Causal contextual bandits with targeted interventions

Chandrasekar Subramanian,
Balaraman Ravindran

Indian Institute of
Technology Madras

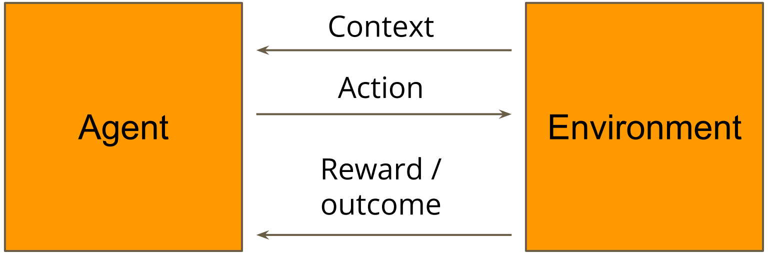
Robert Bosch Centre
for AI & Data Science

Contextual bandits

Contextual bandit agents learn **policies** (maps from a state space to an action space) …

… by repeatedly **interacting*** with an environment …

… in order to minimize some notion of **reward/regret**.



They naturally model **decision making** scenarios (e.g. recommendation systems, marketing campaign allocation, etc.)

* Though offline settings are also an active field of research.

Our framework

We propose a **new contextual bandit framework** where

1. The agent is able to target actions on specific subsets of the population ("**targeted interventions**")

2. The agent has access to **causal side-information** in the form of causal graphs

Why is this useful?

As an example, consider software product experimentation

- Product experiments can often be targeted on specific subgroups of users (e.g. iOS users) → “targeted interventions”

- There might be information* on causal relationships between variables (e.g. `emailopen` causes `click`) → “causal side-information”

We would want the learning agent to utilize these extra levers.

How do we approach this?

Targeted interventions fundamentally change the set of options that the agent has in every round – **necessitating new techniques**. Further, leveraging causal side information in contextual bandit settings hasn't been studied before.

Our contributions include

- **New formalism** for contextual bandits that captures these intricacies
- **New algorithm** based on a **novel entropy-like measure**.
- Theoretical **bound** on regret.
- **Experiments** showing that our algorithm outperforms various baselines

Framework

The agent has a budget of *T* training rounds; evaluated in (*T+1*)th round (simple regret).

In every training round, the agent can interact in one of 2 modes.

$$\boldsymbol{C} = \boldsymbol{C}^{tar} \cup \boldsymbol{C}^{other}$$

Solution approach

The agent faces a tradeoff in each round

- In the standard interaction mode, the agent can learn about the natural distribution of contexts
- In the targeted intervention mode: the agent can learn about rewards given action for a specific subspace of contexts

In the targeted intervention mode, it further faces two choices: choosing to learn about already explored contexts vs. new contexts

What mode to choose in each round? And what intervention to perform?

Unc measure

The Unc measure provides a measure of the expected effect of agent's knowledge of $\mathbb{E}[Y|do(x'), \mathbf{c}^{tar\prime}]$ IF a targeted intervention $(x, \mathbf{c}^{tar})$ is performed.

$$\text{Ent}(\mathbb{P}(V|\mathbf{pa}_V)) \triangleq -\sum_i \left[\frac{\theta_{V|\mathbf{pa}_V}[i]}{\sum_j \theta_{V|\mathbf{pa}_V}[j]} \ln \left(\frac{\theta_{V|\mathbf{pa}_V}[i]}{\sum_j \theta_{V|\mathbf{pa}_V}[j]} \right) \right]$$

$$\begin{aligned}\text{Unc}\left(\mathbb{E}[Y|do(x'), \mathbf{c}^{tar\prime}]|x, \mathbf{c}^{tar}\right) \triangleq \sum_{\mathbf{c}^{other\prime} \in \text{val}(\mathcal{C}^{other})} \Bigg[\sum_{V \in \mathcal{C}^{other}} & \text{Ent}(\mathbb{P}(V|\mathbf{c}'\langle PA_V \rangle)|x, \mathbf{c}^{tar}) + \\ & \text{Ent}(\mathbb{P}(Y|x', \mathbf{c}'\langle PA_Y \rangle)|x, \mathbf{c}^{tar}) \Bigg] \cdot \hat{\mathbb{P}}(\mathbf{c}') \cdot \hat{\mathbb{E}}[Y|\mathbf{c}', do(x')]\end{aligned}$$

Algorithm – high-level idea*

Phase 1 (αT rounds): random exploration

- Perform standard interaction with randomly chosen actions
- Update beliefs about CPDs of graph G

Phase 2 ($(1-\alpha)T$ rounds): allocation of targeted intervention samples

- Find $(x, \mathbf{c}^{tar})$ that minimizes aggregate value of **Unc** calculated using current beliefs
- Perform targeted intervention $(x, \mathbf{c}^{tar})$
- Update beliefs about CPDs of graph G

Return: final beliefs

Evaluation: given test context $\mathbf{c}^{tar}$ chosen from natural distribution, return x that has highest expectation (based on learned beliefs) of reward

* Please refer to the paper for the full algorithm

Regret bound

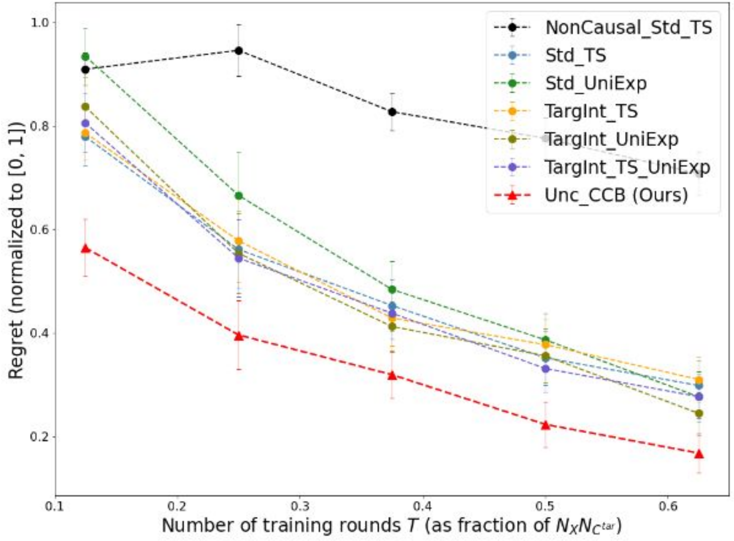
We bound regret to a provide theoretical guard on regret.

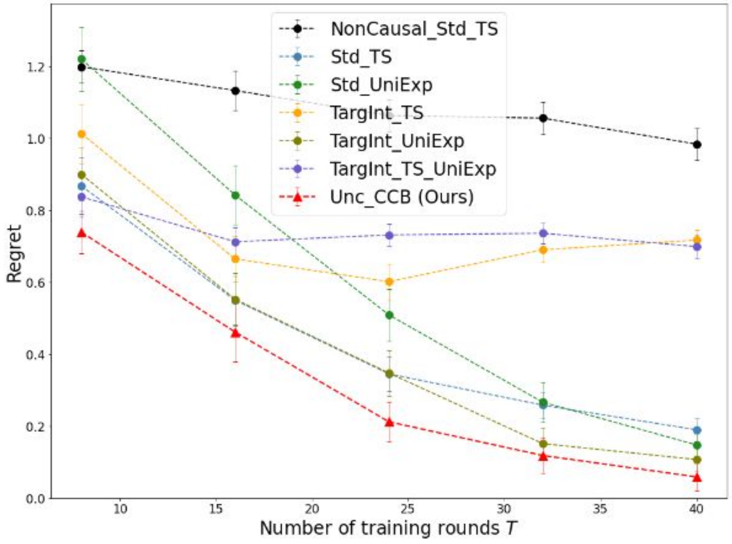
Theorem 3.1. *For any* $0 < \delta < 1$, *with probability* $\geq 1 - \delta$,

$$Regret \leq 3\mathbb{E}_{\boldsymbol{pa}_Y, \boldsymbol{c}^{tar}} \left(\sqrt{\left[\frac{2}{\frac{\alpha T}{N_X} \mathbb{P}(\boldsymbol{pa}_Y, \boldsymbol{c}^{tar}) - \epsilon^T_{X, PA_Y}} \right] \ln \left(\frac{2N_X(N_{\mathcal{C}} + |\mathcal{C}|)}{\delta} \right)} \right)$$

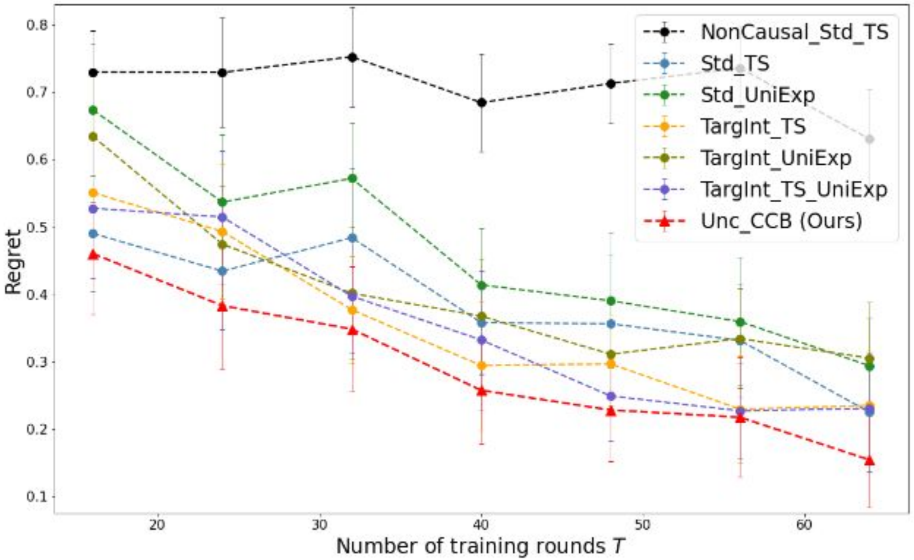
$$+ 3 \sum_{C \in \mathcal{C}^{other}} \mathbb{E}_{\boldsymbol{pa}_C, \boldsymbol{c}^{tar}} \left(\sqrt{\left[\frac{2}{\alpha T \mathbb{P}(\boldsymbol{pa}_C, \boldsymbol{c}^{tar}) - \epsilon^T_{PA_C}} \right] \ln \left(\frac{2(N_{\mathcal{C}} + |\mathcal{C}|)}{\delta} \right)} \right) \quad (1)$$

Experiments – purely synthetic





Experiments – real-world inspired



Thank you