The Annals of Statistics
2022, Vol. 50, No. 4, 2047–2071
https://doi.org/10.1214/22-AOS2171

GENERALIZATION ERROR BOUNDS OF DYNAMIC TREATMENT REGIMES IN PENALIZED REGRESSION-BASED LEARNING

BY EUN JEONG OH[a], MIN QIAN[b] AND YING KUEN CHEUNG[c]

Department of Biostatistics, Columbia University, [a]*eo2366@caa.columbia.edu,* [b]*mq2158@cumc.columbia.edu,* [c]*yc632@cumc.columbia.edu*

A dynamic treatment regime (DTR) is a sequence of decision rules, one per stage of intervention, that maps up-to-date patient information to a recommended treatment. Discovering an appropriate DTR for a given disease is a challenging issue especially when a large set of prognostic variables are observed. To address this problem, we propose penalized regression-based learning methods with l_1 penalty to estimate the optimal DTR that would maximize the expected outcome if implemented. We also provide generalization error bounds of the estimated DTR in the setting of finite number of stages with multiple treatment options. We first examine the relationship between value and Q-functions and derive a finite sample upper bound on the difference in values between the optimal and the estimated DTRs. For practical implementation, we develop an algorithm with partial regularization via orthogonality to construct the optimal DTR. The advantages of the proposed methods are demonstrated with extensive simulation studies and data analysis of depression clinical trials.

1. Introduction. Discovering effective treatment regimes for life-threatening diseases is one of the key goals in medical research. In many trials, a drug that works effectively for one individual may not work or may cause serious adverse reactions for another. The classical "one-size-fits-all" approach is not appropriate if responses to the drug are heterogeneous among individuals. For instance, a significant proportion of treated patients with anti-thrombotic therapy for cardiovascular diseases suffer a new thrombotic event (Marin et al. (2009)), and patients with different levels of psychiatric symptoms show heterogeneity in treatment responses (Piper et al. (1995)). Precision medicine seeks solutions to such challenges by determining optimal patient-tailored treatments for a given disease.

A dynamic treatment regime (DTR) is a sequence of decision rules, one per treatment decision, that provides the mechanism by which patient's key features, called tailoring variables, are translated into dosage level or intervention type. DTRs, also known as adaptive interventions or multistage treatment strategies, operationalize sequential decision making with the goal of improving patient outcome over time. Instead of assigning the same treatment to all individuals, a treatment policy may assign different treatment types or dosages across patients and across time according to patient's evolving status. This concept has been adopted in a variety of health domains, such as depression (Lavori, Dawson and Rush (2000), Murphy et al. (2007), Pineau et al. (2007)), diabetes (Zhao et al. (2020)) and HIV infection (Robins, Orellana and Rotnitzky (2008), Jiang et al. (2017)). A DTR is considered optimal if, when implemented, it optimizes the expected desired cumulative outcome over the study population.

Various statistical methods have been proposed to estimate the optimal DTRs. These methods can be classified into two categories: the indirect approach and the direct approach. In

Received November 2020; revised October 2021.
MSC2020 subject classifications. Primary 62H99, 62J07; secondary 62P10.
Key words and phrases. Personalized medicine, regression-based learning, treatment decision making, variable selection, backward induction.

the indirect approach, one estimates the full or part of the conditional outcome model given the past history at each stage, and then derives a DTR from the estimated conditional outcome model. This includes g-estimation (Robins (1989), Robins (1993), Robins (1997)), Q-learning (Ertefaie and Strawderman (2018), Laber et al. (2014), Moodie, Dean and Sun (2014), Murphy (2005), Song et al. (2015), Wallace and Moodie (2015), Watkins (1989), Ertefaie et al. (2021)) and A-learning (Blatt, Murphy and Zhu (2004), Fan, Lu and Song (2016), Murphy (2003), Robins (2004), Shi et al. (2018)). A comparison of Q- and A-learning can be found in Schulte et al. (2014). In the direct approach, researchers aim to estimate the expected outcome following a DTR using inverse probability weighting methods (Murphy, van der Laan and Robins (2001), Robins (1998), Robins, Orellana and Rotnitzky (2008)), and then choose a DTR that maximizes the estimated expected outcome within a function class. See Zhang et al. (2013) and Zhang and Zhang (2018) for directly searching a DTR that maximizes a doubly robust estimate of the expected outcome, Zhao et al. (2015), Zhou et al. (2017) and Liu et al. (2018) for the outcome weighted learning framework by replacing the indicator loss with a surrogate hinge loss in the objective function, Jiang et al. (2019) for the use of a surrogate binomial deviance loss instead of the indicator loss and Luckett et al. (2020) for an actor-critic V-learning method. Other work along this line includes tree based methods (Foster, Taylor and Ruberg (2011), Laber and Zhao (2015), Lipkovich et al. (2011), Su et al. (2008), Zhu, Zeng and Kosorok (2015)), list-based methods (Zhang et al. (2015), Zhang et al. (2018), Rudin and Ertekin (2018)), and so forth. A detailed discussion of the indirect and direct approaches can be found in Laber et al. (2014).

In this paper, we consider the development of optimal DTRs in the presence of high-dimensional covariates. Our work is motivated by two clinical trials, COPES and CODIACS (Davidson et al. (2010), Davidson et al. (2013)), that compare a centralized depression care approach with standard care for patients with depression after acute coronary syndrome using a stepped care approach. Under the stepped care approach, initial treatments were chosen based on patient's preference or standard care, and then subsequent treatments were assigned based on intermediate symptoms, resulting in different treatment sequences. In the studies, a large number of covariates were collected at baseline (e.g., SF-12 scores, affinity to serotonin), and some were repeatedly recorded over time. We aim to develop an optimal DTR composed of a sequence of intervention decision rules that dynamically map evolving patient information to a recommended treatment over time. A key issue here is to identify features that are useful in tailoring treatments among time-varying covariates and treatment history from the patients.

Constructing an optimal DTR is challenging in high-dimensional data, particularly when a large collection of prognostic factors is measured. In the single-stage decision setting, quite a few methods have been proposed to tackle this problem (see Lu, Zhang and Zeng (2013), Qi and Liu (2018), Qian and Murphy (2011), Shi, Song and Lu (2016), Tian et al. (2014), Zhao et al. (2012), Zhou et al. (2017), Zhu, Zeng and Kosorok (2015), Oh et al. (2020)). The multistage decision-making problem has been discussed in Zhao et al. (2015), Liu et al. (2018), Zhu, Zeng and Song (2019) and Shi et al. (2018). The first two of these considered outcome weighted learning with an l_2 type regularization, and the third focused on the inference of value function with Q-learning. The last considered doubly robust A-learning with variable selection, which is closely related to our work. Specifically, Shi et al. (2018) considered the use of doubly robust estimating equations for A-learning with binary treatment and adopted the Dantzig selector to achieve variable selection. However, the double robustness property is only valid under correct specification of the contrast function in the conditional outcome models, which is unlikely to be satisfied for nonterminal stages.

In the present work, we adopt an indirect regression-based approach, in particular Q- and A-learning, for estimating the optimal DTR. The conditional outcome model at each stage is estimated using (weighted) l_1-penalized least squares, such as Lasso (Tibshirani (2011)) or adaptive Lasso (Zou (2006)), backwards to achieve sparsity. We derive generalization error bounds for the mean outcome of the estimated DTR. Instead of assuming (approximately) correct specification of the conditional mean model as in other papers (e.g., Shi et al. (2018)), we explicitly incorporate potential approximation error due to model misspecification in the error bounds. The upper bounds are composed of minimized sum of the approximation error and estimation error bound of the conditional outcome model at each stage, up to a power depending on the difference in the expected outcome between optimal and suboptimal decisions. The result is further strengthened to include only approximation error of treatment-by-covariate interactions if the propensity score is known or can be consistently estimated. We shall see that in high-dimensional setting, rather than estimating the best approximation model, the goal would be to estimate a linear model that balances the approximation and estimation errors among a set of suitably sparse linear models. With appropriate choices of the tuning parameters, the estimation error achieves the best known convergence rate in existing literature on l_1-penalized regression. The theoretical derivation of our results is valid for an arbitrary number of stages and any number of treatment options at each stage.

The paper is organized as follows. In Section 2, we introduce a general framework of obtaining optimal DTR with l_1-penalized A-learning. In Section 3, we express the reduction in value in terms of Q-functions and derive a finite sample upper bound on the difference in values between the optimal and the estimated DTRs in penalized A-learning. Analogously, the penalized Q-learning framework and the corresponding finite sample upper bound are described in Section 4. In Sections 5 and 6, we compare our proposed methods with other alternative methods through extensive simulation studies and a real data example from the COPES and CODIACS trials. Discussion and conclusions are presented in Section 7. Proofs of theorems are included in the Appendix.

2. A-learning with l_1-penalization.

Consider a finite horizon decision problem with T decision points. Suppose we have data from n independent subjects. For each subject, we observe a time ordered trajectory $\{O_1, A_1, O_2, A_2, \dots, O_T, A_T, O_{T+1}\}$ from a distribution P, where A_t is the treatment assignment at time t for $t = 1, \dots, T$, O_1 contains baseline information, O_t is the information observed after treatment assignment at time $(t-1)$ and prior to time t for $t = 2, \dots, T$ and O_{T+1} is information measured after the last treatment assignment. Denote the history at time t as $H_t = (O_1, A_1, O_2, A_2, \dots, O_t)$, which takes value in space $\mathcal{H}_t$. That is, H_t contains all information available to make decision at time t. Following treatment assignment at each time point $t = 1, \dots, T$, there is a scalar outcome $Y_t = y_t(H_{t+1})$, where y_t is a known function. We assume that A_t takes values in a finite, discrete space $\mathcal{A}_t$, and Y_t is continuous that is coded so that larger values are preferred.

In this setting, a *dynamic treatment regime* (DTR) is a sequence of decision rules $\boldsymbol{\pi} = (\pi_1, \dots, \pi_T)$, where $\pi_t : \mathcal{H}_t \to \mathcal{A}_t$ takes patient's history as input, and returns a treatment as output at time t. Let $E_{\boldsymbol{\pi}}$ denote the expectation with respect to the distribution of a trajectory whereby the DTR $\boldsymbol{\pi}$ is used to determine the treatment assignment at each decision time (i.e., $A_t = \pi_t(H_t)$ for $t = 1, \dots, T$). The *value* of a DTR, denoted by $V(\boldsymbol{\pi}) \triangleq E_{\boldsymbol{\pi}}(\sum_{t=1}^{T} Y_t)$, is the expected cumulative outcome if the entire study population were to follow the regime $\boldsymbol{\pi}$. The *optimal DTR*, denoted by $\boldsymbol{\pi}^o$, is the regime that when implemented will yield the maximal value, $V(\boldsymbol{\pi}^o) = \max_{\boldsymbol{\pi}} V(\boldsymbol{\pi})$.

The goal is to use the observed data to estimate the optimal DTR, $\boldsymbol{\pi}^o$. Denote the vector of treatment decisions past history at time t by $\bar{A}_t = (A_1, \dots, A_t)$. Let $Y_t^*(\bar{A}_t)$ be the potential outcome corresponding to the treatment pattern $\bar{A}_T$. Denote the potential information prior to time interval t of (past) treatments by $O_t^*(\bar{A}_{t-1})$. All subject's potential outcomes are denoted by $\mathcal{W} = \{O_2^*(a_1), \dots, O_{T+1}^*(\bar{a}_T), Y_1^*(a_1), \dots, Y_T^*(\bar{a}_T); \bar{a}_T \in \bar{\mathcal{A}}_T\}$. Throughout the article, we assume the following:

(C1) The stable unit treatment value assumption (SUTVA) holds; that is, $Y_t = Y_t^*(\bar{A}_t)$ and $O_{t+1} = O_{t+1}^*(\bar{A}_t)$, $t = 1, \dots, T$.

(C2) There are no unmeasured confounders, which is also known as sequential ignorability (Murphy (2003), Murphy, van der Laan and Robins (2001), Robins (1997)). It implies $A_t \perp\!\!\!\perp \mathcal{W} | H_t$ for each $t = 1, \dots, T$.

(C3) There is some positive constant S such that the propensity score satisfies $p_t(a_t|h_t) \triangleq P(A_t = a_t | H_t = h_t) \geq S^{-1}$ for all pairs $(h_t, a_t) \in \mathcal{H}_t \times \mathcal{A}_t$, $t = 1, \dots, T$.

Let E denote the expectation with respect to the distribution P. As demonstrated in Murphy (2005), the optimal DTR is related to optimal Q-functions via Bellman optimality equations. Specifically, define the optimal Q-functions

$$Q_T^o(h_T, a_T) = E(Y_T | H_T = h_T, A_T = a_T),$$

and

$$Q_t^o(h_t, a_t) = E\Big[Y_t + \max_{a_{t+1} \in \mathcal{A}_{t+1}} Q_{t+1}^o(H_{t+1}, a_{t+1}) \big| H_t = h_t, A_t = a_t\Big] \quad \text{for } t = T-1, \dots, 1,$$

where Q stands for "quality" of the decision based on the past history. Then, by backward induction, the optimal DTR $\boldsymbol{\pi}^o = (\pi_1^o, \dots, \pi_T^o)$ satisfies

$$\pi_t^o(h_t) = \arg\max_{a_t \in \mathcal{A}_t} Q_t^o(h_t, a_t)$$

for $t = 1, \dots, T$.

Quite a few methods have been proposed based on the above arguments. Q-learning is one of the most popular approaches. It aims to estimate the optimal Q-functions backwards sequentially using regression and construct the optimal DTR by choosing a treatment that maximizes the estimated Q-functions. In contrast to Q-learning, A-learning is motivated by the fact that the optimal decisions only depend on the interaction between history and treatment in the Q-functions. Murphy (2003) and Blatt, Murphy and Zhu (2004) proposed an iterative minimization method to directly estimate the interaction part, and Robins (2004) proposed a g-estimating equation, which can be used to produce consistent estimate of the treatment-by-history interaction if either the main effect of history on outcome or the propensity score model is correctly specified. Details and comparison of the two versions of A-learning can be found in Moodie, Richardson and Stephens (2007).

In this paper, we adopt the framework in Blatt, Murphy and Zhu (2004). Note that the optimal Q-function at each stage can be decomposed as

$$Q_t^o(H_t, A_t) = M_t^o(H_t) + U_t^o(H_t, A_t),$$

where $M_t^o(H_t) = E[Q_t^o(H_t, A_t)|H_t]$ is the main effect of H_t and $U_t^o(H_t, A_t) = Q_t^o(H_t, A_t) - E[Q_t^o(H_t, A_t)|H_t]$ is the centered treatment effect at H_t. Thus, the optimal stage-t decision only depends on U_t^o.

We propose to model $M_t^o(H_t)$ and $U_t^o(H_t, A_t)$ by $\Phi_{t1}^\mathsf{T}(H_t)\boldsymbol{\theta}_{t1}$ and $\Phi_{t2}^\mathsf{T}(H_t, A_t)\boldsymbol{\theta}_{t2}$, respectively, where $\Phi_{t1} \in \mathbb{R}^{J_{t1}}$ is a vector summary of H_t, $\Phi_{t2} \in \mathbb{R}^{J_{t2}}$ is a vector summary of (H_t, A_t), and $\boldsymbol{\theta}_{t1}$ and $\boldsymbol{\theta}_{t2}$ are the corresponding parameters. Since $E[U_t^o(H_t, A_t)|H_t] = 0$, in practical implementation, we center $\Phi_{t2}^\mathsf{T}(H_t, A_t)$ by its conditional mean $E[\Phi_{t2}^\mathsf{T}(H_t, A_t)|H_t]$. This can be easily done in sequentially randomized trials where the propensity score is known. Otherwise, we can plug in the propensity score estimate.

Denote $\Phi_t(H_t, A_t) = (\Phi_{t1}(H_t)^\mathsf{T}, \Phi_{t2}(H_t, A_t)^\mathsf{T})^\mathsf{T}$. This gives a working model for Q-function

$$Q_t(H_t, A_t; \boldsymbol{\theta}_t) = \Phi_t(H_t, A_t)^\mathsf{T}\boldsymbol{\theta}_t = \Phi_{t1}(H_t)^\mathsf{T}\boldsymbol{\theta}_{t1} + \Phi_{t2}(H_t, A_t)^\mathsf{T}\boldsymbol{\theta}_{t2}, \tag{2.1}$$

where $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_{t1}^{\mathsf{T}}, \boldsymbol{\theta}_{t2}^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{J_t}$ is the parameter of interest with $J_t = J_{t1} + J_{t2}$. By the definition of the optimal Q-functions, we can verify that

$$Q_t^o(h_t, a_t) = E\left\{Y_t + \sum_{s=t+1}^{T} \left[Y_s + \max_{a_s \in \mathcal{A}_s} Q_s^o(H_s, a_s) - Q_s^o(H_s, A_s)\right] \Big| H_t = h_t, A_t = a_t\right\}$$

for $t = T-1, \dots, 1$. Thus, the estimate of $\boldsymbol{\theta}_t$ can be obtained by regressing an estimate of $Y_t + \sum_{s=t+1}^{T}[Y_s + \max_{a_s \in \mathcal{A}_s} Q_s^o(H_s, a_s) - Q_s^o(H_s, A_s)]$ against $Q_t(H_t, A_t; \boldsymbol{\theta}_t)$. To address the high-dimensionality problem, we propose to use regression with a Lasso-type penalty. The penalized A-learning algorithm is as follows:

1. At stage T, estimate $\boldsymbol{\theta}_T$ by

$$\hat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}_T} \left\{\mathbb{P}_n[Y_T - \Phi_T(H_T, A_T)^{\mathsf{T}} \boldsymbol{\theta}_T]^2 + \lambda_T \sum_{j=1}^{J_T} w_{Tj} |\theta_{Tj}|\right\},$$

   where $\mathbb{P}_n$ denote the empirical average over n subjects, $w_{Tj} \ge 0$ is the weight for θ_{Tj}, the jth component of $\boldsymbol{\theta}_T$, and λ_T is a tuning parameter that controls model complexity.
2. For $t = T-1, \dots, 1$,

   (a) construct the pseudo outcome

$$\begin{aligned}\tilde{Y}_t &= Y_t + \sum_{s=t+1}^{T} \left[Y_s + \max_{a_s} \Phi_s^{\mathsf{T}}(H_s, a_s)\hat{\boldsymbol{\theta}}_s - \Phi_s^{\mathsf{T}}(H_s, A_s)\hat{\boldsymbol{\theta}}_s\right] \\ &= Y_t + \sum_{s=t+1}^{T} \left[Y_s + \max_{a_s} \Phi_{s2}^{\mathsf{T}}(H_s, a_s)\hat{\boldsymbol{\theta}}_{s2} - \Phi_{s2}^{\mathsf{T}}(H_s, A_s)\hat{\boldsymbol{\theta}}_{s2}\right];\end{aligned}$$

   (b) estimate $\boldsymbol{\theta}_t$ by

$$\hat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}_t} \left\{\mathbb{P}_n[\tilde{Y}_t - \Phi_t^{\mathsf{T}}(H_t, A_t)\boldsymbol{\theta}_t]^2 + \lambda_t \sum_{j=1}^{J_t} w_{tj} |\theta_{tj}|\right\},$$

   where $w_{tj} \ge 0$ is the weight for the jth component of $\boldsymbol{\theta}_t$, and λ_t is a tuning parameter.
3. The estimated DTR is $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \dots, \hat{\pi}_T)$ satisfies

$$\hat{\pi}_t(H_t) \in \arg\max_{a_t}(\Phi_t^{\mathsf{T}}\hat{\boldsymbol{\theta}}_t) = \arg\max_{a_t}(\Phi_{t2}^{\mathsf{T}}\hat{\boldsymbol{\theta}}_{t2}), \quad t = 1, \dots, T.$$

The weights w_{tj}'s in the above algorithm are used to adjust level of penalization on individual variables. For example, the weights can be set to zero for a prespecified set of clinically important variables. Alternatively, the weights could be data dependent. For example, in adaptive Lasso, the weights are set to be inverse proportional to the magnitude of the ordinary least square or elastic net estimate of the coefficients, so that the weights are not too large for truly nonzero coefficients and not too small for zero coefficients.

3. Generalization error bounds for A-learning.

In this section, we provide generalization error bounds of the estimated DTR through l_1-penalized A-learning. Note that the validity of the error bounds do not depend on correct specification of the Q-functions. However, if the Q-functions or the treatment-by-history interaction in the Q-functions are correctly specified, then the upper bounds imply that the estimated DTR is consistent. The upper bounds provide a convergence rate as well. First, we examine the relationship between value and the Q-functions, and then we provide a finite sample upper bound on the difference in values between the optimal and the estimated DTRs.

3.1. *Relationship between value and Q-functions.* For any DTR $\boldsymbol{\pi} = (\pi_1, \dots, \pi_T)$ and any square integrable functions $\{Q_t(H_t, A_t) : t = 1, \dots, T\}$ such that $\pi_t(H_t) \in \arg\max_{a_t} Q_t(H_t, a_t)$, Murphy (2005) showed that

$$V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) \le \sum_{t=1}^{T} 2S^{t/2}\{E[(Q_t(H_t, A_t) - Q_t^o(H_t, A_t))^2]\}^{1/2} \tag{3.1}$$

under conditions (C1)–(C3). The left-hand side of (3.1) is the reduction in value of the DTR $\boldsymbol{\pi}$ as compared to the optimal DTR, and the right-hand side measures the distance between Q_t and the optimal Q-functions.

In the theorem below, we derive several sharper upper bounds under a margin type condition. First, we show that an upper bound with exponent larger than 1/2 can be obtained under a low noise condition, which implicitly implies a faster rate of convergence. Second, as we have discussed previously, the optimal decision only depends on the interaction between treatment and history, and thus our second bound only involves the model for $U^o(H_t, A_t)$ on the right-hand side of the upper bound.

THEOREM 1. *Suppose conditions* (C1)–(C3) *hold. Assume there exist some constants* $C > 0$ *and* $\alpha \ge 0$ *such that*

$$P\Big(\max_{a_t \in \mathcal{A}_t} Q_t^o(H_t, a_t) - \max_{a_t \in \mathcal{A}_t \setminus \arg\max_{a_t} Q_t^o(H_t, a_t)} Q_t^o(H_t, a_t) \le \epsilon\Big) \le C\epsilon_t^{\alpha} \tag{3.2}$$

for all positive ϵ_t *for* $t = 1, \dots, T$. *Then for any dynamic treatment regime* $\boldsymbol{\pi} = (\pi_1, \dots, \pi_T)$ *and sequence of square integrable functions* $\{Q_t(H_t, A_t) : t = 1, \dots, T\}$ *such that* $\pi_t(H_t) \in \arg\max_{a_t} Q_t(H_t, a_t)$, $t = 1, \dots, T$, *we have*

$$V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) \le \sum_{t=1}^{T} C_{1,t}\{E[Q_t(H_t, A_t) - Q_t^o(H_t, A_t)]^2\}^{(1+\alpha)/(2+\alpha)}. \tag{3.3}$$

Furthermore, for any square integrable function $U_t(H_t, A_t)$ *such that* $\arg\max_{a_t} Q_t(H_t, a_t) = \arg\max_{a_t} U_t(H_t, a_t)$ *for* $t = 1, \dots, T$, *we have*

$$V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) \le \sum_{t=1}^{T} C_{1,t}\{E[U_t(H_t, A_t) - U_t^o(H_t, A_t)]^2\}^{(1+\alpha)/(2+\alpha)}, \tag{3.4}$$

where $C_{1,t} = (2+\alpha)[2^{2\alpha}(1+\alpha)^{-(1+\alpha)}S^{(2+\alpha)t-1}C]^{1/(2+\alpha)}$.

REMARK. Condition (3.2) is a margin type condition, which is similar to the margin assumptions that are widely used in the classification context (Tsybakov (2004), Gey (2012)). In a related line of research in contextual bandits, similar conditions are used for the gap between the best and suboptimal arms; see, for example, Lattimore and Munos (2014), Bubeck, Perchet and Rigollet (2013) and references therein. Condition (3.2) measures the difference in mean outcomes between the t-stage optimal action(s) and the t-stage best suboptimal action(s) at H_t. For instance, if $\max_{a_t \in \mathcal{A}_t} Q_t^o(H_t, a_t) - \max_{a_t \in \mathcal{A}_t \setminus \arg\max_{a_t} Q_t^o(H_t, a_t)} Q_t^o(H_t, a_t)$ has bounded continuous density, then condition (3.2) holds with $\alpha = 1$. This condition also holds for positive α in many scenarios where $Q_t^o(H_t, a_t)$ has a mixture of continuous and discrete distributions. Clearly, condition (3.2) always holds for $C = 1$ and $\alpha = 0$. In this case, (3.3) reduces to (3.1) up to a constant; see Qian and Murphy (2011) for more discussion of the condition.

3.2. *Quality of the estimated DTR.* In this section, we provide finite sample upper bounds on the difference between the optimal value and the value obtained by our estimator in terms of the prediction errors resulting from the estimation of Q_t^o or U_t^o for $t = 1, \dots, T$. These upper bounds guarantee that if Q_t^o (or U_t^o) is consistently specified for $t = 1, \dots, T$, the value of the estimated DTR will converge to the optimal value.

Define

$$\boldsymbol{\theta}_T^* = \arg\min_{\boldsymbol{\theta}_T \in \mathbb{R}^{J_T}} E\big[Y_T - \Phi_T^{\mathsf{T}}(H_T, A_T)\boldsymbol{\theta}_T\big]^2$$

$$\text{and} \quad \boldsymbol{\theta}_t^* = \arg\min_{\boldsymbol{\theta}_t \in \mathbb{R}^{J_t}} E\left\{Y_t + \sum_{s=t+1}^{T}\Big[Y_s - \Phi_s^{\mathsf{T}}(H_s, A_s)\boldsymbol{\theta}_s^* + \max_{a_s} \Phi_s^{\mathsf{T}}(H_s, a_s)\boldsymbol{\theta}_s^*\Big] - \Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t\right\}^2 \tag{3.5}$$

for $t = T-1, \dots, 1$. Note that $\Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t^*$ represents the best linear approximation of $Q_t^o(H_t, A_t)$. For expositional simplicity, assume that $\boldsymbol{\theta}_t^*$ is unique for $t = 1, \dots, T$. Results for nonunique $\boldsymbol{\theta}_t^*$'s can be obtained with slight modification of the assumptions as stated in Qian and Murphy (2011). Denote $\bar{w}_{tj} = w_{tj} + 1_{w_{tj}=0}$. Our results rely on the following assumptions:

(A1) Define error terms $\varepsilon_{Ti} = Y_{Ti} - Q_T^o(H_{Ti}, A_{Ti})$, and $\varepsilon_{ti} = Y_{ti} + \sum_{s=t+1}^T [Y_{si} + \max_{a_s} Q_s^o(H_{si}, a_{si}) - Q_s^o(H_{si}, A_{si})] - Q_t^o(H_{ti}, A_{ti})$ for $t = T-1, \dots, 1$. At each stage t, we assume the error terms ε_{ti}, $i = 1, \dots, n$, are independent of (H_{ti}, A_{ti}), $i = 1, \dots, n$ and are i.i.d. with $E(\varepsilon_{ti}) = 0$ and $E[|\varepsilon_{ti}|^l] \le l! c^{l-2}\sigma^2/2$ for some $c, \sigma^2 > 0$ for all $l \ge 2$.

(A2) For $t = 1, \dots, T$, the matrix $E[(\phi_{t1}/\bar{w}_{t1}, \dots, \phi_{tJ_t}/\bar{w}_{tJ_t})^{\mathsf{T}}(\phi_{t1}/\bar{w}_{t1}, \dots, \phi_{tJ_t}/\bar{w}_{tJ_t})]$ is positive definite with the smallest eigenvalue $\tau_t > 0$.

(A3) There exist finite, positive constants η and u such that $\max_{t \in \{1,\dots,T\}} \|Q_t^o - \Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t^*\|_\infty \le \eta$ and $\max_{j \in \{1,\dots,J_t\}, t \in \{1,\dots,T\}} \{\|\phi_{tj}\|_\infty / \bar{w}_{tj}\} \le u$.

(A4) There exists a positive constant b such that $\max_{j \in \{1,\dots,J_t\}, t \in \{1,\dots,T\}} E[\phi_{tj}/\bar{w}_{tj}]^2 \le b^2$.

For any $\boldsymbol{\theta}_t \in \mathbb{R}^{J_t}$, $t = 1, \dots, T$, define the index set

$$I_t(\boldsymbol{\theta}_t) = \big\{j \in \{1, \dots, J_t\} : w_{tj} = 0 \text{ or } \theta_{tj} \ne 0\big\}.$$

Intuitively, $I_t(\boldsymbol{\theta}_t)$ can be viewed as a sparsity measure that indices either nonzero elements in $\boldsymbol{\theta}_t$ or nonpenalized terms.

Further define the set

$$\Theta = \left\{(\boldsymbol{\theta}_1^{\mathsf{T}}, \dots, \boldsymbol{\theta}_T^{\mathsf{T}})^{\mathsf{T}} \in \prod_{t=1}^T \mathbb{R}^{J_t} : \max_{t \in \{1,\dots,T\}} \|\Phi_t^{\mathsf{T}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)\|_\infty \le \eta \right.$$

$$\max_{t \in \{1,\dots,T\}} E\big[\Phi_t^{\mathsf{T}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)\big]^2/\lambda_t^2 \le (21b-8)^{-2} \tag{3.6}$$

$$\left. \text{and} \max_{t \in \{1,\dots,T\}} \big\{|I_t(\boldsymbol{\theta}_t)|/\tau_t\big\} \le \frac{(21b-10)^2}{288bu(21b-8)^2}\sqrt{\frac{n}{2\log[3TJ(J+1)n]}}\right\},$$

where $J = \max_{t \in \{1,\dots,T\}} J_t$ and $|I_t(\boldsymbol{\theta}_t)|$ is the cardinality of $I_t(\boldsymbol{\theta}_t)$.

The set Θ contains sparse parameters that are close to the best $\boldsymbol{\theta}_t^*$'s. Thus, it can be viewed as an oracle parameter set in the sense that parameters in the set have balanced sparsity and prediction performance. Note that Θ is nonempty when sample size n is large as long as J, the maximum number of parameters at each stage, does not grow too fast with n.

Below, we provide finite sample upper bounds for the difference in value of the optimal DTR and the value of the estimated DTR. The first upper bound is presented in terms of the approximation and estimation errors of Q_t^o, the optimal Q-functions. Furthermore, if $E[\Phi_{t2}^{\mathsf{T}}(H_t, A_t)|H_t] = \mathbf{0}$ a.s., then the upper bound can be further strengthened by involving only the approximation and estimation errors in U_t^o, which is the centered treatment effect part in Q_t^o.

THEOREM 2. *Suppose conditions* (*C*1)–(*C*3) *hold, and the margin condition* (3.2) *holds for some* $C > 0$, $\alpha \geq 0$ *and all positive* ϵ_t *for* $t = 1, \dots, T$. *Assume assumptions* (*A*1)–(*A*4) *hold. Suppose the tuning parameters* λ_t, $t = 1, \dots, T$, *satisfy*

$$\lambda_t \geq 96\sqrt{2}[1 + 2(T - t)]b \max\{c, \sigma, \eta\}\sqrt{\frac{\log(12TJ_t n)}{n}}, \tag{3.7}$$

and $\lambda_t^2 \geq c_{t,s}\lambda_s^2$ *for* $t = 1, \dots, T$, $s = t, \dots, T$, *where* $c_{t,t} = 1$, $c_{t,s} = 2(105b - 38)(5S + 3)(T - t)^2 c_{t+1,s}/[9(21b - 8)]$. *Let* Θ *be the set defined in* (3.6) *and assume* Θ *is nonempty. Then for any* $n \geq 8u^2 \log(12TJn)/(9b^2)$, *with the probability at least* $1 - 1/n$ *we have*

$$\begin{aligned} &V(\boldsymbol{\pi}^o) - V(\hat{\boldsymbol{\pi}}) \\ &\quad \leq \min_{(\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_T^\top)^\top \in \Theta} \Biggl[\sum_{t=1}^{T} C_{1,t}\biggl(E[\Phi_t^\top \boldsymbol{\theta}_t - Q_t^o]^2 + K_1 \max_{s \in \{t, \dots, T\}}\biggl\{c_{t,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\biggr\}\biggr)^{(1+\alpha)/(2+\alpha)}\Biggr], \end{aligned}$$

where $J = \max_t J_t$, $C_{1,t} = (2 + \alpha)[2^{2\alpha}(1 + \alpha)^{-(1+\alpha)} S^{(2+\alpha)t-1} C]^{1/(2+\alpha)}$ *and* $K_1 = [64 \times (105b - 38)^2]/[81(21b - 8)^2] + [32b(105b - 38)]/[3(21b - 8)(21b - 10)]$.

Furthermore, suppose $E[\Phi_{t2}^\top(H_t, A_t)|H_t] = \mathbf{0}$ *a.s. for* $t = 1, \dots, T$. *Then with the probability at least* $1 - 1/n$,

$$\begin{aligned} &V(\boldsymbol{\pi}^o) - V(\hat{\boldsymbol{\pi}}) \\ &\quad \leq \min_{(\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_T^\top)^\top \in \Theta} \Biggl[\sum_{t=1}^{T} C_{1,t}\biggl(E[\Phi_{t2}^\top \boldsymbol{\theta}_{t2} - U_t^o]^2 \\ &\qquad + K_2 \max_{s \in \{t, \dots, T\}}\biggl\{\bar{c}_{t,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\biggr\}\biggr)^{(1+\alpha)/(2+\alpha)}\Biggr], \end{aligned} \tag{3.8}$$

where $I_{t2}(\boldsymbol{\theta}_t) = I_t(\boldsymbol{\theta}_t) \cap \{J_{t1} + 1, \dots, J_t\}$, $K_2 = [3 - (21b - 10)^2/[9(21b - 8)^2]]^2 + [2b/(21b - 8)][81(21b - 8)^2/(21b - 10)^2 - 3]$, $\bar{c}_{t,t} = 1$, *and* $\bar{c}_{t,s} = 2(T - t)^2(S + 1)\{81(21b - 8)^2 \max_{s \in \{t+1, \dots, T\}}\{\bar{c}_{t+1,s}/c_{t+1,s}\}/[16(21b - 10)^2] + 1\}\{3 - (21b - 10)^2/[9 \times (21b - 8)^2]\}\bar{c}_{t+1,s}$ *for* $t = 1, \dots, T$, $s = t + 1, \dots, T$.

PROOF. The result follows from the inequalities (3.3) and (3.4) in Theorem 1 and inequalities (C.5) and (C.6) in Theorem 4 in the Appendix with $\varphi = \log(nT)$ and $\gamma = 1/(21b - 8)$, and noticing that

$$\begin{aligned} &\frac{(21b - 10)^2}{288bu(21b - 8)^2}\sqrt{\frac{n}{2\log[3TJ(J + 1)n]}} \\ &\quad \leq \frac{(21b - 10)^2}{144b(21b - 8)^2}\Biggl[\sqrt{\frac{1}{9b^2} + \frac{n}{2u^2\log[3TJ(J + 1)n]}} - \frac{1}{3b}\Biggr] \end{aligned}$$

under the condition $n \geq 8u^2 \log(12TJn)/(9b^2)$. □

REMARKS.

1. Assumption (A1) implies that the error terms do not have heavy tails. Assumptions (A1) and (A3) are needed to show that the sample mean is concentrated around the true mean. Assumption (A2) is used to avoid collinearity. In addition, for any $\boldsymbol{\theta}_t, \acute{\boldsymbol{\theta}}_t \in \mathbb{R}^{J_t}$, one can easily verify that $E[\Phi_t^\top(\acute{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 = (W_t(\acute{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t))^\top M_t W_t(\acute{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t) \geq \tau_t(\sum_{j \in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\acute{\theta}_{tj} - \theta_{tj}|)^2/|I_t(\boldsymbol{\theta}_t)|$ by eigendecomposition and simple algebra, where $W_t = \operatorname{diag}\{\bar{w}_{t1}, \dots, \bar{w}_{tJ_t}\}$

and M_t is the gram-matrix provided in Assumption (A2). Thus, Assumption (A2) is a sufficient condition for

$$E\big[\Phi_t^{\mathsf{T}}(\acute{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)\big]^2 |I_t(\boldsymbol{\theta}_t)| \geq \tau_t\bigg(\sum_{j \in I_t(\boldsymbol{\theta}_t)} \bar{w}_{tj}|\acute{\theta}_{tj} - \theta_{tj}|\bigg)^2, \tag{3.9}$$

for any $\boldsymbol{\theta}_t, \acute{\boldsymbol{\theta}}_t \in \mathbb{R}^{J_t}$; see, for example, van de Geer (2008) for more details. Condition (3.9) is employed in the proofs of Lemmas 1 and 2 in the supplementary material (Oh et al. (2022)). This condition holds if the correlation $|E\phi_{tj}\phi_{tk}|/(\bar{w}_{tj}\bar{w}_{tk})$ is small for all $k \in I_t(\boldsymbol{\theta}_t)$, $j \neq k$. Assumption (A4) is used to ensure

$$\max_j \big|E\big[\Phi_t^{\mathsf{T}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)\phi_{tj}/\bar{w}_{tj}\big]\big| \leq \gamma\lambda_t b, \tag{3.10}$$

for $\boldsymbol{\theta}_t \in \Theta_t$ to derive Theorem 4 in the Appendix. When $\bar{w}_{tj} = (E\phi_{tj}^2)^{1/2}$ as in Qian and Murphy (2011), condition (3.10) is satisfied with $b = 1$.

2. The validity of (3.8) requires that $E[\Phi_{t2}^{\mathsf{T}}(H_t, A_t)|H_t] = \mathbf{0}$ a.s. for $t = 1, \dots, T$. This can be easily achieved in sequential randomized trials, where treatment allocation probabilities $P(A_t = a_t|H_t = h_t)$ are known. In observation studies, similar results can be obtained if the treatment allocation probabilities can be consistently estimated at $\sqrt{n}$ rate.

3. The first term in the generalization error bounds is called the approximation error, and the second term is the estimation error, which provides the convergence rate. When Q_t^o or U_t^o is well approximated, the fastest convergence rate can be achieved by choosing the tuning parameter as $\lambda_s^2 = O(\max_{s' \in \{s,\dots,T\}}(\log(J_{s'}n))/n)$. Also, the convergence rate is affected by $|I_s(\boldsymbol{\theta}_s)|$. To our best knowledge, this is the sharpest convergence rate for Lasso estimators. Also, suppose either $Q_t^o = \Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t^*$ or $U_t^o = \Phi_{t2}^{\mathsf{T}}\boldsymbol{\theta}_{t2}^*$. Then the generalization error bounds imply $V(\boldsymbol{\pi}^o) - V(\hat{\boldsymbol{\pi}}) \leq O_P((\max_{t \in \{1,\dots,T\}} |I_t(\boldsymbol{\theta}_t)|(\log(J_t n))/n)^{(1+\alpha)/(2+\alpha)})$.

4. The number of stages, T, affects the sharpness of the theoretical bounds. Each stage, t, plays a role in $C_{1,t}$, $c_{t,s}$, and $\bar{c}_{t,s}$ given in the upper bounds. $C_{1,t}$ involves a factor of S^t, and both $c_{t,s}$ and $\bar{c}_{t,s}$ in the estimation error involve a factor of $((T-t)!)^2$. Finally, the effect of T is reflected by the summation from $t = 1$ to $t = T$.

5. In the proof, the weights w_{tj} are assumed to be data independent and finite. As discussed before, w_{tj} can also be data dependent such as $w_{tj} = (|\hat{\theta}_{tj}(\text{enet})| + 1/\min\{n, c\})^{-1}$, where $\hat{\theta}_{tj}(\text{enet})$ is the elastic net estimate of θ_{tj} and c is a sufficiently large constant. Similar results can be obtained if the data dependent weight converges to a bounded constant.

4. Q-learning with l_1-penalization.

In what follows, we introduce a framework for obtaining optimal dynamic treatment regimes in penalized Q-learning. It estimates the conditional outcome model using (weighted) l_1-penalized least squares at each stage backwards, which is analogous to penalized A-learning. The main difference between penalized Q- and A-learning is the construction of pseudo-outcome at nonterminal stages. The penalized Q-learning algorithm is given below:

1. At stage T, estimate $\boldsymbol{\theta}_T$ by

$$\hat{\boldsymbol{\theta}}_T^Q = \arg\min_{\boldsymbol{\theta}_T}\Bigg\{\mathbb{P}_n\big[Y_T - \Phi_T(H_T, A_T)^{\mathsf{T}}\boldsymbol{\theta}_T\big]^2 + \lambda_T^Q \sum_{j=1}^{J_T} w_{Tj}|\theta_{Tj}|\Bigg\},$$

   where $\mathbb{P}_n$ denote the empirical average over n subjects, $w_{Tj} \geq 0$ is the weight for θ_{Tj}, the jth component of $\boldsymbol{\theta}_T$, and λ_T^Q is a tuning parameter that controls model complexity.
2. For $t = T-1, \dots, 1$,

(a) construct the pseudo outcome

$$\begin{aligned}\tilde{Y}_t^Q &= Y_t + \max_{a_{t+1}} \Phi_{t+1}^{\mathsf{T}}(H_{t+1}, a_{t+1})\hat{\boldsymbol{\theta}}_{t+1}^Q \\ &= Y_t + \Phi_{(t+1)1}^{\mathsf{T}}(H_{t+1})\hat{\boldsymbol{\theta}}_{(t+1)1}^Q + \max_{a_{t+1}} \Phi_{(t+1)2}^{\mathsf{T}}(H_{t+1}, a_{t+1})\hat{\boldsymbol{\theta}}_{(t+1)2}^Q;\end{aligned}$$

(b) estimate $\boldsymbol{\theta}_t$ by

$$\hat{\boldsymbol{\theta}}_t^Q = \arg\min_{\boldsymbol{\theta}_t}\left\{\mathbb{P}_n\left[\tilde{Y}_t^Q - \Phi_t^{\mathsf{T}}(H_t, A_t)\boldsymbol{\theta}_t\right]^2 + \lambda_t^Q \sum_{j=1}^{J_t} w_{tj}|\theta_{tj}|\right\},$$

where $w_{tj} \geq 0$ is the weight for the jth component of $\boldsymbol{\theta}_t$, and λ_t^Q is a tuning parameter.

3. The estimated DTR is $\hat{\boldsymbol{\pi}}^Q = (\hat{\pi}_1^Q, \dots, \hat{\pi}_T^Q)$ satisfies

$$\hat{\pi}_t^Q(H_t) \in \arg\max_{a_t}\left(\Phi_t^{\mathsf{T}}\hat{\boldsymbol{\theta}}_t^Q\right) = \arg\max_{a_t}\left(\Phi_{t2}^{\mathsf{T}}\hat{\boldsymbol{\theta}}_{t2}^Q\right), \quad t = 1, \dots, T.$$

Next, we show that the finite sample upper bound on the difference in values between the optimal and the estimated DTRs in penalized Q-learning. Note that the pseudo-outcome in Q-learning is $\tilde{Y}_t^Q = Y_t + \max_{a_{t+1}} \Phi_{t+1}^{\mathsf{T}}(H_{t+1}, a_{t+1})\hat{\boldsymbol{\theta}}_{t+1}^Q$, which involves $\Phi_{(t+1)1}^{\mathsf{T}}(H_{t+1})\hat{\boldsymbol{\theta}}_{(t+1)1}^Q$, estimate of the main effect of H_{t+1} at stage $t+1$. Therefore, if the main effect is misspecified, then the pseudo-outcome will be biased. Based on the theorem below, we observe that penalized Q-learning does not have a sharper bound, which involves only the treatment-by-history terms like penalized A-learning.

Define

$$\begin{aligned}&\boldsymbol{\theta}_T^{*Q} = \arg\min_{\boldsymbol{\theta}_T \in \mathbb{R}^{J_T}} E\left[Y_T - \Phi_T^{\mathsf{T}}(H_T, A_T)\boldsymbol{\theta}_T\right]^2 \\ &\text{and}\quad \boldsymbol{\theta}_t^{*Q} = \arg\min_{\boldsymbol{\theta}_t \in \mathbb{R}^{J_t}} E\left[Y_t + \max_{a_{t+1}} \Phi_{t+1}^{\mathsf{T}}(H_{t+1}, a_{t+1})\boldsymbol{\theta}_{t+1}^{*Q} - \Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t\right]^2\end{aligned} \tag{4.1}$$

for $t = T-1, \dots, 1$. Denote $\bar{w}_{tj} = w_{tj} + 1_{w_{tj}=0}$. We state the following assumptions.

(B1) Define error terms $\varepsilon_{Ti}^Q = Y_{Ti} - Q_T^o(H_{Ti}, A_{Ti})$, and $\varepsilon_{ti}^Q = Y_{ti} + \max_{a_{t+1}} Q_{t+1}^o(H_{(t+1)i}, a_{t+1)i}) - Q_t^o(H_{ti}, A_{ti})$ for $t = T-1, \dots, 1$. At each stage t, we assume the error terms ε_{ti}^Q, $i = 1, \dots, n$, are independent of (H_{ti}, A_{ti}), $i = 1, \dots, n$ and are i.i.d. with $E(\varepsilon_{ti}^Q) = 0$ and $E[|\varepsilon_{ti}^Q|^l] \leq l!c^{l-2}\sigma^2/2$ for some $c, \sigma^2 > 0$ for all $l \geq 2$.

(B2) For $t = 1, \dots, T$, the matrix $E[(\phi_{t1}/\bar{w}_{t1}, \dots, \phi_{tJ_t}/\bar{w}_{tJ_t})^{\mathsf{T}}(\phi_{t1}/\bar{w}_{t1}, \dots, \phi_{tJ_t}/\bar{w}_{tJ_t})]$ is positive definite with the smallest eigenvalue $\tau_t > 0$.

(B3) There exist finite, positive constants η and u such that $\max_{t\in\{1,\dots,T\}} \|Q_t^o - \Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t^{*Q}\|_\infty \leq \eta$ and $\max_{j\in\{1,\dots,J_t\}, t\in\{1,\dots,T\}}\{\|\phi_{tj}\|_\infty/\bar{w}_{tj}\} \leq u$.

(B4) There exists a positive constant b such that $\max_{j\in\{1,\dots,J_t\}, t\in\{1,\dots,T\}} E[\phi_{tj}/\bar{w}_{tj}]^2 \leq b^2$.

Further define the set

$$\begin{aligned}\Theta^Q = \Bigg\{&(\boldsymbol{\theta}_1^{\mathsf{T}}, \dots, \boldsymbol{\theta}_T^{\mathsf{T}})^{\mathsf{T}} \in \prod_{t=1}^{T}\mathbb{R}^{J_t} : \max_{t\in\{1,\dots,T\}}\|\Phi_t^{\mathsf{T}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^{*Q})\|_\infty \leq \eta \\ &\max_{t\in\{1,\dots,T\}} E[\Phi_t^{\mathsf{T}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^{*Q})]^2/(\lambda_t^Q)^2 \leq (21b-8)^{-2} \\ &\text{and}\ \max_{t\in\{1,\dots,T\}}\{|I_t(\boldsymbol{\theta}_t)|/\tau_t\} \leq \frac{(21b-10)^2}{288bu(21b-8)^2}\sqrt{\frac{n}{2\log[3TJ(J+1)n]}}\Bigg\},\end{aligned} \tag{4.2}$$

where $J = \max_{t\in\{1,\dots,T\}} J_t$ and $|I_t(\boldsymbol{\theta}_t)|$ is the cardinality of $I_t(\boldsymbol{\theta}_t)$.

The theorem below implies that the upper bound of Q-learning guarantees that the value of the estimated DTR will converge to the optimal value if the entire Q-function Q_t^o is consistently specified for $t = 1, \dots, T$.

THEOREM 3 (Q-learning). *Suppose conditions* (*C*1)–(*C*3) *hold, and the margin condition* (3.2) *holds for some* $C > 0$, $\alpha \ge 0$ *and all positive* ϵ_t *for* $t = 1, \dots, T$. *Assume assumptions* (*B*1)–(*B*4) *hold. Suppose the tuning parameters* λ_t^Q, $t = 1, \dots, T$, *satisfy*

$$\lambda_t^Q \ge 96\sqrt{2}[1 + 2(T - t)]b \max\{c, \sigma, \eta\}\sqrt{\frac{\log(12TJ_t n)}{n}}, \tag{4.3}$$

and $(\lambda_t^Q)^2 \ge \tilde{c}_{t,s}(\lambda_s^Q)^2$ *for* $t = 1, \dots, T$, $s = t + 1$, *where* $\tilde{c}_{t,t} = 1$ *and* $\tilde{c}_{t,s} = (105b - 38)5Sc_{t+1,s}/[9(21b - 8)]$. *Let* Θ^Q *be the set defined in* (4.2) *and assume* Θ^Q *is nonempty. Then for any* $n \ge 8u^2 \log(12TJn)/(9b^2)$, *with the probability at least* $1 - 1/n$ *we have*

$$\begin{aligned} &V(\boldsymbol{\pi}^o) - V(\hat{\boldsymbol{\pi}}^Q) \\ &\quad \le \min_{(\boldsymbol{\theta}_1^\mathsf{T}, \dots, \boldsymbol{\theta}_T^\mathsf{T})^\mathsf{T} \in \Theta^Q} \Biggl[\sum_{t=1}^{T} C_{1,t}\biggl(E[\Phi_t^\mathsf{T}\boldsymbol{\theta}_t - Q_t^o]^2 \\ &\qquad + K_1 \max_{s \in \{t, t+1\}} \biggl\{\tilde{c}_{t,s}\frac{|I_s(\boldsymbol{\theta}_s)|(\lambda_s^Q)^2}{\tau_s}\biggr\}\biggr)^{(1+\alpha)/(2+\alpha)}\Biggr], \end{aligned}$$

where $J = \max_t J_t$, $C_{1,t} = (2 + \alpha)[2^{2\alpha}(1 + \alpha)^{-(1+\alpha)}S^{(2+\alpha)t-1}C]^{1/(2+\alpha)}$, $K_1 = [64(105b - 38)^2]/[81(21b - 8)^2] + [32b(105b - 38)]/[3(21b - 8)(21b - 10)]$, *and* $\max_{s \in \{t,t+1\}}$ *is defined for all* $t = 1, \dots, T$ *by understanding that the stage* $T + 1$ *is not included for the convenience of notation.*

PROOF OF THEOREM 3. The result follows from the inequalities (3.3) and (3.4) in Theorem 1 and inequality (C.5) in Theorem 5 in the supplementary material with $\varphi = \log(nT)$ and $\gamma = 1/(21b - 8)$, and noticing that

$$\begin{aligned} &\frac{(21b - 10)^2}{288bu(21b - 8)^2}\sqrt{\frac{n}{2\log[3TJ(J + 1)n]}} \\ &\quad \le \frac{(21b - 10)^2}{144b(21b - 8)^2}\biggl[\sqrt{\frac{1}{9b^2} + \frac{n}{2u^2\log[3TJ(J + 1)n]}} - \frac{1}{3b}\biggr] \end{aligned}$$

under the condition $n \ge 8u^2 \log(12TJn)/(9b^2)$. □

5. Simulation studies. In this section, we study the performance of the proposed A-learning and Q-learning methods using simulated data. For computations, we apply a technique called partial regularization via orthogonality using the adaptive Lasso (PROaL) to achieve sparsity in the prognostic factors and retain a few key variables, such as treatments, in the model. Thus, we call the methods Alearn-PROaL and Qlearn-PROaL, respectively. To apply the PRO technique, we first divide Φ_t into two parts: those need to be penalized, denoted by $X_t \in \mathbb{R}^{p_{t1}}$, and those left unpenalized, denoted by Z_t. Usually $Z_t \in \mathbb{R}^{p_{t2}}$ is low-dimensional and only includes several key variables. Then we can consider a working model as

$$Q_t(H_t, A_t; \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = Z_t^\mathsf{T}(H_t, A_t)\boldsymbol{\alpha}_t + X_t^\mathsf{T}(H_t, A_t)\boldsymbol{\beta}_t, \tag{5.1}$$

where model (5.1) is equivalent to model (2.1) by letting $\Phi_t = (Z_t, X_t)$ and $\boldsymbol{\theta}_t = (\boldsymbol{\alpha}_t^{\mathsf{T}}, \boldsymbol{\beta}_t^{\mathsf{T}})^{\mathsf{T}}$. The Alearn-PROaL aims to minimize the following stage-t objective function:

$$L_t(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \mathbb{P}_n\big[\tilde{Y}_t - Q_t(H_t, A_t; \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)\big]^2 + \lambda_t \sum_{j=1}^{p_{t1}} w_{tj}|\beta_{tj}|,$$

where $\tilde{Y}_T = Y_T$, $\tilde{Y}_t = Y_t + \sum_{s=t+1}^{T}[Y_s + \max_{a_s} Q_s(H_s, a_s; \hat{\boldsymbol{\alpha}}_s, \hat{\boldsymbol{\beta}}_s) - Q_s(H_s, A_s; \hat{\boldsymbol{\alpha}}_s, \hat{\boldsymbol{\beta}}_s)]$ for $t = T-1, \ldots, 1$, and λ_t is a tuning parameter controlling the amount of penalization at time t. Note that $\boldsymbol{w}_t = (w_{t1}, \ldots, w_{tp_{t1}})$ is a vector of weights adjusting a level of penalization on individual variables at time t. One can adopt $\hat{\boldsymbol{w}}_t = 1/|\tilde{\boldsymbol{\beta}}_t|^{\delta}$ for some $\delta > 0$ with $\tilde{\boldsymbol{\beta}}_t$ being a root-(n/p_{t1}) consistent estimator. In practice, we propose to set $\tilde{\boldsymbol{\beta}}_t$ as perturbed elastic net estimates, following Zou and Zhang (2009). The Alearn-PROaL algorithm, which imposes the adaptive Lasso penalty only on X_t but not on Z_t, is given in Appendix A. The Qlearn-PROaL algorithm can be analogously derived by changing the pseudo-outcome formulation. For comparison, we consider the following competing methods: penalized A-learning (PAL) proposed by Shi et al. (2018) and backward outcome weighted learning (BOWL) proposed by Zhao et al. (2015). For the BOWL method, we consider both linear and radial kernels, which we refer to as BOWL-linear and BOWL-radial, respectively.

We consider five scenarios with two decision points $T = 2$. In all scenarios, the treatment at stage 1, A_1, is randomly generated from Bernoulli(0.5). The baseline covariates O_1 is a p-dimensional standard normal random vector. The intermediate covariate is $O_2 \sim N(O_{11} + A_1 + A_1 O_{11}, 0.5^2)$, where O_{11} is the first component of O_1. The second stage treatments and outcomes are generated as below:

Scenario 1: $P(A_2 = 1) = 0.5$, $Y_1 = 0$, $Y_2 \sim N(A_1 A_2 + A_2(O_{13} - O_{14} + O_2), 1^2)$;
Scenario 2: $P(A_2 = 1) = 0.5$, $Y_1 = 0$, $Y_2 \sim N(A_1 A_2 + A_2(O_{13} - O_{14} + O_2) + A_1(O_{15} - O_{16}), 1^2)$;
Scenario 3: $P(A_2 = 1) = \Pr(N(0,1) \leq (O_1, O_2)^{\mathsf{T}}(\mathbf{0}_{p-2}, 1, -1, 1)^{\mathsf{T}})$, $Y_1 = 0$, $Y_2 \sim N(A_1 A_2 + A_2(O_{13} - O_{14} + O_2), 1^2)$;
Scenario 4: $P(A_2 = 1) = \Pr(N(0,1) \leq (O_1, O_2)^{\mathsf{T}}(\mathbf{0}_{p-2}, 1, -1, 1)^{\mathsf{T}})$, $Y_1 = 0$, $Y_2 \sim N(A_1 A_2 + A_2(O_{13} - O_{14} + O_2) + A_1(O_{15} - O_{16}), 1^2)$;
Scenario 5: $P(A_2 = 1) = 0.5$, $Y_1 \sim N(0.5 O_{13}(2A_1 - 1), 1^2)$, $Y_2 \sim N([(O_{11}^2 + O_{12}^2 - 0.2)(0.5 - O_{11}^2 - O_{12}^2) + Y_1](2A_2 - 1), 1^2)$.

We consider the same set of working models for the Alearn-PROaL and PAL methods. Specifically, we use $\Phi_{21}(H_2) = (1, A_1, O_1, O_2)$ and $\Phi_{22}(H_2, A_2) = \Phi_{21}(H_2) \otimes (A_2 - p_2(A_2|H_2))$ to model $M_2^o(H_2)$ and $U_2^o(H_2, A_2)$, respectively, at stage 2, and $\Phi_{11}(H_1) = (1, O_1)$ and $\Phi_{12}(H_1, A_1) = \Phi_{11}(H_1) \otimes (A_1 - p_1(A_1|H_1))$ to model $M_1^o(H_1)$ and $U_1^o(H_1, A_1)$, respectively, at stage 1. For the Qlearn-PROaL, we consider the same $\Phi_{t1}(H_t)$ as in the Alearn-PROaL, but consider $\Phi_{t2}(H_t, A_t) = \Phi_{t1}(H_t) \otimes A_t$. For the PRO methods, we choose not to penalize the intercept and A_1 in the first stage, and the intercept, A_2 and $A_1 A_2$ in the second stage. The propensity scores are estimated using a logistic regression with Lasso. The PAL method is implemented using the R package provided in Shi et al. (2018). For the implementation of BOWL, we consider the class of linear and radial kernels $f_2(H_2)$ and $f_1(H_1)$ with $H_1 = (1, O_1)$ and $H_2 = (1, A_1, O_1, O_2)$ as input.

Scenarios 1–4 are adopted from Shi et al. (2018). In these scenarios, the models for treatment effect $Q_t^o(H_t, 1) - Q_t^o(H_t, 0)$ are correctly specified at stage 2 but always misspecified at stage 1 because a nonlinear relationship exists between the baseline covariates and the treatment in the Q-function at stage 1. The baseline function at stage 2 is correctly specified in Scenarios 1 and 3, but not in Scenarios 2 and 4, whereas the baseline function at stage 1 is always misspecified. The propensity models at stage 1 in all five scenarios are correctly specified. However, at stage 2, the propensity models are correctly specified in Scenarios 1, 2 and 5, but misspecified in Scenarios 3 and 4. In Scenarios 1 and 3, the active variables are $(A_1, O_{13}, O_{14}, O_2)$ at stage 2 and (O_{11}, O_{13}, O_{14}) at stage 1, since these are associated with A_2 and A_1, respectively. Similarly, in Scenarios 2 and 4, the active variables are $(A_1, O_{13}, O_{14}, O_2)$ at stage 2 and $(O_{11}, O_{13}, O_{14}, O_{15}, O_{16})$ at stage 1. Scenario 5 is adopted from Zhao et al. (2015), where the treatment effect models are misspecified at both stages. The active variables are $(A_1, O_{11}, O_{12}, O_{13})$ at stage 2 and O_{13} at stage 1.

We consider $n = 50/150$ and $p = 60$. Additional results for $p = 200$ are provided in the supplementary material. Table 1 summarizes the performances of the methods based on 1000 replications for $p = 60$. For each replication, we compute the following performance statistics: false positive (FP; the number of inactive variables incorrectly included in the model), false negative (FN; the number of active variables left out of the model) and value function of the estimated optimal treatment regime. The contrast function root-mean-square error (cRMSE) is also calculated for the Alearn- and Qlearn-PROaL as well as for the PAL method. The value function and the cRMSE are assessed using an independent test dataset with sample size of 10,000. In all five scenarios, both Alearn- and Qlearn-PROaL outperform PAL and BOWL-linear in terms of higher value estimates due to not penalizing a few key variables. Although the BOWL-radial also produces a fairly high value estimate when $n = 50$, the value estimate remains very similar with the increased sample size of $n = 150$. On the contrary, the estimated value by Alearn- and Qlearn-PROaL gets very close to the true optimal value in Scenarios 1–4 as the sample size increases. It is worth noting that when $n = 150$, in Scenario 2 where the propensity score is correctly specified but the main effect is misspecified, Alearn-PROaL has a higher value than Qlearn-PROaL. However, in Scenario 3

TABLE 1
Simulation results for $p = 60$. The mean value is reported with the standard deviation in parentheses. The median FP, FN, and cRMSE are recorded with the mean absolute deviation in parentheses. The best results are highlighted in boldface

			Stage 2			Stage 1		
n	Method	Value	FP	FN	cRMSE	FP	FN	cRMSE
Scenario 1								
50	Optimal	2.29						
	Alearn-PROaL	1.99 (0.28)	4 (4.45)	1 (1.48)	1.89 (0.53)	1 (1.48)	3 (0)	1.34 (0.33)
	Qlearn-PROaL	**2.13** (0.21)	2 (2.97)	0 (0)	**1.47** (0.61)	1 (1.48)	2 (1.48)	**0.90** (0.32)
	PAL	1.67 (0.38)	1 (1.48)	2 (1.48)	1.67 (0.58)	2 (1.48)	3 (0)	2.01 (0.55)
	BOWL-linear	0.90 (0.24)	58 (0)	0 (0)	-	57 (0)	0 (0)	-
	BOWL-radial	1.92 (0.36)	-	-	-	-	-	-
150	Optimal	2.29						
	Alearn-PROaL	2.26 (0.02)	1 (1.48)	0 (0)	0.79 (0.25)	0 (0)	2 (0)	0.70 (0.16)
	Qlearn-PROaL	**2.27** (0.02)	0 (0)	0 (0)	0.74 (0.22)	0 (0)	1 (1.48)	**0.56** (0.09)
	PAL	2.21 (0.07)	0 (0)	1 (0)	**0.57** (0.13)	1 (1.48)	2 (0)	0.91 (0.29)
	BOWL-linear	0.96 (0.14)	58 (0)	0 (0)	-	57 (0)	0 (0)	-
	BOWL-radial	1.98 (0.10)	-	-	-	-	-	-
Scenario 2								
50	Optimal	2.48						
	Alearn-PROaL	1.89 (0.33)	3 (4.45)	1 (1.48)	2.27 (0.49)	1 (1.48)	4 (1.48)	1.97 (0.23)
	Qlearn-PROaL	**2.04** (0.28)	2 (2.97)	1 (1.48)	**1.83** (0.62)	2 (2.97)	3 (1.48)	**1.76** (0.23)
	PAL	1.55 (0.42)	1 (1.48)	3 (1.48)	2.11 (0.73)	2 (1.48)	4 (1.48)	2.56 (0.57)
	BOWL-linear	0.92 (0.24)	58 (0)	0 (0)	-	55 (0)	0 (0)	-
	BOWL-radial	1.89 (0.41)	-	-	-	-	-	-

TABLE 1
(*Continued*)

			Stage 2			Stage 1		
n	Method	Value	FP	FN	cRMSE	FP	FN	cRMSE
Scenario 2 (*Continued*)								
150	Optimal	2.48						
	Alearn-PROaL	**2.38** (0.05)	2 (1.48)	0 (0)	0.97 (0.34)	1 (1.48)	2 (0)	**0.92** (0.24)
	Qlearn-PROaL	2.26 (0.02)	1 (1.48)	0 (0)	0.75 (0.27)	2 (0)	2 (1.48)	1.52 (0.04)
	PAL	2.32 (0.10)	0 (0)	1 (0)	**0.74** (0.33)	1 (1.48)	2 (0)	1.02 (0.41)
	BOWL-linear	1.01 (0.15)	58 (0)	0 (0)	-	55 (0)	0 (0)	-
	BOWL-radial	1.97 (0.15)	-	-	-	-	-	-
Scenario 3								
50	Optimal	2.29						
	Alearn-PROaL	1.95 (0.32)	2 (2.97)	1 (1.48)	2.27 (0.75)	1 (1.48)	3 (0)	1.37 (0.28)
	Qlearn-PROaL	**2.16** (0.11)	1 (1.48)	0 (0)	**1.34** (0.56)	1 (1.48)	1 (1.48)	**0.87** (0.30)
	PAL	1.53 (0.39)	1 (1.48)	3 (1.48)	2.30 (0.63)	2 (1.48)	2 (1.48)	2.04 (0.64)
	BOWL-linear	1.12 (0.22)	58 (0)	0 (0)	-	57 (0)	0 (0)	-
	BOWL-radial	1.98 (0.13)	-	-	-	-	-	-
150	Optimal	2.29						
	Alearn-PROaL	2.21 (0.10)	0 (0)	0 (0)	1.29 (0.50)	0 (0)	2 (0)	0.76 (0.22)
	Qlearn-PROaL	**2.26** (0.02)	0 (0)	0 (0)	**0.70** (0.21)	0 (0)	1 (1.48)	**0.57** (0.08)
	PAL	1.93 (0.28)	1 (1.48)	2 (1.48)	1.78 (0.68)	1 (1.48)	2 (0)	1.00 (0.36)
	BOWL-linear	1.17 (0.13)	58 (0)	0 (0)	-	57 (0)	0 (0)	-
	BOWL-radial	1.99 (0.00)	-	-	-	-	-	-
Scenario 4								
50	Optimal	2.48						
	Alearn-PROaL	1.84 (0.38)	2 (2.97)	2 (1.48)	2.56 (0.67)	1 (1.48)	4 (1.48)	1.99 (0.26)
	Qlearn-PROaL	**2.09** (0.15)	2 (2.97)	1 (1.48)	**1.78** (0.61)	2 (2.97)	3 (1.48)	**1.74** (0.24)
	PAL	1.46 (0.43)	1 (1.48)	3 (1.48)	2.58 (0.59)	2 (1.48)	4 (1.48)	2.53 (0.67)
	BOWL-linear	1.11 (0.23)	58 (0)	0 (0)	-	55 (0)	0 (0)	-
	BOWL-radial	1.97 (0.14)	-	-	-	-	-	-
150	Optimal	2.48						
	Alearn-PROaL	**2.26** (0.15)	1 (1.48)	0 (0)	1.61 (0.57)	1 (1.48)	2 (0.74)	**1.05** (0.35)
	Qlearn-PROaL	**2.26** (0.03)	1 (1.48)	0 (0)	**0.76** (0.26)	2 (0)	2 (1.48)	1.52 (0.04)
	PAL	1.94 (0.32)	1 (1.48)	2 (1.48)	2.15 (0.57)	1 (1.48)	2 (0)	1.23 (0.49)
	BOWL-linear	1.19 (0.14)	58 (0)	0 (0)	-	55 (0)	0 (0)	-
	BOWL-radial	1.98 (0.06)	-	-	-	-	-	-
Scenario 5								
50	Optimal	7.21						
	Alearn-PROaL	6.26 (1.48)	0 (0)	3 (0)	18.36 (1.85)	0 (0)	1 (0)	3.19 (2.99)
	Qlearn-PROaL	6.33 (1.41)	0 (0)	3 (0)	**17.06** (0.26)	0 (0)	1 (0)	**0.50** (0.08)
	PAL	3.28 (1.78)	2 (1.48)	4 (0)	21.57 (4.06)	4 (1.48)	1 (0)	16.32 (7.58)
	BOWL-linear	3.41 (1.13)	58 (0)	0 (0)	-	59 (0)	0 (0)	-
	BOWL-radial	**6.72** (0.34)	-	-	-	-	-	-
150	Optimal	7.21						
	Alearn-PROaL	**6.78** (0.00)	0 (0)	3 (0)	18.04 (0.94)	0 (0)	1 (0)	1.93 (1.66)
	Qlearn-PROaL	**6.78** (0.00)	0 (0)	3 (0)	**16.87** (0.08)	0 (0)	0 (0)	**0.33** (0.13)
	PAL	4.99 (1.78)	1 (1.48)	4 (0)	18.94 (1.23)	6 (2.97)	1 (0)	13.40 (3.47)
	BOWL-linear	2.94 (0.58)	58 (0)	0 (0)	-	59 (0)	0 (0)	-
	BOWL-radial	**6.78** (0.00)	-	-	-	-	-	-

where the main effect is correctly specified but the propensity score is misspecified, both Alearn- and Qlearn-PROaL perform similarly in terms of value estimates. This is because the condition $E[\Phi_{t2}^{\mathsf{T}}(H_t, A_t)|H_t] = \mathbf{0}$, which is needed to derive a sharper upper bound for penalized A-learning, may no longer hold in Scenario 3. Scenario 1 is supposedly the best scenario for PAL (Shi et al. (2018)) since both the main effect and propensity score models are correctly specified at stage 2. Under this scenario, PAL yields a smaller RMSE in contrast function estimation at stage 2 as compared to our Q- and A-learning methods. However, their advantage is no longer observed in stage-1 contrast estimation, and thus in the value of the estimated DTR, since the stage-1 contrast function is misspecified. Our Alearn-PROaL method performs better than PAL in terms of higher value estimate, lower cRMSE at stage 1, and comparable or lower cRMSE at stage 2. Furthermore, in all scenarios, FP at stage 1 and FN at stage 2 of our Alearn-PROaL method are smaller than (or at least equal to) that of PAL. In Scenario 5, where the treatment effects are misspecified, both Alearn- and Qlearn-PROaL perform better than PAL and BOWL-linear. In addition, the overall selection performance of these two PROaL methods is better than PAL, and the stage-1 cRMSE by PAL is extremely high in Scenario 5. Although all methods misspecify the relationship, the PRO technique performs favorably against others since it prevents a model from overshrinking by not penalizing several key variables in the model. The BOWL-linear method fails in all scenarios due to a very high FP. In particular, BOWL-radial performs well in Scenario 5 where nonlinear relationships exist, but not in other scenarios with the increased sample size. In overall, both Alearn- and Qlearn-PROaL seem to benefit from the incorporation of weights, which are used to adjust a level of penalization on individual variables.

To explore the results with a larger number of stages, we also consider an extra scenario with three decision points $T = 3$ adopting from Zhao et al. (2015). Since the R package for PAL only works with a two-stage setting, we compare performance of the rest of the methods. In this extra scenario, the Alearn-PROaL has the highest value estimate among all the methods (see Table 4 in the supplementary material). In partciular, the Alearn-PROaL has a higher value estimate than the Qlearn-PROaL. One reason is that the bias in stage-1 pseudo-outcome includes approximation errors due to misspecification of the main effects at stages 2 and 3. Therefore, the advantage of A-learning is apparent as the number of stages increases when the propensity score model is correctly specified. However, the value estimate does not get close to the true optimal value as the sample size increases, which demonstrates one of the limitations of the parametric method for a larger T due to model misspecification. More detailed setting and the results are provided in the supplementary material.

6. Real data application.

We apply the proposed methods to a combined data set from the coronary psychosocial evaluation studies (COPES) and the comparison of depression interventions after acute coronary syndrome (CODIACS) vanguard trial. Both studies were designed to examine the benefits of stepped care approach in post-acute coronary syndrome patients (Davidson et al. (2010), Davidson et al. (2013)). In both studies, patients received either a treatment that contains problem-solving therapy (PST) or not at each stage. As the CODIACS trial was planned as a continuation of the previous trial, COPES, with the same treatment options and population of interest, the data from the two trials were concatenated to increase the sample size. Thus, a total of 281 subjects were used in this study.

In this analysis, the terminal outcome Y_2 is defined as the 6-month reduction in Beck Depression Inventory (BDI), whereas $Y_1 = 0$. The treatment at each stage is coded as $A_t = \{0, 1\}$ for $t = 1, 2$ with 0 indicating non-PST treatment and 1 indicating PST-containing treatment. We consider 29 baseline covariates including patient preference for treatment, age, sex, Hispanic race, Charlson comorbidity index, baseline BDI score and baseline SF-12 score; thus,

TABLE 2
Estimated value and size of DTR in parentheses

Alearn-PROaL	Qlearn-PROaL	PAL	BOWL-linear	BOWL-radial	Observed
9.69 (13)	6.31 (12)	4.49 (19)	5.37 (62)	7.90 (312)	5.30

we have $O_1 \in \mathbb{R}^{29}$. We consider a dichotomized intermediate BDI reduction since initial treatment as the second-stage covariates $O_2 \in \mathbb{R}^1$. Specifically, it is defined by a BDI reduction of at least 3 units (Cheung, Chakraborty and Davidson (2015)), that is, having

$$O_2 = \begin{cases} 1 & \text{if the intermediate BDI reduction} \geq 3, \\ 0 & \text{if the intermediate BDI reduction} < 3. \end{cases}$$

We apply the proposed methods and compare them with other methods as in Section 5. The 5-fold cross-validation is used to estimate the optimal regime and the size of DTR on each training set and evaluate the value of the estimated DTR on each test set. We then take an average for the value and a median for the size of the DTR. The size of DTR is equivalent to the sum of two components: the number of nonzero coefficients involving A_2 at stage 2, and the number of nonzero coefficients involving A_1 at stage 1. It is called the size of DTR since it specifies the number of input variables required to construct the optimal DTR.

Results are presented in Table 2. The last column of the table is the observed outcome where it captures the performance of randomly assigned interventions in the trial. The Alearn-PROaL yields the highest value estimate, followed by the BOWL-radial. The size of DTR for the Alearn-PROaL is comparable to the Qlearn-PROaL and is much smaller than the rest of the methods. Specifically, the Alearn-PROaL increases the change in BDI scores over 6 months by 9.69 on average with fewer variables.

We illustrate the distribution of the estimated optimal DTR using the whole data. Since the stage-2 optimal regime is estimated with $H_2 = (O_1, A_1, O_2)$, it is worthwhile to note that the estimated stage-2 optimal regime is conditioned on patients' stage-1 treatment A_1 not on the estimated stage-1 optimal regime. The Alearn-PROaL recommends PST-containing treatment in about 61.9% and 85.1% in the second and first stage, respectively. The Qlearn-PROaL recommends PST-containing treatment in about 64.4% as the second-stage optimal regime and always recommends PST-containing treatment as the first-stage optimal regime. In contrast, the PAL recommends PST-containing treatment in about 53.4% and 70.8% as the second- and first-stage optimal regime, respectively. The BOWL-linear recommends PST-containing treatment in about 52.7% and 61.2% in the second and first stage, respectively. Lastly, the BOWL-radial recommends PST-containing treatment in about 75.1% as the second-stage optimal regime and always recommends PST-containing treatment as the first-stage optimal regime.

7. Conclusion.

In this paper, we have proposed penalized regression-based learning methods, namely penalized Q- and A-learning, to construct the optimal DTR that would maximize the expected outcome if implemented. The proposed methods place a Lasso-type penalty on some or all variables to find a model that is simple and has a good prediction accuracy. Another advantage of the proposed approaches is that they handle numerous treatment options in a multistage decision problem. We have also provided finite sample upper bounds for the difference between the optimal value and the value obtained by the estimated DTR, which are composed of the sum of approximation error and estimation error of the conditional outcome model at each stage, up to a power depending on the difference in the expected outcome between optimal and suboptimal decisions. The upper bounds guarantee that the value of the estimated DTR will converge to the optimal value if the optimal Q- or treatment effect-functions are correctly specified for penalized A-learning and if the optimal Q-functions are correctly specified for penalized Q-learning. However, the theoretical foundation is based on continuous outcome only. Thus, it remains an interesting task for future studies to generalize it to various types of data, including binary, ordinal and censored outcome.

There are several advantages of our proposed penalized A-learning method over Shi et al. (2018). First, our method and theory apply to any number of stages and treatments per stage. Their framework was formulated under a binary-treatment setting. It is not trivial to generalize their method and proof to a general T (i.e., number of stages) with multiple treatments per stage. Second, the theoretical results in Shi et al. (2018) are derived under the assumption that the contrast functions at noninitial stages are well approximated by the linear models. This assumption plays an important role in Shi et al. (2018) as the double robustness properties that they focused on is only meaningful under this assumption. However, in practice this assumption is likely to be violated as it is challenging to correctly specify the contrast functions at nonterminal stages. Our theoretical results, on the other hand, do not rely on this assumption. Instead, we incorporate potential approximation error of the contrast functions in the generalization error bounds.

Choosing a good representation for Q-functions is important for the proposed methods. The linear basis functions usually work fine at the terminal stage, but choosing a good representation for Φ_t for $t < T$ is challenging since the terms at nonterminal stages are likely to be nonlinear. However, one can check residual plots for diagnosing model misspecification although the patterns do not necessarily indicate in which a problem occurred; that is, whether the problem is in the main or treatment effect terms (Chakraborty and Moodie (2013)). Including higher order terms may be helpful if there is a systematic effect remaining in the residuals, as pointed by Henderson, Ansell and Alshibani (2010).

In practice, we have developed the PROaL algorithms to impose an adaptive Lasso penalty only on a prespecified partial set of variables in each stage for the construction of optimal DTR. A simulation study over different scenarios have shown that Alearn- and Qlearn-PROaL methods produce higher values and better selection performances compared to other competing methods. In the real data analysis, the proposed methods yielded simpler regimes with higher values compared to their counterparts. It is also crucial to recognize that the use of PRO technique mitigates the risk of overshrinkage, which can occur in a completely data-driven regularization method. The optimal DTRs which are estimated from the stable and interpretable model will provide good guidance on medical practitioners and future studies.

APPENDIX A: ALEARN-PROAL ALGORITHM

The Alearn-PROaL algorithm, which imposes the adaptive Lasso penalty only on X_t but not on Z_t, is given in Algorithm 1.

APPENDIX B: PROOF OF THEOREM 1

For any policy $\boldsymbol{\pi} = (\pi_1, \dots, \pi_T)$, denote

$$\Delta Q_t(H_t, A_t) = \max_{a_t} Q_t^o(H_t, a_t) - Q_t^o(H_t, A_t)$$

for $t = 1, \dots, T$. Following Murphy (2005), we have

$$V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) = E_{\boldsymbol{\pi}}\left[\sum_{t=1}^{T}\Big[\max_{a_t} Q_t^o(H_t, a_t) - Q_t^o(H_t, A_t)\Big]\right] = \sum_{t=1}^{T} E_{\boldsymbol{\pi}}\big[\Delta Q_t(H_t, A_t)\big].$$

Algorithm 1 Alearn-PROaL algorithm

```
Require: Data $(O_1, A_1, Y_1, \dots, O_T, A_T, Y_T)$
Ensure: DTR $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \dots, \hat{\pi}_T)$
Set history by $H_1 = O_1$
for $t = 2, \dots, T$ do
    Set history $H_t = (O_1, A_1, O_2, A_2, \dots, O_t)$
end for
Set $\tilde{Y}_T = Y_T$
for $t = T, \dots, 1$ do
    if $t \in \{T-1, \dots, 1\}$ then
        Define $\tilde{Y}_t = Y_t + \sum_{s=t+1}^{T}[Y_s + (\max_{a_s}(X_s^{\mathsf{T}}\hat{\boldsymbol{\beta}}_s + Z_s^{\mathsf{T}}\hat{\boldsymbol{\alpha}}_s)) - (X_s^{\mathsf{T}}\hat{\boldsymbol{\beta}}_s + Z_s^{\mathsf{T}}\hat{\boldsymbol{\alpha}}_s)]$
    end if
    Formulate $X_t$ and $Z_t$ as a function of $H_t$ and $A_t$
    $\hat{\boldsymbol{\upsilon}}_t \leftarrow \arg\min_{\boldsymbol{\upsilon}_t} \mathbb{P}_n(\tilde{Y}_t - Z_t^{\mathsf{T}}\boldsymbol{\upsilon}_t)^2$
    for $j = 1, \dots, p_{t1}$ do
        $\hat{\boldsymbol{\gamma}}_{tj} \leftarrow \arg\min_{\boldsymbol{\gamma}_{tj}} \mathbb{P}_n(X_{tj} - Z_t^{\mathsf{T}}\boldsymbol{\gamma}_{tj})^2$
    end for
    $\hat{\boldsymbol{\Gamma}}_t \leftarrow (\hat{\boldsymbol{\gamma}}_{t1}, \dots, \hat{\boldsymbol{\gamma}}_{tp_{t1}})$
    Construct $\hat{\boldsymbol{w}}_t = |\bar{\boldsymbol{\beta}}_t|^{-\delta}$ for some $\delta > 0$, where $\bar{\boldsymbol{\beta}}_t$ is a root-$(n/p_{t1})$-consistent estimator obtained by regressing the outcome $\tilde{Y}_t - Z_t^{\mathsf{T}}\hat{\boldsymbol{\upsilon}}_t$ on $X_t - Z_t^{\mathsf{T}}\hat{\boldsymbol{\Gamma}}_t$
    Define $(X_t - Z_t^{\mathsf{T}}\hat{\boldsymbol{\Gamma}}_t)^* = (X_t - Z_t^{\mathsf{T}}\hat{\boldsymbol{\Gamma}}_t)/\hat{\boldsymbol{w}}_t$
    Solve the lasso problem for all $\lambda_t$,
        $\hat{\boldsymbol{\beta}}_t^* \leftarrow \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n(\tilde{Y}_t - Z_t^{\mathsf{T}}\hat{\boldsymbol{\upsilon}}_t - ((X_t - Z_t^{\mathsf{T}}\hat{\boldsymbol{\Gamma}}_t)^*)^{\mathsf{T}}\boldsymbol{\beta}_t)^2 + \lambda_t \sum_{j=1}^{p_{t1}} |\beta_{tj}|$
    $\hat{\boldsymbol{\beta}}_t \leftarrow \hat{\boldsymbol{\beta}}_t^*/\hat{\boldsymbol{w}}_t$
    $\hat{\boldsymbol{\alpha}}_t \leftarrow \hat{\boldsymbol{\upsilon}}_t - \hat{\boldsymbol{\Gamma}}_t\hat{\boldsymbol{\beta}}_t$
    $\hat{\pi}_t \in \arg\max_{\boldsymbol{a}_t}(X_t^{\mathsf{T}}\hat{\boldsymbol{\beta}}_t + Z_t^{\mathsf{T}}\hat{\boldsymbol{\alpha}}_t)$
end for
$\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \dots, \hat{\pi}_T)$
```

Define the event

$$\Omega_{\epsilon_t,t} = \Big\{\max_{a_t \in \mathcal{A}_t} Q_t^o(H_t, a_t) - \max_{a_t \in \mathcal{A}_t \setminus \arg\max_{a_t} Q_t^o(H_t, a_t)} Q_t^o(H_t, a_t) \le \epsilon_t\Big\}.$$

Then on the event $\Omega_{\epsilon_t,t}^c$, we have $\Delta Q_t(H_t, A_t) \le [\Delta Q_t(H_t, A_t)]^2/\epsilon_t$. Thus

$$\begin{aligned} V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) &= \sum_{t=1}^{T} E_{\boldsymbol{\pi}}\big[1_{\Omega_{\epsilon_t,t}^C} \Delta Q_t(H_t, A_t) + 1_{\Omega_{\epsilon_t,t}} \Delta Q_t(H_t, A_t)\big] \\ &\le \sum_{t=1}^{T} E_{\boldsymbol{\pi}}\bigg[1_{\Omega_{\epsilon_t,t}^C} \frac{(\Delta Q_t(H_t, A_t))^2}{\epsilon_t} + 1_{\Omega_{\epsilon_t,t}}\bigg(\frac{(\Delta Q_t(H_t, A_t))^2}{\epsilon_t} + \frac{\epsilon_t}{4}\bigg)\bigg] \qquad \text{(B.1)} \\ &= \sum_{t=1}^{T}\bigg[\frac{1}{\epsilon_t} E_{\boldsymbol{\pi}}\big(\Delta Q_t(H_t, A_t)\big)^2 + \frac{\epsilon_t}{4} E_{\boldsymbol{\pi}}(1_{\Omega_{\epsilon_t,t}})\bigg]. \end{aligned}$$

Under conditions (C3) and (3.2),

$$E_{\boldsymbol{\pi}} 1_{\Omega_{\epsilon_t,t}} = E\left[\prod_{s=1}^{t-1} \frac{1_{A_s=\pi_s(H_s)}}{p_s(A_s|H_s)} 1_{\Omega_{\epsilon_t,t}}\right] \le S^{t-1} C\epsilon_t^{\alpha}. \tag{B.2}$$

In addition, note that

$$\begin{aligned} & E_{\boldsymbol{\pi}}[\Delta Q_t(H_t, A_t)]^2 \\ & = E_{\boldsymbol{\pi}}\Big[\max_{a_t} Q_t^o(H_t, a_t) - \max_{a_t} Q_t(H_t, a_t) + Q_t\big(H_t, \pi_t(H_t)\big) - Q_t^o(H_t, A_t)\Big]^2 \\ & \le 2E_{\boldsymbol{\pi}}\Big[\max_{a_t} Q_t^o(H_t, a_t) - \max_{a_t} Q_t(H_t, a_t)\Big]^2 \\ & \quad + 2E_{\boldsymbol{\pi}}\big[Q_t\big(H_t, \pi_t(H_t)\big) - Q_t^o\big(H_t, \pi_t(H_t)\big)\big]^2 \\ & \le 4E_{\boldsymbol{\pi}}\Big(\max_{a_t}\big[Q_t^o(H_t, a_t) - Q_t(H_t, a_t)\big]^2\Big) \\ & = 4E\left(\prod_{s=1}^{t-1} \frac{1_{A_s=\pi_s(H_s)}}{p_s(A_s|H_s)} \frac{1_{A_t \in \arg\max_{a_t}[Q_t^o(H_t,a_t)-Q_t(H_t,a_t)]^2}}{p_t(A_t|H_t)} \right. \\ & \quad \left. \times \big[Q_t^o(H_t, A_t) - Q_t(H_t, A_t)\big]^2\right) \\ & \le 4S^t E\big[Q_t^o(H_t, A_t) - Q_t(H_t, A_t)\big]^2, \end{aligned} \tag{B.3}$$

where the first equality follows since $\pi_t(H_t) \in \arg\max_{a_t} Q_t(H_t, a_t)$ and the last inequality follows from condition (C3). Plugging (B.3) and (B.2) into (B.1) yields

$$V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) \le \sum_{t=1}^{T}\left[\frac{1}{\epsilon_t} 4S^t E\big[Q_t^o(H_t, A_t) - Q_t(H_t, A_t)\big]^2 + \frac{1}{4} S^{t-1} C\epsilon_t^{\alpha+1}\right].$$

By choosing $\epsilon_t = \{16SE[Q_t^o(H_t, A_t) - Q_t(H_t, A_t)]^2/[(1+\alpha)C]\}^{1/(2+\alpha)}$ to minimize the above upper bound, we have

$$V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) \le \sum_{t=1}^{T} C_{1,t}\big\{E\big[Q_t^o(H_t, A_t) - Q_t(H_t, A_t)\big]^2\big\}^{(1+\alpha)/(2+\alpha)},$$

where $C_{1,t} = (2+\alpha)[2^{2\alpha}(1+\alpha)^{-(1+\alpha)} S^{(2+\alpha)t-1} C]^{1/(2+\alpha)}$.

Next, note that

$$Q_t^o(H_t, A_t) - \max_{a_t} Q_t^o(H_t, a_t) = U_t^o(H_t, A_t) - \max_{a_t} U_t^o(H_t, a_t).$$

Thus, using similar arguments, (3.4) can be shown by denoting

$$\Delta U_t(H_t, A_t) = \max_{a_t} U_t^o(H_t, a_t) - U_t^o(H_t, A_t),$$

and $V(\boldsymbol{\pi}^o) - V(\boldsymbol{\pi}) = \sum_{t=1}^{T} E_{\boldsymbol{\pi}}[\Delta U_t(H_t, A_t)]$.

APPENDIX C: UPPER BOUNDS FOR $E[\Phi_t^{\mathsf{T}}\hat{\boldsymbol{\theta}}_t - Q_t^o]^2$ AND $E[\Phi_{t2}^{\mathsf{T}}\hat{\boldsymbol{\theta}}_{t2} - U_t^o]^2$

In this section, we first provide a step-by-step guide for the penalized A-learning theoretical development:

- Theorem 2 in Section 3.2 provides the upper bounds for the value difference between the optimal and estimated DTRs, $V(\boldsymbol{\pi}^o) - V(\hat{\boldsymbol{\pi}})$. The upper bounds are composed of the sum of approximation error and estimation error of the conditional outcome model at each stage, up to a power depending on the difference in the expected outcome between optimal and suboptimal decisions. The result is further strengthened if the propensity score is known or can be consistently estimated. This theorem is a combination of Theorem 1 and Theorem 4. The proof is given in Section 3.2.
- Theorem 4 in Appendix C shows error bounds for Q-functions $E[\Phi_t^\top \hat{\boldsymbol{\theta}}_t - Q_t^o]^2$ and treatment effect $E[\Phi_{t2}^\top \hat{\boldsymbol{\theta}}_{t2} - U_t^o]^2$. We first derive the upper bounds under three sets of events based on Lemmas 1 and 2, and then show that these events hold with high probabilities in Lemmas 3–5 using Bernstein's inequalities.
  - Lemma 1 shows upper bounds for $\sum_{j=1}^{J_t} \bar{w}_{tj} |\hat{\theta}_{tj} - \theta_{tj}|$ and $E[\Phi_t^\top(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2$ on three sets of high probability events. The proof relies on mathematical induction, starting from the terminal stage T and moving backwards ($t = T-1, \dots, 1$).
  - Lemma 2 shows upper bounds for $\sum_{j=J_{t1}+1}^{J_t} \bar{w}_{tj} |\hat{\theta}_{tj} - \theta_{tj}|$ and $E[\Phi_{t2}^\top(\hat{\boldsymbol{\theta}}_{t2} - \boldsymbol{\theta}_{t2})]^2$ on three sets of high probability events similarly.
- The proofs of Lemmas 1–5 are given in the supplementary material.

For any $\varphi > 0$, $0 \le \gamma < 2/(21b - 8)$ and tuning parameter λ_t, define

$$\Theta_t^* = \{\boldsymbol{\theta}_t \in \mathbb{R}^{J_t} : \|\Phi_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)\|_\infty \le \eta \quad \text{and} \quad E[\Phi_t^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)]^2 \le \gamma^2 \lambda_t^2\}$$

for $t = 1, \dots, T$. Denote $J = \max_{t \in \{1,\dots,T\}} J_t$, and

$$\begin{aligned}\Theta = \Biggl\{ (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_T^\top)^\top \in \prod_{t=1}^{T} \Theta_t^* : & \max_{t \in \{1,\dots,T\}} \{|I_s(\boldsymbol{\theta}_s)|/\tau_s\} \\ & \le \frac{(1-2\gamma)^2}{144b} \Biggl[\sqrt{\frac{1}{9b^2} + \frac{n}{2u^2[\log(3J(J+1)) + \varphi]}} - \frac{1}{3b} \Biggr] \Biggr\}.\end{aligned} \tag{C.1}$$

THEOREM 4. *Suppose there exists a constant $S \ge 1$ such that $p_t(a_t|h_t) \ge S^{-1}$ for all (h_t, a_t) pairs for $t = 1, \dots, T$. Assume assumptions (A1)–(A4) hold. For any given $0 \le \gamma < 2/(21b - 8)$ and $\varphi > 0$, suppose the tuning parameters λ_t, $t = 1, \dots, T$, satisfy*

$$\lambda_T \ge \frac{8 \max\{3c, 4\eta\} u [\log(12 J_T) + \varphi]}{[1 - 2\gamma(3b - 2)] n} + \frac{12 \max\{\sigma, 2\eta\} b}{[1 - 2\gamma(3b-2)]} \sqrt{\frac{2[\log(12 J_T) + \varphi]}{n}}, \tag{C.2}$$

$$\begin{aligned}\lambda_t \ge {} & \frac{16 \max\{3c, 4[1 + 2(T-t)]\eta\} u [\log(12 J_t) + \varphi]}{[2 - (21b - 8)\gamma] n} \\ & + \frac{24 \max\{\sigma, 2[1 + 2(T-t)]\eta\} b}{2 - (21b - 8)\gamma} \sqrt{\frac{2[\log(12 J_t) + \varphi]}{n}},\end{aligned} \tag{C.3}$$

$$\textit{and} \quad \lambda_t^2 \ge c_{t,s} \lambda_s^2 \quad \textit{with } c_{t,t} = 1, c_{t,s} = \frac{2}{9}(2\gamma + 5)(5S + 3)(T - t)^2 c_{t+1,s}, \tag{C.4}$$

for $t = 1, \dots, T$, $s = t, \dots, T$. Let Θ be the set defined in (C.1) and assume Θ is nonempty. Then for any $(\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_T^\top)^\top \in \Theta$, we have

$$\begin{aligned}\mathbf{P}\Biggl(\bigcap_{t=1}^{T} \Biggl\{ E[\Phi_t^\top \hat{\boldsymbol{\theta}}_t - Q_t^o]^2 \le E[\Phi_t^\top \boldsymbol{\theta}_t - Q_t^o]^2 + K_{t1} \max_{s \in \{t,\dots,T\}} \Biggl(c_{t,s} \frac{|I_s(\boldsymbol{\theta}_s)| \lambda_s^2}{\tau_s} \Biggr) \Biggr\} \Biggr) \\ \ge 1 - T \exp(-\varphi),\end{aligned} \tag{C.5}$$

where $K_{t1} = [64(2\gamma+5)^2]/81 + [32\gamma b(2\gamma+5)]/[3(1-2\gamma)]$.

Furthermore, if $E[\Phi_{t2}^{\mathsf{T}}(H_t, A_t)|H_t] = \mathbf{0}$ *a.s. Then with probability at least* $1 - T\exp(-\varphi)$,

$$\mathbf{P}\Biggl(\bigcap_{t=1}^{T}\biggl\{E[\Phi_{t2}^{\mathsf{T}}\hat{\boldsymbol{\theta}}_{t2} - U_t^o]^2 \le E[\Phi_{t2}^{\mathsf{T}}\boldsymbol{\theta}_{t2} - U_t^o]^2 + K_{t2}\max_{s\in\{t,\dots,T\}}\biggl(\bar{c}_{t,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\biggr)\biggr\}\Biggr) \tag{C.6}$$
$$\ge 1 - T\exp(-\varphi),$$

where $K_{t2} = \{3 - [(1-2\gamma)^2]/9\}^2 + 6\gamma b\{27/[(1-2\gamma)^2] - 1\}$, $\bar{c}_{t,t} = 1$, *and*

$$\bar{c}_{t,s} = 2(T-t)^2(S+1)\biggl[\frac{81\max_{s\in\{t+1,\dots,T\}}\{\bar{c}_{t+1,s}/c_{t+1,s}\}}{16(1-2\gamma)^2} + 1\biggr]\biggl[3 - \frac{(1-2\gamma)^2}{9}\biggr]\bar{c}_{t+1,s},$$

for $t = 1, \dots, T$, $s = t+1, \dots, T$.

PROOF OF THEOREM 4. For any $(\boldsymbol{\theta}_1^{\mathsf{T}}, \dots, \boldsymbol{\theta}_T^{\mathsf{T}})^{\mathsf{T}} \in \Theta$, we denote

$$\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \dots, \boldsymbol{\theta}_T) = Y_t + \sum_{s=t+1}^{T}\Bigl[Y_s + \max_{a_s}\Phi_s^{\mathsf{T}}(H_s, a_s)\boldsymbol{\theta}_s - \Phi_s^{\mathsf{T}}(H_s, A_s)\boldsymbol{\theta}_s\Bigr] \tag{C.7}$$

when $t = T-1, \dots, 1$, and $\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \dots, \boldsymbol{\theta}_T) \equiv Y_T$ when $t = T$ for the convenience of notation. For $t = 1, \dots, T$, Let $|\mathcal{A}_t|$ be the cardinality of $\mathcal{A}_t$. Define the events

$$\Omega_{t,1}(\boldsymbol{\theta}_t, \dots, \boldsymbol{\theta}_T) = \biggl\{\max_{j,k\in\{1,\dots,J_t\}}\biggl|(E - \mathbb{E}_n)\biggl(\frac{\phi_{tj}\phi_{tk}}{\bar{w}_{tj}\bar{w}_{tk}}\biggr)\biggr| \le \frac{(1-2\gamma)^2}{144\max_{s\in\{t,\dots,T\}}\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}}\biggr\},$$
$$\Omega_{t,2}(\boldsymbol{\theta}_t, \dots, \boldsymbol{\theta}_T) = \biggl\{\max_{j\in\{1,\dots,J_t\}}\biggl|\mathbb{E}_n\biggl[\bigl(\tilde{Y}_t(\boldsymbol{\theta}_{t+1}, \dots, \boldsymbol{\theta}_T) - \Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t\bigr)\frac{\phi_{tj}}{\bar{w}_{tj}}\biggr]\biggr| \le \frac{4\gamma+1}{6}\lambda_t\biggr\},$$
$$\Omega_{t,3}(\boldsymbol{\theta}_t, \dots, \boldsymbol{\theta}_T) = \biggl\{\max_{j,k\in\{1,\dots,J_t\}}\biggl|(E - \mathbb{E}_n)\biggl(\sum_{a_t\in\mathcal{A}_t}\frac{\phi_{tj}(H_t, a_t)\phi_{tk}(H_t, a_t)}{\bar{w}_{tj}\bar{w}_{tk}}\biggr)\biggr|$$
$$\le \frac{(1-2\gamma)^2|\mathcal{A}_t|}{144\max_{s\in\{t,\dots,T\}}\{|I_s(\boldsymbol{\theta}_s)|/\tau_s\}}\biggr\}.$$

By the Cauchy–Schwarz inequality,

$$\max_j\biggl|E\biggl[\Phi_t^{\mathsf{T}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)\frac{\phi_{tj}}{\bar{w}_{tj}}\biggr]\biggr| \le \sqrt{E[\Phi_t^{\mathsf{T}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)]^2\max_j E[\phi_{tj}/\bar{w}_{tj}]^2} \le \gamma b\lambda_t,$$

where the second inequality holds from Assumption (A4). Thus

$$\begin{aligned}
&E[\Phi_t^{\mathsf{T}}\hat{\boldsymbol{\theta}}_t - Q_t^o]^2 \\
&\quad= E[\Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t - Q_t^o]^2 + E[\Phi_t^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 + 2E[\Phi_t^{\mathsf{T}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)][\Phi_t^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)] \\
&\quad\le E[\Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t - Q_t^o]^2 + E[\Phi_t^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 \\
&\qquad+ 2\max_{j\in\{1,\dots,J_t\}}\biggl|E\biggl[\Phi_t^{\mathsf{T}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*)\frac{\phi_{tj}}{\bar{w}_{tj}}\biggr]\biggr|\biggl(\sum_{j=1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj} - \theta_{tj}|\biggr) \\
&\quad\le E[\Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t - Q_t^o]^2 + E[\Phi_t^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)]^2 + 2\gamma b\lambda_t\biggl(\sum_{j=1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj} - \theta_{tj}|\biggr).
\end{aligned}$$

By Lemma 1 given below, on the event $\bigcap_{t=1}^{T}\{\Omega_{t,1}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)\cap\Omega_{t,2}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)\cap\Omega_{t+1,3}(\boldsymbol{\theta}_{t+1},\dots,\boldsymbol{\theta}_T)\}$, we have

$$E\big[\Phi_t^{\mathsf{T}}\hat{\boldsymbol{\theta}}_t-Q_t^o\big]^2\le E\big[\Phi_t^{\mathsf{T}}\boldsymbol{\theta}_t-Q_t^o\big]^2+K_{t1}\max_{s\in\{t,\dots,T\}}\bigg\{c_{t,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\bigg\},$$

for $t=1,\dots,T$, where $\Omega_{T+1,3}(\boldsymbol{\theta}_{T+1})$ is defined as the universe for the convenience of notation.

If $E[\Phi_{t2}^{\mathsf{T}}(H_t,A_t)|H_t]=\mathbf{0}$ a.s., by Lemma 2, on the event $\bigcap_{t=1}^{T}\{\Omega_{t,1}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)\cap\Omega_{t,2}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)\cap\Omega_{t+1,3}(\boldsymbol{\theta}_{t+1},\dots,\boldsymbol{\theta}_T)\}$, we have

$$\begin{aligned}
&E\big[\Phi_{t2}^{\mathsf{T}}\hat{\boldsymbol{\theta}}_{t2}-U_t^o\big]^2\\
&\quad\le E\big[\Phi_{t2}^{\mathsf{T}}\boldsymbol{\theta}_{t2}-U_t^o\big]^2+E\big[\Phi_{t2}^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_{t2}-\boldsymbol{\theta}_{t2})\big]^2+2\gamma b\lambda_t\Bigg(\sum_{j=J_{t1}+1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\Bigg)\\
&\quad\le E\big[\Phi_{t2}^{\mathsf{T}}\boldsymbol{\theta}_{t2}-U_t^o\big]^2+K_{t2}\max_{s\in\{t,\dots,T\}}\bigg\{\bar{c}_{t,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\bigg\},
\end{aligned}$$

for $t=1,\dots,T$.

The conclusion of the theorem follows from the union probability bounds of the events $\Omega_{t,1}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)$, $\Omega_{t,2}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)$, and $\Omega_{t,3}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)$ for $t=1,\dots,T$, provided in Lemmas 3, 4 and 5. □

LEMMA 1. *Assume there exists a constant $S\ge 1$ such that $p_t(a_t|h_t)\ge S^{-1}$ for all (h_t,a_t) pairs. Suppose Assumption* (A2) *and condition* (C.4) *hold. Then, for any $(\boldsymbol{\theta}_1^{\mathsf{T}},\dots,\boldsymbol{\theta}_T^{\mathsf{T}})^{\mathsf{T}}\in\Theta$, on the event $\bigcap_{t=1}^{T}\{\Omega_{t,1}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)\cap\Omega_{t,2}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)\cap\Omega_{t+1,3}(\boldsymbol{\theta}_{t+1},\dots,\boldsymbol{\theta}_T)\}$, we have*

$$\sum_{j=1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\le\frac{16(2\gamma+5)}{3(1-2\gamma)\lambda_t}\max_{s\in\{t,\dots,T\}}\bigg\{c_{t,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\bigg\},\tag{C.8}$$

$$E\big[\Phi_t^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)\big]^2\le\frac{64(2\gamma+5)^2}{81}\max_{s\in\{t,\dots,T\}}\bigg\{c_{t,s}\frac{|I_s(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\bigg\},\tag{C.9}$$

for $t=1,\dots,T$, where $\Omega_{T+1,3}(\boldsymbol{\theta}_{T+1})$ is defined as the universe for the convenience of notation.

LEMMA 2. *Suppose all conditions in Lemma 1 hold. Assume $E[\Phi_{t2}^{\mathsf{T}}(H_t,A_t)|H_t]=\mathbf{0}$ a.s. for $t=1,\dots,T$. Then, for any $(\boldsymbol{\theta}_1^{\mathsf{T}},\dots,\boldsymbol{\theta}_T^{\mathsf{T}})^{\mathsf{T}}\in\Theta$, on the event $\bigcap_{t=1}^{T}\{\Omega_{t,1}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)\cap\Omega_{t,2}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_T)\cap\Omega_{t+1,3}(\boldsymbol{\theta}_{t+1},\dots,\boldsymbol{\theta}_T)\}$, we have*

$$\sum_{j=J_{t1}+1}^{J_t}\bar{w}_{tj}|\hat{\theta}_{tj}-\theta_{tj}|\le\bigg[\frac{81}{(1-2\gamma)^2}-3\bigg]\lambda_t^{-1}\max_{s\in\{t,\dots,T\}}\bigg\{\bar{c}_{t,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\bigg\},\tag{C.10}$$

$$\text{and}\quad E\big[\Phi_{t2}^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_{t2}-\boldsymbol{\theta}_{t2})\big]^2\le\bigg[3-\frac{(1-2\gamma)^2}{9}\bigg]^2\max_{s\in\{t,\dots,T\}}\bigg\{\bar{c}_{t,s}\frac{|I_{s2}(\boldsymbol{\theta}_s)|\lambda_s^2}{\tau_s}\bigg\},\tag{C.11}$$

for $t=1,\dots,T$, where $c_{t,s}$ is defined in Lemma 1, $\bar{c}_{t,t}=1$ and

$$\bar{c}_{t,s}=2(T-t)^2(S+1)\bigg[\frac{81\max_{s\in\{t+1,\dots,T\}}\{\bar{c}_{t+1,s}/c_{t+1,s}\}}{16(1-2\gamma)^2}+1\bigg]\bigg[3-\frac{(1-2\gamma)^2}{9}\bigg]\bar{c}_{t+1,s},$$

for $t=1,\dots,T$, $s=t+1,\dots,T$.

LEMMA 3. *Suppose Assumptions* (*A*3) *and* (*A*4) *hold. Then for any* $\varphi > 0$ *and* $(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_T^\top)^\top \in \Theta$, $\mathbf{P}(\{\Omega_{t,1}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)\}^C) \le \exp(-\varphi)/3$ *for* $t = 1, \ldots, T$.

LEMMA 4. *Suppose Assumptions* (*A*1), (*A*3) *and* (*A*4) *hold. Then for any* $\varphi > 0$, *if* λ_t *satisfies conditions* (C.2), (C.3) *and* (C.4), *then for* $(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_T^\top)^\top \in \Theta$, $\mathbf{P}(\{\Omega_{t,2}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)\}^C) \le \exp(-\varphi)/3$ *for* $t = 1, \ldots, T$.

LEMMA 5. *Suppose Assumptions* (*A*3) *and* (*A*4) *hold. Then for any* $\varphi > 0$ *and* $(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_T^\top)^\top \in \Theta$, $\mathbf{P}(\{\Omega_{t,3}(\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T)\}^C) \le \exp(-\varphi)/3$ *for* $t = 1, \ldots, T$.

Acknowledgments. The authors thank the Associate Editor and referees for their helpful comments.

Funding. This work is partially funded by NIH Grants R01MH109496, R21MH108999 and NSF Grant DMS-2112938.

SUPPLEMENTARY MATERIAL

Supplement to "Generalization error bounds of dynamic treatment regimes in penalized regression-based learning" (DOI: 10.1214/22-AOS2171SUPP; .pdf). The supplementary material contains proofs of the lemmas for penalized A-learning, theorems and proofs for penalized Q-learning and additional simulation results.

REFERENCES

BLATT, D., MURPHY, S. A. and ZHU, J. (2004). A-learning for approximate planning.

BUBECK, S., PERCHET, V. and RIGOLLET, P. (2013). Bounded regret in stochastic multiarmed bandits. In *Proceedings of the* 26*th Annual Conference on Learning Theory*.

CHAKRABORTY, B. and MOODIE, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes*: *Reinforcement Learning*, *Causal Inference*, *and Personalized Medicine*. *Statistics for Biology and Health*. Springer, New York. MR3112454 https://doi.org/10.1007/978-1-4614-7428-9

CHEUNG, Y. K., CHAKRABORTY, B. and DAVIDSON, K. W. (2015). Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program. *Biometrics* **71** 450–459. MR3366249 https://doi.org/10.1111/biom.12258

DAVIDSON, K. W., RIECKMANN, N., CLEMOW, L., SCHWARTZ, J. E., SHIMBO, D., MEDINA, V., ALBANESE, G., KRONISH, I., HEGEL, M. et al. (2010). Enhanced depression care for patients with acute coronary syndrome and persistent depressive symptoms: Coronary psychosocial evaluation studies randomized controlled trial. *Arch. Intern. Med.* **170** 600–608.

DAVIDSON, K. W., BIGGER, J. T., BURG, M. M., CARNEY, R. M., CHAPLIN, W. F., CZAJKOWSKI, S., DORNELAS, E., DUER-HEFELE, J., FRASURE-SMITH, N. et al. (2013). Centralized, stepped, patient preference-based treatment for patients with post-acute coronary syndrome depression: CODIACS vanguard randomized controlled trial. *J. Am. Med. Assoc. Intern. Med.* **173** 997–1004.

ERTEFAIE, A. and STRAWDERMAN, R. L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika* **105** 963–977. MR3877877 https://doi.org/10.1093/biomet/asy043

ERTEFAIE, A., MCKAY, J. R., OSLIN, D. and STRAWDERMAN, R. L. (2021). Robust Q-learning. *J. Amer. Statist. Assoc.* **116** 368–381. MR4227700 https://doi.org/10.1080/01621459.2020.1753522

FAN, A., LU, W. and SONG, R. (2016). Sequential advantage selection for optimal treatment regime. *Ann. Appl. Stat.* **10** 32–53. MR3480486 https://doi.org/10.1214/15-AOAS849

FOSTER, J. C., TAYLOR, J. M. G. and RUBERG, S. J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.* **30** 2867–2880. MR2844689 https://doi.org/10.1002/sim.4322

GEY, S. (2012). Risk bounds for CART classifiers under a margin condition. *Pattern Recognit.* **45** 3523–3534.

HENDERSON, R., ANSELL, P. and ALSHIBANI, D. (2010). Regret-regression for optimal dynamic treatment regimes. *Biometrics* **66** 1192–1201. MR2758507 https://doi.org/10.1111/j.1541-0420.2009.01368.x

JIANG, R., LU, W., SONG, R., HUDGENS, M. G. and NAPRVAVNIK, S. (2017). Doubly robust estimation of optimal treatment regimes for survival data—with application to an HIV/AIDS study. *Ann. Appl. Stat.* **11** 1763–1786. MR3709577 https://doi.org/10.1214/17-AOAS1057

JIANG, B., SONG, R., LI, J. and ZENG, D. (2019). Entropy learning for dynamic treatment regimes. *Statist. Sinica* **29** 1633–1656. MR3970323

LABER, E. B. and ZHAO, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika* **102** 501–514. MR3394271 https://doi.org/10.1093/biomet/asv028

LABER, E. B., LIZOTTE, D. J., QIAN, M., PELHAM, W. E. and MURPHY, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Stat.* **8** 1225–1272. MR3263118 https://doi.org/10.1214/14-EJS920

LATTIMORE, T. and MUNOS, R. (2014). Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems* **27** 550–558.

LAVORI, P. W., DAWSON, R. and RUSH, A. J. (2000). Flexible treatment strategies in chronic disease: Clinical and research implications. *Biol. Psychiatry* **48** 605–614.

LIPKOVICH, I., DMITRIENKO, A., DENNE, J. and ENAS, G. (2011). Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat. Med.* **30** 2601–2621. MR2815438 https://doi.org/10.1002/sim.4289

LIU, Y., WANG, Y., KOSOROK, M. R., ZHAO, Y. and ZENG, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regiments. *Stat. Med.* **37** 3776–3788. MR3869154 https://doi.org/10.1002/sim.7844

LU, W., ZHANG, H. H. and ZENG, D. (2013). Variable selection for optimal treatment decision. *Stat. Methods Med. Res.* **22** 493–504. MR3190671 https://doi.org/10.1177/0962280211428383

LUCKETT, D. J., LABER, E. B., KAHKOSKA, A. R., MAAHS, D. M., MAYER-DAVIS, E. and KOSOROK, M. R. (2020). Estimating dynamic treatment regimes in mobile health using V-learning. *J. Amer. Statist. Assoc.* **115** 692–706. MR4107673 https://doi.org/10.1080/01621459.2018.1537919

MARIN, F., GONZALEZ-CONEJERO, R., CAPRANZANO, P., BASS, T. A., ROLDAN, V. and ANGIOLILLO, D. J. (2009). Pharmacogenetics in cardiovascular antithrombotic therapy. *J. Am. Coll. Cardiol.* **54** 1041–1057.

MOODIE, E. E. M., DEAN, N. and SUN, Y. R. (2014). Q-learning: Flexible learning about useful utilities. *Stat. Biosci.* **6** 223–243.

MOODIE, E. E. M., RICHARDSON, T. S. and STEPHENS, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63** 447–455. MR2370803 https://doi.org/10.1111/j.1541-0420.2006.00686.x

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–366. MR1983752 https://doi.org/10.1111/1467-9868.00389

MURPHY, S. A. (2005). A generalization error for Q-learning. *J. Mach. Learn. Res.* **6** 1073–1097. MR2249849

MURPHY, S. A., VAN DER LAAN, M. J. and ROBINS, J. M. (2001). Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* **96** 1410–1423. MR1946586 https://doi.org/10.1198/016214501753382327

MURPHY, S. A., OSLIN, D. W., RUSH, A. J. and ZHU, J. (2007). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychoarmacology* **32** 257–262.

OH, E. J., QIAN, M. and CHEUNG, Y. K. (2022). Supplement to “Generalization error bounds of dynamic treatment regimes in penalized regression-based learning.” https://doi.org/10.1214/22-AOS2171SUPP

OH, E. J., QIAN, M., CHEUNG, K. and MOHR, D. C. (2020). Building health application recommender system using partially penalized regression. In *Statistical Modeling in Biomedical Research* 105–123. Springer, Berlin.

PINEAU, J., BELLERNARE, M. G., RUSH, A. J., GHIZARU, A. and MURPHY, S. A. (2007). Constructing evidence-based treatment strategies using methods from computer science. *Drug Alcohol Depend.* **88** S52–S60.

PIPER, W. E., BOROTO, D. R., JOYCE, A. S., MCCALLUM, M. and AZIM, H. F. A. (1995). Pattern of alliance and outcome in short-term individual psychotherapy. *Psychotherapy* **32** 639–647.

QI, Z. and LIU, Y. (2018). D-learning to estimate optimal individual treatment rules. *Electron. J. Stat.* **12** 3601–3638. MR3870507 https://doi.org/10.1214/18-ejs1480

QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351 https://doi.org/10.1214/10-AOS864

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

ROBINS, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section* 24–33. Am. Statist. Assoc., Alexandria.

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality* (*Los Angeles*, *CA*, 1994). *Lect. Notes Stat.* **120** 69–117. Springer, New York. MR1601279 https://doi.org/10.1007/978-1-4612-1842-5_4

ROBINS, J. M. (1998). Marginal structural models. In *Proceedings of the American Statistical Association*. *Section on Bayesian Statistical Science* 1–10.

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. *Lect*. *Notes Stat*. **179** 189–326. Springer, New York. MR2129402 https://doi.org/10.1007/978-1-4419-9076-1_11

ROBINS, J., ORELLANA, L. and ROTNITZKY, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Stat*. *Med*. **27** 4678–4721. MR2528576 https://doi.org/10.1002/sim.3301

RUDIN, C. and ERTEKIN, Ş. (2018). Learning customized and optimized lists of rules with mathematical programming. *Math*. *Program*. *Comput*. **10** 659–702. MR3863707 https://doi.org/10.1007/s12532-018-0143-8

SCHULTE, P. J., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statist*. *Sci*. **29** 640–661. MR3300363 https://doi.org/10.1214/13-STS450

SHI, C., SONG, R. and LU, W. (2016). Robust learning for optimal treatment decision with NP-dimensionality. *Electron*. *J*. *Stat*. **10** 2894–2921. MR3557316 https://doi.org/10.1214/16-EJS1178

SHI, C., FAN, A., SONG, R. and LU, W. (2018). High-dimensional A-learning for optimal dynamic treatment regimes. *Ann*. *Statist*. **46** 925–957. MR3797992 https://doi.org/10.1214/17-AOS1570

SONG, R., WANG, W., ZENG, D. and KOSOROK, M. R. (2015). Penalized Q-learning for dynamic treatment regimens. *Statist*. *Sinica* **25** 901–920. MR3409730

SU, X., ZHOU, T., YAN, X., FAN, J. and YANG, S. (2008). Interaction trees with censored survival data. *Int*. *J*. *Biostat*. **4** Art. 2, 28. MR2383729 https://doi.org/10.2202/1557-4679.1071

TIAN, L., ALIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J*. *Amer*. *Statist*. *Assoc*. **109** 1517–1532. MR3293607 https://doi.org/10.1080/01621459.2014.951443

TIBSHIRANI, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **73** 273–282. MR2815776 https://doi.org/10.1111/j.1467-9868.2011.00771.x

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann*. *Statist*. **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann*. *Statist*. **36** 614–645. MR2396809 https://doi.org/10.1214/009053607000000929

WALLACE, M. P. and MOODIE, E. E. M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* **71** 636–644. MR3402599 https://doi.org/10.1111/biom.12306

WATKINS, C. J. (1989). Learning from delayed rewards. Ph.D. thesis, Univ. Cambridge England.

ZHANG, B. and ZHANG, M. (2018). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics* **74** 891–899. MR3860710 https://doi.org/10.1111/biom.12836

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100** 681–694. MR3094445 https://doi.org/10.1093/biomet/ast014

ZHANG, Y., LABER, E. B., TSIATIS, A. and DAVIDIAN, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71** 895–904. MR3436715 https://doi.org/10.1111/biom.12354

ZHANG, Y., LABER, E. B., DAVIDIAN, M. and TSIATIS, A. A. (2018). Interpretable dynamic treatment regimes. *J*. *Amer*. *Statist*. *Assoc*. **113** 1541–1549.

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J*. *Amer*. *Statist*. *Assoc*. **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

ZHAO, Y.-Q., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J*. *Amer*. *Statist*. *Assoc*. **110** 583–598. MR3367249 https://doi.org/10.1080/01621459.2014.937488

ZHAO, Y.-Q., ZHU, R., CHEN, G. and ZHENG, Y. (2020). Constructing dynamic treatment regimes with shared parameters for censored data. *Stat*. *Med*. **39** 1250–1263. MR4098488 https://doi.org/10.1002/sim.8473

ZHOU, X., MAYER-HAMBLETT, N., KHAN, U. and KOSOROK, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *J*. *Amer*. *Statist*. *Assoc*. **112** 169–187. MR3646564 https://doi.org/10.1080/01621459.2015.1093947

ZHU, R., ZENG, D. and KOSOROK, M. R. (2015). Reinforcement learning trees. *J*. *Amer*. *Statist*. *Assoc*. **110** 1770–1784. MR3449072 https://doi.org/10.1080/01621459.2015.1036994

ZHU, W., ZENG, D. and SONG, R. (2019). Proper inference for value function in high-dimensional Q-learning for dynamic treatment regimes. *J*. *Amer*. *Statist*. *Assoc*. **114** 1404–1417. MR4011788 https://doi.org/10.1080/01621459.2018.1506341

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J*. *Amer*. *Statist*. *Assoc*. **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735

ZOU, H. and ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann*. *Statist*. **37** 1733–1751. MR2533470 https://doi.org/10.1214/08-AOS625