למידה במערכות דינמיות אביב תש"פ

תרגיל בית 3

מגישים:

יאיר נחום 034462796

דר ערבה 205874951

* 1. It is easy to see that the optimal policy must be to reach state n as fast as we can since the reward is 0 on every other state. Thus, the policy is to go right on every stage. In stage n we have no other option than to go back to 0.
  2. For the fixed policy we can apply policy evaluation by solving |S| equations (the same as calculating inverse matrix ). Also, we notice that in this case the MDP is deterministic once we know the action by the policy:

We can solve these equations by noticing the connection between the first stage and the last stage:

* 1. According to total probability on the actions, we can calculate the transitions matrix as follows:  
     for each state other than n, the policy is stochastic with 0.5 chance for every action.  
     In order to calculate the stationary distribution of the states, one needs to solve the vector equation:

Or per state probability:  
According to the formula above for the transition probabilities we have n equations and also the demand:

We note that the dynamics are deterministic once the action is selected, thus, the transition probability is actually the policy probability.

Again, by applying the recursion of the equations and using the normalization to a probability vector we get:

* 1. FPVI specifies

In matrix description   
In our case, since we started with then we get:   
When (as we saw the transition matrix can be calculated):  
Therefore, If = 0, then:

From FPVI convergence Proposition 5.2, we get that   
BTW, can be calculated directly from as we proved in Lemma 5.2 in the lecture.

**2.**

**3.**

**4.**

**5.**