

# Markov Decision Processes

*Definitions; Stationary policies; Value improvement algorithm, Policy improvement algorithm, and linear programming for discounted cost and average cost criteria.*

# Markov Decision Process

Let  $X = \{X_0, X_1, \dots\}$  be a *system description* process on state space  $E$  and

let  $D = \{D_0, D_1, \dots\}$  be a *decision* process with action space  $A$ .

The process  $(X, D)$  is a Markov decision process if, for  $j \in E$  and  $n = 0, 1, \dots$ ,  $P\{X_{n+1} = j | X_n, D_n, \dots, X_0, D_0\} = P\{X_{n+1} = j | X_n, D_n\}$

Furthermore, for each  $k \in A$ , let  $\mathbf{f}_k$  be a cost vector and  $\mathbf{P}_k$  be a one-step transition probability matrix. Then the cost  $f_k(i)$  is incurred whenever  $X_n = i$  and  $D_n = k$ , and

$$P\{X_{n+1} = j | X_n = i, D_n = k\} = P_k(i, j)$$

The problem is to determine how to choose a sequence of actions in order to minimize cost.

# Policies

A *policy* is a rule that specifies which action to take at each point in time. Let  $D$  denote the set of all policies.

In general, the decisions specified by a policy may

- depend on the current state of the system description process
- be randomized (depend on some external random event)
- also depend on past states and/or decisions

A *stationary policy* is defined by a (deterministic) action function that assigns an action to each state, independent of previous states, previous actions, and time  $n$ .

Under a stationary policy, the MDP is a Markov chain.

# Cost Minimization Criteria

Since a MDP goes on indefinitely, it is likely that the total cost will be infinite. In order to meaningfully compare policies, two criteria are commonly used:

1. Expected total *discounted* cost computes the present worth of future costs using a discount factor  $\alpha < 1$ , such that one dollar obtained at time  $n = 1$  has a present value of  $\alpha$  at time  $n = 0$ . Typically, if  $r$  is the rate of return, then  $\alpha = 1/(1 + r)$ . The expected total discounted cost is

$$E \left[ \sum_{n=0}^{\infty} \alpha^n f_{D_n} (X_n) \right]$$

2. The long run *average* cost is  $\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{n=0}^{m-1} f_{D_n} (X_n)$

# Optimization with Stationary Policies

If the state space  $E$  is finite, there exists a stationary policy that solves the problem to minimize the discounted cost:

$$v^\alpha(i) = \min_{d \in \mathcal{D}} v_d^\alpha(i), \text{ where } v_d^\alpha(i) = E_d \left[ \sum_{n=0}^{\infty} \alpha^n f_{D_n}(X_n) | X_0 = i \right]$$

If every stationary policy results in an irreducible Markov chain, there exists a stationary policy that solves the problem to minimize the average cost:

$$\varphi^* = \min_{d \in \mathcal{D}} \varphi_d, \text{ where } \varphi_d = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{n=0}^{m-1} f_{D_n}(X_n)$$

# Computing Expected Discounted Costs

Let  $X = \{X_0, X_1, \dots\}$  be a Markov chain with one-step transition probability matrix  $\mathbf{P}$ , let  $f$  be a cost function that assigns a cost to each state of the M.C., and let  $\alpha$  ( $0 < \alpha < 1$ ) be a discount factor. Then the expected total discounted cost

is 
$$g(i) = E \left[ \sum_{n=0}^{\infty} \alpha^n f(X_n) \mid X_0 = i \right] = \left[ (\mathbf{I} - \alpha \mathbf{P})^{-1} f \right](i)$$

Why? Starting from state  $i$ , the expected discounted cost can be found recursively as  $g(i) = f(i) + \alpha \sum_j P_{ij} g(j)$ , or

$$\mathbf{g} = \mathbf{f} + \alpha \mathbf{P} \mathbf{g}$$

Note that the expected discounted cost always depends on the initial state, while for the average cost criterion the initial state is unimportant.

# Solution Procedures for Discounted Costs

Let  $\mathbf{v}^\alpha$  be the (vector) optimal value function whose  $i$ th component is  $v^\alpha(i) = \min_{d \in \mathcal{D}} v_d^\alpha(i)$

For each  $i \in E$ ,  $v^\alpha(i) = \min_{k \in A} \left\{ f_k(i) + \alpha \sum_{j \in E} P_k(i, j) v^\alpha(j) \right\}$

These equations uniquely determine  $\mathbf{v}^\alpha$ .

If we can somehow obtain the values  $\mathbf{v}^\alpha$  that satisfy the above equations, then the optimal policy is the vector  $\mathbf{a}$ , where

$$a(i) = \arg \min_{k \in A} \left\{ f_k(i) + \alpha \sum_{j \in E} P_k(i, j) v^\alpha(j) \right\}$$

“arg min” is “the argument that minimizes”

# Value Iteration for Discounted Costs

Make a guess ... keep applying the optimal value equations until the fixed point is reached.

Step 1. Choose  $\varepsilon > 0$ , set  $n = 0$ , let  $v_0(i) = 0$  for each  $i$  in  $E$ .

Step 2. For each  $i$  in  $E$ , find  $v_{n+1}(i)$  as

$$v_{n+1}(i) = \min_{k \in A} \left\{ f_k(i) + \alpha \sum_{j \in E} P_k(i, j) v_n(j) \right\}$$

Step 3. Let  $\delta = \max_{i \in E} \{v_{n+1}(i) - v_n(i)\}$

Step 4. If  $\delta < \varepsilon$ , stop with  $v^\alpha = v_{n+1}$ . Otherwise, set  $n = n+1$  and return to Step 2.

# Policy Improvement for Discounted Costs

Start myopic, then consider longer-term consequences.

Step 1. Set  $n = 0$  and let  $\mathbf{a}_0(i) = \arg \min_{k \in A} f_k(i)$

Step 2. Adopt the cost vector and transition matrix:

$$f(i) = f_{a_n(i)}(i) \quad P(i, j) = P_{a_n(i)}(i, j)$$

Step 3. Find the value function  $v = (\mathbf{I} - \alpha \mathbf{P})^{-1} f$

Step 4. Re-optimize:  $a_{n+1}(i) = \arg \min_{k \in A} \left\{ f_k(i) + \alpha \sum_{j \in E} P_k(i, j) v(j) \right\}$

Step 5. If  $a_{n+1}(i) = a_n(i)$ , then stop with  $\mathbf{v}^\alpha = v$  and  $\mathbf{a}^\alpha = \mathbf{a}_n(i)$ .

Otherwise, set  $n = n + 1$  and return to Step 2.

# Linear Programming for Discounted Costs

Consider the linear program:

$$\begin{aligned} \max \quad & \sum_{i \in E} u(i) \\ \text{s.t.} \quad & u(i) \leq f_k(i) + \alpha \sum_{j \in E} P_k(i, j) u(j) \text{ for each } i, k \end{aligned}$$

The optimal value of  $u(i)$  will be  $v^\alpha(i)$ , and the optimal policy is identified by the constraints that hold as equalities in the optimal solution (slack variables equal 0).

Note: the decision variables are unrestricted in sign!

# Long Run Average Cost per Period

For a given policy  $d$ , its long run average cost could be found from its cost vector  $f_d$  and one-step transition probability matrix  $\mathbf{P}_d$ :

First, find the limiting probabilities by solving

$$\pi_j = \sum_{i \in E} \pi_i P_d(i, j), \quad j \in E; \quad \sum_{j \in E} \pi_j = 1$$

Then

$$\varphi_d = \lim_{m \rightarrow \infty} \frac{\sum_{n=0}^{m-1} f_{d(X_n)}(X_n)}{m} = \sum_{j \in E} f_d(j) \pi_j$$

So, in principle we could simply enumerate all policies and choose the one with the smallest average cost... not practical if  $A$  and  $E$  are large.

# Recursive Equation for Average Cost

Assume that every stationary policy yields an irreducible Markov chain. There exists a scalar  $\varphi^*$  and a vector  $\mathbf{h}$  such that for all states  $i$  in  $E$ ,

$$\varphi^* + h(i) = \min_{k \in A} \left\{ f_k(i) + \sum_{j \in E} P_k(i, j) h(j) \right\}$$

The scalar  $\varphi^*$  is the optimal average cost and the optimal policy is found by choosing for each state the action that achieves the minimum on the right-hand-side.

The vector  $\mathbf{h}$  is unique up to an additive constant ... as we will see, the difference between  $h(i) - h(j)$  represents the increase in total cost from starting out in state  $i$  rather than  $j$ .

# Relationships between Discounted Cost and Long Run Average Cost

- If a cost of  $c$  is incurred each period and  $\alpha$  is the discount factor, then the total discounted cost is  $v = \sum_{n=0}^{\infty} c\alpha^n = \frac{c}{1-\alpha}$
- Therefore, a total discounted cost  $v$  is equivalent to an average cost of  $c = (1-\alpha)v$  per period, so  $\lim_{\alpha \rightarrow 1} (1-\alpha)v^\alpha(i) = \varphi^*$
- Let  $v^\alpha$  be the optimal discounted cost vector,  $\varphi^*$  be the optimal average cost and  $\mathbf{h}$  be the mystery vector from the previous slide.

$$\lim_{\alpha \rightarrow 1} [v^\alpha(i) - v^\alpha(j)] = h(i) - h(j)$$

# Policy Improvement for Average Costs

Designate one state in  $E$  to be “state number 1”

Step 1. Set  $n = 0$  and let  $\mathbf{a}_0(i) = \arg \min_{k \in A} f_k(i)$

Step 2. Adopt the cost vector and transition matrix:

$$f(i) = f_{a_n(i)}(i) \quad P(i, j) = P_{a_n(i)}(i, j)$$

Step 3. With  $h(1) = 0$ , solve  $\varphi + \mathbf{h} = f + \mathbf{P}\mathbf{h}$

Step 4. Re-optimize: 
$$a_{n+1}(i) = \arg \min_{k \in A} \left\{ f_k(i) + \alpha \sum_{j \in E} P_k(i, j) h(j) \right\}$$

Step 5. If  $\mathbf{a}_{n+1}(i) = \mathbf{a}_n(i)$ , then stop with  $\varphi^* = \varphi$  and  $\mathbf{a}^*(i) = \mathbf{a}_n(i)$ .  
Otherwise, set  $n = n + 1$  and return to Step 2.

# Linear Programming for Average Costs

Consider randomized policies: let  $w_i(k) = P\{D_n = k \mid X_n = i\}$ . A stationary policy has  $w_i(k) = 1$  for each  $k=a(i)$  and 0 otherwise. The decision variables are  $x(i,k) = w_i(k)\pi(i)$ .

The objective is to minimize the expected value of the average cost (expectation taken over the randomized policy):

$$\min \varphi = \sum_{i \in E} \sum_{k \in A} x(i,k) f_k(i)$$

$$\text{s.t. } \sum_{k \in A} x(j,k) = \sum_{i \in E} \sum_{k \in A} x(i,k) P_k(i,j) \text{ for each } j \in E$$

$$\sum_{i \in E} \sum_{k \in A} x(i,k) = 1$$

Note that one constraint will be redundant and may be dropped.