The Fallacy of Minimizing Cumulative Regret in the Sequential Task Setting

Anonymous Author
Anonymous Institution

Abstract

Online Reinforcement Learning (RL) is typically framed as the process of minimizing cumulative regret (CR) through interactions with an unknown environment. However, real-world RL applications usually involve a sequence of tasks, and the data collected in the first task is used to warm-start the second task. The performance of the warm-start policy is measured by **simple regret** (SR). While minimizing both CR and SR is generally a conflicting objective, previous research has shown that in stationary environments, both can be optimized in terms of the duration of the task, T.

In practice, however, in real-world applications, human-in-the-loop decisions between tasks often results in non-stationarity. For instance, in clinical trials, scientists may adjust target health outcomes between implementations. Our results show that task non-stationarity leads to a more restrictive trade-off between CR and SR. To balance these competing goals, the algorithm must explore excessively, leading to a CR bound worse than the typical optimal rate of $T^{1/2}$. These findings are practically significant, indicating that increased exploration is necessary in non-stationary environments to accommodate task changes, impacting the design of RL algorithms in fields such as healthcare and beyond.

1 Introduction

Cumulative regret (CR) minimization has emerged as a central topic in online Reinforcement Learning (RL) research (Auer et al., 2008; Chu et al., 2011; Agrawal and Goyal, 2013; Lattimore and Szepesvári, 2020). Within a stationary environment, prioritizing regret minimization is theoretically sound. A sublinear regret bound ensures a strong lower bound on cumulative rewards, as well as convergence to the optimal policy when the suboptimality gap is non-zero. However, any algorithm with a sublinear regret will stop exploration over time, which may lead to suboptimal performances for downstream tasks in a sequential task setting.

To explore the nature of this suboptimal performance, we consider a simple setting with two tasks, where the agent is restricted to deploy policies from a policy class Π and the goal is to maximize rewards, defined as a mapping f of outcome vectors. Stationarity holds across the two tasks if the reward mapping and policy class are the same. However, often, due to human-in-loop decisions, the choice of the reward mapping and/or policy class changes between tasks. See Figure 1 in which rewards are known functions of outcome vectors. In the first task, the reward is defined by the function $f^{(1)}$, and in the second task, the reward is defined by a different function $f^{(2)}$, where $f^{(2)} \neq f^{(1)}$.

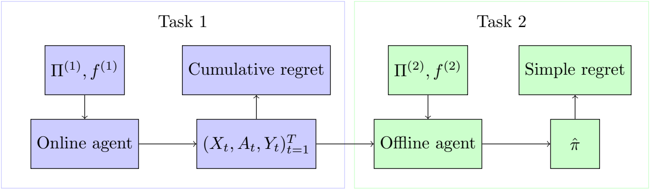

Figure 1: Two-task learning paradigm

1. Task one: The agent observes the policy class $\Pi^{(1)}$, and the reward mapping $f^{(1)}$. It interacts with an unknown environment for T steps. At each step t, the agent observes a context X_t and

Preliminary work. Under review by AISTATS 2025. Do not distribute.

samples an action $A_t \sim \pi_t(X_t)$ from $\Pi^{(1)}$. The environment then generates an outcome vector Y_t, and the agent aims to maximize the cumulative reward $\sum_t f^{(1)}(Y_t)$.

2. Task two: The agent observes the policy class $\Pi^{(2)}$, and the reward mapping $f^{(2)}$. Based on the data collected during the online learning stage, it proposes a policy $\hat{\pi} \in \Pi^{(2)}$. The goal in this task is to maximize $\mathbb{E}[f^{(2)}(Y_t) \mid A_t \sim \hat{\pi}(X_t)]$.

The focus on the fixed policy $\hat{\pi}$ for the second task is motivated by two key concerns. First, in real-world RL deployment, both safety and access to appropriate infrastructure are critical. Indeed implementing an online RL algorithm (i.e., an adaptive policy) in practice presents challenges, such as building the necessary infrastructure to update models in near real-time (Pacchiano et al., 2024). Additionally, deploying online RL algorithms often requires real-time monitoring systems (Trella et al., 2024) to mitigate negative impacts on users or human subjects (Liao et al., 2020), which can be costly. For these and other reasons, a fixed policy is often deployed in the subsequent tasks.

We evaluate the performance of the task one and the task two through cumulative regret (CR) and simple regret (SR), respectively. Simple regret is the expected gap between the reward of the optimal policy in $\Pi^{(2)}$ and that of $\hat{\pi}$. Krishnamurthy et al. (2023) show the trade-off between CR and SR when the environment is stationary, i.e., $\Pi^{(1)} = \Pi^{(2)}, f^{(1)} = f^{(2)}$:

$$\inf_{\text{Algorithm}} \sup_{\text{Instance}} \mathbb{E}[\text{CR}] \times \mathbb{E}[\text{SR}] = \tilde{\Omega}(|\mathcal{X}| \times |\mathcal{A}|) \quad (1)$$

for contextual bandits with context space $\mathcal{X}$ and arm set $\mathcal{A}$ [1]. The above characterizes a weak trade-off between CR and SR in the sense that both CR and SR can be minimax-optimal with rates $\sqrt{|\mathcal{X}||\mathcal{A}|T}$ and $\sqrt{|\mathcal{X}||\mathcal{A}|/T}$, respectively.

Human-in-the-loop between tasks. The lower bound in (1) critically depends on the stationarity between tasks–the agent runs on the same environment across tasks. However, *many real-world RL tasks occur sequentially, with human-in-the-loop interventions between tasks that redefine the task parameters*. This means that although the distribution of the outcome, Y, conditional on context, action (X, A), is the same across tasks, the policy class changes, $\Pi^{(1)} \neq \Pi^{(2)}$ and/or the reward changes, $f^{(1)} \neq f^{(2)}$, leading to non-stationarity in the task. See, for example, applications in mobile health (Liao et al., 2020; Bidargaddi et al., 2020; Trella et al., 2022), and online education (Aleven et al., 2023; Ruan et al., 2023).

For example, in mobile health, an RL agent might initially deliver digital interventions based on the context X_t, such as sleep quality from the previous night. However, in a subsequent implementation, changes in user agreements may prevent the collection of X_t, requiring the agent to make decisions independently of that context. This results in a constrained policy class $\Pi^{(2)} = \{\pi : \pi(x_1) = \pi(x_2), \forall x_1, x_2\}$.

Main contribution. We demonstrate that above changes between tasks, driven by human-in-the-loop decisions, can significantly worsen the trade-off between CR and SR compared to the trade-off implied by (1). Specifically, we instantiate three types of task changes due to human-in-the-loop decisions and show the minimax rate of $\sqrt{\mathbb{E}[\text{CR}]\mathbb{E}[\text{SR}]} = \Omega(1)$ when these changes are present. This lower bound suggests a more restrictive trade-off between CR and SR than (1), since the optimal minimax cumulative regret rate $\sqrt{T}$, the typical rate achieved by most contextual bandit algorithms, results in a suboptimal rate for SR, $\mathbb{E}[\text{SR}] = \Omega(T^{-1/4})$. This rate for SR is suboptimal because pure exploration in the first task gives $\mathbb{E}[\text{SR}] = \mathcal{O}(T^{-1/2})$. We further extend our results to settings with multiple tasks and validate the theory through simulation studies.

The main message of this paper is that *additional exploration is required when unanticipated changes occur between tasks due to human-in-the-loop involvement. Complete exploitation, i.e. use of an algorithm leading to minimax-optimal cumulative regret, within a task can lead to worse performance in later tasks.*

1.1 Related works

In this section, we review two closely related areas of research: non-stationarity RL and multi-task RL.
Previous literature on multi-task RL focuses on transfering the knowledge across tasks with heterogeneity in reward distributions (Cella and Pontil, 2021; Cella et al., 2023). Therefore, these works must make task similarity assumptions to ensure the transferability of knowledge across tasks, and multi-task RL algorithms leverage these assumptions to improve the performance of the second task. Similarly, in the literature on non-stationarity RL, the reward distribution changing across time (Garivier and Moulines, 2008; Raj and Kalyani, 2017; **?**), and the goal in this literature is derive algorithms that are robust to the changes in the reward distribution.
The major difference between our work and the exist-

[1] We presented a stronger version of Theorem 3 in Krishnamurthy et al. (2023) for a easier presentation.

ing literature on multi-task RL and non-stationarity RL is that we focus on the changes in the policy class $\Pi^{(1)} \neq \Pi^{(2)}$ and reward mapping $f^{(1)} \neq f^{(2)}$ between tasks. The outcome distribution P that generates the outcome Y_t does not change across tasks presenting a sharp contrast to existing multi-task RL literature. To our knowledge, the changes in the policy class and reward mapping between tasks have not been considered in any of the existing literature on non-stationarity RL and multi-task RL.

2 Problem Setup

Notations. For a set $\mathcal{X}$, we denote by $\Delta(\mathcal{X})$ the set of all distributions over $\mathcal{X}$. For $N \in \mathbb{Z}$, we let $[N] = \{1, 2, \ldots, N\}$. We use $\mathcal{O}(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$ to denote the big-O, big-Theta and big-Omega notations. The $\tilde{\mathcal{O}}(\cdot)$, $\tilde{\Theta}(\cdot)$, $\tilde{\Omega}(\cdot)$ notation hides all the logarithmic terms. We denote by $D_{\mathrm{KL}}(P \mid Q)$, the KL divergence between two probability measures P and Q with $P \ll Q$.

Two-task contextual bandit learning paradigm. We rigorously introduce the two-task contextual bandit paradigm in Figure 1. The agent operates in a contextual bandit environment with context space $\mathcal{X}$, action set $\mathcal{A}$, and outcome space $\mathcal{Y}$. Here we consider discrete set $\mathcal{X}$. We denote by Π the set of all mappings from $\mathcal{X}$ to $\Delta(\mathcal{A})$, i.e., Markovian policies.

In the first task, the agent is given a policy space $\Pi^{(1)} \subset \Pi$ and reward mapping $f^{(1)} : \mathcal{Y} \mapsto \mathbb{R}$. At each step $t \in [T]$, the environment generates a context $X_t \in \mathcal{X}$ i.i.d. from the context distribution P_X. The agent chooses a policy $\pi_t : \mathcal{X} \mapsto \Delta(\mathcal{A})$, and samples an action $A_t \sim \pi_t(X_t)$, where $\pi_t \in \Pi^{(1)}$. The environment generates a random outcome vector $Y_t \sim P(X_t, A_t)$, with outcome distribution $P : \mathcal{X} \times \mathcal{A} \mapsto \Delta(\mathcal{Y})$. The reward during task 1 is $R_t = f^{(1)}(Y_t)$. The agent minimizes cumulative regret in Definition 1.

Definition 1 (Cumulative regret)**.** *Denote by* $R_\pi^{(1)} := \mathbb{E}_{X_t \sim P_X, A_t \sim \pi(X_t)} \mathbb{E}_{Y_t \sim P(\cdot \mid x,a)} f^{(1)}(Y_t)$ *the mean reward of a policy* π*. The cumulative regret is*

$$\mathrm{CR} := \sum_{t=1}^{T} \left[\max_{\pi \in \Pi^{(1)}} \mathbb{E}[R_\pi^{(1)} - R_{\pi_t}^{(1)}] \right].$$

For the second task, the agent is given a new policy space $\Pi^{(2)} \subset \Pi$ and reward mapping $f^{(2)} : \mathcal{Y} \mapsto \mathbb{R}$. It then learns offline from the data collected during task 1 and proposes a policy $\hat{\pi} \in \Pi^{(2)}$. The goal is to minimize simple regret $\mathrm{SR}(\hat{\pi})$.

Definition 2 (Simple regret)**.** *Define the simple regret of a policy* π *by*

$$\mathrm{SR}(\pi) := \max_{\pi' \in \Pi^{(2)}} R_{\pi'}^{(2)} - R_\pi^{(2)},$$

where $R_\pi^{(2)} := \mathbb{E}_{X_t \sim P_X, A_t \sim \pi(X_t)} \mathbb{E}_{Y_t \sim P(\cdot \mid x,a)} f^{(2)}(Y_t)$.

Note that both CR and SR depend implicitly on the learning algorithm and the environment parameters including $(P, \Pi^{(1)}, \Pi^{(2)}, f^{(1)}, f^{(2)})$. So far we consider **the same outcome distribution** P to focus on the study of changes in policy spaces and reward functions. We extend our discussion to outcome distribution shift in Section 5.

Remark 1. *Policy space and reward mappings* $\Pi^{(i)}, f^{(i)}$ *are part of the task design and are therefore assumed to be known before the agent interacts with the* i*-th task. The agent is learning the underlying outcome distribution* P*, that is considered unknown, and remains the same for two tasks.*

Motivations for changes between tasks. The setup for the two tasks may change in various ways. Below are motivating examples of changes in $(\Pi^{(i)}, f^{(i)})$.

1. **Changes in reward mappings $f^{(i)}$.** In mobile health studies, we observe two outcomes $Y_t = (Y_{t,1}, Y_{t,2})$, where $Y_{t,1}$ is the indicator of healthy behaviors, and $Y_{t,2}$ is the response rate. During the online learning, the reward mapping is $f^{(1)}(Y_t) \equiv Y_{t,1}$. Based on the observed data, a domain expert may decide that the response rate is too low, which could deter participants from using the app. Therefore, they propose a new reward mapping $f^{(2)}(Y_t) = Y_{t,1} - \alpha Y_{t,2}$, where α is a balancing factor.

2. **Changes in policy classes $\Pi^{(i)}$.** In real-world applications, the context space $\mathcal{X}$ can be a high-dimensional vector due to the complexity of real-world observations. During online learning, computational limitations may prevent optimization over the entire space of policies that account for all contexts. Hence, the domain expert might decide to optimize only in the space of context-independent policies $\Pi^{(1)} = \{\pi \in \Pi : \pi(a \mid x_1) = \pi(a \mid x_2)$, for all $(x_1, x_2) \in \mathcal{X}\}$. As more data is collected, the expert may decide that certain components of $\mathcal{X}$ are relevant to the task, which should be included in the offline learning.

3 Minimax Lower Bound

A major result of this paper is the more restrictive trade-off between cumulative regret and simple regret. We demonstrate this trade-off by establishing a minimax lower bound under mild conditions on the problem instance set.

Problem instance. Tuple $(P, \Pi^{(1)}, \Pi^{(2)}, f^{(1)}, f^{(2)})$ uniquely defines the environment the algorithm in which the algorithm operates. Since policy classes and reward mappings are known, we denote a problem instance simply by P, the unknown outcome distribution, and by $\mathcal{P}$ the set of outcome distributions of interest.

Definition 3 (Occupancy measure)**.** *Given a policy π, we define the occupancy measure of π as*

$$\mu_\pi(x, a) = P_X(x)\pi(a \mid x).$$

Note that occupancy measure is independent of bandit instance as we assume the same context distribution.

Definition 4 (Learning algorithm)**.** *We denote a learning algorithm by $L = (L^{(1)}, L^{(2)})$, where $L^{(1)}$ is an online learning algorithm that is a deterministic mapping from history $\mathcal{H}_t = (X_\tau, A_\tau, Y_\tau)_{\tau=1}^{t-1}$ to a policy in $\Pi^{(1)}$. Here, $L^{(2)}$ is an offline learning algorithm that maps $\mathcal{H}_{T+1}$ to a policy $\hat{\pi} \in \Pi^{(2)}$. Note that $L^{(i)}$ can depend on $f^{(i)}$ and $\Pi^{(i)}$ for each $i \in \{1, 2\}$ as they are known knowledge. We denote the set of all such algorithms by $\mathcal{L}$.*

Now we are ready to present our main theorem.

Theorem 1. *For the policies spaces $\Pi^{(1)} = \{\pi : \pi(\cdot \mid x_1) = \pi(\cdot \mid x_2), \text{for all } x_1, x_2 \in \mathcal{X}\}$ and $\Pi^{(2)} = \Pi$, there exists an instance set $\mathcal{P}$, and reward mapping $f^{(1)} = f^{(2)}$, such that,*

$$\inf_{L \in \mathcal{L}} \sup_{P \in \mathcal{P}} \mathbb{E}[\mathrm{SR}(\hat{\pi})]\sqrt{\mathbb{E}[\mathrm{CR}]} = \Omega(1). \tag{2}$$

Similarly, for some $f^{(1)} \neq f^{(2)}$, there exists exists an instance set $\mathcal{P}$ and $\Pi^{(1)} = \Pi^{(2)} = \Pi$, such that

$$\inf_{L \in \mathcal{L}} \sup_{P \in \mathcal{P}} \mathbb{E}[\mathrm{SR}(\hat{\pi})]\sqrt{\mathbb{E}[\mathrm{CR}]} = \Omega(1). \tag{3}$$

Discussion. Theorem 1 demonstrates a more restrictive trade-off between cumulative regret and simple regret than the trade-off shown in (1) (Athey et al., 2022). This more restrictive result arises from allowing either $\Pi^{(1)} \neq \Pi^{(2)}$ or $f^{(1)} \neq f^{(2)}$.

A significant difference between (2), (3) and the existing result in (1) is that the rate for $\mathbb{E}[\mathrm{CR}]$ in (1) becomes square root. To understand the rationale of the square root term, we note that a purely random exploration in the first task results in $\mathbb{E}[\mathrm{CR}] = \Theta(1)$, which leads to a simple regret bound of $\mathbb{E}[\mathrm{SR}] = \Omega(1/\sqrt{T})$. This is consistent with the regular estimation error lower bound for estimating parametric models.

3.1 Lower bound construction

In this section, we discuss the lower bound construction and provide the intuition for the lower bound in Theorem 1.

The overall idea is to construct instances where s specific context-arm pair (x, a) leads to non-zero regret in task 1 whenever (x, a) is visited. In this setting, $\mu_{\pi_1^*}(x, a) = 0$, meaning that the optimal policy for task 1 avoids visiting (x, a) for cumulative regret minimization. However, in task 2, the same pair (x, a) becomes optimal ($\mu_{\pi_2^*}(x, a) > 0$), and the offline learning algorithm will need a dataset with sufficient coverage on (x, a) to learn the optimal policy.

This motivation is closely related to the results in the offline learning literature, where it has been shown that offline learning is fundamentally hard if the single-policy concentrability defined by $\max_{x,a} \mu_{\pi_1^*}(x, a)/\mu_{\pi_2^*}(x, a)$ is unbounded (Chen and Jiang, 2019).

Based on this motivation, we now describe the detailed lower bound construction.

Case of $\Pi^{(1)} \neq \Pi^{(2)}$: add a new feature. During the online learning, we may not have the computational resources to maximize over the space of all policies that takes context into account. As a result, the domain expert might decide to optimize only in the space of context-independent policies, turning the problem into a multi-armed bandit problem. With evidence accumulating, the expert may decide that certain components of $\mathcal{X}$ are relevant to the task, which should be included in the offline learning.

To illustrate this, we consider a set of two-armed contextual bandits with context space $\mathcal{X} = \{x_1, x_2\}$ and outcome space $\mathcal{Y} = \mathbb{R}$. We let $\Pi^{(1)} = \{\pi : \pi(\cdot \mid x_1) = \pi(\cdot \mid x_2), \text{for all } x_1, x_2 \in \mathcal{X}\}$, meaning the policies do not differentiate between contexts. For task 2, the policy class is the set of all policies. We consider $f^{(1)}(y) = f^{(2)}(y) = y$. That is, the reward mappings are identical mappings.

We consider an instance set (outcome distribution set) $\mathcal{P}$ such that the mean of $P(\cdot \mid x, a)$ for each context-arm pair is given by Table 1 for any $\epsilon \in [0, 0.5]$ and $\xi \in [0, 0.25]$. Each choice of ϵ, ξ defines two instances, denoted by $P_{\epsilon,\xi}$, $\bar{P}_{\epsilon,\xi}$. We introduce ξ to ensure the richness of the instance class.

Note that in task 1 the algorithm must ignore context, thus it pulls the arm with the larger marginal mean. In all $P \in \mathcal{P}$, the marginal mean reward for a_1 is strictly larger than that of a_2. This ensures that online learning algorithm must avoid pulling a_2, meaning that they will not collect enough data to distinguish between $P_{\epsilon,\xi}$ from its counterpart $\bar{P}_{\epsilon,\xi}$.

However, in task 2, distinguishing between $P_{\epsilon,\xi}$ and $\bar{P}_{\epsilon,\xi}$ becomes crucial for minimizing simple regret because these two distributions have opposite optimal policies. As a result, we create a fundamental tension between cumulative regret and simple regret: minimizing cumulative regret in task 1 leads to poor data collection, making it difficult to minimize simple regret in task 2.

Table 1: Mean reward for $\Pi^{(1)} \neq \Pi^{(2)}$ case

$P_{\epsilon,\xi}$	x_1	x_2	marginal
a_1	$0.5+\epsilon$	$0.5-\epsilon$	0.5
a_2	$0.5-2\xi$	0.5	$0.5-\xi$
$\bar{P}_{\epsilon,\xi}$	x_1	x_2	marginal
a_1	$0.5+\epsilon$	$0.5-\epsilon$	0.5
a_2	$0.5-2\xi$	$0.5-2\epsilon$	$0.5-\xi-\epsilon$

Case of $f^{(1)} \neq f^{(2)}$: change reward mapping. We consider multi-armed bandit problems with no context. The outcomes vectors are $Y_t = (Y_{t,1}, Y_{t,2})$. The reward mappings are given by $f^{(1)}(Y_t) = Y_{t,1}$ for task 1 and $f^{(2)}(Y_t) = Y_{t,2}$ for task 2.

Consider the instance set $\mathcal{P}$ such that the means of $Y_{t,1}$ and $Y_{t,2}$ for each arm pair is given in Table 2 for any $\epsilon \in [0, 0.5]$, and $\xi \in [0, 0.25]$. Each choice of ϵ and ξ defines two instances, denoted by $P_{\epsilon,\xi}$, $\bar{P}_{\epsilon,\xi}$.

In this setup, it is clear that a_2 is the optimal arm in task 1, while the algorithm must pull a_2 to distinguish $P_{\epsilon,\xi}$ from $\bar{P}_{\epsilon,\xi}$, which is critical to decide the optimal policy for the second task.

Table 2: Mean reward for $f^{(1)} \neq f^{(2)}$ case

$P_{\epsilon,\xi}$	$Y_{t,1}$	$Y_{t,2}$	$\bar{P}_{\epsilon,\xi}$	$Y_{t,1}$	$Y_{t,2}$
a_1	0.5	0.5	a_1	0.5	0.5
a_2	0.5-ξ	0.5-ϵ	a_2	0.5-ξ	0.5+ϵ

4 Optimal Level of Exploration

As implied by Theorem 1, any algorithm that achieves an optimal rate in CR will be suboptimal in SR in terms of the dependence on T. To balance between these two objectives, incorporate additional exploration during the first task. In this section, we characterize the optimal level of additional exploration that minimizes the following weighted objective for different values of T':

$$\mathbb{E}[\mathrm{CR}] + T'\mathbb{E}[\mathrm{SR}], \tag{4}$$

where T' could be interpreted as the number of steps intended to be taken in task 2. Since we primarily focus on the role of horizons, we omit the dependence on $|\mathcal{X}|$ and $|\mathcal{A}|$ throughout the discussions in this section.

Proposition 1 suggests a minimax lower bound for the weighted sum of cumulative regret and simple regret in (4) that is the maximum of three terms: $T'/\sqrt{T}, T'^{2/3}$ and $\sqrt{T}$. The first term $T'/\sqrt{T}$ corresponds to the case where simple regret $T'\mathbb{E}[\mathrm{SR}]$ dominates, and the minimax rate of simple regret in the second for any dataset collected during the first task is $T'/\sqrt{T}$. The second term $T'^{2/3}$ corresponds to the rate characterized by Theorem 1. The last term of $\sqrt{T}$ is the minimax rate of the first task regret minimization. This corresponds to the case when cumulative regret $\mathbb{E}[\mathrm{CR}]$ dominates.

Proposition 1. *Following the same choice of the instance set $\mathcal{P}$ and $\Pi^{(1)}, \Pi^{(2)}, f^{(1)}, f^{(2)}$ in Theorem 1, the following minimax lower bound holds*

$$\inf_{L \in \mathcal{L}} \sup_{P \in \mathcal{P}} \mathbb{E}[\mathrm{CR} + T' \,\mathrm{SR}] = \Omega\left(\max\left\{\frac{T'}{\sqrt{T}}, T'^{2/3}, \sqrt{T}\right\}\right). \tag{5}$$

We show that a simple algorithm that mixes a minimax-optimal online learning algorithm with a purely random exploration policy has upper bounded CR and SR that matches the lower bound in Theorem 1 up to a factor of $|\mathcal{X}||\mathcal{A}|$. This also allows us to achieve any point on the Pareto frontier up to a factor of $|\mathcal{X}||\mathcal{A}|$. The parameter α controls the level of additional exploration in the first task.

Theorem 2. *Let L_0 be an online learning algorithm with a regret bound of $\tilde{\mathcal{O}}(\sqrt{|\mathcal{X}||\mathcal{A}|T})$. Let the algorithm for the first task be $L_\alpha(\tau) = (1-\alpha)L_0(\tau) + \alpha\pi_0$, where τ is any past observations and π_0 is the uniform random policy. For any choice of $\alpha \in [|\mathcal{X}||\mathcal{A}|/\sqrt{T}, 1]$, there exist offline-learning algorithm for the second task such that*

$$\mathbb{E}[\mathrm{CR}] = \tilde{\mathcal{O}}\left(\alpha T\right), \mathbb{E}[\mathrm{SR}] = \tilde{\mathcal{O}}\left(\sqrt{\frac{(|\mathcal{X}||\mathcal{A}|)^2}{\alpha T}}\right). \tag{6}$$

By tuning the exploration rate α, we are able to match the minimax lower bound in (44). In short, there are three regimes of (T, T'), for which we should choose different levels of exploration rate T' to balance $\mathbb{E}[\mathrm{CR}]$ and $\mathbb{E}[\mathrm{SR}]$.

1. **Regime 1: $T \leq T'^{2/3}$.** The first task is too short compared to the weight T' of the second task, and the algorithm should employ pure exploration in the first task ($\alpha = 1$). This regime leads to a global regret of $\mathcal{O}(T'/\sqrt{T})$.

2. **Regime 2: $T'^{2/3} < T \leq T'^{4/3}$.** The algorithm should employ additional exploration compared to these that achieve a minimax optimal rate in a single task. Theorem 2 suggests an additional exploration rate of $\alpha = T'^{2/3}/T$ and a global regret bound of $\mathcal{O}(T'^{2/3})$. Note that under a special case of $T' = T$, the rate of $\alpha = T^{-1/3}$ indicates a regret bound of $\mathcal{O}(T^{2/3})$ in the first task.

3. **Regime 3:** $T > T'^{4/3}$**.** The algorithm should employ $\alpha = 0$, meaning that no excess exploration is needed and the agent in task 1 can minimize the cumulative regret as much as possible. In this regime, the cumulative regret in the first task could achieve the minimax optimal rate of $\sqrt{T}$.

It is often in real-world applications that T' is predetermined and the researcher could decide how many samples to collect in the first task to ensure a good learning in the second one. For instance, in an inventory management context (Madeka et al., 2022), it is determined by the engineering team that how long a learned policy should be deployed for the second task. In such cases, our theory indicates that one should choose $T > T'^{4/3}$, so a greedy cumulative regret minimization for the first task is justified.

5 Study on Changes in P

In real-world implementations, the outcome distribution P often experiences unpredictable shift. Prior research on non-stationary bandits has typically focused on single-task scenarios with potential reward distribution shifts at any step. To manage these shifts, the literature often limits the total variation in distribution shifts, making it possible to establish sublinear regret bounds. However, in a sequential task setting, when the algorithm cannot adaptively learn during the second task, the simple regret is always lower bounded by a constant. This is due to the uncertainty of the optimal policy for the second task, even with full knowledge of the outcome distribution from the first task. To address this challenge, we introduce the concept of *robust simple regret*. We show that the robust simple regret are shown to have a minimax lower bound similar to that shown in Theorem 1.

For simplicity, we consider no change in the policy space or the reward function. Specifically, let $\Pi^{(1)} = \Pi^{(2)} = \Pi$, the set of all policies, and $f^{(1)} = f^{(2)} = f$, the identical mapping in $\mathbb{R}$. We denote by $P^{(1)}$ and $P^{(2)}$ the outcome distributions, aka., reward distribution of the first and second task. We denote a problem instance by $\boldsymbol{P} = (P^{(1)}, P^{(2)})$.

In this setting, the adversary is allowed to choose $P^{(2)}$ from a L_1 ball around $P^{(1)}$. This leads to the instance set $\mathcal{P}(\Delta)$ parametrized by constant Δ such that each $\boldsymbol{P} = (P^{(1)}, P^{(2)})$ satisfies

$$P^{(2)} \in \mathcal{P}(P^{(1)}, \Delta) \coloneqq \left\{ P : \sum_a |P(x,a) - P^{(1)}(x,a)| \le \Delta \text{ for all } x \in \mathcal{X} \right\}, \tag{7}$$

where we abuse the notation for P and let $P(x,a)$ denote the mean reward for (x,a).

Distributionally robust simple regret. When $P^{(2)}$ is allowed to change adversarially from $P^{(1)}$, it is not reasonable to compare against the true optimal policy with respect to the unknown $P^{(2)}$. Instead, we define a robust regret notion. First we define the worst-case simple regret of a policy π:

$$\mathrm{SR}(\pi \mid P^{(1)}, \Delta) \coloneqq \sup_{\substack{P^{(2)} \in \mathcal{P}(P^{(1)}, \Delta) \\ x \in \mathcal{X}}} \left(\max_a P^{(2)}(x,a) - \sum_a P^{(2)}(x,a)\pi(a \mid x) \right). \tag{8}$$

We denote by $\tilde{\pi}_{P^{(1)},\Delta} \coloneqq \inf_{\pi' \in \Pi^{(2)}} \mathrm{SR}(\pi' \mid P^{(1)}, \Delta)$ the optimal robust policy given $(P^{(1)}, \Delta)$. When it is clear from the context, we drop the subscription for $P^{(1)}$ and Δ.

We further define *robust simple regret*, which is the gap between the worst-case simple regret of a given policy and the policy that achieves the lowest worst-case simple regret:

$$\widetilde{\mathrm{SR}}(\pi \mid P^{(1)}, \Delta) \coloneqq \mathrm{SR}(\pi \mid P^{(1)}, \Delta) - \inf_{\pi' \in \Pi^{(2)}} \mathrm{SR}(\pi' \mid P^{(1)}, \Delta). \tag{9}$$

Note that the worst-case regret form over some ambiguity set has been studied in the robust Markov Decision Process literature (Xu and Mannor, 2010; Eysenbach and Levine, 2021; Dong et al., 2022). However, the definition of robust simple regret and the tension between cumulative regret and robust simple regret has not yet been explored.

To illustrate how the trade-off between cumulative regret and simple regret remains relevant in the robust setting, we investigate a simple two-armed, context-free bandit case in Proposition 2. The optimal arm in the the first task is a_1, while the optimal robust policy depends on the gap between the mean reward of both arms. Thus, to reduce the robust simple regret in the second task, the algorithm is forced to have an accurate estimate on the suboptimal arm in the first task.

Proposition 2. *Consider the following two-armed, context-free bandit, with Gap* $\coloneqq P^{(1)}(a_1) - P^{(1)}(a_2) > 0$*. Then the worst-case simple regret is given by*

$$\mathrm{SR}(\pi \mid P^{(1)}, \Delta) = \max\{(\Delta - Gap)\pi(a_1), (\Delta + Gap)\pi(a_2)\}. \tag{10}$$

Assume that $\Delta > Gap$*. The optimal robust policy* $\tilde{\pi}$

w.r.t. the worst-case simple regret has the explicit form

$$\tilde{\pi}_{P^{(1)},\Delta}(a_1) = \frac{\Delta + Gap}{2\Delta}, \ \textit{and } \tilde{\pi}_{P^{(1)},\Delta}(a_2) = \frac{\Delta - Gap}{2\Delta}. \tag{11}$$

Motivated by the instance introduced in Proposition 2, we show the following Theorem that lower bounds the minimax rate of the product between cumulative regret in the first task and the robust simple regret in the second task. Note that robust simple regret does not depend on the actual $P^{(2)}$ of choice, the supremum is only taken over $P^{(1)}$.

Theorem 3. *Assume each $P^{(i)}(\cdot \mid x, a)$ is from a Bernoulli distribution with mean $P^{(i)}(x, a)$ for all $i = 1, 2$ and $x, a \in \mathcal{X} \times \mathcal{A}$. Let $\mathcal{P}^{(1)}$ be the set of such distributions. Then there exists some Δ such that*

$$\inf_{L \in \mathcal{L}} \sup_{P^{(1)} \in \mathcal{P}^{(1)}} \mathbb{E}[\widetilde{\mathrm{SR}}(\hat{\pi} \mid P^{(1)}, \Delta)] \sqrt{\mathbb{E}[\mathrm{CR}]} = \Omega(1). \tag{12}$$

Theorem 3 implies that one should still employ additional exploration for small T, when there is only changes in the outcome distributions.

6 Extending to Multiple Non-linear Tasks

In our two-task study, we highlighted the inherent dilemma between cumulative regret and simple regret. Now we extend our discussion to a sequence of multiple tasks. A significant property about the two-tasks scenario is algorithm's inability to adaptively learn in the second task. This restriction forces the algorithm to "over-explore" in the first task to propose a good policy for the second task, thereby introducing a tension between the regrets in the first and the second task. In fact, such tension can exists between any task $i \in [N]$ and its preceding tasks over a sequence of N tasks in the case that the algorithm is not allow to adaptively learn in all tasks. In this multi-task setting with no adaptivity, we demonstrate that a typical UCB type algorithm is fallacious leading to constant regret over an exponentially long sequence of tasks when the underlying environment is a nonlinear bandit.

Nonlinear contextual bandit. For simplicity, we consider the outcome $\mathcal{Y} = \mathbb{R}$ and a fixed reward function $f^{(i)} = f \equiv x \mapsto x$ for all tasks i, i.e. no rich observations, so we focus on the changes in the policy spaces. Following the setups in Section 2, we now consider potentially large or continuous context and action space $\mathcal{X}$ and $\mathcal{A}$. We define the mean reward as $R(x, a) := \mathbb{E}_{Y \sim P(x,a)}[f(Y)]$. For contextual bandit with nonlinear reward models, we assume that mean reward $R \in \mathcal{F}$ for some known function class $\mathcal{F} : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$.

Complexity for nonlinear bandit. Running UCB on nonlinear bandit is generally hard. Russo and Van Roy (2013) proposed to explore by choosing $A_t \in \arg\max_{a \in \mathcal{A}} \sup_{f \in \mathcal{F}_t} f(X_t, a)$, where $\sup_{f \in \mathcal{F}_t} f(a)$ is an optimistic estimate of $f_\theta(a)$. A choice of $\mathcal{F}_t$ given by Russo and Van Roy (2013) is

$$\mathcal{F}_t = \left\{ f \in \mathcal{F} : \|f - \hat{f}_t^{LS}\|_{2,E_t} \leq \sqrt{\beta_t^\star} \right\}, \tag{13}$$

where $\beta_t^\star$ are constants, $\|g\|_{2,E_t} = \sum_{t=1}^T g^2(X_t, A_t)$ is the empirical 2-norm, and $\hat{f}_t^{LS} \in \inf_{f \in \mathcal{F}} (f(X_t, A_t) - Y_t)^2$ is the empirical risk minimizer. The regret of running UCB with appropriately chosen $\beta_t^\star$ has regret of $\sqrt{\dim_E(\mathcal{F}, T^{-2}) T}$, where $\dim_E(\mathcal{F}, T^{-2})$ is the eluder dimension of the function class $\mathcal{F}$.

Definition 5 (Distributional eluder dimension)**.** *Let $\mathcal{F} : \mathcal{X} \mapsto \mathbb{R}$. A probability measure ν over $\mathcal{X}$ is said to be ϵ-independent of a sequence of probability measures $\{\mu_1, \ldots, \mu_n\}$ w.r.t $\mathcal{F}$ if any pair of functions $f, \bar{f} \in \mathcal{F}$ satisfying $\sqrt{\sum_{i=1}^n (\mathbb{E}_\mu[f(x) - \bar{f}(x)])^2} \leq \epsilon$ also satisfies $|\mathbb{E}_\nu[f(x) - \bar{f}(x)]| \leq \epsilon$. Furthermore, x is ϵ-independent of $\{\mu_1, \ldots, \mu_n\}$ if it is not ϵ-dependent of the sequence.*

The ϵ-eluder dimension $\dim_E(\mathcal{F}, \epsilon)$ is the length of the longest sequence of distributions over $\mathcal{X}$ such that for some $\epsilon' \geq \epsilon$, every distribution is ϵ'-independent of its predecessors.

Recall that the construction of our hard instances in Theorem 1 requires that the new task has the optimal policy whose occupancy measure has no overlap from the occupancy measure of optimal policies in the previous tasks. A generalization of this to the nonlinear case is that a predicted function that minimizes the loss over the dataset collected in the previous tasks may still occur large loss on a new task. Let the optimal policies of tasks $1, \ldots n$ be $\pi_1^\star, \ldots, \pi_n^\star$. Intuitively, as long as n is smaller than $\dim_E(\mathcal{F}, \epsilon)$, we can find a new task with optimal policy $\pi_{n+1}^\star$ for which the occupancy measure $\mu_{\pi_{n+1}^\star}$ is ϵ-independent of $(\mu_{\pi_1^\star}, \ldots, \mu_{\pi_n^\star})$. By the definition of eluder dimension, this implies that the function chosen for task $n+1$ based on the dataset collected by $(\mu_{\pi_1^\star}, \ldots, \mu_{\pi_n^\star})$ may still occur a large error. Note that by running a no-regret online algorithm, the dataset collected during a task will asymptotically distributed as the occupancy measure induced by its optimal policy.

Eluder dimension has been shown to be exponentially large for simple models like one-layer neural network with ReLU activation function (Dong et al., 2021). It is not trivial to show a lower bound directly depending on the eluder dimension. Instead, we provide a concrete example, where UCB described in (13) fails.

Theorem 4. *Consider the hypothesis set $\mathcal{F}$ to be one-hidden layer neural networks with width d. There exists ground-truth reward function and a sequence of tasks of length $\Omega(\exp(d))$ with different $\Pi^{(i)}$, such that the cumulative regret for each task is lower bounded by a constant, even if each $T_i \to \infty$.*

Theorem 4 indicates that even without a change in outcome distributions, there still exists an exponentially long sequence of tasks, for which the tension between local regret minimization and global regret minimization still holds. An UCB algorithm that greedily minimizes local regret fails to provide good guarantees for later tasks.

7 Simulation Studies

A main result of this paper is that the trade-off characterized in (1) can be significantly more restrictive when certain changes present.

Environment from hard instance construction. We validate this theory with two contextual bandit experiments, one focusing on the policy classes change and the other focusing on the reward mappings change based on the hard instance construction discussed in Section 3 to ensure the tension between cumulative regret and simple regret. Specifically, in experiment 1, we choose $\Pi^{(2)} = \Pi$ and $\Pi^{(1)} = \{\pi : \pi(\cdot \mid x_1) = \pi(\cdot \mid x_2), \text{for all } x_1, x_2 \in \mathcal{X}\}$. In experiment 2, we choose $f^{(1)}(\boldsymbol{y}) = y_1$ and $f^{(2)}(\boldsymbol{y}) = y_2$ for $\boldsymbol{y} = (y_1, y_2)$. For each experiment, we run UCB (Auer, 2002) and UCB mixed with probability 0.1 random exploration. We compute the average cumulative regret and simple regret over time.

In both experiments, UCB shows a diminishing average cumulative regret with a constant simple regret for the second task. Mixed with 0.1 random exploration, UCB receives a constant average regret, while the simple regret goes to zero, revealing a strong trade-off between cumulative regrets and simple regret.

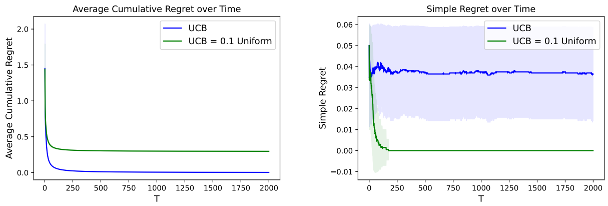

Figure 2: Experiment with changes in policy spaces

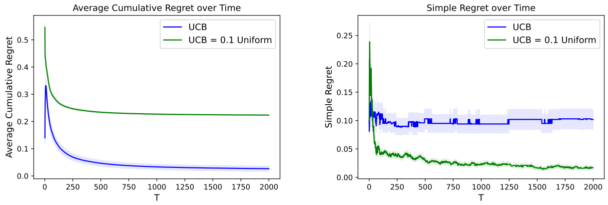

Figure 3: Experiment with changes in reward mappings

Random environment. We further validate our results with a random environment to validate the robustness of our results. In this section, we run d-dimensional linear contextual bandit with both context and the true reward coefficients drawn independently from standard Gaussian distributions. We run Thompson Sampling (Russo et al., 2018) for $T = 500$ steps with different values of $\alpha = 0.1, 0.2, \ldots, 0.6$, and compute the weighted sum of cumulative regret and simple regret $\mathrm{CR}_t + T' \mathrm{SR}_t$ with $T' = 500$ to investigate the optimal additional exploration level as discussed in Section 4. Further, details can be found in Appendix F.

In Figure 4, we plot the average of $\mathrm{CR}_t + T' \mathrm{SR}_t$ over 1000 independent runs. We observe that the optimal α decreases over time, implying that we should increase the level of additional exploration as the horizon for task 1 decreases.

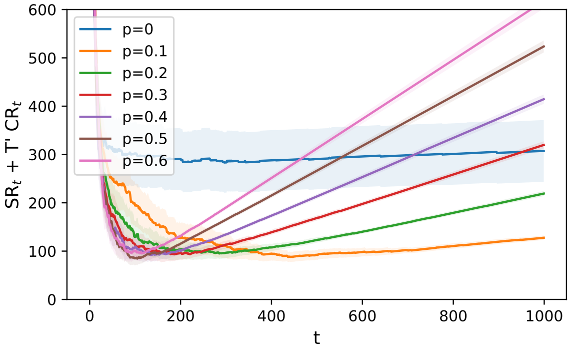

Figure 4: The weighted sum of cumulative regret and simple regret at each time step t based on random environment by running Thompson Sampling with different values of α

8 Discussion

In this paper, we study the trade-off between cumulative regret and simple regret across two contextual bandit tasks. We demonstrate that with changes in the task setups between tasks due to human-in-the-loop decisions, the above trade-off is more restrictive. These changes include changes in policy spaces, reward mappings and outcome distributions, which is of significant novelty. We further

The main implication is that we should employ more random exploration in task 1. More specifically, the results from Section 4 suggest that the optimal level of exploration in task 1 changes according to the the length of two tasks T and T'.

- When $T < T'^{2/3}$, the task 1 is short compared to task 2 and we should employ pure random exploration in task 1.
- When $T'^{4/3} > T > T'^{2/3}$, the task 1 and task 2 are both important and the additional exploration rate is $\alpha = T'^{2/3}/T$.
- When $T > T'^{4/3}$, the task 1 is long compared to task 2 and no additional exploration is needed. The already existing exploration from any minimax optimal online algorithm is sufficient.

Our work provides guidance on the level of exploration in the real-world sequential implementations of RL.

Limitations. Potential limitations of this paper include the scope of theoretical results. As a first step towards the sequential multi-task setting with non-stationarity between tasks, there is a gap in the minimax rate presented in Theorem 1. More specifically, the lower bound should depend on the complexity of the environment $|\mathcal{X}|$ and $|\mathcal{A}|$. Further work should be done to bring this bound to minimax-optimal.

We study minimax rate throughout the paper, which focuses often on the worst case. In reality, some instances are significantly harder to learn than the others. An interesting direction is to propose a theoretical measure of the significance of the trade-off studied in this paper and derive an instance-dependent result.

References

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR.

Aleven, V., Baraniuk, R., Brunskill, E., Crossley, S. A., Demszky, D., Fancsali, S. E., Gupta, S., Koedinger, K., Piech, C., Ritter, S., Thomas, D. R., Woodhead, S., and Xing, W. (2023). Towards the future of ai-augmented human tutoring in math learning. In *International Conference on Artificial Intelligence in Education.*

Athey, S., Byambadalai, U., Hadad, V., Krishnamurthy, S. K., Leung, W., and Williams, J. J. (2022). Contextual bandits in a survey experiment on charitable giving: Within-experiment outcomes versus policy learning. *ArXiv*, abs/2211.12004.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Auer, P., Jaksch, T., and Ortner, R. (2008). Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21.

Bidargaddi, N., Schrader, G., Klasnja, P., Licinio, J., and Murphy, S. (2020). Designing m-health interventions for precision mental health support. *Translational psychiatry*, 10(1):222.

Cella, L., Lounici, K., Pacreau, G., and Pontil, M. (2023). Multi-task representation learning with stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4822–4847. PMLR.

Cella, L. and Pontil, M. (2021). Multi-task and meta-learning with sparse linear bandits. In *Uncertainty in Artificial Intelligence*, pages 1692–1702. PMLR.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

Dong, J., Li, J., Wang, B., and Zhang, J. (2022). Online policy optimization for robust mdp. *arXiv preprint arXiv:2209.13841.*

Dong, K., Yang, J., and Ma, T. (2021). Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *Advances in neural information processing systems*, 34:26168–26182.

Eysenbach, B. and Levine, S. (2021). Maximum entropy rl (provably) solves some robust rl problems. *arXiv preprint arXiv:2103.06257.*

Garivier, A. and Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415.*

Guo, Y., Xu, Z., and Murphy, S. (2024). Online learning in bandits with predicted context. In *International Conference on Artificial Intelligence and Statistics*, pages 2215–2223. PMLR.

Krishnamurthy, S. K., Zhan, R., Athey, S., and Brunskill, E. (2023). Proportional response: Contextual bandits for simple and cumulative regret minimization. *arXiv preprint arXiv:2307.02108.*

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

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22.

Madeka, D., Torkkola, K., Eisenach, C., Luo, A., Foster, D. P., and Kakade, S. M. (2022). Deep inventory management. *arXiv preprint arXiv:2210.03137.*

Pacchiano, A., Lee, J., and Brunskill, E. (2024). Experiment planning with function approximation. *Advances in Neural Information Processing Systems*, 36.

Raj, V. and Kalyani, S. (2017). Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727.*

Ruan, S. S., Nie, A., Steenbergen, W., He, J., Zhang, J., Guo, M., Liu, Y., Nguyen, K. D., Wang, C. Y., Ying, R., Landay, J. A., and Brunskill, E. (2023). Reinforcement learning tutor better supported lower performers in a math task. *ArXiv*, abs/2304.04933.

Russo, D. and Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Simchi-Levi, D. and Wang, C. (2023). Multi-armed bandit experimental design: Online decision-making and adaptive inference. In *International Conference on Artificial Intelligence and Statistics*, pages 3086–3097. PMLR.

Trella, A. L., Zhang, K. W., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., and Murphy, S. A. (2022). Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms*, 15(8):255.

Trella, A. L., Zhang, K. W., Nahum-Shani, I., Shetty, V., Yan, I., Doshi-Velez, F., and Murphy, S. A. (2024). Monitoring fidelity of online reinforcement learning algorithms in clinical trials. *arXiv preprint arXiv:2402.17003.*

Xu, H. and Mannor, S. (2010). Distributionally robust markov decision processes. *Advances in Neural Information Processing Systems*, 23.

Checklist

1. For all models and algorithms presented, check if you include:
   - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   - (b) Complete proofs of all theoretical results. [Yes]
   - (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   - (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   - (b) The license information of the assets, if applicable. [Not Applicable]
   - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   - (d) Information about consent from data providers/curators. [Not Applicable]
   - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Proof Theoerm 1

A.1 Lower Bound When $\Pi^{(1)} \neq \Pi^{(2)}, f^{(1)} = f^{(2)}$

Recall that our lower bound construction is a set of two-armed contextual bandits with context space $\mathcal{X} = \{x_1, x_2\}$ and outcome space $\mathcal{Y} = \mathbb{R}$. We let $\Pi^{(1)} = \{\pi : \pi(\cdot \mid x_1) = \pi(\cdot \mid x_2), \text{for all } x_1, x_2 \in \mathcal{X}\}$, and $\Pi^{(2)}$ be the set of all policies. We consider $f^{(1)}(y) = f^{(2)}(y) = y$. That is, the reward mappings are identical mappings. Additionally, P_X is uniform distribution over $\mathcal{X}$.

We consider the instance set (outcome distribution set) $\mathcal{P}$ such that each $P(\cdot \mid x, a)$ is a Gaussian distribution with mean $p_{x,a}$ and variance σ^2. Specifically, we consider P such that the mean reward for each context-arm pair is given by Table 3 for any $\epsilon \in [0, 0.5]$ and $\xi \in [0, 0.25]$. Note that each ϵ, ξ realizes two instances, denoted by $P_{\epsilon,\xi}$, $\bar{P}_{\epsilon,\xi}$.

Table 3: Case of $\Pi^{(1)} \neq \Pi^{(2)}$

P_ϵ	x_1	x_2	marginal	$\bar{P}_\epsilon$	x_1	x_2	marginal
a_1	$0.5 + \epsilon$	$0.5 - \epsilon$	0.5	a_1	$0.5 + \epsilon$	$0.5 - \epsilon$	0.5
a_2	$0.5 - 2\xi$	0.5	$0.5 - \xi$	a_2	$0.5 - 2\xi$	$0.5 - 2\epsilon$	$0.5 - \xi - \epsilon$

Now we are ready to prove the theorem. Throughout the proof, we use $\mathbb{E}_P^L[]$ for the expectation of random variables of interest given the underlying instance P. Let $T(x, a) = \sum_{t=1}^T \mathbb{1}\{(X_t, A_t) = (x, a)\}$ the random number of visit in the context-arm pair (x, a).

We first note that

$$\begin{aligned}
\mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[\mathrm{CR}] =& \xi \sum_{t=1}^T \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[\mathbb{1}\{A_t = a_2\}] && (14)\\
=& 2\xi \sum_{t=1}^T \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[\mathbb{1}\{A_t = a_2\}]\mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[\mathbb{1}\{X_t = x_2\}] && (15)\\
=& 2\xi \sum_{t=1}^T \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[\mathbb{1}\{A_t = a_2\}\mathbb{1}\{X_t = x_2\}] && (16)\\
=& 2\xi \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(x_2, a_2)], && (17)
\end{aligned}$$

where the second equality is based on the fact that $X_t \sim \mathrm{Unif}(\mathcal{X})$, and the third equality is due to that fact that $A_t \perp X_t$ because $\pi_t \in \Pi^{(1)}$ is context-independent.

For any ϵ, we further provide a lower bound on the sum of squared simple regret of $\hat{\pi}$. Consider a fixed online learning algorithm $L^{(1)}$.

$$\begin{aligned}
\inf_{L^{(2)}} \sup_{P \in \mathcal{P}} \mathbb{E}_P^L[\mathrm{SR}(\hat{\pi})] &\geq \frac{1}{2} \inf_{L^{(2)}} \left(\mathbb{E}^L_{P_{\epsilon,\xi}}[\mathrm{SR}(\hat{\pi})] + \mathbb{E}^L_{\bar{P}_{\epsilon,\xi}}[\mathrm{SR}(\hat{\pi})] \right) && (18)\\
&\geq \frac{\epsilon}{2} \inf_{L^{(2)}} \left(\mathbb{E}^L_{P_{\epsilon,\xi}}[1 - \mu_{\hat{\pi}}(x_2, a_2)] + \mathbb{E}^L_{P_{\epsilon,\xi}}[\mu_{\hat{\pi}}(x_2, a_2)] \right) && (19)\\
&\geq \frac{\epsilon}{4} \inf_{L^{(2)}} \left(\mathbb{P}^L_{P_{\epsilon,\xi}} \left(\mu_{\hat{\pi}}(x_2, a_2) \leq \frac{1}{2} \right) + \mathbb{P}^L_{\bar{P}_{\epsilon,\xi}} \left(\mu_{\hat{\pi}}(x_2, a_2) > \frac{1}{2} \right) \right), && (20)
\end{aligned}$$

where the second inequality is based on the fact that a_1 is an ϵ-suboptimal arm for $P_{\epsilon,\xi}$, while a_2 is an ϵ-suboptimal arm for $\bar{P}_{\epsilon,\xi}$.

Lemma 1 (Bretagnolle–Huber inequality)**.** *For any two probability distributions P, Q on the same measurable space $(\mathcal{X}, \mathcal{F})$, and any event $A \in \mathcal{F}$, we have*

$$P(A) + Q(\bar{A}) \geq \frac{1}{2} \exp(-D_{KL}(P \| Q)).$$

It follows from Lemma 1 and the fact that $P_{\epsilon,\xi}$, $\bar{P}_{\epsilon,\xi}$ only differs in (x_2, a_2) that

$$\inf_{L^{(2)}} \left(\mathbb{P}^L_{P_{\epsilon,\xi}} \left(\mu_{\hat{\pi}}(x_2, a_2) \le \frac{1}{2} \right) + \mathbb{P}^L_{\bar{P}_{\epsilon,\xi}} \left(\mu_{\hat{\pi}}(x_2, a_2) > \frac{1}{2} \right) \right) \tag{21}$$

$$\ge \frac{1}{2} \exp(-\mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(x_2, a_2)] D_{KL}(P_{\epsilon,\xi} \| \bar{P}_{\epsilon,\xi})) \tag{22}$$

$$= \frac{1}{2} \exp\left(-\frac{1}{2} \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(x_2, a_2)] \log\left(\frac{1}{1 - 4\epsilon^2} \right) \right) \tag{23}$$

$$\ge \frac{1}{2} \exp\left(-2\mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(x_2, a_2)] \epsilon^2 \right) \tag{24}$$

where the second inequality holds because $\log(1/(1-x)) \ge x$ for any $x \in [0, 1/2]$.

Combined with (20), we have for any given online learning algorithm $L^{(1)}$, and any $\xi \in [0, 0.25]$,

$$\inf_{L^{(2)}} \sup_{P \in \mathcal{P}} \mathbb{E}^L_P[\mathrm{SR}(\hat{\pi})] \ge \frac{\epsilon}{8} \exp(-2\epsilon^2 \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(x_2, a_2)]) = \frac{\epsilon}{8} \exp(-2\epsilon^2 \mathbb{E}^{L^{(1)}}_{P_{0,\xi}}[T(x_2, a_2)]). \tag{25}$$

The second equality holds because $\mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[\cdot] = \mathbb{E}^{L^{(1)}}_{P_{0,\xi}}[\cdot]$ for any ϵ due to the fact that the online learning algorithm in task 1 learns a multi-armed bandit (MAB), and we have the same (MAB) for all ϵ because the marginal distribution $(P(\cdot \mid x_1, a) + P(\cdot \mid x_2, a))/2$ is independent of ϵ.

Combined with (31), we have

$$\inf_{L^{(2)}} \sup_{P \in \mathcal{P}} \mathbb{E}^L_P[\mathrm{SR}(\hat{\pi})] \ge \frac{\epsilon}{8} \exp(-\epsilon^2 \mathbb{E}^{L^{(1)}}_{P_{0,\xi}}[\mathrm{CR}]/\xi). \tag{26}$$

To finish the proof, we follow a similar argument in Simchi-Levi and Wang (2023). We let $\hat{P}_L = \arg\max_{P \in \mathcal{P}} \mathbb{E}^L_P[\mathrm{CR}]$, and $\epsilon = \sqrt{\xi/\mathbb{E}^L_{\hat{P}_L}[\mathrm{CR}]}$. By symmetry $P_{0,\xi}$ and $\bar{P}_{0,\xi}$ represents the same space, we know that $\mathbb{E}^L_{\hat{P}_L}[\mathrm{CR}] = \mathbb{E}^L_{P_{0,\xi}}[\mathrm{CR}]$ for some ξ. Then for any L, we apply (26)

$$\sup_{P \in \mathcal{P}} \mathbb{E}^L_P[\mathrm{SR}] \ge \frac{\epsilon}{8} \exp(-\epsilon^2 \mathbb{E}^L_{P_{0,\xi}}[\mathrm{CR}]/\xi) \ge \epsilon/(8e). \tag{27}$$

Plugging in ϵ, we have

$$\sup_{P \in \mathcal{P}} \left[\mathbb{E}^L_P[\mathrm{SR}] \sqrt{\mathbb{E}^L_P[\mathrm{CR}]} \right] \ge \left(\mathbb{E}^L_{\hat{P}_L}[\mathrm{SR}] \sqrt{\mathbb{E}^L_{\hat{P}_L}[\mathrm{CR}]} \right) \tag{28}$$

$$= \left(\frac{\epsilon}{8e} \sqrt{\mathbb{E}^L_{\hat{P}_L}[\mathrm{CR}]} \right) \tag{29}$$

$$= \Theta(1). \tag{30}$$

This argument holds for any choice of L, which completes the proof.

A.2 Lower Bound When $f^{(1)} = f^{(2)}, \Pi^{(1)} = \Pi^{(2)}$

The proof for the second statement in Theorem 1 relies on the lower bound construction where the policy space Π remains the same. Recall that we construct a multi-armed bandit environment without context. The outcome vector is a two-dimensional vector $Y_t = (Y_{t,1}, Y_{t,2})$. Table 4 specifies a set of outcome distributions parametrized by ϵ, ξ. Each pair of ϵ, ξ determines the mean rewards for $\mathbb{E}[Y_{t,i} \mid A_t = a_j]$ for all $i, j \in \{1, 2\}$.

Table 4: Mean reward for $f^{(1)} \neq f^{(2)}$ case

$P_{\epsilon,\xi}$	$Y_{t,1}$	$Y_{t,2}$	$\bar{P}_{\epsilon,\xi}$	$Y_{t,1}$	$Y_{t,2}$
a_1	0.5	0.5	a_1	0.5	0.5
a_2	0.5-ξ	0.5-ϵ	a_2	0.5-ξ	0.5+ϵ

Following a similar proof in the previous section, we let $T(a) = \sum_{t=1}^{T} \mathbb{1}\{A_t = a\}$ be the total number of pulls of a in task 1.

We first note that

$$\mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[\mathrm{CR}] = \xi \sum_{t=1}^{T} \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[\mathbb{1}\{A_t = a_2\}] = \xi \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(a_2)]. \tag{31}$$

For any ϵ, we further provide a lower bound on the sum of squared simple regret of $\hat{\pi}$. Consider a fixed online learning algorithm $L^{(1)}$.

$$\begin{aligned}
\inf_{L^{(2)}} \sup_{P \in \mathcal{P}} \mathbb{E}^L_P[\mathrm{SR}(\hat{\pi})] &\geq \frac{1}{2} \inf_{L^{(2)}} \left(\mathbb{E}^L_{P_{\epsilon,\xi}}[\mathrm{SR}(\hat{\pi})] + \mathbb{E}^L_{\bar{P}_{\epsilon,\xi}}[\mathrm{SR}(\hat{\pi})] \right) &(32)\\
&\geq \frac{\epsilon}{2} \inf_{L^{(2)}} \left(\mathbb{E}^L_{P_{\epsilon,\xi}}[1 - \mu_{\hat{\pi}}(a_2)] + \mathbb{E}^L_{P_{\epsilon,\xi}}[\mu_{\hat{\pi}}(a_2)] \right) &(33)\\
&\geq \frac{\epsilon}{4} \inf_{L^{(2)}} \left(\mathbb{P}^L_{P_{\epsilon,\xi}} \left(\mu_{\hat{\pi}}(a_2) \leq \frac{1}{2} \right) + \mathbb{P}^L_{\bar{P}_{\epsilon,\xi}} \left(\mu_{\hat{\pi}}(a_2) > \frac{1}{2} \right) \right), &(34)
\end{aligned}$$

where the second inequality is based on the fact that a_1 is an ϵ-suboptimal arm for $P_{\epsilon,\xi}$, while a_2 is an ϵ-suboptimal arm for $\bar{P}_{\epsilon,\xi}$.

It follows from Lemma 1 and the fact that $P_{\epsilon,\xi}$, $\bar{P}_{\epsilon,\xi}$ only differs in a_2 that

$$\begin{aligned}
&\inf_{L^{(2)}} \left(\mathbb{P}^L_{P_{\epsilon,\xi}} \left(\mu_{\hat{\pi}}(a_2) \leq \frac{1}{2} \right) + \mathbb{P}^L_{\bar{P}_{\epsilon,\xi}} \left(\mu_{\hat{\pi}}(a_2) > \frac{1}{2} \right) \right) &(35)\\
\geq &\frac{1}{2} \exp(-\mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(a_2)] D_{KL}(P_{\epsilon,\xi} \| \bar{P}_{\epsilon,\xi})) &(36)\\
= &\frac{1}{2} \exp\left(-\frac{1}{2} \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(a_2)] \log\left(\frac{1}{1 - 4\epsilon^2} \right) \right) &(37)\\
\geq &\frac{1}{2} \exp\left(-2\mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(a_2)]\epsilon^2 \right) &(38)
\end{aligned}$$

where the second inequality holds because $\log(1/(1-x)) \geq x$ for any $x \in [0, 1/2]$.

Combined with (34), we have for any given online learning algorithm $L^{(1)}$, and any $\epsilon, \xi \in [0, 0.25]$,

$$\inf_{L^{(2)}} \sup_{P \in \mathcal{P}} \mathbb{E}^L_P[\mathrm{SR}(\hat{\pi})] \geq \frac{\epsilon}{8} \exp(-2\epsilon^2 \mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(a_2)]). \tag{39}$$

By symmetry, there exists some $\bar{P}_{0,\xi}$ such that $\mathbb{E}^{L^{(1)}}_{P_{\epsilon,\xi}}[T(a_2)] = \mathbb{E}^{L^{(1)}}_{\bar{P}_{0,\xi}}[T(a_2)]$. To finish the proof, we follow a similar argument in Simchi-Levi and Wang (2023). We let $\hat{P}_L = \arg\max_{P \in \mathcal{P}} \mathbb{E}^L_P[\mathrm{CR}]$, and $\epsilon = \sqrt{\xi / \mathbb{E}^L_{\hat{P}_L}[\mathrm{CR}]}$. By symmetry $\bar{P}_{0,\xi}$ represents the whole set of $\mathcal{P}$, we know that $\mathbb{E}^L_{\hat{P}_L}[\mathrm{CR}] = \mathbb{E}^L_{\bar{P}_{0,\xi}}[\mathrm{CR}]$ for some ξ. Then for any L, we apply (26)

$$\sup_{P \in \mathcal{P}} \mathbb{E}^L_P[\mathrm{SR}] \geq \frac{\epsilon}{8} \exp(-\epsilon^2 \mathbb{E}^L_{\bar{P}_{0,\xi}}[\mathrm{CR}]/\xi) \geq \epsilon/(8e). \tag{40}$$

Plugging in ϵ, we have

$$\begin{aligned}
\sup_{P \in \mathcal{P}} \left[\mathbb{E}^L_P[\mathrm{SR}] \sqrt{\mathbb{E}^L_P[\mathrm{CR}]} \right] &\geq \left(\mathbb{E}^L_{\hat{P}_L}[\mathrm{SR}] \sqrt{\mathbb{E}^L_{\hat{P}_L}[\mathrm{CR}]} \right) &(41)\\
&= \left(\frac{\epsilon}{8e} \sqrt{\mathbb{E}^L_{\hat{P}_L}[\mathrm{CR}]} \right) &(42)\\
&= \Theta(1). &(43)
\end{aligned}$$

This argument holds for any choice of L, which completes the proof.

B Proof of Proposition 1.

Proposition 1. *Following the same choice of the instance set $\mathcal{P}$ and $\Pi^{(1)}, \Pi^{(2)}, f^{(1)}, f^{(2)}$ in Theorem 1, the following minimax lower bound holds*

$$\inf_{L\in\mathcal{L}} \sup_{P\in\mathcal{P}} \mathbb{E}[\mathrm{CR} + T' \,\mathrm{SR}] = \Omega\left(\max\left\{\frac{T'}{\sqrt{T}}, T'^{2/3}, \sqrt{T}\right\}\right). \tag{44}$$

Proof. Since $\inf_{L\in\mathcal{L}} \sup_{P\in\mathcal{P}} \sqrt{\mathbb{E}[\mathrm{CR}]} \times \mathbb{E}[\mathrm{SR}] = \Theta(1)$, we have

$$\inf_{L\in\mathcal{L}} \sup_{P\in\mathcal{P}} \mathbb{E}[\mathrm{CR} + T' \,\mathrm{SR}] = \Omega\left(\inf_{L\in\mathcal{L}} \sup_{P\in\mathcal{P}} \frac{1}{(\mathbb{E}[\mathrm{SR}])^2} + \mathbb{E}[T' \,\mathrm{SR}]\right) = \Omega\left(T'^{2/3}\right). \tag{45}$$

The well-established cumulative regret lower bound (Lattimore and Szepesvári, 2020) gives us

$$\inf_{L\in\mathcal{L}} \sup_{P\in\mathcal{P}} \mathbb{E}[\mathrm{CR}] = \Omega(\sqrt{T}). \tag{46}$$

An information-theoretic lower bound on the estimation error gives us

$$\inf_{L\in\mathcal{L}} \sup_{P\in\mathcal{P}} T'\mathbb{E}[\mathrm{SR}] = \Omega(T'/\sqrt{T}). \tag{47}$$

The proof is finished by combining the above three lower bounds. □

C Proof of Theorem 2

Theorem 2. *Let L_0 be an online learning algorithm with a regret bound of $\tilde{\mathcal{O}}(\sqrt{|\mathcal{X}||\mathcal{A}|T})$. Let the algorithm for the first task be $L_\alpha(\tau) = (1-\alpha)L_0(\tau) + \alpha\pi_0$, where τ is any past observations and π_0 is the uniform random policy. For any choice of $\alpha \in [|\mathcal{X}||\mathcal{A}|/\sqrt{T}, 1]$, there exist offline-learning algorithm for the second task such that*

$$\mathbb{E}[\mathrm{CR}] = \tilde{\mathcal{O}}\left(\alpha T\right), \mathbb{E}[\mathrm{SR}] = \tilde{\mathcal{O}}\left(\sqrt{\frac{(|\mathcal{X}||\mathcal{A}|)^2}{\alpha T}}\right). \tag{48}$$

Proof. The upper bound is straightforward from

$$\mathbb{E}[\mathrm{CR}] = \tilde{O}(\sqrt{|\mathcal{X}||\mathcal{A}|T} + \alpha T) = \tilde{O}(\alpha T),$$

where the second inequality is due to the fact that $\alpha T \geq |\mathcal{X}||\mathcal{A}|\sqrt{T}$. To establish a valid simple regret upper bound, we employ an importance weight estimator:

$$\hat{R}(x,a) := \frac{\sum_{t=1}^{T}\left[\mathbb{1}\{(X_t, A_t) = (x,a)\}\frac{\pi_0(A_t|X_t)}{L_\alpha(\tau_t)(A_t|X_t)} f^{(2)}(Y_t)\right]}{\sum_{t=1}^{T}\left[\mathbb{1}\{(X_t, A_t) = (x,a)\}\frac{\pi_0(A_t|X_t)}{L_\alpha(\tau_t)(A_t|X_t)}\right]}. \tag{49}$$

Theorem 2.1. (Guo et al., 2024) states that it holds for all x, a that

$$|\hat{R}(x,a) - R(x,a)| = \tilde{O}\left(\sqrt{\frac{|\mathcal{X}||\mathcal{A}|^2}{\alpha T}}\right). \tag{50}$$

Then the policy $\hat{\pi}$ defined by

$$\hat{\pi}(a \mid x) = \mathrm{Unif}(\{\arg\max_{a'\in\mathcal{A}} \hat{R}(x,a')\}) \tag{51}$$

achieves the simple regret bound of $\tilde{O}(|\mathcal{X}||\mathcal{A}|/\sqrt{\alpha T})$. □

D Proof of Theorem 3

Theorem 3. *Assume $\mathcal{S}$ is such that $\Pi^{(1)} = \Pi^{(2)} = \Pi$, the set of all policies, and $\Pi^{(1)} = \Pi^{(2)} = f$, the identical mapping in $\mathbb{R}$. Assume each $P^{(i)}(\cdot \mid x, a)$ is from a binomial distribution with mean $P^{(i)}(x, a)$ for all $i = 1, 2$ and $x, a \in \mathcal{X} \times \mathcal{A}$, and $P^{(2)} \in \mathcal{P}(P^{(1)}, B)$. Then there exists some B such that*

$$\inf_{L \in \mathcal{L}_2} \sup_{P^{(1)}} \sqrt{\mathbb{E}[\mathrm{CR}_1]} \mathbb{E}[\widetilde{\mathrm{SR}}(\pi_2 \mid P^{(1)}, B)] = \Omega(1), \tag{52}$$

where π_2 is the random policy chosen by learning algorithm L.

Proof. We construct two hard instances inspired by Proposition 2. Recall that Proposition 2 states that any two-armed, context-free bandit, with $\mathrm{Gap} \coloneqq P^{(1)}(a_1) - P^{(1)}(a_2) > 0$ has the following explicit form of optimal robust policy:

$$\tilde{\pi}(a_1) = \frac{B + \mathrm{Gap}}{2B}, \text{ and } \tilde{\pi}(a_2) = \frac{B - \mathrm{Gap}}{2B}. \tag{53}$$

Let the arm space be $\mathcal{A} = \{a_1, a_2\}$. We construct two instances S and $\bar{S}$ with $P^{(1)}$ and $\bar{P}^{(1)}$, such that $R^{(1)}(a_1) = \bar{R}^{(1)}(a_1) = 1$ and $R^{(1)}(a_2) = 1/2$, while $\bar{R}^{(1)}(a_2) = 1/2 + \epsilon$ for $\epsilon < 1/4$. Let $P^{(i)}(\cdot \mid a)$ and $\bar{P}^{(i)}(\cdot \mid a)$ be Bernoulli distributions of parameter $R^{(i)}(a)$ and $\bar{R}^{(i)}(a)$ for each $i \in \{1, 2\}, a \in \mathcal{A}$. Let $\mathrm{Gap} = R^{(1)}(a_1) - R^{(1)}(a_2) = 1/2$ and $\overline{\mathrm{Gap}} = \bar{R}^{(1)}(a_1) - \bar{R}^{(1)}(a_2) = 1/2 - \epsilon$.

We first connect the cumulative regret in the first task CR_1 with the number of visits in the suboptimal arm a_2 in the first task. It can be shown that $\mathbb{E}_{\boldsymbol{S'}}[\mathrm{CR}_1] \geq 1/4 \mathbb{E}_{\boldsymbol{S'}}[T(a_2)]$, where $T(a_2) \coloneqq \sum_{t=1}^{T_1} \mathbb{1}_{A_{1,t} = a_2}$ for each $S' \in \{S, \bar{S}\}$.

We consider $B = 3/4 > \max\{\mathrm{Gap}, \overline{\mathrm{Gap}}\}$. By Proposition 2, we first lower bound the robust simple regret by

$$\begin{aligned}
&\widetilde{\mathrm{SR}}(\pi \mid P^{(1)}, B) && (54)\\
=&\mathrm{SR}(\pi \mid P^{(1)}, B) - \inf_{\pi'} \mathrm{SR}(\pi' \mid P^{(1)}, B) && (55)\\
=&\max\{(B - \mathrm{Gap})\pi(a_1), (B - \mathrm{Gap})\pi(a_2)\} - \frac{B^2 - \mathrm{Gap}^2}{2B} && (56)\\
=&\left(\pi(a_1) - \frac{B + \mathrm{Gap}}{2B}\right)^+ (B - \mathrm{Gap}) + \left(\pi(a_2) - \frac{B - \mathrm{Gap}}{2B}\right)^+ (B + \mathrm{Gap}) && (57)\\
\geq&(B - \mathrm{Gap})|\pi - \pi^*|/2 \geq |\pi - \pi^*|/8. && (58)
\end{aligned}$$

Similarly, we also have $\widetilde{\mathrm{SR}}(\pi \mid \bar{P}^{(1)}, B) \geq |\pi - \bar{\pi}^*|/8$. Here we let π^*, $\bar{\pi}^*$ be the optimal robust policy for $P^{(2)} \in \mathcal{P}(P^{(1)} \mid B)$ and $P^{(2)} \in \mathcal{P}(\bar{P}^{(1)} \mid B)$, respectively.

Let π_2 be the random policy proposed by the learning algorithm for the second task. We convert the robust learning problem to a testing problem of two instances. Note that $\pi^*(a_1) = 5/6$ and $\bar{\pi}^*(a_1) = 5/6 - 2\epsilon/3$.

The sum of robust simple regrets for two instances can be lower bounded by

$$\begin{aligned}
&\mathbb{E}_{P^{(1)}}[\widetilde{\mathrm{SR}}_2] + \mathbb{E}_{\bar{P}^{(1)}}[\widetilde{\mathrm{SR}}_2] && (59)\\
\geq&\frac{\epsilon}{24} \mathbb{P}_{P^{(1)}} \left(\pi_2(a_1) \leq \frac{5}{6} - \epsilon/3\right) + \frac{\epsilon}{24} \mathbb{P}_{\bar{P}^{(1)}} \left(\pi_2(a_1) > \frac{5}{6} - \epsilon/3\right) && (60)\\
\geq&\frac{\epsilon}{24} \exp(-\epsilon^2 \mathbb{E}_{P^{(1)}}[T(a_2)]) && (61)
\end{aligned}$$

Choosing $\epsilon = \sqrt{1/\mathbb{E}_{P^{(1)}}[T(a_2)]}$ and applying Lemma 1 again, we have

$$\inf_{L \in \mathcal{L}} \sup_{P^{(1)}} \sqrt{\mathbb{E}[\mathrm{CR}_1]} \mathbb{E}[\widetilde{\mathrm{SR}}(\pi_2 \mid P^{(1)}, B)] = \Omega(1).$$

□

E Proof of Theorem 4

Theorem 4 *Consider the hypothesis set $\mathcal{F}$ to be one-hidden layer neural networks with width d. There exists ground-truth reward function and a sequence of tasks of length $\Omega(\exp(d))$ with different $\Pi^{(i)}$, such that the cumulative regret for each task is lower bounded by a constant, even if each $T_i \to \infty$.*

Proof. The construction of the hard instance can be described below. Consider a nonlinear bandit problem with $\mathcal{A} = S^{d-1}$, the d-dimensional sphere. We first define the reward function as

$$R(\theta_2) = \alpha_1 \langle \theta_1, a \rangle + \alpha_2 \left(\langle \theta_2, a \rangle - \epsilon \right)^+ ,$$

where $\theta_1 \in S^{d-1}, \alpha_1 > 0$ and $\alpha_2 > 0$ are known parameters and $\theta_2 \in S^{d-1}$ is unknown. Assume that the true parameter θ_2^* satisfies $\langle \theta_1, \theta_2^* \rangle < 0$, i.e., θ_1 and θ_2^* are on different sphere. Furthermore, let $\alpha_2 = 2\alpha_2/(1-\epsilon)$.

Let $\{\mathcal{A}_1, \dots, \mathcal{A}_N\}$ be an ϵ-pack of the subset $\{a \in \mathcal{A} : \langle \theta_1, a \rangle < 0\}$. Let the allowed policy space for task i be $\Pi^{(i)} = \{\pi : \pi \text{ is supported on } \mathcal{A}_i \cup \{\theta_1\}\}$. Specifically, order $\{\mathcal{A}_1, \dots, \mathcal{A}_N\}$ such that $\theta_2^* \in \mathcal{A}_N$.

To verify, $R(\theta_2)$ is in the family of one-layer neural network with ReLU activation function.

We first observe that the optimal policy for task i is $\pi_i^* = \delta(\theta_1)$ for all $i = 1, \dots N-1$. Note that by running UCB, the algorithm will optimistically choose $a \in \mathcal{A}_i$ as they do not know that whether $\theta_2^* \in \mathcal{A}_i$ for all $i = 1, \dots, N-1$, thus leading to a constant regret for all tasks $i < N$. □

F Simulation Studies

The linear contextual bandit assumes linear reward functions. Specifically, in our simulation studies, we consider the following reward functions:

$$R_t = \langle \theta_{A_t}, X_t \rangle + \epsilon_t,$$

where $X_t \in \mathcal{X} \subset \mathbb{R}^d$ is the context and $\theta_a \in \mathbb{R}^d$ is the true reward coefficient for action a.

Below, we list the hyper-parameters for the environment based on the hard instance and the random environment, respectively.

- **Hard instance: environment 1.** We run a d-dimensional linear contextual bandit with $d = 4$ and the context set is given by $\mathcal{X} = (0,0,0,1), (0,0,1,0), (0,1,0,0), (1,0,0,0) = e_1, e_2, e_3, e_4$, and the true reward coefficients are

$$\theta_1 = \begin{pmatrix} 1 \\ 0.2 \\ 0 \\ 0 \end{pmatrix}, \quad \theta_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \theta_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \theta_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

- **Hard instance: environment 2.** We run a d-dimensional linear contextual bandit with $d = 1$ and the context set is given by $\mathcal{X} = \{1\}$, and the true reward coefficients for task 1 and 2 are

$$\theta_{1:4}^{(1)} = \begin{pmatrix} 1 \\ 0.5 \\ 0.2 \\ 0.1 \end{pmatrix}, \quad \theta_{1:4}^{(2)} = \begin{pmatrix} 0.9 \\ 1.0 \\ 0.2 \\ 0.1 \end{pmatrix}.$$

- **Random environment.** We run a d-dimensional linear contextual bandit with $d = 4$ and both the context and the true reward coefficients are drawn independently from standard Gaussian distributions.