# Interpretable Optimal Treatment Regimes Estimation

Lili Wu, Eric Laber

Department of Statistics, North Carolina State University

## Objectives

- Given state information of a patient, what treatment should be assigned?
- Why does the treatment make sense?
- We are trying to construct interpretable treatment regimes!

## Introduction

*"We can train a model, and it can even give us the right answer. But we cannot just tell the doctor "my neural network says this patient has cancer!" The doctor just won't accept that! They want to know why the neural network says what it says. They want an explanation. They need interpretable models."*
*–"The Doctor Just Won't Accept That!" by Zachary C. Lipton*

- **Motivation**: Precision medicine is of importance. We care about how to formalize a treatment regime, which maps up-to-date patient information to a recommended treatment. Current estimation of optimal treatment regimes may be a 'black-box' of patients' information.
- **Data Structure**: For $n$ independent trajectories
$$\{(\mathbf{S}_i^1, A_i^1, \mathbf{S}_i^2, A_i^2, \cdots, \mathbf{S}_i^{T_i}, A_i^{T_i}, \mathbf{S}_i^{T_i+1})\}_{i=1}^n,$$
where $\mathbf{S}_i^j$ is the state information vector of the $ith$ subject at time point $j$. And $A_i^j$ is the action taken by the $ith$ subject at the state $\mathbf{S}_i^j$. We define the utility function as a known function of states and actions, and denote it by $U_i^t = u(\mathbf{S}_i^t, A_i^t, \mathbf{S}_i^{t+1})$. And we hope utility values as **large** as possible.

NC STATE UNIVERSITY

## Methods

Generate sample data following behavior policy.

1. Use Fitted Q Iteration with Random Forest to get the estimated optimal Q function, $\hat{Q}_N(s,a) := \mathbb{E}\{R^t + \gamma\, max_{a'}\hat{Q}_{N-1}(S^{t+1}, a')|S^t = s, A^t = a\}$, where $N$ is the number of iteration of regression.
2. We restrict $c$ to clauses involving threshold with at most two covariates in order to increase interpretability, hence $c$ is an element of
$$\mathcal{C} = \{\{s_i \le \tau_i\},\ \{s_i > \tau_i\},\ \{s_i \le \tau_i \text{ and } s_j \le \tau_j\},\ \{s_i \le \tau_i \text{ and } s_j > \tau_j\},\ \{s_i > \tau_i \text{ and } s_j \le \tau_j\},\ \{s_i > \tau_i \text{ and } s_j > \tau_j\},\ \{s_i \le \tau_i \text{ or } s_j \le \tau_j\},\ \{s_i \le \tau_i \text{ or } s_j > \tau_j\},\ \{s_i > \tau_i \text{ or } s_j \le \tau_j\},\ \{s_i > \tau_i \text{ or } s_j > \tau_j\} :\ 1 \le i < j \le 8, \tau_i, \tau_j \in \mathbf{R}\},$$
where $\tau_i$ and $\tau_j$ are the thresholds of the $ith$ and $jth$ state variable respectively.
3. **Estimation of the first clause** $(c, a)$: the parameterized decision list is
$$\begin{aligned}&\text{If } \mathbf{s} \in c \text{ then take action } a;\\ &\text{else take action } \arg\max_{a'} \hat{Q}_N(\mathbf{s}, a').\end{aligned} \tag{1}$$
If all samples follow (7), the estimated mean outcome is
$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{I(\mathbf{s}_i^t \in c)\hat{Q}_N(\mathbf{s}_i^t, a) + I(\mathbf{s}_i^t \notin c)\max_{a'}\hat{Q}_N(\mathbf{s}_i^t, a')\right\} \tag{2}$$
In order to make more samples go into region $c$, so as to make the length of decision list shorter, we add a penalty term to the above formula, $\zeta\left\{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} I(\mathbf{s}_i^t \in c)\right\}$, where $\zeta > 0$ is a tuning parameter. So now we can define the estimates $\hat{c}_1$ and $\hat{a}_1$ as
$$(\hat{c}_1, \hat{a}_1) := \arg\max_{c\in\mathcal{C}, a\in\mathcal{A}}\left[\zeta\{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} I(\mathbf{s}_i^t \in c)\}+ \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\{I(\mathbf{s}_i^t \in c)\hat{Q}_N(\mathbf{s}_i^t, a) + I(\mathbf{s}_i^t \notin c)\max_{a'}\hat{Q}_N(\mathbf{s}_i^t, a')\}\right] \tag{3}$$
4. **Estimation of the second clause** $(c, a)$:
$$\begin{aligned}&\text{If } \mathbf{s} \in \hat{c}_1 \text{ then take action } \hat{a}_1;\\ &\text{else if } \mathbf{s} \in c \text{ then take action } a;\\ &\text{else take action } \arg\max_{a'} \hat{Q}_N(\mathbf{s}, a').\end{aligned} \tag{4}$$
If all samples follow the regime 4, we can get the estimated mean outcome. Again, we consider the penalty term. So now we can define the estimates $\hat{c}_2$ and $\hat{a}_2$ as
$$(\hat{c}_2, \hat{a}_2) := \arg\max_{c\in\mathcal{C}, a\in\mathcal{A}}\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\Big[\zeta I(\mathbf{s}_i^t \notin \hat{c}_1, \mathbf{s}_i^t \in c)+ \{I(\mathbf{s}_i^t \notin \hat{c}_1, \mathbf{s}_i^t \in c)\hat{Q}_N(\mathbf{s}_i^t, a) + I(\mathbf{s}_i^t \notin \hat{c}_1, \mathbf{s}_i^t \notin c)\max_{a'}\hat{Q}_N(\mathbf{s}_i^t, a')\Big] \tag{5}$$
Continue this procedure until every sample gets a recommended treatment or the length of the decision list arrives $L_{\max}(= 5)$. If the maximum list length is reached, we apply the treatment $\arg\max_{a'}\sum_{j\in\mathcal{U}}\hat{Q}_N(\mathbf{s}_j, a')$ to the left unassigned states, where $\mathcal{U}$ is the set of unassigned samples.

## Experiments

### Example 1: CartPole

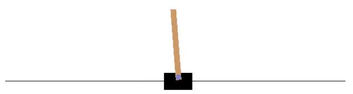

Figure 2: CartPole

- Action space: $\mathcal{A} = \{0, 1\}$, represents {left, right}
- State space: $(S_1, S_2, S_3, S_4) \in \mathbf{R^4}$, represents (position, velocity, angle, angular velocity)
- Goal: Stand for 200 timesteps in each episode. (Large angle or Far-away distance $\rightarrow$ Die! )
- Reward: $R = +1$ until ending the episode

*Solution* :If $s_3 \le 0.0091$ and $s_4 \le 0.0147$, take action 0 ;
else if $s_3 > -0.0337$ and $s_4 > -0.3091$, take action 1;
else if $s_2 > -0.0096$ or $s_3 \le -0.1017$, take action 0;
else if $s_2 \le -0.2543$, take action 1;
else take action 0. (6)

### Example 2: Type I Diabetes

- Action space: $\mathcal{A} = \{0, 1\}$, represents {no treatment, insulin injection}
- Covariates for patient $i$ at time $t$: $(Gl_i^t, Di_i^t, Ex_i^t)$, represents (blood glucose level, dietary intake, total counts of physical activity)
- State space: $(S_1, \cdots, S_8) \in \mathbf{R^8}$, represents $(Gl^t, Di^t, Ex^t, Gl^{t-1}, Di^{t-1}, Di^{t-2}, Di^{t-3}, Ex^{t-1})$
- Utility function: $U_i^t = u(\mathbf{S}_i^t, A_i^t, \mathbf{S}_i^{t+1})$
- Goal: Maximize the mean of cumulative utilities.

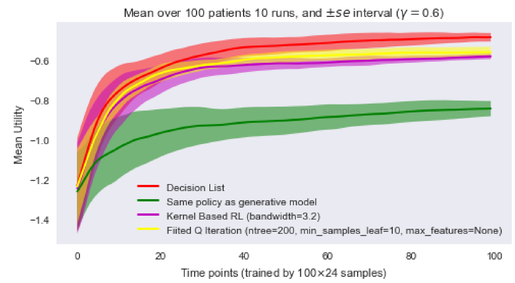


Figure 3: Type 1 Diabetes Performance

Decision List:

If $Gl^t \le 98.19$ and $Di^{t-2} \le 79.87$, take action 0 ;
else if $Gl^t \le 122.19$ and $Ex^t \le 80.53$, take action 1;
else if $Di^t > 68.83$ or $Di^{t-1} \le 0$, take action 1;
else if $Gl^t \le 108.84$ and $Di^t \le 74.28$, , take action 0;
else take action 1. (7)

## Discussion

From the behavior policy, we estimate the optimal policy using decision lists, which are constructed by at most two state variables. This kind of policies increase interpretability and have good performance. The decision lists also can tell us which variables are important to making decisions, we are curious about whether this can also be a way to do variable selection. We are also interested in implementing this framework on more complex environment in the future to check its performance.