibly Non-Unique Optimal Treatment Strategy

Alexander R. Luedtke* Mark J. van der Laan†

*University of California, Berkeley, aluedtke@berkeley.edu

†University of California, Berkeley, laan@berkeley.edu

Statistical Inference for the Mean Outcome Under a Possibly Non-Unique Optimal Treatment Strategy

Alexander R. Luedtke and Mark J. van der Laan

Abstract

We consider challenges that arise in the estimation of the value of an optimal individualized treatment strategy defined as the treatment rule that maximizes the population mean outcome, where the candidate treatment rules are restricted to depend on baseline covariates. We prove a necessary and sufficient condition for the pathwise differentiability of the optimal value, a key condition needed to develop a regular asymptotically linear (RAL) estimator of this parameter. The stated condition is slightly more general than the previous condition implied in the literature. We then describe an approach to obtain root-n rate confidence intervals for the optimal value even when the parameter is not pathwise differentiable. In particular, we develop an estimator that, when properly standardized, converges to a normal limiting distribution. We provide conditions under which our estimator is RAL and asymptotically efficient when the mean outcome is pathwise differentiable. We outline an extension of our approach to a multiple time point problem in the appendix. All of our results are supported by simulations.

1 Introduction

There has been much recent work in estimating optimal dynamic treatment regimes (DTRs) from a random sample. A DTR is an individualized treatment strategy in which treatment decisions for a patient can be based on their measured covariates. Such treatment strategies are commonly used in practice and thus it is natural to want to learn about the best such strategy. The value of a DTR is defined as the population counterfactual mean outcome if the DTR is implemented in the population. The optimal DTR is the DTR which has the maximal value. The value at the optimal DTR is known as the optimal value. In a single time point setting, the optimal DTR can be defined as the sign of the "blip function", defined as the additive effect of a blip in treatment on a counterfactual outcome, conditional on baseline covariates (Robins, 2004). For a general overview of recent work on optimal DTRs, see Chakraborty and Moodie (2013).

Suppose one wishes to know the impact of implementing an optimal DTR in the population, i.e. one wishes to know the optimal value. Inference for the optimal value has been shown to be difficult at exceptional laws, i.e. probability distributions where there exists a strata of the baseline covariates which occurs with positive probability for which treatment is neither beneficial nor harmful (Robins, 2004; Robins and Rotnitzky, 2014). Zhang et al. (2012a) considered inference for the optimal value in restricted classes in which the DTRs are indexed by a finite-dimensional parameter. At non-exceptional laws, they outlined an argument showing that the standard error of their estimator minus the truth, multiplied by root-n, is equal to the standard error of the estimator which estimates the value of the *known* optimal DTR.

Researchers are now focusing on applying machine learning algorithms to estimate the optimal rules from large classes which cannot be described by a finite dimensional parameter (see, e.g., Zhang et al., 2012b; Zhao et al., 2012; Luedtke and van der Laan, 2014). van der Laan and Luedtke (2014b) and van der Laan and Luedtke (2014a) developed inference for the optimal value when the DTR belongs to an unrestricted class. van der Laan and Luedtke (2014a) provide a proof that the efficient influence curve for the parameter which treats the optimal rule as known is equal to the efficient influence curve of the optimal value at non-exceptional laws. One of the contributions of this paper is to present a slightly more precise statement of the condition for the pathwise differentiability of the mean outcome under the optimal rule. We will show that this condition is necessary and sufficient.

Restricting inference to non-exceptional laws can be limiting given that there is often no treatment effect for people in some strata of baseline covariates. Chakraborty et al. (2014) propose using the m-out-of-n bootstrap to obtain inference for the value of an estimated DTR. They work with an inverse probability weighted (IPW) estimator, which yields valid inference when the treatment mechanism is known or is estimated according to a correctly specified parametric model. They also discuss an extension to an augmented inverse probability weighted (AIPW) estimator. The m-out-of-n bootstrap draws samples of size m patients from the data set of size n. In non-regular problems, this method yields valid inference if $m, n \to \infty$ and $m = o(n)$. This approach allows one to obtain confidence intervals for the value of an estimated regime which shrink at a slower than root-n rate, namely a root-m rate. In addition to yielding wide confidence intervals, this approach has the drawback of requiring a choice of the important tuning parameter m for each sample size. The choice of m balances a trade-off between coverage and efficiency. Chakraborty et al. propose

using a double bootstrap to select this tuning parameter.

Goldberg et al. (2014) instead consider the truncating the value function so that only individuals with a clinically meaningful treatment effect contribute to the value, and then proceed with inference for the truncated value function at the optimal DTR. For a fixed truncation level, these authors note that the estimated truncated optimal value minus the true truncated optimal value, multiplied by root-n, converges to a normal limiting distribution. Laber et al. (2014b) propose instead replacing the indicator in the value function with differentiable function. They conjecture about situations in which this estimator minus the truth, multiplied by root-n, has a reasonable limit distribution.

In this work, we develop root-n rate inference for the optimal value under reasonable conditions. Our approach avoids any sort of truncation, and does not require that the estimate of the optimal rule converge to a fixed quantity as the sample size grows. We are able to show that our estimator minus the truth, properly standardized, converges to a standard normal limiting distribution. This allows for the straightforward construction of asymptotically valid confidence intervals for the optimal value. Neither the estimator nor the inference relies on a complicated tuning parameter.

We give conditions under which our estimator is asymptotically efficient among all regular and asymptotically linear (RAL) estimators when the optimal value parameter is pathwise differentiable. The conditions for asymptotic efficiency are very similar to those presented in van der Laan and Luedtke (2014b), but do not require that one knows that the optimal value parameter is pathwise differentiable from the outset. The procedure is computationally efficient and its implementation only requires a minor modification to a typical one-step estimator.

Organization of article

Section 2 formulates the statistical problem of interest. Section 3 gives necessary and sufficient conditions for the pathwise differentiability of the optimal value. Section 4 outlines the challenge of obtaining inference at exceptional laws and gives a thought experiment that motivates our procedure for estimating the optimal value. Section 5 presents an estimator for the optimal value. This estimator represents a slight modification to a recently presented online one-step estimator for pathwise differentiable parameters. Section 6 discusses computationally efficient implementations of our proposed procedure. Section 7 discusses each condition of the key result presented in Section 5. Section 8 describes our simulations. Section 9 gives our simulation results. Section 10 closes with a summary and some directions for future work.

All proofs can be found in Appendix A. We outline an extension of our proposed procedure to the multiple time point setting in Appendix B.

2 Problem formulation

Let $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is a nonparametric model. Let $\mathcal{W}$ denote the range of W. For $P \in \mathcal{M}$, define the treatment mechanism $g(P)(A|W) \triangleq \Pr_P(A|W)$. We will refer to $g(P_0)$ as g_0 and refer to $g(P)$ as g when this will not cause confusion. For a function f, we will use $E_P[f(O)]$ to denote $\int f(o)dP(o)$. We will also use $E_0[f(O)]$ to denote $E_{P_0}[f(O)]$ and $\Pr_0$ to denote the probability of an event under P_0. Let $\Psi : \mathcal{M} \to \mathbb{R}$ be defined by $\Psi(P) \triangleq E_P E_P[Y|A = d(P)(W), W]$ and $d(P) = \text{argmax}_d E_P E_P(Y|A = d(W), W)$ be an optimal treatment rule under

P. We will resolve the ambiguity in the definition of d when the argmax is not unique later in this section. Throughout we assume that $\Pr_0(0 < g_0(1|W) < 1)$ so that $\Psi(P_0)$ is well-defined. Under causal assumptions, $\Psi(P)$ is equal to the counterfactual mean outcome if, possibly contrary to fact, the rule $d(P)$ were implemented in the population. We can also identify $d(P)$ with a causally optimal rule under those same assumptions. We refer the reader to van der Laan and Luedtke (2014b) for a more precise formulation of such a treatment strategy. As the focus of this work is statistical, all of the results will hold when estimating the statistical parameter $\Psi(P_0)$ whether or not the causal assumptions needed for identifiability hold. Define

$$\begin{aligned}\bar{Q}(P)(A,W) &\triangleq E_P[Y|A,W] \\ \bar{Q}_b(P)(W) &\triangleq \bar{Q}(P)(1,W) - \bar{Q}(P)(0,W).\end{aligned}$$

We will refer to $\bar{Q}_b(P)$ the blip function for the distribution P. We will denote to the above quantities applied to P_0 as $\bar{Q}_0$ and $\bar{Q}_{b,0}$, respectively. We will often omit the reliance on P altogether when there is only one distribution P under consideration: $\bar{Q}(A,W)$ and $\bar{Q}_b(W)$. We also define $\Psi_d(P) = E_P\bar{Q}(d(P)(W),W)$. Consider the efficient influence curve of Ψ_d at P:

$$D(d,P)(O) = \frac{I(A = d(W))}{g(A|W)}(Y - \bar{Q}(A,W)) + \bar{Q}(d(W),W) - \Psi_d(P).$$

Let $B(P) \triangleq \{w : \bar{Q}_b(w) = 0\}$. We will refer to $B(P_0)$ as B_0. An exceptional law is defined as a distribution P for which $\Pr_P(W \in B(P)) > 0$ (Robins, 2004). We note that the ambiguity in the definition of $d(P)$ occurs precisely on the set $B(P)$. In particular, $d(P)$ must almost surely agree with some rule in the class

$$\left\{w \mapsto I(\bar{Q}_b(w) > 0)I(w \notin B(P)) + b(w)I(w \in B(P)) : b\right\}, \tag{1}$$

where $b : \mathcal{W} \to \{0,1\}$ is some arbitrary (measurable) function. Consider now the following uniquely defined optimal rule:

$$d^*(P)(W) \triangleq I(\bar{Q}_b(W) > 0).$$

We will let $d_0^* = d^*(P_0)$. We have $\Psi(P) = \Psi_{d^*(P)}(P)$, but now $d^*(P)$ is uniquely defined for all W. More generally, $d^*(P)$ represents a uniquely defined optimal rule. Other formulations of the optimal rule can be obtained by changing the behavior of the rule B_0. Our goal is to construct root-n rate confidence intervals for $\Psi(P_0)$ that maintain nominal coverage, even at exceptional laws. At non-exceptional laws we would like these confidence intervals to belong to and be asymptotically efficient among the class of regular asymptotially linear (RAL) estimators.

3 Necessary and sufficient conditions for pathwise differentiability of Ψ

The pathwise derivative of Ψ at P_0 is defined as $\frac{d}{d\epsilon}\Psi(P_\epsilon)\big|_{\epsilon=0}$ along paths $\{P_\epsilon : \epsilon \in \mathbb{R}\} \subset \mathcal{M}$ that go through P_0 at $\epsilon = 0$, i.e. $P_{\epsilon=0} = P_0$ (Bickel et al., 1993). In particular, these paths are given

by

$$
\begin{aligned}
dQ_{W,\epsilon} &= (1+\epsilon S_W(W))dQ_{W,0}, \\
&\quad \text{where } E_0[S_W(W)] = 0 \text{ and } \sup_w |S_W(w)| < \infty; \\
dQ_{Y,\epsilon}(Y|A,W) &= (1+\epsilon S_Y(Y|A,W))dQ_{Y,0}(Y|A,W), \\
&\quad \text{where } E_0[S_Y|A,W] = 0 \text{ and } \sup_{w,a,y} |S_Y(y|a,w)| < \infty. \qquad (2)
\end{aligned}
$$

Above $Q_{W,0}$ and $Q_{Y,0}$ are respectively the marginal distribution of W and the conditional distribution of Y given A, W under P_0. The parameter Ψ is not sensitive to fluctuations of $g_0(a|w) = \Pr_0(a|w)$, and thus we do not need to fluctuate this portion of the likelihood.

In van der Laan and Luedtke (2014a), we showed that Ψ is pathwise differentiable at P_0 with canonical gradient $D(d_0^*, P_0)$ if P_0 is a non-exceptional law, i.e. $\Pr_0(W \notin B_0) = 1$. Exceptional laws were shown to present problems for estimation of optimal rules indexed by a finite dimensional parameter in Robins (2004), and it was observed in Robins and Rotnitzky (2014) that these laws can also cause problems for unrestricted optimal rules. Here we show that mean outcome under the optimal rule is pathwise differentiable under a slightly more general condition than requiring a non-exceptional law, namely that

$$
\Pr_0\left(W \notin B_0 \text{ or } \max_a \sigma_0^2(a,W) = 0\right) = 1, \qquad (3)
$$

where $\sigma_0(a,w) \triangleq \sqrt{Var_{P_0}(Y|A=a,W=w)}$. The upcoming theorem also gives the converse result, i.e. the mean outcome under the optimal rule is not pathwise differentiable if the above condition does not hold.

Theorem 1. *Assume* $\Pr_0(0 < g_0(1|W) < 1) = 1$*, and* $\Pr_0(|Y| < M) = 1$ *for some* $M < \infty$*. The parameter* $\Psi(P_0)$ *is pathwise differentiable if and only if (3). If* Ψ *is pathwise differentiable at* P_0*, then* Ψ *has canonical gradient* $D(d_0^*, P_0)$ *at* P_0*.*

In the proof of the theorem we construct fluctuations S_W and S_Y such that

$$
\lim_{\epsilon \uparrow 0} \frac{\Psi(P_\epsilon) - \Psi(P_0)}{\epsilon} \neq \lim_{\epsilon \downarrow 0} \frac{\Psi(P_\epsilon) - \Psi(P_0)}{\epsilon} \qquad (4)
$$

when (3) does not hold. It then follows that $\Psi(P_0)$ is not pathwise differentiable. The left- and right-hand sides above are referred to as one-sided directional derivatives in Hirano and Porter (2012).

We note that this condition for the mean outcome differs slightly from that implied for unrestricted rules in Robins and Rotnitzky (2014) in that we still have pathwise differentiability when the $\bar{Q}_{b,0}$ is zero in some strata but the conditional variance of the outcome given covariates and treatment is also zero in all of those strata. This makes sense, given that in this case the blip function could be estimated perfectly in those strata in any finite sample with treated and untreated individuals observed in that strata. Though we do not expect this difference to matter for most data generating distributions encountered in practice, there are cases where it may be relevant. For example, if no

one in a certain strata is susceptible to a disease regardless of treatment status, and researchers are unaware of this *a priori* so that simply excluding those strata from the target population is not an option, then the conditional variance in this strata is zero even though the effect of treatment in this strata is also zero.

In general, however, we expect that the mean outcome under the optimal rule will not be pathwise differentiable under exceptional laws encountered in practice. For this reason, we often refer to "exceptional laws" rather than "laws which do not satisfy (3)" in this work. We do this because the term "exceptional law" is well-established in the literature, and also because we believe that there is likely little distinction between "exceptional law" and "laws which do not satisfy (3)" for many problems of interest.

We remind the reader that an estimator $\hat{\Phi}$ is asymptotically linear for a parameter mapping Φ at P_0 with influence curve IC_0 if

$$\hat{\Phi}(P_n) - \Phi(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC_0(O_i) + o_{P_0}(n^{-1/2}),$$

where $E_0[IC_0(O)] = 0$.

For the definitions of regularity and local unbiasedness we let P_ϵ be as in (2), with g_0 also fluctuated. That is, we let $dP_\epsilon = dQ_{Y,\epsilon} \times g_\epsilon \times dQ_{W,\epsilon}$, where $g_\epsilon(A|W) = (1 + S_A(A|W))g_0(A|W)$ with $E_0[S_A(A|W)|W] = 0$ and $\sup_{a,w} |S_A(a|w)| < \infty$. The estimator $\hat{\Phi}$ of $\Phi(P_0)$ is called regular if the asymptotic distribution of $\sqrt{n}(\hat{\Phi}(P_n) - \Phi(P_0))$ is not sensitive to small fluctuations in the data generating distribution P_0. That is, the limiting distribution of $\sqrt{n}(\hat{\Phi}(P_{n,\epsilon=1/\sqrt{n}}) - \Phi(P_{\epsilon=1/\sqrt{n}}))$ does not depend on S_W, S_A, or S_Y, where $P_{n,\epsilon=1/\sqrt{n}}$ is the empirical distribution $O_1, ..., O_n$ drawn independently and identically distributed (i.i.d.) from $P_{\epsilon=1/\sqrt{n}}$. The estimator $\hat{\Phi}$ is called locally unbiased if the limiting distribution of $\sqrt{n}(\hat{\Phi}(P_{n,\epsilon=1/\sqrt{n}}) - \Phi(P_{\epsilon=1/\sqrt{n}}))$ has mean zero for all fluctuations S_W, S_A, and S_Y, and is called asymptotically unbiased (at P_0) if the bias of $\hat{\Phi}(P_n)$ for the parameter $\Phi(P_0)$ is $o_{P_0}(n^{-1/2})$ at P_0.

We note that the non-regularity of a statistical inference problem does not typically imply the nonexistence of asymptotically unbiased estimators (see Example 4 of Liu and Brown, 1993 and the discussion thereof in Chen, 2004), but rather the nonexistence of *locally* asymptotically unbiased estimators (Hirano and Porter, 2012). It is thus not surprising that we are able to find an estimator $\hat{\Psi}$ that is asymptotically unbiased at a fixed (possibly exceptional) law under mild assumptions. Hirano and Porter also show that there does not exist a regular estimator of the optimal value at any law for which the optimal value is not pathwise differentiable. That is, no regular estimators of $\Psi(P_0)$ exist at laws which satisfy the conditions of Theorem 1 but do not satisfy (3). It follows that one must accept the nonregularity of their estimator when the data is generated according to such laws. Note that this does not rule out the development of locally consistent confidence bounds similar to those presented in Laber and Murphy (2011) and Laber et al. (2014a), though such approaches can be conservative when the estimation problem is nonregular.

In this work we present an estimator $\hat{\Psi}$ for which $\Gamma_n\sqrt{n}(\hat{\Psi}(P_n) - \Psi(P_0))$ converges in distribution to a standard normal distribution for a random standardization term Γ_n under reasonable conditions. Our estimator does not require any complicated tuning parameters, and thus allows one to easily

develop root-n rate confidence intervals for the optimal value. We show that our estimator is RAL and efficient at laws which satisfy (3) under conditions.

4 Inference at exceptional laws

4.1 The challenge

Before presenting our estimator, we discuss the challenge of estimating the optimal value at exceptional laws. Suppose d_n is an estimate of d_0^* and $\hat{\Psi}_{d_n}(P_n)$ is an estimate of $\Psi(P_0)$ relying on the full data set. In van der Laan and Luedtke (2014b) we presented a targeted minimum loss-based estimator (TMLE) $\hat{\Psi}_{d_n}(P_n)$ which satisfies

$$\hat{\Psi}_{d_n}(P_n) - \Psi(P_0) = (P_n - P_0)D(d_n, P_n^*) + \underbrace{\Psi_{d_n}(P_0) - \Psi(P_0)}_{o_{P_0}(n^{-1/2}) \text{ if optimal rule estimated well}} + o_{P_0}(n^{-1/2}),$$

where we use the notation $Pf = E_P[f(O)]$ for any distribution P and the second $o_{P_0}(n^{-1/2})$ term is a remainder from a first-order expansion of Ψ. The term $\Psi_{d_n}(P_0) - \Psi(P_0)$ relies on the optimal rule being estimated well in terms of value and will often prove to be a reasonable condition, even at exceptional laws (see Theorem 7 in Section 7.5). Here P_n^* is an estimate of the components of P_0 needed to estimate $D(d_n, P_0)$. To show asymptotic linearity, one might try to replace $D(d_n, P_n^*)$ with a term that does not rely on the sample:

$$(P_n - P_0)D(d_n, P_n^*) = (P_n - P_0)D(d_0^*, P_0) + \underbrace{(P_n - P_0)(D(d_n, P_n^*) - D(d_0^*, P_0))}_{\text{empirical process}},$$

If $D(d_n, P_n^*)$ belongs to a Donsker class and converges to $D(d_0^*, P_0)$ in $L^2(P_0)$, then the empirical process term is $o_{P_0}(n^{-1/2})$ and $\sqrt{n}(\hat{\Psi}_{d_n}(P_n) - \Psi(P_0))$ converges in distribution to a normal random variable with mean zero and variance $Var_{P_0}(D(d_0^*, P_0))$ (van der Vaart and Wellner, 1996). Note that $D(d_n, P_n^*)$ being consistent for $D(d_0^*, P_0)$ will rely on d_n being consistent for the fixed d_0^* in $L^2(P_0)$ in most cases, which we emphasize is *not* in general implied by $\Psi_{d_n}(P_0) - \Psi(P_0) = o_{P_0}(n^{-1/2})$. Zhang et al. (2012a) make this assumption in the regularity conditions in their Web Appendix A when they consider an analogous empirical process term in deriving the standard error of an estimate of the value function at an optimal rule in a parametric working model. More specifically, Zhang et al. assume a non-exceptional law and consistent estimation of a fixed optimal rule. van der Laan and Luedtke (2014b) also make such an assumption. If P_0 is not an exceptional law, then we likely do not expect d_n to be consistent for any fixed function. Rather, we expect d_n to fluctuate randomly on the set B_0, even as the sample size grows to infinity. In this case the empirical process term considered above is not expected to behave as $o_{P_0}(n^{-1/2})$.

Accepting that our estimates of the optimal rule may not stabilize as sample size grows, we consider an estimation strategy that allows d_n to remain random even as $n \to \infty$.

4.2 A thought experiment

First we give an erroneous estimation strategy which contains the main idea of the approach but is not correct in its current form. A modification is given in the next section. For simplicity, we will

assume that one knows $v_n \triangleq Var_{P_0}(D(d_n, P_0))$ given an estimate d_n and, for simplicity, that v_n is almost surely bounded away from zero. Under reasonable conditions,

$$v_n^{-1/2}\left(\hat{\Psi}_{d_n}(P_n) - \Psi(P_0)\right) = (P_n - P_0)v_n^{-1/2}D(d_n, P_n^*) + o_{P_0}(n^{-1/2}).$$

The empirical process on the right is difficult to handle because d_n and v_n are random quantities that likely will not stabilize to a fixed limit at exceptional laws.

As a thought experiment, suppose that we could treat $\{v_n^{-1/2}D(d_n, P_n^*) : n\}$ as a deterministic sequence, where this sequence does not necessarily stabilize as sample size grows. In this case the Lindeberg-Feller central limit theorem (CLT) for triangular arrays (see, e.g., Athreya and Lahiri, 2006) would allow us to show that the leading term on the right-hand side converges to a standard normal random variable. This result relies on inverse weighting by $\sqrt{v_n}$ so the variance of the terms in the sequence stabilizes to one as sample size gets large.

Of course we cannot treat these random quantities as deterministic. In the next section we will use the general trick of inverse weighting by the standard deviation of the terms over which we are taking an empirical mean, but we will account for the dependence of the estimated rule d_n on the data by inducing a martingale structure that allows us to treat a sequence of estimates of the optimal rule as known (conditional on the past). This will allow us to apply a martingale CLT for triangular arrays to obtain a limiting distribution for a standardized sequence involving our estimator.

5 Estimation of and inference for the optimal value

In this section we present a modified one-step estimator $\hat{\Psi}$ of the optimal value. This estimator relies on estimates of the treatment mechanism g_0, the strata-specific outcome $\bar{Q}_0$, and the optimal rule d_0^*. We first present our estimator, and then present an asymptotically valid two-sided confidence interval for the optimal value under conditions. Next we give conditions under which our estimator is RAL and efficient, and finally we present a (potentially conservative) asymptotically valid one-sided confidence interval which lower bounds the mean outcome under the unknown optimal treatment rule. The one-sided confidence interval uses the same lower bound from the two-sided confidence interval, but does not require a condition about the rate at which the value of the optimal rule converges to the optimal value, or even that the value of the estimated rule is consistent for the optimal value.

The estimators in this section can be extended to a martingale-based TMLE for $\Psi(P_0)$. Because the primary purpose of this paper is to deal with inference at exceptional laws, we will only present an online one-step estimator and leave the presentation of such a TMLE to future work.

5.1 Estimator of the optimal value

Define

$$\tilde{D}(d, \bar{Q}, g)(o) \triangleq \frac{I(a = d(w))}{g(a|w)}(Y - \bar{Q}(a, w)) + \bar{Q}(d(w), w).$$

Let $\{\ell_n\}$ be some sequence of nonnegative integers representing the smallest sample on which the optimal rule is learned. While ℓ_n does constitute a tuning parameter, we show in our simulation that

our procedure is not overly sensitive to the choice of ℓ_n. For each $j = 1, ..., n$, let $P_{n,j}$ represent the empirical distribution of the observations $(O_1, O_2, ..., O_j)$. Let $g_{n,j}$, $\bar{Q}_{n,j}$, and $d_{n,j}$ respectively represent estimates of the g_0, $\bar{Q}_0$, and d_0^* based on (some subset of) the observations $(O_1, ..., O_{j-1})$. We subscript each of these estimates by both n and j because the subsets on which these estimates are obtained may depend on sample size. We give an example of a situaton where this would be desirable in Section 6.1.

Define

$$\tilde{\sigma}^2_{0,n,j} \triangleq Var_{P_0}\left(\tilde{D}(d_{n,j}, \bar{Q}_{n,j}, g_{n,j})(O)\middle| O_1, ..., O_{j-1}\right).$$

Let $\tilde{\sigma}^2_{n,j}$ represent an estimate of $\tilde{\sigma}^2_{0,n,j}$ based on (some subset of) the observations $(O_1, ..., O_{j-1})$. Note that we omit the dependence of $\tilde{\sigma}_{n,j}$ and $\tilde{\sigma}_{0,n,j}$ on $d_{n,j}$, $\bar{Q}_{n,j}$, and $g_{n,j}$ in the notation. Our results apply to any sequence of estimates $\tilde{\sigma}^2_{n,j}$ which satisfies conditions C1) through C5), which are stated later in this section. Also define

$$\Gamma_n \triangleq \frac{1}{n - \ell_n} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}^{-1}_{n,j}.$$

Our estimate $\hat{\Psi}(P_n)$ of $\Psi(P_0)$ is given by

$$\hat{\Psi}(P_n) \triangleq \Gamma_n^{-1} \frac{1}{n - \ell_n} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}^{-1}_{n,j} \tilde{D}_{n,j}(O_j) = \frac{\sum_{j=\ell_n+1}^{n} \tilde{\sigma}^{-1}_{n,j} \tilde{D}_{n,j}(O_j)}{\sum_{j=\ell_n+1}^{n} \tilde{\sigma}^{-1}_{n,j}}, \tag{5}$$

where $\tilde{D}_{n,j} \triangleq \tilde{D}(d_{n,j}, \bar{Q}_{n,j}, g_{n,j})$. We note that the Γ_n^{-1} standardization is used to account for the term-wise inverse weighting so that $\hat{\Psi}(P_n)$ estimates $\Psi(P_0) = E_0[\tilde{D}(d_0^*, \bar{Q}_0, g_0)]$. The above looks a lot like a standard AIPW estimator, but with d_0^* estimated on chunks of data increasing in size and with each term in the sum given a convex weight proportional to an estimate of the conditional variance of that term. Our estimator constitutes a minor modification of the online one-step estimator presented in van der Laan and Lendle (2014). In particular, each term in the sum is inverse weighted by an estimate of the standard deviation of $\tilde{D}_{n,j}$. For ease of reference we will refer to the estimator above as an online one-step estimator of $\Psi(P_0)$.

5.2 Two-sided confidence interval for the optimal value

Define the remainder terms

$$R_{1n} \triangleq \frac{1}{n - \ell_n} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}^{-1}_{n,j} E_0\left[\left(1 - \frac{g_0(d_{n,j}(W)|W)}{g_{n,j}(d_{n,j}(W)|W)}\right)\left(\bar{Q}_{n,j}(d_{n,j}(W), W) - \bar{Q}_0(d_{n,j}(W), W)\right)\right]$$

$$R_{2n} \triangleq \frac{1}{n - \ell_n} \sum_{j=\ell_n+1}^{n} \frac{\Psi_{d_{n,j}}(P_0) - \Psi(P_0)}{\tilde{\sigma}_{n,j}}.$$

The upcoming theorem relies on the following assumptions:

C1) $n - \ell_n$ diverges to infinity as n diverges to infinity.

C2) Lindeberg-like condition: for all $\epsilon > 0$,

$$\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(\frac{\tilde{D}_{n,j}(O)}{\tilde{\sigma}_{n,j}}\right)^2 I\left(\frac{|\tilde{D}_{n,j}(O)|}{\tilde{\sigma}_{n,j}} > \epsilon\sqrt{n-\ell_n}\right)\middle| O_1, ..., O_{j-1}\right] = o_{P_0}(1).$$

C3) $\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n}\frac{\tilde{\sigma}^2_{0,n,j}}{\tilde{\sigma}^2_{n,j}} \to 1$ in probability.

C4) $R_{1n} = o_{P_0}(n^{-1/2})$.

C5) $R_{2n} = o_{P_0}(n^{-1/2})$.

The assumptions are discussed in Section 7. We note that all of our results also hold with R_{1n} and R_{2n} behaving as $o_{P_0}(1/\sqrt{n-\ell_n})$, though we do not expect this observation to be of use in practice as we recommend choosing ℓ_n so that $n - \ell_n$ increases at the same rate as n.

Theorem 2. *Under Conditions C1) through C5), we have that*

$$\Gamma_n\sqrt{n-\ell_n}\left(\hat{\Psi}(P_n) - \Psi(P_0)\right) \rightsquigarrow N(0,1),$$

where we use "$\rightsquigarrow$" to denote convergence in distribution as the sample size converges to infinity. It follows that an asymptotically valid $1-\alpha$ confidence interval for $\Psi(P_0)$ is given by

$$\hat{\Psi}(P_n) \pm z_{1-\alpha/2}\frac{\Gamma_n^{-1}}{\sqrt{n-\ell_n}},$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of a standard normal random variable.

We have shown that, under very general conditions, the above confidence interval yields an asymptotically valid $1-\alpha$ confidence interval for $\Psi(P_0)$. We refer the reader to Section 7 for a detailed discussion of the conditions of the theorem. We note that our estimator is asymptotically unbiased, i.e. has bias of the order $o_{P_0}(n^{-1/2})$, provided that $\Gamma_n = O_{P_0}(1)$ and $n - \ell_n$ grows at the same rate as n.

5.3 Conditions for asymptotic efficiency

We will now show that, if P_0 is a non-exceptional law and $d_{n,j}$ has a fixed optimal rule limit d_0, then we will show that our online estimator is RAL for $\Psi(P_0)$. The upcoming corollary makes use of the following consistency conditions for some fixed rule d_0 which falls in the class of optimal rules given in (1):

$$\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(d_{n,j}(W) - d_0(W)\right)^2\middle| O_1, ..., O_{j-1}\right] = o_{P_0}(1) \tag{6}$$

$$\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(\bar{Q}_{n,j}(d_0(W), W) - \bar{Q}_0(d_0(W), W)\right)^2\middle| O_1, ..., O_{j-1}\right] = o_{P_0}(1) \tag{7}$$

$$\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(g_{n,j}(d_0(W)|W) - g_0(d_0(W)|W)\right)^2\middle| O_1, ..., O_{j-1}\right] = o_{P_0}(1). \tag{8}$$

It also makes use of the following conditions, which are, respectively, slightly stronger than Conditions C1) and C3):

C1') $\ell_n = o(n)$.

C3') $\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n}\left|\frac{\tilde{\sigma}^2_{0,n,j}}{\tilde{\sigma}^2_{n,j}}-1\right| \to 0$ in probability.

Corollary 3. *Suppose that Conditions C1'), C2), C3'), C4), and C5) hold. Also suppose that* $\Pr_0(\delta < g_0(1|W) < 1-\delta) = 1$ *for some* $\delta > 0$*, the estimates* $g_{n,j}$ *are bounded below by some* $\delta > 0$ *with probability* 1*,* Y *is bounded, the estimates* $\bar{Q}_{n,j}$ *are uniformly bounded,* $\ell_n = o(n)$*, and that, for some fixed optimal rule* d_0*, (6), (7), and (8) hold. Finally, assume that* $Var_{P_0}(\tilde{D}(d_0, \bar{Q}_0, g_0)) > 0$ *and that, for some* $\delta_0 > 0$*, we have that*

$$\Pr_0\left(\inf_{j,n}\tilde{\sigma}^2_{n,j} > \delta_0\right) = 1,$$

where the infimum is over natural number pairs (j, n) *for which* $\ell_n < j \leq n$*. Then we have that*

$$\Gamma_n^{-1} \to Var_{P_0}(\tilde{D}(d_0, \bar{Q}_0, g_0)) \text{ in probability as } n \to \infty. \tag{9}$$

Additionally,

$$\hat{\Psi}(P_n) - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} D(d_0, P_0) + o_{P_0}(1/\sqrt{n}). \tag{10}$$

That is, $\hat{\Psi}(P_n)$ *is asymptotically linear with influence curve* $D(d_0, P_0)$*. Under the conditions of this corollary, it follows that* P_0 *satisfies (3) if and only if* $\hat{\Psi}(P_n)$ *is RAL and asymptotically efficient among all such RAL estimators.*

We note that (9) combined with C1') implies that the confidence interval given in Theorem 2 asymptotically has the same width (up to an $o_{P_0}(n^{-1/2})$ term) as the confidence interval which treats (10) and $D(d_0, P_0)$ as known and establishes a typical Wald-type confidence interval about $\hat{\Psi}(P_n)$.

The empirical averages over j in (6), (7), and (8) can easily be dealt with using Lemma 5, presented in Section 7.3. Essentially we have required that $d_{n,j}$, $\bar{Q}_{n,j}$, and $g_{n,j}$ are consistent for d_0, $\bar{Q}_0$, and g_0 as n and j get large, where d_0 is some fixed optimal rule. One would expect such a fixed limiting rule d_0 to exist at a non-exceptional law for which the optimal rule is (almost surely) unique. If g_0 is known then we do not need that $\bar{Q}_{n,j}$ is consistent for $\bar{Q}_0$ to get asymptotic linearity, but rather that $\bar{Q}_{n,j}$ converges to some possibly misspecified fixed limit $\bar{Q}$. The proof of such a result is analogous to the proof of (10) in the above corollary so is omitted.

5.4 Lower bound for the optimal value

It would likely be useful to have a conservative lower bound on the optimal value in practice. If policymakers were to implement an optimal individualized treatment rule whenever the overall benefit is greater than some fixed threshold, i.e. $\Psi(P_0) > v$ for some fixed v, then a one-sided

confidence interval for $\Psi(P_0)$ would help facilitate the decision to implement an individualized treatment strategy in the population.

The upcoming theorem shows that the lower bound from the $1-2\alpha$ confidence interval yields a (potentially conservative) asymptotic $1-\alpha$ confidence interval for the optimal value. If d_0^* is estimated well in the sense of Condition C5), then the asymptotic coverage is exactly $1-\alpha$. Define

$$LB_n(\alpha) \triangleq \hat{\Psi}(P_n) - z_{1-\alpha}\frac{\Gamma_n^{-1}}{\sqrt{n-\ell_n}},$$

where $z_{1-\alpha}$ denotes the $1-\alpha$ quantile of a standard normal random variable. We have the following theorem.

Theorem 4. *Under Conditions C1) through C4), we have that*

$$\liminf_{n\to\infty} \Pr_0\left(\Psi(P_0) > LB_n(\alpha)\right) \geq 1-\alpha.$$

If Condition C5) also holds, then

$$\lim_{n\to\infty} \Pr_0\left(\Psi(P_0) > LB_n(\alpha)\right) = 1-\alpha.$$

The above condition should not be surprising, as we base our confidence interval for $\Psi(P_0)$ on a weighted combination of estimates of $\Psi_{d_{n,j}}(P_0)$ for $j < n$. Because $\Psi_{d_0}(P_0) \geq \Psi_{d_{n,j}}(P_0)$ for all such j, we would expect that the lower bound of the $1-\alpha$ confidence interval given in the previous section provides a valid $1-\alpha/2$ one-sided confidence interval for $\Psi(P_0)$. Indeed this is precisely what we see in the proof of the above theorem.

6 Computationally efficient estimation schemes

Computing $\hat{\Psi}(P_n)$ may initially seem computationally demanding. In this section we discuss two estimation schemes for which estimating $\hat{\Psi}(P_n)$ which yield computationally simple routines. The first option is a straightforward implementation strategy, while the second is a more sophisticated implementation which allows our proposed estimator to scale to large data sets.

6.1 Computing the nuisance functions on large chunks of the data

One can compute the estimates of $\bar{Q}_0$, g_0, and d_0 far fewer than $n-\ell_n$ times. For each j, the estimates $\bar{Q}_{n,j}$, $g_{n,j}$, and $d_{n,j}$ may rely on any subset of the observations $O_1, ..., O_{j-1}$. Thus one can compute these estimators on S increasing subsets of the data, where the first subset consists of observations $O_1, ..., O_{\ell_n}$ and each of the $S-1$ remaining samples adds a $1/S$ proportion of the remaining $n-\ell_n$ observations. Note that this scheme makes use of the fact that, for fixed j, the nuisance function estimates, indexed by n and j, e.g. $d_{n,j}$, may rely on different subsets of observations $O_1, ..., O_{j-1}$ for different sample sizes n.

We note that for $\ell_n \approx n/2$ and $S = 1$, our estimator corresponds to a simple sample split estimator which splits the data in half, learns the rule on the first half, and estimates the expected mean

outcome and a confidence interval for this value on the second half of the sample. The only minor twist is that our estimator estimates the standard error of the estimator on the first rather than the second half of the data set. For large samples, $\ell_n \approx n/2$ makes such an estimator have confidence intervals with width approximately $\sqrt{2}$ times wider than those which use $\ell_n = o(n)$. Thus in practice we recommend choosing $\ell_n = o(n)$, or even bounding ℓ_n so that $\limsup_n \ell_n < \infty$.

6.2 Online learning of the optimal value

As previously discussed, the estimator $\hat{\Psi}$ was inspired by online estimators which can operate on large data sets that will not fit into memory. These estimators use online prediction and regression algorithms which update the intial fit based on previously observed estimates using new observations which were just read into memory. Online estimators of pathwise differentiable parameters were introduced in van der Laan and Lendle (2014). Such estimation procedures often require estimates of nuisance functions, which can be obtained using modern online regression and classification approaches (see, e.g., Zhang, 2004; Langford et al., 2009; Luts et al., 2014). Our estimator constitutes a slight modification of the one-step online estimator presented in van der Laan and Lendle (2014), and thus all discussion of computational efficiency given in that paper also applies to our case.

For our estimator, one could use online estimators of $\bar{Q}_0$, g_0, and d_0, and then update these estimators as the index j in the sum in (5) increases. Calculating the standard error estimate $\tilde{\sigma}_{n,j}$ will typically require access to an increasing subset of the past observations, i.e. as sample size grows one may need to hold a growing number of observations in memory to estimate $\tilde{\sigma}_{0,n,j}$. If one uses a sample standard deviation to estimate $\tilde{\sigma}_{0,n,j}$ based on subset of observations $O_1, ..., O_{j-1}$, the results we present in Section 7.3 will indicate that one really only needs that the number of points on which $\tilde{\sigma}_{0,n,j}$ is estimated grows with j rather than at the same rate as j. This suggest that, if computation time or system memory is a concern for calculating $\tilde{\sigma}_{n,j}$, then one could calculate $\tilde{\sigma}_{n,j}$ based on some $o(j)$ subset of observations $O_1, ..., O_{j-1}$. For example, one could use the sample standard deviation of $\tilde{D}_{n,j}$ calculated on observations $O_1, ..., O_{t(j)}$, where $t(j) \triangleq j - 1$ if j is less than 500, and $t(j) \triangleq 500 + \lfloor \sqrt{j - 500} \rfloor$ otherwise.

We leave deeper consideration of the online estimation of the nuisance functions $\bar{Q}_0$, g_0, and d_0 to future work.

7 Discussion of the conditions of Theorem 2

For ease of notation we will assume that, for all $j > \ell_n$, we do not modify our nuisance function estimates based on the first $j - 1$ data points as the sample size grows. That is, for all sample sizes m, n and all $j \leq \min\{m, n\}$, $d_{n,j} = d_{m,j}$, $\bar{Q}_{n,j} = \bar{Q}_{m,j}$, $g_{n,j} = g_{m,j}$, and $\tilde{\sigma}_{n,j} = \tilde{\sigma}_{m,j}$. One can easily extend all of the discussion in this section to a more general case where, e.g., $d_{n,j} \neq d_{m,j}$ for $n \neq m$. This may be useful if the optimal rule is estimated in chunks of increasing size as was discussed in Section 6.1. To make these object's lack of dependence on n clear, in this section we will denote $d_{n,j}$, $\bar{Q}_{n,j}$, $g_{n,j}$, $\tilde{\sigma}_{n,j}$, and $\tilde{\sigma}_{0,n,j}$ as d_j, $\bar{Q}_j$, g_j, $\tilde{\sigma}_j$ and $\tilde{\sigma}_{0,j}$. This will also help make it clear when o_{P_0} notation refers to behavior as j, rather than n, goes to infinity.

For our discussion we assume there exists a (possibly unknown) $\delta_0 > 0$ such that

$$\Pr_0\left(\inf_{j>\ell_n} \tilde{\sigma}^2_{0,j} > \delta_0\right) = 1, \tag{11}$$

where the probability statement is over the i.i.d. draws $O_1, O_2, \ldots$. The above condition is not necessary, but will make our discussion of the conditions more straightforward. Such a δ_0 may be known if Y is binary and $\bar{Q}_0(a,w) \in (\gamma, 1-\gamma)$ for all a and w and a known $\gamma > 0$. There is not in general any need to actually know this bound.

7.1 Discussion of Condition C1)

We cannot apply the martingale CLT in the proof of Theorem 2 if $n - \ell_n$ does not grow with sample size. Essentially this condition requires that a non-negligible proportion of the data is used to actually estimate the mean outcome under the optimal rule. In practice one would likely like that $n - \ell_n$ grows at the same rate as n grows, which holds if e.g. $\ell_n = pn$ for some fixed proportion p of the data. This will allow our confidence intervals to shrink at a root-n rate. One might even make sure that $\ell_n = o(n)$ so that $\frac{n-\ell_n}{n}$ converges to 1 as sample size grows. In this case we can show that our estimator is asymptotically linear and efficient at non-exceptional laws under conditions, as we did in Corollary 3.

7.2 Discussion of Condition C2)

This is a standard condition that yields a martingale CLT for triangular arrays (Gaenssler et al., 1978). The condition ensures that the variables which are being averaged have sufficiently thin tails. While it is worth stating the condition in general, it is easy to verify that the condition is implied by the following three more straightforward conditions:

- (11) holds.
- Y is a bounded random variable.
- There exists some $\delta > 0$ such that $\Pr_0(\delta < g_j(1|W) < 1-\delta) = 1$ with probability 1 for all j.

7.3 Discussion of Condition C3)

This is a rather weak condition given that $\tilde{\sigma}_{0,j}$ still treats d_j as random. Thus this condition does not require that d_j stabilizes as j gets large. Suppose that

$$\tilde{\sigma}^2_j - \tilde{\sigma}^2_{0,j} = o_{P_0}(1) \tag{12}$$

By (11) and the continuous mapping theorem, it follows that

$$\frac{\tilde{\sigma}^2_{0,j}}{\tilde{\sigma}^2_j} - 1 = o_{P_0}(1). \tag{13}$$

The following general lemma will be useful in establishing Conditions C3), C4), and C5).

Lemma 5. *Suppose that R_j is some sequence of (finite) real-valued random variables such that $R_j = o_{P_0}(j^{-\beta})$ for some $\beta \in [0, 1)$, where we assume that each R_j is measurable with respect to the sigma-algebra generated by $(O_1, ..., O_j)$. Then,*

$$\frac{1}{n}\sum_{j=1}^{n} R_j = o_{P_0}(n^{-\beta}).$$

Applying the above lemma with $\beta = 0$ to (13) shows that Condition C3) holds provided that (11) and (12) hold. We will use the above lemma with $\beta = 1/2$ for when discussing Conditions C4) and C5).

It remains to show that we can construct a sequence of estimators such that (12) holds. Suppose we estimate $\tilde{\sigma}^2_{0,j}$ with

$$\tilde{\sigma}^2_j \triangleq \max\left\{\delta_j, \frac{1}{j-1}\sum_{i=1}^{j-1}\tilde{D}^2_j(O_i) - \left(\frac{1}{j-1}\sum_{i=1}^{j-1}\tilde{D}_j(O_i)\right)^2\right\}, \tag{14}$$

where $\{\delta_j\}$ is a sequence that may rely on j and each $\tilde{D}_{n,j} = \tilde{D}_j$ for all $n \geq j$. We use δ_j to ensure that $\tilde{\sigma}_j^{-2}$ is well-defined (and finite) for all j. If a lower bound δ_0 on $\tilde{\sigma}^2_{0,j}$ is known then one can take $\delta_j = \delta_0$ for all j. Otherwise one can let $\{\delta_j\}$ be some sequence such that $\delta_j \downarrow 0$ as $j \to \infty$.

Note that $\tilde{\sigma}^2_j$ is an empirical process because it involves sums over observations $O_1, ..., O_{j-1}$, and functions $\tilde{D}_j$ which were estimated on those same observations. The following theorem gives sufficient conditions for (12), and thus Condition C3), to hold.

Theorem 6. *Suppose (11) holds and that $\left\{\tilde{D}(d, \bar{Q}, g) : d, \bar{Q}, g\right\}$ is a P_0 Glivenko-Cantelli (GC) class with an integrable envelope function, where d, $\bar{Q}$, and g are allowed to vary over the range of the estimators of d_0^*, $\bar{Q}_0$, and g_0. Let $\tilde{\sigma}^2_j$ be defined as in (14). Then we have that $\tilde{\sigma}^2_j - \tilde{\sigma}^2_{0,j} = o_{P_0}(1)$. It follows that (13) and Condition C3) are satisfied.*

We thus only make the very mild assumption that our estimators of d_0^*, $\bar{Q}_0$, and g_0 belong to GC classes. Note that this assumption is much milder than the typical Donsker condition needed when attempting to establish the asymptotic normality of a (non-online) one-step estimator. An easy sufficient condition for a class to have a finite envelope function is if it is uniformly bounded, which occurs if the conditions discussed in Section 7.2 hold.

7.4 Discussion of Condition C4)

This condition is a weighted version of the typical double robust remainder appearing in the analysis of the AIPW estimator. Suppose that

$$E_0\left[\left(1 - \frac{g_0(d_j(W)|W)}{g_j(d_j(W)|W)}\right)\left(\bar{Q}_j(d_j(W), W) - \bar{Q}_0(d_j(W), W)\right)\right] = o_{P_0}(j^{-1/2}). \tag{15}$$

If g_0 is known (as in an RCT without missingness) and one takes each $g_j = g_0$ then the above ceases to be a condition as the left-hand side is always zero. We note that the only condition on

$\bar{Q}_j$ appears in Condition C4), so that if $R_{1n} = 0$ as in an RCT without missingness then we do not require that $\bar{Q}_j$ stabilizes as j grows. A typical AIPW estimator require the estimate of $\bar{Q}_0$ to stabilize as sample size grows to get valid inference, but here we have avoided this condition in the case where g_0 is known by using the martingale structure and inverse weighting by the standard error of each term in the definition of $\hat{\Psi}(P_n)$.

More generally, Lemma 5 shows that Condition C4) holds if (13) and (15) hold and $\Pr_0(0 < g_j(1|W) < 1) = 1$ with probability 1 for all j. One can apply the Cauchy-Schwarz inequality and take the maximum over treatment assignments to see that (15) holds if

$$\max\left\{\frac{1}{g_j(a|W)}\left\|g_j(a|W) - g_0(a|W)\right\|_{2,P_0}\left\|\bar{Q}_j(a,W) - \bar{Q}_0(a,W)\right\|_{2,P_0} : a = 0,1\right\} = o_{P_0}(j^{-1/2}).$$

If g_0 is not known, the above shows that then (15) holds if g_0 and $\bar{Q}_0$ are estimated well.

7.5 Discussion of Condition C5)

This condition requires that we can estimate d_0^* well as sample size gets large. We now give a theorem which will help us to establish Condition C5) under reasonable conditions. The theorem assumes the following margin assumption for some $\alpha > 0$:

$$\Pr_0(0 < |\bar{Q}_{b,0}(W)| \leq t) \lesssim t^\alpha \quad \forall t > 0, \tag{16}$$

where "$\lesssim$" denotes less than or equal to up to a nonnegative constant. This assumption is a direct restatement of Assumption (MA) from Audibert and Tsybakov (2007) and was considered earlier in Tsybakov (2004). Note that this theorem is similar in spirit to Lemma 1 in van der Laan and Luedtke (2014b), but relies on weaker, and we believe more interpretable, assumptions.

Theorem 7. *Suppose (16) holds for some $\alpha > 0$ and that we have an estimate $\bar{Q}_{b,n}$ of $\bar{Q}_{b,0}$ based on a sample of size n. If $\left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{2,P_0} = o_{P_0}(1)$, then*

$$|\Psi_{d_n}(P_0) - \Psi_{d_0^*}(P_0)| \lesssim \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{2,P_0}^{2(1+\alpha)/(2+\alpha)},$$

where d_n is the function $w \mapsto I(\bar{Q}_{b,n}(w) > 0)$. If $\left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0} = o_{P_0}(1)$, then

$$\begin{aligned}|\Psi_{d_n}(P_0) - \Psi_{d_0^*}(P_0)| &\leq \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0} \Pr_0\left(0 < \bar{Q}_{b,0}(W) \leq \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0}\right) \\ &\lesssim \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0}^{1+\alpha}.\end{aligned}$$

The first part of the above proof is essentially a restatement of Lemma 5.2 in Audibert and Tsybakov (2007). Figure 1 shows various densities which satisfy (16) at different values of α, and also the slowest rate of convergence for the blip function estimates for which Theorem 7 implies Condition C5). As is evident in the figure, $\alpha > 1$ implies that $p_{b,0}(t) \to 0$ as $t \to 0$. Given that we are interested in laws where $\Pr_0(\bar{Q}_{b,0}(W) = 0) > 0$, it is unclear how likely we are to have that $\alpha > 1$ when W contains only continuous covariates. One might, however, believe that the density is bounded near zero so that (16) is satisfied at $\alpha = 1$.

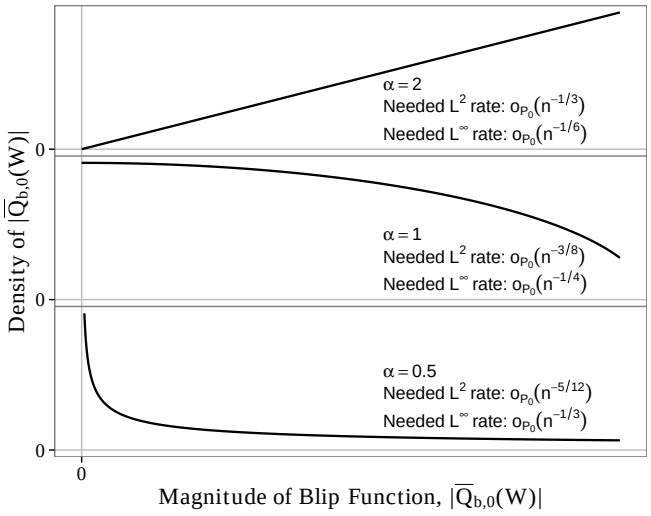

Figure 1: Examples of three densities of $|\bar{Q}_{b,0}(W)|$ whose corresponding cumulative distribution functions satisfy (16). If the rate of convergence of $\bar{Q}_{b,n} - \bar{Q}_{b,0}$ to zero in $L^2(P_0)$ or $L^\infty(P_0)$ attains the rates indicated above indicated above, then condition C5) will be satisfied for the plug-in optimal rule estimate considered in Theorem 7.

We note that if $\left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0} = o_{P_0}(1)$ then the above theorem indicates an arbitrarily fast rate for $\Psi_{d_n}(P_0) - \Psi_{d_0^*}(P_0)$ when there is a margin around zero, i.e. $\Pr_0(0 < |\bar{Q}_{b,0}(W)| \leq t) = 0$ for some $t > 0$. In fact, we have that $\Psi_{d_n}(P_0) - \Psi_{d_0^*}(P_0) = 0$ with probability approaching 1 in this case. A margin will exist for discrete W, and the NPMLE of $\bar{Q}_{b,0}$ will converge in supremum norm. If W is high dimensional then the NPMLE will often not be well-defined for reasonable sample sizes due to the curse of dimensionality and will still not yield a well-behaved estimator when well-defined. In such cases we might hope that smoothing techniques will give an estimate that converges in supremum norm.

Theorem 7 thus shows that $\Psi_{d_j}(P_0) - \Psi_{d_0^*}(P_0) = o_{P_0}(j^{-1/2})$ if $\tilde{\sigma}_{0,j}$ is estimated well in the sense of (13) and the distribution of $|\bar{Q}_{b,0}(W)|$ and our estimates of $\bar{Q}_{b,0}$ satisfy reasonable conditions. In this case an application of Lemma 5 shows that Condition C5) is satisfied.

We note that one does not have to use a plug-in estimator for the blip function to estimate the mean outcome under the optimal rule. One could also use one of the weighted classification approaches, sometimes known as outcome weighted learning (OWL), receintly discussed in the literature to estimate the optimal rule (Qian and Murphy, 2011; Zhao et al., 2012; Zhang et al., 2012b; Rubin and van der Laan, 2012). In some cases we expect these approaches to give better estimates of the optimal rule than methods which estimate the conditional outcomes, so using them may make Condition C5) more plausible. In Luedtke and van der Laan (2014) we describe an ensemble learner that can combine estimators from both the Q-learning and weighted classification frameworks.

Simulation	(n, ℓ_n)
D-E	$(1000, 100)$, $(4000, 100)$
C-NE, C-E, TTP-E	$(250, 25)$, $(1000, 25)$, $(4000, 100)$

Table 1: Primary combinations of sample size (n) and initial chunk size (ℓ_n) considered in each simulation. Different choices of ℓ_n were considered for C-NE and C-E to explore the sensitivity of the estimator to the choice of ℓ_n.

8 Simulation methods

We ran four simulations. Simulation D-E is a point treatment case, where the treatment may rely on a single categorical covariate W. Simulations C-NE and C-E are two different point treatment simulations where the treatment may rely on a single continuous covariate W. Simulation C-NE uses a non-exceptional law, while simulation C-E uses an exceptional law. Simulation TTP-E gives simulation results for a modification of the two time point treatment simulation presented in van der Laan and Luedtke (2014b), where the data generating distribution has been modified so that the second time point treatment has no effect on the outcome. This simulation uses the extension to multiple time point treatments given in Appendix B.

Each simulation setting was run over 2000 Monte Carlo draws to evaluate the performance of our new martingale-based method and a classical (and for exceptional laws incorrect) one-step estimator with Wald-type confidence intervals. Table 1 shows the combinations of sample size (n) and initial chunk size (ℓ_n) considered for each estimator. All simulations were run in `R` (R Core Team, 2014).

8.1 Simulation D-E: discrete W

Data

This simulation uses a discrete baseline covariate W with four levels, a dichotomous treatment A, and a binary outcome Y. The data is generated by drawing i.i.d. samples as follows:

$$\begin{aligned} W &\sim \text{Uniform}\{0, 1, 2, 3\} \\ A|W &\sim \text{Binomial}(0.5 + 0.1W) \\ Y|A, W &\sim \text{Binomial}\left(0.4 + 0.2I(A = 1, W = 0)\right). \end{aligned}$$

The above is an exceptional law because $\bar{Q}_{b,0}(w) = 0$ for $w \neq 0$. The optimal value is 0.45.

Estimation methods

For each $j = \ell_n + 1, ..., n$, we used the nonparametric maximum likelihood estimator (NPMLE) generated by the first $j - 1$ samples to estimate P_0 and the corresponding plug-in estimators to estimate all of the needed features of the likelihood, including the optimal rule. We used the sample standard deviation of $\tilde{D}_{ni}(O_1), ..., \tilde{D}_{ni}(O_{j-1})$ to estimate $\tilde{\sigma}_{0i}$.

8.2 Simulations C-NE and C-E: continuous univariate W

Data

This simulation uses a single continuous baseline covariate W and dichotomous treatment A which are sampled as follows:

$$W \sim \text{Uniform}(-1, 1)$$
$$A|W \sim \text{Binomial}(0.5 + 0.1W)$$

We consider two distributions for the binary outcome Y. The first distribution (C-NE) is a non-exceptional law with $Y|A, W \sim \text{Binomial}(\bar{Q}_0^{\text{n-e}}(A, W))$, where

$$\bar{Q}_0^{\text{n-e}}(A, W) - \frac{3}{10} \triangleq \begin{cases} -W^3 + W^2 - \frac{1}{3}W + \frac{1}{27}, & \text{if } A = 1 \text{ and } W \geq 0 \\ \frac{3}{4}W^3 + W^2 - \frac{1}{3}W + \frac{1}{27}, & \text{if } A = 1 \text{ and } W < 0 \\ 0, & \text{if } A = 0. \end{cases}$$

The optimal value of approximately 0.388 was estimated using 10^8 Monte Carlo draws. The second distribution (C-E) is an exceptional law with $Y|A, W \sim \text{Binomial}(\bar{Q}_0^{\text{e}}(A, W))$, where for $\tilde{W} \triangleq W + 5/6$ we define

$$\bar{Q}_0^{\text{e}}(A, W) - \frac{3}{10} \triangleq \begin{cases} -\tilde{W}^3 + \tilde{W}^2 - \frac{1}{3}\tilde{W} + \frac{1}{27}, & \text{if } A = 1 \text{ and } W < -1/2 \\ -W^3 + W^2 - \frac{1}{3}W + \frac{1}{27}, & \text{if } A = 1 \text{ and } W > 1/3 \\ 0, & \text{otherwise.} \end{cases}$$

The above distribution is an exceptional law because $\bar{Q}_0^{\text{e}}(1, w) - \bar{Q}_0^{\text{e}}(0, w) = 0$ whenever $w \in \left[-\frac{1}{2}, \frac{1}{3}\right]$. The optimal value of approximately 0.308 was estimated using 10^8 Monte Carlo draws.

Estimation methods

To show the flexibility of our estimation procedure with respect to estimators of the optimal rule, we estimated the blip functions using a Nadaraya-Watson estimator, where we behave as though g_0 is unknown when computing the kernel estimate. For the next simulation setting we use the ensemble learner from Luedtke and van der Laan (2014) that we suggest using in practice. Here we estimated

$$\bar{Q}_{b,n}^h(w) \triangleq \frac{\sum_{i=1}^n y_i a_i K\left(\frac{w - w_i}{h}\right)}{\sum_{i=1}^n a_i K\left(\frac{w - w_i}{h}\right)} - \frac{\sum_{i=1}^n y_i (1 - a_i) K\left(\frac{w - w_i}{h}\right)}{\sum_{i=1}^n (1 - a_i) K\left(\frac{w - w_i}{h}\right)},$$

where $K(u) \triangleq \frac{3}{4}(1 - u^2) I(|u| \leq 1)$ is the Epanechnikov kernel and h is the bandwidth. For a candidate blip function estimate $\bar{Q}_b$, define the loss

$$L_{\bar{Q}_0, g_0}(\bar{Q}_b)(o) \triangleq \left(\left[\frac{2a - 1}{g_0(a|w)} (y - \bar{Q}_0(a, w)) + \bar{Q}_0(1, w) - \bar{Q}_0(0, w) \right] - \bar{Q}_b(w) \right)^2.$$

To save computation time we behave as though $\bar{Q}_0$ and g_0 are known when using the above loss. We selected the bandwidth H_n using 10-fold cross-validation with the above loss function to select

from the candidates $h = (0.01, 0.02, ..., 0.20)$. We also behave as though $\bar{Q}_0$ and g_0 are known when estimating each $\tilde{D}_{ni}$, so that the function $\tilde{D}_{ni}$ only depends on $O_1, ..., O_{j-1}$ through the estimate of the optimal rule. This is mostly for convenience, as it saves on computation time and our estimate of the optimal rule d_0^* will still not stabilize, i.e. our optimal value estimators will still encounter the irregularity at exceptional laws. Note that g_0 is known in an RCT, and subtracting and adding $\bar{Q}_0$ in the definition of the loss function will only serve to stabilize the variance of our cross-validated risk estimate. In practice one could substitute an estimate of $\bar{Q}_0$ and expect similar results. We update our estimate $d_{n,j}$ and $\tilde{\sigma}_{0,n,j}$ using the method discussed in Section 6.1, where we let $S = \frac{n-\ell_n}{\ell_n}$.

To explore the sensitivity to the choice of ℓ_n we also considered (n, ℓ_n) pairs $(1000, 100)$ and $(4000, 400)$, where these pairs are only considered where explicitly noted. To explore the sensitivity of our estimators according to permutations in the indices of our data set, we ran our estimator twice on each Monte Carlo draw, with the indices of the observations permuted so that the online estimator sees the data in a different order.

8.3 Simulation TTP-E: two time point simulation

The data generating distribution used in this section was described in Section 8.1.2 of van der Laan and Luedtke (2014b), though here we modify the distribution slightly so that the second time point treatment has no effect on the outcome.

Data

The data is generated as follows:

$$
\begin{aligned}
&L_1(0), L_2(0) \overset{iid}{\sim} Unif(-1, 1) \\
&A(0)|L(0) \sim Bern(1/2) \\
&U_1, U_2|A(0), L(0) \overset{iid}{\sim} Unif(-1, 1) \\
&L_1(1)|A(0), L(0), U_1, U_2 \sim U_1(1.25A(0) + 0.25) \\
&L_2(1)|A(0), L(0), L_1(1), U_1, U_2 \sim U_2(1.25A(0) + 0.25) \\
&A(1)|A(0), \bar{L}(1) \sim Bern(1/2) \\
&Y|\bar{A}(1), \bar{L}(1) \sim Bern\left(0.4 + 0.0345A(0)\, b(L(0))\right),
\end{aligned}
$$

where $b\left(L(0)\right) \triangleq -0.8 - 3\left(\text{sgn}(L_1(0)) + L_1(0)\right) - L_2(0)^2$. The treatment decision at time point 0 may rely on $L(0)$, and the treatment at time point 1 may rely on $L(0)$, $A(0)$, and $L(1)$.

Estimation methods

As in the previous simulation, we assume that the treatment mechanism is known and supply the online estimator with correct estimates of the conditional mean outcome so that $\tilde{D}_{n,j}$ is random only through the estimate of d_0^* (see Appendix B for a definition of $\tilde{D}_{n,j}$ in the two time point case). Given a training sample $O_1, ..., O_j$, our estimator of d_0^* corresponds to using the full candidate library of weighted classification and blip-function based estimators listed in Table 2 of Luedtke and van der Laan (2014), with the weighted log loss function used to determine the convex combination

Figure 2: Coverage of 95% two-sided and one-sided (lower) confidence intervals. The online one-step estimator achieves (close to) nominal coverage for all of the two-sided confidence intervals, and attains better than nominal coverage for the one-sided confidence interval. The classical (non-online) one-step estimator only achieves near-nominal coverage for C-NE. Error bars indicate 95% confidence intervals to account for uncertainty from the finite number of Monte Carlo draws.

of candidates. As for C-E and C-NE, we update our estimate $d_{n,j}$ and $\tilde{\sigma}_{0,n,j}$ using the method discussed in Section 6.1 with $S = \frac{n-\ell_n}{\ell_n}$.

9 Simulation results

Figure 2 shows the coverage attained by the online and classical (non-online) one-step estimates of the optimal value. Note that the two-sided confidence intervals resulting from the online estimator (nearly) attains nominal coverage for all simulations considered. This is in contrast to the non-online estimator, which only (nearly) attains nominal coverage for the non-exceptional law in C-NE. The one-sided confidence intervals from the online one-step estimator attain proper coverage for all simulation settings, in agreement with Theorem 4. The one-sided confidence intervals from the non-online one-step estimates do not (nearly) achieve nominal coverage in any of the simulations considered. This occurs because the rule is estimated on the same observations on which the optimal value is estimated, and thus we expect that we need a very large sample size for the positive bias of the non-online one-step to become small. In van der Laan and Luedtke (2014b) we dealt with this finite sample positive bias at non-exceptional laws by proposing a cross-validated TMLE for the optimal value.

Figure 3 displays the squared bias and mean confidence interval length across the 2000 Monte

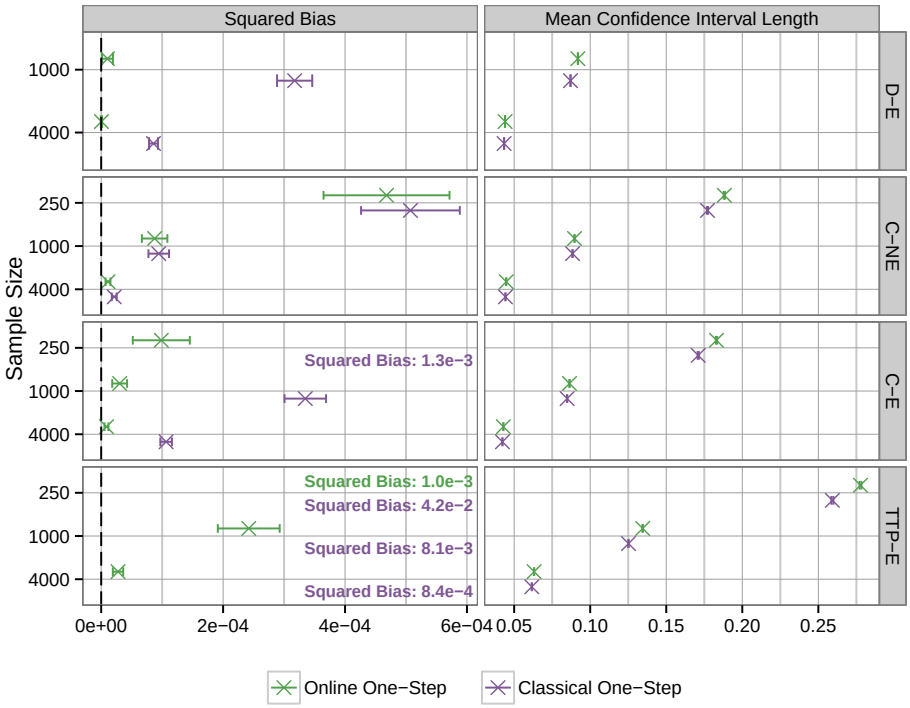

Figure 3: Squared bias and 95% two-sided confidence interval lengths for the online and classical (non-online) one-step estimators, where the mean is taken across 2000 Monte-Carlo draws. The online estimator has lower squared bias than the non-online estimator, while its mean confidence interval length is only slightly longer than those of the non-online estimator on average. Error bars indicate 95% confidence intervals to account for uncertainty from the finite number of Monte Carlo draws.

Carlo draws. The online estimator consistently has lower squared bias across all of our simulations. We also note that the online estimator was negatively biased in all of our simulations, whereas the non-online estimator was positively biased in all of our simulations. This is not surprising: Theorem 4 already implies that the online estimator will in general be negatively biased in finite samples, whereas the non-online estimator will in general be positively biased because d_n is chosen to (approximately) maximize the estimate of the value function.

Next we consider the sensitivity of our estimates to the initial ordering of the data set. Given that our data is i.i.d., we would hope that our estimator is not overly sensitive to the order of the data. Nonetheless, the online estimator we have proposed necessarily relies on an ordering of the data. In particular, data points with lower indices receive more more weight than those with higher indices. Figure 4 demonstrates how the optimal value estimates vary for C-E when the estimator is computed on two permutations of the same data set. We see that our point estimates are somewhat sensitive to the ordering of the data, but that this sensitivity decreases as sample size grows. We computed two-sided confidence intervals based on the two permuted data sets. We found that either both confidence intervals covered or neither confidence interval covered the true optimal value in 93.7%, 93.8%, and 93.1% of the Monte Carlo draws at sample sizes 250, 1000, and 4000, respectively. We performed the same analysis for the C-NE distribution and found similar results. For C-NE either both confidence intervals covered or neither confidence interval covered the true

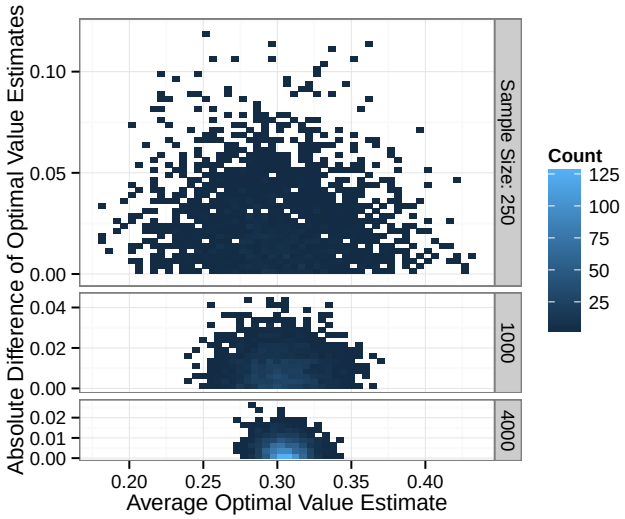

Figure 4: Comparison of optimal value estimates given two different permutations of a data set generated according to C-E, which results in two different estimates of the optimal value. The horizontal axis shows the average of the optimal value estimates across the two permutations, and the vertical axis shows the absolute difference between these two optimal value estimates. Squares represent number of observations (across 2000 Monte Carlo draws) which have a given average-absolute difference combination. The difference between these two estimates decreases as sample size grows.

optimal value in 90.6%, 93.0%, and 94.7% of the Monte Carlo draws at sample sizes 250, 1000, and 4000, respectively.

Different choices of ℓ_n did not greatly affect the coverage in C-E and C-NE. Increasing ℓ_n for C-E decreased the coverage by less than 1% for sample sizes 1000 and 4000. Increasing ℓ_n for C-NE increased the coverage by less than 1% for sample sizes 1000 and 4000. Mean confidence interval length increased predictably based on the increased value of $\sqrt{n-\ell_n}$: for $n = 1000$, increasing ℓ_n from 25 to 100 increased the confidence interval length by a multiplicative factor of $\sqrt{\frac{1000-25}{1000-100}} \approx 1.04$. Similarly, increasing ℓ_n from 100 to 400 increased the confidence interval length by a multiplicative factor of approximately 1.04 for $n = 4000$.

10 Discussion and future work

We have accomplished two primary tasks in this work. The first was to establish conditions under which we would expect that regular root-n rate inference is possible for the mean outcome under the optimal rule. In particular, we completely characterize the pathwise differentiability of the parameter giving the mean outcome under a single time point optimal rule. This characterization on the whole agrees with that implied in Robins and Rotnitzky (2014), but differs in a minor fringe case where the conditional variance of the outcome given covariates and treatment is zero. This fringe case may be relevant if everyone in a strata of baseline covariates is immune to a disease (regardless of treatment status) but are still included in the study because experts are unaware of this immunity *a priori*. In general, however, the two characterizations agree.

The remainder of our work shows that one can obtain an asymptotically unbiased estimate of and a

confidence interval for the optimal value under conditions. This estimator uses a slight modification of the online one-step estimator presented in van der Laan and Lendle (2014). The main condition for the validity of our confidence interval is that the value of one's estimate of the optimal rule converges to the optimal value at a faster than root-n rate, which we show is often a reasonable assumption by referencing the classification literature. The lower bound in our confidence interval is valid even if this condition does not hold. We also showed that one would expect our estimator to be RAL provided that the underlying distribution is a non-exceptional law. Our estimator will be asymptotically efficient among all RAL estimators of the optimal value. Further, the variance estimate used in our online two-sided confidence interval consistently estimates the variance of the influence curve.

We confirmed the validity of our approach using simulations. Our two-sided confidence intervals attained near-nominal coverage for all simulation settings considered, while our lower confidence intervals attained better than nominal coverage (were conservative) for all simulation settings considered. Our confidence intervals were of a comparable length to those attained by the non-online one-step estimator. The non-online one-step estimator only attained near-nominal coverage for the simulation which used a non-exceptional data generating distribution, as would be predicted by theory.

Our approach is designed to scale well to data sets which are too large to read into memory. We briefly outlined how one would implement our estimator using online regression and classification for the nuisance functions so that the data can be seen in chunks rather than all at once. We leave further consideration of the computational efficiency of our estimator to future work.

There is still work to be done in estimating confidence intervals for the optimal rule. While we have shown that the lower bound from our confidence interval maintains nominal coverage under very mild conditions, the upper bound requires the stronger condition that the optimal rule is estimated at a faster than root-n rate. We observed in our simulations that the non-online estimate of the optimal value had positive bias for all settings. Indeed this is to be expected if the optimal rule is chosen to maximize the estimated optimal value. This positive bias can easily be explained analytically under very mild assumptions. It may be worth replacing the upper bound UB_n in our confidence interval by something like $\max\{UB_n, \psi_n(d_n)\}$, where $\psi_n(d_n)$ is a non-online one-step estimate or TMLE of the optimal value. One might expect that the upper bound $\psi_n(d_n)$ will dominate the maximum precisely when the optimal rule is estimated poorly. We leave these considerations to future work.

Finally, we note that our estimation strategy is not limited to unrestricted classes of optimal rules. One could replace our unrestricted class with, e.g., a parametric working model for the blip function and still expect the same general results. This is due to the fact that the pathwise derivative of the function $P \mapsto E_{P_0}[Y_{d(P)}]$, which treats the P_0 in the expectation subscript as known, will typically be zero when $d(P)$ is an optimal rule in some class and does not fall on the boundary of that class (with respect to some metric). Such a result does not rely on $d(P)$ being a unique optimal rule. When the pathwise derivative of $P \mapsto E_{P_0}[Y_{d(P)}]$ is zero, one can often prove something like Theorem 7, which shows that the value of the estimated rule converges to the optimal value at a faster than root-n rate under conditions.

Here we considered the problem of developing a confidence interval for the value of an unknown

optimal DTR. We note that this is complementary to the problem of getting confidence bounds on (the parameters of) the blip functions. As we have shown here, confidence intervals for the optimal value can be developed without directly developing confidence bounds on the blip functions. If one uses a parametric working model for the blip functions, then they could use our proposed approach to estimate the optimal value and obtain confidence bounds on the blip functions using other methods previously proposed in the literature. In a larger (possibly nonparametric) model, estimating confidence bounds on the blip functions becomes a difficult problem, while analyzing our proposed optimal value estimate provides an interpretable and statistically valid approach to gauging the effect of implementing the optimal individualized treatment regime in the population.

Acknowledgement

This research was supported by NIH grant R01 AI074345-06. AL was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. The authors would like to thank Sam Lendle for suggesting the analysis in Figure 4.

References

K B Athreya and S N Lahiri. *Measure theory and probability theory*. Springer, New York Berlin Heidelberg, 2006.

J Y Audibert and A B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

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

P Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9.

B Chakraborty, E B Laber, and Y-Q Zhao. Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials*, 11(4):408–417, 2014.

Bibhas Chakraborty and E E Moodie. *Statistical Methods for Dynamic Treatment Regimes*. Springer, Berlin Heidelberg New York, 2013.

J Chen. Notes on the bias-variance trade-off phenomenon. In *A Festschrift for Herman Rubin*, pages 207–217. Institute of Mathematical Statistics, 2004.

D A Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.

P Gaenssler, J Strobel, and W Stute. On central limit theorems for martingale triangular arrays. *Acta Mathematica Hungarica*, 31(3):205–216, 1978.

Yair Goldberg, Rui Song, Donglin Zeng, Michael R Kosorok, and Others. Comment on Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8: 1290–1300, 2014.

K Hirano and J R Porter. Impossibility results for nondifferentiable functionals. *Econometrica*, 80 (4):1769–1790, 2012.

E Laber and S Murphy. Adaptive confidence intervals for the test error in classification. *J. Am. Stat. Assoc.*, 106:904–913, 2011.

E B Laber, D J Lizotte, M Qian, W E Pelham, and S A Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225, 2014a.

Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, Susan A Murphy, and Others. Rejoinder of Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8:1312–1321, 2014b.

J Langford, L Li, and T Zhang. Sparse online learning via truncated gradient. In *Advances in Neural Information Processing Systems*, pages 905–912, 2009.

Richard C Liu and Lawrence D Brown. Nonexistence of informative unbiased estimators in singular problems. *The Annals of Statistics*, pages 1–13, 1993.

A R Luedtke and M J van der Laan. Super-learning of an optimal dynamic treatment rule. Technical Report available at http://www.bepress.com/ucbbiostat/, Division of Biostatistics, University of California, Berkeley, under review at JCI, 2014.

J Luts, T Broderick, and M P Wand. Real-time semiparametric regression. *Journal of Computational and Graphical Statistics*, 23(3):589–615, 2014.

M Qian and S Murphy. Performance guarantees for individualized treatment rules. *Ann. Stat.*, 39: 1180–1210, 2011.

R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.r-project.org/`.

J M Robins. Optimal structural nested models for optimal sequential decisions. In D Y Lin and Heagerty P, editors, *Proceedings of the Second Seattle Symposium in Biostatistics*, volume 179, pages 189–326, 2004.

J M Robins and A Rotnitzky. Discussion of "Dynamic treatment regimes: Technical challenges and applications". *Electronic Journal of Statistics*, 8(1):1273–1289, 2014.

D B Rubin and M J van der Laan. Statistical issues and limitations in personalized medicine research with clinical trials. *International Journal of Biostatistics*, 8:Issue 1, Article 18, 2012.

W L Steiger. A best possible Kolmogoroff-type inequality for martingales and a characteristic property. *The Annals of Mathematical Statistics*, pages 764–769, 1969.

A B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32 (1):135–166, 2004.

M J van der Laan and S D Lendle. Online Targeted Learning. (available at http://www.bepress.com/ucbbiostat/), 2014.

M J van der Laan and A R Luedtke. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. Technical Report 329, UC Berkeley, 2014a.

M J van der Laan and A R Luedtke. Targeted learning of the mean outcome under an optimal dynamic treatment rule. Technical Report available at http://www.bepress.com/ucbbiostat/, Division of Biostatistics, University of California, Berkeley, in press at JCI, 2014b.

A W van der Vaart and J A Wellner. *Weak convergence and empirical processes*. Springer, Berlin Heidelberg New York, 1996.

Aad van der Vaart and Jon A Wellner. Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In *High dimensional probability II*, pages 115–133. Springer, 2000.

B Zhang, A Tsiatis, M Davidian, M Zhang, and E Laber. A robust method for estimating optimal treatment regimes. *Biometrics*, 68:1010–1018, 2012a.

B Zhang, A A Tsiatis, M Davidian, M Zhang, and E Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 68(1):103–114, 2012b.

T Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the 21st International Conference on Machine Learning*, page 116. ACM, 2004.

Y Zhao, D Zeng, A Rush, and M Kosorok. Estimating individual treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107:1106–1118, 2012.

Appendix

A Proofs

A.1 Proofs of results from Section 3

Proof of Theorem 1. Let $d'(P)$ represent the function $w \mapsto I(\bar{Q}_b(P)(w) > 0)$. For any P, let $\Psi(P) \triangleq E_P E_P[Y|A = d(P)(W), W]$. Note that

$$\Psi(P) - E_P[Y_0] = E_P\left[d^*(P)(W)\bar{Q}_b(P)(W)\right] = E_P\left[d'(P)(W)\bar{Q}_b(P)(W)\right],$$

where we used the fact that $d^*(P)(w) = d'(P)(w)$ on the set where $\bar{Q}_b(P)(w) \neq 0$. Let the fluctuation submodel $\{P_\epsilon : \epsilon\}$ through P_0 be as defined in Section 3 of the main text, where we note that $P_0 = P_{\epsilon=0}$. Telescoping shows that, for fixed ϵ,

$$\begin{aligned}\Psi(P_\epsilon) - \Psi(P_0) =& E_{P_\epsilon}\left[\left(I(\bar{Q}_{b,\epsilon} > 0) - I(\bar{Q}_{b,0} > 0)\right)\bar{Q}_{b,\epsilon}\right] \\ &+ \Psi_{d'_0}(P_\epsilon) - \Psi_{d'_0}(P_0). \qquad \text{(A.1)}\end{aligned}$$

It is well known that $\Psi_d(P) \triangleq E_P E_P[Y|A = d(W), W]$ is pathwise differentiable for fixed d. Thus dividing the second line above by ϵ and taking the limit as $\epsilon \to 0$ yields the pathwise derivative that treats the rule d_0' as known. For a given S_Y, the fluctuated $\bar{Q}_{b,0}$ at $w \in \mathcal{W}$ is given by

$$\begin{aligned}\bar{Q}_{b,\epsilon}(w) &\triangleq \int y \left(dQ_{Y,\epsilon}(y|A=1, W=w) - dQ_{Y,\epsilon}(y|A=0, W=w)\right) \\ &= \bar{Q}_{b,0}(w) + \epsilon \left(E_0\left[Y S_Y(Y|1,W)|A=1, W=w\right] - E_0\left[Y S_Y(Y|0,W)|A=0, W=w\right]\right) \\ &\triangleq \bar{Q}_{b,0}(w) + \epsilon h(w), \end{aligned} \tag{A.2}$$

where we note that $\sup_w |h(w)| < \infty$ because Y and S_Y are uniformly bounded.

Pathwise differentiable if (3).
Suppose (3). Let $B_1 \triangleq \{w : \bar{Q}_{b,0}(w) = 0\}$ and $B_2 \triangleq \{w : \bar{Q}_{b,0}(w) = 0, \max_a \sigma_0(a,w) = 0\}$. Noting that $B_2 \subseteq B_1$ shows

$$\begin{aligned} E_{P_\epsilon}\left[\left(I(\bar{Q}_{b,\epsilon} > 0) - I(\bar{Q}_{b,0} > 0)\right)\bar{Q}_{b,\epsilon}\right] &= \int_{\mathcal{W}\backslash B_1} \left(I(\bar{Q}_{b,\epsilon} > 0) - I(\bar{Q}_{b,0} > 0)\right)\bar{Q}_{b,\epsilon} dQ_{W,\epsilon} \\ &\quad + \int_{B_1\backslash B_2} \left(I(\bar{Q}_{b,\epsilon} > 0) - I(\bar{Q}_{b,0} > 0)\right)\bar{Q}_{b,\epsilon} dQ_{W,\epsilon} \\ &\quad + \int_{B_2} \left(I(\bar{Q}_{b,\epsilon} > 0) - I(\bar{Q}_{b,0} > 0)\right)\bar{Q}_{b,\epsilon} dQ_{W,\epsilon}. \end{aligned} \tag{A.3}$$

Because $\bar{Q}_{b,0} \neq 0$ on $\mathcal{W}\backslash B_2$, the first term above is $o(|\epsilon|)$ by a slight generalization of Lemma 2 in van der Laan and Luedtke (2014a) to finite measures (since $\Pr_0(\mathcal{W}\backslash B_2)$ may be less than 1). The second term is zero because $\Pr_0(B_1\backslash B_2) = 0$ by (3). Let $f(a,w) \triangleq E_0\left[Y S_Y(Y|1,W)|A=1, W=w\right]$. For the third term, note that, for $(a,w) \in \{0,1\} \times B_2$,

$$\begin{aligned} &\int_{B_2} \left(I(\bar{Q}_{b,\epsilon} > 0) - I(\bar{Q}_{b,0} > 0)\right)\bar{Q}_{b,\epsilon} dQ_{W,\epsilon} \\ &\quad = \epsilon \int_{B_2} \left(I(\bar{Q}_{b,\epsilon} > 0) - I(\bar{Q}_{b,0} > 0)\right)\left(f(1,w) - f(0,w)\right) dQ_{W,\epsilon} \end{aligned}$$

Note that $f(a,w) = Cov_{P_0}(Y, S_Y(Y|A,W)|A=a, W=w)$ for $a = 0, 1$ because $E[S_Y|A,W] = 0$, and thus $f(a,w) = 0$ for $(a,w) \in \{0,1\} \times B_2$ since Y has conditional variance 0 given $A = a$ and $W = w$. This shows that the third term in (A.3) is exactly zero. Hence,

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} E_{P_\epsilon}\left[\left(I(\bar{Q}_{b,\epsilon} > 0) - I(\bar{Q}_{b,0} > 0)\right)\bar{Q}_{b,\epsilon}\right] = 0.$$

Thus Ψ has canonical gradient $D(d_0', P_0)$, i.e. the same canonical gradient as the parameter $\Psi_{d_0'}$. Recall that

$$D(d,P)(O) = \frac{I(A = d(W))}{g(A|W)}(Y - \bar{Q}(A,W)) + \bar{Q}(d(W),W) - \Psi_d(P).$$

If (3), then either i) $Y = \bar{Q}(A,W)$ or ii) $d_0^* = d_0'$ with P_0 probability 1. Thus $D(d_0^*, P_0) = D(d_0', P_0)$ almost surely if (3) holds. It follows that Ψ has canonical gradient $D(d_0^*, P_0)$.

Not pathwise differentiable if not (3).
We wish to construct a submodel so that (4) holds. Let $S_W(w) = 0$ for all w. Without loss of generality, suppose that

$$P_0\left(\bar{Q}_{b,0}(W) = 0, \sigma_0(1, W) > 0\right) > 0. \tag{A.4}$$

Let

$$R(w) \triangleq \frac{\Pr_0(Y \leq \bar{Q}_0(1,W)|A=1,W=w)}{\Pr_0(Y > \bar{Q}_0(1,W)|A=1,W=w)},$$

where we let $R(w) = \infty$ when $\Pr_0(Y > \bar{Q}_0(1,W)|A=1,W=w) = 0$. Define S_Y as follows:

$$S_Y(y|a,w) \triangleq \begin{cases} \min\{1, R(w)\}, & \text{if } a = 1 \text{ and } y > \bar{Q}_0(1,w) \\ -\min\{1, 1/R(w)\}, & \text{if } a = 1 \text{ and } y \leq \bar{Q}_0(1,w) \\ 0, & \text{if } a = 0. \end{cases}$$

Above we let $\min\{1, 1/R(W)\} = 0$ when $R(W) = \infty$ and $\min\{1, 1/R(W)\} = 1$ when $R(W) = 0$. Note that $\sup_{w,a,y} |S_Y(y|a,w)| \leq 1$ and $E[S_Y|A=a, W=w] = 0$ for all a, w. We define B_+ and B_- as follows:

$$B_+ \triangleq B_0 \cap \{w : h(w) > 0\}$$
$$B_- \triangleq B_0 \cap \{w : h(w) < 0\},$$

where h is defined in (A.2). By (A.4), $\Pr_0(\bar{Q}_{b,0}(W) = 0, 0 < R(W) < \infty) > 0$, and hence $\Pr_0(B_+) > 0$ and $\Pr_0(B_-) > 0$. Let

$$m(w) \triangleq (I(\bar{Q}_{b,\epsilon}(w) > 0) - I(\bar{Q}_{b,0}(w) > 0))\bar{Q}_{b,\epsilon}(w).$$

The first term in (A.1) yields the following limit from above:

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \int_{\mathcal{W}} m(w) dQ_{W,0}(w) \\
&= \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \int_{B_+} m(w) dQ_{W,0}(w) + \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \int_{B_-} m(w) dQ_{W,0}(w) + \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \int_{\mathcal{W}\setminus(B_+ \cup B_-)} m(w) dQ_{W,0}(w) \\
&= \lim_{\epsilon \downarrow 0} \int_{B_+} I(\epsilon h(w) > 0) h(w) dQ_{W,0}(w) + \lim_{\epsilon \downarrow 0} \int_{B_-} I(\epsilon h(w) > 0) h(w) dQ_{W,0}(w) \\
&\quad + \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \int_{\mathcal{W}\setminus(B_+ \cup B_-)} m(w) dQ_{W,0}(w) \\
&= \int_{B_+} h(w) dQ_{W,0}(w) \\
&> 0,
\end{aligned} \tag{A.5}$$

where the integral over B_- is equal to zero because the indicator in m is 0 for all $\epsilon > 0$ and the integral over $\mathcal{W}\setminus(B_+ \cup B_-)$ is $o(|\epsilon|)$ because

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \int_{\mathcal{W}\setminus(B_+ \cup B_-)} m(w) dQ_{W,0}(w) \\
&= \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \int_{\mathcal{W}\setminus B_0} m(w) dQ_{W,0}(w) + \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \int_{\{w : h(w) = 0\} \cap B_0} m(w) dQ_{W,0}(w) = 0,
\end{aligned}$$

where we used that the first term is 0 by a slight generalization of Lemma 2 in van der Laan and Luedtke (2014a) to finite measures and the second term is 0 because $\bar{Q}_{b,\epsilon} = 0$ on $\{w : h(w) = 0\} \cap B_0$. The inequality in (A.5) is strict because $\Pr_0(B_+) > 0$ and $h > 0$ on B_+. Similarly,

$$\lim_{\epsilon \uparrow 0} \frac{1}{\epsilon} \int m(w) dQ_{W,0}(w) = \int_{B_-} h(w) dQ_{W,0}(w) < 0.$$

Contrasting the above with (A.5) shows that there exists a path about P_0 which results in a fluctuation h for which the limit of the first term in (A.1) divided by ϵ does not exist as $\epsilon \to 0$. But then Ψ cannot be pathwise differentiable: one of the limits in the sum on the right-hand side of (A.1) exists, so the limit on the left-hand side cannot exist. Specifically, suppose c_n has a limit as $n \to \infty$ and $a_n = b_n + c_n$. If b_n does not have a limit, then a_n does not have a limit, since a_n having a limit implies that $b_n = a_n - c_n$ has a limit, contradiction. □

A.2 Proofs of results from Section 5

Proof of Theorem 2. We have that

$$\begin{aligned}
&\Gamma_n \left(\hat{\Psi}(P_n) - \Psi(P_0) \right) \\
&= \frac{1}{n - \ell_n} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}_{n,j}^{-1} \left(\tilde{D}_{n,j}(O_j) - \Psi(P_0) \right) && \text{(A.6)} \\
&= \frac{1}{n - \ell_n} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}_{n,j}^{-1} \left(\left[\tilde{D}_{n,j}(O_j) - \Psi_{d_{n,j}}(P_0) \right] + \left[\Psi_{d_{n,j}}(P_0) - \Psi(P_0) \right] \right) && \text{(A.7)} \\
&= \frac{1}{n - \ell_n} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}_{n,j}^{-1} \left(\tilde{D}_{n,j}(O_j) - \Psi_{d_{n,j}}(P_0) \right) + o_{P_0}(n^{-1/2}) && \text{(A.8)} \\
&= \frac{1}{n - \ell_n} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}_{n,j}^{-1} \left(\tilde{D}_{n,j}(O_j) - E_0 \left[\tilde{D}_{n,j}(O_j) | O_1, ..., O_{j-1} \right] \right) + R_{1n} + o_{P_0}(n^{-1/2}) && \text{(A.9)} \\
&= \frac{1}{n - \ell_n} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}_{n,j}^{-1} \left(\tilde{D}_{n,j}(O_j) - E_0 \left[\tilde{D}_{n,j}(O_j) | O_1, ..., O_{j-1} \right] \right) + o_{P_0}(n^{-1/2}). && \text{(A.10)}
\end{aligned}$$

Above (A.6) is a result of moving the $\Psi(P_0)$ into the summation in the defintion of Γ_n, (A.7) adds zero to the line above, (A.8) follows by C5), (A.9) is a consequence of the fact that $\Psi_d(P_0) = P_0 \tilde{D}(\bar{Q}, g, d) - E_0[(1 - \frac{g_0(d(W)|W)}{g(d(W)|W)})(\bar{Q}(d(W), W) - \bar{Q}_0(d(W), W))]$ for any fixed $\bar{Q}$, g, and d, and (A.10) follows by C4).

For $j = 1, ..., n - \ell_n$, let

$$M_{n,j} \triangleq \frac{1}{\sqrt{n - \ell_n}} \frac{\left(\tilde{D}(d_{n,j+\ell_n})(O_{j+\ell_n}) - E_0 \left[\tilde{D}(d_{n,j+\ell_n})(O_{j+\ell_n}) | O_1, ..., O_{j+\ell_n} \right] \right)}{\tilde{\sigma}_{n,j+\ell_n}}.$$

Note that, for each n, $\{M_{n,j} : j = 1, ..., n - \ell_n\}$ is a discrete-time martingale with respect to the filtration $\mathcal{F}_j$, where each $\mathcal{F}_j$ is the sigma-field generated by $O_1, ..., O_{j+\ell_n}$. In particular,

we have that, for all $j \geq 1$, $E_0[M_{n,j}|\mathcal{F}_{j-1}] = 0$. We also have that $\sum_{j=1}^{n-\ell_n} E_0[M_{n,j}^2|\mathcal{F}_{j-1}] = \frac{1}{n-\ell_n}\sum_{j=1}^{n-\ell_n} \frac{\tilde{\sigma}_{0,n,j+\ell_n}^2}{\tilde{\sigma}_{n,j+\ell_n}^2} \to 1$ by C3). Because the conditional Lindeberg condition in C2) holds, the martingale CLT for triangular arrays (see, e.g., Theorem 2 in Gaenssler et al., 1978) shows that

$$\sum_{j=1}^{n-\ell_n} M_{n,j} \rightsquigarrow N(0,1). \tag{A.11}$$

Plugging this into (A.10) gives that

$$\Gamma_n \sqrt{n-\ell_n}\left(\hat{\Psi}(P_n) - \Psi(P_0)\right) \rightsquigarrow N(0,1).$$

The asymptotically valid $1-\alpha$ confidence interval is now constructed in the usual way. □

Proof of Corollary 3. In this proof we use "$\lesssim$" to denote less than or equal to up to a positive multiplicative constant. Let $\mathcal{F}_j$ represent the sigma-field generated by $O_1, ..., O_j$. Let $\tilde{D}_0 \triangleq \tilde{D}(d_0, \bar{Q}_0, g_0)$ and $s_0^2 \triangleq Var_{P_0}(\tilde{D}(d_0, \bar{Q}_0, g_0)(O))$. The proof can be broken into four parts, which show that: (1) $\tilde{D}_{n,j}$ approximates $\tilde{D}_0$ in mean-square; (2) $\Gamma_n^{-1} \to s_0$ in probability; (3) $\Gamma_n(\hat{\Psi}(P_n) - \Psi(P_0))$ behaves like an empirical mean of the normalized efficient influence curve; (4) $\hat{\Psi}(P_n)$ is RAL and efficient.

Part 1: $\tilde{D}_{n,j}$ approximates $\tilde{D}_0$. Note that

$$\begin{aligned}
&\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(\tilde{D}_{n,j} - \tilde{D}_0\right)^2 \middle| \mathcal{F}_{j-1}\right] \\
&\quad \leq \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(\tilde{D}(d_{n,j}, \bar{Q}_{n,j}, g_{n,j}) - \tilde{D}(d_0, \bar{Q}_{n,j}, g_{n,j})\right)^2 \middle| \mathcal{F}_{j-1}\right] \\
&\qquad + \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(\tilde{D}(d_0, \bar{Q}_{n,j}, g_{n,j}) - \tilde{D}(d_0, \bar{Q}_{n,j}, g_0)\right)^2 \middle| \mathcal{F}_{j-1}\right] \\
&\qquad + \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(\tilde{D}(d_0, \bar{Q}_{n,j}, g_0) - \tilde{D}(d_0, \bar{Q}_0, g_0)\right)^2 \middle| \mathcal{F}_{j-1}\right] \\
&\quad \lesssim \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(d_{n,j}(W) - d_0(W)\right)^2 \middle| \mathcal{F}_{j-1}\right] \\
&\qquad + \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(g_{n,j}(d(W)|W) - g_0(d(W)|W)\right)^2 \middle| \mathcal{F}_{j-1}\right] \\
&\qquad + \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} E_0\left[\left(\bar{Q}_{n,j}(d_0(W), W) - \bar{Q}_0(d_0(W), W)\right)^2 \middle| \mathcal{F}_{j-1}\right] \\
&\quad = o_{P_0}(1)
\end{aligned} \tag{A.12}$$

where the constant in the second inequality relies on the bounds on Y, $\bar{Q}_{n,j}$, g_0, and $g_{n,j}$.

Part 2: $\Gamma_n^{-1} \to s_0$ **in probability.** We have that

$$(\Gamma_n - s_0^{-1})^2 \leq \left(\frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}_{n,j}^{-1} s_0^{-1} |\tilde{\sigma}_{n,j} - s_0| \right)^2 \lesssim \left(\frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} |\tilde{\sigma}_{n,j} - s_0| \right)^2$$
$$\lesssim \frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} (\tilde{\sigma}_{n,j} - s_0)^2, \tag{A.13}$$

where the second inequality on the first line holds by the assumed bounds on $\tilde{\sigma}_{n,j}$ and the final inequality holds by Cauchy-Schwarz. Note that, for any positive real numbers x_1, x_2,

$$(x_1 - x_2)^2 \leq 2|x_1^2 - x_2^2|. \tag{A.14}$$

By the above and Condition C3') , we have that

$$\frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} (\tilde{\sigma}_{n,j} - \tilde{\sigma}_{0,n,j})^2 \lesssim \frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} \left| \tilde{\sigma}_{n,j}^2 - \tilde{\sigma}_{0,n,j}^2 \right| \lesssim \frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} \left| \frac{\tilde{\sigma}_{0,n,j}^2}{\tilde{\sigma}_{n,j}^2} - 1 \right| = o_{P_0}(1).$$

We also have that

$$\frac{1}{n-\ell_n} \sum_{j=1}^{n} (\tilde{\sigma}_{0,n,j} - s_0)^2 \leq \frac{2}{n-\ell_n} \sum_{j=1}^{n} |\tilde{\sigma}_{0,n,j}^2 - s_0^2|$$
$$= \frac{2}{n-\ell_n} \sum_{j=\ell_n+1}^{n} \left| E_0 \left[\tilde{D}_{n,j}^2 - \tilde{D}_0^2 | \mathcal{F}_{j-1} \right] + E_0 \left[\tilde{D}_{n,j} | \mathcal{F}_{j-1} \right]^2 - E_0 \left[\tilde{D}_0 | \mathcal{F}_{j-1} \right]^2 \right|$$
$$\lesssim \frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} E_0 \left[\left| \tilde{D}_{n,j} - \tilde{D}_0 \right| | \mathcal{F}_{j-1} \right] \lesssim \sqrt{\frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} E_0 \left[\left(\tilde{D}_{n,j} - \tilde{D}_0 \right)^2 | \mathcal{F}_{j-1} \right]},$$

where: the first inequality holds by (A.14); the equality holds by the definition of conditional variance; the second inequality holds by twice using that $x_1^2 - x_2^2 = (x_1 + x_2)(x_1 - x_2)$, the strong positivity assumption, and the bounds on Y and $\bar{Q}_{n,j}$; and the final inequality holds by the Cauchy-Schwarz inequality applied to the expectations and the concavity of $x \mapsto \sqrt{x}$. By (A.12), the upper bound above is $o_{P_0}(1)$. By the triangle inequality and the previous two indented equations,

$$\frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} (\tilde{\sigma}_{n,j} - s_0)^2 \leq \frac{1}{n-\ell_n} \sum_{j=\ell_n+1}^{n} \left[(\tilde{\sigma}_{n,j} - \tilde{\sigma}_{0,n,j})^2 + (\tilde{\sigma}_{0,n,j} - s_0)^2 \right] = o_{P_0}(1). \tag{A.15}$$

Plugging this into (A.13) shows that $\Gamma_n = s_0^{-1} + o_{P_0}(1)$. By the continuous mapping theorem, $\Gamma_n^{-1} = s_0 + o_{P_0}(1)$.

Part 3: $\Gamma_n(\hat{\Psi}(P_n) - \Psi(P_0))$ **behaves like an empirical mean.** For each $n > 1$ and $j = \ell_n + 1, ..., n$, define

$$M'_{n,j} \triangleq \frac{\tilde{D}_{n,j}(O_j) - E_0 \left[\tilde{D}_{n,j}(O) | \mathcal{F}_{j-1} \right]}{\tilde{\sigma}_{n,j}} - \frac{\tilde{D}_0(O_j) - E_0 \left[\tilde{D}_0(O) \,|\, \mathcal{F}_{j-1} \right]}{s_0}.$$

We first show that $\frac{1}{\sqrt{n-\ell_n}}\sum_{j=\ell_n+1}^{n} M'_{n,j} \to 0$ in probability. Note that

$$\begin{aligned}
V'_{n,j} \triangleq & Var_{P_0}\left(M'_{n,j}\middle|\,\mathcal{F}_{j-1}\right) = E_0\left[\left(\frac{\tilde{D}_{n,j}(O_j)}{\tilde{\sigma}_{n,j}} - \frac{\tilde{D}_0(O_j)}{s_0}\right)^2\middle|\,\mathcal{F}_{j-1}\right] \\
\leq & E_0\left[\left(\frac{\tilde{D}_{n,j}(O_j)}{\tilde{\sigma}_{n,j}} - \frac{\tilde{D}_0(O_j)}{\tilde{\sigma}_{n,j}}\right)^2\middle|\,\mathcal{F}_{j-1}\right] + E_0\left[\left(\frac{\tilde{D}_0(O_j)}{\tilde{\sigma}_{n,j}} - \frac{\tilde{D}_0(O_j)}{s_0}\right)^2\middle|\,\mathcal{F}_{j-1}\right] \\
\lesssim & E_0\left[\left(\tilde{D}_{n,j}(O_j) - \tilde{D}_0(O_j)\right)^2\middle|\,\mathcal{F}_{j-1}\right] + E_0\left[\left(\tilde{\sigma}_{n,j} - s_0\right)^2\middle|\,\mathcal{F}_{j-1}\right]
\end{aligned}$$

where the constants in the second inequality rely on the bounds on $g_{n,j}$, g_0, $\bar{Q}_{n,j}$, Y, $\tilde{\sigma}_{0,n,j}$, and s_0. By (A.12) and (A.15),

$$\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} V'_{n,j} = o_{P_0}(1). \tag{A.16}$$

Fix $\epsilon, \delta > 0$ and let $v_{\epsilon,\delta} \triangleq \frac{\epsilon^2}{\log(4/\delta)}$. We will show that there exists some N such that

$$\Pr_0\left(\frac{1}{\sqrt{n-\ell_n}}\sum_{j=\ell_n+1}^{n} M'_{n,j} \geq \epsilon\right) < \delta \text{ for all } n \geq N. \tag{A.17}$$

Note that

$$\begin{aligned}
\Pr_0\left(\frac{1}{\sqrt{n-\ell_n}}\sum_{j=\ell_n+1}^{n} M'_{n,j} \geq \epsilon\right) = & \Pr_0\left(\frac{1}{\sqrt{n-\ell_n}}\sum_{j=\ell_n+1}^{n} M'_{n,j} \geq \epsilon, \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} V'_{n,j} \leq v_{\epsilon,\delta}\right) \\
& + \Pr_0\left(\frac{1}{\sqrt{n-\ell_n}}\sum_{j=\ell_n+1}^{n} M'_{n,j} \geq \epsilon, \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} V'_{n,j} > v_{\epsilon,\delta}\right).
\end{aligned}$$

We will bound the terms on the right separately. By our bounding assumptions, there exists some $m^* \in (0,\infty)$ such that $\Pr_0(\sup_{j\leq n}|M_{n,j}| < m^*) = 1$. By Bernstein's inequality for martingale difference sequences with bounded increments (see, e.g, Steiger, 1969; Theorem 1.6 of Freedman, 1975), we have that

$$\begin{aligned}
& \Pr_0\left(\frac{1}{\sqrt{n-\ell_n}}\sum_{j=\ell_n+1}^{n} M'_{n,j} \geq \epsilon, \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} V'_{n,j} \leq v_{\epsilon,\delta}\right) \\
& \leq \Pr_0\left(\frac{1}{\sqrt{n-\ell_n}}\sum_{j=\ell_n+1}^{\tilde{n}} M'_{n,j} \geq \epsilon, \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{\tilde{n}} V'_{n,j} \leq v_{\epsilon,\delta} \text{ for some } \tilde{n} \in \{\ell_n+1,...,n\}\right) \\
& \leq \Pr_0\left(\sum_{j=\ell_n+1}^{\tilde{n}} \frac{M'_{n,j}}{m^*} \geq \frac{\epsilon\sqrt{n-\ell_n}}{m^*}, \sum_{j=\ell_n+1}^{\tilde{n}} \frac{V'_{n,j}}{(m^*)^2} \leq \frac{v_{\epsilon,\delta}(n-\ell_n)}{(m^*)^2} \text{ for some } \tilde{n} \in \{\ell_n+1,...,n\}\right) \\
& \leq \exp\left(-\frac{\epsilon^2\sqrt{n-\ell_n}}{2(m^*\epsilon + v_{\epsilon,\delta}\sqrt{n-\ell_n})}\right) \overset{n\to\infty}{\longrightarrow} \delta/4.
\end{aligned}$$

It follows that there exists some N_1 such that the upper bound above is less then or equal to $\delta/2$ for all $n \geq N_1$. We also have that

$$\Pr_0\left(\frac{1}{\sqrt{n-\ell_n}}\sum_{j=\ell_n+1}^{n} M'_{n,j} \geq \epsilon, \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} V'_{n,j} > v_{\epsilon,\delta}\right) \leq \Pr_0\left(\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} V'_{n,j} \geq v_{\epsilon,\delta}\right).$$

By (A.16), there exists some N_2 so that the upper bound above is no greater than $\delta/4$ for all $n \geq N_2$. Combining the previous two sets of inequalities shows that (A.17) is satisfied for $N \triangleq \max\{N_1, N_2\}$. Thus $\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} M'_{n,j} = o_{P_0}(\sqrt{n-\ell_n})$. Because $\ell_n = o(n)$, $\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} M'_{n,j} = o_{P_0}(n^{-1/2})$. Combining this with (A.10) shows that

$$\begin{aligned}
&\Gamma_n\left(\hat{\Psi}(P_n) - \Psi(P_0)\right) \\
&\quad= \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} \frac{\tilde{D}_{n,j}(O_j) - E_0[\tilde{D}_{n,j}(O)|\mathcal{F}_{j-1}]}{\tilde{\sigma}_{n,j}} + o_{P_0}(n^{-1/2}) \\
&\quad= s_0^{-1}\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n}\left(\tilde{D}_0(O_j) - E_0\left[\tilde{D}_0(O)\right]\right) + \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n} M'_{n,j} + o_{P_0}(n^{-1/2}) \\
&\quad= s_0^{-1}\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n}\left(\tilde{D}_0(O_j) - E_0\left[\tilde{D}_0(O)\right]\right) + o_{P_0}(n^{-1/2}) \\
&\quad= s_0^{-1}\frac{1}{n}\sum_{j=1}^{n}\left(\tilde{D}_0(O_j) - E_0\left[\tilde{D}_0(O)\right]\right) + o_{P_0}(n^{-1/2}),
\end{aligned}$$

where the final equality uses that $\ell_n = o(n)$ and that $\tilde{D}_0$ is bounded.

Part 4: $\hat{\Psi}(P_n)$ is RAL and efficient. Combining Parts 2 and 3 shows that

$$\begin{aligned}
\hat{\Psi}(P_n) - \Psi(P_0) = \Gamma_n^{-1}\Gamma_n\left(\hat{\Psi}(P_n) - \Psi(P_0)\right) &= (s_0 + o_{P_0}(1))\Gamma_n\left(\hat{\Psi}(P_n) - \Psi(P_0)\right) \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(\tilde{D}_0(O_j) - E_0\left[\tilde{D}_0(O)\right]\right) + o_{P_0}(n^{-1/2}).
\end{aligned}$$

Thus $\hat{\Psi}(P_n)$ is an asymptotically linear estimator of $\Psi(P_0)$ with influence curve $D(d_0, P_0) = \tilde{D}_0(O_j) - E_0\left[\tilde{D}_0(O)\right]$. If P_0 satisfies (3) so that $D(d_0, P_0) = D(d_0^*, P_0)$ almost surely, then Theorem 1 shows that $D(d_0^*, P_0)$ is the efficient influence curve of Ψ. By Proposition 1 of Section 3.3 in Bickel et al. (1993), it follows that (3) holds if and only if $\hat{\Psi}(P_n)$ is a RAL estimator and is asymptotically efficient among all RAL estimators. □

Proof of Theorem 4. The below is an abbreviated version of (A.6) through (A.10) and (A.11), with

an added inequality which holds because $R_{2n} \leq 0$:

$$\begin{aligned}
&\sqrt{n-\ell_n}\Gamma_n\left(\hat{\Psi}(P_n) - \Psi(P_0)\right) \\
&= \frac{1}{\sqrt{n-\ell_n}} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}_{n,j}^{-1}\left(\left[\tilde{D}_{n,j}(O_j) - \Psi_{d_{n,j}}(P_0)\right] + \left[\Psi_{d_{n,j}}(P_0) - \Psi(P_0)\right]\right) \\
&\leq \frac{1}{\sqrt{n-\ell_n}} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}_{n,j}^{-1}\left(\tilde{D}_{n,j}(O_j) - \Psi_{d_{n,j}}(P_0)\right) \\
&= \frac{1}{\sqrt{n-\ell_n}} \sum_{j=\ell_n+1}^{n} \tilde{\sigma}_{n,j}^{-1}\left(\tilde{D}_{n,j}(O_j) - E_0\left[\tilde{D}_{n,j}(O_j)|O_1,...,O_{j-1}\right]\right) + o_{P_0}(1) \\
&\rightsquigarrow N(0,1).
\end{aligned}$$

Thus,

$$\liminf_{n\to\infty} \Pr_0\left(\sqrt{n-\ell_n}\Gamma_n\left(\hat{\Psi}(P_n) - \Psi(P_0)\right) \leq z_{1-\alpha}\right) \geq 1-\alpha.$$

The first result follows by rearranging terms in the probability statement. The second result is an immediate corollary of Theorem 2. □

A.3 Proofs of results from Section 7

Proof of Lemma 5. By the almost sure representation theorem (see, e.g., Theorem 1.10.3 in Billingsley, 1999), there exists a probability space $(\Omega', \mathcal{F}', P')$ and a sequence of random variables $R'_n : \Omega' \to \mathbb{R}$ such that $n^\beta R'_n \stackrel{d}{=} n^\beta R_n$ and $n^\beta R'_n(\omega') \to 0$ for all $\omega' \in \Omega'$. Fix $\epsilon > 0$ and $\omega' \in \Omega'$. There exists some N that, for all $n \geq N$, $n^\beta |R'_n(\omega')| < \frac{(1-\beta)\epsilon}{2}$. Also note that

$$\frac{1}{n^{1-\beta}} \sum_{j=1}^{n} j^{-\beta} \leq \frac{1}{n^{1-\beta}} \int_1^n (j-1)^{-\beta} dj = \frac{1}{1-\beta}.$$

Hence, for all $n \geq N$,

$$\begin{aligned}
\frac{1}{n^{1-\beta}} \sum_{j=1}^{n} |R'_j(\omega')| &= \frac{1}{n^{1-\beta}} \sum_{j=1}^{N-1} |R'_j(\omega')| + \frac{1}{n^{1-\beta}} \sum_{j=N}^{n} \frac{1}{j^\beta} j^\beta |R'_j(\omega')| \\
&< \frac{1}{n^{1-\beta}} \sum_{j=1}^{N-1} |R'_j(\omega')| + \frac{(1-\beta)\epsilon}{2n^{1-\beta}} \sum_{j=N}^{n} \frac{1}{j^\beta} \\
&\leq \frac{1}{n^{1-\beta}} \sum_{j=1}^{N-1} |R'_j(\omega')| + \frac{\epsilon}{2}.
\end{aligned}$$

It follows that $\frac{1}{n^{1-\beta}} \sum_{j=1}^{n} |R'_j(\omega')| < \epsilon$ for all n large enough, and thus that $\lim_{n\to\infty} \frac{1}{n^{1-\beta}} \sum_{j=1}^{n} R'_j(\omega') = 0$. Noting that $\frac{1}{n^{1-\beta}} \sum_{j=1}^{n} R_j \stackrel{d}{=} \frac{1}{n^{1-\beta}} \sum_{j=1}^{n} R'_j(\omega')$ for all n, we have that $\frac{1}{n} \sum_{j=1}^{n} R_j = o_{P_0}(n^{-\beta})$. □

Proof of Theorem 6. Let $\tilde{\mathcal{D}}_1 \triangleq \{\tilde{D}(d, \bar{Q}, g) : d, \bar{Q}, g\}$, $\tilde{\mathcal{D}}_2 \triangleq \{\tilde{D}^2(d, \bar{Q}, g) : d, \bar{Q}, g\}$, and $j^* \triangleq \min\{j : \delta_j \leq \delta_0\}$. The class $\tilde{\mathcal{D}}_1$ is P_0 Glivenko-Cantelli (GC) by assumption, and $\tilde{\mathcal{D}}_2$ is GC by Theorem 2 of van der Vaart and Wellner (2000). For all $j \geq j^*$, we have that

$$\begin{aligned}\left|\tilde{\sigma}_j^2 - \tilde{\sigma}_{0,j}^2\right| \leq & \left| \frac{1}{j-1} \sum_{i=1}^{j-1} \tilde{D}_j^2(O_i) - E_0\left[\tilde{D}_j^2(O) \middle| O_1, ..., O_{j-1}\right] \right| \\ & + \left| \left(\frac{1}{j-1} \sum_{k=1}^{j-1} \tilde{D}_j(O_k) \right)^2 - E_0\left[\tilde{D}_j(O) \middle| O_1, ..., O_{j-1}\right]^2 \right|. \end{aligned} \tag{A.18}$$

The first term on the right converges to 0 in probability because $\tilde{\mathcal{D}}_2$ is GC. For the second term, the mean value theorem shows that

$$\begin{aligned}& \left(\frac{1}{j-1} \sum_{k=1}^{j-1} \tilde{D}_j(O_k) \right)^2 - E_0\left[\tilde{D}_j(O) \middle| O_1, ..., O_{j-1}\right]^2 \\ & = 2m_j \underbrace{\left(\frac{1}{j-1} \sum_{k=1}^{j-1} \tilde{D}_j(O_k) - E_0\left[\tilde{D}_j(O) \middle| O_1, ..., O_{j-1}\right] \right)}_{\triangleq \|P_j - P_0\|_{\tilde{\mathcal{D}}_1}},\end{aligned}$$

where m_j is an intermediate value between the two squared values on the first line. Using that $\tilde{\mathcal{D}}_1$ is a GC class, we have that m_j converges to $E_0[\tilde{D}_j(O)|O_1, ..., O_{j-1}]$ in probability and $\|P_j - P_0\|_{\tilde{\mathcal{D}}_1} = o_{P_0}(1)$. Thus the above is $o_{P_0}(1)$, and plugging this into (A.18) shows that $|\tilde{\sigma}_j^2 - \tilde{\sigma}_{0,j}^2| = o_{P_0}(1)$. The continuous mapping theorem shows that (13) is also satisfied. Combining this with Lemma 5 with $\beta = 0$ shows that Condition C3) is satisfied. □

Proof of Theorem 7. In this proof we will omit the dependence of d_0^*, d_n, $\bar{Q}_{b,0}$, and $\bar{Q}_{b,n}$ on W in the notation. Suppose that $\left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{2,P_0} = o_{P_0}(1)$. This part of the proof mimics the proof of Lemma 5.2 in Audibert and Tsybakov (2007). For any $t > 0$,

$$\begin{aligned}|\Psi_{d_n}(P_0) - \Psi_{d_0^*}(P_0)| =& E_0[|\bar{Q}_{b,0}| I(\bar{Q}_{b,0} \neq \bar{Q}_{b,n})] \\ =& E_0[|\bar{Q}_{b,0}| I(\bar{Q}_{b,0} \neq \bar{Q}_{b,n}) I(0 < |\bar{Q}_{b,0}| \leq t)] \\ & + E_0[|\bar{Q}_{b,0}| I(\bar{Q}_{b,0} \neq \bar{Q}_{b,n}) I(|\bar{Q}_{b,0}| > t)] \\ \leq & E_0[|\bar{Q}_{b,n} - \bar{Q}_{b,0}| I(0 < |\bar{Q}_{b,0}| \leq t)] \\ & + E_0[|\bar{Q}_{b,n} - \bar{Q}_{b,0}| I(|\bar{Q}_{b,n} - \bar{Q}_{b,0}| > t)] \\ & \leq \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{2,P_0} \Pr_0(0 < |\bar{Q}_{b,0}| \leq t)^{1/2} + \frac{\left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{2,P_0}^2}{t} \\ & \leq \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{2,P_0} C_0^{1/2} t^{\alpha/2} + \frac{\left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{2,P_0}^2}{t},\end{aligned}$$

where the first inequality holds because $\bar{Q}_{b,0} \neq \bar{Q}_{b,n}$ implies that $|\bar{Q}_{b,n} - \bar{Q}_{b,0}| > |\bar{Q}_{b,0}|$, the second inequality holds by the Cauchy-Schwarz and Markov inequalities, and the third inequality holds

by (16). The first result follows by optimizing over t to find that the upper bound is minimized when $t = C \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{2,P_0}^{2(1+\alpha)/(2+\alpha)}$ for a constant C which depends on C_0 and α.

Now suppose that $\left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0} = o_{P_0}(1)$. Note that

$$\begin{aligned}
\left|\Psi_{d_n}(P_0) - \Psi_{d_0^*}(P_0)\right| &= E_0\left|I(d_n \neq d_0^*)\bar{Q}_{b,0}\right| \\
&\leq E_0\left[I(0 < |\bar{Q}_{b,0}| \leq |\bar{Q}_{b,n} - \bar{Q}_{b,0}|)|\bar{Q}_{b,0}|\right] \\
&\leq E_0\left[I\left(0 < |\bar{Q}_{b,0}| \leq \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0}\right)|\bar{Q}_{b,0}|\right] \\
&\leq \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0} \Pr_0\left(0 < |\bar{Q}_{b,0}| \leq \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0}\right).
\end{aligned}$$

By (16), $|\Psi_{d_n}(P_0) - \Psi_{d_0^*}(P_0)| \lesssim \left\|\bar{Q}_{b,n} - \bar{Q}_{b,0}\right\|_{\infty,P_0}^{1+\alpha}$. □

B Multiple time point case

For simplicity we will consider a two time point treatment with baseline covariates $L(0)$, a treatment $A(0)$, intermediate covariate $L(1)$, a treatment $A(1)$, and an outcome Y which comes after all treatments and covariates. The extension to the more general multiple time point case follows the same general arguments. We use the notation $\bar{A}(1) = (A(0), A(1))$ and $\bar{L}(1) = (L(0), L(1))$. The presentation in this section parallels that given in van der Laan and Luedtke (2014b), and we refer to the reader to that work for a more detailed description of the two time point problem. For the sake of simplicity we do not consider censoring, though censoring can easily be incorporated using the techniques in the referenced paper. The notation is similar in spirit to that of the rest of the paper, though there is some notational overload (e.g. d now used to represent a two time point rule, $\Psi(P_0)$ now the mean outcome under a two time point treatment).

A dynamic rule $d = (d_{A(0)}, d_{A(1)})$ consists of two rules, one for each time point. The first time point rule $d_{A(0)}$ may be a function of $L(0)$, while the second time point rule $d_{A(1)}$ may rely on $L(0)$, $A(0)$, and $L(1)$. Notationally, we use $d(O)$ to mean $(d_{A(0)}(L(0)), d_{A(1)}(A(0), \bar{L}(1)))$. For a rule d, define

$$E_0 Y_d \triangleq E_0 E_0\Big[E_0\left[Y|\bar{A}(1) = d(O), \bar{L}(1)\right]\Big|A(0) = d_{A(0)}(L(0)), L(0)\Big].$$

A (possibly non-unique) optimal rule is given by $d_0^* \triangleq \arg\max_d E_0 Y_{d_0^*}$. Our parameter of interest is $\Psi(P_0) \triangleq E_0 Y_{d_0^*}$. For a distribution P, define the treatment mechanisms $g_{A(0)}(O) \triangleq Pr_P(A(0)|L(0))$ and $g_{A(1)}(O) \triangleq Pr_P(A(1)|A(0), \bar{L}(1))$. Also define

$$\begin{aligned}
\tilde{D}(d, P)(O) \triangleq& D_2^*(d, P)(O) + D_1^*(d, P)(O) \\
&+ E_0\Big[E_0\left[Y|\bar{A}(1) = d(O), \bar{L}(1)\right]\Big|A(0) = d_{A(0)}(L(0)), L(0)\Big],
\end{aligned}$$

where

$$D_1^*(d,P) = \frac{I(A(0)=d_{A(0)}(L(0)))}{g_{A(0)}(O)}\Big(E_P\left[Y \mid \bar{A}(1)=d(O),\bar{L}(1)\right] - E_0\Big[E_0\left[Y|\bar{A}(1)=d(O),\bar{L}(1)\right]\Big|A(0)=d_{A(0)}(L(0)),L(0)\Big]\Big)$$

$$D_2^*(d,P) = \frac{I(\bar{A}(1)=d(O))}{\prod_{k=0}^1 g_{A(k)}(O)}\left(Y - E_P\left[Y \mid \bar{A}(1)=d(O),\bar{L}(1)\right]\right).$$

We can now generalize the confidence interval presented in Section 5 to the two time point case. Let $\{\ell_n\}$ be some sequence of natural numbers. For each $j > \ell_n$, let $\hat{P}_{n,j}$ represent some estimate of P_0 and $d_{n,j}$ some estimate of d_0^*, each based only on the observations $O_1, ..., O_{j-1}$. We really only need estimates of $Pr_P(A(0)|L(0))$, $Pr_P(A(1)|A(0),\bar{L}(1))$, and the two conditional regressions in the definition of $D_1^*(d,P)$. Define

$$\tilde{\sigma}^2_{0,n,j} \triangleq Var_{P_0}\left(\tilde{D}(d_{n,j},\hat{P}_{n,j})\Big| O_1,...,O_{j-1}\right).$$

Let $\tilde{\sigma}^2_{n,j}$ represent an estimate of $\tilde{\sigma}^2_{0,n,j}$ based on (some subset of) the observations $(O_1,...,O_{j-1})$. Also define

$$\Gamma_n \triangleq \frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n}\tilde{\sigma}^{-1}_{n,j}.$$

Define our estimate $\hat{\Psi}(P_n)$ of $\Psi(P_0)$ as

$$\hat{\Psi}(P_n) \triangleq \Gamma_n^{-1}\frac{1}{n-\ell_n}\sum_{j=\ell_n+1}^{n}\tilde{\sigma}^{-1}_{n,j}\tilde{D}_{n,j}(O_j) = \frac{\sum_{j=\ell_n+1}^{n}\tilde{\sigma}^{-1}_{n,j}\tilde{D}_{n,j}(O_j)}{\sum_{j=\ell_n+1}^{n}\tilde{\sigma}^{-1}_{n,j}},$$

where $\tilde{D}_{n,j}(o) \triangleq \tilde{D}(d_{n,j},\hat{P}_{n,j})(o)$. The following $1-\alpha$ confidence interval for $\Psi(P_0)$ is asymptotically valid under conditions similar to C1) through C5) presented in the main text:

$$\hat{\Psi}(P_n) \pm z_{1-\alpha/2}\frac{\Gamma_n^{-1}}{\sqrt{n-\ell_n}}.$$

For an idea of how the conditions change for the two time point case when a non-online estimator is used, see Corollary 2 in van der Laan and Luedtke (2014b). In short, we will see that: conditions like C1) and C2) are still needed to apply the martingale CLT; a condition like C3) is still required to assume that the variance of the terms in the martingale sum stabilizes as sample size grows; a condition like C4) is still needed to ensure that the treatment mechanism and/or conditional mean outcome is estimated well enough to allow the use of the $\tilde{D}$ estimating equation to estimate $\Psi(P_0)$; and a condition like C5) is still needed to ensure that the optimal rule is estimated at a fast enough rate.

The generalization to the general multiple time point problem follows along the same lines as the generalizaion to the two time point problem.