EE625A: Introduction to Reinforcement Learning

# Lecture 10: Bellman Expectation equations and Optimal Policy

13-02-2023

*Lecturer: Prof. Subrahmanya Swamy Peruru* *Scribe: Amit Kumar Yadav*

In Lecture 9 we learnt about Markov decision process, value function and action value function. Towards the end of the class we learnt about Bellman expectation equations.
Today we will prove the uniqueness of the solution of the Bellman expectation equation and we will define the optimal policy and its existence.

## 1 Fixing a policy

Consider a coin toss Markov decision process where head leads to state S+1 if we are in state S and tail leads to state S-1. Now we have two coins blue and red, blue lands in head with probability $p_{blue}$ and red lands in head with probability $p_{red}$. Below is the pictorial view of the MDP.

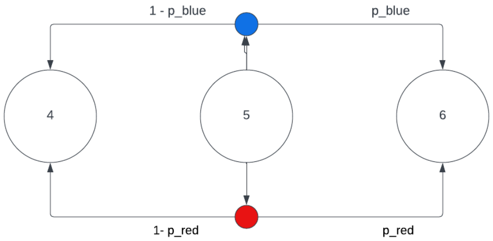

Figure 1: MDP

Any deterministic policy were for each state we have decided the coin we are going to toss the MDP is converted in markov chain. for example if we have decided for state 4 we are going to toss red coin then $p_{45} = p_{red}$ and $p_{43} = 1 - p_{red}$. Similarly for eacch state we can define the transition probability and rewards.

For a stochastic policy if we fix the policy i.e., we know the probability by which we are going to take an action for each state we can find out the transition probability for each state as follows.

$$P^{\pi}_{ss'} = P^{\pi}(s'|s) = \sum_a P(s'|s,a)\pi(a|s) \qquad Transition\,probabilities \tag{1}$$

$$R^a_s = E[R_{t+1}|s_t = s, A_t = a] = \int_r rP(r|s,a)$$

$$R^{\pi}_s = E_{\pi}[R_{t+1}|s_t = s] = \sum_a E[R_{t+1}|s_t, A_t = a]\pi(a|s) \qquad Reward\,for\,policy\,\pi \tag{2}$$

# 2 Vector form of Bellman expectation equation

Let us use $R^{\pi}_{s}$, $P^{\pi}_{ss'}$ notation for vector form.

$$\begin{aligned}
&V_{\pi}(s) = E_{\pi}[G_t|s_t = s] = E_{\pi}[R_{t+1} + \gamma G_{t+1}|s_t = s]\\
&V_{\pi}(s) = E_{\pi}[R_{t+1}|s_t = s] + \gamma E_{\pi}[G_{t+1}|s_t = s]\\
&V_{\pi}(s) = R^{\pi}_{s} + \gamma \sum_{s'} P^{\pi}_{ss'} E_{\pi}[G_{t+1}|s_t = s, s_{t+1} = s']\\
&V_{\pi}(s) = R^{\pi}_{s} + \gamma \sum_{s'} P^{\pi}_{ss'} V_{\pi}(s')
\end{aligned}$$

$$V_{\pi} = \begin{bmatrix} V_{\pi}(s_1) \\ V_{\pi}(s_2) \\ \vdots \\ V_{\pi}(s_n) \end{bmatrix}_{nX1}$$

$$R^{\pi} = \begin{bmatrix} \vdots \\ \vdots \end{bmatrix}_{nX1}$$

$$P^{\pi} = \begin{bmatrix} \cdots \\ \vdots \\ \cdots \end{bmatrix}_{nXn}$$

so the vector form of Bellmann expectation equation is:

$$V_{\pi} = R_{\pi} + \gamma P_{\pi} V_{\pi} \tag{3}$$

# 3 Uniqueness of $V_{\pi}$

If we have 'n' number of states, we have 'n' unknowns: $[V_{\pi}(s)]_{s \in S}$. We have 'n' linear equations. Lets write it in $Ax = b$ form.

$$\begin{aligned}
&V_{\pi} = R_{\pi} + \gamma P_{\pi} V_{\pi}\\
&V_{\pi} - \gamma P_{\pi} V_{\pi} = R_{\pi}\\
&(I - \gamma P_{\pi} V_{\pi}) V_{\pi} = R_{\pi} && (4)\\
&V_{\pi} = (I - \gamma P_{\pi} V_{\pi})^{-1} R_{\pi} && (5)
\end{aligned}$$

Equation 4 is of the form $Ax = b$ where A = $(I - \gamma P_{\pi} V_{\pi})$ and b = $R_{\pi}$. $V_{\pi}$ will have unique solution of A is invertible and A is invertible iff $\lambda \neq 0$ for all eigen values.
We know that $P_{\pi}$ is transition probability matrix where sum of a row is equal to 1 and every element is non negative, this makes $P_{\pi}$ a stochastic matrix.

For a stochastic matrix $\lambda_{max} \leq 1$
Let $x$ is a eigen vector of $P_\pi$, then consider

$$\begin{aligned} A &= (I - \gamma P_\pi V_\pi)x = Ix - \gamma P_\pi V_\pi x \\ A &= x - \gamma \lambda_{p^\pi} x \\ &= (1 - \gamma \lambda_{p^\pi})x \qquad (1 - \gamma \lambda_{p^\pi}) \geq 0 \; as \; \gamma \leq 1 \; and \; \lambda_{p^\pi} \leq 1 \end{aligned} \tag{6}$$

From equation 6 we can see that the eigen value of the matrix A is always positive, hence making the matrix invertible, therefore $V_\pi$ is **unique**

# 4 Discussion on optimal policy

For a given state the optimal policy would the the policy for which the expected reward will be maximum.
Hence for a given state s the optimum policy will be $\underset{x}{\operatorname{argmax}} V_\pi$

## 4.1 Finding shortest path in grid to a Goal G

Consider a grid with one of the cell as goal state. we want to reach the goal cell from any cell with least steps possible. lets formulate this problem as MPD.

- State = Cell
- Action = Up/Down/Left/Right
- The state transition will be deterministic.
- Reward of -1 in each state for any action.
- zero reward in terminal state i.e., Goal state.

For state $S$ the following is an optimal policy.

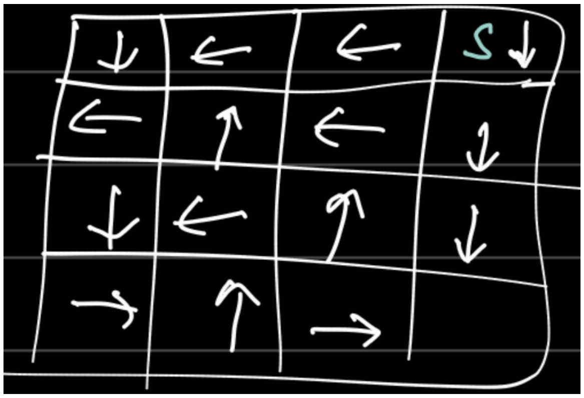

Figure 2: Optimal Policy

- There can be many optimal policy for the state S for example for State (0,0) we can change its action to any other action we will still end up with optimal policy for state S.

- Conceptually we can do the following to find the best policy.
  - Each state 'S' may have multiple policies which are optimal for that state.
  - collect the set of optimal policies for a state s. Repeat this for each state
  - Take intersection of all those sets
  - if intersection is non empty then the policy is the optimal policy.
  - only if that intersection is non empty the optimal policy for whole MDP exists.

## 4.2 Sufficient conditions for existance of optimal policy

- Finite number of states

- Finite action space

- Bounded reward.