Advanced Markov Decision Processes: Constraints, Risks, and Predictions

Postgraduate candidate: Zhihong Wu

Supervisor: Prof. Dr. Jingui Xie

Academic department: School of Management

3. Oktober 2025

1 Slide 1 Markov Decision Process

1.1 Exercise: Machine Replacement

Machine replacement is often used as a benchmark problem in Markov decision processes (MDPs). The state space is $\mathcal{S} = \{1, 2, 3\}$ and corresponds to machine quality, where state 1 denotes the worst performance and state 3 the best.
The action space is $\mathcal{A} = \{1, 2\}$ with 1 = *replace machine* and 2 = *use existing machine*. When action $a = 1$ is chosen, the new machine starts in state 3. When action $a = 2$ is chosen, the machine either remains in the same state or deteriorates by one unit with probability $\varepsilon \in [0, 1]$.
Operating a machine in state x under action $a = 2$ incurs a cost $c(x, 2)$ due to possible productivity loss and poor output quality, while replacing the machine in any state costs R. The aim is to minimize the expected cumulative cost over a specified finite horizon N.

Task. Construct the transition probabilities $P(a)$ for each $a \in \mathcal{A}$ and write down the Bellman backward recursion. *Homework:* Compute the value function and the optimal policy.

Solution

$$\mathbf{P}(1) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{P}(2) = \begin{bmatrix} 1 & 0 & 0 \\ \theta & 1-\theta & 0 \\ 0 & \theta & 1-\theta \end{bmatrix}.$$

The Bellman recursion reads:

- initialize $v_N(i) = 0, \ \forall i \in \mathcal{S}$;
- for $k = N-1, \ldots, 0$,

$$v_k(i) = \min\Big\{ R + v_{k+1}(3), \ c(i,2) + \sum_{j=1}^{3} P_{ij}(2)\, v_{k+1}(j) \Big\}, \qquad i \in \{1, 2, 3\}.$$

Given. Horizon $N = 3$, replacement cost $R = 10$, deterioration probability $\theta = 0.2$. Operating costs under action $a = 2$: $c(1,2) = 5,\ c(2,2) = 3,\ c(3,2) = 2$. So that:

$$\mathbf{P}(2) = \begin{bmatrix} 1 & 0 & 0 \\ 0.2 & 0.8 & 0 \\ 0 & 0.2 & 0.8 \end{bmatrix}.$$

Backward induction ($N = 3$) for

Step $t = 3$.

$$v_3(s_3) = 0, \forall s_3 \in (1,2,3)$$

Step $k = 2$.

$$v_2(1) = \min\{10,\ 5\} = 5,$$
$$v_2(2) = \min\{10,\ 3\} = 3,$$
$$v_2(3) = \min\{10,\ 2\} = 2.$$

Step $t = 1$.

$$v_1(1) = \min\{10 + v_2(3),\ 5 + [1 \cdot v_2(1) + 0 \cdot v_2(2) + 0 \cdot v_2(3)]\} \quad = \min\{12,\ 10\} = 10,$$

$$v_1(2) = \min\{10 + v_2(3),\ 3 + [0.2 \cdot v_2(1) + 0.8 \cdot v_2(2) + 0 \cdot v_2(3)]\}$$
$$= \min\{12,\ 3 + 1 + 2.4\} = 6.4$$

$$v_1(3) = \min\{10 + v_2(3),\ 3 + [0 \cdot v_2(1) + 0.2 \cdot v_2(2) + 0.8 \cdot v_2(3)]\}$$
$$= \min\{12,\ 2 + 0.6 + 1.6\} = 4.2$$

Step $t = 0$.

$$v_0(1) = \min\{10 + v_1(3),\ 5 + [1 \cdot v_1(1) + 0 \cdot v_1(2) + 0 \cdot v_1(3)]\} = \min\{14.2,\ 15\} = 14.2,$$

$$v_0(2) = \min\{10 + v_1(3),\ 3 + [0.2 \cdot v_1(1) + 0.8 \cdot v_1(2) + 0 \cdot v_1(3)]\}$$
$$= \min\{14.2,\ 3 + 2 + 5.12\} = 10.12,$$

$$v_0(3) = \min\{10 + v_1(3),\ 2 + [0 \cdot v_1(1) + 0.2 \cdot v_1(2) + 0.8 \cdot v_1(3)]\}$$
$$= \min\{14.2,\ 2 + 1.28 + 3.36\} = 6.64,$$

Optimal policy (argmin of RHS).

t \ a \ s	1	2	3
0	1	2	2
1	2	2	2
2	2	2	2

1.2 Exercise: POMDP

Homework. The state space is $\mathcal{X} = \{1, 2\}$, where state 1 corresponds to a poorly performing machine while state 2 corresponds to a brand new machine. The action space is $\mathcal{U} = \{1, 2\}$, where action 2 denotes keeping the current machine and action 1 denotes replacing the machine with a brand new one (which starts in state 2).The transition probabilities of the machine are

$$\mathbf{P}(1) = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{P}(2) = \begin{bmatrix} 1 & 0 \\ \theta & 1-\theta \end{bmatrix}, \qquad \theta \in [0, 1],$$

where θ denotes the probability that the machine deteriorates.

Assume that the state of the machine x_t is indirectly observed via the quality of the product $y_t \in \{1, 2\}$ generated by the machine. Let p denote the probability that a machine in the good state produces a high–quality product, and let q denote the probability that a deteriorated machine produces a poor–quality product. The observation probability matrix is

$$\mathbf{B} = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}.$$

Operating the machine in state x_t with action $u_t = 2$ incurs an operating cost $c(x_t, 2)$. Replacing the machine at any state costs R, i.e. $c(x_t, 1) = R$. The aim is to minimize

the expected cumulative cost over a horizon N:

$$\mathbb{E}_{\pi_0}\left[\sum_{k=0}^{N-1} c(x_k, u_k)\right],$$

where π_0 denotes the initial distribution of the machine state at time 0. **Task:** Write the belief–state update function and the Bellman equation.

Solution Let the belief state at time $k-1$ be the column vector $\beta_{k-1} = \begin{bmatrix}\beta_{k-1}(1)\\ \beta_{k-1}(2)\end{bmatrix}$, where $\beta_{k-1}(1) + \beta_{k-1}(2) = 1$. We set

$$\tilde{\beta}_k = P(a_{k-1})^T \beta_{k-1}$$

$$\hat{\beta}_k = B_{o_k}(a_{k-1})\tilde{\beta}_k$$

$$\beta_k = \frac{\hat{\beta}_k}{1_s^T \hat{\beta}_k}$$

(1) Action $a_{k-1} = 1$ (Replace), Observation $o_k =$ high

$$\tilde{\beta}_k = P(1)^T \hat{\beta}_{k-1} = \begin{bmatrix}0 & 0\\ 1 & 1\end{bmatrix}\begin{bmatrix}\hat{\beta}_{k-1}(1)\\ \hat{\beta}_{k-1}(2)\end{bmatrix} = \begin{bmatrix}0\\ \hat{\beta}_{k-1}(1) + \hat{\beta}_{k-1}(2)\end{bmatrix} = \begin{bmatrix}0\\ 1\end{bmatrix}$$

$$\hat{\beta}_k = B_{\text{high}}(1)\tilde{\beta}_k = \begin{bmatrix}1-q & 0\\ 0 & p\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix} = \begin{bmatrix}0\\ p\end{bmatrix}$$

$$\beta_k = \frac{1}{p}\begin{bmatrix}0\\ p\end{bmatrix} = \begin{bmatrix}0\\ 1\end{bmatrix}$$

(2) Action $a_{k-1} = 1$ (Replace), Observation $o_k =$ poor

$$\tilde{\beta}_k = P(1)^T \hat{\beta}_{k-1} = \begin{bmatrix}0 & 0\\ 1 & 1\end{bmatrix}\begin{bmatrix}\hat{\beta}_{k-1}(1)\\ \hat{\beta}_{k-1}(2)\end{bmatrix} = \begin{bmatrix}0\\ \hat{\beta}_{k-1}(1) + \hat{\beta}_{k-1}(2)\end{bmatrix} = \begin{bmatrix}0\\ 1\end{bmatrix}$$

$$\hat{\beta}_k = B_{\text{poor}}(1)\tilde{\beta}_k = \begin{bmatrix}q & 0\\ 0 & 1-p\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix} = \begin{bmatrix}0\\ 1-p\end{bmatrix}$$

$$\beta_k = \frac{1}{1-p}\begin{bmatrix}0\\ 1-p\end{bmatrix} = \begin{bmatrix}0\\ 1\end{bmatrix}$$

(3) Action $a_{k-1} = 2$ (Keep), Observation $o_k =$ high

$$\tilde{\beta}_k = P(2)^T \hat{\beta}_{k-1} = \begin{bmatrix} 1 & \theta \\ 0 & 1-\theta \end{bmatrix} \begin{bmatrix} \hat{\beta}_{k-1}(1) \\ \hat{\beta}_{k-1}(2) \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2) \\ (1-\theta)\hat{\beta}_{k-1}(2) \end{bmatrix}$$

$$\hat{\beta}_k = B_{\text{high}}(2)\tilde{\beta}_k = \begin{bmatrix} 1-q & 0 \\ 0 & p \end{bmatrix} \begin{bmatrix} \hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2) \\ (1-\theta)\hat{\beta}_{k-1}(2) \end{bmatrix} = \begin{bmatrix} (1-q)(\hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2)) \\ p(1-\theta)\hat{\beta}_{k-1}(2) \end{bmatrix}$$

$$\beta_k = \frac{1}{(1-q)(\hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2)) + p(1-\theta)\hat{\beta}_{k-1}(2)} \begin{bmatrix} (1-q)(\hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2)) \\ p(1-\theta)\hat{\beta}_{k-1}(2) \end{bmatrix}$$

(4) Action $a_{k-1} = 2$ (Keep), Observation $o_k =$ poor

$$\tilde{\beta}_k = P(2)^T \hat{\beta}_{k-1} = \begin{bmatrix} 1 & \theta \\ 0 & 1-\theta \end{bmatrix} \begin{bmatrix} \hat{\beta}_{k-1}(1) \\ \hat{\beta}_{k-1}(2) \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2) \\ (1-\theta)\hat{\beta}_{k-1}(2) \end{bmatrix}$$

$$\hat{\beta}_k = B_{\text{poor}}(2)\tilde{\beta}_k = \begin{bmatrix} q & 0 \\ 0 & 1-p \end{bmatrix} \begin{bmatrix} \hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2) \\ (1-\theta)\hat{\beta}_{k-1}(2) \end{bmatrix} = \begin{bmatrix} q(\hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2)) \\ (1-p)(1-\theta)\hat{\beta}_{k-1}(2) \end{bmatrix}$$

$$\beta_k = \frac{1}{q(\hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2)) + (1-p)(1-\theta)\hat{\beta}_{k-1}(2)} \begin{bmatrix} q(\hat{\beta}_{k-1}(1) + \theta\hat{\beta}_{k-1}(2)) \\ (1-p)(1-\theta)\hat{\beta}_{k-1}(2) \end{bmatrix}$$

Bellman Equation

$$V_k(\beta) = \min\{Q_k(\beta, u=1), \quad Q_k(\beta, u=2)\}$$

$$Q_k(\beta, u=1) = R + V_{k+1}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$$

$$\begin{aligned} Q_k(\beta, u=2) = {} & \big(c_1\beta(1) + c_2\beta(2)\big) \\ & + P(y = \text{high} \mid \beta, u=2) \cdot V_{k+1}\big(\tau(\beta, u=2, y=\text{high})\big) \\ & + P(y = \text{poor} \mid \beta, u=2) \cdot V_{k+1}\big(\tau(\beta, u=2, y=\text{poor})\big) \end{aligned}$$

$$Q_k(\beta, u=2) = \big(c_1\beta(1) + c_2\beta(2)\big)$$

$$+ \Big((1-q)(\beta(1)+\theta\beta(2)) + p(1-\theta)\beta(2)\Big)$$
$$\cdot V_{k+1}\left(\begin{bmatrix} \frac{(1-q)(\beta(1)+\theta\beta(2))}{(1-q)(\beta(1)+\theta\beta(2))+p(1-\theta)\beta(2)} \\ \\ \frac{p(1-\theta)\beta(2)}{(1-q)(\beta(1)+\theta\beta(2))+p(1-\theta)\beta(2)} \end{bmatrix}\right)$$

$$+ \Big(q(\beta(1)+\theta\beta(2)) + (1-p)(1-\theta)\beta(2)\Big)$$
$$\cdot V_{k+1}\left(\begin{bmatrix} \frac{q(\beta(1)+\theta\beta(2))}{q(\beta(1)+\theta\beta(2))+(1-p)(1-\theta)\beta(2)} \\ \\ \frac{(1-p)(1-\theta)\beta(2)}{q(\beta(1)+\theta\beta(2))+(1-p)(1-\theta)\beta(2)} \end{bmatrix}\right)$$

2 Slide 2 Constrained Markov Decision Process

2.1 Exercise: Prove

$$lim_{\alpha\to 1} C_\beta(\alpha, u) = C_{ea}(\beta, u)$$

Solution (AI helped) The Expected Average Cost is defined as the long-term average of the expected per-step costs:

$$C_{ea}(\beta, u) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_\beta^u[c(S_t, A_t)]$$

The Normalized Discounted Cost is defined with a normalization factor $(1-\alpha)$:

$$C_\alpha(\beta, u) = (1-\alpha) \sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{E}_\beta^u[c(S_t, A_t)]$$

To simplify the notation in the proof, we let c_t represent the expected cost at step t:

$$c_t := \mathbb{E}^u_\beta[c(S_t, A_t)]$$

Using this shorthand, the equation to be proven can be rewritten as the following equality for the sequence $\{c_t\}$:

$$\lim_{\alpha \to 1^-} (1-\alpha) \sum_{t=1}^{\infty} c_t \alpha^{t-1} = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} c_t$$

Given: A sequence of expected costs $\{c_t\}_{t=1}^{\infty}$, whose arithmetic mean converges to a finite limit L.

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} c_t = L$$

To Prove: The Abel mean of the sequence (i.e., the discounted cost) also converges to the same limit L.

$$\lim_{\alpha \to 1^-} (1-\alpha) \sum_{t=1}^{\infty} c_t \alpha^{t-1} = L$$

Let us define a new sequence $d_t = c_t - L$. The given condition is now equivalent to:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} d_t = 0$$

The proposition we need to prove is equivalent to showing that:

$$\lim_{\alpha \to 1^-} (1-\alpha) \sum_{t=1}^{\infty} d_t \alpha^{t-1} = 0$$

Let $T_n = \sum_{t=1}^{n} d_t$ be the partial sums of the sequence, with $T_0 = 0$. We can rewrite the

series using $d_t = T_t - T_{t-1}$:

$$\begin{aligned}
\sum_{t=1}^{\infty} d_t \alpha^{t-1} &= \sum_{t=1}^{\infty} (T_t - T_{t-1}) \alpha^{t-1} \\
&= \sum_{t=1}^{\infty} T_t \alpha^{t-1} - \sum_{t=1}^{\infty} T_{t-1} \alpha^{t-1} \\
&= \sum_{t=1}^{\infty} T_t \alpha^{t-1} - \sum_{k=0}^{\infty} T_k \alpha^{k} \\
&= \sum_{t=1}^{\infty} T_t \alpha^{t-1} - \left(T_0 \alpha^0 + \sum_{k=1}^{\infty} T_k \alpha^{k} \right) \\
&= \sum_{t=1}^{\infty} T_t \alpha^{t-1} - \sum_{t=1}^{\infty} T_t \alpha^{t} \\
&= \sum_{t=1}^{\infty} \left(T_t \alpha^{t-1} - T_t \alpha^{t} \right) \\
&= \sum_{t=1}^{\infty} T_t \alpha^{t-1} (1 - \alpha) \\
&= (1 - \alpha) \sum_{t=1}^{\infty} T_t \alpha^{t-1}
\end{aligned}$$

The problem becomes to prove:

$$\lim_{\alpha \to 1^-} (1 - \alpha)^2 \sum_{t=1}^{\infty} T_t \alpha^{t-1} = 0$$

Let $B_n = T_n / n$ be the arithmetic mean. We know $\lim_{n \to \infty} B_n = 0$. Using $T_t = t \cdot B_t$, the final goal is to prove:

$$\lim_{\alpha \to 1^-} (1 - \alpha)^2 \sum_{t=1}^{\infty} t B_t \alpha^{t-1} = 0$$

Since $\lim_{n \to \infty} B_n = 0$, by the definition of a limit, for any given $\epsilon > 0$, there exists an integer M such that for all $t > M$, we have $|B_t| < \epsilon$.

We split the infinite sum into two parts: a finite "head"(from $t = 1$ to M) and an infinite "tail"(from $t = M + 1$ to ∞).

$$(1 - \alpha)^2 \sum_{t=1}^{M} t B_t \alpha^{t-1} + (1 - \alpha)^2 \sum_{t=M+1}^{\infty} t B_t \alpha^{t-1}$$

A. The Head (from 1 to M):

$$\lim_{\alpha \to 1^-} (1 - \alpha)^2 \sum_{t=1}^{M} t B_t \alpha^{t-1}$$

Since M is a fixed integer, the summation is a finite polynomial in α. As $\alpha \to 1^-$, the factor $(1-\alpha)^2$ goes to 0, forcing the entire expression to 0.

B. The Tail (from M+1 to ∞): We bound the absolute value of the tail:

$$\begin{aligned}\left|(1-\alpha)^2 \sum_{t=M+1}^{\infty} tB_t\alpha^{t-1}\right| &\leq (1-\alpha)^2 \sum_{t=M+1}^{\infty} t|B_t|\alpha^{t-1} \\ &< (1-\alpha)^2 \sum_{t=M+1}^{\infty} t \cdot \epsilon \cdot \alpha^{t-1} \\ &= \epsilon \cdot (1-\alpha)^2 \sum_{t=M+1}^{\infty} t\alpha^{t-1}\end{aligned}$$

The partial sum is bounded by the full infinite series:

$$\sum_{t=M+1}^{\infty} t\alpha^{t-1} \leq \sum_{t=1}^{\infty} t\alpha^{t-1} = \frac{1}{(1-\alpha)^2}$$

Therefore, we can continue the inequality:

$$\epsilon \cdot (1-\alpha)^2 \sum_{t=M+1}^{\infty} t\alpha^{t-1} \leq \epsilon \cdot (1-\alpha)^2 \cdot \frac{1}{(1-\alpha)^2} = \epsilon$$

This shows that for α sufficiently close to 1, the absolute value of the tail can be bounded by any arbitrarily small ϵ.

For any $\epsilon > 0$, we have shown that as $\alpha \to 1^-$, the absolute value of the entire expression is less than $|\text{head}| + |\text{tail}| < 0 + \epsilon = \epsilon$. Since ϵ can be arbitrarily small, the limit must be 0. This completes the proof of the theorem.

2.2 Exercise: Prove

$$\overline{\text{co}}(\mathcal{L}^{\alpha}_{U_D}(\beta)) \subseteq \mathcal{L}^{\alpha}_{U_S}(\beta)$$

Solution (AI helped) The following proves the more complex converse direction, which is that any extreme point of $\mathcal{L}^{\alpha}_{U_S}(\beta)$ must be generated by a deterministic policy. This is the key step to show that $\mathcal{L}^{\alpha}_{U_S}(\beta) \subseteq \overline{\text{co}}(\mathcal{L}^{\alpha}_{U_D}(\beta))$. The proof is by contradiction.

1. Assumption for Contradiction We assume there exists an extreme point $f \in \mathcal{L}_{U_S}^{\alpha}(\beta)$ that is generated by a stationary policy w which is **randomized**.

Since w is a randomized policy, by definition, there must exist at least one state $s_0 \in S$ and at least two distinct actions $a_1, a_2 \in A(s_0)$ for which the policy assigns a non-zero probability:

$$w_{s_0}(a_1) > 0 \quad \text{and} \quad w_{s_0}(a_2) > 0$$

2. Construction of Two New Policies Let's choose a sufficiently small positive number ϵ such that:

$$0 < \epsilon \leq \min\{w_{s_0}(a_1), w_{s_0}(a_2)\}$$

We now construct two new stationary policies, w_1 and w_2, by slightly perturbing the probabilities in state s_0 and keeping them unchanged in all other states.

For state s_0:

$$w_{1,s_0}(a_1) = w_{s_0}(a_1) + \epsilon \qquad w_{2,s_0}(a_1) = w_{s_0}(a_1) - \epsilon$$

$$w_{1,s_0}(a_2) = w_{s_0}(a_2) - \epsilon \qquad w_{2,s_0}(a_2) = w_{s_0}(a_2) + \epsilon$$

$$w_{1,s_0}(a) = w_{s_0}(a) \qquad w_{2,s_0}(a) = w_{s_0}(a) \quad \text{for } a \notin \{a_1, a_2\}$$

For all other states $s \neq s_0$:

$$w_{1,s}(a) = w_{2,s}(a) = w_s(a) \quad \forall a \in A(s)$$

By construction, w_1 and w_2 are valid stationary policies, so their corresponding occupation measures, $f_1 = f_\alpha(\beta, w_1)$ and $f_2 = f_\alpha(\beta, w_2)$, are both elements of $\mathcal{L}_{U_S}^{\alpha}(\beta)$. Furthermore, since $w_1 \neq w_2$, we have $f_1 \neq f_2$ (assuming state s_0 is reachable).

3. Finding the Contradiction From our construction, it is clear that the original policy w is the midpoint of w_1 and w_2:

$$w = \frac{1}{2}w_1 + \frac{1}{2}w_2$$

The relationship between an occupation measure and its policy is established through the Bellman flow equations, which are linear in the occupation measure vector. Since

the state-transition matrix $P(w)$ depends linearly on w, it can be shown that the occupation measure $f_\alpha(\beta, w)$ also depends linearly on the policy in this context. Therefore, it follows that:

$$f = f_\alpha(\beta, w) = f_\alpha\left(\beta, \frac{1}{2}w_1 + \frac{1}{2}w_2\right) = \frac{1}{2}f_\alpha(\beta, w_1) + \frac{1}{2}f_\alpha(\beta, w_2)$$

This gives us the crucial result:

$$f = \frac{1}{2}f_1 + \frac{1}{2}f_2$$

This result directly contradicts our initial assumption. We assumed that f is an **extreme point** (a vertex) of the convex set $\mathcal{L}^\alpha_{U_S}(\beta)$. However, we have just shown that f can be expressed as a strict convex combination (specifically, the midpoint) of two other distinct points, f_1 and f_2, from the same set. By definition, an extreme point cannot be represented in this way.

4. Conclusion The contradiction implies that our initial assumption—that an extreme point can be generated by a randomized policy—must be false.

Therefore, we conclude that any extreme point f of the set $\mathcal{L}^\alpha_{U_S}(\beta)$ must be generated by a **deterministic** stationary policy. This completes the proof for $\mathcal{L}^\alpha_{U_S}(\beta) \subseteq \overline{\text{co}}(\mathcal{L}^\alpha_{U_D}(\beta))$.

2.3 Exercise: Proof of Equivalence between COP and the LP

Solution Let ρ^* be an optimal solution to the Linear Program $LP^\alpha_1(\beta)$, which exists because the feasible set $Q^\alpha(\beta)$ is compact and the objective is continuous. Let $w(\rho^*)$ be the stationary policy constructed from ρ^* as:

$$w_s(a) := \frac{\rho^*(s,a)}{\sum_{a' \in A(s)} \rho^*(s,a')}$$

We prove that $w(\rho^*)$ is an optimal policy for the COP by showing it is feasible and that it achieves the optimal value.

1. Proof of Feasibility For any constraint k, the cost incurred by the policy $w(\rho^*)$ is $D^k(\beta, w(\rho^*))$.

$$\begin{aligned} D^k(\beta, w(\rho^*)) &= \langle f_\alpha(\beta, w(\rho^*)), d^k \rangle && \text{(By definition of policy cost)} \\ &= \langle \rho^*, d^k \rangle && \text{(Since } f_\alpha(\beta, w(\rho^*)) = \rho^*\text{)} \\ &\leq V_k && \text{(Since } \rho^* \text{ is a feasible solution to the LP)} \end{aligned}$$

This holds for all $k = 1, \ldots, K$. Therefore, the policy $w(\rho^*)$ is feasible for the COP.

2. Proof of Optimality Let $C_\alpha(\beta)$ be the optimal value for the COP. The cost of the policy $w(\rho^*)$ is $C_\alpha(\beta, w(\rho^*))$.

$$\begin{aligned} C_\alpha(\beta, w(\rho^*)) &= \langle f_\alpha(\beta, w(\rho^*)), c \rangle && \text{(By definition of policy cost)} \\ &= \langle \rho^*, c \rangle && \text{(Since } f_\alpha(\beta, w(\rho^*)) = \rho^*\text{)} \\ &= C^* && \text{(By definition, } C^* \text{ is the optimal value of the LP)} \\ &= C_\alpha(\beta) && \text{(From Theorem 3(i), } C^* = C_\alpha(\beta)\text{)} \end{aligned}$$

This shows that the policy $w(\rho^*)$ achieves the optimal cost. Since it is also feasible, $w(\rho^*)$ is an optimal policy for the COP.

2.4 Exercise: Proof The value C_α is the largest super-harmonic function

Solution We prove two parts: (1) C_α is a super-harmonic function, and (2) it is the largest such function.

Part 1: Proof that C_α is a super-harmonic function From the Bellman Optimality Equation, we have:

$$C_\alpha(s) = \min_{a' \in A} \left((1-\alpha)c(s,a') + \alpha \sum_{y \in S} P_{sy}(a') C_\alpha(y) \right)$$

By the definition of the minimum operator, for any action $a \in A$:

$$C_\alpha(s) \leq (1-\alpha)c(s,a) + \alpha \sum_{y \in S} P_{sy}(a) C_\alpha(y)$$

This is the definition of a super-harmonic function. Thus, C_α is super-harmonic.

Part 2: Proof that C_α is the largest super-harmonic function Let ϕ be any super-harmonic function. By its definition, for any $s \in S$:

$$\phi(s) \leq \min_{a' \in A} \left((1-\alpha)c(s,a') + \alpha \sum_{y \in S} P_{sy}(a')\phi(y) \right)$$

Let T be the Bellman Optimality Operator, defined as $(Tv)(s) := \min_{a' \in A}(\dots)$. The inequality above is $\phi \leq T\phi$.

Since T is monotonic, we can apply it repeatedly to the inequality $\phi \leq T\phi$ to obtain the sequence:

$$\phi \leq T\phi \leq T^2\phi \leq \cdots \leq T^n\phi$$

The sequence $T^n\phi$ is known to converge to the unique fixed point of T, which is the optimal value function C_α.

$$\lim_{n \to \infty} T^n\phi = C_\alpha$$

Taking the limit in the inequality chain, we get:

$$\phi \leq \lim_{n \to \infty} T^n\phi \quad \Longrightarrow \quad \phi \leq C_\alpha$$

Since ϕ was an arbitrary super-harmonic function, this proves that C_α is the largest super-harmonic function.

2.5 Exercise: Proof that $DP^\alpha(\beta)$ is the dual of $LP_1^\alpha(\beta)$

Solution The proof is derived by following the standard rules for formulating the dual of a linear program.

First, we write down the primal linear program $LP_1^\alpha(\beta)$ for the unconstrained case ($K = 0$).

$$\begin{aligned}
\min_{\rho} \quad & \sum_{s \in S, a \in A} \rho(s,a)c(s,a) \\
\text{s.t.} \quad & \sum_{y \in S, a \in A} \rho(y,a)(\delta_s(y) - \alpha P_{ys}(a)) = (1-\alpha)\beta(s), && \forall s \in S && (1) \\
& \rho(s,a) \geq 0, && \forall s \in S, a \in A && (2)
\end{aligned}$$

Following the format from the presentation slides:

We introduce a dual variable for each primal constraint.

- For the equality constraint (1): $v(s)$ free, $\forall s \in S$

Based on the rules of duality, the dual problem is formulated as:

$$\begin{aligned} \max_{v \text{ free}} \quad & (1-\alpha)\sum_{s\in S}\beta(s)v(s) \\ \text{s.t.} \quad & v(s) - \alpha\sum_{s'\in S}P_{ss'}(a)v(s') \le c(s,a), \qquad \forall s,a \end{aligned}$$

To match the form of $DP^{\alpha}(\beta)$, we perform a variable substitution. Let $\phi(s) = (1-\alpha)v(s)$. Since $v(s)$ is free, $\phi(s)$ is also free.

The objective function transforms to:

$$(1-\alpha)\sum_{s\in S}\beta(s)v(s) = \sum_{s\in S}\beta(s)\big((1-\alpha)v(s)\big) = \sum_{s\in S}\beta(s)\phi(s) = \langle\beta,\phi\rangle$$

The constraints transform by multiplying by $(1-\alpha) > 0$:

$$\begin{aligned} v(s) - \alpha\sum_{s'\in S}P_{ss'}(a)v(s') &\le c(s,a) \\ (1-\alpha)v(s) - \alpha\sum_{s'\in S}P_{ss'}(a)(1-\alpha)v(s') &\le (1-\alpha)c(s,a) \\ \phi(s) - \alpha\sum_{s'\in S}P_{ss'}(a)\phi(s') &\le (1-\alpha)c(s,a) \end{aligned}$$

Rearranging gives the final form of the constraints:

$$\phi(s) \le (1-\alpha)c(s,a) + \alpha\sum_{y\in S}P_{sy}(a)\phi(y)$$

After substitution, we have the dual problem $DP^{\alpha}(\beta)$:

$$\begin{aligned} \sup_{\phi} \quad & \langle\beta,\phi\rangle \\ \text{s.t.} \quad & \phi(s) \le (1-\alpha)c(s,a) + \alpha\sum_{y\in S}P_{sy}(a)\phi(y), \quad \forall s,a \end{aligned}$$

This completes the proof that $DP^{\alpha}(\beta)$ is the dual of $LP_1^{\alpha}(\beta)$ in the unconstrained case.

3 Slide 3 Markov Decision Process with Prediction

3.1 Homework: Proof of Monotonicity

Solution Let $H(c_{t+1}) = \sum_{\hat{c}_{t+2} \in \mathcal{C}} \tilde{Q}(c_{t+1}, \hat{c}_{t+2}) J^P_{t+1}(c_{t+1}, \hat{c}_{t+2})$. We want to prove that $H(c_{t+1})$ is decreasing in c_{t+1}.

1. Given Properties

- By induction hypothesis, $J^P_{t+1}(c_{t+1}, \hat{c}_{t+2})$ is decreasing in both c_{t+1} and $\hat{c}_{t+2}$.
- The kernel $\tilde{Q}(c_{t+1}, \hat{c}_{t+2})$ is TP2, which implies that for $c'_{t+1} > c_{t+1}$, the distribution $\tilde{Q}(c'_{t+1}, \cdot)$ first-order stochastically dominates $\tilde{Q}(c_{t+1}, \cdot)$.
- A key consequence of first-order stochastic dominance is that for any decreasing function $h(\cdot)$, if $c'_{t+1} > c_{t+1}$, then $\mathbb{E}_{c'_{t+1}}[h] \leq \mathbb{E}_{c_{t+1}}[h]$.

2. Proof Let $c'_{t+1} > c_{t+1}$. We need to show that $H(c'_{t+1}) \leq H(c_{t+1})$. Consider the difference:

$$H(c'_{t+1}) - H(c_{t+1}) = \sum_{\hat{c}_{t+2}} \tilde{Q}(c'_{t+1}, \hat{c}_{t+2}) J^P_{t+1}(c'_{t+1}, \hat{c}_{t+2}) - \sum_{\hat{c}_{t+2}} \tilde{Q}(c_{t+1}, \hat{c}_{t+2}) J^P_{t+1}(c_{t+1}, \hat{c}_{t+2})$$

We add and subtract the term $\sum_{\hat{c}_{t+2}} \tilde{Q}(c_{t+1}, \hat{c}_{t+2}) J^P_{t+1}(c'_{t+1}, \hat{c}_{t+2})$ to split the difference into two parts:

$$H(c'_{t+1}) - H(c_{t+1}) = \underbrace{\left[\sum_{\hat{c}_{t+2}} \tilde{Q}(c'_{t+1}, \hat{c}_{t+2}) J^P_{t+1}(c'_{t+1}, \hat{c}_{t+2}) - \sum_{\hat{c}_{t+2}} \tilde{Q}(c_{t+1}, \hat{c}_{t+2}) J^P_{t+1}(c'_{t+1}, \hat{c}_{t+2}) \right]}_{\text{Part 1}}$$
$$+ \underbrace{\left[\sum_{\hat{c}_{t+2}} \tilde{Q}(c_{t+1}, \hat{c}_{t+2}) J^P_{t+1}(c'_{t+1}, \hat{c}_{t+2}) - \sum_{\hat{c}_{t+2}} \tilde{Q}(c_{t+1}, \hat{c}_{t+2}) J^P_{t+1}(c_{t+1}, \hat{c}_{t+2}) \right]}_{\text{Part 2}}$$

Analyzing Part 1 Let $h(\hat{c}_{t+2}) := J^P_{t+1}(c'_{t+1}, \hat{c}_{t+2})$. Since J^P_{t+1} is decreasing in its second argument, $h(\cdot)$ is a decreasing function. As $c'_{t+1} > c_{t+1}$, first-order stochastic

dominance implies:

$$\sum_{\hat{c}_{t+2}} \tilde{Q}(c'_{t+1}, \hat{c}_{t+2}) h(\hat{c}_{t+2}) \leq \sum_{\hat{c}_{t+2}} \tilde{Q}(c_{t+1}, \hat{c}_{t+2}) h(\hat{c}_{t+2})$$

Therefore, Part 1 ≤ 0.

Analyzing Part 2 We can factor out the common term $\tilde{Q}(c_{t+1}, \hat{c}_{t+2})$:

$$\text{Part 2} = \sum_{\hat{c}_{t+2}} \tilde{Q}(c_{t+1}, \hat{c}_{t+2}) \left[J^P_{t+1}(c'_{t+1}, \hat{c}_{t+2}) - J^P_{t+1}(c_{t+1}, \hat{c}_{t+2}) \right]$$

Since J^P_{t+1} is decreasing in its first argument and $c'_{t+1} > c_{t+1}$, the term in the brackets is ≤ 0. Since $\tilde{Q}(\cdot, \cdot) \geq 0$, the entire summation is a sum of non-positive terms. Therefore, Part 2 ≤ 0.

3. Conclusion Since both Part 1 and Part 2 are less than or equal to zero, their sum is also less than or equal to zero.

$$H(c'_{t+1}) - H(c_{t+1}) \leq 0 \quad \Longrightarrow \quad H(c'_{t+1}) \leq H(c_{t+1})$$

This completes the proof that the expression is decreasing in c_{t+1}.