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

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

$$\tau : S \times A \to \Delta(S) \qquad \text{(the transition fn)}$$

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

MDPs

$$\pi : S \to \Delta(A) \qquad \text{(the policy)}$$

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

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

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, V(\pi) \qquad \text{(the optimisation problem)}$$

Alternative formulation

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

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 do exploration strategies matter?

Why not just do random search?

- Too much exploration and you will take many sub optimal actions, despite knowing better.
- Too little exploration and you will take 'optimal' actions, at least you think they are optimal...

An example: MineRL

Goal: Find and mine a diamond.

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

What do we require from an exploration strategy?

- Non-zero probability of reaching all states, and trying all actions in each state.

Nice to have

- Converges to a uniform distribution over states.
- Scales sub-linearly with states
- Samples states according to their variance. More variance, more samples.

What about goal conditioned exploration?

- ?

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...

Counts / densities

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 have low visitation counts.

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

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

Intrinsic motivation

'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

$$
\begin{aligned}
P^{\pi}(\tau|\pi) &= d_0(s_0)\Pi_{t=0}^{\infty}\pi(a_t|s_t)P(s_{t+1}|s_t, a_t) \\
d^{\pi}(s,t) &= \sum_{\text{all } \tau \text{ with } s = s_t} P^{\pi}(\tau|\pi) \\
d^{\pi}(s) &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d^{\pi}(s,t) \\
\pi^* &= \operatorname*{argmax}_{\pi} \mathop{\mathbb{E}}_{s\sim d^{\pi}}[\log d^{\pi}(s)]
\end{aligned}
$$

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 strageties?
- 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)

Inductive bias

Underconstrained problems.

Occam's Razor and overfitting.

Human bias in Minecraft

Types of prior?

- relational
- visual
- subgoals
- exploration

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.

Relational priors

We know;

- what furnaces are ‘for’ (ore -> metal)
- that coal is needed for heat (furnace + coal -> on(furnace))
- that iron can be profuced via a furnace (on(furnace) + iron ore -> iron)

Visual priors

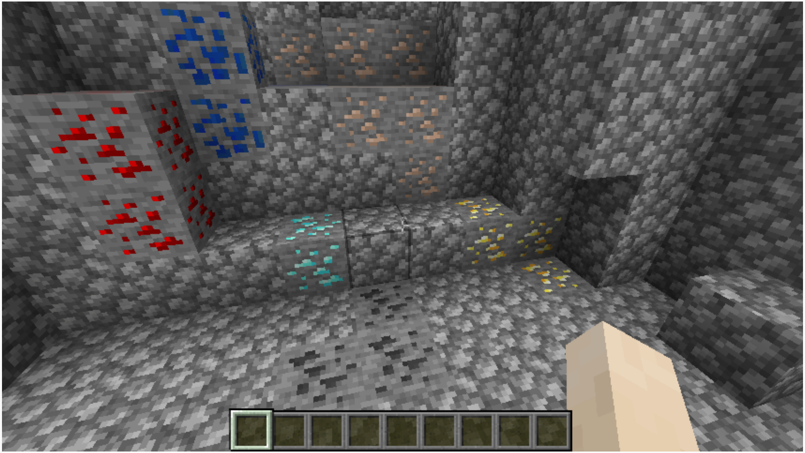

Figure 2: Which one is probably diamond?

(Sub)goal priors

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

A quick aside: Implicit regularisation

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

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

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

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

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

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?

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

For each different RL algol;

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

Thank you!

And questions?