Solutions to Homework 3

1) \(d(s_1) = a_{11} \quad d(s_2) = a_{21}\)
   
   \(s_1\) is recurrent, \(s_2\) is transient

2) \(d(s_1) = a_{11} \quad d(s_2) = a_{22}\)
   
   \(s_1\) is an absorbing state
   \(s_2\) is an absorbing state

3) \(d(s_1) = a_{11} \quad d(s_2) = a_{23}\)
   
   \(s_1\) is recurrent, \(s_2\) is transient

4) \(d(s_1) = a_{21} \quad d(s_2) = a_{21}\)
   
   the chain is irreducible and hence, \(s_1\) and \(s_2\) are recurrent

5) \(d(s_1) = a_{12} \quad d(s_2) = a_{23}\)
   
   \(s_1\) is transient
   \(s_2\) is recurrent

6) \(d(s_1) = a_{12} \quad d(s_2) = a_{23}\)
   
   the chain is irreducible and hence, \(s_1\) and \(s_2\) are recurrent

   Thus, multiple chain and communicating.
2) Suppose that there exists a randomized decision rule \( \pi \) under which the Markov chain is multichain. Then there exist two sets of states \( S_1 \) and \( S_2 \) that are disjoint and recurrent.

Let \( i \in S_1 \) and \( j \in S_2 \). Thus, \( \pi^n(i,j) = 0 \) for all \( n \geq 0 \).

Construct a deterministic decision rule \( \sigma \) as follows. For all \( s \in S \) let \( \sigma(s) = s_0 \) for which \( \pi(d(s) = s_0) > 0 \). Then \( \pi^n(i,j) = 0 \) for all \( n > 1 \) (where \( i \) and \( j \) are defined as above).

But this contradicts the assumption that all deterministic policies are unichain.
3) The problem could be modelled as a Markov decision process.

with state space \( S = \{s_1, s_2, s_3, s_4, T\} \) where \( T \) denotes termination.

Let 0 denote the action of continuing and 1 denote the action of leaving the system. Then

\[
A_0 = \begin{cases} 
1, 1 & \text{if } s = s_1, s_2, s_3, s_4, \\
\text{do nothing} & \text{if } s = T 
\end{cases}
\]

Let \( d_0(s_1) = d_0(s_2) = d_0(s_3) = d_0(s_4) = 0 \quad d_0(T) = \text{do nothing} \)

\[
c_{d_0} = \begin{bmatrix} 
1 \\
2 \\
3 \\
4 \\
0 
\end{bmatrix} \quad P_{d_0} = \begin{bmatrix} 
0.3 & 0.4 & 0.2 & 0.1 & 0 \\
0.2 & 0.3 & 0.5 & 0.0 & 0 \\
0.1 & 0.0 & 0.8 & 0.1 & 0 \\
0.4 & 0.0 & 0.0 & 0.6 & 0 \\
0.0 & 0.0 & 0.0 & 0.0 & 1 
\end{bmatrix}
\]

\[
v_0 = c_{d_0} + \gamma P_{d_0} v_0
\]

\[
v_0(s_1) = 24.0775 \quad v_0(s_2) = 25.5087 \quad v_0(s_3) = 27.3053 \quad v_0(s_4) = 27.5385 \quad v_0(T) = 0
\]

Computing \( d_1 \in \arg \max \{c_d + \gamma P_{d_0} v_0\} \), we have \( d_1 = d_0 \) and the policy iteration terminates.

Thus, \( d^* = d_0 \).