CS533
Intelligent Agents and Decision Making
Infinite Horizon Markov Decision Problems

1. Some MDP formulations use a reward function $R(s, a, s')$ that also depends on the result state s' (we get reward $R(s, a, s')$ when we take action a in s and then transition to s'). Write the Bellman optimality equation with discount factor β for this formulations.

   Solution: Recall that the Bellman equation for the case of a reward function $R(s, a)$ that only depends on the state is as follows:

$$V^*(s) = \max_{a \in A} R(s, a) + \beta \sum_{s' \in S} T(s, a, s') V^*(s')$$

   For the case of a state-action-state reward function $R(s, a, s')$ we need to move the reward into the expectation since it depends on the next state.

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} T(s, a, s') \left(R(s, a, s') + \beta V^*(s') \right)$$

2. In this exercise you will prove that the Bellman Backup operator is a contraction operator.

   (a) Prove that, for any two functions f and g,

$$|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|.$$

   Solution: Assume without loss of generality that $\max_a f(a) \geq \max_a g(a)$ (if not then just interchange f and g, which will not change the absolute value). Let $a^* = \arg\max_a f(a)$. We have that

$$\begin{aligned} |\max_a f(a) - \max_a g(a)| &= f(a^*) - \max_a g(a) \\ &\leq f(a^*) - g(a^*) \text{ (must be non-negative)} \\ &= |f(a^*) - g(a^*)| \\ &\leq \max_a |f(a) - g(a)| \end{aligned}$$

   The first equality follows from our assumption, the second inequality from the definition of max, the third equality from the definition of absolute value, and the fourth inequality from the definition of max.

   (b) Use the above result in order to prove that the Bellman Backup operator $B[\cdot]$ is a contraction mapping. That is, prove that for any two value function V and V',

$$||B[V] - B[V']|| \leq \beta ||V - V'||$$

   where B is the Bellman backup operator, β is the discount factor, and $|| \cdot ||$ is the max norm. By the definition of the max norm, this is equivalent to proving that for any state s,

$$|B[V](s) - B[V'](s)| \leq \beta ||V - V'||.$$

   Solution: We show that for any state s,

$$|B[V](s) - B[V'](s)| \leq \beta ||V - V'||.$$

$$\begin{aligned}
|B[V](s) - B[V'](s)| &= \left|\max_a R(s,a) + \beta\sum_{s'} T(s,a,s')V(s') - \left(\max_a R(s,a) + \beta\sum_{s'} T(s,a,s')V'(s')\right)\right| \\
&\leq \max_a \left|R(s,a) + \beta\sum_{s'} T(s,a,s')V(s') - R(s,a) - \beta\sum_{s'} T(s,a,s')V'(s')\right| \text{ (from part a)} \\
&= \max_a \beta \left|\sum_{s'} T(s,a,s')\left(V(s') - V'(s')\right)\right| \text{ (basic algebra)} \\
&\leq \max_a \beta \sum_{s'} T(s,a,s')|V(s') - V'(s')| \\
&\leq \max_a \beta \sum_{s'} T(s,a,s')||V - V'|| \text{ (by the definition of max norm)} \\
&= \max_a \beta||V - V'|| \text{ (the sum over s' is 1 for any a)} \\
&= \beta||V - V'||
\end{aligned}$$

3. Consider a trivially simple MDP with two states $S = \{s_0, s_1\}$ and a single action $A = \{a\}$. The reward function is $R(s_0, a) = 0$ and $R(s_1, a) = 1$. The transition function is $T(s_0, a, s_1) = 1$ and $T(s_1, a, s_1) = 1$. Note that there is only a single policy π for this MDP that takes action a in both states.

   (a) Using a discount factor $\beta = 1$ (i.e. no discounting), write out the linear equations for evaluating the policy and attempt to solve the linear system. What happens and why?

   Solution: Denote the policy by π and for notational simplicity let $V_0 = V^{\pi}(s_0)$ and $V_1 = V^{\pi}(s_1)$. The linear equations for the the value function are:

$$\begin{aligned}
V_0 &= R(s_0, a) + \beta V_1 = \beta V_1 \\
V_1 &= R(s_1, a) + \beta V_1 = 1 + \beta V_1
\end{aligned}$$

   which for the case of $\beta = 1$ simplifies to the following.

$$\begin{aligned}
V_0 &= V_1 \\
V_1 &= 1 + V_1
\end{aligned}$$

   Clearly this system has no solution, which is an indication that the policy does not have a well defined finite value function.

   (b) Repeat the previous question using a discount factor of $\beta = 0.9$.

   Solution: For $\beta = 0.9$ we get the following system.

$$\begin{aligned}
V_0 &= 0.9V_1 \\
V_1 &= 1 + 0.9V_1
\end{aligned}$$

   This is easily solved to get $V_0 = 9$ and $V_1 = 10$.

This shows how including a discount factor creates a well conditioned system, which is the case for any MDP provided that $\beta \in [0, 1)$.

4. The Bellman Backup operator satisfies the monotonicity property, which states that for any two value functions V and V', if $V \leq V'$, then $B[V] \leq B[V']$. Prove this monotonicity property of B.

   Solution: Recall that $V \leq V'$ implies that for all states s, $V(s) \leq V'(s)$. We now show that if $V \leq V'$, then for any state s, $B[V](s) - B[V'](s) \leq 0$ which is equivalent to saying that $B[V] \leq B[V']$.

$$\begin{aligned}
B[V](s) - B[V'](s) &= \max_a R(s,a) + \beta \sum_{s'} T(s,a,s')V(s') - \left(\max_a R(s,a) + \beta \sum_{s'} T(s,a,s')V'(s') \right) \\
&\leq \max_a R(s,a) + \beta \sum_{s'} T(s,a,s')V(s') - \left(\max_a R(s,a) + \beta \sum_{s'} T(s,a,s')V(s') \text{ (since } V \leq V') \right) \\
&= 0
\end{aligned}$$

5. In class we presented the policy iteration algorithm, which used a "greedy" policy improvement operation. That is, the improved policy π' at each iteration selected the action that maximized the one-step-look ahead value:

$$\pi'(s) = \arg\max_{a \in A} \sum_{s' \in S} T(s,a,s')V_\pi(s')$$

   where π is the current policy.

   Consider a version of policy iteration, which uses a non-greedy policy improvement operator. This operator returns a policy π' that selects an action in each state that improves over the current action selected by π if possible. But we do not require that π' return the best action. More formally, the non-greedy policy improvement operators returns a policy π' such that for any state s,

$$\sum_{s' \in S} T(s,\pi'(s),s')V_\pi(s') \geq \sum_{s' \in S} T(s,\pi(s),s')V_\pi(s')$$

   with strict inequality when possible.

   Prove that the non-greedy policy improvement operator guarantees that $V_{\pi'} \geq V_\pi$ with strict inequality when π is not optimal.

   Solution: This proof can almost exactly follow the proof for the greedy version of policy iteration with a minor modification. First, recall the definition of the restricted Bellman backup, where we restrict the backup to the actions specified by some policy π.

$$B_\pi[V](s) = R(s) + \beta \sum_{s'} T(s,\pi(s),s')V(s')$$

   The above inequality that relates π' and π, implies that for all states s,

$$R(s) + \beta \sum_{s' \in S} T(s,\pi'(s),s')V_\pi(s') \geq R(s) + \beta \sum_{s' \in S} T(s,\pi(s),s')V_\pi(s'),$$

   where we simply multiplied each side by β and added $R(s)$ to each side. This is equivalent to

$$B_{\pi'}[V_\pi] \geq B_\pi[V_\pi].$$

Using this relationships we can derive

$$V_\pi = B_\pi[V_\pi] \leq B_{\pi'}[V_\pi].$$

We can now follow the same proof as for the greedy case. In particular, if we let $B^k_{\pi'}$ represent k applications of $B_{\pi'}$, then we can derive that $V_\pi \leq B^k_{\pi'}[V_\pi]$ for all $k \geq 1$. Since $B^k_{\pi'}[V_\pi] \to V_{\pi'}$ as $k \to \infty$, we have shown that $V_\pi \leq V_{\pi'}$.

The proof of strict inequality when π is sub-optimal follows the exact same argument as shown in the notes for the greedy variant of policy iteration.