TTIC 31180 — Probabilistic Graphical Models

Midterm

May 18, 2020 (Due: 11:59pm May 20, 2020)

Name: _____

1. _____ / 15
2. _____ / 26
3. _____ / 15
4. _____ / 20

Total: _____ / 76

The exam is due by 11:59pm Wednesday May 20th

The exam is open-book and open-notes, however you are not allowed to discuss the exam (or collaborate) with anyone else

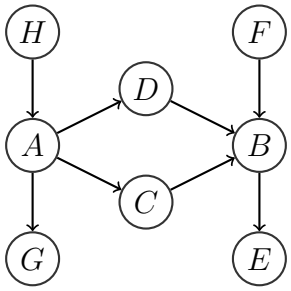
You are free to write your solutions directly on the exam booklet (there are extra pages at the end if you need more space) or to typeset them. Either way, make sure that you show your work and that your answers are clear

1 Independencies [15pts]

In this question, you will investigate independence relationships conveyed by graphical models.

a) Bayesian Network Independencies [5pts]

Determine whether the following independence statements are valid under the Bayesian network shown below. For those that are false, provide an active trail through the graph.



(i) $(H \perp F)$

(ii) $(H \perp D \mid A, G)$

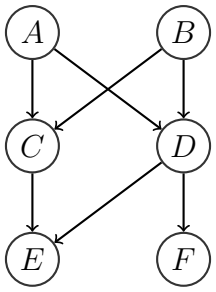
(iii) $(C \perp D \mid A)$

(iv) $(C \perp D \mid A, B)$

(v) $(D \perp F \mid E)$

b) Directed and Undirected Graph Independencies [4pts]

Answer the following questions with regards to the Bayesian network shown below.

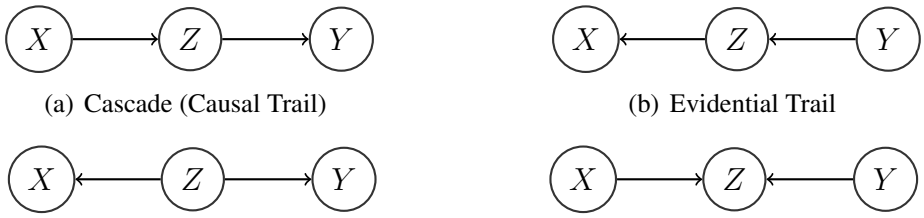


(i) **[2pts]** Assume that the directed graph is a perfect-map for the corresponding distribution. Find the minimal undirected I-map for this distribution.

(ii) **[2pts]** Compare the original directed graph and the undirected graph that you identified in part (i). Are there any independencies that are expressed by the original directed graph, but not the undirected graph? If so, what are they?

c) Markov Blanket [6pts]

The Markov blanket for a node in a Bayesian Network is the set of variables formed by the node's children, parents, and co-parents (i.e., other parents of its children). Recall the four different types of two-edge trails in a Bayesian network provided below:



(a) Cascade (Causal Trail) (b) Evidential Trail

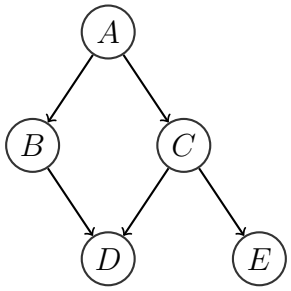
(c) Common Cause (d) Common Effect (v-structure)

Figure 1: The four different types of two-edge trails in Bayesian networks.

Prove informally that *a node in a Bayesian network is conditionally independent of every other set of nodes given its Markov blanket*. Hint: Use the relationship between dependencies and active trails and consider each of the four two-edge trails in Figure 1.

2 Variable Elimination [26pts]

Consider the following Bayesian network.



(i) **[1pts]** Write down a factored expression for the joint distribution.

(ii) **[2pts]** Assuming that each variable can take one of k different values, how many parameters are necessary to specify the joint distribution according to this factorization? How many parameters would be needed if we had no knowledge of independencies?

(iii) **[4pts]** Draw the corresponding MRF associated with this Bayesian network, and specify the clique potentials in terms of the original conditional probability distributions.

(iv) **[4pts]** Draw a clique tree for this MRF where you (i) make explicit the scope of each node (e.g., as ellipsoids) as well as the scope of the sepset (e.g., as boxes); (ii) number each node (not sepset); and (iii) indicate which clique potentials determined above are associated with each node. It's sufficient express these in the form of $\phi(X, Y.Z)$, where X, Y, Z denote the scope (i.e., don't worry about writing the CPDs).

(v) **[8pts]** Suppose that we want to compute $P(D)$. Choosing one node as the root, go through the process of clique tree message passing, indicating which factor(s) are involved at each step (including their scope) and which messages (factors) are created (including their scope). Number these messages in a way that indicates which nodes are involved and the direction in which the messages are sent. Please make these factors and their scope easy to identify (e.g., underline or box them).

(vi) **[1pts]** What is the corresponding variable elimination ordering $\prec$?

(vii) **[2pts]** What is the computational complexity associated with this ordering assuming that each random variable can take on k different values?

(viii) **[2pts]** Suppose that we are now interested in $P(D \mid E)$. What additional messages (if any) are necessary to compute this conditional distribution?

(ix) **[2pts]** Suppose that we used the ordering $\prec = C, A, B, D, E$ (i.e., starting with C). What is the computational complexity associated with this ordering, again assuming that each random variable can take on one of k different values?

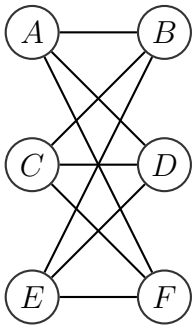
3 Chordal Graphs and Clique Trees [15pts]

This problem considers chordal graphs and clique trees.

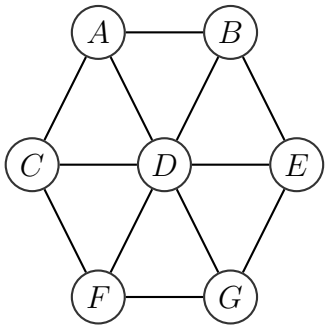
a) Chordal Graphs [3pts]

For each of the graphs below, state whether the graph is chordal. If not, identify a chordless cycle and triangulate the graph, attempting to add as few edges as possible.

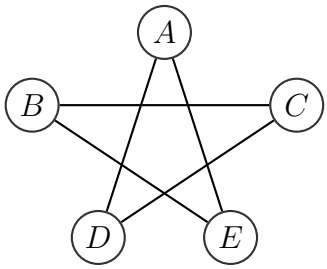
(i) **[1pts]** Bipartite Graph



(ii) **[1pts]** Hexagonal Graph



(iii) **[1pts]** Five-Star Graph



b) Clique Trees and Conditional Probabilities [8pts]

Suppose that we have a clique tree $\mathcal{T}$ for an undirected graph H with initial factors Φ. Let $\boldsymbol{C}_r$ be the root in the tree. Consider some node (clique) $\boldsymbol{C}_j$ and denote its upstream neighbor (i.e., closer to the root) as $\boldsymbol{C}_i$. Let β_j be the potential at $\boldsymbol{C}_j$ after performing the upward pass of sum-product message passing from the leaves up to the root.

Show that β_j encodes the unnormalized conditional distribution $\tilde{P}_\Phi(\boldsymbol{C}_j - \boldsymbol{S}_{ij} \mid \boldsymbol{S}_{ij})$, i.e., letting $\boldsymbol{X} = \boldsymbol{C}_j - \boldsymbol{S}_{ij}$, show that

$$\tilde{P}_\Phi(\boldsymbol{X} \mid \boldsymbol{S}_{ij}) = \frac{\tilde{P}_\Phi(\boldsymbol{X}, \boldsymbol{S}_{ij})}{\tilde{P}_\Phi(\boldsymbol{S}_{ij})} = \cdots = \frac{\beta_j(\boldsymbol{X}, \boldsymbol{S}_{ij})}{\mu_{ij}(\boldsymbol{S}_{ij})}$$

where $\mu_{ij}(\boldsymbol{S}_{ij})$ is the sepset belief determined from β_j. Hint: Write $\beta_j(\boldsymbol{X}, \boldsymbol{S}_{ij})$ and $\mu_{ij}(\boldsymbol{S}_{ij})$ in terms of the initial factors Φ.

c) Fixed Point of Clique Tree Calibration [4pts]

Consider a clique tree $\mathcal{T}$ with cliques $\boldsymbol{C}_i$ for $i \in \{1, 2, \ldots, n\}$. Prove that the clique beliefs $\beta_i(\boldsymbol{C}_i) = P(\boldsymbol{C}_i)$ and sepset beliefs $\mu_{ij} = P(\boldsymbol{S}_{ij})$ correspond to a fixed point of the belief propagation algorithm for this clique tree. In other words, show that if we start belief propagation with these beliefs, they won't be changed by any messages.

4 EM and KL Divergence [20pts]

Consider a domain that involves a set of random variables $\mathcal{X} = \{\boldsymbol{X}, y\}$, where $\boldsymbol{X} \in \mathbb{R}^n$ and y are is a multivariate and scalar random variable, respectively.

Suppose that you have access to a set of training data $\mathcal{D} = \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(m)}\}$, where $\boldsymbol{X}^{(i)} \sim P(\boldsymbol{X})$ are I.I.D. samples drawn from the posterior over $\boldsymbol{X}$, but you don't have any samples of y. However, you'd like to estimate the joint distribution $P(\boldsymbol{X}, Y) = P(\boldsymbol{X} \mid Y)P(Y)$ (e.g., it may be that $P(\boldsymbol{X})$ is particularly complex, but simplifies when you condition on Y).

Without samples of Y, we can't estimate $P(\boldsymbol{X} \mid Y)$ directly from our dataset. The Expectation-Maximization (EM) algorithm provides an iterative solution to the problem of estimation with latent variables. Instead of estimating $P(\boldsymbol{X}, Y)$ directly, EM uses an approximate distribution $Q(\boldsymbol{X}, Y) = Q(y \mid \boldsymbol{X})Q(\boldsymbol{X})$, where $Q(\boldsymbol{X}) = \frac{1}{m}\sum_i \delta(\boldsymbol{X} - \boldsymbol{X}^{(i)})$, and we're free to choose $Q(Y \mid \boldsymbol{X})$.

Given an initial estimate for the desired distribution P_0, the EM algorithm alternates between the following steps:

1. **E-Step**: Update the conditional component of the surrogate distribution $Q(Y \mid \boldsymbol{X})$ by minimizing the following KL-divergence objective:

   $$Q_k = \arg\min_{Q(Y \mid \boldsymbol{X})} D(Q \| P_k)$$

   This amounts to choosing $Q_k(Y \mid \boldsymbol{X}) = P_k(Y \mid \boldsymbol{X})$ for all Y and for each sample $\boldsymbol{X}^{(i)}$.

2. **M-Step**: Update P to minimize the divergence

   $$P_{k+1} = \arg\min_{P} D(Q_k \| P)$$

   The solution amounts to a weighted maximum-likelihood estimation of the parameters associated with P

   $$\begin{aligned} P_{k+1} &= \arg\min_{P} \sum_{i}^{m} \mathbb{E}_{Y \sim Q_k(Y \mid \boldsymbol{X}^{(i)})}[\log P(\boldsymbol{X}^{(i)}, Y)] \\ &= \arg\min_{P} \sum_{i}^{m} \sum_{Y} Q_k(Y \mid \boldsymbol{X}^{(i)}) \log P(\boldsymbol{X}^{(i)}, Y) \end{aligned}$$

Iterating through these steps can be shown to monotonically increase the likelihood of the data $\mathcal{D}$.

(a) **[5pts]** Suppose that $\boldsymbol{X}$ and Y are both discrete random variables, where $\boldsymbol{X}$ takes on one of 2^d values and Y is also binary. Assume that $P(\boldsymbol{X}, Y) = P(\boldsymbol{X} \mid Y)P(Y)$ is unconstrained, i.e., we are free to parameterize $P(\boldsymbol{X} \mid Y)$ and $P(Y)$ arbitrarily. Derive the E- and M-steps. Can you offer any explanation as to why this case is not particularly interesting?

(b) **[10pts]** Now, assume that $\boldsymbol{X}$ is comprised of d binary variables X_j and that the $X_j \in \boldsymbol{X}$ are conditionally independent given Y, i.e., $P(\boldsymbol{X} \mid Y) = \prod_i^d P(X_i \mid Y)$. Derive the E- and M-steps for this problem. What can you say about the optimization problem associated with the M-step? What is the expression for $P(Y)$ and $P(\boldsymbol{X} \mid Y)$ upon convergence? (Hint: You may find Jensen's inequality helpful).

(c) **[5pts]** What is the expression for $Q(Y)$ and $Q(X_j \mid Y)$?

Name: _____ Problem: _____

Name: _____ Problem: _____

Name: _____ Problem: _____