Statistica Sinica Preprint No: SS-2019-0062	
Title	Comment on 'Entropy Learning for Dynamic Treatment Regimes' by Binyan Jiang, Rui Song, et al.
Manuscript ID	SS-2019-0062
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0062
Complete List of Authors	Hongxiang Qiu Alex Luedtke and Mark van der Laan
Corresponding Author	Hongxiang Qiu
E-mail	qiuhx@uw.edu
Notice: Accepted version subject to English editing.	

Comment on 'Entropy Learning for Dynamic Treatment Regimes' by Binyan Jiang, Rui Song, et al.

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

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

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

1. Introduction

We congratulate the authors on this innovative method to estimate dynamic treatment regimes (DTRs). They introduced the entropy learning (E-learning) framework, which circumvents the need to directly model the conditional mean outcome given covariates when estimating an optimal DTR. Their method extended Zhao et al. (2012, 2015); Rubin and van der Laan (2012) by using a smooth surrogate loss function so that they could obtain valid statistical inference about the parameters in the DTR and 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 to circumvent an unpleasant regularity assumption.

Our discussion is organized as follows.

1. We point out two consequences of restricting attention to a linear class of candidate DTRs when an optimal DTR over an unconstrained class does not belong to this class:

   (a) The infinite-sample limit of the proposed E-learning estimator generally depends on the treatment assignment probabilities.

   (b) The estimated optimal value is generally 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 that:

   (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 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 would be needed to apply the test of the null of no individual-level stage-τ treatment effect proposed by the authors. Importantly, we 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 since, 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. We note that introducing the treatment cost in 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

We start by noting that the β_t^0 that 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 $\hat{\beta}_t$, generally 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. As an example, suppose that two clinical trials are run on the same population with different treatment assignment mechanisms. In this case, the optimal linear decision rules in the two trials can be substantially different, even if the sample sizes are very large.

Momentarily, we will provide a simple example in which there is 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 took 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, namely

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

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 will provide a familiar interpretation for this minimizer. When linearity does not hold, the authors noted 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. We first note that the risk function at stage T rewrites 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 that treatments at previous stages are contained in the history S_T, so the previous treatment assignment mechanism $\pi(A_j, S_j)$, $j < T$, influences the marginal distribution of S_T, and hence could influence β_T^0. At stages $t < T$, there is a similar potential for β_t^0 to depend on treatment mechanisms at stages $j < t$. Moreover, the term $\prod_{j=t+1}^{T} \mathbb{1}\{A_j = \mathrm{sgn}(X_j^{*\top}\beta_j^0)\}$ in (2.1) allows β_t^0 to depend on the decision rules β_j^0 at future stages $j > t$, and hence depend on the treatment assignment mechanisms at the current stage and

future stages $\pi(A_j, S_j)$, $t \leq j < T$. By this argument, one 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 in 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 illustrate the meaningful impact that this dependence on treatment mechanism can have on the interpretation of study results through a simple two-stage example. We consider two different data generating mechanisms that are identical in all regards except for their treatment mechanisms. We denote the treatment mechanisms in the two settings by $\pi^{(1)}$ and $\pi^{(2)}$. We will show that the coefficients (removing "value" because it's a special term) 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, $S_1 = X_1$ is a standard normal, $X_2|A_1 = a_1, X_1 = x_1$ is a normal distribution with mean $a_1 x_1$ and variance 1. We consider a setting where the investigator is only interested in maximizing the final reward, so 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 $X_1 > 0.5$, and we let $\pi_2^{(2)} = 0.9$ when $X_2 < 0.5$ and $\pi_2^{(2)} = 0.1$ when $X_2 > 0.5$.

Table 1 presents β_t^0 in this example in two scenarios where only the treatment assignment mechanisms are different. We can clearly see that β_t^0 depends on the treatment assignment mechanism. Imagine that these two β_t^0 parameters were to be estimated from two large clinical trials that were identical in all aspects except the treatment assignment mechanism. On the one hand, based on the results from the first trial, since $\beta_{21}^0(\pi^{(1)})$ and $\beta_{11}^0(\pi^{(1)})$ are very close to 0, policy makers might conclude that the two treatments have very similar effects. On the other hand, based on the results from the second trial, since $\beta_{21}^0(\pi^{(2)}) < 0$ and $\beta_{11}^0(\pi^{(2)}) < 0$, policy makers might conclude that the two treatments have different effects for different people. Consequently, the policy makers might discourage practitioners to collect the variables X_1, X_2 on future patients based on the results from the first trial, but 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

We highlight here that, though the asymptotic normality of $\hat{\beta}_t$ for β_t^0 can be shown to hold even in the case that the true E-learning risk minimizer is

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

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 that is used to obtain decision rules is different from the zero-one loss that is 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 be different 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 drop 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 is different from the value function, namely

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

We define the maximizer of V by $\beta^\dagger$. We note that, because the value function is nonconcave, finding $\beta^\dagger$ in our numerical example is challenging, and therefore we will 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

Table 2: Two linear DTRs and their optimal values. β^0 is the 'true linear DTR' that the estimated DTR using the surrogate loss is consistent for 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

authors $\hat{V}$ is consistent for $V(\beta_0)$, and hence inconsistent for the optimal value that can be possibly obtained from a linear decision rule V^*.

Returning now to the general case, we 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

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

where we refer the reader to the definition of V_t above Eq. 2.10 in the paper under discussion. Hence, $\hat{V}_t$ provides information about whether or not it is worth advocating for a wide application 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 in practice, $D_{t,\text{current}}$, then we would be confident

that we could gain real benefit from implementing the DTR. We also note from (2.3) that 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 valid confidence lower bound for $V_t(\beta_t^0)$ under misspecification.

A question that we have for the authors is: is it possible to derive the asymptotic normality of $\hat{V}_t$ as an estimator of $V_t(\beta_t^0)$ under regularity conditions, which then leads to valid inference?

3. Nonparametric decision rules

3.1 Unconstrained true risk minimizer

When reading this work, we were excited to find that the loss function proposed by the authors yields a (to our knowledge novel) approach to robustly estimate counterfactual log relative risk. To fix ideas, consider the single-stage setting. Consider the population-level E-learning risk

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

Suppose that we aim to minimize this risk, where the form of f is left unrestricted. In this case, the function f^0 minimizing 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 fact leads to a way to estimate 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 relative risk instead of relative reward here because this is a more common measure of effect size in epidemiology. This reminds us of a similar result for the conditional average treatment effect (CATE). Inspired by Rubin and van der Laan (2007), Luedtke and van der Laan (2016c) showed that one can use least-squares with pseudo outcomes $\left[\frac{\mathbb{1}\{A=1\}}{\pi} - \frac{\mathbb{1}\{A=-1\}}{1-\pi}\right] R$, or doubly robust variants thereof, to nonparametrically estimate the CATE.

A question we have for the authors is: 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 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 correct specification of the conditional mean function implies correct specification of the conditional contrast function, it is never

harder to correctly specify the conditional contrast than 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 of treatment effect, 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 $\mathbb{E}[R|A=-1,X]$ will be 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 the log relative reward. Let $\hat{f}$ denote the estimated log relative risk above. The estimated DTR can then be taken to be equal to $x \mapsto -\text{sgn}\{\hat{f}(x)\}$, 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 rather than the relative reward, where the reward is defined as the absence of the event. It can also be easier to compare $\hat{f}$ with results from other studies, especially case-control studies where odds ratios are reported as an approximation to the relative risk.

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

One 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 successfully applied to the least-squares loss and the logistic loss for regression. The approach used in these settings has been termed the Highly Adaptive LASSO (HAL) (van der Laan, 2017; Benkeser and Van Der Laan, 2016). Under conditions, due to a bound on the uniform entropy of the class $\mathcal{F}_M$, these empirical risk minimizers 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 empirical process notation: 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 $L(f^0)$, $[P\{L(f) - L(f^0)\}^2]^{1/2}$,

can be bounded by $P\{L(f) - L(f^0)\}$.

Note that Condition 2 is similar to but different from Condition 3: Condition 2 is about the local behavior of the loss-based dissimilarity $P\{L(f) - L(f^0)\}$ between functions f and f^0, while Condition 3 is about this dissimilarity upper bounding the $L^2(P)$-distance between the loss functions $L(f)$ and $L(f^0)$. We refer the readers to Lemma 1 in van der Laan (2017) for details.

Although the optimization over such a rich function class seems computationally intractable, HAL can be readily implemented. As its name suggests, a HAL estimator can be computed using a LASSO regression. Since 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—that is,

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

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

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, with a nonparametric estimator of the DTR, we can plug it in 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, i.e. $\left[P\{\hat{f}_n - f^0\}^2\right]^{-1/2} = O_p(r_n)$, then the value of the DTR defined via 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 will be

$o_p(n^{-1/4})$. 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 $\alpha = 1$ so 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 HAL is used to estimate f^0.

We remark 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 in the setting where the linearity assumption fails to hold. Second, (3.6) allows the study of the performance of the learned rule under a range of α-dependent margin conditions.

We finally remark that the nonparametric estimation for the decision rule can also be applied in the multi-stage setting. To learn a DTR with 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 of the form similar to (3.5) subject to constraints similar to (3.4). The convergence rate of the estimated optimal value will require more investigation.

4. Non-regularity

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

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

Above, $x_\tau \in \mathbb{R}^{p_\tau}$, $x_\tau^* \equiv (1, x_\tau)$ 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 – we refer the reader 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 $\sqrt{n} \left| \left(x_\tau^{*\top} \hat{\Sigma}_\tau(\hat{\beta}_\tau) x_\tau^* \right)^{-1/2} x_\tau^{*\top} \hat{\beta}_\tau \right|$ exceeds the $(1-\alpha/2)$-quantile of the standard normal distribution, where $\hat{\Sigma}_\tau(\cdot)$ is an estimate of $\Sigma_\tau(\cdot)$.

Here we caution the reader that (4.1) fails to hold in important scenarios that are of scientific interest. The simplest example of this failure of (4.1) 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, and so

(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 great detail, and argues that a condition similar to Condition A3 is essentially necessary for 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 non-regular inference, and therefore the failure of convergence results such as that appearing in (4.1). Informally, exceptional laws arise when the optimal decision for a randomly drawn individual from the population is non-unique at some stage—that is, the same expected reward is attained for this individual regardless of the treatment that they receive.

We note that the validity of (4.1) actually relies on a condition that is slightly weaker than the Condition A3 in the work under discussion. If Condition A3 were strictly required, then this would seem to pose a major issue 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 knew in advance that Condition A3 were valid, then, given a random $X_\tau \sim P$ drawn

independently of the data, the test that rejects the null hypothesis $H_0(X_\tau)$ without looking at the data would 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. Though this weaker condition would continue to require the supposition of Condition A3 to hold for all $t > \tau$, this weaker condition would not require the supposition of Condition A3 to hold 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 or not the optimal treatment for a given individual is unique at some stage—namely, stage τ—it seems problematic to make this *a priori* assumption that this individual's optimal treatment will be unique at all future stages.

One possible approach to mitigate this concern would be to take treatment cost into account when making the stage-τ treatment decision. To fix ideas, 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 large enough additional reward γ_τ so 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 $(X_t^{*\top}\beta_t^0 - \gamma_t)$, where 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, leads to clinical benefit while still satisfying a given cost constraint. Unlike the authors' proposed test that needs to assume that the alternative hypothesis holds at all future stages $t > \tau$, assuming this modification of Condition A3 does not require making this unpleasant assumption, that is, that the expensive treatment is cost-effective at all future stages. The kind of cost-constrained or resource-limited setting that we have discussed here has been previously studied (Luedtke and van der Laan, 2016a; Toth and van der Laan, 2018; VanderWeele et al., 2018). Importantly, in the settings of these works, the standard errors for summaries of the optimal DTR changed in these cost-constrained settings. This is because these works assume that γ_τ is not directly specified, but is instead 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 estimators of the true E-learning risk minimizer would similarly change in this setting.

5. Conclusion

We close by again congratulating the authors on their important contribution to the estimation of and statistical inference for optimal DTRs.

References

Audibert, J. Y. and A. B. Tsybakov (2007). Fast learning rates for plug-in classifiers. *The Annals of statistics 35*(2), 608–633.

Benkeser, D. and M. Van Der Laan (2016). The Highly Adaptive Lasso Estimator. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pp. 689–696. IEEE.

Laber, E. B., D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics 8*(1), 1225.

Luedtke, A. R. and M. J. van der Laan (2016a). Optimal individualized treatments in resource-limited settings. *The international journal of biostatistics 12*(1), 283–303.

Luedtke, A. R. and M. J. van der Laan (2016b, apr). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics 44*(2), 713–742.

Luedtke, A. R. and M. J. van der Laan (2016c). Super-learning of an optimal dynamic treatment

rule. *The international journal of biostatistics 12*(1), 305–332.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pp. 189–326. Springer.

Rubin, D. and M. J. van der Laan (2007). A doubly robust censoring unbiased transformation. *The international journal of biostatistics 3*(1).

Rubin, D. B. and M. J. van der Laan (2012, jul). Statistical issues and limitations in personalized medicine research with clinical trials. *The international journal of biostatistics 8*(1), 18.

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

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*(2).

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

Zhao, Y., D. Zeng, E. B. Laber, and M. R. Kosorok (2015, apr). New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Journal of the American Statistical Association 110*(510), 583–598.

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012, sep). Estimating Individualized

REFERENCES

Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association 107*(499), 1106–1118.

Department of Biostatistics, University of Washington

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

E-mail: laan@berkeley.edu