Risk-Averse Multi-Armed Bandits with Unobserved Confounders: A Case Study in Emotion Regulation in Mobile Health

Yi Shen, Jessilyn Dunn and Michael M. Zavlanos

***Abstract*— In this paper, we consider a risk-averse multi-armed bandit (MAB) problem where the goal is to learn a policy that minimizes the risk of low expected return, as opposed to maximizing the expected return itself, which is the objective in the usual approach to risk-neutral MAB. Specifically, we formulate this problem as a transfer learning problem between an expert and a learner agent in the presence of contexts that are only observable by the expert but not by the learner. Thus, such contexts are unobserved confounders (UCs) from the learner's perspective. Given a dataset generated by the expert that excludes the UCs, the goal for the learner is to identify the true minimum-risk arm with fewer online learning steps, while avoiding possible biased decisions due to the presence of UCs in the expert's data. To achieve this, we first formulate a mixed-integer linear program that uses the expert data to obtain causal bounds on the Conditional Value at Risk (CVaR) of the true return for all possible UCs. We then transfer these causal bounds to the learner by formulating a causal bound constrained Upper Confidence Bound (UCB) algorithm to reduce the variance of online exploration and, as a result, identify the true minimum-risk arm faster, with fewer new samples. We provide a regret analysis of our proposed method and show that it can achieve zero or constant regret. Finally, we use an emotion regulation in mobile health example to show that our proposed method outperforms risk-averse MAB methods without causal bounds.**

I. Introduction

Multi-armed bandit (MAB) problems are sequential decision making problems where an agent sequentially selects arms to pull and receives a random reward in order to learn the reward distributions of all arms and at the same time to find a strategy that maximizes the total expected reward. Many applications, ranging from treatment design [1] and news article recommendation [2] to online marketing [3], can be formulated as MAB problems. However, risk-neutral formulations that only maximize the total expected reward do not always provide desirable solutions. For example, in mobile health-based interventions for emotion regulation (ER) [4], the strategy with the highest average effectiveness rate (i.e., generating the highest possible positive emotions) is not necessarily the best; minimal adverse reactions (e.g., behaviors that cause harm to self or others) are also necessary. To avoid rare but catastrophic outcomes, appropriate risk-averse criteria can be considered during learning, e.g., Conditional Value at Risk (CVaR) [5].

*This work is supported in part by AFOSR under award #FA9550-19-1-0169 and by NSF under award CNS-1932011.

Yi Shen and Michael M. Zavlanos are with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA. Email: {yi.shen478, michael.zavlanos}@duke.edu

Jessilyn Dunn is with the Department of Biomedical Engineering, Duke University, Durham, NC, USA. Email: jessilyn.dunn@duke.edu

Oftentimes, in risk-sensitive applications, such as in the mobile health interventions for ER discussed above, the data collected by experts are available but miss important confounding variables that influence both the dependent and independent variables, causing spurious association effects. This is, e.g., the case when mobile health devices can only collect partial data due to technological limitations and/or privacy concerns. In this situation, when the expert observational data contain contexts that are unobserved confounders (UCs) to the learner, it is well known that the average treatment effects cannot be estimated without bias, regardless of the sample size [6]. Instead, what can be computed using ideas from causal inference are causal bounds on the true treatment effects that include all possible UC realizations, as shown in the seminal work by [7].

Causal bounds computed from expert observational data have been recently used to develop transfer learning methods for bandit problems [8], reinforcement learning problems [9] and imitation learning problems [10]. Specifically, [8] propose a transfer learning method that uses causal bounds computed from expert data that contain UCs to obtain a causal bound constrained Upper Confidence Bound (UCB) algorithm that the learner can use to learn an optimal policy with few new data samples. More recently, [9] extend this framework to reinforcement learning problems by computing causal bounds on value functions and using these bounds to develop a causal bound constrained Q-learning algorithm for the learner. Motivated by these approaches, [10] consider an imitation learning problem, where the expert's policies can be modeled as different arms in a MAB and the learner's goal is to learn the best arm, i.e., policy, and improve it online using only a few new data samples. Common in the above methods is that they have been designed for and are only applicable to risk-neutral problems.

Risk-sensitive bandit problems are studied in [11], [12], [13], [14]. Specifically, [11], [12] extend classic risk-neutral MAB to risk-averse MAB using a mean-variance risk measure. Since in risk-averse MAB the optimal policy is not necessarily a single-arm policy as in risk-neutral MAB, the authors bound the performance gap between the optimal policy and the optimal single arm policy and incorporate this gap into the regret analysis. On the other hand, [13] provide a systematic approach for regret analysis in MAB under different risk criteria such as value-at-risk, Conditional Value at Risk (CVaR), and Sharpe-ratio. Finally, [14] present a distributionally-aware method that adds an exploration term to the estimated cumulative distribution function (CDF) of the rewards and achieves better empirical results compared

to [13] under the CVaR risk measure. Common in all these methods is that they merely focus on online learning problems and do not rely on any existing rich observational data.

In this work, we propose a new transfer learning method for risk-averse MAB that can handle UCs in the expert data. Specifically, we consider an expert and a learner agent, both modeled as contextual MAB [15], [16], and assume that the expert is presented with additional contextual information (e.g., a person's past activities and locations) that can affect both the expert's policy, or the selection of the arms (e.g., the next ER intervention to implement), and the reward function. This contextual information is not recorded in the expert dataset that is provided to the learner (e.g., a person's mobile health device) and, therefore, it constitutes an UC for the learner. The goal of the learner is to identify the optimal arm with fewer online interactions with the environment. To do so, we first formulate a mixed-integer linear program (MIP) that utilizes the observational data to calculate causal bounds on the CVaR values of the true reward function. We then transfer these causal bounds to the learner and propose a causal bound constrained UCB algorithm to avoid risky online exploration and learn the optimal arm without accruing bias from the observational data. We provide a regret analysis that shows that it is possible to achieve zero or constant regret using causal bounds. Finally, we illustrate our proposed method on the mobile health example, which aims to optimize an intervention for ER to increase positive emotions and decrease negative emotions, while avoiding high-risk behaviors that could cause harm to self or others.

To the best of our knowledge, transfer learning methods for risk-averse MAB have not been studied in the literature. Perhaps the most closely related works to the proposed method are [8], [14]. Compared to [8], the calculation of causal bounds proposed here can handle additional assumptions on UCs and, as a result, can return tighter causal bounds. Moreover, the analysis in [8] is tailored to risk-neutral bandits and cannot be directly adapted to risk-averse problems. Compared to [14], here we use the same risk measure but also leverage rich observational data that are available in many risk-averse applications. As a result, we can obtain better online performance even in the presence of UCs.

The rest of the paper is organized as follows. In section II, we discuss the models of the expert and the learner and define the transfer learning problem. In section III, we formulate the optimization problem to compute the causal bounds on CVaR values using observational data and present the proposed causal bound constrained UCB algorithm. In Section IV, we present regret analysis results and show that causal bounds can achieve lower regret under certain conditions. In Section V, we present numerical results on a mobile health-based ER intervention application to illustrate a real-world example application, as well as the effectiveness of the proposed method. Finally, we conclude this work in Section VI.

II. Problem Definition

Consider a contextual MAB problem defined by the tuple $\big(C, X, Y^C(X)\big)$, where C is a random variable that models the context, $X \in \{1, \dots, K\}$ is a random variable that indicates the selection of one of K arms, and $Y^c(x)$ is a random reward function associated with arm $X = x$ given context $C = c$. At each time step, a sample context c is drawn independently from a distribution $P(C)$ and is announced to the agent. Then, an agent chooses one of the arms and the reward associated with this arm is revealed. Without loss of generality, we assume the rewards are non-negative and upper bounded by $U \in \mathbb{R}$. For example, in ER mobile health applications, c can represent the user's demographic information and/or previous activities and X can be the set of all possible treatments to relief anxiety. Then, the psychological clinician that is the expert agent can prescribe one of the possible treatments and observe the outcome $y^c(x)$, i.e., the effectiveness of emotion regulation after the treatment. The goal is to find a context-dependent policy that achieves the best outcome.

Given a contextual MAB problem, we can define a standard MAB problem induced by it as $\big(X, \mathbb{E}_C[Y^C(X)]\big)$, where X is a random variable indicating the selection of an arm as in contextual MAB and $\mathbb{E}_C[Y^C(X)]$ is the expected random reward function with respect to the distribution $P(C)$. We then model the learner's decision process as a standard MAB since the context is not observable to the learner. For example, in the same mobile health application discussed above, due to privacy concerns, the clinician that is the learner agent may not have access to obtain the patients' demographic information; owing to sensor limitations, cheap wearable devices cannot measure the same number of contexts as more expensive ones (the expert agent). In this case, the context C is a random variable that is unobserved by the learner, yet it affects the outcome of the possible treatments. The goal of the learner is to find a context-independent policy that achieves the best outcome. Specifically, we are interested in the case where the learner's performance is evaluated by a risk-averse measure.

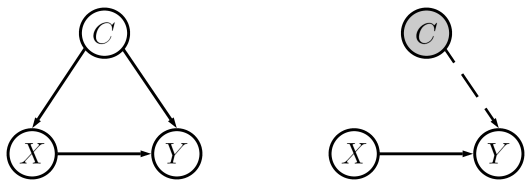
In this paper, we use the CVaR as the risk measure for the learner as in [14]. CVaR measures the expected value of a distribution's tail. Formally, let X be a bounded random variable with CDF $F_X(x) = P(X \leq x)$. The CVaR at level $\alpha \in (0, 1)$ of the random variable X is defined as [17]:

$$\text{CVaR}_\alpha(X) = \sup_\nu \{\nu - \frac{1}{\alpha}\mathbb{E}[(\nu - X)^+]\},$$

where $[x]^+ = \max\{x, 0\}$. We sometimes write $\text{CVaR}_\alpha(F_X)$ for $\text{CVaR}_\alpha(X)$, where F_X is the CDF of the random variable X. Given the conditional value at risk level $\alpha \in (0, 1]$, we define the CVaR regret associated with a MAB at time n, the same as in [14], as

$$R_n^\alpha = n \max_{x \in [K]} \left(\text{CVaR}_\alpha(F_x)\right) - \mathbb{E}\left[\sum_{t=1}^{n} \text{CVaR}_\alpha(F_{X_t})\right], \quad (1)$$

where X_t is the action taken at time step t, $[K] = \{1, \dots, K\}$ and F_x is the CDF of the distribution of rewards of the arm



(a) Expert (b) Learner

Fig. 1. Expert and learner causal graphs. (a) the causal graph of expert; (b) the causal graph of the learner. The gray nodes represent the unobserved confounder C in the learner's model, and the white nodes represent the observable data. The arrows indicate the directions of the causal effects.

x. The goal of the learner is to find the arm x that minimizes the CVaR regret.

Assume now a dataset $\tau_E = \{(x_t, y_t)\}_{t=1}^N$ generated by the expert, where x_t, y_t are the action taken and reward received at time step t. Note that the expert's actions depend on the contextual information c, but this information is not recorded in τ_E. Then, in this paper, we address to solve the following problem.

Problem 1. *Given the observational data τ_E generated by an expert (modeled by a contextual MAB), design a transfer learning algorithm for the learner (modeled by the induced MAB) that leverages the data τ_E to find an optimal arm selection strategy that minimizes the CVaR regret defined in* (1).

Since action-reward pairs are given in the observation data, one might attempt to first calculate the CVaR_α value for each arm and transfer the arm with the highest CVaR_α value to the learner. It turns out that with UCs this approach can return an arm that is sub-optimal or even the worst. See Example 1 in [10] for a case where this naive transfer does not work for risk-neural bandits, that are a special case of CVaR bandits for $\alpha = 1$.

III. Transfer Learning with Unobserved Confounders

In this section, we propose a TL framework to solve Problem 1. We first formulate an optimization problem that calculates causal bounds on CVaR. We then transfer the causal bounds to the learner and propose a causal bound constrained risk-averse MAB algorithm.

A. Causal Bound Optimization

The MAB models introduced in Section II can be seen through a causal lens as depicted in Figure 1. Specifically, the learner's model can be thought of as a special case of the expert's model equipped with the $do(\cdot)$ operation. The intervention (action) $do(X = x)$ represents a model manipulation where the value of X is set to x regardless of how it is originally determined in the model, as discussed in [6] Chapter 3. Defining the action in the expert's model by the $do(X = x)$ operator recovers the learner's model as the action x is selected regardless of contextual information. Note that the reward functions still depend on the contexts. With the above definitions, Problem 1 reduces to finding values of $P(y|do(X = x))$ for all action and reward pairs given observational data $\tau_E = \{(x_t, y_t)\}_{t=1}^N$ that do not contain contextual information. In what follows, we will denote $do(X = x)$ by $do(x)$. We assume that the contexts C, the actions X, and the rewards Y take discrete values.

It is well-known that in general $P(y|do(x)) \neq P(y|x)$ if UCs exist [6]. [18] provide a simple formula for bounding $P(y|do(x))$ in the presence of UCs, that is

$$P(x, y) \leq P(y|do(x)) < 1 - P(x, y'), \tag{2}$$

where y' represents a set of the values that are not equal to y. Note that the probabilities $P(x, y)$ for any action reward pairs can be estimated from the observation data even though $P(y|do(x))$ cannot. However, the obtained causal bounds according to (2) are usually loose since $1-P(x, y')-P(x, y)$ is often close to 1. This limitation can be overcome if additional information is used to develop these bounds that may be available in practice. For example, even though the person's exact current activity such as walking or sitting as a contextual information for ER design may be unknown, the general distribution of possible activities may be known. This information may allow to estimate the UCs' distribution $P(C)$. As suggested in [19], given the distribution of UCs $P(C)$, we can find tighter causal bounds compared to those provided by (2) by formulating an appropriate optimization problem. Note that knowing $P(C)$ does not imply knowing the relationship between the contexts, actions and rewards, i.e., knowing the distributions of the lab results of all patients does not imply a single patient's treatment outcome. Using the back-door criterion as in [6], similar to in [19], we can express $P(y|do(x))$ as

$$P(y|do(x)) = \sum_c \frac{P(x, y, c)P(c)}{P(x, c)}. \tag{3}$$

Note that (3) is not computable since $P(x, y, c)$ and $P(x, c)$ cannot be estimated from the observational data. Nevertheless, we can reformulate (3) as an optimization problem, as shown in the following result.

Theorem 1. *Consider the causal diagram G in Figure 1 (a). Given the probabilities $P(X, Y)$ and $P(C)$, the causal effects $P(Y|do(X))$ of X on Y are bounded as:*

$$LB(y, x) \leq P(y|do(x)) \leq UB(y, x)\ \forall x \in [K], y \in Y, \tag{4}$$

where $LB(y, x)$ and $UB(y, x)$ are the solutions to the optimization problem in (5).

$$\begin{aligned} LB(UB)(y, x) = \min_{a_c, b_c}(\max_{a_c, b_c}) \quad & \sum_c \frac{a_c P(c)}{b_c} \\ \textit{s.t.} \quad & P(c) \geq b_c,\ b_c \geq a_c, \\ & a_c \leq P(x, y),\ b_c \leq P(x), \\ & a_c \geq P(x, y) + P(c) - 1,\ b_c \geq P(x) + P(c) - 1, \\ & a_c, b_c \geq 0,\ \textit{for all } c \in C; \\ & \sum_c a_c = P(x, y),\ \sum_c b_c = P(x). \end{aligned} \tag{5}$$

Note that the optimization problem is well defined as long as b_c is positive for all $c \in C$. Indeed, it is a linear-fractional optimization problem and can be rewritten as a linear programming problem; see Chapter 4.3 in [20] for details. Theorem 1 is a simple modification of Theorem 4 in [19], therefore, its proof is omitted. Theorem 1 provides a way to bound the learner's reward probabilities using the expert's data without introducing any biases caused by the UCs. Recall that the learner's goal is to find the best risk-averse arm that minimizes the CVaR regret defined in (1). Using the bounds on $P(Y|do(X))$ developed in (5), we can calculate the bounds on $\text{CVaR}_\alpha(Y|do(X))$.

Theorem 2. *Assume $Y \in \{y_0, \ldots, y_n\}$ and $y_0 < \ldots < y_n$. Let a_i, b_i the bounds for $P(y_i|do(x))$ obtained by* (5) *for action x, i.e., $a_i \leq P(y_i|do(x)) \leq b_i$. Then, for a given risk level α, the causal bounds on the CVaR can be obtained by the solution of the optimization problem*

$$\begin{aligned} &\textit{CVaR}_\alpha(Y|do(x))_{\min(\max)} = \min(\max) \quad m \qquad (6)\\ &\textit{s.t.} \quad a_i \leq P(y_i|do(x)) \leq b_i \quad \forall i,\\ &\quad \sum_i P(y_i|do(x)) = 1, \textit{ where}\\ &m = \begin{cases} y_0, \textit{ if } p(y_0|do(x)) \geq \alpha;\\ \ldots\\ f(y_0, \ldots, y_k, y_{k+1}), \textit{ if } \sum_{i=0}^{k} P(y_i|do(x)) < \alpha,\\ \textit{and } \sum_{i=0}^{k+1} P(y_i|do(x)) \geq \alpha;\\ \ldots \end{cases}, \qquad (7)\\ &\textit{and } f(y_0, \ldots, y_k, y_{k+1}) = 1/\alpha \cdot \sum_{i=1}^{k} y_i P(y_i|do(x))\\ &+ 1/\alpha \cdot y_{k+1}(\alpha - \sum_{i=0}^{k} P(y_i|do(x))). \end{aligned}$$

Proof. The result follows from CVaR calculations provided in [17], [21]. □

Compared to CVaR calculations as in [17], [21], $P(y_i|do(X))$ is not a fixed number here; thus, we formulate a constrained optimization problem to calculate the CVaR values. The conditional constraints in (7) can be reformulated using binary variables to indicate whether the conditions hold or not; see e.g., [22] Chapter 9. Thus, the optimization in Theorem 2 is indeed a mixed-integer program problem. Next, we present the exact form of this problem when Y is a binary variable.

Corollary 1. *Let Y be binary and denote $Y = 1$ and $Y = 0$ by y and $\bar{y}$. With the same assumptions as in Theorem 2, the causal bounds on the CVaR can be obtained by the solution of the mixed-integer program problem*

Algorithm 1 Causal bound constrained CVaR-UCB

Input: Risk level α, reward range U, horizon n, CVaR causal lower bound l_x and upper bound h_x for all arms.
1: Remove any arm x with $h_x < l_{max}$, where $l_{max} = \max_x\{l_x\}$. Let $[K']$ be the set of remaining arms.
2: Choose each arm $x \in [K']$ once.
3: Set $\hat{F}_x$ as the empirical CDFs of each arm $x \in [K']$ on $[0, U]$,
4: Set $T_x \leftarrow 1$.
5: **for** $t = 1, \ldots, n$ **do**
6: **for** each $x \in [K']$ **do**
7: $\epsilon_x \leftarrow \sqrt{\frac{\ln(2n^2)}{2T_x}}$.
8: $\tilde{F}_x(y) \leftarrow \left(\hat{F}_x(y) - \epsilon_x \mathbb{I}\{y \in [0, U)\}\right)^+$.
9: $\text{UCB}_x^{\text{DKWClip}}(t) \leftarrow \tilde{c}_x^\alpha := \min\{\text{CVaR}_\alpha(\tilde{F}_x), h_x\}$.
10: **end for**
11: Select action $X_t = \arg\max_x \text{UCB}_x^{\text{DKWClip}}(t)$.
12: $T_{X_t} \leftarrow T_{X_t} + 1$.
13: Update the empirical CDF $\hat{F}_{X_t}$ of arm X_t.
14: **end for**

$$\begin{aligned} &\textit{CVaR}_\alpha(Y|do(x))_{\min(\max)} = \min(\max) \quad m - n \qquad (8)\\ &\textit{s.t.} \quad P(\bar{y}|do(x)) \leq \alpha + M(1-m),\\ &\quad -P(\bar{y}|do(x)) \leq -\alpha + Mm,\\ &\quad a_0 \leq P(\bar{y}|do(x)) \leq b_0,\ a_1 \leq P(y|do(x)) \leq b_1,\\ &\quad n \leq Mm, n \leq P(\bar{y}|do(x))/\alpha,\\ &\quad n \geq P(\bar{y}|do(x))/\alpha - M(1-m),\\ &\quad P(y|do(x)) + P(\bar{y}|do(x)) = 1,\\ &\quad n \geq 0, m \in \{0, 1\},\\ &\quad \textit{where } M \textit{ is a constant large number.} \end{aligned}$$

We provide the proof in the online version [23] of this paper.

B. Causal Bound Constrained MAB

Using the above causal bounds on CVaR, we propose a causal bound constrained CVaR-UCB algorithm outlined in Algorithm 1. Specifically, let l_x and h_x be the lower and upper causal bounds on $\text{CVaR}_\alpha(Y|do(x))$ for each action $x \in [K]$. Denote by $\max$ the maximum value of all lower bounds, i.e., $l_{\max} = \max_{x \in [K]} l_x$. Similar to CVaR in [14], we compute the CVaR-UCB for each arm at the beginning of each time step and select the arm with the highest CVaR-UCB. Since the causal bounds on CVaR provide upper bounds on the CVaR values, we can use these causal bounds to clip the CVaR-UCB, denoted as $\text{UCB}_x^{\text{DKWClip}}$, i.e., we take the minimum between CVaR-UCB and h_x for each arm as in step 9 in Algorithm 1. These causal constraints can reduce the variance of the CVaR-UCB estimates, thus, they can help avoid pulling sub-optimal arms that have higher CVaR-UCB values. In addition, any arms with $h_x < l_{\max}$ should not be pulled by the learner since the $l_{\max}$ arm is strictly better than them; see in step 1 in Algorithm 1. In Algorithm 1, we denote by $\hat{F}_x$ the empirical CDF estimate for arm x.

IV. Regret Analysis

In this section, we provide a regret analysis for Algorithm 1 and show that the proposed algorithm can achieve zero or

TABLE I

PROBABILITY TABLES OF GENERATING THE OBSERVATIONAL DATA BY THE EXPERT AND CVAR CAUSAL BOUNDS. (*a*) THE JOINT DISTRIBUTION OF $P(x, c)$; (*b*) THE PROBABILITIES OF BEING EFFECTIVE FOR EVERY CONTEXTS AND STRATEGIES; (*c*) CAUSAL BOUNDS ON $\text{CVaR}(Y|do(X))$ TO THE LEARNER.

(a) $P(X, C)$

	$C=1$	$C=0$
$X=1$	0.2	0.7
$X=0$	0.8	0.3

(b) $P(Y=1|X, C)$

	$C=1$	$C=0$
$X=1$	0.1	0.55
$X=0$	0.3	0.45

(c) $\text{CVaR}(Y|do(X))$

	$\alpha=0.75$	CVaR
$X=0$	[0,0.4]	0.243
$X=1$	[0.29,0.45]	0.328

constant regret under certain conditions.

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

$$R_n^\alpha = \sum_{x=1}^{K} \Delta_x^\alpha \mathbb{E}[T_x(n)], \tag{9}$$

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

Proof. The proof follows from Lemma 4.5 in [24] by replacing the mean sub-optimality gap with CVaR sub-optimality gap. □

The following result shows that the expected number of times that sub-optimal arms are selected by Algorithm 1 is no greater than that provided by the CVaR-UCB algorithm in [14].

Theorem 3. *Let* $\mu^* = \max_x \textit{CVaR}_\alpha(F_x)$*. Then, the expected number of times* $\mathbb{E}[T_x(n)]$ *that any sub-optimal arm* x *is pulled by Algorithm 1 is upper bounded by:*

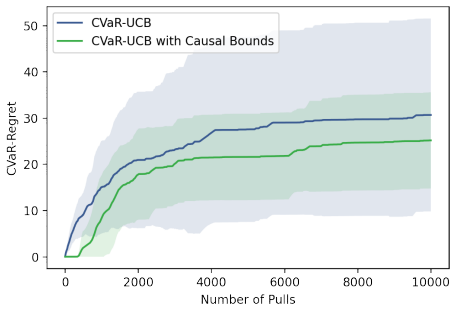
$$\mathbb{E}[T_x(n)] \leq \begin{cases} 0 & h_x < l_{\max} \\ 1 & l_{\max} \leq h_x < \mu^* \\ 3 + \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta_x^{\alpha 2}} & h_x \geq \mu^* \end{cases}.$$

The detailed proof is provided in the online version [23] of this paper.

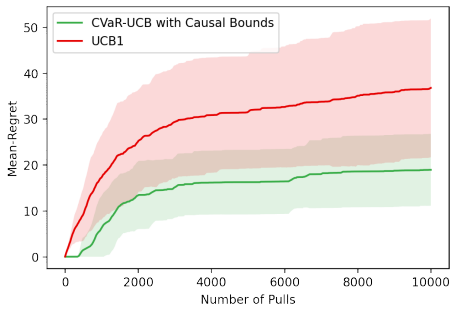
Theorem 3 provides conditions under which causal bounds can help decrease the number of pulls of sub-optimal arms. By multiplying the expected number of pulls of sub-optimal arms with their corresponding sub-optimality gaps in (9), it is straightforward to show that Algorithm 1 achieves lower regret compared to [14]. This is because if there exists an arm that satisfies condition $h_x < l_{\max}$ or $l_{\max} \leq h_x < \mu^*$, then the CVaR regret of Algorithm 1 is lower than that of [14].

V. NUMERICAL EXPERIMENTS

Consider an emotion regulation (ER) intervention design problem for people with high Social Interaction Anxiety Scale (SIAS) [26], who are experiencing moderate to severe social anxiety symptoms and are seeking for rapid and adaptive personalized ER intervention to relieve stress and anxiety. Specifically, we choose Seeking advice/comfort from others (S1) and Accepting thoughts/feelings (S2) as two strategies to help manage people's emotion as in [4]. Note that the first strategy S1 is behavioral while S2 is cognitive since it involves a change in one's thinking. [4] concludes that a user's current state of movement, (e.g., being stationary versus moving) can help to determine which ER strategies would regulate his/her emotions most effectively. However, in some mobile health devices there is no activity detection function due to limited sensors; further, people may not carry the devices all the time or intentionally disable the movement detection due to privacy concerns or battery life. As a result, a person's movement information is an unobserved confounder under these circumstances. Nevertheless, data collected from devices that can detect movement can help those devices without such function using the method as proposed Section in III. We generate a synthetic data to demonstrate this example as follows: we use $C = 1$ to indicate that the person is moving and set $P(C = 1) = 0.12$ to generate the contexts. We assume a binary variable X capturing whether S1 is recommended, i.e., $X = 1$ if S1 is selected and $X = 0$ if S2 is selected; and a binary variable Y capturing whether the person's self-reporting evaluations on the selected ER intervention suggestion is



(a) Cumulative CVaR-regret of our method (green) and CVaR-UCB [14] (blue). The solid lines and shades are averages and standard deviations over 15 runs.



(b) Cumulative regret of our method (green) and standard UCB [25] (red). The solid lines and shades are averages and standard deviations over 15 runs.

Fig. 2. Cumulative regret comparisons

effective or not.. As we assume higher reward is better, we set $Y = 1$ if the ER strategy is effective. As indicated by [4], Seeking advice/comfort from others is more effective for people that are stationary than moving. Thus, when $C = 0$, the strategy S1 is selected more often in the expert policy. The overall context-dependent policy is summarized in Table I(a). The outcomes of the recommendation being effective ($Y = 1$) are generated according to Table I(b). The observational data containing recommendations (the mobile health app suggestion, S1 or S2) and outcomes (user report of effectiveness) but excluding contextual information (movement status) is then transferred to the learner (the mobile health recommendation system). We first apply Theorem 1 to calculate causal bounds on $P(Y|do(X))$. Then, using Theorem 2, we can obtain the CVaR causal bounds for a given level of risk α; Table I (c) shows the CVaR causal bounds for $\alpha = 0.75$ and the true CVaR value for $\alpha = 0.75$. We use Gurobi [27] to solve all the linear and mixed-integer programming problems. We select $\alpha = 0.75$ for our numerical experiments. Specifically, we compare our causal bound constrained CVaR-UCB with CVaR-UCB [14] using CVaR-regret as a performance measure.

The results in Figure 2(a) show that the CVaR-regret of our method is lower than the one without causal bounds, e.g., mobile health users wearing the devices without movement detection benefits from the users with advanced devices by avoiding recommendations with high risk. In addition, causal bounds help to reduce the variance. We further compare our method with the standard UCB algorithm [25] using mean regret as a performance measure to determine whether our proposed risk-averse method can outperform risk-neutral methods using risk-neutral criterion. We observe that, in the two-arm case, our method generates a lower regret and variance compared to the UCB algorithm, as shown in Figure 2(b). This is because the sub-optimality gap in (9) for the CVaR criterion is larger than the gap for the mean criterion. The larger sub-optimality gap for the CVaR criterion makes the best arm identification problem easier.

VI. Conclusion

In this work, 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 provided a regret analysis and showed that our method can achieve zero or constant regret using causal bounds under certain conditions. To illustrate our proposed method, we simulated a mobile health emotion regulation recommender system and demonstrated that interventions can be chosen more appropriately and with lower risk using our method.

References

[1] S. A. Murphy, "Optimal dynamic treatment regimes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 2, pp. 331–355, 2003.

[2] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 661–670.

[3] K. Misra, E. M. Schwartz, and J. Abernethy, "Dynamic online pricing with incomplete information using multiarmed bandit experiments," *Marketing Science*, vol. 38, no. 2, pp. 226–252, 2019.

[4] M. K. Ameko, M. L. Beltzer, L. Cai, M. Boukhechba, B. A. Teachman, and L. E. Barnes, "Offline contextual multi-armed bandits for mobile health interventions: A case study on emotion regulation," in *Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 249–258.

[5] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical finance*, vol. 9, no. 3, pp. 203–228, 1999.

[6] J. Pearl, *Causality*. Cambridge university press, 2009.

[7] A. Balke and J. Pearl, "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, vol. 92, no. 439, pp. 1171–1176, 1997.

[8] J. Zhang and E. Bareinboim, "Transfer learning in multi-armed bandit: a causal approach," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017, pp. 1778–1780.

[9] Y. Zhang and M. M. Zavlanos, "Transfer reinforcement learning under unobserved contextual information," in *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 2020, pp. 75–86.

[10] C. Liu, Y. Zhang, Y. Shen, and M. M. Zavlanos, "Learning without knowing: Unobserved context in continuous transfer reinforcement learning," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 791–802.

[11] A. Sani, A. Lazaric, and R. Munos, "Risk-aversion in multi-armed bandits," *arXiv preprint arXiv:1301.1936*, 2013.

[12] S. Vakili and Q. Zhao, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1093–1111, 2016.

[13] A. Cassel, S. Mannor, and A. Zeevi, "A general approach to multi-armed bandits under risk criteria," in *Conference On Learning Theory*. PMLR, 2018, pp. 1295–1306.

[14] A. Tamkin, R. Keramati, C. Dann, and E. Brunskill, "Distributionally-aware exploration for cvar bandits," in *NeurIPS 2019 Workshop on Safety and Robustness on Decision Making*, 2019.

[15] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," *Advances in neural information processing systems*, vol. 20, no. 1, pp. 96–1, 2007.

[16] A. Slivkins, "Contextual bandits with similarity information," in *Proceedings of the 24th annual Conference On Learning Theory*. JMLR Workshop and Conference Proceedings, 2011, pp. 679–702.

[17] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[18] J. Tian and J. Pearl, "Probabilities of causation: Bounds and identification," *Annals of Mathematics and Artificial Intelligence*, vol. 28, no. 1, pp. 287–313, 2000.

[19] A. Li and J. Pearl, "Bounds on causal effects and application to high dimensional data," *arXiv preprint arXiv:2106.12121*, 2021.

[20] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[21] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of banking & finance*, vol. 26, no. 7, pp. 1443–1471, 2002.

[22] S. P. Bradley, A. C. Hax, and T. L. Magnanti, *Applied mathematical programming*. Addison-Wesley, 1977.

[23] Y. Shen, J. Dunn, and M. M. Zavlanos, "Risk-averse multi-armed bandits with unobserved confounders: A case study in emotion regulation in mobile health," *arXiv:2209.04356*, 2022.

[24] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[25] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[26] R. P. Mattick and J. C. Clarke, "Development and validation of measures of social phobia scrutiny fear and social interaction anxiety," *Behaviour research and therapy*, vol. 36, no. 4, pp. 455–470, 1998.

[27] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2021. [Online]. Available: https://www.gurobi.com