CS 234 Winter 2018
Assignment 1 Solutions

1 Effect of Modeling Errors in Policy Evaluation

Proof. For $i = H$ the statement is trivially true. We assume now it holds for $i+1$ and show it holds also for i. Using only this induction hypothesis and basic algebra, we can write

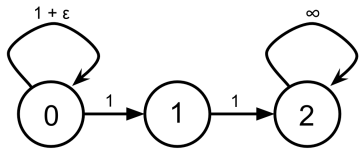
$$
\begin{aligned}
&\hat{V}_i(s_i) - V_i(s_i) \\
&\overset{(a)}{=} \mathbb{E}\left[\hat{r}(s_i, a_i) + \sum_{s' \in \mathcal{S}} \hat{P}(s' \mid s_i, a_i)\hat{V}_{i+1}(s') - r(s_i, a_i) - \sum_{s' \in \mathcal{S}} P(s'|s_i, a_i)V_{i+1}(s') \big| s_i \right] \\
&\overset{(b)}{=} \mathbb{E}\Bigg[\hat{r}(s_i, a_i) - r(s_i, a_i) + \sum_{s' \in \mathcal{S}} \hat{V}_{i+1}(s')(\hat{P}(s'|s_i, a_i) - P(s'|s_i, a_i)) \\
&\quad + \sum_{s' \in \mathcal{S}} P(s'|s_i, a_i)(\hat{V}_{i+1}(s') - V_{i+1}(s')) \big| s_i \Bigg] \\
&\overset{(c)}{=} \mathbb{E}\left[\hat{r}(s_i, a_i) - r(s_i, a_i) + \sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s_i, a_i) - P(s'|s_i, a_i)\right) \hat{V}_{i+1}(s') \big| s_i \right] + \\
&\quad + \mathbb{E}\left[\hat{V}_{i+1}(s_{i+1}) - V_{i+1}(s_{i+1}) \big| s_i \right] \\
&\overset{(d)}{=} \mathbb{E}\Bigg[\hat{r}(s_i, a_i) - r(s_i, a_i) + \sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s_i, a_i) - P(s'|s_i, a_i)\right) \hat{V}_{i+1}(s') + \\
&\quad + \mathbb{E} \sum_{t=i+1}^{H} \left[(\hat{r}(s_t, a_t) - r(s_t, a_t)) + \sum_{s' \in \mathcal{S}} (\hat{P}(s' \mid s_t, a_t) - P(s' \mid s_t, a_t))\hat{V}_{t+1}(s') \Big| s_{i+1} \right] \Big| s_i \Bigg] \\
&\overset{(e)}{=} \mathbb{E} \sum_{t=i}^{H} \left[(\hat{r}(s_t, a_t) - r(s_t, a_t)) + \sum_{s' \in \mathcal{S}} (\hat{P}(s' \mid s_t, a_t) - P(s' \mid s_t, a_t))\hat{V}_{t+1}(s') \Big| s_i \right].
\end{aligned}
$$

In the above passages: (a) follows from Bellman's equation (here the expectation is with respect to the action chosen by the agent a_t), (b) follows from adding and subtracting $\sum_{s' \in \mathcal{S}} P(s' \mid s_i, a_i)\hat{V}_{i+1}(s')$, (c) from the definition of the expectation operator, (d) from the inductive hypothesis and finally (e) is grouping together the terms and using the law of total expectation. □

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 $\varepsilon \in (0, 1)$ 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	$1+\varepsilon$	1	∞
2	$2(1+\varepsilon)$	∞	∞
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 $s \in S$,

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

Then, we can rewrite

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

Equivalently, for any $s \in S$,

$$
\gamma \sum_{s' \in S} P(s'|s, \pi_{\tilde{V}}(s)) L_{\tilde{V}}(s') + 2\varepsilon\gamma \geq L_{\tilde{V}}(s) \,. \tag{1}
$$

We now show $L_{\tilde{V}}(s) \leq \frac{2\gamma\varepsilon}{1-\gamma}$ 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 (1),

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

Then, $M \leq \gamma M + 2\varepsilon\gamma$, or $M \leq \frac{2\varepsilon\gamma}{1-\gamma}$. It follows that $L_{\tilde{V}}(s) \leq \frac{2\varepsilon\gamma}{1-\gamma}$ 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	2	1	0
−5	−4	−3	−2	−1
−6	−5	−4	−3	−2

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

(d) The value function would change but the policy would 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.