COMP 424 A4 TUTORIAL

Jade Yu
lei.yu@mail.mcgill.ca

Hidden Markov Model

- Three fundamental problems for HMMs:

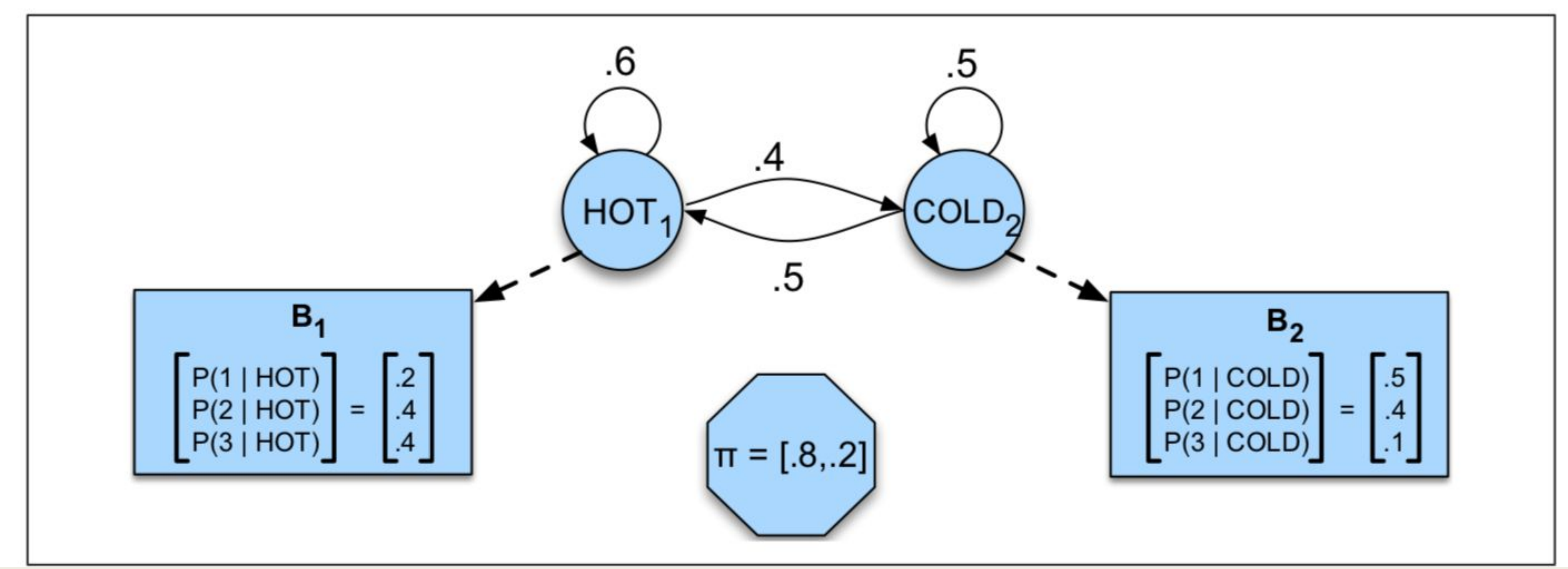
- Likelihood: Given an HMM $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ and an observed sequence O, find the probability $\boldsymbol{Pr}(\boldsymbol{O}|\boldsymbol{\lambda})$.
- Decoding: Given an observation sequence O and HMM $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, find the most probable sequence of hidden states: $\boldsymbol{Q}^* = \operatorname*{argmax}_{\boldsymbol{Q}} \boldsymbol{Pr}(\boldsymbol{O}|\ \boldsymbol{\lambda}, \boldsymbol{Q})$
- Learning: Given an observation sequence O, and sets of all possible hidden/observed states, find a best HMM model:

$$\boldsymbol{\lambda}^* = \boldsymbol{argmax}_{\lambda} \sum_{\boldsymbol{all\ possible\ hidden\ state\ seqs\ Q}} \boldsymbol{Pr}(\boldsymbol{O}|\ \boldsymbol{\lambda}, \boldsymbol{Q}) * \boldsymbol{Pr}(\boldsymbol{Q}|\boldsymbol{\lambda})$$

Problem	Algorithm
Likelihood	Forward Algorithm / Backward Algorithm
Decoding	Viterbi Algorithm (Dynamic Programming)
Learning	Baum-Welch Algorithm (EM)

A toy example by Eisner et al (2002):

- Imagine that you are a climatologist in the year 2799 studying the history of global warming. You cannot find any records of the weather in Baltimore, Maryland, for the summer of 2020, but you do find Jason Eisner's diary, which lists how many ice creams Jason ate every day that summer.
- Construct an HMM for this problem:



Let’s address three fundamental problems through this example:

- **Likelihood:** What is the probability of the observed sequence of ice cream amounts ’1 3 3’ , given the parameters in the previous slide?
- **Decoding:** Given an observed sequence of ice cream amounts ‘1 3 3’, what is the most likely sequence of weathers in 3 days, given the parameters in the previous slide?
- **Learning:** If now the parameters ($\lambda = \{\pi, A, B\}$) are unknown, and we observed a sequence of ice cream amounts ’1 3 3’, what is the best set of parameters for this model?

Utility Theory

What you should know after Jackie’s lecture:

- MEU principle
- Expected utility and maximum expected utility
- Value of information and value of perfect information

Example: St. Petersburg Paradox

- Consider a repeated lottery in which a fair coin (equal probability for heads and tails) is tossed repeatedly. For each round, the player has to pay α dollars at first, and will receive (2^k) dollars if the first head, say, occurs after k tosses of the coin.
  - What's the expected utility from playing this lottery if we assume that the utility always equals to the expected payoff?
  - Is our assumption reasonable?

Example: St. Petersburg Paradox

- Bernoulli’s proposition to address the paradox: people are risk aversive: our utility function is sub-linear (he postulated that if the returned payoff is $N, then the utility would be $log(N))
- The new expected payoff would then be:

$$P_s = \frac{1}{2}\log_{10} 2 + \frac{1}{4}\log_{10} 4 + \ldots + \left(\frac{1}{2}\right)^k \log_{10} 2^k + \ldots ,$$

,which is finite (0.60206).

- To design a similar “St. Petersburg Gamble” on a real scenario, we shall revise the rules so that players will not get infinite expected return (otherwise no bookies would pay off bets).
- We’ll prove that, in our new gambles, rational players will always win (have positive gains).
- The bookmakers are then motivated to earn profit by conspiracy: they’ll covertly use a coin with heavy tail to reduce players’ chances to win.
- What if the conspiracy is exposed? If you’re an “information broker” selling this secret to some players, what is the fair price for this information?

- **A modified version of SPP:** For each round, the player still pays α dollars at first, and will receive $\sqrt{\alpha * 2^k}$ dollars if the first head occurs after k tosses of the coin.
- To simplify our analysis, assume players are risk-neutral (that is, their utility function is equal to their expected net gain), and can always decide an optimal paying strategy given provided information.
- Let β denote the probability that the coin shows its head for each toss. (In our previous slides $\beta = 0.5$)

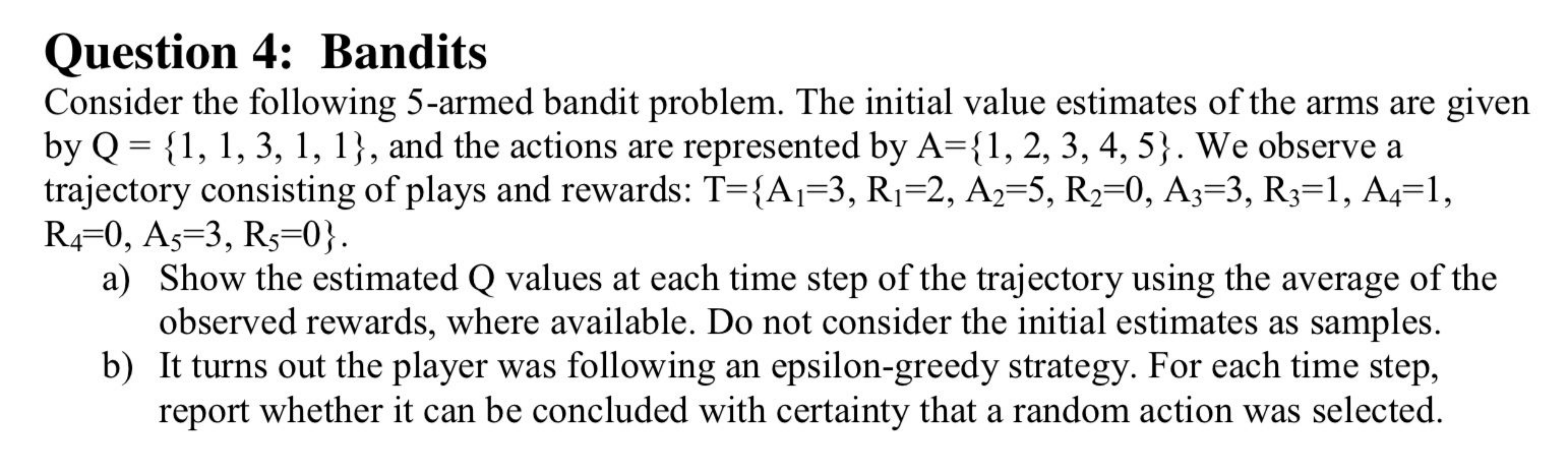
- Given the settings above, for $\beta = 0.5$, find an optimal α^* to maximize the expected utility, and the corresponding maximal utility.
- For any value of β, decide the optimal value $\alpha^*(\beta)$ to maximize the expected utility, and the corresponding maximal utility.
- If the bookmakers conspire to use a coin with heavier tail $(0 < \beta < 0.5)$ without telling the players. For what values of β could the bookies have positive expected gain?

- Assume the playmakers are using a coin with $\beta = 0.3$ (though he claims that the coin is fair).
- If you're a "information broker" who happens to know this conspiracy: **the playmaker is using an heavy-tailed coin with $\boldsymbol{\beta = 0.3}$.**
- Now if you want to sell this information to the players, can you set a fair price for it?

Bandits

- Estimate action values: $Q_n(a) = (r_1 + r_2 + \ldots + r_n) / n$

- Exploration-exploitation trade-off: ε-greedy policy

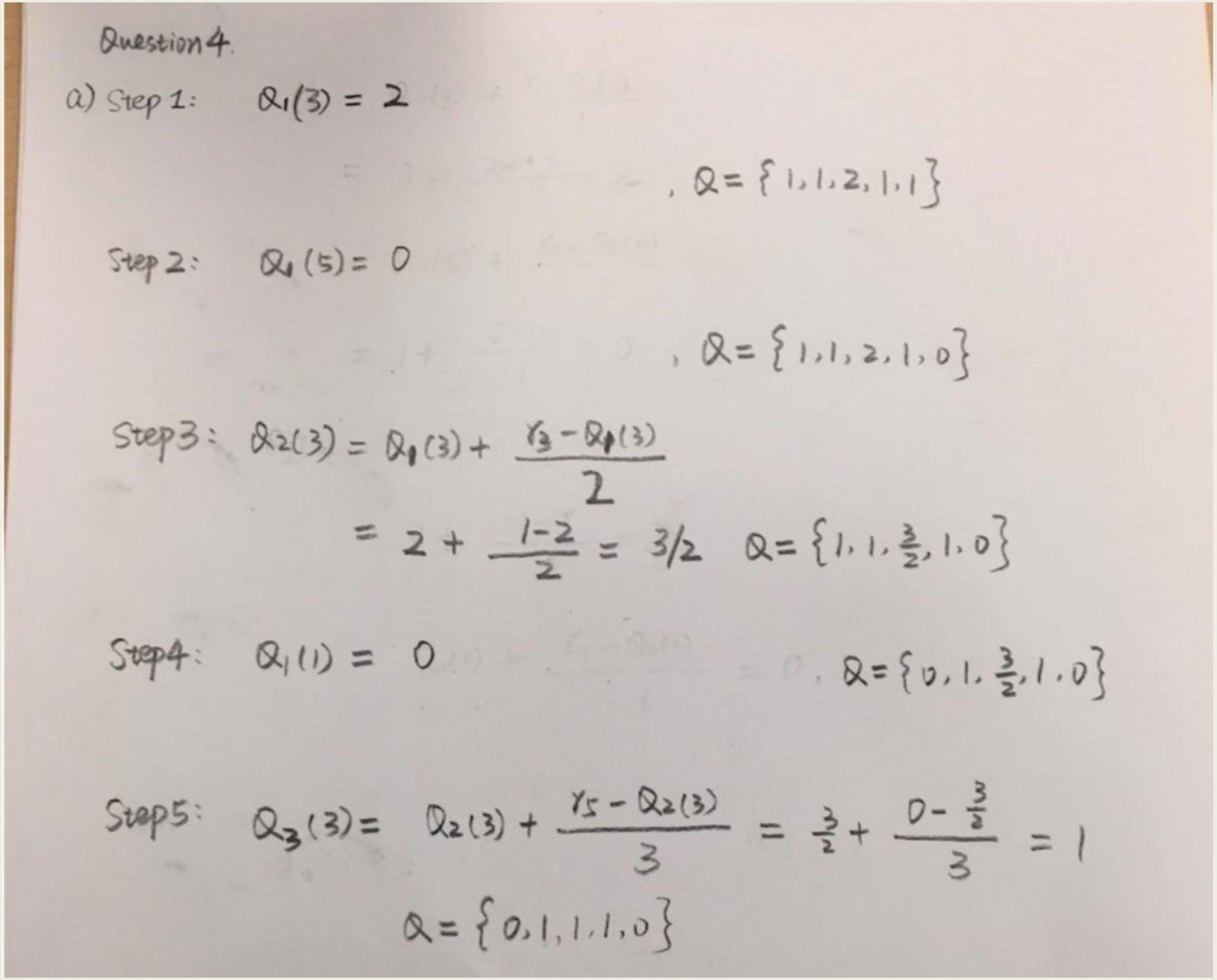


Question 4: Bandits

Consider the following 5-armed bandit problem. The initial value estimates of the arms are given by Q = {1, 1, 3, 1, 1}, and the actions are represented by A={1, 2, 3, 4, 5}. We observe a trajectory consisting of plays and rewards: T={A_1=3, R_1=2, A_2=5, R_2=0, A_3=3, R_3=1, A_4=1, R_4=0, A_5=3, R_5=0}.

a) Show the estimated Q values at each time step of the trajectory using the average of the observed rewards, where available. Do not consider the initial estimates as samples.

b) It turns out the player was following an epsilon-greedy strategy. For each time step, report whether it can be concluded with certainty that a random action was selected.



Question 4.

a) Step 1: $Q_1(3) = 2$

$, Q = \{1,1,2,1,1\}$

Step 2: $Q_1(5) = 0$

$, Q = \{1,1,2,1,0\}$

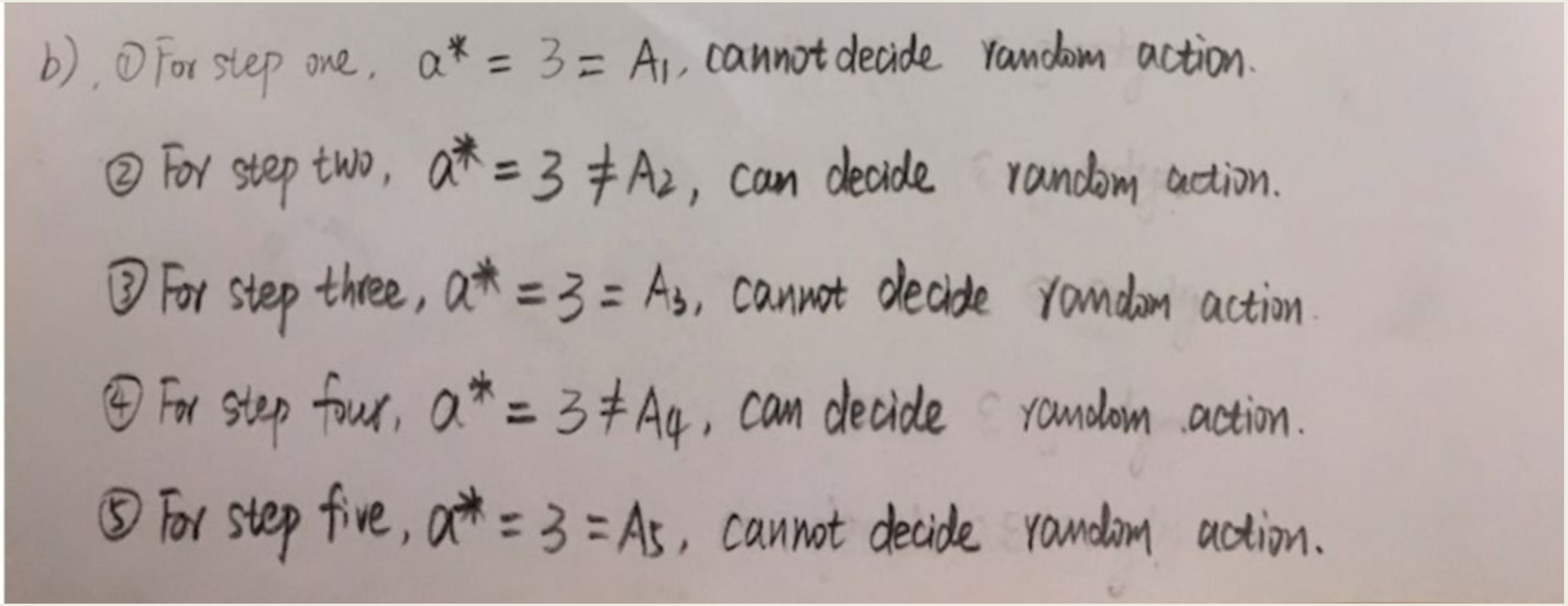
Step 3: $Q_2(3) = Q_1(3) + \dfrac{r_3 - Q_1(3)}{2}$

$= 2 + \dfrac{1-2}{2} = 3/2 \quad Q = \{1,1,\tfrac{3}{2},1,0\}$

Step 4: $Q_1(1) = 0$ $, Q = \{0,1,\tfrac{3}{2},1,0\}$

Step 5: $Q_3(3) = Q_2(3) + \dfrac{r_5 - Q_2(3)}{3} = \tfrac{3}{2} + \dfrac{0 - \frac{3}{2}}{3} = 1$

$Q = \{0,1,1,1,0\}$



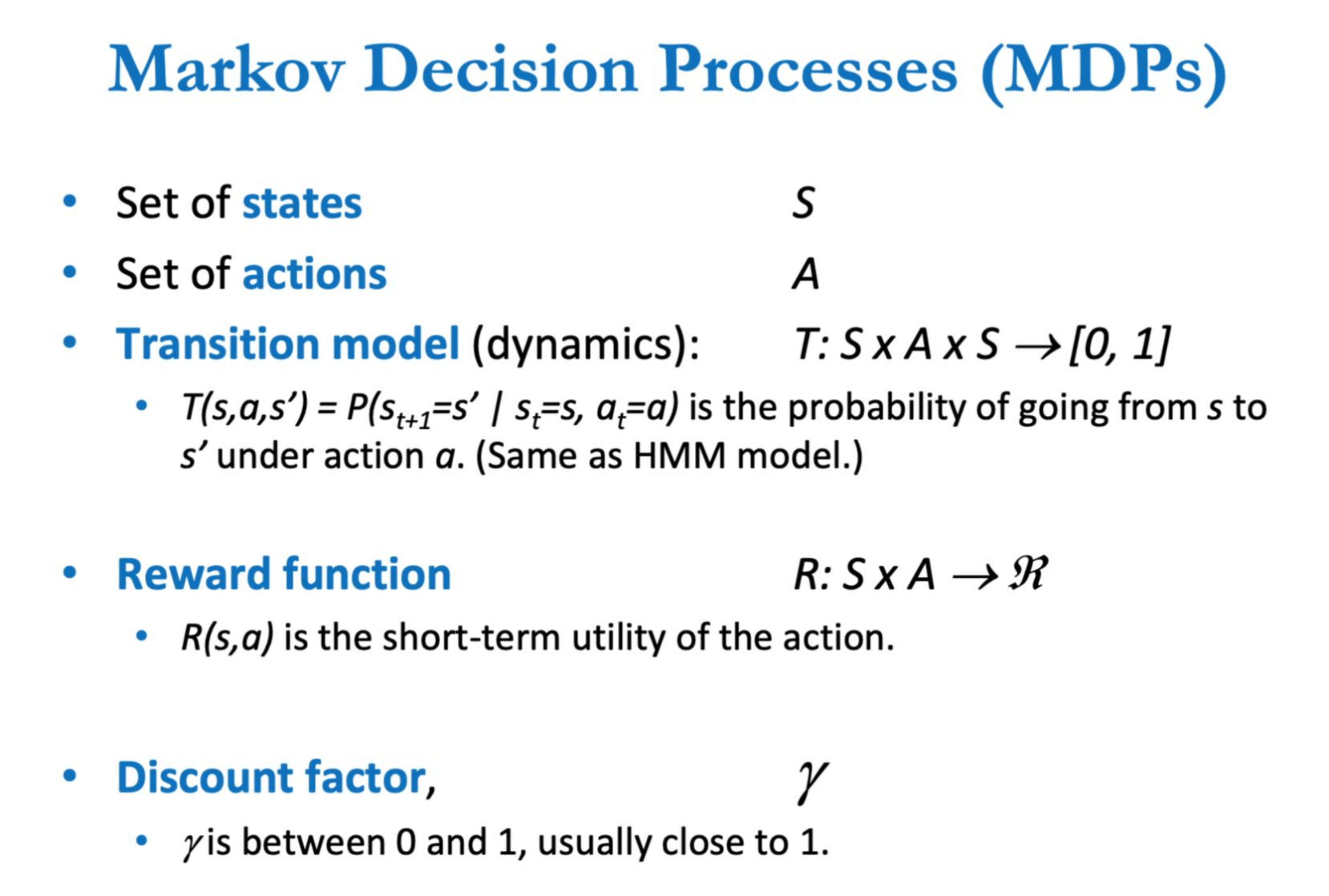
b). ① For step one. $a^* = 3 = A_1$, cannot decide random action.

② For step two, $a^* = 3 \neq A_2$, can decide random action.

③ For step three, $a^* = 3 = A_3$, cannot decide random action.

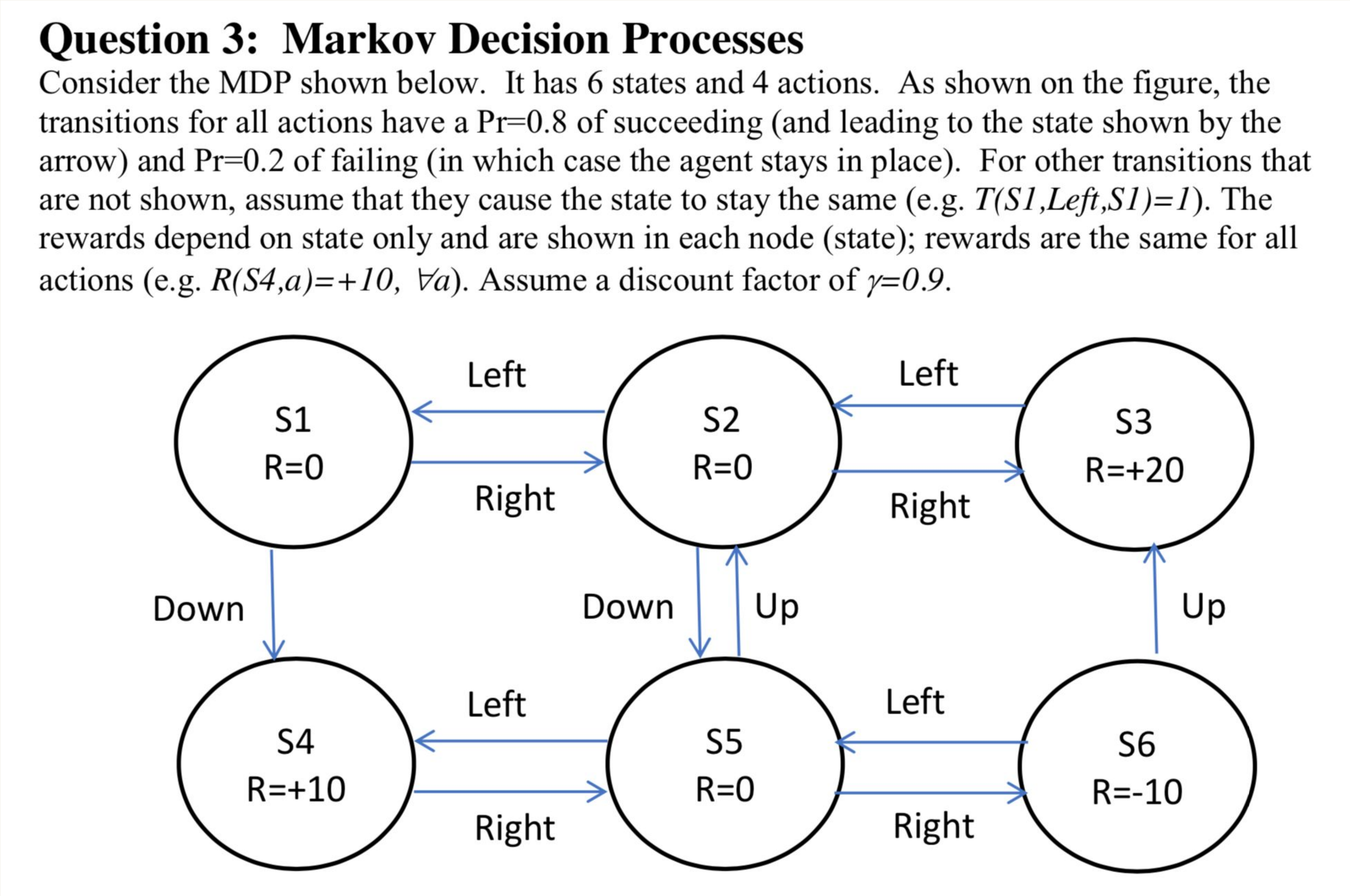
④ For step four, $a^* = 3 \neq A_4$, can decide random action.

⑤ For step five, $a^* = 3 = A_5$, cannot decide random action.



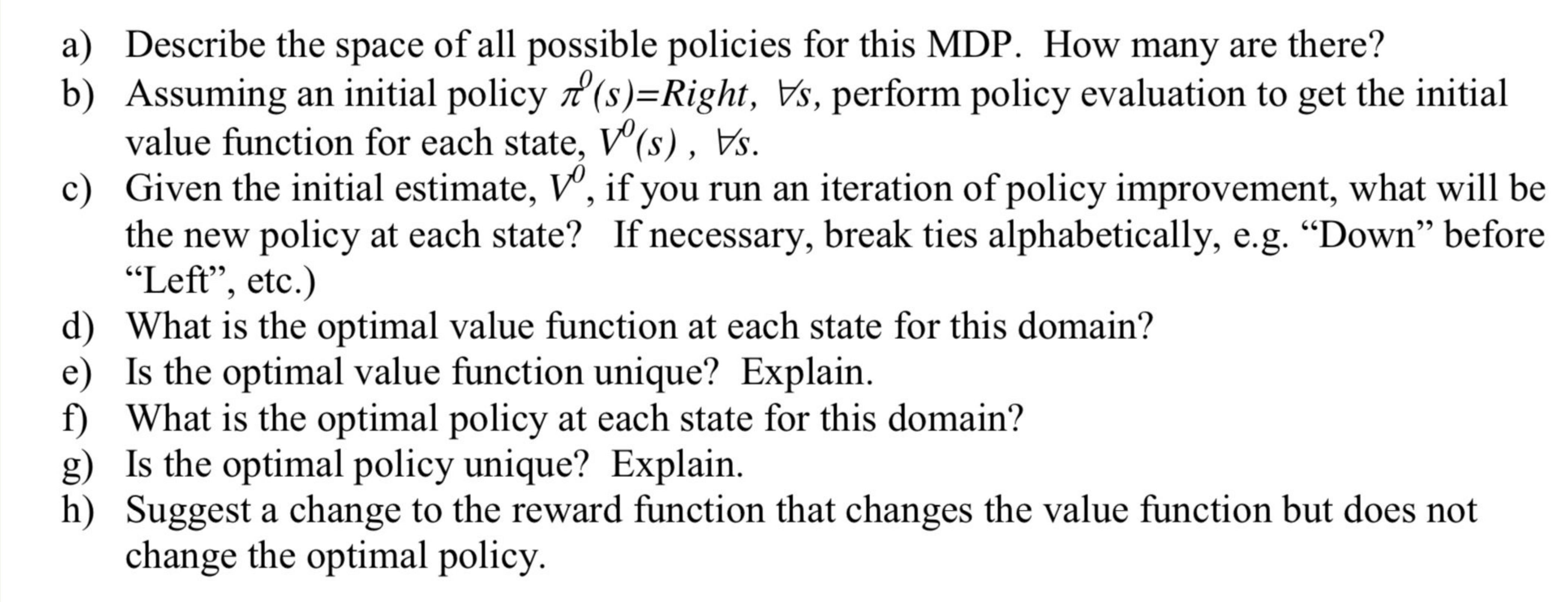
Markov Decision Processes (MDPs)

- Set of **states** S
- Set of **actions** A
- **Transition model** (dynamics): $T: S \text{ x } A \text{ x } S \rightarrow [0, 1]$
  - $T(s,a,s') = P(s_{t+1}=s' \mid s_t=s, a_t=a)$ is the probability of going from *s* to *s'* under action *a*. (Same as HMM model.)
- **Reward function** $R: S \text{ x } A \rightarrow \Re$
  - *R(s,a)* is the short-term utility of the action.
- **Discount factor**, γ
  - γ is between 0 and 1, usually close to 1.



Question 3: Markov Decision Processes

Consider the MDP shown below. It has 6 states and 4 actions. As shown on the figure, the transitions for all actions have a Pr=0.8 of succeeding (and leading to the state shown by the arrow) and Pr=0.2 of failing (in which case the agent stays in place). For other transitions that are not shown, assume that they cause the state to stay the same (e.g. *T(S1,Left,S1)=1*). The rewards depend on state only and are shown in each node (state); rewards are the same for all actions (e.g. *R(S4,a)=+10, ∀a*). Assume a discount factor of $\gamma=0.9$.



a) Describe the space of all possible policies for this MDP. How many are there?

b) Assuming an initial policy $\pi^0(s)=Right,\ \forall s$, perform policy evaluation to get the initial value function for each state, $V^0(s)\ ,\ \forall s$.

c) Given the initial estimate, V^0, if you run an iteration of policy improvement, what will be the new policy at each state? If necessary, break ties alphabetically, e.g. “Down” before “Left”, etc.)

d) What is the optimal value function at each state for this domain?

e) Is the optimal value function unique? Explain.

f) What is the optimal policy at each state for this domain?

g) Is the optimal policy unique? Explain.

h) Suggest a change to the reward function that changes the value function but does not change the optimal policy.