An Introduction to Causal Reinforcement Learning

Elias Bareinboim, Junzhe Zhang, Sanghack Lee

-Present by Jiapeng Zhao

An Introduction to Causal Reinforcement Learning Elias Bareinboim, Junzhe Zhang, Sanghack Lee, Technical Report, Dec 2024

Overview

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Recap - Structural Causal Model (SCM)

Definition 1 (Structural Causal Model (Pearl, 2000)) *A structural causal model (SCM, for short) $\mathcal{M}$ is a 4-tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$ where:*

- *A set of background or exogenous variables $\boldsymbol{U} = \{U_1, U_2, ..., U_k\}$, representing factors outside the model, which nevertheless, affect relationships within the model;*
- *A set of endogenous variables $\boldsymbol{V} = \{V_1, V_2, ..., V_n\}$ representing variables inside the model.*
- *A set $\mathscr{F}$ of structural functions $\{f_i : V_i \in \boldsymbol{V}\}$ s.t. each f_i determines the value of $V_i \in \boldsymbol{V}$,*

$$v_i \leftarrow f_i(\boldsymbol{pa}_i, \boldsymbol{u}_i), \tag{1}$$

where $\boldsymbol{PA}_i \subseteq \boldsymbol{V} \setminus \{V\}$ and $\boldsymbol{U}_i \subseteq \boldsymbol{U}$.

- *A probability distribution over the exogenous variables, $P(\boldsymbol{U})$.* ■

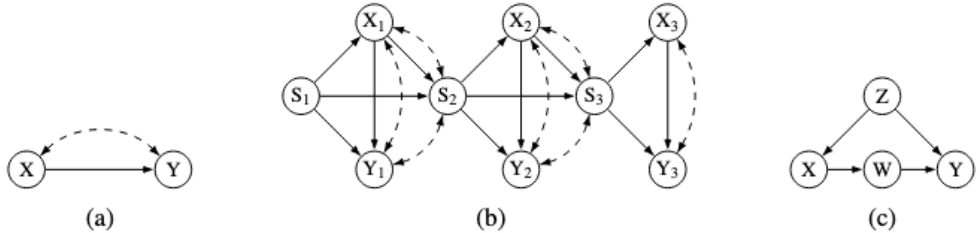


Figure 8: Causal diagrams for (a) a multi-armed bandit (MAB); (b) a Markov decision process (MDP); and (c) an SCM representing a refinement of the MAB environment.

Recap - Reinforcement Learning - MDP

An **MDP** is a 5-tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ where:

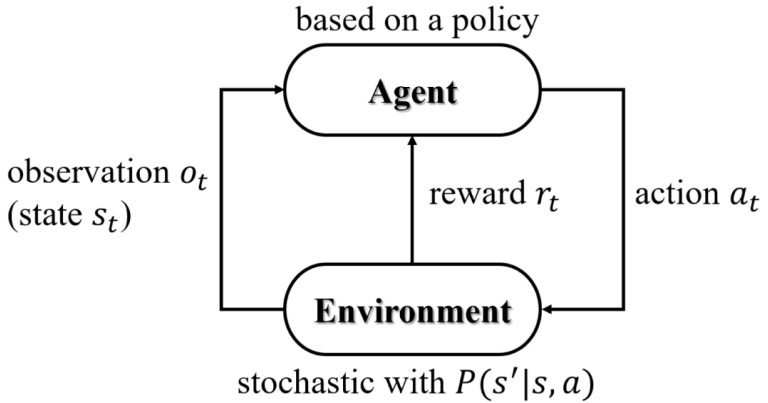
1. $\mathcal{S}$: A (possibly infinite) set of **states**. At each time step t, the agent observes a state $S_t \in \mathcal{S}$.
2. $\mathcal{A}$: A set of **actions**. In each state S_t, the agent chooses an action $A_t \in \mathcal{A}$.
3. P: A **transition probability** function,

$$P(s' \mid s, a) \;=\; \Pr(S_{t+1} = s' \;\mid\; S_t = s,\; A_t = a),$$

   which gives the probability of transitioning to state s' from state s when the agent takes action a.
4. R: A **reward function**, which may be defined in various ways. A common specification is

$$R(s, a) \;=\; \mathbb{E}[\, R_{t+1} \mid S_t = s,\; A_t = a \,],$$

   the expected immediate reward obtained by taking action a in state s. Alternatively, some formulations define $R(s)$ or $R(s, a, s')$.
5. γ: A **discount factor** in $[0, 1]$. It controls how future rewards are weighted relative to immediate rewards. If $\gamma < 1$, future rewards are geometrically discounted by powers of γ.



CRL

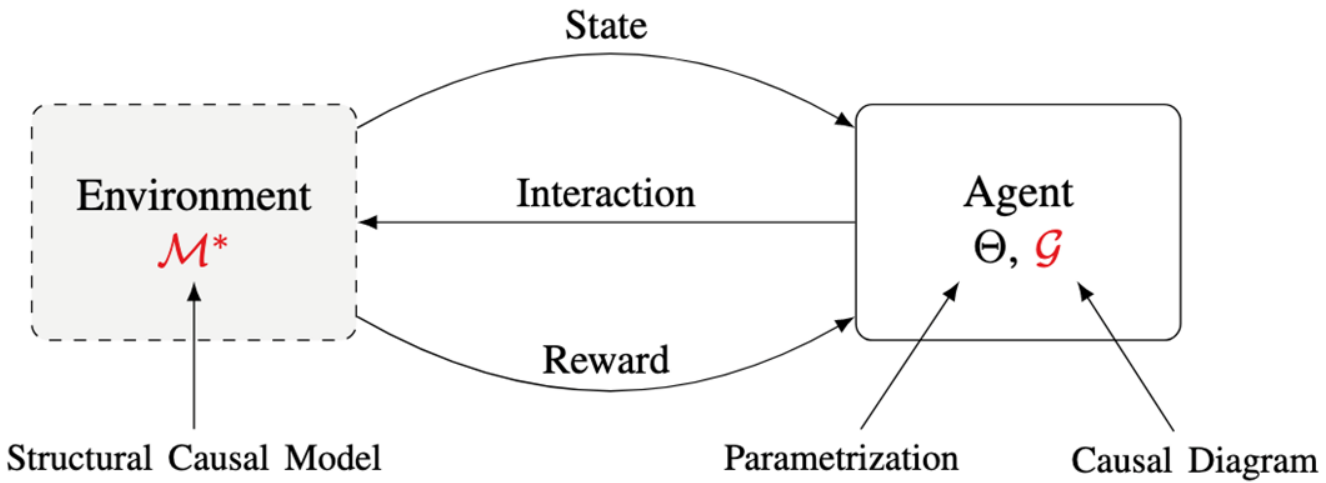


Figure 1: The Agent-Environment interaction from Causal Reinforcement Learning (CRL).

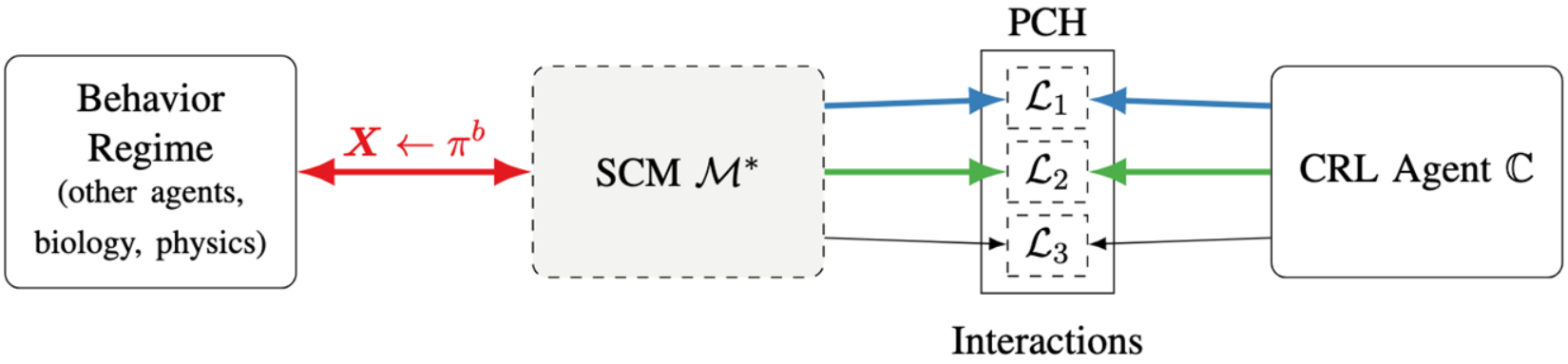


Figure 4: Representation of the CRL agent (right side) interacting with the SCM (middle) through natural (marked in blue) and interventional (green) regimes. Other behavior agents and their interactions are shown in the left side (red).

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Causal Decision Model (CDM)

Definition 11 (Causal Decision Model) *A causal decision model (CDM) is a tuple* $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ *where* $\mathcal{M}^* = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$ *is an SCM,* Π *is a policy space over actions* $\boldsymbol{X} \subseteq \boldsymbol{V}$*, and* $\mathcal{R}$ *is a reward function over reward signals* $\boldsymbol{Y} \subseteq \boldsymbol{V}$. ■

Motivation:

Armed with this new formalism, we can represent many canonical decision-making settings found in the literature within the semantic framework of SCMs.

Canonical RL models:

1. multi-armed bandits (MABs),
2. Markov decision processes (MDPs),
3. dynamic treatment regimes (DTRs).

Policy Space

Definition 9 (Policy Space) *For an SCM* $\mathcal{M}^* = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$*, a policy space* Π *is a collection of policies* π *over actions* $\boldsymbol{X} = \{X_1, \ldots, X_H\}$*. Each policy* π *is a sequence of decision rules* $(\pi_1 (X_1 \mid \boldsymbol{S}_1), \ldots, \pi_H (X_H \mid \boldsymbol{S}_H))$ *such that for every* $i = 1, \ldots, H$*,*[16]

- *Action* X_i *is a non-descendent of* $X_{i+1}, \ldots, X_H$*, i.e.,* $X_i \in \boldsymbol{V} \setminus De(X_{i+1}, \ldots, X_H)$*;*
- *States* $\boldsymbol{S}_i$ *are non-descendants of* $X_i, \ldots, X_H$*, i.e.,* $\boldsymbol{S}_i \subseteq \boldsymbol{V} \setminus De(X_i, \ldots, X_H)$*.*

Henceforth, we will consistently denote such a policy space by $\Pi = \{\langle X_1, \boldsymbol{S}_1 \rangle, \ldots, \langle X_H, \boldsymbol{S}_H \rangle\}$*.* ■

In words, a policy space Π defines:

1. the action space X that the agent could control after being deployed in the environment,
2. the state space $S_1, \ldots, S_H$ that the agent could perceive at the time of intervention for every action $X_1, \ldots, X_H$.

Policy Space

Example: 2-stage dynamic treatment regimes (DTR)

In healthcare, a typical patient is often treated at multiple stages; the physician repeatedly adapts each treatment, tailoring it to the patient's time-varying, dynamic state. SCM definition:

More formally, consider a DTR environment $\mathcal{M}^*_{\text{DTR}}$ *that is a tuple given by*

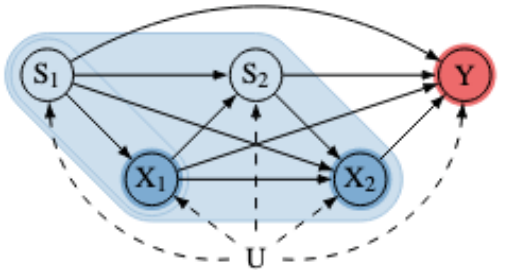
$$\mathcal{M}^*_{\text{DTR}} = \langle \boldsymbol{U} = \{U, U_1, \ldots, U_4\}, \boldsymbol{V} = \{S_1, S_2, X_1, X_2, Y\}, \mathscr{F}_{\text{DTR}}, P(\boldsymbol{U}) \rangle, \tag{77}$$

where the underlying causal mechanisms $\mathscr{F}_{\text{DTR}}$ *are given by,*

$$\mathscr{F}_{\text{DTR}} = \begin{cases} S_1 \leftarrow \mathbb{1}\{U_3 > 0\}, \\ X_1 \leftarrow \mathbb{1}\{3S_1 + \alpha_1 U + U_1 > 0\}, \\ S_2 \leftarrow \mathbb{1}\{0.1 + 0.1S_1 + 0.1X_1 + U_4 > 0\}, \\ X_2 \leftarrow \mathbb{1}\{3S_2 + \alpha_2 U + U_2 > 0\}, \\ Y \leftarrow \mathbb{1}\{3U - 3S_1 - 3X_1 - 3S_1X_1 + 3X_2 - 3S_2X_2 + 3X_1X_2 > 0\}, \end{cases} \tag{78}$$

The exogenous distribution $P(\boldsymbol{U})$ *is defined such that for* $i = 1, \ldots, 4$, $U_i \sim \texttt{Logistic}(0, 1)$ *is an independent variable drawn from a logistic distribution*

$$P(U_i < u) = \frac{1}{1 + e^{-u}} \tag{79}$$



(c) DTR

Policy Space

- **Full Policy Space** $\Pi_{\text{full}} = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1, X_1, S_2\}\rangle\}$
  - Reflects **perfect recall** (Koller & Friedman, 2009).
  - A **general policy** $\pi \in \Pi_{\text{full}}$ is:

$$\pi = \big(\pi_1(X_1 \mid S_1),\ \pi_2(X_2 \mid S_1, X_1, S_2)\big)$$

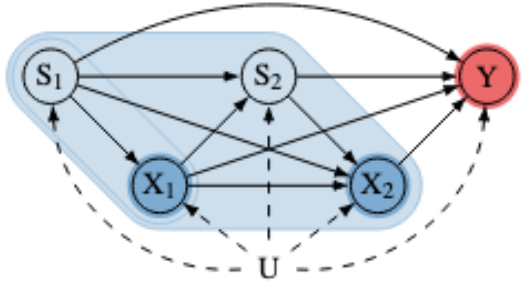
    - π_1 chooses an initial treatment X_1 based on S_1.
    - π_2 decides whether to continue or switch treatments X_2, based on $\{S_1, X_1, S_2\}$.
- **Markov Policy Space** $\Pi_{\text{Markov}} = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_2\}\rangle\}$
  - A **Markov policy** $\pi \in \Pi_{\text{Markov}}$ is:

$$\pi = \big(\pi_1(X_1 \mid S_1),\ \pi_2(X_2 \mid S_2)\big)$$

    - Decisions depend only on the **current** state.
- **Stationary Policy Space** $\Pi_{\text{stationary}} \subset \Pi_{\text{Markov}}$
  - A **stationary policy** π has the **same** decision rule at each stage:

$$\pi_1(X_1 \mid S_1) \ = \ \pi_2(X_2 \mid S_2)$$

- **Containment Relationships**
  - $\Pi_{\text{stationary}} \subset \Pi_{\text{Markov}} \subset \Pi_{\text{full}}$.
  - Fig. 9 shows a Venn diagram illustrating these nested sets.



(c) DTR

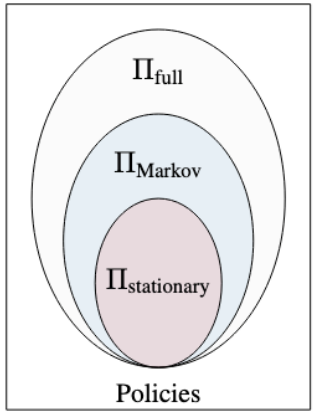


Figure 9: Policy spaces for a 2-stage DTR environment.

Causal Decision Model (CDM)

- **Multi-Armed Bandit (Robbins, 1952)**
  - Induced by a MAB model $\langle \mathcal{M}^*_{\text{MAB}}, \Pi, Y \rangle$.
  - Consists of an arm choice X and a reward Y.
  - **Policy space** $\Pi = \{\langle X, \varnothing \rangle\}$ defines a set of policies π that select values of X following a probability distribution $\pi(X)$.
- **Contextual Bandit (Langford and Zhang, 2008a)**
  - Graphical representation $\langle \mathcal{M}^*_{\text{CMAB}}, \Pi, Y \rangle$.
  - A context variable S is observed (unlike in the standard MAB).
  - **Policy space** $\Pi = \{\langle X, \{S\} \rangle\}$ consists of candidate policies $\pi(X \mid S)$ that select X based on the observed context S.
- **Dynamic Treatment Regime (Murphy et al., 2001a)**
  - A DTR model $\langle \mathcal{M}^*_{\text{DTR}}, \Pi, Y \rangle$.
  - **Policy space** $\Pi = \{\langle X_i, \{S_1, \ldots, S_i, X_1, \ldots, X_{i-1}\} \rangle\}_{i=1}^{H}$, a set of policies $\pi = (\pi_1, \ldots, \pi_H)$.
  - Each $\pi_i(X_i \mid S_1, \ldots, S_i, X_1, \ldots, X_{i-1})$ selects actions at stage i based on the entire history of states and actions.

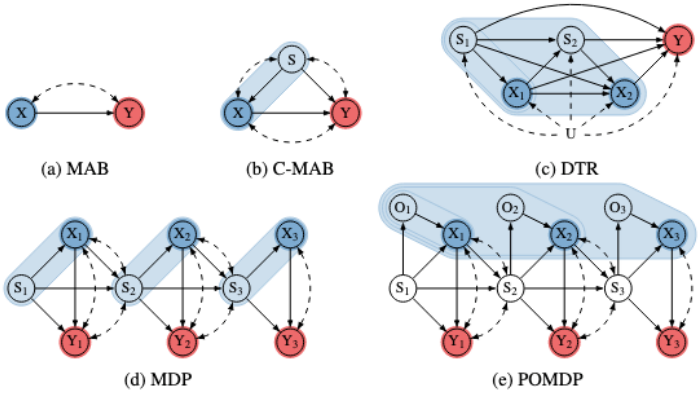


Figure 10: Causal diagrams for CDMs representing canonical decision-making models.

Causal Decision Model (CDM)

- **Markov Decision Process (Bellman, 1966; Puterman, 1994)**
  - MDP model $\langle \mathcal{M}^*_{\text{MDP}}, \Pi, \mathcal{R} \rangle$.
  - Environment $\mathcal{M}^*$ has:
    - States $S = \{S_i^\infty\}_{i=1}$
    - Actions $X = \{X_i^\infty\}_{i=1}$
    - Rewards $Y = \{Y_i^\infty\}_{i=1}$
  - **Policy space** $\Pi = \{\langle X_i, S_i \rangle\}_{i=1}^\infty$, i.e., $\pi = (\pi_i(X_i \mid S_i))_{i=1}^\infty$.

- **Partially Observable MDP (Åström, 1965)**
  - POMDP model $\langle \mathcal{M}^*_{\text{POMDP}}, \Pi, \mathcal{R} \rangle$.
  - States $S = \{S_i^\infty\}_{i=1}$ are hidden; the agent sees observations $O = \{O_i^\infty\}_{i=1}$.
  - Actions $X = \{X_i^\infty\}_{i=1}$.
  - **Policy space** $\Pi = \{\langle X_i, \{O_1, \ldots, O_i, X_1, \ldots, X_{i-1}\} \rangle\}_{i=1}^\infty$.
    - A policy π can be non-Markov: $\pi_i(X_i \mid O_1, \ldots, O_i, X_1, \ldots, X_{i-1})$.

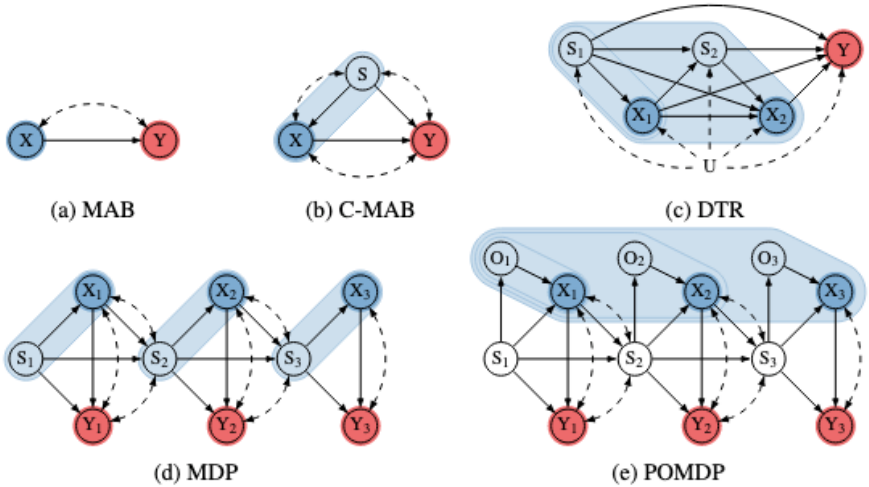


Figure 10: Causal diagrams for CDMs representing canonical decision-making models.

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Task
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

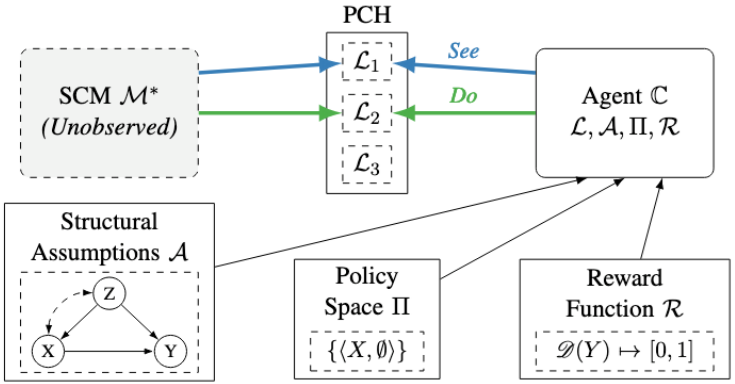
Causal Reinforcement Learning (CRL) Task

Definition 12 (Causal Reinforcement Learning Task) *For an SCM $\mathcal{M}^* = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$, a CRL task $\mathcal{T}$ in the environment $\mathcal{M}^*$ is a 4-tuple $\langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$, where*

1. *$\mathcal{L}$ is a learning regime of an agent's interaction with the SCM $\mathcal{M}^*$, possibly* see *or* do*;*
2. *$\mathcal{A}$ is a set of structural assumptions about the SCM $\mathcal{M}^*$;*
3. *Π is a policy space over actions $\boldsymbol{X}$;*
4. *$\mathcal{R}$ is a reward function over reward signals $\boldsymbol{Y}$.* ■

The goal of the agent is then to find a policy π^* such that

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \,\middle|\, \mathcal{A}, \mathcal{L} \right] \tag{116}$$



General Learning Strategy

Alg. 1 provides pseudocode describing the general learning strategy of a CRL agent to solve a CRL task $\langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$

Algorithm 1 Causal Reinforcement Learning Agent $\mathbb{C}$

Require: CRL task $\mathcal{T} = \langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$
Ensure: a policy estimate $\hat{\pi} \in \Pi$ optimizing a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$.

1: Let data $\mathcal{D} = \{\}$.
2: **for all** every episode $t = 1, \ldots, T$ **do**
3: Interact with the SCM $\mathcal{M}^*$ following regime $\mathcal{L}$ and receive samples $\boldsymbol{V}^{(t)} \sim P_{\mathcal{L}}(\boldsymbol{V}; \mathcal{M}^*)$.
4: Update data $\mathcal{D} = \mathcal{D} \cup \{\boldsymbol{V}^{(t)}\}$.
5: **end for**
6: **return** an empirical estimate $\hat{\pi} \in \Pi$ of the optimal policy π^* from data $\mathcal{D}$ and assumptions $\mathcal{A}$.

CRL Task Example

Example 18 (Off-Policy Learning (Sutton and Barto, 1998), MAB) *Consider first a MAB model* $\langle \mathcal{M}^*, \{\langle X, \emptyset \rangle\}, Y \rangle$ *graphically described in Fig. 10a. A CRL agent* $\mathbb{C}$ *aims to learn an arm*

(a) MAB

$$x^* = \arg\max_x \mathbb{E}_x [Y] \tag{117}$$

with the maximal expected reward. The detailed parametrization of SCM $\mathcal{M}^*$ *is unknown. Instead, the agent could only passively observe the environment and receive the observational distribution* $P(X, Y)$ *evaluated in* $\mathcal{M}^*$*. This leads to an* off-policy learning *task described through the following signature:*

$$\mathcal{T}_{off} = \langle \mathcal{L} = see, \mathcal{A} = NUC, \Pi = \{\langle X, \emptyset \rangle\}, \mathcal{R} = Y \rangle, \tag{118}$$

NUC stands for the assumption of "No Unmeasured Confounder": there is no unobserved confounder affecting the action X *and the reward* Y *simultaneously. All the observed correlations between* X *and* Y *are fully explained by the causal relationships among them, which implies that*

$$P_x(Y) = P(Y \mid X = x) \tag{119}$$

Therefore, the agent could evaluate the expected reward of every arm x *from the observational data* $P(X, Y)$[20]*, and find optimal an optimal arm* x^* *with the maximal empirical reward estimates.* ■

Causal Reinforcement Learning (CRL) Task

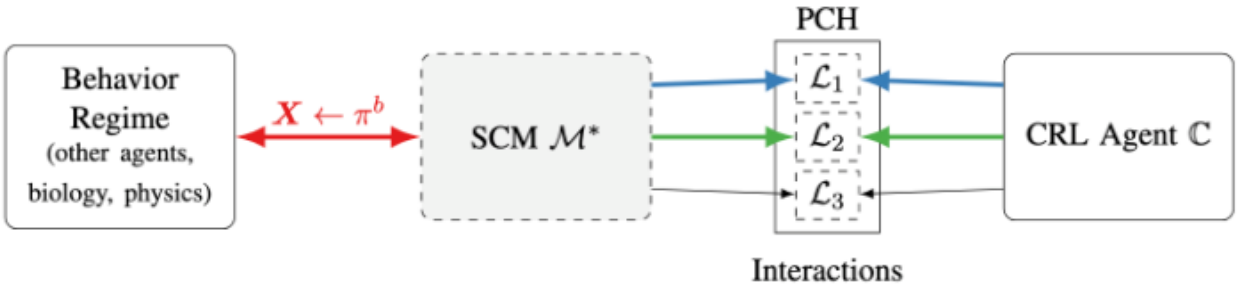
		Signature				
	Task	**Learning Regime** ($\mathcal{L}$)	**Structural Assumptions** ($\mathcal{A}$)	**Policy Space** (Π)	**Reward Function** ($\mathcal{R}$)	**Section**
1	Off-policy Learning	See	NUC	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.1
2	Online Learning	Do	-	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.2
3	Causal Identification	See	DAG $\mathcal{G}$	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.3

Table 4: Summary of causal reinforcement learning tasks investigated in this paper, in terms of their signatures and sections. We highlight in gray the most distinct feature introduced by the task.

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Off-Policy Learning Task

- Off-Policy:
  - an agent attempts to learn an optimal policy from observational data generated by a different behavior policy
  - Can learn from data generated by other policies, including old experiences or experiences from a different source.
  - Generally more sample-efficient and flexible since it can reuse data, but it may be less stable and require careful handling (e.g., replay buffers).
- On-Policy:
  - The agent learns and improves the same policy that it is currently following to select actions.
  - The policy being learned and the policy being used for data collection are the same.
  - Tends to have more stable learning but may be less sample-efficient since it updates based on the current policy's behavior

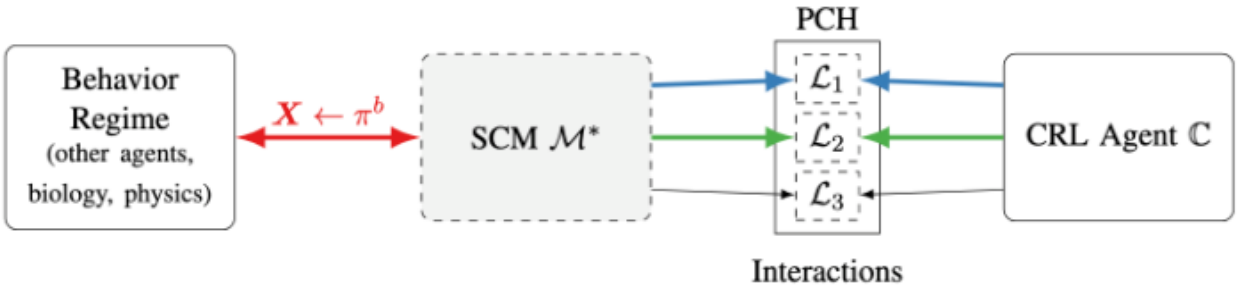


Off-Policy Learning Task

$$\mathcal{T}_{\text{off}} = \left\langle \mathcal{I} = \text{see}, \mathcal{A} = \text{NUC}, \Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}, \mathcal{R} = \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R} \right\rangle. \tag{135}$$

where the agent uses observational data combined with the critical assumption $\mathcal{A} = \text{NUC}$, which means "*No Unmeasured Confounder*". The agent's goal is to obtain an optimal policy estimate from the combination of the observational data and the NUC assumption, i.e.,

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \,\middle|\, \mathcal{A} = \text{NUC}, \; \mathcal{D}_{\text{obs}} \sim P(\boldsymbol{V}) \right]. \tag{136}$$

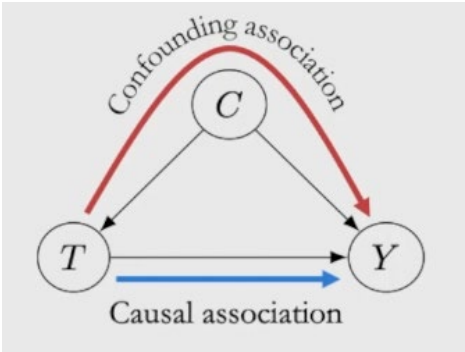


No Unmeasured Confounder (NUC)

Definition 16 (No Unmeasured Confounder) *Let $\mathcal{M}^*$ be a SCM and $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ be a policy space (Def. 9). The "no unmeasured confounder" (for short, NUC) condition holds if for every action $X_i \in \boldsymbol{X}$, its endogenous parents $\boldsymbol{PA}_i$ and exogenous parents $\boldsymbol{U}_i$ satisfy the following conditions:*

1. *Endogenous parents $\boldsymbol{PA}_i$ are contained in the history $\bar{\boldsymbol{X}}_{i-1} \cup \bar{\boldsymbol{S}}_i$, i.e., $\boldsymbol{PA}_i \subseteq \bar{\boldsymbol{X}}_{i-1} \cup \bar{\boldsymbol{S}}_i$;*
2. *Exogenous parents $\boldsymbol{U}_i$ are independent from exogenous noises $\boldsymbol{U}_j$ associated with all the other endogenous variables V_j in the system, i.e., $\boldsymbol{U}_i \perp\!\!\!\perp \{\boldsymbol{U}_j \mid \forall V_j \in \boldsymbol{V} \setminus \{X_i\}\}$.*

In the above definition, Condition 1 says that all endogenous parents of every action X_i are observed, contained in the past states and actions $\bar{\boldsymbol{S}}_i, \bar{\boldsymbol{X}}_{i-1}$; Condition 2 says that given the past history $\bar{\boldsymbol{X}}_{i-1} \cup \bar{\boldsymbol{S}}_i$, values of every action X_i are decided by an independent noise $\boldsymbol{U}_i$. Note that



This NUC assumption is important because

1. We want only the causal association, not the confounding association.
2. If the confounder is not observable, the effect of the variable can incorrectly be (partly) assigned to the treatment, resulting in a biased model.
3. NUC ensures that no other variables generate non-causal variations between the decision X and the outcome Y.

OFF-POLICY EVALUATION (Prediction)

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \,\middle|\, \mathcal{A} = \text{NUC},\ \mathcal{D}_{\text{obs}} \sim P(\boldsymbol{V}) \right].$$

- Motivation:
  - When the expected rewards of candidate policies are computable, the agent could then search over the policy space and obtain an optimal policy estimate using policy gradient (Sutton et al., 1999).
- Goal:
  - given policy π, calculate its expected rewards.
- Methods:
  - Inverse propensity score
  - Dynamic programing

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Inverse Propensity Weighting

Theorem 17 (Inverse Propensity Weighting, under NUC) *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where the policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ and the reward function $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$. If NUC holds, for any $\pi \in \Pi$, the expected reward is computable from the observational distribution $P(\boldsymbol{V})$ as*

$$\mathbb{E}_{\pi}[\mathcal{R}(\boldsymbol{Y})] = \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} \underbrace{\mathbb{E}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H] P(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H)}_{\text{observational distribution}} \underbrace{\prod_{i=1}^{H} \frac{\pi_i(x_i \mid \boldsymbol{s}_i)}{P(x_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_i)}}_{\text{ratio } \pi \text{ and obs. probabilities}} . \tag{138}$$

- Sampling Bias Estimate: Simple average mean is a biased estimate due to sampling bias such as selection bias, missing data, or unobserved variables.
- Intuition: IPW “upweights” or “downweights” individual data points based on how likely (or unlikely) the action taken in the historical data was under the behavior vs. the target policy, which giving us a fairer estimate of how the new policy would perform.
- NUC ensures that behavior probability (propensity score) can be accurately specified, which means no other unobserved variables determines action x_i

$$\mathbb{E}_{\pi}\left[\mathcal{R}(\boldsymbol{Y})\right] = \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} \underbrace{\mathbb{E}\left[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H\right] P\left(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H\right)}_{\textit{observational distribution}} \underbrace{\prod_{i=1}^{H} \frac{\pi_i\left(x_i \mid \boldsymbol{s}_i\right)}{P\left(x_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_i\right)}}_{\textit{ratio } \pi \textit{ and obs. probabilities}} .$$

Upweighting Rare Actions:

If $P(a \mid x)$ is small (the action was rarely chosen by the behavior policy), but $\pi(a \mid x)$ is large (the action is frequently chosen by the new policy), the ratio increases. This upweights the observed reward because the target policy would have chosen that action more often.

Downweighting Overrepresented Actions:

If $P(a \mid x)$ is large (the action was commonly chosen by the behavior policy), but $\pi(a \mid x)$ is small (the action is infrequent under the new policy), the ratio decreases. This downweights the observed reward because the target policy would have chosen that action less often.

Example 27 *Consider again the CDM $\langle \mathcal{M}^*, \Pi, Y \rangle$ described in Eq. 78 where coefficients $\alpha_1 = \alpha_2 = 0$. We will apply IPW estimation to evaluate the effects of the policy $\pi = (X_1 \leftarrow 0, X_2 \leftarrow 1)$ from the observational distribution $P(S_1, X_1, S_2, X_2, Y)$. Applying the estimation formula provided by Thm. 17 gives*

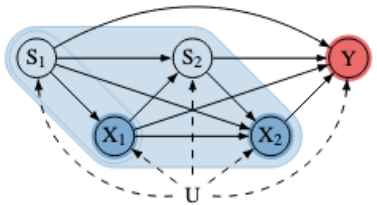
$$\mathbb{E}^{\text{IPW}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = \sum_{s_1, x_1, s_2, x_2} P(s_1, x_1, s_2, x_2, Y = 1) \frac{\mathbb{1}\{x_1 = 0\}}{P(x_1 \mid s_1)} \frac{\mathbb{1}\{x_2 = 1\}}{P(x_2 \mid s_1, x_1, s_2)} \tag{139}$$

$$\mathbb{E}^{\text{IPW}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = \frac{P(S_1 = 0, X_1 = 0, S_2 = 0, X_2 = 1, Y = 1)}{P(X_1 = 0 \mid S_1 = 0) P(X_2 = 1 \mid S_1 = 0, X_1 = 0, S_2 = 0)} \tag{140}$$

$$+ \frac{P(S_1 = 0, X_1 = 0, S_2 = 1, X_2 = 1, Y = 1)}{P(X_1 = 0 \mid S_1 = 0) P(X_2 = 1 \mid S_1 = 0, X_1 = 0, S_2 = 1)} \tag{141}$$

$$+ \frac{P(S_1 = 1, X_1 = 0, S_2 = 0, X_2 = 1, Y = 1)}{P(X_1 = 0 \mid S_1 = 1) P(X_2 = 1 \mid S_1 = 1, X_1 = 0, S_2 = 0)} \tag{142}$$

$$+ \frac{P(S_1 = 1, X_1 = 0, S_2 = 1, X_2 = 1, Y = 1)}{P(X_1 = 0 \mid S_1 = 1) P(X_2 = 1 \mid S_1 = 1, X_1 = 0, S_2 = 1)} \tag{143}$$



(c) DTR

Evaluating the above equation gives $\mathbb{E}^{\text{IPW}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = 0.6757$, which matches the expected reward in Eq. 78, evaluated directly in the SCM $\mathcal{M}^*$. ■

S_1	X_1	S_2	X_2	Y	$P(s_1, x_1, s_2, x_2, y)$	S_1	X_1	S_2	X_2	Y	$P(s_1, x_1, s_2, x_2, y)$
0	0	0	0	0	0.0128	1	0	0	0	0	0.0042
0	0	0	0	1	0.0466	1	0	0	0	1	0.0011
0	0	0	1	0	0.0009	1	0	0	1	0	0.0011
0	0	0	1	1	0.0585	1	0	0	1	1	0.0042
0	0	1	0	0	0.0013	1	0	1	0	0	0.0004
0	0	1	0	1	0.0049	1	0	1	0	1	0.0001

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Dynamic Programming

Theorem 18 (Dynamic Programming) *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ and $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$. If NUC holds, for any $\pi \in \Pi$, the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ is computable from the joint distribution $P(\boldsymbol{V})$ as follows:*

$$\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})] = \mathbb{E}\left[\sum_{x_1} Q_\pi^{(1)}(x_1, \boldsymbol{S}_1)\pi_1(x_1 \mid \boldsymbol{S}_1)\right], \tag{145}$$

where the value function $Q_\pi^{(i)}(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i)$, for $i = 1, \ldots, H-1$, is given by:

$$Q_\pi^{(i)}(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i) = \mathbb{E}\left[\sum_{x_{i+1}} Q_\pi^{(i+1)}(\bar{\boldsymbol{x}}_{i+1}, \bar{\boldsymbol{s}}_i, \boldsymbol{S}_{i+1})\pi_{i+1}(x_{i+1} \mid \boldsymbol{S}_{i+1}) \,\middle|\, \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i\right] \tag{146}$$

$$\text{and} \quad Q_\pi^{(H)}(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H) = \mathbb{E}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H] \tag{147}$$

- **At the final stage H:** $Q_\pi^{(H)}(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H)$ is just the *expected reward* given the final state/action history.
- **At each intermediate stage i:** you compute $Q_\pi^{(i)}$ by taking an *expectation* of the next stage's value function $Q_\pi^{(i+1)}$, weighted by the policy π_{i+1} over possible actions x_{i+1}.

Example 28 *Consider again the CDM $\langle \mathcal{M}^*, \Pi, Y \rangle$ described in Eq. 78 where coefficients $\alpha_1 = \alpha_2 = 0$. We will apply the DP estimation to evaluate the effect of the policy $\pi = (X_1 \leftarrow 0, X_2 \leftarrow 1)$. Thm. 18 allows us to estimate the expected reward $\mathbb{E}_\pi[Y]$ from the observational distribution $P(S_1, X_1, S_2, X_2, Y)$ as follows:*

$$Q_\pi^{(1)}(s_1, x_1) = \sum_{s_2, x_1} Q_\pi^{(2)}(s_1, x_1, s_2, x_2) \mathbb{1}\{x_2 = 1\} P(s_2 \mid x_1, s_1) \tag{148}$$

$$Q_\pi^{(2)}(s_1, x_1, s_2, x_2) = P(Y = 1 | s_1, x_1, s_2, x_2) \tag{149}$$

We compute the parametrization of value functions $Q_\pi^{(1)}(s_1, x_1), Q_\pi^{(2)}(s_1, x_1, s_2, x_2)$ and provide them in Table 7. The expected reward of the policy $\pi = (X_1 \leftarrow 0, X_2 \leftarrow 1)$ is computable as

$$\mathbb{E}^{\text{DP}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = \sum_{s_1} Q_\pi^{(1)}(s_1, x_1) \mathbb{1}\{x_1 = 0\} P(s_1) \tag{150}$$

Evaluating the above equation gives $\mathbb{E}^{\text{DP}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = 0.6757$, which matches the expected reward in Example 12, evaluated in the SCM $\mathcal{M}^*$. ■

S_1	X_1	$Q_\pi^{(1)}$	S_1	X_1	$Q_\pi^{(1)}$
0	0	0.8799	1	0	0.4716
0	1	0.8749	1	1	0.1003

(a) $Q_\pi^{(1)}(s_1, x_1)$

S_1	X_1	S_2	X_2	$Q_\pi^{(2)}$	S_1	X_1	S_2	X_2	$Q_*^{(2)}$
0	0	0	0	0.7851	1	0	0	0	0.2149
0	0	0	1	0.9846	1	0	0	1	0.7851
0	0	1	0	0.7851	1	0	1	0	0.2149
0	0	1	1	0.7851	1	0	1	1	0.2149
0	1	0	0	0.2149	1	1	0	0	0.0008
0	1	0	1	0.9846	1	1	0	1	0.2149
0	1	1	0	0.2149	1	1	1	0	0.0008
0	1	1	1	0.7851	1	1	1	1	0.0154

(b) $Q_\pi^{(2)}(s_1, x_1, s_2, x_2)$

Table 7: Evaluation of value functions $Q_\pi^{(1)}(s_1, x_1), Q_\pi^{(2)}(s_1, x_1, s_2, x_2)$ for the policy $\pi = (X_1 \leftarrow 0, X_2 \leftarrow 1)$ in evaluated in 2-stage DTR environment described in Example 12.

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Online Learning Task

Formally, an online learning task is described by the following signature:

$$\mathcal{T}_{\text{on}} = \left\langle \mathcal{R} = \text{do}, \mathcal{A} = \emptyset, \Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}, \mathcal{R} = \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R} \right\rangle \tag{155}$$

To see the specific optimization in this task, the agent will search for a policy π^* such that

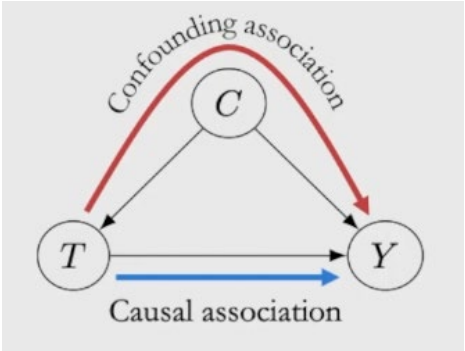
$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \mid \mathcal{D}_{\text{exp}} \sim P_{\boldsymbol{x}}(\boldsymbol{V}) \right], \tag{156}$$

- Online Learning
  - The agent interacts with the environment and learns continuously, updating its policy incrementally as it gathers new data.
- Offline Learning
  - The agent learns from a fixed dataset of experiences without any new interactions with the environment during training.

Experimental NUC

Lemma 19 (Experimental NUC) *Let* $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ *be a CDM where* $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{X_i \in \boldsymbol{X}}$ *and* $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$. *For any policy* $\pi \in \Pi$, *the NUC condition (Def. 16) holds in* $\langle \mathcal{M}^*_\pi, \Pi, \mathcal{R} \rangle$ *induced by intervention* $do(\pi)$. ■

Note that for every policy, in the submodel M induced by intervention do(X), all input covariates S affecting every action X and other variables in the system are observed and measured. This means that the NUC condition is implied in the post-interventional system.



1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Randomized Controlled Trails (RCT)

Algorithm 2 Randomized Controlled Trails (RCT)

Require: the policy space Π, the total number of trials $N \in \mathbb{N}$.

for all episodes $t = 1, 2, \ldots$ **do**
  Choose a policy $\pi^{(t)}$ as follows.
  if $t \leq N$ **then**
    Let $\pi^{(t)}$ be a uniform policy

$$\pi_{\text{UNIF}} = (X_1 \sim \texttt{Unif}(\mathscr{D}(X_1)), \ldots, X_H \sim \texttt{Unif}(\mathscr{D}(X_H))) \,. \tag{160}$$

  else
    Let $\pi^{(t)} = \arg\max_{\pi \in \Pi} \hat{\mathbb{E}}_{\pi}^{(N)}[Y]$.
  end if
  Perform $\text{do}(\pi^{(t)})$ for episode t and receive observations $\boldsymbol{V}^{(t)}$.
end for

- RCT could compute reward estimates $\hat{E}_{\pi}^{(N)}[Y]$ for candidate policies $\pi \in \Pi$ from the experimental data $P_{\pi_{\text{UNIF}}}(V)$ collected during the exploration.
- Finally, RCT selects a policy with the highest empirical reward estimates and commits to it for all future episodes $t > N$.

Cumulative Regret

There are several ways to measure the performance of online learning algorithms. One popular criterion is to study the algorithm's *cumulative regret* (Auer et al., 2002a), which measures its cumulative loss relative to an optimal strategy that always selects the optimal arm x^*. Formally, the cumulative regret for an online learning algorithm in an MAB environment $\mathcal{M}^*$ after T episodes of trials can be defined as:

$$R(T, \mathcal{M}^*) = \underbrace{T\mathbb{E}_{x^*}\left[Y; \mathcal{M}^*\right]}_{\text{Optimal Reward}} - \underbrace{\sum_{t=1}^{T} \mathbb{E}_{X^{(t)}}\left[Y; \mathcal{M}^*\right]}_{\text{Realized}}. \tag{168}$$

- Measures how much reward is *lost* by not always picking the optimal arm.
- Minimizing $R(T, \mathcal{M}^*)$ is equivalent to *maximizing* total reward.

Regret of RCT

- **RCT Algorithm**
  1. **Exploration Phase (N episodes)**:
     - Plays each arm equally often ($\frac{N}{K}$ times per arm).
     - Gathers unbiased estimates of each arm's reward.
  2. **Exploitation Phase ($T - N$ episodes)**:
     - Commits to the arm with the *highest empirical reward* from exploration.
- **Regret Components**
  1. **Exploration Regret**: Paying a cost by *trying all arms*, including suboptimal ones, for N steps.
  2. **Exploitation Regret**: Risk of choosing a suboptimal arm if the empirical estimate is inaccurate.

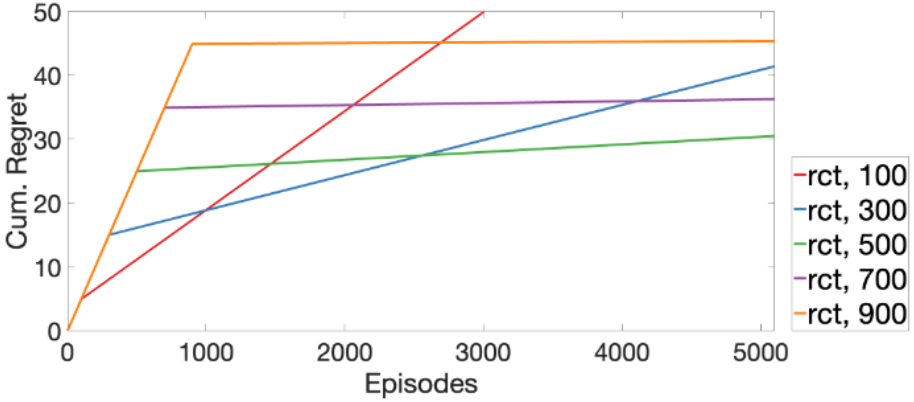


Figure 15: The regret of `RCT` with varying total number of trials.

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Upper Confidence Bound (UCB)

Motivation: RCT requires prior knowledge of the total number of episodes T. UCB is an adaptive randomization strategy which requires no prior knowledge.

$$\mathrm{UCB}_t(x,\delta) = \underbrace{\widehat{E}_x^{(t)}[Y]}_{\text{empirical mean}} + \underbrace{\sqrt{\frac{\log(1/\delta)}{2N_x(t)}}}_{\text{confidence width}},$$

where $\widehat{E}_x^{(t)}[Y]$ is the sample average reward for arm x based on the data so far, $N_x(t)$ is the number of times x has been pulled, and δ governs the probability of bounding the true mean.

Algorithm 3 Upper Confidence Bound (UCB) in MAB

Require: a policy scope $\Pi = \{\langle X, \emptyset \rangle\}$.

1: **for all** episodes $t = 1, 2, \ldots$ **do**
2: Choose an arm $x^{(t)} = \arg\max_x \text{UCB}_{t-1}(x, \delta)$ where $\delta = t^{-4}$.
3: Perform do $\left(x^{(t)}\right)$ for episode t and receive reward $Y^{(t)}$.
4: **end for**

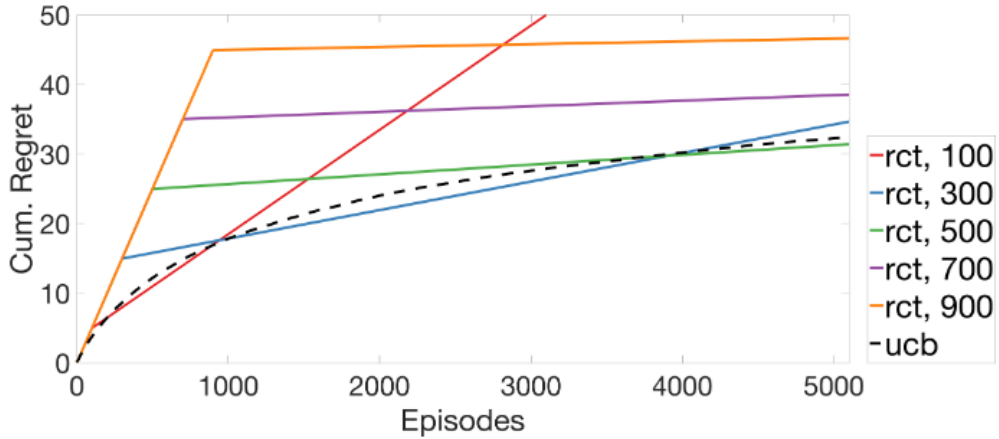


Figure 16: Simulation results comparing performance of online learning algorithms UCB and RCT.

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Causal Identification Task

$$\mathcal{T}_{\text{id}} = \left\langle \mathcal{R} = \text{see}, \mathcal{A} = \mathcal{G}, \Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}, \mathcal{R} = \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R} \right\rangle \tag{186}$$

The optimization in this task goes as follows:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \mid \mathcal{A} = \mathcal{G},\ \mathcal{D}_{\text{obs}} \sim P(\boldsymbol{V}) \right], \tag{187}$$

Motivation:

1. For online learning, randomized experiments are not applicable in all settings
2. For off policy learning, NUC assumption is often fragile in practice.
3. An alternative approach for policy evaluation is to relax the NUC assumption and explore other more general structural assumptions in a causal diagram.

Identifiability

Definition 22 (Identifiability) *Let subsets of variables $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$ and π be a policy over $\boldsymbol{X}$. The interventional distribution $P_\pi(\boldsymbol{Y})$ is identifiable from the structural assumptions $\mathcal{A}$ if $P_\pi(\boldsymbol{Y})$ is uniquely computable from any positive observational distribution $P(\boldsymbol{V})$ in any SCM $\mathcal{M}$ satisfying $\mathcal{A}$. That is, if for every pair of SCMs $\mathcal{M}_1$ and $\mathcal{M}_2$ compatible with structural assumptions $\mathcal{A}$, $P_\pi(\boldsymbol{Y}; \mathcal{M}_1) = P_\pi(\boldsymbol{Y}; \mathcal{M}_2)$ whenever $P(\boldsymbol{V}; \mathcal{M}_1) = P(\boldsymbol{V}; \mathcal{M}_2) > 0$.*

Recall:

1. IPS

$$\mathbb{E}_\pi\left[\mathcal{R}(\boldsymbol{Y})\right] = \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} \underbrace{\mathbb{E}\left[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H\right] P\left(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H\right)}_{\text{observational distribution}} \underbrace{\prod_{i=1}^{H} \frac{\pi_i\left(x_i \mid \boldsymbol{s}_i\right)}{P\left(x_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_i\right)}}_{\text{ratio } \pi \text{ and obs. probabilities}}. \tag{138}$$

1. DP

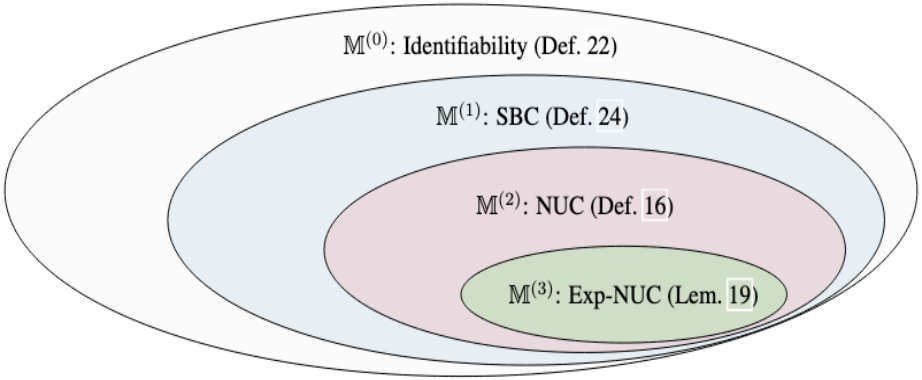
$$\mathbb{E}_\pi\left[\mathcal{R}(\boldsymbol{Y})\right] = \mathbb{E}\left[\sum_{x_1} Q_\pi^{(1)}(x_1, \boldsymbol{S}_1)\pi_1(x_1 \mid \boldsymbol{S}_1)\right],$$

NUC Implies Identification

Corollary 23 *For a policy space* $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ *and reward function* $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$*, let a policy* $\pi \in \Pi$*.* $P_\pi(\boldsymbol{Y})$ *is identifiable from the NUC condition (Def. 16) w.r.t. the policy space* Π*.*

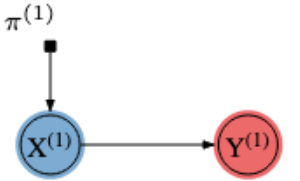
Intuition: Structure assumption NUC don't require a causal diagram. No unobserved confounder (NUC) implicitly enforces all the possible causal diagram in a certain way that only the action variable X can cause reward Y.

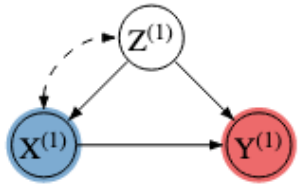
4 Structure Assumptions



Summary of the relationships of structural assumptions about the data generating process discussed so far that permit the evaluation of the effects of candidate policies

1. Online Learning Task
    a. Implies Exp-NUC
    b. Strongest assumption
    c. Could have unobserved confounder
    d. agent goes online in the environment and collects experimental data
2. Off policy Learning Task
    a. Assummes NUC
    b. Weaker assumption
    c. Need to condition on the observable confounder to identify the causal relationship
3. Casual Identification (Causal Diagram Based)
    a. Sequential Backdoor Condition (SBC)
        i. Broader assumption
        ii. Could have unobserved confounders, as long as it was blocked by the backdoor path
    b. Identifiability
        i. Most general assumption
        ii. Complete and sound
        iii. Using Do-calculus and mathematical rules to identify causal effects from the observational distribution





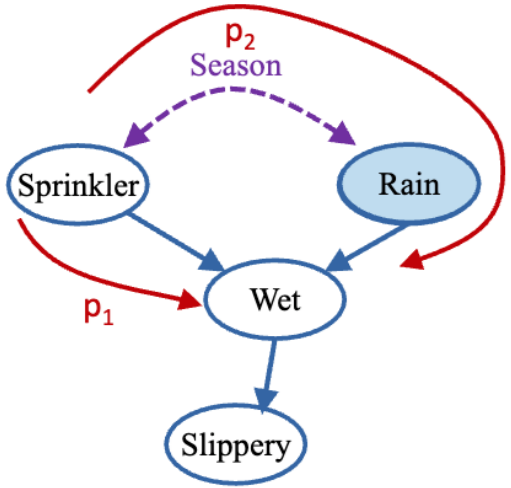
1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Identifiability: Backdoor Criterion

- A set Z satisifies the **backdoor criterion** if
  - No $Z_i \; in \; Z$ is a descendant of X
  - Z blocks every path between X,Y that has an arrow into X
- Then, $p(Y|\mathrm{do}(X)) = \sum_Z p(Y|X,Z)\,p(Z)$

Ex: What if Season is latent?

- Z={Rain} for X=Sprinkler, Y=Wet
  - Conditioning on Rain blocks the non-causal path p_2
  - Leaves the causal path p_1 unaffected!



Sequential Backdoor Condition (SBC)

Definition 24 (Sequential Backdoor Condition (Policy Intervention)) *Let $\mathcal{G}$ be causal diagram and $\boldsymbol{Y} \subseteq \boldsymbol{V}$ be reward signals. A policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ is said to satisfy the* sequential backdoor condition *w.r.t. $\boldsymbol{Y}$ in $\mathcal{G}$ (for short, Π is backdoor admissible) if for every policy $\pi \in \Pi$, the following condition hold: for every $i = 1, \ldots, H$,*

$$\left(X_i \perp\!\!\!\perp \boldsymbol{Y} \mid \bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i\right)_{\mathcal{G}_{\underline{X_i}, \pi_{i+1}, \ldots, \pi_H}} \tag{192}$$

That is, conditioning on nodes $\bar{\boldsymbol{X}}_{i-1} \cup \bar{\boldsymbol{S}}_i$ d-separates all backdoor paths *from action X_i to reward signals $\boldsymbol{Y}$ that contains an arrow pointing into X_i in the manipulated graph $\mathcal{G}_{\pi_{i+1}, \ldots, \pi_H}$.*

1. Recap
    a. Structural Causal Model (SCM)
    b. RL Markov Decision Process (MDP)
2. CRL Definition
    a. Causal Decision Model (CDM)
        i. Policy Space
    b. CRL Tasks
        i. Learning Regimes
        ii. Structure Assumption
3. 3 RL tasks through causal lenses
    a. Off-policy learning
        i. Inverse Propensity Weighting (IPW)
        ii. Dynamic Programming (DP)
    b. Online learning
        i. Randomized Controlled Trials (RCT)
        ii. Upper Confidence Bound (UCB)
    c. Causal Identification
        i. Sequential Backdoor Condition (SBC)
        ii. Do-Calculus learning

Do-calculus Learning

Motivation: There are learning settings where the sequential backdoor criterion does not hold, but the agent could still explore the structural constraints in the causal diagram to recover the effects of candidate policies from the observational data.

Proposition 29 *Rules of Do-calculus, together with standard probability manipulations, are sound and complete for determining the identifiability of all interventional distributions of the form* $P_{\boldsymbol{x}}(\boldsymbol{Y})$ *from a causal diagram* $\mathcal{G}$ *and the available observational and interventional distributions.* ■

Do-calculus Learning

Theorem 28 (Rules of do-calculus (Pearl, 2000)) *Let $\mathcal{G}$ be a causal diagram compatible with a structural causal model $\mathcal{M}$, with endogenous variables $\boldsymbol{V}$. For any disjoint subsets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W} \subseteq \boldsymbol{V}$, the following rules hold for interventional distributions compatible with $\mathcal{G}$:*

Rule 1 *Insertion/deletion of observations:*

$$P_{\boldsymbol{x}}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}) = P_{\boldsymbol{x}}(\boldsymbol{y} \mid \boldsymbol{w}) \text{ if } (\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{W}) \text{ in } \mathcal{G}_{\overline{\boldsymbol{X}}} \tag{215}$$

Rule 2 *Action/observation exchange:*

$$P_{\boldsymbol{x},\boldsymbol{z}}(\boldsymbol{y} \mid \boldsymbol{w}) = P_{\boldsymbol{x}}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}) \text{ if } (\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{W}) \text{ in } \mathcal{G}_{\overline{\boldsymbol{X}}\underline{\boldsymbol{Z}}} \tag{216}$$

Rule 3 *Insertion/deletion of actions:*

$$P_{\boldsymbol{x},\boldsymbol{z}}(\boldsymbol{y} \mid \boldsymbol{w}) = P_{\boldsymbol{x}}(\boldsymbol{y} \mid \boldsymbol{w}) \text{ if } (\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{W}) \text{ in } \mathcal{G}_{\overline{\boldsymbol{X}}\overline{\boldsymbol{Z}(\boldsymbol{W})}} \tag{217}$$

where $\boldsymbol{Z}(\boldsymbol{W})$ is the subset of nodes in $\boldsymbol{Z}$ that are not ancestors of $\boldsymbol{W}$-nodes in $\mathcal{G}_{\overline{\boldsymbol{X}}}$ ■

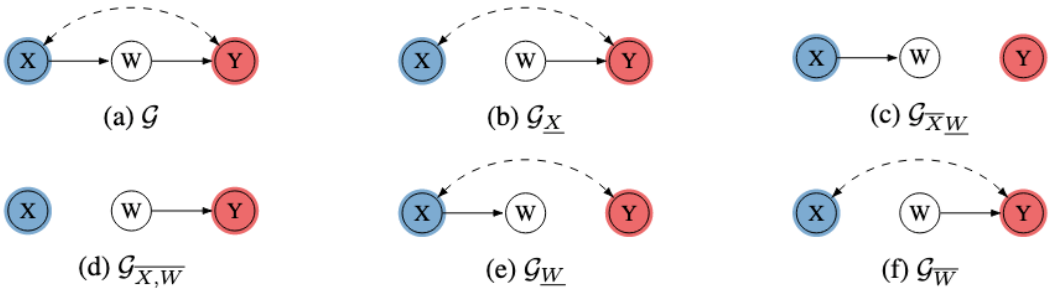


Figure 21: A front-door graph and its manipulated representations.

$$
\begin{aligned}
P_x(y,w) &= P_x(y|w)P_x(w) && \textit{Probability Axioms} && (218)\\
&= P_x(y|w)P(w|x) && \textit{Rule 2 } (W \perp\!\!\!\perp X)_{\mathcal{G}_{\underline{X}}} && (219)\\
&= P_{x,w}(y)P(w|x) && \textit{Rule 2 } (Y \perp\!\!\!\perp W \mid X)_{\mathcal{G}_{\overline{X}\underline{W}}} && (220)\\
&= P_w(y)P(w|x) && \textit{Rule 3 } (Y \perp\!\!\!\perp X \mid W)_{\mathcal{G}_{\overline{X,W}}} && (221)\\
&= P(w|x)\sum_{x'} P_w(y|x')P_w(x') && \textit{Probability Axioms} && (222)\\
&= P(w|x)\sum_{x'} P(y|w,x')P_w(x') && \textit{Rule 2 } (Y \perp\!\!\!\perp W \mid X)_{\mathcal{G}_{\underline{W}}} && (223)\\
&= P(w|x)\sum_{x'} P(y|w,x')P(x') && \textit{Rule 3 } (X \perp\!\!\!\perp W)_{\mathcal{G}_{\overline{W}}} && (224)
\end{aligned}
$$

		Signature				
	Task	**Learning Regime** ($\mathcal{L}$)	**Structural Assumptions** ($\mathcal{A}$)	**Policy Space** (Π)	**Reward Function** ($\mathcal{R}$)	**Section**
1	Off-policy Learning	See	NUC	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.1
2	Online Learning	Do	-	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.2
3	Causal Identification	See	DAG $\mathcal{G}$	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.3
4	Offline-to-Online Learning	See + Do	-	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	5
5	Where to do & What to see	Do	DAG $\mathcal{G}$	Π_{MIX}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	6
6	Counterfactual randomization	Ctf-Do	-	Π_{CTF}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	7
7	Causal Imitation Learning	See	DAG $\mathcal{G}$	Π_{EXP}	-	8

Table 11: Summary of causal reinforcement learning tasks investigated in this paper, in terms of their signatures and sections. We highlight in gray the most distinct feature introduced by the task.

Question

Why online learning implies NUC?

END

Thank You!