# HW 1: MDPs, Policy Iteration, and Value Iteration

EECE 571N – Sequential Decision-Making (EECE 571N)

**Instructor: Cyrus Neary**
**Due: 2025-09-29 at 23:59 PT**

**Name:** Dominic Klukas **Student Number:** 64348378

## Instructions

Submit a single PDF to Canvas. Please include your name and student number at the top of that PDF. For all questions, show your work and clearly justify your steps and thinking. Please feel free to include any code as an attachment at the end of the PDF. State any assumptions. Unless otherwise specified, you may collaborate conceptually but must write up your own solutions independently.

## Grading

Points for each part are indicated. The total number of achievable points is 100. Partial credit is available for incorrect answers with clear reasoning.

## Problem 1

(a) Recall that a trajectory is a sequence $(s_0, a_0, s_1, a_2, s_2, \ldots)$. $\alpha(s)$ is the distribution of the initial states, so $s_0$. Then, given this distribution of initial states, $x_{s,a}$ is the expected discounted time that the trajectories will have $s$ followed by $a$. Discounted, in the sense that earlier times weigh more heavily in this expectation than later times. That is, if we expect more $(\ldots, s, a, \ldots)$ occurances later in the trajectories on average, then this will add less to $x_{s,a}$ than if the starting distribution weighs $s$ heavily and then the policy will very likely take action $a$.

The relation between the variables $x_{s,a}$ and $\pi$ is given by

$$\pi(a|s) = \frac{x_{s,a}}{\sum_{a' \in A} x_{s,a'}}.$$

We can understand this, since $x_{s,a}$ can be thought of as the expected behavior. Then, this fraction can be thought of as the fraction of the time that policy $\pi$ chooses action $a$ when in state $s$, since the sum in the denominator is the occupancy measure of the state $s$ itself, and of those times $x_{s,a}$ is the frequency the policy chooses $a$. This precisely describes the probability distribution of the policy.

(b) From the definition of $V^\pi$, we have

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | \pi, s_0 = s\right].$$

As discussed in class, this expectation is taken over all the different trajectories. Given the powerful conditional probability distribution $Pr(S_t = s, A_t = a | S_0 = s', \pi)$, we can see

that one way of writing the expectation over all trajectories, is by summing this probability distribution over all possible states, actions, and times:

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t \sum_{s \in S} \sum_{a \in A} R(a,s) P(S_t = s, A_t = a | S_0 = s, \pi).$$

Then, the result follows quickly:

$$\begin{aligned}
x^{\pi}_{s,a} &= \sum_{s' \in S} \alpha(s') \sum_{t=0}^{\infty} \gamma^t Pr(S_t = s, A_t = a | S_0 = s', \pi) \\
\sum_{s \in S} \sum_{a \in A} x^{\pi}_{s,a} R(s,a) &= \sum_{s \in S} \sum_{a \in A} \sum_{s' \in S} \alpha(s') \sum_{t=0}^{\infty} \gamma^t Pr(S_t = s, A_t = a | S_0 = s', \pi) \\
\sum_{s \in S} \sum_{a \in A} x^{\pi}_{s,a} R(s,a) &= \sum_{s' \in S} \alpha(s') \sum_{t=0}^{\infty} \gamma^t \sum_{s \in S} \sum_{a \in A} R(s,a) Pr(S_t = s, A_t = a | S_0 = s', \pi) \\
\sum_{s \in S} \sum_{a \in A} x^{\pi}_{s,a} R(s,a) &= \sum_{s' \in S} \alpha(s') V(s').
\end{aligned}$$

## Problem 2

(a) The formula for the updated belief $b'(s')$ is given by:

$$b'(s') = \frac{O(o_{t+1} | s', a_t) \sum_{s \in S} T(s' | s, a_t) b_t(s)}{\sum_{s'' \in S} O(o | s'', a) \sum_{s \in S} T(s'' | s, a) b(s)}.$$

(b) The sum containing the transition probabilities in the numerator denotes the probability of state $s'$ given that we start in state with belief $b(s)$ about our states $s$ and take action $a$, without any extra knowledge of the observation that we make at time $t+1$. That is, simply $P(s' | b(s), a)$. However, belief by definition incorporates the knowledge that we get from the observation we make at time $t+1$, when $s'$ occurs. That is, $b(s') = P(s' | b(s), a, o')$. By Bayes rules for conditional probability, we have:

$$P(s' | b(s), a, o') = \frac{P(o' | s', b(s), a) P(s' | b(s), a)}{P(o' | b(s), a)}.$$

In our case,

$$\begin{aligned}
P(s' | b(s), a) &= \sum_{s \in S} T(s' | s, a_t) b_t(s) \\
\frac{P(o' | s', b(s), a)}{P(o' | b(s), a)} &= \frac{O(o_{t+1} | s', a_t)}{\sum_{s'' \in S} O(o | s'', a) \sum_{s \in S} T(s'' | s, a) b(s)}.
\end{aligned}$$