Statistica Sinica Preprint No: SS-2019-0089	
Title	Discussion of "Entropy Learning for Dynamic Treatment Regimes"
Manuscript ID	SS-2019-0089
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0089
Complete List of Authors	Yichi Zhang and Eric B. Laber
Corresponding Author	Eric B. Laber
E-mail	laber@stat.ncsu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

Discussion of "Entropy Learning for Dynamic Treatment Regimes"

Yichi Zhang[1] and Eric B. Laber[2]

University of Rhode Island[1] *and North Carolina State University*[2]

Key words and phrases: Precision medicine, convex surrogates, outcome weighted learning

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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 et al., 2013; Taylor et al., 2015; Murray et al., 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 et al., 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 et al., 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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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 et al., 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 promis-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

ing 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 et al., 2014a; Ertefaie et al., 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 et al., 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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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 et al., 2014b; 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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 $d(\boldsymbol{x}) = \text{sign}(\boldsymbol{x}^\top \boldsymbol{\beta})$, 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

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

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

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

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 $\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta}} V_0(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} F_0(\boldsymbol{\beta})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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 $\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = a) = \boldsymbol{x}_0^\top \boldsymbol{\gamma}_0 + a\boldsymbol{x}_1^\top \boldsymbol{\gamma}_1$, where $\boldsymbol{x}_0, \boldsymbol{x}_1$ are (possibly nonlinear) features of $\boldsymbol{x}$, 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 $\widehat{m}_n(\boldsymbol{x}) = \boldsymbol{x}_1^\top \widehat{\boldsymbol{\gamma}}_{0,n}$. Subsequently, define

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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

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

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 $\boldsymbol{\beta}=$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

$\boldsymbol{\beta}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a p-dimensional vector, and

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

It follows that

$$\widehat{\boldsymbol{\beta}}^h_{n,\mathrm{s}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}}), \text{ where } \widehat{\boldsymbol{\theta}}^h_{n,\mathrm{s}} = \arg\max_{\boldsymbol{\theta}} \widehat{F}^h_{n,\mathrm{s}}\left\{\boldsymbol{\beta}(\boldsymbol{\theta})\right\}. \qquad (*)$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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 et al., 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 et al. (2017). Denote the negative Hessian matrix of $\widehat{F}_{n,\mathrm{s}}^{\widetilde{h}}\left\{\boldsymbol{\beta}(\boldsymbol{\theta})\right\}$ at $\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{n,\mathrm{s}}^{h}$ as

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

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 $\widetilde{h}(\boldsymbol{x})=c\sigma\left(\boldsymbol{x}^{\top}\widehat{\boldsymbol{\beta}}_{n,\mathrm{s}}^{h}\right)n^{-1/9}$, where c is a tuning parameter chosen so that $\widehat{F}_{n,\mathrm{s}}^{h}\left\{\boldsymbol{\beta}(\boldsymbol{\theta})\right\}\approx\widehat{F}_{n,\mathrm{s}}^{h}\left\{\boldsymbol{\beta}\left(\widehat{\boldsymbol{\theta}}_{n,\mathrm{s}}^{h}\right)\right\}-(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{n,\mathrm{s}}^{h})^{\top}\widehat{\mathbf{H}}_{n,s}^{\widetilde{h}}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{n,\mathrm{s}}^{h})/2$ in a neighborhood of $\widehat{\boldsymbol{\theta}}_{n,\mathrm{s}}^{h}$. In addition, we adjust the diagonal elements of $\widehat{\mathbf{H}}_{n,\mathrm{s}}^{\widetilde{h}}$ to ensure positive definiteness as needed.

The bootstrap procedure is as follows. Sample with replacement from the observed data to obtain a bootstrap sample $\{(\boldsymbol{X}_{i}^{*},A_{i}^{*},R_{i}^{*})\}_{i=1}^{n}$. 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 $\widehat{F}_{n,\mathrm{s}}^{h}\left\{\boldsymbol{\beta}(\boldsymbol{\theta})\right\}$ as

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

Roughly speaking, the first two terms mimic the quadratic behavior of $F_{0}\left\{\boldsymbol{\beta}(\boldsymbol{\theta})\right\}$ near the true value $\boldsymbol{\theta}_{0}$, while the other two terms mimic the random fluctuations of

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

$\widehat{F}^h_{n,\mathrm{s}}\{\boldsymbol{\beta}(\boldsymbol{\theta})\} - F_0\{\boldsymbol{\beta}(\boldsymbol{\theta})\}$. Let

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

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 $X \sim \text{Uniform}[-1,1]$, $A \sim \text{Uniform}\{-1,1\}$, 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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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

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

Minimizing the entropy loss yields a rule of the form $d^{\text{ent}}(x) = \text{sign}\left(\widetilde{\beta}_0 + \widetilde{\beta}_1 x\right)$, 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 1000 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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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.

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 $n = 2000$. 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 2000, 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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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)

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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
2000	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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

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(4):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*, pages 35–44. ACM.

Barto, A. G. and Dietterich, T. G. (2004). Reinforcement learning and its relationship to supervised learning. *Handbook of learning and approximate dynamic programming*, pages 47–64.

Cattaneo, M. D., Jansson, M., and Nagasawa, K. (2017). Bootstrap-based inference for cube root consistent estimators. Technical report. 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(Apr):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(13):2221–2234.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

Ertefaie, A. and Strawderman, R. L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105(4):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)*, pages 2786–2793. IEEE.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a right to explanation. *AI Magazine*, 38(3):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)*, 79(4):1165–1185.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 18(1):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. (2014a). Interactive model building for Q-learning. *Biometrika*, 101(4):831–847.

Laber, E. B., Lizotte, D. J., and Ferguson, B. (2014b). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 70(1):53–61.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

Laber, E. B. and Zhao, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):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)*, pages 424–431.

Lakkaraju, H. and Rudin, C. (2017). Learning cost-effective and interpretable treatment regimes. In *Artificial Intelligence and Statistics*, pages 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(26):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*, (to appear):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(523):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(6):446–462.

Qi, Z., Liu, Y., et al. (2018). D-learning to estimate optimal individual treatment rules. *Electronic Journal of Statistics*, 12(2):3601–3638.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):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(1).

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(524):1698–1709.

Sies, A. and Van Mechelen, I. (2017). Comparing four methods for estimating tree-based treatment regimes. *The international journal of biostatistics*, 13(1).

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement

learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):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*, pages 655–662.

Tao, Y., Wang, L., Almirall, D., et al. (2018). Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *The Annals of Applied Statistics*, 12(3):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(1):267–273.

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

Wang, F. and Rudin, C. (2015). Falling rule lists. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 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)*, pages 65–71. IEEE.

Yang, S. X. and Meng, M. (2000). An efficient neural network approach to dynamic robot motion planning. *Neural networks*, 13(2):143–148.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a). Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(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(4):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(3):681–694.

Zhang, B. and Zhang, M. (2018). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics*, 74(3):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*, pages 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(4):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(510):583–598.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0089

treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):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(4):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*, Just 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(517):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(2):391–400.