Learning When Not to Learn: Risk-Sensitive Abstention in Bandits with Unbounded Rewards

Sarah Liaw*
Harvard University
University of California, Berkeley

Benjamin Plaut*
University of California, Berkeley

Abstract

In high-stakes AI applications, even a single action can cause irreparable damage. However, nearly all of sequential decision-making theory assumes that all errors are recoverable (e.g., by bounding rewards). Standard bandit algorithms that explore aggressively may cause irreparable damage when this assumption fails. Some prior work avoids irreparable errors by asking for help from a mentor, but a mentor may not always be available. In this work, we formalize a model of learning with unbounded rewards without a mentor as a two-action contextual bandit with an abstain option: at each round the agent observes an input and chooses either to abstain (always 0 reward) or to commit (execute a preexisting task policy). Committing yields rewards that are upper-bounded but can be arbitrarily negative, and the commit reward is assumed Lipschitz in the input. We propose a caution-based algorithm that learns when not to learn: it chooses a trusted region and commits only where the available evidence does not already certify harm. Under these conditions and i.i.d. inputs, we establish sublinear regret guarantees, theoretically demonstrating the effectiveness of cautious exploration for deploying learning agents safely in high-stakes environments.

*Equal contribution.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

With artificial intelligence becoming ubiquitous, many learning systems are now deployed in unpredictable, safety-critical domains, such as process control and manufacturing robotics, autonomous driving, and surgical assistance. In these settings, a single ill-chosen action can cause irreparable and lasting damage with no opportunity for subsequent recovery. For instance, a self-driving car cannot compensate for a deadly crash by later driving more safely, nor can a medical robot undo a fatal mistake during surgery. Following Plaut et al. (2025a,b), we refer to such irreparable errors as *catastrophes*.

Despite the risks such deployments pose, there is limited work (and limited theoretical work in particular) on how an agent can learn without ever incurring an irreparable error. The possibility of catastrophes challenges standard frameworks for sequential decision making, especially the familiar notion of *optimism under uncertainty*. Optimism effectively assumes that early mistakes can be offset (or be compensated for) by later gains, an assumption that is inappropriate when errors are irrecoverable. Instead, these settings call for *pessimism under uncertainty*: when evidence is insufficient, prefer inaction to risky action.

One approach to mitigate these problems is to let the agent ask for help from a mentor in unfamiliar or risky situations. Such human-in-the-loop oversight can block unsafe actions and prevent irreparable errors (even if ordinary, recoverable errors still occur). However, this approach depends on the availability of a capable mentor, which can be costly or impractical at scale. This motivates a mentor-free alternative: **can an agent avoid irreparable errors on its own by acting cautiously when inputs appear unfamiliar?**

We propose a model of learning in the presence of irreparable costs without a mentor but with an option to abstain from action. The key question is when to abstain, i.e., when not to learn. To focus on this

question, we assume the agent has previously learned a baseline policy that works well in-distribution but behaves unpredictably elsewhere. This allows us to streamline the model to two actions: *abstain* (do nothing) and *commit* (follow the baseline policy). Abstaining yields a deterministic safe reward $r(x, 0) = 0$, while committing yields a reward $r(x, 1) \in (-\infty, 1]$[1].

We treat the origin as fully "in-distribution" and assume the baseline policy is beneficial there: $r(\mathbf{0}, 1) > 0$. We use the distance from the origin $\|x\|$ as a measure of how out-of-distribution (OOD) an input is. The commit reward is assumed L-Lipschitz, capturing the idea that similar inputs yield similar outcomes. For our main results we focus on a fixed distribution ν; for our impossibility results we also consider a T-dependent distribution ν_T.

We formalize the tension between exploration and safety via two negative results. First, in the worst case, any algorithm that begins by always exploring (i.e., commits on the first round regardless of the input) can suffer infinite expected regret (Thm. 4.1). Second, when every input lies uniformly far OOD, there is no safe way to explore to identify a beneficial committing region, and sublinear regret is impossible (Thm. 4.2). Together, these results delineate both the necessity and the limits of caution.

Motivated by this perspective, we develop a caution-based algorithm that learns only when it can guarantee that an error is not catastrophic (which essentially corresponds to not-too-OOD inputs). This approach yields sublinear expected regret for i.i.d. inputs from any fixed distribution, with bounds that also reflect how often the agent encounters far OOD inputs, while prioritizing the avoidance of irreparable errors. As in standard Lipschitz contextual bandits, however, this regret bound has an exponential dependence on the dimension n, and is therefore most meaningful in low-dimensional settings.

Contributions. Our contributions are:

1. We introduce a formal model of learning with irreparable costs and no external mentor.
2. We prove two impossibility results that delineate the necessity and limits of caution.
3. We develop a caution-based algorithm that achieves sublinear regret for any fixed input distribution.

Organization. §3 introduces the formal model and notation. §4 presents the impossibility results (Thms. 4.1 and 4.2) and their implications for exploration. § 5 describes our caution-based learning algorithm and states the main regret bound. § 6 outlines the proof strategy and supporting lemmas. Appendix A presents the complete proof, and Appendix B provides some numerical simulations.

[1]The asymmetric bounds on the commit reward reflect that a single action can be catastrophic, whereas it is rare for a single action to yield arbitrarily large benefit.

2 RELATED WORK

Most prior work on sequential decision-making and safe exploration focuses on settings where errors are ultimately recoverable; here we contrast this with our setting where individual actions can cause irreparable harm.

2.1 Sequential decision-making when all errors are recoverable

The literature on sequential decision-making is vast, spanning bandit problems, reinforcement learning, and online learning. See Slivkins et al. (2019), Sutton et al. (1998), and Cesa-Bianchi and Lugosi (2006) for introductions to these (somewhat overlapping) topics, respectively. However, nearly all of this work assumes explicitly or implicitly that any error can be recovered from. This assumption enables the agent to ignore risk and simply try all possible behaviors, since no matter how badly it performs in the short term, it can always eventually make up for it. Indeed, most sequential decision-making algorithms with formal regret bounds have this general structure.

This assumption can manifest in different ways. In bandit settings, it suffices to assume that rewards are bounded (or at least have bounded expectation). This assumption implies that the expected regret from any action on any time step is always bounded, which is sufficient for the risk-agnostic exploration mentioned above. In contrast, we allow unbounded negative rewards so that actions can be arbitrarily costly. Indeed, our first negative result (Thm. 4.1) relies on the expected regret for a single action potentially being infinite in our model.

In Markov Decision Processes (MDPs), the agent's actions determine the next state via a transition function, so in addition to bounded rewards, one typically assumes that either the environment is reset at each of each "episode" (e.g., Azar et al., 2017) or that any state is reachable from any other (e.g., Jaksch et al., 2010). The reliance of standard MDP algorithms on these assumptions was observed by Moldovan and Abbeel (2012a); Cohen et al. (2021), among others.

Regardless of the specific form of this assumption, it clearly does not hold in safety-critical contexts where a single action can be catastrophic.

2.2 Safe exploration

These issues have motivated a wide field of safe exploration. A full survey is beyond the scope of this paper (see García and Fernández, 2015; Gu et al., 2024; Krasowski et al., 2023; Tan et al., 2022 for surveys), so we cover only the most relevant prior work. Avoiding irreparable errors while learning has also been studied empirically across multiple domains (e.g., Saunders et al., 2017; Moldovan and Abbeel, 2012b; Wachi et al., 2023; Zhao et al., 2023; Perkins and Barto, 2003), but here we focus on theoretical work, which is most relevant to our setting.

Safe exploration is modeled in two main ways. The first approach is to require the agent to satisfy some sort of constraint in addition to maximizing reward. The constraint can be entirely separate from reward, as in the case of constrained MDPs (Altman, 1999), or they can be related to the reward (e.g., the agent's reward must always exceed some baseline). When zero or near-zero constraint violation is required, these formalisms do capture the possibility of irreparable errors. The second approach treats reward as the sole objective, with safety as a necessary but not sufficient property for maximizing reward. Here, irreparable errors correspond to either unboundedly negative rewards (our work falls into this category) or inescapable "trap" states with poor reward. An agent that obtains very negative rewards or enters trap states clearly cannot obtain high reward.

Both of these models must contend with a fundamental obstacle: how does one learn which actions are catastrophic without trying those actions directly? This can be formalized by the so called "Heaven or Hell problem" (Plaut et al., 2025a). Suppose there are two available actions, where one has unbounded positive reward and the other has unbounded negative reward. In this case, the agent can do no better than simply guessing and can never guarantee good regret. This problem shows that some sort of additional assumption is necessary for any meaningful regret guarantees. Below, we categorize work within safe exploration based on which assumption(s) it uses for this purpose.

Full prior knowledge. Perhaps the simplest approach is to assume that the agent knows the precise safety constraint upfront (see Zhao et al., 2023 for a survey). This immediately resolves the Heaven or Hell problem; indeed, it eliminates the need for the agent to "learn when not to learn" at all. However, full knowledge of the safety constraint may not hold in practice. In contrast, we only assume that the (1) baseline policy performs well in-distribution and (2) the agent can always safely abstain.

Learning constraints using a safe fallback action. There is a growing body of work which shares our assumption of a safe fallback action. Liu et al. (2021); Stradi et al. (2024) use this approach in the constrained MDP model, while Wu et al. (2016); Kazerouni et al. (2017); Lin et al. (2022); Chen et al. (2022) require the reward to exceed a fixed baseline in a bandit model. These papers generally rely on a pair of subtle but crucial assumptions to obtain zero constraint violation: (1) the constraint violation on any given time step is bounded and (2) the baseline policy satisfies the constraints with a known amount of slack (this property is often referred to as *Slater's gap* (Bernasconi et al., 2024), although not all of the above papers use this terminology). This combination of assumptions enables the agent to still explore aggressively with some known probability. Furthermore, the resulting bounds typically depend inversely on Slater's gap.

Our work is complementary to each of these two assumptions. First, rather than assuming global boundedness, we assume that rewards decrease at a bounded *rate*, i.e., rewards are Lipschitz continuous. Second, rather than dependence on the reward or cost function (in the form of Slater's gap), our bounds depend on the input distribution: specifically, our bounds degrade as the agent sees more OOD inputs. Our approach may be more or less realistic depending on the specific context, but it notably differs from the typical way fallback actions are utilized.

Risk-sensitive contextual Lipschitz bandits. We are not aware of prior work on risk-sensitive contextual bandits with Lipschitz rewards. To our knowledge, the Lipschitz contextual bandit literature (e.g., Slivkins, 2011) does not consider risk-sensitivity, and the risk-sensitive bandit literature is either non-contextual (e.g., Wu et al., 2016) or studies linear bandits instead of Lipschitz bandits (e.g., Lin et al., 2022). Because of these differences, it would be difficult to make an apples-to-apples comparison to these algorithms in our simulations, so we do not include these algorithms.

Asking for help. Perhaps the most common approach in this model is relying on external supervision. This is a growing body of work which uses limited queries to a mentor to prove formal regret guarantees in the presence of irreversible dynamics (Cohen et al., 2021; Cohen and Hutter, 2020; Kosoy, 2019; Maillard et al., 2019; Plaut et al., 2025b,a). However, as the number of deployed AI systems continues to grow, it may be impractical for each one to have a human supervisor. Even in cases where external help will eventually become available, the agent may need to behave safely on its own in the short-term. These considerations motivate our study of how to learn safely in the absence of external help.

2.3 Other related work

We briefly discuss some topics that are less directly relevant but still worth mentioning. One is the heavy-tailed bandit model (Bubeck et al., 2013; Agrawal et al., 2021), which studies the case where reward distributions are not subgaussian and thus less predictable. While this model does incorporate elements of safety, as long as the expected reward from any action is bounded, risk-agnostic exploration remains valid (as discussed above). Another topic adjacent to our work is the standard Lipschitz bandit model with bounded rewards and bounded domain (see, e.g., Chapters 4 and 8 of Slivkins, 2011). This work shares some similarities with ours, like the algorithmic use of discretization. However, the core of our paper is removing the boundedness assumptions, which introduces a host of new challenges. Finally, there is complementary work on abstention with bounded rewards (Neu and Zhivotovskiy, 2020; Yang et al., 2024). While this line of work also demonstrates the benefits of abstention, it does not address the possibility of irreparable errors.

3 PRELIMINARIES

We study a two-action contextual bandit model in which, on each round, the agent observes an input and chooses either to commit, thus executing a fixed task policy that may yield risky outcomes, or to abstain, receiving a safe default reward of zero. In this section, we introduce the formal notation and assumptions used throughout.

For $k \in \mathbb{N}$, let $[k] = \{1, \dots, k\}$. Let $\mathcal{X} = \mathbb{R}^n$ be the input space, $T \in \mathbb{N}$ be the time horizon, and $\|\cdot\|$ be the Euclidean norm (though one could also consider a more general metric space). On each time step $t \in [T]$, the agent observes an input $x_t \in \mathcal{X}$, chooses an action $y_t \in \{0, 1\}$, and receives a (noisy) scalar reward; the precise noise assumptions are stated below.

Actions and Rewards. We interpret $y_t = 0$ as "abstaining", a safe default which deterministically yields $r(x_t, 0) = 0$ for any $x_t \in \mathcal{X}$. We interpret $y_t = 1$ as "committing", which executes a preexisting policy whose reward $r(x_t, 1)$ may be arbitrarily negative (catastrophic) but is assumed to have a constant upper bound (rescaled to 1 without loss of generality). This captures the asymmetry of high-stakes settings where catastrophic losses can be unbounded in magnitude, whereas gains typically saturate.

Input models. We assume inputs are i.i.d. draws from an unknown distribution ν on $\mathcal{X}$, i.e., $x_1, \dots, x_T \overset{\text{i.i.d.}}{\sim} \nu$. We typically take ν to be fixed, but in our impossibility results we also consider the case of T-dependent ν (denoted ν_T).

We assume bandit feedback: the agent observes only the realized reward of its chosen action. Abstaining provides no information about the counterfactual commit reward $r(x_t, 1)$, so the agent cannot "learn by abstaining". Formally, at round t the learner observes

$$r_t = r(x_t, y_t) + \eta_t,$$

where $(\eta_t)_{t=1}^T$ are i.i.d. zero-mean σ-subgaussian noise variables, independent of (x_t) and of the learner's internal randomness (specified formally in Def. 6.1).

Regularity. We make two assumptions on the reward function: (i) the commit reward $r(\cdot, 1)$ is L-Lipschitz in the Euclidean norm, i.e., there exists $L > 0$ such that for all $x, x' \in \mathcal{X}$, $|r(x, 1) - r(x', 1)| \le L\|x - x'\|$. This is a standard smoothness condition in Lipschitz bandit models (see, e.g., Slivkins et al., 2019) and captures the intuition that similar inputs yield similar commit rewards. Since $r(x, 0) \equiv 0$, the abstain reward is 0-Lipschitz. (ii) The in-distribution baseline input yields strictly positive reward when committing, i.e. $r(\mathbf{0}, 1) > 0$. This guarantees that committing is beneficial somewhere (at the origin); without it, the optimal policy would be to always abstain and cautious learning would be impossible.

Objective. The agent's goal is to minimize its (expected) regret, which is the difference between its cumulative reward and the optimal cumulative reward. Formally, define

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \Big(\max_{y^\star \in \{0,1\}} r(x_t, y^\star) \; - \; r(x_t, y_t) \Big).$$

We take the expectations over the input process (in the stochastic model), the observation noise, and the learner's internal randomness. The goal is to achieve sublinear expected regret, i.e., $\mathbb{E}[\mathrm{Reg}(T)] = o(T)$, equivalently $\mathbb{E}[\mathrm{Reg}(T)]/T \to 0$ as $T \to \infty$.

4 THE VIRTUES AND LIMITS OF CAUTION

In this section, we provide two impossibility results that demonstrate the importance and limitations of caution in high-stakes, unbounded reward bandits.

First, *caution is necessary*: if an agent commits with non-negligible probability on inputs that are far OOD, catastrophic tail losses dominate—indeed, even a single risk-agnostic exploratory commit can incur infinite expected regret. This kind of "incautious exploration" is exactly how standard bandit algorithms behave when

they begin by pulling every arm at least once. Second, *caution has limits*: when the input stream is uniformly far OOD, there is no way to explore cautiously to identify a beneficial committing region without risking catastrophe. In such settings, sublinear regret is not possible and the optimal strategy is to abstain on every time step.

Theorem 4.1 (The need for caution)**.** *Let ν be any distribution over $\mathcal{X}$ such that $\mathbb{E}_{x\sim\nu}[\|x\|] = \infty$ and assume $x_1, \ldots, x_T \sim \nu$ i.i.d. Then there exists a reward function r such that any algorithm which always commits on the first time step satisfies $\mathbb{E}[\mathrm{Reg}(T)] = \infty$.*

Proof. Define $r(x, 1) = 1 - L\|x\|$ for all $x \in \mathcal{X}$. Then

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\max_{y^*\in\{0,1\}} r(x_t, y^*) - r(x_t, y_t)\right)\right] \\
&\geq \mathbb{E}\left[\max_{y^*\in\{0,1\}} r(x_1, y^*) - r(x_1, y_1)\right] \\
&\geq \mathbb{E}[0 - (1 - L\|x_1\|)] \\
&= L \mathop{\mathbb{E}}_{x\sim\nu}[\|x\|] - 1 \\
&= \infty
\end{aligned}$$

as required. □

The proof can easily be modified to handle the cases where the first commit is taken with constant probability (rather than probability 1) or where the algorithm abstains for a constant number of initial rounds. Essentially, this negative result applies to any algorithm that is not cautious, i.e., that explores without considering how OOD x_t is.

However, caution can only get us so far. While it prevents catastrophic first commits, some exploration is necessary to obtain sublinear regret. If all inputs are far OOD, then there is no safe way to explore, so the agent has no choice but to always abstain. Equivalently, this can be phrased by considering i.i.d. inputs from a T-dependent distribution ν_T supported on $\{x : \|x\| = T\}$.

Theorem 4.2 (The limits of caution)**.** *Let ν_T be any distribution supported on $\{x : \|x\| = T\}$, and suppose $x_1, \ldots, x_T \overset{i.i.d.}{\sim} \nu_T$. Then no algorithm can guarantee $\mathbb{E}[\mathrm{Reg}(T)] \in o(T)$.*

Proof. Define $r^-(x, 1) := 1 - L\|x\|$ and $r^+(x, 1) := 1$, with $r^\pm(x, 0) := 0$. Since we only care about asymptotics, we can restrict our attention to $T > 1/L$. Then for $\|x_t\| = T$, optimal behavior for r^+ is to always commit, while optimal behavior for r^- is to always abstain. We show $\max_{r\in\{r^-,r^+\}} \mathbb{E}[\mathrm{Reg}(T)] \in \Omega(T)$.

To do so, we use a mild version of the probabilistic method. Let $U(r^-, r^+)$ be the uniform distribution over $\{r^-, r^+\}$. It suffices to show $\mathbb{E}_{r\sim U}\,\mathbb{E}[\mathrm{Reg}(T)] \in \Omega(T)$, where the second expectation is over $x_1, \ldots, x_T$ and $y_1, \ldots, y_T$. Let $\mathcal{E}$ be the event that the agent ever commits. If $\mathcal{E}$ holds, there exists $i \in [T]$ with $y_i = 1$. Since y_i is independent of r,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{r}\ \mathbb{E}[\mathrm{Reg}(T) \mid \mathcal{E}] \\
&= \mathop{\mathbb{E}}_{r}\ \mathbb{E}\left[\sum_{t=1}^{T}\left(\max_{y^*\in\{0,1\}} r(x_t, y^*) - r(x_t, y_t)\right) \mid \mathcal{E}\right] \\
&\geq \Pr[r = r^-]\,\mathbb{E}\left[\max_{y^*\in\{0,1\}} r(x_i, y^*) - r(x_i, y_i) \mid r = r^-\right] \\
&= \frac{LT - 1}{2}
\end{aligned}$$

On the other hand, if $\mathcal{E}$ does not occur, then

$$\begin{aligned}
&\mathop{\mathbb{E}}_{r}\ \mathbb{E}[\mathrm{Reg}(T) \mid \neg\mathcal{E}] \\
&\geq \Pr[r = r^+]\,\mathbb{E}\left[\sum_{t=1}^{T}\Big(\max_{y^*\in\{0,1\}} r^+(x_t, y^*) - r^+(x_t, y_t)\Big) \,\Big|\, \neg\mathcal{E}\right] \\
&\geq \frac{T}{2}.
\end{aligned}$$

Then by the law of total expectation,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{r}\ \mathbb{E}[\mathrm{Reg}(T)] \\
&= \Pr[\mathcal{E}] \mathop{\mathbb{E}}_{r}\ \mathbb{E}[\mathrm{Reg}(T) \mid \mathcal{E}] + \Pr[\neg\mathcal{E}] \mathop{\mathbb{E}}_{r}\ \mathbb{E}[\mathrm{Reg}(T) \mid \neg\mathcal{E}] \\
&\geq \min\big(\Pr[\mathcal{E}], \Pr[\neg\mathcal{E}]\big) \min\left(\frac{LT - 1}{2}, \frac{T}{2}\right) \\
&\in \Omega(T)
\end{aligned}$$

as required. □

5 ALGORITHM AND MAIN RESULT

Following the negative results in § 4, we propose an algorithm (Algorithm 1) that operationalizes cautious learning: only learn in regions that are not too far OOD and where the available evidence does not already certify that committing is harmful.

Informally, we define a trusted region around the origin whose radius grows with the time horizon, reflecting the maximum regret we are willing to tolerate—intuitively, this corresponds to allowing mistakes that are bad but not catastrophic. We then discretize the region into bins to exploit Lipschitz continuity. Within each bin, the commit reward cannot vary by more than a Lipschitz discretization error, so it suffices to estimate a

Algorithm 1 Risk-Sensitive Abstention Algorithm

Inputs: $m : \mathbb{N} \to \mathbb{R}_{>0}$, $w : \mathbb{N} \to \mathbb{R}_{>0}$
$\mathcal{H} \leftarrow$ partition of $\mathcal{X}$ into n-cubes of side length $w(T)$
$\mathcal{B} \leftarrow \{B \in \mathcal{H} : \exists x \in B \text{ with } \|x\| \le m(T)\}$
$\sigma_w \leftarrow \sqrt{nL^2 w(T)^2 + \sigma^2}$
$\gamma(k) := \infty$ if $k = 0$ else $\sqrt{\frac{c^{-1}\sigma_w^2 \ln(2T^4)}{k}}$ where c is the absolute constant from Lemma A.2
$(k_B, \hat{\mu}_B) = (0, 0)$ for all $B \in \mathcal{B}$
for $t = 1, \dots, T$ **do**
  if $\exists B \in \mathcal{B}$ s.t. $x_t \in B$ **then**
    ▷ *x_t is not too OOD: it's safe to learn*
    if $\hat{\mu}_B + \gamma(k_B) + L\sqrt{n}w(T) < 0$ **then**
      ▷ *We already know x_t is bad: don't learn*
      Abstain ($y_t = 0$)
    else
      ▷ *x_t might be good: learn*
      Commit ($y_t = 1$)
      $k_B \leftarrow k_B + 1$
      $\hat{\mu}_B \leftarrow \hat{\mu}_B + \frac{r_t - \hat{\mu}_B}{k_B}$
  else
    ▷ *x_t is far OOD: it's too risky to learn*
    Abstain ($y_t = 0$)

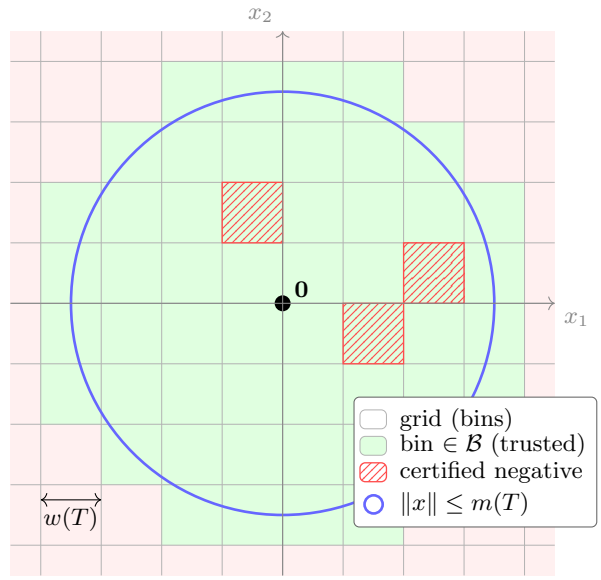

Figure 1: Trusted region of radius $m(T)$ around the origin, partitioned into bins of side $w(T)$. Any square intersecting the ball is shown fully green (bin $\in \mathcal{B}$). Certified negative bins are shown hatched red. The agent abstains outside the ball.

single per-bin mean. The agent always abstains outside the trusted region, and inside it abstains in any bin whose pessimistic upper bound on reward is negative; otherwise it commits to gather information.

More precisely, the algorithm defines a ball of radius $m(T)$ around the origin, treating inputs outside this ball as too OOD to test. The ball is partitioned into n-dimensional hypercubes (bins) of side length $w(T)$. By Lipschitz continuity, the variation of $r(\cdot, 1)$ within any bin B is at most $L\sqrt{n}w(T)$. For each bin B, the algorithm maintains its empirical mean $\hat{\mu}_B$ and a confidence radius $\gamma(k)$ after k commits in B. If $\hat{\mu}_B + \gamma(k) + L\sqrt{n}w(T) < 0$, then B is certified unsafe and the algorithm abstains there permanently. Figure 1 shows a schematic of the algorithm.

We saw in § 4 that the problem is impossible when inputs are too far OOD. A natural way to quantify this is via the amount of probability mass that lies outside a given radius, captured by the *radial survival function*:

Definition 5.1 (Radial survival function)**.** *For any radius $R \ge 0$, the radial survival function of ν is* $\bar{\nu}(R) := \Pr_{x\sim\nu}\left[\|x\| \ge R\right]$.

We are now ready to state our main result.

Theorem 5.2. *In the stochastic setting with $x_t \sim \nu$ i.i.d., Algorithm 1 with $w(T) = T^{-1/(n+2)}$ and $m(T) = \ln T$ satisfies*

$$\mathbb{E}[\mathrm{Reg}(T)] \in O\left((L + \sigma^2)T^{\frac{n+1}{n+2}}(\ln T)^{n+1} + T\bar{\nu}(\ln T)\right).$$

The first term is typical for Lipschitz contextual bandits and reflects the curse of dimensionality (see, e.g., Slivkins et al., 2019, Thms. 4.11–4.12; Plaut et al., 2025a, Thm. 10).

The $T\bar{\nu}(\ln T)$ term is unusual mainly because unbounded domains are unusual: for bounded domains, $\bar{\nu}(\ln T) = 0$ for all large T. Our analysis deals with far OOD inputs directly and the bound necessarily degrades as such inputs become more frequent. This dependence is unavoidable: the construction in Thm. 4.2 sets $x_t = T$ for all $t \in [T]$, hence $\bar{\nu}(\ln T) = 1$ and the bound in Thm. 5.2 becomes linear, matching the impossibility result. By contrast, for any fixed distribution ν, we have $\bar{\nu}(\ln T) \to 0$ as $T \to \infty$, so $T\bar{\nu}(\ln T) = o(T)$ and the overall regret stays sublinear.

For example, if ν is subgaussian with $\bar{\nu}(r) \le e^{-cr^2}$, then $T\bar{\nu}(\ln T) \le Te^{-c(\ln T)^2} = o(1)$. If ν is subexponential with $\bar{\nu}(r) \le e^{-cr}$, then $T\bar{\nu}(\ln T) \le T \cdot T^{-c} = T^{1-c} = o(T)$. If ν has polynomial tails with $\bar{\nu}(r) \asymp r^{-\alpha}$ for $\alpha > 0$, then $T\bar{\nu}(\ln T) \asymp T/(\ln T)^{\alpha} = o(T)$. If one has prior knowledge of ν, the choice of $m(T)$ can be tailored more precisely than our generic setting $m(T) = \ln T$. In particular, for polynomial tails, setting $m(T) = T^c$ for small $c > 0$ improves the bound to $O(T^{1-c\alpha})$.

Thus, the regret decomposes into a geometric/statistical term from discretization and concentration inside the trusted region, and a tail term from far OOD inputs; both are sublinear for any fixed ν.

6 PROOF SKETCH

We now outline the logical structure of the proof of Thm. 5.2; we also provide the intuition behind each step. Full technical details and complete proofs of all lemmas are deferred to Appendix A.

Let $m(T), w(T)$ be as in Algorithm 1, and let $\mathcal{B}$ be the set of bins intersecting the ball of radius $m(T)$. For any $B \in \mathcal{B}$, let $\mu_B = \mathbb{E}_{x\sim\nu}[\, r(x,1) \mid x \in B\,]$ be its true mean commit reward, let $k_B(t)$ be the number of commits taken in B by the end of round t (so $k_B(0) = 0$), and let $\hat{\mu}_B(k)$ be the empirical mean in B after k commits (i.e., the running mean from Algorithm 1 indexed by its commit count).

To control estimation error we define the confidence radius

$$\gamma(k) = \sqrt{\frac{c^{-1}\sigma_w^2 \ln(2T^4)}{k}}, \qquad \sigma_w^2 = nL^2 w(T)^2 + \sigma^2,$$

where $c > 0$ is the absolute constant from Lemma A.2. Here σ_w^2 aggregates the combines observation noise σ^2 with the Lipschitz-induced within-bin variation of $(L\sqrt{n}w(T))^2$.

Definition 6.1 (σ-subgaussian)**.** *A random variable Z is σ-subgaussian if*

$$\mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \;\le\; \exp\!\left(\tfrac{\sigma^2\lambda^2}{2}\right) \quad \textit{for all } \lambda \in \mathbb{R}.$$

Equivalently, $Z - \mathbb{E}[Z]$ has tails that are dominated by a centered Gaussian with variance proxy σ^2.

Define the "good" event under which all per-bin estimates are accurate over the realized commit counts:

$$\mathcal{G} = \Big\{\forall B \in \mathcal{B},\ \forall t \in [T] \text{ s.t. } k_B(t) > 0 : \\ |\hat{\mu}_B(k_B(t)) - \mu_B| \le \gamma(k_B(t))\Big\}.$$

On $\mathcal{G}$, each empirical mean is a reliable proxy for its bin's true mean at every commit count that occurs along the algorithm's trajectory. The analysis conditions on $\mathcal{G}$ (which holds with high probability by a union bound over realized bin–count pairs), and then decomposes regret into: (i) commits inside the trusted region (handled by certification of negative bins plus a margin term), and (ii) abstentions outside the ball of radius $m(T)$ (quantified by the radial survival function).

Lemma 6.2 (Per-bin concentration)**.** *For any bin $B \in \mathcal{B}$ and time step $t \in [T]$ such that $k_B(t) > 0$, $\Pr\left[|\hat{\mu}_B(k_B(t)) - \mu_B| > \gamma(k_B(t))\right] \le T^{-3}$.*

Proof idea. In Algorithm 1, whether we commit depends only on the history (and the bin of the current context). Thus we can use a "reward tape" argument: imagine pre-sampling an infinite list of rewards for each bin and revealing the next one whenever the algorithm commits in that bin. Then the sequence of committed rewards within any fixed bin B is distributionally equivalent to an i.i.d. sequence with $x_i \sim \nu(\cdot \mid x \in B)$ and independent noise.

$$\hat{\mu}_B(k) - \mu_B = \frac{1}{k}\sum_{i=1}^{k}\big(r(x_i,1) - \mu_B\big) + \frac{1}{k}\sum_{i=1}^{k}\eta_i,$$

where the first term captures within-bin reward variation and the second term is σ-subgaussian observation noise. Since all $x_i \in B$, Lem. A.3 gives $|r(x_i,1) - \mu_B| \le L\sqrt{n}w(T)$, so these terms are $(L\sqrt{n}w(T))$-subgaussian; together with the noise, the sum is σ_w-subgaussian. A standard tail bound yields failure probability at most T^{-4} for fixed k. Finally, union bound over $k \in \{1,\dots,T\}$ to handle the random count $k_B(t)$, giving $T \cdot T^{-4} = T^{-3}$. □

Lem. 6.2 gave concentration for each fixed bin and commit count. To extend this guarantee uniformly, we apply a union bound over the at most T bin–time pairs actually realized by the algorithm.

Lemma 6.3 (Uniform concentration bound)**.** *With probability at least $1 - 1/T$, the good event $\mathcal{G}$ holds. Equivalently, $\Pr(\neg\mathcal{G}) \le 1/T$.*

Proof idea. There are at most T^2 relevant pairs (B,t) along the trajectory of the algorithm. Each has failure probability T^{-3} by Lem. 6.2. A union bound yields failure probability at most T^{-1}. □

Recall that the algorithm abstains permanently in any bin once $\hat{\mu}_B(k_B(t)) + \gamma(k_B(t)) + L\sqrt{n}w(T) < 0$. On the good event $\mathcal{G}$, bins with sufficiently negative mean are thus certified unsafe after finitely many commits. (For bins near the decision boundary, certification may not occur, but their per-round regret is $O(L\sqrt{n}w(T))$, so their total contribution is small and accounted for by the margin term later.) We now compute how many commits are needed to certify a negative bin.

Lemma 6.4 (Samples for negative certification)**.** *Consider any $t \in [T]$ and $B \in \mathcal{B}$. On $\mathcal{G}$, if $\mu_B < -L\sqrt{n}w(T)$ and $k_B(t) > \frac{4c^{-1}\sigma_w^2 \ln(2T^4)}{(\mu_B + L\sqrt{n}w(T))^2}$, then bin B is certified negative at time t.*

Proof idea. On the good event $\mathcal{G}$, certification in bin B occurs when $\hat{\mu}_B + \gamma(k_B(t)) + L\sqrt{n}w(T) < 0$. Using the worst-case deviation $\hat{\mu}_B = \mu_B + \gamma(k_B(t))$, this reduces to $\mu_B + 2\gamma(k_B(t)) + L\sqrt{n}w(T) < 0$. Plugging in the definition of $\gamma(k_B(t))$ and solving for k yields the stated number of commits needed for certification. □

Next, we bound the geometry of the trusted region. By construction, $\mathcal{B}$ consists of all bins intersecting

the ball of radius $m(T)$. Consequently, their union $\bigcup_{B\in\mathcal{B}} B$ is contained within a slightly larger ball. This enlarged region will be useful both for bounding how negative rewards can be (via Lipschitz continuity) and for controlling the number of bins (via volume packing).

Let v_1 be the volume of the unit ball $\{x \in \mathcal{X} : \|x\| \le 1\}$. Let $R(T) = m(T) + \sqrt{n}w(T)$, which is the maximum distance from the origin to any point in the trusted region, as shown by the following lemma:

Lemma 6.5 (Trusted region is a slightly larger ball)**.** *Every $x \in \bigcup_{B\in\mathcal{B}} B$ satisfies $\|x\| \le R(T) = m(T) + \sqrt{n}w(T)$.*

We now bound the regret from a truly unsafe bin before it is certified. Let $\Delta_t := \max_{y\in\{0,1\}} r(x_t, y) - r(x_t, y_t)$ be the instantaneous regret at time t.

Lemma 6.6 (Per-bin commit regret)**.** *On $\mathcal{G}$, for any $B \in \mathcal{B}$ with $k_B(T) \ge 1$ and $\mu_B < -(2L\sqrt{n}+1)w(T)$,*

$$\sum_{t:\ x_t\in B,\ y_t=1} \Delta_t \le 2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}.$$

Proof idea. Lem. 6.4 shows that a negative bin is certified after $O(\sigma_w^2/(\mu_B + L\sqrt{n}w(T))^2)$ commits. Each such commit incurs at most $O(|\mu_B|)$ regret, but Lem. 6.5 ensures that $\mu_B \ge -LR(T)$, so the loss per commit is bounded. Multiplying the number of pre-certification commits by the maximum per-step regret yields the stated bound. □

Now that we have controlled the regret contribution of each individual bin, we sum across all bins that are ever visited and include the effect of near-margin bins (those with μ_B close to zero). Such bins may never be certified, but their regret per commit is small, so their total contribution is still controlled.

Lemma 6.7 (Total commit regret inside the trusted region)**.** *On $\mathcal{G}$,*

$$\sum_{t:\ y_t=1} \Delta_t \le \frac{v_1 R(T)^n}{w(T)^n}\left(2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}\right) + (3L\sqrt{n}+1)w(T)T.$$

Proof idea. We partition commits into bins with decisively negative mean and those near the decision boundary. For μ_B well below zero, Lem. 6.6 bounds the regret before certification. Summing over all such bins gives at most $|\mathcal{B}|$ times the per-bin cost, and by the packing bound $|\mathcal{B}|w(T)^n \le v_1 R(T)^n$, this gives the first term. For bins near the margin, the algorithm may continue committing longer, but Lipschitzness bounds the per-round regret by $(3L\sqrt{n}+1)w(T)$, giving the second term after T rounds. □

Lem. 6.7 completes the analysis of commit regret inside the trusted region. Combining this with the abstention regret outside the ball of radius $m(T)$, we yield the final rate in Thm. 5.2 as follows.

Proof idea of Thm. 5.2. Regret decomposes into (i) abstention outside the trusted ball; (ii) commits inside.

For (i), each input with $\|x_t\| > m(T)$ contributes at most 1, giving $T\bar{\nu}(m(T))$, which is sublinear for any fixed ν since $\bar{\nu}(\ln T) \to 0$. For (ii), Lipschitz continuity bounds the within-bin variation, and empirical means stay within confidence radii under the highly likely good event $\mathcal{G}$. Negative bins are certified after $O(\sigma_w^2/\text{margin}^2)$ commits, so each contributes at most $O(LR(T) + \sigma_w^2 \log T/w(T))$ regret. Summing across $O((m(T)/w(T))^n)$ bins gives

$$\tilde{O}\left(R(T)^n\left(LR(T)w(T)^{-n} + \sigma_w^2 w(T)^{-(n+1)}\right)\right),$$

and bins near the decision boundary contribute an additional $O(w(T)T)$. □

If we ignore log factors, the leading terms trade off

$$\underbrace{R(T)^n w(T)^{-(n+1)}}_{\text{variance-driven}} \quad \text{vs.} \quad \underbrace{w(T)T}_{\text{margin-driven}}.$$

Balancing these yields the optimal choice $w(T) \asymp T^{-1/(n+2)}$. Independently, the radius $m(T)$ trades off the abstention term $T\bar{\nu}(m(T))$ against the growth of the volume factor $R(T)^n$. Choosing $m(T) = \ln T$ makes $T\bar{\nu}(m(T))$ sublinear for any fixed ν (since $\bar{\nu}(\ln T) \to 0$) while increasing $R(T)$ only logarithmically.

7 CONCLUSION

In this work, we introduced a formal model for safe learning under distribution shift in contextual bandits with catastrophic tails, provided impossibility results that clarify when sublinear regret is unattainable, and gave a cautious risk-sensitive algorithm with sublinear regret under suitable conditions. Our work has several limitations, which also provide directions for future work. These include relaxing Lipschitz continuity, extending the analysis to non-i.i.d. inputs or worst-case sequences, and incorporating additional structure that could support stronger guarantees or improved rates.

Our regret bound can be close to linear. In Thm. 5.2, the abstention term $T\bar{\nu}(\ln T)$ can dominate for heavy-tailed inputs (e.g., power laws). This is the price of caution: avoiding catastrophic far OOD commits requires systematic abstention in the tails, and the resulting regret can be unavoidable (see the impossibility in Thm. 4.2). Moreover, while the bound is sublinear for every fixed n, the exponent $(n+1)/(n+2) \to 1$

as $n \to \infty$, which is a standard curse of dimensionality in Lipschitz contextual bandits (Slivkins et al., 2019, Thms. 4.11–4.12); see also Plaut et al. (2025a, Thm. 10). While we do not expect to remove these dependencies entirely, future work could improve rates. The simplicity of Algorithm 1 is appealing but ignores useful information: commits inform not only their own bin but also nearby bins via Lipschitz continuity, and certifying a bin as positive could justify expanding the trusted region around it. Additional structural assumptions, such as margin/low-noise conditions, intrinsic low dimensionality, or smoothness beyond Lipschitz, could also help.

Assumptions may not always hold. Our guarantees here rely on i.i.d. inputs and Lipschitz continuity of the commit reward. In practice, inputs may drift or exhibit temporal dependence, and rewards may be only piecewise smooth or even non-smooth. Extending the analysis to weaker smoothness conditions or drifting processes is an important direction. Moreover, Algorithm 1 assumes knowledge of L, σ^2, and T. While knowledge of T can be handled by the standard doubling trick (see Slivkins et al. (2019, §1.5)), L and σ^2 may be unknown. Thus, developing parameter-free (or adaptively tuned) algorithms that remain cautious would increase robustness

No unconditionally irreparable errors. Obtaining regret $-T$ on a single time step is irreparable in the sense that it automatically implies linear regret on that run. However, errors in our model are only irreparable for a fixed T: for any error, there exists a large enough T that the error is no longer catastrophic. It may be worth considering alternative models of catastrophe such as inescapable trap states in MDPs which do allow for errors that are unconditionally catastrophic.

Broader impact. This work is motivated by safety concerns in the deployment of learning systems in high-stakes domains. We provide theoretical justification for abstention as a mechanism for averting catastrophic errors under distribution shift, and abstention is also a practical choice for deployed systems. Agents that can defer action when uncertain may be safer and more trustworthy, but abstention mechanisms must be designed carefully to avoid consequences such as excessive conservatism or over-reliance on human supervision.

Acknowledgments

This work was supported by a gift from Open Philanthropy to the Center for Human-Compatible AI (CHAI) at UC Berkeley. We thank Vamshi Bonagiri and Pavel Czempin for their insightful discussions and valuable feedback.

S. Liaw and B. Plaut contributed equally as co-first authors to this work. S. Liaw contributed to the early development of the project, including exploring several initial ideas and leading the development of the proof structure and intermediate arguments. The project was conceived and supervised by B. Plaut, who developed the final direction of the work and completed the final proof. S. Liaw wrote the first draft of the manuscript, with B. Plaut providing detailed feedback.

References

Agrawal, S., Juneja, S. K., and Koolen, W. M. (2021). Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pages 26–62. PMLR.

Altman, E. (1999). *Constrained Markov Decision Processes*, volume 7. CRC Press.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272. PMLR.

Bernasconi, M., Castiglioni, M., Celli, A., and Fusco, F. (2024). Beyond primal-dual methods in bandits with stochastic and adversarial constraints.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Chen, T., Gangrade, A., and Saligrama, V. (2022). Strategies for safe multi-armed bandits with logarithmic regret and risk.

Cohen, M. K., Catt, E., and Hutter, M. (2021). Curiosity Killed or Incapacitated the Cat and the Asymptotically Optimal Agent. *IEEE Journal on Selected Areas in Information Theory*, 2(2):665–677. Conference Name: IEEE Journal on Selected Areas in Information Theory.

Cohen, M. K. and Hutter, M. (2020). Pessimism About Unknown Unknowns Inspires Conservatism. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 1344–1373. PMLR.

García, J. and Fernández, F. (2015). A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research*, 16(42):1437–1480.

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., and Knoll, A. (2024). A review of safe reinforcement learning: Methods, theories, and applications. 46(12):11216–11235. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(51):1563–1600.

Kazerouni, A., Ghavamzadeh, M., Abbasi-Yadkori, Y., and Roy, B. V. (2017). Conservative contextual linear bandits.

Kosoy, V. (2019). Delegative Reinforcement Learning: learning to avoid traps with a little help. arXiv. arXiv:1907.08461 [cs, stat].

Krasowski, H., Thumm, J., Müller, M., Schäfer, L., Wang, X., and Althoff, M. (2023). Provably safe reinforcement learning: Conceptual analysis, survey, and benchmarking.

Lin, J., Lee, X. Y., Jubery, T., Moothedath, S., Sarkar, S., and Ganapathysubramanian, B. (2022). Stochastic conservative contextual linear bandits.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. (2021). Learning policies with zero or bounded constraint violation for constrained MDPs. *Advances in Neural Information Processing Systems*, 34:17183–17193.

Maillard, O.-A., Mann, T., Ortner, R., and Mannor, S. (2019). Active Roll-outs in MDP with Irreversible Dynamics.

Moldovan, T. M. and Abbeel, P. (2012a). Safe exploration in Markov decision processes. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, pages 1451–1458, Madison, WI, USA. Omnipress.

Moldovan, T. M. and Abbeel, P. (2012b). Safe exploration in markov decision processes.

Neu, G. and Zhivotovskiy, N. (2020). Fast rates for online prediction with abstention.

Perkins, T. J. and Barto, A. G. (2003). Lyapunov design for safe reinforcement learning. *J. Mach. Learn. Res.*, 3(null):803–832.

Plaut, B., Liévano-Karim, J., Zhu, H., and Russell, S. (2025a). Safe learning under irreversible dynamics via asking for help.

Plaut, B., Zhu, H., and Russell, S. (2025b). Avoiding catastrophe in online learning by asking for help. In *Proceedings of the 42nd International Conference on Machine Learning*.

Saunders, W., Sastry, G., Stuhlmueller, A., and Evans, O. (2017). Trial without error: Towards safe reinforcement learning via human intervention.

Slivkins, A. (2011). Contextual Bandits with Similarity Information. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 679–702. ISSN: 1938-7228.

Slivkins, A. et al. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286.

Stradi, F. E., Castiglioni, M., Marchesi, A., and Gatti, N. (2024). Learning adversarial MDPs with stochastic hard constraints. *arXiv preprint arXiv:2403.03672*.

Sutton, R. S., Barto, A. G., et al. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Tan, V. Y. F., L.A., P., and Jagannathan, K. (2022). A survey of risk-aware multi-armed bandits. In Raedt, L. D., editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5623–5629. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Wachi, A., Hashimoto, W., Shen, X., and Hashimoto, K. (2023). Safe exploration in reinforcement learning: A generalized formulation and algorithms.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. (2016). Conservative bandits. In *International Conference on Machine Learning (ICML)*.

Yang, J., Jin, T., and Tan, V. Y. F. (2024). Multi-armed bandits with abstention.

Zhao, W., He, T., Chen, R., Wei, T., and Liu, C. (2023). State-wise Safe Reinforcement Learning: A Survey. volume 6, pages 6814–6822. ISSN: 1045-0823.

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**, § 3 and § 5.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**, § 5 and Appendix A.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Not Applicable**, no code was used for this work.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**, § 3 and § 5.
   (b) Complete proofs of all theoretical results. **Yes**, Appendix A.
   (c) Clear explanations of any assumptions. **Yes**, § 3.

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**, Appendix B.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**, Appendix B.
   (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**, Appendix B.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**, Appendix B.

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**, no use of existing assets or release of new assets.
   (b) The license information of the assets, if applicable. **Not Applicable**, no use of existing assets or release of new assets.
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable**, no use of existing assets or release of new assets.
   (d) Information about consent from data providers/curators. **Not Applicable**, no use of existing assets or release of new assets.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**, no use of existing assets or release of new assets.

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**, no research with human subjects.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**, no research with human subjects.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**, no research with human subjects.

A PROOF OF MAIN RESULT

A.1 Proof notation

The proof will use the following notation:

1. The true mean of bin B is $\mu_B = \mathbb{E}_{x\sim\nu}[r(x,1) \mid x \in B]$.
2. Let $k_B(t)$ denote the value of the variable k_B in Algorithm 1 at the end of time step t.
3. Let $\hat{\mu}_B(k)$ denote the value of the variable $\hat{\mu}_B$ in Algorithm 1 after k rewards from bin B have been observed.
4. Let $\sigma_w = \sqrt{nL^2 w(T)^2 + \sigma^2}$ for brevity.
5. Define the confidence radius $\gamma(k) = \sqrt{\frac{c^{-1}\sigma_w^2 \ln(2T^4)}{k}}$ where c is the absolute constant from Lemma A.2.
6. Define the good event $\mathcal{G} = \{\forall t \in [T], \forall B \in \mathcal{B} \text{ where } k_B(t) > 0 : |\hat{\mu}_B(k_B(t)) - \mu_B| \le \gamma(k_B(t))\}$.
7. A bin B is certified negative at time t if $\hat{\mu}_B(k_B(t)) + \gamma(k_B(t)) + L\sqrt{n}w(T) < 0$.
8. Let $\bar{\nu}$ be the radial survival function of ν. That is, for any $y \in \mathbb{R}_{\ge 0}$, $\bar{\nu}(y) = \Pr_{x\sim\nu}[\|x\| \ge y]$.
9. Let $\Delta_t = \max_{y^*\in\{0,1\}} r(x_t, y^*) - r(x_t, y_t)$ be the single-step regret at time t.
10. Let v_1 be the volume of the unit ball $\{x \in \mathcal{X} : \|x\| \le 1\}$.
11. Let $R(T) = m(T) + \sqrt{n}w(T)$. This will be the maximum distance of any input in $\cup_{B\in\mathcal{B}} B$ from the origin.

Lemma A.1 (Hoeffding's Lemma, Lemma 2.2 in Boucheron et al., 2013)**.** *If Z is a random variable taking values in the bounded interval $[a,b]$, then Z is $(\frac{b-a}{2})$-subgaussian.*

Lemma A.2 (Hoeffding's Inequality, subgaussian version)**.** *Let $X_1, \ldots, X_k$ be independent random variables with mean zero, where each X_i is σ_i-subgaussian for some $\sigma_i > 0$. Then there exists an absolute constant $c > 0$ such that for any $\varepsilon > 0$,*

$$\Pr\left[\left|\sum_{i=1}^{k} X_i\right| > \varepsilon\right] \le 2\exp\left(-\frac{c\varepsilon^2}{\sum_{i=1}^{k}\sigma_i^2}\right)$$

Lemma A.3. *If $x \in B \in \mathcal{B}$, then $|r(x,1) - \mu_B| \le L\sqrt{n}w(T)$.*

Proof. Let $r^- = \inf_{x'\in B} r(x',1)$ and $r^+ = \sup_{x'\in B} r(x',1)$. Then $r^- \le \mu_B \le r^+$ and $r^- \le r(x,1) \le r^+$. Next, for any $\varepsilon > 0$, there exists $x^-, x^+ \in B$ such that $r(x^-,1) - \varepsilon < r^-$ and $r(x^+,1) + \varepsilon > r^+$ (if not, this contradicts r^- and r^+ being the infimum and supremum). Then $r(x,1)$ and μ_B belong to the interval $[r^-, r^+]$, which is a subset of the interval $[r(x^-,1) - \varepsilon, r(x^+,1) + \varepsilon]$.

The maximum distance between two points in an n-cube with side length $w(T)$ is equal to the length of the diagonal of this n-cube, which is $\sqrt{n}w(T)$. Since x^- and x^+ belong to the same n-cube with side length $w(T)$, we have $\|x^- - x^+\| \le \sqrt{n}w(T)$. Then by Lipschitz continuity, $|r(x^+,1) - r(x^-,1)| = r(x^+,1) - r(x^-,1) \le L\sqrt{n}w(T)$. Therefore $r^+ - r^- \le L\sqrt{n}w(T) + 2\varepsilon$. Since $r^- \le r(x,1) \le r^+$ and $r^- \le \mu_B \le r^+$, we have $|r(x,1) - \mu_B| \le L\sqrt{n}w(T) + 2\varepsilon$. Since this holds for all $\varepsilon > 0$, we must have $|r(x,1) - \mu_B| \le L\sqrt{n}w(T)$. □

Lemma 6.2 (Per-bin concentration)**.** *For any bin $B \in \mathcal{B}$ and time step $t \in [T]$ such that $k_B(t) > 0$, $\Pr[|\hat{\mu}_B(k_B(t)) - \mu_B| > \gamma(k_B(t))] \le T^{-3}$.*

Proof. Fix any bin $B \in \mathcal{B}$.

Part 1: The reward tape construction. We will prove equivalence between the learning process and alternative stochastic process where independence is easier to analyze. The idea is to sample upfront a list of all of the rewards we will ever need for bin B. Then, whenever $x_t \in B$ and $y_t = 1$ occurs, the agent simply observes the next unused reward from our list. This is essentially the "reward tape" argument used in Section 1.3.1 of Slivkins et al. (2019).

To formalize this, let K be the maximum possible (over all realizations of randomness) number of time steps the agent commits on an input in B. Before $t = 1$, sample $\eta_B^1, \ldots, \eta_B^K$ i.i.d. from the same distribution as each η_t and sample $x_B^1, \ldots, x_B^K$ i.i.d. from the distribution of ν restricted to B, i.e., $x_B^1, \ldots, x_B^K \sim \nu(x \mid x \in B)$. For each $k \in [K]$, let $q_B^k = r(x_B^k, 1) + \eta_B^k$. For each $t \in [T]$, if $x_t \in B$ and $y_t = 1$, the agent observes $q_B^{k_B(t)}$ instead of the standard reward $r_t = r(x_t, 1) + \eta_t$. If $y_t = 0$, the agent observes reward 0 as usual.

We claim this is equivalent to the normal learning protocol in the following sense. For $k \in [K]$, let r_B^k be the random variable of the kth reward observed from committing on an input in bin B in the normal learning protocol. For brevity, for any $k \in [K]$, let $r_B^{1:k}$ and $q_B^{1:k}$ denote $r_B^1, \ldots, r_B^k$ and $q_B^1, \ldots, q_B^k$ respectively. Let $\mathcal{D}(X)$ denote the distribution of a random variable X. We claim that $\mathcal{D}(r_B^{1:k}) = \mathcal{D}(q_B^{1:k})$. We proceed by induction from $k = 0$ to K (for completeness, define $r_B^0 = q_B^0$ as any constant). The base case at $k = 0$ is trivial, so suppose that $\mathcal{D}(r_B^{1:k-1}) = \mathcal{D}(q_B^{1:k-1})$ for some $i \in [k]$. Let $t_k = \min\{t \in [T] : k_B(t) = k\}$ be the time step on which the kth reward in bin B is observed. Then by definition we have $\mathcal{D}(r_B^k \mid r_B^{1:k-1}) = \mathcal{D}(r(x_t, 1) + \eta_t \mid t = t_k, r_B^{1:k-1})$ for any $t \in [T]$.

We claim that conditioned on $x_t \in B$, $r(x_t, 1)+\eta_t$ is independent of $t = t_k$ and $r_B^{1:k-1}$. Note that the event $\{t = t_k\}$ is equivalent to $\{x_t \in B$ and $y_t = 1$ and $k_B(t-1) = k-1\}$. The key property is that the algorithm's behavior never distinguishes between inputs in the same bin. Formally, conditioned on $x_t \in B$, y_t and x_t are independent. This remains true when also conditioning on the events of prior time steps (in particular, $k_B(t-1)$ and $r_B^{1:k-1}$). Hence

$$\begin{aligned}
\mathcal{D}(r_B^k \mid r_B^{1:k-1}) &= \mathcal{D}(r(x_t, 1) + \eta_t \mid x_t \in B, y_t = 1, r_B^{1:k-1}, k_B(t-1) = k-1) \\
&= \mathcal{D}(r(x_t, 1) + \eta_t \mid x_t \in B, r_B^{1:k-1}, k_B(t-1) = k-1) \\
&= \mathcal{D}(r(x_t, 1) + \eta_t \mid x_t \in B)
\end{aligned}$$

with the last time step because x_t and η_t are independent of the events of prior time steps.

By construction, $\mathcal{D}(x_B^k) = \mathcal{D}(x_t \mid x_t \in B)$ and $\mathcal{D}(\eta_B^k) = \mathcal{D}(\eta_t)$. Since x_B^k and η_B^k are independent (including when conditioned on $x_t \in B$), we have $\mathcal{D}(x_B^k, \eta_B^k) = \mathcal{D}(x_t, \eta_t \mid x_t \in B)$. Applying the transformation $x, \eta \mapsto r(x, 1) + \eta$, we get $\mathcal{D}(r(x_B^k, 1) + \eta_B^k) = \mathcal{D}(r(x_t, 1) + \eta_t \mid x_t \in B)$. We showed above that $\mathcal{D}(r_B^k \mid r_B^{1:k-1}) = \mathcal{D}(r(x_t, 1) + \eta_t \mid x_t \in B)$ and we have $\mathcal{D}(q_B^k) = \mathcal{D}(r(x_B^k, 1) + \eta_B^k)$ by definition. Thus $\mathcal{D}(q_B^k) = \mathcal{D}(r_B^k \mid r_B^{1:k-1})$.

Since $x_B^{1:k}, \eta_B^{1:k}$ are all independent, q_B^k is independent of $q_B^{1:k-1}$. Thus $\mathcal{D}(q_B^k \mid q_B^{1:k-1}) = \mathcal{D}(q_B^k) = \mathcal{D}(r_B^k \mid r_B^{1:k-1})$. Combining with this the inductive hypothesis $\mathcal{D}(q_B^{1:k-1}) = \mathcal{D}(r_B^{1:k-1})$ gives us $\mathcal{D}(q_B^{1:k}) = \mathcal{D}(r_B^{1:k})$, as needed. This completes the induction and shows that $\mathcal{D}(q_B^{1:k}) = \mathcal{D}(r_B^{1:k})$ for any $k \in [K]$.

Recall that $\hat{\mu}_B(k)$ is the mean of the first k rewards observed from committing on an input in B: $\hat{\mu}_B(k) = \frac{1}{k}\sum_{i=1}^k r_B^i$. Therefore for all $k \in [K]$, $\mathcal{D}(k\hat{\mu}_B(k)) = \mathcal{D}(\sum_{i=1}^k q_B^i)$.

Part 2: Applying Lemma A.2. Now fix some $k \in [K]$ and $B \in \mathcal{B}$. For each $i \in [k]$, let $Z_i = r(x_B^i, 1) - \mu_B$. By construction $\mathbb{E}[r(x_B^i, 1)] = \mathbb{E}_{x\sim\nu}[r(x, 1) \mid x \in B] = \mu_B$, so $\mathbb{E}[Z_i] = 0$. Since $x_B^i \in B \in \mathcal{B}$, Lemma A.3 implies that $|r(x_B^i, 1) - \mu_B| \le L\sqrt{n}w(T)$. Thus Z_i belongs to an interval of length $2L\sqrt{n}w(T)$, so Z_i is $(L\sqrt{n}w(T))$-subgaussian. Also, η_B^i is σ-subgaussian with mean 0, since η_B^i has the same distribution as any η_t. Furthermore, our construction ensures that $Z_1, \ldots, Z_k, \eta_B^1, \ldots, \eta_B^k$ are all independent. Therefore by Lemma A.2, for any $\varepsilon > 0$,

$$\begin{aligned}
\Pr\left[\left|\sum_{i=1}^k (r(x_B^i, 1) - \mu_B + \eta_B^i)\right| > \varepsilon\right] &= \Pr\left[\left|\sum_{i=1}^k Z_i + \sum_{i=1}^k \eta_B^i\right| > \varepsilon\right] \\
&\le 2\exp\left(-\frac{c\varepsilon^2}{\sum_{i=1}^k (L\sqrt{n}w(T))^2 + \sum_{i=1}^k \sigma^2}\right) \\
&= 2\exp\left(-\frac{c\varepsilon^2}{k\sigma_w^2}\right)
\end{aligned}$$

Part 3: Converting the inequality to $\hat{\mu}_B(k_B(t))$. By Part 1, $\mathcal{D}(k\hat{\mu}_B(k)) = \mathcal{D}(\sum_{i=1}^k q_B^k)$ for any $k \in [K]$. Since μ_B is a constant, we have

$$\mathcal{D}(k\hat{\mu}_B(k) - k\mu_B) = \mathcal{D}\left(\sum_{i=1}^k q_B^i - k\mu_B\right) = \mathcal{D}\left(\sum_{i=1}^k (r(x_B^i, 1) - \mu_B + \eta_B^i)\right)$$

Therefore

$$\Pr\left[|k\hat{\mu}_B(k) - k\mu_B| > \varepsilon\right] = \Pr\left[\left|\sum_{i=1}^k (r(x_B^i, 1) - \mu_B + \eta_B^i)\right| > \varepsilon\right] \le 2\exp\left(-\frac{c\varepsilon^2}{k\sigma_w^2}\right)$$

Now set $\varepsilon = k\gamma(k) = \sqrt{kc^{-1}\sigma_w^2 \ln(2T^4)}$ to get

$$\Pr\left[|\hat{\mu}_B(k) - \mu_B| > \gamma(k)\right] = \Pr\left[k|\hat{\mu}_B(k) - \mu_B| > \sqrt{kc^{-1}\sigma_w^2 \ln(2T^4)}\right]$$

$$
\begin{aligned}
&= \Pr\left[\left|\sum_{i=1}^{k} Z_i + \sum_{i=1}^{k} \eta_B^i\right| > \sqrt{kc^{-1}\sigma_w^2 \ln(2T^4)}\right] \\
&\leq 2\exp(-\ln(2T^4)) \\
&= 2\exp\left(\ln\left(\frac{1}{2T^4}\right)\right) \\
&= T^{-4}
\end{aligned}
$$

The last step is to convert $\hat{\mu}_B(k)$ to $\hat{\mu}_B(k_B(t))$. Suppose $|\hat{\mu}_B(k_B(t)) - \mu_B| > \gamma(k_B(t))$ for some $B \in \mathcal{B}, t \in [T]$ with $k_B(t) > 0$. Then $\exists k \in [T]$ such that $k_B(t) = k$ and $|\hat{\mu}_B(k) - \mu_B| > \gamma(k)$. Thus by the union bound,

$$
\begin{aligned}
\Pr\left[|\hat{\mu}_B(k_B(t)) - \mu_B| > \gamma(k_B(t))\right] &= \Pr\left[\exists k \in [T] : k_B(t) = k \text{ and } |\hat{\mu}_B(k) - \mu_B| > \gamma(k)\right] \\
&\leq \sum_{k=1}^{T} \Pr\left[k_B(t) = k \text{ and } |\hat{\mu}_B(k) - \mu_B| > \gamma(k)\right] \\
&\leq \sum_{k=1}^{T} \Pr\left[|\hat{\mu}_B(k) - \mu_B| > \gamma(k)\right] \\
&\leq \sum_{k=1}^{T} T^{-4} \\
&= T^{-3}
\end{aligned}
$$

□

Lemma 6.3 (Uniform concentration bound)**.** *With probability at least $1-1/T$, the good event $\mathcal{G}$ holds. Equivalently, $\Pr(\neg\mathcal{G}) \leq 1/T$.*

Proof. Let J be the the number of bins that ever receive at least one commit, i.e., $J = |\{B \in \mathcal{B} : \exists t \in [T] \text{ s.t. } x_t \in B, y_t = 1\}|$. For each $j \in [J]$, let B_j be the jth bin to receive a commit.

$$
\begin{aligned}
\Pr[\neg\mathcal{G}] &= \mathbb{E}[\Pr[\neg\mathcal{G} \mid J, B_1, \ldots, B_J]] && \text{(Law of total expectation)} \\
&= \mathbb{E}\left[\Pr\left[\bigcup_{j=1}^{J} \bigcup_{t:k_{B_j}(t)>0} \{|\hat{\mu}_{B_j}(k_{B_j}(t)) - \mu_{B_j}| > \gamma(k_{B_j}(t))\}\right] \;\middle|\; J, B_1, \ldots, B_J\right] && \text{(Direct negation)} \\
&\leq \mathbb{E}\left[\sum_{j=1}^{J} \sum_{t:k_{B_j}(t)>0} \Pr[|\hat{\mu}_{B_j}(k_{B_j}(t)) - \mu_{B_j}| > \gamma(k_{B_j}(t))] \;\middle|\; J, B_1, \ldots, B_J\right] && \text{(Union bound)} \\
&\leq \mathbb{E}\left[\sum_{j=1}^{J} \sum_{t:k_{B_j}(t)>0} T^{-3} \;\middle|\; J, B_1, \ldots, B_J\right] && \text{(Lemma 6.2)} \\
&\leq \mathbb{E}\left[T^{-1} \;\middle|\; J, B_1, \ldots, B_J\right] && (J \in [T]) \\
&\leq T^{-1} && \text{(Expectation of a constant)}
\end{aligned}
$$

as required. □

Lemma 6.4 (Samples for negative certification)**.** *Consider any $t \in [T]$ and $B \in \mathcal{B}$. On $\mathcal{G}$, if $\mu_B < -L\sqrt{n}w(T)$ and $k_B(t) > \frac{4c^{-1}\sigma_w^2 \ln(2T^4)}{(\mu_B + L\sqrt{n}w(T))^2}$, then bin B is certified negative at time t.*

Proof. Note that $\mu_B < -L\sqrt{n}w(T)$ implies that the denominator is well-defined. By assumption on $k_B(t)$,

$$
\gamma(k_B(t)) = \sqrt{\frac{c^{-1}\sigma_w^2 \ln(2T^4)}{k_B(t)}}
$$

$$
\begin{aligned}
&< \sqrt{\frac{(\mu_B + L\sqrt{n}w(T))^2}{4}} \\
&= \frac{|\mu_B + L\sqrt{n}w(T)|}{2} \\
&= -\frac{\mu_B + L\sqrt{n}w(T)}{2}
\end{aligned}
$$

By definition of $\mathcal{G}$, we have $-\gamma(k_B(t)) \leq \hat{\mu}_B(k_B(t)) - \mu_B \leq \gamma(k_B(t))$, so

$$
\begin{aligned}
\hat{\mu}_B(k_B(t)) + \gamma(k_B(t)) + L\sqrt{n}w(T) &\leq \mu_B + 2\gamma(k_B(t)) + L\sqrt{n}w(T) \\
&< \mu_B - (\mu_B + L\sqrt{n}w(T) + L\sqrt{n}w(T) \\
&= 0
\end{aligned}
$$

so B is certified negative at time t. □

Lemma 6.5 (Trusted region is a slightly larger ball)**.** *Every $x \in \bigcup_{B \in \mathcal{B}} B$ satisfies $\|x\| \leq R(T) = m(T) + \sqrt{n}w(T)$.*

Proof. If $x \in B$ for some $B \in \mathcal{B}$, there must exist $x' \in B$ such that $\|x'\| \leq m(T)$. The maximum distance between any pair of points in an n-cube with side length $w(T)$ is $\sqrt{n}w(T)$. Thus by the triangle inequality, x satisfies

$$\|x\| \leq \|x'\| + \|x - x'\| \leq m(T) + \sqrt{n}w(T) = R(T)$$

as required. □

Lemma 6.6 (Per-bin commit regret)**.** *On $\mathcal{G}$, for any $B \in \mathcal{B}$ with $k_B(T) \geq 1$ and $\mu_B < -(2L\sqrt{n} + 1)w(T)$,*

$$\sum_{t:\ x_t \in B,\ y_t = 1} \Delta_t \leq 2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}.$$

Proof. Since $\mu_B < -(2\sqrt{n}L + 1)w(T) < -L\sqrt{n}w(T)$ and $\mathcal{G}$ holds, Lemma 6.4 implies that B is certified negative on the first time step t such that $k_B(t) > \frac{4c^{-1}\sigma_w^2 \ln(2T^4)}{(\mu_B + L\sqrt{n}w(T))^2}$. Once B is certified negative, we never again commit in B, so the number of commits in B is $|\{t \in [T] : x_t \in B, y_t = 1\}| \leq \lceil \frac{4c^{-1}\sigma_w^2 \ln(2T^4)}{(\mu_B + L\sqrt{n}w(T))^2} \rceil \leq 1 + \frac{4c^{-1}\sigma_w^2 \ln(2T^4)}{(|\mu_B| - L\sqrt{n}w(T))^2}$. Since $|\mu_B| \geq (2L\sqrt{n} + 1)w(T) \geq 2L\sqrt{n}w(T)$, we have

$$|\mu_B| = \frac{|\mu_B|}{2} + \frac{|\mu_B|}{2} \geq \frac{|\mu_B|}{2} + L\sqrt{n}w(T)$$

so $|\mu_B| - L\sqrt{n}w(T) \geq |\mu_B/2|$. Therefore $(|\mu_B| - L\sqrt{n}w(T))^2 \geq \mu_B^2/4$, so $|\{t \in [T] : x_t \in B, y_t = 1\}| \leq 1 + \frac{16c^{-1}\sigma_w^2 \ln(2T^4)}{\mu_B^2}$.

For any $t \in [T]$ such that $y_t = 1$, either $y_t = 1$ is optimal or $y_t = 0$ is optimal. In the former case, the single-step regret Δ_t is 0. In the latter case, $\Delta_t = -r(x_t, 1)$. If $x_t \in B$, Lemma A.3 implies that $r(x_t, 1) \geq \mu_B - L\sqrt{n}w(T)$. Since $\mu_B < -L\sqrt{n}w(T)$, we have $r(x_t, 1) \geq 2\mu_B$. Hence

$$
\begin{aligned}
\sum_{t: x_t \in B, y_t = 1} \Delta_t &\leq \sum_{t: x_t \in B, y_t = 1} (-2\mu_B) \\
&= |\{t \in [T] : x_t \in B, y_t = 1\}| \cdot 2|\mu_B| \\
&\leq \left(1 + \frac{16c^{-1}\sigma_w^2 \ln(2T^4)}{\mu_B^2}\right) \cdot 2|\mu_B| \\
&= 2|\mu_B| + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{|\mu_B|} \\
&\leq 2|\mu_B| + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{(2\sqrt{n}L + 1)w(T)} \\
&\leq 2|\mu_B| + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}
\end{aligned}
$$

with the last step due to $2\sqrt{n}L + 1 \geq 1$. By Lemma 6.5, any $x \in B$ satisfies $\|x\| \leq R(T)$. Thus by Lipschitz continuity, $r(x, 1) \geq r(0, 1) - LR(T) > -LR(T)$ for all $x \in B$. Thus$\mu_B = \mathbb{E}_{x\sim\nu}[r(x, 1) \mid x \in B] \geq \mathbb{E}_{x\sim\nu}[-LR(T)] = -LR(T)$, so

$$\sum_{t:x_t\in B, y_t=1} \Delta_t \leq 2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}$$

as required. □

Lemma 6.7 (Total commit regret inside the trusted region)**.** *On $\mathcal{G}$,*

$$\sum_{t:\ y_t=1} \Delta_t \leq \frac{v_1 R(T)^n}{w(T)^n}\left(2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}\right) + (3L\sqrt{n} + 1)w(T)T.$$

Proof. For each $t \in [T]$, let $B(t)$ denote the bin to which x_t belongs. Note that $B(t)$ is unique by the construction of $\mathcal{H}$ in Algorithm 1 and note that $B(t)$ may not be in $\mathcal{B}$. Partition the time steps with commits into $S_1 = \{t \in [T] : y_t = 1 \text{ and } \mu_{B(t)} < -(2L\sqrt{n}+1)w(T)\}$ and $S_2 = \{t \in [T] : y_t = 1 \text{ and } \mu_{B(t)} \geq -(2L\sqrt{n}+1)w(T)\}$. Let $\mathcal{B}_1 = \{B \in \mathcal{B} : \exists t \in S_1 \text{ s.t. } B(t) = B\}$ be the set of bins associated with time steps in S_1. Then we can write

$$\begin{aligned}\sum_{t\in S_1} \Delta_t &= \sum_{B\in\mathcal{B}_1} \sum_{t\in S_1:B(t)=B} \Delta_t \\ &= \sum_{B\in\mathcal{B}_1} \sum_{t:x_t\in B, y_t=1} \Delta_t\end{aligned}$$

Then by Lemma 6.6,

$$\begin{aligned}\sum_{t\in S_1} \Delta_t &\leq \sum_{B\in\mathcal{B}_1}\left(2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}\right) \\ &\leq |\mathcal{B}_1|\left(2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}\right) \\ &\leq |\mathcal{B}|\left(2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}\right)\end{aligned}$$

By Lemma 6.5, every $x \in \cup_{B\in\mathcal{B}} B$ satisfies $\|x\| \leq R(T)$. Thus $\cup_{B\in\mathcal{B}} B$ is fully contained within an n-ball of radius r. The volume of such a ball is $v_1 R(T)^n$. Each bin in $\mathcal{B}$ is an n-cube with side length $w(T)$ and thus has volume $w(T)^n$. Furthermore, the bins in B have no volume overlap by construction, so the total volume of bins in B is $w(T)^n|\mathcal{B}|$. Then $w(T)^n|\mathcal{B}| \leq v_1 R(T)^n$. Therefore

$$\sum_{t\in S_1} \Delta_t \leq \frac{v_1 R(T)^n}{w(T)^n}\left(2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}\right)$$

Now consider any $t \in S_2$. By definition, $x_t \in B(t) \in \mathcal{B}$, so Lemma A.3 implies that $r(x_t, 1) \geq \mu_{B(t)} - L\sqrt{n}w(T)$. Since $\mu_{B(t)} \geq -(2L\sqrt{n} + 1)w(T)$ by construction of S_2, we have $r(x_t, 1) \geq -(3L\sqrt{n} + 1)w(T)$. Therefore

$$\begin{aligned}\sum_{t\in S_2}\left(\max_{y^*\in\{0,1\}} r(x_t, y^*) - r(x_t, 1)\right) &\leq \sum_{t\in S_2}(3L\sqrt{n} + 1)w(T) \\ &= |S_2|(3L\sqrt{n} + 1)w(T) \\ &\leq (3L\sqrt{n} + 1)w(T)T\end{aligned}$$

Putting it all together,

$$\begin{aligned}\sum_{t:y_t=1} \Delta_t &= \sum_{t\in S_1}\left(\max_{y^*\in\{0,1\}} r(x_t, y^*) - r(x_t, 1)\right) + \sum_{t\in S_2}\left(\max_{y^*\in\{0,1\}} r(x_t, y^*) - r(x_t, 1)\right) \\ &\leq \frac{v_1 R(T)^n}{w(T)^n}\left(2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}\right) + (3L\sqrt{n} + 1)w(T)T\end{aligned}$$

as required. □

Theorem 5.2. *In the stochastic setting with $x_t \sim \nu$ i.i.d., Algorithm 1 with $w(T) = T^{-1/(n+2)}$ and $m(T) = \ln T$ satisfies*

$$\mathbb{E}[\mathrm{Reg}(T)] \in O\left((L+\sigma^2)T^{\frac{n+1}{n+2}}(\ln T)^{n+1} + T\bar{\nu}(\ln T)\right).$$

Proof. First assume $\mathcal{G}$ holds. Let $S_3 = \{t \in [T] : y_t = 1 \text{ and } r(x_t, 1) < r(x_t, 0)\}$ be the time steps where we committed but we should have abstained, and let $S_4 = \{t \in [T] : y_t = 0 \text{ and } r(x_t, 0) < r(x_t, 1)\}$ be the time steps where we should have committed but we abstained. Lemma 6.7 bounds the regret of time steps in S_3. Since we always commit whenever $x_t \in B$ for some $B \in \mathcal{B}$, S_4 can only occur when $x_t \notin B$ for all $B \in \mathcal{B}$. By construction, any such x_t satisfies $\|x_t\| > m(T)$ (otherwise the bin containing x_t would be in $\mathcal{B}$). Also, $r(x_t, 1) - r(x_t, 0) \leq 1$ by assumption. Hence

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T} \Delta_t \\
&= \sum_{t \in S_3} \Delta_t + \sum_{t \in S_4} \Delta_t \\
&\leq \frac{v_1 R(T)^n}{w(T)^n}\left(2LR(T) + \frac{32c^{-1}\sigma_w^2 \ln(2T^4)}{w(T)}\right) + (3L\sqrt{n}+1)w(T)T + \sum_{t=1}^{T} \mathbf{1}(\|x_t\| > m(T)) \\
&= \frac{2Lv_1 R(T)^{n+1}}{w(T)^n} + \frac{32v_1 c^{-1}\sigma_w^2 R(T)^n \ln(2T^4)}{w(T)^{n+1}} + (3L\sqrt{n}+1)w(T)T + \sum_{t=1}^{T} \mathbf{1}(\|x_t\| > m(T))
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T) \mid \mathcal{G}] &\leq \frac{2Lv_1 R(T)^{n+1}}{w(T)^n} + \frac{32v_1 c^{-1}\sigma_w^2 R(T)^n \ln(2T^4)}{w(T)^{n+1}} + (3L\sqrt{n}+1)w(T)T + \sum_{t=1}^{T} \Pr[\|x_t\| > m(T)] \\
&= \frac{2Lv_1 R(T)^{n+1}}{w(T)^n} + \frac{32v_1 c^{-1}\sigma_w^2 R(T)^n \ln(2T^4)}{w(T)^{n+1}} + (3L\sqrt{n}+1)w(T)T + \sum_{t=1}^{T} \bar{\nu}(m(T)) \\
&= \frac{2Lv_1 R(T)^{n+1}}{w(T)^n} + \frac{32v_1 c^{-1}\sigma_w^2 R(T)^n \ln(2T^4)}{w(T)^{n+1}} + (3L\sqrt{n}+1)w(T)T + T\bar{\nu}(m(T))
\end{aligned}$$

Now suppose $\mathcal{G}$ does not hold. Consider an arbitrary $t \in [T]$. If $y_t = 0$, then the regret at time t is at most 1. If $y_t = 1$, we still have $\|x_t\| \leq R(T)$, so by Lipschitz continuity, $r(x_t, 1) \geq -LR(T)$. Therefore $\mathbb{E}[\mathrm{Reg}(T) \mid \neg\mathcal{G}] \leq T + LR(T)T$. Lemma 6.3 implies that $\Pr[\neg\mathcal{G}] \leq 1/T$, so by the law of expectation,

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T)] &= \Pr[\neg\mathcal{G}]\,\mathbb{E}[\mathrm{Reg}(T) \mid \neg\mathcal{G}] + \Pr[\mathcal{G}]\,\mathbb{E}[\mathrm{Reg}(T) \mid \mathcal{G}] \\
&\leq \frac{1}{T}\cdot(T + LR(T)T) + \frac{2Lv_1 R(T)^{n+1}}{w(T)^n} + \frac{32v_1 c^{-1}\sigma_w^2 R(T)^n \ln(2T^4)}{w(T)^{n+1}} + (3L\sqrt{n}+1)w(T)T + T\bar{\nu}(m(T)) \\
&\in O\left(1 + LR(T) + \frac{2Lv_1 R(T)^{n+1}}{w(T)^n} + \frac{32v_1 c^{-1}\sigma_w^2 R(T)^n \ln(2T^4)}{w(T)^{n+1}} + L\sqrt{n}w(T)T + T\bar{\nu}(m(T))\right)
\end{aligned}$$

We now plug in $w(T) = T^{\frac{-1}{n+2}}$ and $m(T) = \ln T$. Since $\lim_{T\to\infty} w(T) = 0$, we have $\sigma_w^2 = nL^2 w(T)^2 + \sigma^2 \in O(\sigma^2)$. Similarly, for any $k \geq 0$, $R(T)^k = (\ln(T) + \sqrt{n}T^{\frac{-1}{n+2}})^k \in O((\ln T)^k)$. Thus

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T)] &\in O\left(\frac{L\ln T}{T} + \frac{L(\ln T)^{n+1}}{T^{\frac{-n}{n+2}}} + \frac{\sigma^2(\ln T)^{n+1}}{T^{\frac{-n-1}{n+2}}} + LT^{\frac{-1}{n+2}}T + T\bar{\nu}(\ln T)\right) \\
&= O\left((L+\sigma^2)T^{\frac{n+1}{n+2}}(\ln T)^{n+1} + T\bar{\nu}(\ln T)\right)
\end{aligned}$$

as required. ∎

B Simulations

In this section, we provide some basic simulations to illustrate the performance of our algorithm. Rigorous experimental validation of our algorithm is beyond the scope of the paper, but we hope that these simulations help provide intuition for the behavior of our algorithm. Our experimental setup is as follows:

1. *Input space.* We use synthetically generated i.i.d. inputs in $\mathbb{R}$.
2. *Input distribution.* We consider two different distributions ν: a Gaussian distribution with $\sigma^2 = 5$ and a Cauchy distribution with $\gamma = 1$, both centered at 0. The former is light-tailed, meaning that even reckless exploration is typically fine since very OOD inputs are unlikely. The latter is heavy-tailed, which is where OOD inputs are relatively common and so caution is crucial. (which is heavy-tailed).
3. *Reward function.* We consider the two extremes of possible reward functions: $r(x, 1) = 1$ and $r(x, 1) = 1 - |x|$. For the former, the task policy generalizes perfectly and the agent should always commit. For the latter, the task policy generalizes poorly and the agent should abstain except for a small region around the origin.
4. *Noise distribution.* In all cases, the noise variables $\eta_1, \ldots, \eta_T$ are sampled from a Gaussian distribution with $\sigma^2 = 1$ centered at 0.
5. *Time horizons.* We tested time horizons T up to 500,000.
6. *Algorithms.* In addition to our algorith, we tested three baselines: (1) always commit, (2) always abstain, and (3) Lipschitz UCB1, a standard non-cautious algorithm for Lipschitz bandits. Specifically, Lipschitz UCB1 is the uniform-discretization algorithm described in Section 8.2 of Slivkins et al. (2019), instantiated with UCB1 (as suggested in Section 8.1 of the same book). We initially planned to also include a risk-sensitive bandit algorithm but could not find one that would admit an apples-to-apples comparison (see § 2 for details).

Our code can be found at `https://github.com/bplaut/abstention-bandits-simulations`. The simulations were not computationally intensive and were run on a personal laptop.

Figure 2 shows the simulation results. For readability, curves with very high regret are truncated. Each algorithm was run 10 times with 10 different seeds. The solid lines are the mean of those runs, and the shaded region indicates the maximum and minimum values across runs. In many cases, the shaded region is not visible because it is so small relative to the scale of the plots.

The combination of input distribution and reward function results in four concrete scenarios, each corresponding to one of the panels in Figure 2. The top row corresponds to $r(x, 1) = 1$ and the bottom row corresponds to $r(x, 1) = 1 - |x|$. The left and right columns correspond to Gaussian inputs and Cauchy inputs respectively.

In the top row, $r(x, 1) = 1$ means that the optimal strategy is to always commit, meaning that there is no benefit to caution. As a result, Lipschitz UCB1 performs well for both Gaussian or Cauchy inputs. Our algorithm also performs well (as it does in all cases). As expected, "Always abstain" performs pessimally.

The bottom row is where things get interesting. The reward function $r(x, 1) = 1 - |x|$ implies that there is a small area of in-distribution inputs near the origin where the agent should commit, but otherwise the agent should abstain. The agent needs to learn the boundaries of this region. "Always commit" performs terribly here for both input distributions. "Always abstain" performs better, but still exhibits linear regret due to never learning to commit near the origin. In the bottom row, committing on far OOD inputs can have catastrophic costs, so caution becomes relevant. However, for a light-tailed input distribution (bottom left), far OOD inputs are rare enough that caution is not necessary; hence the success of Lipschitz UCB1. The crucial scenario is the bottom right: committing on far OOD inputs is catastrophic *and* the agent actually encounters such inputs. In this scenario, Lipschitz UCB1 is abysmal due to its reckless exploration. In contrast, our algorithm explores cautiously and successfully learns where to commit without ever risking catastrophe.

We can gain further insight into the behavior of algorithm by examining the number of abstentions and certified negative bins over time. Figure 3 shows how these quantities evolve over time for the $r(x, 1) = 1 - |x|$ setting. (These quantities are less interesting for $r(x, 1) = 1$, since no bins will ever be certified negative.) Unlike Figure 2 which compares a variety of values of T, Figure 3 studies the single value of $T = 500,000$ and plots the progress of the algorithm across time steps t. The left and right plots in Figure 3 correspond to Gaussian inputs and Cauchy inputs, respectively. In both plots, the number of certified negative bins initially increases quickly as the algorithm explores, and then plateaus. The absention rate also stabilizes as a function of the probability mass outside of the trusted region (i.e., how often we get inputs that are too OOD to even explore).

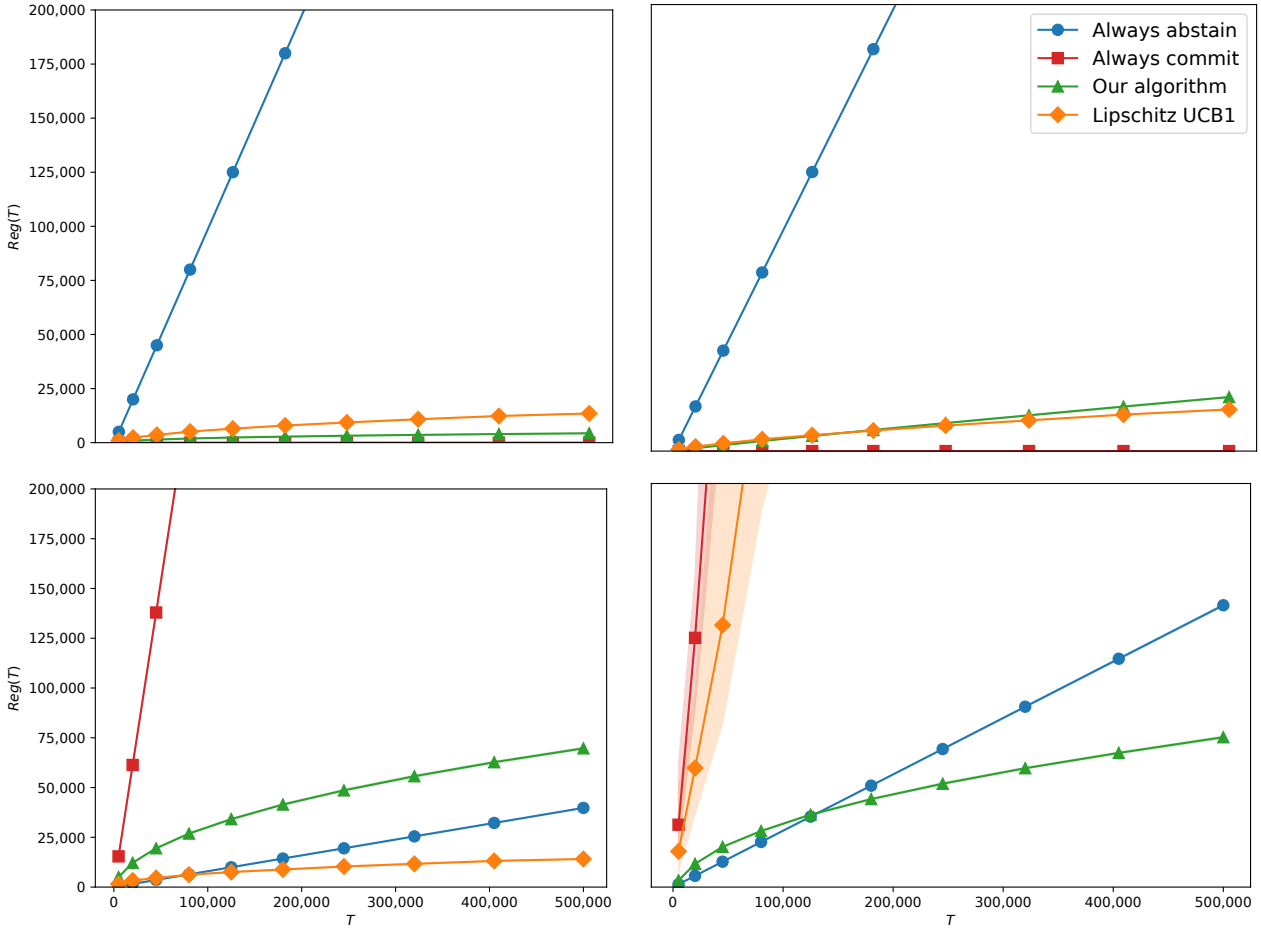

Figure 2: Simulation results. Top row: $r(x, 1) = 1$. Bottom row: $r(x, 1) = 1 - |x|$. Left column: inputs are i.i.d. from a Gaussian distribution. Right column: inputs are i.i.d. from a Cauchy distribution. Each curve is the mean of 10 runs with different seeds; the shaded region indicates the minimum and maximum of those runs.

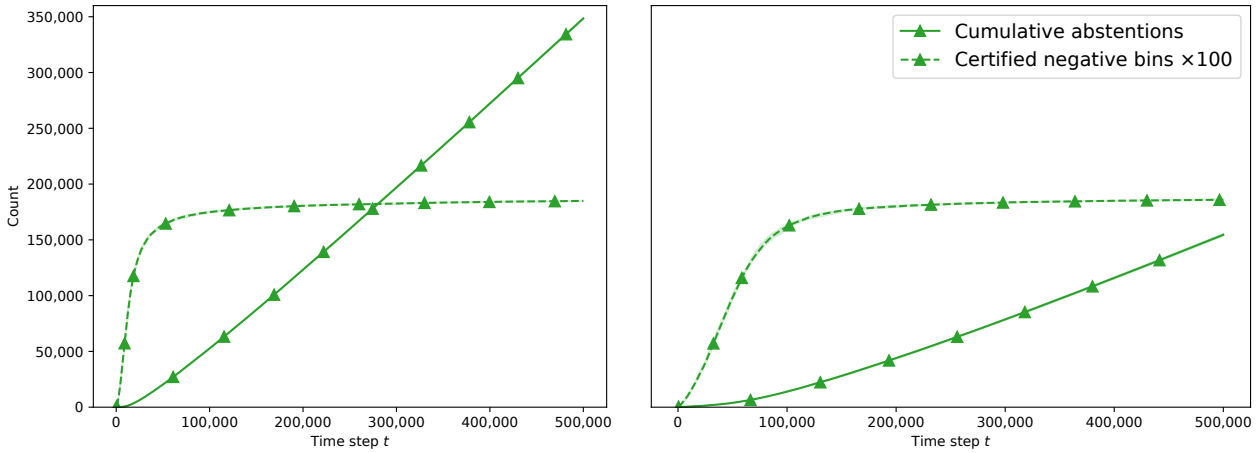

Figure 3: Abstentions and the number of certified negative bins over time. Both plots correspond to $r(x, 1) = 1 - |x|$. The left plot and right plot correspond to Gaussian inputs and Cauchy inputs, respectively.