1. (8 points) **True or False**

   Circle the correct answer for each T/F question. No need to explain the reasoning.

   (a) (1 point) True / False - In reinforcement learning we assume the agent knows the transition model $T(s, a, s')$ and the reward function $R(s, a, s')$.

   (b) (1 point) True / False - Temporal Difference (TD) learning is a form of model-free RL that updates values based on observed transitions without needing a model.

   (c) (1 point) True / False - Two variables X and Y are independent if $P(X, Y) = P(X)P(Y)$.

   (d) (1 point) True / False - In a Bayesian network, conditional independence given parents does not reduce the number of parameters needed in the probability tables.

   (e) (1 point) True / False - Inference by enumeration can produce incorrect results if the Bayes network is dense (has many edges).

   (f) (1 point) True / False - Given no independence assumptions, $P(A \mid B, C) = \frac{P(B|A,C)P(A|C)}{P(B|C)}$.

   (g) (1 point) True / False - The number of parameters in a Bayesian network grows exponentially with the highest in-degree (number of parents) of a node in the network.

   (h) (1 point) True / False - The Markov assumption requires that X_{t+1} depends only on X_t and X_{t-1}, but no earlier states.

2. (10 points) **Short Answer** These short answer questions can be answered with a few sentences each. Be short and precise. No need to expand into too many details.

   (a) (2 points) Briefly describe the difference between model-free reinforcement learning and model-based reinforcement learning.

   (b) (2 points) Briefly describe a situation in which you would use Bayes rule, and why, from the examples we saw in class.

   (c) (2 points) Define the belief state $B_t(X)$ in an HMM. Conceptually, how is $B_t(X)$ updated when you move from time $t-1$ to t and receive a new observation?

   (d) (2 points) Briefly describe the idea behind particle filtering and one advantage it has over exact inference.

   (e) (2 points) In machine learning, explain generalization and over-fitting. Describe an experimental setup that correctly measures generalization. Assume that your algorithm has one hyperparameter that must be set.

3. (12 points) **Markov Chain**

   (a) (3 points) In Markov model, we compute the probability distribution over $X_1, \ldots X_n$ as follows:

   $$P(X_1, \ldots X_n) = P(X_1) \prod_t P(X_t | X_{t-1})$$

   Are there any assumptions required for the formulation above? If so, discuss the assumptions.

   Now consider the following process: You can be employed or unemployed. At each time step, you have a 5% chance of losing your job and a 60% chance of moving from unemployment to employment (finding a new job!). Unless otherwise specified, assume that you have a 50% chance of being employed at time 0.

   (b) (3 points) Model this employment process as a Markov chain.

   (c) (2 points) If you start out employed, what is the probability of being employed after step 2? Show all of your work.

   (d) (2 points) If you are unemployed after time 2, what is the probability that you started out employed at time 0?

   (e) (2 points) What is the stationary distribution of the chain you defined?

4. (7 points) **Hidden Markov Models: Tricky Coins**

   Consider the following random process. A magician has two coins, each of which has an unknown type. They can either be fair coins (50/50 odds of heads vs. tails), or trick coins that either (1) have heads on both sides or (2) have tails on both sides. A priori, each coin is equally likely to be any of the three possible types.

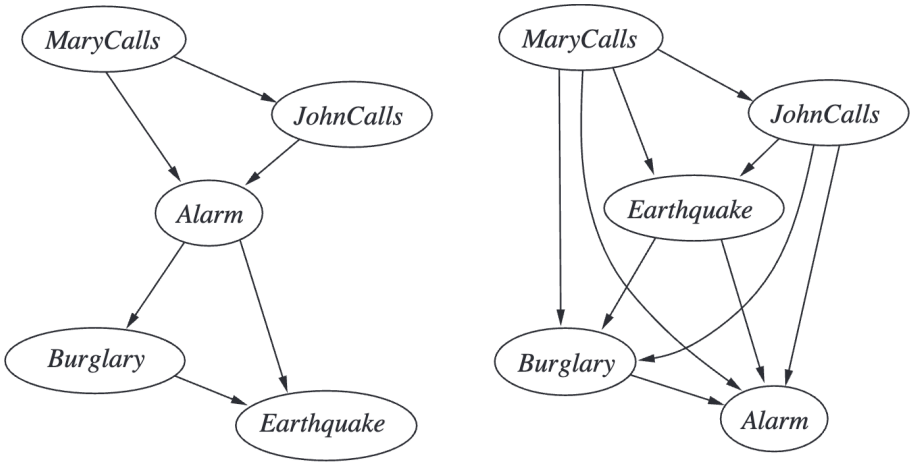
   At every time step, the magician randomly picks a coin (without showing you which one was selected), flips it, and shows you the result. However, unfortunately, the magician only shows you the coin very briefly, and 10% of the time you make a mistake when you observe the true side of the coin (e.g., you see heads when it was actually tails).

   (a) (5 points) Model this process as an HMM. Specify all of the necessary parameters. You do not have to write out all of the probability distributions explicitly, but be careful to specify what values they would have if you did the full enumeration. What conditional independences hold in this HMM?

   (b) (2 points) Consider the Markov model that would result if you ran the process above and always observed heads. What is the stationary distribution of this model?

5. (10 points) **Bayesian Networks**

Consider the following two Bayesian networks, which are variations on the alarm network we discussed in class:



(a) (2 points) Based on the network structure alone, which network above makes the most independence assumptions?

(b) (2 points) Draw a new Bayesian network with the same set of random variables that makes as many independence assumptions as possible.

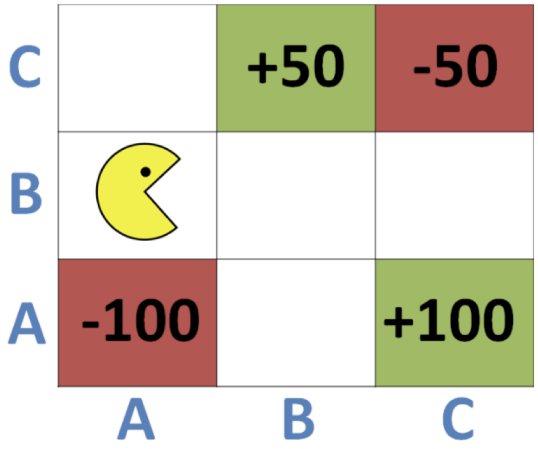
(c) (2 points) Write down two conditional independence assumptions encoded by the structure of network (a). If there are fewer than two, write all of them.

(d) (2 points) Write down two conditional independence assumptions encoded by the structure of network (b). If there are fewer than two, write all of them.

(e) (2 points) If the edge between MaryCalls and Earthquake is removed from network (b), will the class of joint probability distributions that can be represented by the resulting Bayesian network be smaller or larger than that associated with the original network? Briefly explain your answer.

6. (10 points) **Reinforcement Learning**

Consider the grid-world given below and an agent who is trying to learn the optimal policy. States are named as (x-coordinate, y-coordinate) with horizontal axis x and vertical axis y , and the state after exiting is Done. Actions are North, South, East, West, and Exit denoted as N, S, E, W, and X for short. The Exit action can only be taken from shaded states, and Exit is the only action available in the shaded states. Rewards are only awarded for taking the Exit action from one of the shaded states. Taking this action moves the agent to the Done state, and the MDP terminates. Assume $\gamma = 1$ and $\alpha = 0.5$ for all calculations. In Q-Learning, all values are initialized to zero.



Now, assume the agent starts from (A, B) and observes the following sequence of episodes. Each step is a tuple containing (s, a, s', r).

	Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
Step 1	(A,B), N, (A,C), 0	(A,B), E, (B,B), 0	(A,B), E, (B,B), 0	(A,B), E, (B,B), 0	(A,B), E, (B,B), 0
Step 2	(A,C), S, (A,B), 0	(B,B), E, (C,B), 0	(B,B), S, (B,A), 0	(B,B), E, (C,B), 0	(B,B), S, (B,A), 0
Step 3	(A,B), N, (A,C), 0	(C,B), N, (C,C), 0	(B,A), E, (C,A), 0	(C,B), W, (B,B), 0	(B,A), N, (B,B), 0
Step 4	(A,C), E, (B,C), 0	(C,C), X, Done, -50	(C,A), X, Done, +100	(B,B), N, (B,C), 0	(B,B), N, (B,C), 0
Step 5	(B,C), X, Done, +50			(B,C), X, Done, +50	(B,C), X, Done, +50

(a) (4 points) Fill in the following Q-values obtained from direct evaluation from the samples.

$Q((A, B), N) =$

$Q((B, B), E) =$

(b) (3 points) Which Q values are non-zero after running q-learning with the episodes above?

(c) (3 points) Assuming Q-learning continues for many more episodes and eventually converges, what is the optimal action at each of the three shaded states: (A,C), (B,C), and (C,B)?

7. (8 points) **Perceptrons**

We would like to use a perceptron to train a classifier for datasets with 2 features per point and labels +1 or -1 . Consider the following labeled training data:

Features (x_1, x_2)	Label y^*
$(-1, 2)$	1
$(3, -1)$	-1
$(1, 2)$	-1
$(3, 1)$	1

(a) (3 points) Our two perceptron weights have been initialized to $w_1 = 2$ and $w_2 = -2$. After processing the first point with the perceptron algorithm, what will be the updated values for these weights?

(b) (5 points) After how many steps will the perceptron algorithm converge? Write "never" if it will never converge. Note: one steps means processing one point. Points are processed in order and then repeated, until convergence.