

**Infinite-stage Probabilistic
Dynamic Programming: Markov
Decision Processes
(MDP)**

620-362

Recall from Finite-stage Dynamic Programming models...

The state of next period depends only on the state of the current period, and the decision made at the current period. Hence the term Markov. For example, a period begins in state i , and a decision d is chosen, then with probability $p(j | i, d)$, the next period begins in state j .

Recursion Relation: If the stages for the problem have been classified into one of T stages, there must be a recursion that relates the cost or reward earned during stages, $t, t + 1, \dots, T$ to the cost or reward earned from stages $t + 1, t + 2, \dots, T$.

Principle of Optimality: Given the current state, the optimal decision for each of the remaining stages is independent of all previously reached states or previously chosen decisions.

What if...

In previous examples, the *horizon length*, (T), that is, the number of time periods over which the expected profit is maximised is finite.

Problem: Now, what happens if a decision maker is facing a long horizon and is not sure of the horizon length?

Solution: We assume that the horizon length is infinite!

Problem: But then maximising the expected reward earned over an infinite horizon, we may have an unbounded reward!!!

Solution: Use the idea of discounted rewards.

Discounted Rewards

Concept: \$1 reward received during the next period will have the same value as a reward of \$ β dollars ($0 < \beta < 1$) received during the current period.

Let M be the maximum immediate reward that can be received during a single period. The max expected discounted reward that can be received over an infinite period horizon is

$$M + M\beta + M\beta^2 + \dots = \frac{M}{1 - \beta} < \infty.$$

This way the problem of an infinite expected reward is resolved.

Now we can deal with the infinite horizon MDPs.

An example

A Machine Replacement Problem (Winston, Sec 21.5, Ex 11)

At the beginning of each week, a machine is in one of four conditions: excellent (E), good (G), average (A), and bad (B). The weekly revenue earned by a machine in each type of condition is as follows: excellent, \$100; good, \$80; average, \$50; and bad, \$10. After observing the condition of a machine at the beginning of the week, we have the option of instantaneously replacing it with an excellent machine, which costs \$200. The quality of a machine deteriorates over time, as shown in next page.

Task: For this situation, determine the state space, decision sets, transition probabilities, and expected rewards.

Infinite Horizon MDPs - Key Concepts

State space: At the beginning of each period, the MDP is in some state i , where i is a member of $S = \{1, \dots, N\}$ for N finite.

Decision Set: $D(i)$.

Transition Probabilities: Given that a period begins in state i , and decision $d \in D(i)$ is chosen, the probability that the next period starts in state j is given by $p(j|i, d)$.

Expected Rewards: The reward for a period which begins with state i and uses decision $d \in D(i)$ is denoted by r_{id} .

Next Period's States of Machines

Present state	Prob next week's state			
	E	G	A	B
E	0.7	0.3	-	-
G	-	0.7	0.3	-
A	-	-	0.6	0.4
B	-	-	-	1.0

Infinite Horizon MDPs

States: $S = \{E, G, A, B\}$

Let:

R = replace machine at beginning of current period;

NR = do not replace machine at beginning of current period

Decision set:

$D(E) = \{NR\}$; $D(G) = D(A) = D(B) = \{R, NR\}$

Transition probabilities:

$p(j|i, NR)$ for all $i, j \in \{E, G, A, B\}$;

$p(j|i, R)$ for all $j \in \{E, G, A, B\}$, $i \in \{G, A, B\}$

Expected rewards:

$r_{j, NR}$ for all $j \in \{E, G, A, B\}$;

$r_{j, R}$ for all $j \in \{G, A, B\}$;

Infinite Horizon MDPs - More Concepts

A policy is a rule that specifies how each period's decision is chosen.

A policy is a stationary policy if whenever the state is i , the policy δ chooses the same decision, $\delta(i)$, independent of the period.

Let $V_\delta(i)$ be the expected discounted reward earned during an infinite number of periods, given that at beginning of period 1, the state is i and the stationary policy is δ .

In a maximisation problem, we define $V(i) = \max_{\delta \in \text{all policies}} V_\delta(i)$.

If $V_{\delta^*}(i) = \max_{\delta} V_\delta(i) \forall i \in S$, then δ^* is an optimal policy.

The Value Determination Equation

The value determination equation:

$$V_{\delta}(i) = r_{i,\delta(i)} + \beta \sum_{j=1}^N p(j|i, \delta(i))V_{\delta}(j)$$

e.g. $\delta(E) = \delta(G) = NR$, and $\delta(A) = \delta(B) = R$.

$$V_{\delta}(E) = 100 + 0.9(0.7V_{\delta}(E) + 0.3V_{\delta}(G))$$

$$V_{\delta}(G) = 80 + 0.9(0.7V_{\delta}(G) + 0.3V_{\delta}(A))$$

$$V_{\delta}(A) = -100 + 0.9(0.7V_{\delta}(E) + 0.3V_{\delta}(G))$$

$$V_{\delta}(B) = -100 + 0.9(0.7V_{\delta}(E) + 0.3V_{\delta}(G))$$

Solving these equations, we get $V_{\delta}(E) = 687.81$, $V_{\delta}(G) = 572.19$, $V_{\delta}(A) = 487.81$, and $V_{\delta}(B) = 487.81$.

But, how do we determine the optimal stationary policy?

A Linear Programming Approach

Note that:

$$V_i = \max_{d \in D(i)} \left\{ r_{id} + \left(\beta \sum_{j=1}^N p(j|i, d) V_j \right) \right\}, \quad \forall i = 1, \dots, N.$$

Hence an optimal stationary policy for maximisation problem can be found by solving the following LP:

$$\begin{aligned} \min \quad & z = V_1 + V_2 + \dots + V_N \\ \text{s.t.} \quad & V_i \geq r_{id} + \beta \sum_{j=1}^N p(j|i, d) V_j, \quad \forall i = 1, \dots, N \\ & \text{and } \forall d \in d(i) \\ & V_i \text{ urs } \forall i \in S. \end{aligned}$$

The machine replacement example

$$\begin{aligned} & \min_z V_E + V_G + V_A + V_B \\ \text{s.t. } & V_E \geq 100 + 0.9(0.7V_E + 0.3V_G) \quad \{E, NR\} \\ & V_G \geq 80 + 0.9(0.7V_G + 0.3V_A) \quad \{G, NR\} \\ & V_G \geq -100 + 0.9(0.7V_E + 0.3V_G) \quad \{G, R\} \\ & V_A \geq 50 + 0.9(0.7V_A + 0.3V_B) \quad \{A, NR\} \\ & V_A \geq -100 + 0.9(0.7V_E + 0.3V_G) \quad \{A, R\} \\ & V_B \geq 10 + 0.9V_B \quad \{B, NR\} \\ & V_B \geq -100 + 0.9(0.7V_E + 0.3V_G) \quad \{B, R\} \\ & V_i \text{ urs } \forall i \in \{E, G, A, B\}. \end{aligned}$$

Solution: $V_E = 690.23$, $V_G = 575.50$, $V_A = 492.35$, $V_B = 490.23$. Constraints $(\{E, NR\})$, $(\{E, NR\})$, $(\{A, NR\})$, and $(\{B, R\})$ are binding. That is, replace a bad machine and not to replace an excellent, good or average machine.