OPTIMAL POLICY FROM OPTIMAL VALUE FUNCTION

ASHWIN RAO (STANFORD CME 241)

Let us start with the definitions of Optimal Value Function and Optimal Policy (that we covered in the class on Markov Decision Processes).

$$\text{Optimal State Value Function } V_*(s) = \max_{\pi} V_\pi(s) \text{ for all states } s \in \mathcal{S}$$

$$\text{Optimal Action-Value Function } Q_*(s,a) = \max_{\pi} Q_\pi(s,a) \text{ for all states } s \in \mathcal{S}, \text{ for all actions } a \in \mathcal{A}$$

$$\pi_* \text{ is an Optimal Policy if } V_{\pi_*}(s) \geq V_\pi(s) \textbf{ for all policies } \pi \text{ and } \textbf{for all states } s \in \mathcal{S}$$

Let us go beyond these formal definitions and develop an intuitive (and deeper) understanding of the above definitions. The definition of V_* says that for each state $s \in \mathcal{S}$, we go through all policies π and pick out the policy that maximizes $V_\pi(s)$. Because this maximization is done independently for each state $s \in \mathcal{S}$, presumably we could end up with different policies π that maximize $V_\pi(s)$ for different states. The definition of Optimal Policy π_* says that it is a policy that is "better than or equal to" (on the V_π metric) all other policies **for all** states (note that there could be multiple Optimal Policies). So the natural question to ask is whether there exists an Optimal Policy π_* that maximizes $V_\pi(s)$ **for all** states $s \in \mathcal{S}$, i.e., $V_*(s) = V_{\pi_*}(s)$ for all $s \in \mathcal{S}$. On the face of it, this seems like a strong statement. However, this answers in the affirmative. In fact,

Theorem 1. *For any Markov Decision Process*

- *There exists an Optimal Policy π_*, i.e., there exists a Policy π_* such that $V_{\pi_*}(s) \geq V_\pi(s)$ for all policies π and for all states $s \in \mathcal{S}$*
- *All Optimal Policies achieve the Optimal Value Function, i.e. $V_{\pi_*}(s) = V_*(s)$ for all $s \in \mathcal{S}$, for all Optimal Policies π_**
- *All Optimal Policies achieve the Optimal Action-Value Function, i.e. $Q_{\pi_*}(s,a) = Q_*(s,a)$ for all $s \in \mathcal{S}$, for all $a \in \mathcal{A}$, for all Optimal Policies π_**

Proof. First we establish a simple Lemma.

Lemma 1. *For any two Optimal Policies π_1 and π_2, $V_{\pi_1}(s) = V_{\pi_2}(s)$ for all $s \in \mathcal{S}$*

Proof. Since π_1 is an Optimal Policy, from Optimal Policy definition, we have: $V_{\pi_1}(s) \geq V_{\pi_2}(s)$ for all $s \in \mathcal{S}$. Likewise, since π_2 is an Optimal Policy, from Optimal Policy definition, we have: $V_{\pi_2}(s) \geq V_{\pi_1}(s)$ for all $s \in \mathcal{S}$. This implies: $V_{\pi_1}(s) = V_{\pi_2}(s)$ for all $s \in \mathcal{S}$ □

As a consequence of this Lemma, all we need to do to prove the theorem is to establish an Optimal Policy π_* that achieves the Optimal Value Function and the Optimal Action-Value Function. Consider the following Deterministic Policy (as a candidate Optimal Policy) $\pi_* : \mathcal{S} \rightarrow \mathcal{A}$:

$$\pi_*(s) = \arg\max_{a \in \mathcal{A}} Q_*(s, a) \text{ for all } s \in \mathcal{S}$$

First we show that π_* achieves the Optimal Value Function. Since $\pi_*(s) = \arg\max_{a \in \mathcal{A}} Q_*(s, a)$ and $V_*(s) = \max_{a \in \mathcal{A}} Q_*(s, a)$ for all $s \in \mathcal{S}$, π_* prescribes the optimal action for each state (that produces the Optimal Value Function V_*). Hence, following policy π_* in each state will generate the same Value Function as the Optimal Value Function. In other words, $V_{\pi_*}(s) = V_*(s)$ for all $s \in \mathcal{S}$. Likewise, we can argue that: $Q_{\pi_*}(s, a) = Q_*(s, a)$ for all $s \in \mathcal{S}$ and for all $a \in \mathcal{A}$.

Finally, we prove by contradiction that π_* is an Optimal Policy. So assume π_* is not an Optimal Policy. Then there exists a policy π and a state $s \in \mathcal{S}$ such that $V_\pi(s) > V_{\pi_*}(s)$. Since $V_{\pi_*}(s) = V_*(s)$, we have: $V_\pi(s) > V_*(s)$ which contradicts the definition of $V_*(s) = \max_\pi V_\pi(s)$

$\square$