

# 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  $(s_0, a_0, s_1, a_2, s_2, \dots)$ .  $\alpha(s)$  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  $(\dots, s, a, \dots)$  occurrences 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:

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

## Problem 2

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

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

- (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,

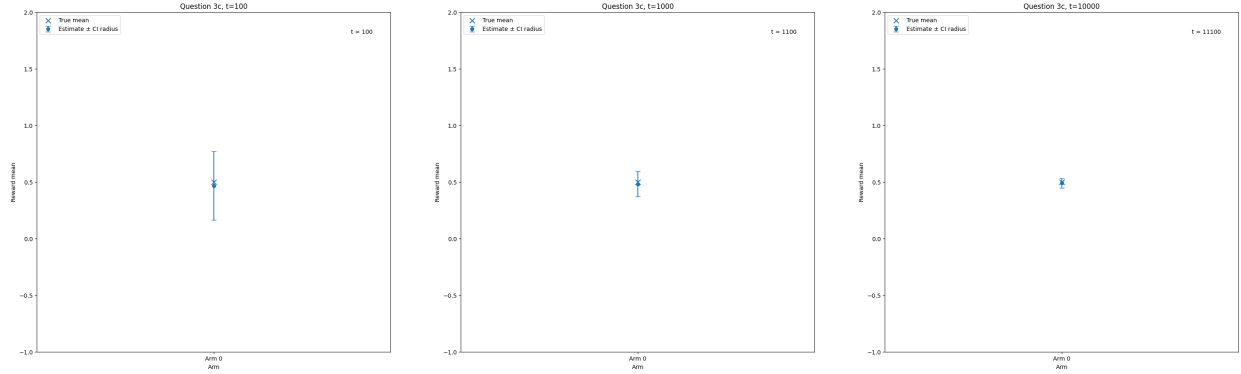
$$\begin{aligned} P(s'|b(s), a) &= \sum_{s \in S} T(s'|s, a_t) b_t(s) \\ \frac{P(o'|s', b(s), a)}{P(o'|b(s), a)} &= \frac{O(o_{t+1}|s', a_t)}{\sum_{s'' \in S} O(o_{t+1}|s'', a) \sum_{s \in S} T(s''|s, a) b(s)}. \end{aligned}$$

That is, the observation factor (the second equation here) conditions the probability on the case that the observation  $o_{t+1}$  was made.

## Problem 3

## Problem 5

**Answer to 5(c): Exploration vs. exploitation in UCB1.** The UCB chooses the arm which maximizes the value of the calculated mean plus the confidence interval. This balances the exploration

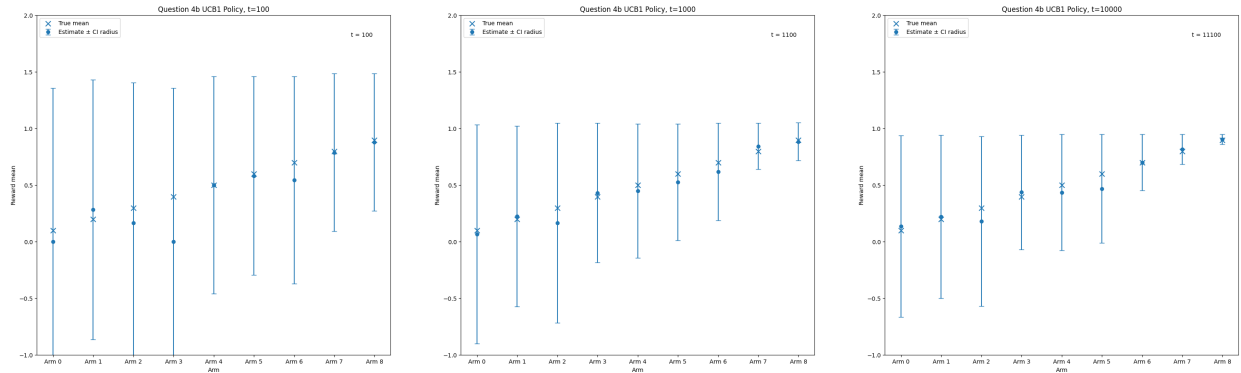


(a)  $t = 100$

(b)  $t = 1000$

(c)  $t = 10000$

Figure 1: Q3: One-armed bandit confidence intervals over time (uniform policy).



(a)  $t = 100$

(b)  $t = 1000$

(c)  $t = 10000$

Figure 2: Q5: **UCB1** on the 9-armed Bernoulli environment (note: assignment text typo refers to “uniform random” here; figure shows UCB1).

and exploitation precisely because, if an arm has high variance it has an advantage, but then as the policy exploits this advantage, the arm’s confidence interval shrinks, so if the mean plus confidence interval of this arm still beats the other arms this will require the mean to increase and be higher than the means of the arms with larger confidence intervals. The value (mean + confidence interval) is precisely the top of the error bar. This is why the policy appears to balance the tops of the error bars, because it chooses the arm with the highest top-of-error-bar, and choosing this arm shrinks the confidence interval thereby lowering it. This is why we can see that all of the tops of the error bars go down uniformly as we progress through the time horizon.

## Problem 6

**Answer to 6(f): Interpreting the regret curves.** We discussed this in class.

- Uniform: the linear shape is as a result of the policy randomly choosing arms throughout the entire horizon, so it’s expected rate of income stays the same throughout the horizon. It

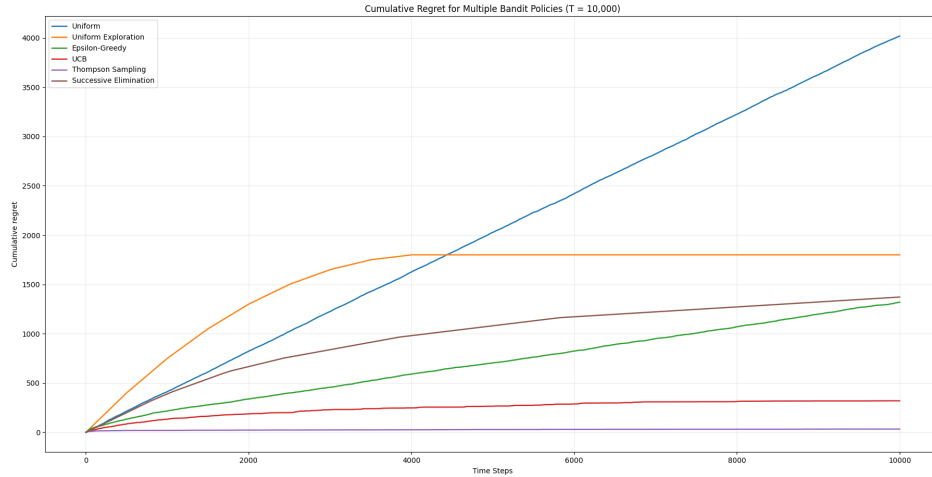


Figure 3: Q6(e): Cumulative regret comparison for all algorithms (horizon  $T = 10,000$ ).

doesn't learn or gain any information from previous trials.

- Uniform Exploration curve: the way I coded the algorithm, it explores each arm 500 times before moving onto its 500 trials on the next arm. Since the order in which the policy explores the arms starts with the  $p = 0.1$  arm and works its way up to  $p = 0.9$ , we can see that the rate of regret decreases as the policy successively selects the more fruitful arms, which explains the curve. After the exploration phase is over, at  $t = 5000$ , the policy has correctly identified the best arm and exploits it for the rest of the time period.
- Successive elimination: we can see the kinks in the shape of the graph, where the policy "drops" the poorest performing arms. At the start, it shares the same slope with the uniform policy when it is still pulling all of the arms, until it drops its first arm at around  $t = 500$  or so.
- Epsilon-Greedy: although the rate of increase goes down, it continues to have an increasing linear shape due to the random sampling with probability  $\varepsilon = 0.3$ .
- UCB and Thompson both perform significantly better than the other policies.