Reinforcement Learning: Assignment #3

Due on Sunday, Mai 07, 2023

Group 4 - Abu El Komboz, Tareq 3405686 | Jain, Likhit 3678905 | Wurm, Marcel 3695946

Task 1

Proofs (5 Points)

(a) Show that the Bellman ***optimality*** *operator T is a γ-contraction. This is similar to but not the same as the Bellman* ***expectation backup*** *operator from lecture 3 slide 20. Be able to explain all the steps! (3P)*

$(Tv)(s) = \max_a \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v(s')\Big]$

$$
\begin{aligned}
&\Big|\Big|(Tv)(s) - (Tv')(s)\Big|\Big|_\infty \\
&= \Big|\Big|\max_a \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v(s')\Big] - \max_{a'} \sum_{s',r} p(s',r|s,a')\Big[r + \gamma v'(s')\Big]\Big|\Big|_\infty \\
&\leq \Big|\Big|\max_a \Bigg[\sum_{s',r} p(s',r|s,a)\Big[r + \gamma v(s')\Big] - \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v'(s')\Big]\Bigg]\Big|\Big|_\infty \\
&= \Big|\Big|\max_a \sum_{s',r} \Bigg[p(s',r|s,a)\Big[r + \gamma v(s')\Big] - p(s',r|s,a)\Big[r + \gamma v'(s')\Big]\Bigg]\Big|\Big|_\infty \\
&= \Big|\Big|\max_a \sum_{s',r} \Bigg[p(s',r|s,a)r + p(s',r|s,a)\gamma v(s') - p(s',r|s,a)r - p(s',r|s,a)\gamma v'(s')\Bigg]\Big|\Big|_\infty \\
&= \Big|\Big|\max_a \sum_{s',r} \Bigg[p(s',r|s,a)\gamma v(s') - p(s',r|s,a)\gamma v'(s')\Bigg]\Big|\Big|_\infty \\
&= \Big|\Big|\max_a \sum_{s',r} p(s',r|s,a)\gamma\Big[v(s') - v'(s')\Big]\Big|\Big|_\infty \\
&= \gamma\Big|\Big|\max_a \sum_{s',r} p(s',r|s,a)\Big[v(s') - v'(s')\Big]\Big|\Big|_\infty \\
&= \gamma \max_s \Big|\max_a \sum_{s',r} p(s',r|s,a)\Big[v(s') - v'(s')\Big]\Big| \\
&\leq \gamma \max_s \Big|\max_a \sum_{s',r} p(s',r|s,a) \max_s \Big[v(s) - v'(s)\Big]\Big| \\
&= \gamma \max_s \Big|\max_a \max_s \Big[v(s) - v'(s)\Big]\Big| \\
&= \gamma \max_s \Big|\max_s \Big[v(s) - v'(s)\Big]\Big| \\
&= \gamma \max_s \Big|v(s) - v'(s)\Big| \\
&= \gamma ||v(s) - v'(s)||_\infty
\end{aligned}
$$

Task 1 continued on next page...

(b) Assuming a general finite MDP (S, A, R, p, γ) where rewards are bounded: $r \in [r_{min}, r_{max}]$ for all $r \in R$. Prove the following equations. (2P)

1. $\frac{r_{\text{min}}}{1-\gamma} \leq v(s) \leq \frac{r_{\text{max}}}{1-\gamma}$

$$\begin{aligned}
v(s) &= \max_a \sum_{s',r} p(s', r|s, a)\Big[r + \gamma v(s')\Big] \\
&= \max_a \sum_{s'} p(s'|s, a)\Big[r(s, a, s') + \gamma v(s')\Big] \\
&= \max_a \sum_{s'} \Big[p(s'|s, a)r(s, a, s') + p(s'|s, a)\gamma v(s')\Big] \\
&= \max_a \Big[\sum_{s'} p(s'|s, a)r(s, a, s') + \sum_{s'} p(s'|s, a)\gamma v(s')\Big] \\
&= \max_a \Big[\sum_{s'} p(s'|s, a)r(s, a, s') + \sum_{s'} p(s'|s, a)\gamma v(s')\Big] \\
&\leq \max_a \Big[\max_{s'} r(s, a, s') + \gamma v(s)\Big] \\
&= \max_a \Big[r_{\text{max}} + \gamma v(s)\Big] \\
&= r_{\text{max}} + \gamma v(s)
\end{aligned}$$

$$\begin{aligned}
v(s) &\leq r_{\text{max}} + \gamma v(s) \qquad | - \gamma v(s) \\
v(s) - \gamma v(s) &\leq r_{\text{max}} \\
(1-\gamma)v(s) &\leq r_{\text{max}} \qquad | : (1-\gamma) \\
v(s) &\leq \frac{r_{\text{max}}}{1-\gamma}
\end{aligned}$$

$$
\begin{aligned}
r_{\text{min}} + \gamma v(s) &= \max_a \Big[r_{\text{min}} + \gamma v(s) \Big] \\
&= \max_a \Big[\min_{s'} r(s,a,s') + \gamma v(s) \Big] \\
&\leq \max_a \Big[\sum_{s'} p(s'|s,a) r(s,a,s') + \sum_{s'} p(s'|s,a) \gamma v(s') \Big] \\
&= \max_a \Big[\sum_{s'} p(s'|s,a) r(s,a,s') + \sum_{s'} p(s'|s,a) \gamma v(s') \Big] \\
&= \max_a \sum_{s'} \Big[p(s'|s,a) r(s,a,s') + p(s'|s,a) \gamma v(s') \Big] \\
&= \max_a \sum_{s'} p(s'|s,a) \Big[r(s,a,s') + \gamma v(s') \Big] \\
&= \max_a \sum_{s',r} p(s',r|s,a) \Big[r + \gamma v(s') \Big] \\
&= v(s)
\end{aligned}
$$

$$
\begin{aligned}
r_{\text{min}} + \gamma v(s) &\leq v(s) \qquad | - \gamma v(s) \\
r_{\text{min}} &\leq v(s) - \gamma v(s) \\
r_{\text{min}} &\leq (1-\gamma) v(s) \qquad | : (1-\gamma) \\
\frac{r_{\text{min}}}{1-\gamma} &\leq v(s)
\end{aligned}
$$

2. $|v(s) - v(s')| \leq \frac{r_{\text{max}} - r_{\text{min}}}{1-\gamma}$

$$
\begin{aligned}
&|v(s) - v(s')| \\
&\leq \left| \frac{r_{\text{max}}}{1-\gamma} - \frac{r_{\text{min}}}{1-\gamma} \right|, \text{because that is the maximal reachable distance} \\
&= \left| \frac{r_{\text{max}} - r_{\text{min}}}{1-\gamma} \right| \\
&= \frac{r_{\text{max}} - r_{\text{min}}}{1-\gamma}, \text{because } r_{\text{max}} > r_{\text{min}}
\end{aligned}
$$

Task 2

Value Iteration (5 points)

(a) Implement the value iteration algorithm (see lecture 3 slide 28) in the function value iteration. Use the values for γ and θ that are given in the code. Initialize the value function $V(s)$ to 0 for all states. How many steps does it need to converge? What is the optimal value function? (3P)

We worked with the standard 4x4 Frozen Lake environment. The Value Iteration algorithm needs 43 loops to converge close to the optimal value function with a maximum error of θ.

The optimal value function is
$$v_* = \begin{pmatrix} v_*(s_0) \\ v_*(s_1) \\ v_*(s_2) \\ v_*(s_3) \\ v_*(s_4) \\ v_*(s_5) \\ v_*(s_6) \\ v_*(s_7) \\ v_*(s_8) \\ v_*(s_9) \\ v_*(s_{10}) \\ v_*(s_{11}) \\ v_*(s_{12}) \\ v_*(s_{13}) \\ v_*(s_{14}) \\ v_*(s_{15}) \end{pmatrix} \approx \begin{pmatrix} 0.015 \\ 0.016 \\ 0.027 \\ 0.016 \\ 0.027 \\ 0 \\ 0.060 \\ 0 \\ 0.058 \\ 0.134 \\ 0.197 \\ 0 \\ 0 \\ 0.247 \\ 0.544 \\ 0 \end{pmatrix}$$

(b) Compute the optimal policy from the value function. (2P)

The optimal policy is
$$\pi_* = \begin{pmatrix} \pi_*(s_0) \\ \pi_*(s_1) \\ \pi_*(s_2) \\ \pi_*(s_3) \\ \pi_*(s_4) \\ \pi_*(s_5) \\ \pi_*(s_6) \\ \pi_*(s_7) \\ \pi_*(s_8) \\ \pi_*(s_9) \\ \pi_*(s_{10}) \\ \pi_*(s_{11}) \\ \pi_*(s_{12}) \\ \pi_*(s_{13}) \\ \pi_*(s_{14}) \\ \pi_*(s_{15}) \end{pmatrix} \approx \begin{pmatrix} 1 \\ 3 \\ 2 \\ 3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 1 \\ 0 \\ 0 \\ 0 \\ 2 \\ 1 \\ 0 \end{pmatrix}$$