CS6700 RL ASSIGNMENT-2

BY: S. VISHAL
CH18B020

Declaration :

I pledge that I have not copied or given any unauthorized assistance on this assignment.

Acknowledgement :

1. Bertsekas DPOC Vol 1 chap 1 (MDP)
2. Class notes

BY: S. VISHAL
CH18B020

CS6700 RL ASSIGNMENT-2

(1) We know that finite-horizon MDPs can be converted to an equivalent SSP

Consider an MDP with state space X, N stages & Action space A

Let the states in the SSP be (i,k)

$i \rightarrow$ ith state in MDP, $i \in X$

$k \rightarrow$ kth stage in MDP, $k \in \{1, \ldots, N\}$

Terminal states : $(i, N) \quad \forall i \in X$

[Because finite-horizon MDP gets over after N stages]

Transitions : $(i,k) \xrightarrow{a} (j, k+1)$ with probability $P_{ij}(a)$

$a \in A_k(i)$

where $A_k(i)$ is the set of available actions in the MDP at stage k in state i

Costs : $g_{SSP}\big((i,N), a, (j,N)\big) = 0 \quad (\because \text{terminal})$

$$g_{SSP}\big((i,k), a, (j,k+1)\big) = g_{MDP}(i, a, j)$$

Now, we can perform our usual policy iteration on the SSP

One more thing to note is All Policies are Proper in this SSP.

Proof: Note that all transitions lead from $(i,k) \to (j,k+1)$ i.e. the second index denoting the stage keeps increasing till it reaches N.

$\Rightarrow$ Irrespective of policy, the terminal state will be reached in N steps.

[this is so, because the finite horizon MDP finishes in N stages]

Policy Iteration

(i) Start with a proper policy π_0 (all policies are proper so start with any policy).

(iv) (ii) Policy evaluation:

Solve $J_{\pi k} = T_{\pi k} J_{\pi k}$; $J_{\pi k}$ will be unique because it is SS & proper

(ii) $J_{\pi_k}(i,r) = \sum_j p((i,r), \bar{\pi}_k((i,r)), (j,r+1)) \times \left(g_{SSP}((i,r), \pi_k(i,r), (j,r+1)) + J_{\pi_k}(j,r+1) \right)$

$p((i,r), \pi_k(i,r), (j,r+1))$ is same as

$p(i, \pi_k(i,r), j)$ MDP in the rth stage

$g_{SSP}((i,r), \pi_k(i,r), (j,r+1)) = g_{MDP}(i, \pi_k(i,r), j)$ in the rth stage)

So we know p, π_k, g. & Just solve for J_{π_k}

(iii) Policy Improvement

$T_{\pi_{k+1}} J_{\pi_k} = T J_{\pi_k}$

i.e. $\pi_{k+1}(i,r) = \underset{a \in A_r(i)}{\arg\min} \sum p((i,r), \pi_k(i,r), (j,r+1)) \times \left(g((i,r), \pi_k(i,r), (j,r+1)) + J_{\pi_k}(j,r+1) \right)$

$A_r(i)$ denote set of all allowed actions in MDP at stage r

P, g, π_k are ~~same~~ defined in same way as policy evaluation step.

$J_{\pi k}$ is obtained from policy evaluation step

(iv) If $J_{\pi k+1}((i,r)) < J_{\pi k}((i,r))$

for at least one state (i,r),

~~continue (go to~~ go to step (ii) & Repeat.

Else stop.

Time complexity analysis

I. DP

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

$j \in X$ → totally n states.

To find the optimal action (a value) of a state at stage k, $n \times m$ operations required (n for each state)

[n multiplications, perform for m actions, choose minimum resulting value]

⇒ For each stage k, $n \times (n \times m) = n^2 m$

(∵ n states at each stage)

⇒ For N stages, $N n^2 m$

Computational requirements of DP: $\boxed{O(N n^2 m)}$

V. Policy Iteration

a. Policy evaluation - required per update,

for state (i, r), n multiplications needed

⇒ totally: $n \times N$ states

∴ $N n^2$ operations / policy evaluate update.

b. Policy update improvement.

For each state (i, r), perform Nn multiplication for m actions. = Nm.

⇒ for all states: $(Nn)(Nnm)$

$= (N n^2 m$ / iteration

(multiplication only for n states because $(i, r-m) \to (i, r-1)$ for other k stages the probability will be 0)

a. Policy ~~improvement~~ evaluation

Solve a system of linear equations with Nn variables

$\Rightarrow (Nn)^3$ operations / iteration

c. Total policies : m actions for each state in each step

$= (m)^{Nn}$

In PI worst case, we iterate over all policies

Worst case regd : $O\left((m)^{Nn}\left((Nn)^3 m + (Nn)^3\right)\right)$

So DP is computationally less intensive than PI for finite horizon MDPs.

(2) $n+1$ states : $\{0\} \cup \{1, \dots, n\}$

↙ Terminal state

$$R = \sum_{t=0}^{T-1} \gamma^t r(x_t), \quad \gamma \in (0,1)$$

Some after T time steps, the terminal state is reached, $r(x_T) = r(x_{T+1}) = \dots = \lim_{k \to \infty} r(x_k) = 0$ (r(0) = 0)

So we can add all those terms to the summation in R → it won't make any difference since all those terms are 0

$$\Rightarrow R = \sum_{t=0}^{T-1} \gamma^t r(x_t) + 0$$

$$= \sum_{t=0}^{T-1} \gamma^t (r(x_t)) + \sum_{t=T}^{\infty} \gamma^t \times 0)$$

$$= \sum_{t=0}^{T-1} \gamma^t r(x_t) + \sum_{t=T}^{\infty} \gamma^t r(0)$$

$\lim_{t \to \infty} \gamma^t \to 0 \quad \because \gamma < 1$

$$= \sum_{t=0}^{T-1} \gamma^t r(x_t) + \sum_{t=T}^{\infty} \gamma^t r(x_t)$$

$$r(x_t) = 0 \quad \forall \ t \geq T$$

$$\Rightarrow R = \sum_{t=0}^{\infty} \gamma^t r(x_t)$$

[did this manipulation because otherwise the sum ends at a random T]

a) $J(x) = E(R | x_0 = x)$

$$= E\left(\sum_{t=0}^{\infty} \gamma^t r(x_t) \;\middle|\; x_0 = x \right)$$

$$= E\left(r(x_0) + \sum_{t=1}^{\infty} \gamma^t r(x_t) \;\middle|\; x_0 = x \right)$$

$$= r(x) + E\left(\sum_{t=1}^{\infty} \gamma^t r(x_t) \;\middle|\; x_0 = x \right)$$

$$\left(\because E(r(x_0) | x_0 = x) = r(x) \right)$$

$$= \gamma E\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(x_{t+1}) \;\middle|\; x_0 = x \right) + r(x)$$

$$= \gamma \mathbb{E}\left(\mathbb{E}\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(x_t) \;\middle|\; x_0 = x,\ x_1 = x_1 \right) \right)$$

over x_1 ; over $x_2, x_3 \ldots$; $\mathbb{E}[r(x)]$

$$= r(x) + \gamma \mathbb{E}\left[\mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r(x_{t+1}) \;\middle|\; x_0 = x,\ x_1 = x_1' \right) \right]$$

[re-indexing — (1) the sum]

note that $P(x_2 \mid \{x_1 = x_1'\}, x_0 = x)$

$$= P(x_2 \mid x_1 = x_1')$$

i.e. future transitions from x_1 don't depend on x_0

[Markov property]

So the sequence of states $\{x_1 = x_1', x_2, x_3, \ldots\}$ depend only on $x_1 = x_1'$

can rewrite it as $\{x_1', x_2', \ldots\}$

So $\mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r(x_{t+1}) \;\middle|\; x_0 = x,\ x_1 = x_1' \right)$

$$= \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r(x_{t+1}') \;\middle|\; x_1 = x_1' \right)$$

can rewrite $\{x_2, x_3, x_4, \dots\}$ as

$$\{x_1', x_2', x_3' \dots\}$$

where transition to x_1' (equally x_2) is governed by x_0' (equally $x' \to \because x_1 = x'$)

$$\therefore E\left(\sum_{t=0}^{\infty} \gamma^t r(x_{t+1}) \mid x_0 = x, x_1 = x' \right)$$

$$= E\left(\sum_{t=0}^{\infty} \gamma^t r(x_{t+1}) \mid x_1 = x' \right)$$

$$= E\left(\sum_{t=0}^{\infty} \gamma^t r(x_t') \mid x_0' = x' = x' \right)$$

$$= E\left(\sum_{t=0}^{\infty} \gamma^t r(x_t') \mid x_0' = x' \right)$$

$$= J(x') \, E(R \mid x_0' = x')$$

$$= J(x') \quad \text{— (2)}$$

$$\therefore J(x) = r(x) + \gamma E\left(J(x_1') \mid x_0 = x \right)$$

$$\boxed{J(x) = r(x) + \gamma \sum_{x' \in X} P(x'|x) J(x')} \quad \text{— (3)}$$

$$M(x) = E(R^2 \mid x_0 = x)$$

$$= E\left(\left(\sum_{t=0}^{\infty} \gamma^t r(x_t)\right)^2 \mid x_0 = x\right)$$

$$= E\left(\left(r(x) + \sum_{t=1}^{\infty} \gamma^t r(x_t)\right)^2 \Big| x_0 = x\right)$$

$$= E\left(r^2(x) + \left(\sum_{t=1}^{\infty} \gamma^t r(x_t)\right)^2 + 2r(x)\sum_{t=1}^{\infty} \gamma^t r(x_t) \Big| x_0 = x\right)$$

$$= r^2(x) + 2r(x)\, E\left(\sum_{t=1}^{\infty} \gamma^t r(x_t) \Big| x_0 = x\right)$$

$$+ E\left(\left(\sum_{t=1}^{\infty} \gamma^t r(x_t)\right)^2 \Big| x_0 = x\right) \quad \text{---(4)}$$

expectation in the second term was then to be equal to $E(J(x_1^*) \mid x_0 = x)$.

$\downarrow$ over x_1^*

And by using markov property i.e. using the equivalence of

$$P(\{x_2, x_3, \ldots\} \mid x_0, x_1) = P(\{x_2, x_3, \ldots\} \mid x_1)$$

$$= \mathbb{E}\left(\left(\sum_{t=1}^{\infty} \gamma^t r(x_t) \right)^2 \mid x_0 = x \right)$$

$$= \gamma^2 \mathbb{E}\left(\mathbb{E}\left(\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(x_t) \right) \mid x_0 = x \right) \right) \quad x_1 = x'$$

$$= \gamma^2 \mathbb{E}\left(\mathbb{E}\left(\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(x_t) \right)^2 \mid x_0 = x, x_1 = x' \right) \right)$$

$$\downarrow \text{over } x_1$$

$$= \gamma^2 \mathbb{E}\left(\left(\mathbb{E}\left(\left(\sum_{t=0}^{\infty} \gamma^t r(x_{t+1}) \right) \mid x_1 = x' \right) \right) \mid x_0 = x \right)$$

$$= \gamma^2 \mathbb{E}\left(\left(\mathbb{E}\left(\left(\sum_{t=0}^{\infty} \gamma^t r(x'_t) \right)^2 \mid x'_0 = x' \right) \right) \mid x_0 = x \right)$$

$$= \gamma^2 \mathbb{E}\left(\sum_{t=0}^{M} J(x'_t) \mid x_0 = x \right)$$

$$= \gamma^2 \mathbb{E}\left(M(x') \mid x_0 = x \right) \quad \text{—— (5)}$$

(4) $\Rightarrow M(x) = r^2(x) + 2\gamma r(x) \mathbb{E}\left(J(x_1) \mid x_0 = x \right) + \gamma^2 \mathbb{E}\left(M(x_1) \mid x_0 = x \right)$

$\Rightarrow$ **$M(x) = r^2(x) + 2\gamma r(x) \sum_{x' \in X} P(x' \mid x) J(x') + \gamma^2 \sum_{x' \in X} P(x' \mid x) M(x')$** —— (6)

b) $V(x) = \mathrm{var}(R \mid x_0 = x)$

$= E(R^2 \mid x_0 = x) - \left(E(R \mid x_0 = x)\right)^2$

$= M(x) - \left(J(x)\right)^2$ ——— (7)

Use (3) & (6) here to subst for $M(x)$ & $J(x)$

$= r^2(x) + 2\gamma r(x) \sum P(x' \mid x) J(x')$

$+ \gamma^2 \sum P(x' \mid x) M(x') - \left(r(x) + \gamma \sum P(x' \mid x) J(x')\right)^2$

using $(a-b)^2 = a^2 + b^2 - 2ab$ and cancelling terms

$= \gamma^2 \sum P(x' \mid x) M(x') - \gamma^2 \left(\sum P(x' \mid x) J(x')\right)^2$

Add and subtract $\gamma^2 \sum P(x' \mid x) \left(J(x')\right)^2$

$= \gamma^2 \left(\sum_{\forall x' \in X} P(x' \mid x) J(x')^2 - \left(\sum_{\forall x' \in X} P(x' \mid x) J(x')\right)^2 \right)$

$+ \gamma^2 \left[\sum_{\forall x' \in X} P(x' \mid x) \left(M(x') - \left(J(x')\right)^2 \right) \right]$

$= \gamma^2 \psi(x) + \gamma^2 \sum_{\forall x' \in X} P(x' \mid x) V(x')$

(from eqn (7) $\mathrm{var}(x) = M(x) - \left(J(x)\right)^2$ & from definition of $\psi(x)$)

(3) a) contraction property $\|f(a) - f(b)\|_\infty \leq \beta \|a-b\|_\infty$

Consider, $\|f(y + aId) - f(y)\|_\infty$

by contraction property of f,

$\|f(y + aId) - f(y)\|_\infty \leq \beta \|y + aId - y\|_\infty$

$= \beta \|aId\|_\infty$

$\|Id\|_\infty = 1$

$\Rightarrow \|f(y + aId) - f(y)\|_\infty \leq \beta|a|$

$\Rightarrow \max_i |(f(y + aId) - f(y))(i)| \leq \beta|a|$

(definition of ∞ norm)

$\Rightarrow (f(y + aId) - f(y))(i) \leq \beta|a|$ — (1)

$\Rightarrow (f(y + aId) - f(y))(i) \geq -\beta|a|$ — (2)

(1) $\Rightarrow f(y + aId) \leq \beta \leq f(y) + \beta|a|Id$ — (3)

Note that $x \leq y + aId$

$\Rightarrow f(x) \leq f(y + aId)$ — (4) ($\because f(\cdot)$ is monotone)

(3) & (4) $\Rightarrow f(x) \leq f(y + aId) \leq f(y) + \beta|a|Id$

$$\boxed{\Rightarrow f(x) \leq f(y) + \beta|a|Id.}$$

b) ∵ Putting $a = 0$, we get

if $x \leq y + 0\,Id$ then $f(x) \leq f(y) + \beta(0) I_d$

$\Rightarrow x \leq y$ then $f(x) \leq f(y)$

∴ if $x \leq y \Rightarrow f(x) \leq f(y)$ —— (1)

This shows that $f(\cdot)$ is monotone.

Consider 2 vectors x & y

we can always find an a scalar a

such that $y - a\,Id \leq x \leq y + a\,Id$. —— (2)

$\left(a \geq \max_k \left(\max_i (y-x)(i_q) \,,\, \max_i (x-y)(i) \right) \right.$

i.e. $a \geq \max_i \left(|(y-x)(i)| \right.$

$\Rightarrow a \geq \|y - x\|_\infty \left. \right)$

So for finite y & x we can always find some scalar a.

Now. apply

$y - a\,Id \leq x$

$\Rightarrow y \leq x + a\,Id$

Apply the given property,

$$f(y) \leq f(x) + \beta|a|I_d.$$

$$\Rightarrow f(x) - f(y) \geq -\beta|a|I_d \quad \text{---(3)}$$

Also $x \leq y + a1_d \Rightarrow f(x) \leq f(y) + \beta|a|I_d$

$$\Rightarrow f(x) - f(y) \leq \beta|a|I_d \quad \text{---(4)}$$

(3) & (4) $\Rightarrow \|f(x) - f(y)\|_\infty \leq \beta|a| \quad \forall a$ ---(5)

$(\because -(f(x) - f(y)) \leq \beta|a|I_d$ (event will)
& $(f(x) - f(y) \leq \beta|a|I_d$
· so all elements' absolute value $< \beta|a|)$

We showed that $a \geq \|x - y\|_\infty$

· By putting $a = \|x - y\|_\infty$

(5) $\Rightarrow \|f(x) - f(y)\|_\infty \leq \beta\|x - y\|_\infty$ ---(6)

$\therefore$ $f(\cdot)$ is a contraction wrt supremum norm

(1) & (6) $\Rightarrow$ $f(\cdot)$ is a monotone contraction

To prove :

c) $x - \frac{1}{1-\beta}\|f(x)-x\|_\infty \mathbb{1}_d \le f(x) - \frac{\beta}{1-\beta}\|f(x)-x\|_\infty \mathbb{1}_d$

— (P1)

$f(x) + \frac{\beta}{1-\beta}\|f(x)-x\|_\infty \mathbb{1}_d \le x + \frac{1}{1-\beta}\cdot\|f(x)-x\|_\infty \mathbb{1}_d$

— (P2)

$f(x) - \frac{\beta}{1-\beta}\|f(x)-x\|_\infty \mathbb{1}_d \le x^* \le f(x) + \frac{\beta}{1-\beta}\|f(x)-x\|_\infty \mathbb{1}_d$

— (P3)

the above are the inequalities that had to be proved.

$x - f(x) \le \|x - f(x)\|_\infty \mathbb{1}_d$ — (1)

$x - f(x) \ge -\|x - f(x)\|_\infty \mathbb{1}_d$ — (2)

} (definition of max norm)

(1) $\Rightarrow x - f(x) \le \left(\frac{1-\beta}{1-\beta}\|x - f(x)\|_\infty\right)\mathbb{1}_d$

(multiply & divide by β)

$\Rightarrow x \le f(x) + \left(\frac{1}{1-\beta} - \frac{\beta}{1-\beta}\right)\left(\|x - f(x)\|_\infty\right)\mathbb{1}_d.$

$$\Rightarrow f(x) \; x - \frac{\|x - f(x)\|_\infty I_d}{1-\beta} \leq f(x) - \frac{\beta}{1-\beta}\|x - f(x)\|_\infty$$

$$\Rightarrow x - \frac{\|x - f(x)\|_\infty I_d}{1-\beta} \leq f(x) - \frac{\beta}{1-\beta}\|x - f(x)\|_\infty I_d \quad \text{— (3)}$$

P1 proved !

Similarly using (2),

$$x - f(x) \geq -\|x - f(x)\|_\infty I_d \left(\frac{1-\beta}{1-\beta}\right)$$

$$\Rightarrow x \geq f(x) - \left(\frac{1}{1-\beta} - \frac{\beta}{1-\beta}\|x - f(x)\|_\infty I_d\right)$$

$$\Rightarrow f(x) + \frac{\beta}{1-\beta}\|f(x) - x\|_\infty I_d \leq x + \frac{1}{1-\beta}\|f(x) - x\|_\infty I_d \quad \text{— (4)}$$

P2 proved !

Consider $\|x - x^*\|_\infty$.

$$\|x - x^*\|_\infty = \|x - f(x) - x^* + f(x)\|$$

$$= \|-(f(x) - x) + (f(x) - x^*)\|_\infty$$

$$\leq \|f(x) - x\| + \|f(x) - x^*\|_\infty \quad \text{(triangle inequality)}$$

But $\|f(x) - f(x)^*\|_\infty \leq \beta\|x - x^*\|_\infty = \|f(x) - f x^*\|_\infty \leq \beta\|x - x^*\|_\infty$

$\because$ x^* is fixed pt. & $f(\cdot)$ is contraction

$$\Rightarrow \|x - x^k\|_\infty \leq \|f(x) - x\|_\infty + \beta \|x - x^k\|_\infty$$

$$\Rightarrow \|x - x^k\|_\infty \leq \frac{1}{1-\beta} \|f(x) - x\|_\infty$$

So $\max_i |x - x^k|(i) \leq \frac{1}{1-\beta} \|f(x) - x\|_\infty$

$$\Rightarrow x - x^k \leq + \frac{1}{1-\beta} \|f(x) - x\|_\infty I_d \quad (5)$$

& $$x - x^k \geq - \frac{1}{1-\beta} \|f(x) - x\|_\infty I_d \quad (6)$$

$$\Rightarrow x \leq x^k + \left(\frac{1}{1-\beta} \|f(x) - x\|_\infty\right) I_d \quad (5)$$

and $$x^k \leq x + \left(\frac{1}{1-\beta} \|f(x) - x\|_\infty\right) I_d \quad (6)$$

these eqns are of the form $x \leq y + a I_d$

where $a = \frac{1}{1-\beta} \|f(x) - x\|_\infty$

Applying the property from part (a) (∵ f is monotone contraction)

$$(5) \Rightarrow f(x) \leq f(x^k) + \beta \left| \frac{1}{1-\beta} \|f(x) - x\|_\infty \right| I_d.$$

But $f(x^k) = x^k$ & $\frac{1}{1-\beta} \|f(x) - x^k\|_\infty > 0$

$$\Rightarrow f(x) \leq x^k + \frac{\beta}{1-\beta} \|f(x) - x^k\|_\infty I_d$$

One again applying $x \le y + a I_d \Rightarrow f(x) \le f(y) + \beta a I_d$

⑥ $\Rightarrow f(x^*) \le f(x) + \beta \left(\frac{1}{1-\beta} \|f(x) - x\|_\infty \right) I_d$

$\Rightarrow x^* \le f(x) + \frac{\beta}{1-\beta} \|f(x) - x\|_\infty I_d$ — ⑧

⑦ $\Rightarrow x^* \ge f(x) - \frac{\beta}{1-\beta} \|f(x) - x\|_\infty I_d$ — ⑨

⑧ & ⑨ $\Rightarrow f(x) - \frac{\beta}{1-\beta} \|f(x) - x\|_\infty I_d \le x^*$

$\le f(x) + \frac{\beta}{1-\beta} \|f(x) - x\|_\infty I_d$ — ⑩

P_3 is proved !

& P_1 & P_2 & P_3 $\Rightarrow$

$x - \frac{1}{1-\beta} \|f(x) - x\|_\infty I_d \le f(x) - \frac{\beta}{1-\beta} \|f(x) - x\|_\infty I_d$

$\le x^* \le f(x) + \frac{\beta}{1-\beta} \|f(x) - x^*\|_\infty I_d$

$\le x + \frac{1}{1-\beta} \|f(x) - x\|_\infty I_d$

(eqns ③, ⑥, ⑫)