Provably Efficient Maximum Entropy Exploration

Elad Hazan, Sham Kakade, Karan Singh, Abby Van Soest

PRINCETON UNIVERSITY

Google AI

UNIVERSITY of WASHINGTON

Motivation

Task Agnostic Exploration

[Lee et. al '19, Fu et. al 2018, …]

Phase 1: Reward-free interactions.

Phase 2: A suite of tasks (with reward).

An exploration policy given a prior $P^*(s)$?

$$\max_{\pi} \{ -KL(d_\pi \,||\, P^*(s)) \}$$

$d_\pi(s)$ is the state distribution under policy π.

Curiosity & Exploration Bonus

[Pathak et. al '17, Bellemare et. al '16, Tang et. al '17…]

Novelty-based exploration bonus.

$$\max_{\pi}\{ E_{d_\pi(s)}[R(s) - \log d_\pi(s)] \}$$

Other Formulations: Downstream task efficiency, Option discovery, Sparse rewards.

No Reward Signal.

Question: What is the agent capable of?

$$\max_{\pi} \{ H(d_\pi) = -\sum_s d_\pi(s) \log d_\pi(s) \}$$

Not a **scalar** reward function.

How to solve this efficiently?

The Setting

- π induces a distribution over states.
  - $d_\pi(s) = (1-\gamma)(P(s_0 = s|\pi) + \gamma P(s_1 = s|\pi) + \gamma^2 P(s_2 = s|\pi) + \ \dots)$
- A policy class Π (infinite).
- Concave functional H, acting on the state distribution.

$$\max_{\pi \in \Pi} H(d_\pi)$$

Proposition

$H(d_\pi)$ is not concave in π.

A Reductions-based Approach:

Reward-based Planning Oracle: Given r, output π with $V_\pi \geq \max_\pi V_\pi - \varepsilon$.

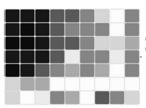
Density Estimation: Given π, output an estimate d'_π so that $|d'_\pi - d_\pi|_\infty \leq \varepsilon$.

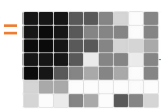
The MaxEnt Algorithm

Concept: Uniform Mixture of Policies $C = (\pi_1, \ldots, \pi_k)$.

Initialization: Start with a 1-policy mixture.

For t = 0, … T-1 **do**

1.  $= DensityEst(mix_t)$.

2. $r_t(s) = \left.\frac{dH(X)}{dX}\right|_{X=}$  $= -(\log d_\pi(s) + 1)$.

3. Compute $\pi_{t+1} = ApproxPlan(r_t, \varepsilon)$.

4. Update the *uniform* mixture to include π_{t+1}.

Estimate State Distribution:
Given π, output d'_π
so that $|d'_\pi - d_\pi|_\infty \leq \varepsilon$.

Reward-based Planning Oracle:
Given r, output π with
$V_\pi \geq \max_\pi V_\pi - \varepsilon$.

Result

Interpret as Conditional Gradient Descent.

Main Theorem

For concave, β-smooth $R(X)$, ie. $|\nabla^2 R(X)| \leq \beta$, the algorithm guarantees

$$H(d_{mix}) \geq \max_{\pi \in \Pi} H(d_\pi) - \varepsilon,$$

As long as

$$T \geq \beta \varepsilon^{-1}$$

Corollary (Entropy)

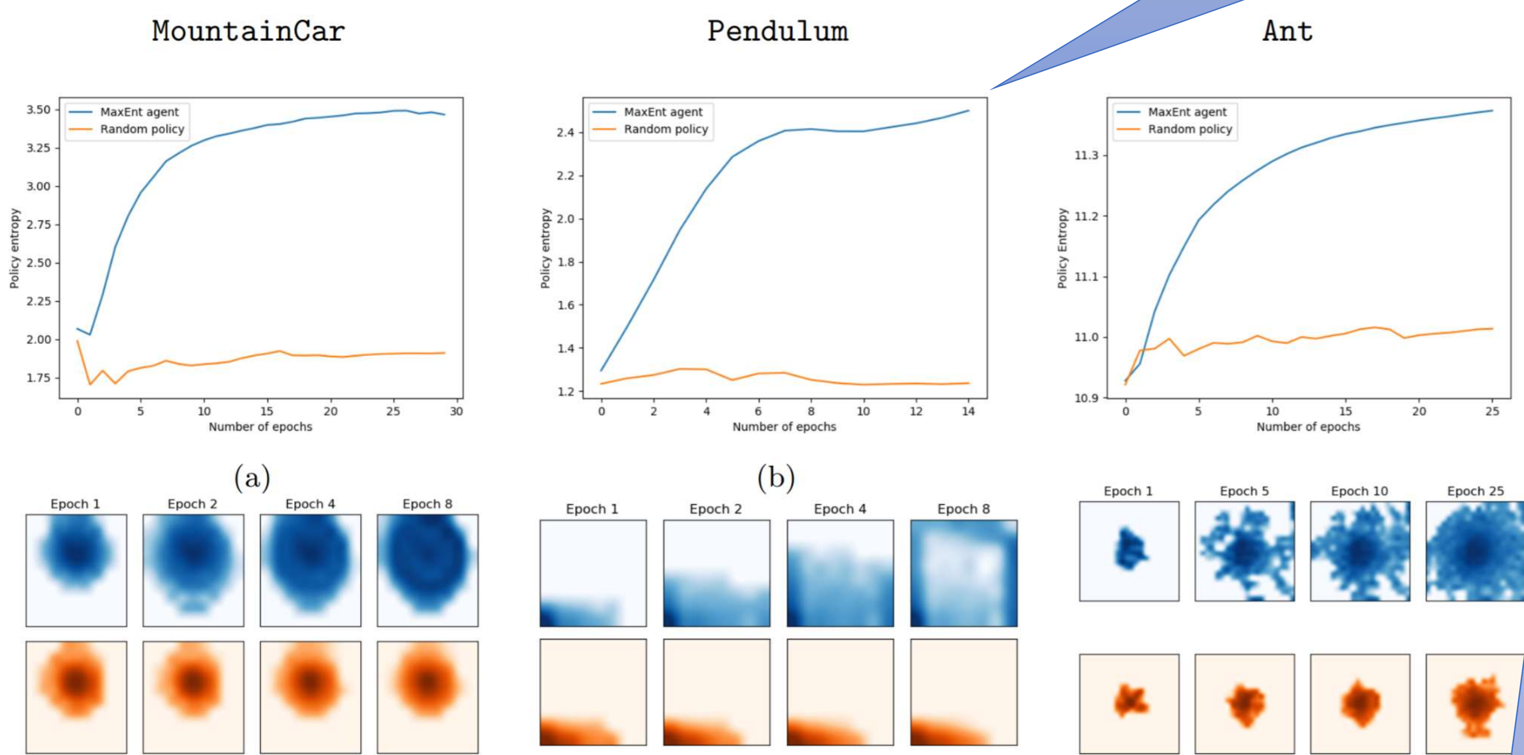
For the entropy objective, tle algorithm needs to run for $T \geq S\varepsilon^{-2}$ steps.

Corollary (Finite MDP; No Oracles)

$O(S^2 A)$ samples suffice to implement the oracles across all iterations.

Prelim. Experiments

Entropy vs. Iterations:
Entropy saturates in a few iterations of the algorithm.



Objective: $\min_{\pi} KL(Unif||d_{\pi})$

Simple, count-based **Density Estimator** (+rand proj).

Planning: Policy Gradient / Actor Critic

State Coverage:
Visits all reachable states in a few iterations.

Pacific Ballroom

#115

The Take-away

Optimize Functions of State Distribution via Blackbox RL algorithms.