Machine Learning **Spring 2021**

Lecture 14: Reinforcement Learning with Known Model

Lecturer: Liwei Wang *Scribe: Jingrui Xing, Qizhi Chen, Weijie Peng, Hao Liu, Yuheng Wu*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

14.1 Reinforcement Learning: Basic Concepts

Simply put, **reinforcement learning** is the process of training models to take a sequence of decisions in an environment that optimize the reward, usually in large-scale problems where dynamic programming is inapplicable. Virtually all RL problems can be formalized as **Markov Decision Process** (MDP).

14.1.1 Markov Decision Process

Definition 14.1 *A Markov Decision Process is a 4-tuple (S, A, P, R), where*

- *S is the set of states;*
- *$A(s)$ is the set of available actions from state $s \in S$;*
- *$P(s'|s,a) = \mathbb{P}(S_{t+1} = s'|S_t = s, a_t = a)$ is the transition probability that, in time t, the action $a \in A$ in state s would lead to state s' in time $t+1$;*
- *$R(s,a)$ is the* ***expected*** *reward received by taking action a at state s;*

In the case when the decision-making process has infinite horizon, this tuple also includes a discouting factor $\gamma \in (0,1)$. For finite horizon, $\gamma = 1$.

The goal of an MDP agent is to maximize the **long term reward**, defined as $R \triangleq R(s_{t+1}, a_{t+1}) + \gamma R(s_{t+2}, a_{t+2}) + \gamma^2 R(s_{t+3}, a_{t+3}) + \cdots$. Yet, in most cases, instead of directly dealing with discrete set of actions, we are dealing with policies.

A **policy** is defined as $\pi : S \to \Delta(A)$, where $\pi(s)$ gives the probability vector of choosing actions in A at state s. For short, we write $\pi(s)$ for a random variable of taking actions with the corresponding probabilities.

There are two tasks in an MDP control problem: how to evaluate a given policy $\pi : S \to A$; how to find the optimal policy π^*. The first task involves two **value functions**. The second task would be discussed in the following chapters.

14.1.2 State Value Function

Definition 14.2 *For a given policy π, at a certain state s, the state value function is defined as*

$$v^\pi(s) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) \Big| s_0 = s\right]$$

This is the expected return when starting from s and following policy π. The randomness comes from the random action distribution and the transition probability.

14.1.3 State-Action Value Function

Definition 14.3 *For a certain state s, a feasible action a in state s, and given policy π, the state-action value function is defined as*

$$Q^\pi(s,a) = \mathbb{E}\left[R(s,a) + \sum_{t=1}^{\infty} \gamma^t R(s_t, \pi(s_t)) \Big| s_0 = s, a_0 = a\right]$$

This is the expected return when starting from s, taking action a, then following policy π.

Note that a does not necessarily follows π. Moreover, $Q^\pi(s,a) = \mathbb{E}_{s_1}\left[R(s,a) + \gamma V^\pi(s_1, a_1) \big| s_0 = s, a_0 = a\right]$. Therefore, if a follows π, then $\mathbb{E}[Q^\pi(s,a)] = V^\pi(s)$.

14.1.4 Optimal Policy

With the definition of the two value functions, we now give a formal definition of the optimal policy.

Definition 14.4 *A optimal policy π^* satisfies: for any policy π and any state s, it holds $V^{\pi^*}(s) \geq V^\pi(s)$.*

In the following chapters, we will discuss the existence of optimal policy and its computation.

14.2 Bellman Expectation Equation and Policy Evaluation

14.2.1 Bellman Expectation Equation

From the definition of the value function, we can generally see that it can be expressed in the form of a transition function. More specifically,

$$\begin{aligned}
v^\pi(s) &= \mathrm{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \mid s_t = s\right] && (14.1)\\
&= \mathrm{E}\left[R_t + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \mid s_t = s\right] && (14.2)\\
&= \mathrm{E}\left[R_t \mid s_t = s\right] + \gamma \mathrm{E}_{s_{t+1} \sim P(s,\pi(s))}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k}\right] && (14.3)\\
&= R(s, \pi(s)) + \gamma \sum_{s'} Pr[s_{t+1} = s' | s_t = s, a_t = \pi(s)] \cdot v^\pi(s') && (14.4)
\end{aligned}$$

where R_t represents reward at time t (which depends on the state and action at time t), and $R(s,a)$ represents the expectation of reward on state s and action a. Thus, we have

$$v^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s' | s, \pi(s)) \cdot v^\pi(s'), \forall s \in S \tag{14.5}$$

which is known as **Bellman Expectation Equation**.

14.2.2 Evaluating the value function

Firstly, we define an operator according to Bellman Expectation Equation.

Definition 14.5 (Bellman Expectation Operator) *The Bellman Expectation Operator Φ is defined as*

$$\begin{aligned} \Phi : \mathbb{R}^N &\to \mathbb{R}^N \\ \mathbf{v} &\mapsto \mathbf{v}' \end{aligned}$$

such that

$$\mathbf{v}'(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) \cdot \mathbf{v}(s'), \forall s \in S \tag{14.6}$$

where $N = |S|$.

To further discuss the property of Φ, we introduce what a γ-contraction mapping is.

Definition 14.6 (α-contraction mapping) *Function $\phi : X \to X$ is a α-contraction mapping w.r.t. $||\cdot||_p$, if $\forall u, v \in X$,*

$$||\phi(u), \phi(v)||_p < \alpha ||u, v||_p \tag{14.7}$$

where $\alpha \in [0, 1)$

Theorem 14.7 (Banach Fixed Point Theorem) *Every contraction mapping has a unique fixed point.*

Now we point out that Φ is a γ-contraction mapping w.r.t. $||\cdot||_\infty$.

Theorem 14.8 *Bellman Expectation Operator Φ is a γ-contracion mapping w.r.t. $||\cdot||_\infty$.*

Proof: $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, let $\mathbf{u}' = \Phi(\mathbf{u}), \mathbf{v}' = \Phi(\mathbf{v})$. We have $\mathbf{u}' - \mathbf{v}' = \gamma \sum_{s'} P(s'|s, \pi(s)) \cdot (\mathbf{u}(s') - \mathbf{v}(s')) \leq \gamma \sum_{s'} P(s'|s, \pi(s)) \cdot ||\mathbf{u} - \mathbf{v}||_\infty = \gamma ||\mathbf{u} - \mathbf{v}||_\infty$. ■

Now we know that Φ converge to a unique point. Thus, we can use the following algorithm to evaluate the value function.

Algorithm 1 Solving Bellman Expectation Equation via Iteration

```
Output: The value function vector $\mathbf{v}^\pi$ w.r.t. a given policy $\pi$
  Stochastically initialize $\mathbf{v}_0$
  $k \leftarrow 0$
  repeat
      $\mathbf{v}_{k+1} \leftarrow \Phi(\mathbf{v}_k)$
      $k \leftarrow k + 1$
  until Convergence. Denote the result as $\mathbf{v}^\pi$
```

14.2.3 Greedy Method

Now given a policy π, we can get $\mathbf{v}^\pi$ through the above algorithm. Then in order to get an improved policy π', we employ a greedy method:

$$\forall s \in S, \pi'(s) = \underset{a}{\operatorname{argmax}} \left[R(S,a) + \sum_{s'} P\left(s' \mid s,a\right) \mathbf{v}^\pi\left(s'\right) \right]$$

Theorem 14.9 $\mathbf{v}^\pi(s) \leq \mathbf{v}^{\pi'}(s)$ *for all* s

Proof:

$$\begin{aligned} \mathbf{v}^\pi(s) &= R(s,\pi(s)) + \gamma \sum_{s'} P\left(s' \mid s,\pi(s)\right) \mathbf{v}^\pi\left(s'\right) \\ &\leq R\left(s,\pi'(s)\right) + \gamma \sum_{s'} P\left(s' \mid s,\pi'(s)\right) \mathbf{v}^\pi\left(s'\right) \end{aligned}$$

Written in matrix form as follows:

$$\mathbf{R}_{\pi'} + \gamma \mathbf{P}_{\pi'} \mathbf{v}^\pi \geq \mathbf{R}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}^\pi = \mathbf{v}^\pi$$

$$\mathbf{R}_{\pi'} \geq \left(\mathbf{I} - \gamma \mathbf{P}_{\pi'}\right) \mathbf{v}^\pi$$

And we have:

$$\|\mathbf{P}\|_\infty = \max_s \sum_{s'} \left|\mathbf{P}_{ss'}\right| = \max_s \sum_{s'} \mathbb{P}\left[s' \mid s,\pi(s)\right] = 1$$

Then $\|\gamma \mathbf{P}\|_\infty = \gamma < 1$. Since the radius of convergence of the power series $(1-x)^{-1}$ is 1, we can use its expansion and write

$$\left(\mathbf{I} - \gamma \mathbf{P}_{\pi'}\right)^{-1} = \sum_{k=0}^{\infty} \left(\gamma \mathbf{P}_{\pi'}\right)^k .$$

Thus, if $\mathbf{Z} = (\mathbf{Y} - \mathbf{X}) \geq \mathbf{0}$, then $\left(\mathbf{I} - \gamma \mathbf{P}_{\pi_{n+1}}\right)^{-1} \mathbf{Z} = \sum_{k=0}^{\infty} \left(\gamma \mathbf{P}_{\pi_{n+1}}\right)^k \mathbf{Z} \geq \mathbf{0}$, which means $\left(\mathbf{I} - \gamma \mathbf{P}_{\pi_{n+1}}\right)^{-1}$ preserves ordering. Then $\mathbf{v}^{\pi'} = \left(\mathbf{I} - \gamma \mathbf{P}_{\pi_{n+1}}\right)^{-1} \mathbf{R}_{\pi_{n+1}} \geq \mathbf{v}^\pi$ ■

Now our goal is to find out if there exists an unique optimal policy π^*, which has the property that: $\forall s, \mathbf{v}^{\pi^*}(s) \geq \mathbf{v}^\pi(s)$

14.3 Finding the Optimal Policy

In the above section we've discussed about the value function. Now, let's turn to the problem of finding the optimal policy.

We've already known the definition of an optimal policy. Now we assume such policy exists, denoted as π^*.

Definition 14.10 (Policy Iteration) *Using the Greedy Method above, we can obtain a method of iteration: start from some policy* π_0*, and compute* v^{π_0}*, get the "greedy policy" of* π_0 *(denoted as* π_1 *), and compute* V^{π_1}*; ... We know the value function of* π_k *should increase during the iteration. This approach is called policy iteration, the value function of* π_k *will converge to* v^**.*

And based on Bellman Expectation Equation, we may guess that,

$$v^{\pi^*}(s) = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) v^{\pi^*}(s') \right], \forall s \in S \tag{14.8}$$

This is called the **Bellman Optimal Equation**. Similarly, we define an op operator according to this equation.

Definition 14.11 (Bellman Operator) *The Bellman Operator* Φ^* *is defined as*

$$\Phi^* : \mathbb{R}^N \to \mathbb{R}^N$$
$$\mathbf{v} \mapsto \mathbf{v}'$$

such that

$$\mathbf{v}'(s) = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) \mathbf{v}(s') \right], \forall s \in S \tag{14.9}$$

where $N = |S|$.

We point out that Φ^* is also a contraction mapping.

Theorem 14.12 *Bellman Operator* Φ^* *is a* γ*-contraction mapping w.r.t.* $||\cdot||_\infty$.

Proof: $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, let $\mathbf{u}' = \Phi^*(\mathbf{u}), \mathbf{v}' = \Phi^*(\mathbf{v})$. For a state $s \in S$, let

$$a = \operatorname*{argmax}_b \left[R(s,b) + \gamma \sum_{s'} P(s'|s,b) \mathbf{v}(s') \right]$$
$$a' = \operatorname*{argmax}_b \left[R(s,b) + \gamma \sum_{s'} P(s'|s,b) \mathbf{u}(s') \right]$$

We have

$$\begin{aligned}
\mathbf{v}'(s) - \mathbf{u}'(s) &= \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) \mathbf{v}(s') \right] - \left[R(s,a') + \gamma \sum_{s'} P(s'|s,a') \mathbf{u}(s') \right] \\
&\leq \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) \mathbf{v}(s') \right] - \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) \mathbf{u}(s') \right] \\
&= \gamma \sum_{s'} P(s'|s,a)(\mathbf{v}(s') - \mathbf{u}(s')) \\
&\leq \gamma \sum_{s'} P(s'|s,a) ||\mathbf{v} - \mathbf{u}||_\infty \\
&= \gamma ||\mathbf{v} - \mathbf{u}||_\infty
\end{aligned}$$

Similarly we show that $\mathbf{u}'(s) - \mathbf{v}'(s) \leq \gamma ||\mathbf{v} - \mathbf{u}||_\infty$. Thus, we have $||\mathbf{v}' - \mathbf{u}'||_\infty \leq \gamma ||\mathbf{v} - \mathbf{u}||_\infty$. ■

Now we know that Φ^* converge to a unique fixed point. So similar to Algorithm 1, we can use a iteration approach to compute the unique fixed point v^* of Φ^*

We have following results.

Algorithm 2 Value Iteration

```
Output: The unique fixed point $\mathbf{v}^*$ of $\Phi^*$
  Stochastically choose a policy $\pi_0$ ,let $\mathbf{v}_0$ be its value function
  $k \leftarrow 0$
  repeat
      $\mathbf{v}_{k+1} \leftarrow \Phi^*(\mathbf{v}_k)$
      $k \leftarrow k+1$
  until Convergence. Denote the result as $\mathbf{v}^*$
```

Theorem 14.13 *Let π_0 be an arbitrary policy and $v_0 = v^{\pi_0}$. Then $(\Phi^*(v_0))(s) \geq v_0(s), \forall s \in S$*

Proof: Note that for all $s \in S$,

$$v_0(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} P(s'|s, \pi(s))v_0(s')$$

by the Bellman Expectation Equation, and

$$(\Phi^*(v_0))(s) = \max_{a \in A}[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)v_0(s')]$$

by the definition of Φ^*.

Comparing the right sides immediately yields $(\Phi^*(v_0))(s) \geq v_0(s)$ ■

Theorem 14.14 *Let π_0 be an arbitrary policy and $v_0 = v^{\pi_0}$ and $v_{k+1} = \Phi^*(v_k)$.Then $v_{k+1}(s) \geq v_k(s), \forall s \in S$*

Proof: we can prove this by induction since we already have $(\Phi^*(v_0))(s) \geq v_0(s), \forall s \in S$

suppose we have $v_k(s) \geq v_{k-1}(s), \forall s \in S$

$$v_{k+1} = (\Phi^*(v_k))(s) = \max_{a \in A}[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)v_k(s')]$$

$$v_k = (\Phi^*(v_{k-1}))(s) = \max_{a \in A}[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)v_{k-1}(s')]$$

Comparing the right sides and it's trivial that $v_{k+1}(s) \geq v_k(s)$ ■

And the following fact is trivial.

Theorem 14.15 *Let π_0 be an arbitrary policy and $v_0 = v^{\pi_0}$ and $v_{k+1} = \Phi^*(v_k)$. Then v_k converges to v^* which satisfies the Bellman Optimal Equation.*

Theorem 14.16 *Let v^* is the fixed point of Φ^*. Then there exists a policy π^* whose value function is v^*.*

Proof: let $\pi^*(s) := \operatorname{argmax}_a[R(s,a) + \gamma \sum_{s'} P(s'|s,a)v^*(s')]$

$$
\begin{aligned}
v_{\pi^*}(s) &= R(s,\pi^*(s)) + \gamma \sum_{s'} P(s'|s,\pi^*(s)) v_{\pi^*}(s') \\
&= R(s,\pi^*(s)) + \gamma \sum_{s'} P(s'|s,\pi^*(s))[R(s',\pi^*(s')) + \gamma \sum_{s''} P(s''|s',\pi^*(s')) v_{\pi^*}(s'')] \\
&= \cdots \\
&= v^*(s)
\end{aligned}
$$

■

Now we have proved the existence of π^* and gives a way to calculate it.And finally we have:

Theorem 14.17 *Let v^* be the value function of the optimal policy, assume $||v_k - v^*||_\infty \leq \delta$. Let π_k be the greedy policy w.r.t. v_k, then $||v^{\pi_k} - v^*||_\infty \leq \frac{\gamma}{1-\gamma}\delta$.*

Proof: According to the assumption $||v_k - v^*||_\infty \leq \delta$ and that the greedy strategy is to choose $\pi_k(s) := \text{argmax}_a[R(s,a) + \gamma \sum_{s'} P(s'|s,a) v_k(s')]$, so we know that if we choose a wrong action due to the $\delta-$difference at state s, time t, and follow the best strategy since then, we would only be $\gamma\delta$ less than the best strategy in expectation.

By the same argument, if we make two mistakes successively, then our expectation value drops at most $\gamma\delta + \gamma^2\delta$

So we get $||v^{\pi_k} - v^*||_\infty \leq (\gamma + \gamma^2 + \cdots)\delta = \frac{\gamma}{1-\gamma}\delta$ ■

References