CS 234 Spring 2017
Assignment 1 Solutions

1 Bellman Operator Properties

(a) For all s:

$$\begin{aligned}(T_\pi V)(s) &= \mathbb{E}_\pi\left[R(s,\pi(s)) + \gamma \sum_{s'\in S} P(s'|s,\pi(s))V(s')\right]\\ &\le \mathbb{E}_\pi\left[R(s,\pi(s)) + \gamma \sum_{s'\in S} P(s'|s,\pi(s))V'(s')\right]\\ &= (T_\pi V')(s)\end{aligned}$$

So,

$$T_\pi V \le T_\pi V' \tag{1}$$

Let π_1 and π_2 be the greedy policies referring to V and V' respectively. So,

$$T_{\pi_1} V = TV, \quad T_{\pi_2} V' = TV'$$

As a result:

$$TV = T_{\pi_1} V \le T_{\pi_1} V'$$

The inequality above results from inequality (1). Further,

$$T_{\pi_1} V' \le T_{\pi_2} V' = TV'$$

Hence,

$$TV \le TV' \tag{2}$$

(b) For the T_π operator,

$$\begin{aligned}
(T_\pi V + c\vec{1})(s) &= \mathbb{E}_\pi\left[R(s,\pi(s)) + \gamma\sum_{s'\in S} P(s'|s,\pi(s))(V(s')+c)\right]\\
&= \mathbb{E}_\pi\left[R(s,\pi(s)) + \gamma\sum_{s'\in S} P(s'|s,\pi(s))V(s') + \gamma c\sum_{s'\in S} P(s'|s,\pi(s))\right]\\
&= \mathbb{E}_\pi\left[R(s,\pi(s)) + \gamma\sum_{s'\in S} P(s'|s,\pi(s))V(s')\right] + \gamma c\\
&= (T_\pi V)(s) + \gamma c
\end{aligned}$$

For the T operator, the solution is similar since the constant c does not affect the max operation.

(c) Let π_1 be the greedy policy referring to $\alpha V + \beta V'$. Hence, for all s:

$$\begin{aligned}
T(\alpha V + \beta V')(s) &= T_{\pi_1}(\alpha V + \beta V')(s)\\
&= \mathbb{E}_{\pi_1}\left[R(s,\pi(s)) + \gamma\sum_{s'\in S} P(s'|s,\pi(s))(\alpha V(s') + \beta V(s'))\right]\\
&= \mathbb{E}_{\pi_1}\left[\alpha R(s,\pi(s)) + \gamma\sum_{s'\in S} P(s'|s,\pi(s))(\alpha V(s'))\right]\\
&+ \mathbb{E}_{\pi_1}\left[\beta R(s,\pi(s)) + \gamma\sum_{s'\in S} P(s'|s,\pi(s))(\beta V(s'))\right]\\
&= (\alpha T_{\pi_1} V)(s) + (\beta T_{\pi_1} V')(s)\\
&\leq (\alpha T V)(s) + (\beta T V')(s)
\end{aligned}$$

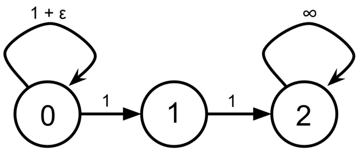
(d) The statement is wrong. Consider the counter-example below:

$$\begin{gathered}
V = \vec{0}\\
R(s,a) > 0 \quad \text{for all } s, a
\end{gathered}$$

This will result in: $TV > V$

2 Value Iteration

(a) Yes, it is possible that the policy changes again with further iterations of value iteration. Consider the following example. There are three states $\{0, 1, 2\}$. In each state, the available actions are indicated by outgoing edges and each is denoted by the next state the action leads to. The corresponding rewards are indicated on the edges. The state transition is deterministic, that is, taking an action always leads to the corresponding state in the next time step. Assume $\gamma = 1$. Note $\epsilon \in (0, 1)$ is some small constant.



Over the iterations of value iteration, the value function can be computed as follows:

Iteration	$V(0)$	$V(1)$	$V(2)$
0	0	0	0
1	$1+\epsilon$	1	∞
2	$2(1+\epsilon)$	∞	∞
3	∞	∞	∞

Over the iterations of value iteration, the current best policy can be computed as follows:

Iteration	$\pi(0)$	$\pi(1)$	$\pi(2)$
0	Arb.	Arb.	Arb.
1	0	2	2
2	0	2	2
3	1	2	2

(b) Note the optimal value function V^* is such that

$$V^*(s) = \max_a \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s') \right\}, \forall s \in S \, .$$

The best policy for $\tilde{V}$ is deterministic and is defined

$$\pi_{\tilde{V}}(s) = \arg\max_a \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \tilde{V}(s') \right\}, \forall s \in S \, ,$$

and the corresponding value function is such that

$$V_{\pi_{\tilde{V}}}(s) = R(s, \pi_{\tilde{V}}(s)) + \gamma \sum_{s' \in S} P(s'|s, \pi_{\tilde{V}}(s)) V_{\pi_{\tilde{V}}}(s'), \forall s \in S \, .$$

We first find relations on $\{L_{\tilde{V}}(s)\}_{s \in S}$. For any $s \in S$,

$$\begin{aligned} V_{\pi_{\tilde{V}}}(s) &= R(s, \pi_{\tilde{V}}(s)) + \gamma \sum_{s' \in S} P(s'|s, \pi_{\tilde{V}}(s)) V_{\pi_{\tilde{V}}}(s') \\ &= \max_a \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \tilde{V}(s') \right\} \\ &\qquad + \gamma \sum_{s' \in S} P(s'|s, \pi_{\tilde{V}(s)}) \left(V_{\pi_{\tilde{V}}}(s') - \tilde{V}(s') \right) \, , \text{ by definition of } \pi_{\tilde{V}} \\ &\geq \max_a \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \left(V^*(s') - \epsilon \right) \right\} \\ &\qquad + \gamma \sum_{s' \in S} P(s'|s, \pi_{\tilde{V}}(s)) \left(V_{\pi_{\tilde{V}}}(s') - (V^*(s') + \epsilon) \right) \, , \text{ since } |V^*(s) - \tilde{V}(s)| \leq \epsilon, \forall s \\ &= \max_a \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s') \right\} \\ &\qquad + \gamma \sum_{s' \in S} P(s'|s, \pi_{\tilde{V}}(s)) \left(V_{\pi_{\tilde{V}}}(s') - V^*(s') \right) - 2\epsilon\gamma \, , \text{ since } \sum_{s' \in S} P(s'|s,a) = 1, \forall s, a \\ &= V^*(s) + \gamma \sum_{s' \in S} P(s'|s, \pi_{\tilde{V}}(s)) \left(V_{\pi_{\tilde{V}}}(s') - V^*(s') \right) - 2\epsilon\gamma \, , \text{ by definition of } V^*. \end{aligned}$$

Then, we can rewrite

$$\gamma \sum_{s' \in S} P(s'|s, \pi_{\tilde{V}}(s)) \left(V^*(s') - V_{\pi_{\tilde{V}}}(s') \right) + 2\epsilon\gamma \geq V^*(s) - V_{\pi_{\tilde{V}}}(s) \, .$$

Equivalently, for any $s \in S$,

$$\gamma \sum_{s' \in S} P(s'|s, \pi_{\tilde{V}}(s)) L_{\tilde{V}}(s') + 2\epsilon\gamma \geq L_{\tilde{V}}(s) \, . \tag{3}$$

We now show $L_{\tilde{V}}(s) \leq \frac{2\gamma\epsilon}{1-\gamma}$ for all $s \in S$. Let $M = \max_s L_{\tilde{V}}(s)$ and the max be achieved at state s^*, i.e., $L_{\tilde{V}}(s^*) = M$. From (3),

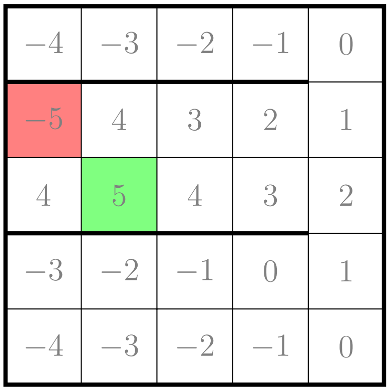
$$\begin{aligned} L_{\tilde{V}}(s^*) &\leq \gamma \sum_{s' \in S} P(s'|s^*, \pi_{\tilde{V}}(s^*)) L_{\tilde{V}}(s') + 2\epsilon\gamma \\ &\leq \gamma \sum_{s' \in S} P(s'|s^*, \pi_{\tilde{V}}(s^*)) M + 2\epsilon\gamma \\ &= \gamma M + 2\epsilon\gamma \,. \end{aligned}$$

Then, $M \leq \gamma M + 2\epsilon\gamma$, or $M \leq \frac{2\epsilon\gamma}{1-\gamma}$. It follows that $L_{\tilde{V}}(s) \leq \frac{2\epsilon\gamma}{1-\gamma}$ for all $s \in S$.

3 Grid Policies

(a) Let all rewards be -1.

(b)



-4	-3	-2	-1	0
-5	4	3	2	1
4	5	4	3	2
-3	-2	-1	0	1
-4	-3	-2	-1	0

(c) No. Changing γ changes the value function but not the relative order.

(d) The value function changes (all states increase by 15), but the policy does not.

4 Frozen Lake MDP

(c) Stochasticity generally increases the number of iterations required to converge. In the stochastic frozen lake environment, the number of iterations for value iteration increases. For policy iteration, depending on the implementation method, the number of iterations could remain unchanged; or policy iteration might not even converge at all. The stochasticity would also change the optimal policy. In this environment, the optimal policy of the stochastic frozen lake is different from the one of the deterministic frozen lake.

5 Frozen Lake Reinforcement Learning

(d) Q-learning is less computationally intensive, since it only needs to update a single Q-value per update. Model-based learning, on the other hand, must keep track of all the counts and attempt to estimate the model parameters. However, model-based learning uses the update data more efficiently, so it may require less experience to learn a good policy (provided the model space is not prohibitively large).