# CS 234 Winter 2018
# Assignment 1
# Due: January 23 at 11:59 pm

Li Quan Khoo (SCPD)

## 1 Optimal Policy for Simple MDP [20 pts]

(a) Since action $a_0$ is always chosen under $\pi^*$, working backwards from state $G = S_n$, where $0 < \gamma < 1$ (i.e. small and strictly positive),

$$\begin{aligned} V^{\pi^*}(G = s_n) &= 1 + \gamma + \gamma^2 + \gamma^3 + ... = \frac{1}{1-\gamma} \\ V^{\pi^*}(s_{n-1}) &= r(s_{n-1}, a_0) + \gamma V^{\pi^*}(s_n) = 0 + \gamma \frac{1}{1-\gamma} \\ V^{\pi^*}(s_{n-2}) &= \gamma^2 \frac{1}{1-\gamma} = \gamma^{n-(n-2)} \frac{1}{1-\gamma} \\ &... \\ V^{\pi^*}(s_1) &= \gamma^{n-1} \frac{1}{1-\gamma} \text{ , by induction.} \end{aligned}$$

Hence for any state $s_i$, its optimal value function is

$$V^{\pi^*}(s_i) = \gamma^{n-i} \frac{1}{1-\gamma}, \text{where } 1 \leq i \leq n$$

(b) Not in this scenario, as long as $\gamma$ remains in the given range of $0 < \gamma < 1$, which makes $\frac{1}{1-\gamma}$ and $\gamma^k$ bounded for any $k \in \mathbb{N}$. For any such value of $\gamma$, consider the situation where the agent is in the state $s_i$, which has two valid actions, i.e. $a_0$ into state $s_{i+1}$, and $a_1$ into state $s_{i-1}$.

$$V^{\pi^*}(s_{i-1}) = \gamma^{n-(i-1)} \frac{1}{1-\gamma} \quad < \quad \gamma^{n-(i+1)} \frac{1}{1-\gamma} = V^{\pi^*}(s_{i+1}) \quad \text{, for all valid } i$$

And, by definition of the value function,

$$\begin{aligned} V^{\pi}(s_i) &= r(s_i, \pi(s)) + \gamma \, \mathbb{E}_{p(s'|s,a)\sim\pi}[V^{\pi}(s')] \\ &= 0 + \gamma \, \mathbb{E}_{\pi}[V^{\pi}(s')] \end{aligned}$$

Since the optimal policy maximizes $V^{\pi}(s_i)$ by definition, action $a_0$ (move right) is chosen every single time for every state $s_i$, since we have shown that for any such $\gamma$, its value function is larger. Hence we have that the optimal policy is independent of $\gamma$.

(c) Deriving in the same manner as in (a),

$$
\begin{aligned}
V^{\pi^*}(G = s_n) &= (1+c) + \gamma(1+c) + \gamma^2(1+c) + ... = \frac{1+c}{1-\gamma} \\
V^{\pi^*}(s_{n-1}) &= R(s_{n-1}, a_0) + \gamma V^{\pi^*}(s_n) = c + \gamma\frac{1+c}{1-\gamma} \\
V^{\pi^*}(s_{n-2}) &= c + \gamma c + \gamma^2\frac{1+c}{1-\gamma} \\
V^{\pi^*}(s_{n-3}) &= c[1+\gamma+\gamma^2] + \gamma^3\frac{1+c}{1-\gamma} \\
&= c\left(\frac{1-\gamma^3}{1-\gamma}\right) + \gamma^3\frac{1+c}{1-\gamma}, \quad \text{by formula of finite geometric sum.} \\
&= c\left(\frac{1-\gamma^{n-(n-3)}}{1-\gamma}\right) + \gamma^{n-(n-3)}\frac{1+c}{1-\gamma} \\
&... \\
V^{\pi^*}(s_1) &= c(\frac{1-\gamma^{n-1}}{1-\gamma}) + \gamma^{n-1}\frac{1+c}{1-\gamma}, \quad \text{by induction.}
\end{aligned}
$$

Hence for any state $s_i$, its optimal value function is

$$
\begin{aligned}
V^{\pi^*}(s_i) &= c(\frac{1-\gamma^{n-i}}{1-\gamma}) + \gamma^{n-i}\frac{1+c}{1-\gamma} \\
&= \frac{c+\gamma^{n-i}}{1-\gamma}
\end{aligned}
$$

Now to show that the optimal policy doesn't change (i.e. move right every time), it is sufficient to show that $V^{\pi^*}(s_{i+1}) > V^{\pi^*}(s_{i-1})\ \forall s_i$, which is to say $V^{\pi^*}(s_{i+1}) - V^{\pi^*}(s_{i-1}) > 0,\ \forall s_i$.

$$
\begin{aligned}
V^{\pi^*}(s_{i+1}) - V^{\pi^*}(s_{i-1}) &= \frac{c+\gamma^{n-(i+1)}}{1-\gamma} - \frac{c+\gamma^{n-(i-1)}}{1-\gamma} \\
&= \frac{1}{1-\gamma}[\gamma^{n-(i+1)} - \gamma^{n-(i-1)}] \\
&= \frac{1}{1-\gamma}[\text{positive term}]
\end{aligned}
$$

Hence we have shown that for $0 < \gamma < 1$, the above expression is always positive, independent of the value of $c$. QED.

(d) By the same strategy,

$$
\begin{aligned}
V^{\pi^*}(G = s_n) &= a(1+c) + \gamma a(1+c) + \gamma^2 a(1+c) + ... = a\frac{1+c}{1-\gamma} \\
V^{\pi^*}(s_{n-1}) &= R(s_{n-1}, a_0) + \gamma V^{\pi^*}(s_n) = a(c+0) + \gamma a\frac{1+c}{1-\gamma} \\
V^{\pi^*}(s_{n-2}) &= a\left[c + \gamma c + \gamma^2\frac{1+c}{1-\gamma}\right]
\end{aligned}
$$

Since the term in brackets was already derived in part (c), we can now see that for any state $s_i$, its optimal value function is now

$$V^{\pi^*}(s_i) = a\frac{c+\gamma^{n-i}}{1-\gamma}$$

Now taking the difference between the value functions of the states adjacent to $s_i$ again,

$$\begin{aligned} V^{\pi^*}(s_{i+1}) - V^{\pi^*}(s_{i-1}) &= a\frac{c+\gamma^{n-(i+1)}}{1-\gamma} - a\frac{c+\gamma^{n-(i-1)}}{1-\gamma} \\ &= a\frac{1}{1-\gamma}[\gamma^{n-(i+1)} - \gamma^{n-(i-1)}] \\ &= a \text{ [positive term] [positive term]} \end{aligned}$$

It is clear from the equation that if $a < 0$, the condition $V^{\pi^*}(s_{i+1}) > V^{\pi^*}(s_{i-1})\ \forall s_i$ no longer holds true, and we have demonstrated in part (c) that it is a necessary condition for the optimal policy to remain the same. $c$ can in fact take any bounded value, since the equation above is still independent of $c$. This is intuitive because, suppose $a = -1$ and $c = 0$; instead of being rewarded in state $G$, the agent is being penalized instead. Hence the optimal policy is to not reach state $G$ at all. In fact, following the policy that was optimal for $a > 0$ in this case results in the agent being penalized forever! In this particular case where $a = -1$ and $c = 0$, the optimum policy is simply to not enter state G, and that gives the agent the maximal reward of 0. It is also easy to see that this policy is not unique, since there are multiple ways to do so, as long as that single transition into G is avoided.

## 2 Running Time of Value Iteration [20 pts]

(a)

$$\begin{aligned} Q^{\pi}(s_0, a) &= r(s_0, a) + \gamma \mathbb{E}_{p(s'|a)\sim\pi}[V^{\pi}(s')], \quad \text{definition of Q-function} \\ Q^{\pi_{a_1}}(s_0, a_1) &= r(s_0, a_1) + \gamma[V^{\pi}(s_1)], \quad \text{since deterministic} \\ Q^{\pi_{a_1}}(s_0, a_1) &= 0 + (\gamma + \gamma^2 + \gamma^3 + ...) = \frac{1}{1-\gamma} \end{aligned}$$

(b)

$$\begin{aligned} Q^{\pi_{a_2}}(s_0, a_2) &= r(s_0, a_2) + \gamma[V^{\pi}(s_2)] \\ Q^{\pi_{a_2}}(s_0, a_2) &= \frac{\gamma^2}{1-\gamma} + 0 \\ Q^{\pi_{a_2}}(s_0, a_2) &= \gamma^2 Q^{\pi_{a_1}}(s_0, a_1) < Q^{\pi_{a_1}}(s_0, a_1), \text{ since } 0 < \gamma < 1 \end{aligned}$$

The optimal action is $a_1$, which has a larger Q-value.

(c) We need to show for what $k \in \mathbb{N}$, $\hat{Q}^\pi(s_0, a_1) \geq Q^\pi(s_0, a_2) = \frac{\gamma^2}{1-\gamma}$

$$
\begin{aligned}
t = 0, \quad & \hat{Q}^\pi(s_0, a_1) = \ 0 \\
t = 1, \quad & \hat{Q}^\pi(s_0, a_1) = \ r(s_0, a_1) + \gamma(V^\pi(s_1)) = 0 + \gamma(0) \\
t = 2, \quad & \hat{Q}^\pi(s_0, a_1) = \ \gamma^2 \\
t = 3, \quad & \hat{Q}^\pi(s_0, a_1) = \ \gamma^2 + \gamma^3 = \gamma(\gamma + \gamma^2) \\
t = k, \quad & \hat{Q}^\pi(s_0, a_1) = \ \gamma(\gamma + \gamma^2 + ... + \gamma^{k-1}) = \gamma\frac{1-\gamma^k}{1-\gamma}, \quad \text{by induction.}
\end{aligned}
$$

$$
\begin{aligned}
\hat{Q}^\pi(s_0, a_1) &\geq Q^\pi(s_0, a_2) \\
\gamma\frac{1-\gamma^{n^*}}{1-\gamma} &\geq \frac{\gamma^2}{1-\gamma} \\
1 - \gamma^{n^*} &\geq \gamma, \quad \text{since } \frac{\gamma}{1-\gamma} > 0 \\
\gamma^{n^*} &\geq 1 - \gamma \\
n^* &\geq \frac{\log(1-\gamma)}{\log\gamma}, \text{QED}
\end{aligned}
$$

# 3 Approximating the Optimal Value Function [35 pts]

Idea about maximal loss state $z$ is with reference to Singh, Yee. *An Upper Bound on the Loss from Approximate Optimal-Value Functions.* Machine Learning, 16, 227-233 (1994).

(a) $||\tilde{Q} - Q^*||_\infty \leq \varepsilon$ means according to some approximating policy $\pi$, some action $b$ must appear at least as good as the optimal action $b^*$, in a state $z$ where this maximum loss occurs. We assign $\varepsilon$ as the maximal difference between the two Q-values, i.e.

$$|\tilde{Q}(z, b) - Q^*(z, b^*)| \leq \varepsilon$$

By definition of the infinity norm (element-wise max), we can conclude that for any state $s$:

$$|Q^*(s, a^*) - \tilde{Q}(s, a)| \ \leq \ |Q^*(z, b^*) - \tilde{Q}(z, b)| \ \leq \ \varepsilon \quad , \forall s \in S$$

Abbreviating $V^*(s),\ Q^*(s, \pi(s)),\ \tilde{Q}(s, \pi(s))$ to $V^*, Q^*, \tilde{Q}$ in just the following section:

$$
\begin{aligned}
V^* - Q^* \leq \ & |V^* - Q^*| \\
= \ & |V^* - \tilde{Q} + \tilde{Q} - Q^*| \\
\leq \ & |V^* - \tilde{Q}| + |\tilde{Q} - Q^*| \\
= \ & |Q^*(s, a^*) - \tilde{Q}| + |\tilde{Q} - Q^*| \\
\leq \ & \varepsilon + \varepsilon \\
= \ & 2\varepsilon \quad , \text{QED}
\end{aligned}
$$

(b)

$$\text{To show: } V^\pi(s) - V^*(s) \leq \frac{2\varepsilon}{1-\gamma}$$

$$\begin{aligned}
& V^\pi(s) - V^*(s) \\
&= V^\pi(s) - Q^*(s,\pi(s)) + Q^*(s,\pi(s)) - V^*(s) \\
&= V^\pi(s) - Q^*(s,\pi(s)) + [Q^*(s,\pi(s)) - Q^*(s,a^*)] \\
&\leq V^\pi(s) - Q^*(s,\pi(s)) + 2\varepsilon \\
&= [r(s,\pi(s)) + \gamma \sum_{s'} p(s'|s,\pi(s)) V^\pi(s')] \; - [r(s,\pi(s)) + \gamma \sum_{s'} p(s'|s,\pi(s)) V^*(s')] + 2\varepsilon \\
&= 2\varepsilon + \gamma \sum_{s'} p(s'|s,\pi(s))[V^\pi(s') - V^*(s')] \\
&= 2\varepsilon + \gamma[V^\pi(s') - V^*(s')] \quad \because \pi(s) \text{ is deterministic (greedy)}
\end{aligned}$$

Now that we have shown that:

$$V^\pi(s) - V^*(s) \; \leq 2\varepsilon + \gamma[V^\pi(s') - V^*(s')]$$

If we now substitute this new state $s'$ into the equation above and and follow the steps exactly, we get the result and advance the state being considered one further. Then:

$$\begin{aligned}
V^\pi(s') - V^*(s') \; &\leq 2\varepsilon + \gamma[V^\pi(s'') - V^*(s'')] \\
\gamma[V^\pi(s') - V^*(s')] \; &\leq 2\varepsilon\gamma + \gamma^2[V^\pi(s'') - V^*(s'')] \\
V^\pi(s) - V^*(s) \; \leq \; 2\varepsilon + \gamma[V^\pi(s') - V^*(s')] \; &\leq 2\varepsilon + 2\varepsilon\gamma + \gamma^2[V^\pi(s'') - V^*(s'')] \\
V^\pi(s) - V^*(s) \; &\leq 2\varepsilon + 2\varepsilon\gamma + \gamma^2[V^\pi(s'') - V^*(s'')]
\end{aligned}$$

Now we show the result by induction:

$$\begin{aligned}
V^\pi(s) - V^*(s) \; &\leq 2\varepsilon + 2\varepsilon\gamma + \gamma^2[V^\pi(s'') - V^*(s'')] \\
V^\pi(s) - V^*(s) \; &\leq 2\varepsilon + 2\varepsilon\gamma + 2\varepsilon\gamma^2 + \gamma^3[V^\pi(s''') - V^*(s''')] \\
&\ldots \\
V^\pi(s) - V^*(s) \; &\leq 2\varepsilon[1 + \gamma + \gamma^2 + \gamma^3 + ...] \\
V^\pi(s) - V^*(s) \; &\leq \frac{2\varepsilon}{1-\gamma} \quad , QED
\end{aligned}$$

(c) By observation, the optimal policy $\pi$ would have $p(\text{"stay"}) = 0$ and $p(\text{"go"}) = 1$ simply because the action "stay" gives strictly less reward.

$$V^*(s_2) = 2\varepsilon + 2\varepsilon\gamma + 2\varepsilon\gamma^2 + ... = \frac{2\varepsilon}{1-\gamma}$$

$$\begin{aligned}
V^*(s_1) &= r(s_1, \text{"go"}) + \gamma V^*(s_2) \quad \because \pi^* \text{ deterministic} \\
V^*(s_1) &= 2\varepsilon + \gamma[2\varepsilon + 2\varepsilon\gamma + 2\varepsilon\gamma^2 + ...] \\
&= \frac{2\varepsilon}{1-\gamma}
\end{aligned}$$

(d) Reusing the idea from part (a), $||\tilde{Q} - Q^*||_\infty \leq \varepsilon$ implies that there exists some maximal loss state $z$ where $|\tilde{Q}(z, \pi(z)) - Q^*(z, \pi^*(z))| < \varepsilon$, which means the inequality holds true for every state $s$.
Case $\pi(s_1) =$ "stay":

$$\begin{aligned} V^\pi(s_1) &= r(s_1, \text{"stay"}) + \gamma \tilde{V}(s_1) \\ &= 0 + \gamma \tilde{V}(s_1) \implies \tilde{V}(s_1) = 0 \\ V^\pi(s_1) - V^*(s_1) &= -\frac{2\varepsilon}{1-\gamma} \quad \text{, QED as required} \end{aligned}$$

For the other case where $\pi(s_1) =$ "go", because it would mean that $\pi$ would be an optimal policy, we would expect the error to be exactly zero.

$$\begin{aligned} \tilde{V}(s_1) &= r(\pi, \text{"go"}) + \gamma \tilde{V}(s_2) \\ &= 2\varepsilon + \gamma \tilde{V}(s_2) \\ \tilde{V}(s_2) &= r(s_2, \pi(s_2)) + \gamma \tilde{V}(s_2) \\ &= 2\varepsilon + \gamma \tilde{V}(s_2) \\ &= \frac{2\varepsilon}{1-\gamma} \\ \tilde{V}(s_1) &= 2\varepsilon + \gamma \frac{2\varepsilon}{1-\gamma} \\ &= 2\varepsilon + \gamma[2\varepsilon + 2\varepsilon\gamma + 2\varepsilon\gamma^2 + ...] \\ &= \frac{2\varepsilon}{1-\gamma} = V^*(s_1) \\ \tilde{V}(s_1) - V^*(s_1) &= 0 \end{aligned}$$

Now the only thing left to do is to demonstrate that $\tilde{Q}$ and $Q^*$ for states $s_1$ and $s_2$ satisfies the given inequality for some greedy policy.

$$|\tilde{Q}(s_2, \pi(s_2)) - Q^*(s_2, \pi^*(s_2))| = |\tilde{V}(s_2) - V^*(s_2)| = 0 \leq \varepsilon$$
$$|\tilde{Q}(s_1, \pi(s_1) = \text{"go"}) - Q^*(s_1, \pi^*(s_1))| = 0 \leq \varepsilon \quad \text{as demonstrated previously}$$

This result is expected since the optimal policy is also a greedy one. QED.

# 4 Frozen Lake MDP [25 pts]

(a) (code)

(b) (code)

(c) The optimal policy is different to the one in the deterministic case, and the number of iterations required before convergence increases. Due to the environment not always responding with the chosen movement, despite having an optimal policy or value function, the agent can still fail to reach the goal despite following such a policy.