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 $\mathcal{D}: \mathcal{X} \rightarrow \{-1,1\}$, 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 $h: \{-1,1\} \times \mathbb{R} \rightarrow \mathbb{R}^+$, 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 $f^{opt}(\boldsymbol{x}) = \arg\min_{f:\mathcal{X}\rightarrow\mathbb{R}} \mathcal{R}_h(f) = \log(\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}, A=1)/\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}, A=-1))$. 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 1. *Suppose R is positive and bounded from above by a constant $B > 0$. Then, for any $f: \mathcal{X} \rightarrow \mathbb{R}$ and $\mathcal{D}: \mathcal{X} \rightarrow \{-1,1\}$, 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 $\psi: \mathbb{R}^+ \rightarrow \mathbb{R}$ 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)