Supplementary Materials for ”Tree-based Reinforcement Learning for Estimating Optimal Dynamic Treatment Regimes”

Yebin Tao[1], Lu Wang[1] and Daniel Almirall[2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

[2]Institute for Social Research, University of Michigan, Ann Arbor, MI 48104, USA

email: yebintao@umich.edu; luwang@umich.edu; dalmiral@umich.edu

Table S1: Simulation results for a single stage and five treatment options (500 replications, $n = 1000$). π is the propensity score model. $\varphi^{(1)}$ and $\varphi^{(2)}$ indicate equal and varying penalties for misclassification. $opt\%$ shows the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their optimal treatments. $\hat{E}\{Y^*(\hat{g}^{opt})\}$ shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime. $E\{Y^*(g^{opt})\} = 8$.

π	Method	$\varphi^{(1)}$		$\varphi^{(2)}$	
		$opt\%$	$\hat{E}\{Y^*(\hat{g}^{opt})\}$	$opt\%$	$\hat{E}\{Y^*(\hat{g}^{opt})\}$
Correct	ACWL-C_1	94.2 (3.5)	7.69 (0.21)	88.7 (5.5)	7.60 (0.22)
	ACWL-C_2	90.4 (6.1)	7.38 (0.40)	86.4 (8.4)	7.36 (0.38)
	T-RL	95.2 (3.1)	7.74 (0.20)	92.9 (3.7)	7.72 (0.18)
Incorrect	ACWL-C_1	92.5 (4.1)	7.60 (0.23)	84.2 (6.7)	7.47 (0.24)
	ACWL-C_2	90.2 (6.0)	7.37 (0.38)	85.6 (8.2)	7.35 (0.36)
	T-RL	95.2 (2.8)	7.74 (0.17)	91.0 (4.3)	7.68 (0.16)

Additional Simulation 1

This simulation follows Scenario 1 in Tao and Wang (2016). Specifically, we have treatment A from $Multinomial(\pi_0/\pi_s, \pi_1/\pi_s, \pi_2/\pi_s, \pi_3/\pi_s, \pi_4/\pi_s)$, with $\pi_0 = 1$, $\pi_1 = \exp(0.5 - 0.5X_1)$, $\pi_2 = \exp(0.5X_1 + 0.2)$, $\pi_3 = \exp(0.5X_5 + 0.1)$, $\pi_4 = \exp(0.5X_5 - 0.1)$, and $\pi_s = \sum_{m=0}^{4} \pi_m$. We set A to take values in $\{0, \ldots, 4\}$ and generate outcomes as

$$Y = \exp[2.06 + 0.2X_3 - |X_1 + X_2|\varphi\{A, g^{opt}(\mathbf{H})\}] + \epsilon,$$

with $\varphi\{A, g^{opt}(\mathbf{H})\}$ taking the form of $\varphi^{(1)} = 3I\{A \neq g^{opt}(\mathbf{H})\}$ or $\varphi^{(2)} = \{A - g^{opt}(\mathbf{H})\}^2$, $g^{opt}(\mathbf{H}) = I(X_1 > -1)\{1 + I(X_2 > -0.4) + I(X_2 > 0.4) + I(X_2 > 1)\}$ and $\epsilon \sim N(0, 1)$.

The results are shown in Table S1.

Table S2: Additional simulation results based on Scenario 1 with five baseline covariates and outcome model indicating arbitrary penalties for misclassification (500 replications, $n = 500$). $E\{Y^*(g^{opt})\} = 4.69$.

π	Method	Tree-type	
		$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
-	RG	69.7 (3.3)	3.71 (0.11)
Correct	OWL	63.3 (10.1)	3.54 (0.37)
	LZ	95.2 (6.5)	4.54 (0.19)
	ACWL-C_1	90.6 (4.7)	4.49 (0.12)
	ACWL-C_2	90.4 (5.3)	4.47 (0.13)
	T-RL	96.0 (5.1)	4.58 (0.14)
Incorrect	OWL	48.6 (8.0)	3.05 (0.34)
	LZ	84.4 (17.9)	4.24 (0.51)
	ACWL-C_1	88.2 (4.1)	4.46 (0.12)
	ACWL-C_2	88.5 (4.9)	4.46 (0.13)
	T-RL	96.0 (7.8)	4.58 (0.21)

Additional Simulation 2

This simulation follows Scenario 1 in the main paper with five baseline covariates, the same treatment model and the same optimal treatment model but different outcome model. The outcome model indicates arbitrary penalties for misclassification, which is

$$Y = \exp[1.5 + 0.3X_4 - |1.5X_1 - 1| I(A \neq g^{opt})\{4I(A = 0) + I(A = 1) + 2I(A = 2)\}] + \epsilon,$$

with $\epsilon \sim N(0, 1)$.

The results are shown in Table S2.

Table S3: Additional simulation results based on Scenario 1 with five baseline covariates, outcome model (b) and non-tree-type optimal treatment regime (500 replications, $n = 500$). $E\{Y^*(g^{opt})\} = 2$.

π	Method	Non-tree-type	
		$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
-	RG	75.5 (3.5)	1.67 (0.09)
Correct	OWL	46.4 (7.6)	0.98 (0.21)
	LZ	78.6 (6.9)	1.72 (0.13)
	ACWL-C_1	81.5 (4.7)	1.76 (0.11)
	ACWL-C_2	83.0 (4.8)	1.81 (0.10)
	T-RL	82.1 (4.3)	1.79 (0.10)
Incorrect	OWL	35.1 (5.7)	0.71 (0.19)
	LZ	75.2 (9.5)	1.67 (0.51)
	ACWL-C_1	81.4 (4.9)	1.77 (0.11)
	ACWL-C_2	82.0 (5.1)	1.80 (0.10)
	T-RL	81.1 (4.9)	1.78 (0.10)

Additional Simulation 3

This simulation follows Scenario 1 in the main paper with five baseline covariates, the same treatment model and the same outcome model (b) (i.e., varying penalties for treatment misclassification) but different optimal treatment model, which has a non-tree-type

$$g^{opt}(\mathbf{H}) = I(X_1 > 0) + I(X_1 + X_2 > 0).$$

The results are shown in Table S3.

Table S4: Additional simulation results comparing DL by Zhang *et al.* (2015) and T-RL based on Scenario 1 with five baseline covariates, outcome model (a) and various optimal treatment regimes (500 replications, $n = 500$). $E\{Y^*(g^{opt})\} = 2$.

g^{opt}	DL		T-RL	
	$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
(a)	92.6 (7.4)	1.85 (0.16)	97.2 (3.3)	1.94 (0.06)
(b)	93.0 (5.1)	1.85 (0.11)	98.8 (1.1)	1.95 (0.05)
(c)	84.6 (5.1)	1.68 (0.11)	89.7 (9.1)	1.77 (0.19)
(d)	83.0 (2.7)	1.64 (0.07)	85.5 (2.9)	1.69 (0.07)

(a) Same as Scenario 1; (b) $g^{opt}(\mathbf{H}) = I(X_1 > 0.5) + 2I(X_1 \leq 0.5$ and $X_2 \leq -0.3)$; (c) $g^{opt}(\mathbf{H}) = I(X_1 > 0$ and $X_2 > -0.5$ and $X_3 > -1) + 2I(X_1 \leq 0$ and $X_4 > -0.5$ and $X_5 > -1)$; (d) $g^{opt}(\mathbf{H}) = I(X_1 > 0) + I(X_1 + X_2 > 0)$.

Additional Simulation 4

This simulation follows Scenario 1 in the main paper with five baseline covariates, the same treatment model and the same outcome model (a) but different optimal treatment models:

(a) $g^{opt}(\mathbf{H}) = I(X_1 \leq 0)I(X_2 > 0.5) + I(X_1 > 0)\{1 + I(X_3 \leq 0.5)\}$,

(b) $g^{opt}(\mathbf{H}) = I(X_1 > 0.5) + 2I(X_1 \leq 0.5$ and $X_2 \leq -0.3)$,

(c) $g^{opt}(\mathbf{H}) = I(X_1 > 0$ and $X_2 > -0.5$ and $X_3 > -1) + 2I(X_1 \leq 0$ and $X_4 > -0.5$ and $X_5 > -1$,

(d) $g^{opt}(\mathbf{H}) = I(X_1 > 0) + I(X_1 + X_2 > 0)$.

The results are shown in Table S4.

References

1. Zhang, Y., Laber, E. B., Tsiatis, A. and Davidian, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71** 895-904.
2. Tao, Y. and Wang, L. (2017). Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics* **73** 145-155.