Adaptively Exploiting d-Separators with Causal Bandits

Blair Bilodeau [1,2] Linbo Wang [1] Daniel M. Roy [1,2]

[1] University of Toronto [2] Vector Institute

Contribution Summary

- First **adaptive regret bounds** with respect to causal assumptions.
- **Impossibility result:** no algorithm can be strictly adaptive.
- Novel **lower bounds** for existing causal bandit algorithms.
- General **algorithmic framework** achieves adaptivity with hypothesis testing.

Causal Multi-Armed Bandits

Standard Multi-Armed Bandits

- Sequentially pick intervention $A_t \in \mathcal{A}$
- Observe reward $Y_t \in [0,1]$
- Goal is to learn optimal action $\arg\max_{a\in\mathcal{A}} \mathbb{E}_a Y$

Bandits with Post-Action Contexts

Also observe $Z_t \in \mathcal{Z}$ *after* A_t.

We have no guarantees that observing Z_t will help us...
...but we would like to exploit it when we can.

An **environment** ν is a collection of distributions on $(\mathcal{Z}, \mathcal{Y})$: one for each $a \in \mathcal{A}$.
A **policy** π maps the observed history to actions.

Regret: $R_{\nu,\pi}(T) = T \cdot \max_{a\in\mathcal{A}} \mathbb{E}_{\nu_a}[Y] - \mathbb{E}_{\nu,\pi}[\sum_{t=1}^{T} Y_t]$.

Existing State of the Art

UCB (Auer et al. 2002): For any ν, $R_{\nu,\text{UCB}}(T) = \tilde{\Theta}\left(\sqrt{|\mathcal{A}|T}\right)$.

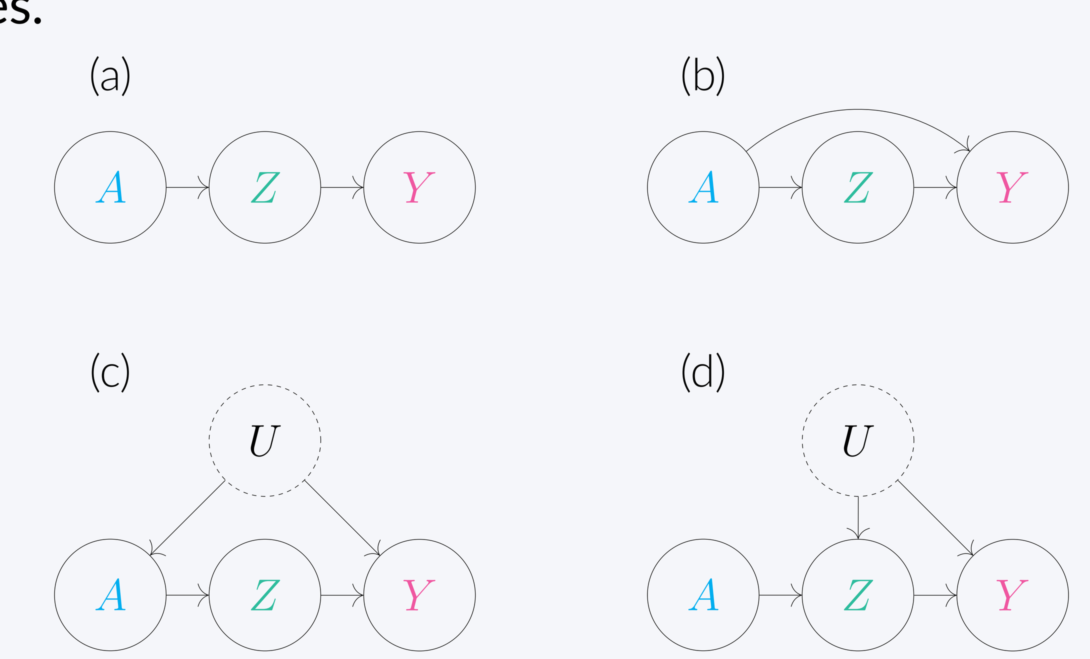
C-UCB (Lu et al. 2020): Under *causal assumptions* on ν, $R_{\nu,\text{C-UCB}}(T) = \tilde{\Theta}\left(\sqrt{|\mathcal{Z}|T}\right)$.

We prove that when causal assumptions fail, C-UCB can incur linear regret!

Conditionally Benign Property

Definition 3.1. An environment ν is conditionally benign if and only if $\nu_a(Y \mid Z)$ is constant as a function of $a \in \mathcal{A}$.

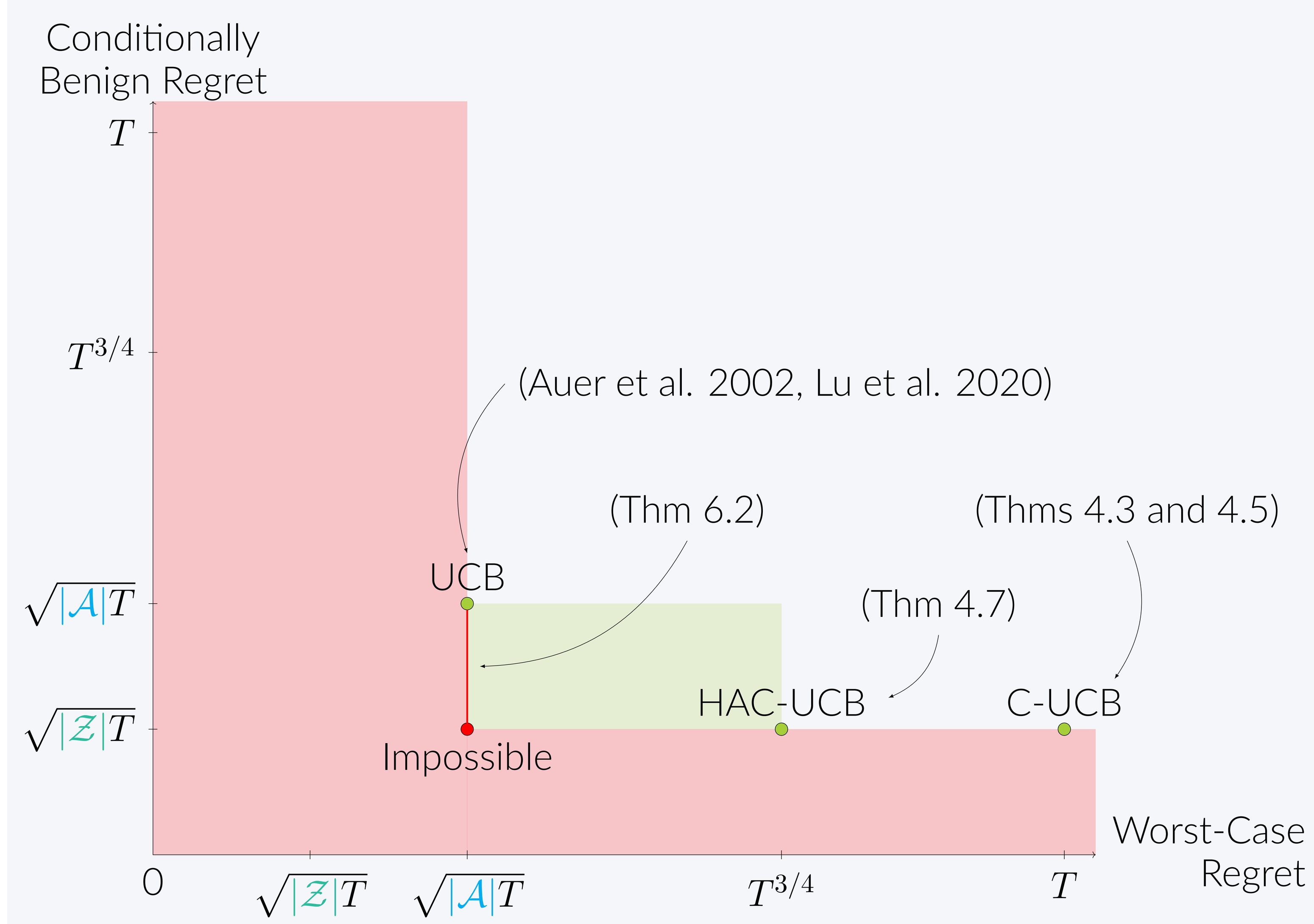
Examples.



A: intervention, Z: post-action context, Y: reward, and U: unobserved variable.
(a) the environment is conditionally benign,
(b) the environment need not be conditionally benign,
(c) the environment is conditionally benign if $\mathcal{A}$ is only *hard interventions*,
(d) the environment need not be conditionally benign.

Ask me how this generalizes d-separation and the front-door criterion!

Pareto Frontier of Causal Bandits



Ask me which open problems remain!

Novel Algorithm: HAC-UCB

Input $\tilde{\nu}$: Initial guess for $(\nu_a(Z))_{a\in\mathcal{A}}$.

Initial Exploration: Uniformly sample $a \in \mathcal{A}$ for $\sqrt{T}/|\mathcal{A}|$ rounds.
Compute MLE estimate $\hat{\nu}$ of $(\nu_a(Z))_{a\in\mathcal{A}}$. If $\sup_{a\in\mathcal{A}} \|\tilde{\nu}_a - \hat{\nu}_a\|_1 \gtrsim T^{-1/4}$, set $\tilde{\nu} \leftarrow \hat{\nu}$.

Optimistic Phase: For each round t...

$\text{UCB}_t(a) \approx \hat{\mathbb{E}}_{\nu_a}[Y] + \sqrt{(\log T)/N_a(t)}$.

$\widetilde{\text{UCB}}_t(a) \approx \sum_{z\in\mathcal{Z}} \left[\hat{\mathbb{E}}_\nu[Y \mid Z=z] + \sqrt{(\log T)/N_z(t)}\right] \tilde{\nu}_a(Z=z)$.

If $\text{UCB}_t(a) \approx \widetilde{\text{UCB}}_t(a)$, play $A_{t+1} = \arg\max_{a\in\mathcal{A}} \widetilde{\text{UCB}}_t(a)$.
Otherwise, switch to Pessimistic Phase.

Pessimistic Phase: For remaining rounds t, play $A_{t+1} = \arg\max_{a\in\mathcal{A}} \text{UCB}_t(a)$.

The key technical challenge is defining $\approx$
to balance optimism and pessimism.

If the conditionally benign assumption holds,
$\text{UCB}_t(a) \approx \widetilde{\text{UCB}}_t(a)$ and the algorithm correctly plays optimistically.
If the conditionally benign assumption fails,
either $\text{UCB}_t(a) \not\approx \widetilde{\text{UCB}}_t(a)$ and the algorithm correctly plays pessimistically,
or the regret incurred from playing optimistically is still sufficiently small.

Worst-Case Lower Bound

Theorem 4.5. For every $\mathcal{A}$ and $\mathcal{Z}$, there exists ν such that

$$\lim_{T\to\infty} \frac{R_{\nu,\text{C-UCB}}(T)}{T} \geq 1/120.$$

Adaptive Upper Bound

Theorem 4.7. For any $\mathcal{A}$, $\mathcal{Z}$, T, ν, and $\tilde{\nu}$,

$$R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}(T^{3/4}).$$

Further, if ν is conditionally benign and $\sup_{a\in\mathcal{A}} \|\tilde{\nu}_a - \nu_a\|_1 \leq \varepsilon$,

$$R_{\nu,\text{HAC-UCB}}(T) \leq \tilde{O}\left(\sqrt{|\mathcal{Z}|T} + \varepsilon T\right).$$

Ask me how this avoids making any causal assumptions!

Impossibility Result

Theorem 6.2. If π is such that $R_{\nu,\pi}(T) \leq O\left(\sqrt{|\mathcal{A}|T}\right)$ for all ν,
there exists ν that is conditionally benign but $R_{\nu,\pi}(T) \geq \Omega\left(\sqrt{|\mathcal{A}|T}\right)$.

Simulation Results