Non-Stationary Structural Causal Bandits

Yeahoon Kwon Yesong Choe Soungmin Park Neil Dhir† Sanghack Lee†

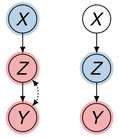
Seoul National University
Graduate School of Data Science

Introduction and Background

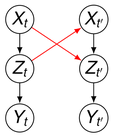
- Dependencies among arms in structural causal bandits (SCM-MABs) extend to temporal dependencies in **non-stationary** environments through the causal lens.
- We introduce **POMIS⁺**, a **temporally-aware** framework that leverages information propagation across time to identify non-myopic optimal intervention sequences.

⟨ **Preliminaries for SCM-MAB** (Lee and Bareinboim, 2018) ⟩

- Minimal Intervention Set (MIS): Minimal variable set among interventions sharing identical rewards. ($\mu_{\mathbf{x},\mathbf{z}} = \mu_{\mathbf{x}}$).
- Possibly-Optimal MIS (POMIS): Let $\mathbf{X}$ be a MIS. If there exists an SCM $\mathcal{M}$ s.t. $\mu_{\mathbf{x}^*} \geq \mu_{\mathbf{z}^*}$, $\forall \mathbf{Z} \in \mathbb{Z} \setminus \{\mathbf{X}\}$, then $\mathbf{X}$ is POMIS.
- Minimal Unobserved Confounder Territory (**MUCT**): A set of endogenous variables governed by a set of unobserved confounders.
- Interventional Border (**IB**): $\mathbf{X} = Pa_{\mathcal{G}}(\mathbf{T}) \setminus \mathbf{T}$, for a MUCT $\mathbf{T}$ w.r.t. Y. (POMIS $\iff$ IB)



Non-Stationary MAB as a Structural Causal Model



$$\mathcal{M}_t: \begin{cases} f_{X_t} = U_{X_t} \\ f_{Z_t} = U_{Z_t} \oplus x_t \\ f_{Y_t} = U_{Y_t} \oplus z_t \end{cases} \quad \mathcal{M}_{t'}: \begin{cases} f_{X_{t'}} = U_{X_{t'}} \oplus z_t \\ f_{Z_{t'}} = U_{Z_{t'}} \oplus x_{t'} \oplus x_t \\ f_{Y_{t'}} = U_{Y_{t'}} \oplus z_{t'} \end{cases}$$

- In the non-stationary SCM-MAB (NS-SCM-MAB), values generated at time t (e.g., z_t, x_t) influence the structural functions at t' through transition edges (in red).

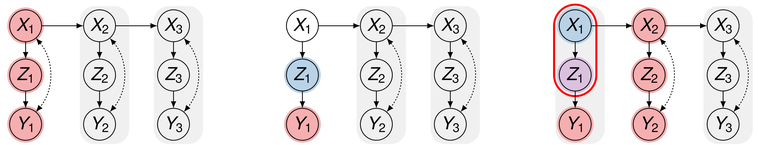
Definition (Non-Stationary SCM-MAB):
A structural causal bandit is **non-stationary** if there exist $t < t'$ such that

$$P\left(Y_t \mid \text{do}(\mathbf{X}_t = \mathbf{x}_t), 1_{t>1} \cdot I_{1:t-1}\right) \neq P\left(Y_{t'} \mid \text{do}(\mathbf{X}_{t'} = \mathbf{x}_{t'}), 1_{t'>1} \cdot I_{1:t'-1}\right)$$

The disparity between reward distributions is referred to as a *reward distribution shift*.

Motivating Example

Sequential intervention selection: ($\mathcal{G}$ for task 1)



- (a–b) At $t = 1$, the local slice shows that $\emptyset$ and $\{Z_1\}$ are the possibily-optimal intervention sets, while X_1 provides no benefit for Y_1.
- (c) At $t = 2$, $\{X_1\}$ is a possibly-optimal intervention set for Y_2 but sub-optimal for Y_1.

POMIS⁺ and Graphical Characterizations

✓ **POMIS with Future Support**

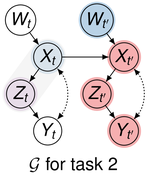
- (Def. 5.2 (**POMIS⁺** simplified ver.)) If $\mathbf{X}_t$ is a POMIS for Y_t and $\mathbf{W}_t$ is a POMIS for $Y_{t'}$, then $(\mathbf{X}_t \cup \mathbf{W}_t)$ can maximize the expected sum of the rewards Y_t and $Y_{t'}$.

✓ **Interventional Border for Future Rewards (IB⁺)**

- IB^+ is defined as the IB for $Y_{t'}$ but in $\mathbf{V}_t$, where $t \leq t'$, it serves as the interventional border for future reward $Y_{t'}$.

✓ **Qualified Interventional Border (QIB)**

- QIB_t is the IB for Y_t preserving the causal influence on Y_t even when $\text{IB}^+_{t,t'}$ is fixed.



$\mathcal{G}$ for task 2

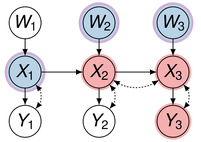
$\Rightarrow$ In other words, IB^+ captures the intervention set for $Y_{t'}$, while QIB captures the intervention set for Y_t. Both are in time step t.

Proposition 6.1 (Composition of POMIS⁺):
Given $\langle \mathcal{G}, \mathbf{Y}_i \rangle$, $\text{IB}^+_{t,t'}(\mathcal{G}[\bigcup_{i=t}^{t'} \mathbf{V}_i], Y_{t'}) \cup \text{QIB}_t(\mathcal{G}[\mathbf{V}_t], Y_t)$ is a $\text{POMIS}^+_{t,t'}$ for $t, t' \in [T]$ and $t < t'$.

Enumeration Algorithm and Experiments

Computing all POMIS⁺ intervention sequences

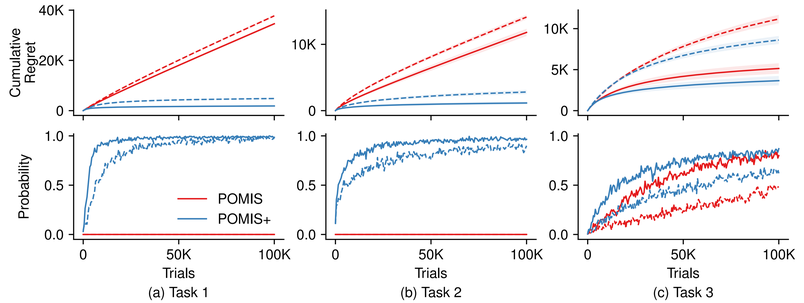
- **Input.** $\mathcal{G}$, $\mathbb{V}$, $\mathbb{Y} = \{Y_1, \ldots, Y_T\}$.
- **Step 1: Recursion in time.** Find the earliest time index t_e where IB^+ appears. If $t_e > 1$, recursively repeat the procedure until $t_e = 1$.
- **Step 2: Combine current and future supports.** When reaching the beginning of the horizon ($t_e = 1$), take the product of IB^+_t and QIB_t (Prop. 6.1) across all time steps: this yields all POMIS^+ intervention sequences.



$\mathcal{G}$ for task 3

Experimental Results

- (**Top**) averaged cumulative regret and (**Bottom**) optimal arm probability
- Thompson Sampling in solid lines, kl-UCB in dashed lines



- Cumulative Regret: POMIS $\geq$ POMIS^+ (the smaller the better)

Discussion and Conclusion

- Introduced NS-SCM-MAB, integrating non-stationary MAB with SCM to capture evolving reward dynamics.
- Formalized **POMIS⁺**, a temporally-extended notion of optimal interventions
- Proposed a recursive algorithm constructing future-aware intervention sequences.

Acknowledgements: This work was partly supported by the IITP (RS-2022-II220953/25%, RS-2025-02263754/25%) and NRF (RS-2023-00211904/25%, RS-2023-00222663/25%) grants funded by the Korean government.