IMS Collections
From Probability to Statistics and Back: High-Dimensional Models and Processes
Vol. 9 (2013) 150–162
© Institute of Mathematical Statistics, 2013
DOI: 10.1214/12-IMSCOLL911

Adaptive Q-learning*

Yair Goldberg[1], **Rui Song**[2,†] **and Michael R. Kosorok**[3,‡]

University of Haifa, North Carolina State University and
University of North Carolina at Chapel Hill

Abstract: Developing an effective multi-stage treatment strategy over time is one of the essential goals of modern medical research. Developing statistical inference, including constructing confidence intervals for parameters, is of key interest in studies applying dynamic treatment regimens. Estimation and inference in this context are especially challenging due to non-regularity caused by the non-smoothness of the problem in the parameters. While various bootstrap methods have been proposed, there is a lack of theoretical validation for most bootstrap inference methods. Recently, Song et al. [Penalized Q-learning for dynamic treatment regimes (2011) Submitted] proposed the penalized Q-learning procedure, that enables valid inference without the need of bootstrapping. As a major drawback, penalized Q-learning can only handle discrete covariates. To overcome this issue, we propose an adaptive Q-learning procedure which is an adaptive version of penalized Q-learning. We show that the proposed method can not only handle continuous covariates, but it can also be more efficient than penalized Q-learning.

1. Introduction

Since treatment of cancer and certain other chronic diseases typically involve a series of therapeutic decisions over time, there has recently been increasing interest on how to choose the best dynamic treatment strategy. The concept of dynamic treatment regimens, or adaptive treatment strategies [7], is currently one of the most promising ideas for developing effective, multi-stage therapeutic regimens. This idea has been utilized in a number of settings, including in drug and alcohol dependency studies. Data collected from such studies can be used to estimate the optimal treatment strategy. The estimation procedures are often formed as multistage decision making problems. Among these procedures, Q-learning [11], a sub-area of machine learning, has gained popularity in estimation for dynamic treatment regimens.

In spite of this increased interest in developing valid inference procedures via Q-learning for optimal dynamic treatment regimens, the research output in the area is rather limited. The difficulty lies in the fact that the treatment effect parameters at any stage prior to the last stage may be non-regular for certain longitudinal data settings, as discussed in Robins [8], and recognized by many other researchers. This

[*] We are grateful to an anonymous reviewer for insightful comments and helpful suggestions.

[†] Rui Song was funded in part by National Science Foundation grant DMS-1007698.

[‡] Michael R. Kosorok was funded in part by grant CA142538 from the National Cancer Institute.

[1] Department of Statistics, University of Haifa, Mount Carmel, Haifa 31905, Israel, e-mail: ygoldberg@stat.haifa.ac.il

[2] Department of Statistics, North Carolina State University, Raleigh NC, 27695, e-mail: rsong@ncsu.edu

[3] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, e-mail: kosorok@unc.edu

AMS 2000 subject classifications: Primary 62G05, 62G20; secondary 62F12

Keywords and phrases: Adaptive estimation, dynamic treatment regimes, penalized estimation, Q-learning

non-regularity arises when the optimal last stage treatment is non-unique for at least some subjects in the population, causing estimation bias and failure of traditional inferential approaches. In the special case that the Q-functions take the form of linear models and the treatment levels take values 1 and -1, the non-regularity problem concentrates on the "indifference" hyperplane of patient covariates where the two treatments have the same effect. A crucial step in performing inference for non-regular parameters is to correctly identify those covariate values which lie on this indifference hyperplane [see 9].

Research addressing this non-regularity problem includes the following: Moodie and Richardson [6] proposed the Zeroing Instead of Plugging In (ZIPI) method; Chakraborty et al. [1] proposed a soft-threshold estimator and implemented several kinds of bootstrap methods; and Laber et al. [5] proposed an adaptive confidence interval for the hard-max method. Due to the non-regularity, the asymptotic distribution of these estimators does not have an explicit form and hence data dependent inference procedure such as the bootstrap method have been proposed. These computational intensive methods can be unaffordably expensive for some multistage and multi-treatment settings. Moreover, the empirical results in the associated papers usually do not perform uniformly well over the evaluated design settings. To resolve this non-regularity while also saving on computational costs, penalized Q-learning was proposed in Song et al. [9] for inference with dynamic treatment regimens. The theoretical validity was verified and the computation was shown to be fast without the additional costs of bootstrapping. As a major drawback, however, penalized Q-learning can only handle discrete covariates.

The proposed adaptive Q-learning procedure is a penalized Q-learning technique that uses special adaptive weights in the penalization. We consider a two-stage decision problem for which the Q-functions have a linear model form. As discussed by Chakraborty et al. [1] and Song et al. [9], identifying the points that are on the second-stage indifference hyperplane is the key to solving the non-regularity issue. Estimation of the second-stage parameters in the adaptive Q-learning is done in two steps. First, a naive estimator for the treatment effect vector is calculated. This can be done, for example, using least-squares minimization. Second, the second-stage parameters are estimated again, this time using least-squares with a penalization term. The goal of the penalization term is to force the estimated treatment effect vector towards the true treatment effect vector. Following Goldberg and Kosorok [3] we use adaptive weights that penalize the angle to observations that are close to or on the estimated hyperplane, as determined by the naive estimator. After estimating the second-stage parameters, the proposed adaptive Q-learning procedure follows the same remaining steps as in standard Q-learning.

We prove a number of theoretical results for the proposed adaptive Q-learning. First we show that the indifference hyperplane is identified for all n large enough, with probability one. Second, we show root-n consistency and asymptotic normality of the estimators for both stages. These two results generalize the results of Song et al. [9] to the continuous covariates case. Third, we show that when the probability of being on the indifference hyperplane is positive, the asymptotic variance of the second stage estimator is the same as it would be if the direction of the indifference hyperplane were known in advance. We refer to this property, together with the identification of the indifference hyperplane, as the oracle property. Finally, we show that when the errors are Gaussian, the estimation is efficient, as if the indifference hyperplane were known in advance.

The outline of the rest of the paper is as follows. In Section 2, we present a two-stage Q-learning problem and describe existing approaches for addressing the non-regularity problem. In Section 3, we present the proposed adaptive Q-learning. In Section 4, we give the asymptotic properties of the adaptive Q-learning procedure. Concluding remarks are presented in Section 5. All proofs are deferred to the Appendix.

2. Q-learning and the non-regularity problem

In the following, we present the framework of a two-stage Q-learning problem. We then discuss the difficulty that arises due to non-smoothness of the presented problem. Finally, we review some existing approaches that address this non-smoothness. A detailed introduction to Q-learning can be found in Sutton and Barto [10]. For more details on the Q-learning framework discussed here, we refer the reader to Chakraborty et al. [1] and Song et al. [9].

Let O_t, $t = 1, 2, 3$, be random vectors that represent the states at the beginning of stage one, beginning of stage two, and at the end of stage two, respectively. For $t = 1, 2$, let $A_t \in \{-1, 1\}$ be the random action taken at stage t. Let R_t be the reward of stage t, where R_t is a function of all history, up to and including stage t, and O_{t+1}. We also denote the information at the beginning of stage t as $\mathbf{S}_t$; specifically, $\mathbf{S}_1 = O_1$ and $\mathbf{S}_2 = (O_1, A_1, R_1, O_2)$.

Let the Q-function for time $t = 1, 2$, be modeled as

$$Q_t(\mathbf{S}_t, A_t; \boldsymbol{\beta}_t, \boldsymbol{\psi}_t) = \boldsymbol{\beta}_t' \mathbf{S}_{t(1)} + (\boldsymbol{\psi}_t' \mathbf{S}_{t(2)}) A_t \,, \tag{1}$$

where $\mathbf{S}_t = (\mathbf{S}_{t(1)}', \mathbf{S}_{t(2)}')'$, and $\mathbf{S}_{t(1)}$ and $\mathbf{S}_{t(2)}$ are random vectors that take values in $\mathbb{R}^{p_t}$ and $\mathbb{R}^{q_t}$, respectively. The parameters of the Q-functions are $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t', \boldsymbol{\psi}_t')'$, where $\boldsymbol{\beta}_t \in \mathbb{R}^{p_t}$ reflects the main effect of the current state on the outcome, and $\boldsymbol{\psi}_t \in \mathbb{R}^{q_t}$ reflects the interaction between the current state and the randomly assigned treatment. Let Y_t denote the optimal pseudo-outcome reward at time t given $\mathbf{S}_t$ and A_t. The optimal pseudo-outcome consists of the current outcome plus the expected sum of all future outcomes when using the optimal treatments at all future stages. More formally, let $Y_2 = R_2$, and $Y_1 = R_1 + \max_{a \in \{-1,1\}} Q_2(\mathbf{S}_2, a; \boldsymbol{\beta}_2^*, \boldsymbol{\psi}_2^*)$, where $\boldsymbol{\theta}_2^* = (\boldsymbol{\beta}_2^{*\prime}, \boldsymbol{\psi}_2^{*\prime})'$ are the true unknown parameter values. The observed data thus consists of n trajectories of the form $(O_{1i}, A_{1i}, R_{1i}, O_{2i}, A_{2i}, R_{2i})$ for $i = 1, \ldots, n$. By definition, the pseudo-outcome allows the delayed effects of possible current treatments to be taken into account.

A standard two-stage Q-learning procedure consists of the following three steps:

Step 1. Estimate the second-stage parameters by least-squares estimation:

$$\hat{\boldsymbol{\theta}}_2 = \operatorname{argmin}_{\boldsymbol{\beta}_2, \boldsymbol{\psi}_2} \sum_{i=1}^{n} \big(Y_{2i} - Q_2(\mathbf{S}_{2i}, A_{2i}; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2)\big)^2 \,. \tag{2}$$

Step 2. Estimate the first-stage optimal pseudo-outcomes $\hat{Y}_{1i}$, where

$$\hat{Y}_{1i} = R_{1i} + \max_a Q_2(\mathbf{S}_{2i}, a; \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\psi}}_2) = R_{1i} + \hat{\boldsymbol{\beta}}_2' \mathbf{S}_{2i(1)} + |\hat{\boldsymbol{\psi}}_2' \mathbf{S}_{2i(2)}| \,. \tag{3}$$

Step 3. Estimate the first-stage parameters by least-squares estimation:

$$\hat{\boldsymbol{\theta}}_1 = \operatorname{argmin}_{\boldsymbol{\beta}_1, \boldsymbol{\psi}_1} \sum_{i=1}^{n} \big(\hat{Y}_{1i} - Q_1(\mathbf{S}_{1i}, A_{1i}; \boldsymbol{\beta}_1, \boldsymbol{\psi}_1)\big)^2 \,. \tag{4}$$

Note that the estimation of the optimal pseudo-outcomes Y_{1i} in Step 2 above involves a non-smooth function of $\hat{\boldsymbol{\psi}}_2$. Since $\hat{\boldsymbol{\theta}}_1$ is a function of $\{\hat{Y}_{11}, \ldots, \hat{Y}_{1n}\}$, it is also a non-smooth function of $\hat{\boldsymbol{\psi}}_2$. As a consequence, the asymptotic distribution of $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ does not converge uniformly over the parameter space of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. More specifically, it can be shown that the distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)$ is normal if $P(\boldsymbol{\psi}_2^{*\prime} S_{2(2)} = 0) = 0$, but is not normal if $P(\boldsymbol{\psi}_2^{*\prime} S_{2(2)} = 0) > 0$. As a result, standard inference methods such as Wald-type confidence intervals may perform poorly [1, 6, 8, 9].

We define this non-regular condition as

(NR) $P(\boldsymbol{\psi}_2^{*\prime} S_{2(2)} = 0) > 0$.

We refer to the hyperplane $\{S_{2(2)} : \boldsymbol{\psi}_2^{*\prime} S_{2(2)} = 0\}$ as the indifference hyperplane since for all points on this hyperplane, the treatment effect $\boldsymbol{\psi}_2^{*\prime} S_{2(2)} A_2$ is zero, independent of the choice of treatment A_2.

Some estimation procedures have been suggested to address the non-smoothness in (3) noted above. Moodie and Richardson [6] suggested *Zeroing Instead of Plugging In (ZIPI)* which replaces $\hat{Y}_{1i}$ of (3) with

$$\hat{Y}_{1i} = R_{1i} + \hat{\boldsymbol{\beta}}_2' \mathbf{S}_{2i(1)} + |\hat{\boldsymbol{\psi}}_2' \mathbf{S}_{2i(2)}| \cdot 1\left\{ \frac{\sqrt{n}|\hat{\boldsymbol{\psi}}_2' \mathbf{S}_{2i(2)}|}{\sqrt{\mathbf{S}_{2i(2)}' \hat{\Sigma}_2 \mathbf{S}_{2i(2)}}} > z_{1-\alpha/2} \right\}, \tag{5}$$

where $\hat{\Sigma}_2$ is the estimated covariance matrix of $\hat{\boldsymbol{\psi}}_2$, and z_α is the α-quantile of a standard normal. Chakraborty et al. [1] suggested the *soft-threshold estimator* in which $\hat{Y}_{1i}$ is replaced by

$$\hat{Y}_{1i} = R_{1i} + \hat{\boldsymbol{\beta}}_2' \mathbf{S}_{2i(1)} + |\hat{\boldsymbol{\psi}}_2' \mathbf{S}_{2i(2)}| \left(1 - \frac{\lambda_i}{|\hat{\boldsymbol{\psi}}_2' \mathbf{S}_{2i(2)}|} \right)_+,$$

where $x_+ = \max\{x, 0\}$, and λ_i is a tuning parameter. Song et al. [9] suggested *penalized Q-learning* that replaces the minimization problem (2) of Step 1 with the following penalized version:

$$\hat{\boldsymbol{\theta}}_2 = \operatorname{argmin}_{\boldsymbol{\beta}_2, \boldsymbol{\psi}_2} \sum_{i=1}^{n} \big(Y_{2i} - Q_2(\mathbf{S}_{2i}, A_{2i}; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2) \big)^2 + \sum_{i=1}^{n} p_{\lambda_n}(|\boldsymbol{\psi}_2' \mathbf{S}_{2i(2)}|), \tag{6}$$

where $p_{\lambda_n}(\cdot)$ is a pre-specified penalty function and λ_n is a tuning parameter.

In general, when condition (NR) holds, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_1$ for these three methods is not known. When $S_{2(2)}$ takes values in some finite set, Song et al. [9] has shown, under some conditions, that both $\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*)$ and $\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^*)$ converge to normal random vectors (see [9], Theorems 3–4).

3. Adaptive Q-learning

In this section we present the adaptive Q-learning procedure. This procedure is an example of penalized Q-learning that uses adaptive weights. We then discuss the choice of weights for the adaptive minimization problem.

Adaptive Q-learning consists of three steps similar to the steps for standard Q-learning as given in Section 2 above. Here we replace the minimization problem (2) of Step 1 in the standard Q-learning procedure, with the following adaptive minimization problem:

$$\Phi_{2n}(\boldsymbol{\theta}_2) = \sum_{i=1}^{n} \big(Y_{2i} - Q_2(\mathbf{S}_{2i}, A_{2i}; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2) \big)^2 + \frac{\lambda_n}{n} \sum_{i=1}^{n} \hat{w}_{ni} |\boldsymbol{\psi}_2' \mathbf{S}_{2i(2)}|, \tag{7}$$

where $\hat{w}_{ni}$ are data-driven weights and λ_n is a tuning parameter. Let $\tilde{\boldsymbol{\theta}}_2$ be the minimizer of Φ_{2n}. The second and third steps of adaptive Q-learning are the same as for the standard Q-learning, after replacing $\hat{\boldsymbol{\theta}}_2$ with $\tilde{\boldsymbol{\theta}}_2$. We denote the minimizer of (4) in adaptive Q-learning by $\tilde{\boldsymbol{\theta}}_1$.

The choice of the weights is the key to achieving the oracle property. Here, the oracle property means that the estimator behaves asymptotically as if the indifference hyperplane was known in advance. Our goal is to find weights that penalize observations that are close to or on the indifference hyperplane. We would also like the weights for points that are far from this hyperplane to converge to zero as the number of trajectories goes to infinity. Such weights enable us to identify the indifference hyperplane and hence solve the non-regularity problem.

Following Goldberg and Kosorok [3], we define the weights $\hat{w}_{ni} = n^{1/2+\alpha}(1 - n^{\alpha}|\hat{\boldsymbol{\psi}}_2'\mathbf{S}_{2i(2)}|)_+$ for some fixed $\alpha \in (0, 1/2)$, and some root-n consistent estimator $\hat{\boldsymbol{\psi}}_2$. $\hat{\boldsymbol{\psi}}_2$ can be obtained, for example, as the second component of the minimizer of (2). Note that the weight $\hat{w}_{ni}$ is large whenever $|\hat{\boldsymbol{\psi}}_2'\mathbf{S}_{2i(2)}|$ is small, i.e., when $\mathbf{S}_{2i(2)}$ is close to the hyperplane $\{S_{2(2)} : \hat{\boldsymbol{\psi}}_2' S_{2(2)} = 0\}$. Since $\hat{\boldsymbol{\psi}}_2 - \boldsymbol{\psi}_2^* = O_p(n^{-1/2})$, the weight is large whenever $\mathbf{S}_{2i(2)}$ is close to the indifference hyperplane $\{S_{2(2)} : \boldsymbol{\psi}_2^{*\prime} S_{2(2)} = 0\}$. Moreover, when $\mathbf{S}_{2i(2)}$ is far from the indifference hyperplane, the weight is small or zero.

More formally, define $M_n = \{i : \boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2i(2)} \neq 0\}$ and let $M_n^c = \{i : \boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2i(2)} = 0\}$. We assume the following:

(A1) $\mathbf{S}_{2(2)}$ is bounded, i.e., there is a constant M such that $\|\mathbf{S}_{2(2)}\| \leq M$ almost surely.

(A2) $P(|\boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2(2)}| < x | \boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2(2)} \neq 0) \leq Cx$ for some constant C and all $x > 0$ small enough.

(A3) $\mathrm{Var}(\mathbf{S}_{2(2)})$ is positive definite.

We remark that Assumption (A1) can be relaxed at the price of complicating the proofs, and that Assumption (A2) holds whenever $\mathbf{S}_{2(2)}|\mathbf{S}_{2(2)}'\boldsymbol{\psi}_2^* \neq 0$ has bounded density.

The following result is due to Goldberg and Kosorok [3] (see Lemma 2.1):

Lemma 3.1. *Assume (A1)–(A3), then*

$$\max_{i \in M_n} \frac{\hat{w}_{ni}}{\sqrt{n}} = o_p(1)\,,$$

and

$$\min_{i \in M_n^c} \frac{\hat{w}_{ni}}{\sqrt{n}} \to_p \infty\,.$$

Remark 3.2. A more general form for the weights is given by

$$\hat{w}_{ni} = n^{1/2+\alpha_1}(1 - n^{\alpha_2}c|\hat{\boldsymbol{\psi}}_2'\mathbf{S}_{2i(2)}|)_+$$

for $\alpha_1, \alpha_2 \in (0, 1/2)$ and $c > 0$. Developing strategies for choosing the parameters for the weights is an important open research question, which is beyond the scope of this paper.

4. Theoretical results

In this section, we establish asymptotic properties for the proposed adaptive Q-learning method.

We have two goals. First, we would like to show that $\sqrt{n}(\tilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*_t)$ converges weakly to some mean zero normal random vector, for $t = 1, 2$. Second, we would like to show that the estimators are oracle. Here, oracle means that when condition (NR) holds, the indifference hyperplane is identified for all n large enough, with probability one; and that the asymptotic variance of the estimators is the same as if the indifference hyperplane was known in advance.

Remark 4.1. The definition of the oracle property is close to the one discussed by Goldberg and Kosorok [3] in the context of classifiers. It generalizes the oracle property discussed by Song et al. [9], which consists of identifying the sets M_n and M_n^c for all n large enough, with probability one.

Remark 4.2. The oracle property discussed above is different from the oracle property in the context of variable selection (see, for example, [2, 12]). In the context of variable selection, the oracle property means that if there are zero components, they are estimated as zero with probability tending to 1, and the nonzero components are estimated as well as if the correct submodel were known.

We first need some notation. For $t = 1, 2$, let $\boldsymbol{\theta}^*_t$ be the minimizer of

$$E\big[\big(Y_t - Q_t(\mathbf{S}_t, A_t; \boldsymbol{\theta}_t)\big)^2\big] \equiv E\big[\big(Y_t - \mathbf{S}'_{t(1)}\boldsymbol{\beta}_t - A_t\mathbf{S}'_{t(2)}\boldsymbol{\psi}_t\big)^2\big].$$

Define the matrices

$$\Omega_t(\boldsymbol{\theta}_t) = E\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}_t}(Y_t - Q_t(\mathbf{S}_t, A_t; \boldsymbol{\theta}_t))^2\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_t}(Y_t - Q_t(\mathbf{S}_t, A_t; \boldsymbol{\theta}_t))^2\right\}'\right],$$
$$H_t(\boldsymbol{\theta}_t) = E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}_t^2}(Y_t - Q_t(\mathbf{S}_t, A_t; \boldsymbol{\theta}_t))^2\right].$$

To simplify the notation, denote $\Omega^*_t \equiv \Omega_t(\boldsymbol{\theta}^*_t)$. Note that $H_t(\boldsymbol{\theta}_t) \equiv 2E[\mathbf{Z}_t\mathbf{Z}'_t]$, where $\mathbf{Z}_t = (\mathbf{S}'_{t(1)}, \mathbf{S}'_{t(2)}A_t)'$, and hence does not depend on the parameters. Moreover, H_t is positive semi-definite for $t = 1, 2$.

We need the following regularity assumptions:

(B1) For $t = 1, 2$, H_t is positive definite.
(B2) For $t = 1, 2$, $\boldsymbol{\theta}^*_t \in \Theta_t$ is an inner point, where $\Theta_t \subset \mathbb{R}^{p_t+q_t}$ is compact.

Note that Assumption (B1) implies that, for $t = 1, 2$, $E[(Y_t - Q_t(\mathbf{S}_t, A_t; \boldsymbol{\theta}_t))^2]$ is strictly convex, and thus has a unique minimizer. Assumption (B2) simplifies the proofs but can be relaxed.

We start by discussing the properties of $\tilde{\boldsymbol{\theta}}_{2n}$. Formally, let $\boldsymbol{v} = \boldsymbol{\psi}^*_{22}/\|\boldsymbol{\psi}^*_{22}\|$. Define

$$\Sigma_V = \begin{cases} \Sigma_2^{(NR)} \equiv P'_V(P_VH_2P'_V)^{-1}P_V\Omega^*_2P'_V(P_VH_2P'_V)^{-1}P_V, \\ \qquad \text{when condition (NR) holds,} \\ H_2^{-1}\Omega^*_2H_2^{-1}, \\ \qquad \text{when condition (NR) does not hold,} \end{cases}$$

where

$$P_V = \begin{pmatrix} I_{p_2\times p_2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{v}' \end{pmatrix}.$$

and where $I_{p_2\times p_2}$ is the $p_2 \times p_2$ identity matrix. The matrix $\Sigma_2^{(NR)}$ is the limiting covariance matrix of the least-squares estimator for $\boldsymbol{\theta}^*_2$ under the assumption that $\boldsymbol{\psi}^*_2$ is known up to scale.

Theorem 4.3. *Assume (A1)–(A3) and (B1)–(B2). Let λ_n be a sequence bounded away from zero and infinity. Then $\tilde{\boldsymbol{\theta}}_{2n} \to_{a.s.} \boldsymbol{\theta}_2^*$, and*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_{2n} - \boldsymbol{\theta}_2^*) \to_d N(\mathbf{0}, \Sigma_V)\,. \tag{8}$$

Moreover, when condition (NR) holds, with probability tending to 1,

$$\tilde{\boldsymbol{\psi}}_{2n} \in \mathrm{span}(\boldsymbol{\psi}_2^*)\,. \tag{9}$$

In other words, the observations that are on the indifference hyperplane are identified for all n large enough, with probability one.

The proof appears in the Appendix.

Before we continue to discuss the distribution of $\tilde{\boldsymbol{\theta}}_1$, consider the specific model

$$Y_{2i} = Q_2(\mathbf{S}_{2i}, A_2; \boldsymbol{\theta}_2^*) + \varepsilon_i = \boldsymbol{\theta}_2^{*\prime}\mathbf{Z}_{2i} + \varepsilon_i, \tag{10}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ are i.i.d. and independent of the $\mathbf{S}_{2i}$. In this case we have the following result:

Corollary 4.4. *Assume the conditions of Theorem 4.3 and that model (10) holds. Then when condition (NR) holds, the estimator $\tilde{\boldsymbol{\theta}}_2$ asymptotically achieves the information bound for the submodel in which $\boldsymbol{\psi}_2^*$ is known up to scale.*

The proof is provided in the Appendix.

Remark 4.5. Theorem 4.3 and Corollary 4.4 improve on Theorems 2–3 of Song et al. [9] in two ways. First, Song et al. [9] ensures identification of the indifference hyperplane only when the set of potential values for $\mathbf{S}_{2(2)}$ is finite. Second, when condition (NR) holds, the asymptotic variance is changed correspondingly to the identification of the span of $\boldsymbol{\psi}_2^*$. Moreover, when the normal model (10) holds, $\tilde{\boldsymbol{\theta}}_2$ is as efficient as if the $\boldsymbol{\psi}_2^*$ is known up to scale. Compare to Theorem 3 of Song et al. [9], where the variance is approximately $n^{-1}H_2^{-1}\Omega_2^* H_2^{-1}$, i.e., it does not change when condition (NR) holds.

We are now ready to discuss the asymptotic behavior of $\tilde{\boldsymbol{\theta}}_1$:

Theorem 4.6. *Assume (A1)–(A3) and (B1)–(B2). Let λ_n be a sequence bounded away from zero and infinity. Let $\bar{\mathbf{Z}}_2 \equiv (\mathbf{S}_{2(1)}', \mathbf{S}_{2(2)}' \,\mathrm{sign}(\boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2(2)}))'$. Then $\sqrt{n}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*) \to -H_1^{-1}X$, where*

$$X \sim N\big(\mathbf{0}, \mathrm{Cov}\{\phi(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*) - 2\{P(\mathbf{Z}_1\bar{\mathbf{Z}}_2')\}F_2(Y_2, \mathbf{Z}_2)\}\big),$$
$$\phi(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*) = -2\big(R_1 + \boldsymbol{\beta}_2^{*\prime}\mathbf{S}_{2(1)} + |\boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2(2)}| - \boldsymbol{\theta}_1^{*\prime}\mathbf{Z}_1\big)\mathbf{Z}_1$$

and $F_2(Y_2, \mathbf{Z}_2)$ is the influence function for $\tilde{\boldsymbol{\theta}}_2$ given in (22) below.

The proof is provided in the Appendix.

The empirical versions of the asymptotic variance matrices can be used as variance estimators whether (NR) condition does or does not hold. A test to check if condition (NR) holds can be performed using (9) of Theorem 4.3, together with Assumption (A2). Details of such a procedure, as well as an empirical study, are deferred for future research.

5. Discussion

In this paper, we have proposed an adaptive Q-learning procedure for dynamic treatment regimens. We showed that the proposed method has oracle properties. We also proved that under certain conditions, the estimation of the second-stage parameters is as efficient as if the indifference hyperplane was known in advance.

The framework studied in this paper is of a two-stage decision problem for which the Q-functions have a linear model form. We believe that the results presented here illustrate an important approach for dealing with inference for non-regular parameters in the multistage decision context. While this work focuses on the two-stage problem, generalization to the multistage decision problem is relatively straightforward. However, the linear model form of the Q-functions presented here may not be sufficiently flexible for certain practical settings, and more research is needed to address such general cases. Further research directions also include more flexible semiparametric or nonparametric modeling to allow general forms for the covariates and diverse data such as ordinal or censored outcomes. Finally, an empirical study is of a great importance in order for both a better understanding of the performance of the proposed adaptive Q-learning as well as for developing data driven strategies for choosing the constants employed in the adaptive weights. Such empirical research is the subject of ongoing efforts.

Appendix A: Proofs

Proof of Theorem 4.3. We first assume that condition (NR) holds. Write $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2', \boldsymbol{\psi}_2')' \equiv (\boldsymbol{\beta}_2^{*\prime} + \frac{\boldsymbol{u}_{(1)}'}{\sqrt{n}}, \boldsymbol{\psi}_2^{*\prime} + \frac{\boldsymbol{u}_{(2)}'}{\sqrt{n}})'$, where $\boldsymbol{u} = (\boldsymbol{u}_{(1)}', \boldsymbol{u}_{(2)}') \in \mathbb{R}^{p_2+q_2}$. Define

$$\begin{aligned}\Gamma_n(\boldsymbol{u}) \equiv& \Phi_{2n}\left(\boldsymbol{\theta}_2^* + \frac{\boldsymbol{u}}{\sqrt{n}}\right) - \Phi_{2n}(\boldsymbol{\theta}_2^*)\\ =&\sum_{i=1}^n \left(\left(Y_{2i} - Q_2\left(\mathbf{S}_{2i}, A_{2i}; \boldsymbol{\theta}_2^* + \frac{\boldsymbol{u}}{\sqrt{n}}\right)\right)^2 - \left(Y_{2i} - Q_2(\mathbf{S}_{2i}, A_{2i}; \boldsymbol{\theta}_2^*)\right)^2\right)\\ &+ \frac{\lambda_n}{n}\sum_{i=1}^n \hat{w}_{ni}\left(\left|\left(\boldsymbol{\psi}_2^* + \frac{\boldsymbol{u}_{(2)}}{\sqrt{n}}\right)'\mathbf{S}_{2i(2)}\right| - |\boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2i(2)}|\right)\\ \equiv& A_n(\boldsymbol{u}) + B_n(\boldsymbol{u}_{(2)})\,. \end{aligned} \tag{11}$$

Fix $\boldsymbol{u}$ and let

$$\begin{aligned}G_{n2}(\boldsymbol{\theta}_2) =& \frac{\partial}{\partial \boldsymbol{\theta}_2}\left(\sum_{i=1}^n (Y_{2i} - Q_2(\mathbf{S}_{2i}, A_{2i}; \boldsymbol{\theta}_2))^2\right)\\ =& -2n\mathbb{P}_n(Y_2 - \boldsymbol{\theta}_2'\mathbf{Z}_2)\mathbf{Z}_2\,,\end{aligned}$$

where $\mathbb{P}_n$ is the empirical measure and $\mathbf{Z}_2 = (\mathbf{S}_{2(1)}', \mathbf{S}_{2(2)}'A_2)'$. It follows from Assumption (B1) that

$$A_n(\boldsymbol{u}) = \frac{\boldsymbol{u}'}{\sqrt{n}}G_{n2}(\boldsymbol{\theta}_2^*) + \frac{1}{2}\boldsymbol{u}'\mathbb{P}_n[2\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{u} + o_p(1) \to_d N\left(\frac{1}{2}\boldsymbol{u}'H_2\boldsymbol{u}, \Omega_2^*\right)\,. \tag{12}$$

Consider now $B_n(\boldsymbol{u}_{(2)})$. Write

$$\begin{aligned} B_n(\boldsymbol{u}_{(2)}) &= \frac{\lambda_n}{n}\sum_{i\in M_n}\hat{w}_{ni}\left(\left|\left(\boldsymbol{\psi}_2^* + \frac{\boldsymbol{u}_{(2)}}{\sqrt{n}}\right)'\mathbf{S}_{2i(2)}\right| - |\boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2i(2)}|\right) \\ &\quad + \frac{\lambda_n}{n}\sum_{i\in M_n^c}\frac{\hat{w}_{ni}}{\sqrt{n}}|\boldsymbol{u}'_{(2)}S_{2i(2)}| \\ &\equiv B_{n1}(\boldsymbol{u}_{(2)}) + B_{n2}(\boldsymbol{u}_{(2)})\,. \end{aligned} \tag{13}$$

Similar arguments to those that appear in the proof of Theorem 3.2 of Goldberg and Kosorok [3] verify that

$$|B_{n1}(\boldsymbol{u}_{(2)})| = o_p(\|\boldsymbol{u}_{(2)}\|);\quad B_{n2}(\boldsymbol{u}_{(2)}) \to \begin{cases} 0, & \boldsymbol{u}_{(2)}\in V, \\ \infty, & \boldsymbol{u}_{(2)}\notin V. \end{cases} \tag{14}$$

Thus,

$$\Gamma_n(\boldsymbol{u}) \to_d \Gamma(\boldsymbol{u}) \equiv \begin{cases} \frac{1}{2}\boldsymbol{u}'H_2\boldsymbol{u} + W'\boldsymbol{u}, & \boldsymbol{u}_{(2)}\in V, \\ \infty, & \boldsymbol{u}_{(2)}\notin V, \end{cases}$$

where $W \sim N(\mathbf{0}, \Omega_2^*)$.

Note that in order to minimize $\Gamma(\boldsymbol{u})$, $\boldsymbol{u}_{(2)}$ must be in V. Write $\boldsymbol{u} = (\boldsymbol{u}'_{(1)}, \alpha\boldsymbol{v}')'$. Thus, we need to minimize

$$\frac{1}{2}(\boldsymbol{u}'_{(1)}, \alpha\boldsymbol{v}')H_2\begin{pmatrix}\boldsymbol{u}_{(1)}\\ \alpha\boldsymbol{v}\end{pmatrix} + (\boldsymbol{u}'_{(1)}, \alpha\boldsymbol{v}')\begin{pmatrix}W_{(1)}\\ W_{(2)}\end{pmatrix} \equiv \frac{1}{2}\bar{\boldsymbol{u}}'\bar{H}_2\bar{\boldsymbol{u}} + \bar{\mathbf{W}}'\bar{\boldsymbol{u}}\,, \tag{15}$$

where

$$\bar{\boldsymbol{u}} = \begin{pmatrix}\boldsymbol{u}_{(1)}\\ \alpha\end{pmatrix};\quad \bar{H}_2 = \begin{pmatrix} I & \mathbf{0}\\ \mathbf{0} & \boldsymbol{v}'\end{pmatrix}H_2\begin{pmatrix} I & \mathbf{0}\\ \mathbf{0} & \boldsymbol{v}\end{pmatrix} \equiv P_V H_2 P_V';\quad \bar{W} = \begin{pmatrix}W_{(1)}\\ \boldsymbol{v}'W_{(2)}\end{pmatrix} = P_V W\,.$$

The minimizer of (15) is $\bar{\boldsymbol{u}}^* \equiv -\bar{H}_2^{-1}\bar{W}$. The covariance matrix of $\bar{\boldsymbol{u}}^*$ is given by $\bar{H}_2^{-1}P_V\Omega_2^*P_V'\bar{H}_2^{-1}$. Write $\boldsymbol{u}^* = P_V'\bar{\boldsymbol{u}}^*$ and note that $\boldsymbol{u}^*$ is the minimizer of $\Gamma(\boldsymbol{u})$. Finally, note that the covariance matrix of $\boldsymbol{u}^*$ is given by $\Sigma_2^{(NR)} \equiv P_V'\bar{H}_2^{-1}P_V\Omega_2^*P_V' \times \bar{H}_2^{-1}P_V$.

Let $\tilde{\boldsymbol{u}}_n = \operatorname{argmin}_{\boldsymbol{u}}\Gamma_n(\boldsymbol{u})$; then $\tilde{\boldsymbol{u}}_n = \sqrt{n}(\tilde{\boldsymbol{\theta}}_{2n} - \boldsymbol{\theta}_2^*)$. We would like to show that $\tilde{\boldsymbol{u}}_n \to_d \boldsymbol{u}^*$. Note that $\Gamma_n(\boldsymbol{u})$, for all $n \geq 1$, and $\Gamma(\boldsymbol{u})$ are stochastic processes indexed by $\mathbb{R}^{p_2+q_2}$. The sample paths of Γ are lower semicontinuous and possess a unique minimum. We would like to show that $\{\tilde{\boldsymbol{u}}_n\}_n$ is uniformly tight. It is enough to show that for any given $\varepsilon > 0$, there exists a constant C_1 such that

$$P\Big(\inf_{\|\boldsymbol{u}\|\geq C_1}\Phi_{2n}(\boldsymbol{\theta}_2^* + \boldsymbol{u}/\sqrt{n}) > \Phi_{2n}(\boldsymbol{\theta}_2^*)\Big) \geq 1 - \varepsilon \tag{16}$$

for all n large enough. To see this, note that

$$\begin{aligned} &\Phi_{2n}\left(\boldsymbol{\theta}_2^* + \frac{\boldsymbol{u}}{\sqrt{n}}\right) - \Phi_{2n}(\boldsymbol{\theta}_2^*) \\ &\quad\geq \frac{\boldsymbol{u}'}{\sqrt{n}}G_{n2}(\boldsymbol{\theta}_2^*) + \frac{1}{2}\boldsymbol{u}'H_2\boldsymbol{u}(1 + \mathbf{o}_p(1)) \\ &\qquad + \frac{\lambda_n}{n}\sum_{i\in M_n}\hat{w}_{ni}\left(\left|\left(\boldsymbol{\psi}_2^* + \frac{\boldsymbol{u}_{(2)}}{\sqrt{n}}\right)'\mathbf{S}_{2i(2)}\right| - |\boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2i(2)}|\right), \end{aligned} \tag{17}$$

where the $o_p(1)$ term is uniform in $\boldsymbol{u}$. The first term of (17) is $O_p(1)$. By Assumption (B1), H_2 is positive definite and hence for C_1 large enough the second term dominates the first term. By (14), the second term of (17) also dominates the third term. Consequently, we obtain that $\{\tilde{\boldsymbol{u}}_n\}_n$ is uniformly tight. Hence, all the conditions of the Argmax Theorem [4, Theorem 14.1] hold, and consequently, $\tilde{\boldsymbol{u}}_n \to_d \boldsymbol{u}^*$. In other words, $\sqrt{n}(\tilde{\boldsymbol{\theta}}_{2n} - \boldsymbol{\theta}_2^*) \to_d N(\mathbf{0}, \Sigma_2^{(NR)})$, which concludes the proof of (8) when condition (NR) holds.

We would like to show that $\tilde{\psi}_{2n} \in V$ with probability that tends to 1. It is sufficient to show that for any sequence $\boldsymbol{\theta}_{2n}$satisfying $\boldsymbol{\theta}_{2n} - \boldsymbol{\theta}_2^* = O_p(n^{-1/2})$ and $\boldsymbol{\psi}_{2n} \notin V$, $\Phi_{2n}(\boldsymbol{\theta}_{2n}) > \Phi_{2n}(T_V\boldsymbol{\theta}_{2n})$, with probability tending to 1 as $n \to \infty$, where $T_V \equiv P_V' P_V$ is the projection matrix that projects the second component of a (p_2+q_2)-vector onto the subspace V.

For a vector $\boldsymbol{u} = (\boldsymbol{u}_{(1)}', \boldsymbol{u}_{(2)}')' \in \mathbb{R}^{p_2+q_2}$, write

$$\boldsymbol{u}_V = \boldsymbol{v}\boldsymbol{v}'\boldsymbol{u}_{(2)}; \quad \boldsymbol{u}_{V^\perp} = (I_{q_2\times q_2} - \boldsymbol{v}\boldsymbol{v}')\boldsymbol{u}_{(2)}; \quad T_V\boldsymbol{u} = (\boldsymbol{u}_{(1)}', \boldsymbol{v}\boldsymbol{v}'\boldsymbol{u}_{(2)})\,.$$

Write $\boldsymbol{\psi}_{2n} = \boldsymbol{\psi}_2^* + \frac{\boldsymbol{u}_V}{\sqrt{n}} + \frac{\boldsymbol{u}_{V^\perp}}{\sqrt{n}}$, and note that $\boldsymbol{\theta}_{2n} - T_V\boldsymbol{\theta}_{2n} = (\mathbf{0}', \frac{\boldsymbol{u}_{V^\perp}}{\sqrt{n}}')'$. Hence

$$\begin{aligned}
&\frac{\Phi_{2n}(\boldsymbol{\theta}_2) - \Phi_{2n}(T_V\boldsymbol{\theta}_2)}{\|\boldsymbol{u}_{V^\perp}\|} \\
&\quad \geq \frac{1}{\|\boldsymbol{u}_{V^\perp}\|} \\
&\qquad \times \left((\boldsymbol{\theta}_2 - T_V\boldsymbol{\theta}_2)' G_{n2}(\boldsymbol{\theta}_2^*) + n\,(\boldsymbol{\theta}_2 - T_V\boldsymbol{\theta}_2)' H_2\,(\boldsymbol{\theta}_2 - T_V\boldsymbol{\theta}_2)\left(1 + o_p(1)\right)\right) \\
&\qquad + \frac{1}{\|\boldsymbol{u}_{V^\perp}\|}\frac{\lambda_n}{n}\sum_{i=1}^{n}\hat{w}_{ni}\left(|\boldsymbol{\psi}_2'\mathbf{S}_{2i(2)}| - |T_V\boldsymbol{\psi}_2'\mathbf{S}_{2i(2)}|\right) \\
&\quad \geq \frac{1}{\sqrt{n}}\frac{\boldsymbol{u}_{V^\perp}}{\|\boldsymbol{u}_{V^\perp}\|}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\psi}_2}\left(Y_{2i} - Q_2(\mathbf{S}_{2i}, A_{2i}; \boldsymbol{\theta}_2^*)\right)^2 + o_p(1) \\
&\qquad + \frac{\lambda_n}{n}\frac{1}{\|\boldsymbol{u}_{V^\perp}\|}\sum_{i\in M_n}\hat{w}_{ni}\left(|\boldsymbol{\psi}_2'\mathbf{S}_{2i(2)}| - \left|\left(\boldsymbol{\psi}_2^* + \frac{\boldsymbol{u}_V}{\sqrt{n}}\right)'\mathbf{S}_{2i(2)}\right|\right) \\
&\qquad + \frac{\lambda_n}{n}\sum_{i\in M_n^c}\hat{w}_{ni}\frac{|\boldsymbol{u}_{V^\perp}{}'\mathbf{S}_{2i(2)}|}{\|\boldsymbol{u}_{V^\perp}\|} \\
&\quad \equiv \widetilde{A}_n + \widetilde{B}_{n1} + \widetilde{B}_{n2} + o_p(1)\,.
\end{aligned}$$

where the $o_p(1)$ term is uniform in $\boldsymbol{u}$. By Assumption (B1), $\widetilde{A}_n = O_p(1)$ uniformly in $\boldsymbol{u}$. Using similar arguments to those used to prove (14), $\widetilde{B}_{n1} = o_p(1)$ uniformly in $\boldsymbol{u}$ and $\widetilde{B}_{n2} \to_p \infty$ uniformly in $\boldsymbol{u}$. Hence $\Phi_{2n}(\boldsymbol{\theta}_2) > \Phi_{2n}(T_V\boldsymbol{\theta}_2)$ with probability that tends to one uniformly in $\boldsymbol{u}$. This concludes the proof of (9).

Consider now the case in which condition (NR) does not hold. Note hat $\Gamma_n(\boldsymbol{u}) = A_n(\boldsymbol{u}) + B_{n1}(\boldsymbol{u}_{(2)})$, where A_n is defined in (12) and B_{n1} is defined in (13). It follows from the same arguments as given above that $A_n(\boldsymbol{u}) \to_d N\left(\frac{1}{2}\boldsymbol{u}'H_2\boldsymbol{u}, \Omega_2^*\right)$ and $B_{n1}(\boldsymbol{u}_{(2)}) \to_p 0$. Hence,

$$\Gamma_n(\boldsymbol{u}) \to_d \frac{1}{2}\boldsymbol{u}'H_2\boldsymbol{u} + W'\boldsymbol{u}\,,$$

where $W \sim N(\mathbf{0}, \Omega_2^*)$. Note that $\boldsymbol{u}^*$, the minimizer of $\frac{1}{2}\boldsymbol{u}'H_2\boldsymbol{u} + W'\boldsymbol{u}$, is $-H_2^{-1}W$ and follows the distribution $N(\mathbf{0}, H_2^{-1}\Omega_2^*H_2^{-1})$. It can be shown that the Argmax theorem conditions now all hold, and thus we obtain that $\sqrt{n}(\tilde{\theta}_{2n}-\boldsymbol{\theta}_2^*) = \tilde{\boldsymbol{u}}_n \to_d \boldsymbol{u}^*$. This proves (8) for the case that (NR) does not hold. □

Proof of Corollary 4.4. First note that when model (10) holds,

$$\ell(Y, \mathbf{S}_{2(1)}, \mathbf{S}_{2(2)}A_2; \boldsymbol{\beta}_2^*, \boldsymbol{\psi}_2^*) = \frac{1}{\sigma^2}(Y_2 - \boldsymbol{\theta}_2^{*\prime}\mathbf{Z}_2)\mathbf{Z}_2\,,$$

and $\mathcal{I}(\boldsymbol{\theta}_2^*) = \sigma^{-2}E[\mathbf{Z}_2\mathbf{Z}_2']$ where $\mathcal{I}(\boldsymbol{\theta}_2^*)$ is the Fisher information matrix in the full model. Note that

$$\begin{aligned}\Omega(\boldsymbol{\theta}_2^*) =&4E\left[(Y_2 - \boldsymbol{\theta}_2^{*\prime}\mathbf{Z}_2)\mathbf{Z}_2\mathbf{Z}_2'(Y_2 - \boldsymbol{\theta}_2^{*\prime}\mathbf{Z}_2)\right]\\ =&4E[\varepsilon^2]E[\mathbf{Z}_2\mathbf{Z}_2'] = 4\sigma^2E[\mathbf{Z}_2\mathbf{Z}_2']\,,\end{aligned} \tag{18}$$

and recall that $H_2 = 2E[\mathbf{Z}_2\mathbf{Z}_2']$. Consider the submodel in which $\boldsymbol{\psi}_2^* = \alpha\boldsymbol{v}$ is known up to the scalar α. In this case, the score for $(\boldsymbol{\beta}_2', \alpha)'$ is given by

$$\ell(Y, \mathbf{S}_{2(1)}, \mathbf{S}_{2(2)}A_2; \boldsymbol{\beta}_2, \alpha) = \frac{1}{\sigma^2}\left(Y - \boldsymbol{\beta}_2'\mathbf{S}_{2(1)} - \alpha\boldsymbol{v}'\mathbf{S}_{2(2)}A_2\right)(\mathbf{S}_{2(1)}', \boldsymbol{v}'\mathbf{S}_{2(2)}A_2)'\,.$$

Hence, the information matrix for this submodel at the true parameter values is

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta}_2^*, \boldsymbol{\psi}_2^{*\prime}\boldsymbol{v}) =&E[\ell(\boldsymbol{\beta}_2^*, \boldsymbol{\psi}_2^{*\prime}\boldsymbol{v})\ell(\boldsymbol{\beta}_2^*, \boldsymbol{\psi}_2^{*\prime}\boldsymbol{v})']\\ =&\frac{1}{\sigma^4}E[\varepsilon^2]P_VE[\mathbf{Z}_2\mathbf{Z}_2']P_V' = \frac{1}{4\sigma^4}P_V\Omega(\boldsymbol{\theta}_2^*)P_V'\,.\end{aligned} \tag{19}$$

By Theorem 4.3, when condition (NR) holds, the limiting covariance matrix of $\sqrt{n}\tilde{\boldsymbol{\theta}}_2$ is given by

$$\Sigma_2^{(NR)} = P_V'(P_VH_2P_V')^{-1}P_V\Omega_2^*P_V'(P_VH_2P_V')^{-1}P_V = P_V'\mathcal{I}^{-1}(\boldsymbol{\beta}_2^*, \boldsymbol{\psi}_2^{*\prime}\boldsymbol{v})P_V, \tag{20}$$

where the last equality follows from (18) and (19). Since the parameter of interest is $\psi((\boldsymbol{\beta}_2', \alpha)') = P_V'(\boldsymbol{\beta}_2', \alpha)' = (\boldsymbol{\beta}_2', \alpha\boldsymbol{v}')'$, the information bound is thus given by

$$\frac{\partial\psi(\boldsymbol{\theta}_2)}{\partial(\boldsymbol{\beta}_2', \alpha)'}\mathcal{I}^{-1}(\boldsymbol{\beta}_2^*, \boldsymbol{\psi}_2^{*\prime}\boldsymbol{v})\frac{\partial\psi(\boldsymbol{\theta}_2)}{\partial(\boldsymbol{\beta}_2', \alpha)'}' = P_V'\mathcal{I}^{-1}(\boldsymbol{\beta}_2^*, \boldsymbol{\psi}_2^{*\prime}\boldsymbol{v})P_V\,,$$

and the result now follows from (20). □

Proof of Theorem 4.6. Note that by taking the derivative, the minimization problem (4) of Step 3 is equivalent to solving an estimating equation. In the following, we use Z-estimation results to prove the asymptotic normality. We remark that since the estimating equation is based on the recursion, the data is no longer i.i.d.

Write

$$\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)(\mathbf{S}_2) = -2\big(R_1 + \boldsymbol{\beta}_2'\mathbf{S}_{2(1)} + |\boldsymbol{\psi}_2'\mathbf{S}_{2(2)}| - \boldsymbol{\theta}_1'\mathbf{Z}_1\big)\mathbf{Z}_1\,,$$

where $\mathbf{Z}_1 = (\mathbf{S}_{1(1)}', \mathbf{S}_{1(2)}'A_1)'$. Note that

$$\begin{aligned}\mathbb{P}_n\phi(\boldsymbol{\theta}_1, \tilde{\boldsymbol{\theta}}_2) &= \frac{\partial}{\partial\boldsymbol{\theta}_1}\left(\frac{1}{n}\sum_{i=1}^n(\hat{Y}_{1i} - Q_1(\mathbf{S}_{1i}, A_{1i}; \boldsymbol{\theta}_1))^2\right);\\ P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*) &= P\frac{\partial}{\partial\boldsymbol{\theta}_1}(Y_1 - Q_1(\mathbf{S}_1, A_1; \boldsymbol{\theta}_1))^2,\end{aligned}$$

where $\mathbb{P}_n$ and P are the expectations with respect to the empirical and true distributions, respectively. Hence $\hat{\boldsymbol{\theta}}_1$ is the solution of the estimating equation $\mathbb{P}_n\phi(\boldsymbol{\theta}_1, \tilde{\boldsymbol{\theta}}_2) = \mathbf{0}$, and $\boldsymbol{\theta}_1^*$ is the unique solution of $P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*) = \mathbf{0}$. Write $\Psi_n(\boldsymbol{\theta}_1) = \mathbb{P}_n\phi(\boldsymbol{\theta}_1, \tilde{\boldsymbol{\theta}}_2)$, $\Psi_n^*(\boldsymbol{\theta}_1) = \mathbb{P}_n\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*)$, and $\Psi(\boldsymbol{\theta}_1) = P\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*)$.

Note that

$$\begin{aligned}
&\sqrt{n}\,(\Psi_n(\boldsymbol{\theta}_1) - \Psi(\boldsymbol{\theta}_1)) \\
&\quad = \sqrt{n}\,(\Psi_n(\boldsymbol{\theta}_1) - \Psi_n^*(\boldsymbol{\theta}_1)) + \sqrt{n}\,(\Psi_n^*(\boldsymbol{\theta}_1) - \Psi(\boldsymbol{\theta}_1)) \\
&\quad = -2\mathbb{P}_n((\tilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)'\mathbf{S}_{2(1)} + |\tilde{\boldsymbol{\psi}}_2'\mathbf{S}_{2(2)}| - |\boldsymbol{\psi}_2^*\mathbf{S}_{2(2)}|)\mathbf{Z}_1 + \sqrt{n}(\mathbb{P}_n - P)\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*) \qquad (21)\\
&\quad = -2\mathbb{P}_n((\tilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)'\mathbf{S}_{2(1)} + \operatorname{sign}(\boldsymbol{\psi}_2^{*\prime}\mathbf{S}_{2(2)})(\tilde{\boldsymbol{\psi}} - \boldsymbol{\psi}_2^*)'\mathbf{S}_{2(2)})\mathbf{Z}_1 + o_p(1) \\
&\qquad + \sqrt{n}(\mathbb{P}_n - P)\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*)\,,
\end{aligned}$$

where the $o_p(1)$ term is uniform over $\boldsymbol{\theta}_1 \in \Theta_1$; and where the last equality follows since, by Theorem 4.3, $\tilde{\boldsymbol{\psi}}_2'\mathbf{S}_{2(2)}$ and $\boldsymbol{\psi}_2^*\mathbf{S}_{2(2)}$ have the same sign with probability tending to one. It follows from the proof of Theorem 4.3, that $\sqrt{n}(\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^*) = \sqrt{n}\mathbb{P}_n F_2(Y_2, \mathbf{Z}_2) + o_p(1)$, where

$$F_2(Y_2, \mathbf{Z}_2) = \begin{cases} -2\bar{H}_2^{-1}P_V(Y_2 - \boldsymbol{\theta}_2'\mathbf{Z}_2)\mathbf{Z}_2, \\ \qquad \text{when condition (NR) holds,} \\ -2H_2^{-1}(Y_2 - \boldsymbol{\theta}_2'\mathbf{Z}_2)\mathbf{Z}_2, \\ \qquad \text{when condition (NR) does not hold,} \end{cases} \qquad (22)$$

is the influence function for $\tilde{\boldsymbol{\theta}}_2$. Hence, we can rewrite (21) as

$$\sqrt{n}(\Psi_n(\boldsymbol{\theta}_1) - \Psi(\boldsymbol{\theta}_1)) = \sqrt{n}(\mathbb{P}_n - P)\big(\phi\big(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^* - 2P(\mathbf{Z}_1\bar{\mathbf{Z}}_2')F_2(Y_2, \mathbf{Z}_2)\big)\big) + o_p(1)\,.$$

Consequently,

$$\sqrt{n}\big(\Psi_n(\boldsymbol{\theta}_1) - \Psi(\boldsymbol{\theta}_1)\big) \to_d X(\boldsymbol{\theta}_1)\,,$$

where $X(\boldsymbol{\theta}_1) \sim N(\mathbf{0}, \operatorname{Cov}\{\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*) - 2\{P(\mathbf{Z}_1\bar{\mathbf{Z}}_2')\}F_2(Y_2, \mathbf{Z}_2)\})$ is Gaussian process. As a Gaussian process, $X(\boldsymbol{\theta}_1)$ has continuous sample paths in $\boldsymbol{\theta}_1$. Note that by Assumption (B1), Ψ is the derivative of a strictly convex function. Hence, for every sequence $\boldsymbol{\theta}_{1n}$ such that $\|\Psi(\boldsymbol{\theta}_{1n})\| \to 0$, $\|\boldsymbol{\theta}_{1n} - \boldsymbol{\theta}_1^*\| \to 0$. Finally, note that by Assumption (B2), Ψ is uniformly bounded over Θ_1. Hence, all the conditions of Corollary 13.6 of Kosorok [4] hold, and we obtain that

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*) \to_d -H_1^{-1}X(\boldsymbol{\theta}_1^*)\,,$$

which concludes the proof. □

References

[1] Chakraborty, B., Murphy, S. and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research* **19** 317–343.

[2] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

[3] Goldberg, Y. and Kosorok, M. R. (2011). Comment on "Adaptive confidence intervals for the test error in classification," by E. B. Labor and S. A. Murphy. *Journal of the American Statistical Association* **106** 920–924.

[4] Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

[5] Laber, E. B., Lizotte, D. J., Qian, M. and Murphy, S. A. (2011). Statistical inference in dynamic treatment regimes. Manuscript.

[6] Moodie, E. E. M. and Richardson, T. S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics* **37** 126–146.

[7] Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B – Statistical Methodology* **65** 331–355.

[8] Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics* (D. Y. Lin and P. Heagerty, eds.). 189–326.

[9] Song, R., Wang, W., Zeng, D. and Kosorok, M. R. (2011). Penalized Q-learning for dynamic treatment regimes. Submitted.

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

[11] Watkins, C. J. C. H. (1989). Learning from delayed rewards. Ph.D. thesis, Cambridge University.

[12] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.