



## Homework 1

Due: November 29, 2023

Questions in this Homework appear *framed and in italics*.

Theoretical questions are marked with the symbol  and are worth 4 points in total. Questions that require implementation are marked with the symbol  and are worth 6 points in total. You should submit a pdf file with your answers to [fmelo@inesc-id.pt](mailto:fmelo@inesc-id.pt). As an appendix to your document, you may include the code used to generate your answers. Note, however, that code is not considered an answer.

Do not be excessively concerned with the evaluation facet of the homework. If you have any difficulties in understanding a question or coming up with an answer, come forward and discuss with me and the rest of the class—treat the homework more as a tool to help you learn than to grade you.

### 1 Multi-armed bandits (3 pts.)

In this part of the homework, you will reproduce some of the results in Chapter 2 of the Sutton and Barto book.


Suppose that you have a 10-armed bandit problem—i.e., an agent has to select among a set of 10 different actions, each of which yields a different (average) reward. In particular, suppose that the *actual* reward associated with each action  $a \in \{1, \dots, 10\}$  is a random variable following a Gaussian distribution with mean  $Q(a)$  and variance 1.

You will compare the average reward received by:

- A greedy policy, where the estimated reward for each action  $a$  is initialized as  $\hat{Q}(a) = 0$ ;
- A greedy policy, where the estimated reward for each action  $a$  is initialized as  $\hat{Q}(a) = 5$ ;
- An  $\varepsilon$ -greedy policy, with  $\varepsilon = 0.1$ , where the estimated reward for each action  $a$  is initialized as  $\hat{Q}(a) = 0$ ;
- An  $\varepsilon$ -greedy policy, with  $\varepsilon = 0.01$ , where the estimated reward for each action  $a$  is initialized as  $\hat{Q}(a) = 0$ ;
- The UCB policy.

To do this,

1. Select the average values  $Q(a), a \in \{1, \dots, 10\}$ , at random from a normal distribution with mean 0 and variance 1;
2. Run the policies listed above for 1,000 steps and compute, for each policy, the reward received at each time step;
3. Repeat the two previous steps 2,000 times;
4. Compute the reward obtained at each time step by each of the policies averaged across the 2,000 trials.

 **Question 1.** Using the results from the procedure outlined above, plot the average reward obtained with each policy during the 1,000 steps in a single plot. Comment the differences observed.


## 2 The gambler problem (3 pts.)

Consider the following problem, introduced in Chapter 4 of the Sutton and Barto book.

A gambler is engaged in a betting game, where he must place bets on the outcomes of a sequence of coin flips. Before each flip, the gambler decides how much to bet on the outcome that the coin will come up heads—note that he can only decide how much to bet, not in which outcome to bet.

If the coin does come up heads, the gambler doubles the money bet on that coin flip—in other words, if the gambler bets \$5 dollars, he will get his \$5 dollars back plus another \$5 dollars. If the coin comes up tails, the gambler loses the money he bet. The game goes on until either the gambler reaches his goal of \$100 dollars, or loses by running out of money. On each flip, the gambler must decide what portion of his capital to stake, in integer numbers of dollars. Suppose that the probability of the coin coming up heads is  $p_H = 0.4$ .

You will analyze the optimal betting policy for the gambler. To do so, model the gambler's decision problem as an MDP, specifying the state and action spaces, the transition probabilities and the reward function. *Make sure to use a reward function that only rewards the gambler for reaching his goal.* Use a discount  $\gamma = 1$ .

 **Question 2.** Using the MDP model for the gambler problem, run value iteration. Plot, in the same plot, the computed estimate for the optimal value function at iterations 1, 2, 3 and final (stop your algorithm when the overall error is smaller than  $10^{-8}$ ). Can you provide an interpretation for the values obtained? Plot also the optimal policy computed.

**Note:** The gambler problem is somewhat numerically unstable, so to compute the optimal policy make sure to rounded up all values to 4 decimal places.

## 3 Convergence of value iteration (2 pts.)

Given a deterministic policy  $\pi$ , the value iteration algorithm to compute  $v_\pi$  relies on the recursion

$$v_\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, \pi(s)) v_\pi(s'), \quad (1)$$

where  $r(s, a)$  is the average reward for taking action  $a$  in state  $s$  and

$$p(s' | s, a) = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a].$$

In particular, given any initial estimate  $v^{(0)}$ , value iteration successively improves such estimate through the update

$$v^{(k+1)}(s) = r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, \pi(s)) v^{(k)}(s'). \quad (2)$$

In this question, you will show that the sequence  $\{v^{(k)}\}$  thus generated indeed converges to the desired result.

To do so, we start by noting that  $v_\pi$  as well as the estimates  $v^{(k)}$  can be written as column vectors  $\mathbf{v}_\pi$  and  $\mathbf{v}^{(k)}$  taking values in  $\mathbb{R}^{|\mathcal{S}|}$ . Similarly, given  $\pi$ , the rewards  $r(s, \pi(s))$  are a function of the state  $s$  only and, therefore, can also be written as a vector  $\mathbf{r}_\pi$ . Finally, the transition probabilities  $p(s' | s, \pi(s))$  depend only on  $s$  and  $s'$ , and can be compactly represented as a matrix  $\mathbf{P}_\pi$ . Using this notation, (1) and (2) can be written as

$$\mathbf{v}_\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}_\pi, \quad \mathbf{v}^{(k+1)} = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}^{(k)},$$

respectively.

To prove the convergence of value iteration, you will use the following simplified version of the *Banach fixed point theorem*.

**Theorem 1** (Banach fixed-point theorem,  $\mathbb{R}^n$  version). *Given a mapping  $\mathsf{T} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ , if  $\mathsf{T}$  is a contraction, i.e.,*

$$\|\mathsf{T}\mathbf{x} - \mathsf{T}\mathbf{y}\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|$$


*for any  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  and  $\gamma < 1$ , then it has a unique fixed point, i.e., a point  $\mathbf{x}^* \in \mathbb{R}^n$  such that*

$$\mathsf{T}\mathbf{x}^* = \mathbf{x}^*.$$

*Moreover, given any  $\mathbf{x}_0 \in \mathbb{R}^n$ , the sequence  $\{\mathbf{x}_k, k \in \mathbb{N}\}$  defined recursively as*

$$\mathbf{x}_{k+1} = \mathsf{T}\mathbf{x}_k, k \geq 0,$$

*converges to the unique fixed point of  $\mathsf{T}$ .*

 **Question 3.** *Using the theorem above, show that value iteration converges to  $v_\pi$ .*

**Suggestion:** *Define an operator  $\mathsf{T}$  that corresponds to the value iteration update and show that it is a contraction. Then use the theorem to establish the desired conclusion.*

## 4 The temporal difference operator (2 pts.)

The value function  $v_\pi$  associated with a policy  $\pi$  is defined for each state  $x \in \mathcal{X}$  as

$$v_\pi(s) = \mathbb{E}_\pi [G_0 | S_0 = s] = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s \right],$$

where  $R_t$  is the (random) reward received at time step  $t$ . From here, we then showed that

$$v_\pi(s) = \mathbb{E}_\pi \left[ R_0 + \sum_{t=1}^{\infty} \gamma^t R_t | S_0 = x \right] = r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, \pi(s)) v_\pi(s'). \quad (3)$$

From this recursive relation, you defined an operator  $\mathsf{T}$  corresponding to the value iteration update and showed that such operator converges. However, there is no particular reason to single out  $R_0$  in (3). In particular, if we write (3) in vector form, we get

$$\mathbf{v}_\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}_\pi,$$

which, by successively replacing  $\mathbf{v}_\pi$ , yields

$$\begin{aligned}
 \mathbf{v}_\pi &= \mathbf{r}_\pi + \gamma \mathbf{P}_\pi [\mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}_\pi] \\
 &= \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{r}_\pi + \gamma^2 \mathbf{P}_\pi^2 \mathbf{v}_\pi, \\
 \mathbf{v}_\pi &= \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{r}_\pi + \gamma^2 \mathbf{P}_\pi^2 \mathbf{r}_\pi + \gamma^3 \mathbf{P}_\pi^3 \mathbf{v}_\pi, \\
 &\vdots \\
 \mathbf{v}_\pi &= \sum_{n=0}^N \gamma^n \mathbf{P}_\pi^n \mathbf{r}_\pi + \gamma^{N+1} \mathbf{P}_\pi^{N+1} \mathbf{v}_\pi \\
 &\vdots
 \end{aligned}$$

Making a geometric combination of the righthand sides, we get, for any  $0 < \lambda < 1$ ,


$$\mathbf{v}_\pi = (1 - \lambda) \sum_{N=0}^{\infty} \lambda^N \left[ \sum_{n=0}^N \gamma^n \mathbf{P}_\pi^n \mathbf{r}_\pi + \gamma^{N+1} \mathbf{P}_\pi^{N+1} \mathbf{v}_\pi \right].$$

Some manipulations yield

$$\mathbf{v}_\pi = \sum_{n=0}^{\infty} (\lambda \gamma \mathbf{P}_\pi)^n [\mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}_\pi - \mathbf{v}_\pi] + \mathbf{v}_\pi.$$

From the recursive relation above, we can now define the operator  $\mathsf{T}^{(\lambda)} : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{|S|}$  as

$$\mathsf{T}^{(\lambda)} \mathbf{v} = \sum_{n=0}^{\infty} (\lambda \gamma \mathbf{P}_\pi)^n [\mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v} - \mathbf{v}] + \mathbf{v}.$$

 **Question 4.** Show that the operator  $\mathsf{T}^{(\lambda)}$  is a contraction with respect to the sup-norm.

**Suggestion:** You may want to use some of the following facts:

$$\begin{aligned}
 \sum_{n=0}^{\infty} c^n &= \frac{1}{1 - c}, \quad \text{for a scalar } c < 1; \\
 \|\mathbf{P}_\pi\| &\leq 1.
 \end{aligned}$$