Causal DAG-informed Reward Design: Constructing a Low-Variance Surrogate Reward

Lutong Zou[1,2], Ziping Xu[1], Daiqi Gao[1], Susan Murphy[1]
[1]*Harvard University,* [2]*Peking University.*

Harvard John A. Paulson **School of Engineering** and Applied Sciences

Introduction

Challenge. Real-world RL environments are often highly noisy, leading to high variance. Leveraging structural information about the environment is essential to accelerate learning.

Solution. Actions are usually mediated by intermediate variables, i.e., mediators, which we use to construct *low-variance* surrogates of true rewards. Below we show a causal DAG of a *contextual bandit environment* with perfect surrogacy where M_t blocks all paths from A_t to R_t.

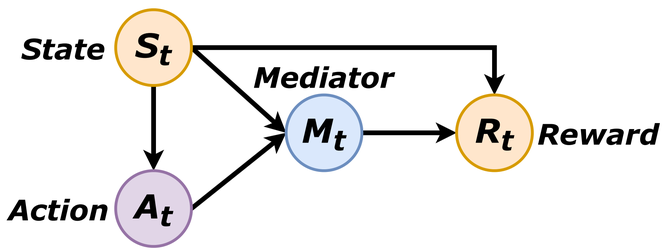

Figure 1: Causal DAG with Perfect Surrogacy

Theorem 1 (Unbiasness and Low-Variance)**.** *The conditional expectation $f^*(M_t, S_t) = \mathbb{E}[R_t \mid M_t, S_t]$ is an unbiased estimation of $\mathbb{E}[R_t \mid A_t, S_t]$ and has strict lower variance.*

Source of Benefits. f^* can be learned on the pooled data across arms, while $\mathbb{E}[R \mid S, A]$ can not.

Contextual Bandit Setup. At each decision time t, given a context $S_t \in \mathbb{R}^{d_S}$ and action A_t, the following linear model generates reward $R_t \in \mathbb{R}$ and mediator $M_t \in \mathbb{R}^{d_M}$

$$M_t = \Gamma_{A_t} S_t + \omega_t,$$

$$R_t = \theta_S^\top S_t + \theta_M^\top M_t + \epsilon_t,$$

- ω_t, ϵ_t Sub-Gaussian noise with variance proxy $\sigma_\omega^2, \sigma_\epsilon^2$
- Γ_a Matrix $\in \mathbb{R}^{d_M \times d_S}$ maps state to mediator for action a
- θ_S/θ_M Vector $\in \mathbb{R}^{d_S}/\mathbb{R}^{d_M}$ maps state/mediator to reward

Method

Algorithm. To leverage established online contextual bandit algorithms, we design a **two-level learning framework**. Given an online learning oracle $\mathcal{O}$ *that takes input of previous reward and current state and outputs an action*, we propose a reward design agent that outputs a surrogate reward R'_t for oracle $\mathcal{O}$ at the each decision time t. See Figure 2.

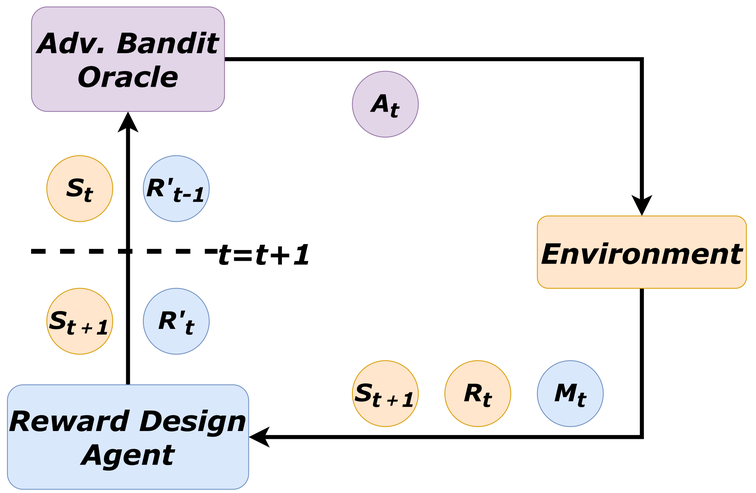

Figure 2: Proposed two-level learning framework.

Reward design agent. At each time t, the reward design agent

1. adds S_{t+1}, M_t, R_t to its observation;
2. gets estimations $\hat{\theta}_{M,t}, \hat{\theta}_{S,t}$ using Ordinary Least Squares;
3. constructs surrogate $R'_t = \hat{\theta}_{M,t}^\top M_t + \hat{\theta}_{S,t}^\top S_t$;
4. passes S_{t+1} and R'_t to the oracle.

Why adversarial? Since $\hat{\theta}_{M,t}, \hat{\theta}_{S,t}$ are updated at each decision time t, the expected reward of an action a evolves accordingly, necessitating the deployment of adversarial bandit oracles. The advantage becomes clearer with higher reward noise (σ_ϵ) and lower mediator noise (σ_ω).

Theoretical Results

Theorem 2 (Regret Bound)**.** *If the oracle has an adversarial regret bound of $\tilde{\mathcal{O}}(\sqrt{d_S|\mathcal{A}|T})$, the true regret bound is*

$$\left(\sigma_\omega\sqrt{d_S|\mathcal{A}|} + \sigma_\epsilon\sqrt{d_M + d_S} + \sigma_\omega\right)\tilde{\mathcal{O}}(\sqrt{T}).$$

- Lower bound **without mediator**: $(\sigma_\epsilon + \sigma_\omega)\tilde{\mathcal{O}}(\sqrt{d_S|\mathcal{A}|T})$
- **Improvement**: $[\sigma_\epsilon(\sqrt{d_S|\mathcal{A}|} - \sqrt{d_M + d_S}) - \sigma_\omega]\tilde{\mathcal{O}}(\sqrt{T})$

Experiments

Algorithm Candidates We compare (1) Reward Design Agent + LinExp3, (2) Reward Design Agent + LinUCB, and (3) LinExp3 against the baseline, LinUCB. Performance is represented as the ratio of reward over the baseline reward minus 1 under varying levels of mediator noise (σ_ω) with a fixed reward noise level ($\sigma_\epsilon = 5$). Positive values imply advantages.

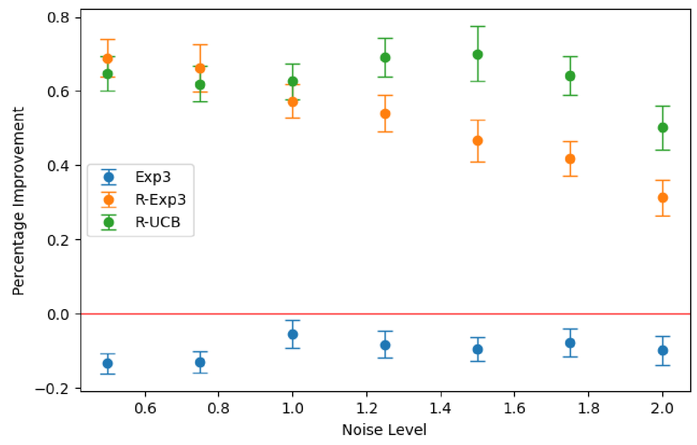

Figure 3: Experiment Result.

Takeaway. Our method significantly enhances performance when (1) the action set is large, (2) rewards are noisy (greater σ_ϵ), and (3) mediator noise (σ_ω) is minimal.