An Efficient Causal Algorithm for Adversarial Bandits

Arvind Raghavan (arvind.raghavan@columbia.edu)

May 30, 2022

Abstract

Adversarial bandits are important in safety-critical or risk-averse settings where reward distributions might change with great detriment to the learner. Recent variants of common algorithms, like the *LinEXP3* variant of the standard *EXP3*, explore how to improve regret guarantees by making parametric assumptions like linearity. We introduce a novel algorithm *CausalEXP3* that leverages knowledge of the underlying structural causal model to demonstrably reduce the cumulative regret incurred during on-line learning. We apply this experimentally in the context of Dynamic Treatment Regimes (DTR). First, we test it on a Lung Cancer DTR, involving heavy confounding between variables. Second, we test it on a Drug-Offences Intervention DTR under adversarial reward shift conditions. In both experiments, we demonstrate it outperforms the same *EXP3* algorithm that does not exploit structural causal knowledge.

1 Introduction

This project focuses on Adversarial Dynamic Treatment Regimes (DTR). DTRs are a powerful model for multi-stage interventions with complex dependencies on all past interventions and historical observations. For instance, in a multi-stage medical intervention such as chemotherapy sessions, the optimal dosage for each visit depends on all the past visits as well as symptoms observed at all past visits, as depicted in Figure 1. Other examples may be airline pricing strategies or prison rehabilitation programs. The outcome is measured after all the interventions: survival rate, airline revenue, 2-year recidivism etc. These can be highly risk-averse settings where we want to reduce worst-case risk and prepare for adverse reward shift through time.

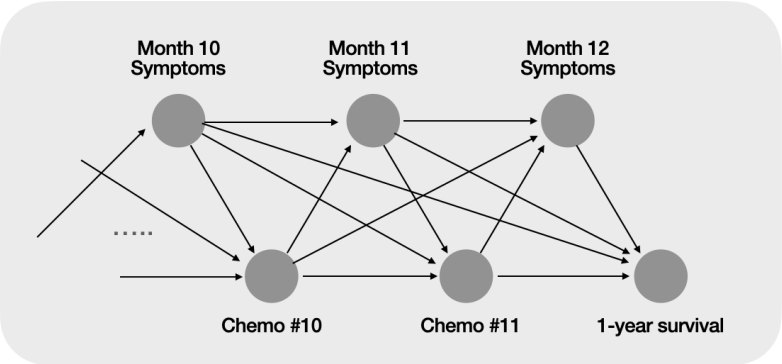

Figure 1: **Example of a medical DTR with complex historical dependencies**

Further, there could be arbitrary causal dependencies or unobserved confounding with past variables. Therefore, running regular Bandit algorithms would require that we fix all the interventions at the very beginning as one vector of intervention values and explore this space of joint actions (e.g. fix a vector of $\langle Chemo\#1, ..., Chemo\#11 \rangle$ at the start and explore this space). Alternatively, if we want to truly optimise the treatment at every stage based on all past variables (i.e. we know previous values when deciding an intervention) then we incur a massive exploration cost because we must account for the product of the cardinalities of domains of all past variables for each intervention.

Fortunately, we can make use of a Structural Causal Model (SCM) (7) to drastically reduce the policy space that we actually need to explore. Intuitively, if an oncologist informs us that last month's blood pressure and inflammation markers are not relevant *given* this month's WBC count, we can leverage that information to explore fewer arms. In doing so, we contribute to the growing field of Causal Reinforcement Learning, where incorporating inductive bias (expert knowledge in the form of a causal diagram) can improve existing algorithms.

2 Contributions

To the best of our knowledge, this project is the first to explicitly use SCMs for Adversarial Bandits, or in the Adversarial DTR setting. This project improves upon the well-known algorithm *EXP3* and introduces a new version, *CausalEXP3*, which uses importance-weighted exploration and the properties of Causal Bayesian Networks (1) to simultaneously achieve

- Better sample complexity to convergence in experiments using simulated data
- Robustness to reward shift over time (e.g. if the cancer metastasises in response to treatment)

Work in progress: a provably better theoretical regret bound than *EXP3*

3 Related Works

DTRs have been studied extensively since they were formulated in 2003. Chakraborty et al (2014) (2) offer a comprehensive survey of the field, including estimation methods.

This has been predominantly studied as a planning problem. If the model is fully known, efficient offline methods exist to compute optimal sequential treatment plans. Also, these methods typically assume there is *no causal confounding* in the MDP representing the DTR, which is a strong (and potentially fatal) assumption in many applications. Wang et al (2012) (9) illustrate how a patient's symptoms and cancer remission are often confounded by unobserved latent variables.

Tian (2008) (8) first proposed methods to compute causal effects in DTRs with unobserved confounding. Zhang & Bareinboim (2019) (10) use this to develop the first on-line algorithm for DTRs,

and in Zhang & Bareinboim (2020) (11) improve upon this by exploiting sparsity in dependencies. However, these only address the *stochastic* setting, and are unlikely to be robust in the face of *adversarial* reward shift. Hu & Kallus (2020) (3) present an interesting DTR Bandit solution for continuous variables, but only under linearity assumption and for a 2-step decision process.

On the Adversarial Bandits side, Neu & Olkhovskaya (2020) (6) recently introduced a series of improvements to the *EXP3* algorithm under the linear realizability assumption. We intend to follow a similar approach to theirs in improving the *EXP3* regret bound, except by incorporating causal assumptions instead of linearity assumptions.

4 Set Up

4.1 Structural Causal Models (SCM)

An SCM (7) is a tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$, where $\mathbf{U}$ refers to exogenous variables, $\mathbf{V}$ refers to endogenous variables and $\mathcal{F}$ refers to the set of functions that determines the values of $V \in \mathbf{V}$ from its parents and exogenous variables. **Bold** letters represent sets of variables. Upper-case letters refer to random variables and lower-case variables are a shorthand for those variables taking a specific value. E.g $P(\mathbf{x}, z_1, z_2)$ is shorthand for $P(\mathbf{X} = \mathbf{x}, Z_1 = z_1, Z_2 = z_2)$.

Each SCM is associated with a Directed Acyclic Graph (DAG), $\mathcal{G}$. Figure 2 shows a DAG for a sample DTR. By convention, we only depict endogenous variables ($\mathbf{V}$) in the DAG. Bidirected edges between nodes indicates *causal confounding*: the unobserved parents of these nodes are correlated (or they share noise variables).

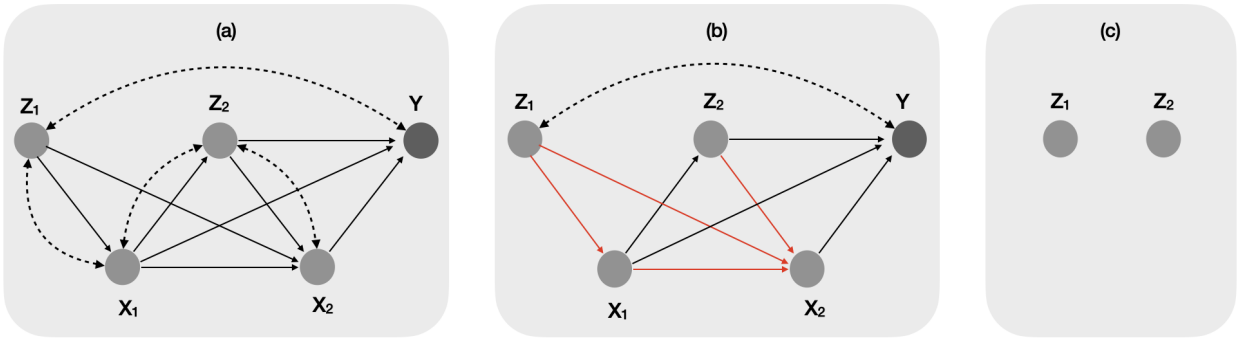

Figure 2: **(a) Graph $\mathcal{G}$ for a sample DTR; (b) Graph $\mathcal{G}_\pi$ for the DTR under a policy π from the policy space $\Pi = \{\langle Z_1 \to X_1 \rangle, \langle Z_1, X_1, Z_2 \to X_2 \rangle\}$; red arrows indicate that the policy decides X_i based on its historical context set; (c) $\mathcal{G}_{[\mathbf{Z}]}$ contains only $\{Z_1, Z_2\}$ and arrows between them**

4.2 Graphs for Dynamic Treatment Regimes (DTR)

The graph $\mathcal{G}$ induces a topological ordering corresponding to temporal ordering, with interventions $X_i \in \mathbf{X}$ indexed by the time-step of the intervention. Y represents the final loss outcome we want to minimize. $Z_i \in \mathbf{Z}$ refers to any covariate that is observed before X_i, and after X_{i-1} (if any). For each X_i, let $\mathbf{H_i}$ refer to the set of all "context" variables (historical covariates and past treatments) that are used to determine the intervention X_i. In Figure 2(a), $\mathbf{H_1} = \{Z_1\}$ and $\mathbf{H_2} = \{Z_1, Z_2, X_1\}$.

We review some graph-theoretic definitions we will use frequently (optional):

- $C(\mathbf{A})$ is the *c-component* (8) containing $\mathbf{A}$ in the graph specified. This includes all the nodes to which there is a path from $\mathbf{A}$ containing only bidirected arrows (including $\mathbf{A}$). E.g., in Figure 2(a) $C(Z_1)$ in $\mathcal{G}$ contains all nodes, while in 2(b) $C(Z_1)$ in $\mathcal{G}_\pi$ contains only $\{Z_1, Y\}$.
- $\mathcal{G}_{[\mathbf{A}]}$ is the sub-graph of $\mathcal{G}$ containing only $\mathbf{A}$ and its arrows. E.g., Figure 2(c) shows $\mathcal{G}_{[\mathbf{Z}]}$.
- $Pa_{\mathbf{A}}$ refers to the *direct parents* (7) of nodes in $\mathbf{A}$ (including $\mathbf{A}$ itself).
- Fix a topological ordering $\prec$ over $\mathcal{G}$. *Extended parents* (1) of a variable $V \in \mathbf{V}$ are defined as

  $$Pa_V^+ = Pa(\{V_i \in C(V) | V_i \preceq V\}) \setminus \{V\}, \text{ with } C(V) \text{ defined in } \mathcal{G}_{[\mathbf{V}]}$$

  In words, take the *c-component* containing V in $\mathcal{G}_{[\mathbf{V}]}$. Consider only the nodes up to V in the partial ordering, and their parents in $\mathcal{G}$. These are the *extended parents* of V. Note that we exclude V. E.g., for the graph $\mathcal{G}$ in Figure 2, $Pa_{Z_1}^+ = \emptyset$ and $Pa_{Z_2}^+ = \{X_1\}$

Ω_V refers to the domain of the variable V, and $|\Omega_V|$ is the cardinality of the domain (we only consider discrete variables in this project).

The policy space Π of the DTR is a set of mappings $\{\langle \Omega_{\mathbf{H}_i} \rightarrow \Omega_{X_i} \rangle\}$, to the domain of each intervention X_i from the domains of its context variables $\mathbf{H}_i$. I.e., this the set of Bandit "arms" to explore. If we are restricting Π to only deterministic policies, we can ignore previous interventions in each subsequent policy search. For Figure 2, a deterministic policy space would be

$$\Pi = \{\langle \Omega_{Z_1} \rightarrow \Omega_{X_1} \rangle, \langle \Omega_{Z_1, Z_2} \rightarrow \Omega_{X_2} \rangle\}$$

5 Efficient Exploration of DTR Policy Space

If we were running *EXP3* naively on the policy space Π the total number of arms we would need to explore is $\prod_i^n |\Omega_{\mathbf{H}_i} \rightarrow \Omega_{X_i}|$, which is massive for even moderately-sized discrete domains. However, we can use *d-separation* and *c-component factorization* to make this more efficient.

The efficient planning algorithms mentioned in Section 3 won't work here, since the model is unknown and we make no parametric assumptions about the data-generating process. We only know the graph $\mathcal{G}$ and the policy space Π, and need to learn on-line, assuming adversarial loss selection.

5.1 Removing Irrelevant Information

Zhang et al (2020) (11) propose a 2-step iterated procedure called *REDUCE* which removes "irrelevant" treatments and covariates that don't add any value, even if they are causal ancestors. These rules essentially exploit independence constraints (e.g., a doctor might say that, *given* last month's chemotherapy dosage and bio-markers, the dosage information for prior visits isn't needed).

The result is a minimal graph and policy space that contains only "relevant" treatments and covariates. In this project, we will assume that *REDUCE* has already been applied to the problem, and that the Π we work with is minimal. We also marginalise from the graph any variables that are not in $\mathbf{Z}$, $\mathbf{X}$ or Y, to get a minimal $\mathcal{G}$. Figure 2 shows one such minimal $\mathcal{G}$ and Π.

5.2 Decomposing the Expected Reward

Lemma 1: *Given a minimal DTR graph $\mathcal{G}$ and a policy π from the minimal policy space Π as defined in previous sections, we can express the expected reward under π as follows,*

$$E[Y|do(\pi)] = \sum_{\mathbf{x},\mathbf{z}} E[Y|do(\mathbf{x}), \mathbf{z}].P(\mathbf{z}|do(\mathbf{x})) \prod_{i=1}^{n} \pi(x_i|\mathbf{h}_i) \tag{1}$$

Refer to Appendix A for proof.

We know the probabilities of each x_i under a given policy π (if we are playing deterministic policies, the product term on the right would be 1 for some $X_i = x_i$ and 0 everywhere else). So we are effectively expressing the expected loss in terms of these quantities:

- $E[Y|do(\mathbf{x}), \mathbf{z}]$; and
- $P(\mathbf{z}|do(\mathbf{x}))$, for different possible values of $\mathbf{x}, \mathbf{z}$

We will proceed with the rest of the project, assuming for ease of exposition that $E[Y|do(\mathbf{x}), \mathbf{z}]$ is known and $P(\mathbf{z}|do(\mathbf{x}))$ is unknown. The same sampling procedure in our *CausalEXP3* algorithm 1 can be easily extended to discover $E[Y|do(\mathbf{x}), \mathbf{z}]$ as well.

Lemma 2: *Given a minimal DTR graph $\mathcal{G}$ and a policy π from the minimal policy space Π as defined in previous sections, we can express the unknown quantity from Equation 1 as follows,*

$$P(\mathbf{z}|do(\mathbf{x})) = \prod_{Z_i \in \mathbf{Z}} P(z_i|do(\mathbf{x}^{i-}), pa_{Z_i}^{+}) \tag{2}$$

where $\mathbf{X}^{i-}$ refers to the interventions temporally preceding Z_i.

This is a straightforward consequence of Tian (2008, Theorem 1) (8) and Zhang et al (2020, Corollary 2) (11). We now believe it is more intuitive to express such decompositions using the *Semi-Markovian factorization* notation in Bareinboim et al (2020, Definition 15) (1), in terms of *extended parent* sets.

6 *CausalEXP3* Algorithm

In the *LinEXP3* algorithm (6), Neu & Olkhovskaya parameterize the loss for each arm as a linear function with an unknown parameter. With each pull of an arm, we get information about other arms in all non-orthogonal directions in the linear parameter space. We follow the same principle with *CausalEXP3*, where each pull of an arm allows us to get more information by belief propagation in the graph via the decomposition given by Eq. 2, at Steps 4 and 5 below.

Algorithm 1 CausalEXP3

Input:

- Minimal graph $\mathcal{G}$, containing only interventions $\mathbf{X}$, covariates $\mathbf{Z}$, loss outcome Y
- Minimal deterministic policy space Π
- Learning rate η

Define:

- $n_t(\mathbf{a})$: empirical frequency of any event $(\mathbf{A} = \mathbf{a})$ up to episode t
- $N = |\Pi|$, the number of policy arms to explore
- $w_t \in \mathbb{R}^N$: a vector of weights assigned to each policy at episode t

Initialize: $w_1 = (1, 1, 1, 1...)$

for episode $t = 1, 2...T$ **do**

1. Let
$$p_t(j) = \frac{w_t(j)}{\sum_{j'=1}^{N} w_t(j')}$$
2. Sample an arm $\pi_t \sim p_t$
3. Perform $do(\pi_t)$ and observe $\mathbf{x}_t, \mathbf{z}_t$
4. For each $Z_i \in \mathbf{Z}$, compute the empirical estimate for the estimands in Eq. 2
$$\hat{P}_t(z_i|do(\mathbf{x}^{i-}), pa^+_{Z_i}) = \frac{n_t(z_i, pa^+_{Z_i})}{\max\{n_t(pa^+_{Z_i}), 1\}}$$
5. For each $\pi \in \Pi$, compute the loss estimate by Eq. 1, and apply importance-weighting
$$l_t(j) = \sum_{\mathbf{x},\mathbf{z}} E[Y|do(\mathbf{x}), \mathbf{z}].\hat{P}(\mathbf{z}|do(\mathbf{x})) \prod_{i=1}^{n} \pi_j(x_i|\mathbf{h}_i)$$
$$\tilde{l}_t(j) = \frac{l_t(j)}{p_t(j)}$$
6. Update $w_{t+1}(j) \leftarrow w_t(j) \exp(-\eta.\tilde{l}_t(j))$

end for

6.1 Comparison with *Naive EXP3*: Policy Search

For a full description of a *Naive EXP3* algorithm for DTRs, refer to Appendix C. We illustrate the difference between the two algorithms using the example DTR in Figure 2. Let us assume all variables are binary.

Fact 3: *Given a **deterministic** policy space Π, for a DTR with n interventions, the number of policy arms to explore (without making any parametric assumptions) is*

$$|\Pi| = \prod_{i=1}^{n} \Omega_{X_i}^{\Omega_{H_i \setminus X^{i-}}}$$

where $(H_i \setminus X^{i-})$ is the set of covariates with arrows into X_i.

For the example DTR in Figure 2, we have:

- Deterministic policy space, $\Pi = \{\langle \Omega_{Z_1} \rightarrow \Omega_{X_1} \rangle, \langle \Omega_{Z_1,Z_2} \rightarrow \Omega_{X_2} \rangle\}$
- Number of policies, $|\Pi| = (|X_1|^{|Z_1|}).(|X_2|^{|Z_1|.|Z_2|}) = (2^2).(2^4) = 256$

Table 1: **Difference in policy search under both algorithms for Figure 2 DTR**

	Naive EXP3	*CausalEXP3*
Policy arms	256	256
Updates per episode	1	256
Updates per episode for	Only π_t	All $\pi \in \Pi$
Update method	Importance-weighting	Importance-weighting
Update using	Actual loss $(l_t \vert \pi_t)$	Expected loss $\hat{E}[l_t \vert \pi]$

6.2 Comparison with *Naive EXP3*: Run-Time

Unfortunately, *CausalEXP3* incurs a hefty cost for updating estimators and the whole weight vector in each episode t, as illustrated in Table 2

6.3 Comparison with *Naive EXP3*: Regret Bounds

Fact 4: *The expected regret bound for Naive EXP3, over a deterministic DTR policy space Π, when setting learning rate $\eta = \sqrt{log|\Pi|/(|\Pi|T)}$ is*

$$\mathbb{E}[\text{TotalRegret}_T] \leq \sqrt{2|\Pi|T.\log|\Pi|}$$

Table 2: **Difference in run-time complexity of each algorithms for Figure 2 DTR**

	Naive EXP3	*CausalEXP3*
Sampling π_t	$\log \lvert\Pi\rvert = 6$	6
Querying DTR's SCM	1 (constant)	1
Updating estimators	-	$\sum_i \lvert\Omega_{Z_i}\rvert.\lvert\Omega_{\mathbf{X}^{i-}}\rvert = 2.1 + 2.2 = 6$
Weight updates	1 (constant)	$\sim \lvert\Pi\rvert.\lvert\Omega_{\mathbf{X}}\rvert.\lvert\Omega_{\mathbf{Z}}\rvert = 256.4.4 = 4096$
Total run-time per t	**8**	$\sim$ **4109**

where $|\Pi| = \prod_{i=1}^{n} \Omega_{X_i}^{\Omega_{H_i \setminus X^{i-}}}$*, when we don't make any parametric assumptions.*

Unfortunately, we have not yet proved a theoretical regret guarantee for *CausalEXP3*. We are working on it and are confident that the regret is provably lower, based on the following experiments.

7 Experiments

We conducted two sets of experiments. Coding a multi-step DTR proved tricky and buggy, so we chose experiments with specific goals in mind.

- The first experiment implements a DTR for **Lung Cancer Treatment**, with the goal of gauging performance when there is heavy confounding among variables (we hide relevant variables to mimic unobserved confounding).
- The second experiment is a DTR for **Drug-Offence Correctional Interventions**. We assume no confounding but rather test performance under an adversarial reward shift, by switching data-generating probabilities midway through an epoch (graph remains the same).

7.1 Lung Cancer Treatment

7.1.1 DTR Description

In this experiment, we implement a popular DTR for lung-cancer, introduced by Nease Jr & Owens (1997) (5). The verbal description of their proposed multi-stage treatment is in Appendix D. This multi-stage proposal was used to build the SCM for this experiment.

Figure 3(a) is the graph for the multi-stage treatment regime described in Appendix D. Table 3 details the variable labels and domains. We consistently use 0 to mean "Yes", 1 to mean "No" and 2 to mean "N/A". For any variable H, we use the notation $P(h_0)$ to refer to $P(H = 0)$.

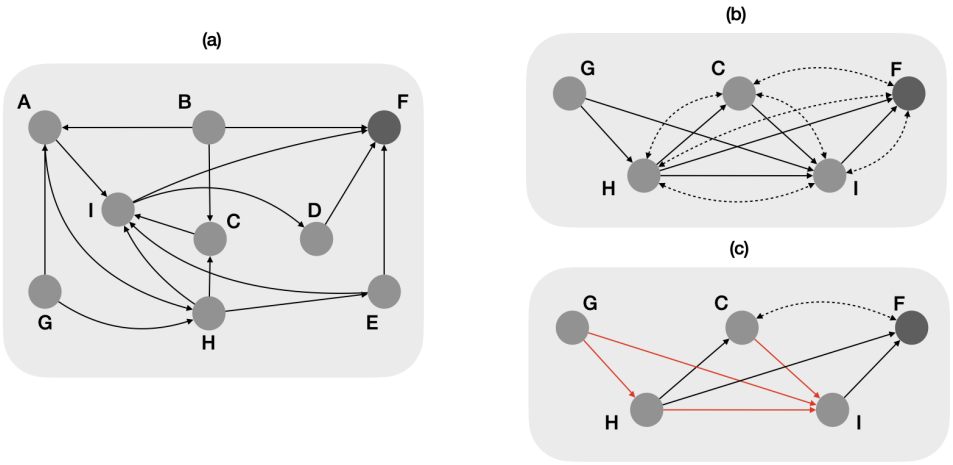

Figure 3: **(a) Graph for Lung Cancer DTR in Nease & Owens (1997), with variables described in Table 3; (b) Marginalized graph with Treatments =** $\{H, I\}$**, Covariates =** $\{G, C\}$**, Outcome =** $\{F\}$**; (c) Graph under policy from deterministic space** $\Pi = \{\langle \Omega_G \to \Omega_H \rangle, \langle \Omega_{G,C} \to \Omega_I \rangle\}$

Table 3: **Variables for the Lung Cancer DTR depicted in Figure 3**

Variable	Description	Domain
A	CT Result	0,1,2
B	Mediastinal Metastases	0,1
C	Mediastinoscopy Result	0,1,2
D	Treatment Death	0,1
E	Mediastinoscopy Death	0,1
F	**Life Expectancy**	0,1
G	CT?	0,1
H	Mediastinoscopy?	0,1
I	Treatment?	0,1

We only retain the **covariates** (G, C), the **intervention** decisions (H, I) and fix Life Expectancy (F) as the **outcome** we want to optimise for. We treat the other variables as confounders, inducing a marginalized DTR graph in Figure 3(b). The candidate policy space is $\Pi = \{\langle \Omega_G \to \Omega_H \rangle, \langle \Omega_{G,C} \to \Omega_I \rangle\}$. Recall that we are only considering deterministic policies in this project, so for the intervention on (I) we don't have to optimize for (H) as it is fully determined by (G). However, (H) still factors in the decision to implement (I) or not. The intervention graph is in Figure 3(c).

7.1.2 Experiment Steps

1. We initialize a random SCM compatible with the graph in Figure 3. For reference, the data-generating probabilities we used are listed in Appendix E, Table 5

2. We define our candidate policy space as
    - $\Pi = \{\langle \Omega_G \to \Omega_H \rangle, \langle \Omega_{G,C} \to \Omega_I \rangle\}$
    - Number of policy "arms", $|\Pi| = 2^2.2^4 = 256$
3. We recognize the optimal policy π^* (according to Appendix E, Table 6) as
    - $\pi^* = do(h_1, i_1)$
    - $E[F|do(\pi^*)] = 0.5891$
4. We define regret of a policy π_t vs. optimal policy as $\text{Regret}_\pi = 0.5891 - E[F|do(\pi_t)]$
5. We run for a total of $T = 10,001$ episodes,
    - *Naive EXP3*, updating exactly one weight value (for our sampled π_t) in each episode
    - *CausalEXP3* with our decomposition estimands $E[F|do(h,i),c], \hat{P}(g), \hat{P}(c|do(h))$
    - We use a learning rate, $\eta = 1/T$
6. We track the following: **one-step reward** and **cumulative regret**, every 500 episodes
7. We repeat this experiment 100 times, and average over the runs to give us stable results

7.1.3 Results and Analysis

The parameters and metrics we track are:

- Number of runs = 100
- Number of episodes per run, $T = 10,001$
- Instantaneous Reward at episode t, $\text{Reward}_t = f_t$
- Cumulative Regret at episode t, $\text{TotalRegret}_t = \sum_t (0.5891 - f_t)$

Note that TotalRegret_t is an unbiased estimator of the total regret of policies chosen until that time-step, $\sum_t \text{Regret}_{\pi_t}$.

Figure 4 shows the Instantaneous Reward plotted over the training sequence (averaged over 100 runs). As we can see, the results are noisy, but the regression line shows no increase in Reward for *Naive EXP3*, and a steady increase for *CausalEXP3*.

Figure 5 shows a clearer illustration of the advantage. Cumulative Regret (averaged over 100 runs) for *Naive EXP3* is **linear** in t, essentially as good as **random guesses** even after 10,000 episodes. However, the same for *CausalEXP3* shows a **sub-linear** dependency on t and appears to flatten out. This means that *CausalEXP3* will very likely **converge to the optimal** policy, given time.

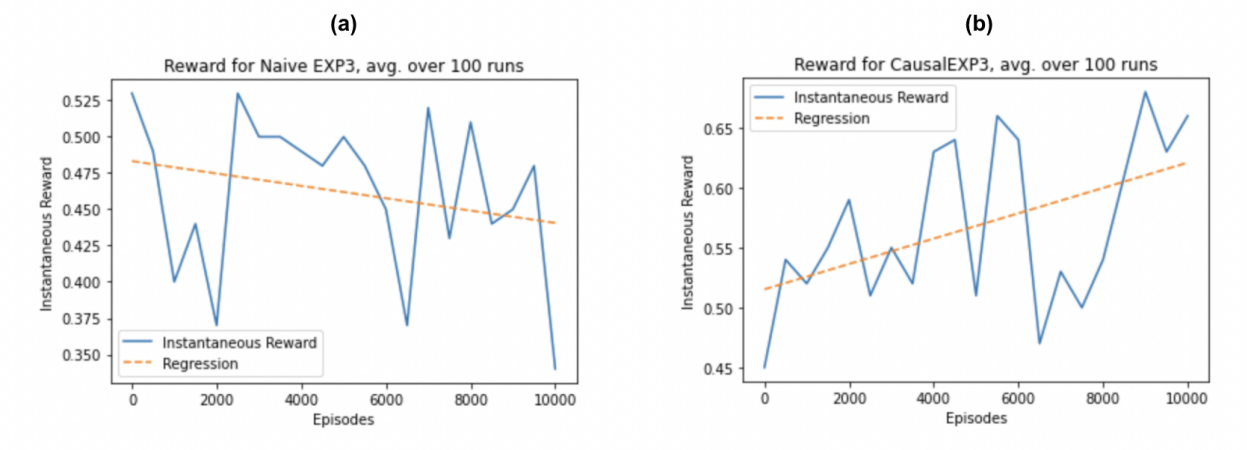

Figure 4: **Reward per episode t, averaged over 100 runs; trend-line shows (a) Reward is not increasing using *Naive EXP3*; (b) Reward is increasing on average using *CausalEXP3***

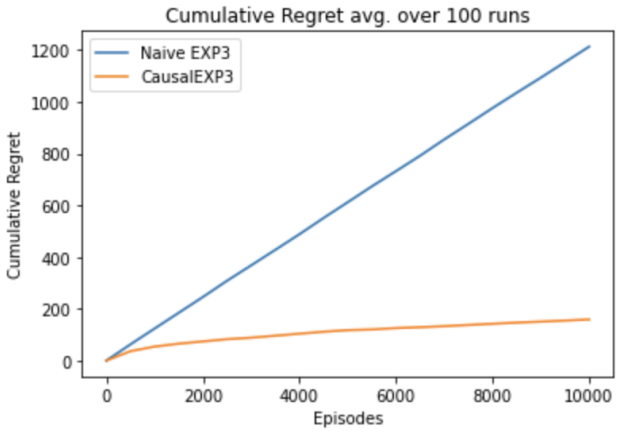

Figure 5: **Cumulative Regret is almost linear in t for Naive *EXP3*, but is sub-linear and heading to convergence for *CausalEXP3***

7.2 Drug-Offence Corrections Program

7.2.1 DTR Description

For this experiment, we devise a DTR for drug-court interventions, based on the adaptive treatment proposed by Marlowe et al (2008) (4). To keep things computationally tractable, we deviate slightly in our SCM design from the description of the original proposal, available in Appendix F.

Figure 6 is the graph for the multi-stage treatment regime described in Appendix F. Table 4 details the variable labels and domains. We consistently use 0 to mean "No", 1 to mean "Yes", except for variable D where numbers index the intervention options. For any variable X, we use the notation $P(x_0)$ to refer to $P(X = 0)$.

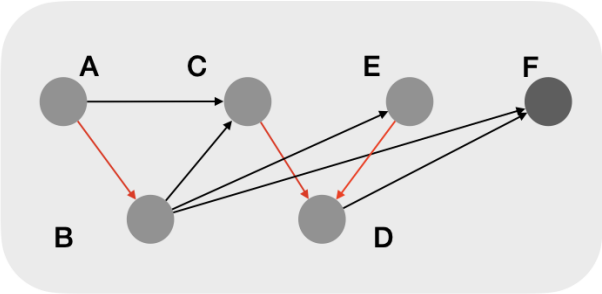

Figure 6: **Graph for Drug Corrections DTR in Marlowe et al (2008), with variables described in Table 4; We assume no confounding, but focus on testing adversarial reward shifts in this experiment (e.g. drug-offenders start gaming the system after feedback from peers)**

Table 4: **Variables for the Drug-Offence Corrections DTR depicted in Figure 6**

Variable	Description	Domain
A	High Risk	0,1
B	Court Hearings?	0,1
C	Counselling Session Compliance	0,1
D	Bi-weekly court sessions, As-needed, or Jeopardy Contract?	0,1,2
E	Responsive to Drug Test	0,1
F	**Rehabilitation within Budget**	0,1

For this DTR, the **covariates** are (A, C, E), the **intervention** decisions are (B, D) and our **outcome** of interest is (F), whether the offender is rehabilitated within the per-person budget. The candidate policy space is $\Pi = \{\langle \Omega_A \to \Omega_B \rangle, \langle \Omega_{C,E} \to \Omega_D \rangle\}$.

7.2.2 Experiment Steps

1. We initialize **two random SCMs** compatible with the graph in Figure 6. For half the training lifetime, we will assume the DTR operates according to SCM1, and half-way through we will **switch to SCM2 to mimic adversarial reward shift** (e.g. the community of drug offenders starts trying to game the system). For reference, the data-generating SCM probabilities are listed in Appendix G, Tables 7, 9

2. We define our candidate policy space as

   - $\Pi = \{\langle \Omega_A \to \Omega_B \rangle, \langle \Omega_{C,E} \to \Omega_D \rangle\}$
   - Number of policy "arms", $|\Pi| = 2^2.3^4 = 324$

3. We recognize the optimal policy π^* (according to Appendix G, Tables 8, 10) as

   - $\pi^* = do(b_1, d_2)$:- impose Court Hearings + Jeopardy Contracts

- $E[F|do(\pi^*)] = 0.7211$

4. We define regret of a policy π_t vs. optimal policy as $\text{Regret}_\pi = 0.7211 - E[F|do(\pi_t)]$
5. We run for a total of $T = 10,001$ episodes,
    - *Naive EXP3*, updating the weight value for our sampled π_t in each episode
    - *CausalEXP3* with estimands $E[F|do(b,d)], \hat{P}(a), \hat{P}(c|do(b),a), \hat{P}(e|do(b))$
    - We use a learning rate, $\eta = 1/T$
    - **For t <=5000, we sample from SCM1; for t > 5000 we sample from SCM2** (simulating adversarial reward shift)
6. We track the **one-step reward** and **cumulative regret**, every 500 episodes
7. We repeat this experiment 50 times, and average over the runs to give us stable results

7.2.3 Results and Analysis

The parameters and metrics we track are:

- Number of runs = 50
- Number of episodes per run, $T = 10,001$
- Instantaneous Reward at episode t, $\text{Reward}_t = f_t$
- Cumulative Regret at episode t, $\text{TotalRegret}_t = \sum_t (0.7221 - f_t)$

Figure 7 shows the Instantaneous Reward plotted over the training sequence (averaged over 50 runs). The results are noisy, but the regression line shows that *Naive EXP3* takes a hit from the adversarial reward shift at $t = 5000$ from which it does not recover. *CausalEXP3* also drops due to the shift, but manages to slowly **climb up** thereafter.

Figure 8 corroborates the causal advantage. Cumulative Regret (averaged over 50 runs) for *Causal-EXP3* is **below that of *Naive EXP3*** after 10,000 episodes. However, it is worrying that *Causal-EXP3*'s regret seems linear in t after the reward shift, instead of sub-linear. This could perhaps be improved by building estimators that detect any change in the transition probabilities we are tracking, so that *CausalEXP3* doesn't continue to use outdated estimates for its updates.

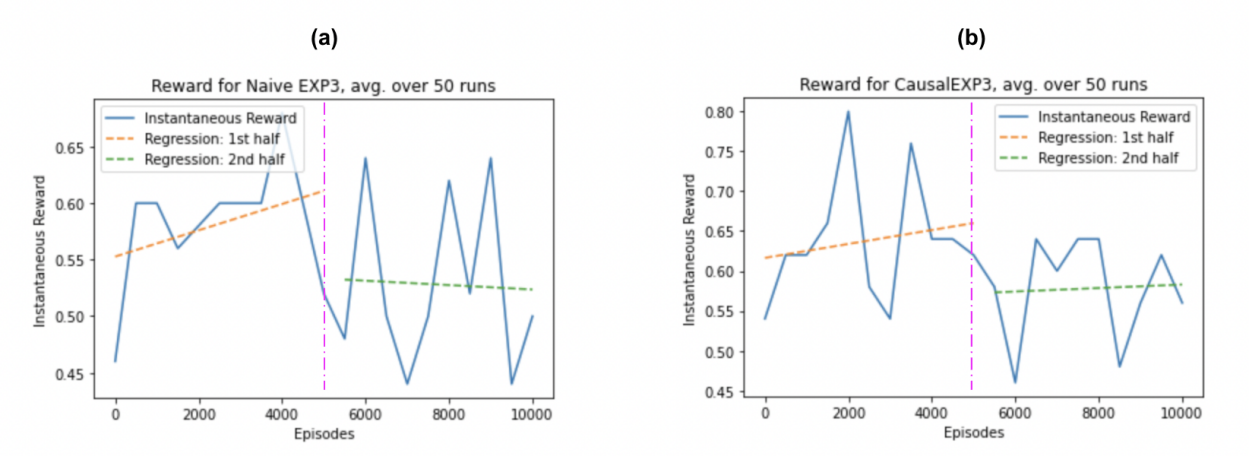

Figure 7: **Reward per episode t, averaged over 50 runs; adversarial reward shift happens at $t = 5000$; trend-line shows (a) Reward drops and decreases in 2nd half for *Naive EXP3*; (b) Reward drops in 2nd half but still increases slowly for *CausalEXP3***

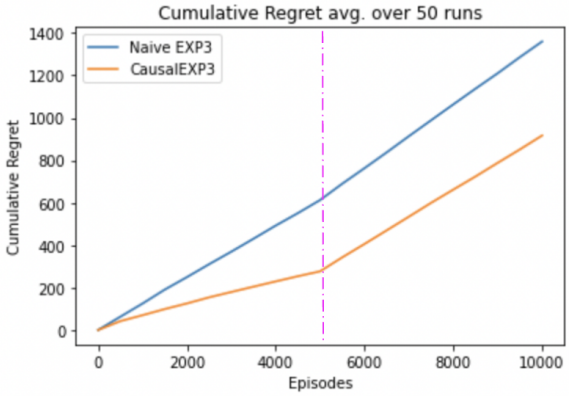

Figure 8: **Cumulative Regret is definitely lower for *CausalEXP3*; however, regret doesn't seem to converge for *CausalEXP3* after reward shift, showing that *CausalEXP3* is still relying on outdated estimates for its weight updates**

8 Limitations and Future Work

Our experiment on adversarial reward shift was limited to just one transition, since this proved challenging to code and was heavy on compute (>2 hours). We intend to continue this exploration of reward shift, especially on trying to build estimators that detect such shifts and refresh the estimates being tracked. The current algorithm may incur heavy bias over time if using estimates that don't re-set occasionally. Another easy extension is applying the causal decomposition to *FTRL* and *FTPL* algorithms which are close cousins of *EXP3* with attractive properties.

More pressingly, the experimental results are still not satisfying for a real-world application (we typically can't experiment with over 2000 patients to hone in on the right dosage of chemotherapy). Infusing parametric assumptions about the functional forms or monotonicity under treatment may yield more powerful performance guarantees.

9 Conclusion

DTRs are a powerful framework for complex, multi-stage interventions such as medicine, epidemiology and finance. However, current on-line algorithms are either very restrictive in their parametric assumptions, limited in time-steps, or assume stable reward distributions. The last assumption in particular is worrying, since applications like epidemiology could well see adversarial reward shift (e.g. a strain of disease developing antibiotic resistance with aggressive treatment).

We propose a novel algorithm, *CausalEXP3*, to address this gap in the literature. *CausalEXP3* uses the constraints in the causal graph to significantly reduce the amount of exploration actually needed. We applied this experimentally to DTRs for Lung Cancer Treatment and Drug-Offence Corrections and showed that it outperformed a naive implementation of the same algorithm, despite heavy confounding and adversarial reward shifts.

References

[1] Bareinboim, Elias, Juan Correa, Duligur Ibeling and Thomas Icard. 2020. "On Pearl's Hierarchy and the Foundations of Causal Inference." In *Probabilistic and Causal Inference: The Works of Judea Pearl, ACM Turing Series*

[2] Chakraborty, Bibhas, and Susan A Murphy. 2014. "Dynamic Treatment Regimes." In *Annual review of statistics and its application* vol. 1: 447-464. doi:10.1146/annurev-statistics-022513-115553

[3] Hu, Yichun and Nathan Kallus. 2020. "DTR Bandit: Learning to Make Response-Adaptive Decisions With Low Regret". arXiv:2005.02791 [stat.ML]. URL: https://doi.org/10.48550/arXiv.2005.02791

[4] Marlowe, Douglas B., David S. Festinger, Patricia L. Arabia, Karen L. Dugosh, Kathleen M. Benasutti, Jason R. Croft and James R. Mackay. 2008. "Adaptive interventions in drug court: a pilot experiment." In *Criminal Justice Review*; 33(3):343–360

[5] Nease, Robert F. Jr. and Douglas K. Owens. 1997. "Use of Influence Diagrams to Structure Medical Decisions." In *Medical Decision Making*;17(3):263-275.

[6] Neu, Gergely and Julia Olkhovskaya. 2020. "Efficient and robust algorithms for adversarial linear contextual bandits." In *Proceedings of the 33rd Annual Conference on Learning Theory*

[7] Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York: Cambridge University Press

[8] Tian, Jin. 2008. "Identifying dynamic sequential plans." In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*

[9] Wang, Lu, Andrea Rotnitzky, Xihong Lin, Randall E. Millikan, and Peter F. Thall. 2012. "Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer." In *Journal of the American Statistical Association*, 107(498)

[10] Zhang, Junzhe and Elias Bareinboim. 2019. "Near-Optimal Reinforcement Learning in Dynamic Treatment Regimes." In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*

[11] Zhang, Junzhe and Elias Bareinboim. 2020. "Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach." In *Proceedings of the 37th International Conference on Machine Learning*

A Proof of Lemma 1

$$
\begin{aligned}
& E[Y|do(\pi)] \\
&= \sum_{\mathbf{x},\mathbf{z}} E[Y, \mathbf{x}, \mathbf{z}|do(\pi)] \\
&= \sum_{\mathbf{x},\mathbf{z}} E[Y|\mathbf{x}, \mathbf{z}, do(\pi)].P(\mathbf{x}, \mathbf{z}|do(\pi)) \\
&= \sum_{\mathbf{x},\mathbf{z}} E[Y|do(\mathbf{x}), do(\pi), \mathbf{z}].P(\mathbf{x}, \mathbf{z}|do(\pi)) && \text{Rule 2: } (Y \perp \mathbf{X}|\mathbf{Z}) \text{ in } \mathcal{G}_{\pi\underline{\mathbf{X}}} \\
&= \sum_{\mathbf{x},\mathbf{z}} E[Y|do(\mathbf{x}), \mathbf{z}].P(\mathbf{x}, \mathbf{z}|do(\pi)) && \text{Rule 3: } (Y \perp \pi|\mathbf{X}, \mathbf{Z}) \text{ in } \mathcal{G}_{\overline{\mathbf{X}}} \\
&= \sum_{\mathbf{x},\mathbf{z}} E[Y|do(\mathbf{x}), \mathbf{z}].P(\mathbf{z}|\mathbf{x}, do(\pi)).P(\mathbf{x}|do(\pi)) \\
&= \sum_{\mathbf{x},\mathbf{z}} E[Y|do(\mathbf{x}), \mathbf{z}].P(\mathbf{z}|do(\mathbf{x}), do(\pi)).P(\mathbf{x}|do(\pi)) && \text{Rule 2: } (\mathbf{Z} \perp \mathbf{X}) \text{ in } \mathcal{G}_{\pi\underline{\mathbf{X}}} \\
&= \sum_{\mathbf{x},\mathbf{z}} E[Y|do(\mathbf{x}), \mathbf{z}].P(\mathbf{z}|do(\mathbf{x})).P(\mathbf{x}|do(\pi)) && \text{Rule 3: } (\mathbf{Z} \perp \pi|\mathbf{X}) \text{ in } \mathcal{G}_{\overline{\mathbf{X}}} \\
&= \sum_{\mathbf{x},\mathbf{z}} E[Y|do(\mathbf{x}), \mathbf{z}].P(\mathbf{z}|do(\mathbf{x})) \prod_{i=1}^{n} \pi(x_i|\mathbf{h}_i) && \text{Markovian factorization}
\end{aligned}
$$

For details of how to apply the rules of *do-calculus*, refer to Pearl (2009) (7).

Graphically, an intervention π can be thought of as a parent-less node pointing into each X_i.

B *CausalEXP3* Regret Bound

Work in progress.

C Naive *EXP3* algorithm for DTRs

Algorithm 2 EXP3 (for DTR policies)

Input:

- Minimal graph $\mathcal{G}$, with only interventions $\mathbf{X}$, covariates $\mathbf{Z}$, outcome $\mathbf{Y}$
- Deterministic policy space Π
- Learning rate η

Define:

- $N = |\Pi|$, the number of policy arms to explore
- $w_t \in \mathbb{R}^N$: a vector of weights assigned to each policy at episode t

Initialize: $w_1 = (1, 1, 1, 1...)$

for episode $t = 1, 2...T$ **do**

1. Let
$$p_t(\pi_j) = \frac{w_t(j)}{\sum_{j'=1}^{N} w_t(j')}$$
2. Sample an arm $\pi_t \sim p_t$
3. Perform $do(\pi_t)$ and observe $\mathbf{x}_t, \mathbf{z}_t, y_t$
4. For each $\pi \in \Pi$, compute the importance weighted loss
$$l_t(\pi_t) = y_t$$
$$\tilde{l}_t(\pi) = \frac{l_t(\pi_t).\mathbb{I}[\pi = \pi_t]}{p_t(\pi)}$$
5. Update $w_{t+1}(j) \leftarrow w_t(j).\exp[-\eta\tilde{l}(\pi_j)]$

end for

D Lung Cancer DTR Description

The below description from Nease Jr & Owens (1997) (5) is used to form the graph and SCM depicted in Figure 3, and the variables in Table 3:

> Consider the case of a patient with a known nonsmall-cell carcinoma of the lung. The primary tumor is 1cm in diameter; a chest x-ray examination suggests that the tumor does not abut the chest wall or mediastinum. Additional workup reveals no evidence of distance metastases. The preferred treatment in such a situation is thoracotomy, followed by lobectomy or pneumonectomy, depending on whether the primary tumor has metastasized to the hilar lymph nodes. Of fundamental importance in the decision to perform thoracotomy is the likelihood of mediastinal metastases. If mediastinal metastases are known to be present, most clinicians would deem thoracotomy to be contraindicated: thoracotomy subjects the patient to a risk of death but confers no health benefit...If mediastinal metastases are known to be absent, thoracotomy offers a substantial survival advantage, so long as the primary tumor has not metastasized to distant organs. There are several diagnostic tests available to assess any involvement of the mediastinum. For this example, we shall focus on computed tomography (CT) of the chest and mediastinoscopy. Our problem involves three decisions. First, should the patient undergo a CT scan? Second, given our decision about CT and any CT results obtained, should the patient undergo mediastinoscopy? Third, given the results of any tests that we have decided to perform, should the patient undergo thoracotomy?

E Lung Cancer SCM Probabilities

Table 5: **"True" probabilities for sample SCM generated for Lung Cancer DTR in Figure 3**

A:	$P(a_0 \mid b_0, g_0)$	=	0.2841	$P(a_1 \mid b_0, g_0)$	=	0.5005
	$P(a_0 \mid b_0, g_1)$	=	0.4862	$P(a_1 \mid b_0, g_1)$	=	0.4792
	$P(a_0 \mid b_1, g_0)$	=	0.4680	$P(a_1 \mid b_1, g_0)$	=	0.4077
	$P(a_0 \mid b_1, g_1)$	=	0.0330	$P(a_1 \mid b_1, g_1)$	=	0.6757
B:	$P(b_0)$	=	0.5417	$P(b_1)$	=	0.4583
C:	$P(c_0 \mid b_0, h_0)$	=	0.4103	$P(c_1 \mid b_0, h_0)$	=	0.1062
	$P(c_0 \mid b_0, h_1)$	=	0.3080	$P(c_1 \mid b_0, h_1)$	=	0.4666
	$P(c_0 \mid b_1, h_0)$	=	0.3997	$P(c_1 \mid b_1, h_0)$	=	0.5083
	$P(c_0 \mid b_1, h_1)$	=	0.3017	$P(c_1 \mid b_1, h_1)$	=	0.3389
D:	$P(d_0 \mid i_0)$	=	0.4328	$P(d_0 \mid i_1)$	=	0.2731
E:	$P(e_1 \mid h_0)$	=	0.1473	$P(e_1 \mid h_1)$	=	0.8849
F:	$P(f_1 \mid b_0, d_0, e_0, i_0)$	=	0.1491	$P(f_1 \mid b_0, d_0, e_0, i_1)$	=	0.9693
	$P(f_1 \mid b_0, d_0, e_1, i_0)$	=	0.0177	$P(f_1 \mid b_0, d_0, e_1, i_1)$	=	0.2382
	$P(f_1 \mid b_0, d_1, e_0, i_0)$	=	0.8229	$P(f_1 \mid b_0, d_1, e_0, i_1)$	=	0.9601
	$P(f_1 \mid b_0, d_1, e_1, i_0)$	=	0.2460	$P(f_1 \mid b_0, d_1, e_1, i_1)$	=	0.8257
	$P(f_1 \mid b_1, d_0, e_0, i_0)$	=	0.0937	$P(f_1 \mid b_1, d_0, e_0, i_1)$	=	0.2567
	$P(f_1 \mid b_1, d_0, e_1, i_0)$	=	0.5303	$P(f_1 \mid b_1, d_0, e_1, i_1)$	=	0.1900
	$P(f_1 \mid b_1, d_1, e_0, i_0)$	=	0.4400	$P(f_1 \mid b_1, d_1, e_0, i_1)$	=	0.3264
	$P(f_1 \mid b_1, d_1, e_1, i_0)$	=	0.6326	$P(f_1 \mid b_1, d_1, e_1, i_1)$	=	0.3320
G:	$P(g_0)$	=	0.2546	$P(g_1)$	=	0.7454
H:	$P(h_1 \mid a_0, g_0)$	=	0.9456	$P(h_1 \mid a_0, g_1)$	=	0.4239
	$P(h_1 \mid a_1, g_0)$	=	0.7273	$P(h_1 \mid a_1, g_1)$	=	0.6931
	$P(h_1 \mid a_2, g_0)$	=	0.4035	$P(h_1 \mid a_2, g_1)$	=	0.4228
I:	$P(i_0 \mid a, c_0, e_0, g_0, h_0)$	=	0.1576	$P(P(i_0 \mid a, c_0, e_0, g_0, h_1))$	=	0.8491
	$P(i_0 \mid a, c_0, e_0, g_1, h_0)$	=	0.4218	$P(P(i_0 \mid a, c_0, e_0, g_1, h_1))$	=	0.6555
	$P(i_0 \mid a, c_0, e_1, g_0, h_0)$	=	0.4854	$P(P(i_0 \mid a, c_0, e_1, g_0, h_1))$	=	0.7577
	$P(i_0 \mid a, c_0, e_1, g_1, h_0)$	=	0.9595	$P(P(i_0 \mid a, c_0, e_1, g_1, h_1))$	=	0.0318
	$P(i_0 \mid a, c_1, e_0, g_0, h_0)$	=	0.9706	$P(P(i_0 \mid a, c_1, e_0, g_0, h_1))$	=	0.9340
	$P(i_0 \mid a, c_1, e_0, g_1, h_0)$	=	0.9157	$P(P(i_0 \mid a, c_1, e_0, g_1, h_1))$	=	0.1712
	$P(i_0 \mid a, c_1, e_1, g_0, h_0)$	=	0.8003	$P(P(i_0 \mid a, c_1, e_1, g_0, h_1))$	=	0.7431
	$P(i_0 \mid a, c_1, e_1, g_1, h_0)$	=	0.6557	$P(P(i_0 \mid a, c_1, e_1, g_1, h_1))$	=	0.2769
	$P(i_0 \mid a, c_2, e_0, g_0, h_0)$	=	0.9572	$P(P(i_0 \mid a, c_2, e_0, g_0, h_1))$	=	0.6787
	$P(i_0 \mid a, c_2, e_0, g_1, h_0)$	=	0.7922	$P(P(i_0 \mid a, c_2, e_0, g_1, h_1))$	=	0.7060
	$P(i_0 \mid a, c_2, e_1, g_0, h_0)$	=	0.1419	$P(P(i_0 \mid a, c_2, e_1, g_0, h_1))$	=	0.3922
	$P(i_0 \mid a, c_2, e_1, g_1, h_0)$	=	0.0357	$P(P(i_0 \mid a, c_2, e_1, g_1, h_1))$	=	0.0462

Table 6: **"True" values for *CausalEXP3* estimands for sample SCM generated for Lung Cancer DTR**

G:	$P(g_0)$	=	0.2546	$P(g_1)$	=	0.7454
C:	$P(c_0\|do(h_0))$	=	0.4055	$P(c_0\|do(h_1))$	=	0.3051
	$P(c_1\|do(h_0))$	=	0.2904	$P(c_1\|do(h_1))$	=	0.4081
	$P(c_2\|do(h_0))$	=	0.3041	$P(c_2\|do(h_1))$	=	0.2868
F:	$E[F\|do(h_0, i_0), c_0]$	=	0.3559	$E[F\|do(h_1, i_0), c_0]$	=	0.3759
	$E[F\|do(h_0, i_0), c_1]$	=	0.4546	$E[F\|do(h_1, i_0), c_1]$	=	0.3707
	$E[F\|do(h_0, i_0), c_2]$	=	0.2677	$E[F\|do(h_1, i_0), c_2]$	=	0.3845
	$E[F\|do(h_0, i_1), c_0]$	=	0.5406	$E[F\|do(h_1, i_1), c_0]$	=	0.5919
	$E[F\|do(h_0, i_1), c_1]$	=	0.3854	$E[F\|do(h_1, i_1), c_1]$	=	0.6303
	$E[F\|do(h_0, i_1), c_2]$	=	0.6794	$E[F\|do(h_1, i_1), c_2]$	=	0.5276

F Drug-Offence Corrections DTR Description

The below description from Nease Jr & Owens (1997) (5) is used to form the graph and SCM depicted in Figure 6, and the variables in Table 4:

> At entry into the program, offenders were classified as high risk for failure if they met the criteria in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, for antisocial personality disorder or had previously attended at least one formal drug abuse intervention, excluding self-help groups; otherwise offenders were classified as low risk. Offenders who were classified as high risk were assigned to biweekly court hearings, whereas offenders classified as low risk were assigned to as-needed court hearings. In addition to the court hearings, all offenders were required to attend weekly group substance abuse counseling sessions and to provide weekly urine specimens.
>
> Each offender's progress in the program was assessed monthly. If at any monthly assessment an offender had missed two or more counseling sessions without an excuse or failed to provide two or more scheduled urine specimens, he/she was classified as noncompliant, and the level of court supervision was increased. In particular, offenders who were assigned to as-needed court hearings but did not comply moved to biweekly court hearings; offenders who were assigned to biweekly court hearings but did not comply were placed on a jeopardy contract in which further violation of the rules of the program resulted in moving into a regular court system. If an offender attended the scheduled counseling sessions, provided the urine specimens and did not commit new infractions, but two or more urine specimens were drug-positive, then the offender was classified as nonresponsive. In this case, the intensity and scope of the drug abuse treatment were altered. More specifically, these offenders entered an intensive case management program in which they were provided twice-weekly individual substance abuse counseling sessions.

G Drug-Offence Corrections SCM Probabilities

Table 7: **"True" probabilities for sample SCM generated for Drug Correction DTR for first half of all episodes**

A:	$P(a_0)$	=	0.4379	$P(a_1)$	=	0.5621
B:	$P(b_0\|a_0)$	=	0.1164	$P(b_0\|a_1)$	=	0.8857
C:	$P(c_0\|a_0, b_0)$	=	0.801	$P(c_0\|a_1, b_0)$	=	0.6638
	$P(c_0\|a_0, b_1)$	=	0.3137	$P(c_0\|a_1, b_1)$	=	0.1879
D:	$P(d_0\|c_0, b_0)$	=	0.3111	$P(d_1\|c_0, b_0)$	=	0.3741
	$P(d_0\|c_0, b_1)$	=	0.7275	$P(d_1\|c_0, b_1)$	=	0.6911
	$P(d_0\|c_1, b_0)$	=	0.6368	$P(d_1\|c_1, b_0)$	=	0.5336
	$P(d_0\|c_1, b_1)$	=	0.3116	$P(d_1\|c_1, b_1)$	=	0.1193
E:	$P(e_0\|b_0)$	=	0.82	$P(e_0\|b_1)$	=	0.1643
F:	$P(f_1\|b_0, d_0)$	=	0.5011	$P(f_1\|b_1, d_0)$	=	0.5602
	$P(f_1\|b_0, d_1)$	=	0.7787	$P(f_1\|b_1, d_1)$	=	0.3759
	$P(f_1\|b_0, d_2)$	=	0.755	$P(f_1\|b_1, d_2)$	=	0.6166

Table 8: **"True" values for *CausalEXP3* estimands for Drug Correction DTR for first half of all episodes**

A:	$P(a_0)$	=	0.4379	$P(a_1)$	=	0.5621
C:	$P(c_0\|do(b_0), a_0)$	=	0.801	$P(c_1\|do(b_0), a_0)$	=	0.199
	$P(c_0\|do(b_0), a_1)$	=	0.6638	$P(c_1\|do(b_0), a_1)$	=	0.3362
	$P(c_0\|do(b_1), a_0)$	=	0.3137	$P(c_1\|do(b_1), a_0)$	=	0.6863
	$P(c_0\|do(b_1), a_1)$	=	0.1879	$P(c_1\|do(b_1), a_1)$	=	0.8121
E:	$P(e_0\|do(b_0))$	=	0.82	$P(e_1\|do(b_0))$	=	0.18
	$P(e_0\|do(b_1))$	=	0.1643	$P(e_1\|do(b_1))$	=	0.8357
F:	$E[F\|do(b_0, d_0)]$	=	0.5011	$E[F\|do(b_1, d_0)]$	=	0.5602
	$E[F\|do(b_0, d_1)]$	=	0.7787	$E[F\|do(b_1, d_1)]$	=	0.3759
	$E[F\|do(b_0, d_2)]$	=	0.755	$E[F\|do(b_1, d_2)]$	=	0.6166

Table 9: **"True" probabilities for sample SCM generated for Drug Correction DTR for second half of all episodes**

A:	$P(a_0)$	=	0.7341	$P(a_1)$	=	0.2659
B:	$P(b_0\|a_0)$	=	0.7231	$P(b_0\|a_1)$	=	0.1903
C:	$P(c_0\|a_0, b_0)$	=	0.4357	$P(c_0\|a_1, b_0)$	=	0.2378
	$P(c_0\|a_0, b_1)$	=	0.7603	$P(c_0\|a_1, b_1)$	=	0.5035
D:	$P(d_0\|c_0, b_0)$	=	0.7008	$P(d_1\|c_0, b_0)$	=	0.6466
	$P(d_0\|c_0, b_1)$	=	0.6593	$P(d_1\|c_0, b_1)$	=	0.4147
	$P(d_0\|c_1, b_0)$	=	0.6767	$P(d_1\|c_1, b_0)$	=	0.1747
	$P(d_0\|c_1, b_1)$	=	0.2991	$P(d_1\|c_1, b_1)$	=	0.8065
E:	$P(e_0\|b_0)$	=	0.2696	$P(e_0\|b_1)$	=	0.6865
F:	$P(f_1\|b_0, d_0)$	=	0.5315	$P(f_1\|b_1, d_0)$	=	0.4102
	$P(f_1\|b_0, d_1)$	=	0.3952	$P(f_1\|b_1, d_1)$	=	0.685
	$P(f_1\|b_0, d_2)$	=	0.5807	$P(f_1\|b_1, d_2)$	=	0.8256

Table 10: **"True" values for *CausalEXP3* estimands for Drug Correction DTR for second half of all episodes**

A:	$P(a_0)$	=	0.7341	$P(a_1)$	=	0.2659
C:	$P(c_0\|do(b_0), a_0)$	=	0.4357	$P(c_1\|do(b_0), a_0)$	=	0.5643
	$P(c_0\|do(b_0), a_1)$	=	0.2378	$P(c_1\|do(b_0), a_1)$	=	0.7622
	$P(c_0\|do(b_1), a_0)$	=	0.7603	$P(c_1\|do(b_1), a_0)$	=	0.2397
	$P(c_0\|do(b_1), a_1)$	=	0.5035	$P(c_1\|do(b_1), a_1)$	=	0.4965
E:	$P(e_0\|do(b_0))$	=	0.2696	$P(e_1\|do(b_0))$	=	0.7304
	$P(e_0\|do(b_1))$	=	0.6865	$P(e_1\|do(b_1))$	=	0.3135
F:	$E[F\|do(b_0, d_0)]$	=	0.5315	$E[F\|do(b_1, d_0)]$	=	0.4102
	$E[F\|do(b_0, d_1)]$	=	0.3952	$E[F\|do(b_1, d_1)]$	=	0.685
	$E[F\|do(b_0, d_2)]$	=	0.5807	$E[F\|do(b_1, d_2)]$	=	0.8256