Statistics in Medicine 000, 000–000 DOI: 000
000 0000

Supplementary Material for Nested Step-adjusted Tree-based Reinforcement Learning (SAT-Learning) for Evaluating Nested Dynamic Treatment Regimes using Test-and-Treat Observational Data

Ming Tang[1]*, Lu Wang[1], Michael A. Gorin[2]*** , and Jeremy M.G Taylor[1]******

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

[2]The James Buchanan Brady Urological Institute and Department of Urology,

Johns Hopkins University School of Medicine, MD 21287, U.S.A

**email:* mingtang@umich.edu

***email:* luwang@umich.edu

****email:* mgorin1@jhmi.edu

*****email:* jmgt@umich.edu

This paper has been submitted for consideration for publication in *Statistics in Medicine*

Supplementary Materials

Algorithm 1: Stopping Rules

Algorithm 1 Stopping Rules

```
if the node size is less than $2n_0$ then
    the node will not be split
end if
if all possible splits of a node result in a child node with size smaller than $n_0$ then
    the node will not be split
end if
if the current tree depth reaches the user-specified maximum depth then
    the node will not be split
end if
Calculate the best split by
    $$\widehat{\omega}^{opt} = arg\max_{\omega} \left[\mathcal{P}_{sj}(\Omega, \omega) : \min\{n\mathbb{P}_n I(\mathbf{H}_{sj} \in \omega), n\mathbb{P}_n I(\mathbf{H}_{sj} \in \omega^c)\} \geqslant n_0\right].$$
if the maximum purity improvement $\mathcal{P}_{sj}(\Omega, \widehat{\omega}^{opt}) - \mathcal{P}_{sj}(\Omega) < \lambda$ then
    the node will not be split
else
    Split $\Omega$ into $\omega$ and $\omega^c$
end if
```

Algorithm 2: SAT-Learning Implementation Algorithm

Algorithm 2 SAT-Learning Implementation Steps

```
    Stage s  Start the algorithm with $s = S$ Within Stage s:

            (1.1) Set $j = 2$ and only use the data with $d_{sj} = 1$
            (1.2) Obtain AIPW estimates $\widehat{\mu}^{AIPW}_{sj,d_{sj}}(\mathbf{H}_{sj}), d_{sj} = 1, \ldots, K_{sj}$
            (1.3) Set $m = 1$ at root node $\Omega_{sj,m}$
            (1.4) At node $\Omega_{sj,m}$, evaluate the Stopping Rules. If stop, assign a single
                  best treatment
                      $$arg \max_{d_{sj} \in \mathcal{D}_{sj}} \mathbb{P}_n[\widehat{\mu}^{AIPW}_{sj,d_{sj}}(\mathbf{H}_{sj}) I(\mathbf{H}_{sj} \in \Omega_{sj,m})]$$
                  to all subject in $\Omega_{sj,m}$. Otherwise, split $\Omega_{sj,m}$ into child nodes $\Omega_{sj,2m}$
                  and $\Omega_{sj,2m+1}$ by $\widehat{\omega}^{opt}$.
            (1.5) Set $m = m + 1$ and repeat (1.4) until all nodes are terminal.
            (2.1) Set $j = 1$ and use the full data and restrict the available nodes'
                  values according to $\mathcal{P}_{s2}(\Omega, \omega)$
            (2.2) Repeat Steps (1.2)-(1.5)

  Next Stage: Set $s = s - 1$ and repeat Stage s:(1.1)-(2.2), stop if $s = 1$.
```

Simulation Data Generating Process for 5.2

Three covariates, X_1 to X_3, generated as baseline covariates follow $N(0,1)$. Two correlated covariates, X_4 and X_5, are generated as time-varying biomarkers which are measured just before the decision time of the test step within each stage. $(X_4, X_5)' \sim N(\mu, \Sigma)$, where $\mu = (0,0)'$ and $\Sigma = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$. After the test step of each stage, the covariates X_{12} and X_{22} mimic the test results that contribute to the treatment decision nested within each test decision with other covariates. Typically, the test results, such as biopsy results, are of great importance to the treatment decision making. X_{12} and X_{22} follow the distribution of $N(0,1)$. To make the rates of taking the curative treatment equal to 5%, 15%, 20% and 25% in both stages, we also modify the parameters in the data generating models. More details of parameter setting are as follows:

Data Generation for Stage 1 The test decision variables, i.e., D_{11} and D_{21} are set to be the values of $\{0,1\}$ at the first step of each stage. For stage 1 step 1, we generate D_{11} from a $Bernoulli(\pi_{11,1})$ distribution with $\pi_{11,1} = \exp(0.2X_3 + X_4 - 0.5)/(1 + \exp(0.2X_3 + X_4 - 0.5))$. The reward of the stage 1 step 1 is generated as $Y_{11} = X_4^2 + (0.5X_3 + 4)^2 \times I[g_{11}^{opt}(\mathbf{H}_{11}) = D_{11}] - 3|X_1|I(D_{11} = 1) + \epsilon_{11}$ with optimal regimes as

$$g_{11}^{opt}(\mathbf{H}_{11}) = I(X_1 > 0.3)I(X_4 \leqslant 1.3)$$

and $\epsilon_{11} \sim N(0,1)$.

For patients who have been assigned the test, i.e., $D_{11} = 1$, we further generate the treatment assignment D_{12} for them as $D_{12} \sim Bernoulli(\pi_{12,1})$ with

$$\pi_{12,1} = \begin{cases} \exp(0.5X_{12} - X_2 - 3.3)/(1\exp(0.5X_{12} - X_2 - 3.3)) & \text{for Treatment rate=5\%} \\ \exp(0.5X_{12} - X_2 - 2.3)/(1\exp(0.5X_{12} - X_2 - 2.3)) & \text{for Treatment rate=15\%} \\ \exp(0.5X_{12} - X_2 - 1.8)/(1\exp(0.5X_{12} - X_2 - 1.8)) & \text{for Treatment rate=20\%} \\ \exp(0.5X_{12} - X_2 - 1.5)/(1\exp(0.5X_{12} - X_2 - 1.5)) & \text{for Treatment rate=25\%} \end{cases}$$

We generate stage 1 step 2 reward as $Y_{12} = I[D_{12} = g_{12}^{opt}(\mathbf{H}_{12})](7 + 2X_4) + 4X_3 + Y_{11}/3 + 3I(D_{12} = 1)[I(g_{12}^{opt}(\mathbf{H}_{12}) = 1) - 1] + I(D_{12} = 1)(X_{12}^2 + 4) + \epsilon_{12}$ with $\epsilon_{12} \sim N(0, 1)$. The tree-type optimal regime at step 2 is specified as

$$g_{12}^{opt}(\mathbf{H}_{12}) = I(X_4 > 0.5)I(X_{12} \leqslant 0.3)$$

Data Generation for Stage 2: In stage 2, we generate the test decision $D_{21} \sim \mathit{Bernoulli}\ (\pi_{21,1})$ with $\pi_{21,1} = \exp(-0.7 + 0.1X_2 + X_5)/(1 + \exp(-0.7 + 0.1X_2 + X_5))$. The reward of stage 2 step 1 is generated as $Y_{21} = X_1^2 + 2X_2^2 + (8 - X_5)I[g_{21}^{opt}(\mathbf{H}_{21}) = D_{21}] - I(D_{21} = 1) + 4.5I(D_{21} = 1)[I(g_{21}^{opt}(\mathbf{H}_{21}) = 1) - 1] + \epsilon_{21}$ with $\epsilon_{21} \sim N(0, 1)$. The optimal regime is specified as

$$g_{21}^{opt}(\mathbf{H}_{21}) = I(X_2 < 0.8)I(X_5 > 0.1)$$

Among the patients who have had the test in the first step of stage 2, i.e., $D_{21} = 1$ we generate their treatment assignment D_{22} for the second step of the second stage. Specifically, we generate treatment $D_{22} \sim \mathit{Bernoulli}(\pi_{22,1})$ with

$$\pi_{22,1} = \begin{cases} \exp(0.5X_{22} - X_2 - 3.3)/(1\exp(0.5X_{22} - X_2 - 3.3)) & \text{for Treatment rate=5\%} \\ \exp(0.5X_{22} - X_2 - 2.3)/(1\exp(0.5X_{22} - X_2 - 2.3)) & \text{for Treatment rate=15\%} \\ \exp(0.5X_{22} - X_2 - 1.8)/(1\exp(0.5X_{22} - X_2 - 1.8)) & \text{for Treatment rate=20\%} \\ \exp(0.5X_{22} - X_2 - 1.5)/(1\exp(0.5X_{22} - X_2 - 1.5)) & \text{for Treatment rate=25\%} \end{cases}$$

The reward of stage 2 step 2 is generated as $Y_{22} = 3I[D_{22} = g_{22}^{opt}(\mathbf{H}_{22})](2X_{22} - X_5)^2 + Y_{21}/3 + (2X_4 + X_1) + \epsilon_{22}$ and $\epsilon_{22} \sim N(0, 1)$. The optimal treatment for stage 2 is specified as

$$g_{21}^{opt}(\mathbf{H}_{21}) = I(X_{22} < 0.3)I(X_5 > 0.5)$$

Data Preprocessing for the active surveillance data

For the prostate cancer data the exclusion criteria were the following: patients who did not have any PSA observations in the first 4 years were excluded and patients who were not followed after year 4 are excluded. For the remaining patients if they did not have a biopsy, the most recent PSA value that was used in the analysis was the last PSA within the time window between year 0 and year 2 for stage 1 and the last PSA value between year 2 and year 4 for stage 2. For patients who had a biopsy test, the most recent PSA for that test is the PSA value right before the date of biopsy. If a patient had more than one biopsy within a stage, we used the last biopsy result.

To assess the sensitivity of the estimated DTR tree in Figure 2 to modifications of the reward, we included an additional discounting factor for the reward of patients who had an especially high risk of future metastatic prostate cancer. Specifically, when a patient had his Gleason score $\geqslant 7$ (4+3) during the first four years after diagnosis, his reward is reduced by a factor of 95%. The new estimated trees were very similar to the estimated optimal DTR shown in Figure 2, the only differences being small changes is the splitting thresholds at each node.