Published in final edited form as:
Stat Methods Med Res. 2010 June ; 19(3): 317–343. doi:10.1177/0962280209105013.

Inference for Nonregular Parameters in Optimal Dynamic Treatment Regimes

Bibhas Chakraborty[1,*], **Victor Strecher**[2], and **Susan Murphy**[1]

[1] Department of Statistics, University of Michigan

[2] Center for Health Communications Research, University of Michigan

Abstract

A dynamic treatment regime is a set of decision rules, one per stage, each taking a patient's treatment and covariate history as input, and outputting a recommended treatment. In the estimation of the optimal dynamic treatment regime from longitudinal data, the treatment effect parameters at any stage prior to the last can be nonregular under certain distributions of the data. This results in biased estimates and invalid confidence intervals for the treatment effect parameters. In this paper, we discuss both the problem of nonregularity, and available estimation methods. We provide an extensive simulation study to compare the estimators in terms of their ability to lead to valid confidence intervals under a variety of nonregular scenarios. Analysis of a data set from a smoking cessation trial is provided as an illustration.

Keywords

dynamic treatment regime; nonregularity; bias; hard-threshold; soft-threshold; empirical Bayes; bootstrap

1 Introduction

Many diseases such as mental illness, HIV infection, and substance abuse are clinically treated in multiple stages, adapting the treatment type and dosage to the ongoing measures of an individual patient's response, adherence, burden, side effects, and preference. Dynamic treatment regimes represent one way to operationalize this sequential decision making. A dynamic treatment regime (DTR) is a sequence of decision rules, one per stage. Each decision rule takes a patient's treatment and covariate history as input, and outputs a recommended treatment. The main motivations for considering sequences of treatments are high variability across patients in response to any one type of treatment, likely relapse, presence or emergence of co-morbidities, time-varying side effect severity, and reduction of costs and burden when intensive treatment is unnecessary[1].

A DTR is said to be optimal if it optimizes the mean outcome at the end of the final stage of treatment. Data for estimating the optimal DTR can come from either an observational longitudinal study or a sequential multiple assignment randomized trial (SMART)[2–5]. In these designs, each patient is followed through stages of treatment and at each stage the patient is randomized to one of the possible treatment options. Experimental designs similar to SMART have been implemented in the treatments of schizophrenia6, depression7, and cancer[8,9].

[*]Address for correspondence: Bibhas Chakraborty, Department of Statistics, 439 West Hall, 1085 South University Avenue, Ann Arbor, MI 48109-1107, USA. E-mail: bibhas@umich.edu.

Estimating the optimal DTR is a problem of sequential, multi-stage decision making. Murphy[10] developed a semiparametric method for estimating the optimal DTR, an efficient version of which was provided by Robins[11]. A nice discussion about the relationship between these two methods can be found in Moodie et al.[12]. Other methods for estimating optimal DTRs in the literature include likelihood-based methods, both frequentist and Bayesian, developed by Thall and colleagues[8,13,14], and the semiparametric methods of Lunceford et al.[15], and Wahed and Tsiatis[9,16].

Robins[11] considered the problem of inference for the parameters of the optimal DTR. As discussed by Robins, the treatment effect parameters at any stage prior to the last can be *nonregular* under certain longitudinal distributions of the data which he called *exceptional laws*. By nonregularity, we mean that the asymptotic distribution of the estimator of the treatment effect parameter does not converge uniformly over the parameter space (see section 2.4 for further details). This technical phenomenon of nonregularity has considerable practical consequences; it often causes bias in estimation, and leads to poor frequentist properties of the confidence intervals. Recently Moodie and Richardson[17] provided a method called *Zeroing Instead of Plugging In* (ZIPI) for correcting the bias in the estimation of the optimal DTRs resulting under exceptional laws.

The main goals of this paper are to illustrate the problem of nonregularity, and to compare available estimation methods that attempt to address this problem. In section 2, we discuss the problem of nonregularity in detail. Section 3 provides a description of different methods that address the problem. We provide an extensive simulation study in section 4 to compare the estimators in terms of their ability to lead to valid confidence intervals using bootstrap. This is followed by an analysis of a data set from a longitudinal smoking cessation trial in section 5; the purpose is to demonstrate the applicability of the estimation methods in a real-life nonregular scenario. Finally an overall discussion is provided in section 6. Throughout this article, we assume that the data come from SMART designs. The main reason for this is to separate the issue of nonregularity from causal inference issues. However the problem of nonregularity also arises when observational data[11,17] are used; and the estimators proposed in section 3 should be applicable to observational data as well.

2 Estimation and Inference via Q-learning

2.1 Notation and Data Structure

For simplicity, we focus on studies with two stages. Longitudinal data on a single patient are given by the trajectory $(O_1, A_1, O_2, A_2, O_3)$, where O_j $(j = 1, 2)$ denotes the covariates measured prior to treatment at the beginning of the j-th stage, O_3 is the observation at the end of stage 2, and A_j $(j = 1, 2)$ is the treatment assigned at the j-th stage subsequent to observing O_j. The data set consists of a random sample of n patients. Define the history at each stage as: $H_1 = O_1$, $H_2 = (O_1, A_1, O_2)$. We consider a SMART design in which there are two possible treatments at each stage, $A_j \in \{-1, 1\}$; here we assume $P[A_j = -1|H_j] = P[A_j = 1|H_j] = \frac{1}{2}$. The study can have either a single primary outcome Y observed at the end of stage 2, or two outcomes Y_1, Y_2 observed at the two stages. Note that the case of a single outcome Y observed at the end can be viewed as a case with $Y_1 \equiv 0$ and $Y_2 = Y$. We assume $Y_1 = f_1(O_1, A_1, O_2)$ and $Y_2 = f_2(O_1, A_1, O_2, A_2, O_3)$, with known functions f_1, f_2. A two-stage DTR consists of two decision rules, say (d_1, d_2), with $d_j(H_j) \in \mathcal{A}_j$, where $\mathcal{A}_j$ is the set of possible treatments at the j-th stage.

One simple method to construct (d_1, d_2) is Q-learning[18–20]. Q-learning, like Robins' *g-estimation of optimal structural nested mean models* (hereafter simply referred to as *Robins' method*), suffers from nonregularity – the common reason being an underlying non-smooth maximization operation. Here we will illustrate the problem due to nonregularity using Q-learning, since it can be viewed as a generalization of the least squares regression to multistage

decision problems, and hence simpler to explain than Robins' semiparametric efficient method. In Lemma 1 below, we provide conditions under which Q-learning is equivalent to an inefficient version of Robins' method.

2.2 Q-learning with Linear Models

First let us define the Q-functions[19,20] for the two stages as follows:

$$Q_2(H_2, A_2)=E\,[\,Y_2|H_2, A_2\,]\,,$$
$$Q_1(H_1, A_1)=E[\,Y_1+\max_{a_2} Q_2(H_2, a_2)|H_1, A_1].$$

If the two Q-functions were known, the optimal DTR (d_1, d_2), using backwards induction (as in dynamic programming) argument, would be

$$d_j(h_j)=\arg\max_{a_j} Q_j(h_j, a_j),\quad j=1, 2. \tag{1}$$

In practice, the true Q-functions are not known and hence must be estimated from the data. Consider a linear model for the Q-functions. Let the stage-j (j = 1, 2) Q-function be modeled as

$$Q_j(H_j, A_j;\beta_j, \psi_j)=\beta_j^T H_{j0}+(\psi_j^T H_{j1})A_j, \tag{2}$$

where H_{j0} and H_{j1} are two (possibly different) summaries of the history H_j, with H_{j0} denoting the "main effect of history" and H_{j1} denoting the part of history that interacts with treatment (H_{j0} and H_{j1} include the intercept term). The Q-learning algorithm is:

1. Stage-2 regression: $(\widehat{\beta}_2, \widehat{\psi}_2)=\arg\min_{\beta_2,\psi_2}\frac{1}{n}\sum_{i=1}^{n}(Y_{2i}-Q_2(H_{2i}, A_{2i};\beta_2, \psi_2))^2$.
2. Stage-2 optimal rule: $\hat{d}_2(h_2)=\arg\max_{a_2} Q_2(h_2, a_2;\hat{\beta}_2,\hat{\psi}_2)$.
3. Stage-1 pseudo-outcome: $\hat{Y}_{1i}=Y_{1i}+\max_{a_2} Q_2(H_{2i}, a_2;\hat{\beta}_2,\hat{\psi}_2)$, $i=1,\ldots,n$.
4. Stage-1 regression: $(\widehat{\beta}_1, \widehat{\psi}_1)=\arg\min_{\beta_1,\psi_1}\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_{1i}-Q_1(H_{1i}, A_{1i};\beta_1, \psi_1)\right)^2$.
5. Stage-1 optimal rule: $\hat{d}_1(h_1)=\arg\max_{a_1} Q_1(h_1, a_1;\hat{\beta}_1,\hat{\psi}_1)$.

The estimated optimal DTR using Q-learning is given by $(\hat{d}_1, \hat{d}_2)$.

The following lemma gives a set of sufficient conditions under which Q-learning is equivalent to an inefficient version of Robins' method.

Lemma 1—Consider linear models for the Q-functions as in (2). Assume that:

- **i.** the parameters in Q_1 and Q_2 are distinct;
- **ii.** A_j has zero conditional mean given the history H_j, j = 1, 2; and
- **iii.** the covariates used in the model for Q_1 are nested within the covariates used in the model for Q_2, *i.e.,* $(H_{10}^T, H_{11}^T A_1)\subset H_{20}^T$.

Then Q-learning is algebraically equivalent to an inefficient version of Robins' method.

The proof is given in Appendix A.

2.3 The Inference Problem

With (2) as the model for Q-functions, the optimal DTR is given by

$$d_j(H_j) = \arg\max_{a_j} (\psi_j^T H_{j1}) a_j = sign(\psi_j^T H_{j1}), \quad j = 1, 2, \tag{3}$$

where $sign(x) = 1$ if $x > 0$, and -1 otherwise. Note that the term $\beta_j^T H_{j0}$ on the right side of (2) does not feature in the optimal DTR. Thus for estimating optimal DTRs, the ψ_j's are the parameters of interest, while β_j's are nuisance parameters. We want to perform inference (e.g., construct confidence intervals) on ψ_j's.

Conducting inference on ψ_j's is important due to the following reasons. First, if the confidence intervals (or hypothesis tests) for ψ_j reveal that there is no evidence that some components of the vector ψ_j are different from zero, then the corresponding components of the history vector H_{j1} need not be collected to make decisions using the optimal DTR. This reduces the cost of data collection in a future implementation of the optimal DTR. Thus in the present context, confidence intervals (or hypothesis tests) can be viewed as a tool for doing variable selection. Second, it is important to know when there is insufficient support in the data to recommend one treatment over another, since in such cases treatment can be chosen according to other considerations like cost, familiarity, burden, preference etc. Third, as discussed by Robins[11], confidence intervals for ψ_j can lead to confidence intervals for d_j. In the following, we discuss the problem of nonregularity in inference.

2.4 Nonregularity in Inference

Note that the stage-1 pseudo-outcome (in the Q-learning algorithm) is

$$\widehat{Y}_{1i} = Y_{1i} + \max_{a_2} Q_2(H_{2i}, a_2; \widehat{\beta}_2, \widehat{\psi}_2) = Y_{1i} + \widehat{\beta}_2^T H_{20,i} + |\widehat{\psi}_2^T H_{21,i}|, \; i = 1, \ldots, n, \tag{4}$$

which is a non-smooth (e.g., non-differentiable at $\widehat{\psi}_2^T H_{21,i} = 0$) function of $\hat{\psi}_2$, because of the maximization operation. Since $\hat{\psi}_1$ is a function of $\hat{Y}_{1i}$, $i = 1,\ldots,n$, it is in turn a non-smooth function of $\hat{\psi}_2$. As a consequence, the asymptotic distribution of $\sqrt{n}(\widehat{\psi}_1 - \psi_1)$ does not converge uniformly[11] over the parameter space of $\psi = (\psi_1, \psi_2)$. More specifically, the asymptotic distribution of $\sqrt{n}(\widehat{\psi}_1 - \psi_1)$ is normal if ψ_2 is such that $P[H_2 : \psi_2^T H_{21} = 0] = 0$, but is non-normal if $P[H_2 : \psi_2^T H_{21} = 0] > 0$. This change in the asymptotic distribution happens abruptly. The (vector) parameter ψ_1 is called a *nonregular* parameter and the estimator $\hat{\psi}_1$ is called a *nonregular* estimator; see Bickel et al.[21] for the precise definition of nonregularity. Because of this nonregularity, given the noise level present in small samples, the estimator $\hat{\psi}_1$ oscillates between the two asymptotic distributions across samples. As a result, usual Wald type confidence intervals perform poorly[11,17].

The issue of nonregularity can be better understood with a toy example discussed by Robins[11] (here is a slightly modified version). Consider the problem of estimating $|\mu|$ based on n i.i.d. observations $X_1,\ldots,X_n$ from $N(\mu, 1)$. Note that $|\bar{X}_n|$is the maximum likelihood estimator of $|\mu|$, where $\bar{X}_n$ is the sample average. It can be shown that the asymptotic distribution of $\sqrt{n}(|\bar{X}_n| - |\mu|)$ for $\mu = 0$ is different from that for $\mu \neq 0$. Thus $|\bar{X}_n|$is a nonregular estimator of $|$

$\mu|$. Also, for $\mu = 0$, $\lim_{n\to\infty} E[\sqrt{n}(|\overline{X}_n| - |\mu|)] = \sqrt{\frac{2}{\pi}}$. Robins referred to this quantity as the *asymptotic bias* of the estimator $|\overline{X}_n|$. This asymptotic bias is one symptom of the underlying nonregularity, as discussed by Moodie and Richardson[17].

In many situations where the asymptotic distribution of an estimator is unavailable, bootstrap is used as an alternative approach to conduct inference. But the success of bootstrap also hinges on the underlying smoothness of the estimator. When an estimator is nonsmooth, the ordinary (n out of n) bootstrap procedure produces an inconsistent bootstrap estimator[22]. Inconsistency of bootstrap in the above simple normal theory example has been discussed by Andrews[23]. As shown by Shao[22], an alternative resampling procedure called "m out of n bootstrap" is consistent in such nonsmooth scenarios. One concern regarding the use of this procedure is the slower rate of convergence than $\sqrt{n}$ even in a regular setting (e.g., when $P[H_2:\psi_2^T H_{21}=0]=0$). Moreover, a data-adaptive choice of the tuning parameter m in the present context of DTRs is not obvious; see however Bickel and Sakov[24] and Hall et al.[25] for data-adaptive choice of m in other contexts.

The above concerns regarding nonregularity led us to investigate possible regularizations of the estimation procedure, and then use bootstrap for inference. In the simulation study to follow, we will investigate the behavior of different types of bootstrap confidence intervals for the parameters ψ_j of the optimal DTR in both regular and nonregular settings.

3 Different Regularized Estimators

In this section, we will present two competing estimators to address the non-regularity problem described above. Limited theoretical results are available at this point, and consequently it is not clear which estimator is better. In this paper, we will study their relative merits and demerits in simulations.

From the discussion on nonregularity above, it is clear that $\hat{\psi}_1$ is a non-regular estimator because the stage-1 pseudo-outcome $\hat{Y}_1$ is a non-smooth function (e.g., absolute value) of $\hat{\psi}_2$. The estimators presented in this section "regularize" the nonregular estimator (sometimes called the "hard-max" estimator because of the maximum operation used in the definition) by shrinking or thresholding the effect of the term involving the maximum, e.g., $|\widetilde{\psi}_2^T H_{21}|$, towards zero.

3.1 Hard-threshold Estimator

Recall that the pseudo-outcome $\widehat{Y}_1 = Y_1 + \widehat{\beta}_2^T H_{20} + |\widetilde{\psi}_2^T H_{21}|$ is non-differentiable in $\hat{\psi}_2$ only when $\widetilde{\psi}_2^T H_{21}=0$, and so the corresponding estimator $\hat{\psi}_1$ is problematic only when the true $\psi_2^T H_{21}$ is close to zero. The general form of the hard-threshold pseudo-outcome is

$$\widehat{Y}_{1i}^{HT} = Y_{1i} + \widehat{\beta}_2^T H_{20,i} + |\widetilde{\psi}_2^T H_{21,i}| \cdot \mathbf{1}\{|\widetilde{\psi}_2^T H_{21,i}| > \lambda_i\},\ i=1,\ldots,n, \tag{5}$$

where λ_i (> 0) is the threshold for the i-th subject in the sample (possibly depending on the variability of the linear combination $\widetilde{\psi}_2^T H_{21,i}$ for that subject). One way to operationalize this is to perform a preliminary test (for each subject in the sample) of the hypothesis $H_{0i}:\psi_2^T H_{21,i}=0$ ($H_{21,i}$ is considered fixed in this test), set $\widehat{Y}_{1i}^{HT} = \widehat{Y}_{1i}$ if H_{0i} is rejected, and replace $|\widetilde{\psi}_2^T H_{21,i}|$ with the "better guess" 0 in case H_{0i} is accepted. Thus the hard-threshold pseudo-outcome can be written as

$$\widehat{Y}_{1i}^{HT}=Y_{1i}+\widehat{\beta}_2^T H_{20,i}+|\widehat{\psi}_2^T H_{21,i}|\cdot \mathbf{1}\left\{\frac{\sqrt{n}|\widehat{\psi}_2^T H_{21,i}|}{\sqrt{H_{21,i}^T \widehat{\Sigma}_2 H_{21,i}}}>z_{\alpha/2}\right\}, i=1,\ldots,n, \tag{6}$$

where $\hat{\Sigma}_2$ is the estimated covariance matrix of $\hat{\psi}_2$. The corresponding estimator of ψ_1, denoted by $\widehat{\psi}_1^{HT}$, will be referred to as the hard-threshold estimator. The hard-threshold estimator is common in many areas like variable selection in linear regression and wavelet shrinkage[26]. Moodie and Richardson[17] proposed this estimator for bias correction in the context of Robins' method, and called it *Zeroing Instead of Plugging In* (ZIPI) estimator.

Note that $\widehat{Y}_1^{HT}$ is still a non-smooth function of $\hat{\psi}_2$ and hence $\widehat{\psi}_1^{HT}$ is a nonregular estimator of ψ_1. However, the problematic term $|\widehat{\psi}_2^T H_{21}|$ is shrunk (thresholded) towards zero, and hence one might expect that the degree of nonregularity is somewhat reduced. Moodie and Richardson[17] showed that this estimator reduces the bias occurring in Robins' method (efficient version of Q-learning). In the simulation study to follow, we will explore if this estimator can be used to construct valid confidence intervals for ψ_1. An important issue regarding the use of this estimator is the choice of significance level α of the preliminary test, which is an unknown tuning parameter. As discussed by Moodie and Richardson[17], this is a difficult problem even in better-understood settings where preliminary test based estimators are used; and no widely applicable data-driven method for choosing α in this setting is currently available.

3.2 Soft-threshold or Shrinkage Estimator

The general form of the soft-threshold pseudo-outcome considered here is

$$\widehat{Y}_{1i}^{ST}=Y_{1i}+\widehat{\beta}_2^T H_{20,i}+|\widehat{\psi}_2^T H_{21,i}|\cdot\left(1-\frac{\lambda_i}{|\widehat{\psi}_2^T H_{21,i}|^2}\right)^+, i=1,\ldots,n, \tag{7}$$

where $x^+ = x\mathbf{1}\{x>0\}$ stands for the positive part of a function, and $\lambda_i\,(>0)$ is a tuning parameter associated with the i-th subject in the sample (again possibly depending on the variability of the linear combination $\widehat{\psi}_2^T H_{21,i}$ for that subject). In the contexts of regression shrinkage [27] and wavelet shrinkage[28], the third term in (7) is generally known as the *nonnegative garrote* estimator. As discussed by Zou[29], the nonnegative garrote estimator is a special case of the *adaptive lasso* estimator. As in the case of hard-threshold estimator, a crucial issue here is to choose a data-driven tuning parameter λ_i. Below we provide a choice following a Bayesian approach.

Like the hard-threshold pseudo-outcome, $\widehat{Y}_1^{ST}$ is also a non-smooth function of $\hat{\psi}_2$ and hence $\widehat{\psi}_1^{ST}$ remains a nonregular estimator of ψ_1. However, the problematic term $|\widehat{\psi}_2^T H_{21}|$ is shrunk (or thresholded) towards zero, and hence one might expect that the degree of nonregularity is somewhat reduced. In the simulation study to follow, we will investigate how much improvement this estimator offers over the "hard-max" estimator, when it comes to constructing confidence intervals. Figure 1 presents the hard-max, the hard-threshold, and the soft-threshold pseudo-outcomes.

3.2.1 Choice of Tuning Parameter—A hierarchical Bayesian formulation of the problem, inspired by the work of Figueiredo and Nowak[30] in the area of wavelet-based image processing, can be used in the context of the soft-threshold estimator to choose λ_i's in a data-driven way.

It turns out that the estimator (7) with $\lambda_i = 3H_{21,i}^T \widehat{\sum}_2 H_{21,i}/n, i=1,\ldots,n$, where $\hat{\Sigma}_2/n$ is the estimated covariance matrix of $\hat{\psi}_2$, is an approximate empirical Bayes estimator. The following lemma will be used to derive the choice of λ_i.

Lemma 2: Let X be a random variable such that $X|\mu \sim N(\mu, \sigma^2)$ with known variance σ^2. Let the prior distribution on μ be given by $\mu|\phi^2 \sim N(0, \phi^2)$,with Jeffrey's noninformative hyper-prior on ϕ^2, e.g., $p(\phi^2) \propto 1/\phi^2$. Then an empirical Bayes estimator of $|\mu|$ is given by

$$\widehat{|\mu|}^{EB} = X\left(1-\frac{3\sigma^2}{X^2}\right)^+\left(2\Phi\left(\frac{X}{\sigma}\sqrt{\left(1-\frac{3\sigma^2}{X^2}\right)^+}\right)-1\right) + \sqrt{\frac{2}{\pi}}\sigma\sqrt{\left(1-\frac{3\sigma^2}{X^2}\right)^+}\exp\left\{-\frac{X^2}{2\sigma^2}\left(1-\frac{3\sigma^2}{X^2}\right)^+\right\}, \tag{8}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

The proof is given in Appendix B.

Clearly, $\widehat{|\mu|}^{EB}$ is a thresholding rule, since $\widehat{|\mu|}^{EB}=0$ for $|X|<\sqrt{3}\sigma$. Moreover, when $|\frac{X}{\sigma}|$ is large, the second term of (8) goes to zero exponentially fast, and

$$\left(2\Phi\left(\frac{X}{\sigma}\sqrt{\left(1-\frac{3\sigma^2}{X^2}\right)^+}\right)-1\right) \approx (2\,I_{\{X>0\}}-1) = sign(X).$$

Consequently, the empirical Bayes estimator is approximated by

$$\widehat{|\mu|}^{EB} \approx X\left(1-\frac{3\sigma^2}{X^2}\right)^+ sign(X) = |X|\left(1-\frac{3\sigma^2}{X^2}\right)^+. \tag{9}$$

Now for $i = 1,\ldots,n$ separately, put $X=\widetilde{\psi}_2^T H_{21,i}$, and $\mu=\widetilde{\psi}_2^T H_{21,i}$ (for fixed $H_{21;i}$); and plug in $\widehat{\sigma}^2 = H_{21,i}^T \widehat{\sum}_2 H_{21,i}/n$ for σ^2. This leads to a choice of λ_i in the soft-threshold pseudo-outcome (7):

$$\widehat{Y}_{1i}^{ST} = Y_{1i} + \widehat{\beta}_2^T H_{20,i} + |\widehat{\psi}_2^T H_{21,i}| \cdot \left(1-\frac{3H_{21,i}^T \widehat{\sum}_2 H_{21,i}}{n|\widehat{\psi}_2^T H_{21,i}|^2}\right)^+, \tag{10}$$

$$= Y_{1i} + \widehat{\beta}_2^T H_{20,i} + |\widehat{\psi}_2^T H_{21,i}| \cdot \left(1-\frac{3H_{21,i}^T \widehat{\sum}_2 H_{21,i}}{n|\widehat{\psi}_2^T H_{21,i}|^2}\right) \cdot \mathbf{1}\left\{\frac{\sqrt{n}|\widehat{\psi}_2^T H_{21,i}|}{\sqrt{H_{21,i}^T \widehat{\sum}_2 H_{21,i}}} > \sqrt{3}\right\}, \quad i=1,\ldots,n. \tag{11}$$

The presence of the indicator function in (11) indicates that $\widehat{Y}_{1i}^{ST}$ is a thresholding rule for small values of $|\widehat{\psi}_2^T H_{21,i}|$, while the term just preceding the indicator function makes $\widehat{Y}_{1i}^{ST}$ a shrinkage rule for moderate to large values of $|\widehat{\psi}_2^T H_{21,i}|$ (for which the indicator function takes the value one). Thus the current Bayesian formulation gives us a data-driven choice of the tuning parameters.

4 Simulation Study

In this section, we consider a simulation study to compare the performances of the hard-max, the hard-threshold, and the soft-threshold estimators under different nonregular scenarios. In this study, we vary the parameters of the generative model, the degree of nonregularity, and the type of bootstrap confidence interval.

Generative Model

Recall that the data consist of n patient trajectories, each of the form $(O_1, A_1, O_2, A_2, O_3)$. Without loss of generality, we assume $Y_1 \equiv 0$ and $Y_2 \equiv Y = O_3$. Let $\mu_Y = E[Y|O_1, A_1, O_2, A_2]$, and ε be the associated error term. Then $Y = \mu_Y + \varepsilon$, where

$$\mu_Y = \gamma_1 + \gamma_2 O_1 + \gamma_3 A_1 + \gamma_4 O_1 A_1 + \gamma_5 A_2 + \gamma_6 O_2 A_2 + \gamma_7 A_1 A_2,$$

and $\varepsilon \sim N(0, 1)$. Next, we consider binary treatments randomized with probability 1/2, e.g., $P[A_j = 1] = P[A_j = -1] = 1/2, j = 1, 2$. Also, the binary covariates O_j's are generated as

$$P[O_1 = 1] = P[O_1 = -1] = 1/2,$$
$$P[O_2 = 1|O_1, A_1] = 1 - P[O_2 = -1|O_1, A_1] = expit(\delta_1 O_1 + \delta_2 A_1),$$

where $expit(x) = \exp(x)/(1 + \exp(x))$. Note that $\gamma_1, \ldots, \gamma_7$ and δ_1, δ_2 are the parameters that specify the generative model. These parameters will be varied in the examples to follow.

Analysis Model

$$Q_2(H_2, A_2) = \beta_{20} + \beta_{21} O_1 + \beta_{22} A_1 + \beta_{23} O_1 A_1 + (\psi_{20} + \psi_{21} O_2 + \psi_{22} A_1) A_2,$$
$$Q_1(H_1, A_1) = \beta_{10} + \beta_{11} O_1 + (\psi_{10} + \psi_{11} O_1) A_1.$$

Two dimensions of nonregularity: p and φ

Nonregularity in stage 1 parameters arises when the optimal stage 2 treatment is non-unique for at least some subjects in the population. With reference to the present generative model, a setting is nonregular if the linear combination $\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1 = 0$ with positive probability. Also one might expect some nonregular behavior as $\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1$ falls in a small neighborhood of zero (even though not exactly zero). In the following, we consider specific examples varying the "degree of nonregularity", e.g., $p = P[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1 = 0]$ and the "standardized effect size" defined as $\phi = |E[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1]| / \sqrt{Var[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1]}$. The quantities p and ϕ, which depend on the distribution of the above linear combination, represent two dimensions of the nonregularity phenomenon. Note that the linear combination $(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)$ can take only four possible values corresponding to the four possible (O_2, A_1) cells. The cell probabilities can be easily calculated; the formulae are provided in Table 1.

It follows that $E[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1] = q_1 f_1 + q_2 f_2 + q_3 f_3 + q_4 f_4$, and $E[(\gamma_5+\gamma_6 O_2+\gamma_7 A_1)^2]=q_1 f_1^2+q_2 f_2^2+q_3 f_3^2+q_4 f_4^2$, where $q_1,\ldots,q_4$ are the cell probabilities given in Table 1. From these two, one can calculate $Var[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1]$, and subsequently the effect size ϕ.

We want to conduct inference on ψ_{10} and ψ_{11}, the analysis model parameters associated with stage 1 treatment A_1. They can be expressed in terms of γ's and δ's, the parameters of the generative model, as follows. It turns out that

$$\psi_{10}=\gamma_3+q_1|f_1| - q_2|f_2|+q_3|f_3| - q_4|f_4|,$$
$$\text{and } \psi_{11}=\gamma_4+q_1'|f_1| - q_2'|f_2| - q_3'|f_3|+q_4'|f_4|,$$

where $q_1'=q_3'=\frac{1}{4}(expit(\delta_1+\delta_2) - expit(-\delta_1+\delta_2))$, and $q_2'=q_4'=\frac{1}{4}(expit(\delta_1 - \delta_2) - expit(-\delta_1 - \delta_2))$. In the following, we consider specific examples for varying p and ϕ. In Examples 1 − 4 below, we use $\delta_1 = \delta_2 = 0.5$. For this choice, we get the following values of the cell probabilities: $q_1 = q_4 = 0.3078$ and $q_2 = q_3 = 0.1922$. This choice of the δ's also makes $q_1'=q_2'=q_3'=q_4'=.0578$.

Example 1 (p = 1, φ undefined)—Consider a setting where there is no treatment effect for any subject (any history) in either stage. This is achieved by setting $\gamma_1 = \ldots = \gamma_7 = 0$, and $\delta_1 = \delta_2 = 0.5$. Then $f_1 = f_2 = f_3 = f_4 = 0$, and hence $\psi_{10} = \psi_{11} = 0$, $p = 1$, and ϕ is undefined (0/0). This is a fully nonregular scenario.

Example 2 (p = 0, φ infinite)—Consider a setting similar to Example 1, where there is a very weak stage 2 treatment effect for every subject (all possible history). This is achieved by setting $\gamma_5 = 0.01$ and $\gamma_j = 0$, $\forall j \neq 5$, and $\delta_1 = \delta_2 = 0.5$. Then $f_1 = f_2 = f_3 = f_4 = 0.01$; $\psi_{10} = \psi_{11} = 0$, $p = 0$, and ϕ is infinite (0.01/0). This is a regular scenario, but close to nonregularity (it is hard to detect the very weak effect given the noise level in the data).

Example 3 ($p=\frac{1}{2}$, φ = 1)—Consider a setting where there is no stage 2 treatment effect for half the subjects in the population, but a reasonably large effect for the other half of subjects. This is achieved by setting $\gamma_1 = \gamma_2 = \gamma_4 = \gamma_6 = 0$, $\gamma_3 = -0.5$, $\gamma_5 = \gamma_7 = 0.5$, and $\delta_1 = \delta_2 = 0.5$. Then $f_1 = f_3 = 1, f_2 = f_4 = 0$, $\psi_{10} = \psi_{11} = 0$, $p=\frac{1}{2}$ and $\phi = 1$. This is a nonregular setting.

Example 4 (p = 0, φ = 1.0204)—Consider a setting where there is a very weak stage 2 treatment effect for half the subjects in the population, but a reasonably large effect for the other half of subjects. This is achieved by setting $\gamma_1 = \gamma_2 = \gamma_4 = \gamma_6 = 0$, $\gamma_3 = -0.5$, $\gamma_5 = 0.5$, $\gamma_7 = 0.49$, and $\delta_1 = \delta_2 = 0.5$. It follows that $f_1 = f_3 = 0.99, f_2 = f_4 = 0.01$, $\psi_{10} = -0.0100$, $\psi_{11} = 0$, $p = 0$, and $\phi = 1.0204$. This regular example is close to the nonregular Example 3.

Example 5 ($p=\frac{1}{4}$, φ = 1.4142)—Consider a setting where there is no stage 2 treatment effect for one-fourth of the subjects in the population, but others have a reasonably large effect. To achieve this, set $\gamma_1 = \gamma_2 = \gamma_4 = 0$, $\gamma_3 = -0.5$, $\gamma_5 = 1$, $\gamma_6 = \gamma_7 = 0.5$, $\delta_1 = 1$, and $\delta_2 = 0$. Then $f_1 = 2, f_2 = f_3 = 1, f_4 = 0$; the cell probabilities are equal, i.e., $q_1=q_2=q_3=q_4=\frac{1}{4}$; and $q_1'=q_2'=q_3'=q_4'=0.1155$. Consequently, $\psi_{10} = \psi_{11} = 0$, $p=\frac{1}{4}$, and $\phi = 1.4142$. This is a nonregular setting.

Example 6 (p = 0, φ = 0.3451)—Consider a completely regular setting where there is a reasonably large stage 2 treatment effect for every subject in the population. This can be achieved by setting $\gamma_1 = \gamma_2 = \gamma_4 = 0$, $\gamma_3 = -0.5$, $\gamma_5 = 0.25$, $\gamma_6 = \gamma_7 = 0.5$, and $\delta_1 = \delta_2 = 0.1$. Then

$f_1 = 1.25, f_2 = f_3 = 0.25$, and $f_4 = -0.75$; the cell probabilities are $q_1 = q_4 = 0.2625$, $q_2 = q_3 = 0.2375$; and $q_1' = q_2' = q_3' = q_4' = 0.0125$. It follows that $\psi_{10} = -0.3688$, $\psi_{11} = 0.0187$, $p = 0$ and $\phi = 0.3451$.

Note that in Example 5, the effect size ϕ is greater than Cohen's[31] benchmark large effect size (=0.8). Such a high effect size can be criticized as being unrealistic, based on the *principle of clinical equipoise*[32], which provides the ethical basis for medical research involving randomization. This principle says that there must be a honest, professional disagreement (high variability) among expert clinicians about the preferred treatment (and thus the standardized effect size of treatment is likely small). Hence this example might be somewhat down-weighted for overall comparison of performance. Furthermore, Example 6 violates the *Hierarchical Ordering Principle*[33] in that the coefficient of the interaction term A_1A_2 (γ_7) is larger than the co-efficient of the main effect A_2 (γ_5). So this example might be given lower weight as well.

Competing Estimators

In the simulation, we will consider four estimators: the hard-max estimator (original Q-learning), the soft-threshold estimator, and the hard-threshold estimator with two values of the tuning parameter α, e.g., 0.2, which was empirically found to be a good choice by Moodie and Richardson[17], and 0.08 which corresponds to the threshold used by the soft-threshold estimator proposed in this paper (from (11), the threshold used by the soft-threshold estimator is $\sqrt{3}=1.7321$; equating this point to $z_{\alpha/2}$ and solving for α, we get $\alpha = 0.0833$).

Different Bootstrap CIs

We consider three types of bootstrap CIs, e.g., percentile, hybrid, and double (percentile) bootstrap CIs. Let $\hat{\theta}$ be an estimator of θ and $\hat{\theta}^*$ be its bootstrap version. Then the $100(1-\alpha)$% percentile bootstrap (PB) CI is given by $\left(\widehat{\theta}^*_{(\frac{\alpha}{2})}, \widehat{\theta}^*_{(1-\frac{\alpha}{2})}\right)$, and the $100(1-\alpha)$% hybrid bootstrap (HB) CI is given by $\left(2\widehat{\theta} - \widehat{\theta}^*_{(1-\frac{\alpha}{2})}, 2\widehat{\theta} - \widehat{\theta}^*_{(\frac{\alpha}{2})}\right)$, where $\widehat{\theta}^*_\gamma$ is the 100γ-th percentile of the bootstrap distribution. The double bootstrap (DB) CI is calculated as follows:

1. Draw B_1 first-stage bootstrap samples from the original data. For each first-stage bootstrap sample, calculate the bootstrap version of the estimator $\hat{\theta}^{*b}$, $b = 1,\ldots,B_1$.
2. Conditional on each first-stage bootstrap sample, draw B_2 second-stage (nested) bootstrap samples and calculate the double bootstrap versions of the estimator, e.g., $\hat{\theta}^{**bm}$, $b = 1,\ldots,B_1$, $m = 1,\ldots,B_2$.
3. For $b = 1,\ldots,B_1$, calculate $u^{*b} = \frac{1}{B_2}\sum_{m=1}^{B_2} \mathbf{1}\{\widehat{\theta}^{**bm} \le \widehat{\theta}\}$, where $\hat{\theta}$ is the estimator based on the original data.
4. The double bootstrap CI is given by $\left(\widehat{\theta}^*_{\widehat{q}(\frac{\alpha}{2})}, \widehat{\theta}^*_{\widehat{q}(1-\frac{\alpha}{2})}\right)$, where $\widehat{q}(\gamma) = u^*_{(\gamma)}$, the 100γ-th percentile of the distribution of u^{*b}, $b = 1,\ldots,B_1$.

See Davison and Hinkley[34] and Nankervis[35] for details about double bootstrap CIs. One disadvantage of these CIs is that they are computationally very intensive.

We use $B = 1000$ bootstrap iterations to calculate the percentile and the hybrid bootstrap CIs. However, the double bootstrap CIs are based on $B_1 = 500$ first-stage and $B_2 = 100$ second-stage bootstrap iterations (due to the increased computational burden). The results in Tables 2 – 3 are based on $N = 1000$ Monte Carlo iterations.

4.1 Results

The simulation study compares the competing estimators on a variety of settings represented by Examples 1 – 6. We considered estimation and inference for both ψ_{10} and ψ_{11}. However in the present examples, the effect of nonregularity turned out to be more pronounced for the parameter ψ_{10} (main effect of A_1) than ψ_{11} (interaction of A_1 with O_1). Hence we included results on ψ_{10} only in Tables 2 and 3. Also in the following discussion, we will focus on ψ_{10}.

In Example 1 (top part of Table 2), where stage 2 effects for all possible histories are zero (i.e., the stage 2 optimal treatment is non-unique for every subject in the population), we see that there is no bias associated with the hard-max estimator; and the mean squared error (MSE) is essentially the same as the variance. However the percentile bootstrap CI (both 95% and 90%) has over-coverage (note that over-coverage translates to lower power of the corresponding hypothesis test), and the hybrid bootstrap CI (95%) has under-coverage compared to the nominal level. We have also studied the Wald type CIs for this setting (not included in this paper) and observed over-coverage (the problem with Wald type CIs in such nonregular settings is well-known11›17). This suggests that the asymptotic distribution of the hard-max estimator has a lighter tail than a comparable normal distribution. However, the double bootstrap CIs have correct coverage. Note that both versions of the hard-threshold estimator fail to rectify the coverage rate, even though neither suffer from bias. However, the soft-threshold estimator offers correct coverage for both types of bootstrap CIs. Moreover, it gives the lowest MSE among the four estimators. Note that the soft-threshold estimator is also non-smooth (nonregular), and consequently the bootstrap distribution is inconsistent for the true asymptotic distribution of this estimator. But in this setting, it reduces the degree of nonregularity just enough so that the bootstrap CIs do not show the problem with coverage.

Even though Example 2 (middle part of Table 2) is a regular setting ($p = 0$), it is very close to Example 1 and hence affected by nonregularity. Results are similar to those in Example 1. Thus the presence of very small effects causes problems with coverage even in regular settings.

Example 3 (bottom part of Table 2) is a setting where the stage 2 optimal treatment is non-unique for half the subjects in the population ($p=\frac{1}{2}$) and is unique for the remaining half, but the overall standardized stage 2 effect size ϕ (= 1) is quite large. Here the hard-max estimator is biased, and hence both the percentile and the hybrid bootstrap CIs under-cover the true value. However the double bootstrap CI gives correct coverage rate. Both versions of the hard-threshold estimator reduce bias and one of them (corresponding to $\alpha = 0.08$) gives correct coverage, while the other also offers substantial improvement of the coverage rate. This is consistent with the findings of Moodie and Richardson[17]. The soft-threshold estimator also reduces bias, gives the lowest MSE among the four estimators, and provides correct coverage with the hybrid bootstrap method but not with the percentile method (even though it offers substantial improvement). Thus in this example, the hard-threshold estimator with $\alpha = 0.08$ emerges as the winner, with the soft-threshold estimator at the second place. However, note that the value 0.08 of the tuning parameter α is not arbitrary – it corresponds to the threshold used by the soft-threshold estimator. If constructing confidence intervals is the main goal (so biased estimation is less of an issue), double bootstrap CI along with the hard-max estimator can also be used in this setting, although it is computationally more expensive.

Example 4 (top part of Table 3) is a regular setting, very similar to the nonregular setting in Example 3. Results are quite similar to those in Example 3. This is consistent with our previous observation (Example 2) that the presence of very small effects causes problems with coverage even in regular settings.

In example 5 (middle part of Table 3), the stage 2 optimal treatment is non-unique for one-fourth of the subjects in the population ($p=\frac{1}{4}$) and the standardized effect size ϕ is very large

(=1.4142). Again, the hard-max estimator is biased, and has low coverage of the CIs (except for double bootstrap). The hard-threshold and the soft-threshold estimators offer improvement in terms of bias as well as coverage. The soft-threshold estimator emerges as the best (lowest MSE and correct coverage rate) in this example.

Example 6 (bottom part of Table 3) is a regular setting ($p = 0$, with no extremely tiny stage 2 effect as in Examples 2 and 4), with the standardized effect size 0.3451. The reason for investigating this setting is to check if the regularized estimators (hard and soft threshold) perform poorly in settings where there is no need to regularize. As expected, the hard-max estimator performs well here. The soft-threshold estimator introduces some bias when there is none in the hard-max estimator and increases MSE; but still manages to provide correct coverage for the percentile bootstrap method. The hard-threshold estimators also give correct coverage for percentile CIs.

To summarize, the hard-max estimator is problematic in nonregular scenarios, except when used with the computationally intensive double bootstrap method for constructing confidence intervals. The hard-threshold estimator, if properly tuned, addresses the problem of bias but not the problem of light tail. The soft-threshold estimator seems to address both problems to a large extent. In the simulation, the soft-threshold estimator consistently produced the lowest MSE among the competing methods across all the nonregular scenarios. Also in all the nonregular settings, either the soft-threshold estimator or the hard-threshold estimator with $\alpha = 0.08$ (this α corresponds to the threshold used by the soft-threshold estimator) emerged as the winner in terms of providing correct coverage rate of the bootstrap CIs. Even though the soft-threshold estimator incurs some bias in regular settings, it manages to provide reasonable coverage rate for small to moderate standardized effect sizes (we have studied up to around 0.35). Across all the scenarios considered here (Examples 1–6), the soft-threshold estimator emerged as more robust than the hard-threshold estimator to the degree of regularity of the underlying data distribution, probably because of its "soft" nature (the soft-threshold estimator is continuous everywhere even though it has two points of non-differentiability, whereas the hard-threshold estimator has two points of discontinuity – see Figure 1). Furthermore, note that overall the hybrid bootstrap CIs performed slightly better than the percentile bootstrap CIs in this simulation study. Hence the hybrid bootstrap CIs will be used in the data analysis to follow.

5 Analysis of Smoking Cessation Data

To demonstrate the occurrence of nonregularity and the use of the soft-threshold method in a real application, here we present the analysis of a data set from a randomized, two-stage, longitudinal, internet-based smoking cessation study conducted by the Center for Health Communications Research at the University of Michigan. The stage 1 of this study (*Project Quit*) was conducted to find an optimal multi-factor behavioral intervention to help adult smokers quit smoking; and the stage 2 (*Forever Free*) was a follow-on study to help those (among the participants of *Project Quit*) who already quit stay quit, and help those who failed at the previous stage with a second chance. Details of the study design and primary analysis of the stage 1 data can be found in Strecher et al.[36]

At stage 1, although there were five two-level treatment factors, only two, e.g., `source` (of online behavioral counseling message) and `story` (of a hypothetical character who succeeded in quitting smoking) were significant in the analysis reported in Strecher et al.[36] For simplicity, we considered only these two treatment factors at stage 1 of our present analysis, which gave a total of 4 treatment combinations at stage 1 corresponding to the 2×2 design. The treatment factor `source` was varied at two levels, e.g., high vs. low personalized, coded 1 and −1; also the factor `story` was varied at two levels, e.g., high vs. low tailoring depth (degree

to which the character in the story was tailored to the individual subject's baseline characteristics), coded 1 and −1. Baseline variables at this stage included subjects' `motivation` to quit (on a 1–10 scale), `selfefficacy` (on a 1–10 scale) and `education` (binary, ≤ high school vs. > high school, coded −1/1). At stage 2, originally there were 4 different treatment groups and a control group; however the 4 treatment groups were combined together for the present analysis because of very little difference between them. This resulted in only two choices of treatment at stage 2; this treatment variable was called `FFarm`, coded −1/1 (1=treatment, −1 = control).

There were two outcomes at the two stages of this study. The stage 1 outcome was binary quit status called `PQ6Quitstatus` (1=quit, 0=not quit) at 6 month from the date of randomization. The stage 2 outcome was binary quit status `FF6Quitstatus` at 6 months from the date of stage 2 randomization (i.e., 12 months from the date of stage 1 randomization).

An example DTR can have the following form: "At stage 1, if a subject's baseline `selfefficacy` is greater than a threshold value (say 7, on a 1–10 scale), then provide the highly-personalized level of the treatment component `source`, and if the subject is willing to continue treatment, then at stage 2 provide treatment if he/she continues to be a smoker at the end of stage 1". Of course characteristics other than `selfefficacy` or a combination of more than one subject characteristics can be used to specify a DTR. To find the optimal DTR, we applied both the hard-max and the soft-threshold estimators within the Q-learning framework. This involved:

1. a stage 2 regression ($n = 281$) of `FF6Quitstatus` using the model:

$$\begin{aligned} FF6Quitstatus = {} & \beta_{20} + \beta_{21}\, motivation + \beta_{22}\, source + \beta_{23}\, selfefficacy \\ & + \beta_{24}\, story + \beta_{25}\, education + \beta_{26}\, PQ6Quitstatus \\ & + \beta_{27}\, source * selfefficacy + \beta_{28}\, story * education \\ & + (\psi_{20} + \psi_{21}\, PQ6Quitstatus) * FFarm + \varepsilon_2; \end{aligned}$$

2. finding both the hard-max pseudo-outcome ($\hat{Y}_1$) and the soft-threshold pseudo-outcome ($\widehat{Y}_1^{ST}$) for the stage 1 regression:

$$\begin{aligned} \widehat{Y}_1 = {} & PQ6Quitstatus + \widehat{\beta}_{20} + \widehat{\beta}_{21}\, motivation + \widehat{\beta}_{22}\, source + \widehat{\beta}_{23}\, selfefficacy \\ & + \widehat{\beta}_{24}\, story + \widehat{\beta}_{25}\, education + \widehat{\beta}_{26}\, PQ6Quitstatus \\ & + \widehat{\beta}_{27}\, source * selfefficacy + \widehat{\beta}_{28}\, story * education \\ & + |\widehat{\psi}_{20} + \widehat{\psi}_{21}\, PQ6Quitstatus|; \end{aligned}$$

$$\begin{aligned} \widehat{Y}_1^{ST} = {} & PQ6Quitstatus + \widehat{\beta}_{20} + \widehat{\beta}_{21}\, motivation + \widehat{\beta}_{22}\, source + \widehat{\beta}_{23}\, selfefficacy \\ & + \widehat{\beta}_{24}\, story + \widehat{\beta}_{25}\, education + \widehat{\beta}_{26}\, PQ6Quitstatus \\ & + \widehat{\beta}_{27}\, source * selfefficacy + \widehat{\beta}_{28}\, story * education \\ & + |\widehat{\psi}_{20} + \widehat{\psi}_{21}\, PQ6Quitstatus| \cdot \left(1 - \frac{3\mathrm{Var}(\widehat{\psi}_{20} + \widehat{\psi}_{21}\, PQ6Quitstatus)}{|\widehat{\psi}_{20} + \widehat{\psi}_{21}\, PQ6Quitstatus|^2}\right)^{+}; \end{aligned}$$

and (3) for each of the two pseudo-outcomes, a stage 1 regression ($n = 1401$) of the pseudo-outcome using a model of the form:

$$\widehat{Y}_1 \text{ or } \widehat{Y}_1^{ST} = \beta_{10} + \beta_{11}\, motivation + \beta_{12}\, selfefficacy + \beta_{13}\, education + \left(\psi_{10}^{(1)} + \psi_{11}^{(1)}\, selfefficacy\right) * source + \left(\psi_{10}^{(2)} + \psi_{11}^{(2)}\, education\right) * story + \varepsilon_1.$$

Note that the sample sizes at the two stages differ because only 281 subjects were willing to continue treatment into stage 2 (as allowed by the study protocol). Our stage 2 analysis was a usual regression analysis. No significant treatment effect was found at this stage, indicating the likely existence of nonregularity. At stage 1, for either estimator, 95% confidence intervals were constructed by hybrid bootstrap using 1000 bootstrap replications. The stage 1 analysis summary is presented in Table 4. In this case, the hard-max and the soft-threshold estimators produced similar results.

The conclusions from the present data analysis can be summarized as follows. We did not find any significant stage 2 treatment effect. So this analysis suggests that the stage 2 behavioral intervention need not be adapted to the smoker's individual characteristics, interventions previously received, or stage 1 outcome. More interesting results are found at stage 1. It is found that subjects with higher level of `motivation` or `selfefficacy` are more likely to quit. The highly personalized level of `source` is more effective for subjects with a higher `selfefficacy` (≥ 7), and deeply tailored level of `story` is more effective for subjects with lower `education` ($\leq$ high school); these two conclusions can be drawn from the interaction plots (with confidence intervals) presented in figure 2. Thus this secondary data analysis suggests that to maximize each individual's chance of quitting over the two stages, the web-based smoking cessation intervention should be designed in future such that: (1) smokers with high `self-efficacy` (≥ 7) are assigned to highly personalized level of `source`, and (2) smokers with lower `education` are assigned to deeply tailored level of `story`.

6 Discussion

In this paper, we have illustrated the problem of nonregularity that arises in the context of DTRs in the estimation of the optimal current treatment rule, when the optimal treatments at subsequent stages are non-unique for at least some proportion of subjects in the population. We have illustrated the phenomenon using Q-learning as the estimation procedure, which is a simpler yet inefficient version of Robins' method; however the problem of nonregularity arises in Robins' method as well[11,17].

For some underlying data-generating models (e.g., Examples 3, 4, 5 in the simulation study), this nonregularity induces bias in the point estimates of the parameters of the optimal DTRs, which in turn causes under-coverage of the bootstrap confidence intervals. In contrast, in case of Examples 1 and 2, this nonregularity causes lightness of tail of the asymptotic distribution but no bias, as seen from the over-coverage of the percentile bootstrap CIs (equivalently conservative tests leading to lower power). The coexistence of these two not-so-well-related issues (they work in opposite directions, e.g., bias tends to make the CIs under-cover, whereas lightness of tail tends to make the CIs over-cover) makes this problem a unique and challenging one.

As mentioned in section 2.4, the phenomenon of nonregularity can be understood more clearly with a simpler problem, e.g., estimating $|\mu|$ (note that ψ_{10} is a linear combination of terms like $|\mu|$; see section 4) by $|\bar{X}_n|$ (similar to the hard-max estimator), where $\bar{X}_n$ is the sample average of n i.i.d. observations $X_1,\ldots,X_n$ from $N(\mu, 1)$. From section 2.4, we know that when $\mu = 0$, $|\bar{X}_n|$ is a biased estimator of $|\mu| = 0$, with $bias = E(|\bar{X}_n|) = \sqrt{\frac{2}{n\pi}}$. Because of this bias (wrong centering), bootstrap CIs exhibited gross under-coverage in a toy example. But once we used

a bias-corrected estimate (corrected by the analytically calculated bias), the percentile bootstrap CI exhibited over-coverage. This suggests that the distribution of $|\bar{X}_n|$ is peaked around its mean, or to put it in another way, has light tails.

In the simulation study to compare the competing estimators of the optimal DTR, we considered estimation of ψ_{10}, which involve linear combinations of $|f_1|$, $|f_2|$, $|f_3|$, and $|f_4|$ (terms like $|\mu|$). Under the non-regular scenarios, some or all (depending on the degree of nonregularity p) of the f_i's are zero; and hence a phenomenon similar to the one described above in the toy example happens for each $|f_i|$ for which $f_i = 0$. Each such term has its associated bias, and each has its own lightness of tail, with bias being the dominant property. In some nonregular scenarios (Example 1), the bias associated with the individual $|f_i|$'s (in the expression for ψ_{10}) cancel each other out (note the opposite signs in front of $|f_i|$'s), and hence the lightness of tail is revealed, resulting in a percentile bootstrap CI that over-covers. In other nonregular examples, however, bias is not canceled out, and hence dominates the property of the hard-max estimator. Hence under-coverage of the bootstrap CIs is observed.

Nonregularity is an issue in the estimation of the optimal DTRs because it arises when there is no treatment effect at subsequent stages (equivalently, there is no unique optimal treatment at subsequent stages). Unfortunately often there is no or very weak treatment effect in the settings we are interested in (e.g., randomized trials on mental illness or substance abuse). Thus we want our estimator to enjoy good statistical properties (e.g., less bias, lower risk or MSE, correct coverage rate of CIs, good power to detect "local" alternatives, etc.) when the optimal treatment at subsequent stages is non-unique. In case of the hard-max estimator, unfortunately the point of non-differentiability coincides with the parameter value such that $\psi_2^T H_{21}=0$ (non-unique optimal treatment at the subsequent stage), which causes nonregularity (bias, higher MSE, low power). But the soft-threshold estimator (also, hard-threshold estimator), in some sense, redistributes the nonregularity from this "null point" to two different points symmetrically placed on either side of the "null point" (see Figure 1). This is one reason why the soft-threshold (also, hard-threshold) estimator works well in nonregular settings.

We have shown that using bootstrap confidence intervals along with the soft-threshold (also, hard-threshold in some cases) estimator reduces the degree of nonregularity, and gives correct coverage rate. Also, the double bootstrap method can be used along with the original hard-max estimator to address the nonregularity. But this method is highly computationally intensive and may be difficult to use in practice. An alternative method to construct CIs for ψ's in nonregular settings is the score method due to Robins[11]. We have not investigated this in our simulation study.

One can consider an alternative Bayesian approach to formulate an estimator similar to the soft-threshold estimator as follows. Let the data distribution $\widehat{\psi}_2^T H_{21}|\psi_2^T H_{21} \sim N(\psi_2^T H_{21}, \sigma^2)$ with known σ^2, and the prior distribution of $\psi_2^T H_{21}$ be a mixture of a point mass at 0 and $N(0, 1)$, with mixing parameter p $(0 < p < 1)$. Then the posterior distribution of $\psi_2^T H_{21}$ is a mixture distribution given by

$$f_{post}(\psi_2^T H_{21}) = \widehat{w} \cdot \mathbf{1}\{\psi_2^T H_{21}=0\} + (1-\widehat{w}) \cdot N\left(\frac{\widehat{\psi}_2^T H_{21}}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right),$$

$$\text{where } \widehat{w} = \left\{1+\frac{1-p}{p}\sqrt{\frac{\sigma^2}{1+\sigma^2}}\exp\left\{\frac{(\widehat{\psi}_2^T H_{21})^2}{2\sigma^2(1+\sigma^2)}\right\}\right\}^{-1}.$$

One can use the median of this posterior distribution in place of $\widehat{\psi}_2^T H_{21}$ in the expression for $\hat{Y}_1$. Thus the Bayes estimator becomes

$$\widehat{Y}_1^{Bayes} = \widehat{\beta}_2^T H_{20} + \text{median of } f_{post}(\psi_2^T H_{21}).$$

For using this, one has to replace σ^2 by $\widehat{\sigma}^2 = H_{21}^T \widehat{\sum}_2 H_{21}/n$, and p by either some empirical estimate or a fixed value (e.g., $\frac{1}{2}$). In place of the above mixture prior, Johnstone and Silverman[37] suggest using the mixture of a point mass and a heavy-tailed distribution (e.g., double-exponential). This is a promising formulation that we want to investigate in future.

In this paper, we have focused on randomized trials only to separate the issue of nonregularity from causal inference issues. However the problem of nonregularity also arises when observational data11›17 are used; and the hard-threshold and the soft-threshold estimators should be applicable in those settings as well. Also, here we have focussed on only two stages for clarity. However, it should be understood that Q-learning can be used for studies with more than two stages as well. In case of many stages, one can think of a scenario where some parameters are shared across stages, in which case a simultaneous version of Q-learning (as opposed to the recursive version discussed in this paper) would be more appropriate. Unfortunately nonregularity does not go away if a simultaneous estimation procedure is used; see Moodie and Richardson[17] for a discussion on this with reference to Robins' method. However, unlike the case of recursive estimation, it is not well understood at this point whether the threshold estimators (hard or soft) can reduce the nonregularity in simultaneous estimation. Moodie and Richardson17 gave a simulated nonregular example showing that hard-threshold or ZIPI estimator is not always better than simultaneous estimator of Robins. We did not investigate this issue in the current paper, but we recognize this as an important avenue of future research.

To conclude, we think in the estimation of optimal DTRs, appropriately tuned hard-threshold estimator and the soft-threshold estimator should be seriously considered as improved versions of Q-learning (and Robins' method of estimation).

Acknowledgments

We acknowledge support for this project from National Institutes of Health grants RO1 MH080015, P50 DA10075, and P50 CA101451.

References

1. Collins LM, Murphy SA, Bierman K. A conceptual framework for adaptive preventive interventions. Prevention Science 2004;5:185–196. [PubMed: 15470938]
2. Lavori PW. A design for testing clinical strategies: biased adaptive within-subject randomization. Journal of the Royal Statistical Society, Series A 2000;163:29–38.
3. Lavori PW, Dawson R. Dynamic treatment regimes: practical design considerations. Clinical Trials 2004;1:9–20. [PubMed: 16281458]
4. Dawson R, Lavori PW. Placebo-free designs for evaluating new mental health treatments: the use of adaptive treatment strategies. Statistics in Medicine 2004;23:3249–3262. [PubMed: 15490427]
5. Murphy SA. An experimental design for the development of adaptive treatment strategies. Statistics in Medicine 2005;24:1455–1481. [PubMed: 15586395]
6. Schneider LS, Tariot PN, Lyketsos CG, Dagerman KS, Davis KL, Davis S. National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE): alzheimer disease trial methodology. American Journal of Geriatric Psychology 2001;9:346–360.

7. Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. Controlled Clinical Trials 2003;25:119–142. [PubMed: 15061154]
8. Thall PF, Millikan RE, Sung HG. Evaluating multiple treatment courses in clinical trials. Statistics in Medicine 2000;30:1011–1128. [PubMed: 10790677]
9. Wahed AS, Tsiatis AA. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomized designs in clinical trials. Biometrics 2004;60:124–133. [PubMed: 15032782]
10. Murphy SA. Optimal dynamic treatment regimes (with discussions). Journal of the Royal Statistical Society, Series B 2003;65:331–366.
11. Robins, JM. Optimal structural nested models for optimal sequential decisions. In: Lin, DY.; Heagerty, P., editors. Proceedings of the Second Seattle Symposium on Biostatistics. New York: Springer; 2004. p. 189-326.
12. Moodie EEM, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics 2007;63:447–455. [PubMed: 17688497]
13. Thall PF, Sung HG, Estey EH. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. Journal of the American Statistical Association 2002;97(457):29–39.
14. Thall PF, Wooten LH, Logothetis CJ, Millikan RE, Tannir NM. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. Statistics in Medicine 2007;26:4687–4702. [PubMed: 17427204]
15. Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. Biometrics 2002;58:48–57. [PubMed: 11890326]
16. Wahed AS, Tsiatis AA. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. Biometrika 2006;93(1):163–177.
17. Moodie EEM, Richardson TS. Estimating optimal dynamic regimes: correcting bias under the null. Scandinavian Journal of Statistics. 2008
18. Watkins, CJCH. PhD Thesis. Cambridge University; 1989. Learning from delayed rewards.
19. Sutton, RS.; Barto, AG. Reinforcement learning: An introduction. Cambridge: MIT Press; 1998.
20. Murphy SA. A generalization error for Q-learning. Journal of Machine Learning Research 2005;6:1073–1097. [PubMed: 16763665]
21. Bickel, P.; Klaassen, C.; Ritov, Y.; Wellner, J. Efficient and adaptive estimation for semiparametric models. Johns Hopkins University Press; 1993.
22. Shao J. Bootstrap sample size in nonregular cases. Proceedings of the American Mathematical Society 1994;122(4):1251–1262.
23. Andrews DWK. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. Econometrica 2000;68(2):399–405.
24. Bickel P, Sakov A. On the choice of *m* in the *m* out of *n* bootstrap and confidence bounds for extrema. Statistica Sinica 2008;18(3):967–985.
25. Hall P, Horowitz J, Jing B. On blocking rules for the bootstrap with dependent data. Biometrika 1995;82:561–574.
26. Donoho DL, Johnstone IM. Ideal spatial adaptation by wavelet shrinkage. Biometrika 1994;81(3): 425–455.
27. Breiman L. Better subset regression using the nonnegative garrote. Technometrics 1995;37(4):373–384.
28. Gao H. Wavelet shrinkage denoising using the nonnegative garrote. Journal of Computational and Graphical Statistics 1998;7:469–488.
29. Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 2006;101(476):1418–1429.
30. Figueiredo M, Nowak R. Wavelet-based image estimation: an empirical Bayes approach using Jeffreys' noninformative prior. IEEE Transactions on Image Processing 2001;10(9):1322–1331. [PubMed: 18255547]
31. Cohen, J. Statistical power for the behavioral sciences. 2. Hillsdale, NJ: Erlbaum; 1988.

32. Freedman B. Equipoise and the ethics of clinical research. The New England Journal of Medicine 1987;317(3):141–145. [PubMed: 3600702]
33. Wu, CFJ.; Hamada, M. Experiments: planning, analysis, and parameter design optimization. New York: Wiley; 2000.
34. Davison, AC.; Hinkley, DV. Bootstrap methods and their application. Cambridge, UK: Cambridge University Press; 1997.
35. Nankervis JC. Computational algorithms for double bootstrap confidence intervals. Computational Statistics & Data Analysis 2005;49:461–475.
36. Strecher V, McClure J, Alexander G, Chakraborty B, Nair V, et al. Web-based smoking cessation components and tailoring depth: results of a randomized trial. American Journal of Preventive Medicine 2008;34(5):373–381. [PubMed: 18407003]
37. Johnstone IM, Silverman BW. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. The Annals of Statistics 2004;32(4):1594–1649.
38. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics 1994;23:2379–2412.

Appendix A: Proof of Lemma 1

Proof

Define the *advantage* at stage j as

$$\mu_j(H_j, A_j)=Q_j(H_j, A_j) - \max_{a_j} Q_j(H_j, a_j),\ j=1, 2.$$

Note that $\mu_j(H_j, A_j)$ represents the expected difference in outcome when using A_j instead of the optimal treatment at stage j, for subjects with treatment and covariate history H_j who receive the optimal DTR at stages subsequent to j. According to Robins[11] (p. 201), this is simply the *blip* function with arg $\max_{a_j} Q_j(H_j, a_j)$ as the reference treatment. Below we will establish the connection between Q-learning and Robins' method using the advantage function; one can derive the connection using other blip functions (other choices of reference treatment) following similar steps. When Q-functions are modeled as in (2), the advantages become

$$\mu_j(H_j, A_j;\psi_j)=\psi_j^T H_{j1} A_j - |\psi_j^T H_{j1}|,\ j=1, 2. \tag{12}$$

Since by condition (i), no parameters are shared across stages, we will proceed stage by stage, starting with stage 2, doing recursive (rather than simultaneous) estimation. The notation $\mathbb{P}_n$ will be used below to denote the empirical average over a sample of size n. Also, define $m_1(H_1) = E[Q_1(H_1, A_1)|H_1]$ and $m_2(H_2) = E[Q_2(H_2, A_2)|H_2]$.

Stage 2

At stage 2, Q-learning is a usual least squares regression problem. Thus the estimating equations are given by

$$\mathbb{P}_n\left[\begin{pmatrix} H_{20} \\ H_{21}A_2 \end{pmatrix}(Y_2 - H_{20}^T\beta_2 - H_{21}^T A_2\psi_2)\right]=0. \tag{13}$$

From (13), it follows that

$$\widehat{\beta}_2=(\mathbb{P}_n(H_{20}H_{20}^T))^{-1}\left[\mathbb{P}_n(H_{20}Y_2)-\mathbb{P}_n(H_{20}H_{21}^TA_2)\widehat{\psi}_2\right] \tag{14}$$

where $\hat{\psi}_2$ is the estimate of ψ_2 satisfying (13). Thus $\hat{\psi}_2$ satisfies the estimating equation

$$\mathbb{P}_n\left[(H_{21}A_2)(Y_2-H_{20}^T\widehat{\beta}_2-H_{21}^TA_2\widehat{\psi}_2)\right]=0.$$

On the other hand, the stage 2 estimating equation for Robins' method (Robins[11], p. 211) is given by

$$\mathbb{P}_n\left[(H_{21}A_2-E[H_{21}A_2|H_2])\,(Y_2-\mu_2(H_2,A_2;\psi_2)-E[Y_2-\mu_2(H_2,A_2;\psi_2)|H_2])\right]=0, \tag{15}$$

where $Var(Y_2-\mu_2(H_2, A_2; \psi_2)-E[Y_2-\mu_2(H_2, A_2; \psi_2)|H_2]|H_2, A_2)$ is omitted (This is one of the reasons why Q-learning is an inefficient version). Note that $E[H_{21}A_2|H_2]=0$, by condition (ii) of the lemma. From (12), $\mu_2(H_2,A_2;\psi_2)=\psi_2^TH_{21}A_2-|\psi_2^TH_{21}|$. Then $E[\mu_2(H_2,A_2;\psi_2)|H_2]=-|\psi_2^TH_{21}|$, again by condition (ii). Also,

$$E[Y_2|H_2]=E[E[Y_2|H_2,A_2]|H_2]=E[Q_2(H_2,A_2)|H_2]=m_2(H_2).$$

Therefore, $Y_2-\mu_2(H_2,A_2;\psi_2)-E[Y_2-\mu_2(H_2,A_2;\psi_2)|H_2]=Y_2-m_2(H_2)-H_{21}^TA_2\psi_2$. Thus, $\hat{\psi}_2$ in Robins' method solves the following reduced version of (15):

$$\mathbb{P}_n\left[(H_{21}A_2)\left(Y_2-m_2(H_2)-H_{21}^TA_2\widehat{\psi}_2\right)\right]=0,$$

for any choice of $m_2(H_2)$ (with the conditional variance omitted). In particular, for $m_2(H_2)=H_{20}^T\widehat{\beta}_2$, where $\hat{\beta}_2$ is given by (14), this estimating equation exactly matches with that of Q-learning.

Stage 1

For Q-learning, the stage 1 pseudo-outcome is

$$\widehat{Y}_1=Y_1+\max_{a_2}Q_2(H_2,A_2)=Y_1+H_{20}^T\widehat{\beta}_2+|\widehat{\psi}_2^TH_{21}|,$$

and so the estimating equations are given by

$$\mathbb{P}_n\left[\begin{pmatrix}H_{10}\\H_{11}A_1\end{pmatrix}\left(Y_1+H_{20}^T\widehat{\beta}_2+|\widehat{\psi}_2^TH_{21}|-H_{10}^T\beta_1-H_{11}^TA_1\psi_1\right)\right]=0. \tag{16}$$

Now from (13)

$$\mathbb{P}_n\left[H_{20}\left(Y_2 - H_{20}^T\widehat{\beta}_2 - H_{21}^T A_2\widehat{\psi}_2\right)\right]=0. \tag{17}$$

Since by condition (iii) of the lemma, $(H_{10}^T, H_{11}^T A_1) \subset H_{20}^T$, it follows that

$$\begin{aligned}&\mathbb{P}_n\left[\begin{pmatrix} H_{10} \\ H_{11}A_1 \end{pmatrix}\left(Y_2 - H_{20}^T\widehat{\beta}_2 - H_{21}^T A_2\widehat{\psi}_2\right)\right]=0,\\ \text{or, } &\mathbb{P}_n\left[\begin{pmatrix} H_{10} \\ H_{11}A_1 \end{pmatrix}(H_{20}^T\widehat{\beta}_2)\right]=\mathbb{P}_n\left[\begin{pmatrix} H_{10} \\ H_{11}A_1 \end{pmatrix}(Y_2 - H_{21}^T A_2\widehat{\psi}_2)\right].\end{aligned} \tag{18}$$

Using (18) in (16), we get

$$\mathbb{P}_n\left[\begin{pmatrix} H_{10} \\ H_{11}A_1 \end{pmatrix}\left(Y_1+Y_2 - H_{21}^T A_2\widehat{\psi}_2+|\widehat{\psi}_2^T H_{21}| - H_{10}^T\beta_1 - H_{11}^T A_1\psi_1\right)\right]=0. \tag{19}$$

Solving for β_1 gives,

$$\widehat{\beta}_1=(\mathbb{P}_n(H_{10}H_{10}^T))^{-1}\left[\mathbb{P}_n\left(H_{10}(Y_1+Y_2 - H_{21}^T A_2\widehat{\psi}_2+|\widehat{\psi}_2^T H_{21}|)\right) - \mathbb{P}_n(H_{10}H_{11}^T A_1)\widehat{\psi}_1\right]. \tag{20}$$

Thus $\hat{\psi}_1$ satisfies

$$\mathbb{P}_n\left[(H_{11}A_1)\left(Y_1+Y_2 - H_{21}^T A_2\widehat{\psi}_2+|\widehat{\psi}_2^T H_{21}| - H_{10}^T\widehat{\beta}_1 - H_{11}^T A_1\widehat{\psi}_1\right)\right]=0.$$

On the other hand for Robins' method, the stage 1 pseudo-outcome (Robins11, p. 208; see also Moodie and Richardson17) is $\tilde{Y}_1 = Y_1 + Y_2 - \mu_2(H_2, A_2)$, and so the stage 1 estimating equation (Robins[11], p. 211) is given by

$$\mathbb{P}_n\left[(H_{11}A_1 - E[H_{11}A_1|H_1])\left(\tilde{Y}_1 - \mu_1(H_1, A_1;\psi_1) - E[\tilde{Y}_1 - \mu_1(H_1, A_1;\psi_1)|H_1]\right)\right]=0, \tag{21}$$

where again the conditional variance $Var(\tilde{Y}_1 - \mu_1(H_1, A_1; \psi_1) - E[\tilde{Y}_1 - \mu_1(H_1, A_1; \psi_1)|H_1]|H_1$, A_1 is omitted. Note that $E[H_{11}A_1|H_1] = 0$, by condition (ii) of the lemma. From (12), $\mu_1(H_1, A_1;\psi_1)=\psi_1^T H_{11}A_1 - |\psi_1^T H_{11}|$. Then $E[\mu_1(H_1, A_1;\psi_1)|H_1]= - |\psi_1^T H_{11}|$, again by condition (ii). Also,

$$
\begin{aligned}
E[\tilde{Y}_1|H_1] &= E[Y_1+Y_2-\mu_2(H_2,A_2)|H_1] \\
&= E[Y_2-Q_2(H_2,A_2)+Y_1+\max_{a_2} Q_2(H_2,a_2)|H_1] \\
&= E\left[E[Y-Q_2(H_2,A_2)|H_2,A_2]|H_1\right]+E[Y_1+\max_{a_2} Q_2(H_2,a_2)|H_1] \\
&= 0+E\left[E[Y_1+\max_{a_2} Q_2(H_2,a_2)|H_1,A_1]|H_1\right] \\
&= E[Q_1(H_1,A_1)|H_1] \\
&= m_1(H_1).
\end{aligned}
$$

Finally, plug in $Y_1+Y_2-\mu_2(H_2,A_2;\hat{\psi}_2)$ for $\tilde{Y}_1$. Thus, $\hat{\psi}_1$ in Robins' method solves the following reduced version of (21):

$$
\mathbb{P}_n\left[(H_{11}A_1)\left(Y_1+Y_2-H_{21}^T A_2\widehat{\psi}_2+|\widehat{\psi}_2^T H_{21}|-m_1(H_1)-H_{11}^T A_1\widehat{\psi}_1\right)\right]=0.
$$

for any choice of $m_1(H_1)$ (again omitting the conditional variance). In particular, for $m_1(H_1)=H_{10}^T\widehat{\beta}_1$, where $\hat{\beta}_1$ is given by (20), this estimating equation exactly matches with that of Q-learning.

In summary, the Q-learning algorithm as presented here is inefficient because: (a) it sets the conditional variances to be constant over (H_j, A_j), and (b) uses $H_{j1}A_j$ instead of the "efficient choice" of the term $S_{\text{eff};j}$ (that attains semiparametric variance bound) in Robins' estimating equation (see Robins11, p. 212; more details in Robins38).

Appendix B: Proof of Lemma 2

Proof

To estimate the hyper-parameter ϕ^2, first integrate out μ to get the marginal likelihood $X|\phi^2 \sim N(0, \phi^2+\sigma^2)$. The corresponding Jeffrey's prior on the variance parameter is $p(\phi^2) \propto 1=(\phi^2+\sigma^2)$. Based on this formulation, the posterior distribution of ϕ^2 is given by

$$
p(\phi^2|X) \propto (\phi^2+\sigma^2)^{-3/2}\exp\left\{-\frac{X^2}{2(\phi^2+\sigma^2)}\right\}.
$$

Hence the posterior mode of ϕ^2 is

$$
\widehat{\phi^2}=\underset{\phi^2\geq 0}{\operatorname{argmax}} p(\phi^2|X)=\left(\frac{X^2}{3}-\sigma^2\right)^+. \tag{22}
$$

Given $\phi^2=\widehat{\phi^2}$, now we will consider the data likelihood $X|\mu \sim N(\mu, \sigma^2)$ along with the prior $\mu|\phi^2 \sim N(0, \phi^2)$ to derive an empirical Bayes estimator for $|\mu|$. It is easy to show that the posterior distribution of μ is given by

$$\mu|X \sim N\left(\frac{X\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}, \frac{\sigma^2\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}\right). \tag{23}$$

Now under the squared error loss, the Bayes estimator of $|\mu|$ is $E_{\mu|X}(|\mu|)$ which can be calculated using (23). If $Y \sim N(\theta, \tau^2)$, then $E|Y|$ is given by:

$$E|Y| = \theta\,(2\Phi(\theta/\tau) - 1) + \sqrt{\frac{2}{\pi}}\tau\, e^{-\theta^2/2\tau^2}. \tag{24}$$

In the present problem,

$$Y = \mu|X, \quad \theta = \frac{X\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}, \quad \tau^2 = \frac{\sigma^2\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}.$$

$$\text{Hence,} \quad \frac{\theta}{\tau} = \frac{X}{\sigma}\sqrt{\frac{\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}}, \quad \frac{\theta^2}{2\tau^2} = \frac{X^2}{2\sigma^2}\left(\frac{\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}\right).$$

From (22), we get

$$\frac{\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2} = \frac{(X^2-3\sigma^2)^+}{X^2} = \left(1-\frac{3\sigma^2}{X^2}\right)^+,$$

$$\theta = X\left(1-\frac{3\sigma^2}{X^2}\right)^+, \quad \tau^2 = \sigma^2\left(1-\frac{3\sigma^2}{X^2}\right)^+,$$

$$\frac{\theta}{\tau} = \frac{X}{\sigma}\sqrt{\left(1-\frac{3\sigma^2}{X^2}\right)^+}, \quad \frac{\theta^2}{2\tau^2} = \frac{X^2}{2\sigma^2}\left(1-\frac{3\sigma^2}{X^2}\right)^+.$$

Thus an empirical Bayes estimator of $|\mu|$ is given by

$$\widehat{|\mu|}^{EB} = X\left(1-\frac{3\sigma^2}{X^2}\right)^+\left(2\Phi\left(\frac{X}{\sigma}\sqrt{\left(1-\frac{3\sigma^2}{X^2}\right)^+}\right)-1\right)$$
$$+\sqrt{\frac{2}{\pi}}\sigma\sqrt{\left(1-\frac{3\sigma^2}{X^2}\right)^+}\exp\left\{-\frac{X^2}{2\sigma^2}\left(1-\frac{3\sigma^2}{X^2}\right)^+\right\}. \tag{25}$$

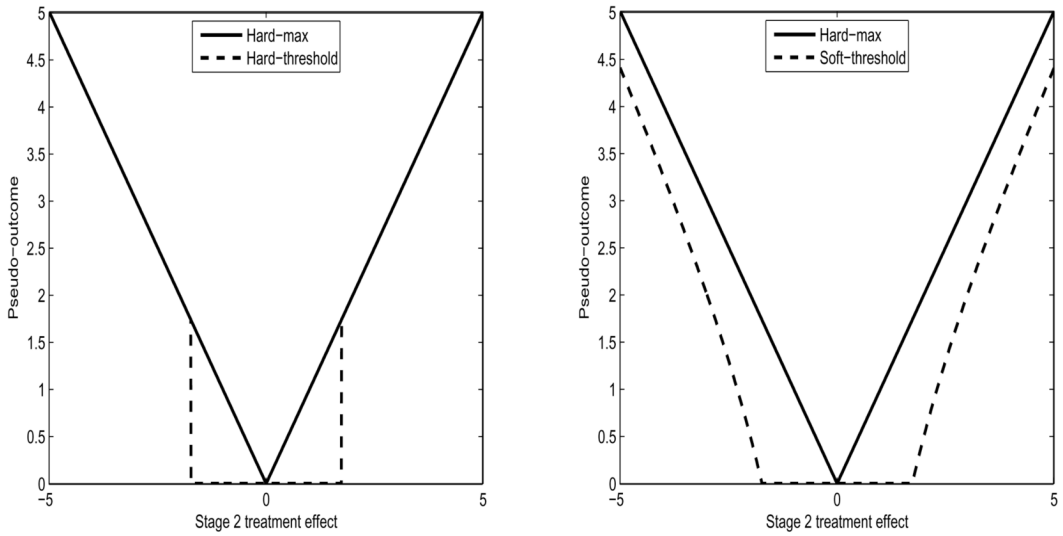

Figure 1.
Hard-threshold and Soft-threshold pseudo-outcomes compared with the Hard-max pseudo-outcome.

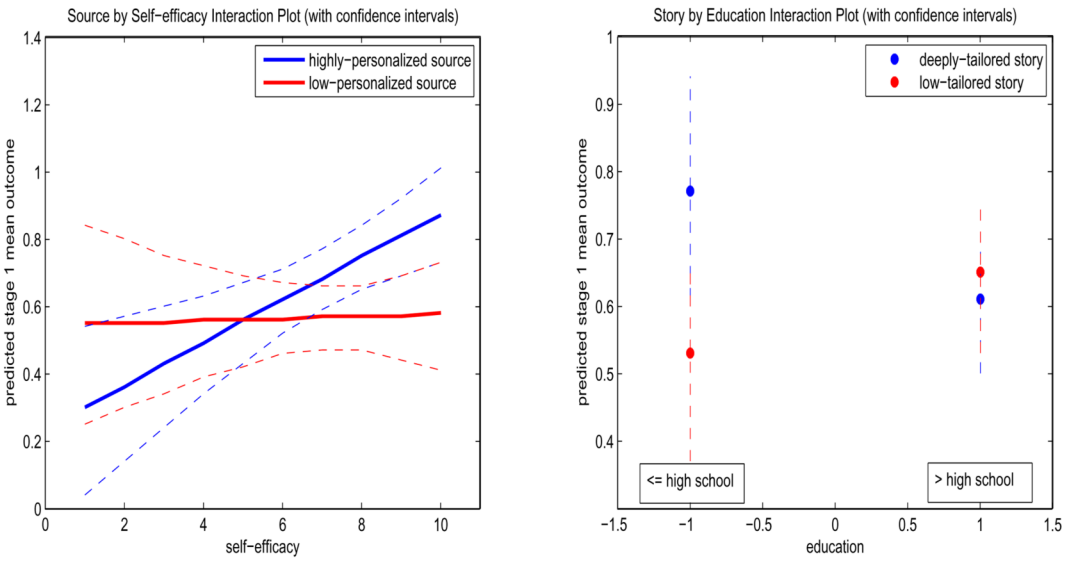

Figure 2.
Interaction plots: (a) source by self-efficacy (left panel), (b) story by education (right panel), along with confidence intervals for predicted stage 1 pseudo-outcome.

Table 1

Distribution of the linear combination $(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)$

(O_2, A_1) cell	cell probability (averaged over O_1)	value of the linear combination
(1, 1)	$q_1 \equiv \frac{1}{4}(expit(\delta_1+\delta_2)+expit(-\delta_1+\delta_2))$	$f_1 \equiv \gamma_5 + \gamma_6 + \gamma_7$
(1, −1)	$q_2 \equiv \frac{1}{4}(expit(\delta_1-\delta_2)+expit(-\delta_1-\delta_2))$	$f_2 \equiv \gamma_5 + \gamma_6 - \gamma_7$
(−1, 1)	$q_3 \equiv \frac{1}{4}(expit(\delta_1-\delta_2)+expit(-\delta_1-\delta_2))$	$f_3 \equiv \gamma_5 - \gamma_6 + \gamma_7$
(−1, −1)	$q_4 \equiv \frac{1}{4}(expit(\delta_1+\delta_2)+expit(-\delta_1+\delta_2))$	$f_4 \equiv \gamma_5 - \gamma_6 - \gamma_7$

Table 2

Summary statistics and coverage rates of 95% and 90% nominal percentile (PB), hybrid (HB), and double (DB) bootstrap CIs for ψ_{10} using the hard-max (HM), the hard-threshold with $\alpha = 0.08$ ($HT_{0.08}$) and $\alpha = 0.2$ ($HT_{0.20}$), and the soft-threshold (ST) estimators. A "*" indicates significantly different coverage rate than the nominal rate, using a test of proportion (Type I error rate = 0.05).

	Summary Statistics			Coverage of 95% CI			Coverage of 90% CI		
Estimator	**Bias**	**Var**	**MSE**	**PB**	**HB**	**DB**	**PB**	**HB**	**DB**
Example 1: $p = 1$ and ϕ undefined ($\psi_{10} = \psi_{11} = 0$)									
HM	0.0003	0.0045	0.0045	96.8*	93.5*	93.6	92.9*	88.2	88.8
$HT_{0.08}$	0.0017	0.0044	0.0044	97.0*	95.0	–	93.7*	90.3	–
$HT_{0.20}$	0.0002	0.0050	0.0050	97.4*	92.8*	–	94.2*	86.9*	–
ST	0.0009	0.0036	0.0036	95.3	96.1	–	91.1	91.4	–
Example 2: $p = 0$ and ϕ infinite ($\psi_{10} = \psi_{11} = 0$)									
HM	0.0003	0.0045	0.0045	96.7*	93.4*	93.6	92.4*	88.2	89.0
$HT_{0.08}$	0.0010	0.0044	0.0044	97.1*	95.3	–	94.0*	90.5	–
$HT_{0.20}$	0.0003	0.0050	0.0050	97.3*	93.5*	–	94.3*	87.1*	–
ST	0.0008	0.0036	0.0036	95.4	95.9	–	90.8	91.5	–
Example 3: $p = \frac{1}{2}$ and $\phi = 1$ ($\psi_{10} = \psi_{11} = 0$)									
HM	−0.0401	0.0059	0.0075	88.4*	92.7*	94.8	81.2*	86.1*	89.0
$HT_{0.08}$	−0.0083	0.0058	0.0059	94.3	94.3	–	88.5	89.0	–
$HT_{0.20}$	−0.0179	0.0062	0.0065	93.5*	93.5*	–	87.0*	88.1*	–
ST	−0.0185	0.0055	0.0058	93.4*	94.9	–	87.1*	89.4	–

Table 3

Summary statistics and coverage rates of 95% and 90% nominal percentile (PB), hybrid (HB), and double (DB) bootstrap CIs for ψ_{10} using hard-max (HM), hard-threshold with $\alpha = 0.08$ ($HT_{0.08}$) and $\alpha = 0.2$ ($HT_{0.20}$), and soft-threshold (ST) estimators. A "*" indicates significantly different coverage rate than the nominal rate, using a test of proportion (Type I error rate = 0.05).

	Summary Statistics			Coverage of 95% CI			Coverage of 90% CI		
Estimator	**Bias**	**Var**	**MSE**	**PB**	**HB**	**DB**	**PB**	**HB**	**DB**
Example 4: $p = 0$ and $\phi = 1.0204$ ($\psi_{10} = -0.01$, $\psi_{11} = 0$)									
HM	−0.0353	0.0059	0.0072	89.6*	93.1*	94.4	82.9*	86.6*	90.2
$HT_{0.08}$	−0.0037	0.0058	0.0058	94.6	94.1	–	88.9	89.0	–
$HT_{0.20}$	−0.0130	0.0062	0.0064	93.9	92.8*	–	87.9*	87.9*	–
ST	−0.0138	0.0055	0.0057	94.1	95.0	–	87.4*	89.7	–
Example 5: $p=\frac{1}{4}$ and $\phi = 1.4142$ ($\psi_{10} = \psi_{11} = 0$)									
HM	−0.0209	0.0069	0.0074	92.7*	93.1*	94.2	87.8*	89.0	88.4
$HT_{0.08}$	−0.0059	0.0070	0.0071	93.9	93.2*	–	89.5	88.2	–
$HT_{0.20}$	−0.0101	0.0072	0.0073	93.3*	93.0*	–	89.3	88.0*	–
ST	−0.0065	0.0069	0.0069	93.8	94.6	–	89.7	89.0	–
Example 6: $p = 0$ and $\phi = 0.3451$ ($\psi_{10} = -0.3688$, $\psi_{11} = 0.0187$)									
HM	0.0009	0.0067	0.0067	95.0	93.8	95.0	89.2	87.4*	88.2
$HT_{0.08}$	0.0003	0.0081	0.0081	95.1	88.5*	–	90.1	82.9*	–
$HT_{0.20}$	0.0011	0.0074	0.0074	94.8	91.2*	–	89.7	86.4*	–
ST	0.0052	0.0074	0.0074	94.8	91.7*	–	89.4	85.3*	–

Table 4

Regression coefficients and 95% hybrid bootstrap confidence intervals at stage 1, using both the hard-max and the soft-threshold estimators.

	Hard-max		Soft-threshold	
Variable	**Coefficient**	**95% CI**	**Coefficient**	**95% CI**
motivation	0.04	(−0.00, 0.08)	0.04	(0.00, 0.08)
selfefficacy	0.03	(0.00, 0.06)	0.03	(0.00, 0.06)
education	−0.01	(−0.07, 0.06)	−0.01	(−0.07, 0.06)
source	−0.15	(−0.35, 0.06)	−0.15	(−0.35, 0.06)
source*selfefficacy	0.03	(0.00, 0.06)	0.03	(0.00, 0.06)
story	0.05	(−0.01, 0.11)	0.05	(−0.01, 0.11)
story*education	−0.07	(−0.13, −0.01)	−0.07	(−0.13, −0.01)