CS 4100 5100
Problem Set 2

Instructions

- All submissions must be typed. No exceptions are made to this rule.
- Hand-drawn figures are acceptable only where specified in the question, and provided all labels are clear and legible. If we can't read your text, we can't assign points. Please scan and insert any such figures in the final PDF document.

Academic Integrity

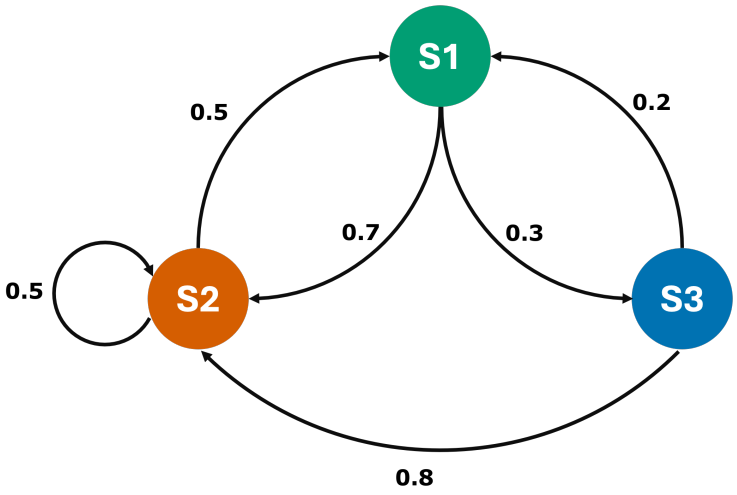
- If you discuss concepts related to this programming assignment with one or more classmates, all parties must declare collaborators in their individual submissions. Such discussions must be kept at a conceptual level, and no sharing of written answers is permitted. Do not discuss the specific problems within this problem set.
- You may not, as a general rule, use generative AI for problem sets. Any use of GenAI that is solely intended to improve writing clarity or sentence structure is acceptable, but must be accompanied by an appendix containing your original, unedited answers. If you use generative AI in this manner, start a new page titled 'Appendix' at the end of the written submission, and paste your original answers here with the corresponding question numbers clearly indicated.
- Failure to disclose collaborations or use of generative AI will be treated as a violation of academic integrity, and penalties listed in the course syllabus will be enforced.

Reach Out!

If at any point you feel stuck with the assignment, please reach out to the TAs or the instructor, and do so early on! This lets us guide you in the right direction in a timely fashion and will help you make the most of your assignment.

Markov Models

Consider the following Markov model with 3 states, **S1**, **S2** and **S3**. The transition probabilities of the model are represented along the edges (note that each edge is directional).



Q1. Given the above Markov model, explain whether the stationary distribution depends on the start state (without actually computing it). (2)

Solution: The chain is *irreducible* (every state communicates with every other) and *aperiodic* (self-loop at S_2). Hence a unique stationary distribution π exists and it is *independent* of the initial state.

Q2. Given the initial probability distribution $p_0 = [0.5,\ 0.1,\ 0.4]$ for states $\mathbf{S1}, \mathbf{S2}, \mathbf{S3}$, compute the stationary distribution π for the Markov model. (2)

Solution:

$$P = \begin{pmatrix} 0 & 0.7 & 0.3 \\ 0.5 & 0.5 & 0 \\ 0.2 & 0.8 & 0 \end{pmatrix}, \qquad \pi = [\pi_1, \pi_2, \pi_3], \qquad \pi P = \pi,\ \sum_{i=1}^{3} \pi_i = 1.$$

From $\pi P = \pi$ we obtain

$$\pi_3 = 0.3\,\pi_1, \qquad 0.94\,\pi_1 = 0.5\,\pi_2 \implies \pi_2 = 1.88\,\pi_1.$$

Normalising, $\pi_1 + \pi_2 + \pi_3 = \pi_1 + 1.88\pi_1 + 0.3\pi_1 = 3.18\pi_1 = 1 \implies \pi_1 = \frac{50}{159}$. Therefore

$$\boxed{\pi = \left[\tfrac{50}{159},\ \tfrac{94}{159},\ \tfrac{15}{159}\right]} \approx [0.3145,\ 0.5912,\ 0.0943].$$

Q3. Probability of the transition sequence $\mathbf{S1} \to \mathbf{S2} \to \mathbf{S2} \to \mathbf{S1}$ (assume the chain is in stationarity). (2)

Solution:

$$\Pr(S_1 \to S_2 \to S_2 \to S_1) = \pi_1 \, P_{1,2} \, P_{2,2} \, P_{2,1} = \frac{50}{159} \times 0.7 \times 0.5 \times 0.5 = \frac{35}{636} \approx 0.0550.$$

Q4. Starting in state $\mathbf{S2}$, what is the probability of being back in $\mathbf{S2}$ after *two* transitions? (2)

Solution:

$$P = \begin{pmatrix} 0 & 0.7 & 0.3 \\ 0.5 & 0.5 & 0 \\ 0.2 & 0.8 & 0 \end{pmatrix}, \qquad P^2 = P \cdot P = \begin{pmatrix} 0.41 & 0.59 & 0.00 \\ 0.25 & 0.60 & 0.15 \\ 0.40 & 0.54 & 0.06 \end{pmatrix}.$$

The required probability is the $(2, 2)$ entry of P^2:

$$\boxed{P^{(2)}_{2,2} = 0.60}.$$

Q5. With starting distribution $p_0 = [0.5,\, 0.1,\, 0.4]$, find the probability of being in $\mathbf{S2}$ after exactly two transitions. (2)

Solution:

$$p_2 = p_0\, P^2 = \begin{bmatrix}0.5,\, 0.1,\, 0.4\end{bmatrix} \begin{pmatrix} 0.41 & 0.59 & 0.00 \\ 0.25 & 0.60 & 0.15 \\ 0.40 & 0.54 & 0.06 \end{pmatrix} = \begin{bmatrix}0.39,\ 0.571,\ 0.039\end{bmatrix}.$$

Hence

$$\boxed{\Pr\{S_2 \text{ after two steps}\} = 0.571}.$$

Hidden Markov Models

Recall Hidden Markov Models (HMM) from class, where we applied this approach to sequence labeling tasks such as parts-of-speech tagging or named entity recognition. Here, your task is to construct and use an HMM model to make inferences about a coin-flipping game with the following rules.

Your professor produces two identical looking coins. However, only one of the coins is a fair coin, and the other is a biased coin that produces an outcome of **Heads** 70% of the time. The professor always knows which coin is the fair one, and will perform three coin flips in total. Between each flip, the professor may swap the coin, following the rule that if a fair coin is flipped in one round, then in the next round, the professor chooses a coin completely at random. If, however, the biased coin is flipped in any round, then the professor is two times as likely to choose the biased coin again in the next round, as compared to the fair coin. As the three flips are performed, you observe the outcomes **Heads**, **Tails** and **Tails** respectively.

Q6. Draw the HMM diagram for this game showing transitions between hidden states, and emissions to outcomes with the corresponding probabilities labeled along the edges. Hand-drawn figures accepted for this question, provided the grader can read everything clearly. (For reference, see the second diagram in our HMM notes, showing the *Very Late*, *Late* and *On Time* hidden states, and the *Happy* and *Sad* outcomes.) (5)

Solution (Q6):

- **Hidden states:** F (fair coin), B (biased coin).
- **Observations:** H (Heads), T (Tails).
- **Transition matrix** $A = \begin{pmatrix} 0.5 & 0.5 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$.
- **Emission probabilities**
  $B_F(H) = 0.5,\ B_F(T) = 0.5;\ B_B(H) = 0.7,\ B_B(T) = 0.3.$

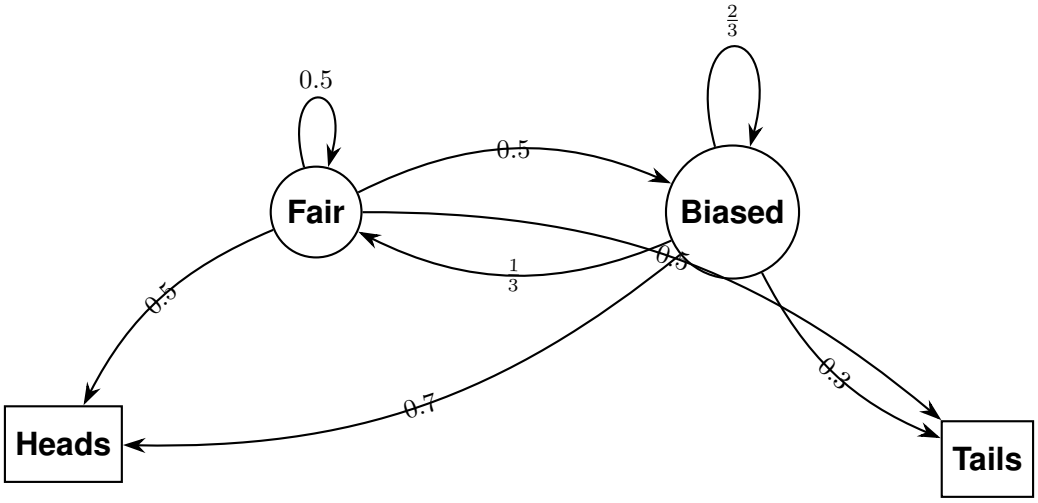


Figure 1: HMM for the coin-flipping game.

Q7. Given the observed outcome sequence, predict which coin was most likely flipped in each of the three turns (i.e., compute the most likely hidden sequence). Show all your intermediate calculations and use the stationary distribution to reason about which coin was flipped in the very first round. (10)

Solution (Q7):

HMM parameters.

$$A = \begin{pmatrix} 0.5 & 0.5 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}, \qquad B_F(H) = 0.5,\ B_F(T) = 0.5,\ B_B(H) = 0.7,\ B_B(T) = 0.3.$$

Stationary prior: $\pi_F = 0.4,\ \pi_B = 0.6$.

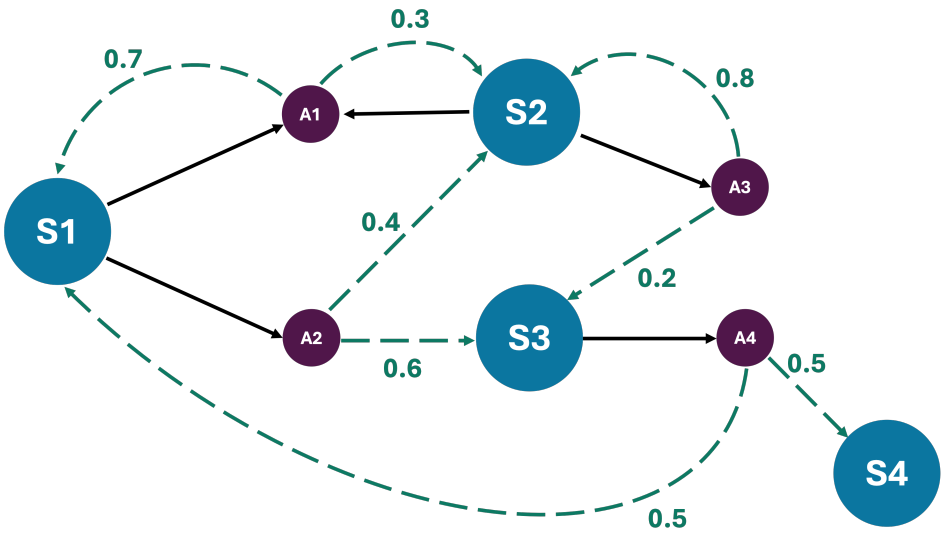
t	$\delta_t(F)$	$\delta_t(B)$	back-pointer ψ_t
1 (H)	$0.4 \times 0.5 = 0.20$	$0.6 \times 0.7 = 0.42$	–
2 (T)	$\max\{0.20 \times 0.5,\ 0.42 \times \frac{1}{3}\} \times 0.5 = 0.07$	$\max\{0.20 \times 0.5,\ 0.42 \times \frac{2}{3}\} \times 0.3 = 0.084$	both from B
3 (T)	$\max\{0.07 \times 0.5,\ 0.084 \times \frac{1}{3}\} \times 0.5 = 0.0175$	$\max\{0.07 \times 0.5,\ 0.084 \times \frac{2}{3}\} \times 0.3 = 0.0168$	$q_3^* = F$

Back-tracking:

$$q_3^* = F,\ q_2^* = \psi_3(F) = F,\ q_1^* = \psi_2(F) = B.$$

$$\boxed{(q_1^*, q_2^*, q_3^*) = (B,\ F,\ F)}$$

Interpretation: the biased coin was most likely used on the first flip, then the fair coin on flips two and three.



Markov Decision Processes and Reinforcement Learning

Consider the following MDP with 4 states, and 4 actions. A dashed line represents a transition from a chosen action to some next state. Transition probabilities are specified along each dashed edge.

Q8. How many unique policies does this MDP have? Explain your reasoning and list all policies. Use **X** to indicate no possible action from a state. (2)

1. Actions available at each state

The black arrows in the diagram show which actions *can be chosen* from each state:

$$\begin{aligned} S_1: &\quad \{A1,\ A2\} \\ S_2: &\quad \{A1,\ A3\} \\ S_3: &\quad \{A4\} \\ S_4: &\quad \text{terminal} \ \ (\mathbf{X}) \end{aligned}$$

2. Counting policies

A deterministic policy selects *one* action for every non-terminal state:

$$|A(S_1)| \times |A(S_2)| \times |A(S_3)| \ = \ 2 \times 2 \times 1 \ = \ 4.$$

3. Listing all 4 policies

Policy	S_1	S_2	S_3	S_4
π_1	$A1$	$A1$	$A4$	**X**
π_2	$A1$	$A3$	$A4$	**X**
π_3	$A2$	$A1$	$A4$	**X**
π_4	$A2$	$A3$	$A4$	**X**

Hence the MDP admits exactly 4 distinct deterministic policies.

Q9. If from any state, all valid actions are equally likely, then what is the total probability of reaching S4 from S1 using paths of at most length 3? List all such paths and compute the total probability. Show your calculations. (An action followed by a transition into a next state counts as a total of one move.) (4)

1. Action sets and transition kernels

State	Actions	Non-zero transitions $P(s' \mid s, a)$
S_1	$A1, A2$	$A1$: $S_1(0.7)$, $S_2(0.3)$
		$A2$: $S_3(0.6)$, $S_2(0.4)$
S_2	$A1, A3$	$A1$: $S_1(0.7)$, $S_2(0.3)$
		$A3$: $S_2(0.8)$, $S_3(0.2)$
S_3	$A4$	$A4$: $S_4(0.5)$, $S_1(0.5)$
S_4	**X** (terminal)	

Uniform action–choice probabilities: $P(A1 \mid S_1) = P(A2 \mid S_1) = \frac{1}{2},\ P(A1 \mid S_2) = P(A3 \mid S_2) = \frac{1}{2},\ P(A4 \mid S_3) = 1$.

Only $A4$ can *reach* S_4, so the last move must be $(S_3,\ A4)$. Hence we first need paths that arrive at S_3 in one or two moves from S_1.

2. Length-2 paths ($S_1 \to S_3 \to S_4$)

$$S_1 \xrightarrow{A2} S_3 \xrightarrow{A4} S_4 :\ \left(\tfrac{1}{2}\right)(0.6) \cdot 1 \cdot 0.5 = 0.15$$

3. Length-3 paths ($S_1 \to \text{—} \to S_3 \to S_4$)

$$S_1 \xrightarrow{A1} S_1 \xrightarrow{A2} S_3 \xrightarrow{A4} S_4 :\ (\tfrac{1}{2})(0.7)(\tfrac{1}{2})(0.6)(1)(0.5) = 0.0525$$

$$S_1 \xrightarrow{A1} S_2 \xrightarrow{A3} S_3 \xrightarrow{A4} S_4 :\ (\tfrac{1}{2})(0.3)(\tfrac{1}{2})(0.2)(1)(0.5) = 0.0075$$

$$S_1 \xrightarrow{A2} S_2 \xrightarrow{A3} S_3 \xrightarrow{A4} S_4 :\ (\tfrac{1}{2})(0.4)(\tfrac{1}{2})(0.2)(1)(0.5) = 0.0100$$

4. Total probability (length ≤ 3)

$$P(S_1 \rightsquigarrow S_4 \text{ in } \leq 3) = \underbrace{0}_{\text{length 1}} + \underbrace{0.15}_{\text{length 2}} + \underbrace{(0.0525 + 0.0075 + 0.0100)}_{\text{length 3}}$$

$$= \boxed{0.22}.$$

Q10. Given that $R(S1, A2, S3) = 5$, $R(S2, A3, S3) = 5$ and $R(S3, A4, S4) = 100$, and that rewards for all other transitions are 0, write and expand the optimal value function equation for $V_{opt}(S1)$. Assume that the discount factor is γ, and leave your final answer in terms of $V_{opt}(S1)$, $V_{opt}(S2)$ and $V_{opt}(S3)$. (4)

1. Bellman optimality (deterministic)

$$V_{\text{opt}}(s) = \max_{a \in A(s)} \sum_{s'} P(s' \mid s, a) \big[R(s, a, s') + \gamma \, V_{\text{opt}}(s') \big].$$

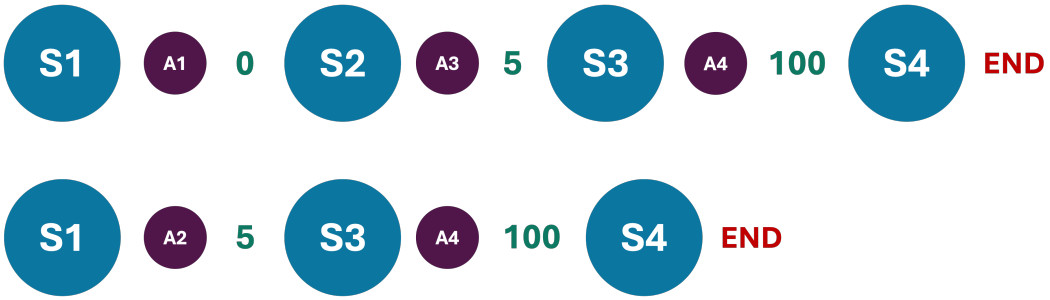
2. Optimal Q–values from S_1

$$\begin{aligned} Q_{\text{opt}}(S_1, A_1) &= 0.7(0 + \gamma V_{\text{opt}}(S_1)) + 0.3(0 + \gamma V_{\text{opt}}(S_2)) \\ &= 0.7\gamma \, V_{\text{opt}}(S_1) + 0.3\gamma \, V_{\text{opt}}(S_2), \\ Q_{\text{opt}}(S_1, A_2) &= 0.4(0 + \gamma V_{\text{opt}}(S_2)) + 0.6(5 + \gamma V_{\text{opt}}(S_3)) \\ &= 3 + 0.4\gamma \, V_{\text{opt}}(S_2) + 0.6\gamma \, V_{\text{opt}}(S_3). \end{aligned}$$

3. Bellman equation for $V_{\text{opt}}(S_1)$

$$\boxed{V_{\text{opt}}(S_1) = \max\Big\{ 0.7\gamma \, V_{\text{opt}}(S_1) + 0.3\gamma \, V_{\text{opt}}(S_2), \; 3 + 0.4\gamma \, V_{\text{opt}}(S_2) + 0.6\gamma \, V_{\text{opt}}(S_3) \Big\}.}$$

Q11. Assume that by simulating this MDP using some exploration policy π, we obtain the two following episodes:



Use Q-learning updates to calculate the agent's final optimal policy given this data stream, and show all intermediate steps. Assume $\gamma = 1$. For your reference, the Q-learning update equation is given by: (10)

$$\eta_{s,a} = \frac{1}{1 + \text{number of updates to } \hat{Q}_{opt}(s,a)}$$

For each observed (s, a, r, s'):

$$\text{Estimate, } \hat{Q}_{opt}^{(new)}(s,a) = (1-\eta_{s,a})\hat{Q}_{opt}^{(old)}(s,a) + \eta_{s,a}[R(s,a,s') + \gamma \hat{V}_{opt}^{(old)}(s')]$$

$$\text{where } \hat{V}_{opt}(s') = \max_{a'} \hat{Q}_{opt}(s',a')$$

$$\underbrace{S_1 \xrightarrow{A_1, r=0} S_2 \xrightarrow{A_3, r=5} S_3 \xrightarrow{A_4, r=100} S_4}_{\text{Episode 1}} \qquad \underbrace{S_1 \xrightarrow{A_2, r=5} S_3 \xrightarrow{A_4, r=100} S_4}_{\text{Episode 2}}$$

Apply one–step *Q-learning* with $\eta_{s,a} = \dfrac{1}{1+N(s,a)}$, $\gamma = 1$, $Q^{\text{new}} = (1-\eta)\, Q^{\text{old}} + \eta\big[r + \gamma \max_{a'} Q^{\text{old}}(s',a')\big]$.

1. Initial tables

$$Q(s,a) = 0 \;\; \forall (s,a), \qquad N(s,a) = 0, \qquad V(s) = \max_a Q(s,a) = 0.$$

The only actions that ever appear in the data are A_1, A_2 at S_1, A_3 at S_2, A_4 at S_3.

2. Episode 1 updates

1. $(S_1, A_1, 0, S_2)$: $\eta = 1$, $V(S_2) = 0$

$$Q(S_1, A_1) = 0, \qquad N(S_1, A_1) = 1.$$

2. $(S_2, A_3, 5, S_3)$: $\eta = 1,\ V(S_3) = 0$

$$Q(S_2, A_3) = 5, \qquad N(S_2, A_3) = 1.$$

3. $(S_3, A_4, 100, S_4)$: $\eta = 1,\ V(S_4) = 0$

$$Q(S_3, A_4) = 100, \qquad N(S_3, A_4) = 1.$$

After episode 1

$$Q(S_1, A_1) = 0,\ Q(S_2, A_3) = 5,\ Q(S_3, A_4) = 100,\ Q(S_1, A_2) = 0.$$

3. Episode 2 updates

1. $(S_1, A_2, 5, S_3)$:
   $V(S_3) = \max_{a'} Q(S_3, a') = 100,\ \eta = 1$

$$Q(S_1, A_2) = 5 + 100 = 105, \qquad N(S_1, A_2) = 1.$$

2. $(S_3, A_4, 100, S_4)$:
   This pair has already been updated once, so $\eta = \frac{1}{1+N(S_3,A_4)} = \frac{1}{2}$. $V(S_4) = 0$.

$$Q(S_3, A_4) = (1 - \tfrac{1}{2})100 + \tfrac{1}{2}\big[100 + 0\big] = 100,$$
$$N(S_3, A_4) = 2.$$

4. Final Q-values

$$Q(S_1, A_1) = 0, \qquad Q(S_1, A_2) = 105, \qquad Q(S_2, A_3) = 5, \qquad Q(S_3, A_4) = 100.$$

(All other pairs remain 0 because they were never visited.)

5. Greedy policy extracted from $\hat{Q}$

$$\boxed{\pi^\star(S_1) = A_2, \quad \pi^\star(S_2) = A_3, \quad \pi^\star(S_3) = A_4, \quad \pi^\star(S_4) = \textbf{X}\ \text{(terminal)}.}$$

Thus, after processing the two episodes, the agent's optimal (greedy) policy is $S_1 \to A_2,\ S_2 \to A_3,\ S_3 \to A_4$.