TUTORIAL 2: EXERCISES

O.K.

MARKOV DECISION PROCESSES

PROBLEM 1 (MARKOV DECISION PROCESSES)

Consider the following game. We have the player A and the dealer B. A has two actions: "play" and "stop". The player A starts with zero tokens.

- If A chooses the action "play", the dealer rolls a three (sic!) sided dice. Let D be the outcome of the roll. The player then gets D tokens. If the player has more than 3 tokens, the game ends and the player does not get any reward.
- If A chooses the action "stop", the game moves into a terminal state and the player receives a reward equal to the number of the player's tokens.

We want to formalize the problem as a Markov decision problem and find an optimal policy for it (but we are noy going to do all of it... see next slides).

PROBLEM 1A, 3 MINUTES

How many states will the MDP have?

PROBLEM 1A - SOLUTION

How many states will the MDP have?

Answer: We will need 5 states, corresponding to "0 tokens", "1 token", ..., "3 tokens", "terminal state".

PROBLEM 1B, 3 MINUTES

What discount factor should we use?

PROBLEM 1B - SOLUTION

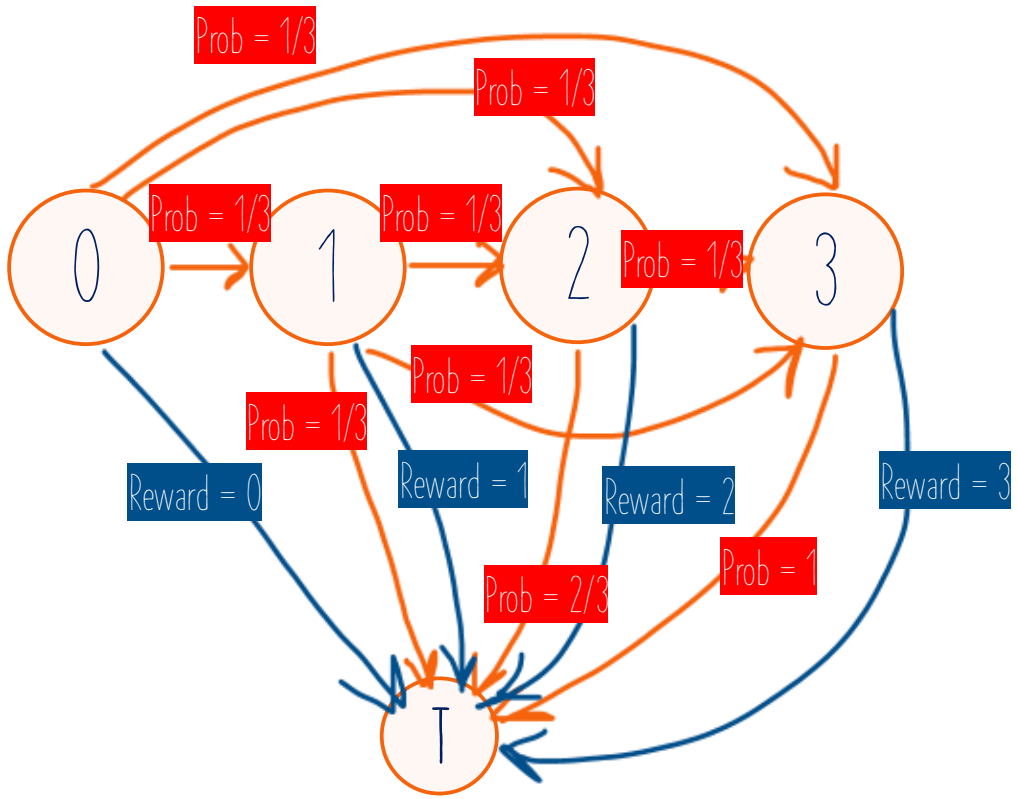
What discount factor should we use?

Answer: We should use discount factor = 1. In this example, there is no reason to do discounting. We want to maximize the sum of rewards and we do not care about how long that will take. Moreover, the game always ends after finitely many steps (so there are also no mathematical obstacles).

PROBLEM 1C, 10 MINUTES

Draw the resulting MDP.

PROBLEM 1C - SOLUTION



Blue lines are the "stop" actions. They always have deterministic effect. Therefore we do not show their probabilities.

Orange lines are the "play" actions. They always lead to 0 immediate reward. Therefore we do not show their rewards.

PROBLEM 1D, 10 MINUTES

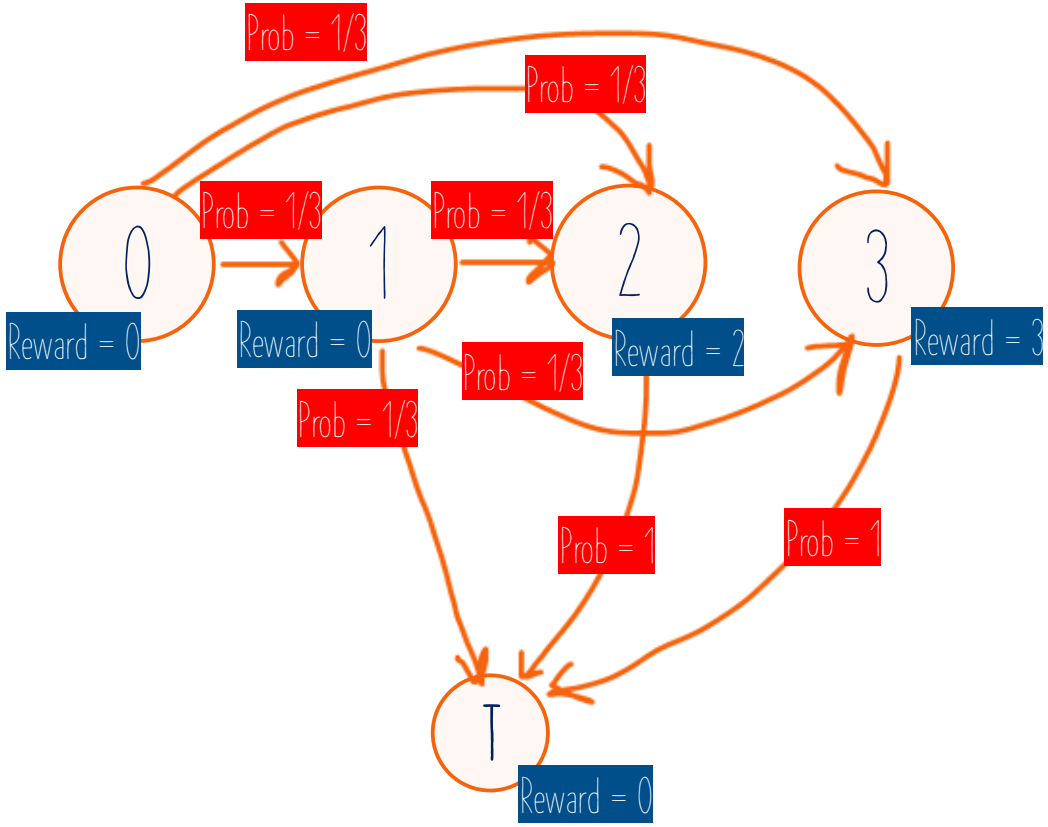
Consider the following deterministic policy:

$\pi(0)$ = "play", $\pi(1)$ = "play", $\pi(2)$ = "stop", $\pi(3)$ = "stop".

Do the following tasks:

- Compute and draw the Markov reward process obtained from our MDP and this policy.
- Compute the state-value function V with respect to this policy.

PROBLEM 1D - SOLUTION (MRP)



PROBLEM 1D - SOLUTION (VALUE FUNCTION)

Similar to one of the previous examples. Time-permitting, we are going to do it on whiteboard.

$$
\begin{aligned}
&(1) \quad V(0) = \tfrac{1}{3}V(1) + \tfrac{1}{3}V(2) + \tfrac{1}{3}V(3) \\
&(2) \quad V(1) = \tfrac{1}{3}V(2) + \tfrac{1}{3}V(3) + \tfrac{1}{3}V(T) \\
&(3) \quad V(2) = 2 + V(T) \\
&(4) \quad V(3) = 3 + V(T) \\
&(5) \quad V(T) = 0
\end{aligned}
$$

$$
\begin{aligned}
V(0) &= \frac{20}{9} \approx 2.2222 \\
V(1) &= \frac{5}{3} \approx 1.6667 \\
V(2) &= 2 \\
V(3) &= 3 \\
V(T) &= 0
\end{aligned}
$$

PROBLEM 1E, 5 MINUTES

Is the policy from Problem 9E optimal?

PROBLEM 1E - SOLUTION

Is the policy from Problem 9E optimal?

Solution: We are going to compute the Q-function with respect to the given policy. If it is already the greedy policy, we know it is optimal. Otherwise, it is not optimal.

We have, e.g.,: Q(0,"play") = 0 + 1/3 * (V(1) + V(2) + V(3)) = 1/3 * (1.6667 + 2 + 3) = 2.222, Q(0,"stop") = 0 + 0 = 0 and so on....

PROBLEM 1F, 15 MINUTES

Start with the initial policy:

$\pi(0)$ = "stop", $\pi(1)$ = "stop", $\pi(2)$ = "stop", $\pi(3)$ = "stop".

Use policy iteration to find the optimal policy. You should do the policy iteration by hand but you can use external tools to solve the systems of equations you get on the way.