9/10

Reinforcement Learning
Exercise 3 - Solution

Jonathan Schnitzler - st166934
Eric Choquet - st160996

May 2, 2024

Proofs

a) Bellman optimality operator is a gamma-contraction We want to show

$$(\mathcal{T}v)(s) = \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v(s')] \tag{1}$$

fullfills the γ-contraction property, namely

$$\|\mathcal{T}v - \mathcal{T}w\|_\infty \leq \gamma \|v - w\|_\infty \tag{2}$$

Inspired by the lecture for the Bellman expectation backup operator, we will similarly use the definition of the infinity norm to show the contraction property

$$\|\mathcal{T}v - \mathcal{T}w\|_\infty = \|\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v(s')] - \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma w(s')]\| \tag{3}$$

$$= \|\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v(s') - (r + \gamma w(s'))]\| \tag{4}$$

$$= \gamma \|\max_a \sum_{s',r} p(s',r|s,a)[v(s') - w(s')]\| \tag{5}$$

$$\leq \gamma \|\max_a \sum_{s',r} p(s',r|s,a)\|v(s') - w(s')\|_\infty\| \tag{6}$$

$$= \gamma \|v(s') - w(s')\|_\infty \|\max_a \sum_{s',r} p(s',r|s,a)\| \tag{7}$$

$$\leq \gamma \|v - w\|_\infty \tag{8}$$

b) Bounding general finite MDPs This is quite simple by imagining, a sequence of actions for which always the best reward $r_{\max}$ or always the worst outcome, i.e. $r_{\min}$ occurs. We can use the geometric sum formular for $\gamma < 1$

Be careful with the equality signs
+3

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t|S_t = s] && (9)\\
&= \mathbb{E}_\pi[\sum_{i=0}^{\infty} \gamma^i R_{t+i+1}|S_t = s] && (10)\\
&\leq \sum_{i=0}^{\infty} \gamma^i r_{\max} && (11)\\
&= r_{\max} \frac{1}{1-\gamma} && (12)
\end{aligned}
$$

which reversly holds for the minimum with a lower bound

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t|S_t = s] && (13)\\
&= \mathbb{E}_\pi[\sum_{i=0}^{\infty} \gamma^i R_{t+i+1}|S_t = s] && (14)\\
&\geq \sum_{i=0}^{\infty} \gamma^i r_{\min} && (15)\\
&= r_{\min} \frac{1}{1-\gamma} && (16)
\end{aligned}
$$

This yields

$$\frac{r_{\min}}{1-\gamma} \leq v(s) \leq \frac{r_{\max}}{1-\gamma} \tag{17}$$

From this we can follow from arbitrary $v(s)$ and $v(s')$ by assuming without loss of generality taht $v(s) \geq v(s')$ (since the naming is arbitrary)

$$
\begin{aligned}
|v(s) - v(s')| &= v(s) - v(s') && (18)\\
&\leq \frac{r_{\max}}{1-\gamma} - v(s') && (19)\\
&\leq \frac{r_{\max}}{1-\gamma} - \frac{r_{\min}}{1-\gamma} && (20)\\
&= \frac{r_{\max} - r_{\min}}{1-\gamma} && (21)
\end{aligned}
$$

which concludes the proof.

+2

Value Iteration

a) & b) Implementation of the value function

The value function is initialized with zero-values

$$V(s) = 0 \quad \forall_{s \in \mathcal{S}} \tag{22}$$

+4
Try to show other optimal policies as well.

and $\gamma = 0.8$, $\theta = 10^{-8}$. It converges in 43 Iterations

0.015	0.016	0.027	0.016
0.027	0.000	0.060	0.000
0.058	0.134	0.197	0.000
0.000	0.247	0.544	0.000

Table 1: Optimal value v_*

$\downarrow$	$\uparrow$	$\rightarrow$	$\uparrow$
$\leftarrow$	H	$\leftarrow$	H
$\uparrow$	$\downarrow$	$\leftarrow$	H
H	$\rightarrow$	$\downarrow$	G

Table 2: Optimal policy π_*