The Annals of Statistics
2018, Vol. 46, No. 3, 925–957
https://doi.org/10.1214/17-AOS1570

HIGH-DIMENSIONAL *A*-LEARNING FOR OPTIMAL DYNAMIC TREATMENT REGIMES

BY CHENGCHUN SHI[1], AILIN FAN, RUI SONG[1] AND WENBIN LU[2]

North Carolina State University

Precision medicine is a medical paradigm that focuses on finding the most effective treatment decision based on individual patient information. For many complex diseases, such as cancer, treatment decisions need to be tailored over time according to patients' responses to previous treatments. Such an adaptive strategy is referred as a dynamic treatment regime. A major challenge in deriving an optimal dynamic treatment regime arises when an extraordinary large number of prognostic factors, such as patient's genetic information, demographic characteristics, medical history and clinical measurements over time are available, but not all of them are necessary for making treatment decision. This makes variable selection an emerging need in precision medicine.

In this paper, we propose a penalized multi-stage *A*-learning for deriving the optimal dynamic treatment regime when the number of covariates is of the nonpolynomial (NP) order of the sample size. To preserve the double robustness property of the *A*-learning method, we adopt the Dantzig selector, which directly penalizes the A-leaning estimating equations. Oracle inequalities of the proposed estimators for the parameters in the optimal dynamic treatment regime and error bounds on the difference between the value functions of the estimated optimal dynamic treatment regime and the true optimal dynamic treatment regime are established. Empirical performance of the proposed approach is evaluated by simulations and illustrated with an application to data from the STAR*D study.

1. Introduction. Precision medicine is a medical paradigm that focuses on finding the most effective treatment decision based on individual patient information. For many chronic diseases, such as cancer, cardiovascular disease and diabetes, treatment decisions need to be tailored over time according to patients' responses to previous treatments. Such an adaptive treatment strategy is referred to as an dynamic treatment regime. Formally speaking, a dynamic treatment regime is a sequence of decision rules, dictating how the treatment will be tailored through time to individual's status. The optimal dynamic treatment regime is defined as the one that yields the most favorable outcome on average.

Received January 2016; revised January 2017.

[1]Supported in part by NSF Grant DMS-1555244 and NCI Grant P01 CA142538.

[2]Supported in part by NCI Grant P01 CA142538.

MSC2010 subject classifications. Primary 62C99; secondary 62J07.

Key words and phrases. *A*-learning, Dantzig selector, NP-dimensionality, model misspecification, optimal dynamic treatment regime, oracle inequality.

Various methods have been proposed to estimate the optimal dynamic treatment regime, including Q-learning [Chakraborty, Murphy and Strecher (2010), Watkins and Dayan (1992)] and A-learning [Murphy (2003), Robins, Hernan and Brumback (2000)]. Both Q-learning and A-learning rely on a backward induction algorithm to find the optimal dynamic treatment regime, however, Q-learning models the conditional mean of the outcome given predictors and treatment while A-learning directly models the contrast function that is sufficient for treatment decision. In particular, A-learning has the so-called doubly robust property, that is, when either the baseline mean function or the propensity score model is correctly specified, the resulting A-learning estimating equation for the contrast function is consistent.

With the fast development of new technology, it becomes possible to gather an extraordinary large number of prognostic factors for each individual, such as patient's genetic information, demographic characteristics, medical history and clinical measurements over time. For such big data, it is important to make effective use of information that is relevant to make optimal individualized treatment decisions, which makes variable selection as an emerging need for implementing precision medicine. In addition, variable selection is an essential tool in making inference for problems in which the number of covariates is comparable or much larger than the sample size. There have been extensive developments of penalized regression methods for variable selection in prediction, for example, LASSO [Tibshirani (2011)], SCAD [Fan and Li (2001)] and the Dantzig selector [Candès and Tao (2007)], to name a few. In contrast to most penalized regression methods, which adds a penalty term to an objective function, the Dantzig selector focuses directly on estimating equations.

Although there is a large amount of work on developing variable selection methods for prediction, variable selection tools for deriving optimal individualized treatment regimes have been less studied, especially when the number of predictors is much larger than the sample size. Qian and Murphy (2011) proposed to estimate the conditional mean response using a L_1-penalized regression and studied the error bound of the value function for the estimated treatment regime. When the number of covariates is fixed, Lu, Zhang and Zeng (2013) introduced a new penalized least squared regression framework and established the oracle property of the estimator, which is robust against the misspecification of the conditional mean function. Shi, Song and Lu (2016) extended this result to the setting allowing NP-dimensionality of covariates. However, all these works only consider studies with a single treatment decision. When moving to multiple-stage studies, the asymptotic properties of the estimated optimal dynamic treatment regime are much harder to derive since it needs to handle model misspecification of the contrast functions in the presence of NP-dimensionality of covariates. Moreover, these methods are not doubly robust.

In this paper, we propose a penalized A-learning method for deriving the optimal dynamic treatment regime when the number of covariates is of NP-order of

the sample size. To preserve the doubly robust property of the A-learning method, we adopt the Dantzig selector [Candès and Tao (2007)], which directly penalizes the A-leaning estimating equations. The technical challenges and advances of the proposed estimators are described as follows.

First, to prove the theoretical properties of the Dantzig estimator in linear regression setting, the uniform uncertainty principle [UUP, Candès and Tao (2007)] or restricted eigenvalue condition [RE, Bickel, Ritov and Tsybakov (2009)] is required on the Gram matrix $X^T X$, where X stands for the design matrix. The UUP condition essentially requires that every principle submatrix with the number of rows or columns less than some specified s behaves like an orthonormal system. The RE condition is the weakest, and hence the most general condition in the literature to ensure the theoretical properties of Lasso and Dantzig estimators. A close connection between these two conditions are discussed in Bickel, Ritov and Tsybakov (2009). In a random design case, Candès and Tao (2007) studied the UUP condition for Gaussian, Bernoulli and Fourier ensembles. Mendelson, Pajor and Tomczak-Jaegermann (2007, 2008) obtained a similar result for a more general class of design matrices, the isotropic sub-Gaussian matrices, based on some empirical process results. These results were further extended by Zhou (2009), where the UUP and RE conditions are developed for sub-Gaussian ensembles with a correlated covariance structure. In the proposed penalized A-learning method, however, such conditions are required on matrices involving estimates, such as

$$X^T \operatorname{diag}\bigl(A \circ (1 - \hat{\pi})\bigr) X, \tag{1.1}$$

where $A = (A_1, \ldots, A_n)^T$ denotes the vector of treatments received by n subjects, $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)$ denotes the corresponding estimated propensity scores and $\circ$ denotes the componentwise product operator. The presence of $\hat{\pi}$ in (1.1) adds extraordinary difficulties in establishing theoretical properties of such a random matrix. We establish the UUP and RE conditions under a proper convergence rate of $\hat{\pi}$, which provides a new theoretical framework for studying random matrices that involve estimates of unknown parameters.

Second, in the proposed penalized A-learning method, we need to estimate the baseline mean function and the propensity score model with NP-dimensionality of covariates. We adopt the penalized regressions with the folded-concave penalties, for example, a linear regression for the baseline mean function and a logistic regression for the propensity score model, with the SCAD penalties. Several difficulties need to be addressed for deriving the theoretical properties of the resulting penalized estimators. First, to our knowledge, penalized regressions with folded-concave penalties have seldom been studied in a random design setting. A major difficulty of adapting the existing results for the fixed design case to the random design case is to control the maximum eigenvalues of some random matrices,

$$\max_j \lambda_{\max}\bigl[{X^M}^T \operatorname{diag}\bigl(|X^j|\bigr) X^M\bigr],$$

where $\lambda_{\max}[K]$ denotes the maximum eigenvalue of a matrix K, M is a given subset of $[1, \ldots, p]$, X^j denotes the jth column of a matrix X, and X^M the submatrix formed by columns in M. Such a problem is not standard since matrix ${X^M}^T \operatorname{diag}(|X^j|)X^M$ does not possess subexponential tail. We derive some concentration inequalities for such random matrices and for summations of subexponential and sub-Gaussian random variables. Based on these results, we establish the weak oracle [Lv and Fan (2009)] properties, that is, sign consistency and L_∞ convergence rate of the estimators under sub-Gaussian ensembles, which is one of our major technical contributions. Moreover, the posited models for the baseline mean function or the propensity score may be misspecified. Therefore, the derivation of the asymptotic properties needs to take into account model misspecification with NP-dimensionality of covariates, which is challenging.

Third, a challenge for extending the results for a single treatment decision to sequential treatment decisions is that the contrast functions are likely to be misspecified in the backward induction algorithm, such as A-learning. This together with the NP-dimensionality of covariates make it extremely hard to study theoretical properties of the value function under the estimated optimal dynamic treatment regime. We overcome this difficulty by first defining population-level least favorable parameters in the misspecified contrast functions. Moreover, we derive the error bounds for the corresponding estimates under the model misspecification, which in turn leads to an error bound for the difference between the value functions of the estimated optimal dynamic treatment regime and the underlying true optimal dynamic treatment regime.

The remainder of the paper is organized as follows. We introduce the proposed penalized A-learning method in Section 2. Some implementation issues are addressed in Section 3, followed by simulation results in Section 4. We apply our method to a data from the STAR*D study in Section 5. Section 6 studies the error bounds of the penalized A-learning estimator and the difference between the value functions of the estimated optimal regime and the true optimal regime, at the second stage. Section 7 characterizes such results for the estimates at the first stage. Section 8 presents the weak oracle properties of the penalized estimators in the propensity score and baseline mean models under a random design setting. Section 9 discusses the UUP and RE condition in the context of A-learning. All technical conditions, lemmas and proofs are given in the Supplementary Material [Shi et al. (2018)].

2. Penalized A-learning. For simplicity of presentation, we only consider a two-stage study where binary treatment decisions are made at time points t_1 and t_2. The data of a subject can be summarized as

$$O = \bigl(S^{(1)}, A^{(1)}, S^{(2)}, A^{(2)}, Y\bigr), \tag{2.1}$$

where $S^{(1)}$ denotes the covariates collected prior to t_1, $A^{(1)} \in \{0, 1\}$ is the treatment received at time t_1, $S^{(2)}$ denotes intermediate covariates collected between time

points t_1 and t_2, $A^{(2)} \in \{0, 1\}$ is the treatment received at time t_2, and Y is the final outcome of interest. It is assumed that a larger value of Y stands for a better clinical outcome. Denote $Y^\star(a_1, a_2)$ the potential response of patient if he/she were given a_1 as the first treatment and a_2 as the second. If a patient follows a given regime (d_1, d_2), we can write the potential outcome

$$Y^\star(d_1, d_2) = \sum_{a_1 \in \{0,1\}, a_2 \in \{0,1\}} Y(a_1, a_2) I(d_1 = a_1, d_2 = a_2),$$

where $I(\cdot)$ denotes the indicator function. Our goal is to find a dynamic treatment regime to maximize the mean potential outcome. Throughout the paper, we make the commonly used assumptions for studying dynamic treatment regimes: stable unit treatment value assumption and sequential randomization assumption [Murphy (2003)].

The observed data from n subjects can be summarized as

$$O_i = \big(S_i^{(1)}, A_i^{(1)}, S_i^{(2)}, A_i^{(2)}, Y_i\big), \qquad i = 1, \ldots, n,$$

which are assumed to be independently and identically distributed copies of O. We assume the following semiparametric regression model for Y:

$$Y_i = h^{(2)}(X_i) + A_i^{(2)} C^{(2)}(X_i) + e_i, \tag{2.2}$$

where $X_i = ((S_i^{(1)})^T, A_i^{(1)}, (S_i^{(2)})^T)^T$ is the vector of covariates for the ith patient, $h^{(2)}(\cdot)$ is an unspecified baseline mean function, $C^{(2)}(\cdot)$ the contrast function, and e_i is an independent error with mean 0. The design matrix is denoted as $X = (X_1, \ldots, X_n)^T$.

Define

$$V_i = \max_{A_i^{(2)}} \mathrm{E}\big(Y_i | S_i^{(1)}, A_i^{(1)}, S_i^{(2)}, A_i^{(2)}\big) = h^{(2)}(X_i) + C^{(2)}(X_i) I\big(C^{(2)}(X_i) > 0\big).$$

At the first stage, we consider the following conditional mean model for $V^{(2)}$:

$$\mathrm{E}\big(V_i | S_i^{(1)}, A_i^{(1)}\big) = h^{(1)}\big(S_i^{(1)}\big) + A_i^{(1)} C^{(1)}\big(S_i^{(1)}\big), \tag{2.3}$$

where $h^{(1)}(\cdot)$ and $C^{(1)}(\cdot)$ are functions of the baseline covariates. To simplify the notation, we use a shorthand S_i for $S_i^{(1)}$ and let $S = (S_1, \ldots, S_n)^T$, the design matrix at the baseline.

It can be shown that the optimal dynamic treatment regime is given by $d^{\mathrm{opt}} = (d_1^{\mathrm{opt}}, d_2^{\mathrm{opt}})$, where

$$d_1^{\mathrm{opt}}(S_i) = I\big\{C^{(1)}(S_i) > 0\big\} \quad \text{and} \quad d_2^{\mathrm{opt}}(X_i) = I\big\{C^{(2)}(X_i) > 0\big\}. \tag{2.4}$$

To estimate d_1^{opt} and d_2^{opt}, we posit the following models for $C^{(1)}(\cdot)$, $C^{(2)}(\cdot)$, $h^{(1)}(\cdot)$, $h^{(2)}(\cdot)$, $\pi^{(1)}(\cdot)$, and $\pi^{(2)}(\cdot)$:

$$\pi^{(1)}(s, \alpha_1) = \exp\big(s^T \alpha_1\big)/\big\{1 + \exp\big(s^T \alpha_1\big)\big\}, \tag{2.5}$$

$$\pi^{(2)}(x,\alpha_2) = \exp(x^T\alpha_2)/\{1+\exp(x^T\alpha_2)\}, \tag{2.6}$$

$$\begin{aligned} h^{(1)}(s) &= s^T\theta_1, \qquad h^{(2)}(x) = x^T\theta_2, \\ C^{(1)}(s) &= s^T\beta_1, \qquad C^{(2)}(x) = x^T\beta_2 \end{aligned} \tag{2.7}$$

and

$$\pi^{(1)}(s) = \Pr\bigl(A_i^{(1)} = 1 | S_i = s\bigr) \quad \text{and} \quad \pi^{(2)}(x) = \Pr\bigl(A_i^{(2)} = 1 | X_i = x\bigr).$$

Models in (2.5)–(2.7) can be misspecified; however, we require that either $h^{(j)}$ or $\pi^{(j)}$ is correct for $j = 1, 2$. For simplicity, we require $C^{(2)}$ to be correctly specified. The general case when $C^{(2)}$ is misspecified can be similarly discussed. We use backward induction to estimate the optimal dynamic treatment regime. At the second decision point, we first estimate the parameters in the posited propensity score and baseline mean models using penalized regressions. Specifically, define

$$\begin{aligned} \hat{\alpha}_2 = \arg\min_{\alpha_2\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \bigl[\log\{1+\exp(X_i^T\alpha_2)\} - A_i^{(2)}X_i^T\alpha_2\bigr] \\ + \sum_{j=1}^p \lambda_{1n}^{(2)}\rho_1^{(2)}\bigl(|\alpha_2^j|, \lambda_{1n}^{(2)}\bigr) \end{aligned}$$

and

$$\hat{\theta}_2 = \arg\min_{\theta_2\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n (1-A_i^{(2)})(Y_i - X_i^T\theta_2)^2 + \sum_{j=1}^p \lambda_{2n}^{(2)}\rho_2^{(2)}\bigl(|\theta_2^j|, \lambda_{2n}^{(2)}\bigr),$$

where $\alpha_2 = (\alpha_2^1, \ldots, \alpha_2^p)^T$, $\theta_2 = (\theta_2^1, \ldots, \theta_2^p)^T$, $\rho_1^{(2)}$ and $\rho_2^{(2)}$ belong to the class of folded-concave penalty functions [Lv and Fan (2009)], such as SCAD [Fan and Li (2001)], and $\lambda_{1n}^{(2)}$, $\lambda_{2n}^{(2)}$ the associated regularization parameters.

Next, we estimate β_2 in (2.2) using the Dantzig selector based on A-learning estimating function [Murphy (2003)], defined by

$$\hat{\beta}_2 = \arg\min_{\beta_2\in\Lambda^{(2)}} \|\beta_2\|_1, \tag{2.8}$$

where

$$\Lambda^{(2)} = \biggl\{\beta_2\in\mathbb{R}^p : \biggl\|\frac{1}{n}X^T\operatorname{diag}\bigl(A^{(2)} - \hat{\pi}^{(2)}\bigr)\bigl\{Y - X\hat{\theta}_2 - A^{(2)}\circ(X\beta_2)\bigr\}\biggr\|_\infty \le \lambda_{3n}^{(2)}\biggr\},$$

$$Y = (Y_1, \ldots, Y_n)^T,$$

$$A^{(2)} = \bigl(A_1^{(2)}, \ldots, A_n^{(2)}\bigr)^T \ \hat{\pi}^{(2)} = \bigl(\pi^{(2)}(X_1, \hat{\alpha}_2), \ldots, \pi^{(2)}(X_n, \hat{\alpha}_2)\bigr)^T,$$

and $\lambda_{3n}^{(2)}$ the regularization parameter.

To estimate the regime at the first decision point, we define the pseudo-outcome $\hat{V}_i$ using the advantage function [Murphy (2003)] by

$$\hat{V}_i = Y_i + X_i^T \hat{\beta}_2 \{ I(X_i^T \hat{\beta}_2 > 0) - A_i^{(2)} \}. \tag{2.9}$$

Similarly, define

$$\hat{\alpha}_1 = \arg \min_{\alpha_1 \in \mathbb{R}^q} \frac{1}{n} \sum_{i=1}^n [\log\{1 + \exp(S_i^T \alpha_1)\} - A_i^{(1)} S_i^T \alpha_1] + \sum_{j=1}^q \rho_1^{(1)}(|\alpha_1^j|, \lambda_{1n}^{(1)})$$

and

$$\hat{\theta}_1 = \arg \min_{\theta_1 \in \mathbb{R}^q} \frac{1}{n} \sum_{i=1}^n (1 - A_i^{(1)})(\hat{V}_i - S_i^T \theta_1)^2 + \sum_{j=1}^q \rho_2^{(1)}(|\theta_1^j|, \lambda_{2n}^{(1)}),$$

where $\alpha_1 = (\alpha_1^1, \ldots, \alpha_1^q)^T$, $\theta_1 = (\theta_1^1, \ldots, \theta_1^q)^T$, and $\rho_1^{(1)}$ and $\rho_2^{(1)}$ are folded-concave penalty functions. Then we estimate β_1 in (2.3) by

$$\hat{\beta}_1 = \arg \min_{\beta_1 \in \Lambda^{(1)}} \|\beta_1\|_1, \tag{2.10}$$

where

$$\Lambda^{(1)} = \left\{ \beta_1 \in \mathbb{R}^q : \left\| \frac{1}{n} S^T \operatorname{diag}(A^{(1)} - \hat{\pi}^{(1)}) \{ \hat{V} - S\hat{\theta}_1 - A^{(1)} \circ (S\beta_1) \} \right\|_\infty \le \lambda_{3n}^{(1)} \right\},$$

$$\hat{V} = (\hat{V}_1, \ldots, \hat{V}_n)^T, \qquad A^{(1)} = (A_1^{(1)}, \ldots, A_n^{(1)})^T$$

and

$$\hat{\pi}^{(1)} = (\pi^{(1)}(S_1, \hat{\alpha}_1), \ldots, \pi^{(1)}(S_n, \hat{\alpha}_1))^T.$$

The estimated optimal dynamic treatment regime is given by

$$\hat{d}_1(S_i) = I(\hat{\beta}_1^T S_i > 0) \quad \text{and} \quad \hat{d}_2(X_i) = I(\hat{\beta}_2^T X_i > 0). \tag{2.11}$$

3. Some implementation issues. When the tuning parameters in optimization problems (2.8) and (2.10) are fixed, the Dantzig selector can be solved by a standard linear programming algorithm. One issue for implementing Dantzig selector is the choice of the tuning parameters. We use a BIC criterion for selecting tuning parameters. For Dantzig selector (2.8), $\lambda_{3n}^{(2)}$ is chosen as the minimizer of

$$\mathrm{BIC}(\lambda) = n \log(\mathrm{RSS}(\lambda)/n) + d(\lambda)\{\log(n) + \log(p+1)\}, \tag{3.1}$$

where $\mathrm{RSS}(\lambda) = \sum_{i=1}^n [\{A_i^{(2)} - \pi^{(2)}(X_i, \hat{\alpha}_2)\}(Y_i^{(2)} - X_i^T \hat{\theta}_2 - A_i^{(2)} X_i^T \hat{\beta}_2)]^2$, and $d(\lambda)$ is the number of nonzero components in $\hat{\beta}_2$. A similar BIC criterion was proposed by Chen and Chen (2008). We use a similar criterion for choosing $\lambda_{3n}^{(1)}$.

It was observed that the Dantzig estimators may underestimate the true values of parameters due to the shrinkage estimation [Candès and Tao (2007)]. Therefore, we use a two-step procedure for practical implementation, which is referred as Gauss–Dantzig selector in Candès and Tao (2007). Specifically, in the first step, we apply the proposed penalized A-learning to select important variables for making an optimal decision, that is, those variables with nonzero estimated coefficients. Then, in the second step, their corresponding coefficients are recalculated by solving the unpenalized A-learning estimating equations with important variables only.

4. Simulation studies.

4.1. *Settings*. To evaluate the numerical performance of the proposed penalized A-learning method, we consider simulation studies with two treatment decision points, based on the following model:

$$Y = A^{(1)}A^{(2)} + A^{(2)}\big(\beta_2^T S^{(1)} + S^{(2)} + \beta_0\big) + A^{(1)}\big(\beta_1^T S^{(1)}\big) + \epsilon, \tag{4.1}$$

where $A^{(j)}$, $j = 1, 2$, is the treatment given at the jth stage, $S^{(j)}$, $j = 1, 2$, denote the covariate information collected before the jth treatment is given, and Y is the final response of interest. The random error ϵ follows a normal distribution with mean 0 and variance 0.25. Here, covariates $S^{(1)} = (S_1^{(1)}, \ldots, S_q^{(1)})^T$ follow a multivariate normal distribution with mean 0 and variance I_q. In addition, the intermediate covariate $S^{(2)}$ is a scalar and generated as $S^{(2)} = S_1^{(1)} + A^{(1)} + A^{(1)}S_1^{(1)} + e$, where e follows a normal distribution with mean 0 and variance 0.25.

We set $\beta_0 = 0$. Based on model (4.1), the optimal treatment regime at stage 2 is $I(A^{(1)} + \beta_2^T S^{(1)} + S^{(2)} > 0)$. Following this optimal treatment regime at stage 2, the Q-function at stage 1 is given by

$$\begin{aligned} Q_1\big(S^{(1)}, A^{(1)}\big) &= E\big\{\big(A^{(1)} + \beta_2^T S^{(1)} + S^{(2)}\big)_+ | S^{(1)}, A^{(1)}\big\} + A^{(1)}\big(\beta_1^T S^{(1)}\big) \\ &= \frac{\beta_2}{\sqrt{8\pi}} \exp\big(-2\mu^2\big) + \mu\big\{1 - \Phi(-2\mu)\big\} + A^{(1)}\big(\beta_1^T S^{(1)}\big), \end{aligned}$$

where $\mu = A^{(1)} + \beta_2^T S^{(1)} + S_1^{(1)} + A^{(1)} + A^{(1)}S_1^{(1)}$ and $a_+ = (|a| + a)/2$. Therefore, the contrast function $C(S^{(1)}) = Q_1(S^{(1)}, 1) - Q_1(S^{(1)}, 0)$ and thus the optimal treatment regime at stage 1 is $I\{C(S^{(1)}) > 0\}$.

To evaluate the double robustness of the proposed method, we consider a variety of scenarios with correctly specified and misspecified baseline mean and/or propensity score models. At stage 2, a linear model with covariates $S^{(1)}$, $S^{(2)}$ and $A^{(1)}$ is fitted for the baseline mean function, while the true baseline mean function is $h^{(2)}(X) = A^{(1)}(\beta_1^T S^{(1)})$. We choose $\beta_1 = 0_q$, for which the baseline mean function is correctly specified, and $\beta_1 = (0_4, 1, -1, 0_{q-6})^T$, for which the baseline mean function is misspecified. At stage 1, a linear model with covariates $S^{(1)}$ is fitted for the baseline mean function, which is always misspecified. Logistic models

TABLE 1
Simulation settings

	Stage	Baseline	Propensity score	Important variables
Setting 1	Stage 2	right	right	$(S^{(2)}, A_1, S_3^{(1)}, S_4^{(1)})$
	Stage 1	wrong	right	$(S_1^{(1)}, S_3^{(1)}, S_4^{(1)})$
Setting 2	Stage 2	wrong	right	$(S^{(2)}, A_1, S_3^{(1)}, S_4^{(1)})$
	Stage 1	wrong	right	$(S_1^{(1)}, S_3^{(1)}, S_4^{(1)}, S_5^{(1)}, S_6^{(1)})$
Setting 3	Stage 2	right	wrong	$(S^{(2)}, A_1, S_3^{(1)}, S_4^{(1)})$
	Stage 1	wrong	right	$(S_1^{(1)}, S_3^{(1)}, S_4^{(1)})$
Setting 4	Stage 2	wrong	wrong	$(S^{(2)}, A_1, S_3^{(1)}, S_4^{(1)})$
	Stage 1	wrong	right	$(S_1^{(1)}, S_3^{(1)}, S_4^{(1)}, S_5^{(1)}, S_6^{(1)})$

are used for estimating the propensity scores, which are correctly specified for the constant model but misspecified for the probit model. The following four settings are considered:

Setting 1: $\beta_1 = 0_q$, $P(A^{(2)} = 1) = 0.5$;
Setting 2: $\beta_1 = (0_4, 1, -1, 0_{q-6})^T$, $P(A^{(2)} = 1) = 0.5$;
Setting 3: $\beta_1 = 0_q$, $P(A^{(2)} = 1) = \Pr(N(0,1) \le S^T\gamma)$;
Setting 4: $\beta_1 = (0_4, 1, -1, 0_{q-6})^T$, $P(A^{(2)} = 1) = \Pr(N(0,1) \le S^T\gamma)$,

where $S = ((S^{(1)})^T, S^{(2)})^T$ and $N(0,1)$ a standard normal random variable. For other parameters, we choose $P(A_1 = 1) = 0.5$, $\beta_2 = (0, 0, 1, -1, 0_{q-4})^T$, $\sigma_1 = \sigma_2 = 0.5$, $d = (d_0, d_1, d_2, d_3)^T = (0, 1, 1, 1)^T$ and $\gamma = (0_{q-2}, 1, -1, 1)^T$. Table 1 summarizes the information of model misspecification for the baseline mean and propensity score models and associated important variables under different settings. In next section, we show simulation results of the four settings with $q = 1000$ and sample size $n = 150/300$ over 500 replications.

4.2. *Competing methods*. We further compare our method with outcome weighted learning [OWL, Zhao et al. (2012)], which is a robust method which estimates individualized treatment rule by directly maximizing the estimated value function. Zhao et al. (2015) further introduced backward outcome weighted learning (BOWL) and simultaneous outcome weighted learning (SOWL) to extend their methods to multiple-stage studies. Here, we consider a double robust version of BOWL (DR-BOWL) for comparison. For a single-stage study, the developed DR-BOWL method is similar to the residual weighted learning method [Zhou et al. (2017)].

Specifically, we first estimate the propensity score $\hat{\pi}^{(2)} = (\hat{\pi}_1^{(2)}, \ldots, \hat{\pi}_n^{(2)})^T$ and baseline $h^{(2)} = X^T\hat{\theta}^{(2)} = (h_1^{(2)}, \ldots, h_n^{(2)})^T$ as in Section 2. We consider the linear

decision rule $I(x^T\beta_{20} > 0)$ and estimate β_{20} by minimizing the following loss function:

$$\tilde{\beta}_2 = \arg\min_{\beta_2} \frac{1}{n}\sum_i \frac{(Y_i - h_i^{(2)})\{1 - (2A_i^{(2)} - 1)X_i^T\beta_2\}_+}{A_i^{(2)}\hat{\pi}_i^{(2)} + (1 - A_i^{(2)})(1 - \hat{\pi}_i^{(2)})} + \lambda_{3n}^{(2)}\|\beta_2\|_1.$$

The penalty term in original OWL is $\lambda_{3n}^{(2)}\|\beta_2\|_2^2$. We replace it with the L_1 norm here to simultaneously select variables. Then we construct the pseudo outcome $\hat{V}_i$ using augmented inverse propensity score estimator [AIPWE, Zhang et al. (2012)],

$$\begin{aligned}\hat{V}_i = {} & \frac{A_i^{(2)}\tilde{d}_2(X_i) + (1 - A_i^{(2)})\{1 - \tilde{d}_2(X_i)\}}{A_i^{(2)}\hat{\pi}_i^{(2)} + (1 - A_i^{(2)})(1 - \hat{\pi}_i^{(2)})}Y_i \\ & - \left(\frac{A_i^{(2)}\tilde{d}_2(X_i) + (1 - A_i^{(2)})\{1 - \tilde{d}_2(X_i)\}}{A_i^{(2)}\hat{\pi}_i^{(2)} + (1 - A_i^{(2)})(1 - \hat{\pi}_i^{(2)})} - 1\right) \\ & \times \left[\hat{h}_i^{(2)}\{1 - \tilde{d}_2(X_i)\} + \hat{\Phi}_i^{(2)}\tilde{d}_2(X_i)\right],\end{aligned}$$

where $\tilde{d}_2(X_i) = I(X_i^T\tilde{\beta}_2 > 0)$, and $\hat{\Phi}_i^{(2)}$ is an estimate of $\Phi_i^{(2)} = \text{Mean}(Y|A = 1, X = X_i)$. Here, we fit a linear model for $\text{Mean}(Y|A = 1, X)$ and use nonconcave penalized regression with SCAD penalty to obtain $\hat{\Phi}_i^{(2)}$. Denoted by $\hat{\pi}^{(1)} = (\hat{\pi}_1^{(1)}, \ldots, \hat{\pi}_n^{(1)})^T$ and $h^{(1)} = S^T\hat{\theta}^{(1)} = (h_1^{(1)}, \ldots, h_n^{(1)})^T$ estimated propensity score and baseline at the first stage, we consider linear treatment regime of the form $I(s^T\beta_1^\star > 0)$ and estimate $\beta_1^\star$ by

$$\tilde{\beta}_1 = \arg\min_{\beta_1} \frac{1}{n}\sum_i \frac{(\hat{V}_i - h_i^{(1)})\{1 - (2A_i^{(1)} - 1)S_i^T\beta_1\}_+}{A_i^{(1)}\hat{\pi}_i^{(1)} + (1 - A_i^{(1)})(1 - \hat{\pi}_i^{(1)})} + \lambda_{3n}^{(1)}\|\beta_1\|_1.$$

Tuning parameters $\lambda_{3n}^{(2)}$ and $\lambda_{3n}^{(1)}$ are obtained by minimizing a value-based BIC criterion.

4.3. *Results*. Table 2 summarizes variable selection results for optimal treatment decisions and the empirical performance of the estimated optimal treatment regime compared with the true optimal regime, using our penalized A-learning method (denoted as PAL) and DR-BOWL, respectively. Specifically, it reports the false negative (FN) rate (the percentage of important variables that are missed) and false positives (FP) rate (the percentage of unimportant variables that are selected), the ratio of value functions (denoted by VR) calculated using the value function of the estimated optimal treatment regime divided by that of the true optimal regime and the error rates (ER) of the estimated optimal treatment regimes for treatment decision making, in both stages. Here, the ER at stage 2 is calculated as the mean of $n^{-1}\sum_{i=1}|I(\hat{\beta}_2^T X_i > 0) - I(\beta_{2,0}^T X_i > 0)|$ and at stage 1 as the

TABLE 2
Variable selection simulation results (%)

		Stage 2				Stage 1			
n	**Method**	**FN**	**FP**	**VR***	**ER**	**FN**	**FP**	**VR**	**ER**
				Setting 1					
150	PAL	12.6	0.1	64.7	6.1	63.8	0.1	98.3	7.0
	DR-BOWL	85.7	0.1	39.0	34.7	99.5	0.1	39.1	48.3
300	PAL	1.1	0.1	65.4	2.6	41.9	0.1	99.7	6.2
	DR-BOWL	78.1	0.1	49.2	27.5	98.0	0.2	50.2	48.3
				Setting 2					
150	PAL	25.9	0.1	57.8	10.4	56.2	0.2	90.8	15.7
	DR-BOWL	86.3	0.1	35.1	35.6	99.0	0.2	35.9	47.2
300	PAL	11.0	0.1	59.6	6.2	32.5	<0.05	97.9	8.0
	DR-BOWL	79.8	0.1	42.4	29.9	97.2	0.2	44.6	47.1
				Setting 3					
150	PAL	33.7	0.3	59.9	13.5	64.5	0.1	93.0	9.1
	DR-BOWL	18.8	1.3	60.2	7.5	72.3	0.5	92.4	24.4
300	PAL	12.3	0.3	64.2	7.2	52.7	<0.05	98.3	6.9
	DR-BOWL	74.9	0.2	55.3	23.2	97.8	<0.01	56.4	48.4
				Setting 4					
150	PAL	55.7	0.2	48.2	22.4	62.2	0.1	79.4	17.7
	DR-BOWL	75.0	0.1	51.0	23.4	99.0	<0.01	51.7	47.2
300	PAL	26.4	0.3	56.2	13.2	36.4	<0.05	94.3	8.4
	DR-BOWL	74.9	0.2	50.9	23.1	97.4	<0.01	52.8	47.0

FN: proportion of related variables with zero coefficients; FP: proportion of unrelated variables with nonzero coefficients; VR: value ratio between estimated and true treatment regimes; ER: error rate of estimated treatment regimes.

mean of $n^{-1}\sum_{i=1}|I(\hat{\beta}_1^T S_i > 0) - I(C(S_i) > 0)|$. The value function of a given treatment regime is calculated using Monte Carlo simulations based on 10,000 replications. The VR at stage 2 (devoted by VR*) is to compare the estimated optimal treatment regime at stage 2 and a randomly assigned treatment at stage 1 as in simulated data with the true optimal dynamic treatment regime for both stages. The VR at stage 1 is to compare the estimated optimal dynamic treatment regime with the true optimal dynamic treatment regime for both stages.

The DR-BOWL methods fail in all settings. Take Setting 1, $n = 300$ as an example, FN $= 78.1\%$ for the second stage where the baseline, propensity score and contrast functions are all correctly specified. It missed approximately $3/4$ of the important variables. Besides, VR $= 50.2$, indicating the poor performance of the estimated treatment rules.

On the other hand, the overall performance of our penalized *A*-learning method is good. We make the following observations. First, the FN rates are much higher

than the FP rates. This suggests that the Dantzig selector tends to have conservative variable selection results, which is commonly seen in the literature. Second, the variable selection results and the error rates of the estimated optimal treatment regime at stage 2 are generally much better than those at stage 1, which is expected since the optimal linear treatment decision rule is correctly specified at stage 2 but not at stage 1. At stage 2, for $n = 150$, over 55% important variables are not selected for all 4 settings. Third, our method requires correct specification of either the propensity score or the baseline model, especially when the sample size is small. This is implied by comparing results in Setting 4 with other three settings. For example, when $n = 150$, the false negative at second stage reaches 55.7%, which is much higher than those FNs in other three settings. Besides, our estimator is very efficient in Setting 1 where both models are correctly specified. Even when $n = 150$, the ratio of the value functions reaches 98.7%, and all error rates are abound 6–7%. These results are even comparable with those under Setting 2 and 3 when $n = 300$. Lastly, the estimation and variable selection performance of the estimated optimal dynamic treatment regimes improves as the sample size increases. In particular, in Settings 1–3 when $n = 300$, the VRs are all above 97.9% and ERs are all below 8%, which implies that the estimated optimal treatment regimes nearly maximize the value functions.

4.4. *Nonregularity.* As suggested by one of the referees, we further examine our methods under settings with different degrees of nonregularity. Specifically, we consider the setting where all covariates in $S^{(1)}$ are independent Rademacher random variables. We set $S^{(2)}$ to be another Rademacher random variable independent of $S^{(1)}$ and $A^{(1)}$.

Denoted by $A^{(1)*} = 2A^{(1)} - 1$, the response Y is generated as follows:

$$Y = 2A^{(2)}\bigl(A^{(1)*} + \delta_1 S_1^{(1)} + S^{(2)} - \delta_2\bigr) + A^{(1)}\bigl(\beta^T S^{(1)}\bigr) + \varepsilon, \tag{4.2}$$

where $\varepsilon \sim N(0, 0.25)$.

For each stage, we fit linear models for the baseline and contrast function, and a logistic regression model for the propensity score. The parameter β in (4.2) determines the baseline function on the second stage. Similar to the regular case discussed in Section 4.1 in the revision, we also consider four settings here:

Setting 1: $\beta = 0_q$, $P(A^{(2)} = 1) = 0.5$;
Setting 2: $\beta = (0_4, 1, -1, 0_{q-6})^T$, $P(A^{(2)} = 1) = 0.5$;
Setting 3: $\beta = 0_q$, $P(A^{(2)} = 1) = \Pr(N(0, 1) \le S^T\gamma)$;
Setting 4: $\beta = (0_4, 1, -1, 0_{q-6})^T$, $P(A^{(2)} = 1) = \Pr(N(0, 1) \le S^T\gamma)$,

where $S = ((S^{(1)})^T, S^{(2)})^T$ and $\gamma = (0_{q-2}, 1, -1, 1)^T$.

TABLE 3
Simulation for nonregular settings

	Stage	Baseline	Propensity score	Important variables
Setting 1	Stage 2	right	right	$(S^{(2)}, A_1, S_1^{(1)})$
	Stage 1	right	right	$(S_1^{(1)})$
Setting 2	Stage 2	wrong	right	$(S^{(2)}, A_1, S_1^{(1)})$
	Stage 1	right	right	$(S_1^{(1)}, S_5^{(1)}, S_6^{(1)})$
Setting 3	Stage 2	right	wrong	$(S^{(2)}, A_1, S_1^{(1)})$
	Stage 1	right	right	$(S_1^{(1)})$
Setting 4	Stage 2	wrong	wrong	$(S^{(2)}, A_1, S_1^{(1)})$
	Stage 1	right	right	$(S_1^{(1)}, S_5^{(1)}, S_6^{(1)})$

Parameters δ_1 and δ_2 in (4.2) controls the degree of nonregularity on the second stage. We consider three choices of δ_1 and δ_2. Set $\delta_1 = 1$, $\delta_2 = 1$, we obtain

$$\Pr\bigl(C^{(2)}(X) = 0\bigr) = \Pr\bigl(A^{(1)*} + S_1^{(1)} + S^{(2)} = 1\bigr) = 0.375.$$

Set $\delta_1 = 1.1$, $\delta_2 = 1.1$, we have

$$\Pr\bigl(C^{(2)}(X) = 0\bigr) = \Pr\bigl(A^{(1)*} + S^{(2)} = 1, S^{(1)} = 1\bigr) = 0.25.$$

Set $\delta_1 = 1$, $\delta_2 = 1.1$, we have

$$\Pr\bigl(C^{(2)}(X) = 0\bigr) = 0.$$

With some calculation, we can show the Q-function on the first stage takes the following form:

$$Q\bigl(S^{(1)}, A^{(1)}\bigr) = A^{(1)}\bigl(\beta^T S^{(1)} + f_1 S_1^{(1)} + f_2\bigr).$$

Hence, the contrast function is correctly specified on the first stage. When $\delta_1 = 1$, $\delta_2 = 1$ or $\delta_1 = 1.1$, $\delta_2 = 1.1$, we have $f_1 = f_2 = 1$. When $\delta_1 = 1$, $\delta_2 = 1.1$, we have $f_1 = f_2 = 0.95$. Information about model specification and important variables in the contrast function are given in Table 3.

We also consider two choices of sample size, $n = 150$ and $n = 300$, respectively. This gives us a total of 24 scenarios. For each scenario, we report FN, FP, VR and ER as Section 4.3. ER for the first and second stage are calculated as

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\bigl|I\bigl(\hat{\beta}_1^T S_i > 0\bigr) - I\bigl(C(S_i) > 0\bigr)\bigr|I\bigl(C(S_i) \neq 0\bigr)\right\} \Big/ \left\{\frac{1}{n}\sum_{i=1}^{n} I\bigl(C(S_i) \neq 0\bigr)\right\}$$

and

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\bigl|I\bigl(\hat{\beta}_2^T X_i > 0\bigr) - I\bigl(\beta_{2,0}^T X_i > 0\bigr)\bigr|I\bigl(\beta_{2,0}^T X_i \neq 0\bigr)\right\} \Big/ \left\{\frac{1}{n}\sum_{i=1}^{n} I\bigl(\beta_{2,0}^T X_i \neq 0\bigr)\right\}.$$

TABLE 4
Variable selection simulation results for nonregular settings (%)

		Stage 2				Stage 1			
n	**Nonregularity**	**FN**	**FP**	**VR***	**ER**	**FN**	**FP**	**VR**	**ER**
				Setting 1					
150	$\delta_1 = 1, \delta_2 = 1$	0	<0.01	53.6	0	4.0	0.3	93.9	5.0
	$\delta_1 = 1.1, \delta_2 = 1.1$	0	<0.01	53.5	0.1	0.5	0.3	95.1	3.2
	$\delta_1 = 1, \delta_2 = 1.1$	0	0.01	52.9	5.0	1.0	0.3	93.8	4.2
300	$\delta_1 = 1, \delta_2 = 1$	0	<0.01	53.3	0	0	0.4	97.3	0.9
	$\delta_1 = 1.1, \delta_2 = 1.1$	0	<0.01	53.9	0	0	0.3	97.9	0.9
	$\delta_1 = 1, \delta_2 = 1.1$	0	<0.01	52.8	2.0	0	0.3	97.0	0.9
				Setting 2					
150	$\delta_1 = 1, \delta_2 = 1$	0	<0.05	46.1	0	14.7	0.3	90.7	5.6
	$\delta_1 = 1.1, \delta_2 = 1.1$	0	<0.05	45.9	2.0	14	0.3	89.5	5.7
	$\delta_1 = 1, \delta_2 = 1.1$	0	<0.05	44.4	11.6	9.7	0.3	89.7	11.5
300	$\delta_1 = 1, \delta_2 = 1$	0	<0.01	45.8	0	0	0.2	97.1	0.4
	$\delta_1 = 1.1, \delta_2 = 1.1$	0	<0.01	45.6	0.5	0	0.2	96.4	0.4
	$\delta_1 = 1, \delta_2 = 1.1$	0	0.01	45.1	6.8	0	0.2	98.1	7.4
				Setting 3					
150	$\delta_1 = 1, \delta_2 = 1$	5.7	0.6	45.0	2.9	19.0	0.2	85.6	4.1
	$\delta_1 = 1.1, \delta_2 = 1.1$	8.2	0.5	45.1	6.6	17.0	0.2	87.3	3.4
	$\delta_1 = 1, \delta_2 = 1.1$	6.3	0.6	44.6	14.6	18.5	0.2	85.8	4.1
300	$\delta_1 = 1, \delta_2 = 1$	0	0.1	53.1	0	0	0.3	97.1	1.0
	$\delta_1 = 1.1, \delta_2 = 1.1$	0	0.1	53.8	1.4	0	0.3	98.0	0.6
	$\delta_1 = 1, \delta_2 = 1.1$	0	0.1	52.9	8.4	0	0.3	97.9	0.8
				Setting 4					
150	$\delta_1 = 1, \delta_2 = 1$	20.7	0.5	25.4	8.6	52	0.2	66.6	14.0
	$\delta_1 = 1.1, \delta_2 = 1.1$	20.8	0.5	25.3	12.4	54.2	0.2	62.5	14.9
	$\delta_1 = 1, \delta_2 = 1.1$	21.5	0.6	23.7	22.6	51.7	0.2	61.7	22.1
300	$\delta_1 = 1, \delta_2 = 1$	0.3	0.2	44.9	0.2	3.3	0.2	95.8	0.7
	$\delta_1 = 1.1, \delta_2 = 1.1$	0	0.2	44.8	3.9	0.2	0.2	97.5	0.4
	$\delta_1 = 1, \delta_2 = 1.1$	0	0.2	43.8	13.2	0.3	0.2	97.1	8.3

Compared to definitions in Section 4.3, error rates here are calculated with respect to those patients with nonzero contrast functions. Such definitions are more meaningful since both two treatments are optimal for these patients. We simulate over 200 replications. Results are reported in Table 4.

Within each setting, most results are similar across different choices of δ_1 and δ_2. This suggests the nonregularity issues do not have a big impact on the variable selection results. Apart from results in Setting 4, false negatives and false positives are all very small. When the sample size increases to 300, false negatives for most scenarios are exactly equal to 0 while false positives for all settings are below 0.4%, demonstrating perfect variables selections performance of our methods. In

Settings 1–3, most error rates are below 7% while the ratios of value function are all above 85%, indicating our estimated optimal treatment regimes are very close to the truth in these scenarios.

5. Application to STAR*D study. We applied the proposed method to a dataset from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study, which was conducted to compare different treatments for patients with major depressive disorder (MDD). There were 4041 participants (ages 18–75) with nonpsychotic MDD enrolled in this study. At the first level, all participants were treated with citalopram (CIT) up to 14 weeks. Subsequently, three more levels of treatments were provided for participants without a satisfactory response to CIT. At each level, participants were randomly assigned to treatment options acceptable to them. At Level 2, participants were eligible for seven treatment options: sertraline (SER), venlafaxine (VEN), bupropion (BUP), cognitive therapy (CT) and augmenting CIT with bupropion (CIT + BUP), buspirone (CIT + BUS) or cognitive therapy (CIT + CT). Participants without a satisfactory response to CT were proceeded to Level 2A for additional medication treatments. All participants who did not respond satisfactorily at Level 2 or 2A were eligible for four treatments at Level 3: medication switch to mirtazapine (MIRT) or nortriptyline (NTP), and medication augmentation with either lithium (Li) or thyroid hormone (THY). Participants without satisfactory response to Level 3 were rerandomized at Level 4 to either tranylcypromine (TCP) or a combination of mirtazapine and venlafaxine (MIRT+VEN). See Fava et al. (2003) and Rush et al. (2004) for more details of the STAR*D study. One goal of the study is to determine which treatment strategies, in what order or sequence, provide the optimal treatment effect.

As an illustration, we focused on a subset of participants who were given treatment BUP or SER at Level 2 and did not receive satisfactory responses, and then were randomized to treatment MIRT or NTP at Level 3. For this study, we considered 381 covariates collected at baseline and intermediate levels as possible relevant predictors. For treatment regime at Level 3, all the 381 covariates as well as the assigned treatment at Level 2 were considered as possible predictors for making optimal treatment decision. For treatment regime at Level 2, 305 covariates that were collected before giving treatment at Level 2 were considered for making optimal treatment decision. Negative 16-item Quick Inventory of Depressive Symptomatology-Clinician-Rated (QIDS-C_{16}) was used as the final response, which is a measurement of symptomatic status of depression. There are 73 participants who had complete records in the subset of data we are interested in. Among these participants, 36 were treated with BUP and 37 were treated with SER at Level 2, and 33 were treated with NTP and 40 were treated with MIRT at Level 3.

The selection and estimation results are summarized as follows. At Level 3, our method selected two covariates: "age" in baseline demographics (AGE), and the suicide risk of the patient (SUICD). The estimated optimal treatment regime is $I(1.459 - 0.091 \times \text{AGE} + 0.158 \times \text{SUICD} \geq 0)$, where 1 represents treatment

NTP and 0 represents treatment MIRT. This optimal treatment regime assigns 27 participants to NTP and the rest 46 participants to MIRT. At Level 2, our method also selected two covariates: age and QIDS-C percent improvement" in clinic visit clinical record form at Level 1 (QCIMP). The estimated optimal treatment regime is $I(-8.600+0.145\times \mathrm{AGE}+0.125\times \mathrm{QCIMP}\geq 0)$, where 1 stands for treatment BUP and 0 stands for treatment SER. This optimal treatment regime assigns 37 participants to BUP and the rest 36 participants to SER.

To further examine the estimated optimal dynamic treatment regime, we compare the estimated value function of our estimated optimal treatment regime with values under those four nondynamic treatment regimes, BUP + NTP, BUP + MIRT, SER + NTP and SER + MIRT. For a given dynamic treatment regime $d=(d^{(1)}, d^{(2)})$, we evaluate its average value function using AIPWE [Zhang et al. (2013)]:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}_{A_i}^{(1)}}{\hat{\pi}_{A_i}^{(1)}}\left(\frac{\mathrm{d}_{A_i}^{(2)}}{\hat{\pi}_{A_i}^{(2)}}Y_i-\frac{\mathrm{d}_{A_i}^{(2)}-\hat{\pi}_{A_i}^{(2)}}{\hat{\pi}_{A_i}^{(2)}}\{d_i^{(2)}(\hat{h}_i^{(2)}+X_i^T\hat{\beta}_2)+(1-d_i^{(2)})\hat{h}_i^{(2)}\}\right)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}_{A_i}^{(1)}-\hat{\pi}_{A_i}^{(1)}}{\hat{\pi}_{A_i}^{(1)}}\{d_i^{(1)}(\hat{h}_i^{(1)}+S_i^T\hat{\beta}_1)+(1-d_i^{(1)})\hat{h}_i^{(1)}\},$$

where $d_{A_i}^{(2)}=A_i^{(2)}d_i^{(2)}+(1-A_i^{(2)})$, $d_{A_i}^{(1)}=A_i^{(1)}d_i^{(1)}+(1-A_i^{(1)})(1-d_i^{(1)})$, $\hat{\pi}_{A_i}^{(2)}=A_i^{(2)}\hat{\pi}_i^{(2)}+(1-A_i^{(2)})(1-\hat{\pi}_i^{(2)})$, $\hat{\pi}_{A_i}^{(1)}=A_i^{(1)}\hat{\pi}_i^{(1)}+(1-A_i^{(1)})(1-\hat{\pi}_i^{(1)})$, $d_i^{(2)}$ and $d_i^{(1)}$ the assigned treatment for the ith patient, according to $d^{(2)}$ and $d^{(1)}$. Based on this formula, we report the estimated value functions of the four nondynamic treatment regimes in Table 5.

Estimating the value of the optimal treatment regime is well known to be a nonregular problem when there is nonzero probability that the contrast function (either at the second or the first stage) is equal to zero. To evaluate the value function under our estimated optimal treatment regime, we consider the online estimator proposed by Luedtke and van der Laan (2016). Specifically, for $i=l_n+1, l_n+2, \ldots, n$, we obtain the estimated optimal dynamic treatment regime

TABLE 5
Estimated values of different treatment regimes

Treatment regime	**Estimated value**
Estimated optimal regime	−9.02
BUP + NTP	−12.86
BUP + MIRT	−12.57
SER + NTP	−12.57
SER + MIRT	−12.28

$\hat{d}^{\mathrm{opt}(i)} = (\hat{d}^{\mathrm{opt}(i)(1)}, \hat{d}^{\mathrm{opt}(i)(2)})$ and its associated parameters $\hat{\beta}_2^{(i)}$, $\hat{\beta}_1^{(i)}$, propensity score function $\hat{\pi}^{(i)(2)}$, $\hat{\pi}^{(i)(1)}$, baseline function $\hat{h}^{(i)(2)}$, $\hat{h}^{(i)(1)}$ based on data from patients 1 to $i-1$, using penalized A-learning. Then we evaluate the value of $\hat{d}^{\mathrm{opt}(i)} = (\hat{d}^{\mathrm{opt}(i)(1)}, \hat{d}^{\mathrm{opt}(i)(2)})$ on the ith patient using [AIPWE, Zhang et al. (2013)]

$$\begin{aligned}\hat{V}_i(i) = {} & \frac{\hat{d}_{A_i}^{\mathrm{opt}(i)(1)}}{\hat{\pi}_{A_i}^{(i)(1)}}\Bigg(\frac{\hat{d}_{A_i}^{\mathrm{opt}(i)(2)}}{\hat{\pi}_{A_i}^{(i)(2)}} Y_i \\ & - \frac{\hat{d}_{A_i}^{\mathrm{opt}(i)(2)} - \hat{\pi}_{A_i}^{(i)(2)}}{\hat{\pi}_{A_i}^{(i)(2)}}\{\hat{d}_i^{\mathrm{opt}(i)(2)}(\hat{h}_i^{(i)(2)} + X_i^T\hat{\beta}_2^{(i)}) + (1-d_i^{(2)})\hat{h}_i^{(i)(2)}\}\Bigg) \\ & - \frac{\hat{d}_{A_i}^{\mathrm{opt}(i)(1)} - \hat{\pi}_{A_i}^{(i)(1)}}{\hat{\pi}_{A_i}^{(i)(1)}}\{\hat{d}_i^{\mathrm{opt}(i)(1)}(\hat{h}_i^{(i)(1)} + S_i^T\hat{\beta}_1^{(i)}) \\ & + (1-d_i^{\mathrm{opt}(i)(1)})\hat{h}_i^{(i)(1)}\}.\end{aligned}$$

The variance of $\hat{V}_i(i)$ conditional on data from patients 1 to $i-1$ is evaluated by

$$\tilde{\sigma}_i^2 = \frac{1}{i-1}\sum_{j=1}^{i-1}\hat{V}_i^2(j) - \left(\frac{1}{i-1}\sum_{j=1}^{i-1}\hat{V}_i(j)\right)^2,$$

where $\hat{V}_i(j)$ is the estimated value of $\hat{d}^{\mathrm{opt}(i)}$ on the jth patient.

The final estimator is given by

$$\hat{V} = \frac{\sum_{j=l_n+1}^{n}\tilde{\sigma}_j^{-1}\hat{V}_j(j)}{\sum_{j=l_n+1}^{n}\tilde{\sigma}_j^{-1}},$$

with the estimated standard error

$$\hat{\sigma} = \frac{\sqrt{n-l_n}}{\sum_{j=l_n+1}^{n}\tilde{\sigma}_j^{-1}}.$$

Since the sample size of our dataset is small, we choose $l_n \approx 2n/3$, that is, $l_n = 49$. The estimated value $\hat{V}$ is equal to -9.02 with an estimated standard error $\hat{\sigma} = 1.66$. From Table 5, we can see the value under our estimated treatment regime is much larger than those under four nondynamic treatment regime.

6. Oracle inequalities for $\hat{\beta}_2$ and the value function of the estimated regime at the second stage. We first introduce some notation. For an arbitrary matrix $\Phi \in \mathbb{R}^{M\times M}$ and an arbitrary vector $\phi \in \mathbb{R}^M$, the superscript Φ^j is used to denote the jth column of Φ, ϕ^j the jth element of ϕ, while the subscript Φ_i denotes the

ith row of Φ. For subsets $J, J' \subset \{1, \ldots, M\}$, let $|J|$ be the cardinality of J, J^c be the complement of J. We denote by ϕ^J the vector in $\mathbb{R}^{|J|}$ that has the same coordinates as ϕ on J, and Φ^J the submatrix formed by columns in J, $\Phi_J^{J'}$ the submatrix formed by rows in J and columns in J'. The support of ϕ is defined by $\operatorname{supp}(\phi) = \{j \in \{1, \ldots, M\} : \phi^j \neq 0\}$. Let $\|\phi\|_p$ be the L_p norm of ϕ, $\|\Phi\|_p$ be the operator norm corresponding to the p-norm vector. If Φ is positive semidefinite, define

$$\rho_{\min}^s(\Phi) = \min_{\substack{\|y|_2=1 \\ |\operatorname{supp}(y)|\le s}} \|\Phi^{1/2} y\|_2 \quad \text{and} \quad \rho_{\max}^s(\Phi) = \max_{\substack{\|y\|_2=1 \\ |\operatorname{supp}(y)|\le s}} \|\Phi^{1/2} y\|_2.$$

Let $\|Y\|_{\psi_p}$ be the Orlicz norm for any random variable Y, defined as

$$\|Y\|_{\psi_p} \triangleq \inf_u \left\{ u > 0 : \mathrm{E} \exp\left(\frac{|Y|}{u} \right)^m \le 2 \right\},$$

for some $p \ge 1$. For any two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \gg b_n$ means $\lim_n b_n/a_n = 0$. Throughout this paper, we use c_0 and $\bar{c}$ to denote some universal constants, whose values may change from place to place.

6.1. *Oracle inequality for* $\hat{\beta}_2$. Recall $C^{(2)}(x) = x^T \beta_2$, according to our assumption. Let $\beta_{2,0}$ denote the true values of β_2. Define $s_{\beta_2} = |M_{\beta_2}| = O(n^{l_6})$ for some $0 \le l_6 < 1$, the nonsparsity size of $\beta_{2,0}$, M_{β_2} the support of $\beta_{2,0}$. We allow the number of covariates p to grow exponentially fast with respect to the sample size n, that is, $\log p = O(n^{a_2})$ for some $0 < a_2 < 1$. To deal with such NP-dimensionality, following Zhou (2009), we assume

$$X = U \Sigma^{1/2}, \qquad \Sigma_{jj} = 1, \qquad \forall j = 1, \ldots, p, \tag{6.1}$$

where $U = (U_1^T, \ldots, U_n^T)^T$ and $U_1, \ldots, U_n$ are i.i.d. copies of a p-dimensional isotropic random vector U_0. More specifically, we require that for any vector $a \in \mathbb{R}^p$,

$$\mathrm{E}(a^T U_0)^2 = a^T a \quad \text{and} \quad \|a^T U_0\|_{\psi_2} \le \omega \|a\|_2, \tag{6.2}$$

for some isotropic constants ω.

REMARK 6.1. The definition of the isotropic random vector was first introduced by Milman and Pajor (2003). Independent normal and independent Rademacher random variables are two most important examples of isotropic random vectors. More generally, coordinates of the isotropic random vector do not need to be independent. They can be distributed uniformly on various convex and symmetric bodies, for example, an appropriate multiple of the unit ball in $\mathbb{R}^p$ equipped with the L_K-norm for any $1 \le K \le \infty$. For these distributions, we denote ω_K as their isotropic constants. It is further shown in [Milman and Pajor (1989)] that ω_K are uniformly bounded for $K \ge 1$. However, it remains unknown whether the isotropic property holds for all uniform distributions on arbitrary symmetric convex bodies with Lebesgue measure 1.

REMARK 6.2. The isotropic formulation requires covariates in U_0 to be uncorrelated, and hence does not allow for correlated Bernoullis. However, according to our definition $X = U\Sigma^{1/2}$, different covariates in the design matrix X can be correlated when $\Sigma_{ij} \neq 0$. Such formulations allows us to impose conditions on the tail of U_0 and the covariance matrix Σ separately.

Since the A-learning estimating equation involves the plug-in estimators $\hat{\alpha}_2$ and $\hat{\theta}_2$, we need some conditions on these two estimators to establish oracle inequalities for $\hat{\beta}_2$. More precisely, we assume that $\hat{\alpha}_2$ and $\hat{\theta}_2$ converge to some $\alpha_2^\star$ and $\theta_2^\star$, respectively. When the propensity score model $\pi^{(2)}$ and the baseline model $h^{(2)}$ are correctly specified, $\alpha_2^\star$ and $\theta_2^\star$ represent the true coefficients in these two models. When the models are misspecified, $\alpha_2^\star$ and $\theta_2^\star$ correspond to the population-level least favorable parameters. Denote M_{α_2} and M_{θ_2} the support of $\alpha_2^\star$ and $\theta_2^\star$, respectively. Let $s_{\alpha_2} = |M_{\alpha_2}|$ and $s_{\theta_2} = |M_{\theta_2}|$, the number of nonzero elements. We assume $s_{\alpha_2} = O(n^{l_4})$ and $s_{\theta_2} = O(n^{l_5})$ for some $0 \leq l_4, l_5 < 1/2$.

CONDITION 1. Assume that there exist some positive constants γ_{α_2} and γ_{θ_2}, such that with probability at least $1 - \bar{c}/(n+p)$,

$$\hat{\alpha}_2^{M_{\alpha_2}^c} = 0, \qquad \big\|\hat{\alpha}_2^{M_{\alpha_2}} - \alpha_2^{\star M_{\alpha_2}}\big\|_\infty = O\big(n^{-\gamma_{\alpha_2}} \log n\big), \tag{6.3}$$

$$\hat{\theta}_2^{M_{\theta_2}^c} = 0, \qquad \big\|\hat{\theta}_2^{M_{\theta_2}} - \theta_2^{\star M_{\theta_2}}\big\|_\infty = O\big(n^{-\gamma_{\theta_2}} \log n\big). \tag{6.4}$$

Moreover, assume $d_{\alpha_2} \gg n^{-\gamma_{\alpha_2}} \log n$ and $d_{\theta_2} \gg n^{-\gamma_{\theta_2}} \log n$, where $d_{\alpha_2} = \min_j |\alpha_2^{\star j}|/2$ and $d_{\theta_2} = \min_j |\theta_2^{\star j}|/2$.

REMARK 6.3. Condition 1 assumes the weak oracle properties of $\hat{\alpha}_2$ and $\hat{\theta}_2$, that is, selection consistency and consistency under L_∞ norm. The weak oracle properties of $\hat{\alpha}_2$ and $\hat{\theta}_2$ are established in Theorems 8.1 and 8.2 of Section 8, respectively.

Define

$$C^{(2)} = E\big\{X_i \pi_i^{(2)*}\big(1 - \pi_i^{(2)*}\big) X_i^T\big\}, \qquad D^{(2)} = \mathrm{E}\big\{X_i X_i^T \big(1 - A_i^{(2)}\big)\big\}$$

and $\pi_i^{(2)*} \equiv \pi^{(2)}(X_i, \alpha_2^\star)$.

CONDITION 2. Assume that matrices $D^{(2)}$, $C^{(2)}$ and Σ satisfy

$$\lambda_{\max}\big(\Sigma_{M_{\alpha_2}}^{M_{\alpha_2}}\big) = O(1), \qquad \lambda_{\max}\big(\Sigma_{M_{\theta_2}}^{M_{\theta_2}}\big) = O(1),$$

$$\liminf_n \lambda_{\min}\big(D_{M_{\theta_2}}^{(2)M_{\theta_2}}\big) > 0, \qquad \liminf_n \lambda_{\min}\big(C_{M_{\alpha_2}}^{(2)M_{\alpha_2}}\big) > 0.$$

Define $\Omega^{(2)}(\alpha_2) = \mathrm{E}[X_i X_i^T A_i^{(2)}\{1-\pi^{(2)}(X_i,\alpha_2)\}]$ and $\Omega_n^{(2)} = n^{-1}\sum_i X_i X_i^T \times A_i^{(2)}(1-\hat{\pi}_i^{(2)})$ with $\hat{\pi}_i^{(2)} = \pi^{(2)}(X_i,\hat{\alpha}_2)$. For any positive semidefinite matrix $\Psi \in \mathbb{R}^{p\times p}$, integer s and positive number c, define function $K(s,c,\Psi)$ as follows:

$$K(s,c,\Psi) = \min_{\substack{J\subset\{1,\ldots,p\}\\ |J|\le s}} \min_{\substack{y\neq 0\\ \|y^{J^c}\|_1 \le c\|y^J\|_1}} \frac{\|\Psi^{1/2}y\|_2}{\|y^J\|_2} > 0.$$

The following condition ensures that the RE condition holds for the matrix $\Omega_n^{(2)}$.

CONDITION 3. Assume that for any $0<\theta_s<1$ and sufficiently large n, we have

$$K\big(s_{\beta_2},1,\Omega_n^{(2)}\big) > (1-\theta_s)\inf_{\alpha_2\in H_{\alpha_2}} K\big(s_{\beta_2},1,\Omega^{(2)}(\alpha_2)\big) > 0, \tag{6.5}$$

where H_{α_2} denotes the set of vectors α_2 that satisfies the weak oracle property (6.3).

REMARK 6.4. It is tedious to verify (6.5) due to the plug-in estimator $\hat{\pi}_i^{(2)}$. The key to prove such a result is that the estimator $\hat{\alpha}_2$ in $\hat{\pi}_i^{(2)}$ should be sparse. That is the reason we use penalized regression with a folded-concave penalty to obtain $\hat{\alpha}_2$, since it can ensure selection consistency of the estimator. We provide a general result characterizing the UUP and RE conditions for the random matrix $\Omega_n^{(2)}$ in Lemmas 9.1 and 9.2 of Section 9.

To establish the oracle inequality for $\hat{\beta}_2$, we first provide an upper bound for

$$\Big\|\frac{1}{n}X^T \operatorname{diag}\big(A^{(2)}-\hat{\pi}^{(2)}\big)\big(Y - X\hat{\theta}_2 - A^{(2)}\circ X\beta_{2,0}\big)\Big\|_\infty,$$

which is given in the following lemma.

LEMMA 6.1. *Assume that Conditions* 1 *and* 2 *hold,* $\|h^{(2)}(X_i) - X_i^T\theta_2^\star\|_{\psi_1} < \infty$, $\|e_i\|_{\psi_2} < \infty$, $a_2 + l_4 < 1$, *and that either* $\pi^{(2)}$ *or* $h^{(2)}$ *is correctly specified. Then, for sufficiently large* n, *there exist some constants* $c^{(2)}$, *such that with probability at least* $1-\bar{c}/(n+p)$,

$$\begin{aligned}&\Big\|\frac{1}{n}X^T \operatorname{diag}\big(A^{(2)}-\hat{\pi}^{(2)}\big)\big(Y - X\hat{\theta}_2 - A^{(2)}\circ X\beta_{2,0}\big)\Big\|_\infty\\ &\quad\le c^{(2)}(E_1+E_2+E_3+E_4),\end{aligned}$$

where

$$\begin{aligned}E_1 &= \sqrt{\log p/n}, \qquad E_2 = s_{\alpha_2}n^{-2\gamma_{\alpha_2}}\log^2 n + s_{\theta_2}n^{-2\gamma_{\theta_2}}\log^2 n,\\ E_3 &= \sigma_3\{\sqrt{s_{\alpha_2}\log n/n} + \sqrt{s_{\alpha_2}}\lambda_{1n}^{(2)}\rho_2^{(1)}(d_{n\alpha_2})\},\\ E_4 &= \sigma_4\{\sqrt{s_{\theta_2}\log n/n} + \sqrt{s_{\theta_2}}\lambda_{2n}^{(2)}\rho_2^{(2)}(d_{n\theta_2})\},\end{aligned}$$

$\sigma_3^2 = \mathrm{E}\{h^{(2)}(X_i) - X_i^T\theta_2^\star\}^2$, *and* $\sigma_4^2 = \mathrm{E}\{\pi^{(2)}(X_i) - \pi_i^{(2)*}\}^2$.

REMARK 6.5. Recall that $\log p = O(n^{a_2})$, $s_{\alpha_2} = O(n^{l_4})$ for some $0 \geq a_2$, $l_4 < 1$. The condition $a_2 + l_4 < 1$ implies $n \gg s_{\alpha_2} \log p$.

REMARK 6.6. Here, E_1 describes how the curse of dimensionality takes effect, E_2 is due to estimation errors of $\hat{\alpha}^{(2)}$ and $\hat{\theta}^{(2)}$, E_3 and E_4 are due to model misspecification. Since we assume that at least one of $h^{(2)}$ and $\pi^{(2)}$ is correctly specified, either E_3 or E_4 is zero.

THEOREM 6.1. *Assume that conditions in Lemma 6.1 and Condition 3 hold, and* $\lambda_{3n}^{(2)} \geq c^{(2)}(E_1 + E_2 + E_3 + E_4)$ *where the constant* $c^{(2)}$ *is defined in Lemma 6.1. Then, for some fixed* $0 < \theta_s < 1$ *and sufficiently large* n, *the following two inequalities hold with probability at least* $1 - \bar{c}/(n+p)$ *for some constant* $\bar{c} > 0$:

$$\|\hat{\beta}_2 - \beta_{2,0}\|_2 \leq \frac{12\lambda_{3n}^{(2)}\sqrt{s_{\beta_2}}}{(1-\theta_s)^2 \inf_{\alpha_2 \in H_{\alpha_2}} K^2(s_{\beta_2}, 1, \Omega^{(2)}(\alpha_2))}, \tag{6.6}$$

$$\|\hat{\beta}_2 - \beta_{2,0}\|_1 \leq \frac{8\lambda_{3n}^{(2)} s_{\beta_2}}{(1-\theta_s)^2 \inf_{\alpha_2 \in H_{\alpha_2}} K^2(s_{\beta_2}, 1, \Omega^{(2)}(\alpha_2))}. \tag{6.7}$$

Moreover, we have $\|\hat{\beta}_2^{M_{\beta_2}^c}\|_1 \leq \|\hat{\beta}_2^{M_{\beta_2}} - \beta_{2,0}^{M_{\beta_2}}\|_1$.

From (6.6), it is immediate to see that $\|\hat{\beta}_2 - \beta_{2,0}\|_2 \xrightarrow{P} 0$ as long as

$$\frac{\sqrt{s_{\beta_2}}(E_1 + E_2 + E_3 + E_4)}{\inf_{\alpha_2 \in H_{\alpha_2}} K^2(s_{\beta_2}, 1, \Omega^{(2)}(\alpha_2))} \to 0, \tag{6.8}$$

which implies the doubly robust property of $\hat{\beta}_2$. We provide a sufficient condition for (6.8) in the following corollary.

COROLLARY 6.1 (Double robustness of $\hat{\beta}_2$). *Assume that conditions in Theorem 6.1 and the following conditions hold*:

$$l_6 < \min(4\gamma_{\theta_2} - 2l_5, 4\gamma_{\alpha_2} - 2l_4), \tag{6.9}$$

$$\lambda_{2n}^{(2)}\rho_2^{(2)}(d_{n\theta_2}) = O\big(n^{-1/2}\big) \quad \textit{and} \quad \lambda_{1n}^{(2)}\rho_1^{(2)}(d_{n\alpha_2}) = O\big(n^{-1/2}\big). \tag{6.10}$$

$$\liminf \inf_{\alpha_2 \in H_{\alpha_2}} K\big(s_{\beta_2}, 1, \Omega^{(2)}(\alpha_2)\big) > 0. \tag{6.11}$$

If either the baseline $h^{(2)}$ *or the propensity score model* $\pi^{(2)}$ *is correctly specified, then* $\|\hat{\beta}_2 - \beta_{2,0}\|_2 \xrightarrow{P} 0$.

REMARK 6.7. Condition (6.9) imposes a constraint between the sparsity of population parameters and the convergence rates of $\hat{\alpha}_2$ and $\hat{\theta}_2$. When $s_{\beta_2} = O(1)$, it requires $\hat{\alpha}_2$ and $\hat{\theta}_2$ to be consistent under L_2 norm. Condition (6.10) automatically holds for SCAD penalty function when $d_{n\theta_2} \gg \lambda_{2n}^{(2)}$ and $d_{n\alpha_2} \gg \lambda_{1n}^{(2)}$.

6.2. *Oracle inequality for the value function of the estimated regime at the second stage.* Now we establish the error bound for the difference between the mean responses (i.e., the value functions) of the estimated optimal regime at the second stage $\hat{d}_2(X_0) = I(X_0^T \hat{\beta}_2 > 0)$ and the true optimal one $d_2^{opt}(X_0) = I(X_0^T \beta_{2,0} > 0)$ for an individual with covariate X_0. Here, X_0 is also assumed to have the form $\Sigma^{1/2}U$ with Σ and U defined in (6.1), independent of $X_i, i = 1, \ldots, n$. In addition, the regime at the first stage is chosen the same as the actually received treatment $A_0^{(1)}$ at the first stage.

Under the assumptions of SUTVA and no unmeasured confounders, the difference of the corresponding value functions is given by

$$\begin{aligned} &\mathrm{E}\{Y_0^\star(A_0^{(1)}, d_2^{\mathrm{opt}})\} - \mathrm{E}\{Y_0^\star(A_0^{(1)}, \hat{d}_2)\} \\ &\quad = \mathrm{E}[X_0^T \beta_{2,0}\{I(X_0^T \beta_{2,0} > 0) - I(X_0^T \hat{\beta}_2 > 0)\}]. \end{aligned} \tag{6.12}$$

Since (6.12) is nonnegative, it suffices to provide an upper bound. Here, we impose the following condition.

CONDITION 4. The probability density function $\mathrm{g}^{(2)}(\cdot)$ of $X_0^T \beta_{2,0}$ exists and is bounded.

Condition 4 is a mild condition on the true optimal decision function, which holds in most cases when at least one of the important covariates (the corresponding component of $\beta_{2,0}$ is nonzero) is continuous.

THEOREM 6.2. *Assume that conditions in Theorem 6.1 and Condition 4 hold. Assume* $\mathrm{E}(X_0^T \beta_{2,0})^2 = O(1)$. *Then, for fixed* $0 < \theta_s < 1$ *and sufficiently large* n,

$$\begin{aligned} &\mathrm{E}[X_0^T \beta_{2,0}\{I(X_0^T \beta_{2,0} > 0) - I(X_0^T \hat{\beta}_2 > 0)\}] \\ &\quad \le \frac{\bar{c}\omega}{n} + \frac{c_0 \omega^2 \rho_{\max}^{s_{\beta_2}}(\Sigma)(\lambda_{3n}^{(2)})^2 s_{\beta_2} \log^2 n}{(1-\theta_s)^4 \inf_{\alpha_2 \in H_{\alpha_2}} K^4(s, 1, \Omega^{(2)}(\alpha_2))}. \end{aligned}$$

REMARK 6.8. Error bound for the difference of the value functions follows from the error bound on $\hat{\beta}_2$ and Condition 4. Since the first term in the upper bound is small, the difference of the value functions is mainly characterized by the second term, which is of the order $O(\rho_{\max}^{s_{\beta_2}}(\Sigma)\|\hat{\beta}_2 - \beta_{2,0}\|_2^2 \log^2 n)$.

7. Error bounds for $\hat{\beta}_1$ and the value function of the estimated dynamic treatment regime.

7.1. *Misspecified contrast function.* In the context of A-learning, a major challenge arising in multi-stage studies is that the contrast functions are likely to be misspecified in backward induction. In order to study the finite sample bounds of $\hat{\beta}_1$, we need to first define least favorable parameters under the misspecification of the contrast function.

Recall that $C^{(1)}(S_i)$ is the true contrast function for the ith patient, which can be a very complex function of S_i due to the backward induction. For notational convenience, we use a shorthand $C(s)$ for $C^{(1)}(s)$. We posit a linear model $S_i^T\beta_1$ for $C(\cdot)$, which is often misspecified. When either the propensity score model $\pi^{(1)}$ or the baseline mean function $h^{(1)}$ is correctly specified, the associated least favorable parameters $\beta_1^\star$ is defined as follows:

$$\beta_1^\star = \arg\min_{\beta_1\in\Lambda^*} \|\beta_1\|_1, \tag{7.1}$$

where

$$\Lambda^* = \{\beta_1 \in \mathbb{R}^q : \|\mathrm{E}[S_i A_i^{(1)}(1-\pi_i^{(1)*})\{C(S_i) - S_i^T\beta_1\}]\|_\infty \le \kappa_0\},$$

$\pi_i^{(1)*} = \pi^{(1)}(S_i, \alpha_1^\star)$ and κ_0 is a nonnegative constant. Define

$$\kappa_0^* = \|\mathrm{E}[S_i A_i^{(1)}(1-\pi_i^{(1)*})\{C(S_i) - S_i^T\beta_1^\star\}]\|_\infty.$$

By simple algebra, we can show $\kappa_0^* \le \min\{\kappa_0, O(\sigma_0)\}$, where $\sigma_0^2 = \mathrm{E}[\{C(S_i) - S_i^T\beta_1^\star\}^2]$, describing the degree of misspecification of the contrast function. Define $s_{\beta_1} = |M_{\beta_1}| = O(n^{l_3})$ for some $0 \le l_3 < 1/2$, where $M_{\beta_1} = \mathrm{supp}(\beta_1^\star)$.

7.2. *Error bound for* $\hat{\beta}_1$. Assume that $\log q = O(n^{a_1})$ for some $0 < a_1 < 1$ and $S_1, \ldots, S_n$ are i.i.d. copies of S_0 that

$$S_0 \overset{d}{=} \Psi^{1/2} V_0, \tag{7.2}$$

where $\Psi \in \mathbb{R}^{q\times q}$ is some positive definite matrix with $\Psi_{jj} = 1$ for $j = 1, \ldots, q$, and V_0 is a q-dimensional isotropic random vector with some isotropic constants ζ.

As in the second stage, we first give conditions on $\hat{\alpha}_1$ and $\hat{\theta}_1$. Assume that these two estimators converge to some $\alpha_1^\star$ and $\theta_1^\star$, respectively, under possible model misspecification. Denote $M_{\alpha_1} = \mathrm{supp}(\alpha_1^\star)$, $M_{\theta_1} = \mathrm{supp}(\theta_1^\star)$, $s_{\alpha_1} = |M_{\alpha_1}| = O(n^{l_1})$, and $s_{\theta_1} = |M_{\theta_1}| = O(n^{l_2})$ for some $0 \le l_1, l_2 < 1/2$.

CONDITION 5. Assume that there exist some positive constants γ_{α_1} and γ_{θ_1}, with probability at least $1 - \bar{c}/(n+p+q)$, the following hold:

$$\hat{\alpha}_1^{M_{\alpha_1}^c} = 0, \qquad \|\hat{\alpha}_1^{M_{\alpha_1}} - \alpha_1^{\star M_{\alpha_1}}\|_\infty = O(n^{-\gamma_{\alpha_1}}\log n), \tag{7.3}$$

$$\hat{\theta}_1^{M_{\theta_1}^c} = 0, \qquad \|\hat{\theta}_1^{M_{\theta_1}} - \theta_1^{\star M_{\theta_1}}\|_\infty = O(n^{-\gamma_{\theta_1}}\log n). \tag{7.4}$$

Moreover, assume $d_{\alpha_1} \gg n^{-\gamma_{\alpha_1}} \log n$ and $d_{\theta_1} \gg n^{-\gamma_{\theta_1}} \log n$, where $d_{\alpha_1} = \min_j |\alpha_1^{\star j}|/2$ and $d_{\theta_1} = \min_j |\theta_1^{\star j}|/2$.

CONDITION 6. Assume that $D^{(1)}$, $C^{(1)}$ and Ψ satisfy

$$\lambda_{\max}\big(\Psi_{M_{\alpha_1}}^{M_{\alpha_1}}\big) = O(1), \qquad \lambda_{\max}\big(\Psi_{M_{\theta_1}}^{M_{\theta_1}}\big) = O(1),$$

$$\liminf_n \lambda_{\min}\big(D_{M_{\theta_1}}^{(1)M_{\theta_1}}\big) > 0, \qquad \liminf_n \lambda_{\min}\big(C_{M_{\alpha_1}}^{(1)M_{\alpha_1}}\big) > 0,$$

where

$$D^{(1)} = \mathrm{E}\big\{S_i S_i^T\big(1 - A_i^{(1)}\big)\big\}, \qquad C^{(1)} = \mathrm{E}\big\{S_i S_i^T \pi_i^{(1)*}\big(1 - \pi_i^{(1)*}\big)\big\},$$

and $\pi^{(1)*} = \pi^{(1)}(S_i, \alpha_1^\star)$.

Since both the propensity score model and the contrast function at the first stage can be misspecified, we need the following condition to control their effect on estimation of $\beta_1^\star$.

CONDITION 7. Assume that

$$\tau_0 \equiv \big\| F^{M_{\alpha_1}}\big[C^{(1)} M_{\alpha_1 M_{\alpha_1}}\big]^{-1} b^{(1)M_{\alpha_1}} \big\|_\infty < \infty, \tag{7.5}$$

where $b^{(1)} = \mathrm{E}\{S_i(A_i^{(1)} - \pi_i^{(1)*})\}$ and

$$F = \mathrm{E}\big[S_i A_i^{(1)} \pi_i^{(1)*}\big(1 - \pi_i^{(1)*}\big)\big\{C(S_i) - S_i \beta_1^\star\big\} S_i^T\big].$$

REMARK 7.1. It is immediate to see $\tau_0 = 0$ when either the contrast function or the propensity score model is correctly specified.

When going back to the first stage, the error bound of $\hat{\beta}_1$ is directly affected by that of $\hat{\beta}_2$. This is because the estimated response $\hat{V}_i$ in the first stage is obtained based on $\hat{\beta}_2$ using the advantage function. To simplify presentation, we introduce the following condition.

CONDITION 8. Assume that with probability at least $1 - \bar{c}/(n + p)$, there exists some constant $\mu_1 > 0$ such that

$$\sqrt{\rho_{\max}^{s_{\beta_2}}(\Sigma)}\|\hat{\beta}_2 - \beta_{2,0}\|_2 = O\big(n^{-\mu_1} \log n\big), \tag{7.6}$$

and $\|\hat{\beta}_2^{M_{\beta_2}^c}\|_1 \le \|\hat{\beta}_2^{M_{\beta_2}} - \beta_{2,0}^{M_{\beta_2}}\|_1$.

A more explicit form of the error bound for (7.6) is given in Theorem 6.1. In the next lemma, we provide an upper bound for the term

$$\big\| S^T \operatorname{diag}\big(A^{(1)} - \hat{\pi}^{(1)}\big)\big(\hat{V} - S\hat{\theta}_1 - A^{(1)} \circ S\beta_1^\star\big)\big\|_\infty / n. \tag{7.7}$$

LEMMA 7.1. *Assume that Conditions* 5–8 *and those in Theorem* 6.1 *hold*, $\|C(S_i) - S_i^T\beta_1^\star\|_{\psi_1} < \infty$, $\|V_i - \mathrm{E}(V_i|S_i, A_i^{(1)})\|_{\psi_2} < \infty$, $a_1 + l_1 < 1$, $n \gg s_{\beta_2}\log p\rho_{\max}^{s_{\beta_2}}(\Sigma)^2/\rho_{\min}^{s_{\beta_2}}(\Sigma)$, *and either* $\pi^{(1)}$ *or* $h^{(1)}$ *is correctly specified. Then, for sufficiently large* n, *with probability at least* $1 - \bar{c}/(n+p+q)$, (7.7) *can be bounded from above by* $c^{(1)}(E_5 + E_6 + E_7 + E_8 + E_9 + E_{10})$ *for some constant* $c^{(1)} > 0$, *where*

$$E_5 = \sqrt{\log q/n}\log^2 n, \qquad E_6 = s_{\alpha_1}n^{-2\gamma_{\alpha_1}}\log^2 n + s_{\theta_1}n^{-2\gamma_{\theta_1}}\log^2 n,$$

$$E_7 = \sigma_1\{\sqrt{s_{\alpha_1}\log n/n} + \sqrt{s_{\alpha_1}}\lambda_{1n}^{(1)}\rho_1^{(1)}(d_{n\alpha_1})\},$$

$$E_8 = \sigma_2\{\sqrt{s_{\theta_1}\log n/n} + \sqrt{s_{\theta_1}}\lambda_{2n}^{(1)}\rho_2^{(1)}(d_{n\theta_1})\},$$

$$E_9 = \sigma_0\{\sqrt{s_{\alpha_1}\log n/n} + \sqrt{s_{\alpha_1}}\lambda_{1n}^{(1)}\rho_1^{(1)}(d_{n\alpha_1}) + \tau_0 + \kappa_0^*\},$$

$E_{10} = n^{-\mu_1}\log n$, $\sigma_0^2 = \mathrm{E}\{C(S_i) - S_i^T\beta_1^\star\}^2$, $\sigma_1^2 = \mathrm{E}(h^{(1)} - S_i^T\theta_1^\star)^2$, *and* $\sigma_2^2 = \mathrm{E}\{\pi_i^{(1)*} - \pi^{(1)}(S_i)\}^2$.

REMARK 7.2. The terms $E_5 - E_8$ have similar interpretations as $E_1 - E_4$ in Lemma 6.1, respectively. The additional term E_{10} is due to the error bound of $\hat{\beta}_2$ in the backward induction, while E_9 is due to the misspecification of the contrast function.

Define $\Omega^{(1)}(\alpha_1) = \mathrm{E}[S_iS_i^TA_i^{(1)}\{1 - \pi^{(1)}(S_i, \alpha_1)\}]$ and $\Omega_n^{(1)} = n^{-1}\sum_i S_iS_i^T \times A_i^{(1)}(1 - \hat{\pi}_i^{(1)})$ with $\hat{\pi}_i^{(1)} = \pi^{(1)}(X_i, \hat{\alpha}_1)$. Similar as in stage 2, we need the following condition to ensure the RE condition for the matrix $\Omega_n^{(1)}$.

CONDITION 9. Assume that for any $0 < \theta_s < 1$ and sufficiently large n, we have

$$K\big(s_{\beta_1}, 1, \Omega_n^{(1)}\big) > (1 - \theta_s)\inf_{\alpha_1 \in H_{\alpha_1}} K\big(s_{\beta_1}, 1, \Omega^{(1)}(\alpha_1)\big) > 0, \tag{7.8}$$

where H_{α_1} denotes the set of vectors α_1 that satisfies the weak oracle property (7.3).

THEOREM 7.1. *Assume that Condition* 9 *and those conditions in Lemma* 7.1 *hold, and* $\lambda_{3n}^{(1)} \geq c^{(1)}\sum_{k=5}^{10}E_k$. *The constant* $c^{(1)}$ *is defined in Lemma* 7.1. *Then there exists a constant* c_8, *such that for sufficiently large* n *and some fixed* $0 < \theta_s < 1$, *with probability at least* $1 - c_8/(n+p+q)$, *the error bounds for* $\hat{\beta}_1$ *are given by*

$$\|\hat{\beta}_1 - \beta_1^\star\|_2 \leq \frac{12\lambda_{3n}^{(1)}\sqrt{s_{\beta_1}}}{(1-\theta_s)^2\inf_{\alpha_1 \in H_{\alpha_1}} K^2(s_{\beta_1}, 1, \Omega^{(1)}(\alpha_1))}, \tag{7.9}$$

$$\|\hat{\beta}_1 - \beta_1^\star\|_1 \le \frac{8\lambda_{3n}^{(1)} s_{\beta_1}}{(1-\theta_s)^2 \inf_{\alpha_1 \in H_{\alpha_1}} K^2(s_{\beta_1}, 1, \Omega^{(1)}(\alpha_1))}. \tag{7.10}$$

7.3. *Error bound for the value function of the estimated dynamic treatment regime*. Under the SUTVA and sequential randomization assumptions, the value function of a given dynamic treatment regime $(d_1(S_0), d_2(X_0))$ is given by

$$\mathrm{E}\{Y_0^\star(d_1, d_2)\} = E\big[h^{(2)}(X_0) + (\beta_{2,0}^T X_0) d_2(X_0) + C(S_0)\{d_1(S_0) - A_0^{(1)}\}\big],$$

where S_0 and X_0 denote the baseline covariates and covariates for the second stage, respectively. Then the difference of the value functions under the estimated optimal dynamic treatment regime (2.4) and the true optimal regime (d_1^{opt}, d_2^{opt}) is given by

$$\begin{aligned}
&\mathrm{E}\{Y_0^\star(d_1^{opt}, d_2^{opt})\} - \mathrm{E}\{Y_0^\star(\hat{d}_1, \hat{d}_2)\} \\
&\quad = \mathrm{E}\big[C(S_0)\{I(C(S_0) > 0) - I(S_0^T \hat{\beta}_1 > 0)\}\big] \\
&\qquad + \mathrm{E}\big[X_0^T \beta_{2,0}\{I(X_0^T \beta_{2,0} > 0) - I(X_0^T \hat{\beta}_2 > 0)\}\big].
\end{aligned}$$

Similar to Condition 4, we impose the following condition.

CONDITION 10. Assume that the probability density function $\mathrm{g}^{(1)}(\cdot)$ of $S_0^T \beta_1^\star$ exists and is bounded.

THEOREM 7.2. *Assume that conditions in Theorem 7.1 and Condition 10 hold. Assume* $\mathrm{E}(X_0^T \beta_{2,0})^2 = O(1)$, $\mathrm{E}(S_0^T \beta_1^\star)^2 = O(1)$. *Then, for some fixed* $0 < \theta_s < 1$ *and sufficiently large* n,

$$\begin{aligned}
0 &\le \mathrm{E}\{Y_0^\star(d_1^{opt}, d_2^{opt})\} - \mathrm{E}\{Y_0^\star(\hat{d}_1, \hat{d}_2)\} \\
&\le \frac{\bar{c}(\omega + \zeta)}{n} + c_0 \sigma_0^{4/3} + \frac{c_0 \omega^2 \rho_{\max}^{s_{\beta_2}}(\Sigma) {\lambda_{3n}^{(2)}}^2 s_{\beta_2} \log^2 n}{(1-\theta_s)^4 \inf_{\alpha_2 \in H_{\alpha_2}} K^4(s_{\beta_2}, 1, \Omega^{(2)}(\alpha_2))} \\
&\quad + \frac{c_0 \zeta^2 \rho_{\max}^{s_{\beta_1}}(\Psi) {\lambda_{3n}^{(1)}}^2 s_{\beta_1} \log^2 n}{(1-\theta_s)^4 \inf_{\alpha_1 \in H_{\alpha_1}} K^4(s_{\beta_1}, 1, \Omega^{(1)}(\alpha_1))}.
\end{aligned} \tag{7.11}$$

REMARK 7.3. Theorem 7.2 suggests that the upper bound for the difference of the value functions come from three major components: the misspecification of the contrast function, described by σ_0^2, and estimation errors of $\hat{\beta}_2$ and $\hat{\beta}_1$.

8. Weak oracle properties of $\hat{\alpha}_j$'s and $\hat{\theta}_j$'s. In order to prove the error bounds of $\hat{\beta}_1$, $\hat{\beta}_2$ and the value functions of the estimated treatment regimes presented in Sections 6 and 7, we need to establish the weak oracle properties of $\hat{\alpha}_j$ and $\hat{\theta}_j$ $(j = 1, 2)$ in the posited models for the propensity score and baseline mean

functions. Here, we prove the results based on a posited logistic regression model for the propensity score and a linear model for the baseline mean function under a random design setting. However, these results can be extended to generalized linear models [McCullagh and Nelder (1989)].

8.1. *Weak oracle properties of $\hat{\alpha}_2$ and $\hat{\theta}_2$.* We assume that $\hat{\alpha}_2$ and $\hat{\theta}_2$ converge to some population parameters $\alpha_2^\star$ and $\theta_2^\star$, respectively. Under Conditions B1–B6 given in the Supplementary Material [Shi et al. (2018)], we establish the weak oracle properties of $\hat{\alpha}_2$ and $\hat{\theta}_2$ in the following two theorems. Recall that $s_{\alpha_2} = |M_{\alpha_2}| = O(n^{l_4})$ for some $0 \le l_4 < 1/2$.

THEOREM 8.1. *Assume that Conditions* B.1–B.3 *hold*, $l_4 + a_2 < 1$ *and* $\lambda_{\max}(\Sigma_{M_{\alpha_2}}^{M_{\alpha_2}}) = O(1)$. *Then, for sufficiently large n, there exist some constants $\gamma_{\alpha_2} > 0$, such that with probability at least $1 - \bar{c}/(n+p)$:*

a. $\hat{\alpha}_2^{M_{\alpha_2}^c} = 0$.
b. $\|\hat{\alpha}_2^{M_{\alpha_2}} - \alpha_2^{\star M_{\alpha_2}}\|_\infty = O(n^{-\gamma_{\alpha_2}} \log n)$.

THEOREM 8.2. *Assume that Conditions* B.4–B.6 *hold*, $\lambda_{\max}(\Sigma_{M_{\theta_2}}^{M_{\theta_2}}) = O(1)$, *and* $\|e_i\|_{\psi_2} < \infty$, *where e_i is defined in* (2.2). *Then there exist some constants $\gamma_{\theta_2} > 0$, such that with probability at least $1 - \bar{c}/(n+p)$:*

a. $\hat{\theta}_2^{M_{\theta_2}^c} = 0$.
b. $\|\hat{\theta}_2^{M_{\theta_2}} - \theta_2^{\star M_{\theta_2}}\|_\infty = O(n^{-\gamma_{\theta_2}} \log n)$.

REMARK 8.1. Theorem 1 in Shi, Song and Lu (2016) established weak oracle results of the penalized estimators for a fixed design setting. This is mainly for technical convenience. Its proofs can be obtained using similar arguments as in Fan and Lv (2011). In this paper, we focus on a random design setting, which is more practical in medical studies. To the best of our knowledge, the weak oracle properties of penalized estimators have not been studied in a random design setting with the NP dimensionality. The major difficulty lies in developing some random matrix theories, such as controlling the maximum eigenvalue of some random matrices. Such results are established in Theorems 8.1 and 8.2.

REMARK 8.2. The condition $l_4 + a_2 < 1$ ensures that for large n

$$\max_{j=1}^{p} \lambda_{\max}\big[\big(X^{M_{\alpha_2}}\big)^T \operatorname{diag}\big(|X^j|\big) X^{M_{\alpha_2}}\big] = O(n), \tag{8.1}$$

with probability approaching 1. A major technical difficulty in deriving (8.1) is that the matrix $(X^{M_{\alpha_2}})^T \operatorname{diag}(|X^j|) X^{M_{\alpha_2}}$ does not have the subexponential tail (see

Definition G.2 in the Supplementary Material [Shi et al. (2018)]). When $s_{\alpha_2} \leq n$, we can bound $\max_{j \in M_{\alpha_2}} |X_i^j|$ from above by $\sqrt{2}\omega \log n$ with probability at least $1 - 2/n$, which ensures the subexponential tail of the truncated matrix. Lemma B.2 in the Supplementary Material proves such a result for a more general case.

8.2. *Weak oracle properties of $\hat{\alpha}_1$ and $\hat{\theta}_1$.* The weak oracle properties of $\hat{\alpha}_1$ can be similarly derived as for $\hat{\alpha}_2$. However, unlike the results for $\hat{\theta}_2$, the weak oracle properties of $\hat{\theta}_1$ depend on $\hat{\beta}_2$ even when the baseline mean function $h^{(1)}$ is correctly specified. This is because the estimated response $\hat{V}_i$ is obtained based on $\hat{\beta}_2$. A necessary condition to ensure $\|\hat{\theta}_1 - \theta_1^\star\|_\infty \xrightarrow{P} 0$ is that $\|\hat{\beta}_2 - \beta_{2,0}\|_2 \xrightarrow{P} 0$, which is established in Corollary 6.1.

THEOREM 8.3. *Assume that Condition 8 and Conditions* B.7–B.12 *in the Supplementary Material hold. Further assume that* $\lambda_{\max}(\Psi_{M_{\alpha_1}}^{M_{\alpha_1}}) = O(1)$, $\lambda_{\max}(\Psi_{M_{\theta_1}}^{M_{\theta_1}}) = O(1)$, $n \gg s_{\beta_2} \log p \{\rho_{\max}^{s_{\beta_2}}(\Sigma)\}^2 / \rho_{\min}^{s_{\beta_2}}(\Sigma)$, $a_1 + l_1 < 1$, $\|e_i\|_{\psi_2} < \infty$ *and* $\|V_i - \mathrm{E}(V_i | S_i^{(1)}, A_i^{(1)})\|_{\psi_2} < \infty$. *Then, for sufficiently large n, there exist some constants $\gamma_{\alpha_1} > 0$ and $\gamma_{\theta_1} > 0$, with probability at least $1 - \bar{c}/(n + q + p)$, such that the estimators $\hat{\alpha}_1$ and $\hat{\theta}_1$ must satisfy*:

a. $\hat{\alpha}_1^{M_{\alpha_1}^c} = 0$, $\hat{\theta}_1^{M_{\theta_1}^c} = 0$,
b. $\|\hat{\alpha}_1^{M_{\alpha_1}} - \alpha_1^{\star M_{\alpha_1}}\|_\infty = O(n^{-\gamma_{\alpha_1}} \log n)$, $\|\hat{\theta}_1^{M_{\theta_1}} - \theta_1^{\star M_{\theta_1}}\|_\infty = O(n^{-\gamma_{\theta_1}} \log n)$.

9. Uniform uncertainty principle and restricted eigenvalue conditions in *A*-learning. In this section, we establish the UUP and RE conditions in the context of A-learning. In our setting, these two conditions are needed on random matrices $\Omega_n^{(2)}$ and $\Omega_n^{(1)}$.

For brevity, we only study the UUP and RE conditions for the random matrix $\Omega_n^{(2)}$. Those for $\Omega_n^{(1)}$ can be similarly derived. Recall that M_{α_2} refers to the support of $\alpha_2^\star$, $M_{\beta_2} = \mathrm{supp}(\beta_{2,0})$, and $s_{\beta_2} = |M_{\beta_2}|$. We assume that the weak oracle properties of $\hat{\alpha}_2$ are achieved such that with probability at least $1 - \bar{c}/(n + p)$:

$$\hat{\alpha}_2^{M_{\alpha_2}^c} = 0 \quad \text{and} \quad \|\hat{\alpha}_2 - \alpha_2^\star\|_\infty = O\big(n^{-\gamma_{\alpha_2}} \log n\big), \tag{9.1}$$

for some $\gamma_{\alpha_2} > 0$. The following lemma establishes the UUP condition for $\Omega_n^{(2)}$.

LEMMA 9.1. *Assume the convergence rate of $\hat{\alpha}_2$ satisfies*

$$\|\hat{\alpha}_2 - \alpha_2^\star\|_2 = O\big(\sqrt{s_{\alpha_2}} n^{-\gamma_{\alpha_2}} \log n\big) = O(1),$$

and the sample size satisfies

$$n \gg \frac{\{\rho_{\max}^{s_{\beta_2}}(\Sigma)\}^2 (s_{\beta_2} \log p + s_{\alpha_2}^2)}{\inf_{\alpha_2 \in H_{\alpha_2}} \rho_{\min}^{s_{\beta_2}}(\Omega^{(2)}(\alpha_2))}. \tag{9.2}$$

Then for any $0 < \theta < 1$, *with probability at least* $1 - \bar{c}/(n+p)$, *we have*

$$\begin{aligned}(9.3)\qquad &\left\| \frac{1}{n} y^T \Omega_n^{(2)} y - y^T \Omega^{(2)} y \right\|_2 \\ &\quad \le \left\{ \theta + \frac{4\omega^2}{n} + \sqrt{2}\omega^2 \|\hat{\alpha}_2 - \alpha_2^\star\|_2 \sqrt{\lambda_{\max}\big(\Sigma_{M_{\alpha_2}}^{M_{\alpha_2}}\big)} \right\} \rho_{\max}^{s_{\beta_2}}(\Sigma) \|y\|_2^2,\end{aligned}$$

for any $y \in \mathbb{R}^p$ *and* $|\operatorname{supp}(y)| \le s_{\beta_2}$.

REMARK 9.1. In our setting, if the following regularity conditions hold,

$$\liminf_{\alpha_2 \in H_{\alpha_2}} \rho_{\min}^{s_{\alpha_2}}\big(\Omega^{(2)}(\alpha_2)\big) > 0 \quad \text{and} \quad \rho_{\max}^{s_{\beta_2}}(\Sigma) = O(1),$$

the requirement on the sample size (9.2) reduces to $n \gg s_{\beta_2} \log p$ since $s_{\alpha_2}^2 = O(n^{2l_4}) \ll n$.

REMARK 9.2. The second term on the right-hand side of (9.3) represents the difference between $y^T \tilde{\Omega}_n^{(2)} y$ and $y^T \Omega_n^{(2)} y$, where $\tilde{\Omega}_n^{(2)}$ is defined as the expectation of the truncated random matrix

$$(9.4)\qquad \frac{1}{n} \sum_i X_i X_i^T A_i^{(2)} \big\{1 - \pi^{(2)}(X_i, \hat{\alpha}_2)\big\} I\big(\|X_i^{M_{\alpha_2}}\|_\infty \le \sqrt{2}\omega \log n\big).$$

This term will vanish as $n \to \infty$. The third term represents the estimation error of $\hat{\alpha}_2$. When $\rho_{\max}^{s_{\beta_2}}(\Sigma) < 2$ and $\sqrt{\lambda_{\max}(\Sigma_{M_{\alpha_2}}^{M_{\alpha_2}})}\|\hat{\alpha}_2 - \alpha_2^\star\|_2 \to 0$, (9.3) proves the UUP condition for $\Omega_n^{(2)}$.

REMARK 9.3. A key assumptions in Lemma 9.1 is the sparsity of $\alpha_2^\star$, which is needed to bound the infinity norm in the indicator function of (9.4). This extra requirement comes from the involvement of the estimated propensity scores in $\Omega_n^{(2)}$, which adds significant difficulties in proving Lemma 9.1.

After some algebra, the RE condition for $\Omega_n^{(2)}$ follows similarly from Lemma 9.1, which is presented below.

LEMMA 9.2. *For any integer* c_0, *assume that* $\|\hat{\alpha}_2 - \alpha_2^\star\|_2 = O(1)$, *and the sample size satisfies*

$$(9.5)\qquad n \gg \frac{\{\rho_{\max}^{s_{\beta_2}}(\Sigma)\}^2 (s_{\beta_2} \log p + s_{\alpha_2}^2)}{\inf_{\alpha_2 \in H_{\alpha_2}} K^2(s_\beta, c_0, \Omega^{(2)}(\alpha_2))}.$$

Then, for any $0 < \theta < 1$ *and sufficiently large* n, *with probability at least* $1 - \bar{c}/(n+p)$, *we have*

$$K\big(s_{\beta_2}, c_0, \Omega_n^{(2)}\big) > (1-\theta) \inf_{\alpha \in H_{\alpha_2}} K\big(s_{\beta_2}, c_0, \Omega^{(2)}(\alpha_2)\big).$$

REMARK 9.4. The sample size requirement (9.5) is stronger than (9.2). To see this, for any positive semidefinite matrix Ψ, and positive integers s and c_0, we have

$$K^2(s, c_0, \Psi) \leq K^2(s, 0, \Psi) = \rho^s_{\min}(\Psi).$$

10. Discussion.

10.1. *Post selection inference*. As pointed by one of the referees, the main goal of constructing optimal DTRs is to find treatments that are significantly superior to other treatment options. This requires addressing a post selection inference issue, that is, the problem of influencing the estimated optimal value function (or the difference between the estimated value and the value function under other treatment options). In the fixed dimension setting, we can use either the empirical average of the advantage function [Murphy (2003)] or the augmented inverse propensity score type estimates [AIPWE, Zhang et al. (2012)] to estimate the optimal value function. Both types of estimators are asymptotically normally distributed. However, the inference based on the advantage function may not be valid in high dimensions. This is because when the number of predictors is large; the parameter estimates in the contrast function may not have oracle property (i.e., model selection consistency and asymptotic normality).

For a single-stage study, assuming a linear interaction form $X^T\beta_0$ for the contrast function. Under certain conditions, we can show AIPWE is asymptotically normal even for NP-dimensionality if (i) $\|\hat{\beta} - \beta_0\|_2 = o_p(n^{-1/4})$, (ii) with probability going to 1, $\|\hat{\beta}^{M^c_\beta} - \beta_0^{M^c_\beta}\|_1 \leq c_0\|\hat{\beta}^{M_\beta} - \beta_0^{M_\beta}\|_1$ for some constant c_0, where M_β is the support of β_0. For our penalized A-learning estimator, Assumption (i) can be achieved assuming certain conditions on the dimension of covariates, sample size and the sparsity of parameters in the contrast, the baseline and propensity score function. Assumption (ii) is typically satisfied for Lasso, Dantzig and folded-concave type estimators. Similar to Theorem 6.1, we can show our estimator satisfies $\|\hat{\beta}^{M^c_\beta} - \beta_0^{M^c_\beta}\|_1 \leq \|\hat{\beta}^{M_\beta} - \beta_0^{M_\beta}\|\hat{\beta}^{M_\beta} - \beta_0^{M_\beta}\|_1$ with probability going to 1. The asymptotic normality of AIPWE therefore follows. Standard error of the value estimator can be similarly obtained as in Zhang et al. (2012). Alternatively, we can use the one step online estimator as in Luedtke and van der Laan (2016). However, the asymptotic variance will be larger since it does not use all the data to construct the estimator. In summary, it is important and interesting to develop statistical inference for the estimated value function under the obtained optimal treatment regime in high dimensions, but it is beyond the scope of the current paper.

10.2. *Tuning parameter selection*. Bayesian information criteria (BIC) is used to tune the penalty functions. BIC has been widely used in model selection for

selecting the tuning parameter when the goal is prediction. In high dimensional regressions, Chen and Chen (2008) proposed an extended BIC for model selection, and showed their BIC is consistent when the number of predictors grows polynomially in sample size. Fan and Tang (2013) proposed a similar criterion and showed its consistency when the number of predictors is in the nonpolynomial order of the sample size. When the goal is to select treatment effect modifiers, Lu et al. (2013) also used a BIC-type criterion, which showed good empirical performance. This motivated us to use a similar BIC-type criterion for selecting the tuning parameter in our method. Our simulations demonstrated that the proposed BIC-type criterion empirically worked well. We conjecture that following similar arguments in the proof of Theorem 1 of Chen and Chen (2008) and the proof of Theorem 3 in Fan and Tang (2013), we can show our proposed BIC-type criterion is also consistent for selecting important variables in the contrast function. This is another interesting topic that needs further investigation.

10.3. *Extensions to multiple stages and general models*. In this paper, we mainly focus on a two-stage study. Extension of results to three-stage studies are provided in the Supplementary Material [Shi et al. (2018)]. It raises additional challenges to establish these results, since the potential model misspecification of contrast functions in the previous two stages can add up and stronger assumptions are needed to guarantee consistency of the parameter estimates. Readers can refer to the Supplementary Material for details.

For technical convenience, we assume a linear interaction form the contrast function on the last stage. More general results when the contrast function is misspecified can be similarly derived as the three-stage studies discussed in the supplementary article.

Acknowledgments. We thank the Editor, Associate Editor and two referees for providing helpful suggestions that significantly improved the quality of the paper. The STAR*D data are provided by National Institute of Mental Health.

SUPPLEMENTARY MATERIAL

Supplement to "High-dimensional *A*-learning for optimal dynamic treatment regimes" (DOI: 10.1214/17-AOS1570SUPP; .pdf). Supplementary material includes some proofs.

REFERENCES

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

CANDÈS, E. and TAO, T. (2007). Rejoinder: "The Dantzig selector: Statistical estimation when p is much larger than n" [Ann. Statist. **35** (2007), 2313–2351; MR2382644]. *Ann. Statist.* **35** 2392–2404. MR2382651

CHAKRABORTY, B., MURPHY, S. and STRECHER, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.* **19** 317–343. MR2757118

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. MR2849368

FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552. MR3065478

FAVA, M., RUSH, A. J., TRIVEDI, M. H., NIERENBERG, A. A., THASE, M. E., SACKEIM, H. A., QUITKIN, F. M., WISNIEWSKI, S., LAVORI, P. W., ROSENBAUM, J. F. et al. (2003). Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study. *Psychiatr. Clin. North Am.* **26** 457–494.

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

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615

LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models. Monographs on Statistics and Applied Probability*, 2nd ed. Chapman & Hall, London. MR3223057

MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.* **17** 1248–1282. MR2373017

MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2008). Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.* **28** 277–289. MR2453368

MILMAN, V. D. and PAJOR, A. (1989). Isotropic position and inertia ellipsoids and zonoids of the unit ball of a normed n-dimensional space. In *Geometric Aspects of Functional Analysis* (1987–1988). *Lecture Notes in Math.* **1376** 64–104. Springer, Berlin. MR1008717

MILMAN, V. D. and PAJOR, A. (2003). Regularization of star bodies by random hyperplane cut off. *Studia Math.* **159** 247–261. MR2052221

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–366. MR1983752

QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiol.* **11** 550–560.

RUSH, A. J., FAVA, M., WISNIEWSKI, S. R., LAVORI, P. W., TRIVEDI, M. H., SACKEIM, H. A., THASE, M. E., NIERENBERG, A. A., QUITKIN, F. M., KASHNER, T. M. et al. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): Rationale and design. *Control. Clin. Trials* **25** 119–142.

SHI, C., SONG, R. and LU, W. (2016). Robust learning for optimal treatment decision with NP-dimensionality. *Electron. J. Stat.* **10** 2894–2921. MR3557316

SHI, C., FAN, A., SONG, R. and LU, W. (2018). Supplement to "High-dimensional A-learning for optimal dynamic treatment regimes." DOI:10.1214/17-AOS1570SUPP.

TIBSHIRANI, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 273–282. MR2815776

WATKINS, C. J. C. H. and DAYAN, P. (1992). Q-Learning. *Mach. Learn.* **8** 279–292.

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. MR3040007

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

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

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

ZHOU, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. Available at arxiv:0912.4045.

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

DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NORTH CAROLINA 27695
USA
E-MAIL: rsong@ncsu.edu