CS 7545: Machine Learning Theory **Fall 2019**

Lecture 25: Reinforcement Learning

Lecturer: Bhuvesh Kumar *Scribes: Alen Polakof, Yinglun Xu*

25.1 Reinforcement Learning Introduction

Problem Setting: The problem is formulated as an agent, which takes actions in an environment changing its state so as to maximize a cumulative reward. Same as the Stochastic Bandit Setting presented in class before, we have the exploration vs. exploitation trade-off.

Problem Formulation: MDP (Markov Decision Process)

- A model of the environment and the interactions of the environment.
- A set S of states.
- A start state s_0.
- A set A of actions.
- Transition Probability: $Pr(s' \mid s, a) \implies s' = (s, a)$. Probability of taking action a in state s and arriving to the new state s'.
- Reward Probability: $Pr(r' \mid s, a) \implies r' = r(s, a)$. Probability of taking action a in state s and obtaining a reward r'.
- Policy: A policy $\pi : s \to A$ or $\pi(s)$ is a mapping from S to probability distributions over the action space A.

We use MDPs because the problem is only dependent on the current state and action taken, not on its history.

Problem Objective: Maximize the reward, where the reward depends on the policy taken. Hence, the objective is to find the policy that produces the highest cumulative reward.

Examples

1. Agent Mario interacts with environment and gets rewards as coins or superpowers. Actions can be moving, or jumping, which change the environment.
2. Alpha Go. Computer program developed by DeepMind that plays the game of Go.

Type of horizons: In Reinforcement Learning there are two different games:

- Finite Horizon Game:

$$\text{Reward} = \sum_{t=0}^{T} r(s_t, \pi(s_t))$$

- Infinite Horizon Game: Let $\gamma \in (0, 1)$ be the discount factor. In other words, γ is a weighting term that represents the importance of future rewards, and decreases the further we look into the future.

$$\text{Reward} = \sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t))$$

Policy Value: The value $V_\pi(s)$, which is the expected reward if $s_0 = s$ and you follow the policy π. As a result, is an expectation of the reward that will be accumulated starting at state s and following the policy π.

- For finite horizon case: $V_\pi(s) = \mathbb{E}[\sum_{t=0}^{T} r(s_t, \pi(s_t))|s_0 = s]$
- For infinite horizon case: $V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t))|s_0 = s]$

(From now no, only focus on infinite horizon cases)

25.2 Bellman Equation:

Theorem 25.1 *For all s in S, where S is the set of all possible states, and s' is the states that can be reached from s. We define the Bellman Equation as:*

$$V_\pi(s) = \mathbb{E}[r(s, \pi(s))] + \gamma \sum_{s' \in S} Pr(s' \mid s, \pi(s)) V_\pi(s')$$

Proof:

$$\begin{aligned}
V_\pi(s) &= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)|s_0 = s)] \\
&= \mathbb{E}[r(s, \pi(s))] + \gamma \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, \pi(s_{t+1}))|s_0 = s] \\
&= \mathbb{E}[r(s, \pi(s))] + \gamma \mathbb{E}[\sum_{s' \in S} \sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, \pi(s_{t+1}))|s_1 = s', s_0 = s] Pr(s_1 = s'|s_0 = s) \\
&= \mathbb{E}[r(s, \pi(s))] + \gamma \sum_{s' \in S} \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, \pi(s_{t+1}))|s_1 = s'] Pr(s'|s, \pi(s)) \\
&= \mathbb{E}[r(s, \pi(s))] + \gamma \sum_{s' \in S} V_\pi(s') Pr(s'|s, \pi(s))
\end{aligned}$$

■

As a result, this demonstrates that $V = R + \gamma P \cdot V$, where:

$$\begin{aligned}
&P \in \mathbb{R}^{|s|x|s|} && P_{ss'} = Pr(s'|s, \pi(s)) \\
&V \in \mathbb{R}^{|s|} && V_s = V_\pi(s) \\
&R \in \mathbb{R}^{|s|} && R_s = \mathbb{E}[r(s, \pi(s))]
\end{aligned}$$

Theorem 25.2 *For finite MDP, the Bellman Equation has a unique solution. Therefore, $V = (I - \gamma P)^{-1} \cdot R$.*
Proof: *Note that:*

$$\|P\|_\infty = max_s \sum_{s' \in S} |P_{s,s'}| = max_s 1 = 1 \text{ and } (I - \gamma P) \cdot V = R$$

$$\begin{aligned}
&\|\gamma P\|_\infty = \gamma \\
&\implies \textit{All eigenvalues of } \gamma P \leq \gamma \\
&\implies \textit{Eigenvalues of } (I - \gamma P) \geq 1 - \gamma > 0 \qquad\qquad \gamma \in (0,1)
\end{aligned}$$

As a result, since all eigenvalues are greater than 0, the matrix $I - \gamma P$ is invertible, and $V = (I - \gamma P)^{-1} R$ ■

25.3 Optimal Policy and Q-Function

A policy π^* is optimal if for all $s \in S$ and for all possible policies π,

$$V_{\pi^*}(s) \geq V_\pi(s)$$

As a result, the value of the optimal policy must be always greater all equal to the value of all other possible policies.

Definition 25.3 (Q Function)

$$Q_\pi(s,a) = \mathbb{E}[r(s,a)] + \gamma \sum_{s'} P[s' \mid s, \pi(s)] V_\pi(s')$$
$$Q_{\pi^*}(s,a) = \mathbb{E}[r(s,a)] + \gamma \sum_{s'} P[s' \mid s, \pi^*(s)] V_{\pi^*}(s')$$

Bellman optimality condition:

$$V_{\pi^*}(s) = \max_a Q_{\pi^*}(s,a)$$

Two Settings:

- Planning: $P[s' \mid s, \pi(s)]$ and $r(s,a)$ are known
- Learning: Don't know anything

Algorithm 1 Value Iteration

while $\|V - \phi(V)\|_\infty \geq \frac{(1-\gamma)\epsilon}{\gamma}$ **do**
  $V = \phi(V)$
  where $\phi(V)(s) = \max_a \{\mathbb{E}[r(s,a)] + \gamma \sum_{s'} P[s' \mid s,a] V(s')\}$
end while
return $\pi(s) = \text{argmax}_a \{\mathbb{E}[r(s,a)] + \gamma \sum_{s'} P[s' \mid s,a] V(s')\}$

Lemma 25.4 *ϕ is $\gamma -$ Lipschitz in $\|\cdot\|_\infty$, i.e.,*

$$\|\phi(V) - \phi(U)\|_\infty \leq \gamma \|V - U\|_\infty .$$

Proof: $\forall s \in S$, Let $a^*(s) = \text{argmax}_a \{\mathbb{E}[r(s,a)] + \gamma \sum_{s'} P[s' \mid s,a] V(s')\}$. Then

$$\begin{aligned}
\phi(V)(s) - \phi(U)(s) &\leq \phi(V)(s) - [\mathbb{E}[r(s,a^*)] + \gamma \sum_{s'} P[s' \mid s, a^*] V(s')] \\
&= \gamma \sum_{s'} P[s' \mid s, a^*][V(s') - U(s')] \\
&\leq \gamma \sum_{s'} P[s' \mid s, a^*] \|V - U\|_\infty \\
&= \gamma \|V - U\|_\infty
\end{aligned}$$

Similarly, Let $a^{*2}(s) = \text{argmax}_a\{\mathbb{E}[r(s,a)] + \gamma \sum_{s'} P[s^{'} \mid s,a]U(s^{'})\}$, it can be proved that

$$\phi(U)(s) - \phi(V)(s) \leq \gamma \left\|V - U\right\|_\infty .$$

Combine both then

$$\left\|\phi(V) - \phi(U)\right\|_\infty \leq \gamma \left\|V - U\right\|_\infty .$$

■

Lemma 25.5 *$V^* = \phi(V^*)$ where $V^* = V_{\pi^*}$*

Proof: (Exercise) Hint: Bellman Optimality ■

Theorem 25.6 *For any V_0, the sequence $V_{n+1} = \phi(V_n) = \phi(V_n)$ converge to V^*.*

Proof:

$$\left\|V^* - v_{n+1}\right\|_\infty = \left\|\phi(V_* - \phi(V_n))\right\|_\infty \leq \gamma \left\|V_* - V_n\right\|_\infty \leq \gamma^n \left\|V_* - V_1\right\|_\infty$$

So V_n converges to V^* ■

Theorem 25.7 *Value Iteration halts in $O(log\frac{1}{\epsilon})$ steps and return π such that $\left\|V_n - V^*\right\|_\infty \leq \epsilon$*

Proof: $\forall n \in N$,

$$\begin{aligned}
\left\|V^* - V_{n+1}\right\|_\infty &\leq \left\|V^* - \phi(V_{n+1})\right\|_\infty + \left\|\phi(V_{n+1}) - V_{n+1}\right\|_\infty \\
&= \left\|\phi(V^*) - \phi(V_{n+1})\right\|_\infty + \left\|\phi(V_{n+1}) - \phi(V_n)\right\|_\infty \\
&\leq \gamma \left\|V^* - V_{n+1}\right\|_\infty + \gamma \left\|\phi(V_n) - V_n\right\|_\infty
\end{aligned}$$

■

Then

$$\left\|V^* - V_{n+1}\right\|_\infty \leq \frac{\gamma}{1-\gamma} \left\|\phi(V_n) - V_n\right\|_\infty = \epsilon$$

Let n be the largest index such that $\left\|\phi(V_n) - v_n\right\|_\infty \leq \frac{1-\gamma}{\gamma}\epsilon$. Since $\left\|\phi(V_n) - V_n\right\|_\infty \leq \gamma^n \left\|V_1 - V_0\right\|_\infty$, then

$$\frac{1-\gamma}{\gamma}\epsilon \leq \gamma^n \left\|V_1 - V_0\right\|_\infty$$

$$n \leq O(log\frac{1}{\epsilon})$$