Figure 2.21: A student MDP


Figure 1: A student MDP

Figure 1 is an example of an MDP representing the day of a university student. There are six possible states: Class 1, Class 2, Class 3, Social, Bed, and Pub. The edges between the states represent state transitions. On the edges, we have the action and the reward, denoted by r. Possible actions are Study, Social, Beer, and Sleep. The initial state, represented by the incoming arrow, is Class 1. The goal of the student is to select the best actions in each state, maximizing their return.

In the following paragraphs, we will discuss some possible strategies for this MDP.

A student agent starts from **Class 1**. They can decide to study and complete all of the lessons. Each study decision comes with a small negative reward, **-2**. If the student decides to sleep after **Class 3**, they will land in the absorbing state, **Bed**, with a high positive reward of **+10**. This represents a very common situation in daily routines. You have to sacrifice some immediate reward in order to obtain a higher reward in the future. In this case, by deciding to study in **Class 1** and **2**, you obtain a negative reward but are compensated by the positive reward after **Class 3**.

Another possible strategy in this MDP is to select a **Social** action right after the **Class 1** state. This action comes with a small negative reward. The student can continue doing the same action, and each time they get the same reward. The student can also decide to **Study** from the **Social** state (notice that **Social** is both a state and an action) by returning to **Class 1**. Feeling guilty, in **Class 1**, the student can decide to study. After having studied a bit, they may feel tired and decide to sleep for a little while, ending up in the **Bed** state. Having performed the **Social** action, the agent has cumulated a negative return.

Let's evaluate the possible strategies for this example. We will assume a discount factor of  formula, that is, no discount:

* Strategy: Good student. The good student strategy was the first strategy that was described. Supposing the student will end in **Class 1**, they can perform the following actions: **Study**, **Study**, and **Study**. The associated sequence of states is thus **Class 1**, **Class 2**, **Class 3**, and **Sleep**. The associated return is, therefore, the sum of the rewards along the trajectory:

Figure 2.22: Return for the good student


Figure 2: Return for the good student

* Strategy: Social student. The social student strategy is the second strategy described. The student can perform the following actions: Social, Social, Social, Study, Study, and Sleep. The associated sequence of states is Class 1, Social, Social, Social, Class 1, Class 2, and Bed. The associated return is, in this case, as follows:

Figure 2.23: Return for the social student


Figure 3: Return for the social student

* By looking at the associated return, we can see that the good student strategy is a better strategy in comparison to the social student strategy, having a higher return.

Let's now implement our understanding of the state- and action-value functions for our student MDP example. In this example, we will use the calculation of the state-value function and the action-value function for the student MDP in Figure 1. We will consider the case of an undecided student, that is, a student with a random policy for each state. This means that the probability of each action for each state is exactly 0.5.

We will examine a different case for a myopic student in the following example.

Solution Visual Representation:

Figure 2.38: State values of the student MDP for  


Figure Sol1: State values of the student MDP for b

Figure 2.39: State values of the student MDP for γ=0


Figure Sol2: State values of the student MDP for γ=0

Figure 2.42: Action values for the student MDP
Figure Sol3: Action values for the student MDP

Figure 2.43: The student MDP best actions
Figure Sol 4: The student MDP best actions

Here, the best action the agent can take, starting from Class 1, is Social as it provides a bigger immediate reward compared to the Study action. The Social action brings the agent to state Social. Here, the best the agent can do is to repeat the Social action, cumulating negative rewards.