**Statistica Sinica Preprint No: SS-2019-0034**

<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>DISCUSSION OF “ENTROPY LEARNING FOR DYNAMIC TREATMENT REGIMES”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manuscript ID</strong></td>
<td>SS-2019-0034</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
<td>10.5705/ss.202019.0034</td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
<td>Min Qian and Bin Cheng</td>
</tr>
<tr>
<td><strong>Corresponding Author</strong></td>
<td>Min Qian</td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
<td><a href="mailto:mq2158@cumc.columbia.edu">mq2158@cumc.columbia.edu</a></td>
</tr>
</tbody>
</table>

Notice: Accepted version subject to English editing.
DISCUSSION OF “ENTROPY LEARNING FOR DYNAMIC TREATMENT REGIMES”

Min Qian and Bin Cheng

Department of Biostatistics, Columbia University

We would like to congratulate Professors Jiang, Song, Li and Zeng (hereafter JSLZ) for a stimulating article on dynamic treatment regimes (hereafter DTR). The authors made 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.

Dynamic treatment regimes concern with the development of treatment decision rules that can be used to tailor treatment based on patients’ need over time. Current methods for estimating DTR can be classified into two branches: the indirect approach (e.g. Q-learning; see Murphy 2005) and the direct approach. In the direct approach, one needs to deal with a non-convex optimization problem due to the existence of an indicator loss, and often a surrogate loss is used (e.g. hinge loss used in Zhao et al. 2015). JSLZ proposed to replace the indicator loss by a smooth surrogate entropy loss,
and obtained asymptotic normality results for the estimated parameters and value functions for the purpose of inference. Below we will first discuss the inference problem and the conditions. We will then look at the problem from the risk bound point of view.

Inference is of critical importance in DTR. It helps researchers to decide the best treatment for each patient with a measure of confidence. However, it is challenging to make inference when data are presented around the decision boundary [Robins 2004 Laber et al. 2014]. In a linear decision boundary setting, following JSLZ’s notations, this means that $|X_t^* \beta_t^0|$ has a non-negligible probability mass around 0. Indeed, the asymptotic normality results in JSLZ rely on a low-noise condition, namely $|X_t^* \beta_t^0|$ 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 $|X_t^* \beta_t^0|$ is bounded away from zero, and non-normal otherwise; an adaptive procedure was proposed to deal with this problem. From the treatment decision point of view, for patient with $X_t^* = x_t^*$, since treatment decision is based on the sign of $x_t^* \beta_t^0$, it is essential to test whether $x_t^* \beta_t^0 = 0$. Thus the behavior of $X_t^* \beta_t$ around zero is of great interest. It would be desirable to see how to address the non-regularity 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, for the entropy loss function, two risk bounds inspired by Bartlett et al. (2006). We demonstrate them in the single stage decision setting. The result in the multi-stage setting would be similar.

Let $X$ 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 $\pi(X) \triangleq P(A = 1|X)$ denote the known treatment randomization probability. The value function for a treatment decision rule $D : \mathcal{X} \to \{-1, 1\}$, namely $V(D)$, is defined as the expected outcome if the study population follows the decision rule. The goal is to estimate the optimal decision rule $D^{opt}$ that maximizes $V(D)$. It is easy to see that

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

Thus maximizing $V(D)$ is equivalent to minimizing $\mathbb{E} \left[ \frac{RI(A \neq D(X))}{(A\pi(X) + (1-A)/2)} \right].$ JSLZ proposed to replace the indicator loss $I(A \neq D(X))$ by 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

$$\mathcal{R}_h(f) = \mathbb{E} \left[ \frac{R h(A, f(X))}{(A\pi(X) + (1-A)/2)} \right].$$
Minimizing $R_h(f)$ yields $f_{\text{opt}}(x) = \arg \min_{f: \mathcal{X} \to \mathbb{R}} R_h(f) = \log(\mathbb{E}(Y|X = x, A = 1)/\mathbb{E}(Y|X = x, A = -1))$. It can be shown that $D_{\text{opt}}(X) = \text{sign}(f_{\text{opt}}(X))$. The following theorem connects the excess value, $V(D_{\text{opt}}) - V(D)$, to the excess entropy risk, $R_h(f) - R_h(f_{\text{opt}})$. The proof is similar to that of Bartlett et al. (2006), and is omitted.

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

$$\psi(V(D_{\text{opt}}) - V(D)) \leq R_h(f) - R_h(f_{\text{opt}}),$$

(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 $\epsilon > 0$,

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

(1.2)

Then we have

$$c' \left\{ V(D_{\text{opt}}) - V(D) \right\}^{\beta/1+\beta} \psi \left( \frac{(V(D_{\text{opt}}) - V(D))^{1/(1+\beta)}}{2c'} \right) \leq R_h(f) - R_h(f_{\text{opt}}).$$

(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 \cite{Qian and Murphy 2011} for indirect learning and in \cite{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, $R_h(f) - R_h(f^{opt})$, characterize the asymptotic behavior of the entropy risk. To see that, we replace $f$ and $\mathcal{D}$ in the above theorem by estimates $\hat{f}(X) \triangleq X^T\hat{\beta}$ and $\hat{\mathcal{D}}(X) \triangleq \text{sign}(X^T\hat{\beta})$, respectively, where $X^* = (1, X^T)^T$, and $\hat{\beta}$ is obtained by minimizing the empirical entropy risk. $R_h(\hat{f}) - R_h(f^{opt})$ can be decomposed as

$$R_h(\hat{f}) - R_h(f^{opt}) = [R_h(\hat{f}) - R_h(f^*)] + [R_h(f^*) - R_h(f^{opt})],$$

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

Due to the convexity of $\psi(\cdot)$, it is easy to verify that the risk bound (1.3) always gives an equivalent or better rate than (1.1). The low-noise
condition (1.2) plays a critical role here. Note that (1.2) is a variant of the 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 faster, and thus a faster rate of convergence.

In summary, when there is non-negligible noise presented 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 inference with machine learning to improve learning efficiency at the decision boundary.

References


REFERENCES

Qian, M. and Murphy, S. (2011). Performance Guarantees for Individualized Treatment Rules, 


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

E-mail: mq2158@cumc.columbia.edu; Phone: (212) 305-6448; Fax: (212) 305-9408

E-mail: bc2159@cumc.columbia.edu