MARKOV DECISION PROCESSES (MDP)

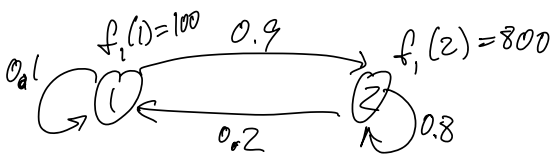
(CHAPTER 12 IN FELDMAN — VALDEZ-FLORES TEXT)

EXAMPLE

$S = E = \{1,2\}$ STATE SPACE

$A = \{1,2\}$ ACTION SPACE

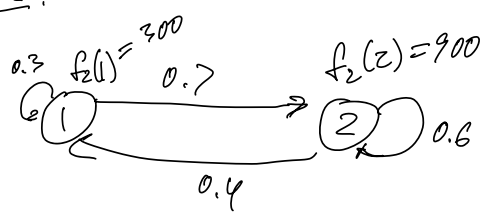
ACTION $k=1$:



$$P_1 = \begin{pmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \end{pmatrix} = (P_1(i,j))$$

$$f_1 = \begin{pmatrix} 100 \\ 800 \end{pmatrix}$$

ACTION $k=2$:



$$P_2 = \begin{pmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{pmatrix} = (P_2(i,j))$$

$$f_2 = \begin{pmatrix} 300 \\ 900 \end{pmatrix}$$

IN THIS COURSE WE ALWAYS ASSUME: E AND A FINITE

GENERAL DEFINITION OF AN MDP:

$X = (X_0, X_1, X_2, \ldots)$ STATE PROCESS, $X_n \in E \equiv S$ = STATE SPACE;

$D = (D_0, D_1, D_2, \ldots)$ ACTION PROCESS, $D_n \in A$ = ACTION SPACE;

ACTION $D_n \in A$ AT TIME n MAY DEPEND ONLY ON $X_n, D_{n-1}, X_{n-1}, \ldots, D_0, X_0$;

$f_k(i)$ = COST INCURRED WHEN ACTION k IS CHOSEN IN STATE i;

$P_k(i,j)$ = PROB. OF TRANSITION FROM i TO j WHEN ACTION k CHOSEN.

PROCESS (X,D) IS AN MDP IF $\forall n$, $\forall i,j \in E$, $\forall k \in A$:

$$\mathbb{P}\{X_{n+1}=j \mid X_0, D_0, \ldots, X_n=i, D_n=k\} = \mathbb{P}\{X_{n+1}=j \mid X_n=i, D_n=k\} = P_k(i,j).$$

FINITE HORIZON PROBLEM

$0 \le \alpha \le 1$ IS A DISCOUNT FACTOR.

$d \in \mathcal{D}$ A GIVEN POLICY; $\mathcal{D}$ = SET OF ALL POLICIES

POLICY d: $a_{(m)}(i) \in A$ ACTION YOU TAKE IN STATE i, WHEN AT TIME-DISTANCE m FROM THE END OF HORIZON

EXPECTED DISCOUNTED COST UNDER A GIVEN POLICY d, OVER TIME HORIZON $0, 1, 2, \dots, m$:

$$v^{\alpha}_{(m),d}(i) = \mathbb{E}\left[\sum_{n=0}^{m} \alpha^n f_{D_n}(X_n) \,\middle|\, X_0 = i\right] =$$

$$= \mathbb{E}\left[\sum_{n=0}^{m} \alpha^n f_{a_{(m-n)}(X_n)}(X_n) \,\middle|\, X_0 = i\right]$$

EXAMPLE ($m = 2$):

$$v^{\alpha}_{(2),d}(i) = \mathbb{E}\left[f_{a_{(2)}(X_0)}(X_0) + \alpha f_{a_{(1)}(X_1)}(X_1) + \alpha^2 f_{a_{(0)}(X_2)}(X_2) \,\middle|\, X_0 = i\right] =$$

$$= f_{a_{(2)}(i)}(i) + \alpha \sum_{j \in E} P_{a_{(2)}(i)}(i,j) \cdot$$

$$\cdot \mathbb{E}\left[f_{a_{(1)}(X_1)}(X_1) + \alpha f_{a_{(0)}(X_2)}(X_2) \,\middle|\, X_0 = i, D_0 = a_{(2)}(i), X_1 = j\right]$$

$$= f_{a_{(2)}(i)}(i) + \alpha \sum_{j \in E} P_{a_{(2)}(i)}(i,j) \cdot$$

$$\cdot \mathbb{E}\left[f_{a_{(1)}(j)}(j) + \alpha f_{a_{(0)}(X_2)}(X_2) \,\middle|\, X_1 = j\right]$$

FINALLY, WE OBTAIN:

$$v^{\alpha}_{(2),d}(i) = f_{a_{(2)}(i)}(i) + \alpha \sum_{j \in E} P_{a_{(2)}(i)}(i,j)\, v^{\alpha}_{(1),d}(j), \quad \forall i$$

IN GENERAL:

$$\begin{cases} v^{\alpha}_{(m),d}(i) = f_{a_{(m)}(i)}(i) + \alpha \sum_{j \in E} P_{a_{(m)}(i)}(i,j)\, v^{\alpha}_{(m-1),d}(j), \quad \forall i \\ \text{INITIAL CONDITION:} \quad v^{\alpha}_{(0),d}(i) = f_{a_{(0)}(i)}(i), \quad \forall i \end{cases}$$

$\Downarrow$

CAN RECURSIVELY COMPUTE ALL $v^{\alpha}_{(m),d}(i)$, $\forall i$, $m = 0, 1, 2, \ldots$

FOR ANY GIVEN POLICY d

PROBLEM: FIND OPTIMAL(MIN) $v^{\alpha}_{(m)}(i)$ AND OPTIMAL $a_{(m)}(i)$!

$$v^{\alpha}_{(m)}(i) = \min_{d \in \mathcal{D}} v^{\alpha}_{(m),d}(i)$$

(d: POLICY)

SOLUTION: USE BELLMAN (DYNAMIC PROGRAMMING) EQUATIONS:

$$v^{\alpha}_{(m)}(i) = \min_{k \in A} \left\{ f_k(i) + \alpha \sum_{j \in E} P_k(i,j)\, v^{\alpha}_{(m-1)}(j) \right\}, \quad \forall i$$

INITIAL CONDITION: $v^{\alpha}_{(0)}(i) = \min_{k \in A} f_k(i), \quad \forall i$

$$a^{\alpha}_{(m)}(i) = \operatorname*{argmin}_{k \in A} \left\{ f_k(i) + \alpha \sum_{j \in E} P_k(i,j)\, v^{\alpha}_{(m-1)}(j) \right\}, \quad \forall i$$

$$a^{\alpha}_{(0)}(i) = \operatorname*{argmin}_{k \in A} f_k(i), \quad \forall i$$

USING BELLMAN EQUATIONS, CAN RECURSIVELY, FOR $m = 0, 1, 2, \cdots$, COMPUTE OPTIMAL VALUE FUNCTION $v^{\alpha}_{(m)}(i)$ AND OPTIMAL POLICY $a_{(m)}(i)$.

INFINITE HORIZON PROBLEM. $\alpha < 1$

TOTAL DISCOUNTED COST (FOR $\alpha < 1$) UNDER A GIVEN POLICY d:

$$v_d^\alpha(i) = \mathbb{E}\left[\sum_{n=0}^{\infty} \alpha^n f_{D_n}(X_n) \mid X_0 = i\right]$$

PROBLEM: FIND A POLICY d MINIMIZING $v_d^\alpha(i)$, $\forall i$.

OPTIMAL VALUE FUNCTION:

$$v^\alpha(i) = \min_{d \in \mathcal{D}} v_d^\alpha(i)$$

$\mathcal{D}$ — SET OF ALL POLICIES

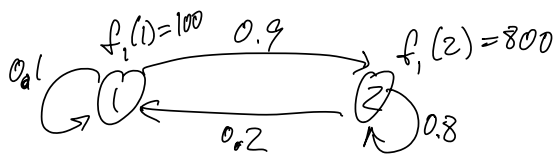
STATIONARY POLICIES

A POLICY S.T. ACTION $a(i)$ DEPENDS ONLY ON THE CURRENT STATE i, BUT NOT ON TIME, IS CALLED STATIONARY. $a = (a(i))$ IS CALLED ACTION FUNCTION.

IF POLICY d IS STATIONARY, GIVEN BY ACTION FUNCTION a, THEN OFTEN ACTION FUNCTION a ITSELF IS CALLED A (STATIONARY) POLICY

EXAMPLE:

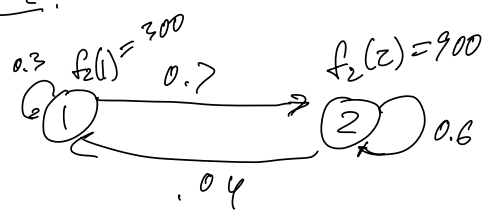
ACTION $k=1$:



$$P_1 = \begin{pmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \end{pmatrix} = (P_1(i,j))$$

$$f_1 = \begin{pmatrix} 100 \\ 800 \end{pmatrix}$$

ACTION $k=2$:



$$P_2 = \begin{pmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{pmatrix} = (P_2(i,j))$$

$$f_2 = \begin{pmatrix} 300 \\ 900 \end{pmatrix}$$

STAT. POLICY: $\{a(1)=2,\ a(2)=1\} \iff a=(2,1)$.

THE PROCESS BECOMES A DTMC:

$$P_a = \begin{pmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{pmatrix}$$ TRANSITION PROBABILITIES

$$f_a = \begin{pmatrix} 300 \\ 800 \end{pmatrix}$$ COSTS

IN GENERAL:

AN MDP UNDER A STATIONARY POLICY BECOMES A DTMC

(WITH A COST ASSOCIATED WITH EACH STATE)

INFINITE HORIZON. DISCOUNT FACTOR $\alpha < 1$

TOTAL DISCOUNTED COST UNDER A GIVEN STATIONARY POLICY a

$$v_a^\alpha(i) = f_a(i) + \alpha \sum_j P_a(i,j) \cdot f_a(j) + \alpha^2 \sum_j P_a^{(2)}(i,j) f_a(j) + \cdots$$

$$\Updownarrow$$

$$v_a^\alpha = \begin{pmatrix} \vdots \\ v_a^\alpha(i) \\ \vdots \end{pmatrix} = f_a + \alpha P_a f_a + \alpha^2 P_a^2 f_a + \cdots =$$

$$= \left[I + \alpha P_a + \alpha^2 P_a^2 + \cdots \right] f_a$$

$$\boxed{v_a^\alpha = \left[I - \alpha P_a \right]^{-1} f_a}$$

INFINITE HORIZON. DISCOUNT FACTOR $\alpha < 1$.

TOTAL DISCOUNTED COST MINIMIZATION

FOR ANY $0 \le \alpha < 1$, AN OPTIMAL POLICY EXISTS, WHICH IS A STATIONARY POLICY.

(OPT. POLICY MAY, AND TYPICALLY WILL, DEPEND ON α)

THE OPTIMAL VALUE FUNCTION $v^\alpha(i)$ SATISFIES BELLMAN EQ:

$$v^\alpha(i) = \min_{u \in A} \left\{ f_u(i) + \alpha \sum_j P_u(i,j) v^\alpha(j) \right\}.$$

MOREOVER, IT IS THE ONLY FUNCTION WHICH SATISFIES THIS EQUATION.

A STATIONARY POLICY a IS OPTIMAL IF AND ONLY IF

$$a(i) = \arg\min_{u \in A} \left\{ f_u(i) + \alpha \sum_j P_u(i,j) v^\alpha(j) \right\}.$$

INFINITE HORIZON. DISCOUNT FACTOR $\alpha < 1$.

TOTAL DISCOUNTED COST MINIMIZATION

FINDING AN OPTIMAL STATIONARY POLICY AND THE OPTIMAL VALUE FUNCTION.

POLICY IMPROVEMENT ALGORITHM
FOR INFINITE HORIZON DISCOUNTED COST
(IN MDP-COMMS FILE)

AVERAGE COST MINIMIZATION PROBLEM

(NO DISCOUNT : $\alpha = 1$)

$d \in \mathcal{D}$ A POLICY. AVE. COST:

$$\varphi_d = \lim_{n \to \infty} E\left\{ \frac{f_{D_0}(X_0) + \dots + f_{D_{n-1}}(X_{n-1})}{n} \right\} =$$

FINITE E, FINITE A

$$= \lim_{n \to \infty} \frac{f_{D_0}(X_0) + \dots + f_{D_{n-1}}(X_{n-1})}{n}$$

AVERAGE COST MINIMIZATION PROBLEM:

FIND POLICY d^*, FOR WHICH

$$\varphi_{d^*} = \min_d \varphi_d$$

$$\varphi^* = \varphi_{d^*} = \text{OPTIMAL (MINIMUM) COST}$$

AVERAGE COST UNDER A GIVEN STATIONARY POLICY $a=(a(i))$

IN THIS COURSE:

WE WILL ALWAYS ASSUME THAT UNDER ANY STAT. POLICY a, THE TRANS. PROB. ARE SUCH THAT THE RESULTING DTMC IS IRREDUCIBLE

TRANS. PROB. MATRIX P_a; COSTS (COLUMN VECTOR) f_a

AVE. COST $\ell_a = \sum_{i \in E} \pi_i f_a(i)$, WHERE

$\pi = (\pi_i)$ IS THE (UNIQUE) STATIONARY DISTR.

$$\begin{cases} \pi = \pi P_a \\ \sum_i \pi_i = 1 \end{cases}$$

UNDER OUR ASSUMPTIONS, THERE ALWAYS EXISTS AN OPTIMAL POLICY α, WHICH IS STATIONARY.

$\exists$ NUMBER φ^* AND A FUNCTION (VECTOR) $h = (h(i))$ SUCH THAT:

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

(ii) φ^* IS THE OPT. COST;

(iii) STAT. POLICY α IS OPTIMAL IF AND ONLY IF

$$a(i) = \operatorname*{argmin}_{k \in A} \left\{ f_k(i) + \sum_{j \in E} P_k(i,j)\, h(j) \right\}, \quad \forall i;$$

(iv) SUCH FUNCTION $h = (h(i))$ IS UNIQUE UP TO AN ADDITIVE CONSTANT.

FOR EXAMPLE, IF YOU FIX $i_0 \in E$, AND SET $h(i_0) = 0$, THEN h IS JUST UNIQUE.

INTUITIVELY :

1) $\varphi^* = \lim_{\alpha \uparrow 1} (1-\alpha)\, v^\alpha(i)$, $\forall i$.

$$(1-\alpha)\, v^\alpha(i) = (1-\alpha)\alpha^0 \mathbb{E} f_{D_0}(X_0) + (1-\alpha)\alpha^1 \mathbb{E} f_{D_1}(X_1) + (1-\alpha)\alpha^2 \mathbb{E} f_{D_2}(X_2) + \cdots$$

2) $h(i) - h(j) = \lim_{\alpha \uparrow 1} \{ v^\alpha(i) - v^\alpha(j) \}$

$$v^\alpha(i) = \min_k \Big\{ f_k(i) + \alpha \sum_j P_k(i,j)\, v^\alpha(j) \Big\} \quad , \forall i$$

$$(1-\alpha)\, v^\alpha(i) = \min_k \Big\{ f_k(i) + \alpha \sum_j P_k(i,j)\, v^\alpha(j) \Big\} - \alpha\, v^\alpha(i) =$$

$$= \min_k \Big\{ f_k(i) + \alpha \sum_j P_k(i,j) \big(v^\alpha(j) - v^\alpha(i) \big) \Big\}$$

$\Big\downarrow \alpha \uparrow 1$

$$\varphi^* = \min_k \Big\{ f_k(i) + \sum_j P_k(i,j) \{ h(j) - h(i) \} \Big\}$$

$$\varphi^* = \min_k \Big\{ f_k(i) + \sum_j P_k(i,j)\, h(j) - h(i) \Big\}$$

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

HOW TO CHECK THAT A GIVEN STATIONARY POLICY α IS OPTIMAL

→ SEE EXAMPLE ON PAGE 338 OF MDP-COMMS

HOW TO FIND AN OPTIMAL STATIONARY POLICY

POLICY IMPROVEMENT ALGORITHM (FOR AVERAGE COST):

PROPOSITION 12.11 ON PAGE 340 OF MDP-COMMS

— SEE EXAMPLE OF THE POLICY IMPROVEMENT ALGORITHM APPLICATION, PAGE 341 OF MDP-COMMS