"Near-Optimal Reinforcement Learning in Dynamic Treatment Regimes" Supplemental Material

Anonymous Author(s)
Affiliation
Address
`email`

Appendix I. Proofs

In this section, we provide proofs for the theoretical results presented in the main text.

Proof of Theorems 1 to 3

We start by introducing necessary notations for the proof. We say an episode t is ϵ-bad if $V_{\boldsymbol{\pi}^*}(M^*) - Y^t \geq \epsilon$. Let T_ϵ be the number of episodes taken by `UC-DTR` that are ϵ-bad. Let L_ϵ denote the indices of the ϵ-bad episodes up to episode T. The cumulative regret $R_\epsilon(T)$ in ϵ-bad episodes up to episode T is defined as $R_\epsilon(T) = \sum_{t \in L_\epsilon} V_{\boldsymbol{\pi}^*}(M^*) - Y^t$. For any $k = 1, \dots, K$, we define event counts $N(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k)$ in total episodes T as $N(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k) = \sum_{t=1}^{T} I_{\bar{\boldsymbol{S}}_k^t = \bar{\boldsymbol{s}}_k, \bar{\boldsymbol{X}}_k^t = \bar{\boldsymbol{x}}_k}$. Finally, we denote by $\boldsymbol{\mathcal{H}}^t$ the history up to episode t, i.e., $\boldsymbol{\mathcal{H}}^t = \{\bar{\boldsymbol{X}}_K^1, \bar{\boldsymbol{S}}_K^1, Y^1, \dots, \bar{\boldsymbol{X}}_K^t, \bar{\boldsymbol{S}}_K^t, Y^t\}$.

Lemma 2. *Fix $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{t \in L_\epsilon} \left(E_{\bar{\boldsymbol{X}}_K^t}[Y | \bar{\boldsymbol{S}}_K^t] - Y^t \right) \leq \sqrt{\frac{T_\epsilon \log(1/\delta)}{2}}.$$

Proof. Let $\boldsymbol{D}^T$ denote the sequence $\{\bar{\boldsymbol{X}}_K^1, \bar{\boldsymbol{S}}_K^1, \dots, \bar{\boldsymbol{X}}_K^T, \bar{\boldsymbol{S}}_K^T\}$. Rewards Y^t are independent variables by conditioning on $\boldsymbol{D}^T = \boldsymbol{d}^T$. Applying Hoeffding's inequality gives:

$$P\left(\sum_{t \in L_\epsilon} \left(E_{\bar{\boldsymbol{x}}_K^t}[Y | \bar{\boldsymbol{s}}_K^t] - Y^t \right) \geq \sqrt{\frac{T_\epsilon \log(1/\delta)}{2}} \mid \boldsymbol{d}^T \right) \leq \delta.$$

We thus have:

$$P\left(\sum_{t \in L_\epsilon} \left(E_{\bar{\boldsymbol{X}}_K^t}[Y | \bar{\boldsymbol{S}}_K^t] - Y^t \right) \geq \sqrt{\frac{T_\epsilon \log(1/\delta)}{2}} \mid \boldsymbol{d}^T \right) \leq \delta \sum_{\boldsymbol{d}^T} P(\boldsymbol{d}^T) = \delta. \qquad \square$$

Lemma 3. *Fix $\epsilon > 0$, $\delta \in (0,1)$. With probability (w.p.) of at least $1-\delta$, it holds for any $T > 1$, $R_\epsilon(T)$ of `UC-DTR` with parameter δ is bounded by*

$$R_\epsilon(T) \leq 12K\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}| T_\epsilon \log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)} + 4K\sqrt{T_\epsilon \log(2T/\delta)}$$

Proof. Let M^* denote the underlying DTR. Recall that $\mathcal{M}_t$ is a set of DTR instances such that for any $M \in \mathcal{M}_t$, its system dynamics satisfy

$$\left\| P^M_{\bar{\boldsymbol{x}}_k}(\cdot | \bar{\boldsymbol{s}}_k) - \hat{P}^t_{\bar{\boldsymbol{x}}_k}(\cdot | \bar{\boldsymbol{s}}_k) \right\|_1 \leq \sqrt{\frac{6 |\mathcal{S}_{k+1}| \log(2K|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k| t/\delta)}{\max\{1, N^t(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k)\}}}, \tag{16}$$

$$\left| E^M_{\bar{\boldsymbol{x}}_K}[Y | \bar{\boldsymbol{s}}_K] - \hat{E}^t_{\bar{\boldsymbol{x}}_K}[Y | \bar{\boldsymbol{s}}_K] \right| \leq \sqrt{\frac{2 \log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}| t/\delta)}{\max\{1, N^t(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)\}}}. \tag{17}$$

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

By union bounds and Hoeffding's inequality (following a similar argument in [4, C.1]),

$$P(M^* \in \mathcal{M}_t) \leq \frac{\delta}{4t^2}.$$

Since $\sum_{t=1}^{\infty} \frac{1}{4t^2} \leq \frac{\pi^2}{24}\delta < \frac{\delta}{2}$, it follows that with probability at least $1 - \frac{\delta}{2}$, $M^* \in \mathcal{M}^t$ for all episodes $t = 1, 2, \dots$.

For the remainder of the proof, we will assume that $M^* \in \mathcal{M}_t$ for all t. Let $E^{M_t}_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K]$ denote the conditional expected reward in the optimistic DTR M_t. We can write $R_\epsilon(T)$ as:

$$\begin{aligned} R_\epsilon(T) = &\sum_{t \in L_\epsilon} \left(V_{\boldsymbol{\pi}^*}(M^*) - E^{M_t}_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K]\right) &(18)\\ &+ \sum_{t \in L_\epsilon} \left(E^{M_t}_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] - E_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K]\right) &(19)\\ &+ \sum_{t \in L_\epsilon} \left(E_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] - Y^t\right). &(20)\end{aligned}$$

We will next derive bounds over $R_\epsilon(T)$ by bounding quantities in Eqs. (18) to (20) separately.

Bounding Eq. (18) For any DTR M and policy $\boldsymbol{\pi}$, let $V_{\boldsymbol{\pi}}(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1}; M) = E^M_{\boldsymbol{\pi}}[Y|\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1}]$ and $V_{\boldsymbol{\pi}}(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k; M) = E^M_{\boldsymbol{\pi}}[Y|\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k]$. Since $M^* \in \mathcal{M}_t$, we must have $V_{\boldsymbol{\pi}^*}(s_1; M^*) \leq V_{\boldsymbol{\pi}_t}(s_1; M_t)$, i.e., the maximal expected reward of the optimal reward in the optimistic M_t is no less than that in the underlying DTR M^* for any initial state s_1. Further, since $\boldsymbol{\pi}_t$ is deterministic, for any stage k and DTR M,

$$V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_k, \bar{\boldsymbol{X}}^t_{k-1}; M) = V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_k, \bar{\boldsymbol{X}}^t_{k-1}; M). \tag{21}$$

We thus have

$$V_{\boldsymbol{\pi}^*}(M^*) - E^{M_t}_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] \leq V_{\boldsymbol{\pi}^*}(M^*) - V_{\boldsymbol{\pi}^*}(\bar{\boldsymbol{S}}^t_1; M^*) + V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_1, \bar{\boldsymbol{X}}^t_1; M^*) - E^{M_t}_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K].$$

Let $M_t(k)$ denote a combined DTR obtained from M^* and M_t such that

- for $i = 0, 1, \dots, k-1$, its transition probability $P^{M_t(k)}_{\bar{\boldsymbol{x}}_i}(s_{i+1}|\bar{\boldsymbol{s}}_i)$ coincides with the transition probability $P_{\bar{\boldsymbol{x}}_i}(s_{i+1}|\bar{\boldsymbol{s}}_i)$ in the real DTR M^*;
- for $i = k, \dots, K-1$, its transition probability $P^{M_t(k)}_{\bar{\boldsymbol{x}}_i}(s_{i+1}|\bar{\boldsymbol{s}}_i)$ coincides with the transition probability $P^{M_t}_{\bar{\boldsymbol{x}}_i}(s_{i+1}|\bar{\boldsymbol{s}}_i)$ in the optimistic M_t

This is, for any $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, the interventional distribution $P^{M_t(k)}_{\boldsymbol{\pi}}(\bar{\boldsymbol{x}}_K, \bar{\boldsymbol{s}}_K, y)$ factorizes as follows:

$$\begin{aligned} P^{M_t(k)}_{\boldsymbol{\pi}}(\bar{\boldsymbol{x}}_K, \bar{\boldsymbol{s}}_K, y) = &P^{M_t}_{\bar{\boldsymbol{x}}_K}(y|\bar{\boldsymbol{s}}_K) \prod_{i=0}^{k-1} P_{\bar{\boldsymbol{x}}_i}(s_{i+1}|\bar{\boldsymbol{s}}_i)\\ &\cdot \prod_{j=k}^{K-1} P^{M_t}_{\bar{\boldsymbol{x}}_j}(s_{i+1}|\bar{\boldsymbol{s}}_j) \prod_{l=1}^{K-1} \pi_{l+1}(x_{l+1}|\bar{\boldsymbol{s}}_{l+1}, \bar{\boldsymbol{x}}_l). \end{aligned} \tag{22}$$

Obviously, $E^{M_t}_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] = V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_K, \bar{\boldsymbol{X}}^t_K; M^{(K)}_t)$ and $V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_1, \bar{\boldsymbol{X}}^t_1; M_t) = V_{\boldsymbol{\pi}_t}(S^t_1, X^t_1; M^{(1)}_t)$. We thus have

$$\begin{aligned} V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_1, \bar{\boldsymbol{X}}^t_1; M_t) - E^{M_t}_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] &= V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_1, \bar{\boldsymbol{X}}^t_1; M^{(1)}_t) - V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_K, \bar{\boldsymbol{X}}^t_K; M^{(K)}_t)\\ &= \sum_{k=1}^{K-1} V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_k, \bar{\boldsymbol{X}}^t_k; M^{(1)}_t) - V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_{k+1}, \bar{\boldsymbol{X}}^t_{k+1}; M^{(K)}_t)\\ &= \sum_{k=1}^{K-1} V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_k, \bar{\boldsymbol{X}}^t_k; M^{(1)}_t) - V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}^t_{k+1}, \bar{\boldsymbol{X}}^t_k; M^{(K)}_t). \end{aligned}$$

The last step is ensured by Eq. (21). We further have:

$$V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_1^t,\bar{\boldsymbol{X}}_1^t;M_t)-E^{M_t}_{\bar{\boldsymbol{X}}_K^t}[Y|\bar{\boldsymbol{S}}_K^t]=\sum_{k=1}^{K-1}V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k)})-V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k+1)})$$
$$+\sum_{k=1}^{K-1}V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k+1)})-V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_{k+1}^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k+1)}).$$

Eq. (18) can thus be written as:

$$\sum_{t\in L_\epsilon}\left(V_{\boldsymbol{\pi}_t}(M_t)-E^{M_t}_{\bar{\boldsymbol{X}}_K^t}[Y|\bar{\boldsymbol{S}}_K^t]\right)=\sum_{k=1}^{K-1}\sum_{t\in L_\epsilon}V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k)})-V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k+1)})+\sum_{t\in L_\epsilon}Z_t,$$

where Z_t is defined as

$$Z_t=V_{\boldsymbol{\pi}^*}(M^*)-V_{\boldsymbol{\pi}^*}(\bar{\boldsymbol{S}}_1^t;M)+\sum_{k=1}^{K-1}V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k+1)})-V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_{k+1}^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k+1)})$$

By Eq. (22) and basic probabilistic operations,

$$\begin{aligned}
&V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k)})-V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k+1)})\\
&=\sum_{s_{k+1}}(P^{M_t}(s_{k+1}|\bar{\boldsymbol{S}}_k,\bar{\boldsymbol{X}}_k)-P(s_{k+1}|\bar{\boldsymbol{S}}_k,\bar{\boldsymbol{X}}_k))V_{\boldsymbol{\pi}_t}(s_{k+1},\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t)\\
&\leq\left\|P^{M_t}_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k)-P_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k)\right\|_1\max_{s_{k+1}}V_{\boldsymbol{\pi}_t}(s_{k+1},\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t)\\
&\leq 2\sqrt{6|\mathcal{S}_{k+1}|\log(2K|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k|T/\delta)}\frac{1}{\sqrt{\max\{1,N^t(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t)}}
\end{aligned}$$

The last step follows from Eq. (16). From results in [4, D], we have

$$\sum_{t\in L_\epsilon}\frac{1}{\sqrt{\max\{1,N^t(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t)\}}}\leq(\sqrt{2}+1)\sqrt{T_\epsilon|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k|}.$$

This implies:

$$\begin{aligned}
&\sum_{t\in L_\epsilon}\sum_{k=1}^{K-1}V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k)})-V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t,\bar{\boldsymbol{X}}_k^t;M_t^{(k+1)})\\
&\leq\sum_{k=1}^{K-1}2(\sqrt{2}+1)\sqrt{6T_\epsilon|\bar{\boldsymbol{\mathcal{S}}}_{k+1}||\bar{\boldsymbol{\mathcal{X}}}_k|\log(2K|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k|T/\delta)}\\
&\leq 2(\sqrt{2}+1)(K-1)\sqrt{6T_\epsilon|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|\log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)}
\end{aligned}\tag{23}$$

Let $\boldsymbol{\mathcal{H}}^t$ denote the history up to episode t, i.e., $\{\bar{\boldsymbol{X}}_K^1,\bar{\boldsymbol{S}}_K^1,Y^1,\dots,\bar{\boldsymbol{X}}_K^t,\bar{\boldsymbol{S}}_K^t,Y^t\}$. Since $|Z_t|\leq K$ and $E[Z_{t+1}|\boldsymbol{\mathcal{H}}_t]=0$, $\{Z_t:t\in L_\epsilon\}$ is a sequence of martingale differences. By Azuma-Hoeffding inequality [3], we have, with probability at least $1-\frac{\delta}{8T^2}$,

$$\sum_{t\in L_\epsilon}Z_t\leq K\sqrt{6T_\epsilon\log(2T/\delta)}\tag{24}$$

Since $\sum_{T=1}^{\infty}\frac{1}{8T^2}\leq\frac{\pi^2}{48}\delta<\frac{\delta}{4}$, the above inequality holds with probability $1-\frac{\delta}{4}$ for all $T>1$. Eqs. (23) and (24) combined give

$$\begin{aligned}
&\sum_{t\in L_\epsilon}\left(V_{\boldsymbol{\pi}^*}(M^*)-E^{M_t}_{\bar{\boldsymbol{X}}_K^t}[Y|\bar{\boldsymbol{S}}_K^t]\right)\\
&\leq 2(\sqrt{2}+1)(K-1)\sqrt{6T_\epsilon|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|\log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)}+K\sqrt{6T_\epsilon\log(2T/\delta)}
\end{aligned}\tag{25}$$

Bounding Eq. (19) Since both M^*, M_t are in the set $\mathcal{M}_t$,

$$\begin{aligned} E^{M_t}_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] - E_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] &\leq \left| E^{M_t}_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K] - \hat{E}^t_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K] \right| + \left| E_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] - \hat{E}^t_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K] \right| \\ &\leq 2\sqrt{2\log(2K|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T/\delta)} \frac{1}{\sqrt{\max\{1, N^t(\bar{\boldsymbol{S}}^t_K, \bar{\boldsymbol{X}}^t_K)\}}} \end{aligned}$$

The last step follows from Eq. (17). From results in [4, D], we have

$$\sum_{t \in L_\epsilon} \frac{1}{\sqrt{\max\{1, N^t(\bar{\boldsymbol{S}}^t_K, \bar{\boldsymbol{X}}^t_K)\}}} \leq (\sqrt{2}+1)\sqrt{T_\epsilon|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|}.$$

This implies

$$\sum_{t \in L_\epsilon} \left(E^{M_t}_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] - E_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] \right) \leq 2(\sqrt{2}+1)\sqrt{2T_\epsilon|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|\log(2K|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T/\delta)} \tag{26}$$

Bounding Eq. (20) By Lem. 2, we have with probability at least $1 - \frac{\delta}{8T^2}$,

$$\sum_{t \in L_\epsilon} \left(E_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] - Y^t \right) \leq \sqrt{\frac{3T_\epsilon \log(2T/\delta)}{2}} \tag{27}$$

Since $\sum_{T=1}^{\infty} \frac{1}{8T^2} \leq \frac{\pi^2}{48}\delta < \frac{\delta}{4}$, the above equation holds with probability $1 - \frac{\delta}{4}$ for any T.

Eqs. (25) to (27) together give that, with probability at least $1 - \frac{\delta}{2} - \frac{\delta}{4} - \frac{\delta}{4} = 1 - \delta$,

$$\begin{aligned} R_\epsilon(T) \leq{}& (K-1)2(\sqrt{2}+1)\sqrt{6T_\epsilon|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|\log(2K|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T/\delta)} + K\sqrt{6T_\epsilon \log(2T/\delta)} \\ &+ 2(\sqrt{2}+1)\sqrt{2T_\epsilon|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|\log(2K|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T/\delta)} + \sqrt{\frac{3T_\epsilon \log(2T/\delta)}{2}}. \end{aligned}$$

A quick simplification gives:

$$R_\epsilon(T) \leq 12K\sqrt{|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T_\epsilon \log(2K|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T/\delta)} + 4K\sqrt{T_\epsilon \log(2T/\delta)}. \qquad \square$$

Theorem 1. *Fix a $\delta \in (0,1)$. With probability (w.p.) of at least $1-\delta$, it holds for any $T > 1$, the regret of `UC-DTR` with parameter δ is bounded by*

$$R(T) \leq 12K\sqrt{|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T \log(2K|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T/\delta)} + 4K\sqrt{T \log(2T/\delta)}.$$

Proof. Fix $\epsilon = 0$. Naturally, $T_\epsilon = T$ and $R_\epsilon(T) = R(T)$. By Lem. 3,

$$R(T) \leq 12K\sqrt{|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T \log(2K|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T/\delta)} + 4K\sqrt{T \log(2T/\delta)}. \qquad \square$$

Theorem 2. *For any $T \geq 1$, with parameter $\delta = \frac{1}{T}$, the expected regret of `UC-DTR` is bounded by*

$$E[R(T)] \leq \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^-} \left\{ \frac{33^2 K^2 |\boldsymbol{S}||\boldsymbol{\mathcal{X}}|\log(T)}{\Delta_{\boldsymbol{\pi}}} + \frac{32}{\Delta^3_{\boldsymbol{\pi}}} + \frac{4}{\Delta_{\boldsymbol{\pi}}} \right\} + 1.$$

Proof. By Lem. 3 and a quick simplification, we have

$$R_\epsilon(T) \leq 23K\sqrt{|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T_\epsilon \log(T/\delta)}.$$

Since $R_\epsilon(T) \geq \epsilon T_\epsilon$, $\epsilon T_\epsilon \leq 23K\sqrt{|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T_\epsilon \log(T/\delta)}$, which implies

$$T_\epsilon \leq \frac{23^2 K^2 |\boldsymbol{S}||\boldsymbol{\mathcal{X}}|\log(T/\delta)}{\epsilon^2}. \tag{28}$$

This implies that, with probability at least $1-\delta$,

$$R_\epsilon(T) \leq 23K\sqrt{|\boldsymbol{S}||\boldsymbol{\mathcal{X}}|T_\epsilon \log(T/\delta)} = \frac{23^2 K^2 |\boldsymbol{S}||\boldsymbol{\mathcal{X}}|\log(T/\delta)}{\epsilon}$$

Let $\Delta = \arg\min_{\boldsymbol{\pi}\in\boldsymbol{\Pi}^-} \Delta_{\boldsymbol{\pi}}$. Fix $\epsilon = \frac{\Delta}{2}$, $\delta = \frac{1}{T}$, we have

$$E[R_{\frac{\Delta}{2}}(T)] \leq \frac{33^2 K^2 |\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|\log(T)}{\Delta} + 1. \tag{29}$$

We now only need to bound the regrets cumulated in the episodes that are not ϵ-bad, which we call ϵ-good. Let $\tilde{R}_\epsilon(T)$ denote the regret in episodes that are ϵ-good. Let $\tilde{T}_\epsilon$ denote the total number of ϵ-good episodes and let $\tilde{L}_\epsilon$ be indices of ϵ-good episodes. Fix $\epsilon = \frac{\Delta}{2}$, for any ϵ-good episode t, we have $V_{\boldsymbol{\pi}_t}(M^*) - Y^t < \epsilon$. Fix event $\tilde{T}_{\frac{\Delta}{2}} = t$,

$$\tilde{R}_\epsilon(T) = \sum_{i\in\tilde{L}_\epsilon} V_{\boldsymbol{\pi}^*}(M^*) - Y^i \leq t\frac{\Delta}{2}.$$

The above inequality is equivalent to

$$\begin{aligned}
&\sum_{i\in\tilde{L}_{\frac{\Delta}{2}}} V_{\boldsymbol{\pi}^*}(M^*) - V_{\boldsymbol{\pi}_i}(M^*) - Y^i \leq t\frac{\Delta}{2} - \sum_{i\in\tilde{L}_{\frac{\Delta}{2}}} V_{\boldsymbol{\pi}_i}(M^*) \\
&\Rightarrow \sum_{i\in\tilde{L}_{\frac{\Delta}{2}}} \Delta_{\boldsymbol{\pi}_i} - Y^i \leq t\frac{\Delta}{2} - \sum_{i\in\tilde{L}_{\frac{\Delta}{2}}} V_{\boldsymbol{\pi}_i}(M^*) \\
&\Rightarrow \sum_{i\in\tilde{L}_{\frac{\Delta}{2}}} \Delta - Y^i \leq t\frac{\Delta}{2} - \sum_{i\in\tilde{L}_{\frac{\Delta}{2}}} V_{\boldsymbol{\pi}_i}(M^*)
\end{aligned}$$

Since $|\tilde{L}_\epsilon| = \tilde{T}_{\frac{\Delta}{2}}$, we have

$$\tilde{T}_{\frac{\Delta}{2}} = t \Rightarrow \sum_{i\in\tilde{L}_{\frac{\Delta}{2}}} V_{\boldsymbol{\pi}_i}(M^*) - Y^i \leq -t\frac{\Delta}{2}. \tag{30}$$

We could thus bound $E[\tilde{R}_{\frac{\Delta}{2}}(T)]$ as

$$E[\tilde{R}_{\frac{\Delta}{2}}(T)] \leq \frac{\Delta}{2} E[\tilde{T}_{\frac{\Delta}{2}}(T)] \leq \frac{\Delta}{2}\sum_{t=1}^{T} tP(\tilde{T}_{\frac{\Delta}{2}} = t)$$

By Eq. (30), we further have

$$E[\tilde{R}_{\frac{\Delta}{2}}(T)] \leq \frac{\Delta}{2}\sum_{t=1}^{T} tP\left(\sum_{i\in\tilde{L}_{\frac{\Delta}{2}}} V_{\boldsymbol{\pi}_i}(M^*) - Y^i \leq -t\frac{\Delta}{2}\right)$$

Let $C_t = V_{\boldsymbol{\pi}_t}(M^*) - Y^t$. Since $|C_t| < 1$ and $E[C_{t+1}|\boldsymbol{\mathcal{H}}^t] = 0$, $\{C_i : i \in \tilde{L}_{\frac{\Delta}{2}}\}$ is a sequence of martingale differences. Applying Azuma-Hoeffding lemma gives,

$$P\left(\sum_{i\in\tilde{L}_{\frac{\Delta}{2}}} C_i \leq -t\frac{\Delta}{2}\right) \leq e^{-\frac{\Delta^2 t}{8}}.$$

Thus

$$E[\tilde{R}_{\frac{\Delta}{2}}(T)] \leq \frac{\Delta}{2}\sum_{t=1}^{T} te^{-\frac{\Delta^2 t}{8}} \leq \frac{\Delta}{2}\frac{64}{\Delta^4}(\frac{\Delta^2}{8} + 1)e^{-\frac{\Delta^2}{8}}$$

which implies

$$E[\tilde{R}_{\frac{\Delta}{2}}(T)] \leq \frac{32}{\Delta^3} + \frac{4}{\Delta}. \tag{31}$$

Eqs. (29) and (31) together give:

$$E[R(T)] = E[R_{\frac{\Delta}{2}}(T)] + E[\tilde{R}_{\frac{\Delta}{2}}(T)] \leq \frac{33^2 K^2 |\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|\log(T)}{\Delta} + \frac{32}{\Delta^3} + \frac{4}{\Delta} + 1$$

The right-hand side of the above inequality is a decreasing function regarding the gap Δ. By a quick simplification, we prove the statement. $\square$

Theorem 3. *For any algorithm $\mathcal{A}$, any natural numbers $K \geq 1$, and $\left|\boldsymbol{\mathcal{S}}^k\right| \geq 2, \left|\boldsymbol{\mathcal{X}}^k\right| \geq 2$ for any $k \in \{1, \dots, K\}$, there is a DTR M with horizon K, state domains $\boldsymbol{\mathcal{S}}$ and action domains $\boldsymbol{\mathcal{X}}$, such that the expected regret of $\mathcal{A}$ after $T \geq |\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|$ episodes is as least*

$$E[R(T)] \geq 0.05\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T}.$$

Proof. The classic results in bandit literature [1, Thm. 5.1] shows that for each state sequence $_K$, there exists a bandit instance such that for any the total regret of any algorithm is lower bound by

$$E[R(T)] \geq 0.05 \sum_{\bar{\boldsymbol{s}}_K} \sqrt{N(\bar{\boldsymbol{s}}_K)|\boldsymbol{\mathcal{X}}|},$$

where $N(\bar{\boldsymbol{s}}_K)$ is the event count $\bar{\boldsymbol{S}}_K = \bar{\boldsymbol{s}}_K$ for all T episodes. The lower bound in Thm. 3 is achieved when all states $_K$ are decided uniformly at random, i.e., $N(\bar{\boldsymbol{s}}_K) = T/|\bar{\boldsymbol{\mathcal{S}}}_K|$. $\square$

Proofs of Theorems 4 to 6, Lemma 1, and Corollary 2

In this section, we provide proofs for the bounds on transition probabilities of DTRs. Our proofs build on the notion of counterfactual variables [6, Ch. 7.1] and axioms of "composition, effectiveness and reversibility" defined in [6, Ch. 7.3.1].

For a SCM M, arbitrary subsets of endogenous variables $\boldsymbol{X}, \boldsymbol{Y}$, the potential outcome of $\boldsymbol{Y}$ to intervention $do(\boldsymbol{x})$, denoted by $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$, is the solution for Y with $\boldsymbol{U} = \boldsymbol{u}$ in the sub-model $M_{\boldsymbol{x}}$. It can be read as the counterfactual sentence "the value that $\boldsymbol{Y}$ would have obtained in situation $\boldsymbol{U} = \boldsymbol{u}$, had $\boldsymbol{X}$ been $\boldsymbol{x}$." Statistically, averaging $\boldsymbol{u}$ over the distribution $P(\boldsymbol{u})$ leads to the counterfactual variables $\boldsymbol{Y}_{\boldsymbol{x}}$. We denote $P(\boldsymbol{Y}_{\boldsymbol{x}})$ a distribution over counterfactual variables $\boldsymbol{Y}_{\boldsymbol{x}}$. We use $P(\boldsymbol{y}_{\boldsymbol{x}})$ as a shorthand for probabilities $P(\boldsymbol{Y}_{\boldsymbol{x}} = \boldsymbol{y})$ when the identify of the counterfactual variables is clear.

We now introduce a family of DTRs which represent the exogenous variables $\boldsymbol{U}$ using partitions defined by the corresponding counterfactual variables. For any $k = 1, \dots, K-1$, let $S_{k+1_{\bar{\boldsymbol{\mathcal{X}}}_k}}$ denote a set of counterfactual variables $\{S_{k+1_{\bar{\boldsymbol{x}}_k}} : \bar{\boldsymbol{x}}_k \in \bar{\boldsymbol{\mathcal{X}}}_k\}$. Similarly, let $Y_{\bar{\boldsymbol{\mathcal{X}}}_K}$ denote a set $\{Y_{\bar{\boldsymbol{x}}_K} : \bar{\boldsymbol{x}}_K \in \bar{\boldsymbol{\mathcal{X}}}_K\}$. Further, we define $\bar{\boldsymbol{S}}_{k+1_{\bar{\boldsymbol{\mathcal{X}}}_k}}$ a set $\{S_1, S_{2_{\bar{\boldsymbol{\mathcal{X}}}_1}}, \dots, S_{k+1_{\bar{\boldsymbol{\mathcal{X}}}_k}}\}$.

Definition 1 (Counterfactual DTR)**.** A counterfactual dynamic treatment regime is a DTR $\langle \boldsymbol{U}, \{\bar{\boldsymbol{X}}_K, \bar{\boldsymbol{S}}_K, Y\}, \boldsymbol{F}, P(\boldsymbol{u})\rangle$ where for $k = 2, \dots, K$,

- The exogenous variables $\boldsymbol{U} = \{\bar{\boldsymbol{X}}_K, \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{\mathcal{X}}}_{K-1}}}, Y_{\bar{\boldsymbol{\mathcal{X}}}_K}\}$;
- Values of $S_1, \bar{\boldsymbol{X}}_K$ are drawn from $P(\bar{\boldsymbol{X}}_K, \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{\mathcal{X}}}_{K-1}}}, Y_{\bar{\boldsymbol{\mathcal{X}}}_K})$;
- Values of S_k are decided by a function $S_k \leftarrow \tau_k(S_{k_{\bar{\boldsymbol{\mathcal{X}}}_{k-1}}}, \bar{\boldsymbol{X}}_{k-1}) = S_{k_{\bar{\boldsymbol{X}}_{k-1}}}$;
- Values of Y are decided by a function $Y \leftarrow r(Y_{\bar{\boldsymbol{\mathcal{X}}}_K}, \bar{\boldsymbol{X}}_K) = Y_{\bar{\boldsymbol{X}}_K}$.

Give observational distribution $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) > 0$, we next construct a family of counterfactual DTRs $\mathcal{M}_{\text{OBS}}$ that are compatible with the observational distribution, i.e., for any $M \in \mathcal{M}_{\text{OBS}}$, $P^M(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) = P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y)$. First, any $M \in \mathcal{M}_{\text{OBS}}$, its exogenous distribution $P^M(\bar{\boldsymbol{X}}_K, \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{\mathcal{X}}}_{K-1}}}, Y_{\bar{\boldsymbol{\mathcal{X}}}_K})$ must satisfy the following decomposition:

$$P^M(\bar{\boldsymbol{X}}_K, \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{\mathcal{X}}}_{K-1}}}, Y_{\bar{\boldsymbol{\mathcal{X}}}_K}) = P^M(s_1) \prod_{\bar{\boldsymbol{x}}_K^y \in \bar{\boldsymbol{\mathcal{X}}}_K} P^M(Y_{\bar{\boldsymbol{x}}_K^y} | \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{X}}_K) P^M(\bar{\boldsymbol{X}}_K | \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{X}}_{K-1})$$
$$\cdot \prod_{k=1}^{K-1} \prod_{\bar{\boldsymbol{x}}_k^{k+1} \in \bar{\boldsymbol{\mathcal{X}}}_k} P^M(S_{k+1_{\bar{\boldsymbol{X}}_k^{k+1}}} | \bar{\boldsymbol{S}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{x}}_k) P^M(\bar{\boldsymbol{X}}_k | \bar{\boldsymbol{S}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{X}}_{k-1}).$$

Among quantities in the above equation, we define factors $P^M(s_1)$ as the observational probabilities $P(s_1)$, i.e, $P^M(s_1) = P(s_1)$. We further define conditional probabilities

$$P^M(y_{\bar{\boldsymbol{x}}_K} | \bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}_K) = P(y | \bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K), \qquad P^M(\bar{\boldsymbol{x}}_K | \bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}_{K-1}) = P(\bar{\boldsymbol{x}}_K | \bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1}),$$
$$P^M(s_{k+1_{\bar{\boldsymbol{x}}_k}} | \bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{x}}_k) = P(s_{k+1} | \bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k), \quad P^M(\bar{\boldsymbol{x}}_k | \bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{x}}_{k-1}) = P(\bar{\boldsymbol{x}}_k | \bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1}).$$

Other factors can be arbitrary conditional probabilities. It is verifiable that for any $M \in \mathcal{M}_{\text{OBS}}$, $P^M(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) = P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y)$. To witness,

$$\begin{aligned}
P^M(\bar{\boldsymbol{S}}_K, \bar{\boldsymbol{X}}_K, Y) &= \sum_{k=1}^{K-1} \sum_{\{Y_{\bar{\boldsymbol{x}}_K^y} : \bar{\boldsymbol{x}}_K^y \neq \bar{\boldsymbol{x}}_K\}} \sum_{\{S_{k+1_{\bar{\boldsymbol{x}}_k^{k+1}}} : \bar{\boldsymbol{x}}_k^{k+1} \neq \bar{\boldsymbol{x}}_k\}} P^M(\bar{\boldsymbol{X}}_K, \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{\mathcal{X}}}_{K-1}}}, Y_{\bar{\boldsymbol{\mathcal{X}}}_K}) \\
&= P^M(s_1) \prod_{\bar{\boldsymbol{x}}_K^y \in \bar{\boldsymbol{\mathcal{X}}}_K} \sum_{\{Y_{\bar{\boldsymbol{x}}_K^y} : \bar{\boldsymbol{x}}_K^y \neq \bar{\boldsymbol{x}}_K\}} P^M(Y_{\bar{\boldsymbol{x}}_K^y} | \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{X}}_K) P^M(\bar{\boldsymbol{X}}_K | \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}_{K-1}) \\
&\cdot \prod_{k=1}^{K-1} \prod_{\bar{\boldsymbol{x}}_k^{k+1} \in \bar{\boldsymbol{\mathcal{X}}}_k} \sum_{\{S_{k+1_{\bar{\boldsymbol{x}}_k^{k+1}}} : \bar{\boldsymbol{x}}_k^{k+1} \neq \bar{\boldsymbol{x}}_k\}} P^M(S_{k+1_{\bar{\boldsymbol{x}}_k^{k+1}}} | \bar{\boldsymbol{S}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{X}}_k) P^M(\bar{\boldsymbol{X}}_k | \bar{\boldsymbol{S}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{X}}_{k-1}) \\
&= P^M(S_1) P^M(Y_{\bar{\boldsymbol{x}}_K} | \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{X}}_K) P^M(\bar{\boldsymbol{X}}_K | \bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{X}}_{K-1}) \\
&\cdot \prod_{k=1}^{K-1} P^M(S_{k+1_{\bar{\boldsymbol{x}}_k}} | \bar{\boldsymbol{S}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{x}}_k) P^M(\bar{\boldsymbol{X}}_k | \bar{\boldsymbol{S}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{X}}_{k-1}).
\end{aligned}$$

By definitions of $\mathcal{M}_{\text{OBS}}$, we thus have that, for any $\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y$,

$$\begin{aligned}
P^M(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) &= P(s_1) P(y | \bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K) P(\bar{\boldsymbol{x}}_K | \bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1}) \prod_{k=1}^{K-1} P(s_{k+1} | \bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k) P(\bar{\boldsymbol{x}}_k | \bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1}) \\
&= P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y).
\end{aligned}$$

We will now use the constructions of $\mathcal{M}_{\text{OBS}}$ to prove the non-identifiability of $P_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y)$ in DTRs.

Theorem 4. *Given $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) > 0$, there exists DTRs M_1, M_2 such that $P^{M_1}(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) = P^{M_2}(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) = P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y)$ while $P^{M_1}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y) \neq P^{M_2}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y)$.*

Proof. We define two counterfactual DTRs $M_1, M_2 \in \mathcal{M}_{\text{OBS}}$ that are compatible with the observational distribution $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y)$. If $K = 1$, for any y, s_1, x_1 and any $x_1^y \neq x_1$, we define

$$P^{M_1}(y_{x_1^y} | s_1, x_1) = 0, \qquad P^{M_2}(y_{x_1^y} | s_1, x_1) = 1$$

It is verifiable that

$$P^{M_1}_{x_1}(s_1, y) = P(s_1, x_1, y), \qquad P^{M_2}_{x_1}(s_1, y) = P(s_1, x_1, y) + (1 - P(x_1 | s_1)) P(s_1)$$

Since $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) > 0$, we have $P^{M_2}_{x_1}(s_1, y) \neq P^{M_1}_{x_1}(s_1, y)$.

We now consider the case where $K > 1$. For any $\bar{\boldsymbol{x}}_K, \bar{\boldsymbol{s}}_K, y$, and any $\bar{\boldsymbol{x}}_K^y \neq \bar{\boldsymbol{x}}_K$, we define

$$P^{M_1}(y_{\bar{\boldsymbol{x}}_K^y} | \bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}_K) = 0 \tag{32}$$

By definitions, $P^{M_1}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y)$ is equal to the counterfactual quantities $P^{M_1}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_K})$. Thus,

$$\begin{aligned}
P^{M_1}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y) &= P^{M_1}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_K}, \bar{\boldsymbol{x}}_K) + \sum_{\bar{\boldsymbol{x}}'_K \neq \bar{\boldsymbol{x}}_K} P^{M_1}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_K}, \bar{\boldsymbol{x}}'_K) \\
&= P^{M_1}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_K}, \bar{\boldsymbol{x}}_K) + \sum_{\bar{\boldsymbol{x}}'_K \neq \bar{\boldsymbol{x}}_K} P^{M_1}(y_{\bar{\boldsymbol{x}}_K} | \bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}'_K) P^{M_1}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}'_K)
\end{aligned}$$

By the composition axiom, $\bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{x}}_{K-1}}} = \bar{\boldsymbol{S}}_K, Y_{\bar{\boldsymbol{x}}_K} = Y$ if $\bar{\boldsymbol{X}}_K = \bar{\boldsymbol{x}}_K$. Thus, $P^{M_1}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_K}, \bar{\boldsymbol{x}}_K) = P^{M_1}(\bar{\boldsymbol{s}}_K, y, \bar{\boldsymbol{x}}_K)$. Since $M_1 \in \mathcal{M}_{\text{OBS}}$, $P^{M_1}(\bar{\boldsymbol{s}}_K, y, \bar{\boldsymbol{x}}_K) = P(\bar{\boldsymbol{s}}_K, y, \bar{\boldsymbol{x}}_K)$. Together with Eq. (32), we can obtain

$$P^{M_1}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y) = P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y).$$

As for M_2, for any $\bar{\boldsymbol{x}}_{K-1}^K \neq \bar{\boldsymbol{x}}_{k-1}$, we define its factor

$$P^{M_2}(s_{K_{\bar{\boldsymbol{x}}_{K-1}^K}} | \bar{\boldsymbol{s}}_{K-1_{\bar{\boldsymbol{x}}_{K-2}}}, \bar{\boldsymbol{x}}_{K-1}) = 0$$

The above equation implies that for any $\bar{\boldsymbol{x}}'_{K-1} \neq \bar{\boldsymbol{x}}_{K-1}$,

$$\begin{aligned}
&P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_K}, \bar{\boldsymbol{x}}'_{K-1}) \\
&= P^{M_2}(y_{\bar{\boldsymbol{x}}_K} | \bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}'_{K-1}) P^{M_2}(s_{K_{\bar{\boldsymbol{x}}_{K-1}}} | \bar{\boldsymbol{s}}_{K-1_{\bar{\boldsymbol{x}}_{K-2}}}, \bar{\boldsymbol{x}}'_{K-1}) P^{M_2}(\bar{\boldsymbol{s}}_{K-1_{\bar{\boldsymbol{x}}_{K-2}}}, \bar{\boldsymbol{x}}'_{K-1}) \\
&= 0
\end{aligned} \tag{33}$$

For any $\bar{\boldsymbol{x}}^y_K \neq \bar{\boldsymbol{x}}_K$, we define

$$P^{M_2}(y_{\bar{\boldsymbol{x}}^y_K} | \bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}_K) = 1 \tag{34}$$

We will now show that the above equation implies that for any $x'_K \neq x_K$,

$$P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_k}, x'_K, \bar{\boldsymbol{x}}_{K-1}) = P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1}). \tag{35}$$

We first write $P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_k}, x'_K, \bar{\boldsymbol{x}}_{K-1})$ as:

$$P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_k}, x'_K, \bar{\boldsymbol{x}}_{K-1}) = P^{M_2}(y_{\bar{\boldsymbol{x}}_k} | \bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, x'_k, \bar{\boldsymbol{x}}_{K-1}) P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, x'_k, \bar{\boldsymbol{x}}_{K-1})$$

It is immediate from Eq. (34) that

$$P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_k}, X_k \neq x_k, \bar{\boldsymbol{x}}_{K-1}) = P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}_{K-1}).$$

By the composition axiom, $\bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{x}}_{K-1}}} = \bar{\boldsymbol{S}}_K$ if $\bar{\boldsymbol{X}}_{K-1} = \bar{\boldsymbol{x}}_{k-1}$. Since $M_2 \in \mathcal{M}_{\text{OBS}}$, we thus have:

$$P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_k}, X_k \neq x_k, \bar{\boldsymbol{x}}_{K-1}) = P^{M_2}(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1}) = P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1}).$$

We now turn our attention to the interventional distribution $P^{M_2}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y)$. By expanding on $\bar{\boldsymbol{X}}_K$,

$$\begin{aligned}
P^{M_2}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y) &= P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_k}, \bar{\boldsymbol{x}}_K) + P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_k}, X_k \neq x_k, \bar{\boldsymbol{x}}_{K-1}) \\
&\quad + P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_k}, \bar{\boldsymbol{X}}_{K-1} \neq \bar{\boldsymbol{x}}_{K-1})
\end{aligned}$$

The above equation, together with Eqs. (33) and (35), gives:

$$P^{M_2}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y) = P^{M_2}(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, y_{\bar{\boldsymbol{x}}_k}, \bar{\boldsymbol{x}}_K) + P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1}).$$

Again, by the composition axiom and $M_2 \in \mathcal{M}_{\text{OBS}}$,

$$P^{M_2}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y) = P^{M_2}(\bar{\boldsymbol{s}}_K, y, \bar{\boldsymbol{x}}_K) + P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1}) = P(\bar{\boldsymbol{s}}_K, y, \bar{\boldsymbol{x}}_K) + P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1}).$$

Since $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1}) > 0$, we have $P^{M_1}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y) \neq P^{M_2}_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K, y)$, which proves the statement. □

Lemma 1. *For a DTR, given $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y)$, for any $k = 1, \dots, K-1$,*

$$P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) - P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k) \leq P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) - P(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k).$$

Proof. Note that $P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1})$ can be written as the counterfactual quantity $P(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}})$. For any set of variables $\boldsymbol{V}$, let $\neg \boldsymbol{v}$ denote an event $\boldsymbol{V} \neq \boldsymbol{v}$. $P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1})$ could thus be written as:

$$P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) = P(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_k) + P(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}}, \neg x_k, \bar{\boldsymbol{x}}_{k-1}) + P(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}}, \neg \bar{\boldsymbol{x}}_{k-1}),$$

By the composition axiom, $\bar{\boldsymbol{S}}_{k+1_{\bar{\boldsymbol{x}}_k}} = \bar{\boldsymbol{S}}_{k+1}$ if $\bar{\boldsymbol{X}}_k = \bar{\boldsymbol{x}}_k$. So,

$$\begin{aligned}
P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) &= P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) + P(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}}, \neg x_k, \bar{\boldsymbol{x}}_{k-1}) + P(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}}, \neg \bar{\boldsymbol{x}}_{k-1}) \\
&\leq P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) + P(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_k}}, \neg x_k, \bar{\boldsymbol{x}}_{k-1}) + P(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_k}}, \neg \bar{\boldsymbol{x}}_{k-1}) \\
&= P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) + P(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_{k-1}) - P(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_k) + P(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_k}}) - P(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_{k-1}) \\
&= P(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_k}}) + P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) - P(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_k).
\end{aligned}$$

Again, by the composition axiom, $\bar{\boldsymbol{S}}_{k_{\bar{\boldsymbol{x}}_k}} = \bar{\boldsymbol{S}}_k$ if $\bar{\boldsymbol{X}}_k = \bar{\boldsymbol{x}}_k$. Since $P(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_k}}) = P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k)$,

$$P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) \leq P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k) + P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) - P(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k)$$

Rearranging the above equation proves the statement. □

Lemma 4. *For a DTR, given $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y)$, for any $k = 0, \dots, K-1$,*

$$P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) \leq \Gamma(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k),$$

where $\Gamma(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) = P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) - P(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k) + \Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1})$ and $\Gamma(s_1) = P(s_1)$.

Proof. We prove this statement by induction.

Base Case: $k = 0$ By definition, $\Gamma(s_1) = P(s_1)$. We thus have $P(s_1) \leq \Gamma(s_1)$.

Induction Step We assume that the statement holds for k, i.e., $P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) \leq \Gamma(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)$. We will prove that the statement holds for $k+1$, i.e., $P_{\bar{\boldsymbol{x}}_{k+1}}(\bar{\boldsymbol{s}}_{k+2}) \leq \Gamma(\bar{\boldsymbol{s}}_{k+2}, \bar{\boldsymbol{x}}_{k+1})$. To begin with,

$$P_{\bar{\boldsymbol{x}}_{k+1}}(\bar{\boldsymbol{s}}_{k+2}) = P_{\bar{\boldsymbol{x}}_{k+1}}(\bar{\boldsymbol{s}}_{k+2}) - P_{\bar{\boldsymbol{x}}_{k+1}}(\bar{\boldsymbol{s}}_{k+1}) + P_{\bar{\boldsymbol{x}}_{k+1}}(\bar{\boldsymbol{s}}_{k+1}).$$

By Lem. 1,

$$P_{\bar{\boldsymbol{x}}_{k+1}}(\bar{\boldsymbol{s}}_{k+2}) \leq P(\bar{\boldsymbol{s}}_{k+2}, \bar{\boldsymbol{x}}_{k+1}) - P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_{k+1}) + P_{\bar{\boldsymbol{x}}_{k+1}}(\bar{\boldsymbol{s}}_{k+1}).$$

Since $\bar{\boldsymbol{S}}_{k+1}$ are non-descendants of X_{k+1}, $P_{\bar{\boldsymbol{x}}_{k+1}}(\bar{\boldsymbol{s}}_{k+1}) = P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1})$. Since $P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) \leq \Gamma(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)$,

$$P_{\bar{\boldsymbol{x}}_{k+1}}(\bar{\boldsymbol{s}}_{k+2}) \leq P(\bar{\boldsymbol{s}}_{k+2}, \bar{\boldsymbol{x}}_{k+1}) - P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_{k+1}) + \Gamma(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) = \Gamma(\bar{\boldsymbol{s}}_{k+2}, \bar{\boldsymbol{x}}_{k+1}). \qquad \square$$

Theorem 5. *For a DTR, given $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) > 0$, for any $k = 1, \dots, K-1$,*

$$\frac{P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)}{\Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1})} \leq P_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) \leq \frac{\Gamma(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)}{\Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1})},$$

Proof. By basic probabilistic operations,

$$P_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) = \frac{P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1})}{P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k)}.$$

By Lem. 1,

$$P_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) \leq 1 + \frac{P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) - P(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k)}{P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k)}.$$

Since $P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) \leq P(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k)$, $P_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k)$ is upper-bounded when $P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k)$ is the maximal. Since $\bar{\boldsymbol{S}}_k$ are non-descendants of X_k, $P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k) = P_{\bar{\boldsymbol{x}}_{k-1}}(\bar{\boldsymbol{s}}_k)$. Together with Lem. 4, the above equation can be further bounded as:

$$P_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) \leq 1 + \frac{P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) - P(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k)}{\Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1})} = \frac{\Gamma(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)}{\Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1})}.$$

By definition, $P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) = P(\bar{\boldsymbol{s}}_{{k+1}_{\bar{\boldsymbol{x}}_k}})$. By basic probabilistic operations,

$$P_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) = \frac{P(\bar{\boldsymbol{s}}_{{k+1}_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_k) + P(\bar{\boldsymbol{s}}_{{k+1}_{\bar{\boldsymbol{x}}_k}}, \neg\bar{\boldsymbol{x}}_k)}{P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k)} \geq \frac{P(\bar{\boldsymbol{s}}_{{k+1}_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_k)}{P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k)}.$$

By the composition axiom, $\bar{\boldsymbol{S}}_{{k+1}_{\bar{\boldsymbol{x}}_k}} = \bar{\boldsymbol{S}}_{k+1}$ if $\bar{\boldsymbol{X}}_k = \bar{\boldsymbol{x}}_k$. Applying Lem. 4 again gives

$$P_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) \geq \frac{P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)}{P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k)} = \frac{P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)}{\Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1})}. \qquad \square$$

Theorem 6. *Given $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) > 0$, for any $k \in \{1, \dots, K-1\}$, let $P_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) \in [a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1}), b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1})]$ denote the bound given by Thm. 5. There exists DTRs M_1, M_2 such that $P^{M_1}(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) = P^{M_2}(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) = P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y)$ while $P^{M_1}_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) = a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1})$, $P^{M_2}_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) = b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1})$.*

Proof. Without loss of generality, we assume that $K > 1$. We consider two counterfactual DTRs $M_1, M_2 \in \mathcal{M}_{\text{OBS}}$ compatible with the observational distribution $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y)$, which we define at the beginning of this section. For all $i = 1, \dots, k-1$, for any $\bar{\boldsymbol{x}}_i^{i+1} \neq \bar{\boldsymbol{x}}_i$, we define that for any $M \in \{M_1, M_2\}$, its factors satisfy:

$$P^M(s_{{i+1}_{\bar{\boldsymbol{x}}_i^{i+1}}} | \bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_{i-1}}}, \bar{\boldsymbol{x}}_i) = 1. \tag{36}$$

Following a similar argument in Lem. 1, we will show that for any $M \in \{M_1, M_2\}$, for any $i = 1, \dots, k-1$,

$$P^M_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}) - P^M_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_i) = P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) - P(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_i). \tag{37}$$

By $P^M_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}) = P^M(\bar{\boldsymbol{s}}_{i+1_{\bar{\boldsymbol{x}}_i}})$ and basic probabilistic operations,

$$P^M_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}) = P^M(\bar{\boldsymbol{s}}_{i+1_{\bar{\boldsymbol{x}}_i}}, \bar{\boldsymbol{x}}_i) + P^M(\bar{\boldsymbol{s}}_{i+1_{\bar{\boldsymbol{x}}_i}}, X_i \neq x_i, \bar{\boldsymbol{x}}_{i-1}) + P^M(\bar{\boldsymbol{s}}_{i+1_{\bar{\boldsymbol{x}}_i}}, \bar{\boldsymbol{X}}_{i-1} \neq \bar{\boldsymbol{x}}_{i-1}).$$

By the composition axiom, $\bar{\boldsymbol{S}}_{i+1_{\bar{\boldsymbol{x}}_i}} = \bar{\boldsymbol{S}}_{i+1}$ if $\bar{\boldsymbol{X}}_i = \bar{\boldsymbol{x}}_i$. Since $M \in \mathcal{M}_{\text{OBS}}$, $P^M(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) = P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i)$. Therefore,

$$\begin{aligned}
P^M_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}) &= P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) + P^M(\bar{\boldsymbol{s}}_{i+1_{\bar{\boldsymbol{x}}_i}}, X_i \neq x_i, \bar{\boldsymbol{x}}_{i-1}) + P^M(\bar{\boldsymbol{s}}_{i+1_{\bar{\boldsymbol{x}}_i}}, \bar{\boldsymbol{X}}_{i-1} \neq \bar{\boldsymbol{x}}_{i-1}), \\
&= P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) + \sum_{x_i' \neq x_i} P^M(s_{i+1_{\bar{\boldsymbol{x}}_i}} | \bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_{i-1}}}, x_i', \bar{\boldsymbol{x}}_{i-1}) P(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_{i-1}}}, x_i', \bar{\boldsymbol{x}}_{i-1}) \\
&\quad + \sum_{\bar{\boldsymbol{x}}_{i-1}' \neq \bar{\boldsymbol{x}}_{i-1}} P^M(s_{i+1_{\bar{\boldsymbol{x}}_i}} | \bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_{i-1}}}, x_i, \bar{\boldsymbol{x}}_{i-1}') P(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_{i-1}}}, x_i, \bar{\boldsymbol{x}}_{i-1}')
\end{aligned}$$

By Eq. (36), $P^M(s_{i+1_{\bar{\boldsymbol{x}}_i}} | \bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_{i-1}}}, x_i', \bar{\boldsymbol{x}}_{i-1}) = P^M(s_{i+1_{\bar{\boldsymbol{x}}_i}} | \bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_{i-1}}}, x_i, \bar{\boldsymbol{x}}_{i-1}') = 1$, which gives

$$\begin{aligned}
P^M_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}) &= P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) + P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}, X_i \neq x_i, \bar{\boldsymbol{x}}_{i-1}) + P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}, \bar{\boldsymbol{X}}_{i-1} \neq \bar{\boldsymbol{x}}_{i-1}) \\
&= P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) + P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}, \bar{\boldsymbol{x}}_{i-1}) - P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}, \bar{\boldsymbol{x}}_i) + P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}) - P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}, \bar{\boldsymbol{x}}_{i-1}) \\
&= P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}) + P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) - P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}, \bar{\boldsymbol{x}}_i)
\end{aligned}$$

Again, by the composition axiom and $M \in \mathcal{M}_{\text{OBS}}$, $P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}, \bar{\boldsymbol{x}}_i) = P(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_i)$. Since $P^M(\bar{\boldsymbol{s}}_{i_{\bar{\boldsymbol{x}}_i}}) = P^M_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_i)$, we have

$$P^M_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}) = P^M_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_i) + P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) - P(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_i).$$

Rearranging the above equation proves Eq. (36). Following a similar induction procedure in the proof of Lem. 4, we have that for any $M \in \{M_1, M_2\}$,

$$P^M_{\bar{\boldsymbol{x}}_{k-1}}(\bar{\boldsymbol{s}}_k) = \Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1}). \tag{38}$$

As for M_1, for any $\bar{\boldsymbol{x}}_k^{k+1} \neq \bar{\boldsymbol{x}}_k$, we define

$$P^{M_1}(s_{k+1_{\bar{\boldsymbol{x}}_k^{k+1}}} | \bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{x}}_k) = 0$$

This implies

$$\begin{aligned}
P^{M_1}_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) &= P^{M_1}(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_k) + \sum_{\bar{\boldsymbol{x}}_k' \neq \bar{\boldsymbol{x}}_k} P^{M_1}(s_{k+1_{\bar{\boldsymbol{x}}_k}} | \bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{x}}_k') P^{M_1}(\bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{x}}_k') \\
&= P^{M_1}(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_k).
\end{aligned}$$

By the composition axiom and $M_1 \in \mathcal{M}_{\text{OBS}}$, $P^{M_1}(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}}, \bar{\boldsymbol{x}}_k) = P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)$, which gives

$$P^{M_1}_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) = P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k).$$

The above equation, together with Eq. (38), gives:

$$P^{M_1}_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) = \frac{P^{M_1}_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1})}{P^M_{\bar{\boldsymbol{x}}_{k-1}}(\bar{\boldsymbol{s}}_k)} = \frac{P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)}{\Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1})} = a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1}).$$

As for M_2, for any $\bar{\boldsymbol{x}}_k^{k+1} \neq \bar{\boldsymbol{x}}_k$, we define

$$P^{M_2}(s_{k+1_{\bar{\boldsymbol{x}}_k^{k+1}}} | \bar{\boldsymbol{s}}_{k_{\bar{\boldsymbol{x}}_{k-1}}}, \bar{\boldsymbol{x}}_k) = 1.$$

Following a similar procedure for proving Eq. (38), we have

$$P^M_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) = \Gamma(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k).$$

Thus,

$$P^{M_2}_{\bar{\boldsymbol{x}}_k}(s_{k+1} | \bar{\boldsymbol{s}}_k) = \frac{P^{M_2}_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1})}{P^{M_2}_{\bar{\boldsymbol{x}}_{k-1}}(\bar{\boldsymbol{s}}_k)} = \frac{\Gamma(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k)}{\Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1})} = b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1}). \qquad \square$$

Corollary 2. *For a DTR, given $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y) > 0$,*

$$\frac{E[Y|\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K]P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)}{\Gamma(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1})} \leq E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_k] \leq 1 + \frac{(E[Y|\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K] - 1)P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)}{\Gamma(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1})}.$$

Proof. By basic probabilistic operations,

$$E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_k] = \frac{E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K]P_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K)}{P_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K)}.$$

Note the counterfactual $Y_{\bar{\boldsymbol{x}}_K, \bar{\boldsymbol{s}}_K}(\boldsymbol{u}) \in [0, 1]$. Following a similar argument as Lem. 1,

$$E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K]P_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K) - P_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K) \leq E[Y|\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K]P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K) - P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K).$$

This implies

$$E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_k] \leq 1 + \frac{(E[Y|\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K] - 1)P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)}{P_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K)}$$

Since $E[Y|\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K] \leq 1$, $E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_k]$ is upper-bounded when $P_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K)$ is the maximal. Since $\bar{\boldsymbol{S}}_K$ are non-descendants of X_K, $P_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K) = P_{\bar{\boldsymbol{x}}_{K-1}}(\bar{\boldsymbol{s}}_K)$. By Lem. 4,

$$E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_k] \leq 1 + \frac{(E[Y|\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K] - 1)P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)}{\Gamma(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_{k-1})}.$$

By definition, $P_{\bar{\boldsymbol{x}}_K}(y, \bar{\boldsymbol{s}}_K) = P(y_{\bar{\boldsymbol{x}}_K}, \bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}})$. By basic probabilistic operations,

$$E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_k] \geq \frac{E[Y_{\bar{\boldsymbol{x}}_K}|\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}_K]P(\bar{\boldsymbol{s}}_{K_{\bar{\boldsymbol{x}}_{K-1}}}, \bar{\boldsymbol{x}}_K)}{P_{\bar{\boldsymbol{x}}_{K-1}}(\bar{\boldsymbol{s}}_K)}.$$

By the composition axiom, $\bar{\boldsymbol{S}}_{K_{\bar{\boldsymbol{x}}_{K-1}}} = \bar{\boldsymbol{S}}_{K-1}, Y_{\bar{\boldsymbol{x}}_K} = Y$ if $\bar{\boldsymbol{X}}_K = \bar{\boldsymbol{x}}_K$. Applying Lem. 4 gives

$$E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_k] \geq \frac{E[Y|\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K]P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)}{P_{\bar{\boldsymbol{x}}_K}(\bar{\boldsymbol{s}}_K)} = \frac{E[Y|\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K]P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)}{\Gamma(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_{K-1})}.$$

□

Proof of Theorems 7 and 8

Lemma 5. *Fix $\epsilon > 0$, $\delta \in (0, 1)$. With probability (w.p.) of at least $1 - \delta$, it holds for any $T > 1$, $R_\epsilon(T)$ of* `UC-DTR` *with parameter δ and causal bounds $\boldsymbol{\mathcal{C}}$ is bounded by*

$$R_\epsilon(T) \leq \min\left\{12K\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T_\epsilon \log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)}, \left\|\boldsymbol{\mathcal{C}}\right\|_1 T_\epsilon\right\} + 4K\sqrt{T_\epsilon \log(2T/\delta)}$$

Proof. Note that causal bounds $\boldsymbol{\mathcal{C}}$ is a set $\{\boldsymbol{\mathcal{C}}_1, \dots, \boldsymbol{\mathcal{C}}_K\}$ where for $k = 1, \dots, K-1$,

$$\begin{aligned} \boldsymbol{\mathcal{C}}_k &= \left\{\forall \bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k : \left[a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1}), b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1})\right]\right\}, \\ \text{and } \boldsymbol{\mathcal{C}}_K &= \left\{\forall \bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K : \left[a_{\bar{\boldsymbol{x}}_K, \bar{\boldsymbol{s}}_K}, b_{\bar{\boldsymbol{x}}_K, \bar{\boldsymbol{s}}_K}\right]\right\}. \end{aligned} \tag{39}$$

$\mathcal{M}^c$ is a set of DTRs such that for any $M \in \mathcal{M}^c$, its causal quantities $P_{\bar{\boldsymbol{x}}_k}(s_{k+1}|\bar{\boldsymbol{s}}_k)$ and $E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K]$ satisfy the causal bounds $\boldsymbol{\mathcal{C}}$, i.e.,

$$P_{\bar{\boldsymbol{x}}_k}(s_{k+1}|\bar{\boldsymbol{s}}_k) \in \left[a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1}), b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1})\right], \quad \text{and } E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K] \in \left[a_{\bar{\boldsymbol{x}}_K, \bar{\boldsymbol{s}}_K}, b_{\bar{\boldsymbol{x}}_K, \bar{\boldsymbol{s}}_K}\right]. \tag{40}$$

Let $\mathcal{M}_t^c = \mathcal{M}_t \cap \mathcal{M}^c$. Since $\mathcal{M}_t^c \subseteq \mathcal{M}_t$, following a similar argument in [4, C.1], we have

$$P(M^* \in \mathcal{M}_t^c) \leq P(M^* \in \mathcal{M}_t) \leq \frac{\delta}{4t^2}. \tag{41}$$

Since $\sum_{t=1}^{\infty} \frac{1}{4t^2} \leq \frac{\pi^2}{24}\delta < \frac{\delta}{2}$, it follows that with probability at least $1 - \frac{\delta}{2}$, $M^* \in \mathcal{M}_c^t$ for all episodes $t = 1, 2, \dots$.

Following the proof of Lem. 3, we have

$$\begin{aligned} R_\epsilon(T) \leq K\sqrt{6T_\epsilon \log(2T/\delta)} &+ \sqrt{\frac{3T_\epsilon \log(2T/\delta)}{2}} \\ &+ \sum_{k=1}^{K-1} \sum_{t \in L_\epsilon} V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k)}) - V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k+1)}) \qquad (42) \\ &+ \sum_{t \in L_\epsilon} \left(E_{\bar{\boldsymbol{X}}_K^t}^{M_t}[Y|\bar{\boldsymbol{S}}_K^t] - E_{\bar{\boldsymbol{X}}_K^t}[Y|\bar{\boldsymbol{S}}_K^t]\right). \qquad (43) \end{aligned}$$

It thus suffices to bound quantities in Eqs. (42) and (43) separately.

Bounding Eq. (42) By Eq. (22) and basic probabilistic operations,

$$
\begin{aligned}
&V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k)}) - V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k+1)}) \\
&= \sum_{s_{k+1}} (P^{M_t}(s_{k+1}|\bar{\boldsymbol{S}}_k, \bar{\boldsymbol{X}}_k) - P(s_{k+1}|\bar{\boldsymbol{S}}_k, \bar{\boldsymbol{X}}_k)) V_{\boldsymbol{\pi}_t}(s_{k+1}, \bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t) \\
&\leq \left\| P_{\bar{\boldsymbol{x}}_k}^{M_t}(\cdot|\bar{\boldsymbol{s}}_k) - P_{\bar{\boldsymbol{x}}_k}(\cdot|\bar{\boldsymbol{s}}_k) \right\|_1 \max_{s_{k+1}} V_{\boldsymbol{\pi}_t}(s_{k+1}, \bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t) \\
&\leq \min\left\{ 2\sqrt{6|\mathcal{S}_{k+1}| \log(2K|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k|T/\delta)} \frac{1}{\sqrt{\max\{1, N^t(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t)}}, \left\|\boldsymbol{\mathcal{C}}_k\right\|_1 \right\}
\end{aligned}
$$

The last step follows from Eqs. (16) and (40). We thus have

$$
\begin{aligned}
&\sum_{t \in L_\epsilon} V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k)}) - V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k+1)}) \\
&\leq \sum_{t \in L_\epsilon} \min\left\{ 2\sqrt{6|\mathcal{S}_{k+1}| \log(2K|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k|T/\delta)} \frac{1}{\sqrt{\max\{1, N^t(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t)}}, \left\|\boldsymbol{\mathcal{C}}_k\right\|_1 \right\} \\
&\leq \min\left\{ \sum_{t \in L_\epsilon} 2\sqrt{6|\mathcal{S}_{k+1}| \log(2K|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k|T/\delta)} \frac{1}{\sqrt{\max\{1, N^t(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t)}}, \sum_{t \in L_\epsilon} \left\|\boldsymbol{\mathcal{C}}_k\right\|_1 \right\} \\
&\leq \min\left\{ 2(\sqrt{2}+1)\sqrt{6T_\epsilon|\bar{\boldsymbol{\mathcal{S}}}_{k+1}||\bar{\boldsymbol{\mathcal{X}}}_k| \log(2K|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k|T/\delta)}, \left\|\boldsymbol{\mathcal{C}}_k\right\|_1 T_\epsilon \right\}
\end{aligned}
$$

The last step follows from results in [4, D] and $|L_\epsilon| = T_\epsilon$. Eq. (42) could thus be written as:

$$
\begin{aligned}
&\sum_{k=1}^{K-1} \sum_{t \in L_\epsilon} V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k)}) - V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k+1)}) \\
&\leq \sum_{k=1}^{K-1} \min\left\{ 2(\sqrt{2}+1)\sqrt{6T_\epsilon|\bar{\boldsymbol{\mathcal{S}}}_{k+1}||\bar{\boldsymbol{\mathcal{X}}}_k| \log(2K|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k|T/\delta)}, \left\|\boldsymbol{\mathcal{C}}_k\right\|_1 T_\epsilon \right\} \\
&\leq \min\left\{ \sum_{k=1}^{K-1} 2(\sqrt{2}+1)\sqrt{6T_\epsilon|\bar{\boldsymbol{\mathcal{S}}}_{k+1}||\bar{\boldsymbol{\mathcal{X}}}_k| \log(2K|\bar{\boldsymbol{\mathcal{S}}}_k||\bar{\boldsymbol{\mathcal{X}}}_k|T/\delta)}, \sum_{k=1}^{K-1} \left\|\boldsymbol{\mathcal{C}}_k\right\|_1 T_\epsilon \right\}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sum_{k=1}^{K-1} \sum_{t \in L_\epsilon} V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k)}) - V_{\boldsymbol{\pi}_t}(\bar{\boldsymbol{S}}_k^t, \bar{\boldsymbol{X}}_k^t; M_t^{(k+1)}) \\
&\leq \min\left\{ (K-1)2(\sqrt{2}+1)\sqrt{6T_\epsilon|\bar{\boldsymbol{\mathcal{S}}}||\bar{\boldsymbol{\mathcal{X}}}| \log(2K|\bar{\boldsymbol{\mathcal{S}}}||\bar{\boldsymbol{\mathcal{X}}}|T/\delta)}, \sum_{k=1}^{K-1} \left\|\boldsymbol{\mathcal{C}}_k\right\|_1 T_\epsilon \right\}.
\end{aligned}
\tag{44}
$$

Bounding Eq. (43) Since both M^*, M_t are in the set $\mathcal{M}_t^c$,

$$
\begin{aligned}
E^{M_t}_{\bar{\boldsymbol{X}}_K^t}[Y|\bar{\boldsymbol{S}}_K^t] - E_{\bar{\boldsymbol{X}}_K^t}[Y|\bar{\boldsymbol{S}}_K^t] &\leq \left| E^{M_t}_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K] - \hat{E}^t_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K] \right| + \left| E_{\bar{\boldsymbol{X}}_K^t}[Y|\bar{\boldsymbol{S}}_K^t] - \hat{E}^t_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K] \right| \\
&\leq \min\left\{ 2\sqrt{2\log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)} \frac{1}{\sqrt{\max\{1, N^t(\bar{\boldsymbol{S}}_K^t, \bar{\boldsymbol{X}}_K^t)\}}}, \left\|\boldsymbol{\mathcal{C}}_K\right\|_1 \right\}
\end{aligned}
$$

Eq. (43) can thus be written as:

$$
\begin{aligned}
&\sum_{t \in L_\epsilon} \left(E^{M_t}_{\bar{\boldsymbol{X}}_K^t}[Y|\bar{\boldsymbol{S}}_K^t] - E_{\bar{\boldsymbol{X}}_K^t}[Y|\bar{\boldsymbol{S}}_K^t] \right) \\
&\leq \sum_{t \in L_\epsilon} \min\left\{ 2\sqrt{2\log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)} \frac{1}{\sqrt{\max\{1, N^t(\bar{\boldsymbol{S}}_K^t, \bar{\boldsymbol{X}}_K^t)\}}}, \left\|\boldsymbol{\mathcal{C}}_K\right\|_1 \right\} \\
&\leq \min\left\{ \sum_{t \in L_\epsilon} 2\sqrt{2\log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)} \frac{1}{\sqrt{\max\{1, N^t(\bar{\boldsymbol{S}}_K^t, \bar{\boldsymbol{X}}_K^t)\}}}, \sum_{t \in L_\epsilon} \left\|\boldsymbol{\mathcal{C}}_K\right\|_1 \right\}.
\end{aligned}
$$

The last step follows from Eqs. (17) and (40). From results in [4, D], we have

$$\begin{aligned}&\sum_{t \in L_\epsilon} \left(E^{M_t}_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] - E_{\bar{\boldsymbol{X}}^t_K}[Y|\bar{\boldsymbol{S}}^t_K] \right) \\ &\leq \min\left\{ 2(\sqrt{2}+1)\sqrt{2T_\epsilon |\bar{\boldsymbol{\mathcal{S}}}||\bar{\boldsymbol{\mathcal{X}}}| \log(2K|\bar{\boldsymbol{\mathcal{S}}}||\bar{\boldsymbol{\mathcal{X}}}|T/\delta)}, \left\|\boldsymbol{\mathcal{C}}_K\right\|_1 T_\epsilon \right\}.\end{aligned} \tag{45}$$

Eqs. (44) and (45) together give:

$$\begin{aligned}R_\epsilon(T) \leq{}& K\sqrt{6T_\epsilon \log(2T/\delta)} + \sqrt{\frac{3T_\epsilon \log(2T/\delta)}{2}} \\ &+ \min\left\{ (K-1)2(\sqrt{2}+1)\sqrt{6T_\epsilon |\bar{\boldsymbol{\mathcal{S}}}||\bar{\boldsymbol{\mathcal{X}}}| \log(2K|\bar{\boldsymbol{\mathcal{S}}}||\bar{\boldsymbol{\mathcal{X}}}|T/\delta)}, \sum_{k=1}^{K-1} \left\|\boldsymbol{\mathcal{C}}_k\right\|_1 T_\epsilon \right\} \\ &+ \min\left\{ 2(\sqrt{2}+1)\sqrt{2T_\epsilon |\bar{\boldsymbol{\mathcal{S}}}||\bar{\boldsymbol{\mathcal{X}}}| \log(2K|\bar{\boldsymbol{\mathcal{S}}}||\bar{\boldsymbol{\mathcal{X}}}|T/\delta)}, \left\|\boldsymbol{\mathcal{C}}_K\right\|_1 T_\epsilon \right\}.\end{aligned} \tag{46}$$

A quick simplification gives:

$$R_\epsilon(T) \leq \min\left\{ 12K\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T_\epsilon \log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)}, \left\|\boldsymbol{\mathcal{C}}\right\|_1 T_\epsilon \right\} + 4K\sqrt{T_\epsilon \log(2T/\delta)}. \qquad \square$$

Theorem 7. *Fix a $\delta \in (0,1)$. With probability of at least $1-\delta$, it holds for any $T > 1$, the regret of* `UC`^{c}`-DTR` *with parameter δ and causal bounds $\boldsymbol{\mathcal{C}}$ is bounded by*

$$R(T) \leq \min\left\{ 12K\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T \log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)}, \left\|\boldsymbol{\mathcal{C}}\right\|_1 T \right\} + 4K\sqrt{T \log(2T/\delta)}.$$

Proof. Fix $\epsilon = 0$. Naturally, $T_\epsilon = T$ and $R_\epsilon(T) = R(T)$. By Lem. 5,

$$R(T) \leq \min\left\{ 12K\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T \log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)}, \left\|\boldsymbol{\mathcal{C}}\right\|_1 T \right\} + 4K\sqrt{T \log(2T/\delta)}. \qquad \square$$

Theorem 8. *For any $T \geq 1$, with parameter $\delta = \frac{1}{T}$ and causal bounds $\boldsymbol{\mathcal{C}}$, the expected regret of* `UC`^{c}`-DTR` *is bounded by*

$$E[R(T)] \leq \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^-_{\boldsymbol{\mathcal{C}}}} \left\{ \frac{33^2 K^2 |\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}| \log(T)}{\Delta_{\boldsymbol{\pi}}} + \frac{32}{\Delta^3_{\boldsymbol{\pi}}} + \frac{4}{\Delta_{\boldsymbol{\pi}}} \right\} + 1.$$

Proof. Let $\tilde{R}_\epsilon(T)$ denote the regret cumulated in ϵ-good episode up to T steps. By Eqs. (41) and (46),

$$\begin{aligned}E[R(T)] &\leq E[R_\epsilon(T) I_{M^* \in \mathcal{M}^c_t}] + E[\tilde{R}_\epsilon(T) I_{M^* \in \mathcal{M}^c_t}] + \sum_{t=1}^{T} P(M \notin \mathcal{M}^c_t) \\ &\leq \min\left\{ 12K\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T \log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T/\delta)}, \left\|\boldsymbol{\mathcal{C}}\right\|_1 T \right\} + 4K\sqrt{T \log(2T/\delta)} \\ &+ E[\tilde{R}_\epsilon(T) I_{M^* \in \mathcal{M}^c_t}] + \frac{\delta}{T} \\ &\leq 23K\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T_\epsilon \log(T/\delta)} + E[\tilde{R}_\epsilon(T) I_{M^* \in \mathcal{M}^c_t}] + \frac{\delta}{T}\end{aligned}$$

Fix $\delta = \frac{1}{T}$, it is immediate from Eq. (28) that

$$E[R(T)] \leq \frac{23^2 K^2 |\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}| \log(T^2)}{\epsilon} + E[\tilde{R}_\epsilon(T) I_{M^* \in \mathcal{M}^c_t}] + 1. \tag{47}$$

Note that when $M^* \in \mathcal{M}^c_t$, the maximal expected reward of any $\boldsymbol{\pi}_t$ over all instances in the family of DTRs $\mathcal{M}^c_t$ must be no less than the true optimal value $V_{\boldsymbol{\pi}^*}(M^*)$. In words, $\boldsymbol{\Pi}^-_{\boldsymbol{C}}$ is the effective policy space of `UC`^{c}`-DTR` procedure. Let $\Delta = \arg\min_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^-_{\boldsymbol{C}}} \Delta_{\boldsymbol{\pi}}$. Fix $\epsilon = \frac{\Delta}{2}$, Eq. (47) implies:

$$E[R(T)] \leq \frac{33^2 K^2 |\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}| \log(T)}{\Delta} + E[\tilde{R}_{\frac{\Delta}{2}}(T) I_{M^* \in \mathcal{M}^c_t}] + 1.$$

Among quantities in the above equation, $E[\tilde{R}_{\frac{\Delta}{2}}(T) I_{M^* \in \mathcal{M}^c_t}]$ can be bounded following a similar procedure in the proof of Thm. 2, which proves the statement. $\square$

Appendix II. Estimation of Causal Bounds

The bounds developed in the main text are functions of the observational distribution $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K, y)$ which is identifiable by the sampling process, and so can be estimated consistently. Bounding causal effects from a finite set of observations is more involved, due to the issues of sampling variability. We now present efficient methods to address these issues.

Given a finite set of observational samples $\{\bar{\boldsymbol{S}}^i_K, \bar{\boldsymbol{X}}^i_K, Y^i\}_{i=1}^n$, let $\hat{P}(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)$ denote the sample mean estimate of $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)$. Fix $\delta \in (0,1)$. W.p. at least $1-\delta$, the L1-deviation of the true distribution $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)$ and the empirical distribution $\hat{P}(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K)$ over state-action domains $\boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{X}}$ from n samples is bounded according to [9] by

$$\left\| P(\cdot) - \hat{P}(\cdot) \right\|_1 \leq \sqrt{2 |\boldsymbol{\mathcal{S}}| |\boldsymbol{\mathcal{X}}| \log(2/\delta)/n}. \tag{48}$$

We could derive confidence bounds of probabilities $P_{\bar{\boldsymbol{x}}_k}(s_{k+1}|\bar{\boldsymbol{s}}_k)$ for all $k = 1, \dots, K-1$ w.p. $1-\delta$ by optimizing the causal bounds $\left[a_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1}), b_{\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{s}}_k}(s_{k+1})\right]$ subject to convex polytope defined in Eq. (48) and probabilistic constraints $P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K) \in [0,1]$ and $\sum_{\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K} P(\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K) = 1$. The objective functions in Eq. (9) are ratios of linear functions, leading to a linear-fractional program (LFP). A LFP can be transformed into an equivalent linear program (LP) by [2], which is solvable using standard LP algorithms. The expected reward $E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K]$ could be bounded following a similar procedure.

Appendix III. Experimental Setup

In this section, we provide details about the setup of experiments in the main text. For all experiments, we test sequentially randomized trials (*rand*), `UC-DTR` algorithm (*uc-dtr*) and the causal `UC-DTR` (*uc*^{c}*-dtr*) with causal bounds derived from 1×10^5 observational samples. Each experiment lasts for $T = 1.1 \times 10^4$ episodes. The parameter $\delta = 1/KT$ for *uc-dtr* and *uc*^{c}*-dtr* where K is the total stages of interventions. For all algorithms, we measure their cumulative regret over 200 repetitions.

Random DTRs We generate 200 instances of the counterfacutal DTR defined in Def. 1. We assume treatments X_1, X_2, states S_1, S_2 and primary outcome Y are all binary variable. The probabilities of the counterfacutal distribution $P(s_1, x_1, s_{2_{x_1}}, x_{2_{x_1}}, y_{\bar{\boldsymbol{x}}_2})$ are drawn uniformly at random over $[0,1]$.

Cancer Treatment We test the survival model of patients inspired by the two-stage clinical trial conducted by the Cancer and Leukemia Group B [5, 8]. Protocol 8923 was a double-blind, placebo controlled two-stage trial reported by [7] examining the effects of infusions of granulocyte-macrophage colony-stimulating factor (GM-CSF) after initial chemotherapy. Patients were randomized initially to GM-CSF or placebo following standard chemotherapy. Later, patients meeting the criteria of complete remission were offered a second randomization to one of two intensification treatments.

We will describe this treatment procedure using the DTR with $K = 2$. $X_1, X_2 \in \{0, 1\}$ represent treatments; $S_1 = \emptyset$ and S_2 indicates the observed remission after the first treatment (0 stands for no remission and 1 for complete remission); Y indicates the survival of patients at the time of recording. The exogenous variable U is the age of patients where $U = 1$ if the patient is old and $U = 0$ otherwise. Values of U are drawn from a distribution $P(u)$ where $P(U = 1) = 0.2358$. Values of S_2 are drawn from a distribution $P_{x_1}(s_2)$ described in Table 1.

	$X_1 = 0$	$X_1 = 1$
$U = 0$	0.8101	0.0883
$U = 1$	0.7665	0.2899

Table 1: Probabilities of the distribution $P(S_2 = 1|u, x_1)$.

Let T_1, T_2 denote the potential survival time induced by treatment X_1, X_2 respectively. Values of T_1, T_2 are decided by functions defined as follows:

$$T_1 \leftarrow \min\{(1 - S_2)T_1^* + S_2(T_2^* + T_3^*), L\}, \quad T_2 \leftarrow \min\{(1 - S_2)T_1^* + S_2(T_2^* + T_4^*), L\}$$

where $L = 1.5$. Let $exp(\beta)$ denote an exponential distribution with mean $1/\beta$. Values of T_1^*, T_2^*, T_3^* are drawn from exponential distributions defined as follows:

$$T_1^* \sim exp(\beta^1_{u,x_1}), \qquad T_2^* \sim exp(\beta^2_{u,x_1}), \qquad T_3^* \sim exp(\beta^3_{u,x_1})$$

Given T_3^*, values of T_4^* are drawn from distribution

$$T_4^* \sim exp(\beta^3_{u,x_1} + \beta^4_{u,x_1} T_3^*).$$

The total survival time T of a patient is decided as follows:

$$T \leftarrow (1 - S_2)T_1 + S_2(1 - X_2)T_1 + S_2 X_2 T_2.$$

The parameters $\boldsymbol{\beta}_{u,x_1} = (\beta^1_{u,x_1}, \beta^2_{u,x_1}, \beta^3_{u,x_1}, \beta^4_{u,x_1})$ are described in Table 2.

		β^1_{u,x_1}	β^2_{u,x_1}	β^3_{u,x_1}	β^4_{u,x_1}
$U = 0$	$X_1 = 0$	4.3063	4.9607	0.8737	4.2538
	$X_1 = 1$	0.8286	8.2074	8.7975	7.6468
$U = 1$	$X_1 = 0$	2.6989	0.0235	5.9835	6.8059
	$X_1 = 1$	3.6036	1.1007	9.4426	7.3960

Table 2: Parameters $\boldsymbol{\beta}_{u,x_1}$.

The primary outcome Y is the survival of the patient at the time of observation $t = 1$. Values of Y are decided by the indicator function $Y \leftarrow I_{T>1}$.

We generate the confounded observational data following a sequence of decision rules $X_1 \sim \pi_1(X_1|U), X_2 \sim \pi_2(X_2|U, X_1, S_2)$. The policy $\pi_1(X_1|U)$ is a conditional distribution mapping from U to the domain of X_1 where $\pi_1(X_1 = 1|U = 0) = 0.5102$ and $\pi_1(X_1 = 1|U = 1) = 0.2433$. Similarly, $\pi_2(X_2|U, X_1, S_2)$ is a conditional distribution mapping from U, X_1, S_2 to the domain of X_2; Table 3 describes its parametrization.

	$X_1 = 0$		$X_1 = 1$	
	$S_2 = 0$	$S_2 = 1$	$S_2 = 0$	$S_2 = 1$
$U = 0$	0.2173	0.8696	0.6195	0.4641
$U = 1$	0.8869	0.0103	0.5314	0.4339

Table 3: Probabilities of $\pi_2(X_2 = 1|U, X_1, S_2)$.

References

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[2] A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research logistics quarterly*, 9(3-4):181–186, 1962.

[3] W. HOEFFDING. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Associ.*, 58(301):13–30, 1963.

[4] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[5] J. K. Lunceford, M. Davidian, and A. A. Tsiatis. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 58(1):48–57, 2002.

[6] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

[7] R. M. Stone, D. T. Berg, S. L. George, R. K. Dodge, P. A. Paciucci, P. Schulman, E. J. Lee, J. O. Moore, B. L. Powell, and C. A. Schiffer. Granulocyte–macrophage colony-stimulating factor after initial chemotherapy for elderly patients with primary acute myelogenous leukemia. *New England Journal of Medicine*, 332(25):1671–1677, 1995.

[8] A. S. Wahed and A. A. Tsiatis. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 60(1):124–133, 2004.

[9] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. 2003.