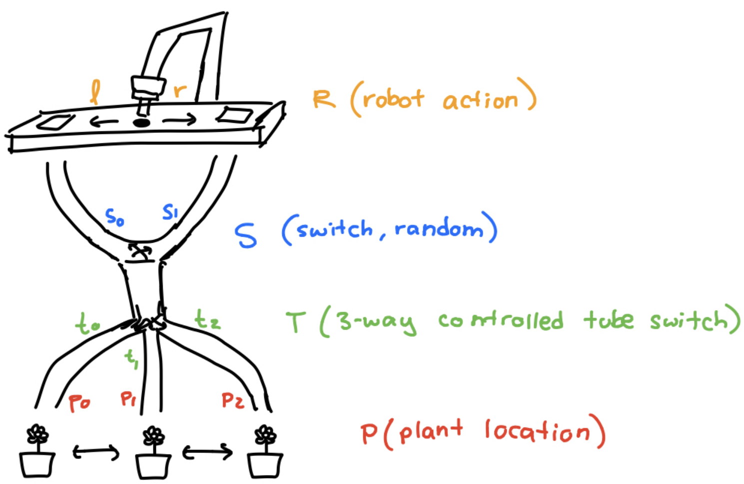
CS 188
Summer 2022

Final Review Utility / RL

Q1. Value of Perfect Information

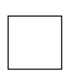
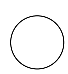
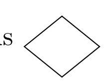
Consider the setup shown in the figure below, involving a robotic plant-watering system with some mysterious random forces involved. Here, there are 4 main items at play.
(1) The robot (R) can choose to move either left (l) or right (r). Its chosen action pushes a water pellet into the corresponding opening.
(2) The random switch (S) is arbitrarily in one of two possible positions $\{s_0, s_1\}$. When in position (s_0), it accepts a water pellet only from the (l) tube. When in position (s_1), it accepts a water pellet only from the (r) tube.
(3) A controllable three-way switch (T) can be chosen to be placed in one of three possible positions $\{t_0, t_1, t_2\}$.
(4) A plant (P) is arbitrarily located in one of three possible locations $\{p_0, p_1, p_2\}$. When in position p_i, it can only be successfully watered if the corresponding tube t_i has been selected **and** if the water pellet was sent in a direction that was indeed accepted by the first switch (S).

Finally, in this problem, utility (U) is 1 when the plant successfully receives the water pellet, and 0 otherwise.



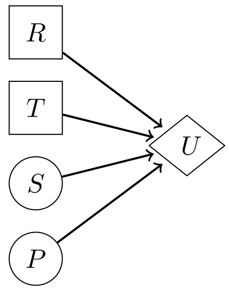
(a) Let's first set this problem up as a decision network.

(i) Which of the following decision networks correctly describe the problem described above? Select all that apply. Recall the conventions from the lecture notes:

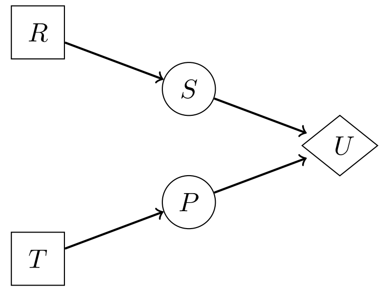
action nodes as rectangles , chance nodes as ovals , and utility nodes as diamonds 

.

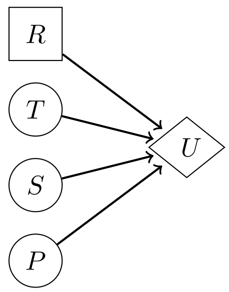
☐ (A)



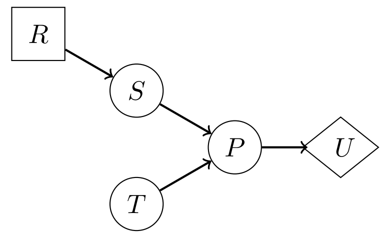
☐ (B)



☐ (C)



☐ (D)



(ii) Fill in the following probability tables, given that there is an equal chance of being at each of their possible locations.

S	$P(S)$

P	$P(P)$

(b) Before selecting your actions, suppose that someone could tell you the value of either S or P. Follow the steps below to calculate the **maximum expected utility (MEU)** when knowing S, or when knowing P. Then, decide which one you would prefer to be told.

(i) What is $MEU(S)$?

○ 0 ○ $\frac{1}{9}$ ○ $\frac{1}{6}$ ○ $\frac{1}{4}$ ○ $\frac{1}{3}$ ○ $\frac{1}{2}$
○ $\frac{2}{3}$ ○ $\frac{3}{4}$ ○ $\frac{5}{6}$ ○ 1 ○ None of the above

(ii) What is $MEU(P)$?

○ 0 ○ $\frac{1}{9}$ ○ $\frac{1}{6}$ ○ $\frac{1}{4}$ ○ $\frac{1}{3}$ ○ $\frac{1}{2}$
○ $\frac{2}{3}$ ○ $\frac{3}{4}$ ○ $\frac{5}{6}$ ○ 1 ○ None of the above

(iii) Would you prefer to be told S or P? ○ S ○ P

(c) **(i)** What is $MEU(S, P)$?

○ 0 ○ $\frac{1}{9}$ ○ $\frac{1}{6}$ ○ $\frac{1}{4}$ ○ $\frac{1}{3}$ ○ $\frac{1}{2}$
○ $\frac{2}{3}$ ○ $\frac{3}{4}$ ○ $\frac{5}{6}$ ○ 1 ○ None of the above

(ii) In this problem, does $VPI(S, P) = VPI(S) + VPI(P)$? ○ Yes ○ No

(iii) In general, does $VPI(a, b) = VPI(a) + VPI(b)$? Select all of the statements below which are true.

○ Yes, because of the additive property.
○ Yes, because the order in which we observe the variables does not matter.
○ Yes, but the reason is not listed.
○ No, because the value of knowing each variable can be dependent on whether or not we know

the other one.
- ◯ No, because the order in which we observe the variables matters.
- ◯ No, but the reason is not listed.

(d) For each of the following new variables introduced to this problem, what would the corresponding VPI of that variable be?

(i) A new variable X indicates the weather outside, which affects the overall health of the plant.
◯ VPI(X)< 0 ◯ VPI(X)$= 0$ ◯ VPI(X)> 0

(ii) A new variable X indicates the weather outside, which affects the metal of switch S such that when it's hot outside, the switch is most likely to remain in position s_0 with probability 0.9 (and goes to s_1 with probability 0.1).
◯ VPI(X)< 0 ◯ VPI(X)$= 0$ ◯ VPI(X)> 0

Q2. Policy Evaluation

In this question, you will be working in an MDP with states S, actions A, discount factor γ, transition function T, and reward function R.

We have some fixed policy $\pi : S \rightarrow A$, which returns an action $a = \pi(s)$ for each state $s \in S$. We want to learn the Q function $Q^\pi(s,a)$ for this policy: the expected discounted reward from taking action a in state s and then continuing to act according to π: $Q^\pi(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma Q^\pi(s', \pi(s'))]$. The policy π will not change while running any of the algorithms below.

(a) Can we guarantee anything about how the values Q^π compare to the values Q^* for an optimal policy π^*?

- ◯ $Q^\pi(s,a) \leq Q^*(s,a)$ for all s, a
- ◯ $Q^\pi(s,a) = Q^*(s,a)$ for all s, a
- ◯ $Q^\pi(s,a) \geq Q^*(s,a)$ for all s, a
- ◯ None of the above are guaranteed

(b) Suppose T and R are *unknown*. You will develop sample-based methods to estimate Q^π. You obtain a series of *samples* $(s_1, a_1, r_1), (s_2, a_2, r_2), \ldots (s_T, a_T, r_T)$ from acting according to this policy (where $a_t = \pi(s_t)$, for all t).

(i) Recall the update equation for the Temporal Difference algorithm, performed on each sample in sequence:

$$V(s_t) \leftarrow (1-\alpha)V(s_t) + \alpha(r_t + \gamma V(s_{t+1}))$$

which approximates the expected discounted reward $V^\pi(s)$ for following policy π from each state s, for a learning rate α.

Fill in the blank below to create a similar update equation which will approximate Q^π using the samples.

You can use any of the terms $Q, s_t, s_{t+1}, a_t, a_{t+1}, r_t, r_{t+1}, \gamma, \alpha, \pi$ in your equation, as well as $\sum$ and max with any index variables (i.e. you could write $\max_a$, or $\sum_a$ and then use a somewhere else), but no other terms.

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\,[\underline{\hspace{8cm}}]$$

(ii) Now, we will approximate Q^π using a linear function: $Q(s,a) = \mathbf{w}^\top \mathbf{f}(s,a)$ for a weight vector $\mathbf{w}$ and feature function $\mathbf{f}(s,a)$.

To decouple this part from the previous part, use Q_{samp} for the value in the blank in part (i) (i.e. $Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha Q_{samp}$).

Which of the following is the correct sample-based update for $\mathbf{w}$?

- ◯ $\mathbf{w} \leftarrow \mathbf{w} + \alpha[Q(s_t, a_t) - Q_{samp}]$
- ◯ $\mathbf{w} \leftarrow \mathbf{w} - \alpha[Q(s_t, a_t) - Q_{samp}]$
- ◯ $\mathbf{w} \leftarrow \mathbf{w} + \alpha[Q(s_t, a_t) - Q_{samp}]\mathbf{f}(s_t, a_t)$
- ◯ $\mathbf{w} \leftarrow \mathbf{w} - \alpha[Q(s_t, a_t) - Q_{samp}]\mathbf{f}(s_t, a_t)$
- ◯ $\mathbf{w} \leftarrow \mathbf{w} + \alpha[Q(s_t, a_t) - Q_{samp}]\mathbf{w}$
- ◯ $\mathbf{w} \leftarrow \mathbf{w} - \alpha[Q(s_t, a_t) - Q_{samp}]\mathbf{w}$

(iii) The algorithms in the previous parts (part i and ii) are:

☐ model-based ☐ model-free