Lecture 20 - Exploration in deep RL

Guiliang Liu

The Chinese University of Hong Kong, Shenzhen

DDA4230: Reinforcement Learning
Course Page: [Click]

Hardness of Exploration

Recall that the Atari game Montezuma's Revenge:

this is easy (mostly)

Why?

this is impossible

Hardness of Exploration

Recall that the Atari game Montezuma's Revenge is played with several factors as

- Getting key $\rightarrow$ reward
- Opening door $\rightarrow$ reward
- Getting killed by skull $\rightarrow$ nothing (is it good? bad?)
- Finishing the game only weakly correlates with rewarding events
- We know what to do because we understand what these sprites mean!

This game is notoriously difficult for reinforcement learning and extensive efforts have been deployed to solve the problem.

Hardness of Exploration

Noisy TV Experiment. Assume that in a maze an agent aims at finding the correct path to escape the maze. In the maze, there is a TV that presents noisy signals An agent that focuses on exploration could be trapped into sitting in front of the TV and watching the TV for good, without solving the task.

Agent in a maze with a noisy TV

Agent in a maze without a noisy TV

The Exploration-Exploitation dilemma

There are two potential definitions of the exploration problem

- How can an agent discover high-reward strategies that require a temporally extended sequence of complex behaviors that, individually, are not rewarding?
- How can an agent decide whether to attempt new behaviors (to discover ones with higher reward) or continue to do the best thing it knows so far?

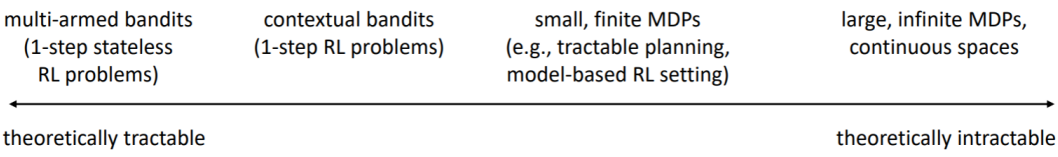
The Exploration-Exploitation dilemma

RL algorithms should balance between exploitation and exploration

- Exploitation: doing what you know will yield the highest reward
- Exploration: doing things you haven't done before, in the hopes of getting even higher reward

Exploration strategies in bandits

Tractability of exploration in a variety of settings.



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Exploration strategies in bandits

Optimistic exploration. Optimism is the fundamental underlying idea of UCB-based methods. While we maintain an empirical mean reward $\mu_{i,t}$ for each arm, instead of picking the arm with the largest empirical mean, we add a exploration bonus $\sigma_{i,t}$ as some sort of variance estimate, for example

$$\sigma_{i,t} = \sqrt{\frac{2 \log t}{N_{i,t}}}.$$

The optimism idea is to keep trying an arm until it is sure that the arm is suboptimal (with probability at least $1 - \delta$, as an intuition by Hoeffding's inequality in UCB). This idea extends to other algorithms as well.

Exploration strategies in bandits

Thompson sampling. The algorithm simply samples one state from the belief space and executes the exploitation process using this state. In bandits, this equals to pretending the Thompson sampling to be the true mean reward and running the greedy algorithm. The algorithm dates back to 1933 but was grants rigorous theoretical guarantees only recently.

Exploration strategies in bandits

Information Theory. Without loss of generality, let z be a latent variable we aim to determine, which can be a^*, V^*, or other latent variables such as goals for goal-conditioned policies. Then, with a new observation y, this entropy is reduced by

$$H(z) - H(z \mid y),$$

where y can be next states and reward signals. Then it is natural to formulate the information gain of z given y as $IG(z, y) = \mathbb{E}_y[H(z) - H(z \mid y)]$. The information gain can be utilized to encourage exploration. For example in bandits, it chooses arm i with the smallest $\widehat{\Delta}_i / IG(\theta_i \mid a_t = i)$.

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Exploration in reinforcement learning

The ideas of exploration in bandits can be extended to reinforcement learning:

1. Optimistic exploration
   - New state = good state;
   - Requires estimating state visitation frequencies or novelty;
   - Typically realized by means of exploration bonuses.
2. Posterior sampling
   - Learn distribution over Q-functions or policies;
   - Sample and act according to sample.
3. Information gain
   - Reason about information gain from visiting new states;
   - Reason about information gain from achieving new goals.

Optimistic exploration in RL

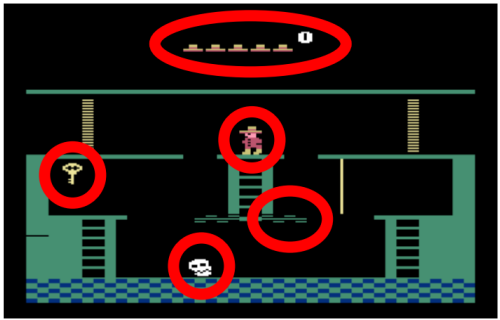
Recall that the UCB algorithm chooses

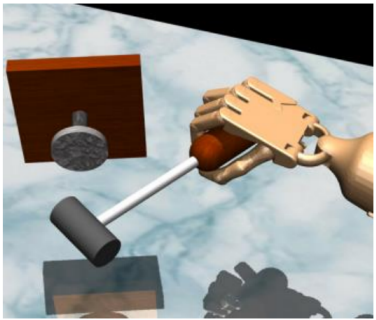
$$a = \arg\max_i \hat{\mu}_{i,t} + \sqrt{\frac{2\log t}{N_{i,t}}}.$$

For RL, it is natural to extend the exploration bonus $\sqrt{\frac{2\log t}{N_{i,t}}}$ with some generalized count-based bonus term $B(N_{s,t})$, where $B(\cdot)$ is a monotonically decreasing function that determines the bonus. This term can be added to the reward which generalizes the mean estimation of bandit arms, as $r_b = r + B(N_{s,t})$. This can be done for any model-free algorithms in discrete state spaces.

Optimistic exploration in RL

A concern is exactly the same state is almost never observed for a second time.





Optimistic exploration in RL

One approach is to fit a generative model $\mathbb{P}_\theta(s)$, which describes the distribution of the states that have been visited so far.

- Once a new state s' is visited, $\mathbb{P}_\theta(s')$ will describe how dissimilar s' is compared to previous experiences. A lower $\mathbb{P}_\theta(s')$ will result in a larger exploration bonus.
- $\mathbb{P}_\theta(s')$ is known as the pseudo-count, which generates the frequency $N_{s,t}/N$ of discrete spaces, where N denotes the number of experiences.

Optimistic exploration in RL

The update of $\mathbb{P}_\theta(s)$ is by incrementally estimating a new model and the count is calculated in analogous to the update of frequency by solving

$$\mathbb{P}_\theta(s) = \frac{N_{s,t}}{N}, \quad \mathbb{P}_{\theta'}(s) = \frac{N_{s,t}+1}{N+1},$$

where $\mathbb{P}_{\theta'}(s)$ is the updated generative model after visiting s. Choices of the bonus term include $B = \sqrt{\frac{2\log t}{N_{s,t}}}$, $B = \sqrt{\frac{1}{N_{s,t}}}$, $B = \frac{1}{N_{s,t}}$, etc.
It is up to the choice of using different generative models $\mathbb{P}_\theta(s)$.

Optimistic exploration in RL

Counting via errors. We can use the (s, a) pair to estimate some function $f(s, a)$ and then treat the estimation error as the novelty of this new (s, a) pair. As a consequence, the exploration bonus is $\|\hat{f}(s, a) - f(s, a)\|$ which is added to the reward.

- It is very natural to choose $f(s, a)$ to be the state transition model which is to predict the next state s'. This target $f(\cdot)$ has an alternative interpretation of information gain.
- It is also possible to simply predict a function $f(\cdot)$ that is randomly determined at the beginning of the learning.

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Posterior sampling

- Thompson sampling can be extended to reinforcement learning via sampling a Q-function from a distribution of Q-functions at each step.
- The reason we can do this is that Q-learning is off-policyand consequently the experience collected at each round can be used in the future no matter which Q-function is used to generate the experiences.

Posterior sampling

The space of distributions over a function is large and intractable in general.

- Ensemble Q-Learning. Train multiple Q-functions using the same set of data. These Q-functions can share some low-level features in the neural network to reduce the computational cost.
- Distributional Q-Learning. the Q-function estimates a distribution over the action value given a state-action pair instead of estimating its mean. Posterior sampling can then sample one action value of that distribution.

Posterior sampling

An important observation is to use randomized value estimations instead of random policies in posterior sampling:

- As exploring with random actions (e.g., epsilon-greedy) can incur oscillation back and forth and might not go to a coherent or interesting state.
- Exploring with random Q-functions can commit to a randomized but internally consistent strategy for an entire episode.

Information Gain

We first need to reason about which variable we should consider for information gain.

- For the information gain of generative models $IG(\mathbb{P}(s))$, an approximation is to use the predictive gain $\log \mathbb{P}_\theta(s) - \log \mathbb{P}_{\theta'}(s)$. This connects with the count-based model described before.
- For information gain of transition models $IG(\mathbb{P}(s' \mid s, a))$, it can be approximated by the KL-divergence $d_{\mathrm{KL}}(\mathbb{P}(s' \mid \theta) \| \mathbb{P}(s' \mid \theta'))$ between the before-update and the after-update prediction of the world model. This connects with counting methods via errors where d_{KL} can be replaced by other measures as well.

Information Gain

More generally, $IG(\mathbb{P}(s' \mid s, a))$ can be written into $d_{\mathrm{KL}}(\mathbb{P}(\theta \mid \tau, x) \| \mathbb{P}(\theta \mid \tau))$, where τ is the trajectory up to the moment and $x = (s, a, r, s')$ is the new experience collected at this step.

- This term is still intractable, but by decomposing $\mathbb{P}(\theta \mid \tau)$ one can use variational inference to construct a surrogate of the objective.
- Optimizing this surrogate up to some further approximations and tricks constitute variational information maximizing exploration, or VIME.

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Question and Answering (Q&A)

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen