1. A robot is trapped in a maze. Initially it has to choose one of two directions. If it goes to the right, then it will wander around in the maze for three minutes and will then return to its initial position. If it goes to the left, then with probability 1/3 it will depart the maze after two minutes of maneuvering, and with probability 2/3 it will return to its initial position after five minutes of travelling. Assuming that the robot is at all times equally likely to go to the left or the right, what is the expected number of minutes that it will be trapped in the maze?

**Answer.** Let $N$ denote the number of minutes in the maze. If $L$ is the event the robot chooses left, and $R$ the event it chooses right, we have by conditioning on the first direction chosen:

$$E[N] = \frac{1}{2}E[N|L] + \frac{1}{2}E[N|R]$$

$$= \frac{1}{2} \left[ \frac{1}{2} + \frac{2}{3}(5+E[N]) \right] + \frac{1}{2} \left[ 3+E[N] \right]$$

$$= \frac{5}{6}E[N] + \frac{21}{6}.$$  


2. Consider the following Markov reward process. Suppose a discount factor of $\gamma = 0.5$.

$$
\begin{align*}
\begin{cases}
  s: & s (r = 6, p = 0.6) \quad ; a (r = -4, p = 0.4), \\
  a: & a (r = 2, p = 0.2) \quad ; c (r = -3, p = 0.8), \\
  b: & s (r = 3, p = 0.5) \quad ; c (r = -1, p = 0.5), \\
  c: & a (r = -1, p = 0.6) \quad ; b (r = 1, p = 0.2) \quad ; c (r = -1, p = 0.2).
\end{cases}
\end{align*}
$$  

Node $s$ is the start node.

(a) Give a state transition diagram, including probabilities and rewards.

(b) Give the probability transition matrix, $P$, and the immediate reward matrix, $R$.

**Answer.**

- **State Transition Diagram:**
  - $s$: Probability of staying $s = 0.6$, leaving to $a = 0.4$.
  - $a$: Probability of staying $a = 0.2$, leaving to $c = 0.8$.
  - $b$: Probability of staying $b = 0.5$, leaving to $c = 0.5$.
  - $c$: Probability of staying $c = 0.6$, leaving to $b = 0.2$.

- **Probability Transition Matrix:** $P = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 \\ 0 & 0.2 & 0 & 0.8 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0.6 & 0.2 & 0.2 \end{pmatrix}$

- **Immediate Reward Matrix:** $R = \begin{pmatrix} 6 & -4 & 3 & 0 \\ 0 & 2 & \frac{2}{3} & 3 \\ 0 & 2 & -3 & 1 \\ 0 & 3 & -1 & -1 \end{pmatrix}$
The order of the nodes is \( s, a, b, \ldots \), so \( r_{00} = r_{s,s}, \ r_{01} = r_{s,a}, \ldots \), and \( p_{00} = p_{s,s}, \ p_{01} = p_{s,a}, \ldots \). Blank entries are irrelevant. (You may give them any value.)

(c) Express the vector of optimal values \( v^* = (s, a, \ldots, c) \) as a solution of a system of linear equations. It is not necessary to simplify this system of linear equations.

**Answer.**

\[
\begin{align*}
  s &= 0.6(6 + \gamma s) + 0.4(-4 + \gamma a), \\
  a &= 0.2(2 + \gamma a) + 0.8(-3 + \gamma c), \\
  b &= 0.5(3 + \gamma s) + 0.5(-1 + \gamma c), \\
  c &= 0.6(-1 + \gamma a) + 0.2(1 + \gamma b) + 0.2(-1 + \gamma c).
\end{align*}
\]

On decision nodes (provided there are any) the probabilities must represent the policy that is to be evaluated.

(d) Perform two steps of value evaluation. Use a table to show the results.

**Answer.** Iterate the Bellman equation

\[
V(i) = \sum_{all \ j} p_{ij}[r_{ij} + \gamma V(j)],
\]

(the one for Markov reward processes) on all nodes.

<table>
<thead>
<tr>
<th>Value evaluation</th>
<th>( s )</th>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start values ( v_0 )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>( v_1 )</td>
<td>2.00</td>
<td>-2.00</td>
<td>1.00</td>
<td>-0.60</td>
</tr>
<tr>
<td>( v_2 )</td>
<td>2.20</td>
<td>-2.44</td>
<td>1.35</td>
<td>-1.16</td>
</tr>
</tbody>
</table>
Scratch paper.
Scratch paper.