Risk-averse Online learning: from Multi-Armed Bandits with Unobserved Confounders to Convex Games

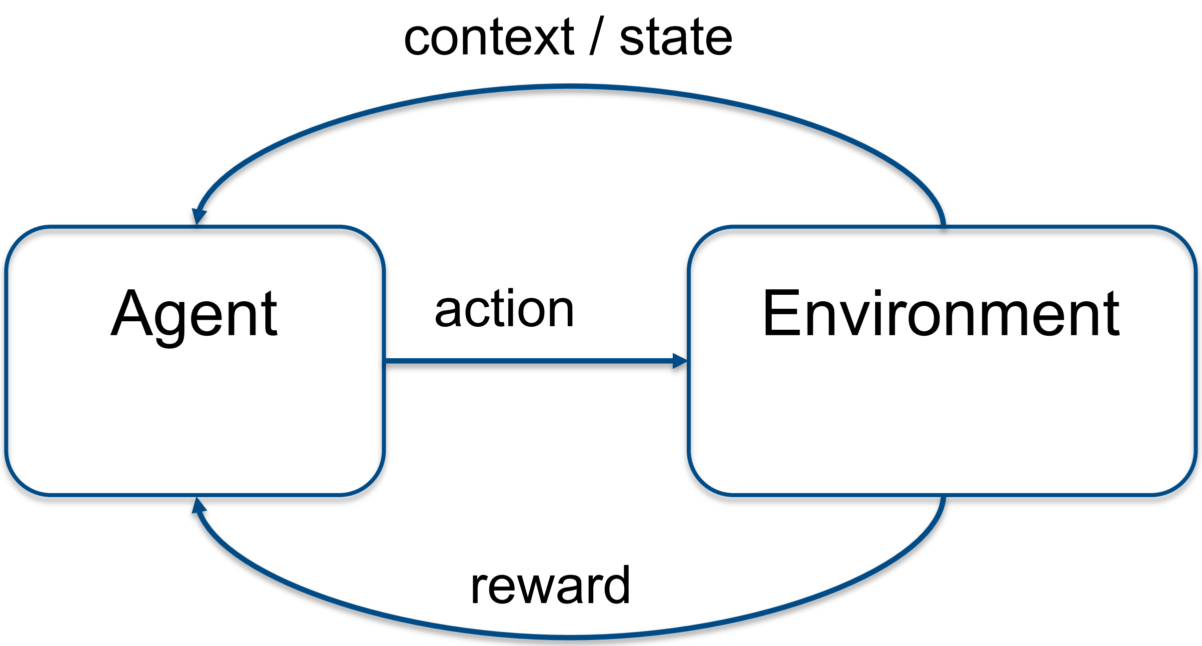
Yi Shen
Mechanical Engineering & Materials Science
Duke University

Joint work with: Jessilyn Dunn, Zifan Wang, Scott Nivison, Zachary I. Bell and Michael M. Zavlanos

Assured Autonomy in Contest Environments (AACE)
Spring 2022 Review
April 7th, 2022

Duke UNIVERSITY

Online Decision Making – Big Picture

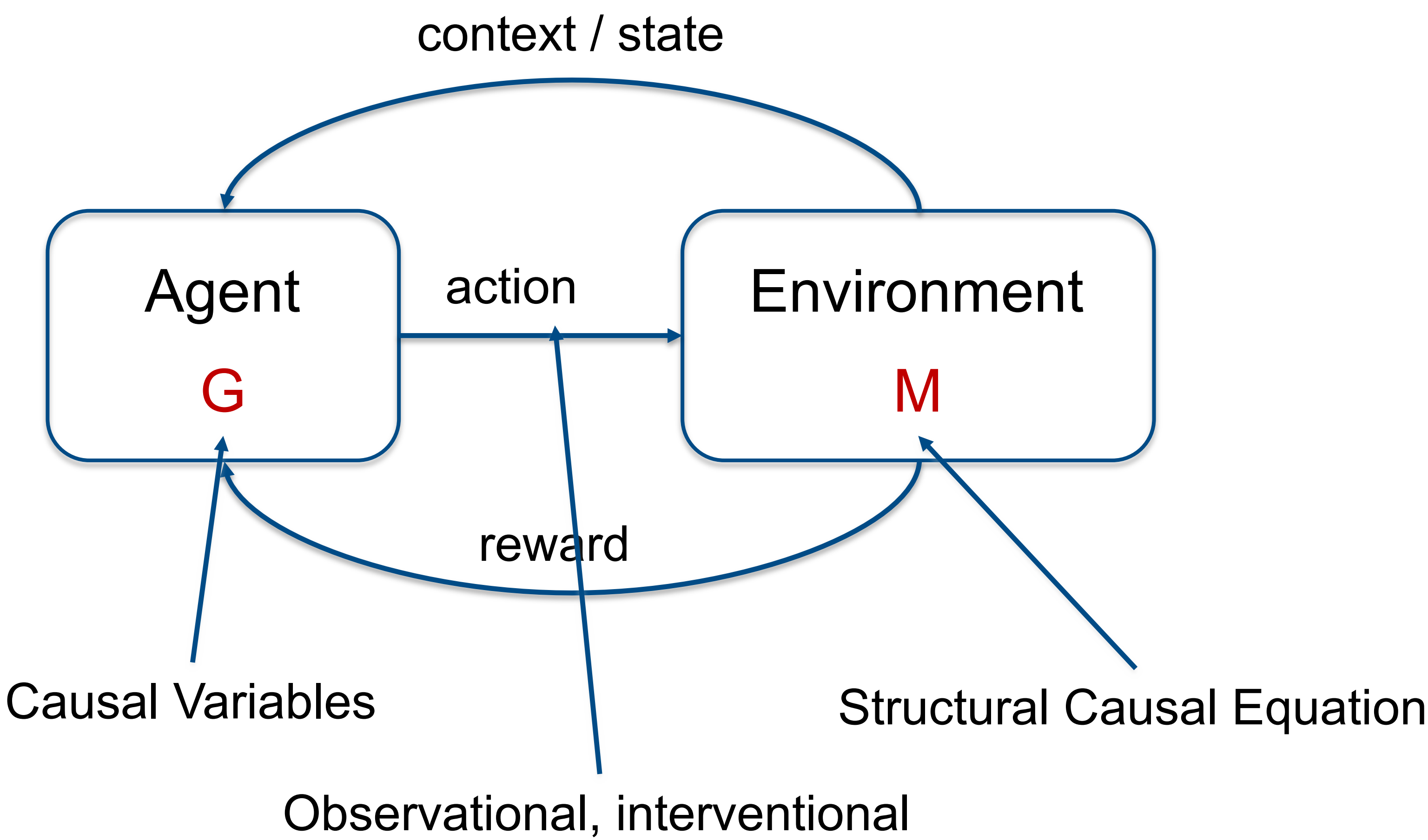


The agent learn aims to choose actions that maximize expected rewards.

Slide adapted from Dr. Bareinboim's ICML 2020 Tutorial

Duke
UNIVERSITY

Causal Online Decision Making – Big Picture



Slide adapted from Dr. Bareinboim's ICML 2020 Tutorial

Duke
UNIVERSITY

Structural Causal Models & Causal Graphs

- Processes

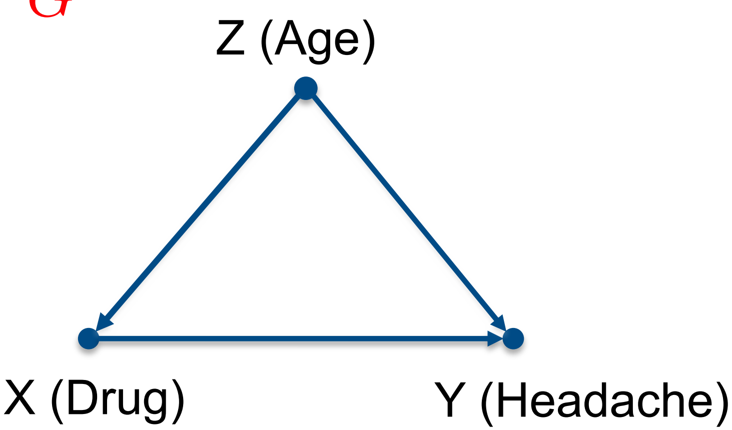
$$\text{Drug} \leftarrow f_D(\text{Age}, U_D)$$

$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

- Intervention

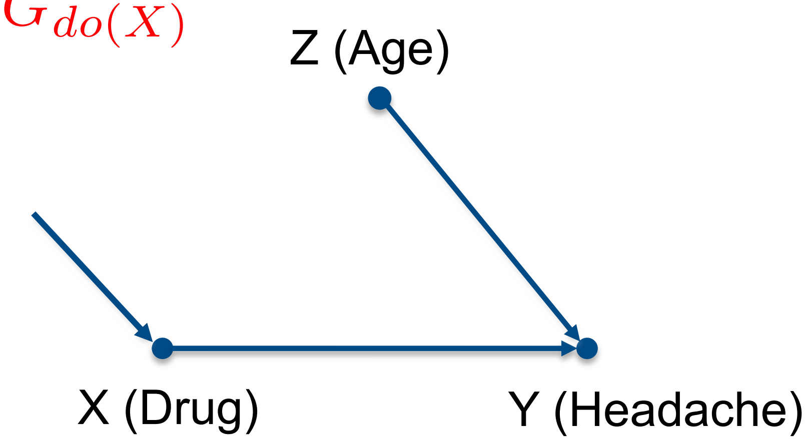
$$\textbf{Drug} \leftarrow \text{Th}(\text{Age})$$

G



P(Y,Z|X)
(observational)

$G_{do(X)}$



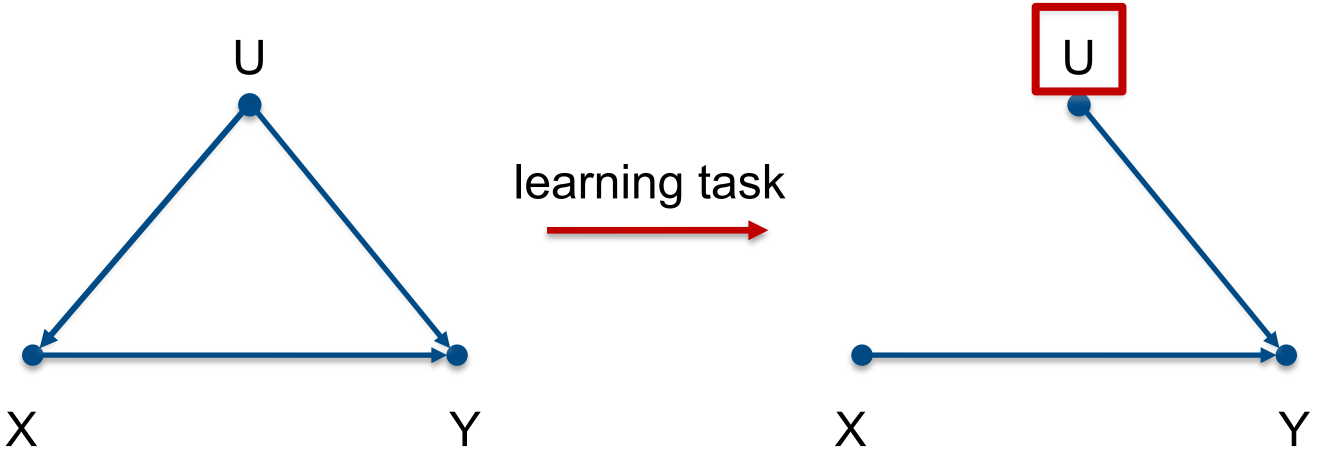
P(Y,Z|do(X=x))
(interventional)

Slide adapted from Dr. Bareinboim's ICML 2020 Tutorial

Duke UNIVERSITY

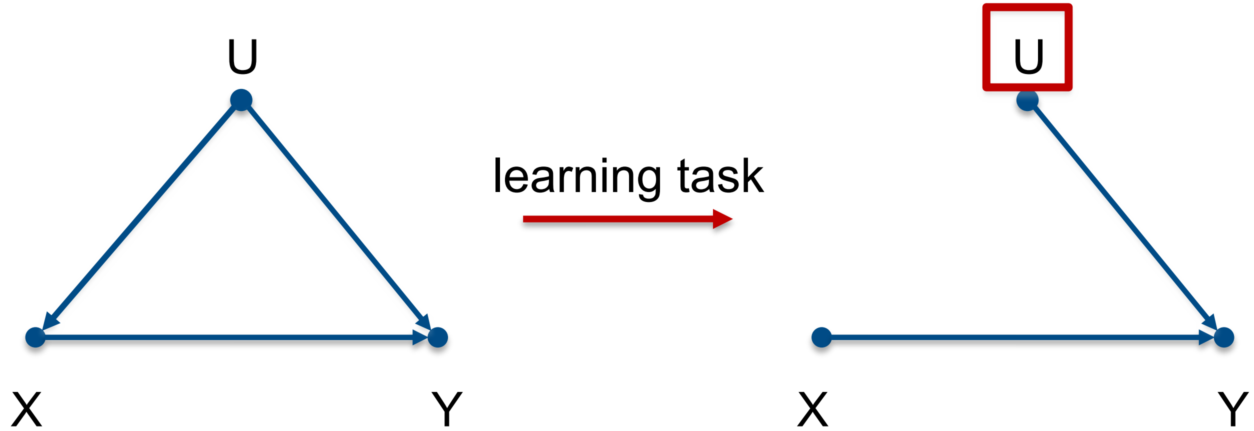
MAB with Unobserved Confounders

- Input: P(x,y), learn: P(y|do(x)).
  - Robotics: learning by demonstration when the expert can observe a richer context (e.g., more accurate sensors)
  - Mobile Health: optimal experimental design from observation data



$$P(Y|X) \neq P(Y|do(X))$$

How to estimate P(Y|do(X))?



$$P(Y|X) \neq P(Y|do(X))$$

$$P(X,Y) = \sum_U P(Y|X,U) \boxed{P(X|U)} P(U)$$

$$P(do(X),Y) = \sum_U P(Y|do(X),U) \boxed{P(do(X)|U)} P(U)$$

$$= \sum_U P(Y|X,U) \boxed{P(X)} P(U)$$

Duke UNIVERSITY

How to estimate P(Y|do(X))?

- Even though we cannot have a point estimate of P(Y|do(x)), bounds on it can be obtained by solving an optimization problem. $P(y|do(x)) = \sum_u \frac{P(x,y,u)P(u)}{P(x,u)}$

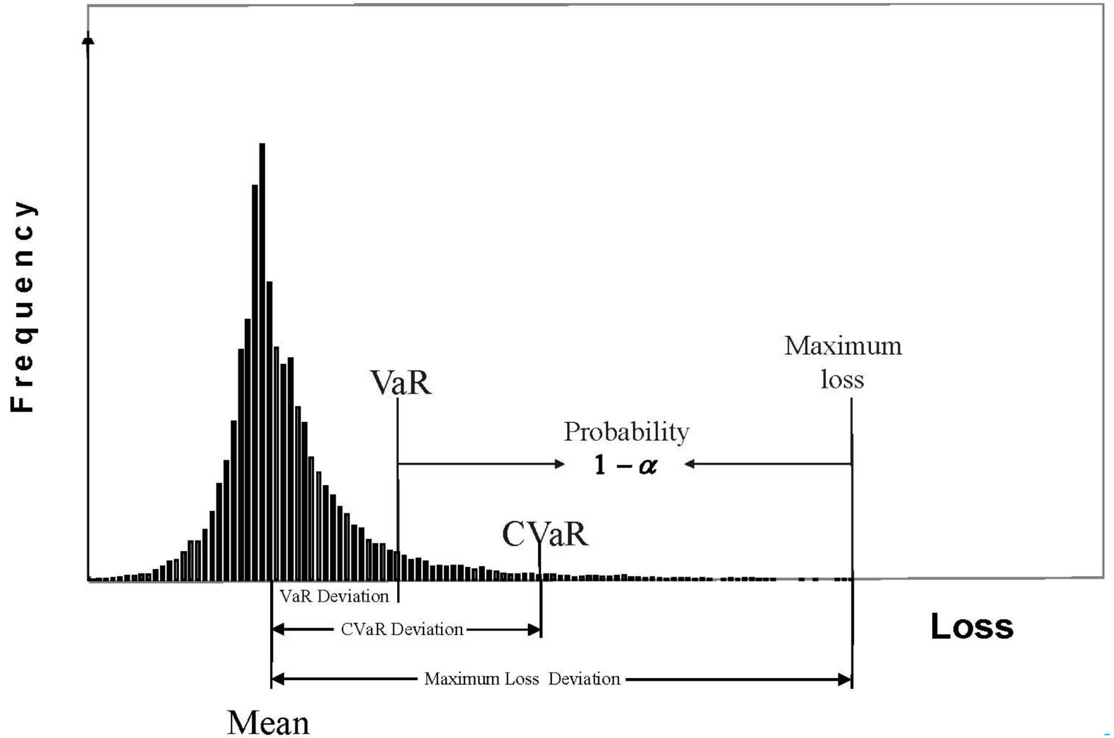
$$LB(UB) \quad P(y|do(x)) = \min_{a_u,b_u} (\max_{a_u,b_u}) \quad \sum_u \frac{a_u P(u)}{b_u}$$

$$\begin{aligned} \text{s.t.} \quad & P(u) \geq b_u,\ b_u \geq a_u, \\ & a_u \leq P(x,y),\ b_u \leq P(x), \\ & a_u \geq P(x,y) + P(u) - 1,\ b_u \geq P(x) + P(u) - 1, \\ & a_u, b_u \geq 0,\ \text{for all } u \in U; \\ & \sum_u a_u = P(x,y),\ \sum_u b_c = P(x). \end{aligned}$$

Linear Programming

Risk-averse Online Learning

- Agents aim to find a policy that maximizes the expected return while avoiding large losses.
- We consider Conditional Value at Risk (CVaR) as the risk measure for our problem.



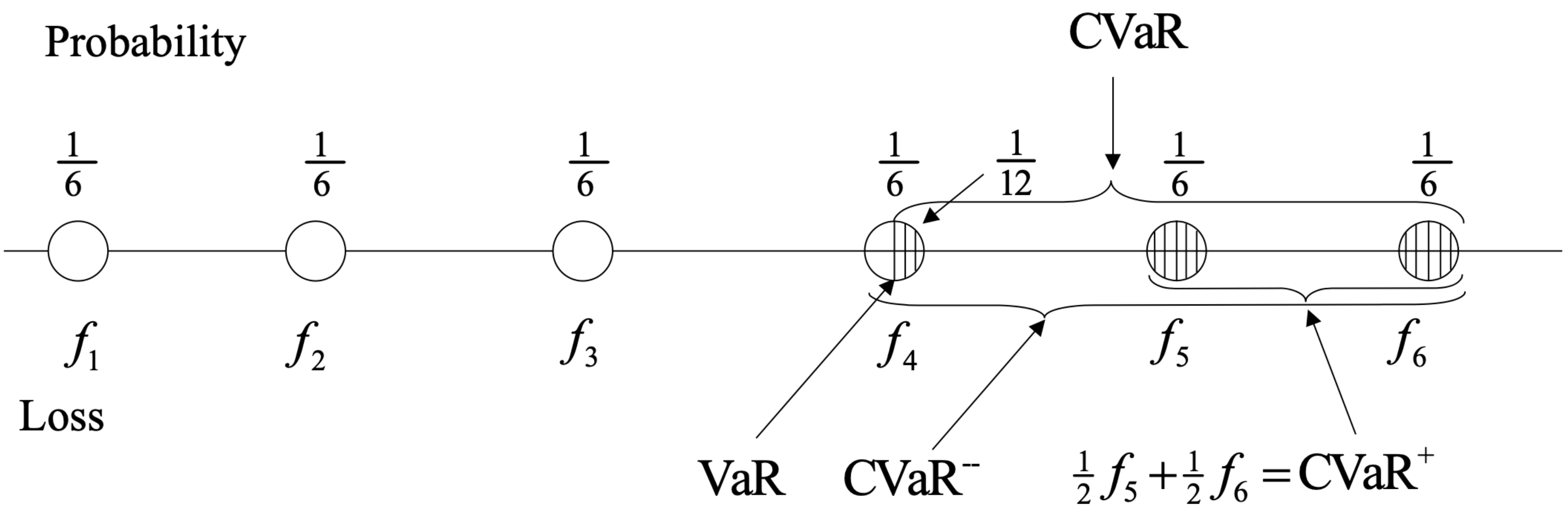
Slide adapted from Dr. Uryasev's 2000 CVaR tutorial

CVaR Calculation: Discrete Distributions

with unobserved confounders, e.g., 0.1<=p1<=0.2

Six scenarios, $p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$, $\alpha = \frac{7}{12}$

$$\text{CVaR} = \tfrac{1}{5}\text{VaR} + \tfrac{4}{5}\text{CVaR}^{+} = \tfrac{1}{5} f_4 + \tfrac{2}{5} f_5 + \tfrac{2}{5} f_6$$



Duke UNIVERSITY

CVaR with Unobserved Confounders

Mixed-integer Programming

$$
\begin{aligned}
\text{CVaR}_\alpha(Y|do(x))_{\min} = \min \quad & m - n \\
\text{s.t.} \quad & P(\bar{y}|do(x)) \leq \alpha + M(1-m), \\
& -P(\bar{y}|do(x)) \leq -\alpha + Mm, \\
& a_0 \leq P(\bar{y}|do(x)) \leq b_0,\ a_1 \leq P(y|do(x)) \leq b_1, \\
& n \leq Mm, n \leq P(\bar{y}|do(x))/\alpha, \\
& n \geq P(\bar{y}|do(x))/\alpha - M(1-m), \\
& P(y|do(x)) + P(\bar{y}|do(x)) = 1, \\
& n \geq 0, m \in \{0,1\}, \\
& \text{where } M \text{ is a constant large number.}
\end{aligned}
$$

How does causal bounds help?

- Causal bounds tell us with probability 1, P(y|do(x)) is contained in the causal bounds.
- In many online learning problems, concentrations bounds tell us with probability $1-\delta$, P(y|do(x)) is contained in the concentration bounds.
- Causal bounds can help us to better estimate all quantities built upon P(y|do(x)) in online learning, e.g., expected returns, UCB type algorithms. As a result, unsafe explorations are avoided.

Causal Bound Constrained Online Exploration

$$\tilde{F}_x(y) \leftarrow \left(\hat{F}_x(y) - \epsilon_x \mathbb{I}\{y \in [0, U)\}\right)^+$$

$$\mathrm{UCB}_x^{\mathrm{DKWClip}}(t) \leftarrow \tilde{c}_x^\alpha := \min\{\mathrm{CVaR}_\alpha(\tilde{F}_x), h_x\}$$

Risk-averse upper confidence bound   causal upper bound

Regret Analysis

Lemma 1 (Regret Decomposition). The CVaR regret satisfies the following identity

$$R_n^\alpha = \sum_{x=1}^{K} \Delta_x^\alpha \mathbb{E}[T_x(n)],$$

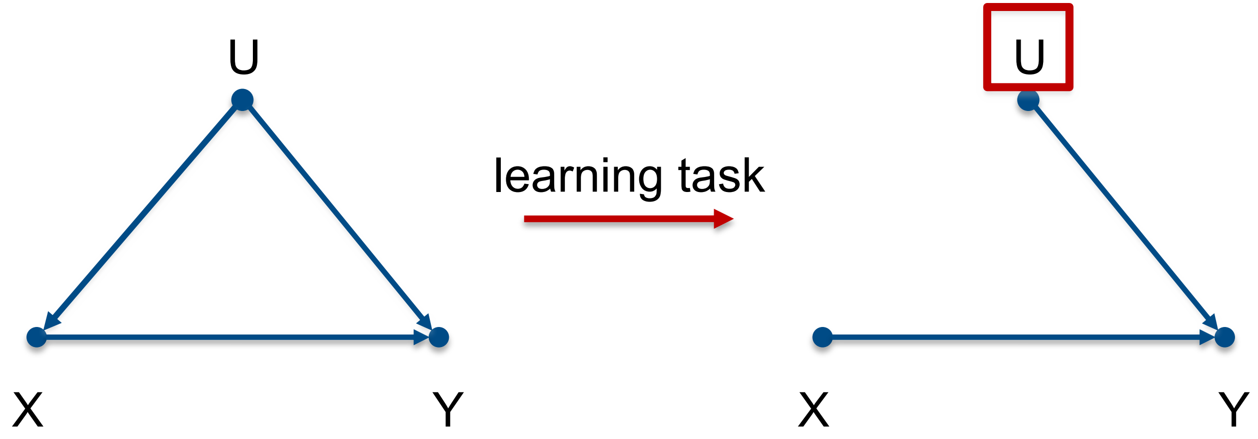
where $\Delta_x^\alpha = \max_i \mathrm{CVaR}_\alpha(F_i) - \mathrm{CVaR}_\alpha(F_x)$ is the sub-optimality gap of arm x with respect to the optimal CVaR arm and T(n) is the number of times arm x has been pulled up to time step n.

Theorem 1. Let. $\mu^* = \max_x \mathrm{CVaR}_\alpha(F_x)$. Then, the expected number of times that any sub-optimal arm x is pulled by Algorithm 1 is upper bounded by:

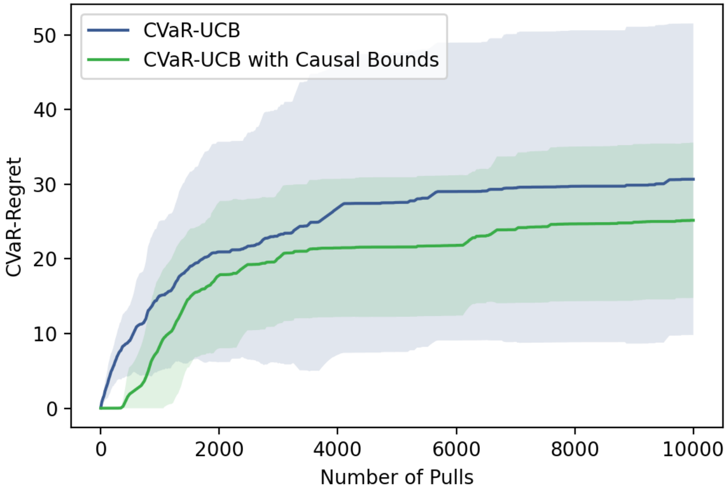
$$\mathbb{E}[T_x(n)] \leq \begin{cases} 0 & h_x < l_{\max} \\ 1 & l_{\max} \leq h_x < \mu^* \\ 3 + \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta_x^{\alpha 2}} & h_x \geq \mu^* \end{cases} .$$

Duke UNIVERSITY

A Case Study in Emotion Regulation in Mobile Health

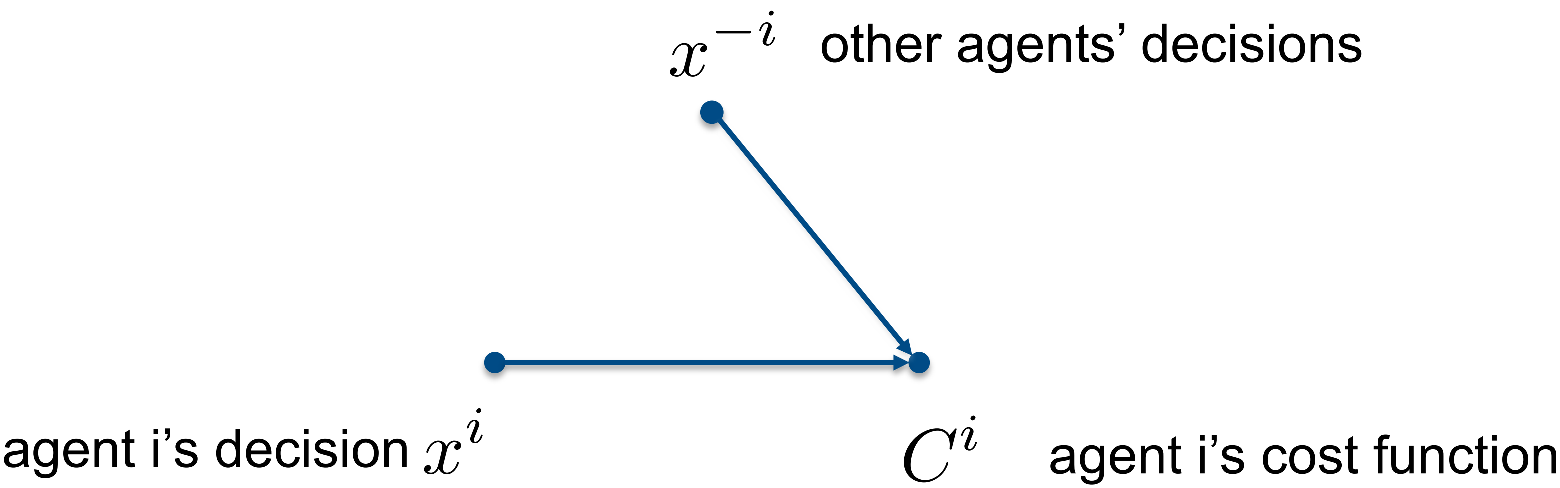


U: motion detection
X: two strategies to relieve stress and anxiety
(S1) Seeking advice/comfort from others
(S2) Accepting thoughts/feelings
Y: user's self-reporting binary evaluations on the selected recommendations



Duke UNIVERSITY

Risk-averse Convex Games



Goal: find an optimal decision that minimize the CVaR value of the cost function with bandit feedback (zeroth-order information).

Challenges in Risk-averse Convex Games

- Individual cost functions depend on joint decisions.
- CVaR values of cost functions cannot be accurately estimated due to finite samples.
- Gradients cannot be accurately measured due to bandit feedback.

Sampling strategy:

$$n_t = \lceil bU^2(T - t + 1)^a \rceil$$

Momentum Method for Risk-Averse Online Convex Games

using past samples

$$\bar{F}_{i,t}(y) = \boxed{\beta \bar{F}_{i,t-1}(y)} + (1-\beta)\hat{F}_{i,t}(y)$$

using previous gradient estimate

$$\bar{g}_{i,t} = \frac{d_i}{\delta}\left(\mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t}] - \boxed{\mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t-1}]}\right) u_{i,t}$$

Reduce Variance

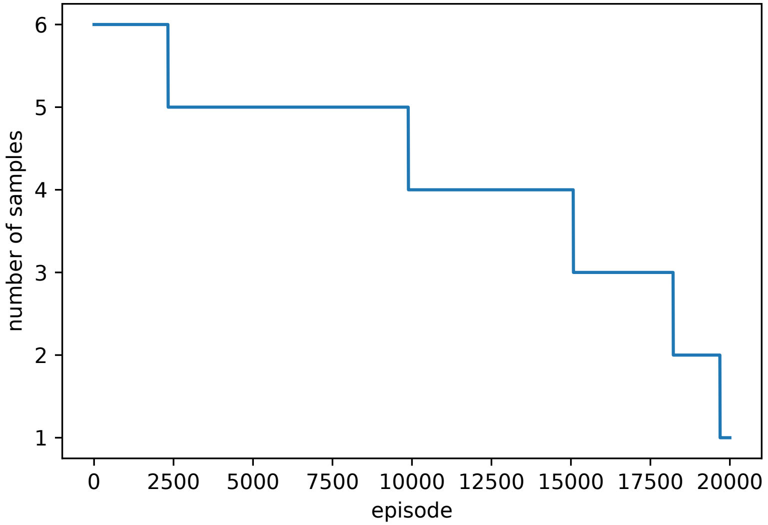
Source: Empirical distribution function, Wikipedia

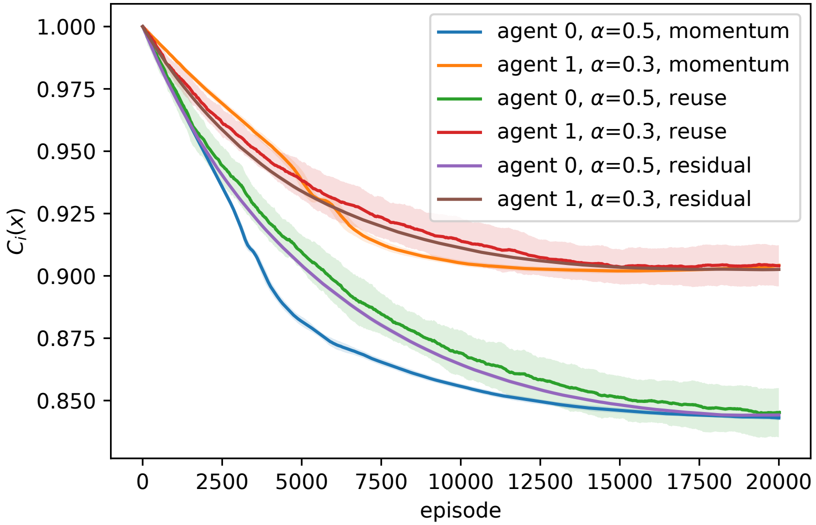
Duke UNIVERSITY

Preliminary Numerical Results

- We consider a Cournot game example.

$$J_i = 1 - (2 - \sum_j x_j)x_i + 0.2x_i + \xi_i x_i$$





Duke UNIVERSITY

Summary

- We proposed a transfer learning method for risk-averse MAB that can handle UCs. Specifically, we formulated a mixed-integer linear program (MIP) that utilizes the observational data to calculate causal bounds on CVaR values. We then transferred these CVaR causal bounds to the learner and proposed a causal bound constrained UCB algorithm to reduce the variance of online learning.

- We proposed a zeroth-order momentum method for online convex games with risk-averse agents.

Duke UNIVERSITY

Support

Thank You for Your Attention !