

Exploration for RL

Inductive biases in exploration strategies

Alexander Telfar

June 30th, 2019

What is RL?

(learning to) make optimal decisions

Context (S), potential actions (A), utility function / reward (r).

Markov decision problems

$M = \{S, A, \tau, r, d_0\}$ (the MDP)

$\tau : S \times A \rightarrow \Delta(S)$ (the transition fn)

$r : S \times A \rightarrow \mathbb{R}^+$ (the reward fn)

$$\pi : S \rightarrow \Delta(A) \quad (\text{the policy})$$

$$s_{t+1} \sim \tau(s_t, a_t), a_t \sim \pi(s_t) \quad (\text{sampled actions and states})$$

$$V(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \quad (\text{value estimate})$$

$$\pi^* = \operatorname{argmax}_{\pi} V(\pi) \quad (\text{the optimisation problem})$$

Alternative formulation

$$V(\pi^*) \equiv \mathbb{E}_{s_0 \sim d_0} \max_{a_0} r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim p(\cdot | s_0, a_0)} \left[\max_{a_1} r(s_1, a_1) + \gamma \mathbb{E}_{s_2 \sim p(\cdot | s_1, a_1)} \left[\max_{a_2} r(s_2, a_2) + \gamma \mathbb{E}_{s_3 \sim p(\cdot | s_2, a_2)} \left[\dots \right] \right] \right]$$

Why are RL problems hard?

Because of three main properties;

1. they allow, **evaluations**, but dont give 'feedback',
2. the observations are sampled **non-IID**,
3. they provide **delayed** credit assignment.

Example: Multi-armed Bandits

The two armed bandit is one of the simplest problems in RL.

- Arm 1: $[10, -100, 0, 0, 30]$
- Arm 2: $[2, 0]$

Which arm should you pick next?

Why not just do random search?

An example: MineRL

Goal: Find and mine a diamond.

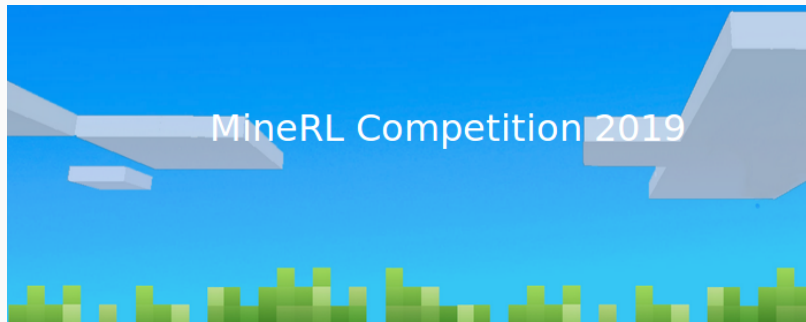


Figure 1: <http://minerl.io/competition/>

What are some existing exploration strategies?

- Injecting noise: Epsilon greedy, boltzman
- Optimism in the face of uncertainty
- Bayesian model uncertainty and Thompson sampling
- Counts / densities and Max entropy
- Intrinsic motivation (Surprise, Reachability, Randomly picking goals)
- Disagreement

Note. They mostly require some form of memory and / or a model of uncertainty. Exploration without memory is just random search. . .

In the simplest setting, we can just count how many times we have been in a state. We can use this to explore states that have low visitation counts.

$$P(s = s_t) = \frac{\sum_{s=s_t} 1}{\sum_{s \in S} 1} \quad (\text{normalised counts})$$

$$a_t = \operatorname{argmin}_a P(s = \tau(s_t, a)) \quad (\text{pick the least freq } s)$$

The policy is rewarded for;

‘Surprise’ (prediction error)

$$r_t = \| s_{t+1} - f_{dec}(f_{enc}(s_t, a_t)) \|_2^2$$

‘Reachability’ (is reachable within k steps?)

$$r_t = \min_{x \in M} D_k(s_t, x)$$

Maximum entropy

$$P^\pi(\xi|\pi) = d_0(s_0)\pi(a_0|s_0)\tau(s_1|s_0, a_0)\pi(a_1|s_1)\tau(s_2|s_1, a_1)\dots$$

(probability of a trajectory)

$$d^\pi(s, t) = \sum_{\text{all } \xi \text{ with } s = s_t} P^\pi(\xi|\pi)$$

(all trajectories with s at t)

$$d^\pi(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t d^\pi(s, t)$$

$$\pi^* = \operatorname{argmax}_{\pi} \mathbb{E}_{s \sim d^\pi} [-\log d^\pi(s)]$$

Inductive biases in exploration strategies

So my questions are;

- do some of these exploration strategies prefer to explore certain states first?
- which inductive biases do we want in exploration strategies?
- how can we design an inductive biases to accelerate learning?
- what is the optimal set of inductive biases for certain classes of RL problem?
- how quickly does the state visitation distribution converge?

(we will come back to this)

*Of the possible candidates, which one should we pick?
Well, we could try each of them, or... just pick the one
we think will work, a priori.*

- The 'simplest'.
- The one most likely to generalise.

A set of priors, that guide the search where data doesn't.

Implicit regularisation

Matrix factorisation ($m \ll d^2, Z \in \mathbb{R}^{d \times d}$)

$$y_i = \langle A_i, W^* \rangle \quad (\text{matrix sensing})$$

$$\mathcal{L}(X) = \frac{1}{2} \sum_{i=1}^m (y_i - \langle A_i, XX^T \rangle)^2 \quad (\text{factorisation from observations})$$

$$X^* = \underset{X}{\operatorname{argmin}} \mathcal{L}(X) \quad (\text{the optimisation problems})$$

When stochastic gradient descent is used to optimise this loss (with initialisation near zero and small learning rate), the solution returned also has minimal nuclear norm

$$X^* \in \{X : \underset{X \in S}{\operatorname{argmin}} \|X\|_*\}, \quad S = \{X : \mathcal{L}(X) = 0\}.$$

Types of prior?

- relational
- visual
- subgoals
- exploration

Relational priors

We know;

- what furnaces are 'for' (ore \rightarrow metal)
- that coal is needed for heat (furnace + coal \rightarrow on(furnace))
- that iron can be profuced via a furnace (on(furnace) + iron ore \rightarrow iron)





Figure 2: Which one is probably diamond?

We can easily generate a curriculum of subgoals;

1. Kill food
2. Find shelter
3. Build tools
4. Get money

Exploration priors

We quickly generalise spatial exploration to be much of the same; trees, rivers, mountains, ... And focus on exploring the many crafting possibilities.



Also;

- we know that diamonds are likely to be found (deep) underground
- we know that pick axes will be useful for exploring underground

The power of priors

Last time I tried to mine a yellow sparkly rock, nothing happened, this time, 1,000 actions later, I got gold. Which action(s) helped?

I took 10,000 actions, now I have an axe. It doesn't appear to help me get diamonds.

After trying all 2,000 different ways of cutting down a tree. I am ready to conclude that you cannot get diamonds from cutting down trees. What about exploding a tree?

How do RL algorithms implicitly regularise exploration?

Exploration via;

Surprise

- Has a bias towards states with more noise in them.

Density

- The approximation of the density may be biased.

Intrinsic motivation

- Highly dependent on its history of samples.

The state visitation distribution

How can we reason, in a principled manner, about bias / regularisation in exploration strategies?

$$d^A(s, t) = (1 - \gamma) \sum_{t=0}^t \gamma^t \Pr^A(s = s_t)$$

For each different RL algol;

- Do state-visitations, $d(s_i, t)$ converge monotonically to $\frac{1}{n}$?
- Which state-visitations, $d(s_i, t)$, converge first?
- What is the difference between the i different convergence rates?
- Does $d(s, t)$ converge to uniform as $t \rightarrow \infty$?

Thank you!

And questions?