1 Finite horizon problem

Given a Markov Decision Process, with state space $S$, action space $A$, transition probability $P_a(x, y)$, cost function $g_a(x)$ and a finite horizon $T$, we need to find policy $u(x, t): S \times R \rightarrow A$ to solve the following optimization problem:

$$\min_{u(\cdot, \cdot)} \mathbb{E} \left[ \sum_{t=0}^{T-1} g_u(x_t, t)(x_t) | x_0 = x \right]$$

For each policy $u$, we define the cost-to-go function at time $t$, state $x$ as:

$$J_u(x, t) = \mathbb{E} \left[ \sum_{t'=t}^{T} g_u(x_{t'}, t')(x_{t'}) | x_t = x \right]$$

then the finite horizon problem (1) can be viewed as a deterministic nonlinear programming problem:

$$\min_{u(\cdot, \cdot) \in \mathcal{U}} J_u(x, 0)$$

where $\mathcal{U}$ is all possible values $u(\cdot, \cdot)$ can take, i.e. $|\mathcal{U}| = |A||S|^{T}$.

A simple example

Consider a simple MDP problem with 3 states in Figure 1. Each state has a cost function $g(i), i = 1, 2, 3$ and you can only take action when you are in state 1. Action $a_1$ makes the system stay at state 1 with probability 1 while action $a_2$ leads the system to transit to state 2 with probability 1. Transition probabilities from state 2 and 3 are shown in the figure.

Consider the time horizon $T = 2$, we want to solve the problem:

$$\max_{u(\cdot, \cdot)} \mathbb{E} \left[ g_u(x_0, 0)(x_0) + g_u(x_1, 1)(x_1) + g_u(x_2, 2)(x_2) | x_0 = i \right]$$

Our approach is illustrated in Figure 2. The main idea is backward induction. First we consider the terminal time ($t = 2$), the cost of being in state $i$ is $g(i)$, so the cost-to-go function $J(i, 2) = g(i)$. Then back to $t = 1$, we can easily calculate $J(i, 1), i = 2, 3$ knowing $J(i, 2), i = 2, 3$ and the transition probability. We only need to make decision at state 1, from the optimality principle, we should take $a_2$, since $J(2, 2) > J(1, 2)$. Knowing $J(i, 1)$, we can follow the same procedure to get $J(i, 0)$ and the optimal action at time 0. Note that using backward induction, the computational complexity is $O(|A| \cdot |S| \cdot T)$, which is far less than enumerating all possible policies.
To formalize the backward induction approach, we need to use the definition of cost-to-go function in 2. Note that

\[
J_u(x; t) = \mathbb{E}_T \left[ \sum_{t'=t}^T g_u(x_{t'}, t') x_t = x \right]
\]

(5)

\[
= g_u(x) + \mathbb{E}_T \left[ \sum_{t'=t+1}^T g_u(x_{t'}, t') x_t = x \right]
\]

(6)

\[
= g_u(x, t) + \sum_{y \in S} P_u(x, y) J_u(y, t + 1)
\]

(7)

Define the optimal cost-to-go function as

\[
J^*(x, t) = \min_{u(\cdot, \cdot)} J_u(x, t)
\]

(8)

then from 5, we have

\[
J^*(x, t) = \min_{u = \{u_t, u_{t+1}, \cdots, u_T\}} \left\{ g_u(x, t) + \sum_{y \in S} P_u(x, y) J_u(y, t + 1) \right\}
\]

(9)

\[
= \min_{u_t} \left\{ g_u(x, t) + \sum_{y \in S} P_u(x, y) \min_{u_{t+1}, \cdots, u_T} J_u(y, t + 1) \right\}
\]

(10)

\[
= \min_{a \in A_x} \left\{ g_a(x, t) + \sum_{y \in S} P_a(x, y) J^*(y, t + 1) \right\}
\]

(11)

and we can get the optimal policy from the optimal cost-to-go function:

\[
u^*(x, t) = \arg \min_{a} \left\{ g_a(x, t) + \sum_{y \in S} P_a(x, y) J^*(y, t + 1) \right\}
\]

(12)

So we have the following theory:
Theorem 1 (Backward Induction) ¹ Let $J^*(x,t)$ be defined by (11) and $u^*(x,t)$ be defined by (12), then
(a) $J^*(x,t)$ is the optimal cost-to-go function starting at state $x$ and time $t$.
(b) $u^*$ is an optimal policy for the problem starting at time $t$ and ending at time $T$.

2 Discounted cost problem

Now we consider infinite time horizon and want to find the optimal policy for

$$\min_{u(\cdot)} E \left[ \sum_{t=0}^{\infty} \alpha^t g_u(x_t,t|x_0) | x_0 = x \right]$$

where $\alpha \in [0,1]$ represents the discount factor.

Let

$$J_u(x) = E \left[ \sum_{t=0}^{\infty} \alpha^t g_u(x_t,t|x_0) | x_0 = x \right]$$

$$= g_u(x) + \alpha \sum_y P_u(x,y) g_u(y) + \alpha^2 \sum_y P(x_{t+2} = y|x_0 = x) g_u(y) + \ldots$$

$$= \left( \sum_0^{\infty} \alpha^t P_u^t g_u \right)(x)$$

and

$$J^* = \min_u \sum_{t=0}^{\infty} \alpha^t P_u^t g_u$$

We can first consider the discounted finite horizon problem with $T = t$, then study the case where $t \to \infty$.

¹Optimal policy got from (12) is always deterministic, although we can have randomized policies. A simple example is that if we have 2 optimal policies, any random combination of them will also be optimal.
Similar as section 2, we define cost-to-go function as

$$J_u(x, t) = E \left[ \sum_{t'=0}^{t} \alpha^{t-t'} g_u(x_t)(x_t) | x_t = x \right]$$

(17)

as shown in Figure 3.

Figure 3: A simple example

Then from Theorem, we have

$$J^*(x, t) = \min_{\alpha \in \mathbb{A}_0} \left\{ g_u(x) + \alpha \sum_{y} P_\alpha(x, y) J^*(y, t-1) \right\}$$

(18)

Note that

$$J_u(x, t) = E \left[ \sum_{t'=0}^{t-1} \alpha^{t-t'} g_u(x_t)(x_t) | x_t = x \right] + \alpha^t E \left[ g_u(x_0) | x_t = x \right]$$

(19)

$$J_u(x, t-1) = E \left[ \sum_{t'=0}^{t-1} \alpha^{t-t'} g_u(x_t)(x_t) | x_t = x \right] + 0$$

(20)

Since $\alpha \in [0, 1]$, as $t \to \infty$, we will have $J_u(x, t) \to J_u(x)$, and (18) will become

$$J^*(x) = \min_{\alpha \in \mathbb{A}_0} \left\{ g_u(x) + \alpha \sum_{y} P_\alpha(x, y) J^*(y) \right\}$$

(21)

which is the famous Bellman’s equation. We will prove it rigorously in the next lecture.