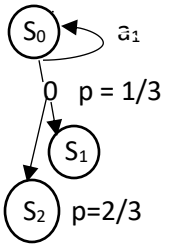
Assignment 5

MDP : r(s,a)

$r_2(s,a,s') = r_1(s,a)$



$r_1(s,a,s_1) = 5$

$r_2(s,a,s') = 10$

$r(s,a_1) = 5.(1/3) + 10.(2/3)$

$= 25/3$

$V(s) = \max_a r(s,a) + \nu \Sigma\ p(s,a,s').r(s')$

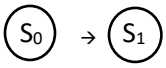
$V(s) = \max_a \Sigma_{s'}\ p(s,a,s').(r(s,a,s') + \nu\ v(s')$

Average Reward MDPs:

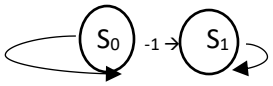
Discounted $(\Sigma_{t=0}\infty)E[r^t R_t]$

Average Reward MDPs

$\text{Lim}_{T->\infty}(1/T)(\Sigma_{t=1}\infty)\ E[R_t]$ (1/1) /(2/2)/(3/3)/….



(-1000/1)/(-991/2)/(-998/3)/….



Always 1 as we eventually move to a state.

0.999

Gain ~ Value function

Gain(s) = average reward starting from state from state s.

$G(s) = \lim_{T->\infty} \Sigma_s\ E[R_t]$

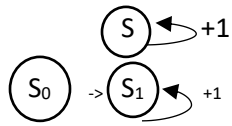
$G(s_1) = 1$

$G(s_0) = 1$

1. Base stock is maximum mandatory value that one has to hold in inventory management. For example, if the base stock is 35, and end of day count is 25, one has to order 10 more to satisfy base stock policy.
2. When we increase the discount factor, the order amount decreases, very marginally.
3. Computation time decreases when we increase the discount factor

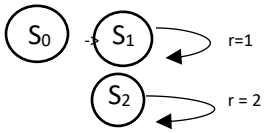
Advantage of TD(0) over Monte-carlo algorithm:

- TD methods do not require a model of the environment, only experience
- TD, but not MC, methods can be fully incremental.



Ergodic Markov chain

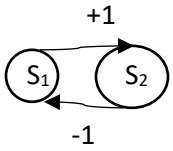
Single chain



multichain $g(s_1) = 1$

$g(S_2) = 2$

$g(s_0) = 1.5$



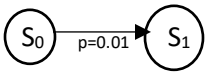
Periodic $g(s_1) = 0$

$g(s_2) = 0$

$P = p_0^T p\infty$ $p^1 = p$

$P^2 = p.p$

$P^3 = p.p.p$



$$P = \begin{matrix} 0.99 & 0.01 \\ 0 & 1 \end{matrix}$$

$\Sigma_{t=0}\infty\ r^t p^t = (I-\nu P)^{-1}$

$P\infty$ $= \lim(1/T)\ \Sigma_{t=0}^T\ p^t = (1/T)\ (I-P)^{-1}$ (Inverse doesn't exist)

Stochastic matrix has an eigen value 1.

$P_s^T = P_s^T P$

$$P_s = \begin{matrix} 0 \\ 1 \end{matrix}$$