Causal Bandits: Learning Good Interventions via Causal Inference

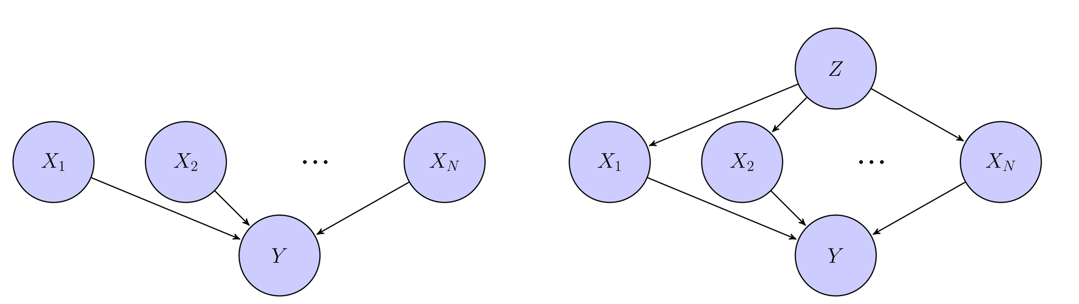
Finnian Lattimore, Tor Lattimore and Mark Reid

Department and University Name

Introduction

We study the problem of using causal models to improve the rate at which good interventions can be learned online in a stochastic environment. Our formalism combines multi-arm bandits and causal inference to model a novel type of bandit feedback that is not exploited by existing approaches. We propose a new algorithm that exploits the causal feedback and prove a bound on its simple regret that is strictly better (in all quantities) than algorithms that do not use the additional causal information.

Setup



(a) Parallel graph (b) Confounded parallel graph

Figure : Causal Models

Parallel Model

-
-
-

Algorithm 1 Parallel Bandit Algorithm

Input: Total rounds T and N.
for $t \in 1, \ldots, T/2$ **do**
 Perform empty intervention $do()$
 Observe $\vec{X}_t$ and Y_t
for $a = do(X_i = x) \in \mathcal{A}$ **do**
 Count times $X_i = x$ seen: $T_a = \sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = x\}$
 Estimate reward: $\hat{\mu}_a = \frac{1}{T_a} \sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = x\} Y_t$
 Estimate probabilities: $\hat{p}_a = \frac{2T_a}{T}$, $\hat{q}_i = \hat{p}_{do(X_i=1)}$
Compute $\hat{m} = m(\vec{q})$ and $A = \{a \in \mathcal{A} \colon \hat{p}_a \leq \frac{1}{\hat{m}}\}$.
Let $T_A := \frac{T}{2|A|}$ be times to sample each $a \in A$.
for $a = do(X_i = x) \in A$ **do**
 for $t \in 1, \ldots, T_A$ **do**
  Intervene with a and observe Y_t
 Re-estimate $\hat{\mu}_a = \frac{1}{T_A} \sum_{t=1}^{T_A} Y_t$
return estimated optimal $\hat{a}^*_T \in \arg\max_{a \in \mathcal{A}} \hat{\mu}_a$

General Graphs

Algorithm 2 General Algorithm

Input: T, $\eta \in [0,1]^{\mathcal{A}}$, $B \in [0, \infty)^{\mathcal{A}}$
for $t \in \{1, \ldots, T\}$ **do**
 Sample action a_t from η
 Do action a_t and observe X_t and Y_t
for $a \in \mathcal{A}$ **do**

$$\hat{\mu}_a = \frac{1}{T} \sum_{t=1}^{T} Y_t R_a(X_t) \mathbb{1}\{R_a(X_t) \leq B_a\}$$

return $\hat{a}^*_T = \arg\max_a \hat{\mu}_a$

$$\textbf{upper-bound,}\ R_T \in \mathcal{O}\left(\sqrt{\frac{m(\eta)}{T} \log(2NT)}\right).$$

$$\eta^* = \arg\min_{\eta} \underbrace{\max_{a \in \mathcal{A}} \mathbb{E}_a \left[\frac{\mathrm{P}\{\mathcal{P}a_Y(X)|a\}}{\sum_{b \in \mathcal{A}} \eta_b \mathrm{P}\{\mathcal{P}a_Y(X)|b\}}\right]}_{m(\eta)}.$$

Experiments

experiment1_20161020_1247.pdf

(a) Simple regret vs $m(\boldsymbol{q})$ for fixed horizon $T = 400$ and number of variables $N = 50$

experiment2_20161020_1249.pdf

(b) Simple regret vs horizon, T, with $N = 50$, $m = 2$ and $\varepsilon = \sqrt{\frac{N}{8T}}$

experiment3_20161020_1252.pdf

(c) Simple regret vs horizon, T, with $N = 50$, $m = 2$ and fixed $\varepsilon = .3$

Figure : Experimental results

Stuff

Stuff

Conclusion

Nunc tempus venenatis facilisis. **Curabitur suscipit** consequat eros non porttitor. Sed a massa dolor, id ornare enim. Fusce quis massa dictum tortor **tincidunt mattis**. Donec quam est, lobortis quis pretium at, laoreet scelerisque lacus. Nam quis odio enim, in molestie libero. Vivamus cursus mi at *nulla elementum sollicitudin*.

Additional Information

Maecenas ultricies feugiat velit non mattis. Fusce tempus arcu id ligula varius dictum.

- Curabitur pellentesque dignissim
- Eu facilisis est tempus quis
- Duis porta consequat lorem

References

Acknowledgements

Nam mollis tristique neque eu luctus. Suspendisse rutrum congue nisi sed convallis. Aenean id neque dolor. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

Contact Information

- Web: http://www.university.edu/smithlab
- Email: john@smith.com
- Phone: +1 (000) 111 1111

PLACEHOLDER LOGO PLACEHOLDER LOGO