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

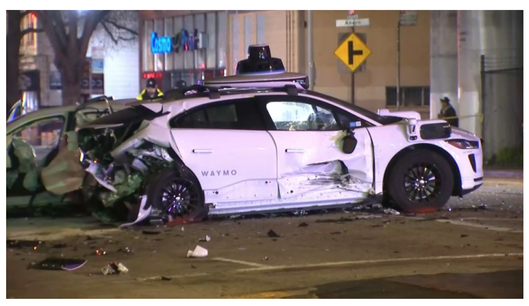
Sarah Liaw* Benjamin Plaut*

UC Berkeley

CHAI Center for Human-Compatible Artificial Intelligence

The Problem: Cautious Learning

Standard reinforcement learning aggressively explores to learn. But what happens when **taking a wrong action is catastrophic?**



Our approach: Stop learning and abstain when OOD!

Formal Setup

- **Input Space**: $x_t \overset{\text{i.i.d.}}{\sim} \nu$ on $\mathbb{R}^n$.
- **Actions**: $y_t \in \{0 \text{ (Abstain)}, 1 \text{ (Commit)}\}$.
- **Rewards**: Observe noisy $r_t = r(x_t, y_t) + \eta_t$ on commit.
- **Noise**: η_t zero-mean subgaussian.
- **Regularity**: $r(\cdot, 1)$ is L-Lipschitz.
- $r(\mathbf{0}, 1) > 0$.
- **Expected Regret Objective**:

$$\mathbb{E}[\mathrm{Reg}(T)] = \mathbb{E}\left[\sum_{t=1}^{T}\left(\max_{y\in\{0,1\}} r(x_t, y) - r(x_t, y_t)\right)\right]$$

The Limits and Virtues of Caution

Why Caution Is Necessary

If the agent always commits initially, the expected regret from a single catastrophic mistake scales to infinity.

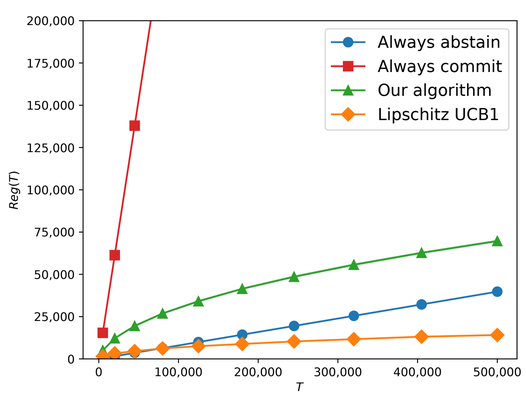
$$\mathbb{E}[\mathrm{Reg}(T)] = \infty$$

Why Caution Has Limits

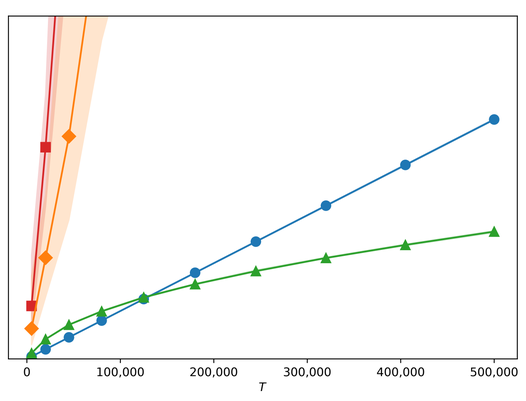
If every input is too far OOD, then sublinear regret is formally impossible regardless of the policy.

$$\mathbb{E}[\mathrm{Reg}(T)] \notin o(T)$$

Regret Bounds



Gaussian



Cauchy

Heavy-tailed settings heavily break standard UCB. Our Cautious Learner completely restricts this regret by recognizing and securely abstaining from untrusted regions.

Algorithm & Trusted Region Schematic

The agent partitions a trusted origin-centered ball $\mathcal{B}$ of radius $m(T)$ into bins $B \in \mathcal{H}$ of side length $w(T)$. For each bin, it maintains an empirical commit reward mean $\hat{\mu}_B$ and a confidence radius $\gamma(k_B) = \sqrt{c^{-1}\sigma_w^2 \ln(2T^4)/k_B}$.

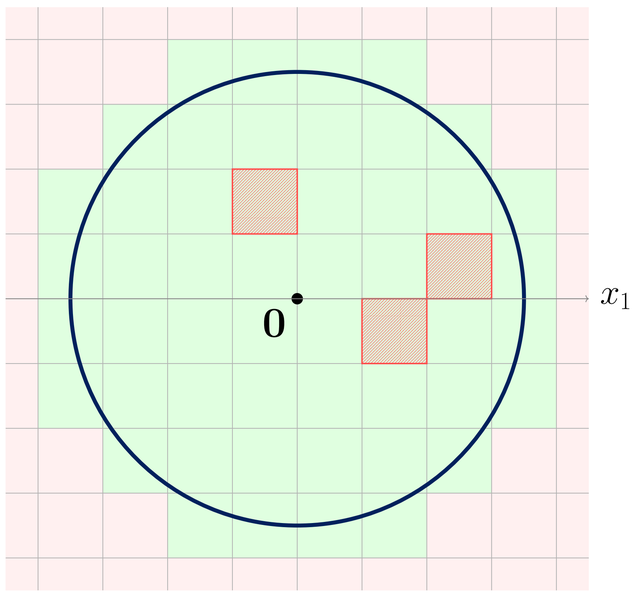
Decision Rule at time t for input x_t:

1. Far OOD (Too Risky)
Condition: $\|x_t\| > m(T)$ $\longrightarrow$ **Abstain**

2. Certified Harmful Bin
Condition: $x_t \in B$, but $\underbrace{\hat{\mu}_B + \gamma(k_B)}_{\text{UCB}} + \underbrace{L\sqrt{n}w(T)}_{\text{dispersion}} < 0$ $\longrightarrow$ **Abstain**

3. Informative/Safe Bin
Condition: Otherwise $\longrightarrow$ **Commit**

$\hookrightarrow$ Update empirical mean $\hat{\mu}_B$ and visit count k_B.

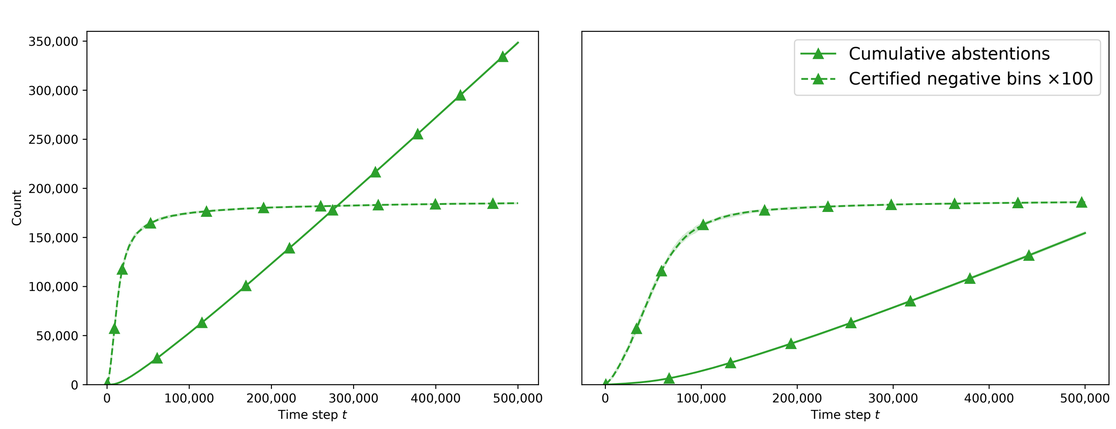


Main Result

Theorem: For any fixed i.i.d. input distribution ν, the algorithm achieves:

$$\mathbb{E}[\mathrm{Reg}(T)] \in O\left(\underbrace{(L+\sigma^2)\, T^{\frac{n+1}{n+2}} (\ln T)^{n+1}}_{\text{discretization error (trusted region)}} + \underbrace{T\,\bar{\nu}(\ln T)}_{\text{tail error (OOD inputs)}}\right)$$

Behavioral Analysis: Commit vs. Abstain



Gaussian **Cauchy**

Conclusion

- We formalize learning under **catastrophic, unbounded risk** and show naive exploration can incur **infinite regret**.
- Our algorithm **learns only in trusted regions** and abstains otherwise.
- **Future work:** non-i.i.d. inputs, and leveraging structure beyond Lipschitzness.

Overall goal: safety guarantees for high-stakes AI systems