Statistica Sinica **29** (2019), 1633-1710
doi:https://doi.org/10.5705/ss.202018.0076

DISCUSSION PAPER

ENTROPY LEARNING FOR DYNAMIC TREATMENT REGIMES

Binyan Jiang, Rui Song, Jialiang Li and Donglin Zeng

The Hong Kong Polytechnic University, North Carolina State University
National University of Singapore and University of North Carolina, Chapel Hill

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.

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, Murphy and Strecher (2010); Goldberg and Kosorok (2012); Laber et al. (2014); Song et al. (2015)) and A-learning (Robins, Hernan and Brumback (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 asymptotic 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, Jordan and McAuliffe (2006) produced a profound conceptual work on classification loss, for a relatively general setting. However, to link their work to recursive or dynamic optimization is not trivial. To do so, we employ a logistic loss. Luedtke and van der Laan (2016b) 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 $h'(1, 0) \le 0$.

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

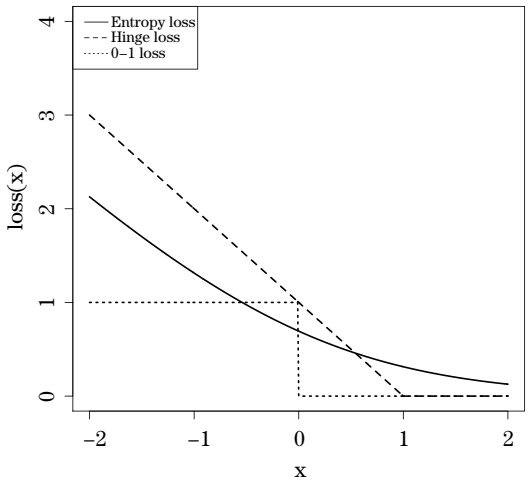

Figure 1. Comparison of loss functions.

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

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

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

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 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 $t = 1, \ldots, T$, and let $\mathbf{S}_t = (\mathbf{X}_1, A_1, \ldots, \mathbf{X}_{t-1}, A_{t-1}, \mathbf{X}_t)$ be the history by t.

A dynamic treatment regime (DTR) is a sequence of deterministic decision rules, $\mathbf{d} = (d_1, \ldots, d_T)$, 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 $(\mathbf{X}_1, A_1, R_1, \ldots, \mathbf{X}_T,$

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 $(\mathbf{X}_{1i}, A_{1i}, R_{1i} \ldots, \mathbf{X}_{Ti}, A_{Ti}, R_{Ti})$, for $i = 1, \ldots, n$, forming n independent and identically distributed (i.i.d.) patient trajectories, and let $\mathbf{S}_{ti} = \{(\mathbf{X}_{1i}, A_{1i}, \ldots, A_{t-1,i}, \mathbf{X}_{ti}) : i = 1, \ldots, n\}$, for $1 \le t \le T$. 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 $t+1, \ldots, T$, denoted as $d^*_{t+1}, \ldots, d^*_T$. Then, the optimal decision rule at stage t, $d^*_t(\mathbf{S}_t)$, should maximize

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

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

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

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

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

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

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

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 $t = T-1, \ldots, 1$, sequentially minimize

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

where $\hat{d}_{t+1}, \ldots, \hat{d}_T$ 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 $\hat{d}_t(\mathbf{S}_t) = \operatorname{sgn}(\hat{f}_t(\mathbf{X_t}))$.

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*

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

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

totic 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 $t = T - 1, \ldots, 1$.

3.1. Parameter estimation

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

$$\mathbf{0} = -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\left(\sum_{j=t}^{T} R_{ji}\right)\prod_{j=t+1}^{T} I(A_{ji} = \hat{d}_j(\mathbf{S}_{ji}))}{\prod_{j=t}^{T}\pi(A_{ji}, \mathbf{S}_{ji})}\right.$$
$$\left.\left[0.5(A_{ti}+1) - \frac{\exp(\mathbf{X}_{ti}^{*\top}\beta_t)}{1+\exp(\mathbf{X}_{ti}^{*\top}\beta_t)}\right]\right\}\mathbf{X}_{ti}^*.$$

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 $\mathbf{D}_t(\beta_t) = \operatorname{diag}\{d_{t1}, \ldots, d_{tn}\}$ with

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

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 $t = T - 1, \ldots, 1$. 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 $\beta_t \in \mathbb{R}^{p_t+1}$, $t = 1, \ldots, T$, where

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

(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 $1 \le t \le T$. 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 $Y_t \le Cn^{-1/2}$ is $o(n^{-1/2})$. 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*

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

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* i*th 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*

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

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

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

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

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

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,*

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

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

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

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 $\mathbf{X}_t^{*\top}\beta_t^0 = 0$. 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 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 $T(\mathbf{X}, A) = 3(0.4 - x_1 - x_2)A$, $Q(\mathbf{X}) = 8 + 2x_1 - x_2 + 0.5x_3$. 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 $\sigma_{ij} = 0.5^{|i-j|}$, for $1 \le i, j \le 3$. We generate the reward R by setting $T(\mathbf{X}, A) = (0.8 - 2x_1 - 2x_2)A$, $Q(\mathbf{X}) = 5 + 0.5x_1^2 + 0.5x_2^2 + 0.5(x_3^2 + 0.5x_3)$. 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: $R_1 = (1 - 5x_{11} - 5x_{12})A_1 + 11.1 + 0.1x_{11} - 0.1x_{12} + 0.1x_{13} + \epsilon$, where x_{11}, x_{12}, x_{13} are generated independently from a uniform distribution in $[-1, 1]$. The Stage 2 outcome R_2 is generated by $R_2 = 0.5A_1A_2 + 3 + (0.2 - x_{21} - x_{22})A_2 + \epsilon$, 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 $x_{2i} = 0.8x_{1i} + 0.2U_i$, 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 2,000 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 replication under each

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

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

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

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 estimation. Throughout this paper, Q-learning and outcome-weighted learning are implemented using the

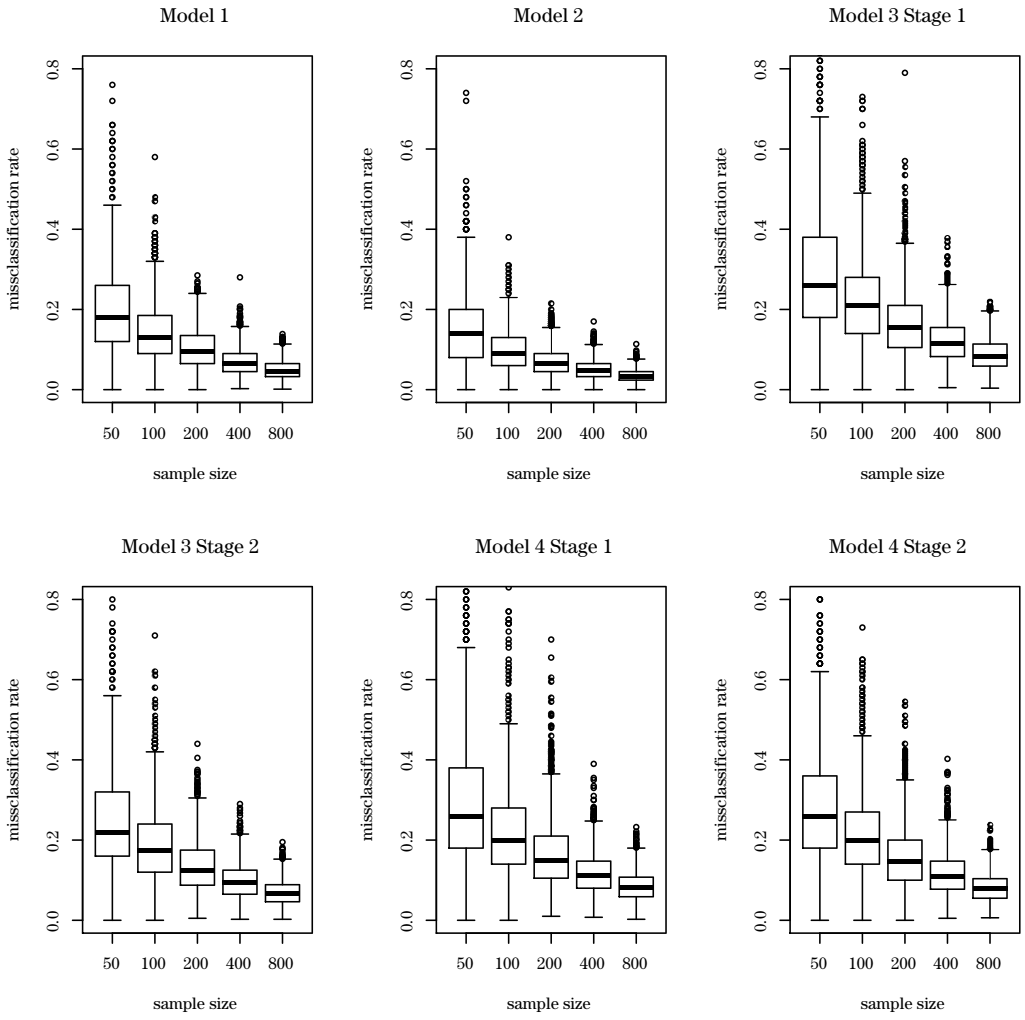

Figure 2. Box plot of misclassification rates over 2,000 replications.

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 $T_1(\mathbf{X}_1) = x_{11} - x_{12} + 2x_{13}^2 +$

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

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(0.0)
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)	-(-)

$2x_{11}x_{12}$, 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.

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

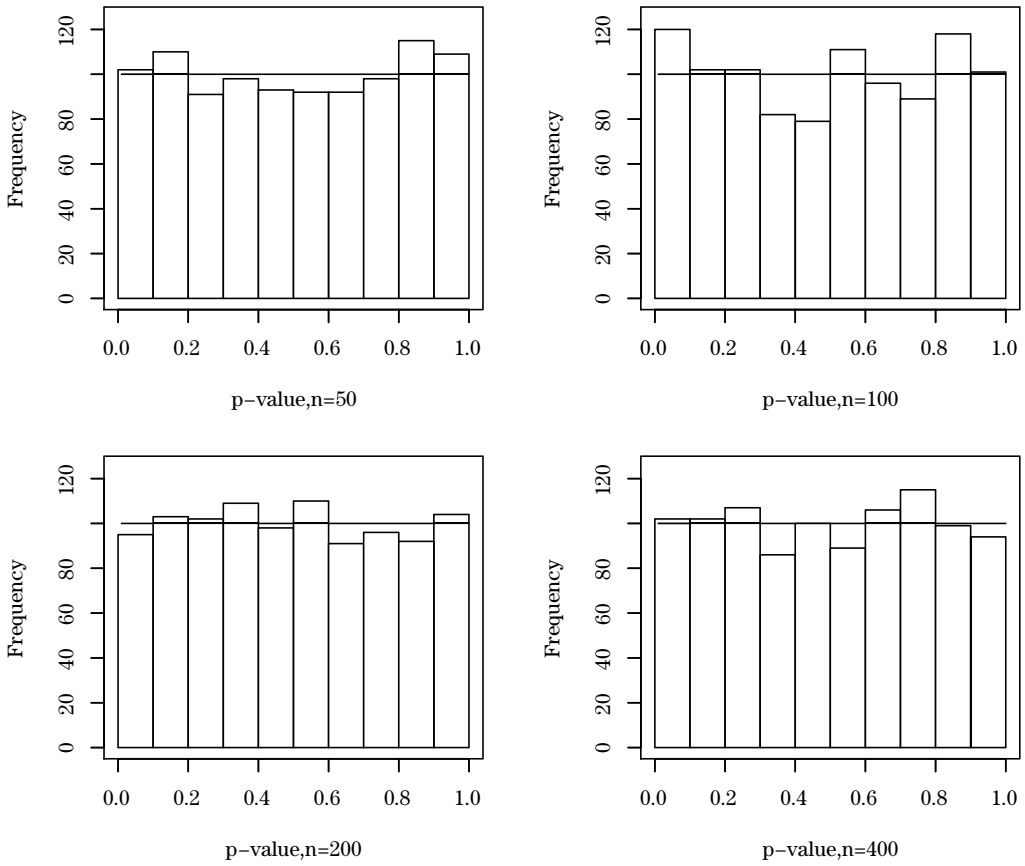

Figure 3. P-value of $X^\top \hat{\beta}_1$ under case 1 over 1,000 replications.

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

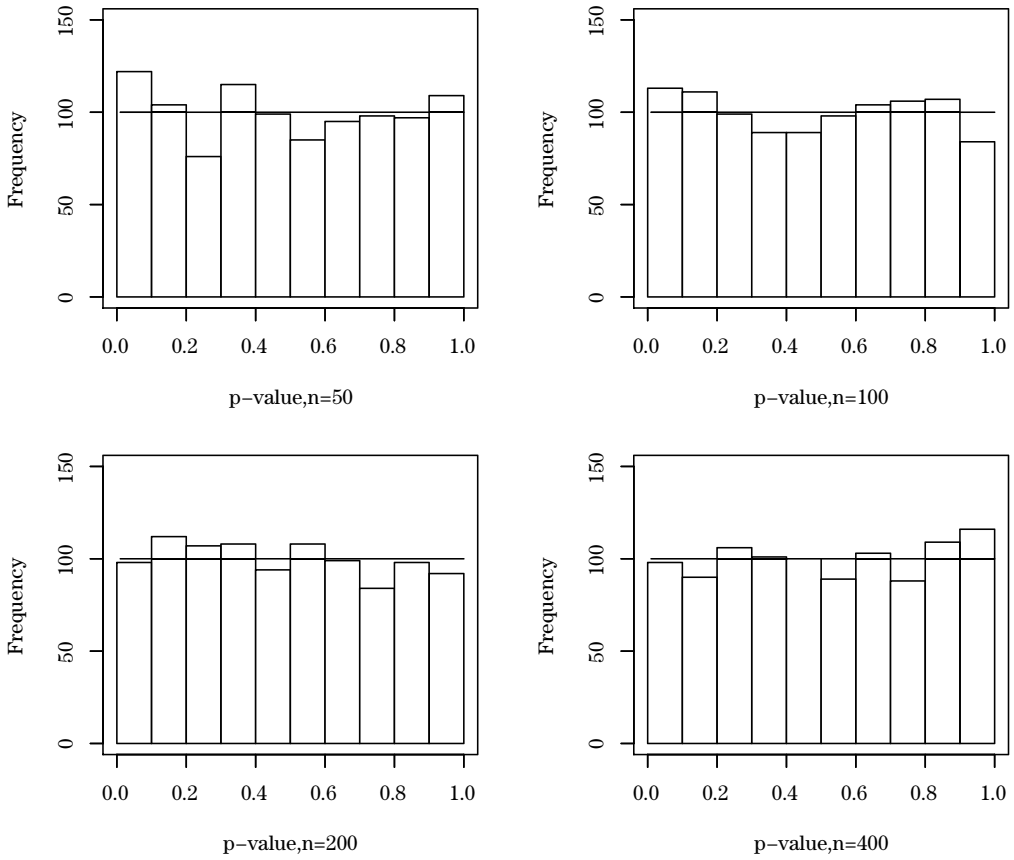

Figure 4. P-value of $X_1^\top \hat{\beta}_1$ under case 2 over 1,000 replications.

1,000 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, Kosorok and Zeng (2009)), the objective functions for the estimation of the parameters become continuous functions, and

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

	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

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} 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 2,000 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, randomized, multistep clinical trial of outpatients with nonpsychotic major depressive disorder; see Rush et al. (2004) and Sinyor, Schaffer and Levitt (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 $R = c - \text{QIDS-SR(16)}$ 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, Li and Rush (2010, 2014)), we consider 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.

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.

	Chronic		Gender		Age		GMC
	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1
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)
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)
	GMC	Anxiety		Week		QIDS-SR(16)	
	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2
Switch	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	0.51(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 6. Entropy learning for the STAR*D study.

	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)	-

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

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.

	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

tervals 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 identified 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 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

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

	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**

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, Schaffer and Levitt (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.

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.

Supplementary Material

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

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.

References

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

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.

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

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

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

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

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

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

Robins, J., Hernan, M. and 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. and 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. and Kosorok, M. (2015). Penalized q-learning for dynamic treatment regimens. *Statist. Sinica* **25**, 901–920.

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

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

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

Zhang, B., Tsiatis, A., Laber, E. and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100**, 681–694.

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

Zhao, Y., Zeng, D., Rush, A. and Kosorok, M. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. 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, Raleigh, NC 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 Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA.

E-mail: dzeng@email.unc.edu

(Received February 2018; accepted October 2018)

DISCUSSION

Wenbin Lu

North Carolina State University

I congratulate the authors for their thoughtful article on using entropy learn-

ing to estimate optimal individualized treatment rules. Their study makes several major contributions. First, it proposes a class of smooth outcome-weighted loss functions for estimating optimal individualized treatment rules. Second, the Fisher consistency and a proper inference for the parameters of the estimated treatment rules can be established in the proposed general framework.

The proposed smooth loss function is motivated by the sign consistency of the derived optimal treatment rule. When the true optimal treatment decision rule is contained in the considered class of treatment rules, can the estimated optimal treatment rule obtained using entropy learning be shown to have sign consistency asymptotically? In addition, given a treatment rule, the proposed smooth loss function is just an approximation to the weighted classification error loss (corresponding to the value function). Is it possible to quantify the difference between the value functions under the derived optimal rule using entropy learning and the true optimal rule? Among the class of proposed smooth loss functions, is it possible to find an optimal loss function that minimizes the value difference?

To derive the asymptotic properties of the parameter estimates in the derived optimal treatment rule, the authors make a few assumptions. In particular, assumption (A3) ensures that the optimal decision is estimable. Does this assumption exclude the possibility of having a nonregular setting, that is $P(X_t^{*\prime}\beta_t^0 = 0) > 0$. Under the nonregular setting, can we establish the asymptotic distributions of the estimators in the derived optimal treatment rule and its associated estimated value function, as in Theorems 1 and 2? Finally, can the proposed entropy learning method be extended to accommodate multiple treatment options at each treatment stage? I would appreciate comments from the authors on these issues.

Department of Statistics, North Carolina State University, 5112 SAS Hall, 2311 Stinson Drive, Raleigh, NC 27695, USA.

E-mail: lu@stat.ncsu.edu

(Received January 2019; accepted January 2019)

DISCUSSION

Xin He[1], Shirong Xu[2] and Junhui Wang[2]

[1]*Shanghai University of Finance and Economics*
and [2]*City University of Hong Kong*

We would like to congratulate Drs. Jiang, Song, Li, and Zeng (JSLZ) for their well-written and thought-provoking work, which bridges machine learning and statistical inferences when estimating optimal individualized treatment rules (ITRs), and opens numerous avenues for future research on related topics. Below we discuss the paper from two aspects: its extension to kernel-based nonparametric ITRs, and inferences for nonparametric ITRs.

1. Kernel-based Nonparametric ITRs

JSLZ assume that the decision functions $f_t(\mathbf{x})$, for $t = 1, \ldots, T$, have a linear form, which facilitates the model fitting and statistical inferences. However, in the machine learning community, much research is being conducted on nonparametric decision functions, for example, in a reproducing kernel Hilbert space (RKHS; see Kimeldorf and Wahba (1971); Shen et al. (2003); Wang and Shen (2007); Wang, Shen and Liu (2008); Zhao et al. (2015); Qi and Liu (2018), and the references therein). An RKHS provides a flexible framework for modeling nonparametric functions without explicitly enumerating the functional basis. It can be fully induced by any symmetric and nonnegative definite kernel function, where the choice of kernel functions relies on the available prior information about f_t. In practice, if no prior information is available, it is a common practice to use the Gaussian kernel, which is known to be universal in the sense that any continuous function can be well approximated by the induced RKHS under the infinity norm Steinwart (2005).

To extend JSLZ to estimate kernel-based nonparametric ITRs, we consider the case of $f_t \in \mathcal{H}_K$, an RKHS induced by some kernel function $K(\cdot, \cdot)$. The formulation of the kernel-based nonparametric ITRs then becomes

$$\min_{f_t \in \mathcal{H}_K} -\frac{1}{n}\Big(\boldsymbol{\omega}_t \odot \big(0.5(\mathbf{A}_t + \mathbf{1}_n) \odot \mathbf{f}_t + \ln \boldsymbol{\xi}_t\big)\Big)^T \mathbf{1}_n + \lambda_n \|f_t\|_K^2, \tag{1.1}$$

where $\boldsymbol{\omega}_t = (\omega_{t1}, \ldots, \omega_{tn})^T$, $\mathbf{A}_t = (A_{t1}, \ldots, A_{tn})^T$, $\mathbf{f}_t = (f_t(\mathbf{x}_1^t), \ldots, f_t(\mathbf{x}_n^t))^T$, $\odot$ denotes a componentwise product, $\boldsymbol{\xi}_t = (\xi_{t1}, \ldots, \xi_{tn})^T$ with $\xi_{ti} = \big(1 + \exp(f_t(\mathbf{x}_i^t))\big)^{-1}$, and $\|f_t\|_K^2 = \langle f_t, f_t \rangle_K$ is the associated RKHS-norm of f_t. By the

Table 1. Comparison of value function of linear and nonparametric ITRs with their standard errors in parenthesis.

Method	Linear ITR	Nonparametric ITR
Value function	1.568(0.010)	1.629(0.008)

representer theorem Kimeldorf and Wahba (1971), the minimizer of (1.1) must have the form

$$f_t(\mathbf{x}^t) = \sum_{i=1}^{n} \alpha_{ti} K(\mathbf{x}_i^t, \mathbf{x}^t) = \boldsymbol{\alpha}_t^T \mathbf{K}_n(\mathbf{x}^t),$$

where $\boldsymbol{\alpha}_t = (\alpha_{t1}, \ldots, \alpha_{tn})^T$ and $\mathbf{K}_n(\mathbf{x}^t) = (K(\mathbf{x}_1^t, \mathbf{x}^t), \ldots, K(\mathbf{x}_n^t, \mathbf{x}^t))^T$. Moreover, let $\mathbf{K} = \big((K(\mathbf{x}_i, \mathbf{x}_j))\big)_{i,j=1}^n$. Then $\mathbf{f}_t = \mathbf{K}\,\boldsymbol{\alpha}_t$ and $\|f_t\|_K^2 = \boldsymbol{\alpha}_t^T\,\mathbf{K}\,\boldsymbol{\alpha}_t$. After substituting these into (1.1), the optimization task with respect to the infinite-dimensional f_t simplifies to an equivalent optimization task with respect to the n-dimensional $\boldsymbol{\alpha}_t$, which can be solved by a slightly modified algorithm, as in JSLZ. It is evident that the kernel-based formulation in (1.1) is fairly similar to the original linear model of JSLZ, while admitting flexible model structures of f_t, thus allowing for general covariate effects on the ITRs.

We now examine the numerical performance of the kernel-based nonparametric ITRs using the simulated example in Qi and Liu (2018), where $R = Q(\mathbf{x}) + T(\mathbf{x}, A) + \epsilon$, with $T(\mathbf{x}, A) = 3.8(0.8 - x_1^2 - x_2^2)A, Q(\mathbf{x}) = 1 + x_1 + x_2 + 2x_3 + 0.5x_4$, and $\epsilon \sim N(0, 1)$. We consider the Gaussian kernel, set the training sample size as 400 and the validation sample size as 200,000, and set the ridge parameter $\lambda_n = 0.001$. The experiment is repeated 100 times, and the averaged value function values are summarized in Table 1.

Clearly, the nonparametric ITR outperforms its linear counterpart in the simulated example with nonlinear decision boundaries. In practice, as pointed out in Qi and Liu (2018), the selection of the kernel function can be regarded as a tuning parameter selection problem, with the optimal function being determined using some data-adaptive selection criterion.

2. Inference for Nonparametric ITRs

Few studies examine inferences related to machine-learning-based methods, partly because of their "parameter-free" frameworks. A similar concern is raised in JSLZ, although their inference results are still developed for linear ITRs. In fact, recent attempts have been made to develop inference tools for kernel-based approaches. For example, Jiang, Zhang and Cai (2008) provides an inference

for the prediction error of a kernel-based support vector machine, and Zhao et al. (2015) conducts an inference for kernel-based approaches to estimate the dynamic treatment regimes. The key idea is to utilize resampling techniques to draw inferences on a criterion about the prediction errors of the machine-learning-based methods. Similar treatments can be extended to inferences for the predicated value function of the nonparametric ITRs.

Given the training sample $\mathcal{O}_n = \{(R_{1i}, A_{1i}, \mathbf{x}_i^1, \ldots, R_{Ti}, A_{Ti}, \mathbf{x}_i^T)\}_{i=1}^n$, the estimated optimal decision functions $\widehat{\mathbf{f}} = (\widehat{f}_1, \ldots, \widehat{f}_T)^T$ can be obtained as in Section 1. Then, for a new observation $\mathcal{O}_0 = (R_{10}, A_{10}, \mathbf{x}_0^1, \ldots, R_{T0}, A_{T0}, \mathbf{x}_0^T)$, we consider the predicted value function $V_t^0(\widehat{\mathbf{f}}_t)$, with $\widehat{\mathbf{f}}_t = (\widehat{f}_t, \ldots, \widehat{f}_T)$, and its estimate $\widehat{V}_t(\widehat{\mathbf{f}}_t, \mathcal{O}_n)$. Here, $V_t^0(\widehat{\mathbf{f}}_t)$ and $\widehat{V}_t(\widehat{\mathbf{f}}_t, \mathcal{O}_n)$ are defined as in JSLZ. We then randomly split $\mathcal{O}_n$ into K disjoint subsets $O_n^{(1)}, \ldots, O_n^{(K)}$ of equal size. For each k, we use all observations not in $O_n^{(k)}$ to obtain $\widehat{\mathbf{f}}^{(-k)} = (\widehat{f}_1^{(-k)}, \ldots, \widehat{f}_T^{(-k)})^T$, as in Section 1, and use $O_n^{(k)}$ to compute the cross-validated value function. The procedure is repeated for $k = 1, \ldots, K$, and the final cross-validated value function is

$$\widehat{\mathcal{V}}_{t,n}^{CV} = \frac{1}{K} \sum_{k=1}^{K} \widehat{V}_t(\widehat{\mathbf{f}}^{(-k)}, \mathcal{O}_n^{(k)}).$$

As shown in Jiang, Zhang and Cai (2008), the asymptotic distribution of $\sqrt{n}\big(\widehat{\mathcal{V}}_{t,n}^{CV} - V_t^0(\widehat{\mathbf{f}}_t)\big)$ is the same as that of $\sqrt{n}\big(\widehat{V}_t(\widehat{\mathbf{f}}_t, \mathcal{O}_n) - V_t^0(\widehat{\mathbf{f}}_t)\big)$.

To approximate the distribution of $\sqrt{n}\big(\widehat{\mathcal{V}}_{t,n}^{CV} - V_t^0(\widehat{\mathbf{f}}_t)\big)$, we consider a perturbed version of (1.1), such that

$$\widetilde{f}_t = \underset{f_t \in \mathcal{H}_K}{\operatorname{argmin}} -\frac{1}{n} \mathbf{G}_t \odot \Big(\boldsymbol{\omega}_t \odot \big(0.5(\mathbf{A}_t + \mathbf{1}_n) \odot \mathbf{f}_t + \ln \boldsymbol{\xi}_t\big)\Big)^T \mathbf{1}_n + \lambda_n \|f_t\|_K^2, \quad (2.1)$$

where $\mathbf{G}_t = (G_{t1}, \ldots, G_{tn})^T$ is drawn from an exponential distribution with unit mean and variance. By sequentially solving (2.1), we obtain $\widetilde{\mathbf{f}} = (\widetilde{f}_1, \ldots, \widetilde{f}_T)^T$. Specifically, for stage t, we calculate

$$\widetilde{W}_t = n^{-1/2} \sum_{i=1}^{n} \big(\widehat{V}_t(\widetilde{\mathbf{f}}_t, o_i) - \widehat{V}_t(\widehat{\mathbf{f}}_t, \mathcal{O}_n)\big) G_{ti}, \quad (2.2)$$

where o_i denotes the ith sample of $\mathcal{O}_n$. Note that, given $\mathcal{O}_n$, the only random variable in (2.1) is G_{ti}. More importantly, the computed $\widetilde{W}_t$ in (2.2) can be regarded as a realization of a random variable whose distribution can approximate the distribution distribution of $\sqrt{n}\big(\widehat{\mathcal{V}}_{t,n}^{CV} - V_t^0(\widehat{\mathbf{f}}_t)\big)$ very well, given $\mathcal{O}_n$. Thus in practice, we generate $\{G_{ti}\}_{i=1}^n$ repeatedly M times, and obtain a large number of realizations $\widetilde{\mathbf{W}}_t = \{\widetilde{W}_{tm}\}_{m=1}^M$ to approximate the distribution of $\sqrt{n}\big(\widehat{\mathcal{V}}_{t,n}^{CV} -$

$V_t^0(\widehat{\mathbf{f}}_t)\big)$. Therefore, the confidence interval for the prediction value function in stage t can be obtained based on the empirical distribution of $\widetilde{\mathbf{W}}_t$.

We now construct an approximate confidence inference for the predicted value function of the nonparametric ITRs in the simulated example in Section 1, with sample size 500 and five-fold cross-validation. We first calculate the true prediction value function by repeatedly generating $\mathcal{O}_n$ independently 1,000 times. Then, for each $\mathcal{O}_n$, we calculate the cross-validated value function $\widehat{\mathcal{V}}_n^{CV}$. Therefore, the true value function can be computed as the average of $\widehat{\mathcal{V}}_n^{CV}$. To obtain the interval estimators, we generate $\mathcal{O}_n$ independently 100 times. For each $\mathcal{O}_n$, we compute $\widehat{\mathcal{V}}_n^{CV}$, generate $\mathbf{G}$ repeatedly to obtain 250 realizations of $\widetilde{W}$, and compute the estimated 95% confidence interval. This leads to a 94% coverage rate, which is comparable to the reported coverage rates in JSLZ for parametric ITRs, and may be improved upon with further computational efforts.

3. Concluding Remarks

We appreciate the opportunity to contribute to the discussion on this excellent paper. JSLZ provide proper statistical inferences for machine-learning-based methods when estimating ITRs, and leave numerous open questions for further research. For example, it is of great interest to investigate the statistical inferences for the kernel-based nonparametric ITRs theoretically, which enjoy model flexibility and can be adjusted based on prior information. We would like to congratulate JSLZ again on their enlightening work, and look forward to seeing similar future research.

References

Jiang, B., Zhang, X. and Cai, T. (2008). Estimating the confidence interval for prediction errors of support vector machine classifiers. *Journal of Machine Learning Research* **9**, 521–540.

Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33**, 82–95.

Qi, Z. and Liu, Y. (2018). D-learning to estimate optimal individual treatment rules. *Electronic Journal of Statistics* **12**, 3601–3638.

Shen, X., Tseng, G., Zhang, X. and Wong, W. (2003). On psi-learning. *Journal of the American Statistical Association* **98**, 724–734.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory* **51**, 128–142.

Wang, J. and Shen, X. (2007). Large margin semi-supervised learning. *Journal of Machine Learning Research* **8**, 1867–1891.

Wang, J. and Shen, X. and Liu, Y. (2008). Probability estimation for large-margin classifiers.

Biometrika **95**, 149–167.

Zhao, Y., Zeng, D., Laber, E. and Kosorok, M. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* **110**, 583–598.

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: he.xin17@mail.shufe.edu.cn

School of Data Science, City University of Hong Kong, Hong Kong, China.

E-mail: shirongxu2-c@my.cityu.edu.hk

School of Data Science, City University of Hong Kong, Hong Kong, China.

E-mail: j.h.wang@cityu.edu.hk

(Received January 2019; accepted February 2019)

DISCUSSION

Min Qian and Bin Cheng

Columbia University

We would like to congratulate Professors Jiang, Song, Li, and Zeng (JSLZ) on their stimulating article on dynamic treatment regimes (DTR), in which they make an interesting connection between the entropy loss and the optimal DTR. We found the article enjoyable to read, and we thank the editors for the opportunity to discuss it.

DTRs employ treatment decision rules that can be used to tailor a treatment based on a patient's needs over time. Current methods for estimating DTRs can be classified into two branches: the indirect approach (e.g., Q-learning; see Murphy (2005)), and the direct approach. The direct approach requires that we deal with a nonconvex optimization problem, owing to the existence of an indicator loss, and a surrogate loss is often used (e.g., the hinge loss used in Zhao et al. (2015)). JSLZ proposed replacing the indicator loss with a smooth surrogate entropy loss, and obtained asymptotic normality results for the estimated parameters and value functions for inferences. Below, we first discuss the inference problem and the conditions. Then, we examine the problem from a risk bound point of view.

Inferences are critical in DTRs, because they help researchers to decide on

the best treatment for each patient with a measure of confidence. However, it is challenging to make inferences when the data present around the decision boundary (Robins (2004); Laber et al. (2014)). In a linear decision boundary setting, following JSLZ's notation, this means that $|\boldsymbol{X}_t^{*T}\beta_t^0|$ has a nonnegligible probability mass around zero. Indeed, the asymptotic normality results in JSLZ rely on a *low-noise condition*, namely that $|\boldsymbol{X}_t^{*T}\beta_t^0|$ is bounded away from zero in probability (Assumption A3). The same problem occurs in the (indirect) Q-learning setting. Laber et al. (2014) showed that the parameters are asymptotically normal when $|\boldsymbol{X}_t^{*T}\beta_t^0|$ is bounded away from zero, and nonnormal otherwise; an adaptive procedure was proposed to solve this problem. From a treatment decision point of view, for a patient with $\boldsymbol{X}_t^* = \boldsymbol{x}_t^*$, because the treatment decision is based on the sign of $\boldsymbol{x}_t^{*T}\beta_t^0$, it is essential to test whether $\boldsymbol{x}_t^{*T}\beta_t^0 = 0$. Thus, the behavior of $\boldsymbol{X}_t^{*T}\hat{\beta}_t$ around zero is of great interest. As such, we wish to address the nonregularity issue in the entropy learning framework.

Interestingly, the low-noise condition is also related to the convergence rate, in terms of the risk bounds. Below, we establish two risk bounds for the entropy loss function, following Bartlett, Jordan and McAuliffe (2006). We demonstrate these bounds in the single-stage decision setting. However, the results for the multi-stage setting are similar.

Let $\boldsymbol{X}$ be a random vector containing patient pre-treatment variables, $A \in \{-1, 1\}$ be the treatment assignment, and R be a positive scalar outcome that is bounded from above. Let $\pi(\boldsymbol{X}) \triangleq P(A = 1|\boldsymbol{X})$ denote the known treatment randomization probability. The value function for a treatment decision rule $\mathcal{D} : \mathcal{X} \to \{-1, 1\}$, namely $V(\mathcal{D})$, is defined as the expected outcome if the study population follows the decision rule. The goal is to estimate the optimal decision rule $\mathcal{D}^{opt}$ that maximizes $V(\mathcal{D})$. It is easy to see that

$$V(\mathcal{D}) = \mathbb{E}\left[\frac{R\, I(A = \mathcal{D}(\boldsymbol{X}))}{(A\pi(\boldsymbol{X}) + (1 - A)/2)}\right].$$

Thus, maximizing $V(\mathcal{D})$ is equivalent to minimizing $\mathbb{E}[R\, I(A \neq \mathcal{D}(\boldsymbol{X}))/(A\pi(\boldsymbol{X}) + (1-A)/2)]$. JSLZ proposed replacing the indicator loss $I(A \neq \mathcal{D}(\boldsymbol{X}))$ with a surrogate entropy loss $h : \{-1, 1\} \times \mathbb{R} \to \mathbb{R}^+$, defined as $h(a, y) = -(a + 1)y/2 + \log(1 + e^y)$. Define

$$\mathcal{R}_h(f) = \mathbb{E}\left[\frac{R\, h(A, f(\boldsymbol{X}))}{(A\pi(\boldsymbol{X}) + (1 - A)/2)}\right].$$

Minimizing $\mathcal{R}_h(f)$ yields $f^{opt}(\boldsymbol{x}) = \arg\min_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_h(f) = \log(\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}, A = 1)/\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}, A = -1))$. It can be shown that $\mathcal{D}^{opt}(\boldsymbol{X}) = sign(f^{opt}(\boldsymbol{X}))$.

The following theorem connects the excess value, $V(\mathcal{D}^{opt}) - V(\mathcal{D})$, to the excess entropy risk, $\mathcal{R}_h(f) - \mathcal{R}_h(f^{opt})$. The proof is similar to that of Bartlett, Jordan and McAuliffe (2006), and thus is omitted.

Theorem 3. *Suppose R is positive and bounded from above by a constant $B > 0$. Then, for any $f : \mathcal{X} \to \mathbb{R}$ and $\mathcal{D} : \mathcal{X} \to \{-1, 1\}$, such that $\mathcal{D}(\boldsymbol{X}) = sign(f(\boldsymbol{X}))$, we have*

$$\psi\left(V(\mathcal{D}^{opt}) - V(\mathcal{D})\right) \leq \mathcal{R}_h(f) - \mathcal{R}_h(f^{opt}), \tag{1.1}$$

where $\psi : \mathbb{R}^+ \to \mathbb{R}$ is defined as

$$\psi(\theta) \triangleq (\theta + 2B)\log\left(\frac{2B}{\theta + 2B}\right) + (\theta + B)\log\left(\frac{\theta + B}{B}\right).$$

Furthermore, if there exists $\beta > 0$ and $c > 0$ such that, for all $\epsilon > 0$,

$$P\left(0 < |\mathbb{E}(Y|\boldsymbol{X}, A = 1) - \mathbb{E}(Y|\boldsymbol{X}, A = -1)| < \epsilon\right) \leq c\epsilon^{\beta}, \tag{1.2}$$

then we have

$$c'\left\{V(\mathcal{D}^{opt}) - V(\mathcal{D})\right\}^{\beta/1+\beta} \psi\left\{\frac{(V(\mathcal{D}^{opt}) - V(\mathcal{D}))^{1/(1+\beta)}}{2c'}\right\} \leq \mathcal{R}_h(f) - \mathcal{R}_h(f^{opt}), \tag{1.3}$$

for some $c' > 0$.

The risk bounds provide a way to evaluate the performance of the estimated decision rules. This type of result has been provided in Qian and Murphy (2011) for indirect learning, and in Zhao et al. (2012, 2015) for direct learning methods. The left-hand side of risk bounds (1.1) and (1.3) characterize the distance between the estimated decision rule and the optimal decision rule in terms of value. The right-hand side, $\mathcal{R}_h(f) - \mathcal{R}_h(f^{opt})$, describes the asymptotic behavior of the entropy risk. To see that, we replace f and $\mathcal{D}$ in the above theorem with the estimates $\hat{f}(\boldsymbol{X}) \triangleq \boldsymbol{X}^{*T}\hat{\beta}$ and $\hat{\mathcal{D}}(\boldsymbol{X}) \triangleq sign(\boldsymbol{X}^{*T}\hat{\beta})$, respectively, where $\boldsymbol{X}^* = (1, \boldsymbol{X}^T)^T$, and $\hat{\beta}$ is obtained by minimizing the empirical entropy risk. Then, $\mathcal{R}_h(\hat{f}) - \mathcal{R}_h(f^{opt})$ can be decomposed as

$$\mathcal{R}_h(\hat{f}) - \mathcal{R}_h(f^{opt}) = [\mathcal{R}_h(\hat{f}) - \mathcal{R}_h(f^*)] + [\mathcal{R}_h(f^*) - \mathcal{R}_h(f^{opt})], \tag{1.4}$$

where $f^*(\boldsymbol{X}) \triangleq \boldsymbol{X}^{*T}\beta^*$ minimizes the entropy risk $\mathcal{R}_h(f)$ in the linear decision space. The second term in (1.4), $\mathcal{R}_h(f^*) - \mathcal{R}_h(f^{opt})$, is the approximation error, which measures the distance between the model and the truth. The first term, $\mathcal{R}_h(\hat{f}) - \mathcal{R}_h(f^*)$, is the estimation error. Using Taylor's expansion, we can verify that $\mathcal{R}_h(\hat{f}) - \mathcal{R}_h(f^*) = O((\hat{\beta} - \beta^*)^2)$, which is $O_p(n^{-1})$, as shown in JSLZ.

Owing to the convexity of $\psi(\cdot)$, it is easy to verify that the risk bound in (1.3) always gives an equivalent or better rate than that in (1.1). The low-

noise condition (1.2) plays a critical role here. Note that (1.2) is a variant of Assumption A3 in JSLZ. Intuitively, when it is less likely to have point mass around the decision boundary, we would expect to learn the optimal decision rule more quickly and thus, experience a faster rate of convergence.

In summary, when a nonnegligible noise presents around the decision boundary (i.e., the low-noise condition is violated), there are difficulties in both learning the optimal decision rules and making statistical inferences under the null for various direct and indirect learning methods. An interesting research direction in this area would be to combine the inference with machine learning in order to improve the learning efficiency at the decision boundary.

Acknowledgements

The research is supported in part by NIH Grant R21MH108999.

References

Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association* **101**, 138–156.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E. and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics* **8**, 1225.

Murphy, S. A. (2005). A generalization error for Q-learning. *Journal of Machine Learning Research* **6**, 1073–1097.

Qian, M. and Murphy, S. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics* **39**, 1180–1210.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In: *Proceedings of the Second Seattle Symposium on Biostatitics*, 189–326. Springer.

Zhao, Y., Zeng, D., Laber, E. B. and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* **110**, 583–598.

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.

Department of Biostatistics, Columbia University, 722 West 168th Street, New York City, NY 10032, USA.

E-mail: mq2158@cumc.columbia.edu

Department of Biostatistics, Columbia University, 722 West 168th Street, New York City, NY 10032, USA.

E-mail: bc2159@cumc.columbia.edu

(Received January 2019; accepted February 2019)

DISCUSSION

Hongxiang Qiu[1], Alex Luedtke[1] and Mark van der Laan[2]

[1] *University of Washington and* [2] *University of California at Berkeley*

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

1. Introduction

We congratulate the authors on their innovative method for estimating dynamic treatment regimes (DTRs) (Jiang et al. (2019)). They introduced the entropy learning (E-learning) framework, which circumvents the need to model the conditional mean outcome directly given the covariates, when estimating an optimal DTR. Their method extended the work of Zhao et al. (2012, 2015) and Rubin and van der Laan (2012) by using a smooth surrogate loss function enabling them to obtain valid statistical inferences about the parameters in the DTR, as well as related quantities. In this discussion, we extend their work to consider model misspecification, the estimation of more flexible DTRs, and the treatment cost in the hypothesis test of no treatment effect in order to circumvent an unpleasant regularity assumption.

Our discussion is organized as follows.

1. We point out two consequences of restricting our attention to a linear class of candidate DTRs when an optimal DTR over an unconstrained class does not belong to this class:
   (a) In general, the infinite-sample limit of the proposed E-learning estimator depends on the treatment assignment probabilities.
   (b) In general, the estimated optimal value is inconsistent for the value under the optimal linear DTR, that is, the maximal mean reward attainable under a linear DTR.
2. We study the estimation of an optimal DTR over an unrestricted class using the loss function proposed by the authors. We show the following:
   (a) The unconstrained true-risk minimizer is the conditional log "relative reward" (RR).

(b) We can estimate the conditional log RR well by optimizing over an essentially unrestricted class, where here, and throughout, we use "essentially unrestricted" to refer to a class $\mathcal{F}_M$ of cádlág functions, with a variation norm bounded by a given $M < \infty$ (van der Laan (2017); Benkeser and Van Der Laan (2016)).

(c) We provide theoretical guarantees under which the value of the estimated DTR, based on estimating the conditional log RR over an essentially unrestricted class, converges to the optimal value at a fast rate.

3. We discuss the conditions that required to apply the test of the null of no individual-level stage-τ treatment effect, as proposed by the authors. Importantly, note that the validity of the proposed test relies on the null of no treatment effect not holding at any future stage $t > \tau$. This requirement seems concerning because, if the null of no effect at time τ is plausible, then it would seem that the null at times $t > \tau$ may also be plausible. Note that introducing a treatment cost to the clinical decision could help mitigate this concern.

2. Consequences of Misspecification of the Linear Model

2.1. Dependence of the infinite-sample limit of the E-learning estimator on the treatment assignment probabilities

Recall that β_t^0 indexes the linear DTR that minimizes the population-level E-learning risk, which represents the infinite-sample limit of the estimated linear decision rule parameters $\hat{\beta}_t$. In this section, we show that, in general, β_t^0 depends on the treatment mechanisms, that is, the probability of receiving a given treatment at each stage, given past covariates. This dependence is of more than academic interest — indeed, it can lead to counterintuitive results in real applications of the proposed method. For example, suppose that two clinical trials are run on the same population but with different treatment assignment mechanisms. In this case, the optimal linear decision rules in the two trials can differ substantially, even if the sample sizes are very large.

Momentarily, we will provide a simple example of such a discrepancy between the estimands, in two settings. Before doing so, we provide a brief analytical argument showing why this dependence of β_t^0 on the treatment mechanism should be expected. Recall that the authors consider the DTR to be determined by a

linear function, namely, $x_t \mapsto x_t^{*\top}\beta_t$, where, for any stage-$t$ covariate x_t, $x_t^* \equiv (1, x_t)$. In particular, the rule recommended by the DTR (-1 or 1) is determined by the sign of $x_t^{*\top}\beta_t$. In this case, the authors showed that $\hat{\beta}_t$ converges to the population-level minimizer of the E-learning risk; that is,

$$\beta_t^0(\pi) = \underset{\beta_t \in \mathbb{R}^{p_t+1}}{\operatorname{argmin}} \mathbb{E}\left[\frac{(\sum_{j=t}^T R_j)\prod_{j=t+1}^T \mathbb{1}\{A_j = \operatorname{sgn}(X_j^{*\top}\beta_j^0)\}}{\prod_{j=t}^T \pi(A_j, S_j)} h(A_t, X_t^{*\top}\beta_t)\right], \tag{2.1}$$

which is defined by iterating backwards through times $t = T, T-1, \ldots, 1$, where $h(a, y) = -(a+1)y + 2\log(1+\exp(y))$, and $\beta_t^0(\pi)$ emphasizes the (potential) dependence of β_t^0 on the treatment assignment probabilities π.

The authors also considered the case when the linearity assumption is not true, that is, when the population-level minimizer of their risk over an unrestricted class is nonlinear; in Section 3.1, we provide a familiar interpretation for this minimizer. When linearity does not hold, the authors note that β_t^0 should be understood as the best approximation of the true population-level minimizer in the collection of linear rules, namely, $\{x_t \mapsto x_t^{*\top}\beta_t : \beta_t\}$. We now argue that β_t^0 depends on the treatment assignment mechanism when the linearity assumption is not true. First note that the risk function at stage T can be expressed as follows:

$$\mathbb{E}\left[\frac{R_T}{\pi(A_T, S_T)} h(A_T, X_T^{*\top}\beta_T)\right] = \mathbb{E}\left\{\mathbb{E}\left[R_T h(A_T, X_T^{*\top}\beta_T)\middle| S_T\right]\right\}.$$

Note too that the treatments at previous stages are contained in the history S_T. Thus the previous treatment assignment mechanism $\pi(A_j, S_j)$, for $j < T$, influences the marginal distribution of S_T and, hence, could influence β_T^0. At any stage $t < T$, there is a similar potential for β_t^0 to depend on the treatment mechanisms at all previous stages $j < t$. Moreover, the term $\prod_{j=t+1}^T \mathbb{1}\{A_j = \operatorname{sgn}(X_j^{*\top}\beta_j^0)\}$ in (2.1) allows β_t^0 to depend on the decision rules β_j^0 at all future stages $j > t$. Therefore, β_t^0 depends on the treatment assignment mechanisms at the current stage and future stages $\pi(A_j, S_j)$, for $t \le j < T$. By this argument, we can show that, for all t, β_t^0 can depend on $\pi(A_j, S_j)$, for all $j = 1, \ldots, T-1$. Consequently, collecting two data sets from the same population, but with different treatment assignment probabilities, can lead to different infinite-sample limits for the E-learning estimators used in the two settings.

We use a simple two-stage example to illustrate how this dependence on the treatment mechanism can affect the interpretation of the study results. We consider two data-generating mechanisms, which are identical in all ways except

Table 1. Population-level parameters β_t^0 indexing an optimal DTR at stage t, β_t^0, in a two-stage example with different treatment assignment mechanisms. These parameter values were obtained via a Monte Carlo approximation with sample size 5×10^6. Note that these parameters—particularly the slopes—are markedly different in the two scenarios.

Setting	Treatment assignment mechanism	First stage, β_1^0		Second stage, β_2^0	
		Intercept, β_{10}^0	Slope, β_{11}^0	Intercept, β_{20}^0	Slope, β_{21}^0
1	$\pi^{(1)}$	0.69	0.00	1.50	0.00
2	$\pi^{(2)}$	0.28	-2.53	0.79	-0.88

for their treatment mechanisms. We denote the treatment mechanisms in the two settings by $\pi^{(1)}$ and $\pi^{(2)}$, respectively. We show that the coefficients in (2.1) vary between the two scenarios. Specifically, we show that $\beta_1^0(\pi^{(1)}) \neq \beta_1^0(\pi^{(2)})$ and $\beta_2^0(\pi^{(1)}) \neq \beta_2^0(\pi^{(2)})$. In both examples, $S_1 = X_1$ follows a standard normal distribution, and $X_2|A_1 = a_1$ and $X_1 = x_1$ follow a normal distribution with mean a_1x_1 and variance one. We consider a setting where the investigator is only interested in maximizing the final reward, such that $R_1 = 0$ and $R = R_2$. The outcome regression is given by $\mathbb{E}[R|S_2 = s_2, A_2 = a_2] = \mathbb{1}\{a_2 = 1\}[2x_1^2\,\mathbb{1}\{a_1 = 1\} + \mathbb{1}\{a_1 = -1\} + 2x_2^2] + \mathbb{1}\{a_2 = -1\}$. We let $\pi_t^{(k)}$ denote $P(A_t = 1|S_t)$ in each scenario k. In the first scenario, we let $\pi_1^{(1)} = \pi_2^{(1)} = 0.5$. In the second scenario, we let $\pi_1^{(2)} = 0.9$ when $X_1 < 0.5$ and $\pi_1^{(2)} = 0.1$ when $X_1 > 0.5$. Similarly, $\pi_2^{(2)} = 0.9$ when $X_2 < 0.5$ and $\pi_2^{(2)} = 0.1$ when $X_2 > 0.5$.

Table 1 presents β_t^0 for the two scenarios in this example where only the treatment assignment mechanisms vary. We can clearly see that β_t^0 depends on the treatment assignment mechanism. Suppose these two β_t^0 parameters are estimated from two large clinical trials that are identical in all aspects, except for their treatment assignment mechanisms. On the one hand, based on the results from the first trial, because $\beta_{21}^0(\pi^{(1)})$ and $\beta_{11}^0(\pi^{(1)})$ are very close to zero, policymakers might conclude that the two treatments have very similar effects. On the other hand, based on the results from the second trial, because $\beta_{21}^0(\pi^{(2)}) < 0$ and $\beta_{11}^0(\pi^{(2)}) < 0$, policymakers might conclude that the two treatments have different effects for different people. Consequently, they might discourage practitioners from collecting the variables X_1, X_2 on future patients, based on the results from the first trial, but might encourage them to do so and use a linear DTR, based on the results from the second trial.

2.2. Inconsistency of the estimated optimal value

Note that although the asymptotic normality of $\hat{\beta}_t$ for β_t^0 can be shown to

hold, even when the true E-learning risk minimizer is nonlinear, a similar result cannot be established for the proposed estimator of the optimal value. In fact, the estimator $\hat{V}_t$ may not even be consistent for $V_t^* \equiv \max_{\beta_t \in \mathbb{R}^{p_t+1}} V_t(\beta_t)$ in this case, which is the optimal value that can be possibly obtained from a linear DTR. This possible inconsistency arises because the surrogate loss used to obtain the decision rules differs from the zero-one loss used to define the optimal value. When the restricted class $\mathcal{F}$ of DTRs does not contain an optimal DTR over an unrestricted class, the DTR that minimizes the population-level surrogate risk over $\mathcal{F}$ may differ from the DTR that maximizes the optimal value over $\mathcal{F}$. Therefore, the value of the estimated DTR need not converge to V_t^*.

We illustrate this possible inconsistency of $\hat{V}_t$ for V_t^* using a single-stage scenario. To simplify the notation, throughout this example, we omit the stage index t. The data are generated as follows: $X \sim \text{Unif}(-1, 1)$, $P(A = 1|X) = 0.5$, $\mathbb{E}[R|A = -1, X = x] = 1$, and $\mathbb{E}[R|A = 1, X = x] = 2x^2$. The population-level E-learning coefficients β^0 maximize the following surrogate for the value function in $\beta = (\beta_0, \beta_1)$:

$$-R(\beta) = \mathbb{E}\left[\frac{R[0.5(A+1)(\beta_0 + \beta_1 X) - \log(1 + \exp(\beta_0 + \beta_1 X))]}{A\pi + (1-A)/2}\right].$$

This quantity differs from the value function,

$$V(\beta) = \mathbb{E}\left[\frac{R\,\mathbb{1}\{A = \text{sgn}(\beta_0 + \beta_1 X)\}}{A\pi + (1-A)/2}\right]. \tag{2.2}$$

We denote the maximizer of V by $\beta^\dagger$. Note that because the value function is nonconcave, finding $\beta^\dagger$ in our numerical example is challenging. Therefore we instead use $\beta^\dagger$ to denote any near maximizer of this function.

As can be seen in Table 2, the value of $\beta^\dagger$ is strictly larger than the value of β^0 in this example. Given that the value of $\beta^\dagger$ is a lower bound on the maximum V^* of (2.2), this fact does not impact our conclusion that $V(\beta^0) < V^*$.

It can be shown that the estimator of the optimal value proposed by the authors $\hat{V}$ is consistent for $V(\beta_0)$. Hence it is inconsistent for the optimal value that can be obtained from a linear decision rule V^*.

Returning now to the general case, note that although $\hat{V}_t$ may be inconsistent for the optimal value V_t^* among the class of linear decision rules, this quantity is always a conservative estimator of the true optimal value, in the sense that

$$V_t(\beta_t^0) \leq \max_{\beta_t \in \mathbb{R}^{p_t+1}} V_t(\beta_t) \equiv V_t^*, \tag{2.3}$$

Refer to the definition of V_t above Eq. 2.10 in the paper under discussion. Hence, $\hat{V}_t$ provides information about whether it is worth advocating a wide application

Table 2. Two linear DTRs and their optimal values. β^0 is the "true linear DTR" for which the estimated DTR using the surrogate loss is consistent and minimizes the population-level surrogate risk. $\beta^\dagger$ is a linear DTR that nearly maximizes the value. Note that $V(\beta^\dagger) > V(\beta^0)$.

Parameter indexing the DTR, β	Value, $V(\beta)$
$\beta^0 = (-0.41, 0.00)$	1.00
$\beta^\dagger = (-2.52, 3.55)$	1.07

of a DTR in a given setting: if $\hat{V}_t$ were very large compared with $V_t(D_{t,\text{current}})$ for the current standard decision rule at stage t, $D_{t,\text{current}}$, then we would be confident of benefiting from implementing the DTR. Furthermore, from (2.3), a $(1-\alpha)$-level confidence lower bound for the limit $V_t(\beta_t^0)$ of $\hat{V}_t$ is also a valid $(1-\alpha)$-level lower confidence bound for V_t^*. Therefore, even if the optimal value V_t^* is of interest, rather than the value of the rule indexed by β_t^0, it is still useful to obtain a valid confidence lower bound for $V_t(\beta_t^0)$ under misspecification.

A natural question that arises is the following: is it possible to derive the asymptotic normality of $\hat{V}_t$ as an estimator of $V_t(\beta_t^0)$ under regularity conditions, thus leading to a valid inference?

3. Nonparametric Decision Rules

3.1. Unconstrained true-risk minimizer

The loss function proposed by the authors yields (to the best of our knowledge) a novel approach to robustly estimating the counterfactual log relative risk. Consider the single-stage setting, with the population-level E-learning risk

$$R(f) = \mathbb{E}\left[\frac{R[-0.5(A+1)f(X) + \log(1+\exp(f(X)))]}{A\pi + (1-A)/2}\right]. \tag{3.1}$$

Our goal is to minimize this risk, where the form of f is left unrestricted. In this case, the function f^0 that minimizes this quantity is the conditional *log relative reward*:

$$f^0(x) = \log\left(\frac{\mathbb{E}[R|A=1, X=x]}{\mathbb{E}[R|A=-1, X=x]}\right). \tag{3.2}$$

This leads to a way of estimating the conditional relative risk (instead of reward) function nonparametrically, without estimating the conditional mean function $(a,x) \mapsto \mathbb{E}[R|A=a, X=x]$. First, let R denote an indicator of the occurrence of an event; next, minimize the risk in (3.1) over a large class of functions. We consider the relative risk instead of the relative reward here, because this is a

more common measure of effect size in epidemiology. This is similar to the result for the conditional average treatment effect (CATE). Inspired by Rubin and van der Laan (2007), Luedtke and van der Laan (2016c) showed that we can use least squares with pseudo outcomes $[\mathbb{1}\{A=1\}/\pi - \mathbb{1}\{A=-1\}/(1-\pi)]\,R$, or doubly robust variants thereof, to nonparametrically estimate the CATE.

A natural question that arises is the following: for any contrast of conditional means $\mathbb{E}[R|A=1,X]$ and $\mathbb{E}[R|A=-1,X]$ (e.g., odds ratio), is it possible to select a surrogate loss function h or, in general, a risk function R that allows us to estimate that conditional contrast function without estimating the conditional mean function? In DTRs, the conditional contrast is of interest. Because a correct specification of the conditional mean function implies correct specification of the conditional contrast function, it is never more difficult to correctly specify the conditional contrast than it is to correctly specify the conditional mean. In many cases, we expect that it will be easier. For example, when a test of treatment effect heterogeneity is conducted, the null hypothesis is often that there is no treatment effect. When there is no heterogeneity in the treatment effect, which is an apparently plausible scenario, given that this is often the null of interest, any contrast between the conditional means $\mathbb{E}[R|A=1,X]$ and $\mathbb{E}[R|A=-1,X]$ is constant. Therefore, to correctly specify this quantity, it suffices to use a learner that is able to learn a constant function. We note that all natural learners satisfy this property.

We conclude by noting that it is possible to estimate an optimal DTR based on the log relative risk, rather than using the log relative reward. Let $\hat{f}$ denote the estimated log relative risk above. The estimated DTR is then $x \mapsto -\text{sgn}\{\hat{f}(x)\}$, where $\hat{f}$ is the estimated conditional log relative risk function. One advantage of "reversing the reward" in this fashion is that, in many cases, the event is rare, and it is more common to model the relative risk for a rare event than it is to model the relative reward, where the reward is defined as the absence of the event. It may also be easier to compare $\hat{f}$ with results from other studies, especially case-control studies, where odds ratios are reported as an approximation of the relative risk.

3.2. Nonparametric estimator of the true-risk minimizer with a bounded total variation norm

A promising approach to flexibly estimating the conditional log RR is to minimize the empirical risk over the function class $\mathcal{F}_M$ of cádlág functions, with total variation norms bounded by some $M < \infty$. Similar approaches have been

applied successfully to least-squares and logistic losses for regressions. The approach used in these settings is termed the highly adaptive LASSO (HAL) (van der Laan (2017); Benkeser and Van Der Laan (2016)). Under certain conditions, owing to a bound on the uniform entropy of the class $\mathcal{F}_M$, these empirical risk minimizers based on a loss function L have been shown to have an $o_p(n^{-1/4})$ convergence rate, even when there are numerous covariates and discontinuities in the true function. We first introduce the notation for an empirical process. For a distribution $\mathbb{P}$ and a function g, $\mathbb{P}g \equiv \int g(o)d\mathbb{P}(o)$, and we use P to denote the true distribution from which we draw the observed data. From a high level, these conditions require that:

1. there is a uniform bound on L,
2. $f \mapsto P\{L(f) - L(f^0)\}$ is locally quadratic for $f \in \mathcal{F}_M$, where f^0 is the true function and L is the loss function,
3. the $L^2(P)$-distance between $L(f)$ and $L(f^0)$, $\left[P\{L(f) - L(f^0)\}^2\right]^{1/2}$, is bounded by $P\{L(f) - L(f^0)\}$.

Note that Condition 2 is similar to, but different from, Condition 3. Condition 2 describes the local behavior of the loss-based dissimilarity $P\{L(f) - L(f^0)\}$ between functions f and f^0, whereas Condition 3 shows how this dissimilarity upper bounds the $L^2(P)$-distance between the loss functions $L(f)$ and $L(f^0)$. Refer to Lemma 1 in van der Laan (2017) for further details.

Although the optimization over such a rich function class seems computationally intractable, the HAL approach can be readily implemented. As its name suggests, a HAL estimator can be computed using a LASSO regression. Because the authors' loss function and linearity assumption on the decision rule correspond to a weighted logistic regression, the corresponding HAL estimator can be computed using a weighted LASSO logistic regression, as follows:

$$\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n} \frac{R_i[-0.5(A_i+1)f_\beta(X_i) + \log(1+\exp(f_\beta(X_i)))]}{A_i\pi + (1-A_i)/2} \tag{3.3}$$

$$\text{subject to} \quad |\beta_0| + \sum_{s\subset\{1,\ldots,p\},s\neq\emptyset} \sum_{k=1}^{n} |\beta_{s,k}| \leq M, \tag{3.4}$$

where

$$f_\beta(x) = \beta_0 + \sum_{s\subset\{1,\ldots,p\},s\neq\emptyset} \sum_{k=1}^{n} \mathbb{1}(X_{k,s} \leq x_s)\beta_{s,k}. \tag{3.5}$$

Here we use the notation in Benkeser and Van Der Laan (2016): for a nonempty

index set s, x_s denotes the entries of $x \in \mathbb{R}^p$ that are in the index set s, and the $\leq$ in $\mathbb{1}(X_{k,s} \leq x_s)$ holds entrywise.

3.3. Guarantees on the value of an essentially unrestricted estimated optimal rule

In the single-stage setting, we can use a nonparametric estimator of the DTR to estimate the optimal value. Using the results in Section 7.5 of Luedtke and van der Laan (2016b), which are based on arguments given in Audibert and Tsybakov (2007), we can show that, under fairly weak conditions, if the $L^2(P)$-convergence rate of the estimated conditional log RR function $\hat{f}_n$ is r_n, that is, $\left[P\{\hat{f}_n - f^0\}^2\right]^{-1/2} = O_p(r_n)$, then the value of the DTR defined using the estimated log RR, $V(\hat{f}_n)$, converges to the true optimal value, $V(f^0) = \max_f V(f)$, at rate $O_p(r_n^{2(\alpha+1)/(\alpha+2)})$, where $\alpha > 0$ is a constant in the following margin condition:

$$\begin{aligned} &P\left(0 < \left|\mathbb{E}[R|A=1,X] - \mathbb{E}[R|A=-1,X]\right| \leq t\right) \\ &= P\left(0 < \mathbb{E}[R|A=-1,X]\left|\exp(f^0(X)) - 1\right| \leq t\right) \\ &\leq Ct^{\alpha}, \end{aligned} \tag{3.6}$$

for all t, where f^0 is defined in (3.2) and $C \geq 0$ is a constant. Under some conditions, the $L^2(P)$-convergence rate of the HAL estimator is $o_p(n^{-1/4})$. If we assume that the density of $\mathbb{E}[R|A=1,X] - \mathbb{E}[R|A=-1,X]$ is bounded near zero when X is drawn from the marginal distribution of the covariates, then we can take $\alpha = 1$, such that the optimal value for the estimated decision rule converges to the true optimal value at rate $o_p(n^{-1/3})$, regardless of the number of covariates used in the DTR when the HAL approach is used to estimate f^0.

Note that (3.6) can be viewed as a more general form of Condition A3 given in the paper under discussion, in two respects. First, (3.6) applies when the linearity assumption fails to hold. Second, (3.6) allows us to study the performance of the learned rule under a range of α-dependent margin conditions.

Finally , note that the nonparametric estimation for the decision rule can also be applied in a multistage setting. To learn a DTR using HAL, we can iterate backwards through stages $t = T, T-1, \ldots, 1$ to minimize the surrogate empirical risk in Eqs. 2.7 and 2.8 in the paper under discussion over functions similar to (3.5), subject to constraints similar to (3.4). The convergence rate of the estimated optimal value requires further investigation.

4. Nonregularity

In Section 3.3 of their paper, the authors present a test of the significance of the treatment effect at stage τ, for $1 \leq \tau \leq T$. Specifically, their proposed test relies on the result from their Theorem 1. That is, for a given stage-τ covariate x_τ, the following distributional convergence holds under the conditions of Theorem 1:

$$\sqrt{n} x_\tau^{*\top} [\hat{\beta}_\tau - \beta_\tau^0] \Rightarrow_d \mathrm{N}(0, x_\tau^{*\top} \Sigma_\tau(\beta_\tau^0) x_\tau^*). \tag{4.1}$$

Here, $x_\tau \in \mathbb{R}^{p_\tau}$, $x_\tau^* \equiv (1, x_\tau)$, and, for $\beta_\tau \in \mathbb{R}^{p_\tau+1}$, $\Sigma_\tau(\beta_\tau) \equiv I_\tau(\beta_\tau)^{-1} \Gamma_\tau I_\tau(\beta_\tau)^{-1}$ is a $(p_\tau + 1) \times (p_\tau + 1)$ matrix; refer to Condition A1 and Theorem 1 of the paper under discussion for the definitions of I_τ and Γ_τ, respectively. To test the null hypothesis $H_0(x_\tau) : x_\tau^{*\top} \beta_\tau^0 = 0$ against the complementary alternative, the authors proposed an α-level test that rejects the null hypothesis if $\sqrt{n} |(x_\tau^{*\top} \hat{\Sigma}_\tau(\hat{\beta}_\tau) x_\tau^*)^{-1/2} x_\tau^{*\top} \hat{\beta}_\tau|$ exceeds the $(1-\alpha/2)$-quantile of the standard normal distribution, where $\hat{\Sigma}_\tau(\cdot)$ is an estimate of $\Sigma_\tau(\cdot)$.

Note that (4.1) fails to hold in important scenarios that are of scientific interest. The simplest example occurs when $\beta_t = (0, 0, \ldots, 0)$, for some $t > \tau$. In this case, Condition A3 of Theorem 1 in the paper under discussion fails to hold; thus (4.1) is not implied by Theorem 1. The inability to establish (4.1) in this setting does not appear to be due to the requirement of a sufficient-but-not-necessary condition in the theorem statement. Indeed, Robins (2004) studies "exceptional laws" of this form in detail, arguing that a condition similar to Condition A3 is essentially necessary for a valid inference. See also Theorem 3.3 in Laber et al. (2014) and Theorem 1 in Luedtke and van der Laan (2016b) for related results. Exceptional laws lead to nonregular inferences and, thus, the failure of convergence results such as those in (4.1). Informally, exceptional laws arise when the optimal decision for an individual randomly drawn from the population is nonunique at some stage; that is, the same expected reward is attained for this individual, regardless of the treatment he or she receives.

Note that the validity of (4.1) actually relies on a condition that is slightly weaker than Condition A3 in the work under discussion. If Condition A3 were strictly required, then this would seem to pose a major problem for the authors' test of a treatment effect at x_τ. Specifically, Condition A3 requires that, with probability one, the stage-τ treatment effect is nonzero at the covariate X_τ, where X_τ is a random stage-τ covariate drawn from the distribution P that generated the data. Therefore, if the user knows in advance that Condition A3 is valid, then, given a random $X_\tau \sim P$ drawn independently of the data, a test that rejects

the null hypothesis $H_0(X_\tau)$ without considering the data will make the correct decision, with probability one, over the draw of $X_\tau \sim P$. Fortunately, a convergence result of the form given in (4.1) can hold under a weaker condition than Condition A3. Although this weaker condition would continue to require that Condition A3 holds for all $t > \tau$, it would not require that Condition A3 holds for $t = 1, \ldots, \tau$. This would allow the user to avoid assuming that $H_0(x_\tau)$ holds P-almost surely over x_τ in order to obtain a valid test of $H_0(x_\tau)$. Nonetheless, the user would still be required to assume that the optimal treatment decisions at all future stages are almost surely unique. Given that the purpose of the authors' proposed test is to test whether the optimal treatment for a given individual is unique at some stage—namely, stage τ—it seems problematic to make an *a priori* assumption that this individual's optimal treatment will be unique at all future stages.

A possible approach to mitigating this concern is to take the treatment cost into account when making the stage-τ treatment decision. Suppose that treatment 1 is more expensive than treatment -1. In this case, for a given patient, it is natural to test whether treatment 1 yields a sufficiently large additional reward γ_τ that it is worth applying this more expensive treatment. This can be formalized by testing the null hypothesis $H_0'(x_\tau) : x_\tau^{*\top}\beta_\tau^0 \leq \gamma_\tau$ against the complementary alternative. In this scenario, the uniqueness of the rule at each stage would be ensured by replacing each instance of $X_t^{*\top}\beta_t^0$ in Condition A3 by $(X_t^{*\top}\beta_t^0 - \gamma_t)$. Here γ_t is the threshold on $X_t^{*\top}\beta_t^0$ at which administering treatment 1 at time t becomes cost-effective; that is, it yields a clinical benefit, while still satisfying a given cost constraint. Unlike the authors' proposed test, which needs to assume that the alternative hypothesis holds at all future stages $t > \tau$, this modification of Condition A3 does not require the unpleasant assumption, that the expensive treatment is cost-effective at all future stages. This kind of cost-constrained or resource-limited setting has been studied previously by Luedtke and van der Laan (2016a), Toth and van der Laan (2018) and VanderWeele et al. (2018). Importantly, in the settings of these works, the standard errors for the summaries of the optimal DTR changed in these cost-constrained settings. This is because these works assume that γ_τ is not specified directly, but instead is specified through a constraint on the expected treatment cost, which, in turn, implies a threshold γ_τ that must be estimated from the data. We suspect that the standard errors of the estimators of the true E-learning risk minimizer would change similarly in this setting.

5. Conclusion

We close by again congratulating the authors on their important contribution to estimations and statistical inferences for optimal DTRs.

References

Audibert, J. Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics* **35**, 608–633.

Benkeser, D. and Van Der Laan, M. (2016). The highly adaptive Lasso estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 689-696. IEEE.

Jiang, B., Song, R., Li, J. and Zeng, D. (2019). Entropy learning for dynamic treatment regimes. *Statistica Sinica* **29**, 1633–1656.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E. and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics* **8**, 1225.

Luedtke, A. R. and van der Laan, M. J. (2016a). Optimal individualized treatments in resource-limited settings. *The International Journal of Biostatistics* **12**, 283–303.

Luedtke, A. R. and van der Laan, M. J. (2016b). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics* **44**, 713–742.

Luedtke, A. R. and van der Laan, M. J. (2016c). Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics* **12**, 305–332.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, 189-326. Springer.

Rubin, D. and van der Laan, M. J. (2007). A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics* **3**.

Rubin, D. B. and van der Laan, M. J. (2012). Statistical issues and limitations in personalized medicine research with clinical trials. *The International Journal of Biostatistics* **8**, 18.

Toth, B. and van der Laan, M. (2018). Targeted learning of optimal individualized treatment rules under cost constraints. In: Biopharmaceutical Applied Statistics Symposium. Springer, pp. 1–22.

van der Laan, M. (2017). A Generally Efficient Targeted Minimum Loss Based Estimator based on the Highly Adaptive Lasso. *The International Journal of Biostatistics* **13**.

VanderWeele, T. J., Luedtke, A. R., van der Laan, M. J. and Kessler, R. C. (2018). Selecting optimal subgroups for treatment using many covariates. *Epidemiology* (in Press), ArXiv Preprint ArXiv:1802.09642.

Zhao, Y., Zeng, D., Laber, E. B. and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* **110**, 583–598.

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.
E-mail: qiuhx@uw.edu
Department of Statistics, University of Washington, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center.
E-mail: aluedtke@uw.edu
Division of Biostatistics, University of California at Berkeley, Berkeley, CA 94720, USA.
E-mail: laan@berkeley.edu

(Received February 2019; accepted February 2019)

DISCUSSION

Stefan Wager

Stanford University

Policy Learning. The problem of policy learning (or learning optimal treatment regimes) has received considerable attention across several fields, including statistics, operations research, and economics. In its simplest setting, we observe a sequence of independent and identically distributed samples (X_i, A_i, R_i), where $X_i \in \mathbb{R}^p$ is a vector of features, $A_i \in \{-1, +1\}$ is a randomly assigned action, and $R_i \in \mathbb{R}$ is a reward. We then seek to learn a good decision rule $d : \mathbb{R}^p \to \{-1, +1\}$ that can be used to assign actions in the future. Following the Neyman–Rubin causal model (Imbens and Rubin (2015)), we assume potential outcomes $R_i(-1)$ and $R_i(+1)$, corresponding to the reward that the i-th subject would have experienced had it been assigned action -1 or $+1$ respectively, such that $R_i = R_i(A_i)$. We write the conditional average treatment effect as $\tau(x) = \mathbb{E}\left[R_i(+1) - R_i(-1) \,\middle|\, X_i = x\right]$.

Given this setting, the expected reward from deploying a decision rule d is $V(d) = \mathbb{E}\left[R_i(d(X_i))\right]$; we refer to this quantity as the value of d. Furthermore, assuming randomization such that $\{R_i(-1), R_i(+1)\} \perp\!\!\!\perp A_i$, we have (Kitagawa and Tetenov (2018); Qian and Murphy (2011))

$$V(d) = \mathbb{E}\left[\frac{1\left(\{A_i = d(X_i)\}\right) R_i}{A_i\pi + (1 - A_i)/2}\right], \quad \pi = \mathbb{P}\left[A_i = 1\right], \tag{1.1}$$

and it is natural to consider learning a decision rule $\hat{d}$ by maximizing an empirical estimate $\hat{V}(d)$ of $V(d)$ over a class $\mathcal{D}$ of candidate decision rules. Contributions to

this problem from the statistics literature, including those of Qian and Murphy (2011) and Zhao et al. (2012), are reviewed by Jiang et al. (2019); related results from neighboring fields, including extensions to observational studies and multi-action settings, are developed by Athey and Wager (2017), Dudík, Langford and Li (2011), Hirano and Porter (2009), Kallus (2018), Kallus and Zhou (2018), Kitagawa and Tetenov (2018), Manski (2004), Stoye (2009, 2012), Swaminathan and Joachims (2015), and Zhou, Athey and Wager (2018).

Jiang et al. (2019) present a thought-provoking approach to statistical inference for the policy learning problem. They start by observing that the empirical analogue $\hat{V}(d)$ of the objective function in (1.1) is discontinuous in the decision rule $d(\cdot)$, and so exact asymptotic analysis is complicated. To avoid this difficulty, they propose first solving a surrogate problem with a smooth loss function (they assume all rewards R to be positive),

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{A_i \pi + (1 - A_i)/2} \left(-\frac{(A_i + 1)}{2} f(X_i) + \log\left(1 + e^{f(X_i)}\right) \right) \right\}, \tag{1.2}$$

and then obtain a decision rule by thresholding $\hat{f}$, that is, a rule of the form $\hat{d}(x) = \operatorname{sign}(\hat{f}(x))$. Jiang et al. (2019) show that the loss function used in (1.2) is Fisher-consistent; this implies that that, under reasonable conditions and if the class $\mathcal{F}$ in (1.2) is unrestricted, then a regularized variant of their procedure is consistent for the maximizer of the value function $V(\cdot)$ in large samples. These results are also extended to the dynamic decision-making context.

Parametrizing Policy Learning. Relative to existing methods, the main advantage of the approach of Jiang et al. (2019) is that, because the "entropy loss" minimized in (1.2) is smooth, we can provide an exact characterization of the asymptotic behavior of $\hat{f}$ using classical second-order theory. Such results are particularly intriguing when we restrict ourselves to a parametric class $f(x) = c + x\beta$, as then we can use the results of Jiang et al. (2019) to quantify the uncertainty in the parameters $\hat{c}$ and $\hat{\beta}$ that underlie the learned decision rule $\hat{d}(x) = \operatorname{sign}(\hat{c} + x\hat{\beta})$.

Jiang et al. (2019) go further still, and propose using a regression table to summarize the uncertainty in $\hat{c}$ and $\hat{\beta}$; Table 1 shows an example based on a simple simulation study described at the end of this note. Looking at Table 1, we may feel inclined to cautiously conclude that the first feature X_1 matters for treatment personalization. Jiang et al. (2019) present a similar table to quantify

Table 1. Regression table for the optimal linear rules in the simulation design (1.5), following the entropy learning approach of Jiang et al. (2019). Here, $n = 4,000$, $p = 5$, standard errors are obtained via the bootstrap, and p-values less than 0.05 are indicated in bold.

	intercept	beta 1	beta 2	beta 3	beta 4	beta 5
point estimate	0.429	−0.177	0.031	−0.022	0.070	0.022
standard error	0.096	0.089	0.069	0.083	0.082	0.078
p-value	**0.000**	**0.047**	0.652	0.794	0.392	0.779

the value of personalized depression treatments in the context of the STAR*D study (Sinyor, Schaffer and Levitt (2010)), and argue that gender, age, and other features are significant in determining the best treatment options. Such regression tables have the potential to have a large impact on practice as they present information about optimal treatment rules in a familiar, easy-to-read format.

This regression table approach presents a marked departure from the standard approach to policy learning based on utilitarian regret (Manski (2004)). For example, using the latter approach, Athey and Wager (2017) and Kitagawa and Tetenov (2018) consider the case where the class $\mathcal{D}$ of allowable decision rules has a finite Vapnik–Chervonenkis dimension $\mathrm{VC}(\mathcal{D})$, and show that the policy $\hat{d}$ learned by empirical maximization of the objective (1.1) over $\mathcal{D}$ has regret bounded by

$$\mathcal{R}\left(\hat{d}\right) = \mathcal{O}_P\left(\sqrt{\frac{\mathrm{VC}\left(\mathcal{D}\right)}{n}}\right), \quad \mathcal{R}\left(d\right) = \sup\left\{V(d') : d' \in \mathcal{D}\right\} - V\left(d\right). \tag{1.3}$$

The key distinction between the results of Jiang et al. (2019) that underlie their regression tables and those of Athey and Wager (2017) and Kitagawa and Tetenov (2018) presented above is that the latter do make any optimality claims about the functional form of $\hat{d}$. Rather, they are only focused on high-level properties of $\hat{d}$, in particular the expected reward from deploying it $\hat{d}$.

Interpreting Regression Tables. A major question left open in the above discussion is how the p-values in Table 1 ought to be used in applied data analysis. If a coefficient in the learned rule has a significant p-value, as in Table 1, how should we take this into account in practice? As discussed in Jiang et al. (2019), we should in general not expect the population minimizer of (1.2) to actually be linear in applications; however, given reasonable model-free assumptions, the confidence intervals above ought to cover the population minimizers

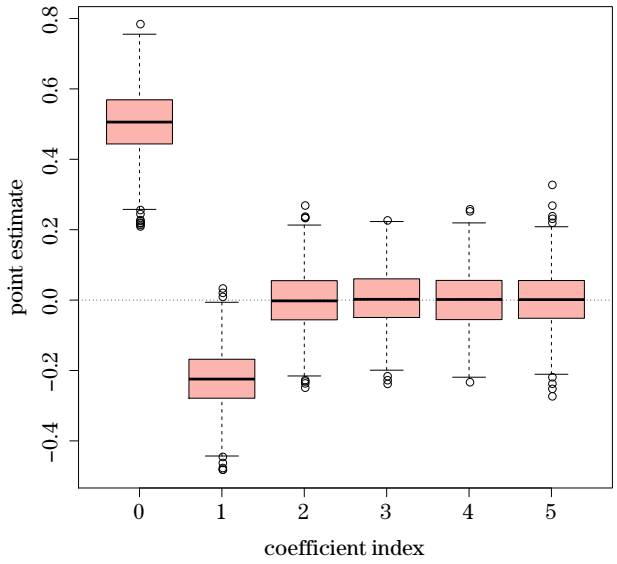

Figure 1. Point estimates for $\hat{c}$ and $\hat{\beta}_1, \ldots, \hat{\beta}_5$ on the design (1.5), aggregated across 1,000 simulation replications.

$$\{c^*, \beta^*\} \tag{1.4}$$
$$= \underset{c,\beta}{\operatorname{argmin}} \left\{ \mathbb{E}\left[\frac{R_i}{A_i\pi + (1-A_i)/2} \left(-\frac{(A_i+1)}{2}(c + X_i\beta) + \log\left(1 + e^{c+X_i\beta}\right)\right)\right]\right\}.$$

The question then becomes one of understanding how to interpret c^* and β^* in a general non-parametric setting. A first encouraging result is that, if there is no treatment effect, then the null model minimizes the above loss:

Proposition 1. *If $\tau(x) = 0$ for all $x \in \mathbb{R}^p$ and A_i is randomized, then $\{c^* = 0, \beta^* = 0\}$ is a minimizer of the population entropy loss* (1.4).

In precision medicine, we are often interested in the more subtle question of whether personalized treatment is useful. One might then hope for a result of the following type: if the treatment effect is constant, i.e., $\tau(x) = \tau$ for all $x \in \mathbb{R}^p$, then $\beta^* = 0$. However, this is *not* true in general. It is possible to design data-generating distributions with no treatment heterogeneity, but where the minimizer β^* in (1.4) is nonzero. Furthermore, it is possible to design settings where where $\mathbb{E}\left[\text{Cov}\left[\tau(X), X_j \,\middle|\, X_{-j}\right]\right]$ is positive but β_j^* is negative, etc.

A Simulation Study. To investigate the extent to which nonzero β^* may arise in a problem without any treatment heterogeneity, we consider a simple simulation example. We generate data as follows, with $n = 4,000$ and $p = 5$:

$$X_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right), \quad A_i \sim \text{Bernoulli}\left(\frac{1}{3}\right), \quad R_i = X_{i1} + \frac{A_i + 1}{3} + E_i, \tag{1.5}$$

where the treatment is randomized, and E_i is an exogenous standard exponential random variable. The treatment effect is obviously constant ($\tau = 2/3$), but a simple calculation shows that β_1^* is nonzero. Figure 1 reports numerical results, and we observe that the point estimate for $\hat{\beta}_1$ is in fact systematically negative. What's going on here is that the baseline expected reward $\mathbb{E}\left[R_i(-1) \,\middle|\, X_i\right]$ changes with X_{i1}, and this affects β_1^* even when there is no treatment heterogeneity.

This simulation was also the basis for the results given in Table 1, which presents bootstrap confidence intervals obtained on a single simulation run. When aggregated over 400 simulation replications, these 95% confidence intervals for $\beta_1^*, \ldots, \beta_5^*$ cover 0 with probabilities 22%, 94%, 96%, 94%, and 96%, respectively. The upshot is that any interpretation of p-values such as those in Table 1 is rather delicate, and a significant p-value for β_j^* cannot necessarily be taken as evidence that variable j is needed for designing optimal personalized treatments. An R script replicating these simulation results is available as Algorithm 1 in the Appendix.

Closing Thoughts. Providing simple and interpretable insights about optimal personalized treatment rules is a challenging task. Existing approaches to policy learning provide utilitarian regret bounds as in (1.3). These bounds require no assumptions on the functional form of the optimal treatment assignment rule. However, one downside of the utilitarian regret approach is that it does not provide much information about the functional form of good treatment assignment rules—rather, in the tradition of learning theory (e.g., Vapnik (2000)), it only seeks to show that $\hat{d}$ is not much worse than the best rule in the class $\mathcal{D}$.

The discussed paper proposes a contrasting approach based on hypothesis testing that allows for simple summaries. However, as discussed above, the resulting p-values are difficult to interpret. In particular, the fact that $\hat{\beta}_j$ is significantly different from 0 does not necessarily imply that X_j is useful for personalized treatment assignment, or that there is any treatment heterogeneity at all. In a general non-parametric setting, results on Fisher consistency of the entropy objective do not translate into a simple characterization the limiting parameters β^* of linear policies obtained via entropy learning.

Interpretable, flexible, and robust significance assessment for policy learning remains an important problem. In a recent advance, Rai (2018) built on the em-

pirical maximization approach (1.3) and proposed confidence sets $\widehat{\mathcal{D}} \subset \mathcal{D}$ for the optimal policy $d^* \in \text{argmax}\,\{V(d') : d' \in \mathcal{D}\}$; however, these confidence sets have a generic shape (they are obtained by inverting a hypothesis test), and so cannot be summarized in a simple way as in Table 1. Finally, Nie and Wager (2017), Zhao, Small and Ertefaie (2017), and others have studied flexible estimation of the treatment effect function $\tau(x)$; however, this statistical task is only indirectly linked to the problem of inference about optimal policies.

Appendix

Algorithm 1 Replication script for simulation results

```
rm(list = ls()); set.seed(1)
# Assume that treatment A is coded as +/- 1.
entropy_treat = function(R, A, X) {
    X.with.intercept = cbind(1, X)
    prob = mean(A == 1)
    loss = function(beta) {
        theta = X.with.intercept %*% beta
        mean(R / (A * prob + (1 - A) / 2) *
                 (-(A + 1) / 2 * theta + log(1 + exp(theta))))
    }
    nlm.out = nlm(loss, rep(0, ncol(X) + 1))
    nlm.out$estimate
}
boot_se = function(R, A, X, B = 100) {
    boot.out = replicate(B, {
        bidx = sample.int(length(A), length(A), replace = TRUE)
        entropy_treat(R[bidx], A[bidx], X[bidx,])
    })
    boot.var = var(t(boot.out))
    boot.se = sqrt(diag(boot.var))
}
n = 4000; p = 5; pi = 1/3
all.results = lapply(1:400, function(idx) {
    A = 2 * rbinom(n, 1, pi) - 1
    X = matrix(rbinom(n * p, 1, 0.5), n, p)
    R = X[,1] + (A + 1) / 3 + rexp(n)
    beta.hat = entropy_treat(R, A, X)
    se.hat = boot_se(R, A, X)
    rbind(beta.hat, se.hat)
})
# For Table 1
beta.hat = all.results[[1]][1,]
standard.errors = all.results[[1]][2,]
```

```
pvalues = 2 * pnorm(-abs(beta.hat) / standard.errors)
# For aggregate coverage
zscores = Reduce(rbind, lapply(all.results,
                 function(lll) (lll[1,] / lll[2,])))
round(colMeans(abs(zscores) < qnorm(0.975))[-1], 2)
# For Figure 1
point.estimates.raw = lapply(1:1000, function(idx){
    A = 2 * rbinom(n, 1, pi) - 1
    X = matrix(rbinom(n * p, 1, 0.5), n, p)
    R = X[,1] + (A + 1) / 3 + rexp(n)
    beta.hat = entropy_treat(R, A, X)
    data.frame(est=beta.hat, coef=0:p)
})
```

References

Athey, S. and Wager, S. (2017). Efficient policy learning. *arXiv preprint arXiv:1702.02896*.

Dudík, M., Langford, J. and Li, L. (2011). Doubly robust policy evaluation and learning. In: *Proceedings of the 28th International Conference on Machine Learning*, 1097–1104.

Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica* **77**, 1683–1701.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jiang, B., Song, R., Li, J. and Zeng, D. (2019). Entropy learning for dynamic treatment regimes. *Statistica Sinica* **29**, 1633–1656.

Kallus, N. (2018). Balanced policy evaluation and learning. *Advances in Neural Information Processing Systems*, 8909–8920.

Kallus, N. and Zhou, A. (2018). Confounding-robust policy improvement. *Advances in Neural Information Processing Systems*, 9289–9299.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* **86**, 591–616.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica* **72**, 1221–1246.

Nie, X. and Wager, S. (2017). Learning objectives for treatment effect estimation. *arXiv preprint arXiv:1712.04912*.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics* **39**, 1180.

Rai, Y. (2018). Statistical inference for treatment assignment policies. Unpublished Manuscript.

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

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics* **151**, 70–81.

Stoye, J. (2012). Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics* **166**, 138–156.

Swaminathan, A. and Joachims, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research* **16**, 1731–1755.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Springer Information Science and Statistics.

Zhao, Q., Small, D. S. and Ertefaie, A. (2017). Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*.

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.

Zhou, Z., Athey, S. and Wager, S. (2018). Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*.

Stanford Graduate School of Business, 655 Knight Way, Stanford, CA 94305, USA.
E-mail: swager@stanford.edu

(Received February 2019; accepted February 2019)

DISCUSSION

Yichi Zhang and Eric B. Laber

University of Rhode Island and North Carolina State University

Key words and phrases: Convex surrogates, outcome weighted learning, precision medicine.

1. Introduction

We congratulate Jiang, Song, Li, and Zeng (JSLZ) on their thought-provoking contribution to the growing literature on classification-based estimation of optimal treatment regimes. We also wish to thank the Editor for organizing this discussion; we are honored to be a part of it. We begin with a discussion of why one might choose to apply a classification-based estimator of an optimal treatment regime and what advantages a surrogate-based approach might offer. Motivated by this discussion, as well as comments made by JSLZ, we then evaluate some of the criticisms leveled against Q-learning and direct search methods, which do not use a convex surrogate. For simplicity, we focus on a single decision; however, the points and methodologies presented here extend readily to the multi-decision setting.

1.1. Classification-based estimators

Classification-based estimators recast the estimation of an optimal treatment regime as a weighted classification problem (Zhang et al. (2012a); Zhao et al. (2012); Rubin and van der Laan (2012); Zhang et al. (2012b, 2013)). Such recasting has the obvious advantage exposing the cache of methodologies and theories already developed for classification to the problem of estimating an optimal treatment regime. Leveraging so-called machine learning methods to improve the quality of estimated optimal regimes has become a major focus of methodological research among both regression-based methods (e.g., Zhao et al. (2011); Moodie, Dean and Sun (2013); Taylor, Cheng and Foster (2015); Murray, Yuan and Thall (2018); Ertefaie and Strawderman (2018); Zhang et al. (2018)) and classification-based methods (e.g., Zhao et al. (2015); Zhou et al. (2017); Zhang and Zhang (2018); Liu et al. (2018); Qi and Liu (2018)). As JSLZ note in their abstract, entropy learning is an example of such research.

By the time the seminal papers on classification-based estimation were published in the statistics literature, the potential benefits of leveraging modern classification methods (as well as modern regression methods) to improve performance in reinforcement learning problems had been known for more than a decade in the computer science literature (see Lagoudakis and Parr (2003); Barto and Dietterich (2004); Ernst, Geurts and Wehenkel (2005) and references therein). In many canonical engineering and computer science applications, the goal is to construct treatment regimes (aka policies or decision strategies) that will be deployed in the field, e.g., to guide the motion of a robot (Singh et al. (1994); Yang and Meng (2000); Finn and Levine (2017)) or to select actions in a strategy game (Silver et al. (2016, 2018)). In such settings, the performance of a learned regime in its target environment is often of paramount importance, whereas factors like intepretability and knowledge generation are secondary. However, in the context of precision medicine, optimal treatment regimes are typically estimated as part of a secondary, i.e., hypothesis-generating, analysis. In such cases, interpretability is key even (or perhaps especially) when the data actually are informing real-time decision support (Nahum-Shani et al. (2017); Tewari and Murphy (2017); Luckett et al. (2018)). Clinicians (rightly) are unwilling to cede their clinical decisions to an unintelligible black-box estimated from a single clinical trial or observational study; indeed, interpretability is now mandated for algorithm-based clinical decision support in the European Union (see Goodman and Flaxman (2017)).

If the goal is to generate new clinical knowledge by means of an interpretable estimated optimal treatment regime, a reasonable approach is to posit a class of acceptable regimes, e.g., those that can be represented as linear thresholds (as in JSLZ, and many others), trees (Zhang et al. (2012a); Laber and Zhao (2015); Zhu et al. (2017); Sies and Van Mechelen (2017); Tao, Wang and Almirall (2018)), or lists (Zhang et al. (2015); Wang and Rudin (2015); Lakkaraju and Rudin (2017); Zhang et al. (2018)). When constructing and evaluating such estimators, we believe that the following factors are key: (F1) consistency for the optimal regime within the class under consideration, (F2) formal inference procedures for the performance of the learned regime, and (F3) diagnostic procedures to identify any loss in performance induced by restricting the class of regimes, e.g., a confidence interval for the difference in value between the optimal regime in the restricted class relative to a larger superclass of regimes.

To the best of our knowledge, (F3) has received little attention in the literature, though it seems critical, especially for highly structured regimes like those representable as lists. With surrogate-based approaches like entropy learning, one potentially promising approach to (F3) would be to consider a confidence interval for the difference between the value of a regime estimated using a nonlinear kernel and that of a linear regime. JSLZ use smoothness of the entropy loss to provide confidence sets for the value of the learned rule, thus addressing (F2). However, entropy-based learning, like Q-learning, need not satisfy (F1). We note that this does not contradict Proposition 1 of JSLZ, as the proposition applies when optimizing over the space of all possible decision rules, not the restricted class of linear decision rules. The lack of (F1) in surrogate-based methods is not a new observation, see Qian and Murphy (2011) and Kosorok and Laber (2019) for examples with squared error loss. In Section 3, we provide an example with entropy loss in which (F1) does not hold, yet the optimal rule is representable as a linear rule. Furthermore, while Q-learning is often criticized by proponents of classification-based methods because of its risk of misspecification and subsequent failure to satisfy (F1), it has the distinct advantage of allowing the use of regression diagnostics to examine model fit, thus mitigating the risk of misspecification (Laber, Linn and Stefanski (2014); Ertefaie, Shortreed and Chakraborty (2016)). We also note that one can separate the class of Q-functions from the class of regimes, i.e., it is not necessary to restrict the class of Q-functions so that the argmax operator induces the desired class of regimes (Taylor, Cheng and Foster (2015); Zhang et al. (2018)). This separation provides greater freedom in modeling the Q-function than presentations of Q-learning sometimes imply.

Methods that directly optimize the inverse probability weighted estimator (IPWE), augumented inverse probability weighted estimator (AIPWE), or other consistent estimators of the value function ensure (F1) under standard conditions (e.g., uniform convergence over the class of regimes, an isolated maximizer, etc.). There appear to be two primary objections to such an approach. The first is that direct optimization of the IPWE/AIPWE is nonconvex and thus potentially computationally burdensome (Section 2.1 JSLZ). However, the application of stochastic optimization algorithms (Zhang et al. (2012b, 2013)), mixed integer programming (Laber, Lizotte and Ferguson (2014); Angelino et al. (2017)), or smoothing with gradient-based procedures with multiple starts (Jiang et al. (2017)) has proved to be successful in a wide variety of precision medicine problems similar to those considered by JSLZ. Nevertheless, such optimization methods may not be feasible in settings with massive data, e.g., electronic health records or billing data, where the convexity in entropy learning and other methods based on convex surrogates may play a critical role (Wang et al. (2016)).

The second criticism leveled against direct optimization of the IPWE/AIPWE is the lack of methodologies for inference. In Section 2, we provide one simple approach that uses an undersmoothed and nonconvex surrogate to retain (F1) while allowing methods for cube-root asymptotics to be used to conduct inference and thereby, we conjecture, satisfy (F2). This approach provides consistently higher value than entropy learning on JSLZ's one-stage simulation examples, while being significantly less variable. Of course, a more thorough examination of this method is needed if any general conclusions are to be made.

2. A Simple Direct Search Estimator

2.1. Framework

For simplicity, we consider data from a single-stage randomized trial; the extension to an observational study is straightforward. We assume that the observed data are $\{(\boldsymbol{X}_i, A_i, R_i)\}_{i=1}^n$, which comprise n i.i.d. copies of $(\boldsymbol{X}, A, R)$, where $\boldsymbol{X} \in \mathbb{R}^{p+1}$ denotes baseline patient covariates, $A \in \{-1, 1\}$ is the assigned treatment, and $R \in \mathbb{R}$ is the outcome coded so that higher values are better. We assume that $\boldsymbol{X}$ has an intercept and that $P(A = 1|\boldsymbol{X}) = P(A = 1) = \pi$ with probability one.

We consider linear decision rules of the form $d(\mathbf{x}) = \text{sign}(\mathbf{x}^\top \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ and $\text{sign}(u) = 1$ if $u > 0$ and $\text{sign}(u) = -1$ otherwise. Define $V_0(\boldsymbol{\beta})$ to be the value of the linear decision rule indexed by $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ so that

$$V_0(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{R}{A\pi + (1-A)/2} I\left\{A = \operatorname{sign}(\boldsymbol{X}^\top \boldsymbol{\beta})\right\}\right],$$

where $I\{\nu\}$ is the indicator that the event ν is true. For any function $m : \mathbb{R}^{p+1} \to \mathbb{R}$, it can be shown (e.g., Laber and Zhao (2015); Zhou et al. (2017)) that

$$V_0(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{R - m(\boldsymbol{X})}{A\pi + (1-A)/2} I\left\{A = \operatorname{sign}(\boldsymbol{X}^\top \boldsymbol{\beta})\right\}\right] + \mathbb{E}\{m(\boldsymbol{X})\}.$$

Define

$$Z = \frac{A\{R - m(\boldsymbol{X})\}}{A\pi + (1-A)/2},$$

so that $V_0(\boldsymbol{\beta}) = F_0(\boldsymbol{\beta}) + D_0$, where $F_0(\boldsymbol{\beta}) = \mathbb{E}\left\{ZI\left(\boldsymbol{X}^\top \boldsymbol{\beta} > 0\right)\right\}$ and $D_0 = -\mathbb{E}\{I(A=-1)Z\} + \mathbb{E}\{m(\boldsymbol{X})\}$. The optimal rule is thus indexed by $\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta}} V_0(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} F_0(\boldsymbol{\beta})$. Because $F_0(\boldsymbol{\beta}) = F_0(k\boldsymbol{\beta})$ for any positive scalar k, we require that $\boldsymbol{\beta}_0^\top \boldsymbol{\beta}_0 = 1$. (Note that a rule indexed by $\boldsymbol{\beta}_0 \equiv 0$ is equivalent to a rule indexed by $\boldsymbol{\beta}_0 = (-1, 0, \ldots, 0)$ and thus there is no loss in generality by assuming a unit norm.)

2.2. Estimation

We begin by describing a plug-in estimator of V_0 and then consider a smoothed variant that is more amenable to gradient-based optimization and inference. To estimate π, we use the sample proportion $\widehat{\pi}_n = n^{-1}\sum_{i=1}^n I(A_i = 1)$. We posit a linear working model of the form $\mathbb{E}(R|\boldsymbol{X} = \mathbf{x}, A = a) = \mathbf{x}_0^\top \boldsymbol{\gamma}_0 + a\,\mathbf{x}_1^\top \boldsymbol{\gamma}_1$, where $\mathbf{x}_0, \mathbf{x}_1$ are (possibly nonlinear) features of $\mathbf{x}$, and $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1$ are unknown coefficients. Let $\widehat{\boldsymbol{\gamma}}_{0,n}$ and $\widehat{\boldsymbol{\gamma}}_{1,n}$ denote the corresponding least squares estimators of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$, and define $\widehat{m}_n(\mathbf{x}) = \mathbf{x}_1^\top \widehat{\boldsymbol{\gamma}}_{0,n}$. Subsequently, define

$$\widehat{Z}_n(\mathbf{x}, a, r) = \frac{a\{r - \widehat{m}_n(\mathbf{x})\}}{a\widehat{\pi}_n + (1-a)/2}.$$

and let $\widehat{Z}_{n,i} = \widehat{Z}_n(\boldsymbol{X}_i, A_i, R_i)$. The plug-in estimator of $F_0(\boldsymbol{\beta})$ is thus

$$\widehat{F}_{n,\mathrm{ns}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \widehat{Z}_{n,i} I\left(\boldsymbol{X}_i^\top \boldsymbol{\beta} > 0\right),$$

where the subscript 'ns' is to indicate that this estimator is non-smooth. As noted in the introduction and by JSLZ, maximizing this objective directly can be difficult and can complicate statistical inference. In the remainder of this discussion, we focus on a smooth alternative to $\widehat{F}_{n,\mathrm{ns}}$.

For each $\boldsymbol{\beta}$, let $p_{\boldsymbol{\beta}}(w, z)$ denote the density of $(\boldsymbol{X}^\top \boldsymbol{\beta}, Z)$. It can be seen that

$$F_0(\boldsymbol{\beta}) = \int zI(w > 0)\,dw\,dz.$$

We consider a kernel density estimator of $p_{\boldsymbol{\beta}}(w,z)$ of the form

$$\widehat{p}^{h}_{\boldsymbol{\beta},n}(w,z)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}\phi\left(\frac{w-\boldsymbol{X}_i^{\top}\boldsymbol{\beta}}{h}\right)\phi\left(\frac{z-\widehat{Z}_{n,i}}{h}\right),$$

where $\phi(t)$ is a Gaussian kernel and $h>0$ is a bandwidth. The smoothed estimator is obtained by replacing $p_{\boldsymbol{\beta}}(w,z)$ with $\widehat{p}^{h}_{\boldsymbol{\beta},n}(w,z)$ to obtain

$$\widehat{F}^{h}_{n,\mathrm{s}}(\boldsymbol{\beta})=\int zI(w>0)\widehat{p}^{h}_{\boldsymbol{\beta},n}(w,z)dwdz=\frac{1}{n}\sum_{i=1}^{n}\widehat{Z}_{n,i}\Phi\left(\frac{\boldsymbol{X}_i^{\top}\boldsymbol{\beta}}{h}\right),$$

where Φ is the CDF of a standard normal random variable. The subscript 's' in $\widehat{F}^{h}_{n,\mathrm{s}}(\boldsymbol{\beta})$ is to indicate that it is smooth. One may also view $\widehat{F}^{h}_{n,\mathrm{s}}(\boldsymbol{\beta})$ as replacing the nonsmooth indicator $I(t>0)$ with the nonconvex surrogate $\Phi(t/h)$ (see Jiang et al. (2017)). In the simulation experiments, we set $h=n^{-1/2}$ to ensure that any asymptotic effects of the smoothing are negligible. To obtain an estimator of $V_0(\boldsymbol{\beta})$, one can use $\widehat{D}_n=n^{-1}\sum_{i=1}^{n}\left\{I(A_i=-1)\widehat{Z}_{n,i}+\widehat{m}_n(\boldsymbol{X}_i)\right\}$ and subsequently define $\widehat{V}^{h}_{n,s}(\boldsymbol{\beta})=\widehat{F}^{h}_{n,\mathrm{s}}(\boldsymbol{\beta})+\widehat{D}_n$.

The estimated optimal regime is indexed by the coefficients

$$\widehat{\boldsymbol{\beta}}^{h}_{n,\mathrm{s}}=\underset{\boldsymbol{\beta}\,:\,\boldsymbol{\beta}^{\top}\boldsymbol{\beta}=1}{\operatorname{argmax}}\,\widehat{V}^{h}_{n,\mathrm{s}}(\boldsymbol{\beta})=\underset{\boldsymbol{\beta}:\boldsymbol{\beta}^{\top}\boldsymbol{\beta}=1}{\operatorname{argmax}}\,\widehat{F}^{h}_{n,\mathrm{s}}(\boldsymbol{\beta}).$$

To facilitate inference, we transform this constrained optimization problem into an unconstrained one by expressing $\boldsymbol{\beta}$ in spherical coordinates. For each $\boldsymbol{\beta}$, write $\boldsymbol{\beta}=\boldsymbol{\beta}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a p-dimensional vector, and

$$\begin{aligned}
\beta_1&=\cos(\theta_1),\\
\beta_2&=\sin(\theta_1)\cos(\theta_2),\\
\beta_3&=\sin(\theta_1)\sin(\theta_2)\cos(\theta_3),\\
&\vdots\ \vdots\ \vdots\\
\beta_p&=\sin(\theta_1)\ldots\sin(\theta_{p-1})\cos(\theta_p),\\
\beta_{p+1}&=\sin(\theta_1)\ldots\sin(\theta_{p-1})\sin(\theta_p).
\end{aligned}$$

It follows that

$$\widehat{\boldsymbol{\beta}}^{h}_{n,\mathrm{s}}=\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}^{h}_{n,\mathrm{s}}),\ \text{where}\ \widehat{\boldsymbol{\theta}}^{h}_{n,\mathrm{s}}=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\,\widehat{F}^{h}_{n,\mathrm{s}}\left\{\boldsymbol{\beta}(\boldsymbol{\theta})\right\}. \qquad (*)$$

Because $\widehat{F}^{h}_{n,\mathrm{s}}\left\{\boldsymbol{\beta}(\boldsymbol{\theta})\right\}$ is not convex in $\boldsymbol{\theta}$, it may have multiple local maximizers. One may employ any of the methods discussed in the introduction to approximate a global maximizer. In the simulations presented in Section 3, we used a gradient descent algorithm with multiple starts.

2.3. Inference

To conduct inference, we work on the $\boldsymbol{\theta}$-scale to avoid the constraint $\boldsymbol{\beta}^\top\boldsymbol{\beta} = 1$. We note that JSLZ appear to avoid this scaling issue by defining the target of inference to be the population minimizer of the *convex surrogate*, which is not scale-invariant but also need not maximize the value over the space of linear decision rules. If the goal is estimation and inference for the linear rule that maximizes the value, the issue of scale invariance may be unavoidable. The proposed estimator resembles the maximum score estimator, and thus the expected rate of convergence is $n^{-1/3}$ rather than $n^{-1/2}$ (Kim and Pollard (1990); Shi, Lu and Song (2018)). It is well known that the standard nonparamteric bootstrap fails for estimators with cube-root convergence (Abrevaya and Huang (2005)); instead, we consider a modified bootstrap procedure as in Cattaneo, Jansson and Nagasawa (2017). Denote the negative Hessian matrix of $\widehat{F}^{\widetilde{h}}_{n,\mathrm{s}}\{\boldsymbol{\beta}(\boldsymbol{\theta})\}$ at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}}$ as

$$\widehat{\mathbf{H}}^{\widetilde{h}}_{n,s} = -\left.\frac{\partial^2 \widehat{F}^{\widetilde{h}}_{n,\mathrm{s}}\{\boldsymbol{\beta}(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}}}.$$

The bandwidth $\widetilde{h}$ used in the construction of the Hessian need not equal the bandwidth used to estimate the value. In our experiments, we used the local bandwidth $\widetilde{h}(\mathbf{x}) = c\sigma(\mathbf{x}^\top\widehat{\boldsymbol{\beta}}^h_{n,\mathrm{s}})n^{-1/9}$, where c is a tuning parameter chosen so that $\widehat{F}^h_{n,\mathrm{s}}\{\boldsymbol{\beta}(\boldsymbol{\theta})\} \approx \widehat{F}^h_{n,\mathrm{s}}\{\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}})\} - (\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}})^\top\widehat{\mathbf{H}}^{\widetilde{h}}_{n,s}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}})/2$ in a neighborhood of $\widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}}$. In addition, we adjust the diagonal elements of $\widehat{\mathbf{H}}^{\widetilde{h}}_{n,\mathrm{s}}$ to ensure positive definiteness as needed.

The bootstrap procedure is as follows. Sample with replacement from the observed data to obtain a bootstrap sample $\{(\boldsymbol{X}^*_i, A^*_i, R^*_i)\}^n_{i=1}$. Let $\widehat{\pi}^*_n, \widehat{m}^*_n$, $\widehat{Z}^*_{n,i}\, i = 1,\ldots,n$, and $\widehat{D}^*_n$ denote the bootstrap analogs of $\pi, \widehat{m}_n$, $\widehat{Z}_{n,i}, \imath = 1,\ldots,n$, and $\widehat{D}_n$. Define the modified bootstrap counterpart to $\widehat{F}^h_{n,\mathrm{s}}\{\boldsymbol{\beta}(\boldsymbol{\theta})\}$ as

$$\widehat{F}^{h*}_{n,\mathrm{s}}\{\boldsymbol{\beta}(\boldsymbol{\theta})\} = \widehat{F}^h_{n,\mathrm{s}}\left\{\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\right\} - \frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}})^\top\widehat{\mathbf{H}}^{\widetilde{h}}_{n,\mathrm{s}}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}}) + \frac{1}{n}\sum_{i=1}^n \widehat{Z}^*_{n,i}\Phi\left\{\frac{\boldsymbol{X}^{*\top}_i\boldsymbol{\beta}(\boldsymbol{\theta})}{h}\right\} - \frac{1}{n}\sum_{i=1}^n \widehat{Z}_{n,i}\Phi\left\{\frac{\boldsymbol{X}^\top_i\boldsymbol{\beta}(\boldsymbol{\theta})}{h}\right\}.$$

Roughly speaking, the first two terms mimic the quadratic behavior of $F_0\{\boldsymbol{\beta}(\boldsymbol{\theta})\}$ near the true value $\boldsymbol{\theta}_0$, while the other two terms mimic the random fluctuations of $\widehat{F}^h_{n,\mathrm{s}}\{\boldsymbol{\beta}(\boldsymbol{\theta})\} - F_0\{\boldsymbol{\beta}(\boldsymbol{\theta})\}$. Let

$$\widehat{\boldsymbol{\theta}}^{h*}_{n,\mathrm{s}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \widehat{F}^{h*}_{n,\mathrm{s}}\{\boldsymbol{\beta}(\boldsymbol{\theta})\},\ \widehat{\boldsymbol{\beta}}^{h*}_{n,\mathrm{s}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}^{h*}_{n,\mathrm{s}}),\ \text{and}\ \widehat{V}^{h*}_{n,\mathrm{s}}(\widehat{\boldsymbol{\theta}}^{h*}_{n,\mathrm{s}}) = \widehat{F}^{h*}_{n,\mathrm{s}}(\widehat{\boldsymbol{\beta}}^{h*}_{n,\mathrm{s}}) + \widehat{D}^*_n.$$

Table 1. The coefficients in the estimated regime (rescaled to have unit norm) and the value of the estimated regime. Monte Carlo standard deviations are in parentheses. For models 1 and 2, the true value for $\boldsymbol{\beta}$ is $(0.272, -0.680, -0.680, 0)$. For models 5 and 6, the value for $\boldsymbol{\beta}$ is not available in closed form because the optimal regime is nonlinear.

Model	Method	β_1	β_2	β_3	β_4	$\mathbb{E}(V_0(\widehat{\boldsymbol{\beta}}))$
1	SIPW	0.275	-0.679	-0.680	0.000	10.303
		(0.010)	(0.011)	(0.011)	(0.018)	(0.006)
	Ent	0.299	-0.724	-0.610	-0.017	10.200
		(0.067)	(0.256)	(0.241)	(0.210)	(0.099)
	QLearn	0.272	-0.680	-0.680	0.000	10.304
		(0.001)	(0.002)	(0.002)	(0.002)	(0.006)
2	SIPW	0.265	-0.677	-0.669	-0.012	9.400
		(0.077)	(0.073)	(0.075)	(0.084)	(0.018)
	Ent	0.334	-0.648	-0.648	0.054	9.350
		(0.064)	(0.206)	(0.224)	(0.183)	(0.063)
	QLearn	0.271	-0.680	-0.679	-0.000	9.412
		(0.038)	(0.063)	(0.065)	(0.052)	(0.012)
5	SIPW	0.255	0.676	-0.675	-0.001	1.846
		(0.061)	(0.054)	(0.055)	(0.109)	(0.014)
	Ent	0.043	0.708	-0.705	0.001	1.787
		(0.041)	(0.187)	(0.186)	(0.142)	(0.023)
	QLearn	-0.123	0.727	-0.727	0.000	1.715
		(0.074)	(0.235)	(0.233)	(0.117)	(0.029)
6	SIPW	0.443	0.613	-0.609	-0.000	4.817
		(0.105)	(0.102)	(0.109)	(0.154)	(0.075)
	Ent	0.194	0.697	-0.698	0.000	4.788
		(0.080)	(0.139)	(0.140)	(0.095)	(0.144)
	QLearn	-0.307	0.692	-0.689	-0.000	4.087
		(0.116)	(0.197)	(0.197)	(0.119)	(0.070)

The empirical percentiles of the forgoing quantities are used to construct confidence sets for the components of $\boldsymbol{\beta}_0$ and $V_0(\widehat{\boldsymbol{\beta}}^h_{n,\mathrm{s}})$.

3. Experiments

3.1. A toy example adopted from Qian and Murphy (2011)

To illustrate the potential impacts of using a surrogate on consistency, we consider the application of entropy learning on the following generative model, which is adapted from Qian and Murphy (2011). Let $X \sim \mathrm{Uniform}[-1, 1]$, $A \sim \mathrm{Uniform}\{-1, 1\}$, and $R = 12 + 5A(X - 1/3)^2 + 0.5\epsilon$, where ϵ is standard normally distributed and independent of X and A. The additive constant of 12 is to ensure that the probability of obtaining a negative reward is vanishingly small. It can be

seen that the optimal decision rule in this case is $d^{\text{opt}}(x) \equiv 1$, which corresponds to the linear estimator $d^{\text{opt}}(x) = \text{sign}\,(\beta_0 + \beta_1 x)$ with $\beta_0 = 1$ and $\beta_1 = 0$. For this generative model, the entropy loss reduces to

$$R(\beta_0, \beta_1) = 12T(\beta_0, \beta_1) - \frac{1}{9}(128\beta_0 - 10\beta_1),$$

where

$$T(\beta_0, \beta_1) = \begin{cases} 2\log(1 + \exp(\beta_0)), & \text{if } \beta_1 = 0, \\ \big\{ \text{Li}_2(-\exp(\beta_0 - \beta_1)) - \text{Li}_2(-\exp(\beta_0 + \beta_1)) \big\}/\beta_1, & \text{if } \beta_1 \neq 0, \end{cases}$$

and $\text{Li}_2(x)$ is the dilogarithm function, defined as

$$\text{Li}_2(x) = \int_x^0 \frac{\log(1-t)}{t} dt.$$

Minimizing the entropy loss yields a rule of the form $d^{\text{ent}}(x) = \text{sign}(\widetilde{\beta}_0 + \widetilde{\beta}_1 x)$, where $\widetilde{\beta}_0 \approx 0.553$ and $\widetilde{\beta}_1 \approx -0.833$. Direct computation shows $V(d^{\text{opt}}) = 14.22$, whereas $V(d^{\text{ent}}) \approx 13.76$ (estimated using 10 million points so that standard errors are on the order of 1×10^{-4}). For comparison, the smoothed estimator proposed in Section 2 has an average value of 14.22, which matches the optimal value up to two significant digits.

3.2. Performance of the estimated regime

We consider models 1, 2, 5, and 6 from JSLZ as these are the one-stage settings. The sample size is fixed at $n = 200$. We compare the regimes obtained by (*) with the regime estimated via entropy learning and Q-learning with a linear model. These three methods are denoted by $SIPW$, Ent and $QLearn$ respectively. To facilitate a fair comparison, we rescale the estimated coefficients $\widehat{\boldsymbol{\beta}}$ in each method so that $\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\beta}} = 1$ and report the Monte Carlo standard deviation of this rescaled version. The value of the estimated regime, $V_0(\widehat{\boldsymbol{\beta}})$, is approximated by generating 10^5 patients following the estimated regime, and its expected value, $\mathbb{E}\{V_0(\widehat{\boldsymbol{\beta}})\}$, is obtained by averaging over 1,000 replications.

The results are given in Table 1. We see that the smoothed method, SIPW, achieves slightly higher value compared to entropy learning on all examples and is considerably less variable. Q-learning is competitive with entropy learning on these examples while also being considerably less variable. In models 1 and 2, where the true values of $\boldsymbol{\beta}_0$ are available analytically, it can be seen that both SIPW and Q-learning exhibit less bias than entropy learning.

Table 2. The coverage rate of 95% confidence intervals for the regime coefficients and its value.

n	Model	Method	β_1	β_2	β_3	β_4	$\mathbb{E}(V_0(\widehat{\boldsymbol{\beta}}))$
200	1	SIPW	1.000	0.999	0.998	0.990	0.949
		Ent	0.998	0.963	0.968	0.970	0.904
		QLearn	0.951	0.938	0.948	0.952	0.952
	2	SIPW	1.000	0.984	0.977	0.982	0.946
		Ent	0.990	0.978	0.981	0.969	0.948
		Qlearn	0.949	0.914	0.929	0.925	0.945
2,000	1	SIPW	0.991	0.988	0.987	0.960	0.945
		Ent	0.968	0.941	0.936	0.949	0.952
		QLearn	0.960	0.942	0.947	0.948	0.949
	2	SIPW	0.996	0.931	0.919	0.972	0.958
		Ent	0.957	0.966	0.961	0.957	0.954
		QLearn	0.946	0.946	0.942	0.965	0.961

3.3. Inference about the coefficients in the estimated regime

We consider models 1 and 2 in JSLZ as these comprise the one-stage settings in which the optimal regime is linear. To explore any large sample effects, we consider sample sizes of $n = 200$ and $n = 2{,}000$. We examine the coverage of a 95% confidence interval for the coefficients indexing the optimal decision rule, as well as the value of the estimated optimal regime. Confidence intervals for Q-learning were based on the (unadjusted) nonparametric bootstrap. The results are given in Table 2. We see that all three methods achieve nominal coverage. The smoothed method, SIPW, gives slightly conservative confidence intervals. As the sample size increases from 200 to 2,000, the coverage rates are closer to the nominal level.

4. Discussion

Entropy learning advances a growing literature on classification-based estimation of optimal treatment regimes. JSLZ are to be commended on an elegant derivation of a class of estimators of which entropy loss is a member. It is interesting to note that entropy loss has been identified as a top performer among convex surrogates in the estimation of optimal treatment regimes using the AIPWE rather than the IPWE as was considered here (Zhao et al. (2019)). We expect such estimators to continue to grow in popularity especially as the computational demands of big data make nonconvex alternatives more difficult to implement.

References

Abrevaya, J. and Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica* **73**, 1175–1204.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M. and Rudin, C. (2017). Learning certifiably optimal rule lists. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*, 35–44.

Barto, A. G. and Dietterich, T. G. (2004). Reinforcement learning and its relationship to supervised learning. *Handbook of Learning and Approximate Dynamic Programming*, 47–64.

Cattaneo, M. D., Jansson, M. and Nagasawa, K. (2017). Bootstrap-based inference for cube root consistent estimators. Tech. rep., arXiv preprint arXiv:1704.08066.

Ernst, D., Geurts, P. and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research* **6**, 503–556.

Ertefaie, A., Shortreed, S. and Chakraborty, B. (2016). Q-learning residual analysis: application to the effectiveness of sequences of antipsychotic medications for patients with schizophrenia. *Statistics in Medicine* **35**, 2221–2234.

Ertefaie, A. and Strawderman, R. L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika* **105**, 963–977.

Finn, C. and Levine, S. (2017). Deep visual foresight for planning robot motion. In: *2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE*, 2786–2793.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* **38**, 50–57.

Jiang, R., Lu, W., Song, R. and Davidian, M. (2017). On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (Statistical Methodology)* **79**, 1165–1185.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics* **18**, 191–219.

Kosorok, M. and Laber, E. (2019). Precision medicine. *Annual Review of Statisticsa and Its Application* **6**, 1–28.

Laber, E. B., Linn, K. A. and Stefanski, L. A. (2014). Interactive model building for Q-learning. *Biometrika* **101**, 831–847.

Laber, E. B., Lizotte, D. J. and Ferguson, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics* **70**, 53–61.

Laber, E. B. and Zhao, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika* **102**, 501–514.

Lagoudakis, M. G. and Parr, R. (2003). Reinforcement learning as classification: Leveraging modern classifiers. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 424–431.

Lakkaraju, H. and Rudin, C. (2017). Learning cost-effective and interpretable treatment regimes. In: *Artificial Intelligence and Statistics*. 166–175.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y. and Zeng, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine* **37**, 3776–3788.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E. and Kosorok, M. R. (2018). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 1–39.

Moodie, E. E. M., Dean, N. and Sun, Y. R. (2013). Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences* **6**, 1–21.

Murray, T. A., Yuan, Y. and Thall, P. F. (2018). A Bayesian machine learning approach for optimizing dynamic treatment regimes. *Journal of the American Statistical Association* **113**, 1255–1267.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A. and Murphy, S. A. (2017). Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* **52**, 446–462.

Qi, Z., Liu, Y. and et al. (2018). D-learning to estimate optimal individual treatment rules. *Electronic Journal of Statistics* **12**, 3601–3638.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics* **39**, 1180–1210.

Rubin, D. B. and van der Laan, M. J. (2012). Statistical issues and limitations in personalized medicine research with clinical trials. *The International Journal of Biostatistics* **8**.

Shi, C., Lu, W. and Song, R. (2018). A massive data framework for M-estimators with cubic-rate. *Journal of the American Statistical Association* **113**, 1698–1709.

Sies, A. and Van Mechelen, I. (2017). Comparing four methods for estimating tree-based treatment regimes. *The International Journal of Biostatistics* **13**.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* **529**, 484.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T. and et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* **362**, 1140–1144.

Singh, S. P., Barto, A. G., Grupen, R. and Connolly, C. (1994). Robust reinforcement learning in motion planning. In: *Advances in Neural Information Processing Systems*. 655–662.

Tao, Y., Wang, L., Almirall, D. and et al. (2018). Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *The Annals of Applied Statistics* **12**, 1914–1938.

Taylor, J. M. G., Cheng, W. and Foster, J. C. (2015). Reader reaction to "a robust method for estimating optimal treatment regimes" by Zhang et al. (2012). *Biometrics* **71**, 267–273.

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In: *Mobile Health. Springer*, 495–517.

Wang, F. and Rudin, C. (2015). Falling rule lists. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. 1013–1022.

Wang, Y., Wu, P., Liu, Y., Weng, C. and Zeng, D. (2016). Learning optimal individualized treatment rules from electronic health record data. In: *2016 IEEE International Conference on Healthcare Informatics (ICHI). IEEE*, 65–71.

Yang, S. X. and Meng, M. (2000). An efficient neural network approach to dynamic robot motion planning. *Neural Networks* **13**, 143–148.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M. and Laber, E. (2012a). Estimating optimal treatment regimes from a classification perspective. *Stat* **1**, 103–114.

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

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.

Zhang, B. and Zhang, M. (2018). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics* **74**, 891–899.

Zhang, Y., Laber, E. B., Davidian, M. and Tsiatis, A. A. (2018). Estimation of optimal treatment regimes using lists. *Journal of the American Statistical Association*, 1–9.

Zhang, Y., Laber, E. B., Tsiatis, A. and Davidian, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71**, 895–904.

Zhao, Y., Zeng, D., Laber, E. B. and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* **110**, 583–598.

Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.

Zhao, Y., Zeng, D., Socinski, M. A. and Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* **67**, 1422–1433.

Zhao, Y.-Q., Laber, E. B., Ning, Y., Saha, S. and Sands, B. (2019). Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *Journal of Machine Learning Research*, accepted.

Zhou, X., Mayer-Hamblett, N., Khan, U. and Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association* **112**, 169–187.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S. and Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics* **73**, 391–400.

257 Tyler Hall, University of Rhode Island, Kingston, RI, 02881, USA.

E-mail: yichizhang@uri.edu

2311 Stinson Drive 5216 SAS Hall, Raleigh, NC 27695-8302 USA.

E-mail: laber@stat.ncsu.edu

(Received March 2019; accepted March 2019)

DISCUSSION

Nathan Kallus

Cornell University

1. Introduction

I would like to congratulate Profs. Binyan Jiang, Rui Song, Jialiang Li,

and Donglin Zeng (JSLZ, henceforth) for an exciting development in conducting inferences on optimal dynamic treatment regimes (DTRs) learned via empirical risk minimization using the entropy loss as a surrogate. JSLZ's ingenuity was to carefully propagate the asymptotic distributions of M-estimators through a backward induction using a roll out of estimated individualized treatment regimes (ITRs) learned by weighted entropy loss minimization. This solved an open problem on how to conduct rigorous inference on DTRs (Laber et al. (2014)).

JSLZ's approach leverages a rejection-and-importance-sampling estimate of the value of a given decision rule based on inverse probability weighting (IPW; see the first unnumbered display equation in JSLZ's Section 2.2) and its interpretation as a weighted (or cost-sensitive) classification, a celebrated reduction (Beygelzimer and Langford (2009); Zhao et al. (2012)). Their use of smooth classification surrogates enables their careful approach to analyzing asymptotic distributions. However, even for evaluation purposes, the IPW estimate is problematic. The estimate is a weighted average of rewards, where, for a horizon of T steps, the weights are the product of T indicators of whether the decision rule's recommendations agree with the observed actions, divided by the product of T propensities for the observed actions. With even just two actions per step, the numerator is most often zero. At the same time, the denominator is invariably tiny, and minor differences in probabilities translate into large differences in their inverse products. The result is weights that discard most of the data and are extremely variable on whatever remains. This renders the estimator practically useless for any horizon T longer than 2–3 and any reasonably sized sample (see also Gottesman et al. (2019)). So, while JSLZ's careful analysis enables us to conduct inferences on DTRs learned by optimizing this estimate (via a surrogate), one might question whether DTRs learned in this way are useful to begin with when $T \geq 3$ and n is realistic, given the unreliable evaluation.

In this comment, I discuss an optimization-based alternative to evaluating ITRs and DTRs, review several connections, and suggest directions forward. In Kallus (2018a), I proposed an approach for evaluating and learning ITRs based on *optimal balance*. Optimal balance – a technique I have also developed for designing controlled experiments (Kallus (2018c)), designing observational studies (Kallus (2017a,b, 2018b); Kallus, Pennicooke and Santacatterina (2018)), and estimating marginal structural models (Kallus and Santacatterina (2018)) – directly targets the error objective of interest by optimally choosing weights that minimize it, rather than relying on plug-in-and-pray approaches that fail for practically sized samples, such as IPW. I show how optimal balance extends to

DTR evaluation and discuss why it holds promise.

2. Balanced Evaluation of ITRs

JSLZ motivate their approach by first considering ITRs; I will do the same. Indeed, using backward induction, evaluating and learning DTRs reduces to evaluating and learning ITRs. In their Eq. (2.1), JSLZ recall the central identity of importance sampling, as applied to ITR evaluation, which I repeat here using potential-outcome notation:

$$V(\mathcal{D} \mid X) \equiv \mathbb{E}\left[\int_{a\in\mathcal{A}} R(a)d\mathcal{D}(a \mid X) \mid X\right] = \mathbb{E}\left[\frac{\mathcal{D}(A \mid X)}{\mathcal{L}(A \mid X)}R \mid X\right], \tag{2.1}$$

where $R(a)$ is the potential reward of action a, for any possible action $a \in \mathcal{A}$ (I make no assumptions on $\mathcal{A}$; it can be discrete or continuous); $X \in \mathcal{X}$ are the prognostic covariates; $\mathcal{D}(a \mid X)$ is the probability (usually Dirac) of the decision rule choosing a when seeing X; A and R are the action and reward, respectively, observed in the data; $\mathcal{L}(a \mid X)$ is the probability of A, given X, in the data; and we assume ignorable assignment: $R(a) \perp\!\!\!\perp A \mid X \; \forall a \in \mathcal{A}$.

Given a sample $\{(X_i, A_i, R_i) : i \leq n\}$, we can operationalize Eq. (2.1) by taking an empirical average of $(\mathcal{D}(A_i \mid X_i)/\mathcal{L}(A_i \mid X_i))R_i$ (e.g., JSLZ's Eq. (2.3)). However, this can prove problematic in practice, because the density ratio $\mathcal{D}(A_i \mid X_i)/\mathcal{L}(A_i \mid X_i)$ can vary wildly, giving some units much higher weight than others and leading to high-variance evaluation. Because of this fundamental problem, there have been many variations and iterations of this basic estimator, including weight normalization and clipping (Swaminathan and Joachims (2015)), "hybrid" clipping using estimates of $\mathbb{E}[R(a) \mid X]$ (Tsiatis and Davidian (2007); Wang, Agarwal and Dudik (2017)), using such estimates as control variates (Dudík, Langford and Li (2011)), optimizing the choice of control variate (Cao, Tsiatis and Davidian (2009); Farajtabar, Chow and Ghavamzadeh (2018)), among others. However, these and other estimators that do not rely completely on extrapolation via outcome modeling need to account for the covariate shift between $\mathcal{L}$ and $\mathcal{D}$ and to weight by the density ratio $\mathcal{D}(A \mid X)/\mathcal{L}(A \mid X)$, and ultimately suffer from its fundamental instability. This is particularly problematic when $\mathcal{D}(A \mid X)$ is Dirac, as is usually the case since optimal policies are deterministic, because it means that any data point that disagrees with $\mathcal{D}$'s recommendation is discarded, even if informative. Smoothing $\mathcal{D}(A \mid X)$ amounts to shrinking the estimate, by linearity. (When A is continuous, this means *all* data points are discarded; smoothing, as in Kallus and Zhou (2018), becomes a

necessity.)

I briefly explain my optimal balancing proposal for ITR evaluation from Kallus (2018a). Given *any* outcome-weighted estimator, $\hat{V} = (1/n)\sum_{i\leq n} W_i R_i$, with $W = W(X_{1:n}, A_{1:n})$, its conditional mean squared error, given the data upon which the weights depend, decomposes to:

$$\mathbb{E}\left[\left(\hat{V} - \frac{1}{n}\sum_{i\leq n} V(\mathcal{D} \mid X_i)\right)^2 \mid X_{1:n}, A_{1:n}\right] = B^2(\mu; W) + \frac{1}{n^2}\sum_{i\leq n} W_i \sigma_i^2,$$

where $\sigma_i^2 = \mathrm{Var}\left(R_i \mid X_i, A_i\right)$, $\mu(x,a) = \mathbb{E}[R_i \mid X_i = x, A_i = a]$, and

$$B(f; W) = \frac{1}{n}\sum_{i\leq n} \int_{a\in\mathcal{A}} f(X_i, a) d(W_i \delta(a - A_i) - \mathcal{D}(a \mid X_i)),$$

which, for every W, is a linear operator on the space of functions $[\mathcal{A}\times\mathcal{X}\to\mathbb{R}]$. (A similar result holds if we augment the weighted estimator with an estimate $\hat{\mu}$, as in AIPW.) Because μ (or the difference $\mu - \hat{\mu}$) is unknown, this suggests seeking weights W that make $B(f; W)$ small for many functions $f \in \mathcal{F}$. Under appropriate conditions,

$$\sup_{f\in\mathcal{F}} B(f; W) = \sup_{\|f\|\leq 1} B(f; W) = \|B(\,\cdot\,; W)\|_*,$$

where $\|\cdot\|$ is the gauge of $\mathcal{F}$ and $\|\cdot\|_*$ its dual. Thus, we seek weights W that make the norm of the operator $B(\,\cdot\,; W)$ small, subject to some 2-norm regularization in order to control the variance. Because setting $W_i = \mathcal{D}(A_i \mid X_i)/\mathcal{L}(A_i \mid X_i)$ makes $B(f; W)$ a sum of independent mean-zero terms, a straightforward empirical process argument (see, e.g., Pollard (1990)) shows that, under appropriate conditions on $\mathcal{F}$, these weights also make $\|B(\,\cdot\,; W)\|_* \to 0$. However, in practice, these plug-in weights still have all the problems of extreme values and being mostly zeros. Instead, my proposal for optimally balanced evaluation of ITRs is to choose weights that directly optimize the error objective of interest:

$$W^* \in \operatorname*{argmin}_{W\geq 0\,:\, 1/n\sum_{i\leq n} W_i = 1} \left\|B(\,\cdot\,; W)\right\|_*^2 + \frac{\lambda}{n^2}\left\|W\right\|_2^2, \tag{2.2}$$

which is a linearly constrained convex optimization problem.

To illustrate how this works, I include an excerpt from Kallus (2018a) in Table 1, where I apply this to an example with $|\mathcal{A}| = 5$, $n = 100$, and low overlap between $\mathcal{L}$ and $\mathcal{D}$. For simplicity, I let $\mathcal{F}$ be the unit ball of the RKHS with kernel $\mathcal{K}((x,a),(x',a')) = \delta(a - a')e^{-\|x-x'\|_2^2}$ and $\lambda = 1$. I include augmented (DR) estimators, using $\hat{\mu}$ fitted by XGBoost, as well as normalized (Hájek) IPW. IPW discards about 86% of the data; the balanced approach only 9%, and cor-

Table 1. ITR evaluation performance in Kallus (2018a, Example 1).

Weights	Outcome Weighting			Augmented OW (DR)			$\|W\|_0$
	RMSE	Bias	SD	RMSE	Bias	SD	
IPW	2.209	−0.005	2.209	4.196	0.435	4.174	13.6 ± 2.9
NIPW	0.519	−0.181	0.487	0.754	0.408	0.634	13.6 ± 2.9
Balanced	**0.280**	0.227	0.163	**0.251**	−0.006	0.251	90.7 ± 3.2

respondingly performs much better.

3. Balanced Evaluation of DTRs

When considering sequential decisions, the fragility of IPW only becomes worse: the weights become even sparser and more extreme, because they are now the ratio of the product of T indicators and the product of T probabilities. Fortunately, the approach to balanced evaluation extends to the case of DTRs, which holds promise for salvaging DTR value estimators that rely on density ratio weighting in any way.

In the sequential setting, we are interested in evaluating the DTR value:

$$V(\mathcal{D}_{1:T}) \equiv \sum_{t \le T} \left\{ V_t(\mathcal{D}_{1:t}) \equiv \mathbb{E} \int_{a_{1:t} \in \mathcal{A}_{1:t}} R_t(a_{1:t}) d\mathcal{D}_{1:t}(a_{1:t} \mid X_{1:t}(a_{1:t-1}), a_{1:t-1}) \right\},$$

where $\mathcal{D}_{1:t}(a_{1:t} \mid X_{1:t}(a_{1:t-1}), a_{1:t-1}) = \prod_{s \le t} \mathcal{D}_s(a_s \mid X_{1:s}(a_{1:s-1}), a_{1:s-1})$ and, for each t and sequence of actions $a_{1:t} \in \mathcal{A}_{1:t} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_t$, we now have potential outcomes for both the reward at time t and the time-dependent covariates at time $t+1$. Our data consist of observations of trajectories $X_{1:T}, A_{1:T}, R_{1:T}$, assuming sequentially ignorable assignment:

$$R_{t:T}(a_{1:T}), X_{t+1:T}(a_{1:T-1}) \perp\!\!\!\perp A_t(a_{1:t-1}) \mid X_{1:t}(a_{1:t-1}), A_{1:t-1}(a_{1:t-2}).$$

As in the case of ITRs, consider estimating $V_t(\mathcal{D}_{1:t})$ by a weighted average of outcomes. To streamline the already cumbersome notation, I discuss this in terms of population averages. Thus, I consider the weighted average of observables $\hat{V}_t = \mathbb{E}[W_{1:t} R_t]$, for some weights $W_{1:t} = \prod_{s \le t} W_s$ where $W_s = W_s(X_{1:s}, A_{1:s})$. Then, iteratively applying sequential ignorability yields a decomposition similar to the ITR case:

$$\hat{V}_t - V_t(\mathcal{D}_{1:t}) = \sum_{s \le t} B_s(\mu_{t,s}; W_s), \tag{3.1}$$

$$B_s(f; W_s) \equiv \mathbb{E} \int_{a_s \in \mathcal{A}_s} f\left(X_{1:s}, A_{1:s-1}, a_s\right) d\left(W_s \delta\left(a_s - A_s\right) - \mathcal{D}_s(a_s | X_{1:s}, A_{1:s-1})\right),$$

Table 2. DTR evaluation performance.

Weights	$T = 3$			$T = 5$			$T = 7$		
	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD
IPW_T	$5e2$	0.96	$5e2$	$4e4$	-42.94	$4e4$	$2e2$	28.61	$2e2$
IPW	$2e2$	0.41	$2e2$	$1e4$	-11.52	$1e4$	$1e4$	-2.08	$1e4$
NIPW_T	11.82	8.39	8.32	38.07	38.01	2.03	63.10	63.09	0.64
NIPW	6.90	4.64	5.10	26.94	26.27	5.96	51.57	51.22	5.98
Bal. $\mathcal{K}_G$	**6.28**	-0.57	6.26	**11.73**	9.69	6.61	**18.65**	17.44	6.61
Bal. $\mathcal{K}_M$	6.87	-0.26	6.87	12.71	10.06	7.78	19.43	17.80	7.78

$$\mu_{t,s}(x_{1:s}, a_{1:s}) \equiv W_{1:s-1}(x_{1:s-1}, a_{1:s-1})\mathbb{E}\left[R^{\mathcal{D}}_{t,s}(a_{1:s}) \mid X_{1:s} = x_{1:s}, A_{1:s-1} = a_{1:s-1}\right],$$
$$R^{\mathcal{D}}_{t,s}(a_{1:s}) \equiv \int_{a_{s+1:t}\in\mathcal{A}_{s+1:t}} R_t(a_{1:t})d\mathcal{D}_{s+1:t}(a_{s+1:t} \mid X_{1:t}(a_{1:t-1}), a_{1:t-1}).$$

This looks rather complicated, but has a simple message: the error is a sum over $s = 1, \ldots, t$ of a particular moment mismatch (B_s) in variables $X_{1:s}, A_{1:s}$ between the weighted data distribution and the distribution induced by deviating and following $\mathcal{D}_s$ at step s. Therefore, to obtain a good estimate, we require weights that make this mismatch small for many functions $f : \mathcal{X}_{1:s} \times \mathcal{A}_{1:s} \to \mathbb{R}$. As before, setting $W_s = \mathcal{D}_s(A_s \mid X_{1:s}, A_{1:s-1})/\mathcal{L}_s(A_s \mid X_{1:s}, A_{1:s-1})$ achieves this at the population level or for very large samples, but can fail horribly in realistically sized samples. (JSLZ actually use weights $\prod_{s=1}^{T} \mathcal{D}_s(A_s \mid X_{1:s}, A_{1:s-1})/\mathcal{L}_s(A_s \mid X_{1:s}, A_{1:s-1})$ on $\sum_{t\le T} R_t$, which is also unbiased, but even more unstable; when estimating the average reward at time t, multiplying by density ratios for times after t is superfluous and just increases the variance.) However, given any sample and some function class $\mathcal{F}_s$, we can seek weights that minimize the (empirical) worst-case mismatches $\|B_s(\,\cdot\,; W_s)\|_{s*}$, subject to some 2-norm regularization to control the variance. Doing so amounts to nothing more than solving Eq. (2.2), for each of $t = 1, \ldots, T$, to obtain W_t, each time considering $X_{1:t}, A_{1:t-1}$ as the "prognostic covariates" being balanced and a_t as the "action." (We could have also placed the $W_{1:s-1}$ term in B_s, rather than in $\mu_{t,s}$, which would have amounted to a simple reweighting of the moment conditions being balanced; however, I focus on the simplest reduction to repeatedly solving problems of the form of Eq. (2.2). We can also apply Eq. (3.1) to the residuals and use an augmented DR-style estimator.)

4. A DTR Evaluation Example

To demonstrate how this works, I include a simple example. Let T vary

and, for $t \leq T$, let $\mathcal{A}_t = \{-1, +1\}$, $\mathcal{X}_t = \mathbb{R}^2$, $R_t(a_{1:t}) = 5a_t + X_{t,1}(a_{t-1}) + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$, $X_{1,j} \sim \mathcal{N}(0,1)$, $X_{t+1,j}(a_{1:t}) = a_t + X_{t,j}(a_{t-1}) + \xi_{t,j}$, $\xi_{t,j} \sim \mathcal{N}(0,1)$, $\mathcal{L}(+1 \mid x_{1:t}, a_{1:t-1}) = \text{expit}(2(X_{t,1} + X_{t,2})A_{t-1})$, and $\mathcal{D}(+1 \mid x_{1:t}, a_{1:t-1}) = \mathbb{I}[(X_{t,1} + X_{t,2})A_{t-1} < 0]$. I consider 2,000 replications of $n = 800$ for each $T \in \{3, 5, 7\}$. To apply balanced evaluation, I let $\mathcal{F}_t$ be the unit ball of the RKHS with kernel $\mathcal{K}((x_{1:t}, a_{1:t}), (x'_{1:t}, a'_{1:t})) = \delta(a_{t-1:t} - a'_{t-1:t})\mathcal{K}_x(x_t, x'_t)$, where $\mathcal{K}_x$ is either the Gaussian ($\mathcal{K}_G$) or Matérn ($\mathcal{K}_M$, $\nu = 5/2$) kernel. I compare this with IPW and normalized IPW. I also include the variation in JSLZ in which we multiply $\sum_{t \leq T} R_t$ by density ratios up to T, referred to as IPW_T.

The results appear in Table 2. The large variance of IPW renders it unusable even with a reasonably sized data set. The variance is so large that it throws off the bias estimated by 2,000 replications (zero in theory). NIPW mitigates this variance, but is actually equal to the uniform weights 37%, 99%, or 100% of the time, for $T = 3, 5, 7$, respectively, and has correspondingly large bias. Balancing has both low bias (indistinguishable from that estimated for IPW) and low variance (comparable to NIPW).

Estimating DTR value when horizons are long is a fundamentally difficult task. Whereas IPW discards most of the data, estimating reward and transition models requires strong modeling assumptions and precarious extrapolations. Balancing could provide a fruitful middle ground: rather than throwing away imperfectly matching trajectories, we imbue the problem with some structure to allow these to be used, while ensuring that our weights achieve the same consistency guarantees afforded by IPW asymptotically (see, e.g., Kallus (2017b, 2018a)).

5. Beyond Evaluation: Learning and Inference

I have argued the merits of using optimal balance to evaluate DTRs. An immediate question is how to use this to learn DTRs. As before, we can optimize the value estimate. Although computationally challenging, this is the approach I took in Kallus (2018a) for ITRs. To apply this to DTRs requires just an application of backward induction with roll out.

With regard to inference (JSLZ's primary concern), this remains open for the balanced approach, but there may be promising directions. Asymptotically, under appropriate conditions on $\mathcal{F}$ and the class of rules being considered, optimal sample weights will uniformly concentrate, so we may consider the distribution when we use the optimal population weights. However, it remains unclear how the estimated rules are distributed (even ITRs). A possible hybrid approach is to use

JSLZ's Eq. (2.8), but to replace $\prod_{s\geq t+1} \mathcal{D}_s(A_s|X_{1:s}, A_{1:s-1})/\mathcal{L}_s(A_s|X_{1:s}, A_{1:s-1})$ with the optimal balancing weights $W^*_{t+1:T}$, while keeping $\mathcal{D}_t(A_t \mid X_{1:t}, A_{1:t-1})/\mathcal{L}_t(A_t \mid X_{1:t}, A_{1:t-1})$ and replacing its numerator with a smooth surrogate. This will at least alleviate issues with longer horizons by limiting IPW to one step, while still being an M-estimator.

While JSLZ's advance is a breakthrough, further advances are necessary. Currently, using IPW and its derivatives to evaluate and learn DTRs when T is moderate and n is realistic is woefully impractical.

References

Beygelzimer, A. and Langford, J. (2009). The offset tree for learning with partial labels. In: *KDD*. 129–138.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–734.

Dudík, M., Langford, J. and Li, L. (2011). Doubly robust policy evaluation and learning. In: *ICML*. 1097–1104.

Farajtabar, M., Chow, Y. and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In: *ICML*. 1446–1455.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F. and Celi, L. (2019). Guidelines for reinforcement learning in healthcare. *Nat Med* **25**, 16–18.

Kallus, N. (2017a). A framework for optimal matching for causal inference. In: *AISTATS*. 372–381.

Kallus, N. (2017b). Generalized optimal matching methods for causal inference. Preprint.

Kallus, N. (2018a). Balanced policy evaluation and learning. In: *NeurIPS*. 8909–8920.

Kallus, N. (2018b). DeepMatch: Balancing deep covariate representations for causal inference using adversarial training. arXiv:1802.05664 [stat.ML]

Kallus, N. (2018c). Optimal a priori balance in the design of controlled experiments. *J Roy Stat Soc B Stat. Methodol.* **80**, 85–112.

Kallus, N., Pennicooke, B. and Santacatterina, M. (2018). More robust estimation of sample average treatment effects using kernel optimal matching in an observational study of spine surgical interventions. arXiv:1811.04274 [stat.ME]

Kallus, N. and Santacatterina, M. (2018). Optimal balancing of time-dependent confounders for marginal structural models. arXiv:1806.01083 [stat.ME]

Kallus, N. and Zhou, A. (2018). Policy evaluation and optimization with continuous treatments. In: *AISTATS*. 1243–1251.

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

Pollard, D. (1990). *Empirical Processes: Theory and Applications*.

Swaminathan, A. and Joachims, T. (2015). The self-normalized estimator for counterfactual learning. In: *NeurIPS*. 3231–3239.

Tsiatis, A. A. and Davidian, M. (2007). Comment: Demystifying double robustness. *Stat. Sci.*

22, 569.

Wang, Y.-X., Agarwal, A. and Dudik, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In: *ICML*. 3589–3597.

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.

School of Operations Research and Information Engineering and Cornell Tech, Cornell University, New York, NY 10044, USA.

E-mail: kallus@cornell.edu

(Received March 2019; accepted March 2019)

REJOINDER

Binyan Jiang, Rui Song, Jialiang Li and Donglin Zeng

The Hong Kong Polytechnic University, North Carolina State University
National University of Singapore and University of North Carolina, Chapel Hill

1. Introduction

We thank Statistica Sinica for providing the venue for this paper and its discussion, and all discussants for their many contributions, insights, and thought-provoking questions. The area of dynamic treatment regimes is developing rapidly, and we hope that our paper and the subsequent discussion will add further momentum to this exciting field. In this rejoinder, we focus on the following four topics: (1) the nonregularity issue, when neither treatment is more beneficial for a nontrivial subgroup (comments by Lu; Qian and Cheng; Qiu et al.); (2) the linear decision boundary (comments by He, Xu, and Wang; Lu; Qiu et al.); (3) extensions that incorporate smooth weights, multiple classes, or a nonconvex loss (comments by Wager; Kallus; Lu; Qian and Cheng; He et al.; Qiu et al.; Zhang and Laber); (4) interpreting the p-value in a real application (comment by Wager).

2. Nonregularity

The nonregularity issue $P(X_t^{*T}\beta_t^0 = 0) > 0$ is a long-standing and challenging inference problem in estimations of dynamic treatment regimes. Our assump-

tion A3 rules out this situation; in particular, we allow a relatively weak condition on the distribution decay near this boundary. Recent attempts to address this issue include finding a probability upper bound, regardless of this nonregularity (Laber et al. (2014)), the m-out-of-n bootstrap method (Chakraborty, Laber and Zhao (2013)), data-adaptive hard-thresholding (Zhu, Zeng and Song (2018)), penalized Q-learning (Song et al. (2014)), and adaptive Q-learning (Goldberg, Song and Kosorok (2012)). However, inferences may be either conservative or unreliable in the case of small sample sizes. Thus, there remains much scope for research on improving inferences with nonregularity.

Although such inferences are theoretically interesting, the impact of nonregularity on practical evaluations of optimal treatment regimes may not be that significant. Essentially, the treatments work very similarly near the boundary. Even if some patients near the decision boundary are allocated to less beneficial treatments, owing to an incorrect inference, the changes to the estimated value function and its inference are practically negligible. This is observed in our numerical studies that demonstrate the robustness of our methods. On the other hand, as suggested by Qiu et al., a more realistic consideration is to test whether the treatment effect exceeds a certain level (i.e., $X_t^{*T}\beta_t^0 \leq \gamma$, for some $\gamma > 0$). Theoretically, we can always choose some γ close to a clinically meaningful threshold such that $P(X_t^{*T}\gamma = 0) = 0$ to void the nonregularity issue.

3. Linear Decision Boundary

Some discussants suggested there may be restrictions on the applicability of the linear form of the treatment decision. Specifically, He et al. suggested nonparametric treatment rules for entropy learning under the RKHS framework, and Qiu et al. obtained nonparametric decision rules using the highly adaptive LASSO approach. Many extensions to our rule are possible, following these suggestions. For example, a simple extension to our linear rule is to incorporate quadratic terms in our estimation to capture possible interactions between the feature covariates. Such ideas emerged recently in the discrimination and regression analysis literature (Jiang, Wang and Leng (2018); Wang, Jiang and Zhu (2019)), and have enjoyed consistency for interaction detection. Furthermore, we may consider smoothing splines to obtain fully nonparametric rules, although the current inference results need to be adapted to reflect the nature of a sieve estimation.

We argue that linear decision rules themselves are still of considerable value

in practice, owing to their simplicity and better interpretability. Several discussants noted that the computational demand could become prohibitively heavy when big data such as electronic transaction records or medical images are present. In this case, the simple form of linear rules coupled with a convex objective function, such as the entropy learning loss in our work, becomes most appealing (Shi et al. (2018)). Finally, partly because of the dichotomous nature of the treatment rule, applying linear rules to derive the value function may not be disadvantageous compared with using rules that are more complex. However, further empirical and theoretical investigation is necessary.

4. Extensions to Incorporate Smooth Weights, Multiclass, or Nonconvex Loss

While many discussants provided helpful suggestions, in this section, we provide brief replies to selected issues; certainly, many deserve a much longer explanation.

Kallus suggested replacing the indicator functions in the estimation equations (e.g., equation (2.8)) with optimal balancing weights to avoid omitting too many samples when T is large. The balanced approach is interesting, and can produce better estimation results than those of outcome-weighted approaches. Here, recent research has led to a greater understanding of the theoretical properties of covariate balancing in causal inferences (Zhao (2019)). However, because the weights are data-driven, it is often difficult to conduct inferences, and the computational complexity might be high for particularly big data. Nevertheless, we agree that it would be meaningful to replace the indicator functions in some early stages with optimal balancing weights. This will enable proper inferences in the later stages, and alleviate the issue of omitting too many samples during the backward estimation procedure. On the other hand, with appropriate smoothness assumptions, it is also possible to obtain valid inferences, with extra effort required to take care of the kernel approximation bias.

Dr. Lu inquired whether E-learning is adaptable to treatments with multiple categories at each stage. Our answer is yes. Note that for the two-class case, the minimizer of (2.4) is $\log(E[R|A=1,\mathbf{X}=\mathbf{x}])/(E[R|A=-1,\mathbf{X}=\mathbf{x}])$, which attains a form similar to that of an odds ratio. Mimicking this form, we may adopt a simple approach to, for example, set the first treatment option as the baseline, and then estimate the pairwise contrast for the other option versus the first option. This operation is similar to the extension of the classical binary

logistic regression model to the multiclass logistic regression model.

In addition to E-learning, proposed in this work, many learning approaches for individual treatment selection have been established under various objective (see the introduction for further examples). Subsequent to this work being accepted for publication, we were informed that C-learning (Zhang and Zhang (2018); Hager, Tsiatis and Davidian (2018)), augmented O-learning (Liu et al. (2018)), concordance assisted learning (Fan et al. (2017); Liang et al. (2018)), maximin projection learning (Shi et al. (2018)), and quantile optimal treatment regimes (Wang et al. (2018)) had since been proposed, among many others. In this discussion, discussants continued to suggest further modifications. Qian and Cheng provided theoretical results for the excess risk and excess value of entropy learning, based on the construction in Bartlett, Jordan and McAuliffe (2006). Qiu et al. studied the behavior of entropy learning under model misspecification, proposing a framework for nonparametric decision rules. Zhang and Laber developed a direct search approach, in which they replace the 0-1 loss with a nonconvex surrogate, to estimate an authentic linear rule that ensures value optimization.

5. Interpretation of p-values

Dr. Wager raised a concern on how to interpret the p-values from the regression tables. We agree that when more than one linear rule leads to the same optimal value, as demonstrated in his numerical example, using a p-value to conclude an important feature for a treatment decision could be misleading.

However, information contained in p-values usually cannot be recovered by other measures. As such, we may not want to completely retire them, for the following detailed reasons:

(a) For an estimated linear rule, such as that in our application, p-values can be used to assess statistical evidence on whether a feature contributes to a rule. However, identifying an important feature does not necessarily imply its utility in the treatment decision for value improvement. This significance is useful in practice when examining the uncertainty of a rule in a finite sample.

(b) The p-values given in the tables provide a computationally simple way to assess the importance of features in the estimated optimal treatment rule. Thus, it is potentially useful for screening out noisy features in the high-dimensional data settings (for example, Zhu, Zeng and Song (2018)). In contrast, using value-based methods to select important features may be computationally intensive or unstable, especially when more than one rule yields the same optimal value.

(c) The p-values given in the tables are associated with the particular surrogate loss (entropy loss) we used. In this sense, each inference used to test a feature's contribution is unique and reliable, in practice. However, value-based inferences are infeasible owing to a lack of uniqueness.

Finally, we believe that the best way to assess the importance of features is a combination of our approach and a value-based method. The former yields an unambiguous treatment rule and associated inference, which is useful in practice. The latter ensures that the selected features truly lead to clinically meaningful benefits.

References

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

Chakraborty, B., Laber, E. and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics* **69**, 714–723.

Fan, C., Lu, W., Song, R. and Zhou, Y. (2017). Concordance-assisted learning for estimating optimal individualized treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79**(5), 1565–1582.

Goldberg, Y., Song, R. and Kosorok, M. R. (2012). Adaptive Q-learning. *IMS Collections: From Probability to Statistics and Back: High-Dimensional Models and Processes* **9**, 150–162,

Hager, R., Tsiatis, A. and Davidian, M. (2018). Optimal two-stage dynamic treatment regimes from a classification perspective with censored survival data. *Biometrics* **74**, 1180–1192.

Jiang, B., Wang, X. and Leng, C. (2018). A direct approach for sparse quadratic discriminant analysis. *J. Mach. Learn. Res.* **19**, 1098–1134.

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

Liang, S., Lu, W., Song, R. and Wang, L. (2018). Sparse concordance-assisted learning for optimal treatment decision. *J. Mach. Learn. Res.* **18**, 1–26.

Liu, Y., Wang, Y., Kosorok, M., Zhao, Y., Zeng, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine* **37**, 3776–3788.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. *Proceedings of The Second Seattle Symposium in Biostatistics*, 189–326. Springer, New York, NY.

Shi, C., Fan, A., Song, R. and Lu, W. (2018). High-dimensional A-learning for optimal dynamic treatment regimes. *Ann. Statist.* **46**, 925–957.

Shi, C., Song, R., Lu, W. and Fu, B. (2018). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80**, 681–702.

Song, R., Wang, W., Zeng, D. and Kosorok, M. R. (2014). Penalized Q-learning for dynamic treatment regimes. *Statistica Sinica* **25**, 901–920.

Wang, C., Jiang, B. and Zhu, L. (2019). Penalized interaction estimation for ultrahigh dimen-

sional quadratic regression. *arXiv preprint arXiv:1901.07147*.

Wang, L., Zhou, Y., Song, R. and Sherwood, B. (2018). Quantile-optimal treatment regimes. *J. Amer. Statist. Assoc.* **113**, 1243–1254.

Zhang, B. and Zhang, M. (2018). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics* **74**, 891–899.

Zhao, Q (2019). Covariate balancing propensity score by tailored loss functions. *Ann. Statist.* **47**, 965–993.

Zhu, W., Zeng, D. and Song, R. (2018). Proper inference for value function in high-dimensional Q-learning for dynamic treatment regimes. *J. Amer. Statist. Assoc.* **113**, 1–14.

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, Raleigh, NC 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 Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA.

E-mail: dzeng@email.unc.edu

(Received May 2019; accepted May 2019)