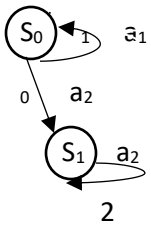
Assignment 3

Bellman optimality equation for a finite-horizon MDP :



T = 1 ; Reward = 1, Optimum Value [1 2]

T = 2; Optimum Value [2 2]

Bellman Optimality Equation for a finite horizon MDP,

$V_t(s) = \max_a (r(s,a) + \sum P(s, a, s') V_{t+1}(S))$, where $\max_a$ = best action

Problem in state, S --> action, a --> goto s'. P(s,a,s')

Backward Function :

t---> t+1

Objective :

R_t – Random variable of reward at time 't'

$E [\sum_{t=1}^{T} R_t]$

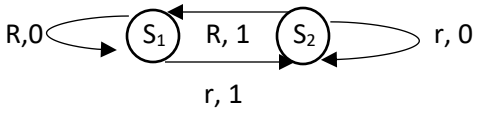
Part b:

Different Policies : π : 1) History Dependent

2) Deterministic Markov policy

$\pi: \mathcal{P}$ --> a (Randomized Policy)

--> To check which policy we are in,



Probability : r (right) : 0 1

l (left) : 0 1

{ ½ r and ½ l } Half the time right and half left, reward greater than '1'.

Randomized policy :

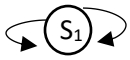
$$\prod : \mathcal{S} \rightarrow \Delta^a$$

$$\mathcal{S} * a\ [0,1]$$

$$V_t(s) = r(s,\prod(s)) + \sum_{s'} P(s, \prod(s), s')\ V_{t+1}(S')$$

If horizon applied to Infinite loops, MDPs applied in discounted state,

Discounted MDPs :



a_1,$1

For example: Day1 Day2 Day3 a_2 is not best since Infinite

$1 $1.25 $1.25^2 + 1.25

$2 $2.50 $2.50^2 + 2.50

Predicting in backward direction (right to left)

$1 <-- $5

$0.2 <-- $1

$E\ [\sum^{\infty}_{t=1} r^t R_t]$

a_1: $\sum^{\infty}_{t=0} r^t = 1/ (1-r)$ r = 0.9, 1 --> 0.9

a_2 : $2 \sum^{\infty}_{t=0} r^t = 2/ (1-r)$ 2 --> 1.8

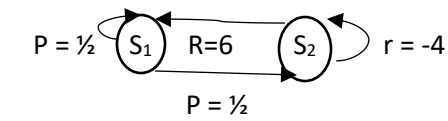
Predicting 10%

Geometric Series, $([1+r+r^2+r^3+r^4] (1-r)) / (1-r)$

Bellman's

$V(s) = \max_a (\nu\ (s,a) + \nu \sum_{s'} p\ (s,a,s')\ v(s'))$, where $\nu = discount$

$\prod(V_{\prod}(S)) = r\ (s, \prod(s) + \nu \sum_{s'} P\ (S\prod(s), s')\ v_{\prod}(s'))$



$V(s_1) = 6 + ½\ V(s_1) + ½\ v(s_2)$

$V(s_2) = -4 + 2/3\ (V(s_1) + 1/3\ v(s_2))$

2 equations, 2 values, 1 V value

Writing in vector-matrix :

From equations, V = V + ν P_v , where v = vector

P = $\begin{matrix} 1/2 & 1/2 \\ 2/3 & 1/3 \end{matrix}$ r = $\begin{matrix} 6 \\ -4 \end{matrix}$

P $\begin{matrix} 1 \\ 1 \end{matrix}$ = $\begin{matrix} 1 \\ 1 \end{matrix}$

Matrix, multiply by 1, should get 1

Backward multiplication,

α = $\begin{matrix} 1 \\ 0 \end{matrix}$ α^TP = [1/2 1/2]

Probability of matrices,

V – ν P_v = r

(I - ν P) ν = r

V = (I - ν p)$^{-1}$ r I = $\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$

Methods include using LV, QR, SVD, Gauss siedel except Inverse which should not be used.

- ⇨ Eigenvalue ≠ 0
- ⇨ (P ∈ (-ν, ν))
- ⇨ (t - νP) ∈ (1- ν, 1+ ν)
- ⇨ ν = $\begin{matrix} 21 \\ 12 \end{matrix}$

P_o = $\begin{matrix} 1 \\ 0 \end{matrix}$ P = 21

P_0 = $\begin{matrix} 0 \\ 1 \end{matrix}$ p = 12

P_0 = $\begin{matrix} 2/3 \\ 1/3 \end{matrix}$ P = p_0^T ν $\sum^{\infty}_{t=0} E[r^t R_t]$, where p_0^T ν = reward

$P_1 = p_0^T P = \begin{matrix} 1/2 \\ 1/2 \end{matrix}$, where $P_1 T_r$ = reward

$P_3 = p_2^T P = P_1 P^3$, where $P_1 P^3$ = reward

$p_0^T \sum^{\infty}_{t=0} r^t p^t r$

$P_0^T r + \nu P_0^T P_r + \nu^2 P_0^T P^2_r$ + ……..

~ $1 + \nu + \nu^2 + \nu^3$ + …….. = $(1 - \nu)^{-1}$

Neumann series, = $P_0^T(I - rp)^{-1}\,\nu$, where $(I - rp)^{-1}\,\nu = P_0^T r$

3) $V - rP_v = r$

Difference between a stationary and Markovian policy

a) Stationary policies are not optimal always. Used if we want to continue taking actions indefinitely. A stationary policy is a mapping from states to actions.

$\prod$: S → A, $\prod$(S) is action to do at state s (regardless of time)

b) Markovian policy is not dependent on history

Pr(St+1|At ,St ,At-1,St-1,..., S0) = Pr(St+1|At ,St) , Next state only depends on current state and current action.

Is the optimal policy in finite-horizon MDP guaranteed to be Markovian or stationary (or both)?

a) To be optimal under finite horizon, a non-stationary policy should be used, hence it is guaranteed to be Markovian.

What happens to the value function when you add a constant to all the rewards?

Only the Relative size of the rewards are vital, for K = 0, we have value of state : $V = \sum^{\infty}_{t=t0} \nu^{(t-t0)}\, r_t$. A value for K gives,

$\sum^{\infty}_{t=t0} \nu^{(t-t0)}\,(r_t + K) = V + (K/(1- \nu))$.

In a sequential task, the sum does not move to infinity, the result is less simple in particular if the length is not fixed.

What happens to the policy. How do the answers differ for finite and infinite-horizon problem?

For the policy, the constant does not change the relative values, only the relative size is vital for all the policies. For example, adding a constant 'c' to all the rewards adds a constant, V_e to all the values of all states and this does not affect the relative values of any states under policies. In an episodal task, The result will be simple in particular if the length of the episodes is not fixed.