Adaptively Exploiting d-Separators with Causal Bandits

Blair Bilodeau[*]
University of Toronto

Linbo Wang
University of Toronto

Daniel M. Roy
University of Toronto

Abstract

Multi-armed bandit problems provide a framework to identify the optimal intervention over a sequence of repeated experiments. Without additional assumptions, minimax optimal performance (measured by cumulative regret) is well-understood. With access to additional observed variables that d-separate the intervention from the outcome (i.e., they are a d-*separator*), recent "causal bandit" algorithms provably incur less regret. However, in practice it is desirable to be agnostic to whether observed variables are a d-separator. Ideally, an algorithm should be *adaptive*; that is, perform nearly as well as an algorithm with oracle knowledge of the presence or absence of a d-separator. In this work, we formalize and study this notion of adaptivity, and provide a novel algorithm that *simultaneously* achieves (a) optimal regret when a d-separator is observed, improving on classical minimax algorithms, and (b) significantly smaller regret than recent causal bandit algorithms when the observed variables are not a d-separator. Crucially, our algorithm does not require any oracle knowledge of whether a d-separator is observed. We also generalize this adaptivity to other conditions, such as the front-door criterion.

1 Introduction

Given a set of interventions (actions) for a specific experiment, we are interested in learning the best one with respect to some outcome of interest. Without knowledge of specific causal structure relating the observed variables, this task is impossible from solely observational data (c.f. Theorem 4.3.2 of [29]). Instead, we seek the most efficient way to sequentially choose interventions for i.i.d. repetitions of the experiment, where the main challenge is that we cannot observe the counterfactual effect of interventions we did not choose. Without any structural assumptions beyond i.i.d., one can always learn the best intervention with high confidence by performing each intervention a sufficient number of times [10]. In the presence of additional structure—such as a causal graph on observables—this strategy may result in performing suboptimal interventions unnecessarily often. However, the presence of such structure is often unverifiable, and incorrectly supposing that it exists may catastrophically mislead the experimenter. Thus, a fundamental question arises: Can we avoid strong, unverifiable assumptions while simultaneously performing fewer harmful interventions when advantageous structure exists?

A natural framework in which to study this question is that of (multi-armed) bandit problems: Over a sequence of interactions with the environment, the experimenter chooses an action using their experience of the previous interactions, and then observes the *reward* of the chosen action. The goal is to achieve comparable performance with what would have been achieved if the experimenter had chosen the (unknown) optimal action in each interaction. Formally, performance is measured by *regret*, which is the difference of the *cumulative reward* incurred by the experimenter compared to the optimal action. In this partial-information setting, regret induces the classical trade-off between *exploration* (choosing potentially suboptimal actions to learn if they're optimal) and *exploitation*

[*]Correspondence to blair.bilodeau@mail.utoronto.ca

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

(choosing the action that empirically appears the best). In contrast, other measures of performance (e.g., only identifying the average treatment effect or the best action at the end of all interactions) do not penalize the experimenter for performing suboptimal actions during exploration, and consequently are insufficient to study our question of interest.

For an *action set* $\mathcal{A}$ and a *time horizon* T, the minimax optimal regret for bandits without any assumptions on the data (worst-case) is $\tilde{\mathcal{O}}(\sqrt{|\mathcal{A}|\,T})$ [6], and is achieved by many algorithms [19]. Recently, Lu et al. [24] showed that, under additional causal structure, a new algorithm (C-UCB) can achieve improved regret. In particular, if the experimenter has access to a variable Z—taking values in a finite set $\mathcal{Z}$—that *d-separates* [29] the intervention and the reward, as well as the interventional distribution of Z for each $a \in \mathcal{A}$, C-UCB achieves $\tilde{\mathcal{O}}(\sqrt{|\mathcal{Z}|\,T})$ regret. However, as we show, the performance of C-UCB when the d-separation assumption fails is orders of magnitude worse than that of UCB. Is *strict adaptation* possible? That is, is there an algorithm that recovers the guarantee of C-UCB when Z is a d-separator and the guarantee of UCB in all other environments, without advance knowledge of whether Z is a d-separator?

As of yet, there is no general theory of adaptivity in the bandit setting. The closest we have to a general method is the Corral algorithm and its offspring [2, 4]. Corral uses online mirror descent to combine "base" bandit algorithms, but requires that each of the base algorithms is "stable" when operating on importance-weighted observations. Unfortunately, while UCB is stable, simulations reveal this not to be the case for C-UCB. This presents a barrier to adapting to causal structure via Corral-like techniques, and raises the question of whether there is a new way to achieve adaptivity.

Contributions. We introduce the *conditionally benign* property for bandit environments: informally, there exists a random variable Z such that the conditional distribution of the reward given Z is the same for each action $a \in \mathcal{A}$. We show that the conditionally benign property is (a) strictly weaker than the assumption of Lu et al. [24] (in their proofs, they actually assume that all causal parents of the reward are observed); (b) equivalent to Z being a d-separator when $\mathcal{A}$ is all interventions; and (c) implied by the front-door criterion [29] when $\mathcal{A}$ is all interventions except the *null intervention* (i.e., a pure observation). We then prove that any algorithm that achieves optimal worst-case regret must incur suboptimal regret for some conditionally benign environment, and hence strict adaptation to the conditionally benign property is impossible. Despite this, we introduce the hypothesis-tested adaptive causal upper confidence bound algorithm (HAC-UCB), which provably (a) achieves non-vacuous (sublinear in T) regret at all times without any assumptions, (b) recovers the improved performance of C-UCB for conditionally benign environments, and (c) performs as well as UCB in certain environments where C-UCB and related algorithms (such as those studied in [25, 27]) incur linear regret. Empirically, we observe these performance improvements on simulated data.

Impact. Recently, multiple works have developed causal bandit algorithms that achieve improved performance in the presence of advantageous causal relationships (initiated by Bareinboim et al. [7] and Lattimore et al. [17]; see Section 7 for more literature). Further, the last decade has seen a flurry of work in bandits on designing algorithms that recover worst-case regret bounds while simultaneously performing significantly better in advantageous settings, without requiring advance knowledge of which case holds [e.g., 9, 31, 34, 26, 1]. However, to the best of our knowledge, no existing work studies algorithms that achieve *adaptive regret guarantees with respect to causal structure*. The present work provides a framework that expands the study of adaptive decision making to the rich domain of causal inference.

2 Preliminaries

2.1 Problem Setting

We consider a general extension of the usual bandit setting where, in addition to a reward corresponding to the action played, the experimenter observes some additional variables *after* choosing their action; we call this the *post-action context*. This is distinct from the contextual bandit problem, where the experimenter has access to side-information *before* choosing their action.

Let $\mathcal{Y} = [0, 1]$ be the *reward space*[2], $\mathcal{Z}$ be a finite set of values for the post-action context to take, and $\mathscr{P}(\mathcal{Z} \times \mathcal{Y})$ denote the set of joint probability distributions. For any $p \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})$ and

[2]Our results hold for $\mathcal{Y} = \mathbb{R}$ using sub-Gaussian rewards with bounded mean at the expense of constants.

$(\mathcal{Z}, \mathcal{Y})$-valued random variable (Z, Y), let $p(Z)$ and $p(Y \mid Z)$ denote the marginal and conditional distributions respectively. Let $\mathbb{E}_p$ and $\mathbb{P}_p$ denote expectation and probability operators under p.

The *stochastic bandit problem with post-action contexts* proceeds as follows: For each round $t \in [T]$, the experimenter selects $A_t \in \mathcal{A}$ while simultaneously $\{(Z_t(a), Y_t(a)) : a \in \mathcal{A}\}$ are independently sampled from the *environment*, which is any family of distributions $\nu = \{\nu_a : a \in \mathcal{A}\} \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ indexed by the action set. The experimenter only observes $(Z_t(A_t), Y_t(A_t))$ and receives reward $Y_t(A_t)$. From a causal perspective, $(Z_t(a), Y_t(a))_{a \in \mathcal{A}}$ corresponds to the potential outcome vector, and under causal consistency, $(Z_t(A_t), Y_t(A_t))$ corresponds to the observed data (Z_t, Y_t) under the chosen intervention A_t.

The *observed history up to round* t is the random variable $H_t = (A_s, Z_s(A_s), Y_s(A_s))_{s \in [t]}$. A *policy* is a sequence of measurable maps from the observed history to the action set, denoted

$$\pi = (\pi_t)_{t \in [T]} \in \Pi(\mathcal{A}, \mathcal{Z}, T) := \prod_{t=1}^{T} \Big\{ (\mathcal{A} \times \mathcal{Z} \times \mathcal{Y})^{t-1} \to \mathcal{A} \Big\}.$$

The experimenter chooses a policy in advance to select their action on each round according to $A_t = \pi_t(H_{t-1})$. Clearly, an environment ν and a policy π together define a joint probability distribution on $(A_t, Z_t(a), Y_t(a))_{t \in [T], a \in \mathcal{A}}$ (which includes the "counterfactuals" not seen by the player). Let $\mathbb{E}_{\nu,\pi}$ denote expectation under this joint distribution. The performance of a policy under an environment is quantified by the *regret*

$$R_{\nu,\pi}(T) = T \cdot \max_{a \in \mathcal{A}} \mathbb{E}_{\nu_a}[Y] - \mathbb{E}_{\nu,\pi} \sum_{t=1}^{T} \mathbb{E}_{\nu_{A_t}}[Y].$$

2.2 Specific Algorithms

Classical bandit algorithms take $|\mathcal{A}|$ and T as inputs. The dependence on T can often be dropped using the doubling trick or more sophisticated techniques such as decreasing learning rates, but we do not focus on these refinements in this work, instead allowing T as an input. In order to account for the additional information provided by the post-action context, we also consider algorithms that take $|\mathcal{Z}|$ as an input. By restricting dependence to only the cardinality of $\mathcal{A}$ and $\mathcal{Z}$, we explicitly suppose that there is no additional structure to exploit on these spaces; much work in the bandit literature has focused on such structure through linear or Lipschitz rewards, but we defer these extensions to future work in favour of focusing on adaptivity. For notational simplicity, we denote algorithmic dependence on $\mathcal{A}$ or $\mathcal{Z}$ even though the dependence is actually through their cardinality (i.e., the labellings of items in the sets are arbitrary).

In the causal bandit literature [17, 24, 27], it is common to suppose that the algorithm also receives distributional information relating actions to intermediate variables. In particular, if the (unknown) environment is ν, prior work supposes that the algorithm has access to the (interventional) marginal distributions $\nu(Z) = \{\nu_a(Z) : a \in \mathcal{A}\}$. In this work, we suppose instead that the algorithm has access to a collection of *approximate* marginal distributions $\tilde{\nu}(Z) = \{\tilde{\nu}_a(Z) : a \in \mathcal{A}\}$; for example, these could be an estimate of $\nu(Z)$ that was learned offline. Ideally, $\tilde{\nu}(Z)$ will be close to $\nu(Z)$, but our novel method is entirely adaptive to this assumption: regardless of how well $\tilde{\nu}(Z)$ approximates $\nu(Z)$, HAC-UCB incurs sublinear regret.

We now introduce additional notation to define the algorithms of interest in this work. Suppose that $\mathcal{A}$, $\mathcal{Z}$, $\tilde{\nu}(Z)$, and T are all fixed in advance, as well as a confidence parameter $\delta = \delta_T \in (0, 1)$. For each $t \in [T]$, $z \in \mathcal{Z}$, and $a \in \mathcal{A}$, define the number of the first t rounds on which z was observed by $\mathbb{T}_t^{\mathcal{Z}}(z) = 1 \vee \sum_{s=1}^{t} \mathbb{I}\{Z_s(A_s) = z\}$ (where $a \vee b = \max\{a, b\}$), and similarly the number of rounds on which a was chosen by $\mathbb{T}_t^{\mathcal{A}}(a) = 1 \vee \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}$. Further, define the empirical mean estimate for the reward under the distribution induced by choosing the action a as $\hat{\mu}_t^{\mathcal{A}}(a) = [\mathbb{T}_t^{\mathcal{A}}(a)]^{-1} \sum_{s=1}^{t} Y_s(A_s)\mathbb{I}\{A_s = a\}$ and the empirical conditional mean estimate for the reward given that z was observed as $\hat{\mu}_t^{\mathcal{Z}}(z) = [\mathbb{T}_t^{\mathcal{Z}}(z)]^{-1} \sum_{s=1}^{t} Y_s(A_s)\mathbb{I}\{Z_s(A_s) = z\}$. Define $\mathrm{UCB}_t^{\mathcal{A}}(a) = \hat{\mu}_t^{\mathcal{A}}(a) + \sqrt{\log(2/\delta)/(2\mathbb{T}_t^{\mathcal{A}}(a))}$, $\mathrm{UCB}_t^{\mathcal{Z}}(z) = \hat{\mu}_t^{\mathcal{Z}}(z) + \sqrt{\log(2/\delta)/(2\mathbb{T}_t^{\mathcal{Z}}(z))}$, and $\widetilde{\mathrm{UCB}}_t(a) = \sum_{z \in \mathcal{Z}} \mathrm{UCB}_t^{\mathcal{Z}}(z)\mathbb{P}_{\tilde{\nu}_a}[Z = z]$.

Using these objects, we define three algorithms, each of which produces actions that are H_t-measurable. The upper confidence bound algorithm (UCB, [6]) is defined by $A_{t+1}^{\mathrm{UCB}} =$

$\arg\max_{a\in\mathcal{A}} \mathrm{UCB}_t^{\mathcal{A}}(a)$, and the causal upper confidence bound algorithm (C-UCB, [24]) is defined by $A_{t+1}^{\mathrm{C}} = \arg\max_{a\in\mathcal{A}} \widetilde{\mathrm{UCB}}_t(a)$, where ties are broken by using some predetermined ordering on $\mathcal{A}$. Finally, we define a new combination of these two methods, which we call the hypothesis-tested adaptive causal upper confidence bound algorithm (HAC-UCB) and describe precisely in Algorithm 1; we denote its actions by A_{t+1}^{HAC}.

Algorithm 1: HAC-UCB($\mathcal{A}$, $\mathcal{Z}$, T, $\tilde{\nu}(Z)$)

---

do Play each $a \in \mathcal{A}$ for $\lceil 4\sqrt{T} \,/\, |\mathcal{A}| \,\rceil$ rounds, and let $\hat{\nu}_a(Z)$ be the MLE of $\nu_a(Z)$
if $\sup_{a\in\mathcal{A}} \sum_{z\in\mathcal{Z}} \Big| \mathbb{P}_{\tilde{\nu}_a}[Z=z] - \mathbb{P}_{\hat{\nu}_a}[Z=z] \Big| > 2T^{-1/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T}$
  replace $\tilde{\nu}(Z) \longleftarrow \hat{\nu}(Z)$
do Play each $a \in \mathcal{A}$ for $\lceil \sqrt{T} \,/\, |\mathcal{A}| \,\rceil$ rounds
set flag = True
while $t \leq T$
  if flag
    `/* Check if either of the two conditions fail */`
    set $\mathrm{D}_{t-1}^{\mathcal{A}}(a) = \mathrm{UCB}_{t-1}^{\mathcal{A}}(a) - \widetilde{\mathrm{UCB}}_{t-1}(a) + \frac{\sqrt{|\mathcal{A}||\mathcal{Z}|\log T}}{T^{1/4}}$
    for $a \in \mathcal{A}$ **do**
      if not $-2\sum_{z\in\mathcal{Z}} \sqrt{\frac{\log T}{\mathbb{T}_{t-1}^{\mathcal{Z}}(z)}}\, \mathbb{P}_{\tilde{\nu}_a}[Z=z] \leq \mathrm{D}_{t-1}^{\mathcal{A}}(a) \leq 2\sqrt{\frac{\log T}{\mathbb{T}_{t-1}^{\mathcal{A}}(a)}} + 2\frac{\sqrt{|\mathcal{A}||\mathcal{Z}|\log T}}{T^{1/4}}$
        set flag = False; **break**
    `/* If conditions pass, play C-UCB */`
    if flag
      set $A_t^{\mathrm{HAC}} = A_t^{\mathrm{C}}$;
    else
      set $A_t^{\mathrm{HAC}} = A_t^{\mathrm{UCB}}$;
  else
    `/* If conditions ever fail, play UCB forever */`
    set $A_t^{\mathrm{HAC}} = A_t^{\mathrm{UCB}}$;

Heuristically, HAC-UCB has an initial exploration period to ensure that $\tilde{\nu}(Z)$ is sufficiently accurate—if not, it is replaced with the maximum likelihood estimate (MLE) of the marginals—and then optimistically plays C-UCB until there is sufficient evidence that the environment is not conditionally benign. The switch from C-UCB to UCB is decided by a hypothesis test performed on each round, which uses the confidence intervals that will hold if the environment is conditionally benign. When the arm mean estimates of UCB and C-UCB disagree, this provides evidence that the environment is not conditionally benign, and the evidence is considered sufficient to switch when the size of the disagreement is large compared to the size of the confidence intervals themselves. As we illustrate in the proof of the regret bounds, with high probability this test will not induce a switch for a conditionally benign environment, and will sufficiently limit the regret incurred by C-UCB if the environment is not conditionally benign.

3 Conditionally Benign Property

We now formalize the main property that HAC-UCB will adaptively exploit.

Definition 3.1. *An environment* $\nu \in \mathscr{P}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}$ is conditionally benign *if and only if there exists* $p \in \mathscr{P}(\mathcal{Z}\times\mathcal{Y})$ *such that for each* $a \in \mathcal{A}$, $\nu_a(Z) \ll p(Z)$ *and* $\nu_a(Y \mid Z) = p(Y \mid Z)$ *p-a.s.*

This definition does not require any causal terminology to define or use for regret bounds, but we now instantiate it for the causal setting. For a collection of finite random variables $\mathbf{V}$ and a (potentially) continuous random variable Y, let $\mathscr{P}_{\mathbf{V}}$ be the set of all joint probability distributions with strictly positive marginal probabilities on $\mathbf{V}$. Fix a DAG $\mathcal{G}$ on $(\mathbf{V}, Y)$ such that Y is a leaf and two disjoint sets $\mathbf{Z} \subseteq \mathbf{V}$ and $\mathbf{A} \subseteq \mathbf{V}$ such that $\mathrm{Pa}_{\mathbf{A}}^{\mathcal{G}} \subseteq \mathbf{V} \setminus \mathbf{Z}$. Let $\mathcal{A}$ be the set of all possible do interventions on $\mathbf{A}$, and for each $a \in \mathcal{A}$ let p_a denote the interventional distribution (Definition A.5). This structure suggests a graphical analogue of the conditionally benign property.

Definition 3.2. *For any DAG $\mathcal{G}$ and $\mathcal{A}' \subseteq \mathcal{A}$,* $(\mathcal{G}, \mathcal{A}')$ is conditionally benign *if and only if for all $p \in \mathscr{P}_{\mathbf{V}}$ that are Markov relative to $\mathcal{G}$, $\{p_a(\mathbf{Z}, Y) : a \in \mathcal{A}'\}$ is conditionally benign.*

We now connect the conditionally benign property to d-separation (Definition A.6) and the front-door criterion (Definition A.8). All proofs are deferred to Appendix A, along with standard notation and definitions from the causal literature.

Theorem 3.3. $\mathbf{Z}$ *d-separates Y from $\mathbf{A}$ on $\mathcal{G}$ if and only if $(\mathcal{G}, \mathcal{A})$ is conditionally benign.*

This equivalence is a strict specialization of the conditionally benign property to the causal setting. In particular, to define conditionally benign, we need not require all possible interventions be allowed. Define $\mathcal{A}_0$ to be $\mathcal{A}$ with the null (observational) intervention removed, and let $\mathcal{G}_{\bar{\mathbf{A}}}$ denote $\mathcal{G}$ with all edges directed into $\mathbf{A}$ removed.

Theorem 3.4. $\mathbf{Z}$ *d-separates Y from $\mathbf{A}$ on $\mathcal{G}_{\bar{\mathbf{A}}}$ if and only if $(\mathcal{G}, \mathcal{A}_0)$ is conditionally benign.*

The benefits of discarding the null intervention are demonstrated by the following fact.

Lemma 3.5. $\mathbf{Z}$ *d-separates Y from $\mathbf{A}$ on $\mathcal{G}_{\bar{\mathbf{A}}}$ if $\mathbf{Z}$ satisfies the front-door criterion for $(\mathbf{A}, Y)$ on $\mathcal{G}$.*

We visualize the preceding results in Figure 1. In graph (a), Z d-separates the intervention from the reward, and hence any Markov relative distribution (Definition A.3) will induce a conditionally benign environment. Graph (b) corresponds to a setting where one cannot hope to always improve performance due to the direct effect of the intervention on the reward, and consequently the environment need not be conditionally benign. In graph (c), the presence of the unobserved confounder U means that Z does not d-separate the intervention from the reward. However, if the null intervention is not considered, the arrow from U to A is never applicable, and hence any Markov relative distribution on the modified DAG will induce a conditionally benign environment. Specifically, graph (c) satisfies the front-door criterion, revealing that the conditionally benign property captures that this setting is still benign for decision-making, even though the conditions assumed by Lu et al. [24] do not hold. Finally, in graph (d), the unobserved confounder U once again violates d-separation, but also the front-door criterion is not satisfied because of the back-door path from Z to Y. Hence, even discarding the null intervention does not guarantee that the environment will be conditionally benign.

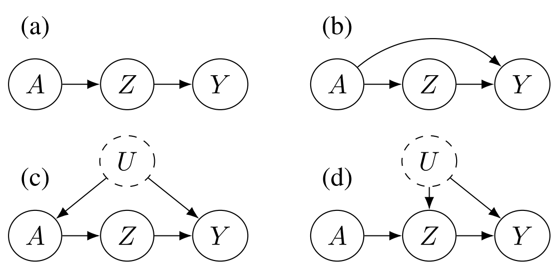

Figure 1: DAGs to illustrate the conditionally benign property. A is the intervention, Z is the post-action context, Y is the reward, and U is an unobserved variable. $(\mathcal{G}, \mathcal{A})$ is conditionally benign for (a) but only $(\mathcal{G}, \mathcal{A}_0)$ is conditionally benign for (c). For (b) and (d) the environment need not be conditionally benign.

4 Analysis of Bandit Algorithms

We now study the impact of the conditionally benign property on regret. All proofs are deferred to Appendix B. First, recall the standard regret bound for UCB, with constants tuned to rewards in $[0, 1]$.

Theorem 4.1 (Theorem 7.2 of [19])**.** *For all $\mathcal{A}$, $\mathcal{Z}$, T, and $\nu \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$, if $\delta = 2/T^2$*

$$R_{\nu,\text{UCB}}(T) \leq 2\,|\mathcal{A}| + 4\sqrt{2\,|\mathcal{A}|\,T \log T}.$$

Second, we generalize the main result of Lu et al. [24] by relaxing two assumptions: we only require that the environment is conditionally benign, and we allow for approximate marginal distributions using the following definition. Later, this will enable us to trade-off approximation error of $\tilde{\nu}(Z)$ with online estimation of $\nu(Z)$ in order to ensure that HAC-UCB *always* incurs sublinear regret.

Definition 4.2. *For any $\varepsilon \geq 0$, $\tilde{\nu}(Z)$ and $\nu(Z)$ are ε-*close *if*

$$\sup_{a\in\mathcal{A}} \sum_{z\in\mathcal{Z}} \Big| \mathbb{P}_{\tilde{\nu}_a}[Z=z] - \mathbb{P}_{\nu_a}[Z=z] \Big| \leq \varepsilon.$$

Theorem 4.3 (Refined Theorem 1 of [24])**.** *For all $\varepsilon > 0$, $\mathcal{A}$, $\mathcal{Z}$, T, and conditionally benign $\nu \in \mathscr{P}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}$, if $\tilde{\nu}(Z)$ and $\nu(Z)$ are ε-close and $\delta = 2/T^2$ then*

$$R_{\nu,\mathrm{C}}(T) \leq 2\,|\mathcal{Z}| + 6\sqrt{|\mathcal{Z}|\,T\log T} + (\log T)\sqrt{2T} + 2\varepsilon(1+\sqrt{\log T})T.$$

Remark 4.4. *Lu et al. [24] assume that $\varepsilon = 0$. Our result implies that an approximation error of $\varepsilon = \sqrt{|\mathcal{Z}|\,/T}$ is sufficient to achieve the optimal rate.* ◁

Next, we motivate our introduction of a new algorithm by showing that C-UCB can catastrophically fail when the environment is *not* conditionally benign, incurring regret that is linear in T (which is as bad as possible for bounded rewards).

Theorem 4.5. *For every $\mathcal{A}$ and $\mathcal{Z}$ with $|\mathcal{A}| \geq 2$, there exists $\nu \in \mathscr{P}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}$ such that even if $\tilde{\nu}(Z) = \nu(Z)$, for all possible settings of the confidence parameters, δ_T,*

$$\lim_{T\to\infty} \frac{R_{\nu,\mathrm{C}}(T)}{T} \geq 1/120.$$

Remark 4.6. *This lower bound is specifically for C-UCB. However, any algorithm that relies on eliminating actions from consideration via assumption rather than data is susceptible to such an issue. In particular, this result is easily modified to apply for C-TS [24] and C-UCB-2 [27]. Other causal algorithms (e.g., Parallel Bandit [17]) also intuitively suffer from the issue that our construction exploits, although a different argument must be made since these rely on different causal structure.* ◁

We now state our regret upper bound for our new algorithm, HAC-UCB. In Theorem 6.2, we will show it is *impossible* to always achieve the optimal regret without knowledge of whether a d-separator is observed, but the following theorem shows *some* adaptivity is always possible. Crucially, HAC-UCB achieves sublinear regret without any assumptions on ν or $\tilde{\nu}(Z)$. For a more detailed breakdown of the constants, see Eq. (B.9).

Theorem 4.7 (Main Result)**.** *For all $\mathcal{A}$, $\mathcal{Z}$, $T \geq 25\,|\mathcal{A}|^2$, $\nu \in \mathscr{P}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}$, and $\tilde{\nu}(Z) \in \mathscr{P}(\mathcal{Z})^{\mathcal{A}}$,*

$$R_{\nu,\mathrm{HAC}}(T) \leq 4\,|\mathcal{A}| + 11\,T^{3/4}(\log T)\sqrt{|\mathcal{A}|\,|\mathcal{Z}|} + 15\sqrt{(|\mathcal{A}|+|\mathcal{Z}|)T\log T} + 5(\log T)\sqrt{T}.$$

For all $\varepsilon \leq T^{-1/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T}$, if ν is conditionally benign and $\tilde{\nu}(Z)$ and $\nu(Z)$ are ε-close then

$$R_{\nu,\mathrm{HAC}}(T) \leq 4\,|\mathcal{A}| + 2\,|\mathcal{Z}| + 6\sqrt{|\mathcal{Z}|\,T\log T} + 4(\log T)\sqrt{T} + 2\varepsilon(1+\sqrt{\log T})T.$$

It is an open problem whether the dependence on $T^{3/4}$ is tight. In Theorem 6.2, we will show that it is impossible to obtain worst-case regret of size $\sqrt{|\mathcal{A}|\,T}$ while still achieving improved regret on conditionally benign environments. However, it may be possible to improve the dependence on T, and the role of logarithmic factors in how much improvement is possible remains to be understood. Towards improving this result, we now show that there exists an environment that forces C-UCB to incur linear regret yet HAC-UCB will switch to following UCB (and hence incurs $\sqrt{|\mathcal{A}|\,T\log T}$ regret at worst). That is, HAC-UCB recovers the improved performance of C-UCB when the conditionally benign property holds, is never worse than C-UCB, and optimally outperforms C-UCB in some settings.

Theorem 4.8. *There exists a constant C such that for any $\mathcal{A}$ and $\mathcal{Z}$ with $|\mathcal{A}| \geq 2$, there exists $\nu \in \mathscr{P}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}$ so that for any δ_T used for C-UCB with $\tilde{\nu}(Z) = \nu(Z)$,*

$$\lim_{T\to\infty} \frac{R_{\nu,\mathrm{C}}(T)}{T} \geq 1/C,$$

and if $\tilde{\nu}(Z) = \nu(Z)$ is used for HAC-UCB then

$$\lim_{T\to\infty} \frac{R_{\nu,\mathrm{HAC}}(T)}{|\mathcal{A}| + |\mathcal{Z}| + \sqrt{|\mathcal{A}|\,T\log T} + (\log T)\sqrt{T}} \leq C.$$

Remark 4.9. *Theorem 4.8 could be stated with $\tilde{\nu}(Z)$ only ε-close to $\nu(Z)$, but for simplicity we have supposed $\varepsilon = 0$ to highlight the role of the conditionally benign property.* ◁

5 Simulations

We now study the empirical performance of these algorithms in two key settings, corresponding to a conditionally benign environment and the lower bound environment from Theorem 4.5. We compare our algorithm HAC-UCB with UCB [5], C-UCB [24], C-UCB-2 [27], and Corral (for which we use the learning rate and algorithm prescribed by [2] with the epoch-based, scale-sensitive UCB prescribed by [4]); for all algorithms, we use the parameters that are optimal as prescribed by existing theory. To focus solely on the impact of the conditionally benign property, we set $\tilde{\nu}(Z) = \nu(Z)$. The results of this section are a representative simulation demonstrating empirically that (a) for worst-case environments, both C-UCB and C-UCB-2 incur linear regret, while HAC-UCB successfully switches to incur sublinear regret to compete with Corral and UCB, and (b) for conditionally benign environments, HAC-UCB and C-UCB enjoy improved performance compared to UCB, Corral, and C-UCB-2, all three of which have regret growing like $\sqrt{|\mathcal{A}|\,T}$. Implementation details are available in Appendix D and code can be found at https://github.com/blairbilodeau/adaptive-causal-bandits.

5.1 Conditionally Benign Environment

First, we consider a conditionally benign environment. Taking the gap $\Delta = \sqrt{|\mathcal{A}|\,(\log T)/T}$, the fixed conditional distribution for $\mathcal{Z} = \{0, 1\}$ is $Y \mid Z \sim \mathrm{Ber}(1/2 + (1 - Z)\Delta)$. Then, for a small ε (we take $\varepsilon = 0.0005$), we set $\mathbb{P}_{\nu_1}[Z = 0] = 1 - \varepsilon$ and $\mathbb{P}_{\nu_a}[Z = 0] = \varepsilon$ for all other $a \in \mathcal{A} \setminus \{1\}$. Thus, $a^*_\nu = 1$, and the actions are separated by Δ. In summary, each $z \in \mathcal{Z}$ has positive probability of being observed, yet each action nearly deterministically fixes Z.

In Figure 2 (left panel) we observe three main effects: (a) C-UCB and HAC-UCB perform similarly (their regret curves overlap), both achieving much smaller regret that remains unchanged by increasing $|\mathcal{A}|$, (b) UCB grows at the worst-case rate of roughly $\sqrt{|\mathcal{A}|\,T \log T}$, not taking advantage of the conditionally benign property, and (c) neither Corral nor C-UCB-2 realize the benefits of the conditionally benign property, since the regret increases with $|\mathcal{A}|$ and empirically they perform worse than UCB. We note that the x-axis starts at $T = 500$ to satisfy the minor requirement of $T > |\mathcal{A}|^2$.

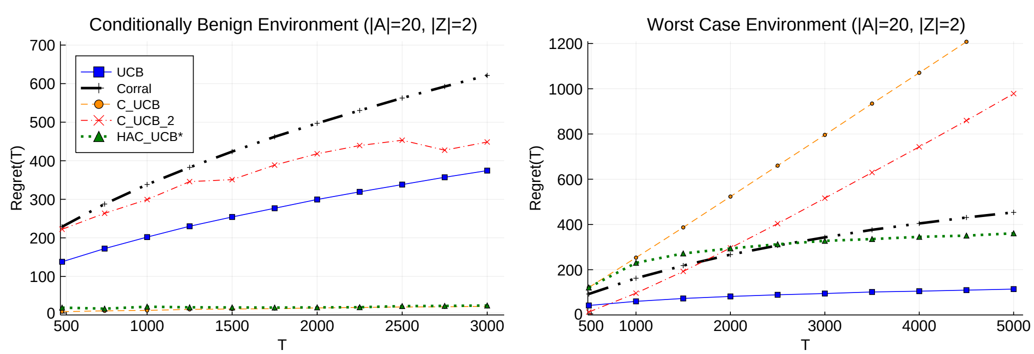

Figure 2: Regret when the conditionally benign property holds (left) and when it fails (right).

5.2 Worst Case Environment

Second, we consider an environment that is *not* conditionally benign. We use the same general environment from our lower bound in Theorem 4.5, where the causal algorithms learn a biased estimate of which $z \in \{0, 1\}$ has a higher conditional mean since Z is not a d-separator, and consequently they concentrate onto a bad action.

In Figure 2 (right panel), we again observe three main effects: (a) both C-UCB and C-UCB-2 incur linear regret, as prescribed by Theorem 4.5, (b) HAC-UCB achieves sublinear regret, although it is worse than UCB (we show in Theorem 6.2 that optimal adaptivity is impossible), and (c) Corral does not suffer linear regret, but appears to still do worse than HAC-UCB for large T.

6 Adaptivity for Causal Structures

Thus far we have analyzed the regret of various algorithms in two cases: environments that either are or are not conditionally benign. Minimax algorithms (UCB) fail to achieve smaller regret for conditionally benign environments, while causal algorithms (C-UCB) fail catastrophically in environments that are not conditionally benign. In this section, we formalize the notion of strict adaptivity (*adaptive minimax optimality*) with respect to the conditionally benign property, show that it is *impossible* to be adaptively minimax optimal with respect to the conditionally benign property, and discuss a relaxed notion of optimal adaptivity based on Pareto optimality.

6.1 Generic Algorithms

In order to describe adaptivity (and its impossibility) in the stochastic bandit problem with post-action contexts, we require a higher level of abstraction than a policy, which we achieve with *algorithms*. It is possible to define algorithms and the corresponding notion of adaptivity in abstract generality. However, we take the same perspective as Bilodeau et al. [8] (who define adaptive minimax optimality with respect to relaxations of the i.i.d. assumption), and sacrifice some generality by defining algorithms using the specific objects we study in this work. Formally, an algorithm is any map from the problem-specific inputs to the space of compatible policies, denoted by

$$\mathfrak{a} : (\mathcal{A}, \mathcal{Z}, T, \tilde{\nu}(Z)) \mapsto \mathfrak{a}(\mathcal{A}, \mathcal{Z}, T, \tilde{\nu}(Z)) \in \Pi(\mathcal{A}, \mathcal{Z}, T).$$

We denote the set of all algorithms by $\mathbb{A}_{\text{C-MAB}}$, and the subset of algorithms that are constant in $(\mathcal{Z}, \tilde{\nu}(Z))$ by $\mathbb{A}_{\text{MAB}}$; this subset contains the classical bandit algorithms that are agnostic to knowledge of post-action contexts, or more specifically, do not exploit causal structure.

6.2 Adaptive Minimax Optimality

Let $\mathfrak{p} : \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}} \mapsto \{0, 1\}$ encode whether a given environment satisfies a certain property of interest; in this work, it is always an indicator for whether ν is conditionally benign. Further, for every $q \in \mathscr{P}(\mathcal{Z})$, denote the set of all environments with marginal q by $\Pi_{\mathcal{A},\mathcal{Z}}(q) = \{\nu \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}} : \nu(Z) = q\}$. There are multiple ways one could define optimal adaptivity: we propose the following notion of strict adaptivity, which requires that the experimenter to do as well as they possibly could have if they had access to $\mathfrak{p}(\nu)$ in advance, but without this knowledge.

Definition 6.1. *An algorithm $\mathfrak{a} \in \mathbb{A}_{\text{C-MAB}}$ is* adaptively minimax optimal *with respect to $\mathfrak{p}$ if and only if there exists $C > 0$ such that for all $\mathcal{A}$, $\mathcal{Z}$, $q \in \mathscr{P}(\mathcal{Z})^{\mathcal{A}}$, and T,*

$$\sup_{\nu \in \Pi_{\mathcal{A},\mathcal{Z}}(q)} R_{\nu,\mathfrak{a}(\mathcal{A},\mathcal{Z},q,T)}(T) \le C \inf_{\pi \in \Pi(\mathcal{A},\mathcal{Z},T)} \sup_{\nu \in \Pi_{\mathcal{A},\mathcal{Z}}(q)} R_{\nu,\pi}(T) \tag{6.1}$$

and

$$\sup_{\nu \in \Pi_{\mathcal{A},\mathcal{Z}}(q),\, \mathfrak{p}(\nu)=1} R_{\nu,\mathfrak{a}(\mathcal{A},\mathcal{Z},q,T)}(T) \le C \inf_{\pi \in \Pi(\mathcal{A},\mathcal{Z},T)} \sup_{\nu \in \Pi_{\mathcal{A},\mathcal{Z}}(q),\, \mathfrak{p}(\nu)=1} R_{\nu,\pi}(T). \tag{6.2}$$

6.3 Impossibility of Strict Adaptivity

We now show it is *impossible* for any algorithm to always realize the benefits of the conditionally benign property while also recovering the worst-case rate of $\sqrt{|\mathcal{A}|\, T}$ (e.g., Theorems 9.1 and 15.2 of [19]), even when the algorithm has access to the true marginals. Our proof strategy is a modification of the finite-time lower bounds from Section 16.2 of Lattimore and Szepesvári [19]. Notably, the lower bounds of Lu et al. [24] already imply that any algorithm that does not take advantage of causal structure cannot be adaptively minimax optimal. We prove a significantly stronger result: even algorithms that use $\mathcal{Z}$ and $\tilde{\nu}(Z) = \nu(Z)$ cannot be adaptively minimax optimal!

Theorem 6.2. *Let $\mathfrak{a} \in \mathbb{A}_{\text{C-MAB}}$ be such that there exists $C > 0$ such that for all $\mathcal{A}$, $\mathcal{Z}$, and T,*

$$\sup_{\nu \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}} R_{\nu,\mathfrak{a}(\mathcal{A},\mathcal{Z},\nu(Z),T)}(T) \le C \sqrt{|\mathcal{A}|\, T}.$$

There exists a constant C' such that for all $\mathcal{A}$, $\mathcal{Z}$, and $T \ge |\mathcal{A}|$, there exists conditionally benign $\nu \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ with

$$R_{\nu,\mathfrak{a}(\mathcal{A},\mathcal{Z},\nu(Z),T)}(T) \ge C' \sqrt{|\mathcal{A}|\, T}.$$

6.4 Pareto-Adaptive Minimax Optimality

In light of this impossibility result, it is of interest to characterize relaxations of strict adaptivity that are achievable. Koolen [15] and Lattimore [18] introduce the *Pareto frontier* of regret, which when applied to the conditionally benign property, is all tuples of regret guarantees such that improving the regret in conditionally benign environments would necessarily force the worst-case regret to increase. We propose that it is desirable for an algorithm to do as well as possible in the worst-case, subject to always realizing smaller regret on conditionally benign environments. Formally, let $\mathbb{A}^{\star}_{\text{C-MAB}}$ be the subset of $\mathfrak{a} \in \mathbb{A}_{\text{C-MAB}}$ that satisfy Eq. (6.2) for some constant C.

Definition 6.3. *An algorithm* $\mathfrak{a}^{\star} \in \mathbb{A}^{\star}_{\text{C-MAB}}$ *is* Pareto-adaptively minimax optimal *with respect to* $\mathfrak{p}$ *if and only if there exists* $C > 0$ *such that for all* $\mathfrak{a} \in \mathbb{A}^{\star}_{\text{C-MAB}}$*,* $\mathcal{A}$*,* $\mathcal{Z}$*,* $q \in \mathscr{P}(\mathcal{Z})^{\mathcal{A}}$*, and* T*,*

$$\sup_{\nu \in \Pi_{\mathcal{A},\mathcal{Z}}(q)} R_{\nu,\mathfrak{a}^{\star}(\mathcal{A},\mathcal{Z},q,T)}(T) \leq C \sup_{\nu \in \Pi_{\mathcal{A},\mathcal{Z}}(q)} R_{\nu,\mathfrak{a}(\mathcal{A},\mathcal{Z},q,T)}(T).$$

It remains an open problem to prove whether HAC-UCB is Pareto-adaptively minimax optimal, and more generally to identify the Pareto frontier for the causal bandit problem.

7 Related Work

Kocaoglu et al. [14] and Lindgren et al. [23] efficiently used interventions to learn a causal graph under standard causal assumptions [29]. Hyttinen et al. [13] identified the minimal number of experiments needed to learn underlying causal structure of variables with only a linear structural assumption. Lee and Bareinboim [20, 22, 21] identified the minimal set of interventions that permit causal effect identification in the presence of known causal structure, while Kumor et al. [16] studied analogues of the *back-door condition* for identifying causal effects with unobserved confounders.

Bareinboim et al. [7] introduced causal bandits with a binary motivating example to demonstrate that empirically better performance can be achieved by exploiting a specific, known causal structure. Lattimore et al. [17] and Yabe et al. [35] studied best-arm identification, where the experimenter does not incur any penalty for exploration rounds. Given knowledge of the causal graph informing the interventions and response, they separately proved that exponential improvements in the dependence of the regret on the action set are possible provided the underlying distribution on the causal graph is sufficiently "balanced".

Sen et al. [32] obtained an instance-dependent regret bound under causal assumptions, but obtained the wrong dependence on the arm gap (Δ^{-2} rather than Δ^{-1}), and consequently in the worst-case the dependence on T may still dwarf the structural benefits. Sen et al. [33] studied an alternative type of intervention, where rather than fixing a node only the conditional probabilities are changed. This notion is easily stated in our notation, since we allow for abstract families of distributions (indexed by abstract "interventions") to define a environment. However, they focused on distribution-dependent guarantees under stronger causal assumptions, and hence our results are not directly comparable.

All of the above regret bounds heavily require assumptions about the causal graph, and without such assumptions the presumed information learned from non-intervening rounds can catastrophically mislead the experimenter in exactly the same way that C-UCB suffers in our Theorem 4.5. Hence, it remains an interesting open problem to study adaptivity in each of these variations of the causal bandit setting, and our work provides a stepping stone to do so.

Prior to the present work, Lu et al. [24] has already been extended in multiple directions. de Kroon et al. [12] observed that C-UCB can be reduced to requiring only a separating set, but only prove the regret is no worse than that of UCB if a separating set is observed. The authors remark that a causal discovery algorithm could in principle be used to learn the separating set online, but observed in their experiments that they obtain biased estimates and hence there are no convergence guarantees. Lu et al. [25] replaced knowledge of the causal graph with the assumption that the causal graph has a tree structure, and incorporated the cost of learning the tree into the full regret bound. Nair et al. [27] provided an instance-dependent regret bound for an alternative algorithm to C-UCB, which they call C-UCB-2, in the presence of the full causal graph. While they demonstrated empirically that C-UCB-2 outperforms C-UCB for certain instances, we find that C-UCB-2 performs much worse when a d-separator is observed, and the provable linear lower bound (Theorem 4.5) also applies to C-UCB-2 when there are no observed d-separators.

8 Discussion

We have demonstrated that the improved regret possible when a d-separator is observed can also be realized in the multi-armed bandit problem by requiring only certain conditional independencies, which we have formalized using the conditionally benign property. We proved that it is impossible to optimally adapt to this property, but provided a new algorithm (HAC-UCB) that simultaneously recovers the improved regret for conditionally benign environments and significantly improves on prior work when the conditionally benign property does not hold. Crucially, our algorithm requires no more assumptions about the world than vanilla bandit algorithms. We expect our results to spur future work on (a) improved adaptation to the conditionally benign property, (b) relaxations of the conditionally benign property for which optimal adaptation is possible, and (c) adaptation in more general partial feedback settings.

In practice, HAC-UCB will be most useful in settings with a large action space and intermediate variables that may plausibly satisfy the conditionally benign property. In passing, we mention that one such example is learning the causal effect of genome edits (interventions) on disease phenotypes. Here, the post-action context could be gene expressions that are sometimes assumed to be a d-separator (e.g., [3]). We leave the implementation of our algorithm in clinical settings and collaboration with practitioners for future work.

Acknowledgements

BB is supported by an NSERC Canada Graduate Scholarship and the Vector Institute. DMR is supported in part by an NSERC Discovery Grant, Canada CIFAR AI Chair funding through the Vector Institute, and an Ontario Early Researcher Award. This material is based also upon work supported by the United States Air Force under Contract No. FA850-19-C-0511. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force. Part of this work was done while BB and DMR were visiting the Simons Institute for the Theory of Computing. We thank Zachary Lipton for suggesting to consider simultaneously adapting to estimation of the marginal distributions, and we thank Chandler Squires, Vasilis Syrgkanis, and Angela Zhou for helpful discussions.

References

[1] Y. Abbasi-Yadkori, P. L. Bartlett, V. Gabillon, A. Malek, and M. Valko. Best of both worlds: Stochastic and adversarial best-arm identification. In *Proceedings of the 31st Annual Conference on Learning Theory*, 2018.

[2] A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Proceedings of the 30th Annual Conference on Learning Theory*, 2017.

[3] H. F. Ainsworth and H. J. Cordell. Using gene expression data to identify causal pathways between genotype and phenotype in a complex disease: Application to genetic analysis workshop 19. *BMC Proceedings*, 10(49), 2016.

[4] R. Arora, T. V. Marinov, and M. Mohri. Corralling stochastic bandit algorithms. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[6] P. Auer, N. Cesa-Bianchi, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[7] E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems 28*, 2015.

[8] B. Bilodeau, J. Negrea, and D. M. Roy. Relaxing the i.i.d. assumption: Adaptively minimax optimal regret via root-entropic regularization, 2020. arXiv:2007.06552.

[9] S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Conference on Learning Theory*, 2012.

[10] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory*, 2009.

[11] C. L. Canonne. A short note on learning discrete distributions, 2020. arXiv:2002.11457.

[12] A. A. de Kroon, D. Belgrave, and J. M. Mooij. Causal bandits without prior knowledge using separating sets. In *Proceedings of the 1st Conference on Causal Learning and Reasoning*, 2021.

[13] A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

[14] M. Kocaoglu, A. G. Dimakis, and S. Vishwanath. Cost-optimal learning of causal graphs. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[15] W. M. Koolen. The Pareto regret frontier. In *Advances in Neural Information Processing Systems 27*, 2013.

[16] D. Kumor, J. Zhang, and E. Bareinboim. Sequential causal imitation learning with unobserved confounders. In *Advances in Neural Information Processing Systems 35*, 2021.

[17] F. Lattimore, T. Lattimore, and M. D. Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems 30*, 2016.

[18] T. Lattimore. The Pareto regret frontier for bandits. In *Advances in Neural Information Processing Systems 29*, 2015.

[19] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[20] S. Lee and E. Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019.

[21] S. Lee and E. Bareinboim. Causal effect identifiability under partial-observability. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[22] S. Lee and E. Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. In *Advances in Neural Information Processing Systems 34*, 2020.

[23] E. M. Lindgren, M. Kocaoglu, A. G. Dimakis, and S. Vishwanath. Experimental design for cost-aware learning of causal graphs. In *Advances in Neural Information Processing Systems 32*, 2018.

[24] Y. Lu, A. Meisami, A. Tewari, and Z. Yan. Regret analysis of bandit problems with causal background knowledge. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 2020.

[25] Y. Lu, A. Meisami, and A. Tewari. Causal bandits with unknown graph structure. In *Advances in Neural Information Processing Systems 35*, 2021.

[26] T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 2018.

[27] V. Nair, V. Patil, and G. Sinha. Budgeted and non-budgeted causal bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

[28] F. Orabona. A modern introduction to online learning, 2019. arXiv:1912.13213.

[29] J. Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

[30] J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

[31] Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[32] R. Sen, K. Shanmugam, M. Kocaoglu, A. Dimakis, and S. Shakkottai. Contextual bandits with latent confounders: An NMF approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[33] R. Sen, K. Shanmugan, A. G. Dimakis, and S. Shakkottai. Identifying best interventions through online importance sampling. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[34] T. S. Thune and Y. Seldin. Adaptation to easy data in prediction with limited advice. In *Advances in Neural Information Processing Systems 31*, 2018.

[35] A. Yabe, D. Hatano, H. Sumita, S. Ito, N. Kakimura, T. Fukunaga, and K.-i. Kawarabayashi. Causal bandits with propogating inference. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not see any negative societal impacts of these theoretical results.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] All computations were done on CPU on a personal laptop computer.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs for Causal Equivalences

We begin with a restating of standard definitions for completeness.

A.1 Standard Results from Causal Literature

To be more explicit about the role of Y in the causal setting (namely, it may be continuous), we introduce some more notation. Let $\mathbf{V} = (V_1, \ldots, V_M, Y)$ denote random variables each taking values in $\mathcal{V}_i = \{v_i^1, \ldots, v_i^{k_i}\}$ and $[0, 1]$ respectively. A *causal bandit graph* is any directed acyclic graph (DAG) $\mathcal{G}$ defined on the nodes $(V_1, \ldots, V_M, Y)$ such that (a) Y is a leaf node and (b) if there is a directed arrow from V_i to $V_{i'}$, then $i < i'$. Let $\mathscr{P}_{\mathbf{V}}$ denote the set of all probability distributions p on $(V_1, \ldots, V_M, Y)$ such that the marginal probabilities over $(V_1, \ldots, V_M)$ are all strictly positive.

Definition A.1 (Markovian parents, Definition 1.2.1 of [29]). *For any $p \in \mathscr{P}_{\mathbf{V}}$ and $i \in [M]$, the* Markovian parents of V_i under p *is the minimum-cardinality subset $\mathbf{V}' \subseteq (V_1, \ldots, V_{i-1})$ such that $V_i \perp (V_1, \ldots, V_{i-1}) \setminus \mathbf{V}' \mid \mathbf{V}'$ under p. We denote this by Pa_i^p. Similarly, Pa_Y^p is the minimum-cardinality subset $\mathbf{V}' \subseteq \mathbf{V}$ such that $Y \perp \mathbf{V} \setminus \mathbf{V}' \mid \mathbf{V}'$ under p.*

Definition A.2 (Graphical parents). *For any causal bandit graph $\mathcal{G}$, the* graphical parents of V_i under *$\mathcal{G}$ is the unique subset $\mathbf{V}' \subseteq \mathbf{V}$ of variables that have a directed arrow into V_i. We denote this by $\mathrm{Pa}_i^{\mathcal{G}}$. Similarly, $\mathrm{Pa}_Y^{\mathcal{G}}$ is the unique subset $\mathbf{V}' \subseteq \mathbf{V}$ of variables that have a directed arrow into Y.*

Definition A.3 (Markov relative, Theorem 1.2.7 of [29]). *A distribution $p \in \mathscr{P}_{\mathbf{V}}$ is* Markov relative to a causal bandit graph $\mathcal{G}$ *if any only if for all $V \in \mathbf{V} \cup \{Y\}$, $\mathrm{Pa}_V^p \subseteq \mathrm{Pa}_V^{\mathcal{G}}$.*

Remark A.4 (Equation 1.33 of [29]). *If $p \in \mathscr{P}_{\mathbf{V}}$ is Markov relative to $\mathcal{G}$, then for all measurable $\mathcal{B} \subseteq \mathcal{Y}$ and $(j_1, \ldots, j_M) \in \prod_{i=1}^M [k_i]$,*

$$p(Y \in \mathcal{B}, V_1 = v_1^{j_1}, \ldots, V_M = v_M^{j_M}) = p(Y \in \mathcal{B} \mid \mathrm{Pa}_Y^{\mathcal{G}} = \mathbf{u}_Y) \prod_{i=1}^{M} p(V_i = v_i^{j_i} \mid \mathrm{Pa}_i^{\mathcal{G}} = \mathbf{u}_i),$$

where the conditioning is understood to be on the event where the parents take the specific values defined by $\mathbf{u} = (v_1^{j_1}, \ldots, v_M^{j_M})$. ◁

Definition A.5 (Causal intervention, Definition 1.3.1 of [29]). *Let $p \in \mathscr{P}_{\mathbf{V}}$ be Markov relative to $\mathcal{G}$. The* interventional distribution *induced on $\mathbf{V}$ by the* intervention $a = \mathrm{do}(V_{i_1} = v_{i_1}^{j_{i_1}}, \ldots, V_{i_\ell} = v_{i_\ell}^{j_{i_\ell}})$ *is*

$$\begin{aligned} &p_a(Y \in \mathcal{B}, V_1 = v_1^{j_1}, \ldots, V_M = v_M^{j_M}) \\ &\quad = \mathbb{I}\{V_{i_1} = v_{i_1}^{j_{i_1}}, \ldots, V_{i_\ell} = v_{i_\ell}^{j_{i_\ell}}\} p(Y \in \mathcal{B} \mid \mathrm{Pa}_Y^{\mathcal{G}} = \mathbf{u}_Y) \prod_{i \notin \{i_1, \ldots, i_\ell\}} p(V_i = v_i^{j_i} \mid \mathrm{Pa}_i^{\mathcal{G}} = \mathbf{u}_i). \end{aligned}$$

Definition A.6 (d-Separated, Definition 2.4.1 of [30]). $\mathbf{Z}$ d-separates Y from $\mathbf{A}$ (on $\mathcal{G}$) *if and only if every path between $\mathbf{A}$ and Y is* blocked*; that is, every path contains either (a) $\bigcirc \to B \to \bigcirc$ or $\bigcirc \leftarrow B \to \bigcirc$ such that $B \in \mathbf{Z}$, or (b) $\bigcirc \to B \leftarrow \bigcirc$ with no descendents of B (including itself) in $\mathbf{Z}$.*

Definition A.7 (Back-Door Path, Section 3.3.1 of [29]). *A path from $\mathbf{Z}$ to $\mathbf{Z}'$ is a* back-door path *if it begins with an arrow directed into $\mathbf{Z}$.*

Definition A.8 (Front-Door Criterion, Definition 3.3.3 of [29]). $\mathbf{Z}$ satisfies the front-door criterion relative to $(\mathbf{A}, Y)$ *on $\mathcal{G}$ if and only if (a) all directed paths from $\mathbf{A}$ to Y pass through $\mathbf{Z}$, (b) there is no unblocked back-door path from $\mathbf{A}$ to $\mathbf{Z}$, and (c) all back-door paths from $\mathbf{Z}$ to Y are blocked by $\mathbf{A}$.*

A.2 Proof of Theorem 3.3

We first state an intuitive result about d-separation that is used often in the causal literature, but we could not find stated or proved precisely as follows.

Lemma A.9. *If $\mathbf{Z}$ d-separates Y from $\mathbf{A}$ and $\mathrm{Pa}_{\mathbf{A}}^{\mathcal{G}} \subseteq \mathbf{V} \setminus \mathbf{Z}$, then $\mathbf{Z}$ d-separates Y from $(\mathbf{A}, \mathrm{Pa}_{\mathbf{A}}^{\mathcal{G}})$.*

Proof of Lemma A.9. First, every path from $\mathrm{Pa}_{\mathbf{A}}^{\mathcal{G}}$ to Y that passes through $\mathbf{A}$ must satisfy one of (a) or (b) since a subpath does. Further, every path from $\mathrm{Pa}_{\mathbf{A}}^{\mathcal{G}}$ to Y that doesn't pass through $\mathbf{A}$ can be extended to a (back-door) path from $\mathbf{A}$ to Y using the edge $\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} \to \mathbf{A}$, but $\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} \subseteq \mathbf{V} \setminus \mathbf{Z}$ (and hence this part of the path cannot satisfy either property), so the original path must satisfy either (a) or (b). $\square$

Next, we recall the probabilistic equivalence of d-separation.

Theorem A.10 (Theorem 1.2.4 of [29]). *Fix a causal bandit graph $\mathcal{G}$ and two disjoint sets $\mathbf{Z} \subseteq \mathbf{V}$ and $\mathbf{A} \subseteq \mathbf{V}$ such that $\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} \subseteq \mathbf{V} \setminus \mathbf{Z}$. Then $\mathbf{Z}$ d-separates Y from $\mathbf{A}$ if and only if $Y \perp \mathbf{A} \mid \mathbf{Z}$ under every distribution $p \in \mathscr{P}_{\mathbf{V}}$ that is Markov relative to $\mathcal{G}$.*

We now turn to the main argument to prove Theorem 3.3. Let $p \in \mathscr{P}_{\mathbf{V}}$ and $a = \mathrm{do}(\mathbf{A} = \mathbf{a})$ be arbitrary. First, we prove the "Causal Effect Rule" [30]. For any $\mathbf{V}' \subseteq \mathbf{V}$,

$$\begin{aligned} p_a(\mathbf{V}' = \mathbf{v}') &= \sum_{\mathbf{u}} p_a(\mathbf{V}' = \mathbf{v}' \mid \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) p_a(\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) \\ &= \sum_{\mathbf{u}} p(\mathbf{V}' = \mathbf{v}' \mid \mathbf{A} = \mathbf{a}, \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) p(\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}), \end{aligned}$$

where the sum is over all possible values that $\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}}$ can take and we have used that (a) conditional on $\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}}$, the interventional and conditional distributions given $\mathbf{A} = \mathbf{a}$ are equivalent, and (b) the marginal distribution of $\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}}$ is unchanged by intervening on $\mathbf{A}$.

Now, suppose $\mathbf{Z}$ d-separates Y from $\mathbf{A}$. Then, it follows that

$$\begin{aligned} p_a(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z}) &= \frac{p_a(Y \in \mathcal{B}, \mathbf{Z} = \mathbf{z})}{p_a(\mathbf{Z} = \mathbf{z})} \\ &= \frac{\sum_{\mathbf{u}} p(Y \in \mathcal{B}, \mathbf{Z} = \mathbf{z} \mid \mathbf{A} = \mathbf{a}, \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) p(\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u})}{\sum_{\mathbf{u}} p(\mathbf{Z} = \mathbf{z} \mid \mathbf{A} = \mathbf{a}, \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) p(\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u})} \\ &= \frac{\sum_{\mathbf{u}} p(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z}, \mathbf{A} = \mathbf{a}, \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) p(\mathbf{Z} = \mathbf{z} \mid \mathbf{A} = \mathbf{a}, \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) p(\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u})}{\sum_{\mathbf{u}} p(\mathbf{Z} = \mathbf{z} \mid \mathbf{A} = \mathbf{a}, \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) p(\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u})} \\ &= \frac{\sum_{\mathbf{u}} p(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z}) p(\mathbf{Z} = \mathbf{z} \mid \mathbf{A} = \mathbf{a}, \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) p(\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u})}{\sum_{\mathbf{u}} p(\mathbf{Z} = \mathbf{z} \mid \mathbf{A} = \mathbf{a}, \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) p(\mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u})} \\ &= p(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z}), \end{aligned}$$

where the second last step uses Lemma A.9.

Conversely, suppose there exists $p \in \mathscr{P}_{\mathbf{V}}$ that is Markov relative to $\mathcal{G}$ and under which $Y \not\perp \mathbf{A} \mid \mathbf{Z}$. This implies (see the remark following Theorem 1.2.4 in [29]) there exists $p \in \mathscr{P}_{\mathbf{V}}$ with $p(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z}, \mathbf{A} = \mathbf{a}, \mathrm{Pa}^{\mathcal{G}}_{\mathbf{A}} = \mathbf{u}) \neq p(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z})$. By the above, this implies that $p_a(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z}) \neq p(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z})$. $\square$

A.3 Proof of Theorem 3.4

If $\mathbf{Z}$ d-separates Y from $\mathbf{A}$ on $\mathcal{G}_{\bar{\mathbf{A}}}$, then by Theorem 3.3 $\{p_a(\mathbf{Z}, Y) : a \in \mathcal{A}\}$ is conditionally benign for every $p \in \mathscr{P}_{\mathbf{V}}$ that is Markov relative to $\mathcal{G}_{\bar{\mathbf{A}}}$, and hence is still conditionally benign when the null intervention is excluded. It remains to observe that for any p that is Markov relative to $\mathcal{G}$, there exists p' that is Markov relative to $\mathcal{G}_{\bar{\mathbf{A}}}$ such that

$$\{p'_a(\mathbf{Z}, Y) : a \in \mathcal{A}_0\} = \{p_a(\mathbf{Z}, Y) : a \in \mathcal{A}_0\}.$$

Conversely, suppose there exists $p \in \mathscr{P}_{\mathbf{V}}$ that is Markov relative to $\mathcal{G}_{\bar{\mathbf{A}}}$ and under which $Y \not\perp \mathbf{A} \mid \mathbf{Z}$. Since necessarily $p_a(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z}) = p(Y \in \mathcal{B} \mid \mathbf{Z} = \mathbf{z})$ when a is the null intervention, it must be some $a \in \mathcal{A}_0$ that realizes the failure from the proof of Theorem 3.3, which means $\{p_a(\mathbf{Z}, Y) : a \in \mathcal{A}_0\}$ is not conditionally benign. $\square$

A.4 Proof of Lemma 3.5

Suppose that $\mathbf{Z}$ satisfies the front-door criterion relative to $(\mathbf{A}, Y)$ on $\mathcal{G}$ and there exists a path from $\mathbf{A}$ to Y on $\mathcal{G}_{\bar{\mathbf{A}}}$ that is unblocked by $\mathbf{Z}$. The path cannot be directed, since by the front-door criterion (a) it would include the subpath $\bigcirc \to \mathbf{Z} \to \bigcirc$, and hence would be blocked. The path also cannot be a back-door path since there are no arrows going into $\mathbf{A}$ on $\mathcal{G}_{\bar{\mathbf{A}}}$. Thus, there must be some part of the path that is of the form $\bigcirc \to B \leftarrow \bigcirc$ for some variable $B \in \mathbf{V}$ and there are no remaining colliders on the subpath from B to Y. Since the path is unblocked, this B must have a descendant (potentially itself) in $\mathbf{Z}$, and this creates a back-door path from $\mathbf{Z}$ to Y. On the portion of the back-door path that is from $\mathbf{Z}$ to B, there can be no colliders since then $\mathbf{Z}$ would not be a descendant of B, and hence this backdoor path contains no colliders. Since the back-door path also does not contain $\mathbf{A}$, it is unblocked by $\mathbf{A}$, which violates the front-door criterion (c). Thus, no path from $\mathbf{A}$ to Y on $\mathcal{G}_{\bar{\mathbf{A}}}$ that is unblocked by $\mathbf{Z}$ can exist, so $\mathbf{Z}$ blocks every path and hence $\mathbf{Z}$ d-separates Y from $\mathbf{A}$ on $\mathcal{G}_{\bar{\mathbf{A}}}$. $\square$

B Proofs for Regret Bounds

B.1 Concentration of Empirical Means

For a fixed $t \in [T]$, $a \in \mathcal{A}$, and $z \in \mathcal{Z}$, define the events

$$E_t^{\mathcal{A}}(a) = \left\{ \left| \hat{\mu}_t^{\mathcal{A}}(a) - \mathbb{E}_{\nu_a}[Y] \right| \leq \sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{A}}(a)}} \right\}$$

and

$$E_t^{\mathcal{Z}}(z) = \left\{ \max_{a \in \mathcal{A}} \left| \hat{\mu}_t^{\mathcal{Z}}(z) - \mathbb{E}_{\nu_a}[Y \mid Z = z] \right| \leq \sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}} \right\}.$$

Let $E^{\mathcal{A}} = \cap_{t \in [T], a \in \mathcal{A}} E_t^{\mathcal{A}}(a)$, $E^{\mathcal{Z}} = \cap_{t \in [T], z \in \mathcal{Z}} E_t^{\mathcal{Z}}(z)$, and $E = E^{\mathcal{A}} \cap E^{\mathcal{Z}}$. Finally, define the event

$$E^{\nu} = \left\{ \sup_{a \in \mathcal{A}} \sum_{z \in \mathcal{Z}} \left| \mathbb{P}_{\hat{\nu}_a}[Z = z] - \mathbb{P}_{\nu_a}[Z = z] \right| \leq \frac{\sqrt{|\mathcal{A}|\,|\mathcal{Z}| \log T}}{T^{1/4}} \right\}.$$

Lemma B.1. *For any $\nu \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ and $\pi \in \Pi(\mathcal{A}, \mathcal{Z}, T)$,*

$$\mathbb{P}_{\nu,\pi}[(E^{\mathcal{A}})'] \leq |\mathcal{A}|\, T\delta.$$

Proof of Lemma B.1. For each $a \in \mathcal{A}$, define the new i.i.d. random variables $Y_1^{\circ}(a), \ldots, Y_T^{\circ}(a) \sim \nu_a$. For any $t \in [T]$, Hoeffding's inequality can be applied to obtain

$$\mathbb{P}_{\nu_a}\left(\left| \frac{1}{t} \sum_{s=1}^{t} Y_s^{\circ}(a) - \mathbb{E}_{\nu_a}[Y] \right| > \sqrt{\frac{\log(2/\delta)}{2t}} \right) \leq \delta.$$

Then, using the i.i.d. property of $(Y_1(a), \ldots, Y_T(a))$,

$$\begin{aligned}
&\mathbb{P}_{\nu,\pi}\left(\exists t \in [T], a \in \mathcal{A}: \quad \left| \hat{\mu}_t^{\mathcal{A}}(a) - \mathbb{E}_{\nu_a}[Y] \right| > \sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{A}}(a)}} \right) \\
&\leq \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \mathbb{P}_{\nu_a}\left(\left| \frac{1}{t} \sum_{s=1}^{t} Y_s^{\circ}(a) - \mathbb{E}_{\nu_a}[Y] \right| > \sqrt{\frac{\log(2/\delta)}{2t}} \right) \\
&\leq |\mathcal{A}|\, T\delta,
\end{aligned}$$

where we have used a union bound over $a \in \mathcal{A}$ and $t \in [T]$. $\square$

Lemma B.2. *For any $\nu \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ that is conditionally benign and π,*

$$\mathbb{P}_{\nu,\pi}[(E^{\mathcal{Z}})'] \leq |\mathcal{Z}|\, T\delta.$$

Proof of Lemma B.2. Since ν is conditionally benign, there exists $p \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})$ such that for each $a \in \mathcal{A}$, $\nu_a(Y \mid Z) = p(Y \mid Z)$. Fix $z \in \mathcal{Z}$, and define joint the distribution $q_z(Y, Z) = p(Y \mid Z)\mathbb{I}\{Z = z\}$. Finally, define the new i.i.d. random variables $V_1^\circ, \dots, V_T^\circ \sim q_z$. For any $t \in [T]$, Hoeffding's inequality can be applied to obtain

$$\mathbb{P}_{q_z}\left(\Big|\frac{1}{t}\sum_{s=1}^{t} V_s^\circ - \mathbb{E}_{q_z}[Y]\Big| \leq \sqrt{\frac{\log(2/\delta)}{2t}}\right) \leq \delta.$$

Then,

$$\begin{aligned}
&\mathbb{P}_{\nu,\pi}\left(\exists t \in [T], z \in \mathcal{Z} : \quad \max_{a \in \mathcal{A}}\Big|\hat{\mu}_t^{\mathcal{Z}}(z) - \mathbb{E}_{\nu_a}[Y \mid Z = z]\Big| > \sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\right) \\
&= \mathbb{P}_{\nu,\pi}\left(\exists t \in [T], z \in \mathcal{Z} : \quad \Big|\hat{\mu}_t^{\mathcal{Z}}(z) - \mathbb{E}_{p}[Y \mid Z = z]\Big| \leq \sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\right) \\
&\leq \sum_{t=1}^{T}\sum_{z \in \mathcal{Z}} \mathbb{P}_{q_z}\left(\Big|\frac{1}{t}\sum_{s=1}^{t} V_s^\circ - \mathbb{E}_{q_z}[Y]\Big| \leq \sqrt{\frac{\log(2/\delta)}{2t}}\right) \\
&\leq |\mathcal{Z}|\, T\delta.
\end{aligned}$$

where we have used a union bound over $z \in \mathcal{Z}$ and $t \in [T]$. □

Theorem B.3 (Theorem 1 of Canonne [11])**.** *Let p be any distribution on $[k]$ for some integer k. For any $\varepsilon, \delta > 0$, if $n \geq \max\{k/\varepsilon^2, (2/\varepsilon^2)\log(2/\delta)\}$ and $X_1, \dots, X_n$ is an i.i.d. sample from p, then the MLE estimator*

$$\hat{p}_n(j) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\{X_t = j\} \quad \forall j \in [k]$$

satisfies

$$\mathbb{P}\Big[\frac{1}{2}\sum_{j \in [k]} |\hat{p}_n(j) - p(j)| > \varepsilon\Big] \leq \delta.$$

Lemma B.4. *If $\hat{\nu}(Z)$ is estimated using uniform exploration of at least $\lceil 4\sqrt{T}/|\mathcal{A}|\rceil$ rounds for each $a \in \mathcal{A}$,*

$$\mathbb{P}_{\nu,\pi}[(E^\nu)'] \leq 2\,|\mathcal{A}|\,/T.$$

Proof. Let $\varepsilon = (1/2)T^{-1/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T}$, $\delta = 2/T$, and n denote the number of exploration rounds used to estimate each $\hat{\nu}_a$. By a union bound and Theorem B.3, if $n \geq \max\{|\mathcal{Z}|\,/\varepsilon^2, (2/\varepsilon^2)\log(2/\delta)\}$ then

$$\begin{aligned}
\mathbb{P}_{\nu,\pi}[(E^\nu)'] &\leq \mathbb{P}_{\nu,\pi}\left[\exists a \in \mathcal{A} : \sum_{z \in \mathcal{Z}}\Big|\mathbb{P}_{\hat{\nu}_a}[Z = z] - \mathbb{P}_{\nu_a}[Z = z]\Big| > 2\varepsilon\right] \\
&\leq \sum_{a \in \mathcal{A}} \mathbb{P}_{\nu,\pi}\left[\frac{1}{2}\sum_{z \in \mathcal{Z}}\Big|\mathbb{P}_{\hat{\nu}_a}[Z = z] - \mathbb{P}_{\nu_a}[Z = z]\Big| > \varepsilon\right] \\
&\leq 2\,|\mathcal{A}|\,/T.
\end{aligned}$$

Then, it remains to observe that when $T \geq 3$ and $|\mathcal{Z}| \geq 2$ (which can be trivially assumed, since $\hat{\nu}$ is known exactly if $|\mathcal{Z}| = 1$),

$$\frac{|\mathcal{Z}|}{\varepsilon^2} = \frac{4\,|\mathcal{Z}|\,\sqrt{T}}{|\mathcal{A}|\,|\mathcal{Z}|\log T} \leq \frac{4\sqrt{T}}{|\mathcal{A}|}$$

and

$$\frac{2}{\varepsilon^2}\log(2/\delta) = \frac{8\sqrt{T}}{|\mathcal{A}|\,|\mathcal{Z}|\log T}(\log T) \leq \frac{4\sqrt{T}}{|\mathcal{A}|}.$$

□

B.2 Bounding Accumulated Regret

Lemma B.5. *For any $\nu \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$, $\pi \in \Pi(\mathcal{A}, \mathcal{Z}, T)$, and $t < t' \in [T]$, it holds almost surely that*

$$\sum_{s=t}^{t'} \frac{1}{\sqrt{\mathbb{T}^{\mathcal{A}}_{s-1}(A_s)}} \le \sqrt{8\,|\mathcal{A}|\,(t'-t)}\,.$$

Proof of Lemma B.5. Using Lemma 4.13 of Orabona [28],

$$\begin{aligned}
\sum_{s=t}^{t'} \frac{1}{\sqrt{\mathbb{T}^{\mathcal{A}}_{s-1}(A_s)}} &= \sum_{s=t}^{t'} \sum_{a\in\mathcal{A}} \frac{\mathbb{I}\{A_s = a\}}{\sqrt{1 \vee \sum_{j=1}^{s-1} \mathbb{I}\{A_j = a\}}} \\
&\le \sum_{s=t}^{t'} \sum_{a\in\mathcal{A}} \frac{\mathbb{I}\{A_s = a\}}{\sqrt{\sum_{j=1}^{t-1} \mathbb{I}\{A_j = a\} + (1/2)\sum_{j=t}^{s} \mathbb{I}\{A_j = a\}}} \\
&\le \sqrt{2} \sum_{a\in\mathcal{A}} \int_{\sum_{j=1}^{t-1} \mathbb{I}\{A_j = a\}}^{\sum_{j=1}^{t'} \mathbb{I}\{A_j = a\}} x^{-1/2} \mathrm{d}x \\
&= \sqrt{8} \sum_{a\in\mathcal{A}} \left(\sqrt{\sum_{j=1}^{t'} \mathbb{I}\{A_j = a\}} - \sqrt{\sum_{j=1}^{t-1} \mathbb{I}\{A_j = a\}} \right) \\
&\le \sum_{a\in\mathcal{A}} \sqrt{8 \sum_{j=t}^{t'} \mathbb{I}\{A_j = a\}} \\
&\le |\mathcal{A}| \sqrt{\frac{8}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \sum_{j=t}^{t'} \mathbb{I}\{A_j = a\}} \\
&= \sqrt{8\,|\mathcal{A}|\,(t'-t)}.
\end{aligned}$$

□

Lemma B.6. *For any $\nu \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$, $\pi \in \Pi(\mathcal{A}, \mathcal{Z}, T)$, and $t < t' \in [T]$,*

$$\mathbb{E}_{\nu,\pi}\left[\sum_{s=t}^{t'} \sum_{z\in\mathcal{Z}} \frac{1}{\sqrt{\mathbb{T}^{\mathcal{Z}}_{s-1}(z)}} \mathbb{P}_{\nu_{A_s}}[Z = z] \right] \le \sqrt{8\,|\mathcal{Z}|\,(t'-t)} + \sqrt{(1/2)(t'-t)\log(t'-t)} + 2.$$

Proof of Lemma B.6. First,

$$\begin{aligned}
&\mathbb{E}_{\nu,\pi}\left[\sum_{s=t}^{t'} \sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{s-1}(z)}\, \mathbb{P}_{\nu_{A_s}}[Z = z] \right] \\
&\quad = \mathbb{E}_{\nu,\pi}\left[\sum_{s=t}^{t'} \sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{s-1}(z)}\, \mathbb{I}\{Z_s(A_s) = z\} \right] \\
&\qquad + \mathbb{E}_{\nu,\pi}\left[\sum_{s=t}^{t'} \sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{s-1}(z)} \Big(\mathbb{P}_{\nu_{A_s}}[Z = z] - \mathbb{I}\{Z_s(A_s) = z\} \Big) \right] \\
&\quad \le \sqrt{8\,|\mathcal{Z}|\,(t'-t)} + \mathbb{E}_{\nu,\pi}\left[\sum_{s=t}^{t'} \sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{s-1}(z)} \Big(\mathbb{P}_{\nu_{A_s}}[Z = z] - \mathbb{I}\{Z_s(A_s) = z\} \Big) \right],
\end{aligned}$$

where we have used the same argument as Lemma B.5 applied to $\mathbb{T}^{\mathcal{Z}}_{s-1}(z)$ rather than $\mathbb{T}^{\mathcal{A}}_{s-1}(a)$. Following the analysis of Lu et al. [24], for $t \le j \le t'$ define the random variable

$$M_j = \sum_{s=t}^{j} \sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{s-1}(z)} \Big(\mathbb{P}_{\nu_{A_s}}[Z = z] - \mathbb{I}\{Z_s(A_s) = z\} \Big).$$

Then,

$$\mathbb{E}_{\nu,\pi}[M_j \mid A_j, H_{j-1}] = M_{j-1} + \sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{j-1}(z)}\, \mathbb{E}_{\nu,\pi}\Big[\mathbb{P}_{\nu_{A_s}}[Z=z] - \mathbb{I}\{Z_j(A_j)=z\} \mid A_j\Big] = M_{j-1}.$$

Further, it holds almost surely that

$$\begin{aligned}
|M_j - M_{j-1}| &= \left|\sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{j-1}(z)}\,\Big[\mathbb{P}_{\nu_{A_s}}[Z=z] - \mathbb{I}\{Z_j(A_j)=z\}\Big]\right| \\
&= \left|\mathbb{E}_{\nu,\pi}\left[\sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{j-1}(z)}\,\mathbb{I}\{Z_j(A_j)=z\} \;\Big|\; A_j, H_{j-1}\right] - \sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{j-1}(z)}\,\mathbb{I}\{Z_j(A_j)=z\}\right| \\
&= \left|\mathbb{E}_{\nu,\pi}\left[\sqrt{1/\mathbb{T}^{\mathcal{Z}}_{j-1}(Z_j(A_j))} \mid A_j, H_{j-1}\right] - \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{j-1}(Z_j(A_j))}\right| \\
&\le 1.
\end{aligned}$$

Then, by Azuma-Hoeffding, for all $x > 0$

$$\mathbb{P}_{\nu,\pi}\left[|M_{t'}| > \sqrt{x(t'-t)\log(t'-t)}\right] \le 2e^{-2x\log(t'-t)}.$$

Thus, since $|M_{t'}| \le t'-t$,

$$\begin{aligned}
&\mathbb{E}_{\nu,\pi}\left[\sum_{s=t}^{t'}\sum_{z\in\mathcal{Z}} \sqrt{1/\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}\,\Big(\mathbb{P}_{\nu_{A_s}}[Z=z] - \mathbb{I}\{Z_s(A_s)=z\}\Big)\right] \\
&\quad \le 2(t'-t)e^{-2x\log(t'-t)} + \sqrt{x(t'-t)\log(t'-t)}.
\end{aligned}$$

Taking $x = 1/2$ gives the result. □

B.3 Proof of Theorem 4.3

First, by Lemma B.2

$$\begin{aligned}
R_{\nu,\mathrm{C}}(T) &= \mathbb{E}_{\nu,\mathrm{C}} \sum_{t=1}^{T} \Big[\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \mathbb{E}_{\nu_{A^{\mathrm{C}}_t}}[Y]\Big] \\
&\le |\mathcal{Z}|\,T^2\delta + \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \mathbb{E}_{\nu_{A^{\mathrm{C}}_t}}[Y]\Big)\Big].
\end{aligned} \tag{B.1}$$

Then, by the conditionally benign property,

$$\begin{aligned}
&\mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \mathbb{E}_{\nu_{A^{\mathrm{C}}_t}}[Y]\Big)\Big] \\
&\quad = \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \widetilde{\mathrm{UCB}}_t(A^{\mathrm{C}}_t) + \widetilde{\mathrm{UCB}}_t(A^{\mathrm{C}}_t) - \mathbb{E}_{\nu_{A^{\mathrm{C}}_t}}[Y]\Big)\Big] \\
&\quad = \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\sum_{z\in\mathcal{Z}}\mathbb{E}_{\nu_{a^*_\nu}}[Y \mid Z=z]\mathbb{P}_{\nu_{a^*_\nu}}[Z=z] - \widetilde{\mathrm{UCB}}_t(A^{\mathrm{C}}_t)\Big)\Big] \\
&\qquad + \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\widetilde{\mathrm{UCB}}_t(A^{\mathrm{C}}_t) - \sum_{z\in\mathcal{Z}}\mathbb{E}_{\nu_{A^{\mathrm{C}}_t}}[Y \mid Z=z]\mathbb{P}_{\nu_{A^{\mathrm{C}}_t}}[Z=z]\Big)\Big].
\end{aligned}$$

We bound these terms separately. First, using the fact that $\tilde{\nu}(Z)$ and $\nu(Z)$ are ε-close, the definition of $E^{\mathcal{Z}}$, and the definition of A_t^{C},

$$\begin{aligned}
&\mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\sum_{z\in\mathcal{Z}}\mathbb{E}_{\nu_{a_\nu^*}}[Y \mid Z=z]\mathbb{P}_{\nu_{a_\nu^*}}[Z=z]-\widetilde{\mathrm{UCB}}_t(A_t^{\mathrm{C}})\Big)\Big]\\
&\leq \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\sum_{z\in\mathcal{Z}}\mathbb{E}_{\nu_{a_\nu^*}}[Y \mid Z=z]\mathbb{P}_{\tilde{\nu}_{a_\nu^*}}[Z=z]-\widetilde{\mathrm{UCB}}_t(A_t^{\mathrm{C}})\Big)\Big]+\varepsilon T\\
&\leq \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\widetilde{\mathrm{UCB}}_t(a_\nu^*)-\widetilde{\mathrm{UCB}}_t(A_t^{\mathrm{C}})\Big)\Big]+\varepsilon T\\
&\leq \varepsilon T.
\end{aligned} \tag{B.2}$$

Second, using the definition of $E^{\mathcal{Z}}$, the fact that $\tilde{\nu}(Z)$ and $\nu(Z)$ are ε-close, the conditionally benign property, and Lemma B.6,

$$\begin{aligned}
&\mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\widetilde{\mathrm{UCB}}_t(A_t^{\mathrm{C}})-\mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y]\Big)\Big]\\
&= \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\sum_{z\in\mathcal{Z}}\Big(\hat{\mu}_t^{\mathcal{Z}}(z)+\sqrt{\log(2/\delta)/(2\mathbb{T}_t^{\mathcal{Z}}(z))}\Big)\mathbb{P}_{\tilde{\nu}_{A_t^{\mathrm{C}}}}[Z=z]-\mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y]\Big)\Big]\\
&\leq \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\sum_{z\in\mathcal{Z}}\Big(\mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y \mid Z=z]+\sqrt{2\log(2/\delta)/\mathbb{T}_t^{\mathcal{Z}}(z)}\Big)\mathbb{P}_{\tilde{\nu}_{A_t^{\mathrm{C}}}}[Z=z]-\mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y]\Big)\Big]\\
&\leq \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\Big(\sum_{z\in\mathcal{Z}}\Big(\mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y \mid Z=z]+\sqrt{2\log(2/\delta)/\mathbb{T}_t^{\mathcal{Z}}(z)}\Big)\mathbb{P}_{\nu_{A_t^{\mathrm{C}}}}[Z=z]-\mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y]\Big)\Big]\\
&\quad+\varepsilon\left(1+\sqrt{2\log(2/\delta)}\right)T\\
&= \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{E^{\mathcal{Z}}\}\sum_{t=1}^{T}\sum_{z\in\mathcal{Z}}\sqrt{2\log(2/\delta)/\mathbb{T}_t^{\mathcal{Z}}(z)}\mathbb{P}_{\nu_{A_t^{\mathrm{C}}}}[Z=z]\Big]+\varepsilon\left(1+\sqrt{2\log(2/\delta)}\right)T\\
&\leq 4\sqrt{2\left|\mathcal{Z}\right|T\log T}+(\log T)\sqrt{2T}+4\sqrt{\log T}+\varepsilon(1+2\sqrt{\log T})T.
\end{aligned} \tag{B.3}$$

where the last line follows by taking $\delta = 2/T^2$. The theorem then follows by combining Eqs. (B.1) to (B.3). □

B.4 Proof of Theorem 4.5

We may assume without loss of generality that

$$\lim_{T\to\infty}\frac{\log(1/\delta_T)}{T}=0.$$

If this is not the case, then for large enough T, it holds that $\log(1/\delta_T) \geq c\,T$ for some c, and hence the instance dependent lower bounds for UCB variants [6] imply that regret grows linearly in T in the worst-case. We may also assume $|\mathcal{Z}| > 1$, for otherwise C-UCB plays the same arm forever (using its arbitrary tie-break rule) and so C-UCB can be forced to incur linear regret in a trivial way.

To illustrate that the lower bound is witnessed by a diversity of environments, we describe the construction in more general terms and then provide an example instantiation at the end of the proof. Let $\mathcal{A}_0$ and $\mathcal{Z}_0$ be arbitrary, nonempty, strict subsets of $\mathcal{A}$ and $\mathcal{Z}$ respectively, and let $\mathcal{A}_1 = \mathcal{A} \setminus \mathcal{A}_0$ and $\mathcal{Z}_1 = \mathcal{Z} \setminus \mathcal{Z}_0$.

We now describe sufficient conditions for an environment to be not conditionally benign and to force C-UCB to incur linear regret. For $p^{\mathcal{A}}(0)$ and $p^{\mathcal{A}}(1)$ in $(0, 1)$, let the marginal distribution be such that for $i \in \{0,1\}$, if $a \in \mathcal{A}_i$ then

$$\mathbb{P}_{\nu_a}[Z = z] = \frac{p^{\mathcal{A}}(i)}{|\mathcal{Z}_1|}\mathbb{I}\{z \in \mathcal{Z}_1\} + \frac{1 - p^{\mathcal{A}}(i)}{|\mathcal{Z}_0|}\mathbb{I}\{z \in \mathcal{Z}_0\}.$$

Similarly, for $\mu^{\mathcal{A},\mathcal{Z}}(0,0)$, $\mu^{\mathcal{A},\mathcal{Z}}(0,1)$, $\mu^{\mathcal{A},\mathcal{Z}}(1,0)$, and $\mu^{\mathcal{A},\mathcal{Z}}(1,1)$ in $(0,1)$, let the conditional distribution be such that for $i, j \in \{0,1\}$, if $a \in \mathcal{A}_i$ and $z \in \mathcal{Z}_j$ then

$$\nu_a(Y \mid Z = z) = \mathrm{Ber}(\mu^{\mathcal{A},\mathcal{Z}}(i,j)).$$

Observe that for $i \in \{0,1\}$, if $a \in \mathcal{A}_i$ then

$$\mathbb{E}_{\nu_a}[Y] = \mu^{\mathcal{A},\mathcal{Z}}(i,1)p^{\mathcal{A}}(i) + \mu^{\mathcal{A},\mathcal{Z}}(i,0)[1 - p^{\mathcal{A}}(i)].$$

We suppose the following conditions:

(1) $\forall a \in \mathcal{A}_0, a' \in \mathcal{A}_1 \quad \mathbb{E}_{\nu_a}[Y] > \mathbb{E}_{\nu_{a'}}[Y]$,

(2) $p^{\mathcal{A}}(0) < p^{\mathcal{A}}(1)$,

(3) $\min\{\mu^{\mathcal{A},\mathcal{Z}}(0,1), \mu^{\mathcal{A},\mathcal{Z}}(1,1)\} > \max\{\mu^{\mathcal{A},\mathcal{Z}}(0,0), \mu^{\mathcal{A},\mathcal{Z}}(1,0)\}$.

By Condition (1), $a^*_\nu \in \mathcal{A}_0$ (note that then all $a \in \mathcal{A}_0$ are equally optimal), so a constant amount of regret is incurred whenever $A^{\mathrm{C}}_t \in \mathcal{A}_1$. We now argue that under Conditions (2) and (3), this happens for a constant fraction of rounds with high probability.

First, we require slightly more notation to understand the behaviour of A^{C}_t. For every $t \in [T]$, $i \in \{0,1\}$, and $z \in \mathcal{Z}$, let

$$\mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(i,z) = \sum_{s=1}^{t} \mathbb{I}\{A^{\mathrm{C}}_s \in \mathcal{A}_i, Z_s = z\}.$$

Define $\mathbb{T}^{\mathcal{A}}_t(i) = \sum_{z \in \mathcal{Z}} \mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(i,z)$, and note that $\mathbb{T}^{\mathcal{Z}}_t(z) = \mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(0,z) + \mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(1,z)$. Further, let

$$\hat{\mu}^{\mathcal{A},\mathcal{Z}}_t(i,z) = \frac{1}{\mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(i,z)} \sum_{s=1}^{t} Y_s(A^{\mathrm{C}}_s)\mathbb{I}\{A^{\mathrm{C}}_s \in \mathcal{A}_i, Z_s = z\},$$

By definition,

$$\hat{\mu}^{\mathcal{Z}}_t(z) = \frac{\mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(0,z)}{\mathbb{T}^{\mathcal{Z}}_t(z)}\hat{\mu}^{\mathcal{A},\mathcal{Z}}_t(0,z) + \frac{\mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(1,z)}{\mathbb{T}^{\mathcal{Z}}_t(z)}\hat{\mu}^{\mathcal{A},\mathcal{Z}}_t(1,z).$$

Define the event F to be the case that for all $t \in [T]$, $i, j \in \{0,1\}$, and $z \in \mathcal{Z}_j$,

$$\left|\hat{\mu}^{\mathcal{A},\mathcal{Z}}_t(i,z) - \mu^{\mathcal{A},\mathcal{Z}}(i,j)\right| \le \sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(i,z)}},$$

and let G be the event that for all $t \in [T]$ and $z \in \mathcal{Z}$,

$$\frac{\mathbb{T}^{\mathcal{Z}}_t(z)}{t} \ge \min_{a \in \mathcal{A}} \mathbb{P}_{\nu_a}[Z = z] - \sqrt{\frac{\log T}{t}}$$

and

$$\frac{\mathbb{T}^{\mathcal{Z}}_t(z)}{t} \le \max_{a \in \mathcal{A}} \mathbb{P}_{\nu_a}[Z = z] + \sqrt{\frac{\log T}{t}}\,.$$

By Hoeffding and a union bound (i.e., the same arguments as Lemmas B.1 and B.2), $\mathbb{P}_\nu([F \cap G]') \le 8\,|\mathcal{Z}|\,/T$.

Now, suppose the event $F \cap G$ holds, and consider a fixed t. Recall that $A^{\mathrm{C}}_t \in \mathcal{A}_1$ is implied by

$$\max_{a_1 \in \mathcal{A}_1}\left\{\sum_{z \in \mathcal{Z}}\left(\hat{\mu}^{\mathcal{Z}}_t(z) + \sqrt{\frac{\log(2/\delta_T)}{2\mathbb{T}^{\mathcal{Z}}_t(z)}}\right)\mathbb{P}_{\nu_{a_1}}[Z = z]\right\}$$

$$> \max_{a_0 \in \mathcal{A}_0}\left\{\sum_{z \in \mathcal{Z}}\left(\hat{\mu}^{\mathcal{Z}}_t(z) + \sqrt{\frac{\log(2/\delta_T)}{2\mathbb{T}^{\mathcal{Z}}_t(z)}}\right)\mathbb{P}_{\nu_{a_0}}[Z = z]\right\}.$$

This is equivalent to

$$0 < \sum_{z\in\mathcal{Z}_1}\left(\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(0,z)+\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(1,z)+\sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\right)\frac{p^{\mathcal{A}}(1)-p^{\mathcal{A}}(0)}{|\mathcal{Z}_1|}$$
$$-\sum_{z\in\mathcal{Z}_0}\left(\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(0,z)+\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(1,z)+\sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\right)\frac{p^{\mathcal{A}}(1)-p^{\mathcal{A}}(0)}{|\mathcal{Z}_0|}.$$

The following two cases hold.

(i) For all $z\in\mathcal{Z}_1$,

$$\begin{aligned}&\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(0,z)+\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(1,z)+\sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\\
&\quad\geq\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\left(\mu^{\mathcal{A},\mathcal{Z}}(0,1)-\sqrt{\frac{\log T}{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}}\right)+\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\left(\mu^{\mathcal{A},\mathcal{Z}}(1,1)-\sqrt{\frac{\log T}{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}}\right)\\
&\quad\geq\min\{\mu^{\mathcal{A},\mathcal{Z}}(0,1),\mu^{\mathcal{A},\mathcal{Z}}(1,1)\}-2\sqrt{\log T}\Big(t\,p^{\mathcal{A}}(0)/|\mathcal{Z}_1|-\sqrt{t\log T}\Big)^{-1/2}.\end{aligned}$$

(ii) For all $z\in\mathcal{Z}_0$,

$$\begin{aligned}&\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(0,z)+\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(1,z)+\sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\\
&\quad\leq\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\left(\mu^{\mathcal{A},\mathcal{Z}}(0,0)+\sqrt{\frac{\log T}{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}}\right)+\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\left(\mu^{\mathcal{A},\mathcal{Z}}(1,0)+\sqrt{\frac{\log T}{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}}\right)+\sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\\
&\quad\leq\max\{\mu^{\mathcal{A},\mathcal{Z}}(0,0),\mu^{\mathcal{A},\mathcal{Z}}(1,0)\}+2\sqrt{\log T}\Big(t(1-p^{\mathcal{A}}(1))/|\mathcal{Z}_0|-\sqrt{t\log T}\Big)^{-1/2}\\
&\qquad+\sqrt{\log(2/\delta)}\Big(t(1-p^{\mathcal{A}}(1))/|\mathcal{Z}_0|-\sqrt{t\log T}\Big)^{-1/2}.\end{aligned}$$

That is,

$$\begin{aligned}&\sum_{z\in\mathcal{Z}_1}\left(\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(0,z)+\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(1,z)+\sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\right)\frac{p^{\mathcal{A}}(1)-p^{\mathcal{A}}(0)}{|\mathcal{Z}_1|}\\
&\qquad-\sum_{z\in\mathcal{Z}_0}\left(\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(0,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(0,z)+\frac{\mathbb{T}_t^{\mathcal{A},\mathcal{Z}}(1,z)}{\mathbb{T}_t^{\mathcal{Z}}(z)}\hat{\mu}_t^{\mathcal{A},\mathcal{Z}}(1,z)+\sqrt{\frac{\log(2/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\right)\frac{p^{\mathcal{A}}(1)-p^{\mathcal{A}}(0)}{|\mathcal{Z}_0|}\\
&\quad\geq\Big[\min\{\mu^{\mathcal{A},\mathcal{Z}}(0,1),\mu^{\mathcal{A},\mathcal{Z}}(1,1)\}-2\sqrt{\log T}\Big(t\,p^{\mathcal{A}}(0)/|\mathcal{Z}_1|-\sqrt{t\log T}\Big)^{-1/2}\\
&\qquad-\max\{\mu^{\mathcal{A},\mathcal{Z}}(0,0),\mu^{\mathcal{A},\mathcal{Z}}(1,0)\}-2\sqrt{\log T}\Big(t(1-p^{\mathcal{A}}(1))/|\mathcal{Z}_0|-\sqrt{t\log T}\Big)^{-1/2}\\
&\qquad-\sqrt{\log(2/\delta)}\Big(t(1-p^{\mathcal{A}}(1))/|\mathcal{Z}_0|-\sqrt{t\log T}\Big)^{-1/2}\Big]\Big(p^{\mathcal{A}}(1)-p^{\mathcal{A}}(0)\Big).\end{aligned}$$

By Condition (3), for large enough T and $t\geq T/2$, this last step can be further lower bounded[3] using

$$\begin{aligned}&\min\{\mu^{\mathcal{A},\mathcal{Z}}(0,1),\mu^{\mathcal{A},\mathcal{Z}}(1,1)\}-2\sqrt{\log T}\Big(t\,p^{\mathcal{A}}(0)/|\mathcal{Z}_1|-\sqrt{t\log T}\Big)^{-1/2}\\
&\qquad-\max\{\mu^{\mathcal{A},\mathcal{Z}}(0,0),\mu^{\mathcal{A},\mathcal{Z}}(1,0)\}-2\sqrt{\log T}\Big(t(1-p^{\mathcal{A}}(1))/|\mathcal{Z}_0|-\sqrt{t\log T}\Big)^{-1/2}\\
&\qquad-\sqrt{\log(2/\delta)}\Big(t(1-p^{\mathcal{A}}(1))/|\mathcal{Z}_0|-\sqrt{t\log T}\Big)^{-1/2}\\
&\quad\geq\Big(\min\{\mu^{\mathcal{A},\mathcal{Z}}(0,1),\mu^{\mathcal{A},\mathcal{Z}}(1,1)\}-\max\{\mu^{\mathcal{A},\mathcal{Z}}(0,0),\mu^{\mathcal{A},\mathcal{Z}}(1,0)\}\Big)/2.\end{aligned}$$

[3] When δ_T is polynomial in T, this can be improved to only require $t\geq(\log T)^2$.

Thus, we have that for sufficiently large T and any $a_0 \in \mathcal{A}_0$ and $a_1 \in \mathcal{A}_1$

$$\begin{aligned}
R_{\nu,\mathrm{C}}(T) &= \mathbb{E}_{\nu,\mathrm{C}} \sum_{t=1}^{T} (\mathbb{E}_{\nu_{a_0}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y]) \\
&\geq \mathbb{E}_{\nu,\mathrm{C}}\Big[\mathbb{I}\{F \cap G\} \sum_{t=1}^{T} (\mathbb{E}_{\nu_{a_0}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y])\Big] \\
&\geq (T/4)\Big(\mathbb{E}_{\nu_{a_0}}[Y] - \mathbb{E}_{\nu_{a_1}}[Y]\Big)(1 - 8\,|\mathcal{Z}|\,/T) \\
&\geq (T/5)\Big(\mathbb{E}_{\nu_{a_0}}[Y] - \mathbb{E}_{\nu_{a_1}}[Y]\Big).
\end{aligned}$$

Finally, a concrete example of an environment that satisfies Conditions (1)–(3) is:

$$\begin{aligned}
&\mu^{\mathcal{A},\mathcal{Z}}(0,0) = 1/6 \qquad && \mu^{\mathcal{A},\mathcal{Z}}(1,0) = 2/6 \\
&\mu^{\mathcal{A},\mathcal{Z}}(0,1) = 5/6 \qquad && \mu^{\mathcal{A},\mathcal{Z}}(1,1) = 4/6 \\
&p^{\mathcal{A}}(0) = 6/8 \qquad && p^{\mathcal{A}}(1) = 7/8.
\end{aligned}$$

For simplicity, we use the constants from this example in the theorem statement. □

B.5 Proof of Theorem 4.7

First, by Lemmas B.1 and B.4 with $\delta = 2/T^2$,

$$\begin{aligned}
R_{\nu,\mathrm{HAC}}(T) &= \mathbb{E}_{\nu,\mathrm{HAC}} \sum_{t=1}^{T} \Big[\mathbb{E}_{\nu_{a_\nu^*}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{HAC}}}}[Y]\Big] \\
&\leq 4\,|\mathcal{A}| + \mathbb{E}_{\nu,\mathrm{HAC}}\Big[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\nu}\} \sum_{t=1}^{T} \Big(\mathbb{E}_{\nu_{a_\nu^*}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{HAC}}}}[Y]\Big)\Big].
\end{aligned} \tag{B.4}$$

Let τ_0^{HAC} be the last round on which the algorithm uniformly explores and τ_1^{HAC} be the last round on which $A_t^{\mathrm{HAC}} = A_t^{\mathrm{C}}$. Since the algorithm is deterministic, τ_0^{HAC} is fixed but τ_1^{HAC} is stochastic. Then,

$$\begin{aligned}
&\mathbb{E}_{\nu,\mathrm{HAC}}\Big[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\nu}\} \sum_{t=1}^{T} \Big(\mathbb{E}_{\nu_{a_\nu^*}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{HAC}}}}[Y]\Big)\Big] \\
&\quad = \tau_0^{\mathrm{HAC}} + \mathbb{E}_{\nu,\mathrm{HAC}}\Big[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\nu}\} \sum_{\tau_0^{\mathrm{HAC}} < t \leq \tau_1^{\mathrm{HAC}}} \Big(\mathbb{E}_{\nu_{a_\nu^*}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y]\Big)\Big] \\
&\qquad + \mathbb{E}_{\nu,\mathrm{HAC}}\Big[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\nu}\} \sum_{t > \tau_1^{\mathrm{HAC}}} \Big(\mathbb{E}_{\nu_{a_\nu^*}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{UCB}}}}[Y]\Big)\Big].
\end{aligned} \tag{B.5}$$

Suppose that ν is conditionally benign and $\tilde{\nu}(Z)$ and $\nu(Z)$ are ε-close for $\varepsilon \leq T^{-1/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T}$. By triangle inequality, on the event E^{ν}

$$\sup_{a \in \mathcal{A}} \sum_{z \in \mathcal{Z}} \Big|\mathbb{P}_{\tilde{\nu}_a}[Z = z] - \mathbb{P}_{\hat{\nu}_a}[Z = z]\Big| \leq \frac{2\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T}}{T^{1/4}},$$

and thus $\tilde{\nu}(Z)$ will *not* be replaced by $\hat{\nu}(Z)$. Further, on $E^{\mathcal{A}} \cap E^{\mathcal{Z}}$ with $\delta = 2/T^2$, for all t it holds that

$$\begin{aligned}
&\mathrm{UCB}^{\mathcal{A}}_{t-1}(a) - \widetilde{\mathrm{UCB}}_{t-1}(a) \\
&\quad \geq \mathbb{E}_{\nu_a}[Y] - \sum_{z \in \mathcal{Z}} \left(\mathbb{E}_{\nu}[Y \mid Z = z] + 2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}} \right) \mathbb{P}_{\tilde{\nu}_a}[Z = z] \\
&\quad = \sum_{z \in \mathcal{Z}} \mathbb{E}_{\nu}[Y \mid Z = z] \mathbb{P}_{\nu_a}[Z = z] - \sum_{z \in \mathcal{Z}} \left(\mathbb{E}_{\nu}[Y \mid Z = z] + 2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}} \right) \mathbb{P}_{\tilde{\nu}_a}[Z = z] \\
&\quad \geq -2 \sum_{z \in \mathcal{Z}} \sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}} \mathbb{P}_{\tilde{\nu}_a}[Z = z] - \frac{\sqrt{|\mathcal{A}|\,|\mathcal{Z}| \log T}}{T^{1/4}}
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{UCB}^{\mathcal{A}}_{t-1}(a) - \widetilde{\mathrm{UCB}}_{t-1}(a) \\
&\quad \leq \mathbb{E}_{\nu_a}[Y] + 2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A}}_{t-1}(a)}} - \sum_{z \in \mathcal{Z}} \mathbb{E}_{\nu}[Y \mid Z = z] \mathbb{P}_{\tilde{\nu}_a}[Z = z] \\
&\quad = \sum_{z \in \mathcal{Z}} \mathbb{E}_{\nu}[Y \mid Z = z] \mathbb{P}_{\nu_a}[Z = z] + 2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A}}_{t-1}(a)}} - \sum_{z \in \mathcal{Z}} \mathbb{E}_{\nu}[Y \mid Z = z] \mathbb{P}_{\tilde{\nu}_a}[Z = z] \\
&\quad \leq 2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A}}_{t-1}(a)}} + \frac{\sqrt{|\mathcal{A}|\,|\mathcal{Z}| \log T}}{T^{1/4}},
\end{aligned}$$

and hence $\tau_1^{\mathrm{HAC}} = T$. Thus, by Lemma B.2 with $\delta = 2/T^2$, we can actually bound Eq. (B.5) using

$$\begin{aligned}
&\mathbb{E}_{\nu,\mathrm{HAC}}\left[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\nu}\} \sum_{t=1}^{T} \left(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{HAC}}}}[Y] \right) \right] \\
&\quad \leq \tau_0^{\mathrm{HAC}} + 2\,|\mathcal{Z}| + \mathbb{E}_{\nu,\mathrm{HAC}}\left[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\mathcal{Z}} \cap E^{\nu}\} \sum_{t=\tau_0^{\mathrm{HAC}}+1}^{T} \left(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{HAC}}}}[Y] \right) \right] \\
&\quad \leq \tau_0^{\mathrm{HAC}} + 2\,|\mathcal{Z}| + \mathbb{E}_{\nu,\mathrm{C}}\left[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\mathcal{Z}}\} \sum_{t=1}^{T} \left(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y] \right) \right] \\
&\quad \leq 5\sqrt{T} + 2\,|\mathcal{Z}| + 4\sqrt{2\,|\mathcal{Z}|\, T \log T} + (\log T)\sqrt{2T} + 4\sqrt{\log T} + \varepsilon(2 + 2\sqrt{\log T})T,
\end{aligned}$$

where the last line follows from Eqs. (B.2) and (B.3) and the definition of τ_0^{HAC}. The statement follows from combining this with Eq. (B.4).

Otherwise, consider when ν is not conditionally benign. By Theorem 4.1,

$$\begin{aligned}
&\mathbb{E}_{\nu,\mathrm{HAC}}\left[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\nu}\} \sum_{t > \tau_1^{\mathrm{HAC}}} \left(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{HAC}}}}[Y] \right) \right] \\
&\quad \leq \mathbb{E}_{\nu,\mathrm{UCB}}\left[\mathbb{I}\{E^{\mathcal{A}}\} \sum_{t=1}^{T} \left(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{UCB}}}}[Y] \right) \right] \\
&\quad \leq 4\sqrt{2\,|\mathcal{A}|\, T \log T}.
\end{aligned} \tag{B.6}$$

It remains to focus on the regret contribution from when $A_t^{\mathrm{HAC}} = A_t^{\mathrm{C}}$. The key observation is that if $t \leq \tau_1^{\mathrm{HAC}}$ then the hypothesis test passed for this round. We decompose the regret incurred using

$$\begin{aligned}
&\mathbb{E}_{\nu,\mathrm{HAC}}\left[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\nu}\} \sum_{\tau_0^{\mathrm{HAC}} < t \leq \tau_1^{\mathrm{HAC}}} \left(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y] \right) \right] \\
&= \mathbb{E}_{\nu,\mathrm{HAC}}\left[\mathbb{I}\{E^{\mathcal{A}} \cap E^{\nu}\} \sum_{\tau_0^{\mathrm{HAC}} < t \leq \tau_1^{\mathrm{HAC}}} \left(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \widetilde{\mathrm{UCB}}_{t-1}(A_t^{\mathrm{C}}) + \widetilde{\mathrm{UCB}}_{t-1}(A_t^{\mathrm{C}}) - \mathbb{E}_{\nu_{A_t^{\mathrm{C}}}}[Y] \right) \right].
\end{aligned}$$

First, on the event $E^{\mathcal{A}} \cap E^{\nu}$ with $\delta = 2/T^2$,

$$\begin{aligned}
&\sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \Big(\mathbb{E}_{\nu_{a^*_\nu}}[Y] - \widetilde{\text{UCB}}_{t-1}(A_t^{\text{C}}) \Big) \\
&\quad \le_{(a)} \sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \Big(\text{UCB}^{\mathcal{A}}_{t-1}(a^*_\nu) - \widetilde{\text{UCB}}_{t-1}(A_t^{\text{C}}) \Big) \\
&\quad \le_{(b)} \sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \left(\widetilde{\text{UCB}}_{t-1}(a^*_\nu) - \widetilde{\text{UCB}}_{t-1}(A_t^{\text{C}}) + 2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A}}_{t-1}(a^*_\nu)}} \right) + 2T^{3/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T} \\
&\quad \le_{(c)} 2 \sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A}}_{t-1}(a^*_\nu)}} + 2T^{3/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T} \\
&\quad \le_{(d)} 2 \sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \sqrt{\frac{|\mathcal{A}|\log T}{\sqrt{T}}} + 2T^{3/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T} \\
&\quad \le 4\,T^{3/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T},
\end{aligned} \tag{B.7}$$

where we have used that (a) $E^{\mathcal{A}}$ holds; (b) the hypothesis test for HAC-UCB passed; (c) $A_t^{\text{C}} = \arg\max_{a \in \mathcal{A}} \widetilde{\text{UCB}}_{t-1}(a)$; and (d) for $\tau_0^{\text{HAC}} < t$, $\mathbb{T}^{\mathcal{A}}_{t-1}(a^*_\nu) \ge \mathbb{T}^{\mathcal{A}}_{\tau_0^{\text{HAC}}}(a^*_\nu) \ge \sqrt{T}/|\mathcal{A}|$.

Second, again on the event $E^{\mathcal{A}} \cap E^{\nu}$ with $\delta = 2/T^2$,

$$\begin{aligned}
&\sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \Big(\widetilde{\text{UCB}}_{t-1}(A_t^{\text{C}}) - \mathbb{E}_{\nu_{A_t^{\text{C}}}}[Y] \Big) \\
&\quad \le_{(a)} \sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \left(\text{UCB}^{\mathcal{A}}_{t-1}(A_t^{\text{C}}) + 2\sum_{z \in \mathcal{Z}} \sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}}\, \mathbb{P}_{\tilde{\nu}_a}[Z = z] + \frac{2\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T}}{T^{1/4}} - \mathbb{E}_{\nu_{A_t^{\text{C}}}}[Y] \right) \\
&\quad \le_{(b)} \sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \left(2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A}}_{t-1}(A_t^{\text{HAC}})}} + 2\sum_{z \in \mathcal{Z}} \sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}}\, \mathbb{P}_{\tilde{\nu}_a}[Z = z] + \frac{\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T}}{T^{1/4}} \right) \\
&\quad \le_{(c)} \sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \left(2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A}}_{t-1}(A_t^{\text{HAC}})}} + 2\sum_{z \in \mathcal{Z}} \sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}}\, \mathbb{P}_{\nu_a}[Z = z] + 7\frac{(\log T)\sqrt{|\mathcal{A}|\,|\mathcal{Z}|}}{T^{1/4}} \right) \\
&\quad \le_{(d)} 4\sqrt{2\,|\mathcal{A}|\,T\log T} + 7T^{3/4}(\log T)\sqrt{|\mathcal{A}|\,|\mathcal{Z}|} + 2\sum_{\tau_0^{\text{HAC}} < t \le \tau_1^{\text{HAC}}} \sum_{z \in \mathcal{Z}} \sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}}\, \mathbb{P}_{\nu_a}[Z = z],
\end{aligned} \tag{B.8}$$

where we have used that (a) the hypothesis test for HAC-UCB passed; (b) $E^{\mathcal{A}}$ holds; (c) $\tilde{\nu}(Z)$ and $\nu(Z)$ are $3T^{-1/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T}$-close (if the original $\tilde{\nu}$ does not satisfy this, then by triangle inequality it is replaced with $\hat{\nu}$ that does satisfy this); and (d) Lemma B.5. It remains to take expectation and apply Lemma B.6, and then combine Eqs. (B.6) to (B.8) to obtain

$$\begin{aligned}
R_{\nu,\text{HAC}}(T) \le\ & 4\,|\mathcal{A}| + 4\sqrt{2\,|\mathcal{A}|\,T\log T} + 5\sqrt{T} + 4\,T^{3/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T} \\
& + 4\sqrt{2\,|\mathcal{A}|\,T\log T} + 7T^{3/4}(\log T)\sqrt{|\mathcal{A}|\,|\mathcal{Z}|} \\
& + 2\sqrt{\log T}\Big[\sqrt{8\,|\mathcal{Z}|\,T} + \sqrt{(T/2)\log T} + 2\Big].
\end{aligned} \tag{B.9}$$

□

B.6 Proof of Theorem 4.8

Supposing the event $E^{\mathcal{A}}$ holds (from Lemma B.1), every $a \in \mathcal{A}_0$ satisfies

$$\mathrm{UCB}^{\mathcal{A}}_{t-1}(a) \geq \mu^{\mathcal{A},\mathcal{Z}}(0,1)p^{\mathcal{A}}(0) + \mu^{\mathcal{A},\mathcal{Z}}(0,0)(1-p^{\mathcal{A}}(0)).$$

We use the same environment construction from the proof of Theorem 4.5 in Appendix B.4. Using the specific choice of δ_T, this argument implies that on the event $F \cap G$, $A^{\mathrm{C}}_t \in \mathcal{A}_1$ for sufficiently large T and $t \geq 2\sqrt{T}$. Recall that

$$\hat{\mu}^{\mathcal{Z}}_t(z) = \frac{\mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(0,z)}{\mathbb{T}^{\mathcal{Z}}_t(z)}\hat{\mu}^{\mathcal{A},\mathcal{Z}}_t(0,z) + \frac{\mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(1,z)}{\mathbb{T}^{\mathcal{Z}}_t(z)}\hat{\mu}^{\mathcal{A},\mathcal{Z}}_t(1,z).$$

Since $\mathbb{T}^{\mathcal{A},\mathcal{Z}}_t(0,z) \leq 2\sqrt{T}$ and $\mathbb{T}^{\mathcal{Z}}_t(z)$ grows linearly in t on the event $F \cap G$, for every $\alpha \in (0,1)$ and $\varepsilon > 0$ there exists T large enough such that for all $t \geq (\log T)\sqrt{T}$ (crucially, the $\log T$ ensures that the proportion of these t where $A^{\mathrm{C}}_t \in \mathcal{A}_1$ tends to 1 as T gets larger) it holds that for $z \in \mathcal{Z}_j$,

$$\hat{\mu}^{\mathcal{Z}}_t(z) \leq \alpha\mu^{\mathcal{A},\mathcal{Z}}(0,j) + (1-\alpha)\mu^{\mathcal{A},\mathcal{Z}}(1,j) + \varepsilon.$$

This implies that for all $\varepsilon > 0$, taking α small enough and T large enough with $t \geq (\log T)\sqrt{T}$ gives

$$\widetilde{\mathrm{UCB}}_{t-1}(a) \leq \mu^{\mathcal{A},\mathcal{Z}}(1,1)p^{\mathcal{A}}(0) + \mu^{\mathcal{A},\mathcal{Z}}(1,0)(1-p^{\mathcal{A}}(0)) + \varepsilon.$$

By the exploration phase,

$$2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A}}_{t-1}(a)}} \leq 2\sqrt{\frac{|\mathcal{A}|\log T}{\sqrt{T}}},$$

and hence can be made arbitrarily small by taking T sufficiently large. Thus, under the assumption

$$[\mu^{\mathcal{A},\mathcal{Z}}(0,1) - \mu^{\mathcal{A},\mathcal{Z}}(1,1)]p^{\mathcal{A}}(0) + [\mu^{\mathcal{A},\mathcal{Z}}(0,0) - \mu^{\mathcal{A},\mathcal{Z}}(1,0)](1-p^{\mathcal{A}}(0)) > 0, \tag{B.10}$$

for large enough T the first condition from Algorithm 1 will fail for some $a \in \mathcal{A}_0$ when $t = (\log T)\sqrt{T}$. Note that Eq. (B.10) is satisfied by the example given in the proof of Theorem 4.5. This implies that, on the event $F \cap G \cap E^{\mathcal{A}}$, for large enough T the HAC-UCB algorithm switches to following UCB when $t = (\log T)\sqrt{T}$. Since this joint event holds with probability larger than $1 - 2(|\mathcal{A}| + |\mathcal{Z}|)/T$, combining the exploration phase regret with the regret bound of Theorem 4.1 gives the result. □

C Proof of Theorem 6.2

Fix $\mathcal{A}$, $\mathcal{Z}$, and T. Let $\mathcal{Z}_0$ be an arbitrary strict subset of $\mathcal{Z}$ and $\mathcal{Z}_1 = \mathcal{Z} \setminus \mathcal{Z}_0$. Fix $\Delta \in (0, 1/20)$ and $\varepsilon \in (0,1)$ to be chosen later. Define the family of marginal distributions

$$q_a[Z \in \mathcal{Z}_0] = \begin{cases} 1/2 + 2\Delta & a = 1 \\ 1/2 & a \neq 1, \end{cases}$$

where probability is evenly spaced within $\mathcal{Z}_0$ and $\mathcal{Z}_1$ respectively. Further, define the Bernoulli conditional response distribution

$$\tilde{p}[Y = 1 \mid Z = z] = \begin{cases} 3/4 & z \in \mathcal{Z}_0 \\ 1/4 & z \in \mathcal{Z}_1. \end{cases}$$

Define the Bernoulli conditionally benign environment $\tilde{\nu} \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ for all $a \in \mathcal{A}$ by

$$\mathbb{P}_{\tilde{\nu}_a}[Y = 1] = \sum_{z \in \mathcal{Z}} \tilde{p}[Y = 1 \mid Z = z]q_a[Z = z].$$

Notice that

$$\mathbb{E}_{\tilde{\nu}_a}[Y] = \begin{cases} 1/2 + \Delta & a = 1 \\ 1/2 & a \neq 1. \end{cases}$$

Then, for every $a_0 \neq 1$, define $\nu^{a_0} \in \mathscr{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ for all $a \in \mathcal{A}$ by

$$\mathbb{P}_{\nu_a^{a_0}}[Y = 1] = \sum_{z \in \mathcal{Z}} p_a^{a_0}[Y = 1 \mid Z = z] q_a[Z = z],$$

where

$$p_a^{a_0}[Y = 1 \mid Z = z] = \begin{cases} 3/4 & a = 1, z \in \mathcal{Z}_0 \\ 1/4 & a = 1, z \in \mathcal{Z}_1 \\ 3/4 + 2\Delta(1+\varepsilon) & a = a_0, z \in \mathcal{Z}_0 \\ 1/4 & a = a_0, z \in \mathcal{Z}_1 \\ 3/4 & a \notin \{1, a_0\}, z \in \mathcal{Z}_0 \\ 1/4 & a \notin \{1, a_0\}, z \in \mathcal{Z}_1. \end{cases}$$

Notice that ν^{a_0} is *not* conditionally benign, and

$$\mathbb{E}_{\nu_a^{a_0}}[Y] = \begin{cases} 1/2 + \Delta & a = 1 \\ 1/2 + \Delta(1+\varepsilon) & a = a_0 \\ 1/2 & a \notin \{1, a_0\}. \end{cases}$$

We now extend Lemma 15.1 of Lattimore and Szepesvári [19]. In particular, let $\pi^q = \mathfrak{a}(\mathcal{A}, \mathcal{Z}, q, T)$ and observe that for any $\nu \in \Pi_{\mathcal{A},\mathcal{Z}}(q)$,

$$\mathrm{d}\mathbb{P}_{\nu,\pi^q}(H_T) = \prod_{t=1}^{T} \pi_t^q(A_t \mid H_{t-1}) \mathbb{P}_\nu[Z_t(A_t), Y_t(A_t) \mid A_t].$$

Thus, for every $a_0 \neq 1$,

$$\begin{aligned} \mathrm{KL}\left(\mathbb{P}_{\tilde{\nu},\pi^q} \parallel \mathbb{P}_{\nu^{a_0},\pi^q}\right) &= \mathbb{E}_{\tilde{\nu},\pi^q}\left[\log \frac{\mathrm{d}\mathbb{P}_{\tilde{\nu},\pi^q}}{\mathrm{d}\mathbb{P}_{\nu^{a_0},\pi^q}}(H_T)\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{\tilde{\nu},\pi^q}\left[\mathbb{E}_{\tilde{\nu},\pi^q}\left[\log \frac{\mathbb{P}_{\tilde{\nu}}[Z_t(A_t), Y_t(A_t)]}{\mathbb{P}_{\nu^{a_0}}[Z_t(A_t), Y_t(A_t)]} \;\Big|\; A_t\right]\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{\tilde{\nu},\pi^q}\left[\mathrm{KL}\left(\mathbb{P}_{\tilde{\nu}_{A_t}} \parallel \mathbb{P}_{\nu_{A_t}^{a_0}}\right)\right] \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}_{\tilde{\nu},\pi^q}[\mathbb{T}_T^{\mathcal{A}}(a)]\, \mathrm{KL}\left(\mathbb{P}_{\tilde{\nu}_a} \parallel \mathbb{P}_{\nu_a^{a_0}}\right). \end{aligned}$$

Since the marginal distribution q is shared, for each $a \in \mathcal{A}$ this simplifies to

$$\begin{aligned} \mathrm{KL}\left(\mathbb{P}_{\tilde{\nu}_a} \parallel \mathbb{P}_{\nu_a^{a_0}}\right) &= \sum_{z \in \mathcal{Z}} q_a(Z = z) \mathrm{KL}\left(\tilde{\nu}_a(Y \mid z) \parallel \nu_a^{a_0}(Y \mid z)\right) \\ &= \begin{cases} 0 & a \neq a_0 \\ (1/2)\mathrm{KL}\left(\mathrm{Ber}(3/4) \parallel \mathrm{Ber}(3/4 + 2\Delta(1+\varepsilon))\right) & a = a_0. \end{cases} \end{aligned}$$

Thus, by Pinsker's inequality (Theorem 14.2 of [19]),

$$\begin{aligned} R_{\tilde{\nu},\pi^q}(T) + R_{\nu^{a_0},\pi^q}(T) &> \frac{T\Delta}{2} \mathbb{P}_{\tilde{\nu},\pi^q}[\mathbb{T}_T^{\mathcal{A}}(1) \leq T/2] + \frac{T\Delta\varepsilon}{2} \mathbb{P}_{\nu^{a_0},\pi^q}[\mathbb{T}_T^{\mathcal{A}}(1) > T/2] \\ &\geq \frac{T\Delta\varepsilon}{4} \exp\{-\mathrm{KL}\left(\mathbb{P}_{\tilde{\nu},\pi^q} \parallel \mathbb{P}_{\nu^{a_0},\pi^q}\right)\} \\ &= \frac{T\Delta\varepsilon}{4} \exp\Big\{ -\mathbb{E}_{\tilde{\nu},\pi^q}[\mathbb{T}_T^{\mathcal{A}}(a_0)]\, (1/2)\mathrm{KL}\left(\mathrm{Ber}(3/4) \parallel \mathrm{Ber}(3/4 + 2\Delta(1+\varepsilon))\right) \Big\}. \end{aligned}$$

Using that $\mathrm{KL}\left(\mathrm{Ber}(3/4) \parallel \mathrm{Ber}(3/4 + 2\Delta(1+\varepsilon))\right) \leq 4x^2$ for $x < 1/10$ and the assumption of the theorem, this implies that for all T,

$$\mathbb{E}_{\tilde{\nu},\pi}[\mathbb{T}_T^{\mathcal{A}}(a_0)] \geq \frac{\log(T\Delta\varepsilon) - \log(8C\sqrt{|\mathcal{A}|\,T})}{(1/2)\mathrm{KL}\left(\mathrm{Ber}(3/4) \parallel \mathrm{Ber}(3/4 + 2\Delta(1+\varepsilon))\right)} \quad \geq \frac{1}{8\Delta^2(1+\varepsilon)^2} \log \frac{\Delta\varepsilon\sqrt{T}}{8C\sqrt{|\mathcal{A}|}}.$$

Finally, we combine this with

$$R_{\tilde{\nu},\pi}(T) = \sum_{a_0 \neq 1} \Delta \mathbb{E}_{\tilde{\nu},\pi}[\mathbb{T}_T^{\mathcal{A}}(a_0)],$$

choose $\varepsilon = 1$, and set

$$\Delta = \frac{16C\sqrt{|\mathcal{A}|}}{\sqrt{T}}.$$

□

D Simulation Details

Here we provide more details for the simulations in Section 5. First, the regret bounds are computed by sampling a new data realization for each horizon T we consider, computing the expected regret (with respect to the data randomness) for this realization, and then averaging this value (i.e., over the algorithm randomness) over $M = 300$ realizations.

C-UCB and UCB are implemented exactly according to Subsection 2.2, HAC-UCB is implemented exactly according to Algorithm 1, and C-UCB-2 is implemented exactly according to Algorithm 3 of Nair et al. [27] (including their time-adaptive confidence bound). For Corral, we use the log-barrier method from Algorithms 1 and 2 of Agarwal et al. [2] with base algorithms UCB and C-UCB. We use the prescribed learning rate from their Theorem 5 of

$$\eta = \frac{1}{40 \cdot \mathcal{R}(T) \log T},$$

where $\mathcal{R}(T)$ is an upper bound on the regret of C-UCB. In order to use UCB and C-UCB with importance-weighted losses, we implement the epoch-based approach of Arora et al. [4] along with their Freedman's inequality confidence bound of

$$\sqrt{\frac{4\rho \log t}{\mathbb{T}_t}} + \frac{4\rho \log t}{3\mathbb{T}_t}$$

for the arm means (UCB) and the post-action context conditional means (C-UCB) respectively, where ρ is an upper bound on the importance-weighted losses.

For Corral, it is typical in experiments (e.g., [4]) to swap out the Freedman's inequality confidence bound for the usual Hoeffding's inequality confidence bound. However, there are no theoretical guarantees for the algorithm then (due to the importance-weighted losses), and we observed in additional experiments that it is still not adaptive (i.e., it does no better than UCB even in conditionally benign environments).