Lecture 3 - *Bellman Optimality Equations*

OBJECTIVE: In this lecture we will derive Bellman's fixed point equations. This is an efficient stochastic dynamic programming strategy for computing the value of states in fully observed MDPs. We will also introduce Bellman operators and show t. This is the basic idea that will be needed to prove contractions and hence derive optimal algorithms in the following lecture.

◇ VALUE FUNCTIONS

The **value function** of an MDP is the expectation under a policy π of the future rewards. We deal with the undiscounted finite horizon case first. When there are N decision

steps to go, the value function is:

$$V_N^\pi(\mathbf{x}_0) = \mathbb{E}\left[\sum_{t=0}^{N-1} r_t(\mathbf{x}_t, \mathbf{a}_t, \mathbf{x}_{t+1}) + r_N(\mathbf{x}_N)\right]$$

This expression assumes that the initial state is known to be $\mathbf{x}_0$. That is, the expectation when dealing with randomized policies is take with respect to the distribution:

$$p(\mathbf{a}_0|\mathbf{x}_0)p(\mathbf{x}_1|\mathbf{x}_0, \mathbf{a}_0)p(\mathbf{a}_1|\mathbf{x}_1)\cdots p(\mathbf{x}_N|\mathbf{x}_{N-1}, \mathbf{a}_{N-1})$$

.

Theorem 1 *The value function V_N^π satisfies the following fixed point recursion:*

$$V_N^\pi(\mathbf{x}_0) = \sum_{\mathbf{a}_0} p(\mathbf{a}_0|\mathbf{x}_0) \sum_{\mathbf{x}_1} p_1(\mathbf{x}_1|\mathbf{x}_0, \mathbf{a}_0)\left[r_1(\mathbf{x}_0, \mathbf{a}_0, \mathbf{x}_1) + V_{N-1}^\pi(\mathbf{x}_1)\right]$$

⋆ Proof:

If the policy is deterministic, the fixed point equation simplifies to:

$$V_N^{\pi}(\mathbf{x}_0) = \sum_{\mathbf{x}_1} p_1(\mathbf{x}_1|\mathbf{x}_0, \mathbf{a}_0) \left[r_1(\mathbf{x}_0, \mathbf{a}_0, \mathbf{x}_1) + V_{N-1}^{\pi}(\mathbf{x}_1) \right]$$

A further simplification is obtained when the rewards depend only on the present state (or equivalently when their expectation is taken):

$$V_N^{\pi}(\mathbf{x}_0) = r_0(\mathbf{x}_0, \mathbf{a}_0) + \sum_{\mathbf{x}_1} p_1(\mathbf{x}_1|\mathbf{x}_0, \mathbf{a}_0) V_{N-1}^{\pi}(\mathbf{x}_1)$$

Another quantity often mentioned in the literature is the ongoing value of the state at step t. We define it as follows for deterministic policies:

$$U_t^\pi(\mathbf{x}_t) = \mathbb{E}\left[\sum_{k=t}^{N-1} r_k(\mathbf{x}_k, \mathbf{a}_k) + r_N(\mathbf{x}_N)\right]$$

where the expectation is taken with respect to:

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t)p(\mathbf{x}_{t+2}|\mathbf{x}_{t+1}, \mathbf{a}_{t+1}) \cdots p(\mathbf{x}_N|\mathbf{x}_{N-1}, \mathbf{a}_{N-1})$$

As before, we have a recursion for this value function:

$$U_t^\pi(\mathbf{x}_t) = r_t(\mathbf{x}_t, \mathbf{a}_t) + \sum_{\mathbf{x}_{t+1}} p_{t+1}(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t) U_{t+1}^\pi(\mathbf{x}_{t+1})$$

⋆ Proof:

When dealing with infinite discounted rewards:

$$V^\pi(\mathbf{x}_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t(\mathbf{x}_t, \mathbf{a}_t, \mathbf{x}_{t+1})\right],$$

we have the following recursion:

$$V^\pi(\mathbf{x}_0) = \sum_{\mathbf{x}_1} p_1(\mathbf{x}_1|\mathbf{x}_0, \mathbf{a}_0)\left[r_1(\mathbf{x}_0, \mathbf{a}_0, \mathbf{x}_1) + \gamma V^\pi(\mathbf{x}_1)\right]$$

⋆ Proof:

◇ BELLMAN'S PRINCIPLE OF OPTIMALITY

In this section, we will show how to compute the optimal value function given that we are at state $\mathbf{x}_0$:

$$V^\star(\mathbf{x}_0) = \max_\pi V^\pi(\mathbf{x}_0) \qquad V_N^\star(\mathbf{x}_0) = \max_\pi V_N^\pi(\mathbf{x}_0)$$

We will derive an expression for the optimal policy $\boldsymbol{\pi}^\star$.

We address the finite horizon case first.

Assume that $N = 0$ is the final stage. Then the optimal value function is:

$$V_0^\star(\mathbf{x}_N) = r_N(\mathbf{x}_N)$$

Now assume that there is one step to go $N = 1$. The optimal value function and policy are:

⋆ Proof:

In general we have:

$$V_N^\star(\mathbf{x}_0) = \max_{\mathbf{a}} \sum_{\mathbf{x}_1} p_1(\mathbf{x}_1|\mathbf{x}_0, \mathbf{a}_0) \left[r_1(\mathbf{x}_0, \mathbf{a}_0, \mathbf{x}_1) + V_{N-1}^\star(\mathbf{x}_1) \right]$$

In the infinite horizon case, we drop the sub-index N. The optimal decision is then:

$$\mathbf{d}_0^\star(\mathbf{x}_0) = \arg\max_{\mathbf{a}} \sum_{\mathbf{x}_1} p_1(\mathbf{x}_1|\mathbf{x}_0, \mathbf{a}_0) \left[r_1(\mathbf{x}_0, \mathbf{a}_0, \mathbf{x}_1) + V_{N-1}^\star(\mathbf{x}_1)\right]$$

To establish this result as well as set the theoretical ground for designing algorithms in the next lecture, we need to introduce compact notation. For simplicity, we will assume that the state space is discrete so that the value function is simply a vector $V = (V(1), \ldots, V(n)) \in \mathbb{R}^n$, where n is the number of states.

The vector TV denotes the application of the Bellman operator on the value vector V:

$$TV(i) = \max_a \sum_{j=1}^n p(j|i,a)\,[r(i,a,j) + V(j)]$$

We also define the operator given a policy d

$$T_d V(i) = \sum_{j=1}^n p(j|i,d(i))\,[r(i,d(i),j) + V(j)]$$

These operations allow us to write Bellman's recursion as a friendly linear system:

$$T_d V = r_d + P_d V$$

The Bellman operator can be applied recursively:

$$T^t V = T(T^{t-1} V) \qquad T^0 V = V$$

Likewise, for a t-stage policy $\pi = (d_0, d_1, \ldots, d_{t-1})$, we have:

$$T_{d_0} T_{d_1} \cdots T_{d_{t-1}} V = T_{d_0}\left[T_{d_1}(\cdots T_{d_{k-1}V})\right]$$

Lemma 1 Monotonicity*: For any n-dimensional vectors V and $\overline{V}$, such that $V(i) \leq \overline{V}(i)$ for $i = 1 : n$, and a* *stationary policy d, we have:*

$$T^t V(i) \leq T^t \overline{V}(i)$$

$$T_d^t V(i) \leq T_d^t \overline{V}(i)$$

for all i and t.

As a final notation point, let $V \leq \overline{V}$ be true when $V(i) \leq \overline{V}(i)$ for all i. Then we have the following theorem (there exist many stronger versions of this result):

Theorem 2 *Assume that the terminal state can be reached under a stationary policy and that $\sum_j p(j|a,i) r(i,a,j) \leq 0$, the optimal value function $V^\star$ is finite and satisfies:*

$$V^\star = T V^\star$$

Furthermore, $V^\star$ is the only solution to the equation $V = TV$ and $T_{d^\star} V^\star = T V^\star$ implies the optimality of $d^\star$.

$\star$ Proof:

⋆ Proof: