Bias Correction and Confidence Intervals for Fitted Q-iteration

Bibhas Chakraborty
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
bibhas@umich.edu

Victor Strecher
School of Public Health
University of Michigan
Ann Arbor, MI 48109
strecher@umich.edu

Susan Murphy
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
samurphy@umich.edu

Abstract

We consider finite-horizon fitted Q-iteration with linear function approximation to learn a policy from a training set of trajectories. We show that fitted Q-iteration can give biased estimates and invalid confidence intervals for the parameters that feature in the policy. We propose a regularized estimator called *soft-threshold* estimator, derive it as an approximate empirical Bayes estimator, and show that it reduces bias and improves the coverage rates of confidence intervals via simulated experiments. We also demonstrate the use of this method in the analysis of data from a randomized smoking cessation study.

1 Introduction

The problem of learning a good treatment policy from a training set of finite-horizon trajectories generated by a known behavior policy often arises in medical applications [1, 2, 3, 4]. Characteristics of these applications are poorly understood system dynamics, possibly high-dimensional state space but finite action space, non-Markovian state transitions, a very small number (say, $2-4$) of stages, and limited amount of data. In these medical settings, *batch* reinforcement learning algorithms [5], where the agent is trained using a series of previously collected trajectories generated by a known behavior policy consisting of state, action, and reward information, are useful. In particular, here we consider the fitted Q-iteration algorithm [6]; the argument of the maximum of the fitted Q-function gives the estimated optimal policy. However, since maximization is a non-smooth operation, it often induces bias in estimates and leads to invalid measures of confidence. Constructing measures of confidence for the parameters involved in the policy is important in the medical applications due to the following reasons. First, if confidence intervals for certain parameters reveal that there is no evidence that these parameters are different from zero, then the corresponding state information need not be collected in future implementations of the policy. This is particularly important when data collection is expensive and time-consuming. Second, it is important to know when there is insufficient support in the data to recommend one treatment (action) over another, since in such cases treatment can be chosen according to other considerations like cost, familiarity, burden, preference etc.

In this paper, we use linear approximators to fit the Q-functions, and illustrate the problems with bias and measures of confidence. However these problems are unrelated to the choice of the approximator, and are present in case of other function approximations as well. To overcome the above problems, we propose an easy-to-compute *soft thresholding* operation instead of the *hard-max* operation in the fitted Q-iteration algorithm. The soft-threshold estimator is an approximate empirical Bayes estimator. We compare the hard-max estimator with the soft-threshold estimator via simulated experiments, and demonstrate its use in analysis of data from a smoking cessation study.

1.1 Preliminaries

We consider a finite horizon decision process. At the t-th stage ($1 \leq t \leq K < \infty$), there is an *observation* O_t that takes values in an observation space $\mathcal{O}$, which is followed by an *action* A_t that takes values in a finite, discrete action space $\mathcal{A}$. Define the history available at the t-th stage as $H_t = (\mathbf{O}_t, \mathbf{A}_{t-1})$, where $\mathbf{O}_t = \{O_1, \ldots, O_t\}$ and $\mathbf{A}_t = \{A_1, \ldots, A_t\}$. This H_t defines the *state*. *Rewards* are functions of the history and action, e.g., $R_t \equiv R_t(h_t, a_t)$. The training set is composed of n finite horizon trajectories, each of the form $\{o_1, a_1, r_1, \ldots, o_K, a_K, r_K\}$. Each trajectory in the training set is generated by some stationary, stochastic policy π_b: $A_t \sim \pi_b(\cdot|H_t)$, where $\pi_b(\cdot|h)$ is a density with $\inf_{(h,a)} \pi_b(a|h) > 0$. For simplicity, in the following we only consider $K = 2$, and binary actions at each stage, coded $\{-1, 1\}$, assigned randomly in the training data.

2 Fitted Q-iteration with Linear Regression

Here we consider the fitted Q-iteration algorithm with linear regression using least squares [6] for a finite-horizon, non-discounted problem. Let the stage-t ($t = 1, 2$) Q-function be modeled as

$$Q_t(H_t, A_t; \beta_t, \psi_t) = \beta_t^T H_{t0} + (\psi_t^T H_{t1}) A_t, \tag{1}$$

where H_{t0} and H_{t1} are two (possibly different) basis functions of the state (history) H_t, with H_{t0} denoting the main effect of state and H_{t1} denoting the part of state information that interacts with action (H_{t0} and H_{t1} include the intercept term). We have separated these two parts because only the second term features in the policy. Thus even though we estimate all the parameters from the training data, our main interest lies in the policy parameters ψ_t's only (β_t's are nuisance parameters). Suppose there are n trajectories. For $K = 2$, the algorithm goes as follows:

1. $(\hat{\beta}_2, \hat{\psi}_2) = \arg\min_{\beta_2, \psi_2} \frac{1}{n} \sum_{i=1}^n \Big(R_{2i} - Q_2(H_{2i}, A_{2i}; \beta_2, \psi_2)\Big)^2$.
2. $\hat{Y}_{1i} \leftarrow R_{1i} + \max_{a \in \mathcal{A}} Q_2(H_{2i}, a; \hat{\beta}_2, \hat{\psi}_2),\ i = 1, \ldots, n.$
3. $(\hat{\beta}_1, \hat{\psi}_1) = \arg\min_{\beta_1, \psi_1} \frac{1}{n} \sum_{i=1}^n \Big(\hat{Y}_{1i} - Q_1(H_{1i}, A_{1i}; \beta_1, \psi_1)\Big)^2$.

The estimated optimal policy $\hat{\pi}_t$ satisfies $\hat{\pi}_t(h_t) \in \arg\max_a \ Q_t(h_t, a; \hat{\beta}_t, \hat{\psi}_t), \forall t$. Using model (1), it follows that $\hat{\pi}_t(h_t) = sign(\hat{\psi}_t^T H_{t1}), \forall t$, which does not depend on $\hat{\beta}_t$.

Bias in Estimation
One important issue regarding the above algorithm is that at stage-1, the "target" function $\hat{Y}_1$ becomes random; this function depends on the same data that is used in the subsequent least squares regression. Also, note that in the expression for $\hat{Y}_1$, the second term $\max_{a \in \mathcal{A}} Q_2(H_{2i}, a; \hat{\beta}_2, \hat{\psi}_2)$ is a plug-in estimate of $\max_{a \in \mathcal{A}} Q_2(H_{2i}, a; \beta_2, \psi_2)$. However, since $\max$ is a non-linear, non-smooth operation,

$$E[\max_{a \in \mathcal{A}} Q_2(H_{2i}, a; \hat{\beta}_2, \hat{\psi}_2)] \neq \max_{a \in \mathcal{A}} Q_2(H_{2i}, a; \beta_2, \psi_2).$$

Thus the plug-in estimate is biased, even if Q_2 is unbiased. Since the subsequent step in the algorithm uses these plug-in estimates, this bias propagates as one moves backward in time.

The issue of bias can be better understood with a simple toy example. Consider the problem of estimating $|\mu|$ based on n i.i.d. observations $X_1, \ldots, X_n$ from a normal distribution $N(\mu, 1)$. Note that $|\mu|$ is a non-smooth function of μ (non-differentiable at $\mu = 0$). Here $|\bar{X}_n|$ is the maximum likelihood estimator of $|\mu|$, where $\bar{X}_n$ is the sample average. It can be shown that $\lim_{n \to \infty} E[\sqrt{n}(|\bar{X}_n| - |\mu|)] = \sqrt{\frac{2}{\pi}}$ for $\mu = 0$, but $\lim_{n \to \infty} E[\sqrt{n}(|\bar{X}_n| - |\mu|)] = 0$ for $\mu \neq 0$. Thus at the point of non-differentiability $\mu = 0$, the estimator $|\bar{X}_n|$ has a bias of magnitude $\sqrt{\frac{2}{n\pi}}$.

Coming back to our original context of finite-horizon fitted Q-iteration, we see that the problem of bias should be observed in the stage-1 estimator $\hat{\psi}_1$. $\hat{Y}_1$ in step-2 of the algorithm can be written as

$$\hat{Y}_{1i} = R_{1i} + \max_a \ Q_2(H_{2i}, a; \hat{\beta}_2, \hat{\psi}_2) = R_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}|;\ i = 1, \ldots, n. \tag{2}$$

This is a non-smooth (e.g., non-differentiable at $\hat{\psi}_2^T H_{21,i} = 0$) function of $\hat{\psi}_2$, because of the maximization operation (notice the similarity of the estimator $|\hat{\psi}_2^T H_{21,i}|$ to $|\bar{X}_n|$ in the above toy example). Now since the stage-1 policy parameter estimate $\hat{\psi}_1$ is a function of $\hat{Y}_{1i}$, $i = 1, \ldots, n$, it is in turn a non-smooth function of $\hat{\psi}_2$. Thus any bias occurring in the estimation of the target function $\hat{Y}_{1i}$ can potentially affect the stage-1 policy parameter estimate $\hat{\psi}_1$. We will illustrate the occurrence of bias in the simulated experiments to follow.

Confidence Intervals

In a regression set-up, usual confidence intervals are constructed by assuming that the estimators are approximately normally distributed. When this assumption does not hold, problems occur. The confidence intervals for the policy parameters ψ_t's can go wrong in two ways. First, the bias described above can make the confidence intervals wrongly centered. As a result, confidence intervals can have significantly different coverage rate than their intended coverage probabilities. Second, due to the underlying lack of smoothness, the distribution of the estimated parameter $\hat{\psi}_t$ can get distorted. To understand this, consider again $\hat{Y}_{1i}$ in (2). As mentioned earlier, this is a non-smooth function of $\hat{\psi}_2$. As a consequence, the approximate distribution of $\sqrt{n}(\hat{\psi}_1 - \psi_1)$ is normal if the distribution of the state H_2 is such that $p \equiv P[\psi_2^T H_{21} = 0] = 0$, but is non-normal if $p > 0$. This change in the distribution happens abruptly. This phenomenon is referred to as *nonregularity* [7]. Note that the nonregularity occurs at and around the point of non-differentiability ($\hat{\psi}_2^T H_{21,i} = 0$). Also, $\hat{\psi}_2^T H_{21,i} \approx 0$ means that the actions do not significantly affect the value of the Q-function. This frequently occurs in medical applications, since usually the effects of the treatments used in a clinical trial are not very different due of ethical considerations. In a general reinforcement learning problem, if the actions are "not too different", this problem should arise as well. Because of this nonregularity, given the noise level present in small training data sets, the estimator $\hat{\psi}_1$ oscillates between the normal and non-normal distributions across training data sets. This causes the confidence intervals to have poor frequentist properties.

3 Soft-threshold Fitted Q-iteration

In this section, we present a way to reduce the nonregularity of the "hard-max" estimator $\hat{\psi}_1$, by shrinking (thresholding) the effect of the term involving maximization (e.g., $|\hat{\psi}_2^T H_{21}|$) towards zero. We call this estimator the *soft-threshold estimator*. As before, consider two stages ($K = 2$) with binary actions and linear approximations to the Q-functions. The form of the soft-threshold target function is

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

where $x^+ = x\mathbf{1}\{x > 0\}$ is the positive part of a function, and λ_i is a tuning parameter associated with the i-th subject in the training data (possibly depending on the variability of the linear combination $\hat{\psi}_2^T H_{21,i}$ for that subject). In the context of regression shrinkage [8], the second term in (3) is generally known as the *nonnegative garrote* estimator. Note that like the hard-max target function $\hat{Y}_1$, the soft-threshold target function $\hat{Y}_1^{ST}$ is also a non-smooth function of $\hat{\psi}_2$ and hence $\hat{\psi}_1^{ST}$ remains a nonregular estimator of ψ_1. However, the problematic term $|\hat{\psi}_2^T H_{21}|$ is shrunk (or, thresholded) towards zero, and hence we expect that the degree of nonregularity will be reduced. In the simulated experiments to follow, we will investigate how much improvement, if any, this estimator offers over the hard-max estimator, in terms of bias correction and constructing confidence intervals. Figure 1 presents the hard-max and the soft-threshold functions. A crucial issue here is how to choose a data-driven tuning parameter λ_i. Below we provide a Bayesian approach for choosing λ_i's, inspired by the work of Figueiredo and Nowak [9] in the area of wavelets. We will evaluate this choice based on bias correction as well as for constructing valid confidence intervals.

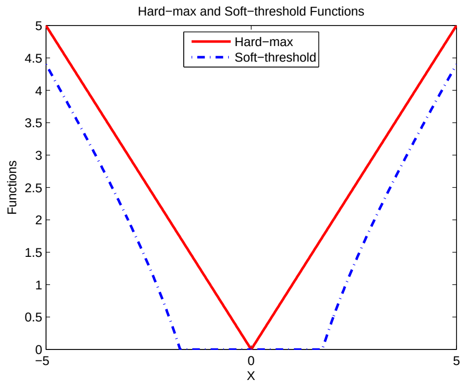

Figure 1: Soft-threshold function compared with the Hard-max function.

3.1 Choice of Tuning Parameter

It can be shown that the estimator (3) with $\lambda_i = 3H_{21,i}^T\hat{\Sigma}_2 H_{21,i}/n,\ i = 1, \ldots, n,$, where $\hat{\Sigma}_2/n$ is the estimated covariance matrix of $\hat{\psi}_2$ is as an approximate empirical Bayes estimator (see below for derivation). Thus the soft-threshold target function (3) becomes

$$\begin{aligned} \hat{Y}_{1i}^{ST} &= \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \left(1 - \frac{3H_{21,i}^T \hat{\Sigma}_2 H_{21,i}}{n|\hat{\psi}_2^T H_{21,i}|^2}\right)^+, \qquad (4) \\ &= \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \left(1 - \frac{3H_{21,i}^T \hat{\Sigma}_2 H_{21,i}}{n|\hat{\psi}_2^T H_{21,i}|^2}\right) \cdot \mathbf{1}\left\{\frac{\sqrt{n}|\hat{\psi}_2^T H_{21,i}|}{\sqrt{H_{21,i}^T \hat{\Sigma}_2 H_{21,i}}} > \sqrt{3}\right\}, \\ & \qquad i = 1, \ldots, n. \qquad (5) \end{aligned}$$

The presence of the indicator function in (5) indicates that $\hat{Y}_{1i}^{ST}$ is a thresholding rule for small values of $|\hat{\psi}_2^T H_{21,i}|$, while the term just preceding the indicator function makes $\hat{Y}_{1i}^{ST}$ a shrinkage rule for moderate to large values of $|\hat{\psi}_2^T H_{21,i}|$ (for which the indicator function takes the value one). The following lemma will be used to derive the choice of λ_i. This lemma follows from the Bayesian framework used in [9].

Lemma 1. *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*

$$\begin{aligned} \widehat{|\mu|}^{EB} &= X\left(1 - \frac{3\sigma^2}{X^2}\right)^+ \left(2\Phi\left(\frac{X}{\sigma}\sqrt{\left(1 - \frac{3\sigma^2}{X^2}\right)^+}\right) - 1\right) \\ &+ \sqrt{\frac{2}{\pi}}\,\sigma\sqrt{\left(1 - \frac{3\sigma^2}{X^2}\right)^+} \exp\left\{-\frac{X^2}{2\sigma^2}\left(1 - \frac{3\sigma^2}{X^2}\right)^+\right\}. \qquad (6) \end{aligned}$$

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

$$\widehat{\phi^2} = \arg\max_{\phi^2 \geq 0} p(\phi^2|X) = \left(\frac{X^2}{3} - \sigma^2\right)^+. \qquad (7)$$

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

$$\mu|X \sim N\left(\frac{X\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}, \frac{\sigma^2\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}\right). \tag{8}$$

Now under the squared error loss, the Bayes estimator of $|\mu|$ is $E_{\mu|X}(|\mu|)$ which can be calculated using (8). If $Y \sim N(\theta, \tau^2)$, then $E|Y|$ is given by:

$$E|Y| = \theta\Big(2\Phi(\theta/\tau) - 1\Big) + \sqrt{\frac{2}{\pi}}\tau\, e^{-\theta^2/2\tau^2}. \tag{9}$$

In the present problem,

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

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

From (7), we get

$$\begin{aligned}
\frac{\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2} &= \frac{(X^2-3\sigma^2)^+}{X^2} = \Big(1-\frac{3\sigma^2}{X^2}\Big)^+,\\
\theta &= X\,\Big(1-\frac{3\sigma^2}{X^2}\Big)^+,\\
\tau^2 &= \sigma^2\,\Big(1-\frac{3\sigma^2}{X^2}\Big)^+,\\
\frac{\theta}{\tau} &= \frac{X}{\sigma}\,\sqrt{\Big(1-\frac{3\sigma^2}{X^2}\Big)^+},\\
\frac{\theta^2}{2\tau^2} &= \frac{X^2}{2\sigma^2}\,\Big(1-\frac{3\sigma^2}{X^2}\Big)^+.
\end{aligned}$$

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

$$\begin{aligned}
\widehat{|\mu|}^{EB} \quad &= \quad X\,\Big(1-\frac{3\sigma^2}{X^2}\Big)^+\left(2\Phi\Big(\frac{X}{\sigma}\,\sqrt{\Big(1-\frac{3\sigma^2}{X^2}\Big)^+}\Big)-1\right)\\
&+ \quad \sqrt{\frac{2}{\pi}}\,\sigma\,\sqrt{\Big(1-\frac{3\sigma^2}{X^2}\Big)^+}\,\exp\left\{-\frac{X^2}{2\sigma^2}\,\Big(1-\frac{3\sigma^2}{X^2}\Big)^+\right\}.
\end{aligned} \tag{10}$$

□

Clearly, $\widehat{|\mu|}^{EB}$ is a thresholding rule, since $\widehat{|\mu|}^{EB} = 0$ for $|X| < \sqrt{3}\sigma$. Moreover, the second term of (6) goes to zero exponentially fast as $|\frac{X}{\sigma}|$ becomes large, and

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

Consequently, the empirical Bayes estimator is approximated by

$$\widehat{|\mu|}^{EB} \approx X\,\Big(1-\frac{3\sigma^2}{X^2}\Big)^+\,sign(X) = |X|\,\Big(1-\frac{3\sigma^2}{X^2}\Big)^+. \tag{11}$$

Now for $i = 1, \ldots, n$ separately, put $X = \hat{\psi}_2^T H_{21,i}$, and $\mu = \psi_2^T H_{21,i}$ (for fixed $H_{21,i}$); and plug in $\hat{\sigma}^2 = H_{21,i}^T \hat{\Sigma}_2 H_{21,i}/n$ for σ^2. This precisely leads to the second term of the soft-threshold pseudo-outcome (3) with $\lambda_i = 3H_{21,i}^T \hat{\Sigma}_2 H_{21,i}/n$. Thus the current Bayesian formulation gives us a data-driven choice of the tuning parameters.

Table 1: Summary statistics and coverage rates of 95% bootstrap confidence intervals for ψ_{10} using the hard-max and the soft-threshold estimators. A "*" indicates significantly different coverage rate than the intended rate.

Example	p	ψ_{10}	Estimator	Bias	Var	MSE	Coverage of 95% CI
1.	1	0	hard-max	0.0003	0.0045	0.0045	93.5*
			soft-threshold	0.0009	0.0036	0.0036	96.1
2.	0	0	hard-max	0.0003	0.0045	0.0045	93.4*
			soft-threshold	0.0008	0.0036	0.0036	95.9
3.	$\frac{1}{2}$	0	hard-max	-0.0401	0.0059	0.0075	92.7*
			soft-threshold	-0.0185	0.0055	0.0058	94.9
4.	$\frac{1}{4}$	0	hard-max	-0.0209	0.0069	0.0074	93.1*
			soft-threshold	-0.0065	0.0069	0.0069	94.6
5.	0	-0.3688	hard-max	0.0009	0.0067	0.0067	93.8
			soft-threshold	0.0052	0.0074	0.0074	91.7*

4 Simulation Experiments

The experiments considered here are simulated using simple generative models. The methods learn the optimal policy using a single training set of size $n = 300$ (a realistic choice in medical applications). We consider a two-stage ($K = 2$) problem, where each trajectory is of the form $(O_1, A_1, R_1, O_2, A_2, R_2)$. Actions are always randomized, e.g., $P[A_t = 1] = P[A_t = -1] = \frac{1}{2}$, $t = 1, 2$. Also, $P[O_1 = 1] = P[O_1 = -1] = \frac{1}{2}$. The variable O_2 is generated as

$$P[O_2 = 1|O_1, A_1] = 1 - P[O_2 = -1|O_1, A_1] = \frac{\exp(\delta_1 O_1 + \delta_2 A_1)}{1 + \exp(\delta_1 O_1 + \delta_2 A_1)},$$

where the parameters δ_1, δ_2 are varied. In the following examples, we always set $R_1 = 0$; however R_2 is varied (so there is only a terminal reward, but no intermediate reward). Also, the parameter $p = P[\psi_2^T H_{21} = 0]$ that governs the underlying nonregularity (see section 2, under confidence intervals) is varied. It takes values $1,\ 0,\ \frac{1}{2}, 0$, and $\frac{1}{4}$ in examples 1-5 respectively.

Example 1: $\delta_1 = \delta_2 = 0.5$, and $R_2 \sim N(0, 1)$.
Example 2: $\delta_1 = \delta_2 = 0.5$, and $R_2 = 0.01A_2 + \epsilon,\ \epsilon \sim N(0, 1)$.
Example 3: $\delta_1 = \delta_2 = 0.5$, and $R_2 = -0.5A_1 + 0.5A_2 + 0.5A_1A_2 + \epsilon,\ \epsilon \sim N(0, 1)$.
Example 4: $\delta_1 = 1,\ \delta_2 = 0$, and $R_2 = -0.5A_1 + A_2 + 0.5O_2A_2 + 0.5A_1A_2 + \epsilon,\ \epsilon \sim N(0, 1)$.
Example 5: $\delta_1 = \delta_2 = 0.1$, and $R_2 = -0.5A_1 + 0.25A_2 + 0.5O_2A_2 + 0.5A_1A_2 + \epsilon,\ \epsilon \sim N(0, 1)$.

Example 1 describes a fully nonregular setting where there is no effect of action at any stage. Example 2 is regular, but is very close to the nonregular Example 1 with a very tiny stage-2 effect (0.01). Example 3 is a nonregular setting where the stage-2 effect is zero for half the subjects in the population. For Example 4, the stage-2 effect is zero for one-fourth of the subjects in the population. Example 5 is a highly regular setting where stage-2 effect sizes are reasonably large. The analysis model is given by (1), and it uses $H_{10} = H_{11} = (1, O_1)$, $H_{20} = (1, O_1, A_1, O_1A_1)$, $H_{21} = (1, O_2, A_1)$. We generated 1000 training data sets each of size $n = 300$, analyzed them by both the hard-max (original) and the soft-threshold fitted Q-iteration, and calculated summary statistics (bias, variance, and mean squared error) for the estimated parameters ($\hat{\psi}_{10},\ \hat{\psi}_{11}$). Also, for each data set, we constructed 95% hybrid bootstrap confidence intervals (based on 1000 bootstrap iterations) for the ψ parameters, and checked using the generative model whether or not the confidence interval captured the true value of the corresponding parameter. Averaging across all the 1000 data sets, we found the Monte Carlo estimate of the coverage probabilities of the confidence intervals. Although we considered both the policy parameters ψ_{10} and ψ_{11} at stage-1 in the analysis, it turned out that the influence of nonregularity is more pronounced on the parameter ψ_{10}; so in Table 1 and subsequent discussion, we focus on ψ_{10} only.

In all these examples except Example 5 (see Table 1), the hard-max estimator performs poorly in terms of coverage rate and/or bias. In Examples 1 and 2, the hard-max estimator does not suffer from bias, yet it gives invalid coverage rate. This is because the shape of the distribution of $\hat{\psi}_{10}$ is distorted due to nonregularity. The soft-threshold estimator corrects this. In examples 3-4, the soft-threshold estimator reduces bias and the mean squared error (a measure of risk), and offers coverage rates closer to the intended coverage probabilities of the confidence intervals. Even though Example 2 is a regular setting ($p = 0$), this is close enough to a nonregular setting (Examples 1), and hence the hard-max estimator shows problems. Example 5 is a highly regular setting (less likely to occur in medical applications) where hard-max has no problem; here soft-threshold estimator induces some bias and shows problem with coverage.

5 Analysis of Smoking Cessation Data

To demonstrate the use of the soft-threshold method, we applied it to the data from a randomized two-stage web-based longitudinal smoking cessation study. The stage-1 of this study (Project Quit) was conducted to find an optimal multi-factor treatment to help adult smokers quit smoking; and the stage-2 (Forever Free) was a follow-on study to help those 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 stage-1 data can be found in [10]. In stage-1, although there were five two-level treatment factors, only two of them, e.g., `source` and `story` came out significant in the analysis reported in [10]. We consider only these two treatment factors at stage-1 of our present analysis, which gives us 4 action choices corresponding to the 2 × 2 design. Baseline variables (state variables) at this stage include subjects' `motivation` to quit (on a 1-10 scale), `selfefficacy` (on a 1-10 scale) and `education` (binary). The outcome (reward) at this stage is binary quit status, e.g., `PQ6Quitstatus` (1=quit, 0=not quit) at 6 month from the date of randomization. In stage-2, originally there were 4 different treatment groups and a control group; however due to small sample sizes resulting from subject dropout and little difference between the treatment groups, the 4 treatment groups were combined together for the present analysis. This resulted in only two action choices at stage-2; this variable is called `FFarm` and is coded $\{-1, 1\}$. The only state variable used at the second stage is the stage-1 `PQ6Quitstatus`. The stage-2 outcome (reward) is again binary quit status `FF6Quitstatus` at 6 month from the date of randomization to stage-2.

To find the optimal treatment policy, we applied both the hard-max and the soft-threshold methods within the fitted Q-iteration framework. This involved: (1) a stage-2 regression of `FF6Quitstatus` on `PQ6Quitstatus` and `FFArm`, allowing for interation (n = 283); (2) finding a stage-1 target function using the hard-max method, and another target function using the soft-threshold method; and (3) for each of the two target functions, a stage-1 regression of the target function on `motivation`, `source` and `selfefficacy` (allowing for interaction), and `story` and `education` (allowing for interaction) (n = 1401). For either method, confidence intervals were constructed by hybrid bootstrap method using 1000 bootstrap replications. At the stage-2, there was no significant treatment effect. Stage-1 regression summaries are presented in Table 2. Both methods produced similar results. But as we have seen in the simulation experiments, when there is no stage-2 effect (just like the present data), the soft-threshold estimator should be preferred – for correcting any bias in the estimates, as well as providing valid confidence intervals. Note that both the hard-max and the soft-threshold methods suggest that the interaction `source:selfefficacy` is significant. Thus in future implementation of the policy, treatments can be individually tailored to the subject's measurement on `selfefficacy`. On the other hand, the main effect of `education` as well as its interaction with `story` are insignificant; this suggests that one need not collect data on subject's education in future implementation of the policy.

6 Discussion

We have shown that the fitted Q-iteration with linear regression provides biased estimates and invalid confidence intervals for the policy parameters, which can be corrected by our proposed soft-threshold estimator. Although we have derived it as an approximate empirical Bayes estimator, no theoretical result about its optimality is known at this point. It would be interesting to investigate if and to what extent the above problems occur in more sophisticated function approximators (e.g., neural network, decision tree etc.), and in other (more standard) reinforcement learning problems.

Table 2: Summary of regression coefficients and bootstrap confidence intervals for the stage-1 variables using both the hard-max and the soft-threshold estimators.

Variable	Method	Coefficient	95% Bootstrap Confidence Interval
`Intercept`	hard-max	0.2599	(0.0662, 0.4442)
	soft-threshold	0.2453	(0.0393, 0.4399)
`motivation`	hard-max	0.0272	(0.0015, 0.0531)
	soft-threshold	0.0271	(0.0000, 0.0513)
`source`	hard-max	-0.1033	(-0.2487, 0.0386)
	soft-threshold	-0.1027	(-0.2421, 0.0357)
`selfefficacy`	hard-max	0.0238	(0.0010, 0.0463)
	soft-threshold	0.0237	(0.0026, 0.0462)
`story`	hard-max	0.0433	(0.0018, 0.0858)
	soft-threshold	0.0430	(0.0019, 0.0877)
`education`	hard-max	-0.0216	(-0.0616, 0.0193)
	soft-threshold	-0.0215	(-0.0616, 0.0248)
`source:selfefficacy`	hard-max	0.0196	(0.0009, 0.0385)
	soft-threshold	0.0194	(0.0023, 0.0380)
`story:education`	hard-max	-0.0202	(-0.0631, 0.0201)
	soft-threshold	-0.0201	(-0.0578, 0.0234)

Acknowledgments

We acknowledge support for this project from National Institute of Health grants RO1 MH080015, P50 DA10075, and P50 CA 101451.

References

[1] L.S. Schneider, P.N. Tariot, C.G. Lyketsos, K.S. Dagerman, K.L. Davis, and S. Davis. National institute of mental health clinical antipsychotic trials of intervention effectiveness (catie): Alzheimer disease trial methodology. *American Journal of Geriatric Psychology*, 9:346–360, 2001.

[2] M. Fava, A.J. Rush, and et al. Background and rationale for the sequenced treatment alternative to relieve depression (star*d) study. *Psychiatric Clinics of North America*, 26(3):457 – 494, 2003.

[3] J. Pineau, M. Bellemare, A.J. Rush, A. Ghizaru, and S. Murphy. Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence*, 88(Supplement 2):S52–S60, 2007.

[4] A. Guez, R. Vincent, M. Avoli, and J. Pineau. Adaptive treatment of epilepsy via batch-mode reinforcement learning. *Innovative Applications of Artificial Intelligence*, 2008.

[5] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, 1998.

[6] A. Antos, R. Munos, and C. Szepesvari. Fitted q-iteration in continuous action-space mdps. *Neural Information Processing Systems*, 2007.

[7] J.M. Robins. Optimal structural nested models for optimal sequential decisions. In D.Y. Lin and P. Heagerty, editors, *Proceedings of the Second Seattle Symposium on Biostatistics*, pages 189–326, New York, 2004. Springer.

[8] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373 – 384, 1995.

[9] M.A.T. Figueiredo and R.D. Nowak. Wavelet-based image estimation: An empirical bayes approach using jeffreys' noninformative prior. *IEEE Transactions on Image Processing*, 10(9):1322 – 1331, 2001.

[10] V. Strecher, J. McClure, G. Alexander, B. Chakraborty, V. Nair, and et al. Web-based smoking cessation components and tailoring depth: Results of a randomized trial. *American Journal of Preventive Medicine*, 34(5):373 – 381, 2008.