Robust Q-Learning

Ashkan Ertefaie‡, James R. McKay§, David Oslin¶, and Robert L. Strawderman∥

‡Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY; §Center on the Continuum of Care in the Addictions, Department of Psychiatry, University of Pennsylvania, Philadelphia, PA; ¶Philadelphia Veterans Administration Medical Center, and Treatment Research Center and Center for Studies of Addictions, Department of Psychiatry, University of Pennsylvania, Philadelphia, PA

ABSTRACT

Q-learning is a regression-based approach that is widely used to formalize the development of an optimal dynamic treatment strategy. Finite dimensional working models are typically used to estimate certain nuisance parameters, and misspecification of these working models can result in residual confounding and/or efficiency loss. We propose a robust Q-learning approach which allows estimating such nuisance parameters using data-adaptive techniques. We study the asymptotic behavior of our estimators and provide simulation studies that highlight the need for and usefulness of the proposed method in practice. We use the data from the “Extending Treatment Effectiveness of Naltrexone” multistage randomized trial to illustrate our proposed methods. Supplementary materials for this article are available online.

1. Introduction

A dynamic treatment strategy is a sequence of decision rules that maps individual characteristics to a treatment option at each decision point (i.e., a specific point in time in which a treatment is to be considered or altered). An optimal dynamic treatment strategy seeks to make these decisions to maximize a particular expected health outcome (Lavori and Dawson 2000; Murphy 2005; Lei et al. 2012; Nahum-Shani et al. 2012a; Davidian, Tsiatis, and Laber 2016). This is similar to clinical decision making whereby care providers tailor the type/dose of treatment over the course of clinical care based on ongoing information regarding patient progress in treatment.

The main goal of precision medicine (i.e., developing an effective dynamic treatment strategy) is to use patient characteristics to inform a personalized treatment plan as a sequence of decision rules that leads to the best possible health outcome for each patient (Nahum-Shani et al. 2012a; Chakraborty and Moodie 2013; Moodie and Kosorok 2015; Butler et al. 2018). Q-learning is a reinforcement learning algorithm that is widely used to estimate an optimal dynamic treatment strategy using data from multistage randomized clinical trials or observational studies (Watkins and Dayan 1992; Nahum-Shani et al. 2012b; Laber et al. 2014). Starting with the final study stage, Q-learning finds the treatment option that optimizes the desired expected outcome. Fixing the optimally chosen treatment at the final stage, Q-learning moves backward to the immediately preceding stage and searches for a treatment option assuming that future treatments will be optimized. The process continues until the first stage is reached. This backward induction procedure is designed to avoid treatment options that appear to be optimal in the short term but may lead to a less desirable long-term outcome (Chakraborty and Moodie 2013; Davidian, Tsiatis, and Laber 2016). Similar to other model-based approaches, model misspecification can seriously affect the result of Q-learning and the problem exacerbates substantially as the number of stages increases. Specifically, it can lead to residual confounding and suboptimal dynamic treatment strategies (Zhao, Kosorok, and Zeng 2009; Ertefaie, Shortreed, and Chakraborty 2016). An alternative approach is A-learning; this backward induction strategy is also model-based, hence subject to the possibility of misspecification, but imposes somewhat less restrictive regression model assumptions through modeling only the contrasts between treatments and the propensity of treatment assignment given the observed patient history (Murphy 2003; Schulte et al. 2014; Shi et al. 2018). However, this extra robustness comes at the price substantially reduced efficiency; indeed, Q-learning may lead to a parameter estimate that can be up to 170% more efficient than A-learning (Schulte et al. 2014). Policy learning methods are another class of methods for estimating an optimal dynamic treatment regime that circumvent the need for the conditional outcome models by directly optimizing the expected outcome among a class of rules (Zhao, Kosorok, and Zeng 2009; Zhao et al. 2011, 2012; Song et al. 2015). Despite this appealing feature, policy learning methods are inefficient and fail to provide reasonable inference for the parameter estimates that define non-smooth decision rules (e.g., indicator or max operators) due to slow rates of convergence (Moodie and Kosorok 2015; Zhao et al. 2015).

Doubly robust estimators that are based on modeling both the treatment and outcome processes have also been proposed for policy learning and for structural nested models (Zhang et al. 2012; Bai, Tsiatis, and O’Brien 2013; Zhang et al. 2013;
In general terms, consistency of the doubly robust estimators is guaranteed as long as either the treatment assignment mechanism or the postulated conditional mean outcome models are correctly specified, and semiparametric efficiency follows when both models are correctly specified (Rotnitzky, Robins, and Scharfstein 1998; van der Laan and Robins 2003; Tsiatis 2007). However, while doubly robust estimators give two routes for consistent estimation, the performance of these estimators depends critically on the modeling choice for the indicated treatment and mean outcome parameters. In practice, finite-dimensional models are used that are too restrictive and likely to be misspecified (Wallace and Moodie 2015). Kang and Schafer (2007) showed that doubly robust estimators can have poor performance when both models are misspecified. To mitigate this problem, bias reduction techniques have been proposed (Cao, Tsiatis, and Davidian 2009; Vermeulen and Vansteelandt 2015, 2016). An alternative is to use flexible learning-based methods that may reduce the chance of inconsistency (Benkeser et al. 2017).

In this article, we consider the problem of Q-learning for the setting of a two-stage dynamic binary treatment choice regime. A typical approach involves postulating linear models for both the first and second stage Q-functions that have stage-specific main effects involving pretreatment variables and interactions of these variables with the stage-specific treatment choice (e.g., Laber et al. 2014); however, only the interaction terms in these models directly influence the corresponding treatment decision functions that are optimized as part of the Q-learning process. Due to the nature of backward induction, the first stage model is likely to be a complicated function of the relevant covariates; misspecification of the main effect models in either stage can induce non-ignorable residual confounding. To increase robustness, we therefore consider the indicated main effects as unknown nuisance parameters and we adapt an approach originally proposed in Robinson (1988) for partial linear models that allows us to eliminate these parameters from the Q-functions. In particular, these hard-to-estimate parameters are replaced with models for the treatment assignment probability and mean outcome given pretreatment covariates that can be more easily estimated using, for example, nonparametric regression methods or related statistical learning methods (e.g., random forests). The resulting transformation process leads to consistent and asymptotically normal estimators of the so-called first- and second-stage blip functions that are robust to misspecification of the main effect nuisance parameters.

### 2. Notation and Formulation

Consider a two-stage study where binary treatment decisions are made at each time point. Let \( O = (X_1, A_1, X_2, A_2, Y) \) follow some probability distribution \( P_0 \) and suppose we observe \( N \) independent, identically distributed trajectories of \( O \). The vector \( X_1 \in X_1 \subset \mathbb{R}^{p_1} \) consists of all available baseline covariates measured before treatment at the first decision point \( A_1 \in \{0, 1\} \) and the vector \( X_2 \in X_2 \subset \mathbb{R}^{p_2} \) consists of all available intermediate covariates measured before treatment at the second decision point \( A_2 \in \{0, 1\} \). For notational convenience we define \( S_1^0 = (X_{i1}^T, A_{i1}, X_{i2}^T) \in S \subset \mathbb{R}^{p_1+p_2+1} \) and \( W_i^0 = X_{i1} \in X_1 \subset \mathbb{R}^{p_1} \). For later use, we also define variables \( S_i \) and \( W_i, i = 1, \ldots, n \); respectively, each represents some finite dimensional function of the variables in \( S_1^0 \) and \( W_0^0 \). Hence, knowledge of \( S_1^0 \) and \( W_0^0 \) respectively, implies knowledge of \( S_i \) and \( W_i \); however, the reverse may not hold. The observed outcome \( Y_i \in \mathbb{R} \) (measured after \( A_2 \)) is assumed continuous, with a larger value of \( Y_i \) indicating a better clinical outcome.

### 3. A Robust Formulation of Q-Learning

The outcome \( Y_i \) is assumed to satisfy the model

\[
Y_i = \eta_2(S_i^0) + A_{2i} \cdot \Delta_2(S_i^0) + \epsilon_{2i},
\]

\[
\mathbb{E}(\epsilon_{2i}|S_i^0, A_{2i}) = 0, \quad i = 1, \ldots, N,
\]

where the deterministic, unknown real-valued functions \( \eta_2(\cdot) \) and \( \Delta_2(\cdot) \) are defined on \( S \). Because the treatment variable \( A_2 \) is binary, the additive error model (1) places no parametric constraints on the conditional mean function.

In Q-learning, backward induction is used to characterize the optimal dynamic treatment regime. Define the second stage Q-function

\[
Q_2(s^0, a_2) = \mathbb{E}(Y|S = s^0, A_2 = a_2) = \eta_2(s^0) + a_2 \Delta_2(s^0); \tag{2}
\]

this measures "quality" when treatment \( a_2 \) is assigned to a patient with characteristics \( s^0 \) at the second stage (Laber et al. 2014). Similarly, define the first stage Q-function as

\[
Q_1(w^0, a_1) = \max_{a_2} \{ Q_2(S_1^0, a_2)|W = w^0, A_1 = a_1 \}, \tag{3}
\]

where, for \( a_2 \in \{0, 1\} \),

\[
\max_{a_2} Q_2(s^0, a_2) = \eta_2(s^0) + \Delta_2(s^0) I[\Delta_2(s^0) > 0].
\]

Analogously to \( Q_2(s^0, a_2) \), \( Q_1(w^0, a_1) \) measures "quality" when treatment \( a_1 \) is assigned to a patient with characteristics \( w^0 \) at baseline, assuming the optimal treatment choice is also made in the second stage. Because \( A_1 \) is binary, \( Q_1(w^0, a_1) \) can without loss of generality be written as \( \eta_1(w^0) + a_1 \Delta_1(w^0) \), where the real-valued functions \( \eta_1(\cdot) \) and \( \Delta_1(\cdot) \) are defined on \( X_1 \). It then follows that

\[
\max_{a_1} Q_1(w^0, a_1) = \eta_1(w^0) + \Delta_1(w^0) I[\Delta_1(w^0) > 0].
\]

Taken together, the optimal dynamic treatment regime is given by \( \delta^{opt}(s^0) = \{ d_{1pt}^{opt}(w^0), d_{2pt}(s^0) \} \), where \( d_{1pt}^{opt}(w^0) = I[\Delta_1(w^0) > 0] \) and \( d_{2pt}(s^0) = I[\Delta_2(s^0) > 0] \). We note that, under standard causal assumptions and formulated appropriately, \( \Delta_1(w^0) \) and \( \Delta_2(s^0) \) are commonly referred to as the first- and second-stage blip functions.

A widely used convention in the literature on dynamic treatment regimes is to, respectively, model \( Q_2(s^0, a_2; b_2) = b_{20} s_{20} + a_2 b_{21} s_{21} \) and \( Q_1(w^0, a_1; b_1) = b_{10} w_{10} + a_1 b_{11} w_{11} \), where \( s_{2k}, k = 0, 1 \) are sets of variables derived from \( s^0 \) (i.e., a realization of \( S^0 \)) and \( w_{1k}, k = 0, 1 \) are sets of variables derived from \( w^0 \) (i.e., a realization of \( W^0 \)). These model formulations impose restrictive assumptions on both \( \eta(\cdot) \) and \( \Delta(\cdot) \), \( j = 1, 2 \). However, the decision functions of interest only depend on the \( \Delta(j, \cdot) \), \( j = 1, 2 \), or per the indicated linear models, on the \( b_{jk} \)'s.
Because misspecification of the models for the nuisance parameters \( \eta_j(\cdot), j = 1, 2 \) can induce residual confounding and affect the causal interpretation of the interaction terms, we propose a novel modification of the Q-learning approach that eliminates the need to directly model \( \eta_j(\cdot), j = 1, 2 \). In particular, we adapt techniques originally introduced by Robinson (1988) for root-n consistent inference in semiparametric regression models, specifically partially linear models, to the Q-learning problem.

### 3.1. Regression Model for Stage 2 Decision Function

Define \( \mu_{2Y}(s^0) = \mathbb{E}[Y|S^0 = s^0] \) and \( \mu_{2A}(s^0) = \mathbb{E}(A_2|S^0 = s^0) \); for now, we will proceed as if these two functions are known. Under (1), or equivalently (2), \( \mu_{2Y}(s^0) = \eta_2(s^0) + \mu_{2A}(s^0)\Delta_2(s^0) \), implying that

\[
Y_i - \mu_{2Y}(S_i^0) = (A_{2i} - \mu_{2A}(S_i^0)) \cdot \Delta_2(S_i^0) + \epsilon_{2i}, \quad i = 1, \ldots, N. 
\]

Define the centered second stage Q-function

\[
Q_{2c}(s^0, a_2) = \mathbb{E}\{Y - \mu_{2Y}(S^0)|S^0 = s^0, A_2 = a_2\} = \{a_2 - \mu_{2A}(s^0)\} \cdot \Delta_2(s^0); 
\]

observe that \( Q_{2c}(s^0, a_2) = Q_2(s^0, a_2) - \mu_{2Y}(s^0) \). Although \( Q_{2c}(s^0, a_2) \) differs from \( Q_2(s^0, a_2) \), the action \( a_2 \) that maximizes \( Q_{2c}(s^0, a_2) \) is the same as which maximizes \( Q_2(s^0, a_2) \) because \( \mu_{2Y}(s^0) \) does not depend on the value of \( a_2 \). With \( \eta_2(s^0) \) eliminated from \( Q_{2c}(s^0, a_2; \beta_2) \), we propose to model (6) via

\[
Q_{2c}(s^0, a_2; \beta_2) = \{a_2 - \mu_{2A}(s^0)\} \cdot \Delta_2(s^0),
\]

where \( s \) denotes the realization of \( S \) (i.e., some finite set of variables derived from \( S^0 \)). Because

\[
\mathbb{E}\{Y - \mu_{2Y}(S^0)|Q_{2c}(s^0, A_2; \beta_2)\} = \{a_2 - \mu_{2A}(s^0)\} \cdot \Delta_2(s^0) - \beta_2, 
\]

it is easily shown that

\[
\beta_2 = \arg\min_{\beta_2} \mathbb{E}\left\{\{Y - \mu_{2Y}(S^0)|Q_{2c}(s^0, A_2; \beta_2)\}^2\right\} = \arg\min_{\beta_2} \mathbb{E}\left\{\{A_2 - \mu_{2A}(S^0)\}^2\{\Delta_2(s^0) - \beta_2\}^2\right\}. 
\]

The second expression shows that \( S^T \beta_2^2 \) is the best (weighted) linear predictor of \( \Delta_2(S^0) \). For data \( (Y_i, S_i^0), i = 1, \ldots, N \), the above developments further show that one can estimate \( \beta_2^2 \) by \( \beta_{2N} \), calculated as the minimizer of

\[
\sum_{i=1}^N \{Y_i - \mu_{2Y}(S_i^0) - Q_{2c}(S_i^0, A_{2i}; \beta_2)\}^2.
\]

Finally, maximizing \( Q_{2c}(s^0, a_2; \beta_2^2) \) for \( a_2 \in \{0, 1\} \) gives \( \beta_{2N}^2 = \mathbb{I}\{s^0|\beta_2^2 > 0\} \) as the optimal model-based treatment decision in Stage 2, with a corresponding estimated decision rule of \( \beta_{2N}^2 = \mathbb{I}\{s^0|\beta_{2N}^2 > 0\} \) (i.e., assuming \( \mu_{2Y}(\cdot) \) and \( \mu_{2A}(\cdot) \) are known).

The calculations above evidently rely on the availability of \( \mu_{2Y}(\cdot) \) and \( \mu_{2A}(\cdot) \). The assumption that \( \mu_{2Y}(\cdot) \) is known is particularly unrealistic; hence, in Section 4, we establish the properties of the corresponding least squares estimator when these functions are estimated from the available data using suitable consistent nonparametric estimators, such as those derived from random forests (e.g., Scornet, Biau, and Vert 2015) or Super Learner (van der Laan, Polley, and Hubbard 2007).

### 3.2. Regression Model for Stage 1 Decision Function

By construction, the relevant first stage Q-function depends on the model for (2). Using the partially linear model of Section 3.1, the model-based analog of the second stage Q-function (2) is given by

\[
Q_{2c}(s^0, a_2; \beta_2^2) = \mu_{2Y}(s^0) + Q_{2c}(s^0, a_2; \beta_2^2),
\]

where the second term is defined in (7). In view of (3), we therefore redefine the first stage Q-function of interest as

\[
Q_1(w^0, a_1) = \mathbb{E}\{\max_{a_2} Q_{2c}(s^0, a_2; \beta_2^2)|W^0 = w^0, A_1 = a_1\}. 
\]

Without loss of generality, and arguing similarly to the previous section, the fact that \( A_1 \) is binary means

\[
Q_1(w^0, a_1) = \eta_1(w^0) + a_1 \Delta_1(w^0)
\]

is a saturated nonparametric model for (9); here, \( a_1 \in \{0, 1\} \) and the real-valued functions \( \eta_1(\cdot) \) and \( \Delta_1(\cdot) \) are defined on \( X_1 \) and are not necessarily assumed to be the same functions that were initially used to define \( Q_1(w^0, a_1) \) at the beginning of Section 3.

Define

\[
\tilde{Y}^1 = \max_{a_2} Q_{2c}(s^0, a_2; \beta_2^2);
\]

then, it is not difficult to show that

\[
\tilde{Y}^1 = \mu_{2Y}(s^0) + S^T \beta_2^2 [I(S^T \beta_2^2 > 0) - \mu_{2A}(s^0)].
\]

Under (10), \( \mathbb{E}\{\tilde{Y}^1|W = w^0, A_1 = a_1\} = \eta_1(w^0) + a_1 \Delta_1(w^0); \)

Similarly, \( \mu_{1Y}(w^0) = \mathbb{E}\{\tilde{Y}^1|W = w^0\} = \eta_1(w^0) + \mu_{1A}(w^0) \Delta_1(w^0) \),

where \( \mu_{1A}(w^0) = \mathbb{E}(A_1|W = w^0) \). Similarly to (6), we can write

\[
Q_{1c}(w^0, a_1) = \mathbb{E}\{\tilde{Y}^1 - \mu_{1Y}(w^0)|W = w^0, A_1 = a_1\} = \{a_1 - \mu_{1A}(w^0)\} \Delta_1(w^0)
\]

and, considering (7), can model (12) via

\[
Q_{1c}(w^0, a_1; \beta_1) = \{a_1 - \mu_{1A}(w^0)\} w^T \beta_1,
\]

where \( w \) is defined analogously to \( s \). Together, these results imply that

\[
\mathbb{E}\{\tilde{Y}^1 - \mu_{1Y}(w^0) - Q_{1c}(w^0, A_1; \beta_1)|W = w^0, A_1 = a_1\} = \{a_1 - \mu_{1A}(w^0)\} \Delta_1(w^0) - w^T \beta_1.
\]

Similarly to the second stage problem, it now follows that

\[
\beta_1 = \arg\min_{\beta_1} \mathbb{E}\{\tilde{Y}^1 - \mu_{1Y}(w^0) - Q_{1c}(w^0, A_1; \beta_1)\}^2 
\]

\[
= \arg\min_{\beta_1} \mathbb{E}\{\{A_1 - \mu_{1A}(w^0)\}^2\{\Delta_1(w^0) - w^T \beta_1\}^2\}
\]

the latter implying that \( W^T \beta_1^2 \) is the best (weighted) linear predictor of \( \Delta_1(w^0) \). For data \( (Y_i, S_i^0), i = 1, \ldots, N \), and assuming that \( \beta_2^2, \mu_{1Y}(\cdot) \) and \( \mu_{1A}(\cdot) \) are all known, the above developments further imply that one can estimate \( \beta_1^2 \) using

\[
\beta_{1N} = \arg\min_{\beta_1} \sum_{i=1}^N \{\tilde{Y}^1_i - \mu_{1Y}(w_i^0) - Q_{1c}(w_i^0, A_{1i}; \beta_1)\}^2.
\]
Parallel to the second stage problem, the optimal model-based treatment decision in Stage 1, assuming the optimal model-based treatment is also given in Stage 2, would be \( d_1^{opt}(w, \beta_1^*) = I(w^T \beta_1^* > 0) \), and may be estimated by \( \hat{d}_1^{opt}(w) = I(w^T \hat{\beta}_{1N} > 0) \).

Of course, none of \( \beta_2^* \), \( \mu_{JY}(\cdot) \) and possibly \( \mu_{JA}(\cdot), j = 1, 2 \) are known in practice; in Section 4, we establish the properties of the corresponding least squares estimator when these quantities are all estimated.

Remark: As an alternative to (11), one can substitute
\[
\hat{Y} = Y + S^T \beta_{2n}^* I(S^T \beta_2^* > 0) - A_2
\]
for (11) when calculating \( \hat{\beta}_{1N} \); this follows directly from the equalities \( E(\hat{Y}|W^0 = w^0, A_1 = a_1) = E(Y|W^0 = w^0, A_1 = a_1) \) and \( E(\hat{Y}|W^0 = w^0) = E(Y|W^0 = w^0) \).

4. Robust Q-Learning: Estimation in Practice and Corresponding Theory

The two-stage procedure described in the previous section leads to a class of decision rules indexed by finite-dimensional parameter vectors, that is, \( d_1^{opt}(W, \beta_1^*) = I(W^T \beta_1^* > 0) \) and \( d_1^{opt}(S; \beta_2^*) = I(S^T \beta_2^* > 0) \) (e.g., Chakraborty and Moodie 2013). Although not explicit in prior developments, the variable selections \( S \) and \( W \) are each assumed to contain a column of ones, so that the main effects of treatment at each stage can be included as part of the decision rule. The proposed approach eliminates the nuisance parameters \( \eta_j(\cdot), j = 1, 2 \) from the first and second stage decision rules at the expense of introducing the four additional unknown functions \( \mu_{JY}(\cdot) \) and \( \mu_{JA}(\cdot), j = 1, 2 \). The advantage of the proposed approach is that the indicated functions depend on observables and can be easily estimated using any nonparametric regression or statistical learning method having sufficiently good prediction performance. Importantly, in the case of a sequentially randomized clinical trial, the functions \( \mu_{JA}(\cdot), j = 1, 2 \) are known and correct models are easily formulated.

4.1. Estimation in Practice

The developments in the next two subsections assume that the original sample, with \( N \) elements independently and identically distributed as \( P_0 \), has been randomly split into two disjoint and independent samples, say \( D_{E_1} \) and \( D_{E_2} \), with \( n = 0(N) \) (e.g., \( n = N/2 \)) and where \( I_E \) and its complement \( I_{E^c} \) form a partition of the index set \([1, \ldots, N]\). The induced nuisance parameters \( \hat{\mu}_{JY}(\cdot), \hat{\mu}_{2A}(\cdot), \hat{\mu}_{1Y}(\cdot), \) and \( \hat{\mu}_{1A}(\cdot) \) are to be estimated as described earlier using the data in \( D_{E_2} \); the finite dimensional parameters of interest are then estimated using the data \( D_{E_1} \), treating \( \hat{\mu}_{JY}(\cdot), \hat{\mu}_{2A}(\cdot), \hat{\mu}_{1Y}(\cdot), \) and \( \hat{\mu}_{1A}(\cdot) \) as known functions. As described above, the use of such sample-splitting is a particularly simple form of cross-fitting and can be generalized easily (Chernozhukov et al. 2018); our use of sample splitting as described above will be sufficient to establish the main ideas for both estimation and asymptotics without unnecessarily complicating notation. Generalization to cross-fitting is straightforward and will be discussed at the end of Section 4.

Let \( \hat{\mu}_{JY}(\cdot), \hat{\mu}_{1Y}(\cdot), \hat{\mu}_{2A}(\cdot), \) and \( \hat{\mu}_{1A}(\cdot) \) denote suitable estimates of \( \mu_{JY}(\cdot) \) and \( \mu_{2A}(\cdot), j = 1, 2 \) derived from the data in \( D_{E_2} \). Backward induction, implemented as described earlier with obvious modifications, can be used to estimate the optimal dynamic treatment regime. In particular, for the second stage, we compute
\[
\hat{\beta}_{2n} = \arg\min_{\beta_2} \sum_{i \in I_E} \left[ Y_i - \hat{\mu}_{JY}(S_i^0) - \{A_{2i} - \hat{\mu}_{2A}(S_i^0)\} \cdot S_i^0 \beta_2^* \right]^2.
\]
To estimate the first stage parameters, we first calculate the estimated first stage pseudo-outcome
\[
\hat{\tilde{Y}}_i = Y_i + S_i^T \hat{\beta}_{2n} I(S_i^T \hat{\beta}_{2n} > 0) - A_2;
\]
and then compute
\[
\hat{\beta}_{1n} = \arg\min_{\beta_1} \sum_{i \in I_E} \left[ \hat{\tilde{Y}}_i - \hat{\mu}_{1Y}(W_i^0) - \{A_{1i} - \hat{\mu}_{1A}(W_i^0)\} \cdot W_i^0 \beta_1^* \right]^2.
\]

The notation in (15) and (17) emphasizes the fact that the nuisance parameters \( \mu_{JY}(\cdot), \mu_{2A}(\cdot), \mu_{1Y}(\cdot), \) and \( \mu_{1A}(\cdot) \) are estimated using the outcome and full set of either second or first stage covariates, whereas the linear specifications used for modeling the centered Q-functions might not use all available covariate information.

As defined, the pseudo-outcomes are nonsmooth functions of the data, hence so is \( \hat{\beta}_{1n} \); this can cause non-regularity problems for \( \hat{\beta}_{1n} \) (Laber et al. 2014). In particular, when \( Pr(\{S_j^T \beta_1^* > 0\}) = 0 \), that is, there exists a strata of the covariates \( S \) used to model the Q-function that occurs with positive probability and for which treatment is neither beneficial nor harmful, the estimators of first stage regression coefficients become non-regular due to the nondifferentiability of the indicator function in the definition of the pseudo-outcome.

The proposed Q-learning models essentially utilize the propensity score regression approach of Robins, Mark, and Newey (1992) to eliminate the problem of mismodeling hard-to-estimate infinite-dimensional parameters (i.e., \( \eta_j(\cdot), j = 1, 2 \)) on the estimators of the \( \beta_j^* \)'s. The resulting estimator of \( \beta_j^* \) is consistent and asymptotically normal under suitable conditions on \( \hat{\mu}_{JA}(\cdot) \) and \( \hat{\mu}_{jY}(\cdot), j = 1, 2 \). In particular, the estimate of \( \beta_j^* \) is robust to misspecification of \( \mu_{jY}(\cdot) \) provided that \( \mu_{jA}(\cdot) \) is consistently estimated where \( \beta_j^* \) represents the parameters of the best linear approximation of the unknown \( \Delta_j(\cdot) \). In practice, we recommend using ensemble learning methods such as Super Learner (van der Laan, Polley, and Hubbard 2007) for estimating both \( \hat{\mu}_{jA}(\cdot) \) and \( \hat{\mu}_{jY}(\cdot) \). Asymptotically, Super Learner performs as well as the best convex combination of the base learners in the chosen library, in the sense of minimizing the difference in risk compared to the corresponding oracle estimator. Moreover, the size of the library can grow at a polynomial rate compared with the sample size without affecting its oracle performance (Dudoit and van der Laan 2003; van der Laan and Dudoit 2003; van der Vaart, Dudoit, and van der Laan 2006). For these reasons, it is recommended that the library consist of a large and diverse set of regression modeling procedures (i.e., nonparametric, semi-parametric, parametric). Importantly, these theoretical results only imply that Super Learner can match the performance of...
the (unknown) best possible convex combination of choices in the specified library. Thus, consistency is not guaranteed unless the corresponding oracle estimator is consistent and converges sufficiently fast. However, with the use of a sufficiently flexible library, Super Learner clearly improves one’s ability to construct a consistent estimator because it eliminates the need to select and subsequently rely on a single method of estimation.

### 4.2. Theoretical Results

To further simplify notation, let \( \Delta_2 = \Delta_2(S^0) \), \( \Delta_1 = \Delta_1(W^0) \), \( \mu_{2A} = \mu_{2A}(S^0) \), \( \mu_{2A} = \mu_{2A}(S^0) \), \( \mu_{1A} = \mu_{1A}(W^0) \), and \( \mu_{1A} = \mu_{1A}(W^0) \). In addition, with \( x^{\otimes 2} = xx^\top \) for any vector \( x \), define the matrices

\[
V_{2n} = \frac{1}{n} \sum_{i \in I_n} (A_{2i} - \mu_{2A})^2 S^{\otimes 2}_i \quad \text{and} \quad \hat{V}_{2n} = \frac{1}{n} \sum_{i \in I_n} (A_{2i} - \hat{\mu}_{2A})^2 S^{\otimes 2}_i ,
\]

\[
V_{1n} = \frac{1}{n} \sum_{i \in I_n} (A_{1i} - \hat{\mu}_{1A})^2 W^{\otimes 2}_i \quad \text{and} \quad \hat{V}_{1n} = \frac{1}{n} \sum_{i \in I_n} (A_{1i} - \hat{\mu}_{1A})^2 W^{\otimes 2}_i .
\]

Let \( \|x\|_q \) denote the usual Q-norm of a vector \( x \) for \( q = 1, 2, \infty \). Also, for \( Z \sim P \) for some probability measure \( P \), suppose \( f(\cdot) \) is any real-valued, P-measurable function; then, we define the \( L^2(P) \) norm of \( f(\cdot) \) as \( \|f(Z)\|_2 = \{ f(\omega) \, dP(\omega) \}^{1/2} \). For a real-valued function \( h(S^0; D_i^n) \) defined for \( s^0 \in S \) whose calculation may depend on the data contained in \( D_i^n \), we can also define the random norm \( \|h(S^0; D_i^n)\|_{P_n,2} \) as the square-root of

\[
\|h(S^0; D_i^n)\|_{P_n,2}^2 = E \{ \|h(S^0; D_i^n)\|_{P_n,2}^2 \}
\]

where \( P_n \) denotes the empirical measure on \( D_i^n \).

Our results are established under the following assumptions.

**Assumption 1.** (i) The support of \( W^0 \) and the conditional treatment effect \( \Delta_1(W^0) \) are uniformly bounded; (ii) the support of \( S^0 \) and the conditional treatment effect \( \Delta_2(S^0) \) are uniformly bounded; and, the supports of \( S \) and \( W \) are uniformly bounded.

**Assumption 2.** (i) \( \|\hat{\mu}_{1A}(W^0; D_i^n) - \mu_{1A}(W^0)\|_{P_n,2} = o_p(n^{-1/2}) \); (ii) \( \|\hat{\mu}_{2A}(S^0; D_i^n) - \mu_{2A}(S^0)\|_{P_n,2} = o_p(n^{-1/2}) \).

**Assumption 3.** (i) \( \|\hat{\mu}_{1Y}(W^0; D_i^n) - \mu_{1Y}(W^0)\|_{P_n,2} = o_p(1) \); (ii) \( \|\hat{\mu}_{2Y}(S^0; D_i^n) - \mu_{2Y}(S^0)\|_{P_n,2} = o_p(1) \).

**Assumption 4.** (i) \( \|\hat{\mu}_{1Y}(W^0; D_i^n) - \mu_{1Y}(W^0)\|_{P_n,2} = o_p(n^{-1/2}) \); (ii) \( \|\hat{\mu}_{2Y}(S^0; D_i^n) - \mu_{2Y}(S^0)\|_{P_n,2} = o_p(n^{-1/2}) \).

**Assumption 5.** There exists \( 1 \leq n_0 < \infty \) such that \( V_{jn} \) and \( \hat{V}_{jn}, j = 1, 2 \) are positive definite for \( n \geq n_0 \).

**Assumption 6.** \( P(\{s^0 \neq \beta^*_2 \}) = 0 \).

Assumption 1 requires no discussion. Assumptions 2–4 impose reasonable conditions on the estimators of the nuisance parameters estimated using cross fitting that are satisfied by many machine learning algorithms; see Chernozhukov et al. (2018) for further discussion. Assumption 5 imposes reasonable conditions on the existence and uniqueness of the least squares estimators (15) and (17). Assumption 5 combined with independent, identically distributed sampling ensures that the limiting matrices

\[
V_2 = E \{ \var{A_2|S^0} S^{\otimes 2} \} \quad \text{and} \quad V_1 = E \{ \var{A_1|W^0} W^{\otimes 2} \}
\]

both exist and are positive definite. Finally, Assumption 6 is imposed to avoid nonregular asymptotic behavior in the first stage least squares estimator (17). Inferences for the parameters that define the estimated optimal dynamic treatment regime \( \hat{d}_2^{opt}(s) = I(s^\top \hat{\beta}_{2n} > 0) \) and \( d_1^{opt}(w) = I(w^\top \hat{\beta}_{1n} > 0) \) can now be derived using the results in the following theorem.

**Theorem 1.** Suppose that Assumptions 1–6 hold.

(a) Let \( \hat{\beta}_{2n} \) be given by (15). Then, \( \sqrt{n}(\hat{\beta}_{2n} - \beta^*_2) \overset{d}{\to} N(0, V_2^{-1} Q_2 V_2^{-1}) \) where the matrices \( V_2 = E \{ \var{A_2|S^0} S^{\otimes 2} \} \) and \( Q_2 = E \{ J_2^{\otimes 2} \} \) for

\[
J_2 = [A_2 - \mu_{2A}(S^0)] \{ Y - \mu_{2Y}(S^0) \} - [A_2 - \mu_{2A}(S^0)] S^\top \beta^*_2 .
\]

(b) Let \( \hat{\beta}_{1n} \) be given by (17). Then, \( \sqrt{n}(\hat{\beta}_{1n} - \beta^*_1) \overset{d}{\to} N(0, V_1^{-1} Q_1 V_1^{-1}) \) where the matrices \( V_1 = E \{ \var{A_1|W^0} W^{\otimes 2} \} \) and \( Q_1 = E \{ (J_1 + KV_2^{-1} J_2^{\otimes 2}) \} \) for

\[
J_1 = [A_1 - \mu_{1A}(W^0)] W \{ Y - \mu_{1A}(W^0) \} - [A_1 - \mu_{1A}(W^0)] W^\top \beta^*_1
\]

and

\[
K = E \{ [A_1 - \mu_{1A}(W^0)] \{ I(S^\top \beta^*_2 > 0) - A_2 \} WS^\top \} .
\]

The following corollary to Theorem 1 shows that Assumption 6 is not required to establish the results in part (b) in certain settings, in contrast to the standard form of Q-learning.

**Corollary 1.** Suppose Assumptions 1–5 hold. In addition, suppose \( E \{ A_{1i} - \mu_{1Ai} | W_i, S_i, I_n \} = 0, i \in I_n \). Then, Theorem 1, part (b) remains true.

The set of variables \( S \) used for modeling the second stage decision rule are those thought to be potential effect modifiers for the second stage treatment assignment \( A_2 \); hence, a sufficient
condition for the Corollary to hold is that $A_1$ is independent of $S$, conditionally on the set of pretreatment covariates $W$ included in the first stage model. Note that this does not preclude the possibility that $A_1$ affects variables in $S^0$ that are not part of $S$. Due to the way in which $A_{ij} = \mu_{1A}(W)$ enters the estimation equation of Wallace and Moodie (2015, eq. (4)), it is unclear whether their approach avoids nonregularity under the same conditions as Corollary 1 even when the propensity model $\mu_{1A}(W^0)$ has been correctly specified.

4.3. Generalization to Cross-Fitting

Sample splitting, as used in the previous two sections, does not make use of the full sample of $N$ observations to estimate the finite-dimensional regression parameters, and this can negatively impact efficiency. We now describe an alternative approach, cross-fitting, that uses the full sample to estimate the desired target parameters.

Suppose that $N = nK$ for some integer $n$ and some integer $K \geq 2$. Using an extension of previous notation, we first randomly split the original sample into disjoint (hence independent) samples $(D_{nk})^K_{k=1}$ such that the size of each sample is $n = N/K$ and $I_{nk}, k = 1, \ldots, K$ partition the indices $\{1, \ldots, N\}$. Analogously to before, define $I_{nk}$ as the set of sample indices that are not included in $I_{nk}$; that is, $I_{nk} = \{1, 2, \ldots, N\} \setminus I_{nk}, k = 1, \ldots, K$. Then, for each $k = 1, 2, \ldots, K$, estimate the nuisance parameters $\mu_{2Y}(\cdot), \mu_{2A}(\cdot), \mu_{1Y}(\cdot), \mu_{1A}(\cdot)$ using the data in $D_{nk}$; we, respectively, denote these estimators $\hat{\mu}_{2Y}(\cdot; D_{nk}), \hat{\mu}_{2A}(\cdot; D_{nk}), \hat{\mu}_{1Y}(\cdot; D_{nk}), \text{and } \hat{\mu}_{1A}(\cdot; D_{nk}), k = 1, \ldots, K$. Finally, we define

$$\hat{\beta}_{2n} = \arg\min_{\hat{\beta}_2} \sum_{k=1}^K \sum_{i \in I_{nk}} \left[ Y_i - \hat{\mu}_{2Y}(S^0_i; D_{nk}) \right]^2 - \left\{ A_{2i} - \hat{\mu}_{2A}(S^0_i; D_{nk}) \right\} \cdot S^0_i \cdot \hat{\beta}_2 \right)^2 \tag{19}$$

and

$$\hat{\beta}_{1n} = \arg\min_{\hat{\beta}_1} \sum_{k=1}^K \sum_{i \in I_{nk}} \left[ Y_i - \hat{\mu}_{1Y}(W^0_i; D_{nk}) \right]^2 - \left\{ A_{1i} - \hat{\mu}_{1A}(W^0_i; D_{nk}) \right\} \cdot W^0_i \cdot \hat{\beta}_1 \right]^2 \tag{20}$$

This form of cross-fitting essentially corresponds to “DML2” as described in Chernozhukov et al. (2018, Def. 3.2). Like sample splitting, cross-fitting helps to guarantee that some of the remainder terms in the asymptotic linearity expansion converge to zero at an appropriately fast rate. However, in contrast to sample splitting, cross-fitting as described above is also capable of asymptotically achieving the same efficiency as in the case where estimators of the regression parameters are computed using all $N$ observations (i.e., with $\mu_{jA}(\cdot)$ and $\mu_{jY}(\cdot)$ $j = 1, 2$ being known).

5. Simulation Studies

We examined the performance of our proposed Q-learning method under different simulation scenarios with various functional complexities and degrees of nonregularity (i.e., violation of Assumption 6).

The main simulation in the regular setting uses the following data generation mechanism. Let $X_1 = (X_{11}, X_{12}, X_{13}, X_{14}, X_{15})^T$ be a five-dimensional vector of baseline covariates independently generated and uniformly distributed on $[-0.5, 0.5]$. Let $X_2 = (X_{21}, X_{22}, X_{23}, X_{24}, X_{25})^T$, $X_3 = X_{11} + U_i, i = 1, 2, 3; X_{24} = 0.35X_{15} + U_i$; and, $X_{25} = U_5$, where $U_i, i = 1, \ldots, 5$ are independent and uniformly distributed on $[-0.5, 0.5]$. It is assumed that only nonresponders to the first stage treatment will receive the second stage treatment. This nonresponse indicator $R$ equals 1 if $X_{24}$ is less than its median value and is 0 otherwise. Finally, the first and second stage treatments $A_j$ are generated from a Bernoulli distribution with success probability $\mu_{jA}(\cdot) = [1 + \exp(-\lambda_{jA}(\cdot))]^{-1}$, where $\lambda_{jA}(\cdot)$ depends on either $S^0 = (X^0_1, A_1, X^0_2)^T$ (j=2) or $W^0 = X_1 (j=1)$; see Section 5.1.

5.1. Performance: Regular Setting

In this case, we consider performance for models that satisfy Assumption 6. To implement our proposed method, we used the R package SuperLearner (Polley et al. 2019) to estimate $\mu_{1Y}(\cdot), \mu_{2Y}(\cdot), \mu_{1A}(\cdot)$, and $\mu_{2A}(\cdot)$. The library used for SuperLearner included generalized linear models (i.e., glm), generalized additive models (i.e., gam; Hastie 2019), multivariate adaptive regression splines (i.e., earth; Milborrow 2019), random forests (i.e., randomForest; Liaw and Wiener 2002), and support vector machines (i.e., svm from the R package e1071; Meyer et al. 2019); estimation was implemented with all tuning parameters set to their respective default values. This simulation study uses four different functional forms for the treatment assignment model $\mu_{jA}(\cdot) = [1 + \exp(-\lambda_{jA}(\cdot))]^{-1}$, $j = 1, 2$:

- **Randomized**: $\lambda_{jA}(\cdot) = 0$
- **Linear**: $\lambda_{jA}(\cdot) = 2X_{11} + 2X_{22} + X_{33} + 0.1X_{44} + 0.1X_{55}$
- **Quadratic**: $\lambda_{jA}(\cdot) = 1.4((X_{11} - 0.5)^2 + (X_{22} - 0.5)^2 + 0.6X_{15} - 0.5)^2 + 0.5X_{14} - 0.5)^2 + 0.5X_{15} - 0.5)^2 + X_{11} + X_{22} + 0.6X_{33} + 0.5X_{44} + 0.5X_{55} = 2$
- **InterQuad**: $\lambda_{jA}(\cdot) = 1.4((X_{11} - 0.5)^2 + (X_{22} - 0.5)^2 + 0.6X_{15} - 0.5)^2 + 0.5X_{14} - 0.5)^2 + 0.5X_{15} - 0.5)^2 + X_{11} + X_{22} + 0.6X_{33} + 0.5X_{44} + 0.5X_{55} + X_{12} + X_{23} - 2$

The Randomized model corresponds to a SMART-like trial where simple randomization is used at baseline and then simple rerandomization occurs among the set of non-responders at the first stage. The randomization model, part of the trial design, is therefore known and the inclusion of an appropriate glm model in the Super Learner library should ensure that $\mu_{jA}(\cdot)$ can be consistently estimated at the usual parametric rate. The other three settings are meant to correspond to increasingly complex observational data settings, where the “assignment” mechanism by which patients follow a particular treatment regimen is covariate-dependent, not randomized, and is not
considered to be known by design. Hence, the analyst cannot knowingly select a correctly specified parametric model a priori. In the case of the Linear model, the inclusion of a glm model in the Super Learner library again ensures that \( \mu_{ja}(\cdot) \) can be consistently estimated at the usual parametric rate. For the other two models, the inclusion of methods such as gam and randomForest will help to mitigate, but not necessarily, the possibility of inconsistent estimation. These observations highlight the importance of using flexible methods when modeling \( \mu_{ja}(\cdot), j = 1, 2 \), particularly in observational data settings.

The outcome models are given by

- **Linear**: \( Y = X_j^T \alpha_1 + X_j^T \alpha_2 + A_1X_j^T \theta_1 + A_2RX_j^T \theta_2 + \epsilon \) where \( \alpha_1 = (1, 0.1, 0.1, 0.1, 0.1)^T \), \( \theta_1 = (0, 0, 0, 0, 0)^T \) and \( \theta_2 = (1, 1, 0, 0, 0)^T \);
- **FGS**: \( Y = f(X_1) + f(X_2) + A_1X_j^T \theta_1 + A_2Rg(X_2) + \epsilon \) where \( \theta_1 = (0, 0, 0, 0, 0)^T \) and for \( x = (x_1, x_2, x_3, x_4, x_5)^T \), we set \( g(x) = 2\sin(\pi x_1 x_2^2) + 2(x_2 - 0.5)^2 \) and
  \[
  f(x) = -1.5 + \sin(\pi x_1 x_2) + 2(x_3 - 0.5)^2 + x_4 + 1.5 \frac{x_1}{|x_2| + |x_3|} + 2x_1(x_2 + x_3).
  \]

The noise variable \( \epsilon \) is generated from \( N(0, \sigma = 0.5) \).

In the connecting the above Linear and outcome model specification with earlier notation, we have \( S^0 = (X_j^T, A_1, X_j^T)^T \), \( \eta_2(S^0) = X_j^T \alpha + X_j^T \alpha_2 + A_1X_j^T \theta_1 \) and \( \Delta_2(S^0) = RX_j^T \theta_2 \), where \( R \) is a function of \( X_{24} \); only we further have \( S^0 = X_1 \). We, respectively, use \( S = (1, X_{21}, X_{22}, X_{23})^T \) and \( W = (1, X_{11}, X_{12})^T \) for modeling the relevant Q-functions. In this case, the target of estimation \( \beta^*_{1} = (0, \theta_{21}, \theta_{22}, \theta_{23})^T \) and it can additionally be shown that (6) coincides with (7). However, for the FGS outcome model, \( \eta_2(S^0) = f(X_1) + f(X_2) + A_1X_j^T \theta_1 \) and \( \Delta_2(S^0) = Rg(X_2) \); here, (6) does not coincide with (7) since the linear parametric specification used in the latter is not equal to \( \Delta_2(S) \). In this case \( S^0 \beta^* \) still exists as the best linear projection of \( \Delta_2(S^0) \) on to the linear space spanned by \( S \); however, its value for this simulation study must be determined numerically (e.g., through simulation).

In general, it is not similarly straightforward to characterize the functions \( \eta_1(W^0) \) and \( \Delta_1(W^0) \), or the value of \( \beta^*_1 \) in the first stage models, without appealing to numerical methods. However, in the current simulation setting, the value of \( \beta^*_1 \) can be determined exactly for both the Linear and FGS outcome model specifications. Specifically, neither model involves an interaction between \( A_1 \) and \( A_2 \); more generally, there is no correlation between \( A_1 \) and the second stage variables \( A_2 \) and \( X_3 \). As a result, the linear term \( A_1X_j^T \theta_1 \) that appears in both the Linear and FGS outcome model specifications accurately describes the interaction between treatment \( A_1 \) and \( X_1 \) in the true first stage Q-function (i.e., \( \Delta_1(W^0) = W^T \theta_1 = X_j^T \theta_1 = 0 \)). It follows that \( \beta^*_1 = 0 \) and hence that expression (12) also coincides with (13).

In our main simulation study, there are 8 possible model combinations represented by the outcome and treatment assignment models, and within each setting we compare the performance of the proposed method for estimating \( \beta^*_1, j = 1, 2 \) to the standard form of Q-learning \( (Q_{N,N}) \) and also to the weighted least squares \( (dWOLS_{N,N}) \) estimator proposed by Wallace and Moodie (2015). The subscripts on these latter two estimators denote the fact that standard errors would normally be calculated using the N-out-of-N bootstrap (i.e., in the regular setting). In the case of dWOLS\(_{N,N}\), linear models are used for the relevant Q-function model specification and logistic regression models are used for estimating the treatment assignment probabilities. The estimation of \( \beta^*_1 \) for the proposed method under the Linear\(_{K} \) outcome model specification. However, there is a possibility of such bias under the FGS\(_{K} \) in the case of \( Q_{N,N} \) and dWOLS\(_{N,N} \). To be more specific, residual confounding bias under the FGS\(_{K} \) outcome model is expected for \( Q_{N,N} \) regardless of the treatment assignment model. For dWOLS\(_{N,N} \), the Randomized and Linear first and second stage treatment assignment models are correctly specified and easily modeled. Hence, under the FGS\(_{K} \) outcome model specification, a significant potential for bias arises only under the Quadratic or InterQuad treatment assignment rules. For the proposed method, residual confounding bias when estimating \( \beta^*_1 \) is not anticipated provided that \( \mu_{ja}(\cdot), j = 1, 2 \) are sufficiently well-estimated.

We generate 500 datasets of size 2000 to examine the performance of our proposed method and use cross-fitting as described in Section 4.3 with \( K = 2 \) to estimate the desired target parameters. Tables 1 and 2 show the empirical absolute bias and standard deviations of the second and first stage parameter estimates (i.e., standard errors). The values of \( \beta^*_{21,1} \approx 0 \) and \( \beta^*_{22,1} \approx -2 \) are determined by simulation. As expected, standard Q-learning performs poorly except under the Randomized treatment assignment model. The proposed method and dWOLS\(_{N,N} \) also perform similarly well under the Randomized and Linear treatment assignment models for estimating the first and second stage parameters. However, under the FGS\(_{K} \) outcome model, the proposed method exhibits similar biases and substantially smaller standard errors. For the Quadratic and InterQuad treatment assignment mechanism, both of which are mismodeled in the case of dWOLS\(_{N,N} \), the corresponding estimators show substantial bias in some of the parameters, whereas those for the proposed method remain comparatively low. For example, under the InterQuad treatment assignment model and FGS\(_{K} \) outcome model, the proposed method, respectively, results in estimators for \( \beta^*_{21,1} \) and \( \beta^*_{22,1} \) with absolute biases of 0.070 and 0.019; in contrast, those for the dWOLS\(_{N,N} \) estimators are 0.758 and 0.455, respectively. We again see a substantial reduction in standard errors; in this same example, the standard errors under the proposed method are 0.612 and 0.271, whereas for dWOLS\(_{N,N} \) these are, respectively, 0.916 and 0.402, the degree of reduction exceeding 30%. Overall, the proposed method is observed to be more robust, typically producing less biased estimators with smaller standard errors compared with the other two approaches.

The performance of our proposed method was also assessed using smaller sample sizes. Tables S1–S6 in the supplementary materials, respectively, show the results for \( \beta^*_1 \) and \( \beta^*_2 \) with sample sizes of 1000, 500, and 250. Overall, the proposed method continues to outperform both \( Q_{N,N} \) and dWOLS\(_{N,N} \), particularly when the underlying treatment assignment and the outcome models are both nonlinear (i.e., settings in which bias can be expected for both \( Q_{N,N} \) and dWOLS\(_{N,N} \)). However, the performance of the proposed method is also affected by sample
Table 1. Performance of the proposed Q-learning method for estimating the second stage parameters under different model complexities.

| Outcome | $\beta_1^2$ | $\beta_2^2$ |
|---------|-------------|-------------|
|         | Randomized | Proposed   | dWOLS | Randomized | Proposed | dWOLS |
|         | $Q_{N,N}$  | $Q_{N,N}$  | $Q_{N,N}$ | $Q_{N,N}$  | $Q_{N,N}$  | $Q_{N,N}$ |
| LinearR | 0.003      | 0.003      | 0.002 | 0.001      | 0.004      | 0.002 |
| FGS     | 0.041      | 0.055      | 0.024 | 0.022      | 0.037      | 0.021 |
| LinearR | 0.004      | 0.004      | 0.006 | 0.003      | 0.000      | 0.002 |
| FGS     | 2.500      | 0.660      | 0.500 | 2.527      | 0.064      | 0.055 |
| LinearR | 0.006      | 0.005      | 0.007 | 0.004      | 0.001      | 0.003 |
| FGS     | 0.797      | 0.586      | 0.811 | 0.012      | 0.017      | 0.022 |
| LinearR | 0.000      | 0.014      | 0.007 | 0.002      | 0.019      | 0.002 |
| FGS     | 0.749      | 0.612      | 0.758 | 0.442      | 0.271      | 0.455 |

NOTE: The true parameters for the linear and FGS outcome models are, respectively, $\beta_1^2 = 1$, $\beta_2^2 = 1$, and $\beta_1^1 \approx 0, \beta_2^2 \approx -2$.

Table 2. Performance of the proposed Q-learning method for estimating the first stage parameters under different model complexities.

| Outcome | $\beta_1^1$ | $\beta_1^2$ |
|---------|-------------|-------------|
|         | Randomized | Proposed   | dWOLS | Randomized | Proposed | dWOLS |
|         | $Q_{N,N}$  | $Q_{N,N}$  | $Q_{N,N}$ | $Q_{N,N}$  | $Q_{N,N}$  | $Q_{N,N}$ |
| LinearR | 0.007      | 0.001      | 0.005 | 0.000      | 0.001      | 0.000 |
| FGS     | 0.000      | 0.001      | 0.001 | 0.000      | 0.000      | 0.000 |
| LinearR | 0.173      | 0.001      | 0.005 | 0.163      | 0.002      | 0.003 |
| FGS     | 1.915      | 0.058      | 0.043 | 1.657      | 0.010      | 0.003 |
| LinearR | 2.404      | 0.015      | 0.003 | 2.076      | 0.003      | 2.000 |
| FGS     | 7.235      | 0.009      | 0.034 | 1.631      | 0.004      | 1.000 |
| LinearR | 2.316      | 0.001      | 0.006 | 2.340      | 0.012      | 2.000 |
| FGS     | 7.500      | 0.036      | 0.862 | 2.470      | 0.070      | 1.000 |

NOTE: The true parameters are $\beta_1^1 = \beta_1^2 = 0$.

size. For example, under the Linear treatment assignment and FGS outcome model with a sample size of $N = 250$, the proposed method shows unacceptably high bias when estimating $\beta_2^*$ when compared to dWOLS, see Table S5. We conjecture that this occurs because the information available for estimating the second stage propensity model is limited to the set non-responders (i.e., 50% of the sample) that are rerandomized. The value functions for the estimated rules in all cases were calculated for all sample sizes and show that the proposed method, followed by dWOLS, typically results in value functions that are closest to optimal; see Section 2.4 of the supplementary materials.

The supplementary materials include simulation results in which SuperLearner is replaced by alternative data adaptive techniques. Specifically, in Tables S7 and S8 in the supplementary materials, the columns RF-RF and GAM-GAM represent modeling approaches in which RandomForest and gam are used for both the marginalized outcome (i.e., $\mu_{1Y}(\cdot)$ and $\mu_{2Y}(\cdot)$) and treatment assignment models (i.e., $\mu_{1A}(\cdot)$ and $\mu_{2A}(\cdot)$). The column RF-GAM instead uses randomForest for the outcome model and gam for the treatment assignment model. Comparing these results with those summarized in Tables 1 and 2 shows that the use of SuperLearner improves performance.

5.2. Performance: Non-Regular Setting

The treatment assignment models considered here are, respectively, Randomized, Linear, and InterQuad, defined as in Section 5.1. Additionally, define $\tilde{X}_2 = (X_{21}, X_{22}, X_{23}, X_{24}, X_{25})$ where $X_{21}$ is generated from a Bernoulli distribution with success probability $[1 + \exp(-(2X_{11} + 2X_{12} - 1))]^{-1}$, $X_{22}$ is
generated from a Bernoulli distribution with success probability $[1 + \exp(-(2X_{12} + X_{21} - 1))]^{-1}$, $X_{23} = U_1$, $X_{24} = 0.35X_{15} + U_2$, and $X_{25} = U_3$, where $U_1, U_2, \ldots, U_5$ are independent and uniformly distributed on [-0.5,0.5]. We consider the following outcome models:

- **Linear** model: $Y = X_1^\top \alpha_1 + X_2^\top \alpha_2 + A_1X_1^\top \theta_1 + A_2(\theta_2R_{X_{21}}) + \epsilon$ where $\alpha_1 = \alpha_2 = (1, 0.1, 0.1, 0.1, 0.1)^\top$, $\theta_1 = (0, 0, 0, 0, 0)^\top$ and $\theta_2 = 2\sigma$.
  
- **Nonlinear** model: $Y = f(X_1) + A_1X_1^\top \theta_1 + A_2 \cdot (\theta_2R_{X_{21}}) + \epsilon$ where $\theta_1 = (0, 0, 0, 0, 0)^\top$, $\theta_2 = 2\sigma$, and for $x = (x_1, x_2, x_3, x_4, x_5)^\top$, we set
  
  $$f(x) = -1.5 + \sin(\pi x_1x_2) + 2(x_3 - 0.5)^2 + x_4 + 1.5 \frac{x_1}{|x_2| + |x_3|} + 2x_1(x_2 + x_3).$$

The noise variable $\epsilon$ is generated from N(0, $\sigma = 0.5$) and the constant $\sigma \in \{0, 1\}$ specifies the degree of non-regularity, as will be discussed further below. In the above models, $S^2 = (X_1^\top, A_1, X_2^\top)^\top$, $W^2 = X_1, A_2(S^2) = \theta_2R_{X_{21}}, \theta_2(S_0)$ is determined by the remaining model terms, and $R$ is a function of $X_{24}$ only. The first and second stage Q-functions are, respectively, modeled as linear functions of $S = R(1, \tilde{X}_{21}, \tilde{X}_{22}, \tilde{X}_{23})^\top$ and $W = (1, X_{11}, X_{13})^\top$. For both models, it is not difficult to show that $\beta_{21}^T = (0, 2\sigma, 0, 0, 0)^\top$ and that $\beta_{11}^T = 0$.

In both scenarios, for each subject $i$, the first-stage pseudo outcome is defined as in (14) and estimated by substituting in $\tilde{\beta}_{21}$ for $\beta_{11}^T$. The construction of the pseudo-outcome, specifically the projection $S^\top \tilde{\beta}_{21}$, violates Assumption 6. In particular, $\sigma = 0$ corresponds to no second-stage effect modifier, implying that $P(S^\top \tilde{\beta}_{21} = 0) = 1$ because $\beta_{21}^T = 0$. Setting $\sigma = 1$ instead implies that there is no second-stage treatment effect when $R\tilde{X}_{21} = 0$, and a reasonably strong effect when $R\tilde{X}_{21} = 1$; in this case, $0 < P(S^\top \tilde{\beta}_{21} = 0) < 1$. However, the conditions of Corollary 1 hold in each case because $S$ does not include $A_1$, resulting in regular asymptotic behavior for the proposed method.

Because these simulations focus on coverage rather than bias and standard error, we simulate 1000 datasets of size $N = 2000$. In the nonregular setting considered here, neither $Q_{m,N}$ nor dWLOS$_{m,N}$ can necessarily be expected to perform well; hence, we compare our proposed method to a modified version of standard Q-learning and doubly robust weighted least squares in which the first stage confidence intervals are, respectively, constructed using a $m$-out-of-N bootstrap technique as developed in Chakraborty and Moodie (2013) (i.e., $Q_{m,N}$) and Siminou et al. (2018) (i.e., dWLOS$_{m,N}$). In both of these approaches, the tuning parameter $\kappa \in \{0, 1\}$ determines the bootstrap sample size $m$; here, $\kappa = 0.05$. Table 3 summarizes the results; for comparison, results obtained using the $N$-out-of-$N$ bootstrap in the first stage are provided in Table S9 in the supplementary materials. In these tables, empirical coverages that are significantly over or under the nominal level 0.95 are indicated with a dagger, with significance being assessed using a binomial test.

The performance of both $Q_{m,N}$ and $Q_{m,N}^{\kappa=0.05}$ relies heavily on the correct specification of the outcome model. In those case where both methods are observed to exhibit reasonable performance, Tables 3 and S9, respectively, show that $Q_{m,N}^{\kappa=0.05}$ typically over-covers whereas $Q_{m,N}$ either under-covers or has close to nominal coverage; in contrast, when $Q_{m,N}$ is observed to under-cover to a very significant extent, so does $Q_{m,N}^{\kappa=0.05}$.

In general, both dWLOS$_{m,N}^{\kappa=0.05}$ and the proposed method lead to significant improvements in performance. Indeed, the proposed method produces valid confidence intervals with coverages close to the nominal level throughout Tables 3 and S9. This can be readily explained by the fact that each setting satisfies the assumptions of Corollary 1 despite violating Assumption 6. Similarly, we see that dWLOS$_{m,N}^{\kappa=0.05}$ performs reasonably well regardless of the outcome model for both the Randomized and Linear treatment assignment models, since in these two cases the latter can be consistently estimated at a parametric rate. However, compared to the

| Models | $\beta_{11}^T$ | $\beta_{21}^T$ |
|--------|----------------|----------------|
|        | $Q_{m,N}^{\kappa=0.05}$ | $Q_{m,N}^{\kappa=0.05}$ |
| Linear | $0.976(0.31)^\dagger$ | $0.956(0.42)$ | $0.969(0.48)^\dagger$ | $0.976(0.31)^\dagger$ | $0.959(0.40)$ | $0.964(0.48)$ |
| Randomized treatment assignment model |
| $0.988(1.14)^\dagger$ | $0.965(1.40)^\dagger$ | $0.964(1.63)$ | $0.981(0.53)^\dagger$ | $0.952(0.56)^\dagger$ | $0.975(0.65)^\dagger$ |
| Linear | $0.965(0.51)^\dagger$ | $0.962(0.46)$ | $0.980(0.82)^\dagger$ | $0.960(0.51)^\dagger$ | $0.966(0.45)^\dagger$ | $0.964(0.83)$ |
| Linear | $0.984(1.13)^\dagger$ | $0.966(1.41)^\dagger$ | $0.960(1.71)$ | $0.963(0.59)$ | $0.946(0.60)$ | $0.964(0.82)$ |
| $Q_{m,N}^{\kappa=0.05}$ | $Q_{m,N}^{\kappa=0.05}$ |
| Linear treatment assignment model |
| $0.961(0.32)$ | $0.965(0.45)^\dagger$ | $0.971(0.53)^\dagger$ | $0.954(0.32)$ | $0.948(0.44)$ | $0.968(0.53)^\dagger$ |
| Linear | $0.521(1.15)^\dagger$ | $0.949(1.58)$ | $0.955(1.83)$ | $0.190(0.58)^\dagger$ | $0.955(0.65)^\dagger$ | $0.975(0.78)^\dagger$ |
| Linear | $0.907(0.51)^\dagger$ | $0.953(0.51)$ | $0.968(0.89)^\dagger$ | $0.900(0.51)^\dagger$ | $0.955(0.50)^\dagger$ | $0.965(0.89)^\dagger$ |
| Linear | $0.450(1.14)^\dagger$ | $0.952(1.61)$ | $0.957(1.93)$ | $0.163(0.63)^\dagger$ | $0.948(0.70)$ | $0.974(0.93)^\dagger$ |
| $Q_{m,N}^{\kappa=0.05}$ | $Q_{m,N}^{\kappa=0.05}$ |
| InterQuadrature treatment assignment model |
| $0.982(0.34)^\dagger$ | $0.966(0.46)^\dagger$ | $0.982(0.54)^\dagger$ | $0.975(0.34)^\dagger$ | $0.957(0.45)^\dagger$ | $0.969(0.53)^\dagger$ |
| Linear | $0.918(1.21)^\dagger$ | $0.959(1.45)$ | $0.846(1.70)^\dagger$ | $0.950(0.61)^\dagger$ | $0.962(0.63)^\dagger$ | $0.866(0.74)^\dagger$ |
| Linear | $0.967(0.56)^\dagger$ | $0.964(0.52)$ | $0.972(0.91)^\dagger$ | $0.959(0.56)$ | $0.950(0.51)^\dagger$ | $0.964(0.91)$ |
| Linear | $0.912(1.20)^\dagger$ | $0.964(1.47)$ | $0.871(1.79)^\dagger$ | $0.943(0.66)$ | $0.967(0.68)^\dagger$ | $0.910(0.92)^\dagger$ |

Numbers in parentheses correspond to average confidence interval length.
proposed method, the coverages tend to be slightly conservative, with longer confidence intervals. In these same cases, dWLOS\(_{N,N}\) also performs reasonably, though does have a tendency to under-cover. Under the InterQuad treatment assignment model, the performance of both dWLOS\(_{m,N}^{x=0.05}\) and dWLOS\(_{N,N}\) declines due to misspecification of the treatment assignment model, and in the Nonlinear\(^{\text{NR,m}}\) setting, also the outcome model. For example, dWOLS\(_{m,N}^{x=0.05}\) exhibits coverage rates as low as 87%. We conjecture that the combination of non-regularity, model misspecification and residual confounding are the main reasons for the poor performance of Q\(_{N,N}\), Q\(_{m,N}^{x=0.05}\) and, where observed to be poor, both dWLOS\(_{m,N}^{x=0.05}\) and dWLOS\(_{N,N}\). In comparing the two approaches to bootstrapping for both standard Q-learning and dWOLS, our results further suggest that tuning \(m\) differently (i.e., increasing \(m\)) may result in better agreement with the nominal coverage level in cases where the relevant models are appropriately specified.

Finally, we conducted a related simulation study in which both Assumption 6 and the conditions of Corollary 1 are violated. Unlike the simulation settings above, this example considers a case where the first and second stage treatments interact with each other. This modified study is described in Section 2.3 of the supplementary materials, where we compare the proposed approach with dWLOS\(_{m,N}^{x=0.05}\); the results are summarized in Table S10. Overall, the methods perform as expected. In particular, the proposed method demonstrates either nominal or modest undercoverage for the first stage regression parameters and dWLOS\(_{m,N}^{x=0.05}\) demonstrates conservative coverage except in cases where the required conditions for consistency are violated.

6. Application

We use the data from the Extending Treatment Effectiveness of Naltrexone (ExTENd) clinical trial to illustrate our method. Naltrexone (NTX) is an opioid receptor antagonist used in the prevention of relapse to alcoholism. Even though NTX has been shown to be efficacious in those that adhere to treatment, its use by clinicians has been limited, at least in some cases, because adherence rates are often negatively impacted by the fact that NTX diminishes the pleasurable effects of alcohol use.

In the ExTENd study (Figure 1), at the first decision stage, patients were randomized to one of two definitions of non-response while receiving NTX: (1) Stringent: a patient is a non-responder if (s)he has two or more heavy drinking days in the first 8 weeks \((A_1 = 0)\); (2) Lenient: a patient is a non-responder if (s)he has five or more heavy drinking days in the first 8 weeks \((A_1 = 1)\). At the second decision stage, the treatment assignment mechanism depends on response status. Specifically, define \(A_2 = 1\) if the current treatment (NTX) is augmented, and zero otherwise; in addition, we let \(\bar{R}\) denote the indicator of response to treatment. Then, among responders \((\bar{R} = 1)\), patients are randomized (with equal probability) to augment NTX with telephone disease management (NTX+TDM; \(A_2 = 1\)) or to maintain NTX alone \((A_2 = 0)\). For nonresponders \((\bar{R} = 0)\), patients are instead randomized (with equal probability) to augment NTX with combined behavioral intervention (NTX+CBI; \(A_2 = 1\)) or to CBI alone \((A_2 = 0)\). In the latter case, maintenance on NTX alone is replaced with an alternative treatment due to nonresponse. The primary outcome is the proportion of abstinence days over 24 weeks. The list of baseline and time varying variables that are used in our analyses are given in Table 4. There are multiple measurements of time-varying

![Figure 1. ExTENd study design. The * notation represents instances of randomization; the N values in this figure represent the subsequent number of patients assigned to each treatment option.](image-url)
to earlier notation, the first stage covariate vector $W^0$ consists of the predictors gender, race, alcyr0, intox0, and ocds0, and the second stage covariate vector $S^0$ consists of all the predictors listed in Table 4, along with response status. First stage regression models are fit using $W = W^0$. The description of the second stage model predictor $S$ is more involved. Specifically, let $Z = (Z_1, \ldots, Z_6)^T$ contain the variables gender, A1, intox0, ocds0, pacs1, and mcs1; then, we define $S = (R, (1 - R)Z_1, (1 - R)Z_2, (1 - R)Z_3)^T$. As specified, the second stage model allows the set of possible effect modifiers to differ between responders and non-responders, with some overlap in the case of gender and A1. We used SuperLearner to estimate $\mu_1(Y(W^0))$ and $\mu_2(Y(S^0))$, employing the same library as we did in the simulation study and, respectively, using $W^0$ and $S^0$ for general confounding control. In view of the fact that the randomization mechanism is known, and mostly successful in view of the overall degree of balance observed in Figure 2, the treatment propensities $\mu_{1A}(w)$ and $\mu_{2A}(s)$ are estimated using logistic regression models. Specifically, the former is estimated as a function of gender, and the latter is estimated using gender, response status, and the interaction between gender and response status. The parameters of the Q-functions used by dWOLS are assumed to follow linear models (i.e., including the main effects). Similarly, for standard Q-learning, linear working models, respectively, replace $\eta_2(S^0)$ and $\eta_1(W^0)$.

As shown in Table 5, the signs of all predictor effects are the same for all methods, though magnitudes and confidence intervals differ. None of the effect modifiers in the second stage are deemed statistically significant among responders using any of the 3 methods. The proposed Q-learning method suggests that both ocds0 and mcs1 are significant effect modifiers of A2 among nonresponders; specifically, individuals with higher ocds0 and mcs1 would benefit from CBI. Similarly, dWOLS identifies mcs1 as a significant effect modifier among nonresponders, whereas none of the effect modifiers are identified as significant using standard Q-learning. For the first stage model, the proposed Q-learning method shows that the years of drinking (i.e., alcyr0) and gender significantly modify the effect of A1. In particular, female individuals and those with more years of drinking would benefit from being treated under a stringent definition of non-response. This makes sense because, for example, individuals with more years of drinking at baseline likely have a higher craving for alcohol and require more immediate attention and
rescue treatments (i.e., A2 for non-responders). In contrast, neither standard Q-learning nor dWOLS detects any effect modifiers. With the exception of the interaction between A1 and gender, the first stage point estimates are rather similar across the three methods, highlighting the fact that the differences in significance stem from the tighter confidence intervals obtained using the proposed methods (i.e., compared to those produced using the m-out-of-N bootstrap).

Indeed, for both stages, the dWOLS and standard Q-learning methods yield point estimates that are mostly similar to each other. These similarities are expected for two reasons. First, under successful randomization, we would not generally expect misspecification of the functional form of the main effects in the Q-function to bias the estimate of interaction terms (i.e., β^2). Second, the linear models being used in the Q-models are identical in both cases; the only difference is that parameter estimation is carried out using weighted versus unweighted least squares. Figure 2 demonstrates the presence of random confounding among non-responders in the second stage for gender, ethnicity, A1, pacs0 and pacs1, whereas there is good balance among the first stage predictors. Comparing dWOLS and standard Q-learning, we see that the largest differences in point estimates occur in the second stage model among non-responders for A1 and pacs1.

Comparing the proposed method to both dWOLS and standard Q-learning, we observe somewhat greater disparity in point estimates. These differences occur primarily among non-responders in the second stage, and include interactions between A2 and each of gender, A1 and pacs1; as noted above, gender, A1 and pacs1 all demonstrate substantial imbalance among non-responders in the second stage. The largest difference among regression coefficients in the first stage model occurs for gender, consistent with the disparities observed in the second stage model as well as propagation of those differences through the backward induction process. We conjecture that modeling the true main effects (i.e., μ_j(j), j = 1, 2) using Super Learner may help to reduce small sample biases when compared to the more restrictive linear models used by both dWOLS and standard Q-learning.

### Table 5. ExTEND data.

| Q-function | Proposed | dWOLS^x=0.05 | Q^x=0.05 |
|------------|----------|--------------|-----------|
| Models     | Est 95% CI | Est 95% CI   | Est 95% CI |
| Stage 1    |          |              |            |
| Responders |          |              |            |
| A1         | -0.06    | (-0.18, 0.06)| -0.05      | (-0.21, 0.08)| -0.06 | (-0.20, 0.09) |
| A1 : gender| -0.24*   | (-0.46, -0.01)| -0.18      | (-0.43, 0.14)| -0.17 | (-0.37, 0.14) |
| A1 : race  | 0.08     | (-0.05, 0.21)| 0.08       | (-0.07, 0.21)| 0.08  | (-0.08, 0.24)|
| A1 : alcyr | -0.07*   | (-0.14, 0.00)| -0.06      | (-0.15, 0.03)| -0.06 | (-0.16, 0.03) |
| A1 : intox | 0.13     | (-0.03, 0.29)| 0.14       | (-0.06, 0.33)| 0.15  | (-0.10, 0.33)|
| A1 : ocdl  | 0.02     | (-0.05, 0.09)| 0.01       | (-0.07, 0.09)| 0.01  | (-0.08, 0.08)|

| Stage 2    |          |              |            |
| Responders |          |              |            |
| A2         | 0.02     | (-0.07, 0.11)| 0.01       | (-0.07, 0.09)| 0.01  | (-0.07, 0.09)|
| A2 : gender| 0.07     | (-0.11, 0.24)| 0.11       | (-0.08, 0.29)| 0.11  | (-0.08, 0.28)|
| A2 : A1    | 0.01     | (-0.11, 0.13)| 0.01       | (-0.11, 0.12)| 0.01  | (-0.11, 0.12)|
| Non-responders |          |              |            |
| A2         | -0.07    | (-0.23, 0.10)| -0.02      | (-0.28, 0.22)| -0.02 | (-0.28, 0.23) |
| A2 : gender| 0.27     | (-0.18, 0.71)| 0.13       | (-0.43, 0.81)| 0.14  | (-0.43, 0.79) |
| A2 : intox | 0.09     | (-0.16, 0.35)| 0.06       | (-0.21, 0.46)| 0.06  | (-0.22, 0.49) |
| A2 : ocdl  | -0.19*   | (-0.34, -0.03)| -0.19      | (-0.43, 0.08)| -0.18 | (-0.43, 0.11) |
| A2 : A1    | -0.21    | (-0.45, 0.01)| -0.23      | (-0.57, 0.26)| -0.16 | (-0.55, 0.30) |
| A2 : pacs1 | 0.10     | (-0.02, 0.21)| 0.07       | (-0.12, 0.25)| 0.03  | (-0.18, 0.23) |
| A2 : mcs1  | -0.18*   | (-0.29, -0.06)| -0.17*     | (-0.37, 0.00)| -0.17 | (-0.39, 0.01) |

**Note:** Significance coefficients at a Type I error rate of 5%.

### 7. Discussion

Much of the current work on Q-learning continues to involve parametric working models despite the fact that finite-dimensional models are generally too restrictive to permit consistent estimation of nuisance parameters. We proposed a robust Q-learning approach where the working models need not all be linear and, specifically, where the main effects that do not influence the optimal decision rules are estimated using data-adaptive approaches. Our simulation studies highlight the value of our proposed approach compared with existing Q-learning methods. The proposed method also performed relatively well in simulations when key regularity assumption (i.e., Assumption 6) is violated; however, we cannot expect this in all scenarios, as the underlying theory and simulation results show otherwise.

An important advantage of the proposed method is that it does not suffer from the curse of dimensionality, as it produces root-n consistent estimators even when \( \mu_{jA}(\cdot) \) and \( \hat{\mu}_{jY}(\cdot) \) (\( j = 1, 2 \)) are estimated at rates slower than root-n. This important property facilitates the use of nonparametric methods like Super Learner for estimating these unknown functions, substantially reducing the chance of model misspecification. A second important feature of the proposed approach is that consistent estimation of the treatment models leads to consistent estimation of the blip function parameters, whether or not these models or those for \( \mu_{jY}(\cdot) \), \( j = 1, 2 \) are correctly specified. However, the proposed estimators are not doubly robust, in that we require that the \( \mu_{jA}(\cdot) \)'s are consistently estimated at a sufficiently fast
rate. True double robustness under (1) for \( \beta^\ast \) requires that one either consistently estimates \( \mu_{2A}(S^0) \) or the treatment-free conditional mean model \( E[Y - A_2S^0 \beta^\ast_{12}(S^0)] = \eta_2(S^0) \); similarly, for \( \beta^\ast_1 \) one must either consistently estimate \( \mu_{1A}(W^0) \) or \( E[Y - A_1W^0 \beta^\ast_{11}(W^0)] = \eta_1(W^0) \). Because the expectation operator is linear, correct specification of both treatment-free models essentially relies on both \( \mu_{1Y}(\cdot) \) and \( \mu_{1A}(\cdot), j = 1, 2 \) being correctly specified. This limitation on the practicality of finding a truly doubly robust estimator applies to the proposed approach as well as that taken in Wallace and Moodie (2015). Further research on doubly robust estimation in this class of problems is merited.

Although data-adaptive estimation methods reduce the risk of inconsistency, there is still a chance that one or more nuisance parameters will be estimated inconsistently. Further research is needed to study the behavior of the proposed methods under inconsistent estimation of a nuisance parameter. In particular, Benkeser et al. (2017) showed that when nuisance parameters are estimated using data-adaptive approaches, inconsistently estimating one nuisance parameter may lead to an irregular estimator having a convergence rate slower than root-\( n \). These authors proposed a targeted minimum loss-based approach to resolve the issue (van der Laan 2014). Generalization of the method of Benkeser et al. (2017) to a multistage decision making process would be an interesting topic for future research. Studying the asymptotic behavior of an appropriate version of the bootstrap in our proposed Q-learning method is also of interest as it can potentially resolve the non-regularity issues in settings where both Assumption 6 fails and Corollary 1 fail (e.g., Chakraborty, Laber, and Zhao 2013). Finally, in practice, there are often many candidate variables to be considered when constructing a decision rule. The inclusion of spurious variables in these analyses can substantially reduce the quality of the estimated decision rules. Although one can adapt the proposed methods to obtain regularized estimators of the target parameters, valid post-selection inference remains a challenge and merits further research (Berk et al. 2013; Fithian, Sun, and Taylor 2014).

**Supplementary Materials**

Supplementary material includes proofs of presented results as well as additional numerical experiments.

**References**

Austin, P. C. (2009), “Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research,” *Communications in Statistics—Simulation and Computation*, 38, 1228–1234. [378]

Bai, X., Tsatis, A. A., and O’Brien, S. M. (2013), “Doubly-Robust Estimators of Treatment-Specific Survival Distributions in Observational Studies With Stratified Sampling,” *Biometrics*, 69, 830–839. [368]

Benkeser, D., Carone, M., van der Laan, M., and Gilbert, P. (2017), “Doubly Robust Nonparametric Inference on the Average Treatment Effect,” *Biometrika*, 104, 863–880. [369,380]

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), “Valid Post-Selection Inference,” *Annals of Statistics*, 41, 802–837. [380]

Butler, E. L., Laber, E. B., Davis, S. M., and Kosorok, M. R. (2018), “Incorporating Patient Preferences Into Estimation of Optimal Individualized Treatment Rules,” *Biometrics*, 74, 18–26. [368]

Cao, W., Tsiatis, A. A., and Davidian, M. (2009), “Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean With Incomplete Data,” *Biometrika*, 96, 723–734. [369]

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013), “Inference for Optimal Dynamic Treatment Regimes Using an Adaptive m-Out-of-n Bootstrap Scheme,” *Biometrics*, 69, 714–723. [380]

Chakraborty, B., and Moodie, E. (2013), *Statistical Methods for Dynamic Treatment Regimes*, New York: Springer. [368,371,376]

Chernozhukov, V., Chetverikov, D., Demirer, M., Dufo, E., Hansen, C., Newey, W., and Robins, J. (2018), “Double/Debiased Machine Learning for Treatment and Structural Parameters,” *The Econometrics Journal*, 21, C1–C68. [371,372,373]

Davidian, M., Tsiatis, A., and Laber, E. (2016), “Dynamic Treatment Regimes” (Chapter 13), in *Cancer Clinical Trials: Current and Controversial Issues in Design and Analysis*, eds. S. George, X. Wang, and H. Pang, Boca Raton, FL: CRC Press, pp. 409–446. [368]

Dudoit, S., and van der Laan, M. J. (2003), “Asymptotics of Cross-Validated Risk Estimation in Model Selection and Performance Assessment,” Tech. Rep., Working Paper 126, Division of Biostatistics, University of California at Berkeley. [371]

Ertefaie, A., Shortreed, S., and Chakraborty, B. (2016), “Q-Learning Residual Analysis: Application to the Effectiveness of Sequences of Antipsychotic Medications for Patients With Schizophrenia,” *Statistics in Medicine*, 35, 2221–2234. [368]

Fithian, W., Sun, D., and Taylor, J. (2014), “Optimal Inference After Model Selection,” arXiv no. 1410.2597. [380]

Hastie, T. (2019), “gam: Generalized Additive Models,” R Package Version 1.16.1. [373]

Kang, J. D., and Schafer, J. L. (2007), “Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean From Incomplete Data,” *Statistical Science*, 22, 523–539. [369]

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014), “Dynamic Treatment Regimes: Technical Challenges and Applications,” *Electronic Journal of Statistics*, 8, 1225. [368,369,371]

Lavori, P. W., and Dawson, R. (2000), “A Design for Testing Clinical Strategies: Biased Adaptive Within-Subject Randomization,” *Journal of the Royal Statistical Society*, Series A, 163, 29–38. [368]

Lee, I., Nahum-Shani, I., Lynch, K., Oslin, D., and Murphy, S. (2012), “A ‘SMART’ Design for Building Individualized Treatment Sequences,” *Annual Review of Clinical Psychology*, 8, 21–48. [368]

Liaw, A., and Wiener, M. (2002), “Classification and Regression by Random Forest,” *R News*, 2, 18–22. [373]

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019), “e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien,” R Package Version 1.7-2. [373]

Milborrow, S. (2019), “earth: Multivariate Adaptive Regression Splines,” R Package Version 5.1.1. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller’s Fortran utilities with Thomas Lumley’s leaps wrapper. [373]

Moodie, E. E. M., and Kosorok, M. R., eds. (2015), *Adaptive Treatment Strategies in Practice*, ASA-SIAM Series on Statistics and Applied Mathematics, Philadelphia, PA: Society for Industrial and Applied Mathematics. [368]

Murphy, S. A. (2003), “Optimal Dynamic Treatment Regimes,” *Journal of the Royal Statistical Society*, Series B, 65, 331–355. [368]

(2005), “An Experimental Design for the Development of Adaptive Treatment Strategies,” *Statistics in Medicine*, 24, 1455–1481. [368]

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., Waxmonsky, J. G., Yu, J., and Murphy, S. A. (2012a), “Experimental Design and Primary Data Analysis Methods for Comparing Adaptive Interventions,” *Psychological Methods*, 17, 457. [368]

(2012b), “Q-Learning: A Data Analysis Method for Constructing Adaptive Interventions,” *Psychological Methods*, 17, 478. [368]

Polley, E., LeDell, E., Kennedy, C., and van der Laan, M. (2019), “SuperLearner: Super Learner Prediction,” R Package Version 2.0-25. [373]

Robins, J. M., Mark, S. D., and Newey, W. K. (1992), “Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders,” *Biometrics*, 48, 479–495. [371]
Robinson, P. M. (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954. [369,370]

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998), "Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339. [369]

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014), "Q- and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Statistical Science*, 29, 640–661. [368]

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998), "Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339. [369]

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014), "Q- and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Statistical Science*, 29, 640–661. [368]

Shi, C., Fan, A., Song, R., and Lu, W. (2018), "High-Dimensional A-Learning for Optimal Dynamic Treatment Regimes," *Annals of Statistics*, 46, 925–957. [368]

Simoneau, G., Moodie, E. E., Platt, R. W., and Chakraborty, B. (2018), "Non-Regular Inference for Dynamic Weighted Ordinary Least Squares: Understanding the Impact of Solid Food Intake in Infancy on Childhood Weight," *Biostatistics*, 19, 233–246. [376]

Song, R., Kosorok, M., Zeng, D., Zhao, Y., Laber, E., and Yuan, M. (2015), "On Sparse Representation for Optimal Individualized Treatment Selection With Penalized Outcome Weighted Learning," *Stat*, 4, 59–68. [368]

Tsiatis, A. (2007), *Semiparametric Theory and Missing Data*, Springer Series in Statistics, New York: Springer. [369]

van der Laan, M. J. (2014), "Targeted Estimation of Nuisance Parameters to Obtain Valid Statistical Inference," *The International Journal of Biostatistics*, 10, 29–57. [380]

van der Laan, M. J., and Dudoit, S. (2003), "Unified Cross-Validation Methodology for Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples," Tech. Rep. Working Paper 130, Division of Biostatistics, University of California at Berkeley. [371]

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007), "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, 6, 25. [370,371]

van der Laan, M. J., and Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, Springer Series in Statistics, New York: Springer. [369]

van der Vaart, A. W., Dudoit, S., and van der Laan, M. J. (2006), "Oracle Inequalities for Multi-Fold Cross Validation," *Statistics and Decisions*, 24, 351–371. [371]

Vermeulen, K., and Vansteelandt, S. (2015), "Bias-Reduced Doubly Robust Estimation," *Journal of the American Statistical Association*, 110, 1024–1036. [369]

Vermeulen, K., and Vansteelandt, S. (2015), "Bias-Reduced Doubly Robust Estimation," *The International Journal of Biostatistics*, 12, 253–282. [369]

Wallace, M. P., and Moodie, E. E. (2015), "Doubly-Robust Dynamic Treatment Regimen Estimation via Weighted Least Squares," *Biometrics*, 71, 636–644. [369,373,374,380]

Watkins, C. J., and Dayan, P. (1992), "Q-Learning," *Machine Learning*, 8, 279–292. [368]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012), "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018. [368]

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009), "Reinforcement Learning Design for Cancer Clinical Trials," *Statistics in Medicine*, 28, 3294–3315. [368]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [368]

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011), "Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer," *Biometrics*, 67, 1422–1433. [368]

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015), "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 110, 583–598. [368]