Statistica Sinica Preprint No: SS-2018-0076

Title	Entropy Learning for Dynamic Treatment Regimes
Manuscript ID	SS-2018-0076
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0076
Complete List of Authors	Binyan Jiang Rui Song Jialiang Li and Donglin Zeng
Corresponding Author	Jialiang Li
E-mail	stalj@nus.edu.sg

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

Statistica Sinica

Entropy Learning for Dynamic Treatment Regimes

Binyan Jiang[1], Rui Song[2], Jialiang Li[3] and Donglin Zeng[4]

The Hong Kong Polytechnic University[1], *North Carolina State University*[2]
National University of Singapore[3] *and University of North Carolina*[4]

Abstract: Estimating optimal individualized treatment rules (ITRs) in single- or multi-stage clinical trials is a key element of personalized medicine and, as a result, is receiving increasing attention within the statistical community. Recent works have suggested that machine learning approaches can provide significantly better estimations than those of model-based methods. However, a proper inference for estimated ITRs has not been well established for machine learning-based approaches. In this paper, we propose an entropy learning approach for estimating optimal ITRs. We obtain the asymptotic distributions for the estimated rules in order to provide a valid inference. The proposed approach is demonstrated to perform well through extensive simulation studies. Finally, we analyze data from a multi-stage clinical trial for depression patients. Our results offer novel findings not revealed by existing approaches.

Key words and phrases: Dynamic treatment regime, entropy learning, personalized medicine.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

1. Introduction

An important goal of personalized medicine is to develop a decision support system to provide adequate management for individual patients with specific diseases. Estimating individualized treatment rules (ITRs) using evidence from single- or multi-stage clinical trials is a key element of such a system. As a result, estimation methods are receiving increasing attention within the statistical community. The methods for estimating ITRs include Q-learning (Watkins and Dayan, 1992; Murphy, 2005; Chakraborty et al., 2010; Goldberg and Kosorok, 2012; Laber et al., 2014; Song et al., 2015) and A-learning (Robins et al., 2000; Murphy, 2003). Q-learning models the conditional mean of the outcome, given historical covariates and treatments using a well-constructed statistical model. A-learning models the contrast function that is sufficient for a treatment decision.

Recently, Zhao et al. (2012) discovered that it is possible to cast the estimation of the optimal regime into a weighted classification problem. Based on this, Zhao et al. (2012, 2015) proposed an outcome-weighted learning (OWL) directly optimizes the approximate expected clinical outcome, where the objective function is a hinge loss, weighted by individual outcomes. This method has been shown to outperform the model-based approaches, such as Q- and A-learning, in numerical studies, and the asymp-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

totic behavior might be established, owing to its convexity Hjort and Pollard (2011). However there is no valid inference procedure for the parameters in the optimal treatment rules, owing to the nondifferentiability of the hinge loss near the decision boundary. Furthermore, the minimization operator is more or less heuristic.

In this paper, we propose a class of smooth-loss-based outcome-weighted learning methods for estimating optimal ITRs, among which, one special case of the proposed losses is a weighed entropy loss (Murphy, 2012). By using continuously differentiable loss functions, we not only maintain the Fisher consistency of the derived treatment rule, but also obtain a proper inference for the parameters in the derived rule. Furthermore, we quantify the uncertainty of the value function under the estimated treatment rule, which is potentially useful for designing future trials and comparing the results with those of other, nonoptimal treatment rules. Numerically, in contrast to existing inferences for the model-based approaches, such as the bootstrap approach for Q-learning, our inference procedure does not require tuning parameters. In addition, the proposed method yields a more accurate inference in finite-sample numerical studies.

Note that Bartlett et al. (2006) produced a profound conceptual work on classification loss, for a relatively general setting. However, to link their

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

work to recursive or dynamic optimization is not trivial. To do so, we employ a logistic loss. Luedtke and van der Laan (2016) tried to create a unified surrogate loss function for outcome-dependent learning. Their method of showing the validity of their approach differs from our derivation. Our justification is more intuitive and our algorithm is also different. Whereas super learning is a general and powerful method, a logistic regression can be implemented easily and fits our needs directly. Moreover, the asymptotic properties of our estimators are established in order to conduct a proper inference, which is not addressed in the above-mentioned studies.

The paper is structured as follows. In Section 2, we introduce the proposed entropy learning method for single- and multi-stage settings. In Section 3, we provide the asymptotic properties of our estimators. In Section 4, simulation studies are conducted to assess the performance of our methods. In Section 5, we apply entropy learning to the well-known STAR*D study. We conclude the paper in Section 6. Technical proofs are provided in the Supplementary Material.

2. Method

2.1 Smooth surrogate loss for outcome-weighted learning

To motivate our approach of choosing a smooth surrogate loss to learn the optimal ITRs, we first consider data from a single-stage randomized trial with two treatment arms. A treatment assignment is denoted by $A \in \mathcal{A} = \{-1, 1\}$. A patient's prognostic variables are denoted as a p-dimensional vector $\mathbf{X}$. We use R to denote the observable clinical outcome, also called the reward, and assume that R is positive and bounded from above, with larger values of R being more desirable. Data consist of $\{(\mathbf{X}_i, A_i, R_i) : i = 1 \ldots, n\}$.

For a given treatment decision $\mathcal{D}$, which maps $\mathbf{X}$ to $\{-1, 1\}$, we denote $\mathbb{P}^{\mathcal{D}}$ as the distribution of $(\mathbf{X}, A, R)$, given that $A = \mathcal{D}(\mathbf{X})$. Then, an optimal treatment rule is one that maximizes the value function

$$\mathbb{E}^{\mathcal{D}}(R) = \mathbb{E}\left\{ R \frac{I(A = \mathcal{D}(\mathbf{X}))}{A\pi + (1 - A)/2} \right\}, \tag{2.1}$$

where $\pi = P(A = 1|\mathbf{X})$. Following Qian and Murphy (2011), it can be shown that the maximization problem is equivalent to the problem of minimizing

$$\mathbb{E}\left\{ R \frac{I(A \neq \mathcal{D}(\mathbf{X}))}{A\pi + (1 - A)/2} \right\}. \tag{2.2}$$

The latter is a weighted classification error that can be estimated using the observed sample, as follows:

$$n^{-1}\sum_{i=1}^{n}\left\{R_i\frac{I(A_i\neq\mathcal{D}(\mathbf{X}_i))}{A_i\pi+(1-A_i)/2}\right\}. \tag{2.3}$$

Owing to the discontinuity and nonconvexity of the 0-1 loss on the right-hand side of (2.2), the direct minimization of (2.3) is difficult and a parameter inference is infeasible. To resolve this problem, the hinge loss from the support vector machine (SVM) was proposed as a substitute for the 0-1 loss (Zhao et al., 2012, 2015). However, owing to the nondifferentiability of the hinge loss, the inference remains challenging. This motivates us to seek a smoother surrogate loss function for estimation.

Consider an arbitrary surrogate loss $h(a,y):\{-1,1\}\times\mathcal{R}\mapsto\mathcal{R}$. Then, by replacing the 0-1 loss with this surrogate loss, we estimate the treatment rule by minimizing

$$R_h(f)=\mathbb{E}\left\{R\frac{h(A,f(\mathbf{X}))}{A\pi+(1-A)/2}\right\}. \tag{2.4}$$

To prevent nonconvexity, we require that $h(a,y)$ be convex in y. Furthermore, simple algebra gives

$$\begin{aligned}
&\mathbb{E}\left\{\frac{R}{A\pi+(1-A)/2}h(A,f(\mathbf{X}))\Big|\mathbf{X}=\mathbf{x}\right\}\\
=\ &\mathbb{E}[R|\mathbf{X}=\mathbf{x},A=1]h(1,f(\mathbf{x}))+\mathbb{E}[R|\mathbf{X}=\mathbf{x},A=-1]h(-1,f(\mathbf{x}))\\
=\ &a_{\mathbf{x}}h(1,f(\mathbf{x}))+b_{\mathbf{x}}h(-1,f(\mathbf{x})),
\end{aligned}$$

where $a_{\mathbf{x}} = \mathbb{E}[R|\mathbf{X} = \mathbf{x}, A = 1]$ and $b_{\mathbf{x}} = \mathbb{E}[R|\mathbf{X} = \mathbf{x}, A = -1]\}$. Hence, for any given $\mathbf{x}$, the minimizer for $f(\mathbf{x})$, denoted by $y_{\mathbf{x}}$, solves the equation

$$a_{\mathbf{x}} h'(1, y) + b_{\mathbf{x}} h'(-1, y) = 0,$$

where $h'(a, y)$ is the first derivative of $h(a, y)$ with respect to y. To ensure that the surrogate loss still leads to the correct optimal rule, which is equivalent to $\text{sgn}(a_{\mathbf{x}} - b_{\mathbf{x}})$, we require that the solution have the same sign as $(a_{\mathbf{x}} - b_{\mathbf{x}})$. On the other hand, because $a_{\mathbf{x}} h'(1, y) + b_{\mathbf{x}} h'(-1, y)$ is nondecreasing in y, we conclude that for $a_{\mathbf{x}} > b_{\mathbf{x}}$, if $a_{\mathbf{x}} h'(1, 0) + b_{\mathbf{x}} h'(-1, 0) \le 0$, then the solution $y_{\mathbf{x}}$ should be positive; however, for $a_{\mathbf{x}} < b_{\mathbf{x}}$, if $a_{\mathbf{x}} h'(1, 0) + b_{\mathbf{x}} h'(-1, 0) \ge 0$, then the solution $y_{\mathbf{x}}$ should be negative. In other words, a sufficient condition to ensure the Fisher consistency is

$$(a_{\mathbf{x}} - b_{\mathbf{x}})(a_{\mathbf{x}} h'(1, 0) + b_{\mathbf{x}} h'(-1, 0)) \le 0.$$

However, because $a_{\mathbf{x}}$ and $b_{\mathbf{x}}$ can be arbitrary nonnegative values, this condition holds if and only if

$$h'(1, 0) = -h'(-1, 0) \quad \text{and} \quad h'(1, 0) \le 0.$$

In conclusion, the choice of $h(a, y)$ should satisfy the following:

(I) For $a = -1$ and 1, $h(a, y)$ is twice differentiable and convex in y;

(II) $h'(1, 0) = -h'(-1, 0)$ and $h'(1, 0) \le 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

Many loss functions satisfy the above two conditions. Here, we consider loss functions of the form $h(a, y) = -ay + g(y)$. Then, the first condition automatically holds if g is twice differentiable and convex. The first equation in the second condition also holds. Finally, because $h'(1, 0) = -1 + g'(0)$, the second part holds if we choose g such that $g'(0) = 0$. A special case is to choose

$$g(y) = 2\log(1 + \exp(y)) - y,$$

with the corresponding loss function,

$$h(a, y) = -(a + 1)y + 2\log(1 + \exp(y)),$$

which corresponds to the entropy loss for a logistic regression (Figure 1). Henceforth, we use this loss function, although the results apply to any general smooth loss that satisfies these two conditions. Correspondingly, (2.4) becomes

$$R(f) = \mathbb{E}\left\{\frac{R}{A\pi + (1 - A)/2}[-0.5(A + 1)f(\mathbf{X}) + \log(1 + \exp(f(\mathbf{X})))]\right\} (2.5)$$

2.2 Learning optimal ITRs using the entropy loss

Now, suppose the randomized trial involves T stages, where patients might receive different treatments across the multiple stages. With some abuse

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

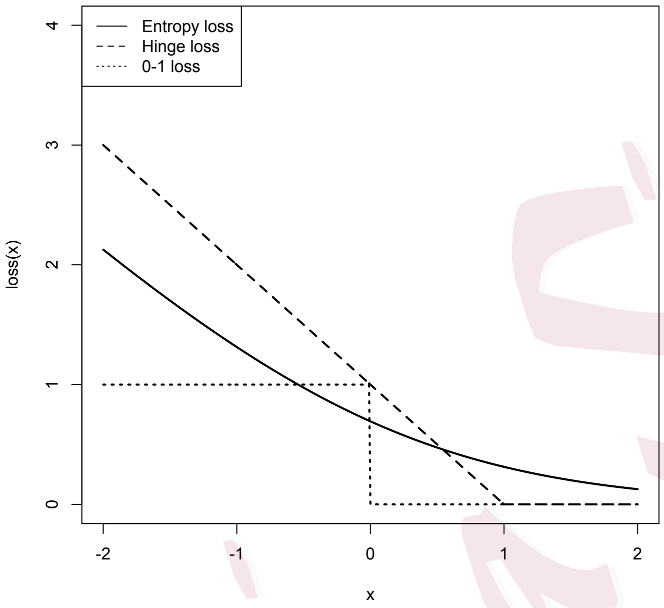

Figure 1: Comparison of loss functions.

of notation, we use $\mathbf{X}_t$, R_t, and A_t to denote the set of covariates, clinical outcome, and corresponding treatment, respectively, at stage $t = 1, \cdots, T$, and let $\mathbf{S}_t = (\mathbf{X}_1, A_1, \cdots, \mathbf{X}_{t-1}, A_{t-1}, \mathbf{X}_t)$ be the history by t.

A dynamic treatment regime (DTR) is a sequence of deterministic decision rules, $\mathbf{d} = (d_1, \cdots, d_T)$, where d_t is a map from the space of history information $\mathbf{S}_t$, denoted by $\mathcal{S}_t$, to the action space of available treatments $\mathcal{A}_t = \{-1, 1\}$. The optimal DTR maximizes the expected total value function $\mathbb{E}^{\mathbf{d}}(\sum_{t=1}^{T} R_t)$, where the expectation is taken with respect to the distribution of $(\mathbf{X}_1, A_1, R_1, \cdots, \mathbf{X}_T, A_T, R_T)$, given the treatment assignment $A_t = d_t(\mathbf{S}_t)$.

DTRs aim to maximize the expected cumulative rewards; hence, the optimal treatment decision at the current stage must depend on subsequent decision rules. This motivates a backward recursive procedure that first estimates the optimal decision rule at future stages. Then, it determines the optimal decision rule at the current stage by restricting the analysis to those subjects who have followed the estimated optimal decision rules. Assume that we observe data $(\mathbf{X}_{1i}, A_{1i}, R_{1i} \cdots, \mathbf{X}_{Ti}, A_{Ti}, R_{Ti})$, for $i = 1, \cdots, n$, forming n independent and identically distributed (i.i.d.) patient trajectories, and let $\mathbf{S}_{ti} = \{(\mathbf{X}_{1i}, A_{1i}, \cdots, A_{t-1,i}, \mathbf{X}_{ti}) : i = 1, \ldots, n\}$, for $1 \leq t \leq T$. Denote $\pi(A_t, \mathbf{S}_t) = A_t\pi_t - (1 - A_t)/2$, where $\pi_t = P(A_t = 1|\mathbf{S}_t)$, for $t = T, \ldots, 1$. Suppose that we already possess the optimal regimes at stages $t + 1, \cdots, T$, denoted as $d^*_{t+1}, \cdots, d^*_T$. Then, the optimal decision rule at stage t, $d^*_t(\mathbf{S}_t)$, should maximize

$$\mathbb{E}\left\{\Big(\sum_{j=t}^{T} R_j\Big)\frac{\prod_{j=t+1}^{T} I(A_j = d^*_j(\mathbf{S}_j))}{\prod_{j=t}^{T} \pi(A_j, \mathbf{S}_j)} I(A_t = d_t(\mathbf{S}_t))\big|\mathbf{S}_t\right\},$$

where we assume all subjects have followed the optimal DTRs after stage t. Hence, d^*_t is a map from $\mathcal{S}_t$ to $\{-1, 1\}$ that minimizes

$$\mathbb{E}\left\{\Big(\sum_{j=t}^{T} R_j\Big)\frac{\prod_{j=t+1}^{T} I(A_j = d^*_j(\mathbf{S}_j))}{\prod_{j=t}^{T} \pi(A_j, \mathbf{S}_j)} I(A_t \neq d_t(\mathbf{S}_t))\big|\mathbf{S}_t\right\}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

Following (2.5), we consider an entropy learning framework in which the decision function at stage t is given as

$$d_t(\mathbf{S}_t) = 2I\{(1+\exp(-f_t(\mathbf{X}_t)))^{-1} > 1/2\} - 1 = \operatorname{sgn}\{f_t(\mathbf{X}_t)\}, \qquad (2.6)$$

for some function $f_t(\cdot)$. Here, for simplicity, as defined in equation (2.6), the decision rule is assumed to depend on the history information $\mathbf{S}_t$ through $\mathbf{X}_t$ only. Although $\mathbf{S}_t = \mathbf{S}_{t-1} \cup \{A_{t-1}, \mathbf{X}_t\}$, any elements in $\mathbf{S}_{t-1}$ and A_{t-1} can be included as one the covariates in $\mathbf{X}_t$. Hence, this assumption is not stringent at all. In particular, our method remains valid when $\mathbf{X}_t$ is set to $\mathbf{S}_t$. Given the observed samples, we obtain estimators for the optimal treatments using the following backward procedure.

Step 1. Minimize

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_{Ti}}{\pi(A_{Ti},\mathbf{S}_{Ti})}\left[0.5(A_{Ti}+1)f_T(\mathbf{X}_{Ti}) - \log(1+\exp(f_T(\mathbf{X}_{Ti})))\right]\right\} (2.7)$$

to obtain the stage-T optimal treatment regime. This is the same as the single-stage treatment selection procedure. Let $\hat{f}_T$ be the estimator of f_T obtained by minimizing (2.7). Then, for a given $\mathbf{S}_T$, the estimated optimal regime is given by $\hat{d}_T(\mathbf{S}_T) = \operatorname{sgn}(\hat{f}_T(\mathbf{X}_T))$.

Step 2. For $t = T-1, \cdots, 1$, sequentially minimize

$$-n^{-1}\sum_{i=1}^{n}\left\{\frac{(\sum_{j=t}^{T} R_{ji})\prod_{j=t+1}^{T} I(A_{ji} = \hat{d}_j(\mathbf{S}_{ji}))}{\prod_{j=t}^{T}\pi(A_{ji},\mathbf{S}_{ji})}\big[0.5(A_{ti}+1)f_t(\mathbf{X}_{ti}) - \log(1+\exp(f_t(\mathbf{X}_{ti})))\big]\right\}, \quad (2.8)$$

where $\hat{d}_{t+1}, \cdots, \hat{d}_T$ are obtained prior to stage t. Let $\hat{f}_t$ be the estimator of f_t obtained by minimizing (2.8). Then, for a given $\mathbf{S}_t$, the estimated optimal regime is given by $\hat{d}_t(\mathbf{S}_t) = \text{sgn}(\hat{f}_t(\mathbf{X_t}))$.

Let $\mathcal{H}_{p_t}$ be the set of all functions from $\mathcal{R}^{p_t}$ to $\mathcal{R}$. As outlined in Section 2.1, the following proposition justifies the validity of our approach.

Proposition 1. *Suppose*

$$f_t = \arg\max_{f \in \mathcal{H}_{p_t}} \mathbb{E}\Big\{ \frac{(\sum_{j=t}^T R_j) \prod_{j=t+1}^T I(A_j = \text{sgn}(f_j(\mathbf{X}_j)))}{\prod_{j=t}^T \pi(A_j, \mathbf{S}_j)} \Big[0.5(A_t+1) f(\mathbf{X}_t) - \log(1 + \exp(f(\mathbf{X}_t)))\Big]\Big\}, \tag{2.9}$$

backward through $t = T, T-1, \ldots, 1$. *We have* $d_j^*(\mathbf{S}_j) = \text{sgn}(f_j(\mathbf{X}_j))$, *for* $j = 1, \ldots, T$.

Let $V_t = \mathbb{E}^{(d_t^*, \ldots, d_T^*)} \sum_{i=t}^T R_i$ be the maximal expected value function at stage t. After obtaining the estimated decision rules $\hat{d}_T, \ldots, \hat{d}_t$, for simplicity, we estimate V_t by

$$\hat{V}_t = n^{-1} \sum_{i=1}^n \left\{ \frac{(\sum_{j=t}^T R_{ji}) \prod_{j=t+1}^T I(A_{ji} = \hat{d}_j(\mathbf{S}_{ji}))}{\prod_{j=t}^T \pi(A_{ji}, \mathbf{S}_{ji})} I(A_{ti} = \hat{d}_t(\mathbf{S}_{ti})) \right\} \tag{2.10}$$

Note that our results also fit into the more general and robust estimation framework constructed by Zhang et al. (2012), Zhang et al. (2013).

3. Asymptotic Theory for Linear Decisions

Suppose the vector of stage-t covariates $\mathbf{X}_t$ is of dimension p_t, for $1 \leq t \leq T$, and assume that the function $f_t(\mathbf{X}_t)$ in (2.7) and (2.8) is of the linear form $f_t(\mathbf{X}_t) = (1, \mathbf{X}_t^\top)\beta_t$, for some $\beta_t \in \mathbb{R}^{p_t+1}$. Then, (2.7) and (2.8) can be carried out as a weighted logistic regression. In this section, we establish the asymptotic distributions of the estimated parameters and value functions under the aforementioned linear decision assumption. Note that when the true unknown solution is nonlinear, similarly to other linear learning rules, our approach can be understood only as finding the best approximation of the true solution (2.9) in the linear space.

We consider the multi-stage case only, because the results for the single-stage case are the same as those for stage T. For the multi-stage case, denote $\mathbf{X}_t^* = (1, \mathbf{X}_t^\top)^\top$ and the observations $\mathbf{X}_{ti}^* = (1, \mathbf{X}_{ti}^\top)^\top$, for $t = 1, \ldots, T$ and $i = 1, \ldots, n$. Then, the $n \times (p_t + 1)$ design matrix for stage t is given by $\mathbf{X}_{t,1:n} = (\mathbf{X}_{t1}^*, \ldots, \mathbf{X}_{tn}^*)^\top$. Let $\beta_t^0 = (\beta_{t0}^0, \beta_{t1}^0, \ldots, \beta_{tp_t}^0)^\top$ be the solution to (2.9) at stage t, and let $\hat{\beta}_t = (\hat{\beta}_{t0}, \hat{\beta}_{t1}, \ldots, \hat{\beta}_{tp_t})^\top$ be its estimator, obtained by solving (2.7) when $t = T$ and (2.8) when $t = T - 1, \ldots, 1$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

3.1 Parameter estimation

By setting the first derivative of (2.8) to zero for stage t, where $1 \leq t \leq T-1$, we have

$$\mathbf{0} = -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(\sum_{j=t}^{T} R_{ji})\prod_{j=t+1}^{T} I(A_{ji}=\hat{d}_j(\mathbf{S}_{ji}))}{\prod_{j=t}^{T}\pi(A_{ji},\mathbf{S}_{ji})}\left[.5(A_{ti}+1)-\frac{\exp(\mathbf{X}_{ti}^{*\top}\beta_t)}{1+\exp(\mathbf{X}_{ti}^{*\top}\beta_t)}\right]\right\}\mathbf{X}_{ti}^{*}.$$

The Hessian matrix of the left-hand side of the above equation is:

$$\mathbf{H}_t(\beta_t) = \frac{1}{n}\mathbf{X}_{t,1:n}^{\top}\mathbf{D}_t(\beta_t)\mathbf{X}_{t,1:n},$$

where $\mathbf{D}_t(\beta_t) = \text{diag}\{d_{t1},\dots,d_{tn}\}$ with

$$d_{ti} = \frac{(\sum_{j=t}^{T} R_{ji})\prod_{j=t+1}^{T} I(A_{ji}=\hat{d}_j(\mathbf{S}_{ji}))}{\prod_{j=t}^{T}\pi(A_{ji},\mathbf{S}_{ji})}\cdot\frac{\exp(\mathbf{X}_{ti}^{*\top}\beta_t)}{(1+\exp(\mathbf{X}_{ti}^{*\top}\beta_t))^2}.$$

Because R_{ti} is positive, $\mathbf{H}_t(\beta_t)$ is positive-definite with probability one. Consequently, the objective function in (2.8) is strictly convex, implying the existence and uniqueness of $\hat{\beta}_t$, for $t = T-1,\dots,1$. This is also true for $t = T$, using a similar argument. To obtain the asymptotic distribution of the estimators, we need the following regularity conditions:

(A1) $\mathbf{I}_t(\beta_t)$ is finite and positive-definite for any $\beta_t \in \mathbb{R}^{p_t+1}$, $t = 1,\dots,T$, where

$$\mathbf{I}_t(\beta_t) = \mathbb{E}\frac{(\sum_{j=t}^{T} R_j)\prod_{j=t+1}^{T} I(A_j = d_j(\mathbf{S}_j))}{\prod_{j=t}^{T}\pi(A_j,\mathbf{S}_j)}\cdot\frac{\exp(\mathbf{X}_t^{*\top}\beta_t)\mathbf{X}_t^{*}\mathbf{X}_t^{*\top}}{(1+\exp(\mathbf{X}_t^{*\top}\beta_t))^2}.$$

(A2) There exists a constant B_T, such that $R_t < B_T$, for $t = 1, \ldots, T$. In addition, we assume that $\mathbf{X}_{t1i}, \ldots, \mathbf{X}_{tni}$ are i.i.d. random variables with bounded support, for $i = 1, \ldots, p_t$. Here, $\mathbf{X}_{tij}$ is the jth element of $\mathbf{X}_{ti}$.

(A3) Denote $Y_t = \mathbf{X}_t^{*\top}\beta_t^0$ and let $g_t(y)$ be the density function of Y_t, for $1 \leq t \leq T$. We assume that $y^{-1}g_t(y) \to 0$ as $y \to 0$. In addition, we assume that there exists a small constant b, such that for any positive constant C and $\beta \in \mathcal{N}_{t,b} := \{\beta : |\beta - \beta_t^0|_\infty < b\}$, $P(|\mathbf{X}_t^{*\top}\beta| < Cy) = O(y)$ as $y \to 0$.

(A4) There exist constants $0 < c_{t1} < c_{t2} < 1$, such that $c_{t1} < \pi_t < c_{t2}$, for $t = 1, \ldots, T$, and $P(\prod_{j=1}^T I(A_j = d_j^*(\mathbf{S}_j)) = 1) > 0$.

Remark 1. By definition, $\mathbf{I}_t(\beta_t)$ is positive semidefinite. In A1, we assume that $\mathbf{I}_t(\beta_t)$ is positive-definite to ensure that the true optimal treatment rule is unique and estimable. The boundedness assumption, A2, can be relaxed further using truncation techniques. Assumption A3 indicates that the probability of $Y_t \leq Cn^{-\frac{1}{2}}$ is $o(n^{-\frac{1}{2}})$. This is necessary to ensure that the optimal decision is estimable, and is essential to establishing asymptotic normality without an additional Bernoulli point mass, as in Laber et al. (2014). Assumption A4 ensures that the treatment design is valid, such that the probability of a patient being assigned to the unknown optimal treatments is nonnegligible.

Theorem 1. *Under assumptions A1–A4, for $t = T, \ldots, 1$, and any constant $\kappa > 0$, there exists a large enough constant C_t,*

$$P\left(|\hat{\beta}_t - \beta_t^0|_\infty > C_t\sqrt{\frac{\log n}{n}}\right) = o\left(\frac{\log n}{n}\right), \tag{3.1}$$

and given $\mathbf{X}_t^*$, for any $x > 1$ and $x = o(\sqrt{n})$, we have

$$P\left(|\mathbf{X}_t^{*\top}(\beta_t^0 - \hat{\beta}_t)| > \frac{xW_t}{\sqrt{n}}\Big|\mathbf{X}_t^*\right) = \left\{1 + O\Big(\frac{x^3}{\sqrt{n}}\Big)\right\}\Phi(-x) + O\Big(\frac{\log n}{\sqrt{n}}\Big), \tag{3.2}$$

where $W_t^2 = \mathbb{V}ar(\mathbf{X}_t^{*\top}(\beta_t^0 - \hat{\beta}_t))$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. In addition, for the ith sample, we have

$$\mathbb{E}\left|\prod_{j=t}^{T} I(A_{ji} = \hat{d}_j(\mathbf{S}_{ji})) - \prod_{j=t}^{T} I(A_{ji} = d_j^*(\mathbf{S}_{ji}))\right| = o\left(\frac{\log n}{n}\right). \tag{3.3}$$

Furthermore, we have,

$$\sqrt{n}\mathbf{I}_t(\beta_t^0)(\hat{\beta}_t - \beta_t^0) \to N(\mathbf{0}, \mathbf{\Gamma}_t), \tag{3.4}$$

where $\mathbf{\Gamma}_t = (\gamma_{tjk})_{1\le j,k\le p+1}$ with

$$\begin{aligned}\gamma_{tjk} \quad = \quad & \mathbb{E}\left[\frac{(\sum_{j=t}^{T} R_{ji})\prod_{j=t+1}^{T} I(A_{ji} = d_j(\mathbf{S}_{ji}))}{\prod_{j=t}^{T}\pi(A_{ji}, \mathbf{S}_{ji})}\right]^2 \\ & \cdot\left[0.5(A_{ti} + 1) - \frac{\exp(\mathbf{X}_{ti}^{*\top}\beta_t^0)}{1 + \exp(\mathbf{X}_{ti}^{*\top}\beta_t^0)}\right]^2 \mathbf{X}_{tij}^*\mathbf{X}_{tik}^*,\end{aligned}$$

and $\mathbf{X}_{tij}^*$ is the jth element of $\mathbf{X}_{ti}^*$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

Remark 2. The proof of Theorem 1 is not straightforward because, for stage $t < T$, the n terms in the summation of the objective function (2.8) are weakly dependent on each other. Note that the estimation errors of the indicator functions in (2.8) might aggregate when the estimators are obtained sequentially. Thus, we need to show that the estimation errors of these indicator functions are well controlled. By establishing Bernstein-type concentration inequalities (3.1) and large deviation results (3.2) for the parameter estimation, we establish error bounds (3.3) for the estimation of these indicator functions. This enables us to establish the asymptotic distribution of the estimators. Detailed proofs are provided in the Supplementary Material. On the other hand, from the proofs, we can see that the asymptotic results in the above theorem would also hold if other loss functions satisfying the two conditions discussed in Section 2.1 are used, with some corresponding modifications to Condition (A1) and the covariance matrix.

In practice, we estimate $\mathbf{\Gamma}_t$ in Theorem 1 by $\hat{\mathbf{\Gamma}}_t = (\hat{\gamma}_{tjk})_{1\le j,k\le p_t+1}$, with

$$\begin{aligned}\hat{\gamma}_{tjk} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(\sum_{j=t}^{T} R_{ji})\prod_{j=t+1}^{T} I(A_{ji}=\hat{d}_j(\mathbf{S}_{ji}))}{\prod_{j=t}^{T}\pi(A_{ji},\mathbf{S}_{ji})}\right]^2 \\ &\quad \cdot\left[0.5(A_{ti}+1)-\frac{\exp(\mathbf{X}_{ti}^{*\top}\hat{\beta}_t)}{1+\exp(\mathbf{X}_{ti}^{*\top}\hat{\beta}_t)}\right]^2 \mathbf{X}_{tij}^{*}\mathbf{X}_{tik}^{*}.\end{aligned}$$

The covariance matrix of $\sqrt{n}(\hat{\beta}_t - \beta_t^0)$ can be estimated by: $\hat{\mathbf{\Sigma}}_t = \mathbf{H}_t^{-1}(\hat{\beta}_t)\hat{\mathbf{\Gamma}}_t\mathbf{H}_t^{-1}(\hat{\beta}_t)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

3.2 Estimating the optimal value function

In this subsection, we establish the asymptotic normality of the estimated maximal expected value function defined in (2.10) when $f(\mathbf{x})$ is a linear function of $\mathbf{x}$.

Theorem 2. *Under the same assumptions as Theorem 1, we have*

$$\sqrt{n}(\hat{V}_t - V_t) \to N(0, \Sigma_{V_t}), \quad t = 1, \ldots, T,$$

where $\hat{V}_t$ *is defined as in* (2.10) *and,*

$$\begin{aligned}\Sigma_{V_t} \;&=\; \mathbb{E}\left\{\frac{(\sum_{j=t}^T R_j)\prod_{j=t+1}^T I(A_j = d_j(\mathbf{S}_j))}{\prod_{j=t}^T \pi(A_j, \mathbf{S}_j)} I(A_t = d_t(\mathbf{S}_t))\right\}^2 - \\ &\left\{\mathbb{E}\frac{(\sum_{j=t}^T R_j)\prod_{j=t+1}^T I(A_j = d_j(\mathbf{S}_j))}{\prod_{j=t}^T \pi(A_j, \mathbf{S}_j)} I(A_t = d_t(\mathbf{S}_t))\right\}^2.\end{aligned}$$

When conducting inferences, Σ_{V_t} can be estimated using the empirical estimators,

$$\begin{aligned}\hat{\Sigma}_{V_t} \;&=\; \frac{1}{n}\sum_{i=1}^n\left\{\frac{(\sum_{j=t}^T R_{ji})\prod_{j=t+1}^T I(A_{ji} = \hat{d}_j(\mathbf{S}_{ji}))}{\prod_{j=t}^T \pi(A_{ji}, \mathbf{S}_{ji})} I(A_{ti} = \hat{d}_t(\mathbf{S}_{ti}))\right\}^2 - \\ &\left\{\frac{1}{n}\sum_{i=1}^n\frac{(\sum_{j=t}^T R_{ji})\prod_{j=t+1}^T I(A_{ji} = \hat{d}_j(\mathbf{S}_{ji}))}{\prod_{j=t}^T \pi(A_{ji}, \mathbf{S}_{ji})} I(A_{ti} = \hat{d}_t(\mathbf{S}_{ti}))\right\}^2.\end{aligned}$$

3.3 Testing treatment effects

In practice, treatments in some stages might not be effective for some patients. When the true optimal treatment rule is linear in $\mathbf{X}_t$, a nonsignificant treatment effect on stage t, for some $1 \le t \le T$, is equivalent to

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

$\mathbf{X}_t^{*\top}\beta_t^0 = 0$. Here, $\mathbf{X}_t^* = (1, \mathbf{X}_t^\top)^\top$. From Theorem 1 we immediately have that, given $\mathbf{X}_t$, $\mathbf{X}_t^{*\top}\hat{\beta}_t \to N(\mathbf{X}_t^{*\top}\beta_t^0, \frac{1}{n}\mathbf{X}_t^{*\top}\mathbf{I}_t(\beta_t^0)^{-1}\mathbf{\Gamma}_t\mathbf{I}_t(\beta_t^0)\mathbf{X}_t^*)$. Therefore, we can use $\mathbf{X}_t^{*\top}\hat{\beta}_t$ as a test statistic when testing the significance of the treatment effects: for a realization $\mathbf{x}_t^*$ and a given significance level α, we reject $H_0 : \mathbf{x}_t^{*\top}\beta_t^0 = 0$ if $\sqrt{n}|(\mathbf{x}_t^{*\top}\hat{\mathbf{I}}_t(\hat{\beta}_t)^{-1}\hat{\mathbf{\Gamma}}_t\hat{\mathbf{I}}_t(\hat{\beta}_t)\mathbf{x}_t^*)^{-1/2}\mathbf{x}_t^{*\top}\hat{\beta}_t| > \Phi(1-\alpha/2)$, where $\hat{\mathbf{I}}_t(\hat{\beta}_t), \hat{\mathbf{\Gamma}}_t(\hat{\beta}_t)$ are empirical estimators of $\mathbf{I}_t, \mathbf{\Gamma}_t$, respectively, evaluated at $\hat{\beta}_t$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Before we proceed to the numerical studies, note that the theoretical results obtained here are still valid if the model is mis-specified. However, the parameters we are estimating are the maximizer of (2.5) under the linear space, not the parameters in the optimal decision rules.

4. Simulation Study

We conduct numerical studies to assess the performance of our proposed methods.

One-stage. The treatment A is generated uniformly from $\{-1, 1\}$ and is independent of the prognostic variables $\mathbf{X} = (x_1, \ldots, x_p)^\top$. We set the reward $R = Q(\mathbf{X}) + T(\mathbf{X}, A) + \epsilon$, where $T(\mathbf{X}, A)$ reflects the interaction between the treatment and the prognostic variables, and ϵ is a random

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

variable such that $\epsilon = |Y|/10$, where Y follows a standard normal distribution. This folded normal error is chosen because R is restricted to be positive. We consider the following models.

Model 1. x_1, x_2, x_3 are generated independently and uniformly in $[-1, 1]$. We generate the reward $R = Q(\mathbf{X}) + T(\mathbf{X}, A) + \epsilon$ by setting $T(\mathbf{X}, A) = 3(.4 - x_1 - x_2)A$, $Q(\mathbf{X}) = 8 + 2x_1 - x_2 + .5x_3$. In this case, the decision boundary is determined only by x_1 and x_2.

Model 2. $\mathbf{X} = (x_1, x_2, x_3)^\top$ is generated from a multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})_{3\times 3}$, where $\sigma_{ij} = .5^{|i-j|}$, for $1 \leq i, j \leq 3$. We generate the reward R by setting $T(\mathbf{X}, A) = (.8 - 2x_1 - 2x_2)A$, $Q(\mathbf{X}) = 5 + .5x_1^2 + .5x_2^2 + .5(x_3^2 + .5x_3)$. The decision boundary of this case is also determined by x_1 and x_2.

Next, we consider multi-stages cases. The treatments A_t are generated independently and uniformly from $\{-1, 1\}$, and are independent of the p-dimensional vector of prognostic variables $\mathbf{X}_t = (x_{t1}, \ldots, x_{tp})^\top$, for $t = 1, \ldots, T$. ϵ is generated in the same way as in the single stage.

Two-stage.

Model 3. The Stage 1 outcome R_1 is generated as follows: $R_1 = (1 - 5x_{11} - 5x_{12})A_1 + 11.1 + .1x_{11} - .1x_{12} + .1x_{13} + \epsilon$, where x_{11}, x_{12}, x_{13} are generated independently from a uniform distribution in $[-1, 1]$. The Stage

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

2 outcome R_2 is generated by $R_2 = .5A_1A_2 + 3 + (.2 - x_{21} - x_{22})A_2 + \epsilon$, where $x_{2i} = x_{1i}$, for $i = 1, 2, 3$. In this case, the covariates from the two stages are identical.

Model 4. We use the same setting as that in Model 3, except that we set $x_{2i} = .8x_{1i} + .2U_i$, for $i = 1, 2, 3$, where U_i is randomly generated from $U[-1, 1]$. In this case, the covariates from the two stages are different and correlated.

4.1 Estimation and classification performance

We first examine the performance of the estimated coefficient parameters, the corresponding value functions, and the classification accuracy.

For stage t, given the sample size n, we repeat the simulation 2000 times. Then, we compute the coverage rate CR_{tj}, which is the proportion that $[\hat{\beta}_{tj} - 1.96\hat{\sigma}_{tjj}, \hat{\beta}_{tj} + 1.96\hat{\sigma}_{tjj}]$ covers the true parameter β_{tj}, for $j = 0, \ldots, p$, where $\hat{\sigma}_{tjj}$ is the (j, j)th element of $\hat{\Sigma}_t$. CR_{V_t} is defined similarly for the coverage rate of the value function. A validation set with 100,000 observations is simulated to compute the oracle values and assess the performance of our approach.

We set the sample size to $n = 50, 100, 200, 400,$ and 800. The coverage rates under Models 1−4 are given in Tables 1 and 2. For each replica-

	Model 1					Model 2				
n	CR_{V_1}	CR_{10}	CR_{11}	CR_{12}	CR_{13}	CR_{V_1}	CR_{10}	CR_{11}	CR_{12}	CR_{13}
50	0.927	0.948	0.950	0.938	0.945	0.946	0.944	0.937	0.931	0.924
100	0.936	0.950	0.947	0.949	0.944	0.942	0.947	0.949	0.945	0.940
200	0.942	0.954	0.947	0.955	0.952	0.951	0.950	0.950	0.953	0.947
400	0.940	0.949	0.960	0.954	0.944	0.946	0.963	0.952	0.949	0.933
800	0.944	0.944	0.953	0.947	0.943	0.951	0.955	0.952	0.954	0.943

Table 1: Coverage rates of the expected value function and coefficient parameters under Models 1 and 2.

tion under each model, we also compute the misclassification rate at each stage. Figure 2 gives the box plots of the misclassification rates over 2000 replications for all four models.

From Tables 1 and 2, we observe that the coverage rates are close to the nominal level (95%), and improve as the sample size increases, indicating that the asymptotic normality of our estimators is well established. In particular, the coverage rates of the coefficient parameter estimators are very close to 95%, even when the sample size is as small as 50. The box plots in Figure 2 also indicate that the misclassification rate of the estimated decision rule decreases toward zero as the sample size increases.

Note that the ultimate goal of dynamic treatment regimes is to maximize the value functions. Next we compare our entropy learning with Q-learning and outcome-weighted learning in terms of the value function esti-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

Model 3	Stage 1				Stage 2				
n	CR_{V_1}	CR_{10}	CR_{11}	CR_{12}	CR_{13}	CR_{V_2}	CR_{20}	CR_{21}	CR_{22}	CR_{23}
50	0.872	0.946	0.937	0.945	0.947	0.912	0.949	0.939	0.951	0.951
100	0.928	0.949	0.956	0.953	0.948	0.941	0.952	0.956	0.954	0.940
200	0.936	0.947	0.942	0.942	0.951	0.950	0.950	0.946	0.948	0.935
400	0.941	0.943	0.948	0.943	0.950	0.943	0.948	0.952	0.948	0.956
800	0.957	0.944	0.955	0.945	0.941	0.954	0.939	0.951	0.952	0.952
Model 4	Stage 1					Stage 2				
50	0.865	0.948	0.944	0.941	0.947	0.908	0.942	0.948	0.942	0.942
100	0.908	0.951	0.939	0.954	0.940	0.942	0.955	0.943	0.947	0.949
200	0.941	0.940	0.943	0.951	0.948	0.948	0.948	0.954	0.954	0.951
400	0.945	0.944	0.946	0.956	0.952	0.948	0.943	0.951	0.947	0.950
800	0.954	0.949	0.946	0.957	0.953	0.951	0.950	0.963	0.952	0.950

Table 2: Coverage rates of the expected value function and coefficient parameters under Models 3 and 4.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

mation. Throughout this paper, Q-learning and outcome-weighted learning are implemented using the R package "DTRlearn." In addition to Models 1−4, we also consider the following nonlinear cases.

Model 5. x_1, x_2, x_3 are generated independently and uniformly in $[-1, 1]$. We generate the reward $R = Q(\mathbf{X}, A) + \epsilon$ with $Q(\mathbf{X}, A) = [-T(\mathbf{X})(A + 1) + 2\log(1 + \exp(T(\mathbf{X})))]^{-1}$, where $T(\mathbf{X}) = (x_1 - x_2 + 2x_1x_2)$.

Model 6. This is identical to Model 5, except that x_1, x_2, x_3 are discrete variables generated independently and uniformly in $\{-1, 0, 1\}$.

Model 7. The Stage 1 outcome R_1 is generated as follows: $R_1 = [0.2 - T_1(\mathbf{X}_1)(A_1 + 1) + 2\log(1 + T_1(\mathbf{X}_1))]^{-1} + \epsilon$, where $T_1(\mathbf{X}_1) = x_{11} - x_{12} + 2x_{13}^2 + 2x_{11}x_{12}$, with x_{11}, x_{12}, x_{13} generated independently from a uniform distribution in $[-1, 1]$. The Stage 2 outcome R_2 is generated by $R_2 = [0.05 + (1 + A_2)(1 + A_1)/4 - T_2(\mathbf{X}_2)(A_2 + 1) + 2\log(1 + T_2(\mathbf{X}_2))]^{-1} + \epsilon$, where $x_{2i} = x_{1i}$, for $i = 1, 2, 3$, and $T_2(\mathbf{X}_2) = x_{21} - x_{22} + 2x_{23}^2 + 2x_{21}x_{22}$.

Model 8. This is identical to Model 7, except that x_{11}, x_{12}, x_{13} are discrete variables generated independently and uniformly in $\{-1, 0, 1\}$.

For each model, we generate 200 random samples and the corresponding estimated treatment rules used to compute the value function using (2.3), with a validation set of size $n = 500,000$. The above procedure is repeated 100 times; the results are reported in table 3.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

Model	E-Learning	Q-Learning	OW-Learning
Model 1	10.2(0.1)	10.3(0.0)	10.3(0.0)
Model 2	9.4(0.1)	9.4(0.0)	9.4(0.0)
Model 3 Stage 2	3.7(0.1)	3.7(0.0)	3.7(0.0)
Model 3 Stage 1	14.5(0.4)	15.0(0.0)	15.0(0.0)
Model 4 Stage 2	3.6(0.1)	3.6(0.0)	3.6(0.00)
Model 4 Stage 1	14.5(0.6)	15.0(0.0)	15.0(0.0)
Model 5	1.8(0.0)	1.7(0.0)	1.8(0.0)
Model 6	4.8(0.1)	4.1(0.1)	-(-)
Model 7 Stage 2	1.5(0.0)	1.5(0.0)	1.5(0.0)
Model 7 Stage 1	1.1(0.1)	1.0(0.0)	1.1(0.1)
Model 8 Stage 2	3.0(0.1)	2.8(0.2)	-(-)
Model 8 Stage 1	1.9(0.3)	0.9(0.2)	-(-)

Table 3: Comparison of value functions using entropy learning (E-learning), Q-learning, and outcome-weighted learning (OW-Learning) under Models 1–8.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

From Table 3, we note the value functions of our entropy learning method are comparable with those of Q-learning and outcome-weighted learning under Models 1−4. However, under Models 5 and 7, where the true treatment regimes are nonlinear, the value functions of entropy learning and outcome-weighted learning are very similar, and seem to be slightly better than those of Q-learning. However, when we consider discrete covariates in Models 6 and 8, outcome-weighted learning barely produces a result, owing to the large condition number when solving a system of equations.

4.2 Testing $\mathbf{X}_t^{*\top}\beta_t^0 = 0$

In the dynamic treatment regime literature, the nonregularity condition $P(\mathbf{X}_t^{*\top}\beta_t^0 = 0) = 0$ is usually required (e.g., in Q-learning) to enable parameter inferences. Here, we examine the performance the entropy learning approach when testing $\mathbf{X}_t^{*\top}\beta_t^0 = 0$.

- Case 1: Test $\mathbf{X}^{*\top}\beta_1^0 = 0$ under model 1. Let $\mathbf{X}^* = (1, x_1, x_2, x_3)^\top$ be the covariate of a new observation and $\beta_1^0 = (\beta_{10}^0, \beta_{11}^0, \beta_{12}^0, \beta_{13}^0)^\top$ be the true parameters. By setting $x_1 = x_3 = 1$ and $x_2 = -(\beta_{10}^0 + x_1\beta_{11}^0 + x_3\beta_{13}^0)/\beta_{12}^0$, we have $\mathbf{X}^{*\top}\beta_1^0 = 0$.

- Case 2: Test $\mathbf{X}_1^{*\top}\beta_1^0 = 0$ under model 4. We set $x_{11} = x_{13} = -1$ and $x_{12} = -(\beta_{10}^0 + x_{11}\beta_{11}^0 + x_{13}\beta_{13}^0)/\beta_{12}^0$.

We set $n = 50, 100, 200, 400$. Note that

$$\mathbf{X}_t^{*\top}\hat{\beta}_t \to N(\mathbf{X}_t^{*\top}\beta_t^0, \mathbf{X}_t^{*\top}\mathbf{I}_t(\beta_t^0)^{-1}\mathbf{\Gamma}_t\mathbf{I}_t(\beta_t^0)\mathbf{X}_t).$$

We use $\mathbf{X}_t^{*\top}\hat{\mathbf{I}}_t(\hat{\beta}_t)^{-1}\hat{\mathbf{\Gamma}}_t\hat{\mathbf{I}}_t(\hat{\beta}_t)\mathbf{X}_t^*$ to estimate the variance of $\mathbf{X}_t^{*\top}\hat{\beta}_t$, where $\hat{\mathbf{I}}_t$ and $\hat{\mathbf{\Gamma}}_t$ are the empirical estimators of $\mathbf{I}_t$ and $\mathbf{\Gamma}_t$. For each case, we run the simulation 1000 times, and for each replication, we compute the p-value of $\mathbf{X}_t^{*\top}\hat{\beta}_t$. P-value plots are given in Figures 3 and 4. We can see that the p-values follow a uniform distribution in $[0, 1]$, indicating that our tests perform well in detecting nonsignificant treatment effects.

4.3 Type-I error comparison with Q-learning

We next assess the performance of the hypothesis tests, because it is often of interest to investigate the significance of the coefficient parameters. Note that in Models 3 and 4, we have $\beta_{13} = \beta_{23} = 0$. We then compute the type-I error to test $\beta_{13} = 0$ and $\beta_{23} = 0$. In the optimization problems (2.7) and (2.8), the decisions A_i are formularized as the weights of a weighted negative log-likelihood. Consequently, unlike Q-learning (Zhao et al. (2009)), the objective functions for the estimation of the parameters become continuous functions, and parameter inferences become feasible, even without the nonregularity condition. For comparison, we compute the same quantities using the bootstrap scheme for Q-learning. Note that, in general, β_{ij}

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

	Model 3				Model 4			
	$H_0 : \beta_{13} = 0$		$H_0 : \beta_{23} = 0$		$H_0 : \beta_{13} = 0$		$H_0 : \beta_{23} = 0$	
n	Elearn	Qlearn	Elearn	Qlearn	Elearn	Qlearn	Elearn	Qlearn
50	0.063	0.069	0.050	0.057	0.060	0.054	0.055	0.056
100	0.044	0.063	0.054	0.056	0.044	0.057	0.043	0.055
400	0.049	0.043	0.055	0.043	0.047	0.053	0.047	0.046
800	0.050	0.059	0.044	0.064	0.047	0.053	0.054	0.055

Table 4: Type-I error comparison using entropy learning and Q-learning, where "Elearn" refers to entropy learning and "Qlearn" refers to Q-learning.

in entropy learning differs from the β_{ij} in Q-learning. However, in Models 3 and 4, x_{13} and x_{23} are not involved in the treatment selection part; hence, the true β in both entropy learning and Q-learning is zero. Here, the significance level α is set to 0.05, and we consider $n = 50, 100, 400, 800$. The simulation is repeated 2000 times, and the results are given in Table 4. Most of the type-I errors using entropy learning are closer to $\alpha = 0.05$, indicating that our learning method can be more appropriate for testing the significance of covariates.

5. Application to STAR*D

We consider a real-data example extracted from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study funded by the National Institute of Mental Health. STAR*D is a multisite, prospective,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

randomized, multistep clinical trial of outpatients with nonpsychotic major depressive disorder; see Rush et al. (2004) and Sinyor et al. (2010) for further details on the study. The complete trial involved four sequential treatment stages (or levels), and patients were encouraged to participate in the next level of treatment if they failed to achieve remission or experience an adequate reduction in symptoms.

During the first level of the STAR*D study, patients initially took the antidepressant citalopram, a selective serotonin reuptake inhibitor (SSRI). Those who did not experience a remission of symptoms for up to 14 weeks had the option of continuing to level 2 of the trial, where they could explore additional treatment options designed to help them become symptom-free (Rush et al. (2006)). Because there was one single treatment for all patients in level 1, we do not discuss these data further.

Level 2 of the study offered seven treatments: four "switched" options, in which study participants changed from citalopram to a new medication or talk therapy; and three "augmented" options, in which patients added a new medication or talk therapy to the citalopram they were already receiving. Data taken from Level 2 are treated as first-stage observations, and we define $A_1 = -1$ if the treatment option is a switch, and $A_1 = 1$ if the treatment option is an augmentation.

During levels 1 and 2 of the STAR*D trial, which started with 2,876 participants, about half of all patients became symptom-free. The other half were then eligible to enter level 3, where as in level 2, patients were given the choice of either switching medications or adding to their existing medication (Fava et al. (2006)). Data taken from level 3 of this trial are treated as second-stage observations, and we define $A_2 = -1$ if the treatment option is a switch, and $A_2 = 1$ if the treatment option is an augmentation.

After excluding cases with missing values, we obtain a sample of 316 patients whose medical information from the two stages are available. Of the 316 patients, 119 are assigned to the augmentation group, and 197 are assigned to the switch group in Stage 1. Then, 115 are assigned to the augmentation group, and 201 are assigned to the switch group in Stage 2. The 16-item Quick Inventory of Depressive Symptomatology-Self-Report (QIDS-SR(16)) scores were obtained during treatment visits for the patients, and are considered the primary outcome variable in this study. To accommodate our model, where the reward is positive and "larger is better," we used $R = c -$ QIDS-SR(16) as the reward at each level, where c is a constant that bounds the empirical QIDS-SR(16) scores. In this study, we simply set $c = 30$ so that all QIDS-SR(16) scores are positive.

Following earlier analysts (e.g., Kuk et al. (2010, 2014)), we consider

	Chronic		Gender		Age		GMC		Anxiety		Week		QIDS-SR(16)	
	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2
Switch	0.29 (0.03)	0.29(0.03)	0.51 (0.04)	0.46(0.04)	43.99 (0.88)	45.78(0.84)	0.59 (0.04)	0.62(0.03)	0.76 (0.03)	0.74(0.03)	9.21 (0.30)	7.48 (0.34)	14.96 (0.29)	14.54 (0.31)
Augmentation	0.26 (0.04)	0.26(0.04)	0.49 (0.05)	0.57(0.05)	44.76 (1.05)	41.65(1.11)	0.56 (0.05)	0.51(0.05)	0.70 (0.04)	0.73(0.04)	9.64 (0.40)	9.35 (0.46)	13.45 (0.37)	12.77 (0.37)

Table 5: Summary statistics for the covariates in the STAR*D study: for continuous variables, we report the means and standard deviations; for dichotomous variables, we report proportions and standard deviations.

the following set of clinically meaningful covariates: (i) chronic depression indicator, equal to one if the chronic episode $>$ 2 years, and 0 otherwise; (ii) gender, where male$=0$ and female$=1$; (iii) patient age (years); (iv) the general medical condition (GMC), defined as one in presence of one or more general medical conditions, and zero otherwise; (v) the anxious feature, defined as one if the Hamilton Depression Rating Scale anxiety/somatization factor score ≥ 7, and zero otherwise (Fava et al. (2008)). In addition, we consider (vi) week, the number of weeks patients spent in the corresponding stage when the QIDS-SR(16) scores at exit were determined, and (vii) the baseline QIDS-SR(16) scores at the corresponding stages. These covariates are summarized in Table 5.

We applied the methods introduced in this paper to estimate the covari-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

ate effects on the optimal treatment allocation for the patients in this study. The fitted results under the entropy learning approach are given in Table 6. The table shows that the baseline QIDS-SR(16) score is a significant predictor of whether a patient should be treated using the switch option or the argumentation option in both stages. More specifically, given other covariates, if the patient has a higher baseline score, adopting a switch option might have better medical outcome. In addition, for the Stage 2 analysis, the baseline score, gender, age, and the treatment time are all significant when determining the best treatment options. Interestingly, the treatment time is significant and has a positive sign, indicating that, given other covariates, treatment argumentation might benefit the patients for a longer term.

For comparison, using the same sets of covariates, the estimation results based on Q-learning are given in Table 7, where the estimated confidence intervals are obtained using the bootstrap procedure. Table 7 shows that gender is identified as the only important factor for the treatment selection at stage 2. This method may be less powerful than our proposed entropy learning method, because it may miss potentially useful markers. Consequently, Q-learning may not be able to achieve the most appropriate treatment allocation using a set of important personalized characteristics identi-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

	Stage 1		Stage 2	
	coefficient(sd)	p-value	coefficient(sd)	p-value
Entropy learning				
intercept	0.855 (0.987)	0.386	0.452 (0.792)	0.569
chronic	-1.231 (0.455)	**0.007**	0.103 (0.314)	0.742
gender	-0.604 (0.340)	0.859	0.702 (0.269)	**0.009**
age	0.001 (0.016)	0.950	-0.028 (0.012)	**0.022**
gmc	0.089 (0.359)	0.805	-0.121 (0.274)	0.658
anxious	0.095 (0.373)	0.799	0.235 (0.298)	0.431
week	0.066 (0.036)	0.071	0.089 (0.029)	**0.002**
qctot	-0.084 (0.044)	0.056	-0.111 (0.034)	**0.001**
A_1	-	-	0.925 (0.273)	**0.001**
$\hat{V}_i$	59.617 (5.485)	-	25.697 (1.325)	-

Table 6: Entropy learning for the STAR*D study.

fied from a significance study. To compare the performance of the proposed method with Q-learning in terms of the value function, we also compute the estimated mean and standard deviation of the value functions, using the fitted regimes obtained using our method and the Q-learning method; see the $\hat{V}_i$ values in Tables 6 and 7. We observe larger mean value functions for our entropy learning approach, indicating that our treatment regime is outperforming that of Q-learning in this data set.

The entropy learning approach may be incorrectly interpreted by some practitioners. The fitted regression model should not be confused with an ordinary association study, in which we fit unweighted logistic regression

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

	Stage 1			Stage 2		
	coefficient	Lower	Upper	coefficient	Lower	Upper
Q-learning						
intercept	0.99	-4.28	5.50	-2.17	-5.32	0.73
chronic	-0.48	-2.31	1.33	-0.63	-1.75	0.48
gender	0.66	-0.80	2.24	**1.30**	0.37	2.28
age	-0.03	-0.09	0.04	0.02	-0.03	0.07
gmc	0.06	-1.48	1.59	0.26	-0.83	1.39
anxious	1.35	-0.32	3.00	0.62	-0.45	1.65
week	-0.14	-0.31	0.04	-0.07	-0.16	0.02
qctot	-0.06	-0.24	0.14	-0.02	-0.16	0.11
A_1	-	-	-	0.11	-0.44	0.66
$\hat{V}_i$	40.34	32.08	48.60	20.54	17.83	23.25

Table 7: Bootstrap confidence interval of Q-learning for the STAR*D study. Lower: lower bound of the 95% confident interval; Upper: upper bound of the 95% confident interval.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

models to the two stage data (see Table 8). In fact, the significant findings from Table 8 only establish how covariates affect the likelihood of being observed in a treatment, in lieu of the likelihood of being allocated the most appropriate treatment.

Finally, because the original design at level 2 of the STAR*D trial was an equipoise-stratified design, one potential source of confounding effects could be due to a patients preference for the strata in the design. A further examination of this issue should include a patients preference in the treatment estimation strategies if we trust that patients selection of treatment options (between switch and augmentation) within each stratum is random, as assumed in the original equipoise-stratified design (Sinyor et al., 2010).

6. Discussion

Many open questions can be addressed using our proposed method. First, the linear specification of the treatment allocation rule may be replaced with a nonparametric formulation, such as a partly linear model or an additive regression model. The implementation of such methods is now widely available in most statistical packages. More effort is required to establish similar theoretical properties to those discussed here, and to achieve interpretable results.

	Stage 1		Stage 2	
	coefficient(sd)	p-value	coefficient(sd)	p-value
intercept	0.210 (0.740)	0.777	0.182 (0.772)	0.813
chronic	-0.183 (0.279)	0.511	0.141 (0.296)	0.635
gender	0.012 (0.242)	0.961	0.537 (0.260)	**0.039**
age	0.010 (0.011)	0.352	-0.030 (0.012)	**0.012**
gmc	-0.093 (0.259)	0.719	-0.219 (0.275)	0.425
anxious	-0.117 (0.275)	0.671	0.096 (0.295)	0.744
week	0.032 (0.029)	0.269	0.098 (0.027)	$<$ **0.001**
qctot	-0.091 (0.030)	**0.003**	-0.104 (0.031)	**0.001**
A_1	-	-	0.851 (0.260)	**0.001**

Table 8: Ordinary association study for the STAR*D data using logistic regression models.

Second, to carry out the clinical study and select the best treatment using our approach, it is necessary to evaluate the required sample size at the designing stage. Applying our theoretical results attained, we can calculate the total number of subjects for every treatment group. However, more empirical studies on various types of settings and data distributions can provide stronger support for the suggestion based on the asymptotic results.

Finally, missing values are quite common in a multi-stage analysis. Most analysts follow the standard practice of excluding cases with missing observations, under the missing-at-random assumption. It is a difficult task to investigate why data are missing, and an even more difficult task to address the problem when missing is not at random. We encourage further research in this direction.

7. Acknowledgements

We thank the Editor, the Associate Editor, and two reviewers for their instructive comments. The work was partly supported by Academic Research Funds R-155-000-174-114 and R-155-000-195-114, and Tier 2 MOE funds in Singapore MOE2017-T2-2-082, R-155-000-197-112 (Direct cost) and R-155-000-197-113 (IRC), and Hong Kong RGC grant PolyU 253023/16P.

Supplementary Material

The Supplement Material provides the technical proofs for the propositions and theorems.

References

Bartlett, P., Jordan, M. & McAuliffe, J. (2006). Convexity, classification, and risk bounds. *J. Am. Statist. Assoc.* **101**, 138–156.

Chakraborty, B., Murphy, S. & Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods. Med. Res.* **19**, 317–343.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

Fava, M. et al. (2006) A comparison of mirtazapine and nortriptyline following two consecutive failed medication treatments for depressed outpatients: A star*d report. *Am. J. Psychiatry* **163**, 1161–1172.

Fava, M. et al. (2008) Difference in treatment outcome in outpatients with anxious versus nonaxious depression: a star*d report. *Am. J. Psychiatry* **165**, 342–351.

Goldberg, Y. & Kosorok, M. (2012). Q-learning with censored data. *Ann. Statist.* **40**, 529.

Hjort, N. & Pollard, D. (2011). Asymptotics for minimisers of convex processes. Preprint series. ArXiv preprint arXiv:1107.3806.

Kuk, A., Li, J. & Rush, A. (2010). Recursive subsetting to identify patients in the star* d: a method to enhance the accuracy of early prediction of treatment outcome and to inform personalized care. *J. Clin. Psychiatry* **71**, 1502–1508.

Kuk, A., Li, J. & Rush, A. (2014). Variable and threshold selection to control predictive accuracy in logistic regression. *J. R. Statist. Soc.* C **63**, 657–672.

Laber, E., Lizotte, D., Qian, M., Pelham, W., & Murphy, S. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Stat.*, **8**, 1225–1272.

Luedtke, A. & van der Laan, M. (2016). Super-learning of an optimal dynamic treatment rule. *Int. J. Biostat.* **12**, 305–332.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective.* MIT press.

Murphy, S. (2003). Optimal dynamic treatment regimes. *J. R. Statist. Soc.* B **65**, 331–366.

Murphy, S. (2003). A generalization error for Q-learning. *J. Mach. Learn. Res.* **6**, 1073–1097.

Qian, M. & Murphy, S. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39**, 1180.

Robins, J., Hernan, M. & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiol.* **11**, 550–560.

Rush, A. et al. (2004). Sequenced treatment alternatives to relieve depression (star* d): rationale and design. *Control. Clin. Trials* **25**, 119–142.

Rush, A. et al. (2006). Bupropion-sr, sertraline, or venlafaxine-xr after failure of ssris for depression. *N. Engl. J. Med.* **354**, 1231–1242.

Sinyor, M., Schaffer, A. & Levitt, A. (2010). The sequenced treatment alternatives to relieve depression (star* d) trial: a review. *Can. J. Psychiatry* **55**, 126–135.

Song, R., Wang, W., Zeng, D. & Kosorok, M. (2015). Penalized q-learning for dynamic treatment regimens. *Stat. Sin.* **25**, 901–920.

Watkins, C. & Dayan, P. (1992). Q-learning. *Mach. Learn.* **8**, 279–292.

Zhao, Y., Kosorok, M. & Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Stat. Med.* **28**, 3294–3315.

Zhang, B., Tsiatis, A., Laber, E. & Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68**, 1010–1018.

Zhang, B., Tsiatis, A., Laber, E. & Davidian, M. (2012). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100**, 681–694.

Zhao, Y., Zeng, D., Laber, E. & Kosorok, M. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Am. Stat. Assoc.* **110**, 583–598.

Zhao, Y., Zeng, D., Rush, A. & Kosorok, M. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.* **107**, 1106–1118.

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China.

E-mail: by.jiang@polyu.edu.hk

Department of Statistics, North Carolina State University, North Carolina 27695, USA.

E-mail: rsong@ncsu.edu

Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore.

E-mail: stalj@nus.edu.sg

Department of Statistics, North Carolina State University, North Carolina 27695, USA.

E-mail: dzeng@email.unc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

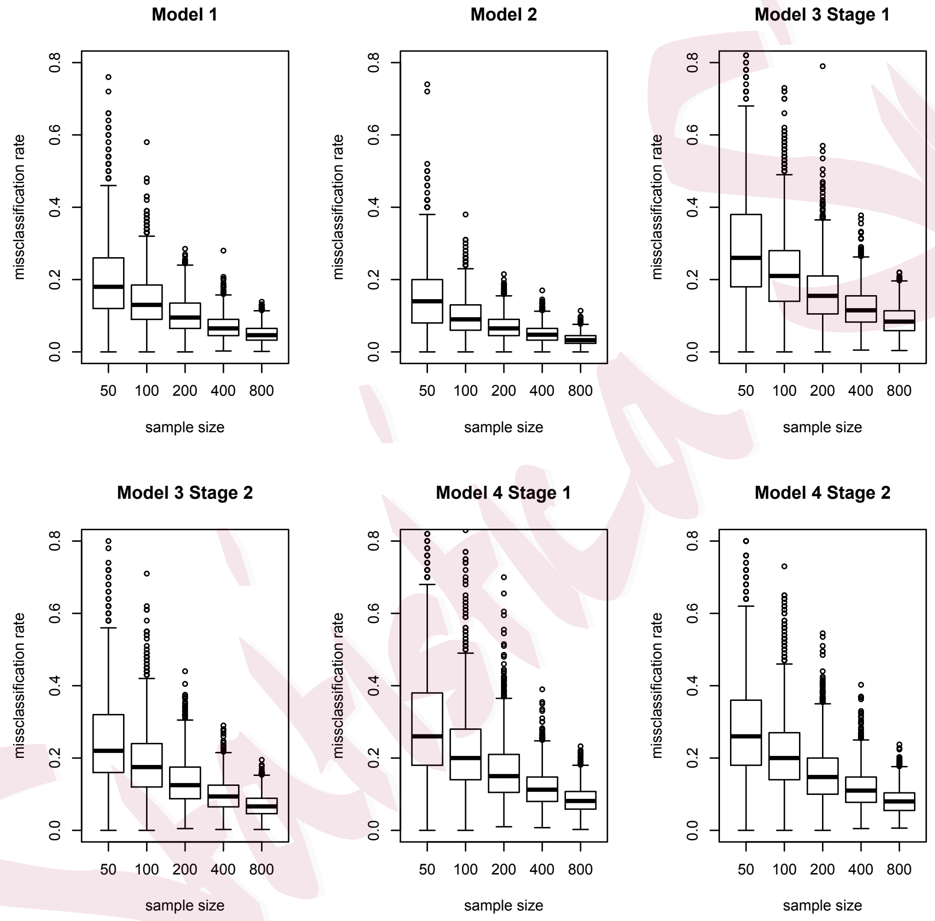

Figure 2: Box plot of misclassification rates over 2000 replications.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

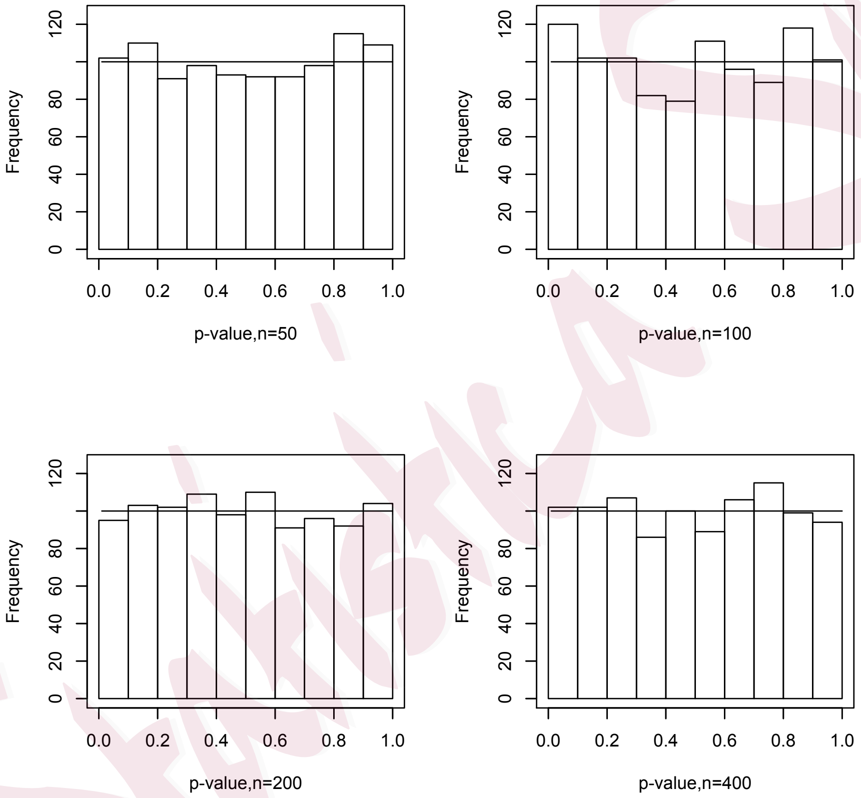

Figure 3: P-value of $X^\top\hat{\beta}_1$ under case 1 over 1000 replications.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0076

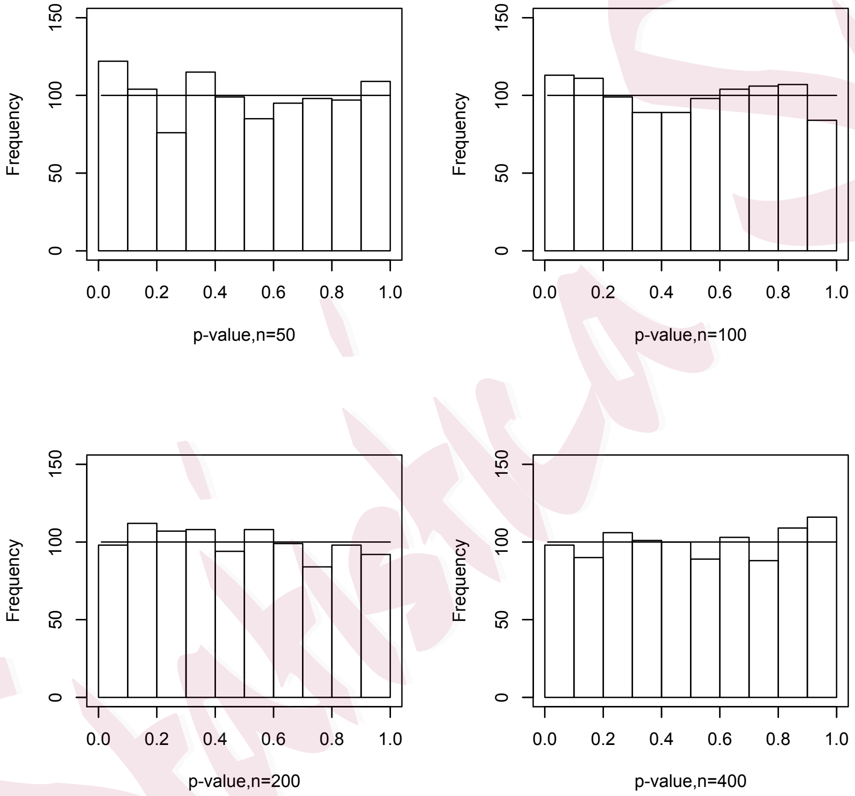

Figure 4: P-value of $X_1^\top \hat{\beta}_1$ under case 2 over 1000 replications.