Seoul National University
Graduate School of Data Science

On Transportability for Structural Causal Bandits

Min Woo Park **Sanghack Lee**

Graduate School of Data Science, Seoul National University

Multi-Armed Bandits

A **Multi-armed bandit** (**MAB**) is a classic sequential decision-making problem.

Actions A set of actions **A** to choose from;
each action is associated with a reward distribution.

Play At each round $t \in [T]$, the learner chooses an action.

Reward A reward is drawn from the reward distribution of the chosen action.

Goal Minimize cumulative regret over T rounds.

Structural Causal Bandits (SCB)

- Structural causal bandits view the MAB problem through a **causal lens**.
- **Structural Causal Bandits** (**SCB**): the environment is an SCM $\mathcal{M}$ with a reward variable $Y \in \mathbf{V}$.

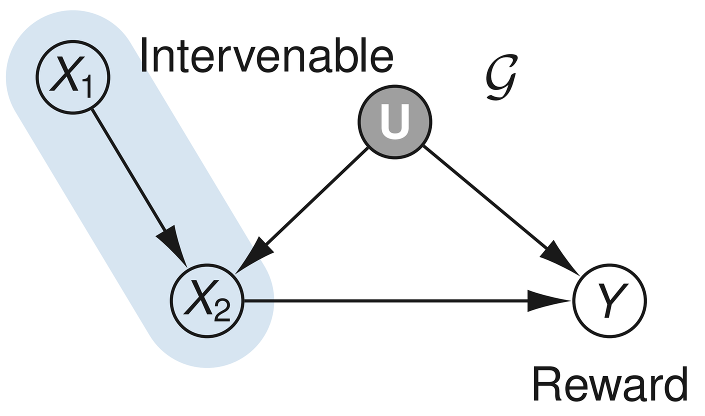
  Intervention sets (ISs) correspond to subsets of $\mathbf{V} \setminus \{Y\}$.

  Actions correspond to **interventions** that assign specific values to variables in an **IS**, i.e., $do(\mathbf{X} = \mathbf{x})$.

  The **reward distribution** is the distribution of Y under intervention $do(\mathbf{x})$, i.e., $P(Y \mid do(\mathbf{x}))$.

  The mean reward is $\mu_{\mathbf{x}} := \mathbb{E}[Y \mid do(\mathbf{x})]$.
- **Assumption**. The causal diagram $\mathcal{G}$, representing the environment class of $\mathcal{M}$, is accessible.

  Example. Suppose we can intervene on $\underbrace{\text{2 binary variables}}$, X_1 and X_2.



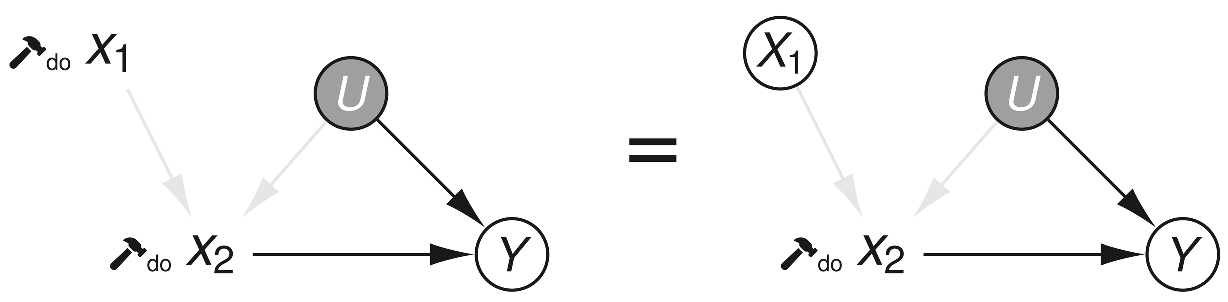
$$\mathcal{M} = \begin{cases} \mathbf{U} = \{U\} \\ \mathbf{V} = \{X_1, X_2, Y\} \\ \mathcal{F} = \begin{cases} f_{X_2}(x_1, u) = x_1 \oplus u \\ f_Y(x_2, u) = x_2 \oplus u \end{cases} \\ P(\mathbf{U}) = \{U \sim \mathtt{Bern}(0.5)\}. \end{cases}$$

Intervention sets (ISs) (**4**): $\emptyset$, $\{X_1\}$, $\{X_2\}$, $\{X_1, X_2\}$.

Actions (**9**): $do(\emptyset)$, $do(X_1 = 0)$, $do(X_1 = 1)$, $do(X_2 = 0)$, $\cdots$, $do(x_1 = 1, x_2 = 1)$.

Structural Properties

Property 1 (Equivalence among Actions). Two actions can share the same reward distribution.



$$\mu_{x_1, x_2} = \mu_{x_2}$$

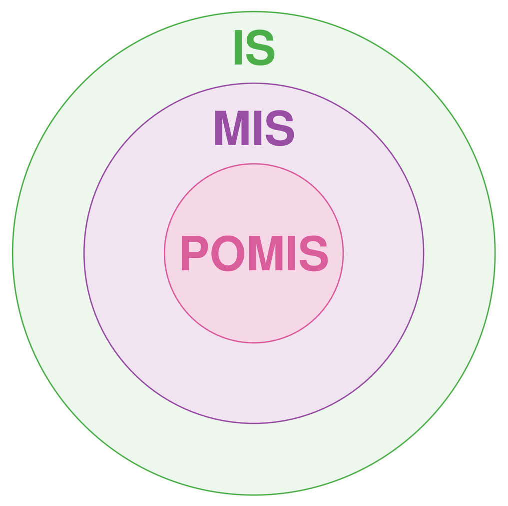
Minimal Intervention Set (**MIS**): A minimal set of variables among **ISs** sharing the same reward distribution.

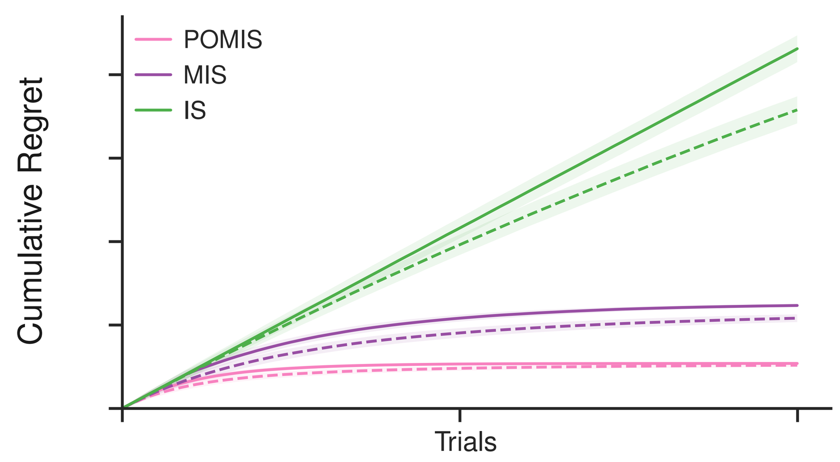
Property 2 (Partial Ordering). The maximum achievable expected rewards can be partially ordered.

$$\mu_{\emptyset} = \sum_{x_2} \mu_{x_2} P(x_2) \leq \sum_{x_2} \mu_{x_2^*} P(x_2) = \mu_{x_2^*}.$$

Possibly-Optimal Minimal Intervention Set (**POMIS**): An **MIS** that can achieve an optimal expected reward in some SCM $\mathcal{M}$.

$\therefore$ It suffices to explore only **POMISs**! $\Rightarrow do(X_1 = 0), do(X_1 = 1), do(X_2 = 0), do(X_2 = 1)$.





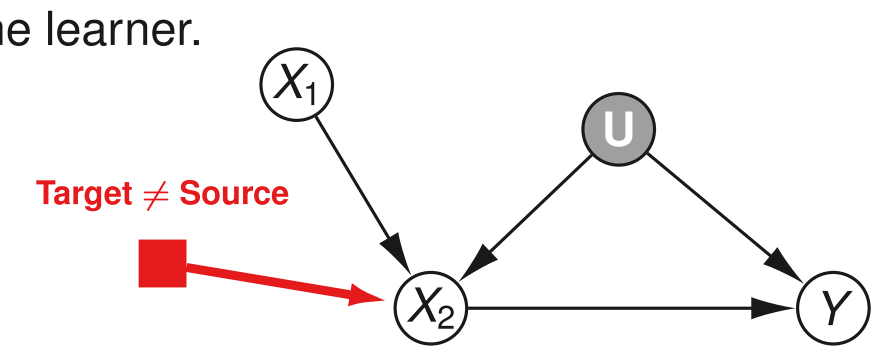
Inclusion: **ISs** $\supseteq$ **MISs** $\supseteq$ **POMISs** $\Rightarrow \emptyset, \{X_1\}, \{X_2\}, \{X_1, X_2\}$.
Cumulative Regret: **ISs** $\geq$ **MISs** $\geq$ **POMISs** (smaller is better).

Quick start. Given a causal diagram $\mathcal{G}$:

1. **Offline**: Identify **POMISs** and remove the other actions.
2. **Online**: Run a MAB solver (e.g., UCB, TS) over the **POMISs**.

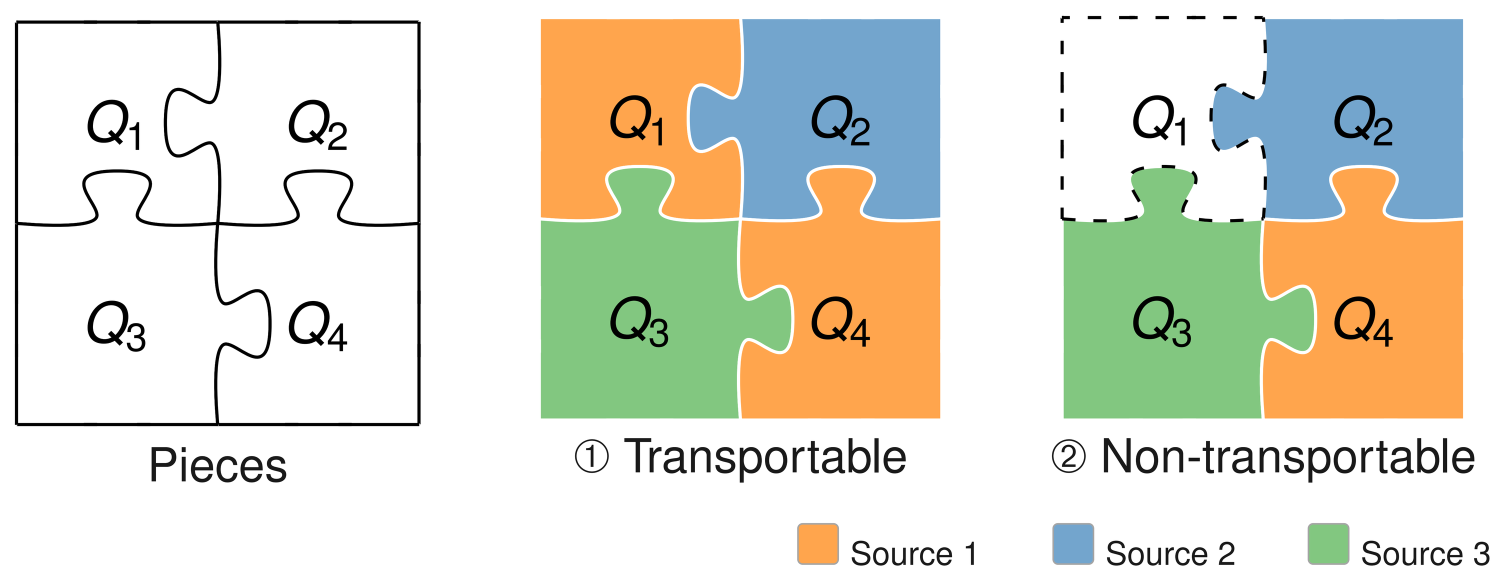
On Transportability for SCB

- Even after restricting exploration to **POMISs**, a learner must estimate their expected rewards in the **target environment** from scratch—*"cold-start"* problem.
- If we instead have **data** from related **source** environments, we can use it to **reduce online exploration** in the **target** environment.
- **Assumption**. The **discrepancy**-augmented causal diagram and **source data** are available to the learner.



❶ **Causal bounds.** For each action $do(\mathbf{X} = \mathbf{x})$, we can compute $[\ell_{\mathbf{x}}, u_{\mathbf{x}}]$.

$$\mu_{\mathbf{x}}^{\text{target}} = \sum_{\mathbf{y}^+} y \underbrace{\prod_{q=1}^{m} Q^{\text{target}}[\mathbf{C}_q]}_{\text{Product of pieces}}.$$

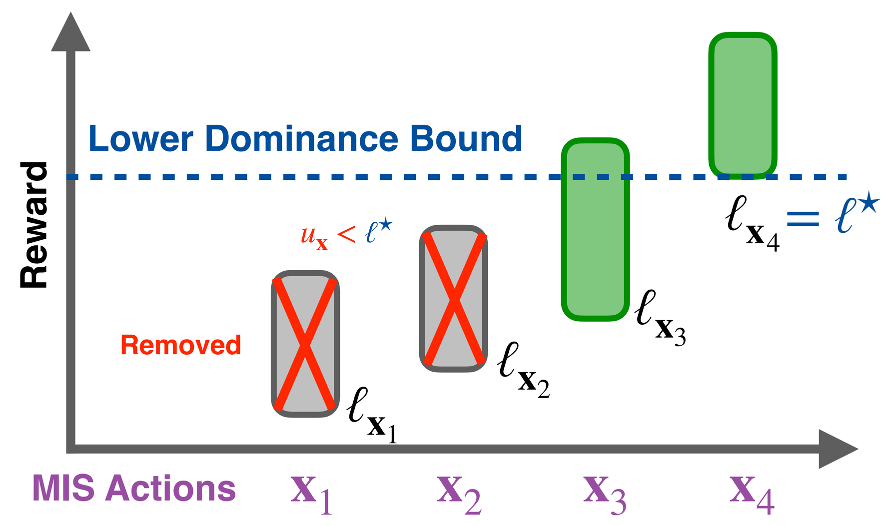


① We can obtain a **point estimate** $\mu_{\mathbf{x}}^{\text{target}} = \sum_{\mathbf{y}^+} y \cdot Q_1 \cdot Q_2 \cdot Q_3 \cdot Q_4$.
② We can obtain an offline **interval estimate** $\mu_{\mathbf{x}}^{\text{target}} \in [\ell_{\mathbf{x}}, u_{\mathbf{x}}]$.

❷ **Dominance bounds.** Dominance relationships yield an interval $[\ell^*, u^*]$.

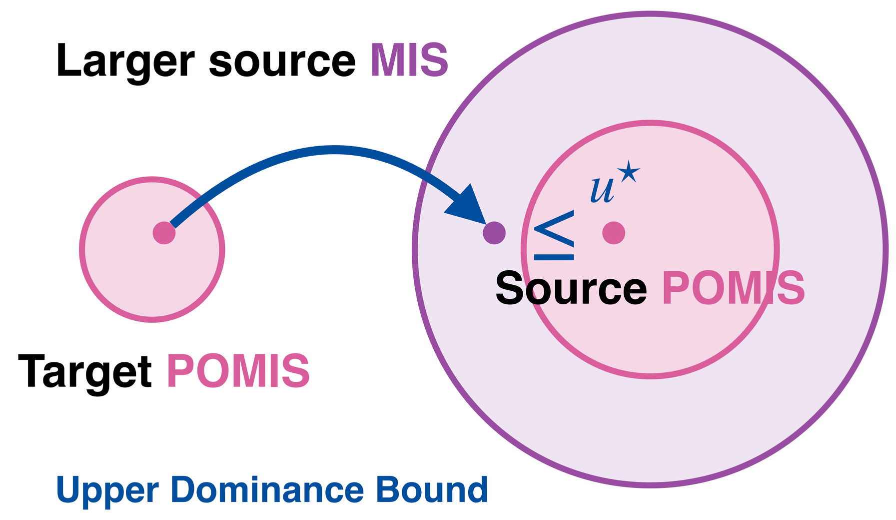
◇ **Lower dominance bound.** $\ell^* = \max_{\text{MIS}} \ell_{\mathbf{x}}$

$\Rightarrow$ Any action $\mathbf{x} \in$ **POMIS** satisfying $u_{\mathbf{x}} < \ell^*$ can never be optimal.
$\Rightarrow$ Such actions can be safely eliminated.



◇ **Upper dominance bound.** A **POMIS** in the **target** environment is an **MIS** in a **source** environment with a larger intervention set, i.e. $\mathbf{V}_{\text{target}} \subsetneq \mathbf{V}_{\text{source}}$.

$\Rightarrow$ **POMISs** in the **source** environment serve as an upper bound on **POMISs** in the **target** environment.



Online: Using the bounds from ❶ + ❷, the UCB index can be simply *clipped* to achieve lower cumulative regret than exploring all **POMISs**.

Related Work

Structural Causal Bandits: Where to Intervene?
Lee and Bareinboim. **NeurIPS 2018**.

Structural Causal Bandits under Markov Equivalence
Park, Arditi, Bareinboim and Lee. **NeurIPS 2025**.

On Structural Causal Bandit with Non-manipulable Variables
Lee and Bareinboim. **AAAI 2019**.

Counterfactual Structural Causal Bandits
Park and Lee. **ICLR 2026**.

Acknowledgements. This work was supported by the IITP (RS-2022-II220953/50%) and NRF (RS-2023-00211904/50%) grants funded by the Korean government.