Cost Efficient Exploration for Reinforcement Learning

Pankayaraj Pathmanathan, Christabel Acquaye, Prannoy Namala

UNIVERSITY OF MARYLAND

FEARLESSLY FORWARD

Motivation

Background

Exploration and Exploitation Dilemma

RL maintains this balance

Information Theory guided exploration

Motivation

Minimize the cost of exploration while achieving good performance

Related Works

Literature Review

- Curiosity Based
  - Induce Exploration by Leveraging Uncertainty about the Environment
  - Paper: Formal theory of creativity, fun, and intrinsic motivation: Schmidhuber [2010]; Curiosity-driven exploration in deep reinforcement learning via bayesian neural networks;Houthooft et al. [2016]
- Information Theory Based
  - Occupancy Information Ratio

Occupancy Information Ratio: Infinite-Horizon, Information-Directed, Parameterized Policy Search

- Information-directed exploration to infinite-horizon, parameterized policy search problems.

$$\rho(\theta) = \frac{J(\theta)}{\kappa + H(d_\theta)}$$

Cost

Entropy of Steady State Estimate

Goals

→ Incorporate exploration cost in OIR

→ OIR Validation

→ Comparative Study

Methodology

Incorporating exploration cost in OIR

- OIR induces a uniform state occupancy limited by the reward maximization
- Non uniform policy induces over exploration in certain states
- Penalize over exploration

$$\rho(\theta) = \frac{J(\theta) + \alpha . \mathbb{1}_{\sum_s N^t(\theta,s)/N^t_{tot} > N} \cdot \sum_t p}{\kappa + H^t\left(d_\theta\right)}$$

Intuition

$$D_{\mathrm{TV}}(P,Q) := \sup_{A\in\mathcal{F}} |P(A) - Q(A)|$$

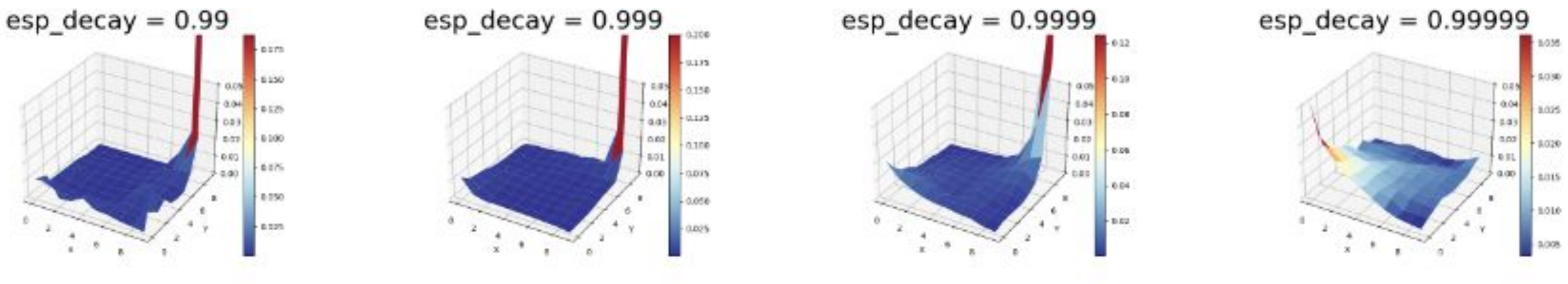
$$d_{mix}(t) = \sup_{s\in\Omega} D_{\mathrm{TV}}\left(P^t(s,\cdot),\mu\right)$$
$$\tau_{mix}(\epsilon) = \inf t : d_{mix}(t) \leq \epsilon$$

Limit on number of samples

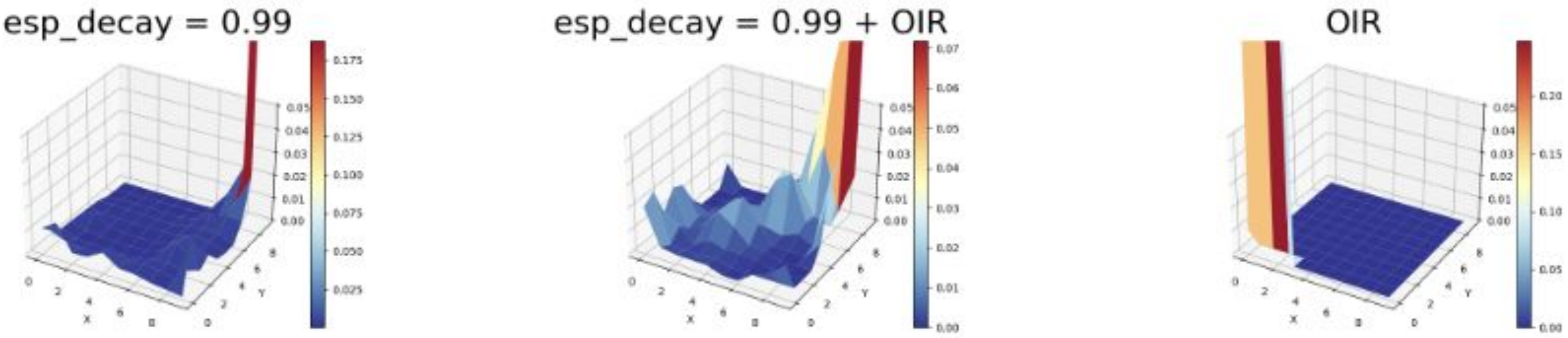
$$P(|d_\theta - d_\theta^*| > \delta) \leq 2e^{-2n\delta^2}$$

$$\begin{aligned} P(|d_\theta - d_\theta^*| > \delta) &\leq \alpha \\ 2e^{-2n\delta^2} &\leq \alpha \\ -2n\delta^2 &\leq \ln \frac{\alpha}{2} \\ n &\geq \frac{1}{2\delta^2} \ln \frac{2}{\alpha} \end{aligned}$$

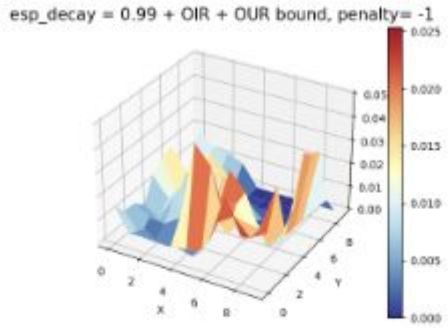
Exploration vs State Occupancy

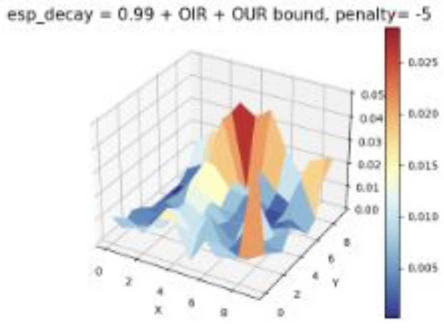


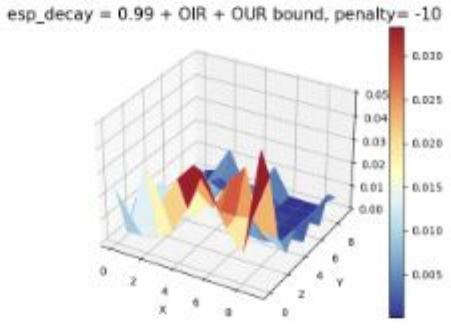
State Occupancy in OIR



State Occupancy in our work







UNIVERSITY OF MARYLAND FEARLESSLY FORWARD

OIR Validation and Comparative Study

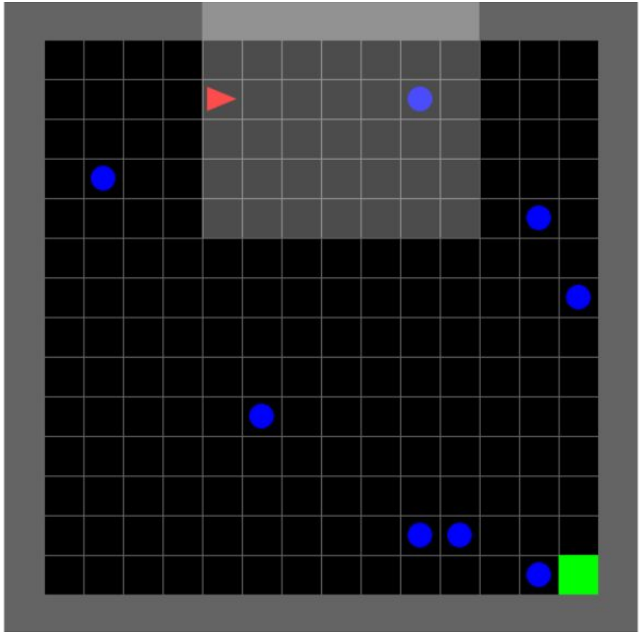
OIR Validation

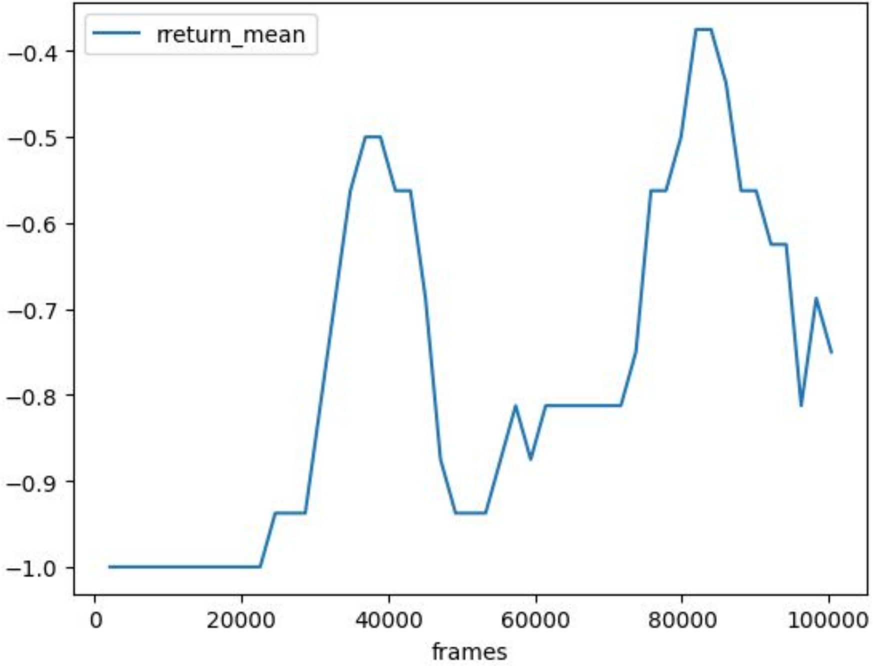
- GridWorld
- IDAC

Comparative Study

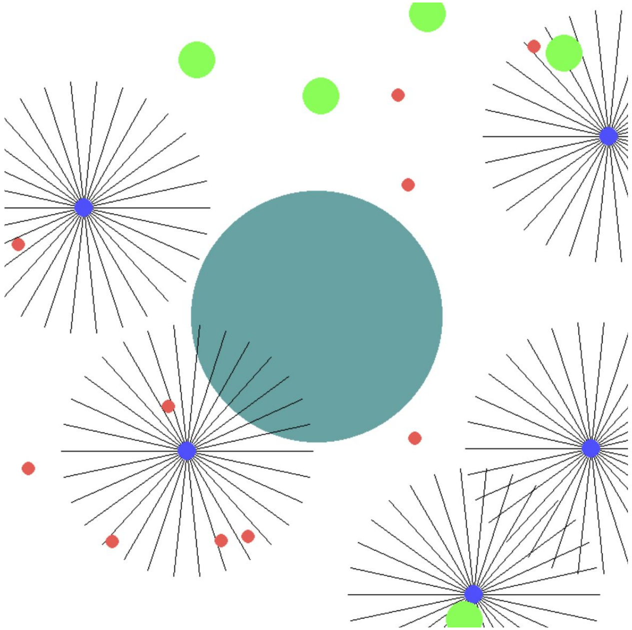
- WaterWorld
- PPO and DDPG
- Both vanilla and OIR versions

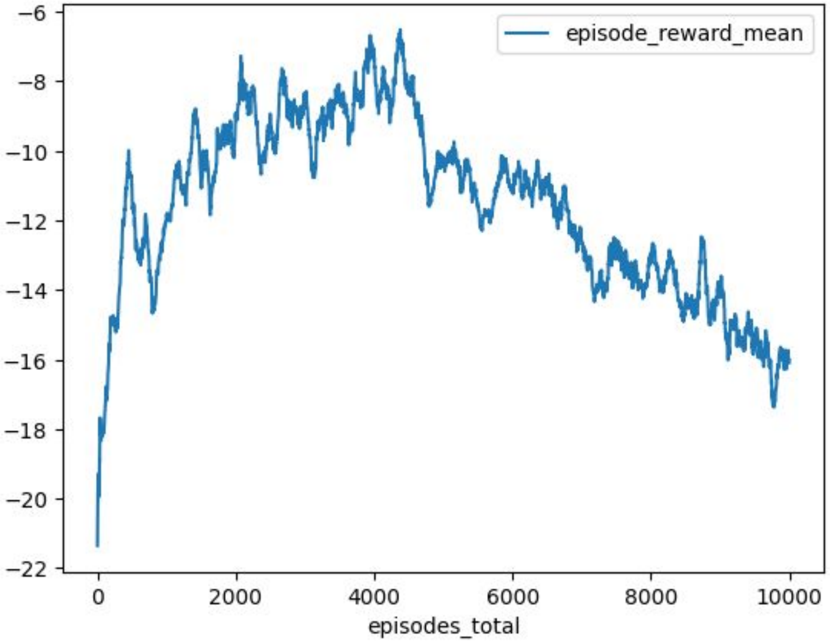
A2C Algorithm for Gridworld



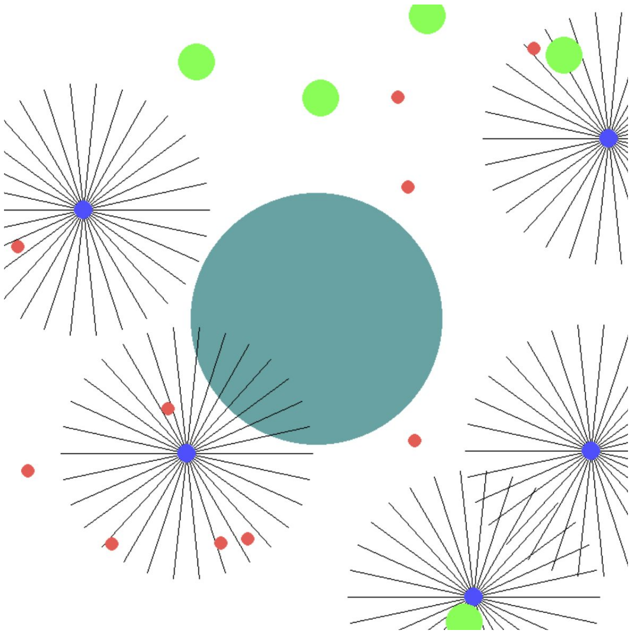


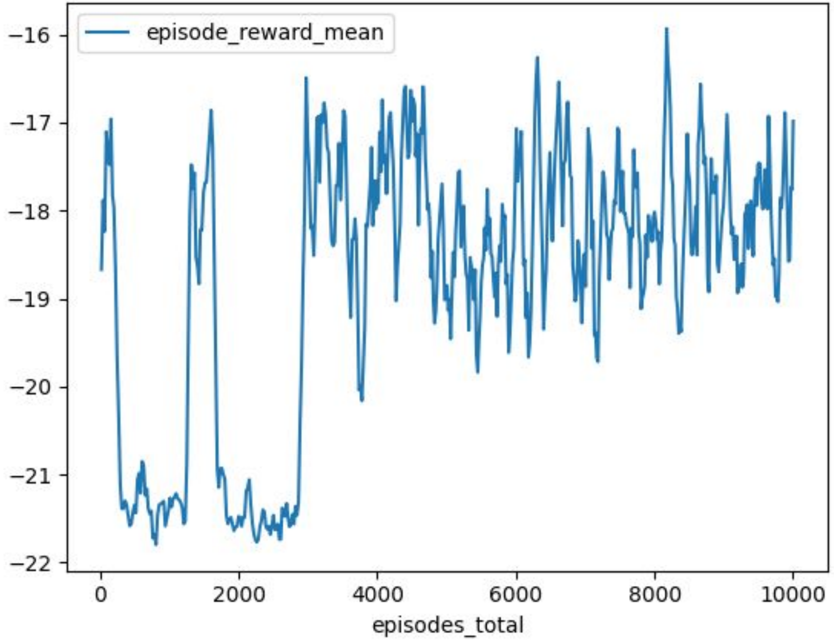
PPO Algorithm for WaterWorld





APEX DDPG Algorithm for Waterworld





Unachieved Objectives

Density Estimation

- Learning based method
- Kernel based Density Estimate
- Multi-Agent Implementation

$$\rho(\theta) = \frac{J(\theta)}{\kappa + H(d_\theta)}$$

Future Work

- Completing implementation of Density Estimator
- Applying OIR to Heterogeneous case (Pursuit Evasion)
- Extend to Large Scale Environment for ORI cost

Questions?