# CS 188 Spring 2010 — Introduction to Artificial Intelligence — Written 4 MDPs and Probability

**Due:** Thursday 3/11 in 283 Soda Drop Box by 11:59pm (no slip days)

**Policy:** Can be solved in groups (acknowledge collaborators) but must be written up individually

| | |
|---|---|
| First name | |
| Last name | |
| SID | |
| Login | |
| Collaborators | |

**For staff use only:**

| | |
|---|---|
| Q1. Eat or Play? | /5 |
| Q2. Reward Shaping | /9 |
| Q3. Medical Diagnosis | /11 |
| Total | /25 |

# Q1. [5 pts] Eat or Play?

At the cs188 casino, there are two things to do: Eat buffet and Play cs188-Blackjack. You start out Poor and Hungry, and would like to leave the casino Rich and Full. If you Play while you are Full you are more likely to become Rich, but if you are Poor you may have a hard time becoming Full on your budget.

We can model your decision making process as the following MDP:

State Space: {PoorHungry, PoorFull, RichHungry, RichFull}
Actions: {Eat, Play}
Initial State: PoorHungry
Terminal State: RichFull

Transition Model:

| $s$ | $a$ | $s'$ | T(s,a,s') |
|---|---|---|---|
| PoorHungry | Play | PoorHungry | 0.8 |
| PoorHungry | Play | RichHungry | 0.2 |
| PoorHungry | Eat | PoorHungry | 0.8 |
| PoorHungry | Eat | PoorFull | 0.2 |
| PoorFull | Play | PoorFull | 0.5 |
| PoorFull | Play | RichFull | 0.5 |
| RichHungry | Eat | RichHungry | 0.2 |
| RichHungry | Eat | RichFull | 0.8 |

Rewards:

| $s'$ | $R(\cdot, \cdot, s')$ |
|---|---|
| PoorHungry | -1 |
| PoorFull | 1 |
| RichHungry | 0 |
| RichFull | 5 |

**(a)** [3 pts] Complete the table for the first 3 iterations of Value Iteration. Assume $\gamma = 1$. Use the *batch* version of Value Iteration.

| State | $i = 0$ | $i = 1$ | $i = 2$ | $i = 3$ |
|---|---|---|---|---|
| PoorHungry | 0 | | | |
| PoorFull | 0 | | | |
| RichHungry | 0 | | | |
| RichFull | 0 | 0 | 0 | 0 |

**(b)** [2 pts] Assuming that we are acting for three time steps, what is the optimal action to take from the starting state? Justify your answer.

# Q2. [9 pts] Reward Shaping

Consider an arbitrary MDP with state set $S$, transition function $T(s, a, s')$, and discount factor $\gamma$. Suppose we are given an arbitrary *potential function* $\phi(s)$ that maps each state $s \in S$ to a real number $\phi(s)$, with the constraint that $\phi(s) = 0$ for all terminal states $s \in S$. Based on $\phi(s)$, we define a reward function

$$R_\phi(s, a, s') = \phi(s) - \gamma\phi(s').$$

**(a)** [2 pts] Prove that any policy is optimal for the MDP $(S, T, \gamma, R_\phi)$. Hint: It is sufficient to prove for any policy $\pi$ that $V^\pi(s) = \phi(s)$ satisfies the Bellman equation.

Consider again an arbitrary MDP with state set $S$, transition function $T(s, a, s')$, and discount factor $\gamma$. Suppose we are given two alternative reward functions $R_1(s, a, s')$ and $R_2(s, a, s')$ and that furthermore, there exists a single policy function $\pi^*(s)$ that is optimal for both the MDP $(S, T, \gamma, R_1)$ and also the MDP $(S, T, \gamma, R_2)$. We define a new reward function

$$R_{1+2}(s, a, s') = R_1(s, a, s') + R_2(s, a, s').$$

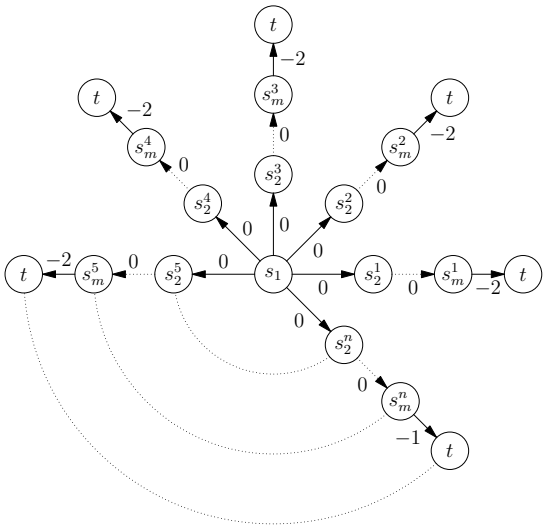
**(b)** [2 pts] Prove that $\pi^*(s)$ is optimal for the MDP $(S, T, \gamma, R_{1+2})$. Hint: It is sufficient to prove that $V^*_{1+2}(s) = V^*_1(s) + V^*_2(s)$ satisfies the Belman equation with policy $\pi^*$, where $V^*_1(s)$ and $V^*_2(s)$ are the optimal value functions for the MDPs $(S, T, \gamma, R_1)$ and $(S, T, \gamma, R_2)$, respectively. Also note that if $\arg\max_{x \in X} f(x) = \arg\max_{x \in X} g(x)$, then $\max_{x \in X}[f(x) + g(x)] = \max_{x \in X} f(x) + \max_{x \in X} g(x)$.

Suppose we are given an arbitrary MDP $(S, T, \gamma, R)$ *and additionally* a potential function $\phi(s)$ (with the constraint that $\phi(s) = 0$ for all terminal states $s \in S$).

**(c)** [2 pts] Prove that the modified MDP $(S, T, \gamma, R')$, where $R'(s, a, s') = R(s, a, s') + \phi(s) - \gamma\phi(s')$, has the same set of optimal policies as the original MDP. Hint: Use the previous results to prove that any optimal policy for the original MDP is optimal for the modified MDP, and any optimal policy for the modified MDP is optimal for the original MDP. Also note that $-R_\phi = R_{-\phi}$.

Modifying the rewards of an MDP in this way is called *reward shaping*. Given a potential function $\phi$, a reinforcement learning agent in the original MDP can pretend that it is operating in the modified MDP by pretending to have received a reward of $r + \phi(s) - \gamma\phi(s')$ whenever it actually receives a reward of $r$ for transitioning from $s$ to $s'$. Note that it is not necessary for the agent to have access to the transition model $T$. Because the modified MDP leaves the optimal policy unchanged, Q-learning on the modified MDP can be used to learn the optimal policy for the original MDP. The following questions show why this is useful.

Consider the MDP specified by the following star-shaped state graph.




The discount factor $\gamma$ is 1. All actions are deterministic and are indicated by directed edges. Each edge is labeled by the corresponding reward. There are a total of $n \cdot (m-1)+2$ states: $m-1$ distinct states in each of the $n$ branches, a start state $s_1$, and a terminal state $t$. All transitions to non-terminal states have a reward of 0, and all transitions to the terminal state $t$ have a reward of $-2$ except for the $s_m^n \to t$ transition which has a reward of $-1$.

**(d)** [1 pt] Assuming optimal exploration, what is the minimum number of episodes required for Q-learning to converge to the correct Q-values? Briefly explain your answer. Each episode begins at the start state $s_1$ and explores using an arbitrary sequence of actions (i.e. a sequence of your choosing) until the terminal state $t$ is reached. Assume that the Q-values are all initialized to 0. Since all transitions are deterministic, you can assume $\alpha = 1$.

**(e)** [2 pts] Specify a potential function $\phi(s)$ (with $\phi(t) = 0$) for this MDP such that a minimum number of Q-learning episodes (in the modified MDP) are required to converge to the correct Q-values (for the modified MDP). Also specify the number of episodes required. Assume that the Q-values are all initialized to 0, and that $\alpha = 1$. Briefly explain your answer.

# Q3. [11 pts] Medical Diagnosis

The Center for Disease Control has detected a new outbreak of a virus ($V$). Among those that have recently traveled ($T$), there is a 1/100 chance of contracting the virus, while among those that have not recently traveled, there is a 1/1000 chance of contracting the virus. 1/10 of the population has recently traveled. A symptom ($S$) is shown by 80% of those that have contracted the virus, and is also shown by 15% of those that have not contracted the virus. A blood test ($B$) for detecting infection by the virus is also available. Among those that have recently traveled, the test has a false negative rate of 1/100 and a false positive rate of 1/80. Among those that have not recently traveled, the test has a false negative rate of 1/50 and a false positive rate of 1/10. Note that the false positive rate is the probability of the test returning positive in the case that the virus is not actually present, and the false negative rate is the probability of the test returning negative in the case that the virus is present.

**(a)** [2 pts] Specify a *minimal* Bayesian network over the four binary variables $V$ (indicating presence of the virus), $T$ (indicating recent travel), $S$ (indicating presence of the symptom), and $B$ (indicating a positive detection by the blood test). The network should be minimal in the sense of making as many independence assumptions as possible. Also specify the conditional probability tables for each node.

**(b)** [3 pts] Compute the full joint distribution over $V$, $T$, $S$, and $B$. Calculate to 6 decimal places. Briefly show your work if you want partial credit (in the case that you make a trivial error). If you decide to write a program to compute this table, you can attach your source code.

| $V$ | $T$ | $S$ | $B$ | $\Pr(V,T,S,B)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 1 | |
| 0 | 0 | 1 | 0 | |
| 0 | 0 | 1 | 1 | |
| 0 | 1 | 0 | 0 | |
| 0 | 1 | 0 | 1 | |
| 0 | 1 | 1 | 0 | |
| 0 | 1 | 1 | 1 | |

| $V$ | $T$ | $S$ | $B$ | $\Pr(V,T,S,B)$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | |
| 1 | 0 | 0 | 1 | |
| 1 | 0 | 1 | 0 | |
| 1 | 0 | 1 | 1 | |
| 1 | 1 | 0 | 0 | |
| 1 | 1 | 0 | 1 | |
| 1 | 1 | 1 | 0 | |
| 1 | 1 | 1 | 1 | |

**(c)** [2 pts] A patient is showing the symptom ($S$) and the blood test ($B$) came up positive. It is unknown whether he has traveled recently (the patient is non-responsive). What is the probability that he has contracted the virus ($V$)? Briefly show your work.

Untreated, the virus has a fatality rate of 1/2, independent of any other factors. A treatment is available which is guaranteed to prevent fatality by the virus. Unfortunately, the treatment itself has side effects that have a fatality rate of 1/10, independent of whether the virus is actually present or any other factors.

**(d)** [2 pts] Should the treatment be administered to the patient, given that he is showing the symptom ($S$) and the blood test ($B$) came up positive? Briefly show your work.

**(e)** [2 pts] The hospital staff have learned that the patient mentioned recent travels to a relative. From experience, the relative knows that the patient tells him 90% of the time when he *has* traveled, but makes up stories about traveling 1% of the time that he *has not* traveled. Given this additional information, should the treatment be administered? Briefly show your work.