CS6700 – RL – ASSIGNMENT - 1

BY: S. VRSHAL
CH18B020

③ a) For an optimal policy π^* having cost J^*

$$T_{\pi^*} J^* = TJ^* \Rightarrow J^* = TJ^*$$

(prop 2)

and J^* is the unique solution to the above equation ——— (1)

additionally, J^* is the unique solution to

$$T_{\pi^*} J^* = J \quad \text{——} (2)$$

Upon introducing a new action a' in a state s', we get a new MDP. Let the optimal policy of such an MDP be π^*_{new} and cost be J^*_{new}.

From (2), since we don't have a change in transition probabilities & $\pi^*(i)$ is also unchanged

$$T_{\pi^*} J^* = J^* \text{ still holds} \quad \text{——} (3)$$

We still need to check if

$$J^* = TJ^*$$ to know whether π^* is indeed π^*_{new}.

[Assuming minimisation problem for part a) – although it really doesn't matter here]

$\therefore \pi^*$ was optimal for the older MDP

$(T U^*)(i) = J^*(i)$ is older MDP

$$\Rightarrow J^*(i) = \min_a \sum_{\forall j} \left(P_{ij}(a) \left(g(i,a,j) + J^*(j) \right) \right)$$

In the new MDP,

$$U^{**}(i) = \min_a \sum \left(P_{ij}(a) \left(g(i,a,j) + J^*(j) \right) \right)$$

holds for all $i \neq s'$ ——— (4)

(because nothing has changed
— no new actions have been added)

claim①: If $J^*(s') = \min_a \sum_{\forall j} \left(P_{s'j}(\pi^*(s')) \left(g(s', \pi^*(s'), j) + J^*(j) \right) \right)$

$$= \min_a \sum \left(P_{s'j}(a) \left(g(s', a, j) + J^*(j) \right) \right)$$

then, $\pi^{**} = \pi^*$ optimal & $J^{**} = J^*_{\text{optimal}}$

Proof: Procedure to test:

Assume that compare $J^*(s')$ with

$$\sum_{\forall j} P_{s'j}(a') \left(g(s', a', j) + J'(j) \right)$$

if $J^*(s') \neq \sum_{\forall j} P_{s'j}(a') \left(g(s', a', j) + J'(j) \right)$

then $\pi^*(s')$ is still the solution to the eqn because substituting $\pi^*(s')$ gives us a lower cost than a''

$\Rightarrow$ $\pi^*(s')$ is still optimal

& hence the claim is satisfied.

$\Rightarrow \pi^* = \pi^*_{new}$ & $J^* = J^*_{new}$

proof for the claim①:

If $J^*(s) = \min_a \sum P_{s'j}(a)\,(g(s',a,j) + J^*(j))$

then combining with eqn ④, we have

(for new MDP),

$J^* = J^*$ $\Rightarrow$ J^* is still the optimal cost

& we already showed that expected cost of policy π^* is J^*

$\Rightarrow J^*_{new} = J^*$ & $\pi^*_{new} = \pi^*$.

Summary: For the s' state equation, check if a' produces a minimum in RHS, if it doesn't the old policy remains the optimal policy

claim ②: If a' gives a lower value in T^*J^* (note that J^* is the optimum in old MDP, it is not so in the new MDP) for s' state; then for that state s', the action a' is optimal.

proof:

Suppose a' is not the optimal action, then it is as good as dealing with an MDP without that action

$\Rightarrow$ this is the old MDP

$\Rightarrow$ π^* & J^* are optimal in the new MDP then.

$\Rightarrow$ $J^*(s') = \min_a \sum_j P_{s'j}\,(g(s',a,j) + J(j))$

has solution $a = \pi^*(a)$ instead of a'

this is a contradiction because $J^*(s') \neq \sum_j P_{s'j}(\cdot)$

to find new policy $\Rightarrow$ the claim is true.

From claim ② we can infer that a' is the optimal action at s' if $\pi^* \neq \pi^+$ new

Let $\pi_0(i) = \begin{cases} \pi^*(i) & \forall i \neq s' \\ a' & i = s' \end{cases}$

Using this policy we can perform policy iteration & find the new optimal policy

[note that $\pi_0(i)$ is proper since

$J^* \leq T_{\pi_0} J^*$ — (if a' is optimal)

as shown in that eqn in claim ①]

Since we already found one optimal action, policy iteration might converge faster

b) Assuming it is a maximization problem we are given that the cumulative reward $J_{\pi^*}(s^t)$ decreases by Δ

$J_{\pi^*}(s^t) = E\left(g(s^t, \pi^*(s^t), j) + J_{\pi^*}(j) \right)$

Since no other information is given, I would like to interpret it as $J_{\pi^*}(s^+)$ has changed because of a Δ decrease in average expected transition cost from s^+

that $\hat{E}(g_{new}(s^+, \pi^*(s^+), j))$

$= E(g_{new}(s^+, \pi^*(s^+), j)) - \Delta$

Because of this change in transition cost from s^+, there are 2 implications

i) J_{π^*} might not be solution to $J = T_{\pi^*} J$ i.e. it may not be the expected cost of policy π^*

ii) π^* may not be optimal policy (because returns from s^+ has reduced, intuitively there is a possibility that a policy that reduces has lesser number of transitions to s^+ may be the new optimal policy, π^*_{new})

Checking if π^* is optimal

Step ①: Perform policy evaluation

$$J(i) = \sum_j P_{ij}(\pi_k(i)) \, g(i, \pi_k(i), j) + J(j)$$

where $g(i, \pi_k(i), j) = \begin{cases} g_{original}(i, \pi_k(i), j) & \forall i \neq s^+ \\ g_{original}(s^+, a^+, j) - & \end{cases}$

$$J(i) = E\left(g_{original}(i, \pi_k(i), j) + J(j) \right)$$

& for $i = s^+$

$$J(s^+) = E\left(g_{original}(s^+, \pi_k(s^+), j) + J(s^+) \right) \quad \text{— } \Delta$$

where $g_{original}$ is the transition cost of original MDP. Let soln to these eqns be J_{π^*}

Step ②: Perform policy improvement.

$$T_{\pi_{new}} J_{\pi^*} = T J_{\pi^*}$$

$$\Rightarrow (T_{\pi_{new}} J_{\pi^*})(i) = \max_a E\left(g(i, a, j) + J_{\pi^*}(j) \right)$$

Get $J_{\pi_{new}}$

Get π_{new} & evaluate $J_{\pi_{new}}$.

if $J_{\pi_{new}} = J_{\pi^*}$ then π^* is the optimal policy

If $J_{new} < J_{\pi^+}$, π^+ is not optimal.

In order to find the new optimal policy we can do policy iteration with J_{new} as the initial proper policy.

The expected cost will also be obtained via policy iteration in policy evaluation step.