Dynamic treatment regimes: Technical challenges and applications[*]

Eric B. Laber[†]

North Carolina State University
Raleigh, NC 27696-8203
e-mail: laber@stat.ncsu.edu

Daniel J. Lizotte

University of Waterloo
Waterloo, Ontario N2L 3G1
e-mail: dlizotte@uwaterloo.ca

Min Qian

Columbia University
New York, NY 10032
e-mail: mq2158@columbia.edu

William E. Pelham

Florida International University
Miami, FL 33199
e-mail: wpelham@fiu.edu

and

Susan A. Murphy

University of Michigan
Ann Arbor, MI 48106-1248
e-mail: samurphy@umich.edu

Abstract: Dynamic treatment regimes are of growing interest across the clinical sciences because these regimes provide one way to operationalize and thus inform sequential personalized clinical decision making. Formally,

a dynamic treatment regime is a sequence of decision rules, one per stage of clinical intervention. Each decision rule maps up-to-date patient information to a recommended treatment. We briefly review a variety of approaches for using data to construct the decision rules. We then review a critical inferential challenge that results from nonregularity, which often arises in this area. In particular, nonregularity arises in inference for parameters in the optimal dynamic treatment regime; the asymptotic, limiting, distribution of estimators are sensitive to local perturbations. We propose and evaluate a locally consistent Adaptive Confidence Interval (ACI) for the parameters of the optimal dynamic treatment regime. We use data from the Adaptive Pharmacological and Behavioral Treatments for Children with ADHD Trial as an illustrative example. We conclude by highlighting and discussing emerging theoretical problems in this area.

Keywords and phrases: Personalized medicine, data-driven decision making, nonregular inference, adaptive confidence intervals.

Contents

1. Introduction

The increasing ability to collect, manipulate, and access patient-level data combined with a rapidly growing interest in personalized medicine among clinical scientists has created an unprecedented opportunity to improve the quality of

healthcare using data. The potential gains of a data-driven treatment are especially high in the context of chronic illness in which the treatments must be adaptive to the uniquely evolving health status of each patient. Dynamic treatment regimes (DTRs), also called treatment policies, adaptive interventions or adaptive treatment strategies, were created to inform the development of health-related interventions composed of sequences of individualized treatment decisions. These regimes formalize sequential individualized treatment decisions via a sequence of decision rules that map dynamically evolving patient information to a recommended treatment. An optimal DTR optimizes the expectation of a desired cumulative outcome over a population of interest. An estimated optimal DTR has the potential to: improve patient outcomes through adaptive personalized treatment; reduce treatment burden and resource demands by recommending treatment only if, when, and to whom it is needed; and to generate new scientific hypotheses about heterogeneous treatment effects over time both across and within patients.

We consider the development of DTRs using data from Sequential, Multiple, Assignment Randomized Trials (SMART) (Lavori and Dawson, 2000; Murphy, 2005a; Nahum-Shani et al., 2012a; Lei et al., 2012). In these studies, each participant moves through stages of treatment (usually one to three). At each stage, each participant is randomized among treatments. See PSU Methodology Center (2014b) for a partial list of such studies. The Adaptive Pharmacological and Behavioral Treatments for Children with ADHD Trial (W. Pelham (PI); Nahum-Shani et al., 2012a; Lei et al., 2012) exemplifies a common SMART; we use this study for illustration. In the first stage of treatment, children are uniformly randomly assigned to either a low dose of methylphenidate (a psychostimulant drug) or a low intensity of behavioral modification therapy. Beginning at 2 months and monthly thereafter (for the remainder of the 8 month study), each child is assessed for nonresponse; nonresponse occurs if two different teacher ratings concerning the child's school behavior falls below a prespecified criterion. If nonresponse occurs the child is re-randomized uniformly between two tactics: intensify current treatment or augment the current treatment with the other treatment (for example, augment methylphenidate with behavioral modification therapy). As long as the child does not meet the criterion for nonresponse the child remains on current treatment. See Figure 1 for a schematic of this trial.

The estimation of optimal DTRs presents a number of interesting technical challenges and exciting open problems, one of which is inference for nonregular parameters (Bickel et al., 1993). In particular, if an estimated optimal DTR is to inform clinical decisions or guide future research, it is essential to have reliable measures of uncertainty for the estimated regime. However, many of the most commonly used approaches to estimate an optimal DTR involve estimation and inference for parameters that are nonsmooth functionals of the underlying generative distribution. Consequently, estimators of these quantities are necessarily nonregular and asymptotically biased (Van der Vaart, 1991; Robins, 2004; Hirano and Porter, 2009); standard asymptotic approximations to the sampling distributions of these estimators cannot be used directly to form reliable confi-
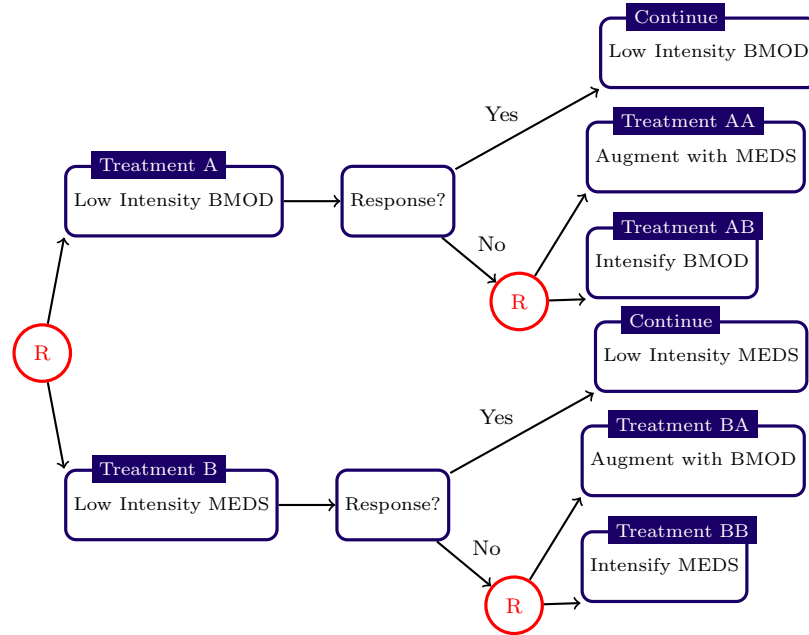
FIG 1. *Schematic describing the Adaptive Pharmacological and Behavioral Treatments for Children with ADHD SMART [W. Pelham (PI)].*

dence intervals or to carry out hypothesis testing. The primary purpose of this paper is to discuss the bias and other inferential problems related to this non-regularity and to offer potential solutions for these problems in the context of DTR research.

In Section 2, we briefly review different methods for constructing optimal DTRs and provide greater detail for one such method, *Q*-learning. In Section 3, we discuss the problem of asymptotic bias and use local alternatives to show that bias-correcting shrinkage methods may perform infinitely worse than uncorrected methods. In Section 4, we discuss interval estimation and propose a locally consistent confidence interval for parameters indexing the optimal DTR. In Section 5, we examine the finite sample performance of the proposed confidence interval using simulated data. In section 6, we perform an analysis of data from a clinical trial involving school-aged children with ADHD. We use this trial to illustrate open problems in model selection and high-dimensional modeling for DTRs that arise even in relatively simple settings. Section 7, provides a general discussion of some open problems relating to estimation and inference of DTRs.

2. Review of methods for constructing dynamic treatment regimes

Throughout, we consider the setting in which there are two stages of binary treatment; this simple setting is sufficient to illustrate the salient theoretical

challenges. Furthermore many SMARTs including the ADHD study described above involve two stages of binary treatment. On each subject, we observe a time-ordered trajectory $(X_1, A_1, X_2, A_2, X_3)$ where: X_1 denotes baseline (pre-randomization) subject information; A_1 denotes the initial treatment, coded to take values in $\{0, 1\}$; X_2 denotes subject information collected during the course of the first treatment but prior to the second treatment; A_2 denotes the second treatment, coded to take values in $\{0, 1\}$; X_3 denotes subject information collected during the course of the second treatment. The initial treatment A_1 is randomly assigned with probability possibly depending on X_1. The second treatment A_2 is randomly assigned with probability possibly depending on (X_1, A_1, X_2). In the ADHD study both A_1 and A_2 are randomized with probability $1/2$ between the binary alternatives. Let H_t, $t = 1, 2$ denote the information available to the decision maker at time t. Thus, $H_1 = X_1$ and $H_2 = (X_1^\intercal, A_1, X_2^\intercal)$. In the ADHD study, H_1 contains baseline ADHD severity, an indicator of oppositional defiant disorder, and an indicator of prior exposure to ADHD medication; H_2 contains H_1, the initial treatment assignment, an indicator of adherence to initial treatment, and month of non-response to initial treatment. There may be an outcome at each stage, say Y_1 and Y_2 given by $Y_1 = y_1(X_1, A_1, X_2)$ and $Y_2 = y_2(X_1, A_1, X_2, A_1, X_3)$ where y_1 and y_2 are known functions. Here we assume that both Y_1 and Y_2 are continuous variables that are coded so that higher values are better. In this case, define $Y \triangleq Y_1 + Y_2$ to be the total outcome. Alternately there may only be an end of study outcome, also denoted by Y. This is the case in the ADHD example, in which Y is a reverse-coded rating of the child's impairment at the end of the last month of the study.

In this two stage setting, a DTR is a pair of decision rules $\pi = (\pi_1, \pi_2)$, where $\pi_t : \text{dom}(H_t) \to \text{dom}(A_t)$ so that a patient presenting at time t with $H_t = h_t$ is assigned treatment $\pi_t(h_t)$. The *value* of a DTR π, denoted $\mathbb{E}^\pi Y$, is the expected outcome under the restriction that $A_t = \pi_t(H_t)$, $t = 1, 2$. The optimal DTR, π^{opt}, satisfies $\mathbb{E}^{\pi^{\text{opt}}} Y = \sup_\pi \mathbb{E}^\pi Y$. Alternately, the optimal DTR can be characterized by Q-functions. In the two stage setting, the Q-functions (Sutton and Barto, 1998; Murphy, 2005b) are

$$
\begin{aligned}
Q_2(h_2, a_2) &\triangleq \mathbb{E}(Y | H_2 = h_2, A_2 = a_2), \\
Q_1(h_1, a_1) &\triangleq \mathbb{E}\left(\max_{a_2} Q_2(H_2, a_2) \big| H_1 = h_1, A_1 = a_1\right),
\end{aligned}
\tag{1}
$$

so that $Q_2(h_2, a_2)$ measures the quality of assigning treatment a_2 to a patient presenting with h_2 at the second stage, and $Q_1(h_1, a_1)$ measures the quality of assigning treatment a_1 to a patient presenting with h_1 at baseline assuming optimal treatment selection at the second stage. The optimal DTR is given by the dynamic programming solution, $\pi_t^{\text{dp}}(h_t) = \arg\max_{a_t} Q_t(h_t, a_t)$ (Bellman, 1957).

Methods for estimating optimal DTRs from data can be broadly classified as either indirect or direct estimation methods (Barto and Dieterich, 2004). Indirect estimation methods use approximate dynamic programming with parametric, semiparametric, or nonparametric methods to first estimate models for

the conditional mean or other aspects of the conditional distributions of the outcomes Y_1, Y_2, Y, (for example, models for the Q-function) and then from these models infer the optimal DTR. Indirect estimation includes g-estimation in structural nested models (Robins, 1989, 1993, 1997) and its variations (Murphy, 2003; Robins, 2004). These methods were originally developed for use with observational data, thus much of the work focused on causal inference. See Robins (1986, 1987) for a formal theory of causal inference for estimating the causal effect of a DTR. Other indirect methods include Q-learning (Murphy, 2005b; Chakraborty and Moodie, 2013; Qian et al., 2013; Chakraborty and Murphy, 2014; Laber et al., 2014), A-learning (Murphy, 2003; Robins, 2004), and regret-regression (Henderson et al., 2009). We provide a detailed discussion of Q-learning below.

Direct estimation methods, also known as policy search methods, directly estimate the marginal mean, $\mathbb{E}^\pi Y$ for all DTRs in a pre-specified class and then maximize the estimator to obtain an estimated DTR. These methods do not require models for conditional means or other aspects of the conditional distributions of the the outcomes (Y_1, Y_2, Y). Recent statistical work in this area includes marginal structural mean models (Robins et al., 2008; Orellana et al., 2010), augmented value maximization (Zhang et al., 2012, 2013), and outcome weighted learning (Zhao et al., 2012, 2013). Again, many direct methods were developed for use with observational data; these methods employ inverse probability weighting methods (Robins, 1998, 1999; Murphy et al., 2001; van der Laan, 2006; van der Laan and Petersen, 2007; Robins et al., 2008; Orellana et al., 2010).

One potential advantage of indirect methods is that the requisite outcome models can be built using standard statistical models (generalized regression models, etc.), which can be checked for goodness of fit. This is particularly attractive when scientific theory and/or expert opinion can be used in forming the outcome model. A potential drawback is that the estimator of the optimal DTR requires that the outcome models are correctly specified. In contrast, most direct estimation methods employ a non- or semi-parametric estimator of $\mathbb{E}^\pi Y$. For example, Zhao et al. (2012) used an inverse probability of treatment weighted estimator of $\mathbb{E}^\pi Y$ whereas Zhang et al. (2012, 2013) used an augmented inverse probability of treatment weighted estimator. Both of these estimators do not require correctly specified conditional outcome models to be consistent and as a result are robust to model misspecification. However, direct estimation methods generally produce estimators of the parameters (in an DTR) with higher variance than indirect estimation methods. This fact has been recognized for some time in the computer science literature with efforts there focused on using outcome models in combination with direct methods so as to reduce variance (Sutton et al., 1999; Konda and Tsitsiklis, 2003). Indeed, there is a vast literature concerning both indirect and direct methods for constructing optimal policies, (i.e., dynamic treatment regimes) in the field of reinforcement learning with many good introductory books (Sutton and Barto, 1998; Si et al., 2004; Busoniu et al., 2010; Szepesvári, 2010; Wiering and van Otterlo, 2012). However, the focus of this work is on algorithms for estimation. To the best of our

knowledge, inference, e.g., confidence intervals or test statistics, that can be used to discuss the level of confidence concerning the constructed DTR with clinical scientists, are absent.

To illustrate and discuss inferential challenges, we consider estimators constructed using Q-learning. Q-learning is attractive to statistical practitioners because Q-learning can be viewed as a multi-stage extension of regression (Nahum-Shani et al., 2012b), thus enabling much of the intuition developed in that area to be (somewhat) easily translated to the area of DTRs. Q-learning is an indirect method of constructing a DTR from data; in Appendix A, we illustrate a direct method, outcome-weighted learning, and illustrate that the use of this method poses the same inferential challenges as Q-learning. The problems we identify with Q-learning apply to many of the aforementioned estimators.

Q-learning provides estimators of the Q-functions. Owing to the max-operator in (1), Q_1 is a nonsmooth functional of the underlying generative distribution; hence, the estimand is also nonsmooth. Next we use Q-learning to illustrate how this nonsmoothness impacts the sampling distributions of DTR estimators.

2.1. Q-learning

Q-learning estimates the optimal DTR by postulating regression models for the Q-functions and then taking the plug-in dynamic programming solution. Consider linear models for the Q-functions of the form $Q_t(h_t, a_t; \beta_t) = h_{t,0}^\mathsf{T} \beta_{t,0} + a_t h_{t,1}^\mathsf{T} \beta_{t,1}$, where $h_{t,0}$ and $h_{t,1}$ are known feature vectors constructed from h_t and $\beta_t = (\beta_{t,0}^\mathsf{T}, \beta_{t,1}^\mathsf{T})^\mathsf{T}$. These feature vectors might contain splines or other nonlinear basis expansions. As an aside, we note that an open problem in DTR research is the development of a principled feature construction method. The above linear model highlights a crucial difference between constructing features for prediction and constructing features for decision making. To see this, note that from the linear model for the Q-function, only the features $h_{t,1}$ will be used by decision rule π_t^{dp}. Thus, high quality features for decision making (as opposed to prediction) should interact with the treatment a_t sufficiently strongly so that the $\pi_t^{\mathrm{dp}}(h_t)$ varies by $h_{t,1}$. At this time, research focused on discovering features for decision making has been in the one-step setting (see Gunter et al., 2011; Foster et al., 2011; Dusseldorp and Van Mechelen, 2013; Janes et al., 2013); the multistage setting is essentially open.

The parameters indexing the Q-functions are estimated using least squares. Let $\mathbb{P}_n$ denote empirical expectation, for example $\mathbb{P}_n f(Z) = n^{-1} \sum_{i=1}^n f(Z_i)$, where $\{Z_i\}_{i=1}^n$ is a random sample. One version of the Q-learning algorithm is as follows.

1. Stage 2 regression: $\widehat{\beta}_2 = \arg\min_{\beta_2} \mathbb{P}_n (Y_2 - Q_2(H_2, A_2; \beta_2))^2$.
2. Predicted second stage outcome: $\widetilde{Y} = Y_1 + \max_{a_2} Q_2(H_2, a_2; \widehat{\beta}_2)$.
3. Stage 1 regression: $\widehat{\beta}_1 = \arg\min_{\beta_1} \mathbb{P}_n (\widetilde{Y} - Q_1(H_1, A_1; \beta_1))^2$.

The Q-learning estimator of π^{dp} is thus $\widehat{\pi}_t(h_t) = \arg\max_{a_t} Q_t(h_t, a_t; \widehat{\beta}_t)$. The second stage coefficients, $\widehat{\beta}_2$, are ordinary least squares estimators and are

thus regular and asymptotically normal under mild conditions (see Section 4). However, the first stage coefficients depend on the maximized second stage Q-function; because the max operator is nonsmooth, estimated coefficients $\widehat{\beta}_1$ are in turn a nonsmooth function of the data.

For notational simplicity from here until Section 6, $Y_1 \equiv 0$ so that $Y = Y_2$, and thus we will omit any subscripts on Y. Define the following population analogs of the estimators used in Q-learning:

$$\beta_2^* \triangleq \arg\min_{\beta_2} P\left(Y - Q_2(H_2, A_2; \beta_2)\right)^2,$$

$$\widetilde{Y}^* \triangleq \max_{a_2} Q_2(H_2, a_2; \beta_2^*) = H_{2,0}^\intercal \beta_{2,0}^* + \left[H_{2,1}^\intercal \beta_{2,1}^*\right]_+,$$

$$\beta_1^* \triangleq \arg\min_{\beta_1} P\left(\widetilde{Y}^* - Q_1(H_1, A_1; \beta_1)\right)^2,$$

where P denotes expectation with respect to the distribution of $(X_1, A_1, X_2, Y_1, A_2, X_3, Y_2)$, and the second line follows from the fact that $a_2 \in \{0, 1\}$. In addition, define $B_t \triangleq (H_{t,0}^\intercal, A_t H_{t,1}^T)^\intercal$, $\Sigma_{t,\infty} \triangleq P B_t B_t^\intercal$ for $t = 1, 2$, and $\widehat{\Sigma}_t \triangleq \mathbb{P}_n B_t B_t^\intercal$. We assume $\widehat{\Sigma}_t$ is invertible. Then, $\widehat{\beta}_1 = \widehat{\Sigma}_1^{-1} \mathbb{P}_n B_1 \widetilde{Y}$, $\beta_1^* = \Sigma_{1,\infty}^{-1} P B_1 \widetilde{Y}^*$ so that $\sqrt{n}(\widehat{\beta}_1 - \beta_1^*) = \widehat{\Sigma}_1^{-1} \sqrt{n} \mathbb{P}_n B_1 (\widetilde{Y} - B_1^\intercal \beta_1^*)$. It is useful to decompose $\widehat{\Sigma}_1^{-1} \sqrt{n} \mathbb{P}_n B_1 (\widetilde{Y} - B_1^\intercal \beta_1^*)$ as

$$\mathbb{S}_n + \widehat{\Sigma}_1^{-1} \mathbb{P}_n B_1 \mathbb{U}_n, \tag{2}$$

where

$$\mathbb{S}_n = \widehat{\Sigma}_1^{-1} \sqrt{n} \mathbb{P}_n B_1 \left[\left(H_{2,0}^\intercal \beta_{2,0}^* + \left[H_{2,1}^\intercal \beta_{2,1}^*\right]_+ - B_1^\intercal \beta_1^*\right) + H_{2,0}^\intercal \left(\hat{\beta}_{2,0} - \beta_{2,0}^*\right) \right],$$

$$\mathbb{U}_n = \sqrt{n} \left(\left[H_{2,1}^\intercal \hat{\beta}_{2,1}\right]_+ - \left[H_{2,1}^\intercal \beta_{2,1}^*\right]_+ \right).$$

The term $\mathbb{S}_n$ is smooth and asymptotically normal but $\mathbb{U}_n$ is nonsmooth in $\widehat{\beta}_{2,1}$. To understand the implications of this nonsmoothness, fix $H_{2,1} = h_{2,1}$. If $h_{2,1}^\intercal \beta_{2,1}^* \neq 0$, then $\mathbb{U}_n\big|_{H_{2,1}=h_{2,1}}$ is asymptotically normal with mean zero. However, if $h_{2,1}^\intercal \beta_{2,1}^* = 0$ then $\mathbb{U}_n\big|_{H_{2,1}=h_{2,1}} = [h_{2,1}^\intercal \sqrt{n}(\widehat{\beta}_{2,1} - \beta_{2,1}^*)]_+$, which converges to the positive part of a mean zero normal random variable. Thus, the limiting distribution of $\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$ depends in a nonuniform manner on the value of $\beta_{2,1}^*$ and the distribution of $H_{2,1}$.

If $H_{2,1}$ is composed only of continuous variables, then some sceptism is natural because $P[H_{2,1}^\intercal \beta_{2,1}^* = 0] = 0$. However, in most clinical trials, the effect of treatment can be expected to be small ($H_{2,1}^\intercal \beta_{2,1}^*$ is the effect of stage 2 treatment) relative to the noise level. Thus, even though $H_{2,1}^\intercal \beta_{2,1}^*$ may not be 0, its estimator can be expected to be near 0 with high probability. And, as we shall see, the small sample behavior of $\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$ is poorly approximated by fixed-parameter asymptotic results that assume $P[H_{2,1}^\intercal \beta_{2,1}^* = 0] = 0$ (see discussion of bias in Section 3 and evaluation of confidence intervals in Section 5). Formally, $\widehat{\beta}_1$ is a non-regular estimator because under moving-parameter (order $1/\sqrt{n}$ in

our setting) local asymptotics, its limiting distribution depends on the direction along which the local parameter approaches $\beta_{2,1}^*$ (Tsiatis, 2006). We show formally that $\widehat{\beta}_1$ is non-regular in Section 4. Moving-parameter asymptotic results better reflect small sample behavior because they allow $H_{2,1}^{\mathsf{T}}\beta_{2,1}^*$ to remain near to zero even in infinite samples (Leeb and Pötscher, 2003, 2005; Andrews and Guggenberger, 2009; Laber and Murphy, 2011).

3. Asymptotic bias

In the study of nonregular estimators, much attention has been given to asymptotic bias; characterized here as bias that is $O(1/\sqrt{n})$. Because asymptotic bias may be indicative of bias in small samples, incorrect Type I error levels in hypothesis testing, and poor coverage rates of confidence intervals (e.g., Blumenthal and Cohen, 1968; Casella and Strawderman, 1981; Bickel, 1981; Robins, 2004; Marchand and Strawderman, 2004; Chakraborty et al., 2009; Moodie et al., 2010), there is great interest in characterizing and reducing asymptotic bias. This interest certainly rises in the construction of DTRs because if $H_{2,1}^{\mathsf{T}}\beta_{2,1}^*$ is near 0 with high probability, then one anticipates problems in approximating the distribution of $\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$. Considering the ADHD example, one might be concerned about asymptotic bias because in most clinical trials the treatment effects, if non-zero, will be small.

Here, we (i) characterize the points in the parameter space at which asymptotic bias occurs in the first stage Q-learning estimator; (ii) show that, given the correct amount of shrinkage, one can reduce the asymptotic bias by using a shrinkage estimator; (iii) prove that shrinking too aggressively can lead to arbitrarily bad performance in finite samples; and (iv) illustrate, via a toy example, that one can find points in the parameter space that are local to the nonregular parameter values and at which the shrinkage increases the bias relative to no-shrinkage. These results lead us to Section 4 in which we develop a method for constructing confidence intervals (for the first stage parameters) that does not involve shrinkage for bias reduction.

We use $\mathbb{E}$ to denote expectation over P (the distribution of the observed data). Let $c \in \mathbb{R}^{\dim(\beta_1^*)}$ be fixed. For any $\sqrt{n}$-consistent estimator $\tilde{\beta}_1$ of β_1^* with $\sqrt{n}(\tilde{\beta}_1 - \beta_1^*)$ converging in distribution to $\mathbb{M}$, define the c-directional asymptotic bias of $\tilde{\beta}_1$ as

$$\text{Bias}(\tilde{\beta}_1, c) \triangleq \mathbb{E}c^{\mathsf{T}}\mathbb{M}.$$

Define

$$
\begin{aligned}
g_2(B_2, Y; \beta_2^*) &\triangleq B_2(Y - B_2^{\mathsf{T}}\beta_2^*), \\
g_1(B_1, H_2; \beta_1^*, \beta_2^*) &\triangleq B_1\left(H_{2,0}^{\mathsf{T}}\beta_{2,0}^* + \left[H_{2,1}^{\mathsf{T}}\beta_{2,1}^*\right]_+ - B_1^{\mathsf{T}}\beta_1^*\right).
\end{aligned}
$$

Throughout we assume:

(A1) Histories H_2, features B_1, and outcomes Y satisfy the moment inequalities $P||H_2||^2\,||B_1||^2 < \infty$ and $PY^2||B_2||^2 < \infty$.
(A2) Matrices $\Sigma_{t,\infty}$ and $\text{Cov}\,(g_1, g_2)$ are strictly positive definite.

Assumptions (A1)–(A2) are quite mild, requiring only full rank design matrices and some moment conditions. Using standard methods, it can be shown that $\mathbb{V}_n \triangleq \sqrt{n}(\hat{\beta}_2 - \beta_2^*)$ is asymptotically normal with mean zero and variance-covariance $\Omega = (PB_2B_2^\intercal)^{-1}PB_2B_2^\intercal(Y - B_2^\intercal\beta_2^*)^2(PB_2B_2^\intercal)^{-1}$. Let $\Sigma_{21,21}$ denote the submatrix of Ω corresponding to the limiting asymptotic covariance of $\sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*)$ and $\hat{\Sigma}_{21,21}$ the corresponding plug-in estimator. The following result is proved in Appendix B.

Theorem 3.1. *Assume (A1) and (A2) and let $c \in \mathbb{R}^{\dim(\beta_1^*)}$ be fixed. Then:*

$$\mathrm{Bias}(\widehat{\beta}_1, c) = \frac{c^\intercal\Sigma_{1,\infty}^{-1}P\left[B_1\sqrt{H_{2,1}^\intercal\Sigma_{21,21}H_{2,1}}\mathbb{1}_{H_{2,1}^\intercal\beta_{2,1}^*=0}\right]}{\sqrt{2\pi}}.$$

The preceding results show that the asymptotic bias of Q-learning is: (i) only nonzero when the second stage treatment effect, $H_{2,1}^\intercal\beta_{2,1}^*$, satisfies $P(H_{2,1}^\intercal\beta_{2,1}^* = 0) > 0$; (ii) bounded under (A1); and (iii) is a weighted average of $c^\intercal\Sigma_{1,\infty}^{-1}B_1$ with weights $\sqrt{nVar(H_{2,1}^\intercal\widehat{\beta}_{2,1}\mathbb{1}_{H_{2,1}^\intercal\beta_{2,1}^*=0}|H_{2,1})}/\sqrt{2\pi} = \sqrt{H_{2,1}^\intercal\Sigma_{2,1}H_{2,1}}\mathbb{1}_{H_{2,1}^\intercal\beta_{2,1}^*=0}/\sqrt{2\pi}$, these weights will be large when the estimated treatment effect has high variance for subjects with a null treatment effect. From (iii) it can be seen that the asymptotic bias might be reduced if the variability of $H_{2,1}^\intercal\widehat{\beta}_{2,1}$ is reduced for histories that satisfy $H_{2,1}^\intercal\beta_{2,1}^* = 0$. Shrinkage methods have been proposed as one way to reduce this variability.

A common strategy for reducing asymptotic bias in Q-learning is to shrink the predicted outcome $\widetilde{Y}$. Moodie et al. (2010) proposed a hard-thresholding approach; Chakraborty et al. (2009) proposed a soft-thresholding estimator; and more recently Song et al. (2011) proposed a penalized version of Q-learning. We use the soft-thresholding estimator proposed by Chakraborty et al. (2009) as an illustrative example. Chakraborty et al. (2009) illustrate, using simulation studies, that soft-thresholding reduces bias in small samples. Define

$$\widetilde{Y}^\sigma \triangleq \widehat{\beta}_{2,0}^\intercal H_{2,0} + \left[H_{2,1}^\intercal\widehat{\beta}_{2,1}\right]_+\left(1 - \frac{\sigma H_{2,1}^\intercal\widehat{\Sigma}_{21,21}H_{2,1}}{n(\widehat{\beta}_{2,1}^\intercal H_{2,1})^2}\right)_+, \qquad (3)$$

where σ is a nonnegative constant. For positive values of σ, the soft-thresholding estimator shrinks the nonsmooth part of the predicted outcome toward zero. The first stage soft-thresholding estimators are given by

$$\widehat{\beta}_1^\sigma \triangleq \arg\min_{\beta_1}\mathbb{P}_n\left(\widetilde{Y}^\sigma - Q_1(H_1, A_1; \beta_1)\right)^2.$$

The following result is proved in Appendix B.

Theorem 3.2. *Assume (A1) and (A2) and let $c \in \mathbb{R}^{p_1}$ be fixed. Then:*

  1. $\left|\mathrm{Bias}(\widehat{\beta}_1^\sigma, c)\right| \leq \left|\mathrm{Bias}(\widehat{\beta}_1, c)\right|$ *for any $\sigma \geq 0$.*

2. *If* $\text{Bias}(\widehat{\beta}_1, c) \neq 0$ *then for* $\sigma > 0$

$$\frac{\text{Bias}(\widehat{\beta}_1^\sigma, c)}{\text{Bias}(\widehat{\beta}_1, c)} = \exp\{-\sigma/2\} - \sigma \int_{\sqrt{\sigma}}^{\infty} \frac{1}{x} \exp\{-x^2/2\} \mathrm{d}x.$$

Chakraborty et al. (2009) recommend $\sigma = 3$, which corresponds to an approximate empirical Bayes estimator; plugging $\sigma = 3$ into the above expression shows an approximate 13-fold reduction in asymptotic bias. The soft-thresholding estimator has smaller asymptotic bias than Q-learning, and the preceding result seems to suggest that larger values of σ are preferred. Indeed, if $\sigma \to \infty$ the asymptotic bias of the soft-thresholding estimator converges to zero. These results are point-wise in the parameter space for (β_1, β_2); that is, for any fixed true parameter value of (β_1, β_2) the asymptotic bias converges to zero.

While it appears that these methods reduce asymptotic bias, it is known that the methods cannot completely remove the asymptotic bias without driving the mean squared error to infinity (see, for example, Liu and Brown, 1993; Chen, 2004). Furthermore, even considering just the bias, if we evaluate the bias in a uniform (across the parameter space) manner the situation looks quite different. In fact, from this viewpoint, we see that soft-thresholding may actually incur significantly more bias in finite samples than Q-learning, especially for large values of σ. Intuitively reducing bias at one point in the parameter space leads to increased bias at other points. We illustrate the bias both from a theoretical viewpoint and provide a toy example that highlights the bias.

Local or moving-parameter asymptotics play an important role in the theoretical study of nonsmooth estimators, such as $\widehat{\beta}_1$. Local asymptotics provide a way to understand and study the behavior of a nonsmooth estimator in a more uniform manner across the parameter space, in particular by using generative models that are arbitrarily 'close' to the problematic nonsmooth points in the parameter space. Consider the following local asymptotic framework.

(A3) For any $s \in \mathbb{R}^{\dim(\beta_{2,1}^*)}$, there exists a sequence of local alternatives P_n converging to P in the sense that:

$$\int \left[\sqrt{n} \left(dP_n^{1/2} - dP^{1/2} \right) - \frac{1}{2} v_s dP^{1/2} \right]^2 \to 0,$$

for some real-valued measurable function v_s for which

- if $\beta_{2,n}^* \triangleq \arg\min_\beta P_n(Y - Q_2(H_2, A_2; \beta))^2$, then $\beta_{2,1,n}^* \triangleq \beta_{2,1}^* + s/\sqrt{n} + o(1/\sqrt{n})$ and
- $P_n \|H_2\|^2 \|B_1\|^2$, $P_n Y_2^2 \|B_2\|^2$ are bounded sequences.

See the Appendix for the relationship between v_s and s. Define $\tilde{Y}_n^* = H_{2,0}^\mathsf{T} \beta_{2,0,n}^* + [H_{2,1}^\mathsf{T} \beta_{2,1,n}^*]_+$ and $\beta_{1,n}^* \triangleq \arg\min_\beta P_n(\tilde{Y}_n^* - Q_1(H_1, A_1; \beta))^2$. For any estimator $\widetilde{\beta}_1$ of β_1^* for which $\sqrt{n}(\widetilde{\beta}_1 - \beta_{1,n}^*)$ converges in distribution under P_n to a random vector indexed by s, say $\mathbb{M}(s)$, define the c-directional asymptotic bias under P_n as

$$\text{Bias}(\widehat{\beta}_1, c, s) \triangleq \mathbb{E} c^\mathsf{T} \mathbb{M}(s).$$

The following result is proved in Appendix B.

Theorem 3.3. *Assume (A1)–(A3) and let $c \in \mathbb{R}^{\dim(\beta_1^*)}$ be fixed. Further assume that $P1_{H_{2,1}^\mathsf{T} \beta_{2,1}^* = 0} > 0$. Then:*

1. $\sup_{s \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} \left| \text{Bias}(\widehat{\beta}_1, c, s) \right| \leq \frac{||c^\mathsf{T} \Sigma_{1,\infty}^{-1}|| |P[||B_1|| \sqrt{H_{2,1}^\mathsf{T} \Sigma_{21,21} H_{2,1}} 1_{H_{2,1}^\mathsf{T} \beta_{2,1}^* = 0}]}{\sqrt{2\pi}} + o(1)$.

2. $\sup_{s \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} \left| \text{Bias}(\widehat{\beta}_1^\sigma, c, s) \right| \to \infty$ *as* $\sigma \to \infty$.

The preceding suggests that thresholding too aggressively may lead to large bias in finite samples; results of this type are anticipated by Liu and Brown (1993); Hirano and Porter (2012).

Next we consider a toy example which more clearly illuminates the effect of thresholding on bias. Consider data $\{(A_i, Y_i)\}_{i=1}^n$ from a two-arm randomized study where: $A \in \{0, 1\}$ denotes a randomly assigned binary treatment; and $Y \in \mathbb{R}$ denotes the outcome coded so that higher values are better. Assume subjects are randomized with equal probability so that $P(A = 1) = 1/2$. Define $\mu_a^* \triangleq \mathbb{E}(Y|A = a)$, and $\theta^* \triangleq \max(\mu_0^*, \mu_1^*)$ so that θ^* denotes mean outcome if all subjects are assigned treatment $\arg\max_a \mu_a^*$. Let $\widehat{\mu}_a \triangleq \mathbb{P}_n Y 1_{A=a} / \mathbb{P}_n 1_{A=a}$, then the plug-in estimator of θ^* is

$$\widehat{\theta} = \max(\widehat{\mu}_0, \widehat{\mu}_1) = \frac{\widehat{\mu}_0 + \widehat{\mu}_1}{2} + \frac{|\widehat{\mu}_0 - \widehat{\mu}_1|}{2},$$

which is the sum of a smooth term, $(\widehat{\mu}_0 + \widehat{\mu}_1)/2$, and a non-smooth term $|\widehat{\mu}_0 - \widehat{\mu}_1|/2$. In this example, the problematic area of the parameter space is $\Theta_{\text{Bad}} = \{(\mu_1, \mu_2) \in \mathbb{R}^2 : \mu_1 = \mu_2\}$; under mild regularity conditions it can be seen that if $\theta^* \notin \Theta_{\text{Bad}}$, then $\sqrt{n}(\widehat{\theta} - \theta^*)$ converges in distribution to mean zero normal random variable, whereas if $\theta^* \in \Theta_{\text{Bad}}$, then $\sqrt{n}(\widehat{\theta} - \theta^*)$ converges in distribution to $(Z_0 + Z_1)/2 + |Z_0 - Z_1|/2$ where Z_0, Z_1 are independent mean zero normal random variables. Thus, when $\theta^* \in \Theta_{\text{Bad}}$, since $\mathbb{E}|Z_0 - Z_1| \geq 0$ with equality only when both Z_0 and Z_1 are degenerate, $\widehat{\theta}$ has positive asymptotic bias.

One approach to reducing the asymptotic bias of $\widehat{\theta}$ is by thresholding the nonsmooth term in $\widehat{\theta}$. Assume that $\text{Var}(Y|A = a) = 1$ for $a = 0, 1$. For $\sigma > 0$, define

$$\widehat{\theta}^\sigma \triangleq \frac{\widehat{\mu}_0 + \widehat{\mu}_1}{2} + \frac{|\widehat{\mu}_0 - \widehat{\mu}_1|}{2} \left(1 - \frac{4\sigma}{n(\widehat{\mu}_0 - \widehat{\mu}_1)^2} \right)_+, \tag{4}$$

so that (4) is analogous to (3). In fact, (4) is a special case of (3) and is the resulting estimator of the mean response at the first stage when there are no stage 2 covariates (except for the treatment indicator). Thus, analogous arguments to those in the preceding section show that, for $\theta^* \in \Theta_{\text{Bad}}$, $\widehat{\theta}^\sigma$ has smaller asymptotic bias than $\widehat{\theta}$, and that this asymptotic bias decreases as σ increases. Similarly, a local asymptotic analysis suggests that aggressive shrinkage may lead to large bias in finite samples.

We now illustrate the small sample behavior of $\widehat{\theta}^\sigma$ using simulated data. We assume $Y|A = a \sim \text{Normal}(\mu_a, 1)$ and that treatment assignment is perfectly balanced. We use 1000 Monte Carlo replications to estimate bias for each parameter setting. The leftmost plot in Figure 2 shows the bias as a function of
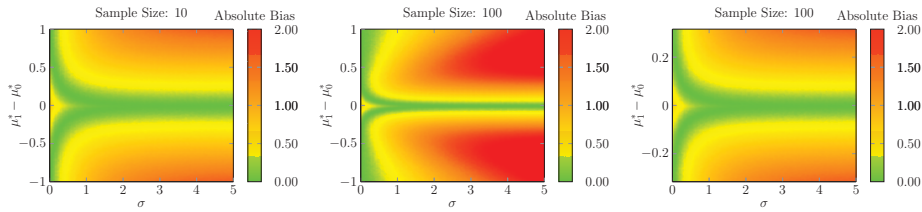
FIG 2. **Left:** *Bias, in units of $1/\sqrt{n}$, as a function of effect size $\mu_1^* - \mu_0^*$ and tuning parameter σ for $n = 10$.* **Center:** *Bias, in units of $1/\sqrt{n}$, as a function of effect size $\mu_1^* - \mu_0^*$ and tuning parameter σ for $n = 100$;* **Right:** *Same as center plot after rescaling y-axis.*

the treatment effect $\mu_1^* - \mu_0^*$ and tuning parameter σ for $n = 10$. Note that when $n = 10$ a standard normal 90% confidence interval for $\mu_1^* - \mu_0^*$ has a width of about two. Thus, the y-axis has been scaled to roughly correspond to a 90% confidence interval centered around the problematic point 0. From the plot it is clear that *if $\mu_1^* - \mu_0^* = 0$*, larger values of σ correspond to lower bias; however, as anticipated from the local asymptotic analysis, large values of σ cause the bias to increase dramatically as $\mu_1^* - \mu_0^*$ moves away from zero but stays within the confidence interval. As the data do not contain sufficient information to differentiate between different parameter values within the confidence interval, an adaptive shrinkage strategy based on the estimated treatment difference $\widehat{\mu}_1 - \widehat{\mu}_0$ is not possible. The middle plot in Figure 2 shows the same bias plot for $n = 100$ displayed with the same y-axis as the $n = 10$ case; the very small yellow-red cross-section above the region around $\sigma = 0$ is anticipated by the fixed asymptotic analysis which states for if $\mu_1^* - \mu_0^* \neq 0$ the bias decreases as the sample size increases. However, the rightmost plot in Figure 2 shows the bias for $n = 100$ after rescaling the y-axis to reflect power (i.e., now the range of the y-axis corresponds to the length of a standard normal 90% confidence interval for $\mu_1^* - \mu_0^*$ when $n = 100$); the figure is *essentially identical* to the leftmost ($n = 10$) plot. The similarity of these plots after rescaling exemplifies the insights gained from a local asymptotics approach which allows notions of 'closeness' to persist as the sample size increases.

4. Confidence intervals

If estimated optimal DTRs are to be used to inform clinical decision making or future research it is essential that they be accompanied by reliable measures of uncertainty. For example, confidence intervals for the parameters in Q-learning can be used to determine which subject covariates are significant, i.e., important for choosing treatment at time t. If a covariate is not important for choosing treatment then it may be possible to forgo collecting it thereby potentially reducing cost and patient burden. Furthermore, a confidence interval for a linear combination of parameters in Q-learning can be used to construct a confidence set for $\arg\max_{a_t} Q_t(h_t, a_t; \beta_t^*)$; if this confidence interval equals $\{0, 1\}$ then there is insufficient evidence to recommend a single treatment for a patient presenting

with $H_t = h_t$ and other factors, e.g., clinical judgment, patient preference, cost, and availability, should be used to choose a treatment.

Constructing valid confidence intervals from nonregular estimators is difficult because it is impossible to uniformly consistently estimate the sampling distribution of a nonregular estimator (Van der Vaart, 1991; Andrews, 2000; Leeb and Poetscher, 2003; Hirano and Porter, 2012). Estimators that reduce asymptotic bias, for example thresholding (Chakraborty et al., 2009) and singular penalization (Song et al., 2011; Goldberg et al., 2012), were originally suggested as methods for constructing high-quality confidence intervals for parameters in Q-learning. However, these methods involve additional nonsmooth operations of the data and it can be shown that the confidence intervals proposed with these estimators are inconsistent under local alternatives. Furthermore, asymptotic bias only reflects the mean of the sampling distribution whereas confidence intervals require estimation of the tails of the sampling distribution. Thus, in general reducing asymptotic bias is not sufficient for valid inference.

On the other hand, confidence intervals that deliver the desired level of confidence can be used to conduct inference even in the presence of bias on the order $1/\sqrt{n}$. In this section we: (i) review a projection interval proposed by Robins (2004); and (ii) propose a new procedure that is adaptive and locally consistent. Additional discussion and potential extensions of the methods proposed here are provided in Section 7.

4.1. A projection interval

Recall that $h_{2,1}^\mathsf{T} \beta_{2,1}^*$ is the second stage treatment effect (see Section 2.1) for feature vector $h_{2,1}$; small sample inferential problems occur when this second stage treatment effect is small with positive probability (e.g., small sample bias, poor coverage properties of standard CIs). Robins (2004) using ideas similar to those of Berger and Boos (1994) proposed a projection confidence interval. A projection region for β_1^* is constructed as follows. If $\beta_{2,1}^*$ were known then it is possible to form a regular asymptotically normal estimator of β_1^*, say $\widehat{\beta}_1(\beta_{2,1}^*)$ (given below), and use this estimator to form a valid $(1-\alpha) \times 100\%$ confidence region for β_1^*, say $\mathbb{I}_{n,\alpha}(\beta_{2,1}^*)$. However, since $\beta_{2,1}^*$ is unknown, a conservative approach is to first form a $(1-\eta) \times 100\%$ confidence region for $\beta_{2,1}^*$, say $\zeta_{n,\eta}$. Second for each $\beta_{2,1} \in \zeta_{n,\eta}$ form a confidence region for $\widehat{\beta}_1(\beta_{2,1})$, say $\mathbb{I}_{n,\alpha}(\beta_{2,1})$. The projection confidence region for β_1^* is the union of $\mathbb{I}_{n,\alpha}(\beta_{2,1})$ over all where $\beta_{2,1} \in \zeta_{n,\eta}$. There are two chances to 'miss' with this approach: (i) $\zeta_{n,\eta}$ may fail to contain $\beta_{2,1}^*$ which occurs with probability no more than η; and (ii) $\beta_{2,1}^* \in \zeta_{n,\eta}$ but $\mathbb{I}_n(\beta_{2,1}^*)$ fails to contain β_1^* which occurs with probability no more than α. Thus, the coverage of the foregoing procedure is at least $(1 - \alpha - \eta) \times 100\%$.

In the context of Q-learning this idea is implemented as follows. For any $\beta_{2,1}$ define $\widetilde{Y}(\beta_{2,1}) \triangleq \max_{a_2} Q_2(H_2, a_2; (\widehat{\beta}_{2,0}^\mathsf{T}, \beta_{2,1}^\mathsf{T}))$ and $\widetilde{Y}^*(\beta_{2,1}) \triangleq \max_{a_2} Q_2(H_2, a_2; (\beta_{2,0}^{*\mathsf{T}}, \beta_{2,1}^\mathsf{T}))$; subsequently define $\widehat{\beta}_1(\beta_{2,1}) \triangleq \arg\min_{\beta_1} \mathbb{P}_n(\widetilde{Y}_1(\beta_{2,1}) - Q_1(H_1, A_1; \beta_1))^2$ and $\beta_1^*(\beta_{2,1}) \triangleq \arg\min_{\beta_1} P(\widetilde{Y}^* - Q_1(H_1, A_1; \beta_1))^2$. Note that $\beta_1^* = \beta_1^*(\beta_{2,1}^*)$.

For $\beta_{2,1}$ fixed, it follows from standard arguments that $\sqrt{n}(\widehat{\beta}_1(\beta_{2,1}) - \beta_1^*(\beta_{2,1}))$ is regular, asymptotically normal with mean zero. Let $C(\beta_{2,1})$ denote the asymptotic variance-covariance matrix of $\sqrt{n}(\widehat{\beta}_1(\beta_{2,1}) - \beta_1^*(\beta_{2,1}))$ and let $\widehat{C}(\beta_{2,1})$ denote a consistent estimator of $C(\beta_{2,1})$. A Wald-type asymptotic $(1-\alpha) \times 100\%$ confidence region for $\beta_1^*(\beta_{2,1})$ is therefore

$$
\mathbb{I}_{n,\alpha}(\beta_{2,1}) \triangleq \left\{ \beta_1 \in \mathbb{R}^{\dim(\beta_1^*)} : \right.
$$
$$
\left. n\left(\widehat{\beta}_1(\beta_{2,1}) - \beta_1\right)^\mathsf{T} \widehat{C}^{-1}(\beta_{2,1})\left(\widehat{\beta}_1(\beta_{2,1}) - \beta_1\right) \leq \chi^2_{1-\alpha,\dim(\beta_1^*)} \right\},
$$

where $\chi^2_{\alpha,d}$ is the $(1-\alpha) \times 100$ percentile of a χ^2-distribution with d degrees of freedom. In particular, $\mathbb{I}_{n,\alpha}(\beta_{2,1}^*)$ is a valid asymptotic $(1-\alpha) \times 100\%$ confidence interval for $\beta_1^*(\beta_{2,1}^*) = \beta_1^*$. Of course, $\beta_{2,1}^*$ is unknown, but $\hat{\beta}_{2,1}$ is a regular asymptotically normal estimator of $\beta_{2,1}^*$ and thus standard methods for constructing confidence sets, e.g., the bootstrap or Taylor series arguments, can be used to construct a valid $(1-\eta) \times 100\%$ for $\beta_{2,1}^*$, say $\zeta_{n,\eta}$. Then, the union

$$
\bigcup_{\beta_{2,1} \in \zeta_{n,\eta}} \mathbb{I}_{n,\alpha}(\beta_{2,1}), \tag{5}
$$

is a valid $(1-\alpha-\eta) \times 100\%$ confidence region for β_1^*. To see this, note that

$$
P\left(\beta_1^* \notin \bigcup_{\beta_{2,1} \in \zeta_{n,\eta}} \mathbb{I}_{n,\alpha}(\beta_{2,1})\right) = P\left(\beta_1^* \notin \bigcup_{\beta_{2,1} \in \zeta_{n,\eta}} \mathbb{I}_{n,\alpha}(\beta_{2,1}), \beta_{2,1}^* \notin \zeta_{n,\eta}\right)
$$
$$
+ P\left(\beta_1^* \notin \bigcup_{\beta_{2,1} \in \zeta_{n,\eta}} \mathbb{I}_{n,\alpha}(\beta_{2,1}), \beta_{2,1}^* \in \zeta_{n,\eta}\right),
$$

which is bounded above by $P(\beta_{2,1}^* \notin \zeta_{n,\eta}) + P(\beta_1^*(\beta_{2,1}^*) \notin \mathbb{I}_{n,\alpha}(\beta_{2,1}^*)) \leq \eta + \alpha + o_P(1)$. This confidence interval is appealing for its simplicity but may be conservative especially when $H_{2,1}^\mathsf{T}\beta_{2,1}^*$ is bounded away from zero with high probability. One approach to reduce conservatism is to first test $H_0 : \beta_{2,1}^* \equiv 0$, if the test rejects then $\mathbb{I}_{n,\alpha}(\widehat{\beta}_{2,1})$ is used, if the test fails to reject then the projection interval (5) is used (Robins, 2004). This pretesting approach is adaptive at the population level but may be conservative when the distribution of $H_{2,1}^\mathsf{T}\beta_{2,1}^*$ has mass both near to and far from zero. A potentially less conservative approach is to partition the observed sample into two groups according to the (estimated) magnitude of $H_{2,1}^\mathsf{T}\beta_{2,1}^*$ and apply a conservative procedure only to observations for which $H_{2,1}^\mathsf{T}\beta_{2,1}^*$ is small. We now discuss such a procedure.

4.2. Adaptive confidence intervals

In this section we construct a regular, i.e., locally consistent, confidence interval for linear combinations of the first stage coefficients. Note that confidence intervals for the second stage coefficients can be obtained using standard methods for least squares estimators. Let $\widehat{\Sigma}_1 \triangleq \mathbb{P}_n B_1 B_1^\mathsf{T}$ so that $\widehat{\beta}_1 = \widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \widetilde{Y}$ and

$\beta_1^* = \Sigma_{1,\infty}^{-1} P B_1 \widetilde{Y}^*$. Recall that it is not possible in general to construct a uniformly convergent estimator of the limiting distribution of $\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$ (Van der Vaart, 1991; Hirano and Porter, 2009). For a given constant $c \in \mathbb{R}^{\dim(\beta_1^*)}$, our approach is to bound $c^{\mathsf{T}}\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$ between two regular, uniformly convergent, upper and lower bounds. Because these bounds are smooth, we can bootstrap them to form a confidence set for $c^{\mathsf{T}}\beta_1^*$. This strategy is similar to the work of Laber and Murphy (2011) on classification but differs in that here the functional of interest is a fixed (rather than data-dependent) parameter and the functional is more complicated. We present the two-stage binary-treatment case here; extensions to the case of an arbitrary number of treatments and stages of treatment can be found in a technical report (Laber et al., 2010).

Recall that for any $c \in \mathbb{R}^{\dim(\beta_1^*)}$ $c^{\mathsf{T}}\sqrt{n}(\widehat{\beta}_1 - \beta_1^*) = c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1(\widetilde{Y} - B_1^{\mathsf{T}}\beta_1^*)$ can be decomposed as $c^{\mathsf{T}}\mathbb{S}_n + c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n$, where the term $\mathbb{S}_n$ is smooth and asymptotically normal but $\mathbb{U}_n$ is nonsmooth. Also recall that $\mathbb{U}_n = \sqrt{n}([H_{2,1}^{\mathsf{T}}\widehat{\beta}_{2,1}]_+ - [H_{2,1}^{\mathsf{T}}\beta_{2,1}^*]_+)$. Our goal is to form smooth upper and lower bounds on $c^{\mathsf{T}}\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$. To limit conservatism, these bounds are based on the nonsmooth term $\mathbb{U}_n$ and only involve subjects with small second stage treatment effects, i.e., those subjects with histories $h_{2,1}$ with $h_{2,1}^{\mathsf{T}}\beta_{2,1}^* \approx 0$. We partition the observed data into two groups: (Group 1) subjects for whom $h_{2,1}^{\mathsf{T}}\beta_{2,1}^*$ cannot be distinguished from zero; and (Group 2) subjects for whom $h_{2,1}^{\mathsf{T}}\beta_{2,1}^*$ is unlikely to be near zero. This partitioning is based on a "pretest" (see Olshen, 1973; Andrews, 2001b; Andrews and Soares, 2007; Cheng, 2008; Andrews and Guggenberger, 2009). The pretest is based on $\widehat{T}(h_{2,1})$ which is a test statistic that diverges to $+\infty$ when $h_{2,1}^{\mathsf{T}}\beta_{2,1}^*$ is nonzero but is bounded in probability when $h_{2,1}^{\mathsf{T}}\beta_{2,1}^* = 0$. The pretest assigns a subject with $H_{2,1} = h_{2,1}$ to Group 1 if $\widehat{T}(h_{2,1}) \leq \lambda_n$ and Group 2 otherwise; λ_n is a tuning parameter. In what follows we assume $\widehat{T}(h_{2,1}) = n(h_{2,1}^{\mathsf{T}}\widehat{\beta}_{2,1})^2/h_{2,1}^{\mathsf{T}}\widehat{\Sigma}_{21,21}h_{2,1}$ where $\widehat{\Sigma}_{21,21}$ is the submatrix of $(\mathbb{P}_n B_2 B_2^{\mathsf{T}})^{-1}\mathbb{P}_n B_2 B_2^{\mathsf{T}}(Y - B_2^{\mathsf{T}}\widehat{\beta}_{2,1})^2(\mathbb{P}_n B_2 B_2^{\mathsf{T}})^{-1}$ corresponding to the plug-in estimator of the asymptotic variance of $\mathbb{V}_n \triangleq \sqrt{n}(\widehat{\beta}_{2,1} - \beta_{2,1}^*)$.

The upper bound on $c^{\mathsf{T}}\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$ is given by

$$\mathcal{U}(c) \triangleq c^{\mathsf{T}}\mathbb{S}_n + c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n 1_{\widehat{T}(H_{2,1})>\lambda_n}$$
$$+ \sup_{\gamma \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \left(\left[H_{2,1}^{\mathsf{T}}(\mathbb{V}_n + \gamma)\right]_+ - \left[H_{2,1}^{\mathsf{T}}\gamma\right]_+ \right) 1_{\widehat{T}(H_{2,1})\leq\lambda_n}. \quad (6)$$

A lower bound, say $\mathcal{L}(c)$, is obtained by replacing sup with inf in the above display. The intuition behind this upper bound is as follows. Notice that the second term in (2), namely $c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n$, is equal to $c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n 1_{\widehat{T}(H_{2,1})>\lambda_n} + c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n 1_{\widehat{T}(H_{2,1})\leq\lambda_n}$. Rewrite the $\mathbb{U}_n$ in $c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n 1_{\widehat{T}(H_{2,1})\leq\lambda_n}$ as $[H_{2,1}^{\mathsf{T}}(\mathbb{V}_n + \sqrt{n}\beta_{2,1}^*)]_+ - [H_{2,1}^{\mathsf{T}}\sqrt{n}\beta_{2,1}^*]_+$. Thus $c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n$, is equal to

$$c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n 1_{\widehat{T}(H_{2,1})>\lambda_n}$$
$$+ c^{\mathsf{T}}\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \left(\left[H_{2,1}^{\mathsf{T}}(\mathbb{V}_n + \sqrt{n}\beta_{2,1}^*)\right]_+ - \left[H_{2,1}^{\mathsf{T}}\sqrt{n}\beta_{2,1}^*\right]_+ \right) 1_{\widehat{T}(H_{2,1})\leq\lambda_n}. \quad (7)$$

The quantity, $[H_{2,1}^{\intercal}\sqrt{n}\beta_{2,1}^{*}]_{+}$ characterizes the degree of nonregularity of $\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ (see Theorem 4.2 below). Replacing $\sqrt{n}\beta_{2,1}^{*}$ with γ and taking the supremum over all $\gamma \in \mathbb{R}^{\dim(\beta_{2,1}^*)}$ is one way of making the second term in (7) insensitive to local perturbations of $\beta_{2,1}^{*}$.

To use the bounds to construct a $(1-\alpha) \times 100\%$ confidence interval for $c^{\intercal}\beta_1^*$, first note that $c^{\intercal}\hat{\beta}_1 - \mathcal{U}(c)/\sqrt{n} \leq c^{\intercal}\beta_1^* \leq c^{\intercal}\hat{\beta}_1 - \mathcal{L}(c)/\sqrt{n}$. We approximate the distribution of the bounds using the nonparametric bootstrap. Let $\hat{u}$ denote the $(1 - \alpha/2) \times 100$ percentile of the bootstrap distribution of $\mathcal{U}(c)$, and let $\hat{l}$ denote the $(\alpha/2) \times 100$ percentile of the bootstrap distribution of $\mathcal{L}(c)$. Then, $[c^{\intercal}\hat{\beta}_1 - \hat{u}\sqrt{n}, c^{\intercal}\hat{\beta}_1 - \hat{l}/\sqrt{n}]$ is the proposed confidence interval for $c^{\intercal}\beta_1^*$. We term this confidence interval an adaptive confidence interval (ACI) for reasons that will become clear shortly.

Remark. In the ACI λ_n is a potentially important tuning parameter. In Section 5 we demonstrate that the double bootstrap is an effective strategy for constructing a data-driven choice of λ_n.

4.2.1. *Theoretical results*

In this section we describe the limiting behavior of the bounds $\mathcal{L}(c)$ and $\mathcal{U}(c)$ and relate them to the limiting distribution of $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$. We assume:

(A4) With probability one the sequence λ_n satisfies $\lambda_n \to \infty$ and $\lambda_n/n \to 0$ as $n \to \infty$.

Theorem 4.1 (Validity of population bounds). *Assume (A1)–(A2) and (A4) and fix $c \in \mathbb{R}^{\dim(\beta_1^*)}$.*

1. *$c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \beta_1^*) \rightsquigarrow c^{\intercal}\mathbb{S}_{\infty} + c^{\intercal}\Sigma_{1,\infty}^{-1}P(B_1 H_{2,1}^{\intercal}\mathbb{V}_{\infty}1_{H_{2,1}^{\intercal}\beta_{2,1}^*>0}) + c^{\intercal}\Sigma_{1,\infty}^{-1}PB_1[H_{2,1}^{\intercal}\mathbb{V}_{\infty}]_{+}1_{H_{2,1}^{\intercal}\beta_{2,1}^*=0}.$*

2. *If for each n, the underlying generative distribution is P_n, which satisfies (A3), then the limiting distribution of $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \beta_{1,n}^*)$ is equal to*

$$c^{\intercal}\mathbb{S}_{\infty} + c^{\intercal}\Sigma_{1,\infty}^{-1}P\left(B_1 H_{2,1}^{\intercal}\mathbb{V}_{\infty}1_{H_{2,1}^{\intercal}\beta_{2,1}^*>0}\right)$$
$$+ c^{\intercal}\Sigma_{1,\infty}^{-1}P\left[B_1\left(\left[H_{2,1}^{\intercal}(\mathbb{V}_{\infty}+s)\right]_{+} - \left[H_{2,1}^{\intercal}s\right]_{+}\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^*=0}\right]. \quad (8)$$

3. *The limiting distribution of $\mathcal{U}(c)$ under both P and under P_n is equal to*

$$c^{\intercal}\mathbb{S}_{\infty} + c^{\intercal}\Sigma_{1,\infty}^{-1}P\left(B_1 H_{2,1}^{\intercal}\mathbb{V}_{\infty}1_{H_{2,1}^{\intercal}\beta_{2,1}^*>0}\right)$$
$$+ \sup_{\gamma \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} c^{\intercal}\Sigma_{1,\infty}^{-1}P\left[B_1\left(\left[H_{2,1}^{\intercal}(\mathbb{V}_{\infty}+\gamma)\right]_{+} - \left[H_{2,1}^{\intercal}\gamma\right]_{+}\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^*=0}\right],$$
$$(9)$$

where $(\mathbb{S}_{\infty}^{\intercal}, \mathbb{V}_{\infty}^{\intercal})$ is asymptotically multivariate normal with mean zero.

See the Appendix for a proof and the formula for the $\mathrm{Cov}(\mathbb{S}_\infty, \mathbb{V}_\infty)$. Notice that limiting distributions of $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ and $\mathcal{U}(c)$ (or equivalently $\mathcal{L}(c)$) are equal in the case $H_{2,1}^\intercal \beta_{2,1}^* \neq 0$ with probability one. That is, when there is a large treatment effect for almost all patients then the upper (or lower) bound is tight. However, when there is a non-null subset of patients for whom there is no treatment effect, then the limiting distribution of the upper bound is stochastically larger than the limiting distribution of $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$. Thus, the ACI adapts to the setting in which all patients experience a treatment effect.

Because the distribution of (8) depends on the local alternative, s, $\hat{\beta}_1$ is a nonregular estimator (Van der Vaart and Wellner, 1996). One might hope to construct an estimator of the distribution of (8) and use this estimator to approximate the distribution of $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$. However, a consistent estimator of the distribution of (8) does not exist because P_n is contiguous with respect to P (by assumption (A3)). To see this, let $F_s(u)$ be the distribution of (8) evaluated at a point, u. If a consistent estimator, say $\hat{F}_n(u)$, existed, that is $\hat{F}_n(u)$ converges in probability to $F_s(u)$ under P_n, then the contiguity implies that $\hat{F}_n(u)$ converges in probability to $F_s(u)$ under P. This is a contradiction (at best $\hat{F}_n(u)$ converges in probability to $F_0(u)$ under P). Because we cannot consistently estimate s and we do not know the value of s, the tightest estimable upper bound on (8) is given by (9). As we shall next see, we are able to consistently estimate the distribution of (9).

In order to form confidence sets, the bootstrap distributions of $\mathcal{U}(c)$ and $\mathcal{L}(c)$ are used. The next result regards the consistency of these bootstrap distributions. Let $\hat{\mathbb{P}}_n^{(b)}$ denote the bootstrap empirical measure, that is, $\hat{\mathbb{P}}_n^{(b)} \triangleq n^{-1} \sum_{i=1}^n M_{n,i} \delta_{\mathcal{T}_i}$ for $(M_{n,1}, M_{n,2}, \ldots, M_{n,n}) \sim \mathrm{Multinomial}(n, (1/n, \ldots, 1/n))$. We use the superscript (b) to denote that a functional has been replaced by its bootstrap analog, so that if $\omega \triangleq f(\mathbb{P}_n)$ then $w^{(b)} \triangleq f(\hat{\mathbb{P}}_n^{(b)})$. Denote the space of bounded Lipschitz-1 functions on $\mathbb{R}^2$ by $BL_1(\mathbb{R}^2)$. Furthermore, let $\mathbb{E}_M$ and P_M denote the expectation and probability with respect to the bootstrap weights. The following results are proved in the Appendix.

Theorem 4.2. *Assume (A1)–(A2), (A4) and fix $c \in \mathbb{R}^{\dim(\beta_1^*)}$. Then $(\mathcal{U}(c), \mathcal{L}(c))$ and $(\mathcal{U}^{(b)}(c), \mathcal{L}^{(b)}(c))$ converge to the same limiting distribution in probability. That is,*

$$\sup_{v \in BL_1(\mathbb{R}^2)} \left| \mathbb{E}v\left((\mathcal{U}(c), \mathcal{L}(c))\right) - \mathbb{E}_M v\left(\left(\mathcal{U}^{(b)}(c), \mathcal{L}^{(b)}(c)\right)\right) \right|$$

converges in probability to zero.

Corollary 4.3. *Assume (A1)–(A2), (A4) and fix $c \in \mathbb{R}^{\dim(\beta_1^*)}$. Let $\hat{u}$ denote the $(1 - \alpha/2) \times 100$ percentile of $\mathcal{U}^{(b)}(c)$ and $\hat{l}$ denote the $(\alpha/2) \times 100$ percentile of $\mathcal{L}^{(b)}(c)$. Then*

$$P_M\left(c^\intercal \hat{\beta}_1 - \hat{u}/\sqrt{n} \leq c^\intercal \beta_1^* \leq c^\intercal \hat{\beta}_1 - \hat{l}/\sqrt{n}\right) \geq 1 - \alpha + o_P(1).$$

Furthermore, if $P(H_{2,1}^\mathsf{T}\beta_{2,1}^* = 0) = 0$, then the above inequality can be strengthened to equality.

The preceding results show that the ACI can be use to construct valid confidence intervals regardless of the underlying parameters or generative model. Moreover, in settings where there is a treatment effect for almost every patient, the ACI delivers asymptotically exact coverage. See Section 5 for discussion of the choice of the tuning parameter λ_n.

5. Experiments

In this section we examine the small sample performance of the adaptive confidence interval (ACI) proposed in the Section 4.2 where performance is measured in terms of coverage and average interval width. We consider both fixed and data-driven choices for the tuning parameter λ_n. For a fixed value we choose $\lambda_n = \sqrt{\log \log n}$; additional simulations taken over a range of λ_n values are provided in the Appendix. These simulations show that the method is potentially sensitive to the choice of λ_n. Consequently, we also consider a data-driven choice of λ_n, tuned using the double-bootstrap (Davison and Hinkley, 1997). In particular, we consider a range of values of λ_n of the form $\lambda_n = \tau\sqrt{\log \log n}$ where $\tau \in [m, M]$ where $0 < m < M < \infty$ are fixed constants. See the Appendix for the specifics of the double bootstrap algorithm. Note that the theoretical properties of the ACI continue to hold with this adaptive scheme for choosing λ_n since $m\sqrt{\log \log n} \le \tau\sqrt{\log \log n} \le M\sqrt{\log \log n}$ so that λ_n satisfies (A4).

We compare the empirical performance of the ACI with λ_n fixed to equal $\sqrt{\log \log n}$ (FACI) and λ_n chosen using the double-bootstrap (DACI) with the following methods: the centered percentile bootstrap (CPB); the centered percentile bootstrap of the soft-thresholding (ST) method of Chakraborty et al. (2009) as described in Section 3; and the adaptive m-out-of-n (MOFN) bootstrap with data-driven tuning of Chakraborty et al. (2013). We also implemented and tested the projection interval described in Section 4.1 with $\eta = 0.01, \alpha = 0.04$; results are not shown in the tables as they were too wide to be useful. The projection interval always covered at least at the nominal level (and frequently much more – in 6 of 18 experiments it covered 100% of the time) but it was between 1.46 and 2.07 times wider than the DACI, which also achieves or exceeds nominal coverage. The hard-thresholding method of Moodie et al. (2010) and the penalized approach of Song et al. (2011) are similar in both theory and performance to the soft-thresholding approach and thus are omitted from our experiments.

Nine generative models are used in these evaluations; each of these generative models has two stages of treatment and two treatments at each stage. Generically, each of the models can be described as follows:

- $X_t \in \{-1, 1\}$, $A_t \in \{-1, 1\}$ for $t \in \{1, 2\}$
- $P(A_1 = 1) = P(A_1 = -1) = 0.5$, $P(A_2 = 1) = P(A_2 = -1) = 0.5$
- $X_1 \sim \text{Bernoulli}(0.5)$, $X_2|X_1, A_1 \sim \text{Bernoulli}(\text{expit}(\delta_1 X_1 + \delta_2 A_1))$
- $Y = \gamma_1 + \gamma_2 X_1 + \gamma_3 A_1 + \gamma_4 X_1 A_1 + \gamma_5 A_2 + \gamma_6 X_2 A_2 + \gamma_7 A_1 A_2 + \epsilon, \epsilon \sim N(0, 1)$

TABLE 1
Parameters indexing the example models

Example	γ	δ	Type	Regularity Measures	
1	$(0,0,0,0,0,0,0)^{\intercal}$	$(0.5,0.5)^{\intercal}$	nonregular	$p=1$	$\phi=0/0$
2	$(0,0,0,0,0.01,0,0)^{\intercal}$	$(0.5,0.5)^{\intercal}$	near-nonregular	$p=0$	$\phi=\infty$
3	$(0,0,-0.5,0,0.5,0,0.5)^{\intercal}$	$(0.5,0.5)^{\intercal}$	nonregular	$p=1/2$	$\phi=1.0$
4	$(0,0,-0.5,0,0.5,0,0.49)^{\intercal}$	$(0.5,0.5)^{\intercal}$	near-nonregular	$p=0$	$\phi=1.02$
5	$(0,0,-0.5,0,1.0,0.5,0.5)^{\intercal}$	$(1.0,0.0)^{\intercal}$	nonregular	$p=1/4$	$\phi=1.41$
6	$(0,0,-0.5,0,0.25,0.5,0.5)^{\intercal}$	$(0.1,0.1)^{\intercal}$	regular	$p=0$	$\phi=0.35$
A	$(0,0,-0.25,0,0.75,0.5,0.5)^{\intercal}$	$(0.1,0.1)^{\intercal}$	regular	$p=0$	$\phi=1.035$
B	$(0,0,0,0,0.25,0,0.25)^{\intercal}$	$(0,0)^{\intercal}$	nonregular	$p=1/2$	$\phi=1.00$
C	$(0,0,0,0,0.25,0,0.24)^{\intercal}$	$(0,0)^{\intercal}$	near-nonregular	$p=0$	$\phi=1.03$

where $\operatorname{expit}(x) = e^x/(1+e^x)$. This class is parameterized by nine values $\gamma_1, \gamma_2, \ldots, \gamma_7, \delta_1, \delta_2$. The analysis model uses feature vectors defined by:

$$H_{2,0} = (1, X_1, A_1, X_1 A_1, X_2)^{\intercal}, \quad H_{2,1} = (1, X_2, A_1)^{\intercal},$$
$$H_{1,0} = (1, X_1)^{\intercal}, \qquad\qquad\qquad H_{1,1} = (1, X_1)^{\intercal}.$$

Our analysis models are given by $Q_2(H_2, A_2; \beta_2) \triangleq H_{2,0}^{\intercal}\beta_{2,0} + H_{2,1}^{\intercal}\beta_{2,1}A_2$ and $Q_1(H_1, A_1; \beta_1) \triangleq H_{1,0}^{\intercal}\beta_{1,0} + H_{1,1}^{\intercal}\beta_{1,1}A_1$. Below the analysis models are correctly specified (match the generative models). This avoids conflating poor performance of confidence intervals due to misspecification with poor performance due to nonregularity. We use a contrast encoding for A_1 and A_2 to allow for a comparison with Chakraborty et al. (2009).

The form of this class of generative models is useful as it allows us to influence the degree of nonregularity present in our example problems through the choice of the γ and δ, and in turn evaluate performance in these different scenarios. Recall that in Q-learning, nonregularity occurs when more than one stage-two treatment produces nearly the same optimal expected reward for a set of patient histories that occur with positive probability. In the model class above, this occurs if the model generates histories for which $\gamma_5 A_2 + \gamma_6 X_2 A_2 + \gamma_7 A_1 A_2 \approx 0$, i.e., if it generates histories for which Q_2 depends weakly or not at all on A_2. By manipulating the values of γ_i and δ_i, we can control i) the probability of generating a patient history such that $\gamma_5 A_2 + \gamma_6 X_2 A_2 + \gamma_7 A_1 A_2 = 0$, and ii) a standardized effect size $E[(\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1)/\sqrt{\operatorname{Var}(\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1)}]$. Each of these quantities, denoted by p and ϕ, respectively, can be thought of as measures of nonregularity.

Table 1 provides the parameter settings; the first six settings were considered by Chakraborty et al. (2009), and are described by them as "nonregular", "near-nonregular", and "regular". To these six, we have added three additional examples labeled A, B, and C. Example A is an example of a strongly regular setting. Example B is an example of a nonregular setting where the nonregularity is strongly dependent on the stage 1 treatment. In example B, for histories with $A_1 = 1$, there is a moderate effect of A_2 at the second stage. However, for histories with $A_1 = -1$, there is no effect of A_2 at the second stage, i.e., both treatments at the second stage are equally optimal. In example C, for histories with $A_1 = 1$, there is a moderate effect of A_2, and for histories with $A_1 = -1$,

TABLE 2

Monte Carlo estimates of coverage probabilities of confidence intervals for the main effect of treatment, $\beta^*_{1,1,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with ∗. There is no ST or MOFN method when there are three treatments at Stage 2. Examples are designated NR = nonregular, NNR = near-nonregular, R = regular

Two txts at stage 2	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B NR	Ex. C NNR
CPB	0.934*	0.935*	0.930*	0.933*	0.938	0.928*	0.939	0.925*	0.928*
FACI	0.989	0.987	0.967	0.969	0.954	0.952	0.950	0.962	0.962
DACI	0.968	0.971	0.958	0.961	0.949	0.943	0.949	0.953	0.953
MOFN	0.965	0.966	0.957	0.958	0.952	0.945	0.949	0.954	0.959
ST	0.948	0.945	0.938	0.942	0.952	0.943	0.919*	0.759*	0.762*

All three of the FACI, DACI, and MOFN methods deliver nominal coverage on all of the examples. The FACI in particular is conservative on examples one and two. The average interval diameters are shown in Table 3; this is to be expected given that it is based on upper and lower bounds. However, we note that the DACI, whose λ_n is tuned using the double bootstrap, has a much smaller width than the FACI, particularly in the three-treatment examples. It is the narrowest among the methods that cover in all examples.

Three txts at stage 2	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B NR	Ex. C NNR
CPB	0.933*	0.938	0.915*	0.921*	0.931*	0.907*	0.940	0.885*	0.895*
FACI	0.999	0.999	0.967	0.968	0.963	0.969	0.958	0.969	0.969*
DACI	0.987	0.987	0.952	0.955	0.957	0.945	0.953	0.940	0.945

there is a small effect of A_2. Thus example C is a 'near-nonregular' setting that behaves similarly to example B. In addition to these new examples, we give extensions of all nine examples to a setting with three treatments at the second stage; details are given in Appendix C.

We first provide confidence intervals for the coefficient of A_1 (the treatment variable), $\beta^*_{1,1,1}$ in settings in which there are two or three treatments at stage 2. (The three-treatment version of the ACI is given by Laber et al. (2010).) Note that given the working models and generative models defined by the parameter settings in Table 10, we can determine the exact value of any parameter $c^\mathsf{T}\beta^*_1$ of interest to set the ground truth for our experiments. Table 2 shows the estimated coverage for the coefficient of A_1, $\beta^*_{1,1,1}$. This simulation uses a sample size of 150, a total of 1000 Monte Carlo replications and 1000 bootstrap samples. Target coverage is 0.95. The CPB fares poorly in terms of coverage, falling significantly below nominal coverage on seven of nine examples. The ST method fails to cover for examples A, B and C. Recall that the ST method has not been developed for the setting in which there are more than two treatments at the second stage.

The coefficient of A_1 is perhaps most relevant from a clinical perspective. However, from a methodological point of view, other contrasts can be illuminating. Table 4 shows the estimated coverage for the intercept using the same generative models. The coverage of the CPB and ST methods is quite poor; the CPB attains nominal coverage on only two of the nine examples, and the ST never achieves nominal coverage. Particularly disturbing is that the ST method

TABLE 3

Monte Carlo estimates of the mean width of confidence intervals for the main effect of treatment $\beta^*_{1,1,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Models have two treatments at each of two stages. Widths with corresponding coverage significantly below nominal are marked with $*$. There is no ST or MOFN method when there are three treatments at Stage 2. Examples are designated NR = nonregular, NNR = near-nonregular, R = regular

Two txts at stage 2	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B NR	Ex. C NNR
CPB	0.385*	0.385*	0.430*	0.430*	0.457	0.436*	0.451	0.428*	0.428*
FACI	0.490	0.490	0.481	0.481	0.483	0.471	0.474	0.484	0.484
DACI	0.442	0.441	0.470	0.470	0.482	0.469	0.474	0.473	0.473
MOFN	0.443	0.443	0.474	0.474	0.489	0.486	0.482	0.488	0.488
ST	0.339	0.339	0.426	0.427	0.469	0.436	0.480*	0.426*	0.424*

Three txts at stage 2	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B NR	Ex. C NNR
CPB	0.446*	0.446	0.518*	0.518*	0.567*	0.518*	0.557	0.508*	0.507*
FACI	0.700	0.700	0.652	0.652	0.637	0.632	0.617	0.661	0.662
DACI	0.564	0.564	0.590	0.590	0.617	0.591	0.604	0.596	0.597

TABLE 4

Monte Carlo estimates of coverage probabilities of confidence intervals for the coefficient of the intercept, $\beta^*_{1,0,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Examples are designated NR = nonregular, NNR = near-nonregular, R = regular

Two txts at stage 2	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B NR	Ex. C NNR
CPB	0.892*	0.908*	0.924*	0.925*	0.940	0.930*	0.936	0.925*	0.931*
FACI	0.952	0.962	0.952	0.954	0.950	0.953	0.947	0.952	0.954
DACI	0.940	0.946	0.946	0.948	0.947	0.945	0.951	0.952	0.947
MOFN	0.944	0.947	0.948	0.948	0.952	0.942	0.951	0.950	0.950
ST	0.935*	0.930*	0.889*	0.878*	0.891*	0.620*	0.687*	0.686*	0.663*

falls more than 30% below nominal levels. In contrast, the FACI and DACI deliver nominal coverage on all examples. Table 5 shows the average interval widths; the DACI is the narrowest among the covering methods.

6. Analysis of the ADHD study

In this section we illustrate the use of the ACI on data from the Adaptive Pharmacological and Behavioral Treatments for Children with ADHD Trial (Nahum-Shani et al. 2012a; Lei et al. 2012). The ADHD data we use here consists of $n = 138$ trajectories which are a subset of the original $N = 155$ observations. This subset was formed by removing the $N - n = 17$ subjects who were either never randomized to an initial treatment (14 subjects), or had massive item missingness (3 subjects). A description of each of the variables is provided in Table 6. Notice that the outcomes Y_1 and Y_2 satisfy $Y_1 + Y_2 \equiv Y$, where Y is the teacher reported TIRS5 score after 32 weeks, i.e. at the end of the last month of the study (month 8).

TABLE 5

*Monte Carlo estimates of the mean width of confidence intervals for the coefficient of the
intercept, $\beta^*_{1,0,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of
size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Models have
two treatments at each of two stages. Widths with corresponding coverage significantly below
nominal are marked with ∗. Examples are designated NR = nonregular, NNR =
near-nonregular, R = regular*

Two txts at stage 2	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B NR	Ex. C NNR
CPB	0.404*	0.404*	0.430*	0.429*	0.457	0.449*	0.450	0.428*	0.428*
FACI	0.506	0.506	0.481	0.481	0.483	0.490	0.474	0.490	0.490
DACI	0.459	0.459	0.466	0.466	0.481	0.482	0.473	0.473	0.473
MOFN	0.475	0.476	0.469	0.470	0.488	0.486	0.477	0.483	0.483
ST	0.344*	0.344*	0.427*	0.427*	0.466*	0.469*	0.474*	0.430*	0.428*

The first step in using Q-learning is to estimate a regression model for the second stage; this analysis only uses data from subjects that were re-randomized during the 8 month study. Of the $n = 138$ subjects, 81 were re-randomized prior to the end of the study. The feature vectors at the second stage are $H_{2,0} \triangleq (1, X_{1,1}, X_{1,2}, X_{1,2}, X_{1,3}, X_{2,1}, A_1)^{\mathsf{T}}$ and $H_{2,1} \triangleq (1, X_{2,1}, A_1)^{\mathsf{T}}$. Thus, the Q-function $Q_2(H_2, A_2; \beta_2) \triangleq H_{2,0}^{\mathsf{T}}\beta_{2,0} + H_{2,1}^{\mathsf{T}}\beta_{2,1}A_2$ contains an interaction term between the second stage action A_2 and a subject's initial treatment A_1, an interaction between A_2 and adherence to their initial medication $X_{2,1}$, a main effect for A_2, and main effects for all the other terms. Table 7 provides the second stage least squares coefficients along with centered percentile bootstrap interval estimates. Examination of the residuals (not shown here) showed no obvious signs of model misspecification. In short, the linear model described above seems to fit the data reasonably well.

Recall that the dependent variable in the first stage regression model is the predicted future outcome $\tilde{Y}_1 \triangleq Y_1 + \max_{a_2 \in \{-1,1\}} Q_2(H_2, a_2; \hat{\beta}_2)$. Since the predictors used in the first stage must predate the assignment of first treatment, the available predictors in Table 6 are baseline ADHD symptoms $X_{1,1}$, diagnosis of ODD at baseline $X_{1,2}$, indicator of a subject's prior exposure to ADHD medication $X_{1,3}$, and first stage treatment A_1. The feature vectors for the second stage are $H_{1,0} \triangleq (1, X_{1,1}, X_{1,2}, X_{1,3})$ and $H_{1,1} \triangleq (1, X_{1,3})$, so that the first stage Q-function $Q_1(H_1, A_1; \beta_1) \triangleq H_{1,0}^{\mathsf{T}}\beta_{1,0} + H_{1,1}^{\mathsf{T}}\beta_{1,1}A_1$ contains an interaction term between the first stage action A_1 and a subject's prior exposure to ADHD medication $X_{1,3}$, a main effect for A_1, and main effects for all other covariates. The first stage regression coefficients are estimated using least squares $\hat{\beta}_1 \triangleq \arg\min_{\beta_1} \mathbb{P}_n(\tilde{Y}_1 - Q_1(H_1, A_1; \beta_1))^2$. Table 8 provides the least squares coefficients along with interval estimates formed using the DACI. Plots of the residuals for this model (not shown here) show no obvious signs of model misspecification. Again a linear model seems to provide a reasonable approximation to the Q-function in the first stage.

To construct an estimate of the optimal DTR, recall that for any $H_t = h_t$, $t = 1, 2$ the estimated optimal DTR $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2)$ satisfies $\hat{\pi}_t(h_t) \in \arg\max_{a_t} Q(h_t, a_t; \hat{\beta}_t)$. The coefficients in Table 7 and the form of the second stage Q-function re-

TABLE 6
Features, treatments and the outcome for the ADHD study

$X_{1,1} \in [0,3]$	:	Baseline symptoms. Teacher-reported mean ADHD symptom score. Measured at the end of the school year preceding the study.
$X_{1,2} \in \{0,1\}$	:	ODD diagnosis. Indicator of a diagnosis of ODD (oppositional defiant disorder) at baseline, coded so that 0 corresponds to no such diagnosis.
$X_{1,3} \in \{0,1\}$	:	Prior med. exposure. Indicator that subject received ADHD medication in the prior year, coded so that 0 corresponds to no ADHD medication.
$A_1 \in \{-1,1\}$	:	1st stage treatment. Coded so that -1 corresponds to medication while 1 corresponds to behavioral modification therapy.
1_{NonRsp}	:	Indicator of non-response, i.e. that a patient was re-randomized to a second-stage treatment during the study. Non-response was determined on the basis of two measures the Impairment Rating Scale (IRS) (Fabiano et al. 2006) and an individualized list of target behaviors (ITB) (e.g., Pelham et al. 1992). The criterion for nonresponse at each month was an average performance of less than 75 on the ITB and a rating of impairment in at least one domain on the IRS. These were measured beginning in week 8 of the study, and montly thereafter.
$Y_1 \triangleq Y \cdot (1 - 1_{\mathrm{NonRsp}})$	:	First stage outcome of responders, i.e. those who were *not* re-randomized (see definition of Y and $\tilde{Y}$ below).
$X_{2,1} \in \{0,1\}$	:	Adherence. Indicator of subject's adherence to their initial treatment. Adherence is coded so that a value of 0 corresponds to low adherence (taking less than 100% of prescribed medication or attending less than 75% of therapy sessions) while a value of 1 corresponds to high adherence.
$X_{2,2} \in \{2,8\}$	:	Month of non-response. Month during school year of observed non-response and re-randomization (not used for responders) Two subjects did not follow protocol and were re-randomized during month 8.
$A_2 \in \{-1,1\}$	:	2nd stage treatment. Coded so that $A_2 = -1$ corresponds to augmenting the initial treatment with the treatment *not* received initially, and $A_2 = 1$ corresponds to enhancing (increasing the dosage of) the initial treatment.
$Y \in \{1,2,\ldots,5\}$	:	Teacher-reported Impairment Rating Scale - Effect on Classroom(TIRS5) item score 8 months (32 weeks) after initial randomization to treatment (Fabiano et al. 2006). The TIRS5 is coded so that higher values correspond to better clinical outcomes.
$Y_2 \triangleq Y \cdot 1_{NonRsp}$	:	Second stage outcome. Only used for non-responders, i.e. subjects who *were* re-randomized.

veal that the second stage decision rule $\hat{\pi}_2$ is quite simple. In particular, $\hat{\pi}_2$ prescribes treatment enhancement to subjects with high adherence to their initial medication and it prescribes treatment augmentation to subjects with low adherence to their initial medication. The first stage decision rule $\hat{\pi}_1$ is equally simplistic. The coefficients in Table 8 show that the first stage decision rule, $\hat{\pi}_1$ prescribes medication to subjects who have had prior exposure to medication, and behavioral modification to subjects who have not had any such prior exposure.

TABLE 7
Least squares coefficients and 90% CPB interval estimates for second stage regression

Term		Coeff.	Estimate	Lower (5%)	Upper (95%)
1	Intercept	$\beta_{2,0,1}$	1.36	0.48	2.26
$X_{1,1}$	Baseline symptoms	$\beta_{2,0,2}$	0.94	0.48	1.39
$X_{1,2}$	ODD diagnosis	$\beta_{2,0,3}$	0.92	0.46	1.41
$X_{1,3}$	Prior med. exposure	$\beta_{2,0,4}$	-0.27	-0.77	0.21
$X_{2,1}$	Adherence	$\beta_{2,0,5}$	0.17	-0.28	0.66
$X_{2,2}$	Month of non-response	$\beta_{2,0,6}$	0.02	-0.20	0.20
A_1	1st stage txt	$\beta_{2,0,7}$	0.03	-0.18	0.23
A_2	2nd stage txt	$\beta_{2,1,1}$	-0.72	-1.13	-0.35
$A_2 : X_{2,1}$	2nd stage txt : Adherence	$\beta_{2,1,2}$	0.97	0.48	1.52
$A_2 : A_1$	2nd stage txt : 1st stage txt	$\beta_{2,1,3}$	0.05	-0.17	0.27

TABLE 8
Least squares coefficients and 90% DACI interval estimates for first stage regression

Term		Coeff.	Estimate	Lower (5%)	Upper (95%)
1	Intercept	$\beta_{1,0,1}$	2.61	2.13	3.05
$X_{1,1}$	Baseline symptoms	$\beta_{1,0,2}$	0.72	0.47	1.00
$X_{1,2}$	ODD diagnosis	$\beta_{1,0,3}$	0.75	0.37	1.08
$X_{1,3}$	Prior med. exposure	$\beta_{1,0,4}$	-0.37	-0.80	0.01
A_1	Initial txt	$\beta_{1,1,1}$	0.17	-0.02	0.36
$A_1 : X_{1,3}$	Initial txt : Prior med. exposure	$\beta_{1,1,2}$	-0.32	-0.59	-0.07

The prescriptions given by the estimated optimal DTR $\hat{\pi}$ are excessively decisive. That is, they recommend one and only one treatment regardless of the amount of evidence in the data to support that the recommended treatment is in fact optimal. When there is insufficient evidence to recommend a single treatment as best for a given patient history, it is preferred to leave the choice of treatment to the clinician. This allows the clinician to recommend treatment based on cost, local availability, patient individual preference, and clinical experience. One way to assess if there is sufficient evidence to recommend a unique optimal treatment for a patient is to construct a confidence interval for the predicted difference in mean response across treatments. In the case of binary treatments, for a fixed patient history $H_t = h_t$, one would construct a confidence interval for the difference $Q_t(h_t, 1; \beta_t^*) - Q_t(h_1, -1; \beta_t^*) = c^\intercal \beta_t^*$ where $c = (\mathbf{0}^\intercal, 2h_{t,1}^\intercal)^\intercal$. If this confidence interval contains zero then one would conclude that there is insufficient evidence at the nominal level for a unique best treatment.

In this example, the subject features that interact with treatment are categorical. Consequently, we can construct confidence intervals for the predicted difference in mean response across treatments for every possible subject history. These confidence intervals are given in Table 9. The 90% confidence intervals suggest that there is insufficient evidence at the first stage to recommend a unique best treatment for each subject history. Rather, we would prefer not to make a strong recommendation at stage one, and leave treatment choice solely at the discretion of the clinician. Conversely, in the second stage, the 90% confidence intervals suggest that there is evidence to recommend a unique best treatment when a subject had low adherence—knowledge that is important for evidence-based clinical decision making.

TABLE 9
Confidence intervals for the predicted difference in mean response across treatments for each possible patient history. Intervals are at the 90% leve. Confidence intervals that contain zero indicate insufficient evidence for recommending a unique best treatment for patients with the given history

Stage	History	Contrast for $\beta_{t,1}$	Lower (5%)	Upper (95%)	Conclusion
1	Had prior med.	(2 2)	-0.88	0.28	Insufficient evidence
1	No prior med.	(2 0)	-0.04	0.72	Insufficient evidence
2	High adherence and BMOD	(2 2 2)	-0.17	1.39	Insufficient evidence
2	Low adherence and BMOD	(2 0 2)	-2.21	-0.57	Sufficient evidence
2	High adherence and MEDS	(2 2 -2)	-0.37	1.26	Insufficient evidence
2	Low adherence and MEDS	(2 0 -2)	-2.51	-0.60	Sufficient evidence

7. Summary, open problems, and the future of DTRs

DTRs have the potential to produce better patient outcomes while simultaneously reducing cost and patient burden. Furthermore, estimated optimal DTRs can provide important scientific insight by revealing interactions between treatments and patient history and delayed treatment effects. As a result, SMART studies for use in constructing DTRs are of much interest in the clinical community. A large number of SMART studies are completed or currently in the field (see PSU Methodology Center, 2014b). Recent funding calls for SMART designs from the National Institutes of Health (NIH, PSU Methodology Center, 2014a) and a commitment to funding for personalized medicine from both the NIH and US Food and Drug Administration (Hamburg and Collins, 2010) suggest that the number of SMART studies will continue to grow.

As has been discussed here, the construction of optimal DTRs from data can involve the vexing problem of nonregularity. When the treatment effect is small relative to noise in the data, nonregularity can make it difficult to conduct basic statistical analyses essential to informing clinical practice or future research. We have shown that these difficulties arise even in the simple setting of two-stages of binary treatment and linear working models. We expect the impact of nonregularity to grow with both the number of treatment options as well as the number of treatment stages. Thus, new methods for estimation and inference, especially those intended for the many-stage, many-treatment setting, must account for the effects of nonregularity. Simulated experiments demonstrating the performance of methods for estimation of and inference for DTRs should include generative models with a mix of treatment effect sizes including those that are zero, small but nonzero, and far from zero, where proximity to zero is measured in terms of statistical power.

We discussed how nonregularity leads to asymptotic bias and complicates inference. It is natural to think that reducing the asymptotic bias (via shrinkage methods) in the estimated β's will assist in the construction of valid confidence

intervals; we conjecture that bias reduction techniques will be very difficult, if not impossible, to tune thus leading to situations in which over-shrinkage occurs. Indeed as shown here, such over-shrinkage can be infinitely worse than no shrinkage at all. Instead of attempting to reduce the bias, we used regular upper and lower bounds on $c^\mathsf{T}\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$ to form an adaptive confidence interval for first stage parameters in Q-learning. However, a potentially less conservative strategy would be to form bounds on the $(\alpha/2) \times 100$ and $(1 - \alpha/2) \times 100$ *percentiles* of the sampling distribution of $c^\mathsf{T}\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$. For example, one could define $\mathcal{B}(c, \gamma) = c^\mathsf{T}\mathbb{S}_n + c^\mathsf{T}\hat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n 1_{\hat{T}(H_{2,1}) > \lambda_n} + c^\mathsf{T}\hat{\Sigma}_1^{-1}\mathbb{P}_n B_1([H_{2,1}^\mathsf{T}(\mathbb{V}_n + \gamma)]_+ - [H_{2,1}^\mathsf{T}\gamma]_+)1_{\hat{T}(H_{2,1}) \leq \lambda_n}$. Then, for any fixed γ and level η one could use the bootstrap to estimate the $\eta \times 100$ percentile of $\mathcal{B}(c, \gamma)$, say, $\widehat{q}_\eta^{(b)}(\gamma)$. The final confidence interval would be

$$\left(c^\mathsf{T}\widehat{\beta}_1 - \sup_{\gamma \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} \widehat{q}_{1-\alpha/2}^{(b)}(\gamma), c^\mathsf{T}\widehat{\beta}_1 - \inf_{\gamma \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} q_{\alpha/2}^{(b)}(\gamma) \right).$$

See (Andrews, 2001a; Cheng, 2008) and references therein for bounding probabilities rather than statistics. It would be interesting to compare this approach with the ACI.

Technological advances are continually improving the efficiency with which both observational and SMART study data can be collected, stored, and accessed. DTR methodologies must adapt with these changes. Here we discuss three areas where current DTR methodology is insufficient. These areas present unique estimation, inference, and computational challenges.

Infinite horizon problems. In settings where number of treatment stages is large (e.g., hundreds or thousands) it may be appropriate to approximate the decision problems as having an infinite number of time points. An important area where such decision problems arise is mobile-health (mHealth) where interventions are delivered using smartphones or other mobile devices (see, for example, Kelly et al., 2012). Mobile devices present unprecedented opportunity for collecting patient information and delivering interventions *in situ*, and thereby potentially narrowing the so-called research-practice gap (Bickman et al., 2012). However, the breadth of opportunities presented by mHealth are matched by their technical challenges. As the number of decision points grows large it becomes infeasible to have separate models for the Q-function at each decision point, in this case additional structure, for example, that the generative model can be characterized as a stationary Markov Decision Process (MDP, Putterman, 1994), is useful. Existing methods for estimating an optimal DTR in the MDP setup (Sutton and Barto, 1998) are highly algorithmic and their statistical properties are largely unknown. There are tremendous opportunities for translating these algorithms into a statistical framework and characterizing their statistical properties, e.g., convergence rates and limiting distribution theory.

Feature Construction   An important open problem is the development of formal feature extraction and construction techniques for DTRs. In the ADHD

study X_1 contains more than 25 variables, some discrete and some continuous, and X_t, $t = 2, 3$ contains more than 40 measurements collected each month; thus, over the course of the eight month study the protocol dictated the collection of more then 360 measurements per subject. In general X_t, $t = 1, 2, 3$ will contain a large number of repeated measurements. The current state-of-the-art is that these measurements are summarized into low-dimensional summaries motivated by clinical judgment, exploratory analyses and convenience; this is certainly the case in the ADHD example. However, in many practical examples, thousands or more than thousands of sparsely observed and irregularly spaced measurements may be collected. By design, information is accumulating over time; if one uses linear models nested inside the sequence of treatments received, then the model size will grow exponentially in the number of treatment stages. Principled, i.e., data-driven, methods for feature construction and extraction are needed. On approach would be to extend dimensionality-reduction methods from machine learning (e.g., isomap, ICA, etc.) or functional data analysis (e.g., functional principle components) to DTRs.

Spatial decision processes   In some applications, for example, adaptive wild-life management, separate treatments must be administered across a series of spatial locations at each time point. The treatment assignment at one spatial location may affect the outcomes at neighboring locations. Furthermore, the total number of treatments than can be administered across all the spatial locations is often limited by budget or other resource constraints. Thus, it is not feasible to estimate a separate DTR at each spatial location but rather a single large DTR recommending treatments for all spatial locations simultaneously is needed. That is, a DTR in this setting is a sequence of functions mapping up-to-date information at all spatial locations to a treatment recommendation at every spatial location. Q-learning, as described, cannot be applied as the dimension of the model grows exponentially in the number of spatial locations. Suppose, for example, that there are S spatial locations, K treatment options available at each location, and a p-dimensional feature vector at each spatial location; a linear model with a main effect of feature, a main effect for treatment, and an interaction between treatments and features would contain $p \times K^S$ terms. Furthermore, even if the Q-functions were known exactly, simply computing the argmax over all K^S possibilities is computationally intractable for moderate values of S and K.

Appendix A: Outcome weighted learning

Recall that the value of a DTR π, $\mathbb{E}^\pi Y$, is the expected outcome of Y under the restriction that $A_t = \pi_t(H_t)$. For expositional simplicity, assume $P(A_t = 1|H_t) = 1/2$ and that Y is coded so that $Y \geq 0$ in this section. Then a change of measure implies that the value $\mathbb{E}^\pi Y = 4P(Y1_{A_1=\pi_1(H_1)}1_{A_2=\pi_2(H_2)})$; the empirical analog is $4\mathbb{P}_n(Y1_{A_1=\pi_1(H_1)}1_{A_2=\pi_2(H_2)})$. Note the resemblance to the classification rate. As in classification, directly maximizing the empirical value over a

class of DTRs is a discrete optimization problem and is usually computationally burdensome. Zhao et al. (2013) solve a concave relaxation of this problem by replacing the nonsmooth indicator functions with concave surrogates. Consider decision rules of the form $\pi_t(h_t) = 1_{h_{t,1}^\mathsf{T} \psi_{t,1} \geq 0}$ where $h_{t,1}$ is a known feature of h_t. Note that $1_{A_t = \pi_t(H_t)} = 1_{(2A_t - 1)H_{t,1}^\mathsf{T} \psi_{t,1} \geq 0}$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a concave function that satisfies $\phi(z) \leq k + 1_{z \geq 0}$ for all z where k is a constant. A version of the algorithm is as follows.

1. Stage 2 optimization: $\widehat{\psi}_{2,1} = \arg\max_{\psi_{2,1}} \mathbb{P}_n Y \phi((2A_2 - 1)H_{2,1}^\mathsf{T} \psi_{2,1})$.
2. Stage 1 optimization:
   $\widehat{\psi}_{1,1} = \arg\max_{\psi_{1,1}} \mathbb{P}_n Y 1_{(2A_2 - 1)H_{2,1}^\mathsf{T} \widehat{\psi}_{2,1} \geq 0} \phi((2A_1 - 1)H_{1,1}^\mathsf{T} \psi_{1,1})$.

The estimator of the optimal DTR is thus $\widehat{\pi}_t(h_t) = 1_{h_{t,1}^\mathsf{T} \widehat{\psi}_{t,1} \geq 0}$. For illustration we use $\phi(z) = 1 - (1-z)^2$. Define the population parameters:

$$\psi_{2,1}^* \triangleq \arg\min_{\psi_{2,1}} P\left[Y(1 - (2A_2 - 1)H_{2,1}^\mathsf{T} \psi_{2,1})^2 \right],$$

$$\psi_{1,1}^* \triangleq \arg\min_{\psi_{1,1}} P\left[Y 1_{(2A_2 - 1)H_{2,1}^\mathsf{T} \psi_{2,1}^* \geq 0}(1 - (2A_1 - 1)H_{1,1}^\mathsf{T} \psi_{1,1})^2 \right].$$

In addition, define $\Psi_1 \triangleq PYH_{1,1}H_{1,1}^\mathsf{T} 1_{(2A_2 - 1)H_{2,1}^\mathsf{T} \psi_{2,1}^* \geq 0}$ and the corresponding plugin estimator $\widehat{\Psi}_1 \triangleq \mathbb{P}_n Y H_{1,1}H_{1,1}^\mathsf{T} 1_{(2A_2 - 1)H_{2,1}^\mathsf{T} \widehat{\psi}_{2,1} \geq 0}$, which we assume is invertible. Then $\sqrt{n}(\widehat{\psi}_{1,1} - \psi_{1,1}^*) = \sqrt{n}\widehat{\Psi}_1^{-1} \mathbb{P}_n Y H_{1,1}(2A_1 - 1)1_{(2A_1 - 1)H_{2,1}^\mathsf{T} \widehat{\psi}_{2,1} \geq 0}(1 - (2A_1 - 1)H_{1,1}^\mathsf{T} \psi_{1,1}^*)$ which can be decomposed as

$$\mathbb{T}_n + \sqrt{n}\widehat{\Psi}_1^{-1} \mathbb{P}_n Y H_{1,1}(2A_1 - 1)(1 - (2A_1 - 1)H_{1,1}^\mathsf{T} \psi_{1,1}^*)\mathbb{L}_n,$$

where

$$\mathbb{T}_n = \widehat{\Psi}_1^{-1} \sqrt{n}(\mathbb{P}_n - P)$$
$$\left[Y H_{1,1}(2A_1 - 1)(1 - (2A_1 - 1)H_{1,1}^\mathsf{T} \psi_{1,1}^*)1_{(2A_2 - 1)H_{2,1}^\mathsf{T} \psi_{2,1}^* \geq 0} \right],$$

$$\mathbb{L}_n = 1_{(2A_2 - 1)H_{2,1}^\mathsf{T} \widehat{\psi}_{2,1} \geq 0} - 1_{(2A_2 - 1)H_{2,1}^\mathsf{T} \psi_{2,1}^* \geq 0}.$$

The term $\mathbb{T}_n$ is smooth and asymptotically normal under mild conditions whereas $\mathbb{L}_n$ is nonsmooth. If $h_{2,1}$ satisfies $h_{2,1}^\mathsf{T} \psi_{2,1}^* = 0$ then $\mathbb{L}_n\big|_{H_{2,1} = h_{2,1}}$ converges in distribution to a Bernoulli random variable with probability of success equal to $1/2$. On the other hand, if $h_{2,1}^\mathsf{T} \psi_{2,1}^* \neq 0$ then $\mathbb{L}_n\big|_{H_{2,1} = h_{2,1}}$ converges in probability to zero. Thus, in parallel with the Q-learning case, the limiting distribution of $\sqrt{n}(\widehat{\psi}_{1,1} - \psi_{1,1}^*)$ depends abruptly on both the value of $\psi_{2,1}^*$ and the distribution of $H_{2,1}$. Therefore the same theoretical challenges as in Q-learning occur in outcome-weighted learning.

Appendix B: Proofs

B.1. Proof of Theorems in Section 3

Lemma B.1. *If $\omega \sim \text{Normal}(0, \nu^2)$ then $\mathbb{E}[\omega]_+ = \nu/\sqrt{2\pi}$.*

Proof. Let ϕ denote the density of a standard normal random variable. Then

$$\mathbb{E}\left[\omega\right]_+ = \int_{\mathbb{R}} \left[\omega\right]_+ \phi(\omega/\nu)/\nu d\omega = \int_{\mathbb{R}_+} \omega\phi(\omega/\nu)/\nu d\omega = \nu/\sqrt{2\pi}.$$

$\square$

Proof of Theorem 3.1. Using Theorem 4.2, part I, it follows that $\mathrm{Bias}(\hat{\beta}_1, c)$ is equal to

$$\mathbb{E}\left(c^\intercal \Sigma_{1,\infty}^{-1} P B_1 \left[H_{2,1}^\intercal \mathbb{V}_\infty\right]_+ 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0}\right).$$

Exchanging expectations and applying Lemma B.1 gives the result. $\square$

Lemma B.2. *If $z \sim \mathrm{Normal}(0,1)$ and $\sigma > 0$ then*

$$\mathbb{E}\left[z\right]_+ \left(1 - \sigma/z^2\right)_+ = \frac{1}{\sqrt{2\pi}}\left\{\exp\{-\sigma/2\} - \sigma \int_{\sqrt{\sigma}}^\infty \exp{-z^2/2}/z dz\right\}.$$

Proof. Let ϕ denote the density of a standard normal random variable, then

$$\begin{aligned}
\mathbb{E}\left[z\right]_+ \left(1 - \frac{\sigma}{z^2}\right)_+ &= \int_{\sqrt{\sigma}}^\infty z\left(1 - \frac{\sigma}{z^2}\right)\phi(z)dz \\
&= \frac{1}{\sqrt{2\pi}}\left\{\exp\{-\sigma/2\} - \sigma \int_{\sqrt{\sigma}}^\infty \frac{1}{z}\exp(-z^2/2)dz\right\}.
\end{aligned}$$

$\square$

Proof of Theorem 3.2. Notice that $\sqrt{n}(\hat{\beta}_1^\sigma - \beta_1^*) = \widehat{\Sigma}_1^{-1}\sqrt{n}\mathbb{P}_n B_1(\tilde{Y}^\sigma - B_1^\intercal\beta_1^*)$ which can be decomposed as

$$\widehat{\Sigma}_1^{-1}\sqrt{n}(\mathbb{P}_n - P)B_1(\tilde{Y}^* - B_1^\intercal\beta_1^*) + \widehat{\Sigma}_1^{-1}\sqrt{n}\mathbb{P}_n B_1(\tilde{Y}^\sigma - \tilde{Y}^*),$$

where we have used $PB_1(\tilde{Y}^* - B_1^\intercal\beta_1^*) = 0$. The first term in the above display is asymptotically normal with mean zero and thus does not contribute to the asymptotic bias. The second term in the above display is equal to

$$\widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1 H_{2,0}^\intercal \sqrt{n}(\hat{\beta}_{2,0} - \beta_{2,0}^*)$$
$$+ \widehat{\Sigma}_1^{-1}\sqrt{n}\mathbb{P}_n B_1\left(\left[H_{2,1}^\intercal\hat{\beta}_{2,1}\right]_+\left(1 - \frac{\sigma H_{2,1}^\intercal\widehat{\Sigma}_{21,21}H_{2,1}}{n(\hat{\beta}_{2,1}^\intercal H_{2,1})^2}\right)_+ - \left[H_{2,1}^\intercal\beta_{2,1}^*\right]_+\right)1_{H_{2,1}^\intercal\beta_{2,1}^* \neq 0}$$
$$+ \widehat{\Sigma}_1^{-1}\mathbb{P}_n B_1\left[H_{2,1}^\intercal\sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*)\right]_+\left(1 - \frac{\sigma H_{2,1}^\intercal\widehat{\Sigma}_{21,21}H_{2,1}}{(H_{2,1}^\intercal\sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*))^2}\right)_+ 1_{H_{2,1}^\intercal\beta_{2,1}^* = 0}.$$

The first two terms can be shown to have asymptotic mean zero and thus they do not contribute the asymptotic bias. The last term converges in distribution to

$$\Sigma_{1,\infty}^{-1}P\left[B_1\left[\mathbb{Z}\right]_+ \sqrt{H_{2,1}^\intercal\Sigma_{21,21}H_{2,1}}\left(1 - \frac{\sigma}{\mathbb{Z}^2}\right)_+ 1_{H_{2,1}^\intercal\beta_{2,1}^* = 0}\right],$$

where $\mathbb{Z}$ is a standard normal random variable. Exchanging expectations and applying Lemma B.2 gives the result. $\square$

Proof of Theorem 3.3. From Theorem 4.2 part 2 it follows that $\text{Bias}(\hat{\beta}_1, c, s)$ is equal to

$$\mathbb{E}\left(c^\mathsf{T}\Sigma_{1,\infty}^{-1}PB_1\left(\left[H_{2,1}^\mathsf{T}(\mathbb{V}_\infty + s)\right]_+ - \left[H_{2,1}^\mathsf{T}s\right]_+\right)1_{H_{2,1}^\mathsf{T}\beta_{2,1}^* = 0}\right),$$

taking absolute values and applying the Cauchy-Schwarz and triangle inequalities gives the first result of the theorem.

It can be shown that $c^\mathsf{T}\sqrt{n}(\hat{\beta}_1^\sigma - \beta_1^*)$ converges in distribution to

$$c^\mathsf{T}\Sigma_{1,\infty}^{-1}PB_1\left(\left[H_{2,1}^\mathsf{T}(\mathbb{V}_\infty + s)\right]_+\left(1 - \frac{\sigma H_{2,1}^\mathsf{T}\Sigma_{21,21}H_{2,1}}{(H_{2,1}^\mathsf{T}(\mathbb{V}_\infty + s))^2}\right)_+ - \left[H_{2,1}^\mathsf{T}s\right]_+\right)1_{H_{2,1}^\mathsf{T}\beta_{2,1}^* = 0}.$$

Recall that $H_{2,1}$ is assumed to have an intercept. Let e_1 denote the first column of an $\dim(\beta_{2,1}^*) \times \dim(\beta_{2,1}^*)$ identity matrix, and choose $s = -\mathbb{V}_\infty + e_1 \log \sigma$ then as $\sigma \to \infty$ the above term behaves as

$$c^\mathsf{T}\Sigma_{1,\infty}^{-1}PB_11_{H_{2,1}^\mathsf{T}\beta_{2,1}^* = 0}\log(\sigma),$$

which tends to ∞ in magnitude. Thus, the supremum over s, of $|\text{Bias}(\hat{\beta}_1^\sigma, c, s)|$ must be at least as large. $\qquad\square$

B.2. Proof of Theorems in Section 4

In the main body we assumed a single terminal reward Y, here, to cover a more general case we assume that an intermediate reward, Y_1 may be observed at the end of the first stage as well as a terminal reward Y_2. Thus, one seeks to maximize $\mathbb{E}^\pi(Y_1 + Y_2)$ where $\mathbb{E}^\pi$ denotes expectation with respect to the joint distribution of the trajectory under the restriction that $A_t = \pi_t(H_t), t = 1, 2$. Throughout this section, let K denote a sufficiently large positive constant that may vary from line to line. Let D_p denote the space of $p \times p$ symmetric positive definite matrices equipped with the spectral norm, and for any $k \in (0, 1)$, let D_p^k denote the subset of D_p with members having eigenvalues in the range $[k, 1/k]$. For any class of real-valued functions $\mathcal{F}$, let $\rho_P(f) \triangleq (P(f - Pf)^2)^{1/2}$ denote the centered L_2-norm on $\mathcal{F}$, $l^\infty(\mathcal{F})$ denote the space of uniformly bounded real-valued functions on $\mathcal{F}$ equipped with the sup norm, and $C_b(\mathcal{F})$ denote the subspace of $l^\infty(\mathcal{F})$ of continuous and bounded functions from $\mathcal{F}$ into $\mathbb{R}$, respectively. Furthermore, let $\mathbb{G}_n \triangleq \sqrt{n}(\mathbb{P}_n - P)$, $\mathbb{G}_n^{(b)} \triangleq \sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)$, and P_M denote probability taken with respect to the bootstrap weights defining the bootstrap empirical measure, respectively.

B.2.1. Results for second stage parameters

In this section we will characterize the limiting distributions of the second stage parameters under fixed and local alternatives. We will also derive the limiting distribution of the bootstrap analog of the second stage parameters. For convenience, let $p_{t0} \triangleq \dim(\beta_{t,0}^*)$, $p_{t1} \triangleq \dim(\beta_{t,1}^*)$, and $p_t \triangleq \dim(\beta_t^*) = p_{t0} + p_{t1}$ for $t = 1, 2$.

Theorem B.3. *Assume (A1) and (A2) and fix $a \in \mathbb{R}^{p_2}$, then*

1. $a^\mathsf{T} \sqrt{n}(\hat{\beta}_2 - \beta_2^*) \rightsquigarrow_P a^\mathsf{T} \mathbb{Z}_\infty$,
2. $a^\mathsf{T} \sqrt{n}(\hat{\beta}_2^{(b)} - \hat{\beta}_2) \rightsquigarrow_{P_M} a^\mathsf{T} \mathbb{Z}_\infty$ *in P-probability; and*
3. *if in addition (A3) holds,* $a^\mathsf{T} \sqrt{n}(\hat{\beta}_2 - \beta_{2,n}^*) \rightsquigarrow_{P_n} a^\mathsf{T} \mathbb{Z}_\infty$,

where $\mathbb{Z}_\infty$ is a mean zero normal random vector with covariance matrix $\Sigma_{2,\infty}^{-1} P[B_2 B_2^\mathsf{T}(Y_2 - B_2^\mathsf{T} \beta_2^*)^2] \Sigma_{2,\infty}^{-1}$.

Proof. Define the class of functions $\mathcal{F}_2$ as

$$\mathcal{F}_2 \triangleq \{f(b_2, y_2; a, \beta_2) \triangleq a^\mathsf{T} b_2(y_2 - b_2^\mathsf{T} \beta_2) \,:\, a, \beta_2 \in \mathbb{R}^{p_2}, ||a|| \leq K, ||\beta_2|| \leq K\}, \tag{10}$$

and the function $w_2 : D_{p_2} \times l^\infty(\mathcal{F}_2) \times \mathbb{R}^{p_2} \times \mathbb{R}^{p_2} \to \mathbb{R}$ as

$$w_2(\Sigma, \mu, \beta_2, a) \triangleq \mu\left(a^\mathsf{T} \Sigma^{-1} B_2(Y_2 - B_2^\mathsf{T} \beta_2)\right). \tag{11}$$

Since the estimated covariance matrices $\hat{\Sigma}_2 = \mathbb{P}_n B_2 B_2^\mathsf{T}$ and $\hat{\Sigma}_2^{(b)} = \hat{\mathbb{P}}_n^{(b)} B_2 B_2^\mathsf{T}$ are weakly consistent (by Lemma B.5), we will avoid additional notation by assuming they are nonsingular for all n without loss of generality. Thus

$$a^\mathsf{T} \sqrt{n}(\hat{\beta}_2 - \beta_2^*) = w_2(\hat{\Sigma}_2, \mathbb{G}_n, \beta_2^*, a),$$
$$a^\mathsf{T} \sqrt{n}(\hat{\beta}_2^{(b)} - \hat{\beta}_2) = w_2(\hat{\Sigma}_2^{(b)}, \mathbb{G}_n^{(b)}, \hat{\beta}_2, a),$$
$$\text{and } a^\mathsf{T} \sqrt{n}(\hat{\beta}_2 - \beta_{2,n}^*) = w_2(\hat{\Sigma}_2, \sqrt{n}(\mathbb{P}_n - P_n), \beta_{2,n}^*, a).$$

In addition, note that $a^\mathsf{T} \mathbb{Z}_\infty = w_2(\Sigma_{2,\infty}, \mathbb{G}_\infty, \beta_2^*, a)$ in distribution, where $\mathbb{G}_\infty$ is a tight Gaussian process in $l^\infty(\mathcal{F}_2)$ with covariance function $\text{Cov}(\mathbb{G}_\infty f_1, \mathbb{G}_\infty f_2) = P(f_1 - Pf_1)(f_2 - Pf_2)$. Results 1 and 3 follow from Lemmas B.4–B.7 and the continuous mapping theorem [Theorem 1.3.6 of Van der Vaart and Wellner 1996]. Result 2 follows from the bootstrap continuous mapping theorem [Theorem 10.8 of Kosorok 2008] together with Lemmas B.4–B.8. □

Lemma B.4. *Under (A1), the function w_2 defined in (11) is continuous at points in $D_{p_2} \times C_b(\mathcal{F}) \times \mathbb{R}^{p_2} \times \mathbb{R}^{p_2}$.*

Proof. Let $\epsilon > 0$ be arbitrary and let $(\Sigma, \mu, \beta_2, a)$ be an element of $D_{p_2} \times C_b(\mathcal{F}) \times \mathbb{R}^{p_2} \times \mathbb{R}^{p_2}$. In addition, let $(\Sigma', \mu', \beta_2', a')$ be an element of $D_{p_2} \times l^\infty(\mathcal{F}) \times \mathbb{R}^{p_2} \times \mathbb{R}^{p_2}$. From the form of $\mathcal{F}$ and the moment assumptions in (A1) we see that if $\Sigma - \Sigma'$, $a - a'$, and $\beta_2 - \beta_2'$ are small then so must $\rho_P(f - f')$ be small, where

$$f(B_2, Y_2) = a^\mathsf{T} \Sigma^{-1} B_2(Y_2 - B_2^\mathsf{T} \beta_2),$$
$$f'(B_2, Y_2) = a'^\mathsf{T} \Sigma'^{-1} B_2(Y_2 - B_2^\mathsf{T} \beta_2').$$

In particular, we can choose $\delta > 0$ sufficiently small so that $||\Sigma - \Sigma'|| + ||a - a'|| + ||\beta_2 - \beta_2'|| < \delta$ implies that $\rho_P(f - f')$ is small enough to guarantee, by appeal to the continuity of μ, that $|\mu(f) - \mu(f')| \leq \epsilon/2$. Finally, note that

$$\left|w_2(\Sigma, \mu, \beta_2, a) - w_2(\Sigma', \mu', \beta_2', a')\right| \leq |\mu(f) - \mu(f')| + ||\mu - \mu'||_{\mathcal{F}_2}.$$

Let $\delta' = \min(\delta, \epsilon/2)$, then $||\Sigma - \Sigma'|| + ||\mu - \mu'||_{\mathcal{F}_2} + ||\beta_2 - \beta_2'|| + ||a - a'|| < \delta'$ implies that $|w_2(\Sigma, \mu, \beta_2, a) - w_2(\Sigma', \mu', \beta_2', a')| \leq \epsilon$. Thus, the desired result is proved. □

Having established the continuity of w_2 the next step will be to characterize the limiting behavior of $\beta_{2,n}^*$, $\hat{\beta}_2$, $\hat{\Sigma}_2$, $\hat{\Sigma}_2^{(b)}$, and the limiting distributions of $\mathbb{G}_n$, $\sqrt{n}(\mathbb{P}_n - P_n)$, and $\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)$. These limits are established in a series of lemmas. Once this has been accomplished we will be able to apply the continuous mapping theorem to obtain the limiting distributions of $\sqrt{n}(\hat{\beta}_2 - \beta_2^*)$, $\sqrt{n}(\hat{\beta} - \beta_{2,n}^*)$, and $\sqrt{n}(\hat{\beta}_2^{(b)} - \hat{\beta}_2)$.

Lemma B.5. *Assume (A1)–(A2), then $\hat{\Sigma}_2 \to_P \Sigma_{2,\infty}$ and $\hat{\Sigma}_2^{(b)} \to_{P_M} \Sigma_{2,\infty}$ in P-probability as $n \to \infty$. Furthermore, if (A3) holds, then $\hat{\Sigma}_2 \to_{P_n} \Sigma_{2,\infty}$ as $n \to \infty$.*

Proof. The first two claims follow from weak law of large numbers [Bickel and Freedman 1981; Csörgő and Rosalsky 2003]. For the third claim, note that $\hat{\Sigma}_2 - \Sigma_{2,\infty} = (\hat{\Sigma}_2 - \Sigma_{2,n}) + (\Sigma_{2,n} - \Sigma_{2,\infty})$ and $\hat{\Sigma}_2 - \Sigma_{2,n} \to_{P_n} 0$ by law of large numbers. Below we show that $\Sigma_{2,n} \to \Sigma_{2,\infty}$. This will complete the proof.

let $c \in \mathbb{R}^{p_2}$ be arbitrary and define $\nu \triangleq c^\mathsf{T} B_2 B_2^\mathsf{T} c$. We will show that $\int \nu(dP_n - dP) = o(1)$. First, note that

$$\int \nu(dP_n - dP) = \int \nu(dP_n^{1/2} + dP^{1/2})(dP_n^{1/2} - dP^{1/2}).$$

Furthermore, the absolute value of the foregoing expression is bounded above by

$$\int |\nu||(dP_n^{1/2} + dP^{1/2})|(dP_n^{1/2} - dP^{1/2})$$
$$\leq \sqrt{\int \nu^2(dP_n^{1/2} + dP^{1/2})^2}\sqrt{\int (dP_n^{1/2} - dP^{1/2})^2},$$

where the last inequality is simply Hölder's inequality. Next, note that owing to the inequality $(\sqrt{a} + \sqrt{b})^2 \leq 2a + 2b$ it follows that

$$\int \nu^2(dP_n^{1/2} + dP^{1/2})^2 \leq 2\int \nu^2 dP_n + 2\int \nu^2 dP = O(1),$$

by appeal to (A3). Now write

$$\int (dP_n^{1/2} - dP^{1/2})^2 = n^{-1}\left\{\int \left(\sqrt{n}(dP_n^{1/2} - dP^{1/2}) - \frac{1}{2}v dP^{1/2}\right)^2 \right.$$
$$\left. - \frac{1}{4}\int v^2 dP + \sqrt{n}\int v dP^{1/2}(dP_n^{1/2} - dP^{1/2})\right\}.$$

The right hand side of the preceding display is equal to

$$O(1/n) + n^{-1/2} \int v dP^{1/2}(dP_n^{1/2} - dP^{1/2})$$

$$\leq O(1/n) + n^{-1/2}\sqrt{\int v^2 dP}\sqrt{\int (dP_n^{1/2} - dP^{1/2})^2},$$

which is $o(1)$. Thus $\Sigma_{2,n} \to \Sigma_{2,\infty}$. $\qquad\square$

Lemma B.6. *Under (A1) and (A2), $\hat{\beta}_2 \to_P \beta_2^*$ as $n \to \infty$. If, in addition (A3) holds, then $\lim_{n\to\infty} \sqrt{n}(\beta_{2,n}^* - \beta_2^*) = \Sigma_2^{-1} P v B_2(Y_2 - B_2^\intercal \beta_2^*)$.*

Proof. $\hat{\beta}_2 \to_P \beta_2^*$ follows from weak law of large numbers and Slutsky's lemma.

Recall that $0 = P_n B_2(Y_2 - B_2^\intercal \beta_{2,n}^*)$ which we can write as

$$\sqrt{n}(P_n - P)B_2(Y_2 - B_2^\intercal \beta_2^*) - \Sigma_{2,n}\sqrt{n}(\beta_2^* - \beta_{2,n}^*),$$

so that for sufficiently large n it follows that $\sqrt{n}(\beta_{2,n}^* - \beta_2^*) = \Sigma_{2,n}^{-1}\sqrt{n}(P_n - P)B_2(Y_2 - B_2^\intercal \beta_2^*)$. By appeal to (A3) it follows that for any vector $a \in \mathbb{R}^{p_2}$ we have $\sup_n P_n(a^\intercal B_2(Y_2 - B_2^\intercal \beta_2^*))^2 < \infty$. Theorem 3.10.12 of Van der Vaart and Wellner (1996) ensures that

$$\sqrt{n}(P_n - P)B_2(Y_2 - B_2^\intercal \beta_2^*) \to P v B_2(Y_2 - B_2^\intercal \beta_2^*)$$

as $n \to \infty$. This completes the proof. $\qquad\square$

Lemma B.7. *Assume (A1)–(A2), then*

1) $\mathbb{G}_n \rightsquigarrow_P \mathbb{G}_\infty$ in $l^\infty(\mathcal{F}_2)$, where $\mathcal{F}_2$ is defined in (10), and $\mathbb{G}_\infty$ is a tight Gaussian process in $l^\infty(\mathcal{F}_2)$ with covariance function $\mathrm{Cov}(\mathbb{G}_\infty f_1, \mathbb{G}_\infty f_2) = P(f_1 - P f_1)(f_2 - P f_2)$; and

2) $\sup_{\omega \in BL_1} \left| \mathbb{E}_M \omega(\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)) - \mathbb{E}\omega(\mathbb{G}_\infty) \right| \to_{P^*} 0$ in $l^\infty(\mathcal{F}_2)$.
If, in addition (A3) holds, then
3) $\sqrt{n}(\mathbb{P}_n - P_n) \rightsquigarrow_{P_n} \mathbb{G}_\infty$ in $l^\infty(\mathcal{F}_2)$.

Proof. First note that $\mathcal{F}_2$ is a subset of the pairwise product of the linear classes $\{a^\intercal b_2 : a \in \mathbb{R}^{p_2}\}$ and $\{y_2 - b_2^\intercal \beta_2 : \beta \in \mathbb{R}^{p_2}\}$ each of which is VC-subgraph of index no more than $p_2 + 1$ and P-measurable. Under (A1), the envelope of $\mathcal{F}_2$, $F_2(B_2, Y_2) = K||B_2||(|Y_2| + K||B_2||)$, is square integrable. This implies that $\mathcal{F}_2$ is P-Donsker, and 1) follows immediately. 2) follows from Theorem 3.6.1 of Van der Vaart and Wellner (1996). For 3), note that from (A3) it follows that $\sup_f |P_n f|$ is a bounded sequence. The result follows from theorem 3.10.12 of Van der Vaart and Wellner (1996). $\qquad\square$

Lemma B.8. *The space $C_b(\mathcal{F}_2)$ is a closed subset of $l^\infty(\mathcal{F}_2)$ and $\mathrm{P}(\mathbb{G}_\infty \in C_b(\mathcal{F}_2)) = 1$.*

Proof. Let $\{\mu_n\}_{n=1}^{\infty}$ be a convergent sequence of elements in $C_b(\mathcal{F}_2)$ and μ_0 the limiting element. For the first claim, we only need to show that $||\mu_0||_{\mathcal{F}_2} = \sup_{f \in \mathcal{F}_2} |\mu_0(f)|$ is bounded, and for any $f \in \mathcal{F}$ and $\epsilon > 0$, there exists some positive δ depending on f so that $|\mu_0(f') - \mu_0(f)| < \epsilon$ for all $f' \in \mathcal{F}_2$ and $\rho_P(f', f) < \delta$. The boundedness argument follows by noticing that $||\mu_0||_{\mathcal{F}_2} \leq ||\mu_n||_{\mathcal{F}_2} + ||\mu_n - \mu_0||_{\mathcal{F}_2}$ for any n; in particular, for some fixed large enough n, $||\mu_n||_{\mathcal{F}_2}$ is bounded by the fact $\mu_n \in C_b(\mathcal{F}_2)$, and $||\mu_n - \mu_0||_{\mathcal{F}_2}$ is bounded above by a constant due to the convergence of μ_n to μ_0. For continuity, note that since μ_n converges to μ_0, we can choose n^* so that $||\mu_n - \mu_0|| < \epsilon/4$ for all $n \geq n^*$. In addition, by the continuity of μ_{n^*}, there exists some $\delta > 0$ so that $|\mu_{n^*}(f') - \mu_{n^*}(f)| < \epsilon$ for all $\rho_P(f', f) < \delta$. Thus

$$
\begin{aligned}
|\mu_0(f') - \mu_0(f)| &\leq |\mu_0(f) - \mu_{n^*}(f)| + |\mu_{n^*}(f') - \mu_0(f')| + |\mu_{n^*}(f) - \mu_{n^*}(f')| \\
&\leq 2||\mu_0 - \mu_{n^*}||_{\mathcal{F}_2} + |\mu_{n^*}(f) - \mu_{n^*}(f')| \\
&\leq 3\epsilon/4.
\end{aligned}
$$

This implies that $C_b(\mathcal{F})$ is closed.

Next note that $\mathbb{G}_\infty$ is a tight Gaussian process in $l^\infty(\mathcal{F}_2)$. By the argument in section 1.5 of van de Van der Vaart and Wellner (1996), almost all sample paths $f \to \mathbb{G}_\infty(f, \omega)$ are uniformly ρ_2-continuous, where $\rho_2(f_1, f_2) = [P(\mathbb{G}_\infty f_1 - \mathbb{G}_\infty f_2)^2]^{1/2}$ is a semimetric on $\mathcal{F}$. Since $\rho_2(f_1, f_2) = [Var(f_1 - f_2)]^{1/2} \leq \rho_P(f_1, f_2)$, the continuity of the sample paths of $\mathbb{G}_\infty$ follows immediately. $\square$

B.2.2. *A characterization of the first stage coefficients and the upper bound $\mathcal{U}(c)$*

In this section we present the proofs for Theorems 4.1 and 4.2. We first derive an expansion for the first stage coefficients and two useful expansions of the upper bound $\mathcal{U}(c)$. The terms in the forementioned expansions will be treated individually in subsequent sections. We will make use of the following functions.

1. $w_{11} : D_{p_1} \times D_{p_1 \times p_{20}} \times l^\infty(\mathcal{F}_{11}) \times l^\infty(\mathcal{F}_{11}) \times \mathbb{R}^{p_2} \times \mathbb{R}^{p_1 + p_2} \to \mathbb{R}$ is defined as

$$
\begin{aligned}
&w_{11}(\Sigma_1, \Sigma_{12}, \mu, \omega, \nu, \beta) \\
&\quad \triangleq c^\intercal \Sigma_1^{-1} \Sigma_{12} \nu_0 + \omega \left(c^\intercal \Sigma_1^{-1} B_1 H_{2,1}^\intercal \nu_1 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} \right) \\
&\quad + \mu \left[c^\intercal \Sigma_1^{-1} B_1 \left(Y_1 + H_{2,0}^\intercal \beta_{2,0} + \left[H_{2,1}^\intercal \beta_{2,1} \right]_+ - B_1^\intercal \beta_1 \right) \right], \quad (12)
\end{aligned}
$$

where $D_{p_1 \times p_{20}}$ is the space of $p_1 \times p_{20}$ matrices equipped with the spectral norm, and $\mathcal{F}_{11} = \Big\{ f(b_1, y_1, h_{2,0}, h_{2,1}) = a_1^\intercal b_1 \big(y_1 + h_{2,0}^\intercal \beta_{2,0} + [h_{2,1}^\intercal \beta_{2,1}]_+ - b_1^\intercal \beta_1 \big) + a_2^\intercal b_1 (h_{2,1}^\intercal \nu_1) 1_{h_{2,1}^\intercal \beta_{2,1}^* > 0},: \beta = (\beta_1^\intercal, \beta_{2,0}^\intercal, \beta_{2,1}^\intercal)^\intercal \in \mathbb{R}^{p_1 + p_2}, \nu = (\nu_0^\intercal, \nu_1^\intercal)^\intercal \in \mathbb{R}^{p_2}, a_1, a_2 \in \mathbb{R}^{p_1}, \max\{||a_1||, ||a_2||, ||\beta||, ||\nu||\} \leq K \Big\}$.

2. $w_{12} : D_{p_1} \times l^\infty(\mathcal{F}_{12}) \times \mathbb{R}^{p_{21}} \times \mathbb{R}^{p_{21}} \to \mathbb{R}$ is defined as

$$
\begin{aligned}
&w_{12}(\Sigma_1, \mu, \nu, \gamma) \\
&\triangleq \mu \left[c^\mathsf{T} \Sigma_1^{-1} B_1 \left(\left[H_{2,1}^\mathsf{T} \nu + H_{2,1}^\mathsf{T} \gamma \right]_+ - \left[H_{2,1}^\mathsf{T} \gamma \right]_+ \right) 1_{H_{2,1}^\mathsf{T} \beta_{2,1}^* = 0} \right], \quad (13)
\end{aligned}
$$

where $\mathcal{F}_{12} = \left\{ f(b_1, h_{2,1}) = a^\mathsf{T} b_1 ([h_{2,1}^T \nu + h_{2,1}^\mathsf{T} \gamma]_+ - [h_{2,1}^\mathsf{T} \gamma]_+) 1_{h_{2,1}^\mathsf{T} \beta_{2,1}^* = 0} : a \in \mathbb{R}^{p_1}, \gamma, \nu \in \mathbb{R}^{p_{21}}, \max\{||a||, ||\nu||\} \leq K \right\}$.

3. $\rho_{11} : D_{p_1} \times D_{p_{21}}^k \times l^\infty(\widetilde{\mathcal{F}}_{11}) \times \mathbb{R}^{p_{21}} \times \mathbb{R}^{p_{21}} \times \mathbb{R}^{p_{21}} \times \mathbb{R} \to \mathbb{R}$, is defined as

$$
\begin{aligned}
&\rho_{11}(\Sigma_1, \Sigma_{21,21}, \mu, \nu, \eta, \gamma, \lambda) \\
&\triangleq \mu \Bigg[c^\mathsf{T} \Sigma_1^{-1} B_1 \left(\left[H_{2,1}^\mathsf{T} \nu + H_{2,1}^\mathsf{T} \gamma \right]_+ - [H_{2,1}^\mathsf{T} \gamma]_+ \right) \\
&\qquad\qquad \times \left(1_{\frac{(H_{2,1}^\mathsf{T} \nu + H_{2,1}^\mathsf{T} \eta)^2}{H_{2,1}^\mathsf{T} \Sigma_{21,21} H_{2,1}} \leq \lambda} - 1_{H_{2,1}^\mathsf{T} \beta_{2,1}^* = 0} \right) \Bigg], \quad (14)
\end{aligned}
$$

where $\widetilde{\mathcal{F}}_{11} = \Big\{ f(b_1, h_{2,1}) = a^\mathsf{T} b_1 ([h_{2,1}^\mathsf{T} \nu - h_{2,1}^\mathsf{T} \gamma]_+ - [h_{2,1}^\mathsf{T} \gamma]_+) \times (1_{\frac{(h_{2,1}^\mathsf{T} \nu + h_{2,1}^\mathsf{T} \eta)^2}{h_{2,1}^\mathsf{T} \Sigma_{21,21} h_{2,1}} \leq \lambda} - 1_{h_{2,1}^\mathsf{T} \beta_{2,1}^* = 0}), : a \in \mathbb{R}^{p_1}, \nu, \eta, \gamma \in \mathbb{R}^{p_{21}}, \max\{||a||, ||\nu||\} \leq K, \lambda \in \mathbb{R}, \Sigma_{21,21} \in D_{p_{21}}^k \Big\}$.

4. $\rho_{12} : D_{p_1} \times l^\infty(\widetilde{\mathcal{F}}_{12}) \times \mathbb{R}^{p_{21}} \times \mathbb{R}^{p_{21}} \to \mathbb{R}$, defined as

$$
\begin{aligned}
&\rho_{12}(\Sigma_1, \mu, \nu, \eta) \\
&\triangleq \mu \Bigg[c^\mathsf{T} \Sigma_1^{-1} B_1 \left(\left[H_{2,1}^\mathsf{T} \nu + H_{2,1}^\mathsf{T} \eta \right]_+ - [H_{2,1}^\mathsf{T} \eta]_+ - H_{2,1}^\mathsf{T} \nu \right) 1_{H_{2,1}^\mathsf{T} \beta_{2,1}^* > 0} \\
&\qquad + c^\mathsf{T} \Sigma_1^{-1} B_1 \left(\left[H_{2,1}^\mathsf{T} \nu + H_{2,1}^\mathsf{T} \eta \right]_+ - [H_{2,1}^\mathsf{T} \eta]_+ \right) 1_{H_{2,1}^\mathsf{T} \beta_{2,1}^* < 0} \Bigg], \quad (15)
\end{aligned}
$$

where $\widetilde{\mathcal{F}}_{12} = \{ a^\mathsf{T} b_1 ([h_{2,1}^\mathsf{T} \nu + h_{2,1}^\mathsf{T} \eta]_+ - [h_{2,1}^\mathsf{T} \eta]_+ - h_{2,1}^\mathsf{T} \nu) 1_{h_{2,1}^\mathsf{T} \beta_{2,1}^* > 0} - a^\mathsf{T} b_1 \times ([h_{2,1}^\mathsf{T} \nu + h_{2,1}^\mathsf{T} \eta]_+ - [h_{2,1}^\mathsf{T} \eta]_+) 1_{h_{2,1}^\mathsf{T} \beta_{2,1}^* < 0} : a \in \mathbb{R}^{p_1}, \nu \in \mathbb{R}^{p_{21}}, \max\{||a||, ||\nu||\} \leq K, \eta \in \mathbb{R}^{p_{21}} \}$.

Using the foregoing functions, we have the following expressions for the first stage parameters:

$$
\begin{aligned}
c^\mathsf{T} \sqrt{n}(\hat{\beta}_1 - \beta_1^*) &= w_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{12}, \mathbb{G}_n, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_2 - \beta_2^*), (\beta_1^{*\mathsf{T}}, \beta_2^{*\mathsf{T}})^\mathsf{T}) \\
&\quad + w_{12}(\hat{\Sigma}_1, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*), \sqrt{n}\beta_{2,1}^*) \\
&\quad + \rho_{12}(\hat{\Sigma}_1, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*), \sqrt{n}\beta_{2,1}^*); \quad (16) \\
\sqrt{n}(\hat{\beta}_1 - \beta_{1,n}^*) &= w_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{12}, \sqrt{n}(\mathbb{P}_n - P_n), \mathbb{P}_n, \sqrt{n}(\hat{\beta}_2 - \beta_{2,n}^*), (\beta_{1,n}^{*\mathsf{T}}, \beta_{2,n}^{*\mathsf{T}})^\mathsf{T}) \\
&\quad + w_{12}(\hat{\Sigma}_1, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1,n}^*), \sqrt{n}\beta_{2,1,n}^*) \\
&\quad + \rho_{12}(\hat{\Sigma}_1, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1,n}^*), \sqrt{n}\beta_{2,1,n}^*), \quad (17)
\end{aligned}
$$

where $\hat{\Sigma}_{12} = \mathbb{P}_n B_1 H_{2,0}^\mathsf{T}$. Similarly, we can express the upper bound $\mathcal{U}(c)$ in terms of the above functions:

$$
\begin{aligned}
\mathcal{U}(c) \;=\; & w_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{12}, \mathbb{G}_n, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_2 - \beta_2^*), (\beta_1^{*\mathsf{T}}, \beta_2^{*\mathsf{T}})^\mathsf{T}) \\
& + \rho_{12}(\hat{\Sigma}_1, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*), \sqrt{n}\beta_{2,1}^*) \\
& - \rho_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{21,21}, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*), \sqrt{n}\beta_{2,1}^*, \sqrt{n}\beta_{2,1}^*, \lambda_n) \\
& + \sup_{\gamma \in \mathbb{R}^{p_{2,1}}} \Big\{ w_{12}(\hat{\Sigma}_1, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*), \gamma) \\
& \qquad\qquad + \rho_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{21,21}, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*), \sqrt{n}\beta_{2,1}^*, \gamma, \lambda_n) \Big\}. \quad (18)
\end{aligned}
$$

We will also make use of the following alternative expression for the upper bound $\mathcal{U}(c)$ under P_n:

$$
\begin{aligned}
\mathcal{U}(c) \;=\; & w_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{12}, \sqrt{n}(\mathbb{P}_n - P_n), \mathbb{P}_n, \sqrt{n}(\hat{\beta}_2 - \beta_{2,n}^*), (\beta_{1,n}^{*\mathsf{T}}, \beta_{2,n}^{*\mathsf{T}})^\mathsf{T}) \\
& + \rho_{12}(\hat{\Sigma}_1, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1,n}^*), \sqrt{n}\beta_{2,1,n}^*) \\
& - \rho_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{21,21}, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1,n}^*), \sqrt{n}\beta_{2,1,n}^*, \sqrt{n}\beta_{2,1,n}^*, \lambda_n) \\
& + \sup_{\gamma \in \mathbb{R}^{p_{21}}} \Big\{ w_{12}(\hat{\Sigma}_1, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1,n}^*), \gamma) \\
& \qquad\qquad + \rho_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{21,21}, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1,n}^*), \sqrt{n}\beta_{2,1,n}^*, \gamma, \lambda_n) \Big\}. \quad (19)
\end{aligned}
$$

Similarly, we will make use of following expression for the bootstrap analog of the upper bound:

$$
\begin{aligned}
\hat{\mathcal{U}}^{(b)}(c) \;=\; & w_{11}(\hat{\Sigma}_1^{(b)}, \hat{\Sigma}_{12}^{(b)}, \sqrt{n}(\mathbb{P}_n^{(b)} - \mathbb{P}_n), \mathbb{P}_n^{(b)}, \sqrt{n}(\hat{\beta}_2^{(b)} - \hat{\beta}_2), (\hat{\beta}_1^\mathsf{T}, \hat{\beta}_2^\mathsf{T})^\mathsf{T}) \\
& + \rho_{12}(\hat{\Sigma}_1^{(b)}, \hat{\mathbb{P}}_n^{(b)}, \sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1}), \sqrt{n}\hat{\beta}_{2,1}) \\
& - \rho_{11}(\hat{\Sigma}_1^{(b)}, \hat{\Sigma}_{21,21}^{(b)}, \hat{\mathbb{P}}_n^{(b)}, \sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1}), \sqrt{n}\hat{\beta}_{2,1}, \sqrt{n}\hat{\beta}_{2,1}, \lambda_n) \\
& + \sup_{\gamma \in \mathbb{R}^{p_{21}}} \Big\{ w_{12}(\hat{\Sigma}_1^{(b)}, \hat{\mathbb{P}}_n^{(b)}, \sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1}), \gamma) \\
& \qquad\qquad + \rho_{11}(\hat{\Sigma}_1^{(b)}, \hat{\Sigma}_{21,21}^{(b)}, \hat{\mathbb{P}}_n^{(b)}, \sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1}), \sqrt{n}\hat{\beta}_{2,1}, \gamma, \lambda_n) \Big\}. \quad (20)
\end{aligned}
$$

The lower bound $\mathcal{L}(c)$ and its bootstrap analog $\hat{\mathcal{L}}^{(b)}(c)$ can be expressed in a similar fashion by replacing the sup with an inf in the expression of $\mathcal{U}(c)$ and $\hat{\mathcal{U}}^{(b)}(c)$, respectively.

By Lemmas B.9 and B.11 below, ρ_{11} is negligible, and w_{11} and w_{12} are continuous at desired points. The negligibility of ρ_{12} can be obtained in a similar fashion. Note that the convergence of $\hat{\Sigma}_1$ and $\hat{\Sigma}_1^{(b)}$ to Σ_1 and the convergence of $\hat{\Sigma}_{12}$ and $\hat{\Sigma}_{12}^{(b)}$ to $PB_1 H_{2,0}^\mathsf{T}$ can be obtained using similar proof techniques as in Lemma B.5. This together with Theorem B.3, Lemmas B.5–B.8, and the

continuous mapping theorems as presented in the previous section, implies that the conclusions of Theorems 4.1 and 4.2 hold with

$$\mathbb{S}_\infty = \Sigma_{1,\infty}^{-1}\big[\mathbb{G}_\infty\big(B_1(Y_1 + H_{2,0}^\mathsf{T}\beta_{2,0}^* + [H_{2,1}^\mathsf{T}\beta_{2,1}^*]_+ - B_1^\mathsf{T}\beta_1^*)\big) + PB_1H_{2,0}^\mathsf{T}\mathbb{Z}_{\infty,0}\big]$$

and $\mathbb{V}_\infty = \mathbb{Z}_{\infty,1}$,

where $\mathbb{Z}_{\infty,0}^\mathsf{T} \in \mathbb{R}^{p_{20}}$, $\mathbb{Z}_{\infty,1}^\mathsf{T} \in \mathbb{R}^{p_{21}}$, and $\mathbb{Z}_\infty = (\mathbb{Z}_{\infty,0}^\mathsf{T}, \mathbb{Z}_{\infty,1}^\mathsf{T})^\mathsf{T} = \Sigma_{2,\infty}^{-1}\mathbb{G}_\infty[B_2(Y_2 - B_2^\mathsf{T}\beta_2^*)]$.

Lemma B.9. *Assume (A1), (A2) and (A4). Then*

1. $\sup_{\gamma \in \mathbb{R}^{p_{21}}} \left|\rho_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{21,21}, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*), \sqrt{n}\beta_{2,1}^*, \gamma, \lambda_n)\right| \to_P 0$, *and*
2. $\sup_{\gamma \in \mathbb{R}^{p_{21}}} \left|\rho_{11}(\hat{\Sigma}_1^{(b)}, \hat{\Sigma}_{21,21}^{(b)}, \hat{\mathbb{P}}_n^{(b)}, \sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1}), \sqrt{n}\hat{\beta}_{2,1}, \gamma, \lambda_n)\right| \to_{P_M} 0$ *almost surely P.*

If, in addition, we assume (A3), then

3. $\sup_{\gamma \in \mathbb{R}^{p_{21}}} \left|\rho_{11}(\hat{\Sigma}_1, \hat{\Sigma}_{21,21}, \mathbb{P}_n, \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1,n}^*), \sqrt{n}\beta_{2,1,n}^*, \gamma, \lambda_n)\right| \to_{P_n} 0$.

Proof. First it is easy to verify that $|[H_{2,1}^\mathsf{T}\nu - H_{2,1}^\mathsf{T}\gamma]_+ - [H_{2,1}^\mathsf{T}\gamma]_+| \le |h_{2,1}^\mathsf{T}\nu|$. Thus for any probability measure μ in $l^\infty(\widetilde{\mathcal{F}}_{11})$,

$$|\rho_{11}(\Sigma_1, \Sigma_{21,21}, \mu, \nu, \eta, \gamma, \lambda)| \le K\left\{\mu\left(||B_1|| \, ||H_{2,1}|| \, 1_{H_{2,1}^\mathsf{T}\beta_{2,1}^*=0, \frac{H_{2,1}^\mathsf{T}\eta}{||H_{2,1}||} > \sqrt{\lambda k} - K}\right)\right.$$

$$+ \mu\left(||B_1|| \, ||H_{2,1}|| \, 1_{H_{2,1}^\mathsf{T}\beta_{2,1}^*=0, \frac{H_{2,1}^\mathsf{T}\eta}{||H_{2,1}||} < -\sqrt{\lambda k} - K}\right)$$

$$\left. + \mu\left(||B_1|| \, ||H_{2,1}|| \, 1_{H_{2,1}^\mathsf{T}\beta_{2,1}^*\neq 0, -\sqrt{\lambda/k}-K \le \frac{H_{2,1}^\mathsf{T}\eta}{||H_{2,1}||} \le \sqrt{\lambda/k}+K}\right)\right\}$$

for a sufficiently large constant $K > 0$ and a sufficiently small constant $k \in (0, 1)$. Since k is held constant there is no loss in generality taking $k = 1$. Define $\rho'_{11} : l^\infty(\mathcal{F}'_{11}) \times \mathbb{R}^{p_{21}} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as

$$\rho'_{11}(\mu, \eta, \delta, \delta') = \mu\left(||B_1|| \, ||H_{2,1}|| \, 1_{H_{2,1}^\mathsf{T}\beta_{2,1}^*=0, \frac{H_{2,1}^\mathsf{T}\eta}{||H_{2,1}||} > \delta}\right)$$

$$+ \mu\left(||B_1|| \, ||H_{2,1}|| \, 1_{H_{2,1}^\mathsf{T}\beta_{2,1}^*=0, \frac{H_{2,1}^\mathsf{T}\eta}{||H_{2,1}||} < \delta'}\right)$$

$$+ \mu\left(||B_1|| \, ||H_{2,1}|| \, 1_{H_{2,1}^\mathsf{T}\beta_{2,1}^*\neq 0, \delta' \le \frac{H_{2,1}^\mathsf{T}\eta}{||H_{2,1}||} \le -\delta'}\right), \quad (21)$$

where

$$\mathcal{F}'_{11} = \left\{f(b_1, h_{2,1}) = ||b_1|| \, ||h_{2,1}||1_{h_{2,1}^\mathsf{T}\beta_{2,1}^*=0, \frac{h_{2,1}^\mathsf{T}\eta}{||h_{2,1}||} > \delta}\right.$$

$$+ \, ||b_1|| \, |||h_{2,1}||1_{h_{2,1}^\intercal \beta_{2,1}^* = 0, \frac{H_{2,1}^\intercal \eta}{||h_{2,1}||} < \delta'} + \, ||b_1|| \, ||h_{2,1}||1_{h_{2,1}^\intercal \beta_{2,1}^* \neq 0, \delta' \le \frac{h_{2,1}^\intercal \eta}{||h_{2,1}||} \le -\delta'},$$

$$\eta \in \mathbb{R}^{p_{21}}, \max\{||\eta||, ||\delta||, ||\delta'||\} \le K \Big\}.$$

Then

$$|\rho_{11}(\Sigma_1, \Sigma_{21,21}, \mu, \nu, \eta, \gamma, \lambda)| \le K \rho_{11}' \left(\mu, \eta/\sqrt{n}, (\sqrt{\lambda} - K)/\sqrt{n}, -(\sqrt{\lambda} + K)/\sqrt{n} \right)$$

for $\mu \in l^\infty(\widetilde{\mathcal{F}}_{11})$. In particular for n sufficiently large,

$$\big| \rho_{11}(\hat{\Sigma}_1^{(b)}, \hat{\Sigma}_{21,21}^{(b)}, \hat{\mathbb{P}}_n^{(b)}, \sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1}), \sqrt{n}\hat{\beta}_{2,1}, \gamma, \lambda_n) \big|$$
$$\le K \, \rho_{11}' \left(\hat{\mathbb{P}}_n^{(b)}, \hat{\beta}_{2,1}, (\sqrt{\lambda_n} - K)/\sqrt{n}, -(\sqrt{\lambda_n} - K)/\sqrt{n} \right)$$
$$+ \, ||c|| \, ||\hat{\Sigma}_1^{(b)}|| \, ||\sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1})|| \, \hat{\mathbb{P}}_n^{(b)} \left(||B_1|| \, ||H_{2,1}|| \right) 1_{||\sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1})|| > K},$$

where we have assumed, without loss of generality, that $\hat{\Sigma}_{21,21}^{(b)}$ is the identity matrix. By part 2 of Lemma B.10 below, we see that the first term on the right hand side of the above display is $o_{P_M}(1)$ almost surely P. To deal with the second term, for any $\epsilon, \delta > 0$, let K sufficiently large so that $P_M(||\sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1})|| > K) < \delta$ for sufficiently large n for almost all sequences P. Then

$$P_M \left(||c|| \, ||\hat{\Sigma}_1^{(b)}|| \, ||\sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1})|| ||\hat{\mathbb{P}}_n^{(b)}|| B_1|| \, ||H_{2,1}||1_{|\sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1})| > K} > \epsilon \right)$$
$$\le P_M \left(||\sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1})|| > K \right) \le \delta,$$

almost surely P. This completes the proof of result 2. Similar arguments can be used to prove results 1 and 3, and are omitted. $\qquad \square$

Lemma B.10. *Let ρ_{11}' be defined in (21). Assume (A1), (A2) and (A4), then*

1. $\rho_{11}'(\mathbb{P}_n, \beta_{2,1}^*, (\sqrt{\lambda_n} - K)/\sqrt{n}, (-\sqrt{\lambda_n} - K)/\sqrt{n}) \to_P 0$, *and*
2. $\rho_{11}'(\hat{\mathbb{P}}_n^{(b)}, \hat{\beta}_{2,1}, (\sqrt{\lambda_n} - K)/\sqrt{n}, (-\sqrt{\lambda_n} - K)/\sqrt{n}) \to_{P_M} 0$, *$P$-almost surely.*

If, in addition, we assume (A3), then

3. $\rho_{11}'(P_n, \beta_{2,1,n}^*, (\sqrt{\lambda_n} - K)/\sqrt{n}, (-\sqrt{\lambda_n} - K)/\sqrt{n}) \to_{P_n} 0$.

Proof. The class $\mathcal{F}_{11}'$ is P-Donsker and measurable by Theorem 8.14 in Anthony and Bartlett (1999) and Donkser preservation results (for example, see Theorem 2.10.6 in Van der Vaart and Wellner 1996). Note that by (A1) and (A3) $\sup_{f \in \mathcal{F}_{11}'} |Pf^2| < \infty$ and $\sup_{f \in \mathcal{F}_{11}'} |P_n f^2|$ is a bounded sequence. Thus, it follows that (i) $||\mathbb{P}_n - P|| \to 0$ almost surely under P in $l^\infty(\mathcal{F}_{11}')$, (ii) $||\hat{\mathbb{P}}_n^{(b)} - P|| \to 0$ almost surely P_M for almost all sequences P (Lemma 3.6.16 in Van der Vaart and Wellner 1996), and (iii) $||\mathbb{P}_n - P_n|| \to 0$ almost surely under P_n in $l^\infty(\mathcal{F}_{11})$ (Theorem 3.10.12 in Van der Vaart and Wellner 1996). Additionally, the argument in the proof of Lemma (B.5) shows that $\hat{\Sigma}_1$ is convergent to Σ_1 under P_n,

and the weak law of large numbers establishes convergence under P. The bootstrap strong law shows that $\hat{\Sigma}_1^{(b)}$ converges to Σ_1 in P_M probability for almost all sequences P.

Next we show that ρ'_{11} is continuous at the point $(P, \beta^*_{2,1}, 0, 0)$. Let $\mu_n \to P$ in $l^\infty(\mathcal{F}'_{11})$, $\eta_n \to \beta^*_{2,1}$, $\delta_n \to 0$, and $\delta'_n \to 0$. We have

$$
\left| \rho'_{11}(\mu_n, \eta_n, \delta_n, \delta'_n) - \rho'_{11}(P, \beta^*_{2,1}, 0, 0) \right|
$$
$$
\leq \left| \rho'_{11}(P, \eta_n, \delta_n, \delta'_n) - \rho'_{11}(P, \beta^*_{2,1}, 0, 0) \right| + \|\mu_n - P\|,
$$

which converges to zero by the dominated convergence theorem. The results follow from the continuous mapping theorems and the fact that $\rho'_{11}(P, \beta^*_{2,1}, 0, 0) = 0$. $\qquad\square$

Lemma B.11. *Assume (A1) and (A2). Then*

1. *w_{11} is continuous at points in $(\Sigma_{1,\infty}, \Sigma_{12,\infty}, C_b(\mathcal{F}_{11}), P, \mathbb{R}^{p_2}, (\beta^{*\mathsf{T}}_1, \beta^{*\mathsf{T}}_2)^{\mathsf{T}})$;*
2. *$w_{12}(\cdot, \cdot, \cdot, \sqrt{n}\beta^*_{2,1})$ and $w_{12}(\cdot, \cdot, \cdot, \sqrt{n}\beta^*_{2,1,n})$ are continuous at points in $(\Sigma_{1,\infty}, P, \mathbb{R}^{p_{21}})$; and*
3. *$w'_{12}(\Sigma_1, \mu, \nu) \triangleq \sup_{\gamma \in \mathbb{R}^{p_{21}}} w_{12}(\Sigma_1, \mu, \nu, \gamma)$ is continuous at points in $(\Sigma_{1,\infty}, P, \mathbb{R}^{p_{21}})$.*

Proof. To prove the desired continuity of w_{12} and w'_{12}, we will establish the stronger result that w_{12} is continuous at points $(\Sigma_{1,\infty}, P, \mathbb{R}^{p_{21}}, \gamma)$ uniformly in γ. That is, for any $\Sigma_n \to \Sigma_{1,\infty}$, probability measures $\mu_n \to P$ and $\nu_n \to \nu$, we have

$$
\sup_\gamma \left| w_{12}(\Sigma_n, \mu_n, \nu_n, \gamma) - w_{12}(\Sigma_1, P, \nu, \gamma) \right| \to 0.
$$

Note that

$$
\left| w_{12}(\Sigma_n, \mu_n, \nu_n, \gamma) - w_{12}(\Sigma_1, P, \nu, \gamma) \right|
$$
$$
\leq \left| w_{12}(\Sigma_n, \mu_n, \nu_n, \gamma) - w_{12}(\Sigma_n, \mu_n, \nu, \gamma) \right|
$$
$$
+ \left| w_{12}(\Sigma_n, P, \nu, \gamma) - w_{12}(\Sigma_1, P, \nu, \gamma) \right|
$$
$$
+ \left| w_{12}(\Sigma_n, \mu_n, \nu, \gamma) - w_{12}(\Sigma_n, P, \nu, \gamma) \right|
$$
$$
\leq \mu_n \left(\left| c^\mathsf{T} \Sigma_n^{-1} B_1 | H^\mathsf{T}_{2,1}(\nu_n - \nu)| \right| \right) + P \left(|c^\mathsf{T}(\Sigma_n^{-1} - \Sigma_{1,\infty}^{-1})B_1| \, |H^\mathsf{T}_{2,1}\nu| \right)
$$
$$
+ \left| (\mu_n - P) \left(c^\mathsf{T} \Sigma_n^{-1} B_1([H^\mathsf{T}_{2,1}\nu + H^\mathsf{T}_{2,1}\gamma]_+ - [H^\mathsf{T}_{2,1}\gamma]_+) 1_{H^\mathsf{T}_{2,1}\beta^*_{2,1}=0} \right) \right|
$$

By (A2), we have that $\|\Sigma_n^{-1}\|$ is bounded above for sufficiently large n, where $\|\cdot\|$ of a matrix denotes the spectral norm of the matrix. Thus the first term in the above display is bounded by $\|c\| \, \|\Sigma_n^{-1}\| \mu_n(\|B_1\| \, \|H_{2,1}\|) \, \|\nu_n - \nu\| = o(1)$, and the second term in the above display is bounded by $\|c\| \, \|\Sigma_1^{-1} - \Sigma_n^{-1}\| \, P(\|B_1\| \, \|H_{2,1}\|) \|\nu\| = o(1)$. For the third term, note that if $\|\nu\| = 0$, then it is zero. Otherwise,

$$
\left| (\mu_n - P) \left(c^\mathsf{T} \Sigma_n^{-1} B_1([H^\mathsf{T}_{2,1}\nu + H^\mathsf{T}_{2,1}\gamma]_+ - [H^\mathsf{T}_{2,1}\gamma]_+) 1_{H^\mathsf{T}_{2,1}\beta^*_{2,1}=0} \right) \right|
$$

TABLE 10
Parameters indexing the example models

Example	ξ	δ	Regularity
1	$(0,0,0,0,0,0,0,0,0,0)^{\intercal}$	$(0.5,0.5)^{\intercal}$	$p=1, \phi=0/0$
2	$(0,0,0,0,0.01,0.01,0,0,0,0)^{\intercal}$	$(0.5,0.5)^{\intercal}$	$p=0, \phi=\infty$
3	$(0,0,-0.5,0,0.5,0.5,0,0,0.5,0.5)^{\intercal}$	$(0.5,0.5)^{\intercal}$	$p=1/2, \phi=1.0$
4	$(0,0,-0.5,0,0.5,0.5,0,0,0.49,0.49)^{\intercal}$	$(0.5,0.5)^{\intercal}$	$p=0, \phi=1.0204$
5	$(0,0,-0.5,0,1.00,1.00,0.5,0.5,0.5,0.5)^{\intercal}$	$(1.0,0.0)^{\intercal}$	$p=1/4, \phi=1.4142$
6	$(0,0,-0.5,0,0.25,0.25,0.5,0.5,0.5,0.5)^{\intercal}$	$(0.1,0.1)^{\intercal}$	$p=0, \phi=0.3451$
A	$(0,0,-0.25,0,0.75,0.75,0.5,0.5,0.5,0.5)^{\intercal}$	$(0.1,0.1)^{\intercal}$	$p=0, \phi=1.035$
B	$(0,0,0,0,0.25,0.25,0,0,0.25,0.25)^{\intercal}$	$(0,0)^{\intercal}$	$p=1/2, \phi=1.00$
C	$(0,0,0,0,0.25,0.25,0,0,0.24,0.24)^{\intercal}$	$(0,0)^{\intercal}$	$p=1/2, \phi=1.00$

$$\leq \left|(\mu_n - P)\left(c^{\intercal}\Sigma_n^{-1}B_1([H_{2,1}^{\intercal}\nu/||\nu|| + H_{2,1}^{\intercal}\gamma/||\nu||]_+ \right.\right.$$
$$\left.\left. -[H_{2,1}^{\intercal}\gamma/||\nu||]_+)1_{H_{2,1}^{\intercal}\beta_{2,1}^*=0}\right)\right|\,||\nu|| \leq ||\mu_n - P||_{\mathcal{F}_{12}}||\nu|| = o(1).$$

This established the continuity of w_{12} and hence w'_{12}. The continuity of w_{11} can be established through similar arguments and is therefore omitted. □

Appendix C: Definitions of three-treatment models

Here, we present a suite of example models similar to those of Chakraborty et al. (2009), but that have three possible treatments at the second stage. These models are defined as follows:

- $X_i \in \{-1,1\}$ for $i \in \{1,2\}$,
  $A_1 \in \{-1,1\}$, and $A_2 \in \{(0,-0.5)^{\intercal}, (-1,0.5)^{\intercal}, (1,0.5)^{\intercal}\}$
- $P(A_1=1) = P(A_1=-1) = 1/2$,
  $P(A_2=(0,-1)^{\intercal}) = P(A_2=(-1,0.5)^{\intercal}) = P(A_2=(1,0.5)^{\intercal}) = 1/3$
- $P(X_1=1) = P(X_1=-1) = 1/2$,
  $P(X_2=1|X_1,A_1) = \mathrm{expit}(\delta_1 X_1 + \delta_2 A_1)$
- $Y_1 \triangleq 0$,
  $Y_2 = \xi_1 + \xi_2 X_1 + \xi_3 A_1 + \xi_4 X_1 A_1 + (\xi_5,\xi_6)A_2 + X_2(\xi_7,\xi_8)A_2 + A_1(\xi_9,\xi_{10})A_2 + \epsilon, \epsilon \sim N(0,1)$

where $\mathrm{expit}(x) = \mathrm{e}^x/(1+\mathrm{e}^x)$. This class is parameterized by twelve values $\xi_1, \xi_2, \ldots, \xi_{10}, \delta_1, \delta_2$. The analysis model uses histories defined by:

$$H_{2,0} = (1, X_1, A_1, X_1 A_1, X_2)^{\intercal} \tag{22}$$
$$H_{2,1} = (1, X_2, A_1)^{\intercal} \tag{23}$$
$$H_{1,0} = (1, X_1)^{\intercal} \tag{24}$$
$$H_{1,1} = (1, X_1)^{\intercal}. \tag{25}$$

Our working models are given by $Q_2(H_2, A_2; \beta_2) \triangleq H_{2,0}^{\intercal}\beta_{2,0} + H_{2,1}^{\intercal}\beta_{2,1,1}A_{2,1} + H_{2,1}^{\intercal}\beta_{2,1,2}A_{2,2}$ and $Q_1(H_1, A_1; \beta_1) \triangleq H_{1,0}^{\intercal}\beta_{1,0} + H_{1,1}^{\intercal}\beta_{1,1}A_1$. In Table 10, for each of these models we give the probability p of generating a history where

TABLE 11

Monte Carlo estimates of coverage probabilities for the ACI method at the 95% nominal level for different choices of λ_n. Here, $\beta_{1,1,1}$ denotes the main effect of treatment and $\beta_{1,0,1}$ denotes the intercept. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. No coverage estimates are significantly below 0.95 at the 0.05 level. Models have two treatments at each of two stages. Examples are designated NR = nonregular, NNR = near-nonregular, R = regular

$\beta_{1,1,1}$ $\lambda_n =$	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B R	Ex. C R
$\sqrt{\log \log n}$	0.989	0.987	0.967	0.969	0.954	0.952	0.950	0.962	0.962
$\log \log n$	0.992	0.992	0.968	0.972	0.957	0.955	0.950	0.964	0.965
$\log n$	0.993	0.994	0.975	0.976	0.962	0.966	0.959	0.969	0.972
$\sqrt{n}$	0.994	0.995	0.975	0.976	0.967	0.972	0.968	0.973	0.975
n	0.994	0.995	0.975	0.976	0.969	0.972	0.968	0.975	0.976

$\beta_{1,0,1}$ $\lambda_n =$	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B R	Ex. C R
$\sqrt{\log \log n}$	0.952	0.962	0.952	0.954	0.950	0.953	0.947	0.952	0.954
$\log \log n$	0.956	0.964	0.954	0.955	0.950	0.957	0.948	0.956	0.957
$\log n$	0.970	0.974	0.961	0.964	0.950	0.966	0.959	0.965	0.968
$\sqrt{n}$	0.971	0.975	0.963	0.968	0.954	0.973	0.965	0.974	0.978
n	0.971	0.975	0.987	0.987	0.979	0.980	0.975	0.983	0.984

each of the three possible treatments at the second stage have exactly the same effect. This is analogous to having the second stage treatment show no effect in a binary model. Furthermore, because of the Helmert encoding we have used in our analysis models, and because of the structure of ξ, it happens that the standardized effect size of treatment 1 versus treatment 2, treatment 1 versus treatment 3, and treatment 2 versus treatment 3 are all exactly equal in our examples. We report this as ϕ in Table 10.

Appendix D: Additional empirical results

Here we present additional empirical results. Tables 11 and 12 show the estimated coverage and interval diameter of the ACI across the nine generative models with two stages and two treatments per stage. The results appear stable across choices of λ_n for which the ACI is consistent. However, the ACI becomes quite conservative when λ_n is allowed to grow faster than $\sqrt{\log \log n}$.

Appendix E: The double bootstrap algorithm for selecting λ

Our algorithmic approach to choosing λ_n is similar to that used by Chakraborty et al. (2013) to choose m for their m-out-of-n bootstrap method. To select λ_n, we first draw r bootstrapped datasets $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(r)}$ from the original dataset $\mathcal{D}$. We take each of these in turn and compute an ACI bootstrap confidence interval at level $1 - \alpha$ with parameter $\lambda_n = \tau\sqrt{\log \log n}$ for $\tau \in \{0.125, 0.25, 0.5, 1, 2, 4\}$. (Because the ACI uses the bootstrap itself, it actually uses double-bootstraps of $\mathcal{D}$ to compute each interval.) Using the parameters estimated by Q-learning

TABLE 12

Monte Carlo estimates of mean width of the ACI method at the 95% nominal level for different choices of λ_n. Here, $\beta_{1,1,1}$ denotes the main effect of treatment and $\beta_{1,0,1}$ denotes the intercept. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. No corresponding estimated coverages are significantly below 0.95 at the 0.05 level. Models have two treatments at each of two stages. Examples are designated NR = nonregular, NNR = near-nonregular, R = regular

$\beta_{1,1,1}$ $\lambda_n =$	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B R	Ex. C R
$\sqrt{\log\log n}$	0.490	0.490	0.481	0.481	0.483	0.471	0.474	0.484	0.484
$\log\log n$	0.502	0.502	0.488	0.488	0.487	0.475	0.477	0.491	0.491
$\log n$	0.557	0.557	0.518	0.518	0.503	0.495	0.492	0.523	0.523
$\sqrt{n}$	0.583	0.582	0.533	0.533	0.513	0.514	0.511	0.540	0.540
n	0.586	0.586	0.538	0.538	0.525	0.521	0.519	0.543	0.543

$\beta_{1,0,1}$ $\lambda_n =$	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. A R	Ex. B R	Ex. C R
$\sqrt{\log\log n}$	0.506	0.506	0.481	0.481	0.483	0.490	0.474	0.490	0.490
$\log\log n$	0.518	0.518	0.487	0.487	0.486	0.494	0.476	0.497	0.498
$\log n$	0.574	0.574	0.517	0.517	0.502	0.517	0.493	0.540	0.541
$\sqrt{n}$	0.596	0.596	0.536	0.536	0.515	0.543	0.519	0.571	0.572
n	0.598	0.598	0.576	0.576	0.565	0.586	0.565	0.579	0.579

on the original $\mathcal{D}$ as ground truth, we compute for each value of τ the number of bootstrapped datasets $\kappa(\tau)$ for which the ACI covers. We then select τ^* to be the smallest τ that satisfies $\kappa(\tau)/r > 1 - \alpha$, and apply the ACI to the original dataset $\mathcal{D}$ using $\lambda = \tau^*\sqrt{\log\log n}$. In our experiments we used $r = 100$.

References

ANDREWS, D.W. and SOARES, G., Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection. *SSRN eLibrary*, 2007.

ANDREWS, D.W.K., Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000. MR1748009

ANDREWS, D.W.K., Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69:683–734, 2001a. MR1828540

ANDREWS, D.W.K., Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69(3):683–734, 2001b. MR1828540

ANDREWS, D.W.K. and GUGGENBERGER, P., Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators. *Journal of Econometrics*, 152(1):19–27, 2009. MR2562760

ANTHONY, M. and BARTLETT, P.L., *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. MR1741038

BARTO, A.G. and DIETERICH, T., Reinforcement learning and its relation to supervised learning. *Handbook of Learning and Approximate Dynamic Programming*, pages 45–63, 2004.

BELLMAN, R.E., *Dynamic Programming*. Princeton University Press, 1957. MR0090477

BERGER, R.L. and BOOS, D., P values maximized over a confidence set for the nuisance parameters. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

BICKEL, P.J., Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *The Annals of Statistics*, 9(6):1301–1309, 1981. MR0630112

BICKEL, P.J. and FREEDMAN, D.A., Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pages 1196–1217, 1981. MR0630103

BICKEL, P.J., KLAASSEN, A.J., RITOV, Y., and WELLNER, J.A., *Efficient and Adaptive Inference in Semi-Parametric Models*. Johns Hopkins University Press, Baltimore, 1993.

BICKMAN, L., KELLEY, S.D. and ATHAY, M., The technology of measurement feedback systems. *Couple and Family Psychology: Research and Practice*, 1(4):274–284, 2012.

BLUMENTHAL, S. and COHEN, A., Estimation of the larger of two normal means. *Journal of the American Statistical Association*, pages 861–876, 1968. MR0231477

BUSONIU, L., BABUSKA, R., DE SCHUTTER, B., and ERNST, D., *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, 2010.

CASELLA, G. and STRAWDERMAN, W.E., Estimating a bounded normal mean. *The Annals of Statistics*, pages 870–878, 1981. MR0619290

CHAKRABORTY, B., LABER, E.B., and ZHAO, Y., Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, TBA(TBA):TBA, 2013. MR3106599

CHAKRABORTY, B., MURPHY, S., and STRECHER, V., Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3), 2009. MR2757118

CHAKRABORTY, B. and MOODIE, E.E.M., *Statistical Methods for Dynamic Treatment Regimes*. Springer, 2013. MR3112454

CHAKRABORTY, B. and MURPHY, S.A., Dynamic treatment regimes. *Annual Review of Statistics and Its Application*, 1(1):null, 2014. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-022513-115553.

CHEN, J., Notes on the bias-variance trade-off phenomenon. *A Festschrift for Herman Rubin: Institute of Mathematical Statistics*, 45:207–217, 2004. MR2126898

CHENG, X., Robust confidence intervals in nonlinear regression under weak identification. *Job Market Paper*, 2008.

CSÖRGŐ, S. and ROSALSKY, A., A survey of limit laws for bootstrapped sums. *International Journal of Mathematics and Mathematical Statistics*, 45:2835–2861, 2003.

DAVISON, A.C. and HINKLEY, D.V., *Bootstrap Methods and Their Application*, volume 1. Cambridge university press, 1997. MR1478673

DUSSELDORP, E. and VAN MECHELEN, I., Qualitative interaction trees:

A tool to identify qualitative treatment–subgroup interactions. *Statistics in Medicine*, 2013.

FOSTER, J.C., TAYLOR, J.M.G. and RUBERG, S.J. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011. MR2844689

GOLDBERG, Y., SONG, R., and KOSOROK, M.R., Adaptive q-learning. *From Probability to Statistics and Back: High-Dimensional Models and Processes*, page 150, 2012. MR3186754

GUNTER, L., ZHU, J., and MURPHY, S.A., Variable selection for qualititative interactions. *Statistical Methodology*, 8(1):42–55, 2011. MR2741508

HAMBURG, M.A. and COLLINS, F.S., The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

HENDERSON, R., ANSELL, P., and ALSHIBANI, D., Regret-regression for optimal dynamic treatment regimes. *Biometrics*, 66(4), 2009. MR2758507

HIRANO, K. and PORTER, J., Impossibility results for nondifferentiable functionals. Mpra paper, University Library of Munich, Germany, 2009. URL http://econpapers.repec.org/RePEc:pra:mprapa:15990.

HIRANO, K. and PORTER, J.R., Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4):1769–1790, 2012. MR2977437

JANES, H., BROWN, M.D., PEPE, M., and HUANG, Y., Statistical methods for evaluating and comparing biomarkers for patient treatment selection, 2013.

KELLY, J., GOODING, P., PRATT, D., AINSWORTH, J., WELFORD, M., and TARRIER, N., Intelligent real-time therapy: Harnessing the power of machine learning to optimise the delivery of momentary cognitive-behavioural interventions. *Journal of Mental Health*, 21(4):404–414, 2012.

KONDA, V.R. and TSITSIKLIS, J.N., Onactor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003. MR2044789

KOSOROK, M.R., *Introduction to Empirical Processes and Semiparametric Inference.* Springer, 2008. MR2724368

LABER, E.B., LINN, L.A., and STEFANSKI, L.A., Interactive model-building for q-learning. *Biometrika*, to appear, 2014.

LABER, E., QIAN, M., LIZOTTE, D.J., and MURPHY, S.A., Statistical inference in dynamic treatment regimes. arXiv preprint arXiv:1006.5831, 2010.

LABER, E.B. and MURPHY, S.A., Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association*, 106(495):904–913, 2011. MR2894746

LAVORI, P.W. and DAWSON, R., A design for testing clinical strategies: Biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38, 2000.

LEEB, H. and POETSCHER, B.M., The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142, 2003. MR1965844

LEEB, H. and PÖTSCHER, B.M., The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142, 2003. MR1965844

H. LEEB and PÖTSCHER, B.M., Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01):21–59, 2005. MR2153856

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., and Murphy, S.A., A "smart" design for building individualized treatment sequences. *Annual Review of Clinical Psychology*, 8:21–48, 2012.

Liu, R.C. and Brown, L.D., Nonexistence of informative unbiased estimators in singular problems. *Annals of Statistics*, 21(1):1–13, 1993. MR1212163

Marchand, E. and Strawderman, W.E., Estimation in restricted parameter spaces: A review. *Lecture Notes-Monograph Series*, pages 21–44, 2004. MR2126884

Moodie, E.E.M., Richardson, T.S., and Stephens, D.A., Estimating optimal dynamic regimes: Correcting bias under the null. *Biometrics*, 63(2):447–455, 2010. MR2370803

Murphy, S.A., An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005a. MR2137651

Murphy, S.A., Van Der Laan, M.J., and Robins, J.M., Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001. MR1946586

Murphy, S.A., Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, 65(2):331–366, 2003. MR1983752

Murphy, S.A., A generalization error for Q-learning. *Journal of Machine Learning Research*, 6:1073–1097, Jul 2005b. MR2249849

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W.E., Gnagy, B., Fabiano, G.A., Waxmonsky, J.G., Yu, J., and Murphy, S.A., Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological methods*, 17(4):457, 2012a.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W.E., Gnagy, B., Fabiano, G.A., Waxmonsky, J.G., Yu, J., and Murphy, S.A., Q-learning: A data analysis method for constructing adaptive interventions. *Psychological methods*, 17(4):478, 2012b.

Olshen, R.A., The conditional level of the F-test. *Journal of the American Statistical Association*, 68(343):692–698, 1973. MR0359198

Orellana, L., Rotnitzky, A., and Robins, J., Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. *Int. Jrn. of Biostatistics*, 6(2), 2010. MR2602551

The PSU Methodology Center, Nih program announcements, January 2014a. URL https://methodology.psu.edu/ra/adap-inter/NIHfunding.

The PSU Methodology Center, Smart studies, January 2014b. URL http://methodology.psu.edu/ra/adap-inter/projects.

Putterman, M.L., *Markov Decision Processes*. John Wiely and Sons, New York, 1994.

Qian, M., Nahum-Shani, I., and Murphy, S.A., Dynamic treatment regimes. In *Modern Clinical Trial Analysis*, pages 127–148. Springer, 2013.

Robins, J., A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986. MR0877758

Robins, J.M., Addendum to "A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the

healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9):923–945, 1987. MR0922792

ROBINS, J.M., The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: A Focus on AIDS*, 113:159, 1989.

ROBINS, J.M., Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, volume 24, page 3. American Statistical Association, 1993.

ROBINS, J.M., Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, pages 69–117. Springer, 1997. MR1601279

ROBINS, J.M., Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. *Computation, Causation, and Discovery*, pages 349–405, 1999. MR1696459

ROBINS, J.M., Marginal structural models. 1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science, pp. 1–10, 1998.

ROBINS, J.M., Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, 2004. MR2129402

ROBINS, J.M., ORELLANA, L., and ROTNITZKY, A., Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, pages 4678–4721, 2008. MR2528576

RUBIN, D.B., Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978. MR0472152

SCHULTE, P.J., TSIATIS, A.A., LABER, E.B., and DAVIDIAN, M., Q- and a-learning methods for estimating optimal dynamic treatment regimes. Technical Report, arXiv:1202.4177v2, arXiv.org, 2013.

SI, J., BARTO, A.G., POWELL, W.B., WUNSCH, D.C., et al., *Handbook of Learning and Approximate Dynamic Programming*. IEEE Press Los Alamitos, 2004.

SONG, R., WANG, W., ZENG, D., and KOSOROK, M., Penalized q-learning for dynamic treatment regimes. Technical Report, arXiv:1108.5338v1, arXiv.org, 2011.

SUTTON, R.S., MCALLESTER, D.A., SINGH, S.P., and MANSOUR, Y., Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, pages 1057–1063, 1999.

SUTTON, R.S. and BARTO, A.G., *Reinforcment Learning: An Introduction*. The MIT Press, 1998.

SZEPESVÁRI, C., Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010.

TSIATIS, A.A., *Semiparametric Theory and Missing Data*. Springer Verlag, 2006. MR2233926

VAN DER LAAN, M.J., Causal effect models for intention to treat and realistic individualized treatment rules. 2006.

van der Laan, M.J. and Petersen, M.L., Causal effect models for realistic individualized treatment and intention to treat rules. *International Journal of Biostatistics*, 3(1):3, 2007. MR2306841

Van der Vaart, A., On differentiable functionals. *The Annals of Statistics*, pages 178–204, 1991. MR1091845

Van der Vaart, A. and Wellner, J., *Weak Convergence and Empirical Processes: With Application to Statistics*. Springer, 1996. MR1385671

Watkins, C.J.C.H. and Dayan, P., Q-learning. *Machine Learning*, 8(3):279–292, 1992.

Wiering, M. and van Otterlo, M., *Reinforcement Learning: State-of-the-art*, volume 12. Springer, 2012.

Zhang, B., Tsiatis, A.A., Laber, E.B., and Davidian, M., Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, To appear, 2013. MR3094445

Zhang, B., Tsiatis, A.A., Laber, E.B., and Davidian, M., A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012. MR3040007

Zhao, Y., Zeng, D., Laber, E.B., and Kosorok, M.R., New statistical learning methods for estimating optimal dynamic treatment regimes. *Under Review*, 107(499):1106–1118, 2013.

Zhao, Y., Zeng, D., Rush, A.J., and Kosorok, M.R., Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012. MR3010898