EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 14 - 3/4/2021

Lecture 14: Introduction to Reinforcement Learning

Lecturer: Jiantao Jiao *Scribe(s): Allen Shen, Lingfeng Sun*

In this lecture, we introduce reinforcement learning (RL) and talk about its various formulations. Reinforcement learning is a very complicated topic, and many people do not necessarily agree on its formulations. Furthermore, it is not necessarily the best method to solve easier problems since we can potentially use simpler algorithms to exploit problem structure.

1 Markov Decision Process

1.1 Infinite Horizon Discounted MDPs

Most of the applied RL community focuses on infinite horizon discounted MDPs. Meanwhile, the theory community has been recently focusing on the episodic case since they have trouble solving the infinite horizon case. An infinite horizon discounted MDP is specified by $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$:

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is the space of probability distributions over $\mathcal{S}$ (i.e., non-negative $|\mathcal{S}|$-dim vector summing to 1). This function tells us the probability distribution of the next state given the current state and action.
- Reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. In general, the reward function can also depend on the next state, but we can reduce a more general reward function to this special case. We will view this function as deterministic for now, but in general it can stochastic..
- Discount factor $\gamma \in [0, 1)$. We do not consider the $\gamma = 1$ case because this defeats the purpose of discounting. (Although there exists a formulation called average case MDP with $\gamma = 1$, it is not widely used in the RL literature.)

1.2 Interaction model

Suppose we start at some initial state $s_1 \in \mathcal{S}$ and take action a_1. (These can be generated in various different ways, but we will not spend too much time talking about data generation.) Then, we receive some instantaneous reward $r = R(s_1, a_1)$. (Assume for now that R is deterministic.) We transition to the next state under a probability distribution $s_2 \sim P(s_1, a_1)$. We can also denote this distribution as $P(\cdot|s_1, a_1)$ where the dot represents s_2 or $P(s'|s, a)$ where s' represents the next state. Then, we can take some action a_2, receive reward $R(s_2, a_2)$ and so on. For an infinite horizon MDP, this type of interaction continues forever.

Policy and value

In general, there is some sort of objective or evaluation. A **policy** tell us how to choose actions, and **value** is defined as the expected total reward. (Value is also sometimes known as return, utility, or long-term reward.) We can explore other objectives besides expected reward, but this is an open area of research. Unlike bandits, regret is not clearly defined in MDPs. We can often use results from bandits to better understand results in RL, but we first need a good understanding of bandit literature.

Stochastic rewards

The reward can also be stochastic in a more general setting. According to Sutton and Barto [**?**], a more general formulation is that the reward and the next state follows a joint distribution conditioned on the current state and action: $(r, s')|(s, a)$. The deterministic reward function is a special case of this general formulation. Some RL papers make a simplifying assumption that r and s' are conditionally independent given (s, a). Furthermore, sometimes the action space depends on the current state. We can denote the actions available at a current state s as A_s. This is usually not an issue because we can union all possible actions in the action space.

1.3 Special cases of MDP

We discuss a few more subtleties which are not usually discussed in RL papers to gain a more holistic view of the field.

- Single action: At every state, there is only one single action available, so given any state we know the distribution of the reward and next state. In this case, it reduces to a infinite horizon **Markov Chain** associated with a reward function. (Note that classical Markov chains do not usually have associated rewards.)
- Deterministic transition: s' is completely determined by a (s, a) pair. In this case, the MDP can be interpreted as a **Directed Graph** with actions and rewards. The MDP evolution becomes walk on this graph; each node represents a state, and each action represents an edge on the graph.

MDP is a super complicated subject, and there are lots of open problems even in these reduced cases. As such, if possible we would like to take advantage of problem structure so that we do not have to solve a generic MDP problem.

1.4 Grid-world Example

Let us consider another special case of a MDP. Consider a grid-world such as:

P		
		G

Here, "P" and "G" represent the player position and terminal/goal position respectively. The MDP in this example can be defined as:

- State: grid number/coordinates.
- Actions: North, South, East, West.
- Transition function: Deterministic transition. At the boundaries of the grid, we might have a restricted set of actions. For example, at the top left state, we cannot move north.
- Reward: 0 for goal state, -1 everywhere else.
- Discount factor: We will choose $\gamma = 0.99$, but we could easily choose a different discount factor.
- The terminal state is an "absorbing" state. This means that once we reach the terminal state we stay there forever, and we do not accumulate any more rewards.

Discounting

The following two questions may arise from the previous scenario:

- Why do we do discounting at all?
- Why do you need to run the MDP for an infinite number of time steps instead of a finite number of time steps?

Note that in this infinite horizon MDP, if we run the MDP forever with no discounting ($\gamma = 1$) and we always stay in the bad intermediate states, the total reward would accumulate to $-\infty$. This could break down the equations and the programs in practice. There are several reasons supporting the discounting formulation:

- Grid-world special case: When $\gamma \approx 1$, the discounted reward is still close to the total reward with no discounting if we stop early.
- Mathematical convenience: Infinite horizon and discounting give rise to a stationary optimal policy, which drastically simplifies mathematical equations. Usually not all optimal policies are stationary and can depend on the time step.
- Computational efficiency: In some cases, it is computationally faster to use a smaller γ for planning even when the true γ is large.
- Economic interpretation (i.e. money, mortgage, etc.).

  In some scenarios, we might want to consider augmenting our state space (i.e. to incorporate time).

Effective horizon

It is usually not evident how we should pick γ, although we might have some intuition in some economic settings (i.e. interest). If you believe the MDP should terminate in H (which is sometimes called the effective horizon) steps, then we can set $\gamma : H = \frac{1}{1-\gamma}$. The total return/value is defined by $\mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t]$, and the discounting at horizon H is roughly $\gamma^H = \gamma^{\frac{1}{1-\gamma}} = (1-(1-\gamma))^{\frac{1}{1-\gamma}}$. Note that $\lim_{n\to\infty}(1 - \frac{1}{n})^n = e^{-1}$; in this case we have $n = \frac{1}{1-\gamma}$. This means that when γ is close to 1, we have not discounted enough up until the effective horizon H since γ^H is effectively a constant. At future time steps, discounting destroys the accumulated rewards. Through this formulation, we are able to obtain the benefits of infinite horizon discounted MDPs (i.e. a stationary optimal policy), and the total return is not ruined too much by discounting before the horizon.

Total reward/value

The total return/value is defined by $V = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t]$. The value depends on the initial state. If we assume bounded rewards: $r_t \in [0, R_{\max}]$, it immediately implies that $V \in [0, \frac{R_{\max}}{1-\gamma}]$ or $V \in [0, HR_{\max}]$ where H is the effective horizon. In this sense, we are trying to spread out $HR_{\max}$ rewards over the infinite horizon. (This is more of an intuition rather than a rigorous proof.)

1.5 Stationary Policy

A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ represents a conditional transition probability kernel $\pi(\cdot|s)$ which tells the distribution of actions given state s. It is called stationary since this policy does not depend on the time t. Normally, the MDP evolves by $s_1, a_1 \sim \pi_1(\cdot|s_1), r_1 = R(s_1, a_1), s_2 \sim P(\cdot|s_1, a_1), a_2 \sim \pi_2(\cdot|s_2), ...$ The policy at each step can be different (i.e. $\pi_1(\cdot|s_1)$ does not have to be the same as $\pi_2(\cdot|s_2)$). The nice part of the discounted infinite horizon result is that there exists an optimal stationary policy which does not depend on time and maximizes the total discounted reward. This means that we only need to search for stationary

policies if we want to find an optimal policy; this helps significantly limit the search space. In practice, we can choose π in whatever way we want, and the optimal policy is not necessarily unique.

Now we introduce a more rigorous definition for the value function for an initial state s and policy π:

$$V^\pi(s) \triangleq \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_1 = s, \pi]$$

In this case, the evolution of the MDP follows the steps: $s_1, a_1 \sim \pi(\cdot|s_1), r_1 = R(s_1, a_1), s_2 \sim P(\cdot|s_1, a_1)$, $a_2 \sim \pi(\cdot|s_2), ...$ Here, the whole joint distribution of $(s_1, a_1, r_1, s_2, a_2, r_2, ...)$ is completely determined by s and π.

1.6 Bellman Equation

The Bellman equation is unique to RL and is not present in bandits. In **policy evaluation**, we are given a MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and a deterministic policy π, and we want to know the value $V^\pi(s)$ for every state s. To compute this, we introduce the Bellman equation:

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_1 = s, \pi] && \text{(Definition of } V^\pi(s)\text{)} \\
&= \mathbb{E}[r_1 + \gamma \sum_{t=2}^{\infty} \gamma^{t-2} r_t | s_1 = s, \pi] && \text{(Pulling } r_1 \text{ and } \gamma \text{ out of the summation)} \\
&= R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_1 = s', \pi] \\
&= R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) V^\pi(s') && \text{(Definition of } V^\pi(s')\text{)}
\end{aligned}$$

In the third equality, we changed the index of the summation inside the expectation from t to $t+1$, and we took advantage of conditional independence from the Markov property. For this derivation, we know that $\pi(s) \in \mathcal{A}$ is well defined since we assume that π is a deterministic policy. We can also generalize this derivation for stochastic policies, which would introduce another summation.

Matrix-vector form of the Bellman equation

We can also derive a matrix-vector form of the Bellman equation by defining the following variables:

- V^π as $|\mathcal{S}| \times 1$ dimension vector $[V^\pi(s)]_{s \in \mathcal{S}}$
- R^π as $|\mathcal{S}| \times 1$ dimension vector $[R(s, \pi(s))]_{s \in \mathcal{S}}$
- P^π as $|\mathcal{S}| \times |\mathcal{S}|$ dimension matrix $[P(s'|s, \pi(s))]_{s, s' \in \mathcal{S}}$ (Here, the rows are indexed by s while the columns are indexed by s'.)

This leads to the matrix vector form of Bellman equation (note that the dimensions match up):

$$\underset{|\mathcal{S}| \times 1}{V^\pi} = \underset{|\mathcal{S}| \times 1}{R^\pi} + \gamma \underset{|\mathcal{S}| \times |\mathcal{S}|}{P^\pi} \cdot \underset{|\mathcal{S}| \times 1}{V^\pi}$$

which implies $(I - \gamma P^\pi) V^\pi = R^\pi$ and $V^\pi = (I - \gamma P^\pi)^{-1} R^\pi$. We can show that the matrix $I - \gamma P^\pi$ is always invertible for $\gamma \in [0, 1)$, and we can derive a similar result if π is stochastic.