Causal Bandits: The Pareto Optimal Frontier of Adaptivity, a Reduction to Linear Bandits, and Limitations around Unknown Marginals

Ziyi Liu [*1 2] **Idan Attias** [*2 3] **Daniel M. Roy** [1 2]

Abstract

In this work, we investigate the problem of adapting to the presence or absence of causal structure in multi-armed bandit problems. In addition to the usual reward signal, we assume the learner has access to additional variables, observed in each round after acting. When these variables d-separate the action from the reward, existing work in causal bandits demonstrates that one can achieve strictly better (minimax) rates of regret (Lu et al., 2020). Our goal is to adapt to this favorable "conditionally benign" structure, if it is present in the environment, while simultaneously recovering worst-case minimax regret, if it is not. Notably, the learner has no prior knowledge of whether the favorable structure holds. In this paper, we establish the Pareto optimal frontier of adaptive rates. We prove upper and matching lower bounds on the possible trade-offs in the performance of learning in conditionally benign and arbitrary environments, resolving an open question raised by Bilodeau et al. (2022). Furthermore, we are the first to obtain instance-dependent bounds for causal bandits, by reducing the problem to the linear bandit setting. Finally, we examine the common assumption that the marginal distributions of the post-action contexts are known and show that a nontrivial estimate is necessary for better-than-worst-case minimax rates.

1. Introduction

In real-world decision making, we often want strong worst-case guarantees as well as the ability to adapt to favorable properties of real-world scenarios. Adaptive sequential decision-making offers a framework to design algorithms to achieve these objectives.

In this paper, we explore adaptivity in multi-armed bandit problems. In standard multi-armed bandits, the learner (policy) takes an action, receives a reward, and then this process repeats over a number of rounds. The learner's regret is the difference between its cumulative reward and the cumulative reward of the single best action in hindsight. Can we work to identify high-reward actions while minimizing regret?

In this work, we assume there is *post-action context*, i.e., there may be additional information available to the learner after taking an action, beyond the reward signal. In a worst-case analysis, however, the learner can ignore the post-action context and still achieve minimax rates of regret: the worst-case environment will not offer useful information. However, many real-world settings possess the structure of multi-armed bandit problems with post-action context and, in those cases, this additional information is useful towards minimizing regret.

One way that post-action context can be useful is if we can assume causal structure relating the action (i.e., an intervention) to the reward and post-action (post-intervention) context. Several authors have studied models in this vein (Bareinboim et al., 2015; Lattimore et al., 2016). In this work, we build on the framework of Lattimore et al. (2016), wherein the post-action context is assumed to d-separate each intervention from its associated reward.

Under d-separation, the intervention and reward are independent, conditional on the post-intervention context. Bilodeau et al. (2022) formalized this structure in general terms: a bandit environment is *conditionally benign* whenever the conditional distribution of the reward, given the post-action context, does not depend on the action.

Minimax regret is well understood for both the classical and causal variant of multi-armed bandits. Notably, algorithms tailored to conditionally benign environments can achieve lower rates of regret, scaling with the number of post-action contexts, rather than the potentially much larger set of actions (Lu et al., 2020; Bilodeau et al., 2022).

[*]Equal contribution [1]Department of Statistical Sciences, University of Toronto, Canada [2]Vector Institute, Canada [3]Department of Computer Science, Ben-Gurion University, Israel. Correspondence to: Ziyi Liu <kevind.liu@mail.utoronto.ca>, Idan Attias <idanatti@post.bgu.ac.il>, Daniel M. Roy <daniel.roy@utoronto.ca>.

Workshop on Foundations of Reinforcement Learning and Control at the 41st International Conference on Machine Learning, Vienna, Austria. Copyright 2024 by the author(s).

Exploiting causal structure is not without its pitfalls. Bilodeau et al. (2022) showed that C-UCB, a minimax optimal causal bandit algorithm, suffers linear regret in some non-benign environments. This raised a natural question: Can we achieve strict adaptivity, i.e., obtain minimax rates simultaneously in the class of conditionally benign environments and in the class of all environments, without knowing in advance which class of environments we will face?

Bilodeau et al. proved that *strict* adaptivity was impossible, but showed some level of adaptivity was possible. They designed a new algorithm, termed HAC-UCB, and proved that it simultaneously achieves minimax optimal rates on the class of benign environments and always achieves (suboptimal, though sublinear) $T^{3/4}$ rates. In light of this result, Bilodeau et al. raised an open problem, asking whether HAC-UCB was, in a sense, Pareto optimal, implying that the slower rate was the price of adaptivity. More generally, we ask:

What is the Pareto optimal frontier of simultaneously achievable rates of regret in the classes of benign and arbitrary environments, and what algorithms achieve these optimal tradeoffs?

In this paper, we address the above question by providing a complete characterization of the Pareto optimal frontier (up to log factors) as well as the achieving algorithms. Besides adaptation, we also study the complexity of causal bandit problems from other perspectives. More specifically, we find a novel reduction from causal bandits to linear bandits, which facilitates the first instance-dependent regret bound for causal bandits and enables the applications of some linear bandit algorithms to causal bandits. We also investigate dropping the common assumption that we have perfect knowledge of "the marginals", i.e., the distribution of the post-action context variable, under each action. On one hand, we show that it is impossible for any algorithm to enjoy improved minimax regret in benign environments without any knowledge of the true marginals. On the other hand, we identify cases where approximate knowledge of the marginal distributions suffices. Our contributions are explained in more details as follows.

- In Section 3, we establish near-optimal Pareto regret frontiers for the setting of causal bandits, resolving an open problem raised by Bilodeau et al. (2022). Utilizing a dynamic balancing method introduced by Cutkosky et al. (2021), we derive the upper bound and also prove near-optimal matching lower bounds. Remarkably, we introduce a phenomenon we call *the price of adaptivity*, to capture the extra regret that one *must* incur when attempting to adapt to the presence or lack of causal structure. Consequently, we demonstrate that the model selection method introduced by Cutkosky et al. (2021) cannot be generally improved, for any nontrivial general improvement would decrease the price of adaptivity beyond our lower bound.
- In Section 4, we present a novel reduction from causal bandits to linear bandits with conditional sub-Gaussian noise. Utilizing a phased elimination technique (Lattimore et al., 2020), we identify a new dimension measuring the inherent complexity of causal bandits. It allows us to establish the first instance-dependent regret bound and a strictly tighter worst-case regret bound for causal bandits for conditionally benign environments. Additionally, we prove instance-dependent bounds for stochastic linear bandits, which are novel to the best of our knowledge.
- In Section 5, we study the situation where we have limited knowledge of the marginal distributions over post-action contexts. We provide a lower bound indicating that no algorithm can utilize the causal structure to achieve improved minimax rates without such prior knowledge. This partly justifies the common assumption in the causal bandits literature that algorithms are given the marginals. On the other side, we give a regret upper bound for the phased elimination algorithm with access to approximate marginals. This result shows that partial knowledge of the marginals suffices in some regimes.

1.1. Related Work

Causal bandits. The causal bandit model was introduced by Lattimore et al. (2016), where their objective was to identify the best intervention. Such pure exploration problem has been extensively studied since then (Sen et al., 2017; Xiong & Chen, 2022), while some other works focused on regret minimization (Lu et al., 2020; Nair et al., 2021; Bilodeau et al., 2022). Another interesting topic is to relax the causal assumptions. For example, the assumption of known causal graph can be relaxed (Lu et al., 2021; Malek et al., 2023). Our work mainly builds on the study by Bilodeau et al. (2022) regarding adapting to the existence of causal structures as well as approximate marginals.

Model selection. To achieve adaptivity, a natural idea is to apply some model selection algorithm on top of a group of base learners. There is an extending line of works studying such corralling strategies in the bandit setting (Agarwal et al., 2017; Pacchiano et al., 2020a;b; Cutkosky et al., 2020; Arora et al., 2021; Cutkosky et al., 2021). Agarwal et al. (2017) required certain stability conditions on the base learners, making their algorithm quite restricted. In contrast, some recently proposed general-purpose model selection algorithms for stochastic bandit problems (Pacchiano et al., 2020b; Cutkosky et al., 2020; 2021) are better candidates in our setting, since they only necessitate mild assumptions on the base learners.

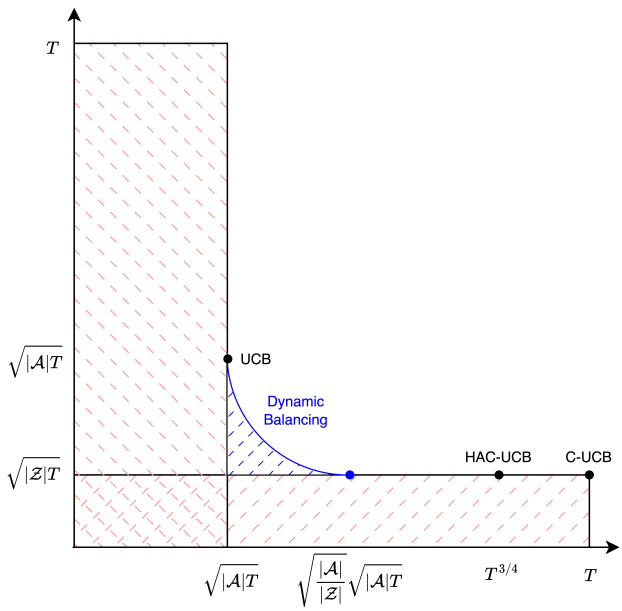

Figure 1. The Pareto-optimal frontier of simultaneously achievable rates of regret in (left axis) the class of conditionally benign environments and (bottom axis) the class of all environments. Shaded regions are unobtainable. All rates are determined up to log terms. Among algorithms that achieve minimax rates on conditionally benign environments, the previously best known algorithm (HAC-UCB) is dominated by an instance of Dynamic Balancing, which our results also demonstrate is Pareto optimal.

Pareto optimal frontier. When we have multiple performance metrics but are unable to achieve the best under all of them simultaneously, the Pareto optimal frontier becomes a common objective to pursue subsequently. Problems with several competing benchmarks are abundant in bandit literature (Koolen, 2013; Lattimore, 2015; Marinov & Zimmert, 2021; Zhu & Nowak, 2022).

2. Problem Setup

We consider the problem of stochastic bandit with post-action contexts, as defined by Bilodeau et al. (2022) and follow their notations. Let $\mathcal{A}$ be the finite action space, $\mathcal{Z}$ be the finite context space and $\mathcal{Y} = [0, 1]$ be the reward space. For any set K, we use $\mathcal{P}(K)$ to denote the set of all probability distributions supported on K. For any $p \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})$, we use $p(Z)$ to denote its marginal distribution over $\mathcal{Z}$, and use $p(Y|Z)$ to denote its the conditional distribution over $\mathcal{Y}$ conditioning on the $Z-$component.

In this bandit problem, a learner interacts with the stochastic environment for T rounds. The role of the environment is instantiated with a family of distributions $\nu = \{\nu_a : a \in \mathcal{A}\} \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ indexed by actions in $\mathcal{A}$. For each round $t \in [T]$, the learner picks an action A_t from $\mathcal{A}$ and then receives a context-reward pair (Z_t, Y_t) which is independently sampled from $\nu_{A_t} \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})$.

To model learner's strategy, we need to formalize the information that can be used for learner's prediction. Let $H_t = (A_s, Z_s, Y_s)_{s\in[t]}$ denote the observed history up to round t, which is a random variable valued in $\mathcal{H}_t := (\mathcal{A} \times \mathcal{Z} \times \mathcal{Y})^t$. A policy π by the learner could be modeled as a sequence of measurable maps from $\mathcal{H}_t$'s to $\mathcal{A}$

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

where $\Pi(\mathcal{A}, \mathcal{Z}, T)$ is the space of all policies compatible with $(\mathcal{A}, \mathcal{Z}, T)$. Then the learner follows this policy by selecting $A_t = \pi_t(H_{t-1})$ for each round t. Indeed, the distribution of all outcomes over T rounds, i.e. $(A_t, Z_t, Y_t)_{t\in[T]}$, is determined by the environment ν and the player's policy π together. We will always highlight the ambient joint distribution by the subscript on probabilistic operators $\mathbb{P}$ and $\mathbb{E}$, say $\mathbb{E}_{\nu_a}$ and $\mathbb{E}_{\nu,\pi}$. Additionally, we denote the expected reward for action a and the optimal action a^* by

$$\mu^{\mathcal{A}}(a) := \mathbb{E}_{\nu_a}[Y], \quad a^* := \operatorname*{argmax}_{a\in\mathcal{A}} \mu^{\mathcal{A}}(a). \tag{1}$$

The goal of the learner is to choose some policy π that maximizes her expected cumulative reward $\mathbb{E}_{\nu,\pi}[\sum_{t=1}^{T} Y_t]$, or equivalently minimizes her expected pseudo-regret

$$\begin{aligned}\mathbb{E}_{\nu,\pi}[\mathrm{Reg}(T)] &:= \mathbb{E}_{\nu,\pi}[\sum_{t=1}^{T} \max_{a\in\mathcal{A}} \mathbb{E}_{\nu_a}[Y] - Y_t] \\ &= T \cdot \mu^{\mathcal{A}}(a^*) - \mathbb{E}_{\nu,\pi}[\sum_{t=1}^{T} \mu^{\mathcal{A}}(A_t)],\end{aligned}$$

with $\mathrm{Reg}(t) := t \cdot \mu^{\mathcal{A}}(a^*) - \sum_{s=1}^{t} \mu^{\mathcal{A}}(A_s), t \in [T]$ being the realized regret, which is stochastic.

Conditionally benign property and d-separation. Under certain structures, the post-action context variable Z enables more efficient exploration and hence smaller regret. One special structure that can be exploited for better regret guarantee in our setting is called *conditionally benign property*, introduced by Bilodeau et al. (2022).

Definition 2.1. (Bilodeau et al., 2022, Definition 3.1) An environment $\nu \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ is *conditionally benign* if and only if there exists $p \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})$ such that for each $a \in \mathcal{A}$, $\nu_a(Z) \ll p(Z)$ and $\nu_a(Y|Z) = p(Y|Z)$ p-a.s. We further denote the space of all conditionally benign environments by $\mathcal{P}_{\mathrm{Benign}}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$.

The conditional benign property is quite general in the sense that it is equivalent to or weaker than some well-studied

causal assumptions (Bilodeau et al. 2022). In particular, the conditionally benign property is the same thing as the context variable Z being a d-separator when $\mathcal{A}$ is all interventions. To leverage this benign structure, the causal UCB (C-UCB) algorithm recently proposed by Lu et al. (2020) achieves $\tilde{O}(\sqrt{|\mathcal{Z}|T})$ regret, while non-causal algorithms that is unaware of this structure would still incur the possibly worse regret of $\tilde{O}(\sqrt{|\mathcal{A}|T})$.

2.1. Adaptivity

A natural question is whether we can compete with C-UCB when the environment is conditionally benign while at the same time still maintain the worst-case $\tilde{O}(\sqrt{|\mathcal{A}|T})$ regret guarantee, without prior knowledge of the nature of the environment. Unfortunately algorithms designed specific to the benign setting may fail drastically in non-benign settings. For instance, C-UCB provably incurs linear regret in some non-benign environments (Bilodeau et al., 2022). To remedy this, Bilodeau et al. (2022) devised HAC-UCB by adding a hypothesis test in each round, which is used for switching away from C-UCB to UCB irreversibly whenever it detects a deviation from conditionally benign property. HAC-UCB is able to recover the $\tilde{O}(\sqrt{|\mathcal{Z}|T})$ regret in benign settings and achieve sublinear $\tilde{O}(T^{3/4})$ regret in the worst case.

Prior to this work, we do not know if HAC-UCB is optimal. Indeed, Bilodeau et al. (2022) showed that *strict adaptation*, meaning that always achieving the worst-case $O(\sqrt{|\mathcal{A}|T})$ regret while still being able to perform as good as C-UCB when causal structure exists, is impossible. But this does not rule out the possibility of improving the worst-case $\tilde{O}(T^{3/4})$ regret of HAC-UCB unilaterally. In this paper we will show that such improvement is indeed feasible and thus obtain an algorithm that dominates HAC-UCB. Further we will show that our regret guarantee is not improvable through the lens of Pareto optimality.

Remark 2.2. Regarding optimal rate of regret under the presence of causal structure, it is easy to show a $\Omega(\sqrt{|\mathcal{Z}|T})$ regret lower bound, nearly matching existing $\tilde{O}(\sqrt{|\mathcal{Z}|T})$ regret upper bounds. Whether the log-factors can be shaved from the upper bound is unknown. However, the lower bound of Bilodeau et al. (2022) still implies that strict adaptation is impossible for general $\mathcal{A}$ and $\mathcal{Z}$, since when $|\mathcal{A}|/|\mathcal{Z}|$ is, say, $\Omega(T^{1/5})$, the $\Omega(\sqrt{|\mathcal{A}|T})$ lower bound in benign settings rules out a $\tilde{O}(\sqrt{|\mathcal{Z}|T})$ upper bound.

Generic algorithms. For rigorous treatment of adaptivity, we adopt the definition of algorithms as maps from Bilodeau et al. (2022). Specifically, an algorithm $\mathfrak{a}$ is any map from problem-specific inputs to the space of compatible policies

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

where $q \in \mathcal{P}(\mathcal{Z})^{\mathcal{A}}$ is the marginal distribution accessed by this algorithm as prior knowledge. When talking about algorithm-induced policies, by default we mean $\mathfrak{a}(\mathcal{A}, \mathcal{Z}, T, \nu(Z))$ if not stated otherwise, following the common assumption in the literature of causal bandits. We will also deal with the case of imperfect prior knowledge in Section 5, where q may not be the exact $\nu(Z)$. For notation simplicity, we will use $\mathfrak{a}$ to denote its induced policy $\mathfrak{a}(\mathcal{A}, \mathcal{Z}, T, q)$ when the problem-specific inputs are clear from context. For example, $\mathbb{E}_{\nu,\mathfrak{a}}$ is the same thing as $\mathbb{E}_{\nu,\mathfrak{a}(\mathcal{A},\mathcal{Z},T,q)}$.

3. The Pareto Regret Frontier

To formalize our notion of Pareto regret frontier, we need the following definition:

Definition 3.1. A pair of rate functions $(R_1(T; \mathcal{A}, \mathcal{Z}), R_2(T; \mathcal{A}, \mathcal{Z}))$ is said to be *realizable* if there is an algorithm $\mathfrak{a}$ such that for all $\mathcal{A}$, $\mathcal{Z}$ and T,

$$\sup_{\nu \in \mathcal{P}_{\text{Benign}}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}} \mathbb{E}_{\nu,\mathfrak{a}}[\text{Reg}(T)] \le R_1(T; \mathcal{A}, \mathcal{Z}),$$
$$\sup_{\nu \in \mathcal{P}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}} \mathbb{E}_{\nu,\mathfrak{a}}[\text{Reg}(T)] \le R_2(T; \mathcal{A}, \mathcal{Z}).$$

A pair $(R_1(T; \mathcal{A}, \mathcal{Z}), R_2(T; \mathcal{A}, \mathcal{Z}))$ is *reasonable* if $R_1(T; \mathcal{A}, \mathcal{Z}) \ge \sqrt{|\mathcal{Z}|T}$ and $R_2(T; \mathcal{A}, \mathcal{Z}) \ge \sqrt{|\mathcal{A}|T}$.

In the following we elide the dependence of rates R_i on $\mathcal{A}$ and $\mathcal{Z}$ below for clarity. We can now describe the Pareto regret frontier, i.e., the set of optimal realizable pairs of rates.

Theorem 3.2. *There exists universal constants $C, c, c' > 0$ such that*

1. *Upper bound: If $(R_1(T), R_2(T))$ is reasonable and $R_1(T)R_2(T) \ge |\mathcal{A}|T$, then $(CR_1(T)\log T, CR_2(T)\log T)$ is realizable;*
2. *Lower bound: For all realizable $(R_1(T), R_2(T))$, we have $R_2(T) > c'T$ or $R_1(T)R_2(T) \ge c|\mathcal{A}|T$.*

Both upper and lower bounds will be extensively discussed in the following sections.

3.1. Upper Bounds

In this section, we show that our upper bound can be obtained by applying the algorithmic principle of *dynamic balancing* (DB) in Cutkosky et al. (2021) to the stochastic bandit problem with post-action contexts. This method is motivated by the fact that, under mild assumptions, it can always achieve $\tilde{O}(\sqrt{T})$ regret when it is running on top of a collection of $\tilde{O}(\sqrt{T})$ regret base learners. So the dependence on T in $\tilde{O}(T^{3/4})$ regret by HAC-UCB in Bilodeau et al. (2022) is easily improved. The use of dynamic balancing in our bandit setting can be justified by the fact that

Algorithm 1 Dynamic balancing (DB) w/ two base learners

Input: Two base learners, $\{\mathfrak{a}_i\}_{i=1,2}$, factor $d_i(\cdot)$ of candidate regret bound, reward bias $b_i(\cdot)$ and scaling coefficient v_i (hyper-parameters) for each base learner $i \in \{1,2\}$, and confidence level $\delta \in (0,1)$.

1. Set $U_i(0) = n_i(0) = 0$ for all $i \in \{1,2\}$ and let the set of active learners be $\mathcal{I}_1 = \{1,2\}$
2. **For** $t = 1, 2, ..., T$ **do**
   (a) Select learner from the active set: $i_t \in \operatorname{argmin}_{i \in \mathcal{I}_t} v_i d_i(\delta)\sqrt{n_i(t-1)}$
   (b) Play action A_t of learner $\mathfrak{a}_{i_t}$ and receive reward Y_t and context Z_t
   (c) Update learner $\mathfrak{a}_{i_t}$ with Z_t and Y_t
   (d) Update $n_i(\cdot)$ and $U_i(\cdot)$:
   $U_i(t) \leftarrow U_i(t-1) + Y_t \mathbb{I}\{i = i_t\}$
   $n_i(t) \leftarrow n_i(t-1) + \mathbb{I}\{i = i_t\}$
   (e) Compute adjusted average reward $\eta_i(t)$ and confidence band $\gamma_i(t)$ for all $i \in \{1,2\}$:
   $\eta_i(t) \leftarrow \frac{U_i(t)}{n_i(t)} - b_i(t)$
   $\gamma_i(t) \leftarrow 3\sqrt{\frac{\log(2\log n_i(t)/\delta)}{n_i(t)}}$
   (f) Update the set of active learners:
   $$\mathcal{I}_{t+1} \leftarrow \Big\{ i \in \{1,2\} : \eta_i(t) + \gamma_i(t) + \frac{d_i(\delta)}{\sqrt{n_i(t)}} \geq \max_{j=1,2} \eta_j(t) + \gamma_j(t) \Big\}$$

dynamic balancing does not rely on what kind of (stochastic) contextual information can be observed in the underlying bandit problem. See Appendix A for a detailed explanation.

Note that dynamic balancing algorithm (Algorithm 1) is input by a set of user-specified candidate regret bounds for each base learner i (which takes the form of $d_i\sqrt{t}$ in our setting). In each round, DB merely picks the base learner with minimal candidate regret bound, and performs a test to identify and deactivate the learners that seem to violate their candidate regret bounds. As long as there is one base learner whose candidate regret is valid, DB is able to compete with the best of such base learners. A more comprehensive exposition of the idea behind dynamic balancing can be found in Cutkosky et al. (2021).

So naturally, we need one base learner that is favorable in benign instances and another base learner that remains robust to non-benign instances. For example, we can pick C-UCB and UCB, but note that any other algorithm with similar regret bound can be applied as well. Formally we characterize base learners that enjoys certain regret bound in certain type of environments by the following definition:

Definition 3.3. Let $d : (0,1) \to \mathbb{R}_{>0}$. A family of learners $\mathfrak{a} = (\mathfrak{a}_\delta)_{\delta \in (0,1)}$ is a *d-benign family* if, for all $\delta \in (0,1)$, for all benign instances, with probability at least $1 - O(\delta)$, for all $t \in [T]$, $\mathfrak{a}_\delta$ has regret no larger than $d(\delta)\sqrt{t}$. Similarly, a learner $\mathfrak{a}$ is a *d-arbitrary learner* if, for all $\delta \in (0,1)$, for all instances, with probability at least $1 - O(\delta)$, for all $t \in [T]$, $\mathfrak{a}$ has regret no larger than $d(\delta)\sqrt{t}$.

Let C-UCB $= (\text{C-UCB}(\delta))_{\delta \in (0,1)}$ and UCB $= (\text{UCB}(\delta))_{\delta \in (0,1)}$ be the families of instances of the C-UCB and UCB algorithms, respectively, whose confidence band is scaled by $\Theta(\sqrt{\log(1/\delta)})$. See Appendix C for details.

Proposition 3.4. C-UCB *is a d-benign family for* $d(\delta) = O((\sqrt{|\mathcal{Z}|} + \sqrt{\log(T/\delta)})\sqrt{\log(|\mathcal{Z}|T/\delta)})$ *and* UCB *is a d'-arbitrary family for* $d'(\delta) = O(\sqrt{|\mathcal{A}|\log(|\mathcal{A}|T/\delta)})$.

Note that the above result for UCB is folklore, but the result for C-UCB is new. The following result describes the adaptive regret of dynamic balancing acting on a benign family and an arbitrary family, which validates the upper bound in Theorem 3.2. What is more impressive is that to realize every point on the Pareto regret frontier (up to log factors), we need only tune the hyper-parameters in DB accordingly. We elide the dependence of rates R_i on $\mathcal{A}$ and $\mathcal{Z}$ below for clarity. See Appendix A.2 for the proof.

Theorem 3.5. *Let $\mathfrak{a}_1$ be a d_1-benign family and let $\mathfrak{a}_2$ be a d_2-arbitrary family of learners, where* $d_1(\delta) = O\Big((\sqrt{|\mathcal{Z}|} + \sqrt{\log(T/\delta)})\sqrt{\log(|\mathcal{Z}|T/\delta)}\Big)$, $d_2(\delta) = O(\sqrt{|\mathcal{A}|\log(|\mathcal{A}|T/\delta)})$. *For every pair of reasonable rate functions $R_1(T), R_2(T)$ such that $R_1(T)R_2(T) \geq |\mathcal{A}|T$, there exist hyper-parameters $b_i(\cdot), v_i$, $i = 1,2$, such that, for all instances ν, the policy* DB(δ)*, for $\delta = 1/T$, given by Algorithm 1 with $\mathfrak{a}_1, \mathfrak{a}_2$ and d_1, d_2, satisfies*

$$\mathbb{E}_{\nu,\text{DB}(\delta)}[\text{Reg}(T)] = \tilde{O}(R_1(T)); \ \nu \text{ is conditionally benign,}$$
$$\mathbb{E}_{\nu,\text{DB}(\delta)}[\text{Reg}(T)] = \tilde{O}(R_2(T)); \ \nu \text{ is arbitrary.}$$

Corollary 3.6. *Taking $R_1(T) = \sqrt{|\mathcal{Z}|T}$ and $R_2(T) = \sqrt{|\mathcal{A}|/|\mathcal{Z}|} \cdot \sqrt{|\mathcal{A}|T}$, the conclusion of Theorem 3.5 is*

$$\mathbb{E}_{\nu,\text{DB}(\delta)}[\text{Reg}(T)] = \tilde{O}(\sqrt{|\mathcal{Z}|T}); \ \nu \text{ is conditionally benign,}$$
$$\mathbb{E}_{\nu,\text{DB}(\delta)}[\text{Reg}(T)] = \tilde{O}(\sqrt{|\mathcal{A}|/|\mathcal{Z}|} \cdot \sqrt{|\mathcal{A}|T}); \ \nu \text{ is arbitrary.}$$

Corollary 3.6 indicates that we need to pay an extra factor of $\sqrt{|\mathcal{A}|/|\mathcal{Z}|}$ in the worst-case regret for adaptivity, and it already improves over the one by HAC-UCB in terms of worst-case regret. Moreover, our regret analysis does not require their cumbersome assumption that $T \geq 25|\mathcal{A}|^2$. Such improvement may be explained as follows. Both dynamic balancing and HAC-UCB play with two base learners and decide which to pick in each round. However, DB is operating in a more reasonable way: DB alternates between two

base learners and never deactivates any of them permanently, whereas HAC-UCB first plays the optimistic base learner persistently up to some point and then switches to UCB for the remaining rounds. Thus the regret of HAC-UCB incurred by running the optimistic base learner improperly may be dominant.

3.2. Lower Bounds

In this section we elaborate on the lower bound in Theorem 3.2 in the following Theorem 3.7, which is a generalization of (Bilodeau et al., 2022, Theorem 6.2). The proof of Theorem 3.7 closely follows that of the original, but we are able to derive a continuum of lower bounds that constitute the Pareto regret frontier. For completeness, we provide the full proof in Appendix D.1.

Theorem 3.7. *There exists constants $c, c' > 0$ such that, for all MAB algorithms $\mathfrak{a}$, rate functions $R(T; \mathcal{A}, \mathcal{Z})$, if, for all $\mathcal{A}, \mathcal{Z}, T$*

$$\sup_{\nu} \mathbb{E}_{\nu,\mathfrak{a}}[\mathrm{Reg}(T)] \leq R(T; \mathcal{A}, \mathcal{Z}),$$

then, for all $\mathcal{A}$, $\mathcal{Z}$ and T, there exists a conditionally benign environment ν such that either $R(T; \mathcal{A}, \mathcal{Z}) > c'T$, or there exists a conditionally benign environment ν such that

$$\mathbb{E}_{\nu,\mathfrak{a}}[\mathrm{Reg}(T)] \geq c \cdot \frac{|\mathcal{A}|T}{R(T; \mathcal{A}, \mathcal{Z})}.$$

Theorem 3.7 shows that any pair of realizable rates must have their product lower bounded by $|\mathcal{A}|T$ unless the worst-case regret bound is vacuously large. Combining Theorem 3.5 with Theorem 3.7, we have justified the Pareto optimality of dynamic balancing. As a corollary, we have found a problem of adaptation where model selection method can be optimal and the price of adaptivity is witnessed by the additional multiplicative factor of $\sqrt{|\mathcal{A}|/|\mathcal{Z}|}$ in the regret bound.

4. Instance-Dependent Bounds via Phased Elimination Algorithm

Besides achieving Pareto optimal regret bounds in Theorem 3.5 that are worst-case in nature, the dynamic balancing algorithm can also enjoy $O(\log T)$ instance-dependent regret at the same time under additional assumptions on the base learners. In particular, C-UCB may not be our best choice for the benign base learner. To leverage the strength of dynamic balancing, we propose a new causal bandit algorithm that enjoys $\tilde{O}(\sqrt{|\mathcal{Z}|T})$ worst-case regret and a novel logarithmic instance-dependent regret in benign settings in this section. We are the first to pursue instance-dependent results in conditionally benign environments for algorithms that are minimax optimal (up to log factors).

Our new algorithm is built upon the idea of *phased elimination with G-optimal design* from linear bandits (Lattimore & Szepesvári, 2020; Lattimore et al., 2020). Our regret analysis hinges on a novel reduction from causal bandits to linear bandits. This reduction enables the use of a broad family of linear bandit algorithms in conditionally benign environments, whose regret guarantees remain intact.

Finally, we will discuss the possibilities and challenges regarding adaptive $O(\log T)$ instance-dependent regret.

4.1. Reduction to Linear Bandits

We need additional notations to illustrate our causal-to-linear reduction. For benign instance ν, define the mean reward vector $\mu^{\mathcal{Z}} \in [0,1]^{|\mathcal{Z}|}$ by $\mu^{\mathcal{Z}}(z) = \mathbb{E}_{\nu_a}[Y|Z = z], \forall z \in \mathcal{Z}$. Also, in this section we use ν_a to denote its associated marginal distribution vector $\nu_a(Z) \in \mathcal{P}(\mathcal{Z}) \subset \mathbb{R}^{|\mathcal{Z}|}$, and we won't distinguish between an action a and its associated marginal vector ν_a.

Recall that in each round t we play some action A_t and then observe context Z_t and reward Y_t. By simply ignoring the realized contexts Z_t, we can write $Y_t = \sum_{z\in\mathcal{Z}} \mu^{\mathcal{Z}}(z) \cdot \nu_{A_t}(z) + \eta_t^{\mathcal{A}} = \langle \mu^{\mathcal{Z}}, \nu_{A_t} \rangle + \eta_t^{\mathcal{A}}$, where $\eta_t^{\mathcal{A}}$ is conditionally 1-sub-Gaussian since $\mathbb{E}[\eta_t^{\mathcal{A}} | (A_s, Y_s)_{s\leq t-1}, A_t] = 0$ and $\eta_t^{\mathcal{A}} \in [-1, 1]$. So now we may think of the game to be linear bandit with actions being ν_a and the unknown mean reward vector being $\mu^{\mathcal{Z}}$. Therefore, any linear bandit algorithm that allows such conditionally sub-Gaussian noise condition should be able to operate in our benign setting by ignoring the realized contexts. More importantly, its regret analysis will go through without change, and hence its regret bounds are retained without loss.

4.2. Phased Elimination and its Regret Bound

Among all valid linear bandit algorithms that can be applied in conditionally benign environments, we opt for the phase elimination algorithm (PE) over others due to its superior performance whenever our action set is finite. Its pseudo-code is summarized in Algorithm 2, which is essentially the same as Lattimore et al. (2020). However, the regret guarantees we present for PE are novel. Our first result is an anytime worst-case regret bound, which qualifies PE for being a base learner of dynamic balancing. Again, $\mathrm{PE} = (\mathrm{PE}(\delta))_{\delta\in(0,1)}$ is the family of instances of phased elimination algorithm, indexed by the confidence level δ.

Theorem 4.1 (Worst-case regret bound for PE)**.** *For all $\delta \in (0,1)$, the policy $\mathrm{PE}(\delta)$ given by Algorithm 2 satisfies the following regret bound for all conditionally benign environments ν,*

$$\mathrm{Reg}(t) \leq C\sqrt{d_\nu \log\left(\frac{|\mathcal{A}| \log T}{\delta}\right) t}, \quad \forall t \in [T]$$

with probability at least $1-\delta$*, where* $d_\nu = \dim(\mathrm{span}\{\nu_a : a \in \mathcal{A}\})$ *and* $C > 0$ *is a universal constant. Note that* $d_\nu \le |\mathcal{Z}|$ *and it could be* $|\mathcal{Z}|$ *in the worst-case. In particular, after taking* $\delta = 1/T$*, we obtain the expected regret bound*

$$\mathbb{E}_{\nu,\mathrm{PE}(\delta)}[\mathrm{Reg}(T)] = O\Big(\sqrt{d_\nu T \log(|\mathcal{A}|T)}\Big).$$

Corollary 4.2. PE *is a* d*-benign family for* $d(\delta) = O\Big(\sqrt{|\mathcal{Z}|\log(|\mathcal{A}|\log T/\delta)}\Big)$*. Therefore, the expected regret bound in Theorem 3.5 can also be achieved by* DB *with* PE *and* UCB *as base learners.*

See Appendix B for the proof. Thanks to our reduction, Theorem 4.1 only depends on d_ν (up to log factors) rather than $|\mathcal{Z}|$. This indicates that the intrinsic complexity of causal bandit problem is not $|\mathcal{Z}|$ and can be further reduced to d_ν, which is not captured by the $\tilde{O}(\sqrt{|\mathcal{Z}|T})$ regret bound of C-UCB.

Next we give an instance-dependent regret bound for PE. Notice that this bound is even new for stochastic linear bandits (with finite action sets). See Appendix B for the proof.

Theorem 4.3 (Instance-dependent regret bound for PE)**.** *For all* $\delta \in (0,1)$*, the policy* $\mathrm{PE}(\delta)$ *given by Algorithm 2 satisfies the following regret for all conditionally benign environments* ν*,*

$$\mathrm{Reg}(T) \le C \cdot \frac{d_\nu \log(|\mathcal{A}|\log T/\delta)}{\Delta_{\min}(\nu)}$$

with probability at least $1-\delta$*, where* $\Delta_{\min}(\nu) := \min_{a\neq a^*} \mu^{\mathcal{A}}(a^*) - \mu^{\mathcal{A}}(a)$ *is the minimal sub-optimality gap of instance* ν *and* $C > 0$ *is a universal constant. In particular, taking* $\delta = 1/T$*,*

$$\mathbb{E}_{\nu,\mathrm{PE}(\delta)}[\mathrm{Reg}(T)] = O\left(\frac{d_\nu \log(|\mathcal{A}|T)}{\Delta_{\min}(\nu)}\right).$$

Remark 4.4. During the implementation of Algorithm 2, it is possible that $\mathcal{A}_\ell$ cannot span $\mathbb{R}^{|\mathcal{Z}|}$ for some ℓ such that $V(\pi_\ell)$ is singular for any $\pi_\ell \in \mathcal{P}(\mathcal{A}_\ell)$. For example, in later phases $|\mathcal{A}_\ell|$ can be smaller than $|\mathcal{Z}|$. Let's say $\dim(\mathrm{span}\{\nu_a : a \in \mathcal{A}_\ell\}) = r < |\mathcal{Z}|$. One workaround is to apply some invertible matrix $X \in \mathbb{R}^{|\mathcal{Z}|\times|\mathcal{Z}|}$ to every $a \in \mathcal{A}_\ell$ such that $X\nu_a$ can be decomposed to a dim-r vector $(X\nu_a)_{[r]}$ and a tail of $(|\mathcal{Z}|-r)$ zeros, and $\big\{(X\nu_a)_{[r]} : a \in \mathcal{A}_\ell\big\}$ can span $\mathbb{R}^r$. Now we use $\big\{(X\nu_a)_{[r]} : a \in \mathcal{A}_\ell\big\}$ as our active set in phase ℓ and the analysis would go through.

4.3. Roadblocks: Instance-Dependent Bounds

Unlike adaptive worst-case regret studied in Section 3, adaptive instance-dependent regret is less understood and a general theory is still absent in the literature. In particular, we do not know if $O(\log T)$ regret can always be achieved, and whenever achieved, whether it is tight. These issues are illustrated for model selection methods in the following. First, it is easy to see that $O(\log T)$ regret can always be achieved in benign environments, e.g., by corralling PE and UCB using dynamic balancing, because in this case both base learners admit logarithmic regret. However, the regret bound of UCB is dominant and thus naive calculation only leads to a $O(|\mathcal{A}|\log T/\Delta_{\min})$ regret for DB. It remains open whether we can adapt to the smaller regret achieved by PE in benign environments. Second, $O(\log T)$ regret is not always granted by model selection in non-benign instances. The only exception we are aware of in the literature is the case where the causal base learner is assumed to incur linear regret whenever its candidate regret bound fails (Cutkosky et al., 2021, Theorem 31). If this type of "algorithm gap" holds, DB will only choose the causal base learner on a $O(\log T)$ number of rounds, and hence enjoy logarithmic regret. Moreover, without changing the parameter setting, DB is able to realize the Pareto optimal rates $(\sqrt{|\mathcal{Z}|T}, |\mathcal{A}|\sqrt{T}/\sqrt{|\mathcal{Z}|})$ up to log factors. However, the "algorithmic gap" requirement on the causal base learner is so stringent that we do not know if it is met by any algorithm

Algorithm 2 Phased Elimination (PE) in Causal Bandit

Input: Action set $\mathcal{A}$, marginals $\{\nu_a : a \in \mathcal{A}\}$, $d_\nu = \dim(\mathrm{span}\{\nu_a : a \in \mathcal{A}\})$, and confidence level $\delta \in (0,1)$

1. Set $\ell = 1$ and let the initial active set $\mathcal{A}_1$ be $\mathcal{A}$
2. Find some near-optimal design $\pi_\ell \in \mathcal{P}(\mathcal{A}_\ell)$ with $\max_{a\in\mathcal{A}_\ell} \|\nu_a\|^2_{V(\pi_\ell)^{-1}} \le 2d_\nu$ and $|\mathrm{supp}(\pi_\ell)| \le 4d_\nu \log\log(d_\nu) + 16$, where $V(\pi_\ell) = \sum_{a\in\mathcal{A}_\ell} \pi_\ell(a)\nu_a\nu_a^\top$
3. Let $m_\ell = 2^{\ell-1}(4d_\nu \log\log(d_\nu) + 16)$
   Compute $T_\ell(a) = \lceil m_\ell \pi_\ell(a)\rceil$ and $T_\ell = \sum_{a\in\mathcal{A}_\ell} T_\ell(a)$
4. Play each action $a \in \mathcal{A}_\ell$ exactly $T_\ell(a)$ times and we call these T_ℓ rounds *phase* ℓ. We also observe corresponding context-reward pairs $(Z_t, Y_t)_{t\in\text{phase }\ell}$
5. Compute the empirical estimate:
   $\hat{\mu}_\ell^{\mathcal{Z}} = V_\ell^{-1} \sum_{t\in\text{phase }\ell} \nu_{A_t} Y_t$ where
   $V_\ell = \sum_{a\in\mathcal{A}_\ell} T_\ell(a)\nu_a\nu_a^\top$
6. Eliminate low rewarding actions and update the active set:
   $\mathcal{A}_{\ell+1} = \Big\{a \in \mathcal{A}_\ell : \max_{b\in\mathcal{A}_\ell} \langle \hat{\mu}_\ell^{\mathcal{Z}}, \nu_b - \nu_a\rangle \le 2\sqrt{\frac{4d_\nu}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)}\Big\}$
7. $\ell \leftarrow \ell + 1$ and **Goto 2**

in every instance. In Appendix E, we show that a version of PE incurs linear regret on some instances.

5. Limited Knowledge of the Marginal Distributions over Context Variables

So far, we have assumed that algorithms knows the marginal distribution over the post-action context for each arm. Of course, perfect knowledge of these marginals may not hold in practice. What is the effect of only having access to approximate marginals on achievable rates of regret?

In this section, we study this question. We give a lower bound indicating that, with zero access to the marginals, it is impossible for any algorithm to exploit the causal structure and beat the minimax rate of an arbitrary environment. To model this setting, recall that algorithms are defined as mappings taking $(\mathcal{A}, \mathcal{Z}, T, q)$ to policies. So naturally, algorithms considered agnostic to the marginals should be constant in $q \in \mathcal{P}(\mathcal{Z})^{\mathcal{A}}$, leading to the following definition:

Definition 5.1. An algorithm $\mathfrak{a}$ is said to be *agnostic to marginals* if, for any $\mathcal{A}$, $\mathcal{Z}$ and T, the map

$$\mathfrak{a}_{\mathcal{A},\mathcal{Z},T} : q \mapsto \mathfrak{a}(\mathcal{A}, \mathcal{Z}, T, q)$$

is constant over $\mathcal{P}(\mathcal{Z})^{\mathcal{A}}$. We denote the set of all such algorithms by $\mathbb{A}_{\text{agnostic}}$.

Examples of algorithms from $\mathbb{A}_{\text{agnostic}}$ include not only heuristic non-causal algorithms like UCB, but also versions of causal algorithms that are always input by the same marginals. For all algorithm $\mathfrak{a} \in \mathbb{A}_{\text{agnostic}}$, we will write the policy it induces given $\mathcal{A}, \mathcal{Z}, T$ as $\mathfrak{a}(\mathcal{A}, \mathcal{Z}, T, \cdot)$ to highlight its independence on the q component. Our lower bound shows that, under this zero-marginal-knowledge regime, we cannot do better than the optimal non-causal algorithm.

Theorem 5.2. *For all $\mathcal{A}, \mathcal{Z}, T \geq |\mathcal{A}|$ and MAB algorithms $\mathfrak{a} \in \mathbb{A}_{\text{agnostic}}$, there exists a conditionally benign environment $\nu \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ such that*

$$\mathbb{E}_{\nu,\mathfrak{a}}[\text{Reg}(T)] \geq c\sqrt{|\mathcal{A}|T},$$

where $c > 0$ is a universal constant.

See Appendix D.2 for the proof.

Remark 5.3. Our lower bound improves on Lu et al. (2020, Theorem 4), which is of the form of $C_\varepsilon \sqrt{|\mathcal{A}|}T^{1/2-\varepsilon}, \forall \varepsilon > 0$ and only holds for *some* set of non-causal algorithms, which is a strict subset of $\mathbb{A}_{\text{agnostic}}$.

5.1. Phased Elimination with Approximate Marginals

Despite the negative result Theorem 5.2, we now argue that some level of misspecification is allowed in the prior knowledge of marginals. Upon interacting with environment ν, suppose we are given some marginal $\tilde{\nu}(Z) \in \mathcal{P}(\mathcal{Z})^{\mathcal{A}}$ which may deviate from the true $\nu(Z)$ to some extent. Now we show that even instantiating PE with the possibly non-accurate $\tilde{\nu}(Z)$ may yield $\tilde{O}(\sqrt{T})$ regret, following a similar result for C-UCB by Bilodeau et al. (2022). First we need the the following definition to measure the amount of deviation of $\tilde{\nu}(Z)$ from $\nu(Z)$.

Definition 5.4. (Bilodeau et al., 2022, Definition 4.2) For any $\varepsilon \geq 0$, $\tilde{\nu}(Z)$ and $\nu(Z)$ are said to be ε-close if

$$\sup_{a \in \mathcal{A}} \sum_{z \in \mathcal{Z}} |\tilde{\nu}_a(z) - \nu_a(z)| \leq \varepsilon.$$

Due to our reduction in Section 4.1, we can find that causal bandits with misspecified marginals is reduced to the well-studied misspecified linear bandits, which yields the following regret bound that subsumes Theorem 4.1. The proof is largely based on the analysis of phased elimination in Lattimore et al. (2020, Proposition 5.1), with necessary modifications for handling conditionally sub-gaussian noises and providing an anytime regret bound. See Appendix B for details.

Theorem 5.5 (Worst-case regret bound, with approximate marginal distributions)**.** *In any conditionally environment ν suppose we instantiate $\text{PE}(\delta)$ with $\tilde{\nu}(Z)$. If $\tilde{\nu}(Z)$ and $\nu(Z)$ are $\varepsilon-$close, then with probability at least $1-\delta$, the regret of $\text{PE}(\delta)$ is bounded for all rounds $t \in [T]$ by*

$$\text{Reg}(t) \leq C\left(\sqrt{d_{\tilde{\nu}} \log\left(\frac{|\mathcal{A}| \log T}{\delta}\right) t} + \varepsilon t \sqrt{d_{\tilde{\nu}}} \log T\right),$$

where $C > 0$ is a universal constant and $d_{\tilde{\nu}} = \dim(\text{span}\{\tilde{\nu}_a : a \in \mathcal{A}\})$

It is implied that $\varepsilon = \tilde{O}(\sqrt{1/T})$ suffices to recover all aforementioned regret guarantees of phased elimination and dynamic balancing. On the other hand, such numerical requirement on ε is almost necessary for us to avoid the lower bound in Theorem 5.2: from the proof of Theorem 5.2 we will find that when $\varepsilon = \Omega(\sqrt{|\mathcal{A}|/T})$, for any algorithm there exists a conditionally benign environment ν and approximate marginal $\tilde{\nu}(Z)$ such that $\tilde{\nu}(Z)$ and $\nu(Z)$ are ε-close, but this algorithm would incur $\Omega(\sqrt{|\mathcal{A}|T})$ regret on ν when it is input by $\tilde{\nu}(Z)$.

It is worth mentioning that the $\sqrt{d_{\tilde{\nu}}}$ factor in the misspecification term cannot be improved in many regimes for linear bandit algorithms (Lattimore et al., 2020). However, C-UCB is able to shave this factor off (Bilodeau et al., 2022, Theorem 4.3) by utilizing realized contexts rather than the least-square estimate of the mean reward vector $\mu^{\mathcal{Z}}$. From this perspective, we see there is a price for pursuing better instance-dependent result by ignoring the context information.

6. Conclusions and Discussions

We provide a comprehensive characterization of the Pareto regret frontier for the bandit problem in the context of adapting to causal structure whenever feasible. We also give the first instance-dependent regret bound under conditionally benign environments, based on our novel causal-to-linear reduction. Finally, we show that the common assumption that we have access to the true marginals is necessary in general but still can be relaxed in some cases.

For future works, it would be important to focus on the design of algorithms that are easier to implement compared to running dynamic balancing over some base learners. On the theoretical side, it would be interesting to investigate other causal bandit scenarios involving adaptivity in light of our Pareto regret frontier. For example, we may define a series of "semi-benign" settings interpolating conditionally benign environments and non-benign environments and study the Pareto regret frontier thereof.

Acknowledgements

ZL is supported by the Vector Research Grant at the Vector Institute. IA is supported by the Vatat Scholarship from the Israeli Council for Higher Education. DMR is supported by an NSERC Discovery Grant and funding through his Canada CIFAR AI Chair at the Vector Institute. The authors would like to thank Tomer Koren, Blair Bilodeau and Csaba Szepesvári for helpful discussions at different stages of this work.

References

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pp. 12–38. PMLR, 2017.

Arora, R., Marinov, T. V., and Mohri, M. Corralling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 2116–2124. PMLR, 2021.

Bareinboim, E., Forney, A., and Pearl, J. Bandits with unobserved confounders: A causal approach. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28, 2015.

Bilodeau, B., Wang, L., and Roy, D. Adaptively exploiting d-separators with causal bandits. *Advances in Neural Information Processing Systems*, 35:20381–20392, 2022.

Cutkosky, A., Das, A., and Purohit, M. Upper confidence bounds for combining stochastic bandits. *arXiv preprint arXiv:2012.13115*, 2020.

Cutkosky, A., Dann, C., Das, A., Gentile, C., Pacchiano, A., and Purohit, M. Dynamic balancing for model selection in bandits and rl. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2276–2285. PMLR, 2021.

Koolen, W. M. The pareto regret frontier. *Advances in Neural Information Processing Systems*, 26, 2013.

Lattimore, F., Lattimore, T., and Reid, M. D. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.

Lattimore, T. The pareto regret frontier for bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

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

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in RL with a generative model. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5662–5670. PMLR, 2020.

Lu, Y., Meisami, A., Tewari, A., and Yan, W. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pp. 141–150. PMLR, 2020.

Lu, Y., Meisami, A., and Tewari, A. Causal bandits with unknown graph structure. *Advances in Neural Information Processing Systems*, 34:24817–24828, 2021.

Malek, A., Aglietti, V., and Chiappa, S. Additive causal bandits with unknown graph. *arXiv preprint arXiv:2306.07858*, 2023.

Marinov, T. V. and Zimmert, J. The pareto frontier of model selection for general contextual bandits. *Advances in Neural Information Processing Systems*, 2021.

Nair, V., Patil, V., and Sinha, G. Budgeted and non-budgeted causal bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2017–2025. PMLR, 2021.

Pacchiano, A., Dann, C., Gentile, C., and Bartlett, P. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020a.

Pacchiano, A., Phan, M., Abbasi Yadkori, Y., Rao, A., Zimmert, J., Lattimore, T., and Szepesvari, C. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33, 2020b.

Sen, R., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, pp. 3057–3066. PMLR, 2017.

Xiong, N. and Chen, W. Pure exploration of causal bandits. *arXiv preprint arXiv:2206.07883*, 2022.

Zhu, Y. and Nowak, R. Pareto optimal model selection in linear bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

A. Regret Analysis for Dynamic Balancing

In this section we show that the regret guarantees of dynamic balancing in Cutkosky et al. (2021) can be generalized to our problem and provide a proof of our main upper bound Theorem 3.5.

Notations. For base learner i, we use $\mathrm{CandidReg}_i(t)$ to denote its candidate anytime regret bound that is expected to hold in its favorable settings. Throughout we consider $\mathrm{CandidReg}_i(t)$ with the form of $d_i\sqrt{t}$, where d_i implicitly depends on the confidence parameter δ. Let i_t be the index of the base learner selected in round t. $U_i(t) = \sum_{s=1}^{t} Y_s \mathbb{I}\{i = i_s\}$ is the observed cumulative reward in the first t rounds where i is picked, and $n_i(t) = \sum_{s=1}^{t} \mathbb{I}\{i = i_s\}$ is the number of rounds i is picked by the end of round t. The local regret of i up to round t is $\mathrm{Reg}_i(t) = n_i(t)\mu^{\mathcal{A}}(a^*) - U_i(t)$. We say learner i is *well-specified* if $\mathrm{Reg}_i(t) \le \mathrm{CandidReg}_i(n_i(t)) = d_i\sqrt{n_i(t)}, \forall t \in [T]$ and otherwise it is *misspecified*. We use $i_\star$ to denote any well-specified learner.

A.1. Preliminaries

Roughly speaking, in each round t, dynamic balancing works by (1) running a misspecification test to temporarily de-activate misspecified base learners and (2) picking the learner i_t with minimal putative regret $d_i\sqrt{n_i(t)}$ among all active learners i in this round. In this way, the regret incurred by DB is comparable to that of the best well-specified learner.

Notice that dynamic balancing was initiated with stochastic contextual bandits (where contexts are revealed *prior to* actions) in Cutkosky et al. (2021). To see that DB can also be applied in stochastic bandits with post-action contexts, it is worth identifying several important features of DB:

1. First of all, the meta decision by DB on each round t only depends on the global information, i.e. $U_i(t)$ and $n_i(t)$ (as well as user-specified d_i, b_i and v_i). In particular, it does not need any information regarding context variables or internal states of base learners.
2. Second, DB only updates the selected base learner i_t in each round t, and the update only uses the reward and contextual information observed in this round, where the context can be either pre-action or post-action, or both. Thus the regret guarantees of DB would hold regardless of the nature of contexts given that the internal updates of base learners are not affected.

Therefore, the essence of dynamic balancing does not rely on what kind of (stochastic) contextual information can be observed in the underlying (stochastic) bandit problem due to above observations.

Now we state the worst-case regret bound of DB in Cutkosky et al. (2021) adapted to our setting. First define the good event

$$\mathcal{E}(\delta) = \left\{\forall i \in \{1,2\}, \forall t \in T : |n_i(t)\mu^{\mathcal{A}}(a^*) - U_i(t) - \mathrm{Reg}_i(t)| \le c\sqrt{n_i(t)\log\left(\frac{2\log n_i(t)}{\delta}\right)}\right\}$$

on which we are able to control the regret of DB. According to the analysis of Cutkosky et al. (2021, Lemma 5), we can fix c to be some absolute constant (which can be actually set to 3 in our setting) such that $\mathbb{P}_{\nu,\pi}[\mathcal{E}(\delta)] \ge 1 - \delta$ for any $\nu \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ and $\pi \in \Pi(\mathcal{A}, \mathcal{Z}, T)$. Conditioning on $\mathcal{E}(\delta)$, we have the following regret bound:

Proposition A.1 (Adapted version of Theorem 22 in Cutkosky et al. (2021))**.** *Let $\mathfrak{a}_1$ be a d_1-benign family and let $\mathfrak{a}_2$ be a d_2-arbitrary family of learners. Let Z_1, Z_2 be arbitrary positive real numbers. For all $\delta \in (0,1)$, we can set hyper-parameters*

$$b_i(t) = \max\left\{\frac{2Z_i}{\sqrt{t}}, \frac{3\sqrt{2\log(2\log t/\delta)}}{\sqrt{t}}\right\}, \quad v_i = \sqrt{\frac{Z_i}{d_i(\delta)^3}}$$

in dynamic balancing such that, the policy $\mathrm{DB}(\delta)$ given by dynamic balancing with $\mathfrak{a}_1, \mathfrak{a}_2, d_1, d_2$ satisfies the following: for all instances ν, conditioning on $\mathcal{E}(\delta)$ and the existence of a well-specified base learner $i_\star$, the regret of $\mathrm{DB}(\delta)$ is bounded by

$$\mathrm{Reg}(T) \le \mathrm{Reg}_{i_\star}(T) + C'\left(\sqrt{\log\left(\frac{\log T}{\delta}\right)} + Z_{i_\star} d_{i_\star}(\delta) + \sum_{i \ne i_\star} \frac{d_i(\delta)}{Z_i}\right)\sqrt{T},$$

where $C' > 0$ is a universal constant.

It is straightforward to see that Proposition A.1 is obtained by taking $M = 2, C = 1, c = 3, \beta = 1/2$, and $W_1 = W_2 = \sqrt{2}$ in Cutkosky et al. (2021, Theorem 22).

A.2. Proof of Theorem 3.5

Theorem 3.5 is the immediate consequence of the following regret bound, which is derived by instantiating Z_1, Z_2 in Proposition A.1 with specific values.

Proposition A.2. *For every pair of reasonable rate functions $R_1(T), R_2(T)$ such that $R_1(T)R_2(T) \geq |\mathcal{A}|T$, we can instantiate Proposition A.1 with $Z_1 = 1, Z_2 = \frac{R_2(T)}{\sqrt{|\mathcal{A}|T}}$ such that for all $\delta \in (0,1)$, the policy $\mathrm{DB}(\delta)$ with the same setup as Proposition A.1 satisfies the following: for all instances ν, with probability at least $1 - O(\delta)$, the regret of $\mathrm{DB}(\delta)$ is bounded by*

$$\mathrm{Reg}(T) \leq C'\left(d_1(\delta) + \sqrt{\log\left(\frac{\log T}{\delta}\right)} + \frac{d_2(\delta)}{\sqrt{|\mathcal{A}|T}} R_1(T)\right)\sqrt{T}, \quad \textit{if } \nu \textit{ is conditionally benign;}$$

$$\mathrm{Reg}(T) \leq C'\left(d_1(\delta) + \sqrt{\log\left(\frac{\log T}{\delta}\right)} + \frac{d_2(\delta)}{\sqrt{|\mathcal{A}|T}} R_2(T)\right)\sqrt{T}, \quad \textit{if } \nu \textit{ is arbitrary,}$$

where $C' > 0$ is a universal constant.

Now we can see that our main upper bound Theorem 3.5 is proved immediately after taking $d_1(\delta) = O\Big((\sqrt{|\mathcal{Z}|} + \sqrt{\log(T/\delta)})\sqrt{\log(|\mathcal{Z}|T/\delta)}\Big)$, $d_2(\delta) = O(\sqrt{|\mathcal{A}|\log(|\mathcal{A}|T/\delta)})$ and $\delta = 1/T$.

Proof of Proposition A.2. By Definition 3.3, we know that for all conditionally instances ν, with probability at least $1-O(\delta)$, learner $\mathfrak{a}_1$ is well-specified with $\mathrm{CandidReg}_1(t) = d_1(\delta)\sqrt{t}$ and the regret bound in Proposition A.1 holds with $i_\star = 1$. Plugging in $Z_1 = 1, Z_2 = \frac{R_2(T)}{\sqrt{|\mathcal{A}|T}}$, the regret of $\mathrm{DB}(\delta)$ is bounded by

$$\mathrm{Reg}(T) \leq C'\left(2d_1(\delta) + \sqrt{\log\left(\frac{\log T}{\delta}\right)} + d_2(\delta)\frac{\sqrt{|\mathcal{A}|T}}{R_2(T)}\right)\sqrt{T}.$$

Similarly for all instances ν, with probability at least $1 - O(\delta)$, learner $\mathfrak{a}_2$ is well-specified with $\mathrm{CandidReg}_2(t) = d_2(\delta)\sqrt{t}$ and the regret bound in Proposition A.1 holds with $i_\star = 2$, which is

$$\mathrm{Reg}(T) \leq C'\left(d_2(\delta) + \sqrt{\log\left(\frac{\log T}{\delta}\right)} + d_2(\delta)\frac{R_2(T)}{\sqrt{|\mathcal{A}|T}} + d_1(\delta)\right)\sqrt{T}.$$

By our assumption that $(R_1(T), R_2(T))$ is reasonable and $R_1(T)R_2(T) \geq |\mathcal{A}|T$, we have that $R_2(T) \geq \sqrt{|\mathcal{A}|T}$ and $\frac{|\mathcal{A}|T}{R_2(T)} \leq R_1(T)$. Hence the regret of $\mathrm{DB}(\delta)$ for all instances ν is further bounded by

$$\mathrm{Reg}(T) \leq C'\left(d_1(\delta) + \sqrt{\log\left(\frac{\log T}{\delta}\right)} + \frac{d_2(\delta)}{\sqrt{|\mathcal{A}|T}} R_1(T)\right)\sqrt{T}, \quad \text{if } \nu \text{ is conditionally benign;}$$

$$\mathrm{Reg}(T) \leq C'\left(d_1(\delta) + \sqrt{\log\left(\frac{\log T}{\delta}\right)} + \frac{d_2(\delta)}{\sqrt{|\mathcal{A}|T}} R_2(T)\right)\sqrt{T}, \quad \text{if } \nu \text{ is arbitrary,}$$

which completes the proof. □

B. Regret analysis of phased elimination

In this section we will prove Theorem 4.3 and Theorem 5.5, while Theorem 4.1 is implied by taking $\varepsilon = 0$ in Theorem 5.5. Recall that the proof of Theorem 5.5 is based on the analysis of phased elimination in Lattimore et al. (2020, Proposition 5.1).

For simplicity we will use $\mathbb{P}$ and $\mathbb{E}$ to denote the probabilistic operators determined jointly by the underlying conditionally benign environment ν and the phased elimination algorithm. Also we use $\Delta_a, \Delta_{\min}$ to denote the true sub-optimality gap $\Delta_a(\nu) = \mu^{\mathcal{A}}(a^*) - \mu^{\mathcal{A}}(a)$ and minimal sub-optimality gap $\Delta_{\min}(\nu) = \min_{a\in\mathcal{A}} \mu^{\mathcal{A}}(a^*) - \mu^{\mathcal{A}}(a)$ respectively with regards to the underlying instance ν.

B.1. Prerequisite

Lemma B.1. *(In-phase concentration) For any phase ℓ, let*

$$E_\ell^{\text{phase}}(\delta) = \left\{ |\langle \hat{\mu}_\ell^{\mathcal{Z}} - \mu^{\mathcal{Z}}, \tilde{\nu}_a\rangle| \le 2\varepsilon\sqrt{d_{\tilde{\nu}}} + \sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)}, \forall a \in \mathcal{A}_\ell \right\}$$

and $\mathscr{F}_\ell$ be the $\sigma-$algebra generated by the history up to the start of phase ℓ. Then $\mathbb{P}[E_\ell^{\text{phase}}(\delta)|\mathscr{F}_\ell] \ge 1 - \frac{\delta}{\log_2(T)}$.

Proof of Lemma B.1. Let $b_a = \langle \nu_a - \tilde{\nu}_a, \mu^{\mathcal{Z}}\rangle, \forall a \in \mathcal{A}$ be the error term due to the use of inaccurate marginals, then we know that $|b_a| \le \varepsilon, \forall a \in \mathcal{A}$ since $\nu(Z)$ and $\tilde{\nu}(Z)$ are $\varepsilon-$close. Observe that

$$\begin{aligned}
\langle \hat{\mu}_\ell^{\mathcal{Z}} - \mu^{\mathcal{Z}}, \tilde{\nu}_a\rangle &= \langle V_\ell^{-1} \sum_{t\in\text{phase }\ell} \tilde{\nu}_{A_t}\tilde{\nu}_{A_t}^\top \mu^{\mathcal{Z}}, \tilde{\nu}_a\rangle - \langle \mu^{\mathcal{Z}}, \tilde{\nu}_a\rangle \\
&\quad + \langle V_\ell^{-1} \sum_{t\in\text{phase }\ell} \tilde{\nu}_{A_t}\eta_t^{\mathcal{A}}, \tilde{\nu}_a\rangle + \langle V_\ell^{-1} \sum_{t\in\text{phase }\ell} \tilde{\nu}_{A_t} b_{A_t}, \tilde{\nu}_a\rangle \\
&= \sum_{t\in\text{phase }\ell} \langle V_\ell^{-1}\tilde{\nu}_{A_t}, \tilde{\nu}_a\rangle \eta_t^{\mathcal{A}} + \sum_{t\in\text{phase }\ell} \langle V_\ell^{-1}\tilde{\nu}_{A_t}, \tilde{\nu}_a\rangle b_{A_t}.
\end{aligned}$$

Using Cauchy-Schwarz inequality and the fact that for all $a \in \mathcal{A}_\ell$, $\|\tilde{\nu}_a\|_{V_\ell^{-1}}^2 \le \frac{1}{m_\ell}\|\tilde{\nu}_a\|_{V(\pi_\ell)^{-1}}^2 \le \frac{2d_{\tilde{\nu}}}{m_\ell}$, the second term on the RHS of the above equality can be bounded by

$$\begin{aligned}
\left| \sum_{t\in\text{phase }\ell} \langle V_\ell^{-1}\tilde{\nu}_{A_t}, \tilde{\nu}_a\rangle b_{A_t} \right| &\le \varepsilon \sum_{t\in\text{phase }\ell} \left|\langle V_\ell^{-1}\tilde{\nu}_{A_t}, \tilde{\nu}_a\rangle\right| \\
&\le \varepsilon \sqrt{\left(\sum_{t\in\text{phase }\ell} 1\right)\left(\sum_{t\in\text{phase }\ell} \langle V_\ell^{-1}\tilde{\nu}_{A_t}, \tilde{\nu}_a\rangle^2\right)} \\
&= \varepsilon\sqrt{T_\ell \|\tilde{\nu}_a\|_{V_\ell^{-1}}^2} \le \varepsilon\sqrt{2m_\ell \frac{2d_{\tilde{\nu}}}{m_\ell}} = 2\varepsilon\sqrt{d_{\tilde{\nu}}}.
\end{aligned}$$

To bound the first term, notice that $(A_t)_{t\in\text{phase }\ell}, V_\ell$ are fixed given the history prior to the start of phase ℓ. Hence $(\eta_t^{\mathcal{A}})_{t\in\text{phase }\ell}$ are independent conditioned on $\mathscr{F}_\ell$ and bounded by $[-1, 1]$. By standard concentration bounds, we have that with probability at least $1 - \frac{\delta}{|\mathcal{A}|\log_2(T)}$,

$$\left| \sum_{t\in\text{phase }\ell} \langle V_\ell^{-1}\nu_{A_t}, \tilde{\nu}_a\rangle \eta_t^{\mathcal{A}} \right| \le \sqrt{2\sum_{t\in\text{phase }\ell} \langle V_\ell^{-1}\tilde{\nu}_{A_t}, \tilde{\nu}_a\rangle^2 \log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)},$$

where the RHS can be rewritten as

$$\sqrt{2\|\tilde{\nu}_a\|_{V_\ell^{-1}}^2 \log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)} \le \sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)}.$$

Combining the two upper bounds above and taking a union bound over all $a \in \mathcal{A}_\ell$, we have that with probability at least $1 - \frac{\delta}{\log_2(T)}$,

$$|\langle \hat{\mu}_\ell^{\mathcal{Z}} - \mu^{\mathcal{Z}}, \tilde{\nu}_a\rangle| \le 2\varepsilon\sqrt{d_{\tilde{\nu}}} + \sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)}, \quad \forall a \in \mathcal{A}_\ell,$$

which finishes the proof. □

Since the marginal distributions $\tilde{\nu}_a$ are possibly not accurate, we may not be able to show that the optimal action a^* is never eliminated with high probability. So what we can hope for is that actions that are near-optimal *relative to* the best action in $\mathcal{A}_\ell$ are retained in the end of the phase ℓ. To be concrete, define $a_\ell^* \in \operatorname{argmin}_{a\in\mathcal{A}_\ell} \Delta_a$ to be the true optimal action within $\mathcal{A}_\ell$. Then we can show that $\Delta_a - \Delta_{a_\ell^*}$ is rather small for any a that is not eliminated in the end of phase ℓ.

Lemma B.2. *Conditioning on event $E_\ell^{\text{phase}}(\delta)$, for any action a not eliminated in the end of phase ℓ, it has relative sub-optimality gap $\langle \mu^{\mathcal{Z}}, \nu_{a_\ell^*} - \nu_a\rangle = \Delta_a - \Delta_{a_\ell^*} \le 2\varepsilon(1+2\sqrt{d_{\tilde{\nu}}}) + 4\sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)}$.*

Proof of Lemma B.2. According to the rule of updating active set, whenever $a \in \mathcal{A}_\ell$ is not eliminated at the end of phase ℓ, it holds

$$\langle \hat{\mu}_\ell^{\mathcal{Z}}, \tilde{\nu}_{a_\ell^*} - \tilde{\nu}_a\rangle \le \max_{b\in\mathcal{A}_\ell}\langle \hat{\mu}_\ell^{\mathcal{Z}}, \tilde{\nu}_b - \tilde{\nu}_a\rangle \le 2\sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)}.$$

It implies that

$$\begin{aligned}\langle \mu^{\mathcal{Z}}, \tilde{\nu}_{a_\ell^*} - \tilde{\nu}_a\rangle &= \langle \mu^{\mathcal{Z}} - \hat{\mu}_\ell^{\mathcal{Z}}, \tilde{\nu}_{a_\ell^*} - \tilde{\nu}_a\rangle + \langle \hat{\mu}_\ell^{\mathcal{Z}}, \tilde{\nu}_{a_\ell^*} - \tilde{\nu}_a\rangle \\ &\le 2\Big(\sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)} + 2\varepsilon\sqrt{d_{\tilde{\nu}}}\Big) + 2\sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)} \\ &= 4\sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)} + 4\varepsilon\sqrt{d_{\tilde{\nu}}}.\end{aligned}$$

where we use the fact that we are conditioning on $E_\ell^{\text{phase}}(\delta)$ in the inequality. Hence under the true marginals ν,

$$\langle \mu^{\mathcal{Z}}, \nu_{a_\ell^*} - \nu_a\rangle \le 4\sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)} + 4\varepsilon\sqrt{d_{\tilde{\nu}}} + 2\varepsilon.$$

□

Now we need to track $\Delta_{a_\ell^*}$, the sub-optimality of the best active action in each phase. Observe that $\Delta_{a_\ell^*} = \sum_{k=1}^{\ell-1}(\Delta_{a_{k+1}^*} - \Delta_{a_k^*})$ since $\Delta_{a_1^*} = \Delta_{a^*} = 0$. Then it suffices to control each $\Delta_{a_{k+1}^*} - \Delta_{a_k^*}, k \le \ell-1$, to control the growth of $\Delta_{a_\ell^*}$.

Lemma B.3. *Conditioning on event $E_\ell^{\text{phase}}(\delta)$, we have $\Delta_{a_{\ell+1}^*} - \Delta_{a_\ell^*} \le 2\varepsilon(1+2\sqrt{d_\nu})$.*

Proof of Lemma B.3. Suppose $E_\ell^{\text{phase}}(\delta)$ happens. Notice that the results holds trivially if a_ℓ^* is not eliminated in the end of phase ℓ, because in this case $a_{\ell+1}^* = a_\ell^*$. On the other hand, if a_ℓ^* is eliminated, define $\hat{a}_\ell \in \operatorname{argmax}_{a\in\mathcal{A}_\ell}\langle \hat{\mu}_\ell^{\mathcal{Z}}, \tilde{\nu}_a\rangle$ to be the empirically best action in the end of phase ℓ and then we have

$$\langle \hat{\mu}_\ell^{\mathcal{Z}}, \tilde{\nu}_{\hat{a}_\ell} - \tilde{\nu}_{a_\ell^*}\rangle > 2\sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)},$$

according to the test performed. In the meantime, recall that due to in-phase concentration and $\varepsilon-$closeness between $\tilde{\nu}$ and ν,

$$\begin{aligned}\langle \hat{\mu}_\ell^{\mathcal{Z}}, \tilde{\nu}_{\hat{a}_\ell} - \tilde{\nu}_{a_\ell^*}\rangle &\le \langle \mu^{\mathcal{Z}}, \tilde{\nu}_{\hat{a}_\ell} - \tilde{\nu}_{a_\ell^*}\rangle + 4\varepsilon\sqrt{d_{\tilde{\nu}}} + 2\sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)} \\ &\le \langle \mu^{\mathcal{Z}}, \nu_{\hat{a}_\ell} - \nu_{a_\ell^*}\rangle + 2\varepsilon + 4\varepsilon\sqrt{d_{\tilde{\nu}}} + 2\sqrt{\frac{4d_{\tilde{\nu}}}{m_\ell}\log\Big(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\Big)}.\end{aligned}$$

Hence we get

$$\Delta_{\hat{a}_\ell} - \Delta_{a_\ell^*} = \langle \mu^{\mathcal{Z}}, \nu_{a_\ell^*} - \nu_{\hat{a}_\ell}\rangle \le 2\varepsilon + 4\varepsilon\sqrt{d_{\tilde{\nu}}}$$

and

$$\Delta_{a_{\ell+1}^*} - \Delta_{a_\ell^*} \le \Delta_{\hat{a}_\ell} - \Delta_{a_\ell^*} \le 2\varepsilon + 4\varepsilon\sqrt{d_{\tilde{\nu}}}.$$

□

Corollary B.4. *For any $\ell \geq 2$ and conditioning on $\bigcap_{k\leq \ell-1} E_k^{\text{phase}}(\delta)$, we have that $\Delta_{a_\ell^*} \leq 2\varepsilon(\ell-1)(1+2\sqrt{d_{\tilde{\nu}}})$ and $\Delta_a \leq 2\varepsilon(1+2\sqrt{d_{\tilde{\nu}}}) + 4\sqrt{\frac{4d_{\tilde{\nu}}}{m_{\ell-1}}\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)} + 2\varepsilon(\ell-2)(1+2\sqrt{d_{\tilde{\nu}}})$ for all $a \in \mathcal{A}_\ell$.*

Proof of Corollary B.4. By conditioning on the intersection of all $E_k^{\text{phase}}(\delta), k \leq \ell-1$, we have that

$$\Delta_{a_{k+1}^*} - \Delta_{a_k^*} \leq 2\varepsilon(1+2\sqrt{d_{\tilde{\nu}}}), \forall k \leq \ell-1,$$

which implies that $\Delta_{a_s^*} = \sum_{k=1}^{s-1}(\Delta_{a_{k+1}^*} - \Delta_{a_k^*}) \leq 2\varepsilon(s-1)(1+2\sqrt{d_{\tilde{\nu}}}), \forall s \leq \ell$. In particular, there is

$$\Delta_{a_\ell^*} \leq 2\varepsilon(\ell-1)(1+2\sqrt{d_{\tilde{\nu}}}).$$

Since every action $a \in \mathcal{A}_\ell$ passes the test in the end of $(\ell-1)-$th phase and hence is not eliminated, by Lemma B.2 we know

$$\Delta_a - \Delta_{a_{\ell-1}^*} \leq 2\varepsilon(1+2\sqrt{d_{\tilde{\nu}}}) + 4\sqrt{\frac{4d_{\tilde{\nu}}}{m_{\ell-1}}\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)}.$$

Therefore, for all $a \in \mathcal{A}_\ell$,

$$\Delta_a = \Delta_a - \Delta_{a_{\ell-1}^*} + \Delta_{a_{\ell-1}^*} \leq 2\varepsilon(1+2\sqrt{d_{\tilde{\nu}}}) + 4\sqrt{\frac{4d_{\tilde{\nu}}}{m_{\ell-1}}\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)} + 2\varepsilon(\ell-2)(1+2\sqrt{d_{\tilde{\nu}}}).$$

□

B.2. Proof of Theorem 5.5

Now we are prepared to prove Theorem 5.5.

Proof of Theorem 5.5. Let $\ell_{\max}(t)$ be the index of the phase where round t is located. It's easy to see that $\ell_{\max}(T) \leq \log_2(T)$. In the following we condition on the event $\bigcap_{\ell\leq \ell_{\max}(T)} E_\ell^{\text{phase}}(\delta)$, which happens with probability at least $1-\delta$ due to Lemma B.1.

Notice that phase $\ell_{\max}(t)$ is not necessarily completed in the end of round t, but we can always round $\text{Reg}(t)$ to the regret incurred in the first $\ell_{\max}(t)$ complete phases. That is,

$$\text{Reg}(t) \leq \sum_{\ell=1}^{\ell_{\max}(t)} \sum_{a\in\mathcal{A}_\ell} T_\ell(a)\cdot\Delta_a.$$

Since we have controlled sub-optimality of all active actions in Corollary B.4, it holds with probability at least $1-\delta$ that

$$\begin{aligned}
\text{Reg}(t) &\leq \sum_{\ell=1}^{\ell_{\max}(t)} \sum_{a\in\mathcal{A}_\ell} T_\ell(a)\cdot\Delta_a \\
&\leq 2m_1 + C\sum_{\ell=2}^{\ell_{\max}(t)} m_\ell\left(\sqrt{\frac{d_{\tilde{\nu}}}{m_{\ell-1}}\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)} + \varepsilon\ell\sqrt{d_{\tilde{\nu}}}\right) \\
&\leq 2m_1 + C\sum_{\ell=2}^{\ell_{\max}(t)} \sqrt{m_\ell\cdot d_{\tilde{\nu}}\cdot\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)} + C\varepsilon\sqrt{d_{\tilde{\nu}}}\sum_{\ell=2}^{\ell_{\max}(t)} m_\ell\ell \\
&\leq 2m_1 + C\sqrt{m_{\ell_{\max}(t)}\cdot d_{\tilde{\nu}}\cdot\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)} + C\varepsilon\sqrt{d_{\tilde{\nu}}}m_{\ell_{\max}(t)}\log_2(T) \\
&\leq C\left(\sqrt{d_{\tilde{\nu}}t\log\left(\frac{2|\mathcal{A}|\log T}{\delta}\right)} + \varepsilon t\sqrt{d_{\tilde{\nu}}}\log T\right),
\end{aligned}$$

where $C > 0$ is an absolute constant that can vary from line to line. Thus we have finished the proof. □

B.3. Proof of Theorem 4.3

Now we go back to the setting where $\varepsilon = 0$. The only modification needed to work out Theorem 4.3 is an instance-dependent control over the number of phases for which sub-optimal arms are not entirely eliminated.

Proof of Theorem 4.3. Again suppose $E_\ell^{\text{phase}}(\delta)$ happens for all ℓ. From Corollary B.4 we know that every suboptimal action a can only be played in those phase $\ell \geq 2$ s.t. $\Delta_a \leq 4\sqrt{\frac{4d_\nu}{m_{\ell-1}} \log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)}$ in addition to the first phase. Let

$$\ell_a = \max\left\{\ell \geq 2 : \Delta_a \leq 4\sqrt{\frac{4d_\nu}{m_{\ell-1}} \log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)}\right\}$$

be the maximal number of phases where a can be played. It is easy to see that

$$\ell_a = 2 + \left\lfloor \log_2\left(\frac{64d_\nu}{m_1\Delta_a^2}\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)\right)\right\rfloor.$$

Hence there are at most $\ell_{\max} = 2 + \left\lfloor \log_2\left(\frac{64d_\nu}{m_1\Delta_{\min}^2}\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)\right)\right\rfloor$ number of phases before all suboptimals are eliminated and $\mathrm{Reg}(T)$ can be controlled more carefully:

$$\begin{aligned}
\mathrm{Reg}(T) &\leq \sum_{\ell=1}^{\ell_{\max}} \sum_{a\in\mathcal{A}_\ell} T_\ell(a)\cdot\Delta_a \\
&\leq 2m_1 + C\sqrt{m_{\ell_{\max}}\cdot d_\nu \cdot \log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)} \\
&= 2m_1 + C\sqrt{2^{\ell_{\max}}\cdot m_1\cdot d_\nu\cdot\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)} \\
&\leq 2m_1 + C\sqrt{\frac{d_\nu\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)}{m_1\Delta_{\min}^2}\cdot m_1\cdot d_\nu\log\left(\frac{2|\mathcal{A}|\log_2(T)}{\delta}\right)} \\
&\leq C\cdot\frac{d_\nu\log(|\mathcal{A}|\log T/\delta)}{\Delta_{\min}},
\end{aligned}$$

where $C > 0$ is an absolute constant that can vary from line to line. Again the above regret bound holds with probability at least $1-\delta$ so we are done. □

C. Anytime Regret Bounds for UCB and C-UCB

In this section we verify Proposition 3.4 for UCB and C-UCB algorithms for completeness. Note that our anytime regret bound for C-UCB is new in the literature.

C.1. Preliminaries

For each $t \in [T], a \in \mathcal{A}$ and $z \in \mathcal{Z}$, define $\mathbb{T}_t^{\mathcal{A}}(a) = 1 \vee \sum_{s=1}^t \mathbb{I}\{A_s = a\}$ to be the number of action a being chosen in the first t rounds, and define $\mathbb{T}_t^{\mathcal{Z}}(z) = 1 \vee \sum_{s=1}^t \mathbb{I}\{Z_s(A_s) = z\}$ to be the number of context z being observed up to the first t rounds. Further define the mean reward estimates $\hat{\mu}_t^{\mathcal{A}}(a), \hat{\mu}_t^{\mathcal{Z}}(z)$ by

$$\hat{\mu}_t^{\mathcal{A}}(a) = \frac{1}{\mathbb{T}_t^{\mathcal{A}}(a)}\sum_{s=1}^t Y_s(A_s)\mathbb{I}\{A_s = a\}$$

$$\hat{\mu}_t^{\mathcal{Z}}(z) = \frac{1}{\mathbb{T}_t^{\mathcal{Z}}(z)}\sum_{s=1}^t Y_s(A_s)\mathbb{I}\{Z_s(A_s) = z\}$$

Then we introduce the upper confidence bounds used by the UCB-type algorithms under consideration. Given any prescribed confidence parameter $\delta \in (0,1)$, define $\mathrm{UCB}_t^{\mathcal{A}}(a) = \hat{\mu}_t^{\mathcal{A}}(a) + \sqrt{\frac{\log(2|\mathcal{A}|T/\delta)}{2\mathbb{T}_t^{\mathcal{A}}(a)}}$, $\mathrm{UCB}_t^{\mathcal{Z}}(z) = \hat{\mu}_t^{\mathcal{Z}}(z) + \sqrt{\frac{\log(2|\mathcal{Z}|T/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}$ and $\widetilde{\mathrm{UCB}}_t(a) = \sum_{z\in\mathcal{Z}} \mathrm{UCB}_t^{\mathcal{Z}}(z)\mathbb{P}_{\nu_a}[Z=z]$ for each $t \in [T], a \in \mathcal{A}$ and $z \in \mathcal{Z}$. Furthermore, we use $\mathrm{UCB}(\delta)$ and C-$\mathrm{UCB}(\delta)$ to denote the standard UCB algorithm and C-UCB algorithm (Lu et al. 2020) which run by playing actions A_t^{UCB} and $A_t^{\mathrm{C-UCB}}$ at each round t respectively, according to:

$$A_t^{\mathrm{UCB}} = \operatorname*{argmax}_{a\in\mathcal{A}} \mathrm{UCB}_{t-1}^{\mathcal{A}}(a)$$
$$A_t^{\mathrm{C-UCB}} = \operatorname*{argmax}_{a\in\mathcal{A}} \widetilde{\mathrm{UCB}}_{t-1}(a).$$

Before analyzing the regret of $\mathrm{UCB}(\delta)$ and C-$\mathrm{UCB}(\delta)$, let's finally define some high-probability events on which we can control the regret. For any given confidence parameters δ, δ', define

$$E^{\mathcal{A}}(\delta) = \left\{\forall t \in [T], a \in \mathcal{A}, |\hat{\mu}_t^{\mathcal{A}}(a) - \mu^{\mathcal{A}}(a)| \le \sqrt{\frac{\log(2|\mathcal{A}|T/\delta)}{2\mathbb{T}_t^{\mathcal{A}}(a)}}\right\},$$

and in conditionally benign environments we additionally define

$$E^{\mathcal{Z}}(\delta) = \left\{\forall t \in [T], z \in \mathcal{Z}, |\hat{\mu}_t^{\mathcal{Z}}(z) - \mu^{\mathcal{Z}}(z))| \le \sqrt{\frac{\log(2|\mathcal{Z}|T/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}\right\},$$
$$E^{\mathrm{MG}}(\delta') = \left\{\forall t \in [T], \sum_{s=1}^{t}\sum_{z\in\mathcal{Z}} \frac{1}{\sqrt{\mathbb{T}_{s-1}^{\mathcal{Z}}(z)}}(\mathbb{P}_{\nu_{A_s}}[Z=z] - \mathbb{I}\{Z_s = z\}) \le \sqrt{2t\log(T/\delta')}\right\},$$

where we recall $\mu^{\mathcal{Z}}(z) = \mathbb{E}_{\nu_a}[Y|Z=z]$ is well-defined here. First we can see that $E^{\mathcal{A}}(\delta)$ and $E^{\mathcal{Z}}(\delta)$ happen with probability at least $1-\delta$ regardless the underlying environment and chosen policy:

Lemma C.1 (Lemma B.1 and B.2 in Bilodeau et al. 2022)**.** *For any $\nu \in \mathcal{P}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}$ and $\pi \in \Pi(\mathcal{A},\mathcal{Z},T)$,*

$$\mathbb{P}_{\nu,\pi}[(E^{\mathcal{A}}(\delta))^c] \le \delta,$$

and for any $\nu \in \mathcal{P}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}$ that is conditionally benign and $\pi \in \Pi(\mathcal{A},\mathcal{Z},T)$,

$$\mathbb{P}_{\nu,\pi}[(E^{\mathcal{Z}}(\delta))^c] \le \delta.$$

To get our new anytime regret bound for C-$\mathrm{UCB}(\delta)$, we need to further condition on $E^{\mathrm{MG}}(\delta')$ which happens with probability at least $1-\delta'$:

Lemma C.2. *For any $\nu \in \mathcal{P}(\mathcal{Z}\times\mathcal{Y})^{\mathcal{A}}$ and $\pi \in \Pi(\mathcal{A},\mathcal{Z},T)$,*

$$\mathbb{P}_{\nu,\pi}[(E^{\mathrm{MG}}(\delta'))^c] \le \delta'$$

Proof of Lemma C.2. Define

$$M_t = \sum_{s=1}^{t}\sum_{z\in\mathcal{Z}} \frac{1}{\sqrt{\mathbb{T}_{s-1}^{\mathcal{Z}}(z)}}(\mathbb{P}_{\nu_{A_s}}[Z=z] - \mathbb{I}\{Z_s=z\}), \forall t \in [T],$$
$$M_0 = 0.$$

Then $E^{\mathrm{MG}}(\delta') = \left\{\forall t \in [T], M_t \le \sqrt{2t\log(T/\delta')}\right\}$ and it is easy to find that $\{M_t\}_{t\ge 0}$ is a martingale sequence with respect to $\mathcal{F}_t = \sigma(A_t, H_{t-1})$. To see this,

$$\mathbb{E}_{\nu,\pi}[M_t|A_t,H_{t-1}] = M_{t-1} + \sum_{z\in\mathcal{Z}} \frac{1}{\sqrt{\mathbb{T}_{t-1}^{\mathcal{Z}}(z)}}\mathbb{E}_{\nu,\pi}[\mathbb{P}_{\nu_{A_t}}[Z=z] - \mathbb{I}\{Z_t=z\}|A_t] = M_{t-1}.$$

Also,

$$
\begin{aligned}
|M_t - M_{t-1}| &= \left| \sum_{z \in \mathcal{Z}} \frac{1}{\sqrt{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}} \big(\mathbb{P}_{\nu_{A_t}}[Z = z] - \mathbb{I}\{Z_t = z\} \big) \right| \\
&= \left| \mathbb{E}_{\nu,\pi} \Big[\sum_{z \in \mathcal{Z}} \frac{1}{\sqrt{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}} \mathbb{I}\{Z_t = z\} | A_t, H_{t-1} \Big] - \sum_{z \in \mathcal{Z}} \frac{1}{\sqrt{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}} \mathbb{I}\{Z_t = z\} \right| \\
&= \left| \mathbb{E}_{\nu,\pi} \Big[\frac{1}{\sqrt{\mathbb{T}^{\mathcal{Z}}_{t-1}(Z_t)}} | A_t, H_{t-1} \Big] - \frac{1}{\sqrt{\mathbb{T}^{\mathcal{Z}}_{t-1}(Z_t)}} \right| \\
&\leq 1.
\end{aligned}
$$

Then by Azuma-Hoeffding,

$$
\begin{aligned}
\mathbb{P}_{\nu,\pi}[M_t > \sqrt{2t \log(T/\delta')}] &= \mathbb{P}_{\nu,\pi}[M_t - M_0 > \sqrt{2t \log(T/\delta')}] \\
&\leq \exp\left(-\frac{2t \log(T/\delta')}{2t} \right) = \delta'/T,
\end{aligned}
$$

and we get $\mathbb{P}_{\nu,\pi}[(E^{\mathrm{MG}}(\delta'))^c] \leq \delta'$ after taking a union bound over $t \in [T]$. □

C.2. Anytime High-probability Regret Bound

Now we provide our high-probability regret bounds for $\mathrm{UCB}(\delta)$ and C-$\mathrm{UCB}(\delta)$ that will lead to Proposition 3.4.

Theorem C.3. *In any environment ν, the regret of* $\mathrm{UCB}(\delta)$ *is bounded by*

$$
\mathrm{Reg}(t) = O\Big(\sqrt{|\mathcal{A}| \log(|\mathcal{A}|T/\delta) t} \Big)
$$

for all $t \in [T]$, conditioning on event $E^{\mathcal{A}}(\delta)$ which happens with probability at least $1 - \delta$.

Proof of Theorem C.3. In event $E^{\mathcal{A}}(\delta)$, we have that $\mu^{\mathcal{A}}(a) \leq \mathrm{UCB}^{\mathcal{A}}_t(a) \leq \mu^{\mathcal{A}}(a) + 2\sqrt{\frac{\log(2|\mathcal{A}|T/\delta)}{2\mathbb{T}^{\mathcal{A}}_t(a)}}$ for all $a \in \mathcal{A}, t \in [T]$. Hence conditioned on $E^{\mathcal{A}}(\delta)$, the regret of $\mathrm{UCB}(\delta)$ up to any round $t \in [T]$ holds

$$
\begin{aligned}
\mathrm{Reg}(t) &= \sum_{s=1}^{t} \mu^{\mathcal{A}}(a^*) - \mu^{\mathcal{A}}(A_s) \\
&= \sum_{s=1}^{t} (\mu^{\mathcal{A}}(a^*) - \mathrm{UCB}^{\mathcal{A}}_{s-1}(A_s)) + (\mathrm{UCB}^{\mathcal{A}}_{s-1}(A_s) - \mu^{\mathcal{A}}(A_s)) \\
&\leq \sum_{s=1}^{t} (\mathrm{UCB}^{\mathcal{A}}_{s-1}(a^*) - \mathrm{UCB}^{\mathcal{A}}_{s-1}(A_s)) + (\mathrm{UCB}^{\mathcal{A}}_{s-1}(A_s) - \mu^{\mathcal{A}}(A_s)) \\
&\leq \sum_{s=1}^{t} (\mathrm{UCB}^{\mathcal{A}}_{s-1}(A_s) - \mu^{\mathcal{A}}(A_s)) \\
&\leq \sum_{s=1}^{t} \sqrt{\frac{2 \log(2|\mathcal{A}|T/\delta)}{\mathbb{T}^{\mathcal{A}}_{s-1}(A_s)}} \\
&= \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} \sqrt{\frac{2 \log(2|\mathcal{A}|T/\delta)}{\mathbb{T}^{\mathcal{A}}_{s-1}(A_s)}} \mathbb{I}\{A_s = a\} \\
&\leq \sum_{a \in \mathcal{A}} \sqrt{8 \log(2|\mathcal{A}|T/\delta) \mathbb{T}^{\mathcal{A}}_{t-1}(a)} \\
&\leq \sqrt{8 \log(2|\mathcal{A}|T/\delta) |\mathcal{A}| t},
\end{aligned}
$$

where we use $A_s = A_s^{\mathrm{UCB}}$ throughout to simplify our notation. □

Theorem C.4. *In any conditionally benign environment ν, the regret of* C-UCB(δ) *is bounded by*

$$\mathrm{Reg}(t) = O\Big(\sqrt{\log(|\mathcal{Z}|T/\delta)}\Big(\sqrt{|\mathcal{Z}|} + \sqrt{\log(T/\delta')}\Big)\sqrt{t}\Big)$$

for all $t \in [T]$, conditioning on event $E^{\mathcal{Z}}(\delta) \cap E^{\mathrm{MG}}(\delta')$ which happens with probability at least $1 - \delta - \delta'$.

Proof of Theorem C.4. Similarly in event $E^{\mathcal{Z}}(\delta)$ we have $\mu^{\mathcal{Z}}(z) \le \mathrm{UCB}_t^{\mathcal{Z}}(z) \le \mu^{\mathcal{Z}}(z) + 2\sqrt{\frac{\log(2|\mathcal{Z}|T/\delta)}{2\mathbb{T}_t^{\mathcal{Z}}(z)}}$ for all $z \in \mathcal{Z}, t \in [T]$. Additionally,

$$\begin{aligned}
\mu^{\mathcal{A}}(a^*) &= \sum_{z\in\mathcal{Z}} \mu^{\mathcal{Z}}(z)\mathbb{P}_{\nu_{a^*}}[Z = z] \\
&\le \sum_{z\in\mathcal{Z}} \mathrm{UCB}_{t-1}^{\mathcal{Z}}(z)\mathbb{P}_{\nu_{a^*}}[Z = z] \\
&= \widetilde{\mathrm{UCB}}_{t-1}(a^*) \le \widetilde{\mathrm{UCB}}_{t-1}(A_t), \forall t \in [T],
\end{aligned}$$

where $A_t = A_t^{\mathrm{C-UCB}}$ is the action played by C-UCB(δ). Therefore we can control the cumulative the regret of C-UCB(δ) in the first t rounds as follows

$$\begin{aligned}
\mathrm{Reg}(t) &= \sum_{s=1}^{t}(\mu^{\mathcal{A}}(a^*) - \widetilde{\mathrm{UCB}}_{s-1}(A_s)) + (\widetilde{\mathrm{UCB}}_{s-1}(A_s) - \mu^{\mathcal{A}}(A_s)) \\
&\le \sum_{s=1}^{t} \widetilde{\mathrm{UCB}}_{s-1}(A_s) - \mu^{\mathcal{A}}(A_s) \\
&= \sum_{s=1}^{t}\sum_{z\in\mathcal{Z}} (\mathrm{UCB}_{s-1}^{\mathcal{Z}}(z) - \mu^{\mathcal{Z}}(z))\mathbb{P}_{\nu_{A_s}}[Z = z] \\
&\le \sum_{s=1}^{t}\sum_{z\in\mathcal{Z}} \sqrt{\frac{2\log(2|\mathcal{Z}|T/\delta)}{\mathbb{T}_{s-1}^{\mathcal{Z}}(z)}}\mathbb{P}_{\nu_{A_s}}[Z = z] \\
&= \sum_{s=1}^{t}\sum_{z\in\mathcal{Z}} \sqrt{\frac{2\log(2|\mathcal{Z}|T/\delta)}{\mathbb{T}_{s-1}^{\mathcal{Z}}(z)}}\mathbb{I}\{Z_s = z\} + \sum_{s=1}^{t}\sum_{z\in\mathcal{Z}} \sqrt{\frac{2\log(2|\mathcal{Z}|T/\delta)}{\mathbb{T}_{s-1}^{\mathcal{Z}}(z)}}(\mathbb{P}_{\nu_{A_s}}[Z = z] - \mathbb{I}\{Z_s = z\}) \\
&\le \sqrt{8\log(2|\mathcal{Z}|T/\delta)|\mathcal{Z}|t} + \sum_{s=1}^{t}\sum_{z\in\mathcal{Z}} \sqrt{\frac{2\log(2|\mathcal{Z}|T/\delta)}{\mathbb{T}_{s-1}^{\mathcal{Z}}(z)}}(\mathbb{P}_{\nu_{A_s}}[Z = z] - \mathbb{I}\{Z_s = z\}),
\end{aligned}$$

where in the last inequality we use the same argument as in the proof of Theorem C.3, and the remaining summation term can be controlled by $\sqrt{4\log(2|\mathcal{Z}|T/\delta)\log(T/\delta')t}$ immediately after we further condition on $E^{\mathrm{MG}}(\delta')$. Therefore, we get

$$\mathrm{Reg}(t) \le \sqrt{\log(2|\mathcal{Z}|T/\delta)}\Big(\sqrt{8|\mathcal{Z}|} + \sqrt{4\log(T/\delta')}\Big)\sqrt{t}, \forall t \in [T],$$

in event $E^{\mathcal{Z}}(\delta) \cap E^{\mathrm{MG}}(\delta')$. □

Combining Theorem C.3 with Theorem C.4 and taking $\delta' = \delta$, we thus verfy Proposition 3.4.

D. Proofs of Lower Bounds

In this section we give the full proof of Theorem 3.7 and Theorem 5.2. Note that our proof of Theorem 3.7 mainly adopts but also largely generalizes the one of Bilodeau et al. (2022, Theorem 6.2).

D.1. Proof of Theorem 3.7

Proof of Theorem 3.7. Fix $\mathcal{A}$, $\mathcal{Z}$ and T. Let $\mathcal{Z}_0$ be an arbitrary proper subset of $\mathcal{Z}$ and $\mathcal{Z}_1 = \mathcal{Z} \setminus \mathcal{Z}_0$. Fix $\Delta \in (0, 1/20)$ to be chosen later. Define the family of marginals for all instances appearing in this proof

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

where probability is evenly spread within $\mathcal{Z}_0$ and $\mathcal{Z}_1$ respectively. Then define a conditionally benign environment $\nu \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}$ by

$$\mathbb{P}_{\nu_a}[Y=1] = \sum_{z \in \mathcal{Z}} p[Y=1|Z=z] q_a[Z=z], \quad \forall a \in \mathcal{A},$$

where $p[Y|Z]$ is a Bernoulli conditional distribution such that

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

Now we define some non-benign instances. For every $a_0 \neq 1$, define ν_{a_0} by

$$\mathbb{P}_{\nu_a^{a_0}}[Y=1] = \sum_{z \in \mathcal{Z}} p_a^{a_0}[Y=1|Z=z] q_a[Z=z], \quad \forall a \in \mathcal{A},$$

where $p_a^{a_0}[Y|Z]$ is a Bernoulli conditional distribution such that

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

For any MAB algorithm $\mathfrak{a}$, let $\pi^q = \mathfrak{a}(\mathcal{A}, \mathcal{Z}, q, T)$ be the actual policy implemented by $\mathfrak{a}$ when it's interacting with ν and ν^{a_0}. Then by the divergence decomposition formula and Bretagnolle-Huber inequality,

$$\begin{aligned}
\mathbb{E}_{\nu,\pi^q}[\mathrm{Reg}(T)] + \mathbb{E}_{\nu^{a_0},\pi^q}[\mathrm{Reg}(T)] &\geq \frac{T\Delta}{2}\mathbb{P}_{\nu,\pi^q}[\mathbb{T}_T^{\mathcal{A}}(1) \leq T/2] + \frac{T\Delta}{2}\mathbb{P}_{\nu^{a_0},\pi^q}[\mathbb{T}_T^{\mathcal{A}}(1) > T/2] \\
&\geq \frac{T\Delta}{4}\exp(-\mathrm{KL}(\mathbb{P}_{\nu,\pi^q} \parallel \mathbb{P}_{\nu^{a_0},\pi^q})) \\
&= \frac{T\Delta}{4}\exp\left(-\frac{1}{2}\mathbb{E}_{\nu,\pi^q}[\mathbb{T}_T^{\mathcal{A}}(a_0)]\mathrm{KL}(\mathrm{Ber}(3/4) \parallel \mathrm{Ber}(3/4+4\Delta))\right) \\
&\geq \frac{T\Delta}{4}\exp\left(-\mathbb{E}_{\nu,\pi^q}[\mathbb{T}_T^{\mathcal{A}}(a_0)] \cdot 32\Delta^2\right),
\end{aligned}$$

where in the last step we use $\mathrm{KL}(\mathrm{Ber}(3/4) \parallel \mathrm{Ber}(3/4+4\Delta)) \leq 64\Delta^2$ for $\Delta < 1/40$. Combined with the worst-case regret upper bound $\mathbb{E}_{\nu,\pi^q}[\mathrm{Reg}(T)] + \mathbb{E}_{\nu^{a_0},\pi^q}[\mathrm{Reg}(T)] \leq 2R(T; \mathcal{A}, \mathcal{Z})$, it implies that

$$\mathbb{E}_{\nu,\pi^q}[\mathbb{T}_T^{\mathcal{A}}(a_0)] \geq \frac{1}{32\Delta^2}\log\left(\frac{T\Delta}{8R(T;\mathcal{A},\mathcal{Z})}\right), \forall a_0 \neq 1.$$

Realizing $\mathbb{E}_{\nu,\pi^q}[\mathrm{Reg}(T)] = \sum_{a_0 \neq 1} \Delta \mathbb{E}_{\nu,\pi^q}[\mathbb{T}_T^{\mathcal{A}}(a_0)]$, we have

$$\mathbb{E}_{\nu,\pi^q}[\mathrm{Reg}(T)] \geq \frac{|\mathcal{A}|-1}{32\Delta}\log\left(\frac{T\Delta}{8R(T;\mathcal{A},\mathcal{Z})}\right).$$

So there exists absolute constants $c = \log 2/1024, c' = 1/641$ such that whenever $R(T; \mathcal{A}, \mathcal{Z}) \leq c'T$, the choice of $\Delta = \frac{16R(T;\mathcal{A},\mathcal{Z})}{T}$ satisfies $\Delta < 1/40$ and

$$\mathbb{E}_{\nu,\pi^q}[\mathrm{Reg}(T)] \geq c \cdot \frac{|\mathcal{A}|T}{R(T;\mathcal{A},\mathcal{Z})},$$

which completes the proof. □

D.2. Proof of Theorem 5.2

Proof of Theorem 5.2. Fix $\mathcal{A}, \mathcal{Z}$ and T. Let $\mathcal{Z}_0$ be an arbitrary proper subset of $\mathcal{Z}$ and $\mathcal{Z}_1 = \mathcal{Z} \setminus \mathcal{Z}_0$. Fix $\Delta \in (0, \frac{1}{40})$ to be chosen later. For all conditionally benign instances ν in this proof, we consider $\mathbb{P}_{\nu_a}[Y|Z]$ to be the Bernoulli distribution given by

$$\mathbb{P}_{\nu_a}[Y = 1|Z = z] = p[Y = 1|Z = z] = \begin{cases} 3/4 & z \in \mathcal{Z}_0 \\ 1/4 & z \in \mathcal{Z}_1, \end{cases}$$

which implies that contexts from $\mathcal{Z}_0$ are more rewarding than those from $\mathcal{Z}_1$.

Now define conditionally benign environments $\nu, \nu^{a_0} \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})^{\mathcal{A}}, \forall a_0 \neq 1$, through their marginals

$$\mathbb{P}_{\nu_a}[Y = 1] = \sum_{z \in \mathcal{Z}} p[Y = 1|Z = z] q_a[Z = z],$$
$$\mathbb{P}_{\nu_a^{a_0}}[Y = 1] = \sum_{z \in \mathcal{Z}} p[Y = 1|Z = z] q_a^{a_0}[Z = z], \forall a \in \mathcal{A}$$

where

$$q_a[Z \in \mathcal{Z}_0] = \begin{cases} 1/2 + 2\Delta & a = 1 \\ 1/2 & a \neq 1 \end{cases} \quad \text{and} \quad q_a^{a_0}[Z \in \mathcal{Z}_0] = \begin{cases} 1/2 + 2\Delta & a = 1 \\ 1/2 + 4\Delta & a = a_0 \\ 1/2 & a \neq 1, a_0, \end{cases}$$

where probability is evenly spaced within $\mathcal{Z}_0$ and $\mathcal{Z}_1$. So clearly action 1 is the only optimal action in ν and action a_0 is the only optimal action in ν^{a_0}, with sub-optimality gap $\Delta_{\min}(\nu) = \Delta_{\min}(\nu^{a_0}) = \Delta$.

Fix algorithm $\mathfrak{a} \in \mathbb{A}_{\text{agnostic}}$ with $\tilde{\pi} = \mathfrak{a}(\mathcal{A}, \mathcal{Z}, T, \cdot)$ be the actual policy implemented by $\mathfrak{a}$. By the divergence decomposition formula (Bilodeau et al. 2022), we have that for every $a_0 \neq 1$,

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_{\nu,\tilde{\pi}} \parallel \mathbb{P}_{\nu^{a_0},\tilde{\pi}}) &= \sum_{a \in \mathcal{A}} \mathbb{E}_{\nu,\tilde{\pi}}[\mathbb{T}_T^{\mathcal{A}}(a)] \mathrm{KL}(\mathbb{P}_{\nu_a} \parallel \mathbb{P}_{\nu_a^{a_0}}) \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}_{\nu,\tilde{\pi}}[\mathbb{T}_T^{\mathcal{A}}(a)] \mathrm{KL}(q_a \parallel q_a^{a_0}) \\ &= \mathbb{E}_{\nu,\tilde{\pi}}[\mathbb{T}_T^{\mathcal{A}}(a_0)] \mathrm{KL}(q_{a_0} \parallel q_{a_0}^{a_0}) \\ &= \mathbb{E}_{\nu,\tilde{\pi}}[\mathbb{T}_T^{\mathcal{A}}(a_0)] \mathrm{KL}(\mathrm{Ber}(1/2) \parallel \mathrm{Ber}(1/2 + 4\Delta)). \end{aligned}$$

By Bretagnolle–Huber inequality,

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

Now we pick $a_0 \in \operatorname{argmin}_{a \neq 1} \mathbb{E}_{\nu,\tilde{\pi}}[\mathbb{T}_T^{\mathcal{A}}(a)]$ which implies that $\mathbb{E}_{\nu,\tilde{\pi}}[\mathbb{T}_T^{\mathcal{A}}(a_0)] \leq \frac{T}{|\mathcal{A}|-1}$. Also $\mathrm{KL}(\mathrm{Ber}(1/2) \parallel \mathrm{Ber}(1/2 + 4\Delta)) \leq 4(4\Delta)^2 = 64\Delta^2$ for $\Delta < 1/40$. So

$$\mathbb{E}_{\nu,\tilde{\pi}}[\mathrm{Reg}(T)] + \mathbb{E}_{\nu^{a_0},\tilde{\pi}}[\mathrm{Reg}(T)] \geq \frac{T\Delta}{4} \exp\left(-\frac{64T\Delta^2}{|\mathcal{A}| - 1} \right).$$

Taking $\Delta = \frac{1}{40}\sqrt{\frac{|\mathcal{A}|-1}{T}}$, we know that $\max\{\mathbb{E}_{\nu,\tilde{\pi}}[\mathrm{Reg}(T)], \mathbb{E}_{\nu^{a_0},\tilde{\pi}}[\mathrm{Reg}(T)]\} \geq \frac{1}{2}(\mathbb{E}_{\nu,\tilde{\pi}}[\mathrm{Reg}(T)] + \mathbb{E}_{\nu^{a_0},\tilde{\pi}}[\mathrm{Reg}(T)]) \geq c\sqrt{|\mathcal{A}|T}$ for some absolute constant $c > 0$, which yields the claim. □

In the above proof, it is easy to see that $\nu(Z)$ and $\nu^{a_0}(Z)$ are $\varepsilon-$close, where $\varepsilon = c \cdot \sqrt{|\mathcal{A}|/T}$ for some absolute constant c. So for any algorithm input by $\tilde{\nu}(Z) = \nu(Z)$ when interacting with $\bar{\nu} \in \{\nu, \nu^{a_0}, a_0 \neq 1\}$, it is satisfied that $\tilde{\nu}(Z)$ and $\bar{\nu}(Z)$ are always $\varepsilon-$close, but the algorithm incurs $\Omega(\sqrt{|\mathcal{A}|T})$ regret in some instance from $\{\nu, \nu^{a_0}, a_0 \neq 1\}$.

E. Instances where PE incurs linear regret

In this section we give an example for PE to illustrate that to merely force linear regret on a causal bandit algorithm, we need to construct non-benign instances carefully and re-code the algorithm to ensure its erratic behavior in those instances. In particular, we construct a non-benign environment ν for every $\Delta \in (0,1)$ such that $\Delta_{\min}(\nu) = \Delta$ while the re-coded PE never plays the optimal arm.

Proposition E.1. *Suppose we modify Algorithm 2 such that, in each phase, we always choose an exact-optimal design whenever feasible. For any $\mathcal{A}$, $\mathcal{Z}$ and T with $|\mathcal{A}| > |\mathcal{Z}| \geq 3$ and $\Delta \in (0,1)$, there exists a non-benign environment ν such that $\Delta_{\min}(\nu) = \Delta$, while Algorithm 2 will never play the optimal arm, hence incurring linear regret,*

$$\mathrm{Reg}(t) \geq \Delta_{\min}(\nu) \cdot t = \Delta \cdot t, \quad \forall t \in [T].$$

Proof of Proposition E.1. For any $\mathcal{A}$ and $\mathcal{Z}$ with $|\mathcal{A}| > |\mathcal{Z}| \geq 3$, suppose we index the contexts in arbitrary way such that $\mathcal{Z} = \{z_1, ..., z_{|\mathcal{Z}|}\}$, and we pick $(|\mathcal{Z}|+1)$ number of arms from $\mathcal{A}$ and denote them by $a^*, a_1, ..., a_{|\mathcal{Z}|}$. Construct marginals ν_a as follows:

$$\nu_{a_i}(Z) = \delta_{\{z_i\}} =: e_i, \quad i \in [|\mathcal{Z}|],$$
$$\nu_{a^*}(Z) = \frac{1}{2}(\delta_{\{z_1\}} + \delta_{\{z_2\}}) = \frac{1}{2}(e_1 + e_2),$$

where we write marginal distributions over $\mathcal{Z}$ as vectors in $\mathbb{R}^{|\mathcal{Z}|}$ according to context indices. Then define conditional distributions $\mathbb{P}_{\nu_a}(Y|Z)$:

$$\mathbb{P}_{\nu_{a_i}}[Y|Z = z_i] = \begin{cases} \delta_{\{0\}} & i \in [|\mathcal{Z}|-1] \\ \delta_{\{1-\Delta\}} & i = |\mathcal{Z}| \end{cases}$$

and

$$\mathbb{P}_{\nu_{a^*}}[Y|Z = z_1] = \mathbb{P}_{\nu_{a^*}}[Y|Z = z_2] = \delta_{\{1\}}.$$

In other words, playing arm a_i yields context z_i and deterministic reward, while we could observe z_1 or z_2 with equal probability and always get the optimal reward by playing arm a^*. So the only optimal arm for ν is a^* with $\Delta_{\min}(\nu) = \Delta$. We can treat all other $a \in \mathcal{A}$ as dummy actions by identifying each of them with one of $a^*, a_i, i \in [|\mathcal{Z}|]$ arbitrarily.

Next we will verify the following facts. (1) When no action is eliminated and $\mathcal{A}_\ell = \mathcal{A}$, any exact G-optimal design $\pi_\ell \in \mathcal{P}(\mathcal{A}_\ell)$ does not have positive mass over a^*. (2) Whenever any action is eliminated in the end of phase ℓ, it must be that all actions except for $a_{|\mathcal{Z}|}$ are eliminated as well. Then PE would just play $a_{|\mathcal{Z}|}$ till the end. Combining these two facts we can conclude that PE never picks a^* during the interaction with ν.

No G-optimal design is supported on a^*. Recall that any G-optimal design π_ℓ maximizes $f(\pi) = \log\det V(\pi)$, where $V(\pi) = \sum_{a \in \mathcal{A}_\ell} \pi(a)\nu_a\nu_a^\top$ over $\pi \in \mathcal{P}(\mathcal{A}_\ell)$ (Lattimore & Szepesvári, 2020, Theorem 21.1). When $\mathcal{A}_\ell = \mathcal{A}$, $\det V(\pi)$ can be computed as

$$\det V(\pi) = \left(\pi(a_1)\pi(a_2) + \frac{\pi(a^*)}{4}(\pi(a_1) + \pi(a_2))\right)\pi(a_3)\cdots\pi(a_{|\mathcal{Z}|}).$$

Then we can find that any maximizing π should have $\pi(a^*) = 0$ after realizing that $\pi(a_1) = \pi(a_2)$ for such π. Moreover, there is only one G-optimal design in this case, which is $\pi_\ell = Unif(a_1, ..., a_{|\mathcal{Z}|})$.

All actions other than $a_{|\mathcal{Z}|}$ would be eliminated at the same time. If the first elimination happens in the end of phase ℓ, then we must have $\hat{\mu}_\ell^{\mathcal{Z}} = (0, ..., 0, 1-\Delta)^\top$ due to that $\pi_\ell = Unif(a_1, ..., a_{|\mathcal{Z}|})$ and rewards are deterministic. So $\max_{b \in \mathcal{A}_\ell}\langle \hat{\mu}_\ell^{\mathcal{Z}}, \nu_b - \nu_a\rangle$ is $1-\Delta$ for all $a \neq a_{|\mathcal{Z}|}$ and 0 for $a = a_{|\mathcal{Z}|}$. Then the elimination must happen within $a^*, a_1, ..., a_{|\mathcal{Z}|-1}$, and thus every one of it should be eliminated simultaneously. □