« Alireza Gargoori Motlagh _ 98102176 »

« Homework 1 _ Reinforcement Learning »

1

a) $R(\cdot) \geq 0$

$$V_k^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{k} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$

$$\leq \sum_{t=0}^{k} \gamma^t R_{max} = \frac{1-\gamma^k}{1-\gamma} R_{max}$$

$$\Rightarrow V_k^*(s) \leq \frac{1-\gamma^k}{1-\gamma} R_{max}$$

b)

$$V_{k+1}^{\pi}(s) = V_k^*(s) + \gamma \sum_{s'} \overbrace{R(s, s', \pi(s))}^{\geq 0} \geq V_k^*(s)$$

$$V_{k+1}^*(s) = \max_{\pi} \left(V_k^*(s) + \gamma \sum_{s'} R(s, s', \pi(s)) \right) \geq V_{k+1}^{\pi}(s) \geq V_k^*(s)$$

$\Rightarrow$ $V_k^*(s)$ is a bounded and increasing sequence in k $\Rightarrow$

$V^*(s)$ converges.

c) $$V_k(s) = \max_a \sum_{s'} \mathbb{P}(s' \mid s,a)\left(R(s,s',a) + \gamma V_{k-1}(s')\right)$$

$$\Longrightarrow V_\infty(s) = \lim_{k\to\infty} V_k(s) = \lim_{k\to\infty} \max_a \sum_{s'} \mathbb{P}(s' \mid s,a)\left(R(s,s',a) + \gamma V_{k-1}(s')\right)$$

$$\overset{*}{=} \max_a \sum_{s'} \mathbb{P}(s' \mid a,s)\left(R(s,s',a) + \gamma \lim_{k\to\infty} V_{k-1}(s')\right)$$

$$= \max_a \sum_{s'} \mathbb{P}(s' \mid a,s)\left(R(s,s',a) + \gamma V_\infty(s')\right)$$

$$\Longrightarrow V_\infty(s) = \max_a \sum_{s'} \mathbb{P}(s' \mid a,s)\left(R(s,s',a) + \gamma V_\infty(s')\right)$$

which is the bellman optimality eqation $\Rightarrow V^*(s) = V_\infty(s)$

$$\boxed{V^*(s) = \max_a \sum_{s'} \mathbb{P}(s' \mid a,s)\left(R(s,s',a) + \gamma V^*(s')\right)}$$

* using convergence & continuity of $V_k(s)$ as a function of $V_{k-1}(s')$

d) $R'(s,s',a) = R(s,s',a) + r_0$

$$V_k^{*(new)}(s) = \max_\pi \mathbb{E}\left[\sum_{t=0}^{k} \gamma^t R'(s_t, s_{t+1}, a_t) \mid \pi, s_0 = s\right] =$$

$$\max_\pi \mathbb{E}\left[\sum_{t=0}^{k} \gamma^t \left(R(s_t, s_{t+1}, a_t) + r_0\right) \mid \pi, s_0 = s\right]$$

$$= \max_\pi \mathbb{E}\left[\sum_{t=0}^{k} \gamma^t R(s_t, s_{t+1}, a_t) \mid \pi, s_0 = s\right] + \mathbb{E}\left[\sum_{t=0}^{k} \gamma^t r_0 \mid \pi, s_0 = s\right]$$

$$= \max_\pi \mathbb{E}\left[\sum_{t=0}^{k} \gamma^t R(s_t, s_{t+1}, a_t) \mid \pi, s_0 = s\right] + r_0 \frac{1-\gamma^k}{1-\gamma}$$

$$= r_0 \frac{1-\gamma^k}{1-\gamma} + \max_\pi \mathbb{E}\left[\sum_{t=0}^{k} \gamma^t R(s_t, s_{t+1}, a_t) \mid \pi, s_0 = s\right]$$

$$= r_0 \frac{1-\gamma^k}{1-\gamma} + V_k^*(s) \Longrightarrow$$

$$\boxed{V_k^{*(new)}(s) = V_k^*(s) + r_0 \frac{1-\gamma^k}{1-\gamma}}$$

$$\pi_k^{*(new)}(s) = \arg\max_\pi V_k^{*(new)}(s) = \arg\max_\pi V_k^*(s) + \underbrace{r_0 \frac{1-\gamma^k}{1-\gamma}}_{}$$ second term is constant

$$= \arg\max_\pi V_k^*(s) = \pi_k^*(s) \Longrightarrow$$

$$\boxed{\pi_k^{*(new)}(s) = \pi_k^*(s)}$$

So if $\min_{s,s',a} R(s,s',a) = r^- < 0$, we can define a new reward for each state as $R'(s,s',a) = R(s,s',a) + |r^-|$ and we have proved that optimal policy remains unchanged. Also, since we have proved Value Iteration converges to the optimal Value function for $R' \geq 0$, we conclude that for any reward, Value Iteration converges to optimal Value function and the optimal policy does not change.

e) Since if we have terminating states, we are not able to write the sum as we wrote in the previous part to prove the following theorem about the negative rewards as well.

2

a) $$\pi_{t+1}(s) = \arg\max_{a} \sum_{s'} \mathbb{P}(s'|\pi_t(s), s)\left(R(s', \pi_t(s), s) + \gamma V_\infty^{\pi_t}(s')\right)$$

$$V_\infty^{\pi_{t+1}}(s) \geq V_\infty^{\pi_t}(s) \quad (\text{as assumed}) \quad \Longrightarrow$$

$$V_\infty^{\pi_{t+1}}(s) = \sum_{s'} \mathbb{P}(s'|\pi_{t+1}(s), s)\left(R(s', \pi_{t+1}(s), s) + \gamma V_\infty^{\pi_{t+1}}(s')\right) \geq$$

$$V_\infty^{\pi_t}(s) = \sum_{s'} \mathbb{P}(s'|\pi_t(s), s)\left(R(s', \pi_t(s), s) + \gamma V_\infty^{\pi_t}(s')\right)$$

$$V_\infty^{\pi_{t+1}}(s) = V_\infty^{\pi_t}(s) \Longrightarrow \sum_{s'} \gamma V_\infty^{\pi_{t+1}}(s')\left(\mathbb{P}(s'|\pi_{t+1}(s), s) - \mathbb{P}(s'|\pi_t(s), s)\right)$$

$$+ \sum_{s'} \mathbb{P}(s'|\pi_{t+1}(s), s) R(s', \pi_{t+1}(s), s) - \mathbb{P}(s'|\pi_t(s), s) R(s', \pi_t(s), s)$$

$$= 0 \quad \overset{R(\cdot)}{\underset{\geq 0}{\Longleftrightarrow}} \quad \mathbb{P}(s'|\pi_{t+1}(s), s) - \mathbb{P}(s'|\pi_t(s), s) = 0 \quad \Longrightarrow$$

$$\boxed{\pi_{t+1}(s) = \pi_t(s) \quad \forall s \in S}$$

which is the definition of convergence.

b) $$n_{\text{pol. eval.}} \leq (\text{number of actions})^{(\text{number of states})} = |A|^{|S|}$$

So the number of policy evaluation step is at most the number of different policies, which is $|A|^{|S|}$.

Since in each step Value function increases, after iteration as many as times required, at most $|A|^{|S|}$, by definition, in every step policy improves. This means that a given policy can be encountered at most once $\Longrightarrow$ Gaurantee to converge.

Also, at convergence $\pi_{t+1}(s) = \pi_t(s) \quad \forall s$. This simply leads to

$$\forall s: \; U^{\pi_t}(s) = \max_a \sum_{s'} P(s' | \pi_t(s), s)\left(R(s, \pi_t(s), s') + \gamma U^{\pi_t}(s')\right)$$

Hence, $U^{\pi_t}(s)$ satisfies Bellman Optimality eq. $\Longrightarrow \forall s: \; U^{\pi_t}(s) = V^*(s)$

$$\pi_t(s) = \pi^*(s)$$

c) Policy iteration generally converges faster than Value iteration ;

Since the VI runs through all possible actions at each iteration to find the maximum action value *, but PI only has an arg max in its policy improvement steps & does not need to iterate over all actions in the policy evaluation step. Both algorithm are gauranteed to converge but PI is less computationally expensive .

1. Initialization
   $v(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Repeat
   - $\Delta \leftarrow 0$
   - For each $s \in \mathcal{S}$:
     - $temp \leftarrow v(s)$
     - $v(s) \leftarrow \sum_{s'} p(s'|s,\pi(s))\Big[r(s,\pi(s),s') + \gamma v(s')\Big]$
     - $\Delta \leftarrow \max(\Delta, |temp - v(s)|)$

   until $\Delta < \theta$ (a small positive number)

3. Policy Improvement
   $policy\text{-}stable \leftarrow true$
   For each $s \in \mathcal{S}$:
   - $temp \leftarrow \pi(s)$
   - $\pi(s) \leftarrow \arg\max_a \sum_{s'} p(s'|s,a)\Big[r(s,a,s') + \gamma v(s')\Big]$
   - If $temp \neq \pi(s)$, then $policy\text{-}stable \leftarrow false$

   If $policy\text{-}stable$, then stop and return v and π; else go to 2

Figure 4.3: Policy iteration (using iterative policy evaluation) for v_*. This algorithm has a subtle bug, in that it may never terminate if the policy continually switches between two or more policies that are equally good. The bug can be fixed by adding additional flags, but it makes the pseudocode so ugly that it is not worth it. :-)

finding optimal value function

Initialize array v arbitrarily (e.g., $v(s) = 0$ for all $s \in \mathcal{S}^+$)

Repeat
- $\Delta \leftarrow 0$
- For each $s \in \mathcal{S}$: *
  - $temp \leftarrow v(s)$
  - $v(s) \leftarrow \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma v(s')]$
  - $\Delta \leftarrow \max(\Delta, |temp - v(s)|)$

until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, π, such that
$\boxed{\pi(s) = \arg\max_a \sum_{s'} p(s'|s,a)\Big[r(s,a,s') + \gamma v(s')\Big]}$

Figure 4.5: Value iteration.

one policy update (extract policy from the optimal value function

(Image taken from Stackoverflow (which itself is SB book))

d) $$V_0^{\pi_{t+1}}(s) = \sum_{s'} \mathbb{P}(s' \mid \pi_{t+1}(s), s)\left[R(s', \pi_{t+1}(s), s) + \gamma V_\infty^{\pi_t}(s')\right]$$

Since $V_0^{\pi_{t+1}}(s)$ is derived from $\pi_{t+1}(s)$ which is an improved policy than $\pi_t(s)$ $\forall s$, it is simple to conclude $V_0^{\pi_{t+1}}(s) \geq V_\infty^{\pi_t}(s)$

Since we have the assumption of : $\forall s$ $V_{k+1}^{\pi_{t+1}}(s) \geq V_k^{\pi_{t+1}}(s)$

we can infer $V_\infty^{\pi_{t+1}}(s) \geq V_0^{\pi_{t+1}}(s)$.

Also, we stated that $V_0^{\pi_{t+1}}(s) \geq V_\infty^{\pi_t}(s) \Longrightarrow$

$$\boxed{V_\infty^{\pi_{t+1}}(s) \geq V_0^{\pi_{t+1}}(s) \geq V_\infty^{\pi_t}(s)}$$

We can have the same method for the proof $V_\infty^{\pi_{t+1}}(s) \geq V_\infty^{\pi_t}(s)$ for the unchanged case as well.

3

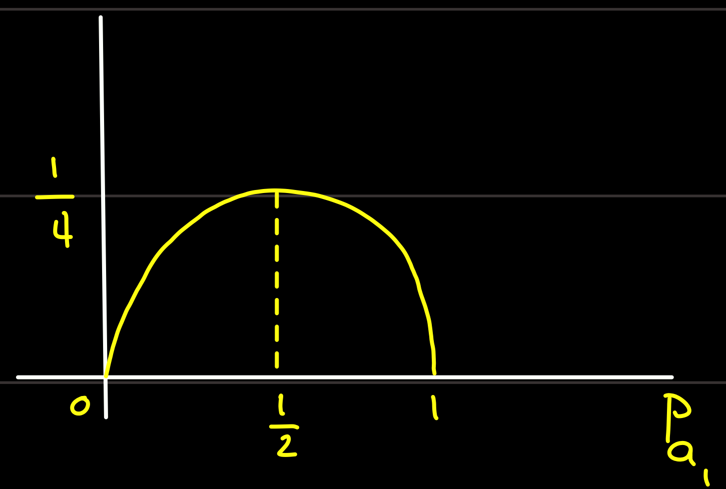
a) If we add the gini index to our objective function, we know for a 2-action state, we know the maximum will occur when

$P_{a_1} = P_{a_2} = \frac{1}{2}$:

$$\max \sum_{i=1}^{2} P_{a_i}(1-P_{a_i})$$

$$\text{s.t.} \quad P_{a_1} + P_{a_2} = 1$$

$$\Rightarrow \sum_{i=1}^{2} P_{a_i}(1-P_{a_i}) = P_{a_1}(1-P_{a_1}) + (1-P_{a_1})(1-1+P_{a_1}) = 2P_{a_1}(1-P_{a_1})$$



As we see, Gini Index gets larger when the difference in the probability distribution is minimized.

b)

$$\max_{\pi_A} \ \mathbb{E}_{\pi_A}\left[r(a)\right] + \beta \, Gini(\pi_A)$$

$$\text{s.t.} \quad \sum_{a_i \in A} \pi_{a_i} = 1$$

$$\pi_A \geq 0 \quad (\pi_a \geq 0 \quad \forall a \in A)$$

$$\mathcal{L} = \mathbb{E}_{\pi_A}\left[r(a)\right] + \beta \, Gini(\pi_A) + \lambda\left(1 - \sum_{a_i \in A} \pi_{a_i}\right) - \sum_{a_i \in A} \nu_{a_i} \pi_{a_i}$$

$$\nu_i \geq 0 \quad \forall i \in \{1, \dots, |A|\}$$

c)

$$\mathcal{L} = \sum_{a \in G} \pi_a \, r(a) + \beta \sum_{a \in G} \pi_a (1 - \pi_a) + \lambda\left(1 - \sum_{a \in G} \pi_a\right) - \sum_{a \in G} \nu_a \pi_a$$

Complementry Slackness: $\nu_a \pi_a = 0 \quad \forall a \in G \xrightarrow{\pi_a \neq 0} \nu_a = 0 \rightarrow \vec{\nu} = 0$

$$\frac{\partial \mathcal{L}}{\partial \pi_a} = 0 \Rightarrow r(a) + \beta\left((1 - \pi_a) - \pi_a\right) - \lambda = 0 \Rightarrow$$

$$r(a) + \beta(1 - 2\pi_a) = \lambda \qquad \forall a \in G$$

$$\Rightarrow \pi_a = \frac{\beta - \lambda + r(a)}{2\beta} \qquad \forall a \in G \qquad (1)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \Rightarrow \sum_{a \in G} \pi_a = 1 \qquad (2)$$

$$\xrightarrow[(2)]{(1)\,\&} \sum_{a\in G} \frac{\beta-\lambda+r(a)}{2\beta} = 1 \Rightarrow \sum_{a\in G} \beta-\lambda+r(a) = 2\beta \Rightarrow$$

$$|G|(\beta-\lambda) + \sum_{a\in G} r(a) = 2\beta \Rightarrow$$

$$\boxed{\lambda = \frac{\beta(|G|-2) + \sum_{a\in G} r(a)}{|G|}} \quad *$$

where $|G|$ is the size of the set G which is the number of actions with non-zero probability.

$$\xrightarrow[(1)]{*\,\&} \quad \pi_a = \frac{1}{2\beta}\left(\beta - \frac{\beta(|G|-2) + \sum_{a\in G} r(a)}{|G|} + r(a)\right)$$

$$\Rightarrow \boxed{\pi_a = \frac{2\beta - \sum_{a\in G} r(a) + |G|\, r(a)}{2\beta|G|}}$$

d)

$$\max_{\pi_A} \ \mathbb{E}_{\pi_A}[r(a)] + \beta \, \mathrm{Gini}(\pi_A)$$

$$\text{s.t.} \quad \sum_{a \in A} \pi_a = 1$$

$$\pi_a \geq 0 \quad \forall a \in A$$

$$f(\pi_A) = \sum_{a \in A} \pi_a r(a) = \beta \sum_{a \in A} \pi_a (1 - \pi_a) = \sum_{a \in A} \pi_a r(a) + \beta \underbrace{\sum_{a \in A} \pi_a}_{=1} - \beta \sum_{a \in A} \pi_a^2$$

$$= \beta + \sum_{a \in A} \pi_a r(a) - \beta \sum \pi_a^2 = \beta + \pi_A^T r(A) - \beta \pi_A^T \pi_A$$

β is a constant; so we can rewrite the optimization problem as:

$$\Longrightarrow \quad \min_{\pi_A} \ \beta \pi_A^T \pi_A - \pi_A^T r(A)$$

$$\text{s.t.} \quad 1^T \pi_A = 1$$

$$\pi_A \geq 0$$

which is a QP and can be solved efficiently using qp-solvers.

After solving, we have: $G: \{a \in A\} \mid \pi_a \neq 0 \}$

4

a)

$$\begin{cases} E_{-1}(s) = 0 \\ E_t(s) = \gamma\lambda E_{t-1}(s) + I_{ss_t} \end{cases} \qquad I_{ss_t} = \begin{cases} 0 & s \neq s_t \\ 1 & s = s_t \end{cases}$$

$$E_t(s) = \gamma\lambda\left(\gamma\lambda E_{t-2}(s) + I_{ss_{t-1}}\right) + I_{ss_t} =$$

$$\gamma\lambda\left(\gamma\lambda\left(\gamma\lambda E_{t-3}(s) + I_{ss_{t-2}}\right) + I_{ss_{t-1}}\right) + I_{ss_t} = \dots$$

$$= (\gamma\lambda)^3 E_{t-3} + (\gamma\lambda)^2 I_{ss_{t-2}} + \gamma\lambda I_{ss_{t-1}} + (\gamma\lambda)^0 I_{ss_t} = \dots$$

$$= (\gamma\lambda)^{t+1} \underbrace{E_{-1}(s)}_{0} + \sum_{k=0}^{t} (\gamma\lambda)^k I_{ss_{t-k}} = \sum_{k=0}^{t} (\gamma\lambda)^{t-k} I_{ss_k}$$

$$\Longrightarrow \boxed{E_s(t) = \sum_{k=0}^{t} (\gamma\lambda)^{t-k} I_{ss_k}}$$

b) $$\sum_{t=0}^{T-1} \Delta V_t^{TD}(s) = \sum_{t=0}^{T-1} \alpha \delta_t E_t(s) = \sum_{t=0}^{T-1} \alpha \delta_t \sum_{k=0}^{t} (\gamma\lambda)^{t-k} I_{ss_k}$$

$$= \sum_{k=0}^{T-1} \alpha \sum_{t=0}^{k} (\gamma\lambda)^{k-t} I_{ss_t} \delta_k = \sum_{t=0}^{T-1} \alpha \sum_{k=t}^{T-1} (\gamma\lambda)^{k-t} I_{ss_t} \delta_k =$$

$$\sum_{t=0}^{T-1} \alpha I_{ss_t} \sum_{k=t}^{T-1} (\gamma\lambda)^{k-t} \delta_k \Rightarrow$$

$$\boxed{\sum_{t=0}^{T-1} \Delta V_t^{TD}(s) = \sum_{t=0}^{T-1} \alpha I_{ss_t} \sum_{k=t}^{T-1} (\gamma\lambda)^{k-t} \delta_k}$$

c) $$\frac{1}{\alpha} \Delta V_t^{\lambda}(s_t) = R_t^{\lambda} - V_t(s_t) = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} R_t^{(n)} - V_t(s) =$$

$$- V_t(s_t) + (1-\lambda)\lambda^0 [r_{t+1} + \gamma V_t(s_{t+1})] +$$
$$(1-\lambda)\lambda [r_{t+1} + \gamma r_{t+2} + \gamma^2 V_t(s_{t+2})] +$$
$$(1-\lambda)\lambda^2 [r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 V_t(s_{t+3})] + \dots$$

$$= \left[(1-\lambda)\sum_{n=0}^{\infty} \lambda^n\right] r_{t+1} + \left[\gamma(1-\lambda)\sum_{n=1}^{\infty} \lambda^n\right] r_{t+2} + \dots + (1-\lambda)\sum_{n=1}^{\infty} \gamma^n \lambda^{n-1} V_t(s_{t+n})$$
$$- V_t(s_t)$$

$$* \quad (1-\lambda)\sum_{n=i}^{\infty}\lambda^n \underset{\lambda<1}{=} (1-\lambda)\times\frac{\lambda^i}{1-\lambda} = \lambda^i \Longrightarrow$$

$$\frac{1}{\alpha}\Delta V_t^{\lambda}(s_t) = -V_t(s_t) + (\gamma\lambda)^0\left[r_{t+1} + \gamma V_t(s_{t+1}) - \gamma\lambda V_t(s_{t+1})\right]$$
$$+ (\gamma\lambda)^1\left[r_{t+2} + \gamma V_t(s_{t+2}) - \gamma\lambda V_t(s_{t+2})\right]$$
$$+ (\gamma\lambda)^2\left[r_{t+3} + \gamma V_t(s_{t+3}) - \gamma\lambda V_t(s_{t+3})\right]$$
$$+ \cdots$$

taking $-V_t(s_t)$ into the first parantheses & replacing $-\gamma\lambda V_t(s_{t+1})$ into to next & etc. :

$$\frac{1}{\alpha}\Delta V_t^{\lambda}(s_t) = (\gamma\lambda)^0\left[r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t)\right]$$
$$+ (\gamma\lambda)^1\left[r_{t+2} + \gamma V_t(s_{t+2}) - V_t(s_{t+1})\right]$$
$$+ (\gamma\lambda)^2\left[r_{t+3} + \gamma V_t(s_{t+3}) - V_t(s_{t+2})\right] + \cdots$$

$\Rightarrow$

$$\boxed{\frac{1}{\alpha}\Delta V_t^{\lambda}(s_t) = \sum_{k=t}^{\infty}(\gamma\lambda)^{k-t}\left[r_{k+1} + \gamma V_t(s_{k+1}) - V_t(s_k)\right]}$$

d) In order to have equality we must have $V_t(s)$ be fixed for all s within an episode and not to be updated as soon as an increment is computed within an episode. Hence, Offline Update acheives this equality due to the previous sentence. Therefore, in Offline update we have:

$$\frac{1}{\alpha} \Delta V_t^{\lambda}(s_t) = \sum_{k=t}^{\infty} (\gamma\lambda)^{k-t} \delta_k = \sum_{k=t}^{T-1} (\gamma\lambda)^{k-t} \delta_k$$

We can change the index from ∞ to $T-1$, since all δ_k after the terminal state are zero:

$$\sum_{t=0}^{T-1} \Delta V_t^{TD}(s) = \sum_{t=0}^{T-1} \alpha I_{ss_t} \sum_{k=t}^{\infty} (\gamma\lambda)^{k-t} \delta_k = \sum_{t=0}^{T-1} \Delta V_t^{\lambda}(s_t) I_{ss_t}$$

Offline Forward $TD(\lambda) \equiv$ Backward $TD(\lambda)$