**Homework 2**

Ido Glanz – 302568936, Matan Weksler – 302955372

**Question 1 - Markov Chains**

* 1. The state transition diagram for this process would be:

To show the chain is recurrent (meaning in other words all states will be re-visited with a probability tending to 1 at some point) we need to show there exists a stationary distribution and then as shown in class for countable Markov chains the chain is recurrent.

More formally:

Plugging the given probabilities:

Giving us the solution:

Hence the probability of re-visiting a state is greater than 0 (always) and at some point in time the state will be reached again (meaning the chain is indeed recurrent).

* 1. First let us prove is a Markov process (meaning it’s state depends only on the state preceding it and the action if relevant)

Applying Bayes rule, we obtain:

as depends solely on and as does not depend on (as it is Markovian) we can write:

using Bayes rule again we derive:

Proving is indeed only depended on and is a Markov Process.

**Claim 2 –is transient**

Intuitively, as is the number of times state 0 was visited before arriving to state , and as reacing again would necessarily require a visit to state 0 (the probability of reaching it from a later state is zero) – the next time we would visit , would necessarily increase by 1 and reaching the same state is not possible. Hence there isn’t a feasible way to revisit a state and the process is hence transient.

* 1. In this state, we are actually in a null-recurrent process (randomly walking) and it doesn’t have a stationary distribution.

Trying to obtain when we derive:

To be a stationary distribution it would need to satisfy:

and as , must tend to 0 to satisfy the equation and a stationary distribution doesn’t exist.

**Question 2**

* + 1. What is the expected return assuming Jacks’ action sequence is 212:

To evaluate the reward for each action at any given state (0,1,2) we will use the expected reward as follows:

Giving the probably and reward matrices:

Hence the expected reward after the 3 rounds (given S0=0) would be:

* + 1. Now Jack changes his strategy as so at each state (starting from 0) he chooses the next one with a uniform probability of 0.5. Hence the probability of reaching a state i would be:  
       And the transition matrix and reward:  
         
       Therefore the expected reward will be:
    2. Next, we’ll solve the Bellman's equation of the problem for 3 iterations:

In general the equations are as follows:

now we’ll run the equations for 3 iteration starting from 0 and initializing V0(s)=0:

For the second iteration:

And for the last step:

Hence, we can now determine the policy for 3 iterations (choosing the best action to act upon at each state, based on the max reward):

* + 1. Now we add a discount factor with the probability of being thrown out of the casino (hence ending the game). Accounting for the discount factor, the expected reward will be:  
       Hence the discount factor will be and the process will die out quicker as beta decreases.
    2. The new Bellman's equations will be:

**Question 3 (Modeling an Inventory MDP)**

* 1. In this section, we can model each day as an independent day since all the hot-dogs are thrown away at the end of the day.

Time (in hours) will be marked as:

Where 0 is the opening time and 13 is the last hour of the day.  
The state will represent the number of hot-dogs in the stand:  
The action corresponds to the number of hot dogs to order at each hour:  
The reward is our selling profit – meaning the difference between the number of hot dogs we have now to what we had an hour ago, subtracting the number of hot dogs we ordered (an hour and dt ago):

The transition probabilities will be then:  
As we know the distribution of the number of hot dogs bought (by customers) is   
Hence:

And for each day our total gain will be the sum of all the estimated rewards.

In this section in addition to previous section we will have to add time indication for each state (at least 2 hours back since the hot dogs are then thrown)  
The state space will be as follow:

* 1. In this case will state the same except in the rush hours that the demand will double and the transition distribution function will be as follows for the rush hours only:
  2. In the case of a 24 hour stand where hot dogs aren’t thrown away at all, but there is a yearly inflation ratio of 3%, we can now look at a t as an increasing whole number in hours and the performance criteria will be:
  3. Last section is in general a combination of sections d and c – we unfortunately didn’t have time to explicitly write it due to being stuck on other sections...

**Question 4: Prove the following equality:**

let:

As r is independent of time (stationary) and so is the policy – and we get:

**Question 5: Second moment and variance of return**

1. *We will show that satisfies a ’Bellman like’ set of equations:*

1. How many equations are needed to solve in order to calculate for all ?

meaning equations in total

1. We define as follow:

How may be calculated?

will be calculated as seen in section a,b