Assignment 2

Gautham Bellamkonda

September 23, 2023

Problem 1

a

Recall that the Bellman optimality operator on a value function v is given by

$$L(v) = \max_{a \in A} [\mathcal{R}^a + \gamma \mathcal{P}^a v]$$

where, $\mathcal{R}^a$ is the reward function for action a (vector of dimensions $n \times 1$) and $\mathcal{P}^a$ is the transition matrix for action a (matrix of dimensions $n \times n$).

$$\begin{aligned}
\|L(v) - L(u)\|_\infty &= \| \max_{a \in A} [\mathcal{R}^a + \gamma \mathcal{P}^a v] - \max_{a \in A} [\mathcal{R}^a + \gamma \mathcal{P}^a u] \|_\infty \\
&\leq \max_{a \in A} [\|\mathcal{R}^a + \gamma \mathcal{P}^a v - \mathcal{R}^a - \gamma \mathcal{P}^a u\|_\infty] \\
&= \max_{a \in A} [\gamma \|\mathcal{P}^a (v - u)\|_\infty] \\
&\leq \gamma \max_{a \in A} \|\mathcal{P}^a\|_\infty \|v - u\|_\infty \\
&\leq \gamma \|v - u\|_\infty
\end{aligned}$$

Here we have used the fact that, for two sequences of vectors $\{X_i\}$ and $\{Y_i\}$, with $i \in I$, we have

$$\begin{aligned}
\| \max_{i \in I} X_i - \max_{j \in I} Y_j \|_\infty &\leq \| \max_{i \in I} (X_i - Y_i) \|_\infty \\
&\leq \| \max_{i \in I} \|X_i - Y_i\|_\infty \|_\infty \\
&= \max_{i \in I} \|X_i - Y_i\|_\infty
\end{aligned}$$

and, that for every $x \in \mathbb{R}^n$, and A is a $m \times n$ matrix, then $\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty$

Hence, the Bellman optimality operator is a contraction mapping with contraction factor γ.

b

We saw in the class that the Bellman evaluation operator L^π is a γ-contraction.

$$\begin{aligned}
\|L^\pi(v) - L^\pi(u)\|_\infty &= \|\mathcal{R}^\pi + \gamma\mathcal{P}^\pi v - \mathcal{R}^\pi - \gamma\mathcal{P}^\pi u\|_\infty \\
&= \gamma\|\mathcal{P}^\pi(v-u)\|_\infty \\
&\leq \gamma\|\mathcal{P}^\pi\|_\infty\|v-u\|_\infty \\
&\leq \gamma\|v-u\|_\infty
\end{aligned}$$

Take $u = V^\pi$ and $v = V_k^\pi$. We have

$$\begin{aligned}
\|L^\pi(V_k^\pi) - L^\pi(V^\pi)\|_\infty &\leq \gamma\|V_k^\pi - V^\pi\|_\infty \\
\Rightarrow \|V_{k+1}^\pi - V^\pi\|_\infty &\leq \gamma\|V_k^\pi - V^\pi\|_\infty
\end{aligned}$$

Similarly, we have

$$\|V_k^\pi - V^\pi\|_\infty \leq \gamma\|V_{k-1}^\pi - V^\pi\|_\infty$$

So, we have

$$\begin{aligned}
\|V_{k+1}^\pi - V^\pi\|_\infty &\leq \gamma\|V_k^\pi - V^\pi\|_\infty \\
&\leq \gamma^2\|V_{k-1}^\pi - V^\pi\|_\infty \\
&\vdots \\
&\leq \gamma^k\|V_1^\pi - V^\pi\|_\infty
\end{aligned}$$

We know that $\gamma < 1$. Hence, the iterative policy evaluation algorithm converges to V^π geometrically.

c

We saw from problem **a** that the Bellman optimality operator L is a γ-contraction. So, we have, for any $k \geq 1$,

$$\begin{aligned}
\|L(V_{k+1}) - L(V_k)\|_\infty &\leq \gamma\|V_{k+1} - V_k\|_\infty \\
\Rightarrow \|V_{k+2} - V_{k+1}\|_\infty &\leq \gamma\|V_{k+1} - V_k\|_\infty
\end{aligned}$$

If $\|V_{k+1} - V_k\|_\infty < \epsilon$, it follows that

$$\begin{aligned}
\|V_{k+2} - V_{k+1}\|_\infty &\leq \gamma\|V_{k+1} - V_k\|_\infty < \gamma\epsilon \\
\|V_{k+3} - V_{k+2}\|_\infty &\leq \gamma\|V_{k+2} - V_{k+1}\|_\infty < \gamma^2\epsilon \\
&\vdots
\end{aligned}$$

So, we have

$$\begin{aligned}
\|V^* - V_{k+1}\|_\infty &= \|(V_{k+2} - V_{k+1}) + (V_{k+3} - V_{k+2}) + \cdots\|_\infty \\
&\le \|V_{k+2} - V_{k+1}\|_\infty + \|V_{k+3} - V_{k+2}\|_\infty + \cdots \\
&\le \gamma\epsilon + \gamma^2\epsilon + \cdots \\
&\le \frac{\gamma\epsilon}{1-\gamma}
\end{aligned}$$

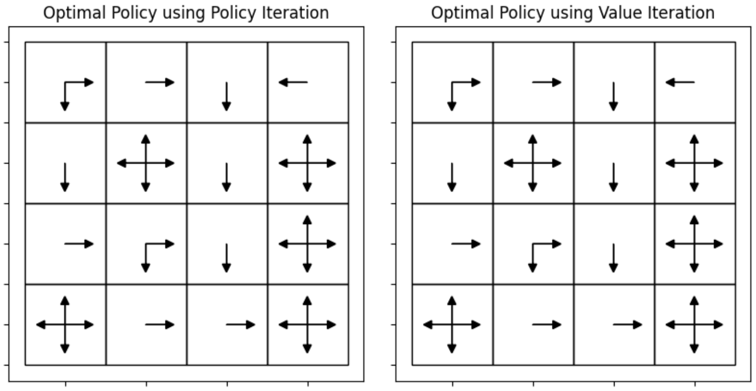
Problem 2

a

Implemented in jupyter notebook.

b

Both policy iteration and value iteration converge to the same optimal policy. The number of iterations required for convergence using policy iteration is 3, while the number of iterations required for value iteration is 7. Snapshots of the optimial policies, from the policy iteration and value iteration algorithms are shown below.


Optimal Policy using Policy Iteration
Optimal Policy using Value Iteration

c

Yes, there are stochastic optimal policies. The snapshot included above is an example of a stochastic policy. The starting state (top left) can either go down or right. The optimal policy says to go down with probability 0.5 and right with probability 0.5. The greedy algorithm that I have implemented picks the action with the highest Q value. If there are multiple actions with the same Q value, it randomly picks one of those actions with equal probability. This is the reason why the optimal policy is stochastic.

d

(i)

Implemented in jupyter notebook.

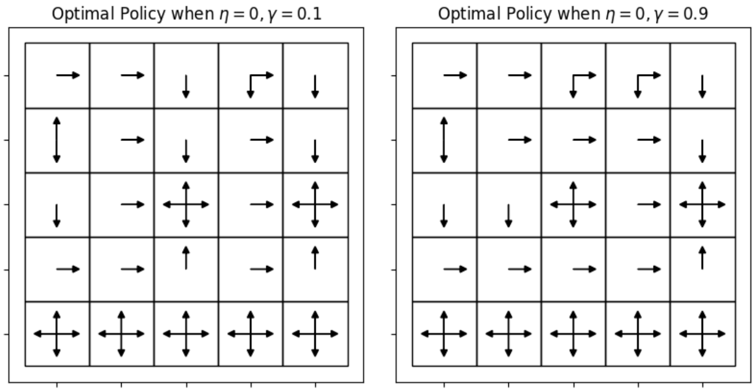
(ii)

The following observations were made by running the code for different values of η and γ.

1. For $\eta = 0, \gamma < 1$, the optimal policy says to prefer the riskier path along the cliff.

   - The agent prefers the smaller goal reward of +1 over the larger goal reward of +10 if γ is small (e.g., $\gamma = 0.1$).
   - The agent prefers the larger goal reward of +10 over the smaller goal reward of +1 if γ is large (e.g., $\gamma = 0.9$).

   These observations can be explained intuitively. Firstly, $\eta = 0$ means that the actions recommended by the policy will be followed determininstically. So, the policy can afford to recommend a riskier path along the cliff. Secondly, if γ is small, the agent is more myopic and prefers the smaller goal reward of +1 over the larger goal reward of +10. If γ is large, the agent is less myopic and prefers the larger goal reward of +10 over the smaller goal reward of +1.



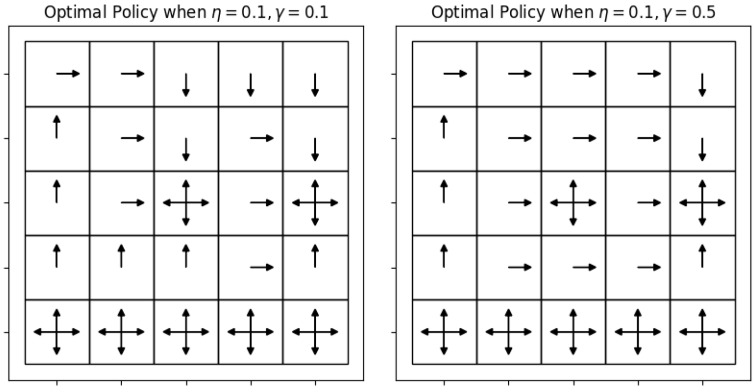
Optimal Policy when $\eta = 0, \gamma = 0.1$

Optimal Policy when $\eta = 0, \gamma = 0.9$

2. For $\eta = 0.1, \gamma < 1$, the optimal policy says to prefer the safer path away from the cliff.

   - The agent prefers the smaller goal reward of +1 over the larger goal reward of +10 if γ is small (e.g., $\gamma = 0.1$).
   - The agent prefers the larger goal reward of +10 over the smaller goal reward of +1 if γ is not too large or too small (e.g., $\gamma = 0.5$).

Firstly, $\eta = 0.1$ means that the actions recommended by the policy may or may not be followed all the time. Thus, by taking the path along the cliff, there are chances that the agent could fall into the cliff, and this results in decreasing the value of the states along the cliff. Secondly, if γ is small, the agent is more myopic and prefers the smaller goal reward of +1 over the larger goal reward of +10. If γ is not too large or too small, the agent is less myopic and prefers the larger goal reward of +10 over the smaller goal reward of +1. It is also observed that the agent prefers the riskier path along the cliff if γ is large (e.g., $\gamma = 0.9$).


Optimal Policy when $\eta = 0.1, \gamma = 0.1$
Optimal Policy when $\eta = 0.1, \gamma = 0.5$

Also, note that the agent *might* end up taking some other path than the paths noted above, when $\eta > 0$. This is because the environment is stochastic. Hence, we cannot say for sure that the agent will always take the path away from the cliff towards the smaller goal if $\gamma = 0.1$, and the path away from the cliff towards the larger goal if $\gamma = 0.5$. It happens with a high probability, but not always.

(iii)

Yes, γ plays a crucial role on the optimal policy, as noted in my solutions to the previous assignment. We can see a similar dependency of γ on the optimal policy even here.