Lecture 11

- Recall that, given initial state x_0, we wond to find policy $\pi = \{\mu_0, \mu_1, \dots\}$ where $\mu_k : S \mapsto C$, $\mu_k(x_k) \in U(x_k)$, for all $x_k \in S$, $k = 0, 1, \dots$, that minimizes the cost function

$$J_\pi(x_0) = \lim_{N\to\infty} \mathop{\mathbb{E}}_{\substack{w_k, \\ k=0,1,\dots}} \left[\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right]$$

where the cost per stage, $g : S \times U \times D \to \mathbb{R}$ is given, $\alpha \in (0,1)$. Let Π be the set of all admissible policies $\pi = \{\mu_0, \mu_1, \dots\}$ with $\mu_k : S \to U$, $\mu_k(x) \in U(x)$ $\forall x \in S$, $k \in \mathbb{Z}_+$. The optimal cost function J^* is defined by

$$J^*(x) = \min_{\pi \in \Pi} J_\pi(x), \quad x \in S$$

* Notations & Monotonicity:

For any function $J : S \to \mathbb{R}$, we define two operators, T and T^μ as follows.

$$(TJ)(x) = \min_{u \in U(x)} \mathbb{E}\left[g(x,u,w) + \alpha J(f(x,u,w)) \right], \quad x \in S$$

and

$$(T_\mu J)(x) = \mathbb{E}\left[g(x, \mu(x), w) + \alpha J(f(x, \mu(x), w)) \right], \quad x \in S$$

- We denote k times composition of T by T^k, i.e.,

$$(T^k J)(x) = (T(T^{k-1}J))(x)$$

and $(T^0 J)(x) = J(x)$ $\forall x \in S$

Similarly T_μ^k is defined (applying $\{\mu, \mu, \dots \mu\}$ for k times)

If we do not apply stationary policy, i.e., $\pi = \{\mu_0, \mu_1, \dots \mu_{k-1}\}$, then $(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{k-1}} J)(x)$ is defined recursively for $i = 0, \dots, k-2$ by

$i = 0, \ldots, k-2$ by

$$\left(T_{\mu_i} T_{\mu_{i+1}} \cdots T_{\mu_{k-1}} J\right)(x) = \left(T_{\mu_i}\left(T_{\mu_{i+1}} \cdots T_{\mu_{k-1}} J\right)\right)(x)$$

and represents the cost of policy π for the k^{th} stage, α-discounted problem with initial state x, cost per stage g and terminal cost function $\alpha^k J$.

Lemma 1 (Monotonicity lemma). For any functions $J: S \mapsto \mathbb{R}$ and $J': S \to \mathbb{R}$ such that

$$J(x) \leq J'(x) \quad \forall\, x \in S$$

and for any stationary policy $\mu: S \to U$, we have

$$\left(T^k J\right)(x) \leq \left(T^k J'\right)(x) \qquad \forall\, x \in S,\ \forall\, k \in \mathbb{N}$$

$$\left(T_\mu^k J\right)(x) \leq \left(T_\mu^k J'\right)(x) \qquad \forall\, x \in S,\ \forall\, k \in \mathbb{N}.$$

Proof: The result is true for $k = 0$ by hypothesis.

Let for some $k \in \mathbb{Z}_+$

$$\left(T^k J\right)(x) \leq \left(T^k J'\right)(x) \qquad \forall\, x \in S,\ \forall\, k \in \mathbb{N}$$

$$\left(T_\mu^k J\right)(x) \leq \left(T_\mu^k J'\right)(x) \qquad \forall\, x \in S,\ \forall\, k \in \mathbb{N}.$$

Now,

$$\begin{aligned}\left(T^{k+1} J\right)(x) &= \min_{u \in U(x)} \mathbb{E}\left[g(x,u,w) + \alpha\left(T^k J\right)\left(f(x,u,w)\right)\right] \\ &\leq \min_{u \in U(x)} \mathbb{E}\left[g(x,u,w) + \alpha\left(T^k J'\right)\left(f(x,u,w)\right)\right] \\ &= \left(T^{k+1} J'\right)(x)\end{aligned}$$

and by similar argument $\left(T_\mu^{k+1} J\right)(x) \leq \left(T_\mu^{k+1} J'\right)(x)$. ▨

Lemma 2: For every k, function $J: S \to \mathbb{R}$, stationary policy μ and scalar r we have

$$\left(T^k\left(J + re\right)\right)(x) = \left(T^k J\right)(x) + \alpha^k r, \quad \forall\, x \in S$$

$$\left(T_\mu^k\left(J + re\right)\right)(x) = \left(T_\mu^k J\right)(x) + \alpha^k r, \quad \forall\, x \in S.$$

Proof: Due to recursive definition of T^k and T_μ^k, it suffices to show this for $k=1$.

$$\big(T(J+re)\big)(x) = \min_{u\in U(x)} \mathbb{E}\Big[g(x,u,w) + \alpha\big(J(f(x,u,w)) + r\big)\Big]$$
$$= (TJ)(x) + \alpha r \qquad \blacksquare$$

* Now we will show the 3 desired results of infinite horizon problems.

1. $J^*(x) = \lim_{k\to\infty} (T^k J)(x) \quad \forall x \in S$
2. $J^* = TJ^*$
3. If policy μ minimizes r.h.s. of Bellman equation i.e. $u = \mu(x)$ minimizes $\mathbb{E}[g(x,u,w) + \alpha J^*(f(x,u,w))]$ for every $x \in S$ then the stationary policy μ is optimal.

- But before going there we will address an obvious question that should have come to our mind. Why the policies we are considering are picking action based on current state x_k, ie. $u_k = \mu_k(x_k)$ while we have all past info available.

We will see now that Markov policies are adequate.

Proposition 1 (Adequacy of Markov Policies): Assume that the control space is countable, and consider an initial state distribution that takes values over a countable set. The probability distribution of each pair (x_k, u_k) and the expected cost of each stage corresponding to a randomized history dependent policy can also be obtained with a randomized Markov policy.

Proof: Let $\pi = \{\mu_0, \mu_1, \dots\}$ be a randomized history dependent policy and let $\xi_k(x_k)$ and $\zeta_k(x_k, u_k)$ be the

dependent policy and let $\xi_k(x_k)$ and $\zeta_k(x_k, u_k)$ be the corresponding distributions of x_k and (x_k, u_k) respectively. Consider a randomized Markov policy $\bar{\pi} = \{\bar{\mu}_0, \bar{\mu}_1, \dots\}$ where is defined $\forall\, x_k$ with $\xi_k(x_k) > 0$ by

$$\bar{\mu}_k(u_k \mid x_k) = \frac{\zeta(x_k, u_k)}{\xi_k(x_k)}$$

Let $\bar{\xi}_k(x_k)$ and $\bar{\zeta}_k(x_k, u_k)$ be the distributions of x_k and (x_k, u_k) corresponding to $\bar{\mu}_k$.

We will show by induction that $\forall\, k, x_k, u_k$ we have

$$\bar{\xi}_k(x_k) = \xi_k(x_k) \text{ and } \bar{\zeta}_k(x_k, u_k) = \zeta_k(x_k, u_k).$$

Indeed, for $k=0$, $\xi_0(x_0)$ and $\bar{\xi}_0(x_0)$ are both equal to the distribution of initial state.

$$\bar{\zeta}_0(x_0, u_0) = \bar{\xi}_0(x_0)\, \bar{\mu}_0(u_0 \mid x_0) = \bar{\xi}_0(x_0) \cdot \frac{\zeta_0(x_0, u_0)}{\xi_0(x_0)}$$

$$= \zeta_0(x_0, u_0).$$

Assuming induction hypothesis for k, we have

$$\bar{\xi}_{k+1}(x_{k+1}) = \sum_{x_k, u_k} \bar{\zeta}_k(x_k, u_k)\, P(x_k, u_k, x_{k+1})$$

$$= \sum_{x_k, u_k} \bar{\xi}_k(x_k) \cdot \bar{\mu}_k(u_k \mid x_k) \cdot P(x_k, u_k, x_{k+1})$$

$$= \sum_{x_k, u_k} \bar{\xi}_k(x_k)\, \frac{\zeta_k(x_k, u_k)}{\xi_k(x_k)} \cdot P(x_k, u_k, x_{k+1})$$

$$= \sum_{x_k, u_k} \zeta_k(x_k, u_k) \cdot P(x_k, u_k, x_{k+1})$$

$$= \xi_{k+1}(x_{k+1})$$

Furthermore,

$$\overline{\zeta}_{k+1}(x_{k+1}, u_{k+1}) = \overline{\xi}_{k+1}(x_{k+1}) \cdot \overline{\mu}_k(u_{k+1} \mid x_{k+1})$$
$$= \zeta_{k+1}(x_{k+1}, u_{k+1})$$

So, π and $\overline{\pi}$ generates the same state-control dist. It follows from there that the expected cost is also the same. ▨