A Omitted results

A.1 Tables of the example in Section 2

(U, W)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\mathbb{E}[Y \| do(A=0)]$	0	0	1	1
$\mathbb{E}[Y \| do(A=1)]$	10	10	0.9	0.9

Table 3: Reward table

(A, W)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\mathbb{P}(A, W)$	0.05	0.05	0.45	0.45

Table 4: Observation probability for agents

A.2 Incorporating estimation error

Since $\theta_{ijk} \in [0,1]$, the variance of one sample is at most $\frac{1}{4}$. If θ_{ijk} has n i.i.d. samples, then its largest variance of the known $\hat{\theta}_{ijk}$ is $\frac{1}{4n}$. We can simply set $\epsilon = \frac{1}{2\sqrt{n}}$ if n is given. For example, if one want to reflect the concentration property, then one can choose the truncate Gaussian distribution $\mathcal{N}(\hat{\theta}_x, \frac{1}{4n_x})$ for $x = ijk$ or $x = l$. If one expects the worst cases, then one can choose the uniform distribution for θ_x with discrepancy parameter ϵ.

A.3 Referred algorithms

Algorithm 4 MC for causal bound

```
Input: cumulative distribution functions $\hat{F}(a, y, w)$ and $\hat{F}(u)$, discrepancy parameter $\epsilon$, sampling distribution $F_s$, batch size $B$
Discrete the variable domain $\mathcal{A}, \mathcal{Y}, \mathcal{W}, \mathcal{U}$
Select a linearly independent variable index set $S$ with size $n_{\mathcal{A}}n_{\mathcal{Y}}n_{\mathcal{W}}n_{\mathcal{U}} - n_{\mathcal{A}}n_{\mathcal{Y}}n_{\mathcal{W}} - n_{\mathcal{U}} + 1$ for linear equations (2)
Compute each $\hat{\theta}_{ijk}$ and $\hat{\theta}_l$ according to (5)
for $n = 1, 2 \cdots, B$ do
    Sample $\theta_x$ from the uniform distribution supported on $[\max\{\hat{\theta}_x - \epsilon, 0\}, \max\{\hat{\theta}_x + \epsilon, 1\}]$ for all $x = ijk$ or $x = l$
    Sequentially solve LP (4) to find support $[l_{ijkl}, h_{ijkl}]$ for each $x_{ijkl}$ with $(i, j, k, l) \in S$ and sample a value from $F_s$ truncated to $[l_{ijkl}, h_{ijkl}]$
    Solving remaining $x_{ijkl}$ by linear equations (2) for all $(i, j, k, l) \notin S$
    Compute the causal effect $b_n(a)$ by $\hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a)]$ for each $a$
For each $a$, sort $B$ valid causal bound and get $b_{(1)}(a), b_{(2)}(a), \cdots, b_{(B)}(a)$
Output: $l(a) = b_{(1)}(a)$ and $h(a) = b_{(B)}(a)$ for $a \in \mathcal{A}$
```

A.4 Referred theorems

Theorem A.1 *Given a causal diagram $\mathcal{G}$ and a distribution compatible with $\mathcal{G}$, let $\{W, U\}$ be a set of variables satisfying the back-door criterion in $\mathcal{G}$ relative to an ordered pair (A, Y), where $\{W, U\}$ is partially observable, i.e., only probabilities $\hat{F}(a, y, w)$ and $\hat{F}(u)$ with the maximum estimation error ϵ, the causal effects of A on Y are then bounded as follows:*

$$l(a_0) \leq \mathbb{E}[Y|do(a_0)] \leq h(a_0),$$

Algorithm 5 Transfer learning in multi-armed bandit

```
Input: time horizon $T$, causal bound $l(a)$ and $h(a)$
Remove the arm $a$ for $h(a) < \max_{i \in \mathcal{A}} l(i)$ and denote the remaining arm set as $\mathcal{A}^*$
Initialize reward vector $\hat{\mu}_a(1)$ and the number of pull $n_a(1)$ to zero, for all $a \in \mathcal{A}^*$
for round $t = 1, 2, \cdots, T$ do
    Compute the upper confidence bound $U_a(t) = \min\{1, \hat{\mu}_a(t) + \sqrt{\frac{2 \log T}{n_a(t)}}\}$
    Truncate $U_a(t)$ to $\hat{U}_a(t) = \min\{U_a(t), h(a)\}$ for all $a \in \mathcal{A}^*$
    Choose the action $a_t = \arg\max_{a \in \mathcal{A}^*} \hat{U}_a(t)$ and observe a reward $y_t$
    Update the empirical mean $\hat{\mu}_{a_t}(t+1) = \frac{\hat{\mu}_{a_t} n_{a_t}(t) + y_t}{n_{a_t}(t) + 1}$ and the number of pulling $n_{a_t}(t+1) = n_{a_t}(t) + 1$
    For $a \neq a_t$, update $\hat{\mu}_a(t+1) = \hat{\mu}_a(t)$ and $n_a(t+1) = n_a(t)$
```

Algorithm 6 MC for causal bound with w

```
Input: cumulative distribution functions $\hat{F}(a, y, w)$ and $\hat{F}(u)$, discrepancy parameter $\epsilon$, sampling distribution $F_s$, batch size $B$
Discrete the variable domain $\mathcal{A}, \mathcal{Y}, \mathcal{W}, \mathcal{U}$
Select a linearly independent variable index set $S$ with size $n_{\mathcal{A}} n_{\mathcal{Y}} n_{\mathcal{W}} n_{\mathcal{U}} - n_{\mathcal{A}} n_{\mathcal{Y}} n_{\mathcal{W}} - n_{\mathcal{U}}$ for linear equations (2)
Compute each $\hat{\theta}_{ijk}$ and $\hat{\theta}_l$ according to (5)
for $n = 1, 2, \cdots, B$ do
    Sample $\theta_x$ from the uniform distribution supported on $[\max\{\hat{\theta}_x - \epsilon, 0\}, \max\{\hat{\theta}_x + \epsilon, 1\}]$ for all $x = ijk$ or $x = l$
    Sequentially solve LP (4) to find support $[l_{ijkl}, h_{ijkl}]$ for each $x_{ijkl}$ with $(i, j, k, l) \in S$ and sample a value from $F_s$ supported on $[l_{ijkl}, h_{ijkl}]$
    Solving remaining $x_{ijkl}$ by linear equations (2) for all $(i, j, k, l) \notin S$
    Compute the causal effect $b_{cnt}(w, a)$ by $\hat{\mathbb{E}}_{\mathcal{M}}[Y | do(a), w]$ for each $a$ and $w$
For each $w, a$, sort $B$ valid causal bound and get $b_{(1)}(w, a), b_{(2)}(w, a), \cdots, b_{(B)}(w, a)$
Output: $l(w, a) = b_{(1)}(w, a)$ and $h(w, a) = b_{(B)}(w, a)$ for $(w, a) \in \mathcal{W} \times \mathcal{A}$
```

where $l(a_0)$ and $h(a_0)$ are solutions to the following functional optimization problem for any given a_0

$$
\begin{aligned}
l(a_0) &= \inf \int_{w \in \mathcal{W}, u \in \mathcal{U}} \int_{y \in \mathcal{Y}} y dF(y | a_0, w, u) dF(w, u) \\
h(a_0) &= \sup \int_{w \in \mathcal{W}, u \in \mathcal{U}} \int_{y \in \mathcal{Y}} y dF(y | a_0, w, u) dF(w, u) \\
s.t. & \int_{u \in \mathcal{U}} dF(a, y, w, u) = F(a, y, w), \forall (a, y, w) \in \mathcal{A} \times \mathcal{W} \times \mathcal{U} \\
& \int_{a \in \mathcal{A}, y \in \mathcal{Y}, w \in \mathcal{W}} dF(a, y, w, u) = F(u), \forall u \in \mathcal{U} \\
& \int_{y \in \mathcal{Y}} dF(a, y, w, u) = F(a, w, u), \forall (a, w, u) \in \mathcal{A} \times \mathcal{W} \times \mathcal{U} \\
& \int_{a \in \mathcal{A}} dF(a, w, u) du = F(w, u), \forall (w, u) \in \mathcal{W} \times \mathcal{U} \\
& F(y | a, w, u) F(a, w, u) = F(a, y, w, u), \forall (a, y, w, u) \in \mathcal{A} \times \mathcal{Y} \times \mathcal{W} \times \mathcal{U} \\
& |F(a, y, w) - \hat{F}(a, y, w)| \leq \epsilon, \forall (a, y, w) \in \mathcal{A} \times \mathcal{Y} \times \mathcal{W} \\
& |F(u) - \hat{F}(u)| \leq \epsilon, \forall u \in \mathcal{U}.
\end{aligned}
$$

Here the inf/sup is taken with respect to all unknown cumulative distribution functions $F(a, y, w, u)$, $F(a, w, u)$, $F(y | a, w, u)$, $F(w, u)$, $F(a, y, w)$ and $F(u)$.

Theorem A.2 *Consider a $|\mathcal{A}|$-MAB problem with rewards bounded in $[0, 1]$. For each arm $a \in \mathcal{A}$, its expected reward μ_a is bounded by $[l(a), h(a)]$. Then in the Algorithm 5, the number of draws $\mathbb{E}[N_a(T)]$ for any sub-optimal arm is upper bounded as:*

$$\mathbb{E}[N_a(T)] \leq \begin{cases} 0, h(a) < \max\limits_{i \in \mathcal{A}} l(i) \\ \dfrac{\pi^2}{6}, \max\limits_{i \in \mathcal{A}} l(i) \leq h(a) < \mu^* \\ \dfrac{\log T}{\Delta_a^2}, h(a) \geq \mu^*. \end{cases}$$

Theorem A.3 *Consider a contextual bandit problem with $|\mathcal{A}| < \infty$ and $|\mathcal{W}| < \infty$. Denote*

$$\widetilde{\mathcal{A}^*}(x) = \mathcal{A} - \{a \in \mathcal{A} | h(x, a) < \mu_w^*\}.$$

Then the regret of Algorithm 2 satisfies

$$\limsup_{T \to \infty} \frac{\mathbb{E}[Reg(T)]}{\sqrt{T \log T}} \leq \sum_{w \in \mathcal{W}} \sqrt{8(|\widetilde{\mathcal{A}^*}(w)| - 1)\mathbb{P}(W = w)}.$$

A.5 Implementation details of Algorithm 3

In Algorithm 3, one needs to compute $\mathcal{F}^*$ and $\mathcal{A}^*(w)$. A naive way costs $\mathcal{O}(|\mathcal{F}|)$ time complexity, which becomes inefficient for large $|\mathcal{F}|$ and infeasible for infinite $|\mathcal{F}|$. Actually, we can implicitly compute $\mathcal{F}^*$ by clipping, i.e., using $\min\{\max\{\hat{f}_m(w, a), l(w, a)\}, h(w, a)\}$ as the estimator at the epoch m. As $\hat{f}_m$ gets closer to the true reward function f^*, which is within the causal bounds, the causal bounds gradually lose their constraint effect. For computing $\mathcal{A}^*(w)$, we refer readers to the section 4 of (Foster et al., 2020), where a systematic method for computing $\mathcal{A}^*(w)$ within a given accuracy is provided.

Another option to implement Algorithm 3 is to compute $\mathbb{E}_W[|\mathcal{A}^*(W)|]$ using expert knowledge $F(a, y, w)$. We can set $\gamma_t = \sqrt{\frac{\eta \mathbb{E}_W[|\mathcal{A}^*(W)|]\tau_{m-1}}{\log(2\delta^{-1}|\mathcal{F}^*| \log T)}}$, so that γ_t remains constant within an epoch. Our proof still holds for this option, and the regret order is the same as in Theorem 3.3. Intuitively, $|\mathcal{A}(w_t)|$ is a sample from an induced distribution with a mean of $\mathbb{E}_W[|\mathcal{A}^*(W)|]$, so on average, the regrets of both options are of the same order.

It is worth noting that Algorithm 3 and Theorem 3.3 can be easily extended to handle infinite $\mathcal{F}$ using standard learning-theoretic tools such as metric entropy. Suppose $\mathcal{F}$ is equipped with a maximum norm $\|\cdot\|$. We can consider an ϵ-covering $\mathcal{F}_\epsilon^*$ of $\mathcal{F}^*$ under maximum norm. Since $|\mathcal{F}_\epsilon^*|$ is finite, we can directly replace $\mathcal{F}^*$ with $\mathcal{F}_\epsilon^*$ and do not change any algorithmic procedure. Thanks to the property of ϵ-covering, there exists a function $f_\epsilon^* \in \mathcal{F}_\epsilon^*$ such that $\|f_\epsilon^* - f^*\| \leq \epsilon$. Hence, the regret can be bounded by

$$Reg(T) \leq 8\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]T \log(2\delta^{-1}|\mathcal{F}_\epsilon^*| \log T)} + \epsilon T.$$

By replacing the dependence on $\log|\mathcal{F}^*|$ in the algorithm's parameters with $\log|\mathcal{F}_\epsilon^*|$ and setting $\epsilon = \frac{1}{T}$, we obtain a similar result as Theorem 3.3 up to an additive constant of 1.

Definition A.1 ((Fan, 1953)) *Let $(\mathcal{F}, \|\cdot\|)$ be a normed space. The set $\{f_1, \cdots, f_N\}$ is an ϵ-covering of $\mathcal{F}$ if $\forall f \in \mathcal{F}$, there exists $i \in [N]$ such that $\|f - f_i\| \leq \epsilon$. The covering number $N(\mathcal{F}, \|\cdot\|, \epsilon)$ is defined as the minimum cardinality N of the covering set over all ϵ-coverings of $\mathcal{F}$.*

It clear that $Reg(T)$ scales with $\sqrt{\log N(\mathcal{F}^*, \|\cdot\|, \epsilon)}$. Note that $N(\mathcal{F}^*, \|\cdot\|, \epsilon) \leq N(\mathcal{F}, \|\cdot\|, \epsilon)$ as $\mathcal{F}^* \subset \mathcal{F}$. The covering number shows clearly how extra causal bounds help improve the algorithm performance by shrinking the search space. Let $m = \inf_{w,a} l(w, a)$ and $M = \sup_{w,a} h(w, a)$. These bounds chip away the surface of the unit sphere and scoop out the concentric sphere of radius m. Therefore, the transfer learning algorithm only needs to search within a spherical shell with a thickness of at most $M - m$.

A.6 Numerical setup

Causal bounds. To evaluate the effectiveness of our proposed Algorithm 4, we compare it with the method proposed by Li and Pearl (2022) when all variables are binary. Specifically, we randomly generate distributions $\mathbb{P}(a, y, w)$, as shown in Table 2, and set $\mathbb{P}(U = 1) = 0.1$. We implement Algorithm 4 with a batch size of 20000, and set $\epsilon = 0$ as Li and Pearl (2022) assume the given distributions are accurate.

(A, Y, W)	$(0, 0, 0)$	$(0, 0, 1)$	$(0, 1, 0)$	$(0, 1, 1)$	$(1, 0, 0)$	$(1, 0, 1)$	$(1, 1, 0)$	$(1, 1, 1)$
$\mathbb{P}(a, y, w)$	0.2328	0.1784	0.1351	0.1467	0.0304	0.1183	0.0149	0.1433

sample space for x_{ijkl}	valid sample proportion
$[0, 1]$	≈ 0
$[\max\{0, \theta_{ijk} + \theta_l - 1\}, \min\{\theta_{ijk}, \theta_l\}]$	$< 10^{-4}$
support found by LP(3)	0.3%
Algorithm 4	100%

Table 5: Valid sample proportion with different sample spaces for the given example in Section 4.

Transfer learning in MAB. We perform simulation for 5-armed Bernoulli bandits with probability $0.1, 0.2, 0.3, 0.6, 0.8$. Simulations are partitioned into rounds of $T = 10^5$ trials averaged over 50 repetitions. For each task, we collect 1000 samples generated by a source agent and compute the empirical joint distribution. The estimated causal bounds without the knowledge of $F(u)$ (CUCB in Figure 2 (Zhang and Bareinboim, 2017)) are $h(a) = 0.9, 0.35, 0.92, 0.96, 0.92$, and $l(a) = 0, 0.08, 0.1, 0.3, 0.4$. The estimated causal bounds with the knowledge of $F(u)$ are $h(a) = 0.2, 0.25, 0.77, 0.7, 0.9$, and $l(a) = 0.01, 0.08, 0.19, 0.38, 0.71$.

Transfer learning in contextual bandits. We generate a function space $\mathcal{F} = \{(w - w_0)^\top (a - a_0)\}$ with a size of 50 by sampling parameters w_0, a_0 in $\mathbb{R}^d$ from $\mathcal{N}(0, 0.1)$, where $d = 10$. We then randomly choose a function as the true reward function f^* from the first 5 functions, and generate the reward as $Y = f^*(W, A) + \mathcal{N}(0, 0.1)$, where the context W is drawn i.d.d. from standard normal distributions and A is the selected action. The whole action set $\mathcal{A}$ is randomly initialized from $[-1, 1]^d$ with a size of 10. We repeat each instance 50 times to obtain a smooth regret curve.

A.7 Related work

Partially Observable Markov decision process (POMDP), including general partially observable dynamical systems (Uehara et al., 2022), also shares the similarity with our setting. Researchers have developed various methods to address causal inference in POMDPs. For example, Guo et al. (2022) use instrumental variables to identify causal effects, while Shi et al. (2022); Lu et al. (2023) extend this approach to general proxy variables in offline policy evaluation. In online reinforcement learning, Jin et al. (2019); Wang et al. (2021) use the backdoor criterion to explicitly adjust for confounding bias when confounders are fully observable. They also incorporate uncertainty from partially observable confounders into the Bellman equation and demonstrate provably optimal learning with linear function approximation in both fully and partially observable tasks. However, due to the complexity of reinforcement learning, transfer learning in POMDPs with the general function approximation still remains unknown. In our task 3, we address the problem of partially observable contextual bandit with the general function approximation under realizability assumption, which shows the potential to generalize to POMDPs and other related settings.

B Deferred Proofs

B.1 Proof of mentioned facts

Fact B.1.1 *Given a series of known observational distributions $F^1, \cdots, F^n$, consider an optimization problem for causal effects:*

$$\inf / \sup CE(\mathcal{M})$$
$$F^i_{\mathcal{M}} = F^i, i = 1, \cdots, n,$$

where $CE(\mathcal{M})$ is the desired casual effect and $F^i_{\mathcal{M}}$ is a distribution in the model $\mathcal{M}$. Here, the inf/sup is taken with respect to all compatible causal models $\mathcal{M}$. Then, a sufficient and necessary condition to identify $CE(\mathcal{M})$ is $LB = UB$, where LB and UB are the lower and upper bound solutions to the optimization problem.

Proof. If $LB = HB$, then for any compatible model $\mathcal{M}_1$ and $\mathcal{M}_2$, we have

$$LB = CE(\mathcal{M}_1) = CE(\mathcal{M}_2) = UB.$$

According the definition of causal identification, the required causal effect $CE(\mathcal{M})$ can be fully identified.

On the contrary, suppose $CE(\mathcal{M})$ is causal identifiable. Then for any compatible model pair $\mathcal{M}_1$ and $\mathcal{M}_2$, we have $CE(\mathcal{M}_1) = CE(\mathcal{M}_2)$. Traveling over all compatible models immediately yields

$$LB = CE(\mathcal{M}_1) = CE(\mathcal{M}_2) = UB.$$

□

Fact B.1.2 *During discretization, equality constraints in Theorem A.1 are automatically satisfied in the sense of integration.*

Proof. The first constraint has been checked.

For the second equality, we integrate over $\mathcal{U}_l$ and have

$$\begin{aligned}
&\int_{u\in\mathcal{U}_l} dF(u)du \\
=&\int_{a\in\mathcal{A},y\in\mathcal{Y},w\in\mathcal{W},u\in\mathcal{U}_l} dF(a,y,w,u)du \\
=&\sum_{ijk}\int_{a\in\mathcal{A}_i,\in\mathcal{Y}_j,w\in\mathcal{W}_k,u\in\mathcal{U}_l} dF(a,y,w,u)du \\
=&\sum_{ijk} x_{ijkl}.
\end{aligned}$$

Hence, the second equality constraint holds in the sense of integration.

For the third and the fourth equality constrains, we can do integration over corresponding blocks and check the equality in the same way.

The conditional distribution in the fifth equality can be approximated by x_{ijkl}. See details in the proof of approximating objective $\hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a)]$.

□

Fact B.1.3 *The object in Theorem A.1 after discretization is approximately equal to $\hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a)]$.*

Proof. We use the average values to approximate distributions at certain points. First, we only need to consider the value

$$\int_{y\in\mathcal{Y}_j,w\in\mathcal{W}_k,u\in\mathcal{U}_l} y\frac{dF(a,y,w,u)dF(w,u)}{dF(a,w,u)},$$

as summing over j, k, l can yield the object.

Suppose the given $a \in \mathcal{A}_i$ and let $vol(\cdot)$ denote the volume of the given block. For the values of distributions in $\mathcal{A}_i \times \mathcal{Y}_j \times \mathcal{W}_k \times \mathcal{U}_l$, we have

$$\begin{aligned} dF(a,y,w,u) \approx & \frac{dadydwdu}{vol(\mathcal{A}_i)vol(\mathcal{Y}_j)vol(\mathcal{W}_k)vol(\mathcal{U}_l)} \int_{a \in \mathcal{A}_i, y \in \mathcal{Y}_j, w \in \mathcal{W}_k, u \in \mathcal{U}_l} dF(a,y,w,u) \\ = & \frac{x_{ijkl} dadydwdu}{vol(\mathcal{A}_i)vol(\mathcal{Y}_j)vol(\mathcal{W}_k)vol(\mathcal{U}_l)}, \end{aligned}$$

$$\begin{aligned} dF(a,w,u) \approx & \frac{dadwdu}{vol(\mathcal{A}_i)vol(\mathcal{W}_k)vol(\mathcal{U}_l)} \int_{a \in \mathcal{A}_i, w \in \mathcal{W}_k, u \in \mathcal{U}_l} dF(a,w,u) \\ = & \frac{dadwdu}{vol(\mathcal{A}_i)vol(\mathcal{W}_k)vol(\mathcal{U}_l)} \int_{a \in \mathcal{A}_i, y \in \mathcal{Y}, w \in \mathcal{W}_k, u \in \mathcal{U}_l} dF(a,y,w,u) \\ = & \frac{\sum_{j'} x_{ij'kl} dadwdu}{vol(\mathcal{A}_i)vol(\mathcal{W}_k)vol(\mathcal{U}_l)}, \end{aligned}$$

$$\begin{aligned} dF(w,u) \approx & \frac{dwdu}{vol(\mathcal{W}_k)vol(\mathcal{U}_l)} \int_{w \in \mathcal{W}_k, u \in \mathcal{U}_l} dF(w,u) \\ = & \frac{dwdu}{vol(\mathcal{W}_k)vol(\mathcal{U}_l)} \int_{a \in \mathcal{A}, y \in \mathcal{Y}, w \in \mathcal{W}_k, u \in \mathcal{U}_l} dF(a,y,w,u) \\ = & \frac{\sum_{i',j'} x_{i'j'kl} dwdu}{vol(\mathcal{W}_k)vol(\mathcal{U}_l)}. \end{aligned}$$

Plugging all above equalities yields

$$\begin{aligned} & \int_{y \in \mathcal{Y}_j, w \in \mathcal{W}_k, u \in \mathcal{U}_l} y \frac{dF(a,y,w,u)dF(w,u)}{dF(a,w,u)} \\ \approx & \frac{x_{ijkl} \sum_{i',j'} x_{i'j'kl}}{\sum_{j'} x_{ij'kl}} \int_{y \in \mathcal{Y}_j, w \in \mathcal{W}_k, u \in \mathcal{U}_l} \\ = & \frac{x_{ijkl} \sum_{i',j'} x_{i'j'kl}}{\sum_{j'} x_{ij'kl}} \int_{y \in \mathcal{Y}_j, w \in \mathcal{W}_k, u \in \mathcal{U}_l} ydydydu/vol(\mathcal{Y}_j)vol(\mathcal{W}_k)vol(\mathcal{U}_l) \\ \approx & \frac{y_j x_{ijkl} \sum_{i',j'} x_{i'j'kl}}{\sum_{j'} x_{ij'kl}}. \end{aligned}$$

If y_j is chosen to be $\frac{\int_{y \in \mathcal{Y}_j} ydy}{\int_{y \in \mathcal{Y}_j} dy}$, then the last symbol of approximation can be replaced with the symbol of equal. If $\mathcal{Y}_j$ is an interval, y_j can be chosen as the midpoint of $\mathcal{Y}_j$. □

The step of approximating dF is crucial in reducing the approximation error. For absolutely continuous cumulative distribution functions, the approximation error will converge to zero as the diameter of each block approaches zero. Furthermore, if all random variables are discrete, the approximation error can be exactly zero when using the natural discretization. In this case, the original objective can be expressed as

$$\sum_{y \in \mathcal{Y}, w \in \mathcal{W}, u \in \mathcal{U}} y \frac{\mathbb{P}(A=a, Y=y, W=w, U=u)\mathbb{P}(W=w, U=u)}{\mathbb{P}(A=a, W=w, U=u)}.$$

Our discretization method can be regarded as approximating probability mass functions.

Fact B.1.4 *If the estimator $\hat{\theta}_x$ for $\theta_x \in [0,1]$ has n_x i.i.d. samples, then the variance of $\hat{\theta}_x$ is at most $\frac{1}{4n_x}$.*

Proof. Since $\theta_x^2 \le \theta_x$, then

$$Var(\theta_x) = \mathbb{E}[\theta_x^2] - (\mathbb{E}[\theta_x])^2 \le \mathbb{E}[\theta_x] - (\mathbb{E}[\theta_x])^2 \le \frac{1}{4}.$$

Hence, we have

$$Var(\hat{\theta}_x) = \frac{1}{n_x} Var(\theta_x) \leq \frac{1}{4n_x}.$$

□

B.2 Proof of Theorem A.1

Proof. Since W and U satisfies the back-door criterion, we can condition on W, U to identify the causal effect $\mathbb{E}[Y|do(a_0)]$ We have

$$\begin{aligned}\mathbb{E}[Y|do(a_0)] &= \int_{w\in\mathcal{W}, u\in\mathcal{U}} \mathbb{E}[Y|do(a_0), w, u] dF(w,u)\\ &= \int_{w\in\mathcal{W}, u\in\mathcal{U}} \mathbb{E}[Y|a_0, w, u] dF(w,u)\\ &= \int_{w\in\mathcal{W}, u\in\mathcal{U}} \int_{y\in\mathcal{Y}} y dF(y|a_0, w, u) dF(w,u).\end{aligned}$$

The equalities come from the normalization properties of distribution functions. The inequalities come from the estimation error. □

B.3 Proof of convergence results of Algorithm 4

We first prove the following lemma to show that our sampling algorithm can cover all values in the feasible region $\mathcal{D}$. We denote the truncated distribution to $[l,h]$ from the user-given distribution F_s when x_i is given in sequential LPs as $F_s(x|x_i,[l,h])$.

Lemma B.1 *The Algorithm 7 induces a distribution on the given simplex $\mathcal{D}$.*

Proof. We need to prove the sample generated by Algorithm 7 can exactly cover the region of $\mathcal{D}$. On the one hand, for any output $\mathbf{x}$, the feasibility of each component of $\mathbf{x}$ indicates that $\mathbf{x}$ must lie in $\mathcal{D}$. On the another hand, for any $\hat{\mathbf{x}} \in \mathcal{D}$, we show that this point can be generated by solving sequential LPs. Since $\hat{\mathbf{x}} \in \mathcal{D}$, $\hat{\mathbf{x}}$ is a feasible solution to the first LP

$$\begin{gathered}\min / \max x_1\\ s.t. A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}.\end{gathered}$$

One can check the feasibility of $\hat{\mathbf{x}}$ in the following LPs

$$\begin{aligned}&\min / \max x_i\\ s.t. &A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\\ &x_1 = \hat{x}_1, \cdots, x_j = \hat{x}_j, j = 1, 2, \cdots, i-1,\end{aligned}$$

because $\hat{\mathbf{x}} \in \mathcal{D}$ and the previous i components are exactly equal to those of $\hat{\mathbf{x}}$.

Suppose that the number of components of $\hat{\mathbf{x}}$ is d. Then the induced distribution is

$$F(\hat{\mathbf{x}}) = F_s(x_1|[l_1,h_1]) \prod_{i=2}^{d} F_s(x_i|[l_i,h_i], x_j, j = 1, 2, \cdots, i-1).$$

□

We now give the proof for Proposition B.1.

Proposition B.1 *Assume that the sampling measure $\mathbb{P}_s$ satisfies $\forall \mathbf{x} \in \mathcal{D}$, and $\forall \delta > 0$,*

$$\mathbb{P}_s(\mathcal{B}(\mathbf{x}, \delta) \cap \mathcal{D}) > 0,$$

where $\mathcal{B}(\mathbf{x}, \delta)$ is a ball centered at $\mathbf{x}$ with radius δ. If the discrepancy parameter is set to 0, then $b_{(1)}(a)$ converges to $\hat{l}(a)$ in probability and $b_{(B)}(a)$ converges to $\hat{h}(a)$ in probability for $B \to \infty$.

Proof. The discretization optimization problem (6) has one-to-one correspondence between $\mathbf{x}$ and each causal model where all random variables are discrete. From Lemma B.1, we know the one-to-one correspondence between each model and $\mathbf{x}$. Therefore, $[\hat{l}(a), \hat{h}(a)]$ is the support of the induced distribution on casual bounds.

As shown in $\hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a)]$, the object can be regarded as a function of $\mathbf{x}$. We define $\phi = \hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a)]$ which is a continuous mapping from $\mathcal{D}$ to $[0,1]$. The continuity of ϕ is clear for $\sum j' x_{ij'kl} > 0$. When $\sum_{j'} x_{ij'kl} = 0$, the non-negativity of x_{ijkl} implies $\sum_{i',j'} x_{i'j'kl} = 0$. In this case, we can set the value of $\frac{y_j x_{ijkl} \sum_{i',j'} x_{i'j'kl}}{\sum_{j'} x_{ij'kl}}$ to be 0 to maintain continuity.

Given that $\mathcal{D}$ is a compact set, there exists $\mathbf{x}_l$ such that $\phi(\mathbf{x}_l) = \hat{l}(a)$. The continuity indicates that $\forall \epsilon > 0$, there exists $\delta > 0$ such that $\phi(\mathbf{x}) < \hat{l}(a) + \epsilon$ for all $\mathbf{x} \in \mathcal{B}(x_h, \delta)$, then

$$\mathbb{P}(b(a) < \hat{l}(a) + \epsilon) = \mathbb{P}_s\left(\bigcup_{b < \hat{l}(a)+\epsilon} \{\mathbf{x} \in \mathcal{D} | \phi(\mathbf{x}) = b\}\right) \geq \mathbb{P}_s(\mathcal{B}(\mathbf{x}_h, \delta)) > 0.$$

This implies that

$$\mathbb{P}(b_{(1)}(a) < \hat{l}(a) + \epsilon) = 1 - (1 - \mathbb{P}(b(a) < \hat{l}(a) + \epsilon))^B \to 1$$

as $B \to \infty$. Since $b_{(1)}(a)$ is a feasible solution to the discrete optimization problem, we have

$$\mathbb{P}(\hat{l}(a) \leq b_{(1)}(a) < \hat{l}(a) + \epsilon) = 1 - (1 - \mathbb{P}(b(a) < \hat{l}(a) + \epsilon))^B \to 1$$

which implies $b_{(1)}(a) \to \hat{l}(a)$ in probability.

Similarly, we can prove that $\mathbb{P}(b_{(B)}(a) > \hat{h}(a) - \epsilon) < 1$ and thus $b_{(B)}(a) \to \hat{h}(a)$ in probability.

□

Proof. The one-to-one correspondence between $\mathbf{x}$ and each causal model has been proved in Lemma B.1. Therefore, $[\hat{l}(a), \hat{h}(a)]$ is the support of the induced distribution on casual bounds. As shown in the proof of Proposition B.1, the defined $\phi = \hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a)]$ is a continuous mapping from $\mathcal{D}$ to $[0,1]$.

Given that $\mathcal{D}$ is a compact set, there exists $\mathbf{x}_l$ such that $\phi(\mathbf{x}_l) = \hat{l}(a)$. The property of **OPT** implies that there exists $\delta > 0$ such that

$$\mathbf{OPT}(min/max, \hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a)], \mathcal{D}, \mathbf{x}) = \mathbf{x}_l, \forall \mathbf{x} \in \mathcal{B}(\mathbf{x}_l, \delta) \cap \mathcal{D}.$$

Hence, we have

$$\mathbb{P}(b(a) = \hat{l}(a)) \geq \mathbb{P}_s(\mathcal{B}(\mathbf{x}_l, \delta) \cap \mathcal{D}) > 0.$$

Due to Borel-Cantelli lemma we can prove that $b_{(1)}(a) \to \hat{l}(a)$ almost surely, because each $b_i(a)$ is independently sampled by Algorithm 1. Similarly, we can prove that $b_{(B)}(a) \to \hat{h}(a)$ almost surely.

□

B.4 Proof of regrets in MAB

We first prove Theorem A.2.

Proof. **Case 1**: $h(a) < \max_{i \in \mathcal{A}} l(i)$

From the algorithmic construction, we know that such arm a is removed and thus

$$\mathbb{E}[N_a(T)] = 0.$$

Case 2: $\max_{i \in \mathcal{A}} l(i) \leq h(a) < \mu^*$.

Let $a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[Y|do(a)]$ be the optimal action with respect to w. Define the following event

$$\mathcal{E}(t) = \left\{\hat{\mu}_a \in [\mu_a - \frac{\log t}{n_a(t)}, \mu_a + \frac{\log t}{n_a(t)}], \forall a \in \mathcal{A}\right\}$$

Then the Chernoff's bound yields

$$\mathbb{P}(\overline{\mathcal{E}(t)}) \leq \sum_{a \in \mathcal{A}} \exp(-2n_a(t) \times \frac{\log t}{n_a(t)}) \leq \frac{|\mathcal{A}|}{t^2}$$

For any given w, the event $\{A_t = a\}$ implies $\hat{U}_a(t) > \hat{U}_{a^*}(t)$. However,

$$\mu^* > h(a) \geq \hat{U}_a(t)$$

and

$$\hat{U}_{a^*}(t) \geq \mu^*$$

if $\mathcal{E}(t)$ holds. This leads to contradiction. Therefore,

$$\begin{aligned}
\mathbb{E}[N_a(T)] &= \sum_{t=1}^{T} \mathbb{P}(A_t = a) \\
&= \sum_{t=1}^{T} \mathbb{P}(A_t = a | \mathcal{E}(t)) \mathbb{P}(\mathcal{E}(t)) + \mathbb{P}(A_t = a | \overline{\mathcal{E}(t)}) \mathbb{P}(\overline{\mathcal{E}(t)}) \\
&\leq \sum_{t=1}^{T} \mathbb{P}(\overline{\mathcal{E}(t)}) \\
&\leq \sum_{t=1}^{T} \frac{|\mathcal{A}|}{t^2} \\
&\leq \frac{|\mathcal{A}|\pi^2}{6}.
\end{aligned}$$

Case 3: $h(a) \geq \mu^*$

We reuse the notation $\mathcal{E}(t)$ in the case 2. Condition on the event $\bigcap_{t=1}^{T} \mathcal{E}(t)$, if $n_a(t) \geq \frac{8 \log T}{\Delta_a^2}$, then

$$\hat{U}_a(t) \leq U_a(t) = \mu_a + \sqrt{\frac{2 \log t}{n_a(t)}} \leq \mu_a + \frac{1}{2}\Delta_a = \mu^* \leq \hat{U}_{a^*}(t),$$

so Algorithm 2 will not choose the action a at the round t. Therefore,

$$\mathbb{E}[N_a(T)] \leq \mathbb{E}\left[N_a(T) \middle| \bigcap_{t=1}^{T} \mathcal{E}(t) \right] + T\mathbb{P}\left(\overline{\bigcap_{t=1}^{T} \mathcal{E}(t)} \right) \leq \frac{8 \log T}{\Delta_a^2} + T\mathbb{P}\left(\overline{\bigcup_{t=1}^{T} \mathcal{E}(t)} \right).$$

We conclude the proof by showing

$$T\mathbb{P}\left(\overline{\bigcup_{t=1} \mathcal{E}(t)} \right) \leq T \sum_{t=1}^{T} \frac{|\mathcal{A}|}{t^2} < T \times \frac{|\mathcal{A}|}{T} = |\mathcal{A}|.$$

□

Actually, the proof is just a simple modification of that in Theorem 3.1, because MAB can be regarded as a special case of contextual bandits.

Theorem B.1 *Consider a MAB bandit problem with $|\mathcal{A}| < \infty$. Denote*

$$\widetilde{\mathcal{A}^*} = \mathcal{A} - \{a \in \mathcal{A} | h(a) < \mu^*\}.$$

Then the regret of Algorithm 5 is upper bounded by

$$\mathbb{E}[Reg(T)] \leq \sqrt{8(|\widetilde{\mathcal{A}^*}(w)| - 1)T \log T}.$$

Proof. Theorem A.2 shows that

$$\begin{aligned}\mathbb{E}[Reg(T)] &= \sum_{a\in\mathcal{A}} \Delta_a \mathbb{E}[N_a(T)] \\ &\le \sum_{a\in\widetilde{\mathcal{A}^*}} \frac{8\log T}{\Delta_a} \mathbf{I}\left\{\Delta_a \ge \Delta\right\} + T\Delta + \mathcal{O}(|\mathcal{A}|) \\ &\le \frac{8(|\widetilde{\mathcal{A}^*}|-1)\log T}{\Delta} + T\Delta + \mathcal{O}(|\mathcal{A}|).\end{aligned}$$

Specifying $\Delta = \sqrt{\frac{8(|\widetilde{\mathcal{A}^*}|-1)\log T}{T}}$ concludes the proof. □

Denote the contextual bandit instances with prior knowledge $l(a)$ and $h(a)$ as

$$\mathfrak{M} = \{\text{MAB bandit instances with } l(a) \le \mu_a \le h(a), \forall a \in \mathcal{A}\}.$$

Theorem B.2 *Suppose $|\mathcal{A}| < \infty$. Then for any algorithm A, there exists an absolute constant $c > 0$ such that*

$$\min_{\mathsf{A}} \sup_{\mathfrak{M}} Reg(T) \ge \frac{1}{27}\sqrt{(|\widetilde{\mathcal{A}^*}|-1)T}.$$

Proof. This is a direct corollary of MAB regret lower bound, because any arm in $\widetilde{\mathcal{A}^*}$ cannot be the optimal one. □

B.5 Omitted theorems in Task 3

From the rule of do-calculus, we have

$$\begin{aligned}\mathbb{E}[Y|do(a), w] &= \int_{u\in\mathcal{U}} \mathbb{E}[Y|do(a), w, u] dF(u) \\ &= \int_{u\in\mathcal{U}} \mathbb{E}[Y|a, w, u] dF(u).\end{aligned}$$

The last equality is due to the rule of do-calculus as W and U is sufficient to block all back-door paths from A to Y.

Theorem B.3 *Given a causal diagram $\mathcal{G}$ and a distribution compatible with $\mathcal{G}$, let $\{W, U\}$ be a set of variables satisfying the back-door criterion in $\mathcal{G}$ relative to an ordered pair (A, Y), where $\{W, U\}$ is partially observable, i.e., only probabilities $\hat{F}(a, y, w)$ and $\hat{F}(u)$ with the maximum estimation error ϵ, the causal effects of $A = a_0$ on Y when $W = w_0$ occurs are then bounded as follows:*

$$l(w_0, a_0) \le \mathbb{E}[Y|do(a_0), w_0] \le h(w_0, a_0),$$

where $l(w_0, a_0)$ and $h(w_0, a_0)$ are solutions to the following functional optimization problem for any given a_0 and w_0

$$\begin{aligned}& l(w_0, a_0) = \inf \int_{w\in\mathcal{W}, u\in\mathcal{U}} \int_{y\in\mathcal{Y}} y dF(y|a_0, w_0, u) dF(u) \\ & h(w_0, a_0) = \sup \int_{w\in\mathcal{W}, u\in\mathcal{U}} \int_{y\in\mathcal{Y}} y dF(y|a_0, w_0, u) dF(u) \\ & s.t. F(y|a, w, u)F(a, w, u) = F(a, y, w, u), \forall (a, y, w, u) \in \mathcal{A}\times\mathcal{Y}\times\mathcal{W}\times\mathcal{U} \\ & \int_{y\in\mathcal{Y}} dF(a, y, w, u) = F(a, w, u), \forall (a, w, u) \in \mathcal{A}\times\mathcal{W}\times\mathcal{U} \\ & \int_{a\in\mathcal{A}, y\in\mathcal{Y}, w\in\mathcal{W}} dF(a, y, w, u) = F(u), \forall u \in \mathcal{U} \\ & \int_{u\in\mathcal{U}} dF(a, y, w, u) = F(a, y, w), \forall (a, y, w) \in \mathcal{A}\times\mathcal{W}\times\mathcal{U} \\ & |F(a, y, w) - \hat{F}(a, y, w)| \le \epsilon, \forall (a, y, w) \in \mathcal{A}\times\mathcal{Y}\times\mathcal{W} \\ & |F(u) - \hat{F}(u)| \le \epsilon, \forall u \in \mathcal{U}.\end{aligned}$$

Here the inf/sup is taken with respect to all unknown cumulative distribution functions $F(a, y, w, u)$, $F(a, y, w)$, $F(a, w, u)$, $F(u)$.

Proof. The object is shown at the beginning of this subsection. The equalities come from the normalization properties, and inequalities follow from estimation error. □

Denote the following optimization problem

$$\begin{aligned} &\max / \min \hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a), w] \\ &\sum_{l} x_{ijkl} = \theta_{ijk}, \forall i \in [n_{\mathcal{A}}], j \in [n_{\mathcal{Y}}], k \in [n_{\mathcal{W}}] \\ &\sum_{ijk} x_{ijkl} = \theta_l, \forall l \in [n_{\mathcal{U}}]. \end{aligned} \tag{8}$$

where the objective $\hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a), w]$ after discretization is defined as

$$\hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a), w] = \sum_{jkl} \frac{y_j \theta_l x_{ijkl}}{\sum_{j'} x_{ij'kl}}. \tag{9}$$

Denote the solutions to (8) as $\hat{l}(w, a)$ and $\hat{h}(w, a)$. Note that the optimization problem (7) shares the same feasible region with that of (1).

Proposition B.2 *Assume that the sampling measure $\mathbb{P}_s$ satisfies $\forall \mathbf{x} \in \mathcal{D}$, and $\forall \delta > 0$,*

$$\mathbb{P}_s(\mathcal{B}(\mathbf{x}, \delta) \cap \mathcal{D}) > 0,$$

where $\mathcal{B}(\mathbf{x}, \delta)$ is a ball centered at $\mathbf{x}$ with radius δ. If the discrepancy parameter is set to 0, then $b_{(1)}(w, a)$ converges to $\hat{l}(w, a)$ in probability and $b_{(B)}(w, a)$ converges to $\hat{h}(w, a)$ in probability for any given (w, a) and $B \to \infty$.

Proof. As shown in $\hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a), w]$, the object can also be regarded as a function of $\mathbf{x}$. We define $\phi = \hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a), w]$ which is a continuous mapping from $\mathcal{D}$ to $[0, 1]$. The continuity of ϕ holds similarly.

Given that $\mathcal{D}$ is a compact set, there exists $\mathbf{x}_l$ such that $\phi(\mathbf{x}_l) = \hat{l}(w, a)$. The continuity indicates that $\forall \epsilon > 0$, there exists $\delta > 0$ such that $\phi(\mathbf{x}) < \hat{l}(w, a) + \epsilon$ for all $\mathbf{x} \in \mathcal{B}(x_h, \delta)$, then

$$\mathbb{P}(b(w, a) < \hat{l}(w, a) + \epsilon) = \mathbb{P}_s\left(\bigcup_{b < \hat{l}(w,a)+\epsilon} \{\mathbf{x} \in \mathcal{D} | \phi(\mathbf{x}) = b\}\right) \geq \mathbb{P}_s(\mathcal{B}(\mathbf{x}_h, \delta)) > 0.$$

This implies that

$$\mathbb{P}(b_{(1)}(w, a) < \hat{l}(w, a) + \epsilon) = 1 - (1 - \mathbb{P}(b(w, a) < \hat{l}(w, a) + \epsilon))^B \to 1$$

as $B \to \infty$. Since $b_{(1)}(a)$ is a feasible solution to the discrete optimization problem, we have

$$\mathbb{P}(\hat{l}(w, a) \leq b_{(1)}(w, a) < \hat{l}(w, a) + \epsilon) = 1 - (1 - \mathbb{P}(b(w, a) < \hat{l}(w, a) + \epsilon))^B \to 1$$

which implies $b_{(1)}(w, a) \to \hat{l}(w, a)$ in probability.

Similarly, we can prove that $\mathbb{P}(b_{(B)}(w, a) > \hat{h}(w, a) - \epsilon) < 1$ and thus $b_{(B)}(w, a) \to \hat{h}(w, a)$ in probability.

□

We can also incorporate the optimization procedure **OPT**. Replacing the objective $\hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a)]$ in Algorithm 1 with $\hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a), w]$, we can also prove the similar almost surely convergence result.

Proposition B.3 *Assume that the sampling measure $\mathbb{P}_s$ satisfies $\forall \mathbf{x} \in \mathcal{D}$, and $\forall \delta > 0$,*

$$\mathbb{P}_s(\mathcal{B}(\mathbf{x}, \delta) \cap \mathcal{D}) > 0,$$

*where $\mathcal{B}(\mathbf{x}, \delta)$ is a ball centered at $\mathbf{x}$ with radius δ. Given a deterministic procedure **OPT** which satisfies for any local optima $\mathbf{x}_{loc}$, there exists $\delta > 0$ such that for any initial guess $\mathbf{x}_0 \in \mathcal{B}(\mathbf{x}_{loc}, \delta) \cap \mathcal{D}$, **OPT** can output $\mathbf{x}_{loc}$ as a result. If the discrepancy parameter is set to 0, then $b_{(1)}(w, a)$ and $b_{(B)}(w, a)$ converge almost surely to $\hat{l}(w, a)$ and $\hat{h}(w, a)$ for $B \to \infty$, respectively.*

Proof. The one-to-one correspondence between $\mathbf{x}$ and each causal model has been proved in Lemma B.1. Therefore, $[\hat{l}(w,a), \hat{h}(w,a)]$ is the support of the induced distribution on casual bounds. As shown in the proof of Proposition B.1, the defined $\phi = \hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a), w]$ is a continuous mapping from $\mathcal{D}$ to $[0,1]$.

Given that $\mathcal{D}$ is a compact set, there exists $\mathbf{x}_l$ such that $\phi(\mathbf{x}_l) = \hat{l}(w,a)$. The property of **OPT** implies that there exists $\delta > 0$ such that

$$\mathbf{OPT}(min/max, \hat{\mathbb{E}}_{\mathcal{M}}[Y|do(a)], \mathcal{D}, \mathbf{x}) = \mathbf{x}_l, \forall \mathbf{x} \in \mathcal{B}(\mathbf{x}_l, \delta) \cap \mathcal{D}.$$

Hence, we have

$$\mathbb{P}(b(w,a) = \hat{l}(w,a)) \geq \mathbb{P}_s(\mathcal{B}(\mathbf{x}_l, \delta) \cap \mathcal{D}) > 0.$$

Due to Borel-Cantelli lemma we can prove that $b_{(1)}(w,a) \to \hat{l}(w,a)$ almost surely, because each $b_i(w,a)$ is independently sampled by Algorithm 1. Similarly, we can prove that $b_{(B)}(w,a) \to \hat{h}(w,a)$ almost surely.

□

B.6 Proof of Theorem 3.1

Proof. We consider any given w in the following proof.

Case 1: $h(w,a) < \max_{i \in \mathcal{A}} l(w,i)$

From the algorithmic construction, we know that such arm a is removed and thus

$$\mathbb{E}[N_a(T_w)] = 0.$$

Case 2: $\max_{i \in \mathcal{A}} l(w,i) \leq h(w,a) < \mu_w^*$.

Let $a_w^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[Y|do(a), w]$ be the optimal action with respect to w. Define the following event

$$\mathcal{E}_w(t) = \left\{ \hat{\mu}_{w,a} \in [\mu_{w,a} - \frac{\log t}{n_{w,a}(t)}, \mu_{w,a} + \frac{\log t}{n_{w,a}(t)}], \forall a \in \mathcal{A} \right\}$$

Then the Chernoff's bound yields

$$\mathbb{P}(\overline{\mathcal{E}_w(t)}) \leq \sum_{a \in \mathcal{A}} \exp(-2 n_{w,a}(t) \times \frac{\log t}{n_{w,a}(t)}) \leq \frac{|\mathcal{A}|}{t^2}$$

For any given w, the event $\{A_t = a\}$ implies $\hat{U}_{w,a}(t) > \hat{U}_{w,a_w^*}(t)$. However,

$$\mu_w^* > h(w,a) \geq \hat{U}_{w,a}(t)$$

and

$$\hat{U}_{w,a_w^*}(t) \geq \mu_w^*$$

if $\mathcal{E}_w(t)$ holds. This leads to contradiction. Therefore,

$$\begin{aligned}
\mathbb{E}[N_a(T_w)] &= \sum_{t=1}^{T_w} \mathbb{P}(A_t = a | w_t = w) \\
&= \sum_{t=1}^{T_w} \mathbb{P}(A_t = a | w_t = w, \mathcal{E}_w(t)) \mathbb{P}(\mathcal{E}_w(t)) + \mathbb{P}(A_t = a | w_t = w, \overline{\mathcal{E}_w(t)}) \mathbb{P}(\overline{\mathcal{E}_w(t)}) \\
&\leq \sum_{t=1}^{T_w} \mathbb{P}(\overline{\mathcal{E}_w(t)}) \\
&\leq \sum_{t=1}^{T_w} \frac{|\mathcal{A}|}{t^2} \\
&\leq \frac{|\mathcal{A}| \pi^2}{6}.
\end{aligned}$$

Case 3: $h(w,a) \geq \mu_w^*$

We reuse the notation $\mathcal{E}_w(t)$ in the case 2. Condition on the event $\bigcap_{t=1}^{T_w} \mathcal{E}_w(t)$, if $n_{w,a}(t) \geq \frac{8 \log T_w}{\Delta_{w,a}^2}$, then

$$\hat{U}_{w,a}(t) \leq U_{w,a}(t) = \mu_{w,a} + \sqrt{\frac{2 \log t}{n_{w,a}(t)}} \leq \mu_{w,a} + \frac{1}{2}\Delta_{w,a} = \mu_w^* \leq \hat{U}_{w,a_w^*}(t),$$

so Algorithm 2 will not choose the action a at the round t. Therefore,

$$\mathbb{E}[N_a(T_w)] \leq \mathbb{E}\left[N_a(T_w), \bigcap_{t=1}^{T_w} \mathcal{E}_w(t)\right] + T_w \mathbb{P}\left(\overline{\bigcap_{t=1}^{T_w} \mathcal{E}_w(t)}\right) \leq \frac{8 \log T_w}{\Delta_{w,a}^2} + T_w \mathbb{P}\left(\overline{\bigcup_{t=1}^{T_w} \mathcal{E}_w(t)}\right).$$

We conclude the proof by showing

$$T_w \mathbb{P}\left(\overline{\bigcup_{t=1}^{T_w} \mathcal{E}_w(t)}\right) \leq T_w \sum_{t=1}^{T_w} \frac{|\mathcal{A}|}{t^2} < T_w \times \frac{|\mathcal{A}|}{T_w} = |\mathcal{A}|.$$

□

B.7 Proof of Theorem A.3

Proof. Let Δ_w be the constant with respect to w that we will specify later. From the proof of Theorem 3.1, we know that the expected regret can be upper bounded as

$$\begin{aligned}\mathbb{E}[Reg(T)] &= \sum_{w \in \mathcal{W}} \sum_{a \in \mathcal{A}} \Delta_{w,a} \mathbb{E}[N_a(T_w)] \\ &\leq \sum_{w \in \mathcal{W}} \left(\sum_{a \in \widetilde{\mathcal{A}^*}(w)} \frac{8 \log T_w}{\Delta_{w,a}} \mathbf{I}\left\{\Delta_{w,a} \geq \Delta_w\right\} + T\Delta_w \right) + \mathcal{O}(|\mathcal{A}|) \\ &\leq \sum_{w \in \mathcal{W}} \left(\frac{8(|\widetilde{\mathcal{A}^*}(w)| - 1) \log T}{\Delta_w} + T_w \Delta_w \right) + \mathcal{O}(|\mathcal{A}|).\end{aligned}$$

We select $\Delta_w = \sqrt{\frac{8|\widetilde{\mathcal{A}^*}(w)| \log T}{T_w}}$ so

$$\mathbb{E}[Reg(T)] \leq \sum_{w \in \mathcal{W}} \sqrt{8(|\widetilde{\mathcal{A}^*}(w)| - 1) T_w \log T}.$$

By strong law of large numbers, we have

$$\begin{aligned}\liminf_{T \to \infty} \frac{\mathbb{E}[Reg(T)]}{\sqrt{T \log T}} &\leq \sum_{w \in \mathcal{W}} \sqrt{8(|\widetilde{\mathcal{A}^*}(w)| - 1) \liminf_{T \to \infty} \frac{T_w}{T}} \\ &= \sum_{w \in \mathcal{W}} \sqrt{8(|\widetilde{\mathcal{A}^*}(w)| - 1) \mathbb{P}(W = w)}.\end{aligned}$$

□

Proof. Consider $|\mathcal{W}|$ MAB instances. For any given context w, the set that the optimal arm will be in is $\mathcal{A}^*(w)$. For any algorithm A, let A_w be the induced algorithm of A when w occurs. From the minimax theorem for MAB instances (Lattimore and Szepesvári, 2020), we know that there exists a MAB instance for each w such that the regret of A_w is at least $\frac{1}{27}\sqrt{(|\widetilde{\mathcal{A}^*}(w)| - 1)T_w}$, where T_w is the number of occurrence of w. Hence,

$$Reg(T) \geq \sum_{w \in \mathcal{W}} \frac{1}{27} \sqrt{(|\widetilde{\mathcal{A}^*}(w)| - 1) T_w}.$$

and almost surely,

$$\begin{aligned}\liminf_{T\to\infty}\frac{Reg(T)}{\sqrt{T}} &\geq \frac{1}{27}\sum_{w\in\mathcal{W}}\sqrt{(|\widetilde{\mathcal{A}^*}(w)|-1)\cdot\liminf_{T\to\infty}\frac{T_w}{T}}\\ &=\frac{1}{27}\sum_{w\in\mathcal{W}}\sqrt{(|\widetilde{\mathcal{A}^*}(w)|-1)\mathbb{P}(W=w)}.\end{aligned}$$

□

B.8 Proof of Theorem 3.3

The framework presented in (Simchi-Levi and Xu, 2021; Foster et al., 2020) provides a method to analyze contextual bandit algorithms in the universal policy space Ψ. In this paper, we mainly focus on a subspace of Ψ shaped by causal bounds. We demonstrate that the action distribution p_m selected in Algorithm 3 possesses desirable properties that contribute to achieving low regrets.

For each epoch m and any round t in epoch m, for any possible realization of γ_t, $\hat{f}_m$, we define the universal policy space of Ψ:

$$\Psi = \prod_{w\in\mathcal{W}}\mathcal{A}^*(w).$$

With abuse of notations, we define

$$\mathcal{R}(\pi)=\mathbb{E}_W[f^*(W,\pi(W))] \text{ and } Reg(\pi)=\mathcal{R}(\pi_{f^*})-\mathcal{R}(\pi).$$

The above quantities do not depend on specific values of W. The following empirical version of above quantities are defined as

$$\widehat{\mathcal{R}}_t(\pi)=\hat{f}_{m(t)}(w,\pi(w)) \text{ and } \widehat{Reg}_t(\pi)=\mathbb{E}_W[\widehat{\mathcal{R}}_t(\pi_{\hat{f}_{m(t)}})-\widehat{\mathcal{R}}_t(\pi)],$$

where $m(t)$ is the epoch of the round t.

Let $Q_m(\cdot)$ be the equivalent policy distribution for $p_m(\cdot|\cdot)$, i.e.,

$$Q_m(\pi)=\prod_{w\in\mathcal{W}}p_m(\pi(w)|w), \forall\pi\in\Psi.$$

The existence and uniqueness of such measure $Q_m(\cdot)$ is a corollary of Kolmogorov's extension theorem. Note that both Ψ and $Q_m(\cdot)$ are $\mathcal{H}_{\tau_{m-1}}$-measurable, where $\mathcal{H}_t$ is the filtration up to the time t. We refer to Section 3.2 of (Simchi-Levi and Xu, 2021) for more detailed intuition for $Q_m(\cdot)$ and proof of existence. By Lemma 4 of (Simchi-Levi and Xu, 2021), we know that for all epoch m and all rounds t in epoch m, we can rewrite the expected regret in terms of our notations as

$$\mathbb{E}[Reg(T)]=\sum_{\pi\in\Psi}Q_m(\pi)Reg(\pi).$$

For simplicity, we define an epoch-dependent quantities

$$\rho_1=1,\rho_m=\sqrt{\frac{\eta\tau_{m-1}}{\log(2\delta^{-1}|\mathcal{F}^*|\log T)}}, m\geq 2,$$

so $\gamma_t=\sqrt{|\mathcal{A}^*(w_t)|}\rho_{m(t)}$ for $m(t)\geq 2$.

Lemma B.2 *(Implicit Optimization Problem). For all epoch m and all rounds t in epoch m, Q_m is a feasible solution to the following implicit optimization problem:*

$$\sum_{\pi\in\Psi}Q_m(\pi)\widehat{Reg}_t(\pi)\leq\mathbb{E}_W[\sqrt{|\mathcal{A}^*(W)|}]/\rho_m \tag{10}$$

$$\mathbb{E}_W\left[\frac{1}{p_m(\pi(W)|W)}\right]\leq\mathbb{E}_W[\mathcal{A}^*(W)]+\mathbb{E}_W[\sqrt{|\mathcal{A}^*(W)|}]\rho_m\widehat{Reg}_t(\pi),\forall\pi\in\Psi. \tag{11}$$

Proof. Let m and t in epoch m be fixed. Denote $\mathcal{P}(\mathcal{W})$ as the context distribution. We have

$$\begin{aligned}
&\sum_{\pi\in\Psi} Q_m(\pi)\widehat{Reg_t}(\pi) \\
=&\sum_{\pi\in\Psi} Q_m(\pi)\mathbb{E}_{w_t}\left[(\hat{f}_m(w_t,\pi_{\hat{f}_m}(w_t)) - \hat{f}_m(w_t,\pi(w_t)))\right] \\
=&\mathbb{E}_{w_t\sim\mathcal{P}(\mathcal{W})}\left[\sum_{a\in\mathcal{A}^*(w_t)}\sum_{\pi\in\Psi}\mathbf{I}\left\{\pi(w_t)=a\right\}Q_m(\pi)(\hat{f}_m(w_t,\pi_{\hat{f}_m}(w_t)) - \hat{f}_m(w_t,a))\right] \\
=&\mathbb{E}_{w_t\sim\mathcal{P}(\mathcal{W})}\left[\sum_{a\in\mathcal{A}^*(w_t)} p_m(a|w_t)(\hat{f}_m(w_t,\pi_{\hat{f}_m}(w_t)) - \hat{f}_m(w_t,a))\right].
\end{aligned}$$

The first and second equalities are the definitions of $\widehat{Reg_t}(\pi)$ and $Q_m(\pi)$, respectively.

Now for the context w_t, we have

$$\begin{aligned}
&\sum_{a\in\mathcal{A}^*(w_t)} p_m(a|w)(\hat{f}_m(w_t,\pi_{\hat{f}_m}(w_t)) - \hat{f}_m(w_t,a)) \\
=&\sum_{a\in\mathcal{A}^*(w_t)-\{\pi_{\hat{f}_m}(w_t)\}}\frac{\hat{f}_m(w_t,\pi_{\hat{f}_m}(w_t)) - \hat{f}_m(w_t,a)}{|\mathcal{A}^*(w_t)| + \gamma_t(\hat{f}_m(w_t,\pi_{\hat{f}_m}(w_t)) - \hat{f}_m(w_t,a))} \\
\leq&\ [|\mathcal{A}^*(w_t)|-1]/\gamma_t \\
\leq&\ \sqrt{|\mathcal{A}^*(w_t)|}/\rho_m.
\end{aligned}$$

We plug in the above term and apply the i.d.d. assumption on w_t to conclude the proof of the first inequality.

For the second inequality, we first observe that for any policy $\pi \in \Psi$, given any context $w \in \mathcal{W}$,

$$\frac{1}{p_m(\pi(w)|w)} = |\mathcal{A}^*(w)| + \gamma_t(\hat{f}_m(w,\pi_{\hat{f}_m}(w)) - \hat{f}_m(w,a)),$$

if $a \neq \pi_{\hat{f}_m}(w)$, and

$$\frac{1}{p_m(\pi(w)|w)} \leq \frac{1}{1/|\mathcal{A}^*(w)|} = |\mathcal{A}^*(w)| + \gamma_t(\hat{f}_m(w,\pi_{\hat{f}_m}(w)) - \hat{f}_m(w,a)),$$

if $a = \pi_{\hat{f}_m}(w)$. The result follows immediately by taking expectation over w. □

Compared with IOP in (Simchi-Levi and Xu, 2021), the key different part is that $\mathbb{E}_W[|\mathcal{A}^*(W)|]$ is replaced by the cardinality $|\mathcal{A}|$ of the whole action set. Another different part is the universal policy space Ψ. We define Ψ as $\prod_{w\in\mathcal{W}}\mathcal{A}^*(w)$ rather than $\prod_{w\in\mathcal{W}}\mathcal{A}$. These two points highlight the adaptivity to contexts and show how causal bound affects the action selection.

Define the following high-probability event

$$\Gamma = \left\{\forall m \geq 2, \frac{1}{\tau_{m-1}}\sum_{t=1}^{\tau_{m-1}}\mathbb{E}_{w_t,a_t}[(\hat{f}_{m(t)}(w_t,a_t) - f^*(w_t,a_t))^2|\mathcal{H}_{t-1}] \leq \frac{1}{\rho_m^2}\right\}.$$

The high-probability event and its variants have been proved in literatures (Foster et al., 2018; Simchi-Levi and Xu, 2021; Foster et al., 2020). Our result is slightly different from them as the whole function space is eliminated to $\mathcal{F}^*$. Since these results share the same form, it is straightforward to show Γ holds with probability at least $1-\delta/2$. This is the result of the union bound and the property of the **Least Square Oracle** that is independent of algorithm design.

Our setting do not change the proof procedure of the following lemma (Simchi-Levi and Xu, 2021), because this lemma does not explicitly involve the number of action set. This lemma bounds the prediction error between the true reward and the estimated reward.

Lemma B.3 *Assume Γ holds. For all epochs $m > 1$, all rounds t in epoch m, and all policies $\pi \in \Psi$, then*

$$\left|\widehat{\mathcal{R}}_t(\pi) - \mathcal{R}_t(\pi)\right| \leq \frac{1}{2\rho_m}\sqrt{\max_{1\leq m'\leq m-1} \mathbb{E}_W\left[\frac{1}{p_{m'}(\pi(W)|W)}\right]}.$$

The third step is to show that the one-step regret $Reg_t(\pi)$ is close to the one-step estimated regret $\widehat{Reg}_t(\pi)$. The following lemma states the result.

Lemma B.4 *Assume Γ holds. Let $c_0 = 5.15$. For all epochs m and all rounds t in epoch m, and all policies $\pi \in \Psi$,*

$$Reg(\pi) \leq 2\widehat{Reg}_t(\pi) + c_0\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}/\rho_m, \tag{12}$$

$$\widehat{Reg}_t(\pi) \leq 2Reg(\pi) + c_0\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}/\rho_m, . \tag{13}$$

Proof. We prove this lemma via induction on m. It is easy to check

$$Reg(\pi) \leq 1, \widehat{Reg}_t(\pi) \leq 1,$$

as $\gamma_1 = 1$ and $c_0\mathbb{E}_W\big[\mathcal{A}^*(W)\big] \geq 1$. Hence, the base case holds.

For the inductive step, fix some epoch $m > 1$ and assume that for all epochs $m' < m$, all rounds t' in epoch m', and all $\pi \in \Psi$, the inequalities (12) and (13) hold. We first show that for all rounds t in epoch m and all $\pi \in \Psi$,

$$Reg(\pi) \leq 2\widehat{Reg}_t(\pi) + c_0\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}/\rho_m.$$

We have

$$\begin{aligned}
&Reg(\pi) - \widehat{Reg}_t(\pi)\\
=&[\mathcal{R}(\pi_{f^*}) - \mathcal{R}(\pi)] - [\widehat{\mathcal{R}}_t(\pi_{\hat{f}_m}) - \widehat{\mathcal{R}}_t(\pi)]\\
\leq&[\mathcal{R}(\pi_{f^*}) - \mathcal{R}(\pi)] - [\widehat{\mathcal{R}}_t(\pi_{f^*}) - \widehat{\mathcal{R}}_t(\pi)]\\
\leq&|\mathcal{R}(\pi_{f^*}) - \widehat{\mathcal{R}}_t(\pi_{f^*})| + |\mathcal{R}(\pi) - \widehat{\mathcal{R}}_t(\pi)|\\
\leq&\frac{1}{\rho_m}\sqrt{\max_{1\leq m'\leq m-1} \mathbb{E}_W\left[\frac{1}{p_{m'}(\pi_{f^*}(W)|W)}\right]} + \frac{1}{\rho_m}\sqrt{\max_{1\leq m'\leq m-1} \mathbb{E}_W\left[\frac{1}{p_{m'}(\pi(W)|W)}\right]}\\
\leq&\frac{\max_{1\leq m'\leq m-1} \mathbb{E}_W\left[\frac{1}{p_{m'}(\pi_{f^*}(W)|W)}\right]}{5\rho_m\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}} + \frac{\max_{1\leq m'\leq m-1} \mathbb{E}_W\left[\frac{1}{p_{m'}(\pi(W)|W)}\right]}{5\rho_m\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}} + \frac{5\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}}{8\rho_m}.
\end{aligned}$$

The last inequality is by the AM-GM inequality. There exists an epoch i such that

$$\max_{1\leq m'\leq m-1} \mathbb{E}_W\left[\frac{1}{p_{m'}(\pi(W)|W)}\right] = \mathbb{E}_W\left[\frac{1}{p_i(\pi(W)|W)}\right].$$

From Lemma B.2 we know that

$$\mathbb{E}_W\left[\frac{1}{p_i(\pi(W)|W)}\right] \leq \mathbb{E}_W[\mathcal{A}^*(W)] + \mathbb{E}_W[\sqrt{|\mathcal{A}^*(W)|}]\rho_i\widehat{Reg}_t(\pi),$$

holds for all $\pi \in \Psi$, for all epoch $1 \leq i \leq m-1$ and for all rounds t in corresponding epochs.

Hence, for epoch i and all rounds t in this epoch, we have

$$\begin{aligned}
&\frac{\max\limits_{1\leq m'\leq m-1}\mathbb{E}_W\left[\frac{1}{p_{m'}(\pi(W)|W)}\right]}{5\rho_m\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}}\\
=&\frac{\mathbb{E}_W\left[\frac{1}{p_i(\pi_{f^*}(W)|W)}\right]}{5\rho_m\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}},\ \text{(Lemma B.2:(13))}\\
\leq&\frac{\mathbb{E}_W[\mathcal{A}^*(W)]+\mathbb{E}_W[\sqrt{|\mathcal{A}^*(W)|}]\rho_i\widehat{Reg}_t(\pi)}{5\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}\rho_m},\ \text{(inductive assumption)}\\
\leq&\frac{\mathbb{E}_W[\mathcal{A}^*(W)]+\mathbb{E}_W[\sqrt{|\mathcal{A}^*(W)|}]\rho_i[2Reg(\pi)+c_0\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}/\rho_i]}{5\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}\rho_m},\ \text{(Jensen's inequality)}\\
\leq&\frac{\mathbb{E}_W[\mathcal{A}^*(W)]+\sqrt{\mathbb{E}_W[|\mathcal{A}^*(W)|]}\rho_i[2Reg(\pi)+c_0\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}/\rho_i]}{5\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}\rho_m},\ (\rho_i\leq\rho_m\ \text{for}\ i\leq m)\\
\leq&\frac{2}{5}Reg(\pi)+\frac{1+c_0}{5\rho_m}\sqrt{\mathbb{E}_W[|\mathcal{A}^*(W)|]}.
\end{aligned}$$

We can bound $\frac{\max_{1\leq m'\leq m-1}\mathbb{E}_W\left[\frac{1}{p_{m'}(\pi(W)|W)}\right]}{5\rho_m\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}}$ in the same way.

Combing all above inequalities yields

$$\begin{aligned}
Reg(\pi)-\widehat{Reg}_t(\pi)\leq&\frac{2(1+c_0)\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}}{5\rho_m}+\frac{4}{5}\widehat{Reg}_t(\pi)+\frac{5\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}}{8\rho_m}\\
\leq&\widehat{Reg}_t(\pi)+(\frac{2(1+c_0)}{5}+\frac{5}{8})\frac{\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}}{\rho_m}\\
\leq&\widehat{Reg}_t(\pi)+c_0\frac{\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}}{\rho_m}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\widehat{Reg}_t(\pi)-Reg(\pi)\\
=&[\widehat{\mathcal{R}}_t(\pi_{\hat{f}_m})-\widehat{\mathcal{R}}_t(\pi)]-[\mathcal{R}(\pi_{f^*})-\mathcal{R}(\pi)]\\
\leq&[\widehat{\mathcal{R}}_t(\pi_{\hat{f}_m})-\widehat{\mathcal{R}}_t(\pi)]-[\mathcal{R}(\pi_{\hat{f}_m})-\mathcal{R}(\pi)]\\
\leq&|\mathcal{R}(\pi_{\hat{f}_m})-\widehat{\mathcal{R}}_t(\pi_{\hat{f}_m})|+|\mathcal{R}(\pi)-\widehat{\mathcal{R}}_t(\pi)|.
\end{aligned}$$

We can bound the above terms in the same steps.

□

Proof. Our regret analysis builds on the framework in (Simchi-Levi and Xu, 2021).

Step 1: proving an implicit optimization problem for Q_m in Lemma B.2.

Step 2: bounding the prediction error between $\widehat{\mathcal{R}}_t(\pi)$ and $\mathcal{R}_t(\pi)$ in Lemma B.3. Then we can show that the one-step regrets $\widehat{Reg}_t(\pi)$ and $Reg(\pi)$ are close to each other.

Step 3: bounding the cumulative regret $Reg(T)$.

By Lemma 4 of (Simchi-Levi and Xu, 2021),

$$\mathbb{E}[Reg(T)]=\sum_{t=1}^{T}\sum_{\pi\in\Psi}Q_{m(t)}(\pi)Reg(\pi).$$

From Lemma B.4, we know

$$Reg(\pi) \leq 2\widehat{Reg}_t(\pi) + c_0\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}/\rho_m$$

so

$$\begin{aligned}
\mathbb{E}[Reg(T)] =&\sum_{t=1}^{T}\sum_{\pi\in\Psi} Q_{m(t)}(\pi)Reg(\pi)\\
\leq&2\sum_{t=1}^{T}\sum_{\pi\in\Psi} Q_{m(t)}(\pi)\widehat{Reg}_t(\pi) + \sum_{t=1}^{T} c_0\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}/\rho_{m(t)}\\
\leq&(2+c_0)\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}\sum_{t=1}^{T}\frac{1}{\rho_{m(t)}}\\
\leq&(2+c_0)\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}\sum_{m=1}^{\lceil \log T\rceil}\sqrt{\log(2\delta^{-1}|\mathcal{F}^*|\log T)\tau_{m-1}/\eta}\\
\leq&(2+c_0)\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]}\sum_{m=1}^{\lceil \log T\rceil}\sqrt{\log(2\delta^{-1}|\mathcal{F}^*|\log T)\tau_{m-1}/\eta}\\
\leq&(2+c_0)\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]\log(2\delta^{-1}|\mathcal{F}^*|\log T)\sum_{m=1}^{\lceil \log T\rceil}\tau_{m-1}/\eta}\\
\leq&(2+c_0)\sqrt{\mathbb{E}_W[\mathcal{A}^*(W)]\log(2\delta^{-1}|\mathcal{F}^*|\log T)T/\eta}.
\end{aligned}$$

□

B.9 Proof of Theorem 3.2

Proof. We first consider $|\mathcal{W}| < \infty$. Since the agent have knowledge about causal bound, any function in $\mathcal{F} - \mathcal{F}^*$ can not be the true reward function. For any given context w, the set that the optimal arm will be in is $\mathcal{A}^*(w)$. For any algorithm A, let A_w be the induced algorithm of A when w occurs. Namely, the agent has access to a function space $\mathcal{F}_w = \{f(w,\cdot)|\forall f \in \mathcal{F}^*\}$ and an action set $\mathcal{A}^*(w)$.

From the minimax theorem 5.1 in (Agarwal et al., 2012), we know that there exists a contextual bandit instance such that the regret of A_w is at least $\sqrt{\mathcal{A}^*(w)T_w\log|\mathcal{F}_w|} = \sqrt{\mathcal{A}^*(w)T_w\log|\mathcal{F}^*|}$, where T_w is the number of occurrence of w. Hence,

$$\begin{aligned}
Reg(T) \geq& \sum_{w\in\mathcal{W}}\sqrt{|\mathcal{A}^*(w)|T_w\log|\mathcal{F}^*|}\\
\geq& \sqrt{\sum_{w\in\mathcal{W}}|\mathcal{A}^*(w)|T_w\log|\mathcal{F}^*|}.
\end{aligned}$$

and almost surely,

$$\begin{aligned}
\liminf_{T\to\infty}\frac{Reg(T)}{\sqrt{T}} \geq& \sqrt{\sum_{w\in\mathcal{W}}|\mathcal{A}^*(w)|\log|\mathcal{F}^*|\cdot\liminf_{T\to\infty}\frac{T_w}{T}}\\
=& \sqrt{\sum_{w\in\mathcal{W}}|\mathcal{A}^*(w)|\log|\mathcal{F}^*|\mathbb{P}(W=w)}\\
=& \sqrt{\mathbb{E}_W[|\mathcal{A}^*(W)|]\log|\mathcal{F}^*|}.
\end{aligned}$$

Now assume $|\mathcal{W}| = \infty$. Thanks to Glivenko-Cantelli theorem, the empirical distribution converges uniformly to the true reward distribution. We conclude the proof by applying the dominated convergence theorem and the Fubini's theorem, because $\mathcal{A}^*(w)$ is uniformly bounded by $|\mathcal{A}|$.

□

C RELATED MATERIALS

Definition C.1 (Back-Door Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph $\mathcal{G}$, a set of variables $\mathbf{Z}$ satisfies the back-door criterion relative to (X, Y), if no node in $\mathbf{Z}$ is a descendant of X, and $\mathbf{Z}$ blocks every path between X and Y that contains an arrow into X.*

Definition C.2 (d-separation) *In a causal diagram $\mathcal{G}$, a path $\mathcal{P}$ is blocked by a set of nodes $\mathbf{Z}$ if and only if*

1. *$\mathcal{P}$ contains a chain of nodes $A \leftarrow B \leftarrow C$ or a fork $A \rightarrow B \leftarrow C$ such that the middle node B is in $\mathbf{Z}$ (i.e., B is conditioned on), or*
2. *$\mathcal{P}$ contains a collider $A \leftarrow B \rightarrow C$ such that the collision node B is not in $\mathbf{Z}$, and no descendant of B is in $\mathbf{Z}$.*

If $\mathbf{Z}$ blocks every path between two nodes X and Y , then X and Y are d-separated conditional on $\mathbf{Z}$, and thus are independent conditional on $\mathbf{Z}$.

If X is a variable in a causal model, its corresponding intervention variable I_X is an exogenous variable with one arrow pointing into X. The range of I_X is the same as the range of X, with one additional value we can call "off". When I_X is off, the value of X is determined by its other parents in the causal model. When I_X takes any other value, X takes the same value as I_X, regardless of the value of X's other parents. If X is a set of variables, then I_X will be the set of corresponding intervention variables. We introduce the following do-calculus rules proposed in (Pearl, 2009).

Rule 1 (Insertion/deletion of observations)

$$\mathbb{P}(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z}, \mathbf{W}) = \mathbb{P}(\mathbf{Y}|do(\mathbf{X}), \mathbf{W})$$

if $\mathbf{Y}$ and $I_{\mathbf{Z}}$ are d-separated by $\mathbf{X} \cup \mathbf{W}$ in $\mathcal{G}^*$, the graph obtained from $\mathcal{G}$ by removing all arrows pointing into variables in $\mathbf{X}$.

Rule 2 (Action/observation exchange)

$$\mathbb{P}(\mathbf{Y}|do(\mathbf{X}), do(\mathbf{Z}), \mathbf{W}) = \mathbb{P}(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z}, \mathbf{W})$$

if $\mathbf{Y}$ and $I_{\mathbf{Z}}$ are d-separated by $\mathbf{X} \cup \mathbf{Z} \cup \mathbf{W}$ in $\mathcal{G}^\dagger$, the graph obtained from $\mathcal{G}$ by removing all arrows pointing into variables in $\mathbf{X}$ and all arrows pointing out of variables in $\mathbf{z}$.

Rule 3 (Insertion/deletion of actions)

$$\mathbb{P}(\mathbf{Y}|do(\mathbf{X}), do(\mathbf{Z}), \mathbf{W}) = \mathbb{P}(\mathbf{Y}|do(\mathbf{X}), \mathbf{W})$$

if $\mathbf{Y}$ and $I_{\mathbf{Z}}$ are d-separated by $\mathbf{X} \cup \mathbf{W}$ in $\mathcal{G}^*$, the graph obtained from $\mathcal{G}$ by removing all arrows pointing into variables in $\mathbf{X}$.

D SAMPLING ON GENERAL SIMPLEX

We generalize the sampling method for general simplex $\mathcal{D}$

$$A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}.$$

in Algorithm 7. Without loss of generality, we always assume $\mathcal{D}$ is not empty.

E CONCLUSIONS

In this paper, we investigate transfer learning in partially observable contextual bandits by converting the problem to identifying or partially identifying causal effects between actions and rewards. We derive causal bounds with the existence of partially observable confounders using our proposed Monte-Carlo algorithms. We formally prove and empirically demonstrate that our causally enhanced algorithms outperform classical bandit algorithms and achieve orders of magnitude faster convergence rates.

Algorithm 7 A sampling algorithm for the given simplex

```
Input: a simplex $\mathcal{D}$, a sampling distribution $F_s$ supported on $[0, 1]$
Denote the number of components of $\mathbf{x}$ as $d$
Solving the following LP
    $\min / \max x_1$
    $A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}.$
to find the bound $[l_1, h_1]$ for $x_1$
Sample a value $\hat{x}_1$ from the truncated $F_s$ supported on $[l_1, h_1]$
for $i = 2, \cdots, d$ do
    Solving the follow LP
        $\min / \max x_i$
        $A\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}$
        $x_j = \hat{x}_j, j = 1, \cdots, i - 1.$
    to find the bound $[l_i, h_i]$ for $x_1$
    Sample a value $\hat{x}_i$ from the truncated $F_s$ supported on $[l_i, h_i]$
Output: a valid sample $\hat{\mathbf{x}} = (\hat{x}_1, \cdots, \hat{x}_d) \in \mathcal{D}$
```

There are several future research directions we wish to explore. Firstly, we aim to investigate whether the solutions to discretization optimization converge to those of the original problem. While the approximation error can be exactly zero when all referred random variables are discrete, it is still unclear which conditions for general random variables and discretization methods can lead to convergence. We conjecture that this property may be related to the sensitivity of the non-linear optimization problem.

Lastly, we aim to extend our IGW-based algorithm to continuous action settings. IGW has been successfully applied to continuous action settings and has shown practical advantages in large action spaces. This extension may be related to complexity measures in machine learning.