CS 228, Winter 2018
Final Exam

This exam is worth 100 points. You have 3 hours to complete it. You are allowed to consult notes, books, and use a laptop but no communication or network access is allowed. Good luck!

The Honor Code is the University's statement on academic integrity written by students in 1921. It articulates University expectations of students and faculty in establishing and maintaining the highest standards in academic work:

- The Honor Code is an undertaking of the students, individually and collectively:
  - that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
  - that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.
- The faculty on its part manifests its condence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.
- While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

Signature

I attest that I have not given or received aid in this examination, and that I have done my share and taken an active part in seeing to it that others as well as myself uphold the spirit and letter of the Stanford University Honor Code.

Name / SUnetID:

Signature:

Question	Score	Question	Score
1	/ 21	5	/ 17
2	/ 6	6	/ 11
3	/ 12	7	/ 14
4	/ 19		

Total score: **/ 100**

Note: Partial credit will be given for partially correct answers. Zero points will be given to answers left blank.

1. **[21 points total] Conceptual Short Answers**

   Each question requires an answer of about one sentence. Longer answers will not positively affect your grade. No explanation is required for multiple-choice questions.

   (i) **[2 points]** An oil exploration company is using decision trees to predict the presence/absence of oil at several locations, using a set of complex geological features f_i for each location. After taking CS 228, you think you can improve on their algorithm using a graphical model to jointly predict presence/absence at multiple locations, capturing spatial correlations among nearby locations. You have access to a large amount of labeled data, but you don't have a lot of domain knowledge in oil exploration. Which graphical model is more suitable for this? A Bayes Net, a MRF or a CRF? Why?

   Answer: CRF, reasonable to assume dependencies are symmetric, don't want to model the complex features (always observed).

   (ii) **[2 points]** Let $p(X_1, X_2, X_3)$ be a joint probability distribution specified by a graphical model G. If G is undirected, is it possible that X_1 and X_3 are (marginally) independent, but not conditionally independent given X_2? What if G is directed?

   Answer: No, Yes (v structure)

   (iii) **[2 points total]** Which of the following is true in graph G?

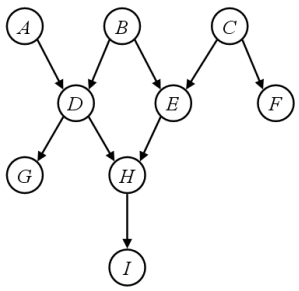


Figure 1: Graph G

   (a) d-sep$_G(F; I|E)$
   (b) d-sep$_G(G; E|\{I, B\})$
   (c) d-sep$_G(\{A, G\}; \{C, H\}|\{B, D\})$
   (d) none of the above

   Answer (c)

   (iv) **[2 points total]** Given a minimimal I-map G for a distribution p, after adding a single edge to G:

   (a) G will always be a minimal I-map for p
   (b) G will sometimes be a minimal I-map for p
   (c) G will never be a minimal I-map for p, but will always be an I-map for p
   (d) G will never be a minimal I-map for p, but will sometimes be an I-map for p
   (e) G will never be a minimal I-map for p and will never be an I-map for p
   (f) none of the above

   Answer (c)

   (v) **[2 points total]** Consider a naive Bayes model where $X_1, \ldots, X_n$ are evidence variables and C is the class variable. Suppose $|Val(X_i)| = l$ for each i and $|C| = k$. How many independent parameters are required to specify the naive Bayes model?

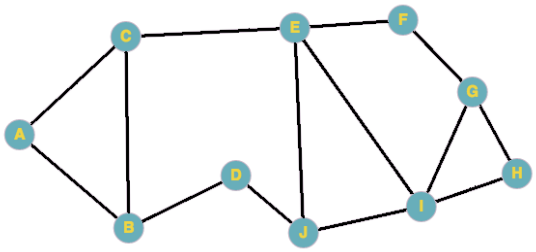
   (a) $k - 1 + n(l - 1)$
   (b) $k + nkl - 1$

(c) $k - 1 + nk(l - 1)$

(d) $k + nkl$

Answer: c

(vi) **[2 points total]** What is the treewidth of the following graph?



Answers 2

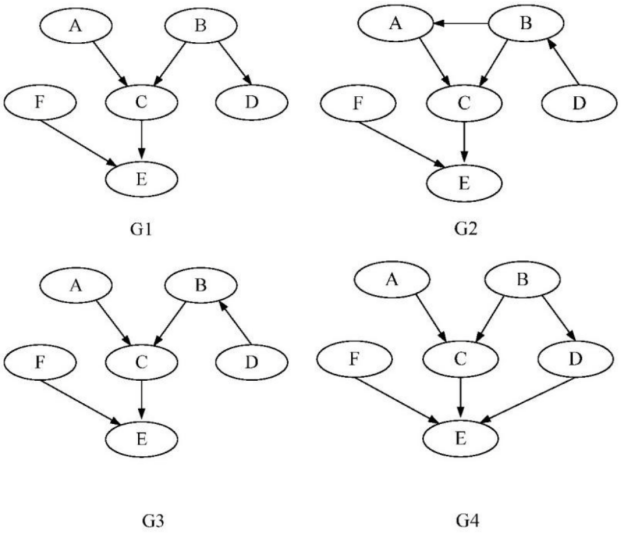
(vii) **[2 points total]** Bob wants to compute some marginal probabilities with respect to an intractable probability distribution P. Since P is intractable, Bob uses importance sampling with a proposal distribution Q. Results are not great, so Alice suggests the following. Sample from Q, and use this sample to initialize a Gibbs sampler (with P as stationary distribution). After t iterations of Gibbs, Bob gets samples from a probability distribution R^t. Intuitively, R^t is "closer" to P than Q. In fact, we know that R^t converges to P as $t \to \infty$. Alice's idea is to do importance sampling using R^T as a proposal (with $T = 100$), instead of Q. Is this a good idea? Why?

Answer: No, cannot be done because it's not possible to evaluate probabilities and compute importance weights with respect to R^T.

(viii) **[2 points]** Suppose that you're sampling from a CRF as in assignment 4 (image denoising), but you're working with very low precision floating point numbers, so values less than some nonzero ϵ are rounded to 0. What problems might this pose for your Gibbs sampling, if any?

Answers The chain is no longer ergodic: the state space may now be disconnected.

(ix) **[2 points total]** Which of the following statements about the BIC scores of the different graphs



is true?

(a) $Score_{BIC}(G1 : D) \geq Score_{BIC}(G2 : D)$ for every dataset D
(b) $Score_{BIC}(G1 : D) \geq Score_{BIC}(G4 : D)$ for every dataset D
(c) $Score_{BIC}(G2 : D) \neq Score_{BIC}(G3 : D)$ for every dataset D
(d) $Score_{BIC}(G1 : D) = Score_{BIC}(G3 : D)$ for every dataset D
(e) None of the above

Answer (d)

(x) Suppose you have a *bipartite* undirected Markov random field over disjoint sets of variables U and V (bipartite means that there are only edges between the U and V variables in the graph), specifying a joint probability distribution $p(u, v)$.

i. **[1.5 points]** You decide to use a mean field approximation for $p(u, v)$. In general, is there any theoretical guarantee for how accurate the mean field approximation will be? Why? Does the initialization of the mean field inference algorithm matter? **Answer:** can be very inaccurate, yes it matters.

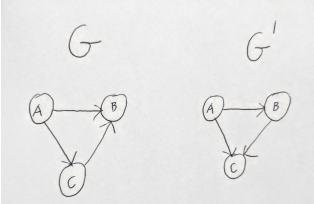
ii. **[1.5 points]** You decide to use a mean field approximation for the posterior $p(u|v)$. In general, is there any theoretical guarantee for how accurate the mean field approximation will be? Why? Does the initialization of the mean field inference algorithm matter?

Answer: it will be exact. initialization does not matter

2. **[6 points total] Representation**

In lecture 3. slide 30, we state the following theorem about Bayes Nets: If G, G' have the same skeleton and same v-structures, then $I(G) = I(G')$. We then state that the converse does not hold. Using no more than 3 variables, provide a counterexample to the converse.

Answer



In both G and G' every pair of variables is directly connnected, thus $I(G) = I(G') = \emptyset$. However, while they have the same skeleton, they do not have the same v-structures. Thus this is a counterexample, to the converse.

3. **[12 points] MRF conversion**

Give a procedure to convert any Markov random field on discrete variables $X = (X_1, \cdots, X_n)$ into a *pairwise* Markov random field. In particular, given an MRF distribution $p(X)$, specify a new distribution $p'(X, Y)$ as a *pairwise* MRF, such that $p(x) = \sum_y p'(x, y)$, where Y are any new variables added. Assume that the input $p(X) = \prod_c \phi_c(X_c)$ is described as full tables specifying the value of each factor ϕ_c for every assignment to the variables X_c in the scope of each factor. The new pairwise MRF must have a description which is polynomial in the size of the original MRF.

Answers For each clique c with variables $\mathbf{X}_c$ and potential ϕ_c, we add a single node Y_c connected to each of the variables $X_{cj} \in \mathbf{X}_c$ in a pairwise fashion. The domain size of Y_c is $|Val(\mathbf{X}_c)| = \prod_j |Val(X_{cj})|$, where each value of Y_c corresponds to a joint assignments of the variables $\mathbf{X}_c$ in the clique (thus Y_c could be represented as a vector $(Y_{c1}, Y_{c2}, \ldots, Y_{c,\dim(\mathbf{X}_c)})$). The pairwise potentials are

$$\phi'_{cj}(X_{cj}, Y_c) = 1\{Y_{cj} = X_{cj}\}$$

The potential is one if the component of Y_c corresponding to variable X_{cj} matches the value of X_{cj}. We also add a singleton potential on Y_c

$$\psi_c(Y_c) = \phi_c(Y_{c1}, Y_{c2}, \ldots, Y_{c,\dim(\mathbf{X}_c)})$$

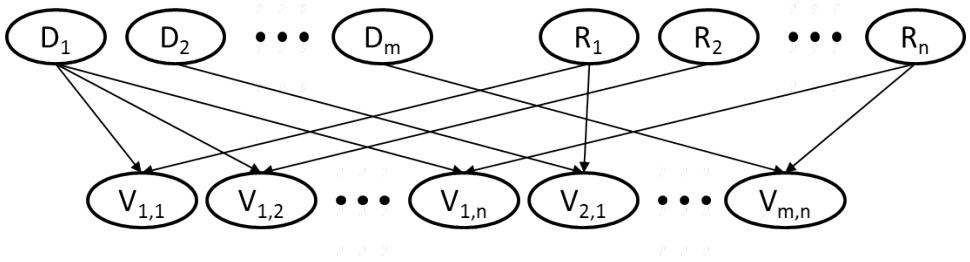
i.e., we evaluate the original potential ϕ_c at the corresponding assignment of the original $\mathbf{X}_c$ variables.

$$\begin{aligned}
\sum_Y P'(X, Y) &= \sum_Y \prod_c \psi_c(Y_c) \prod_j \phi'_{cj}(X_{cj}, Y_c) \\
&= \sum_Y \prod_c \phi_c(Y_{c1}, Y_{c2}, \ldots, Y_{c,\dim(\mathbf{X}_c)}) \prod_j 1\{Y_{cj} = X_{cj}\} \\
&= \prod_c \sum_{Y_c} \phi_c(Y_{c1}, Y_{c2}, \ldots, Y_{c,\dim(\mathbf{X}_c)}) \prod_j 1\{Y_{cj} = X_{cj}\} \\
&= \prod_c \phi_c(\mathbf{X}_c)
\end{aligned}$$

because for each value of $\mathbf{X}_c$, there is a single assignment to Y_c with non zero weight (the one matching $\mathbf{X}_c$ in each component).

4. **[19 points] Inference**

Bayesian Bob's small hometown is gearing up for the next midterm election season, with volunteers for both the Democrats and the Republicans going door-to-door to talk to the residents. Bob is the campaign manager for an independent candidate, and as part of his strategizing, he'd like to predict how the votes will turn out in his hometown. Being Bayesian Bob, he models this process with the following Bayesian network $\mathcal{G}$:



The votes are affected by the quality of the two sets of volunteers: $\mathbf{D} = \{D_1, \ldots, D_m\}$ for the Democrat campaign and $\mathbf{R} = \{R_1, \ldots, R_n\}$ for the Republican campaign; each of these variables are **binary**: $D_i \in \{0, 1\}$ for $i = 1, \cdots, m$, $R_j \in \{0, 1\}$ for $j = 1, \cdots, n$.

For simplicity, there are $m \times n$ voters in the town, represented by **binary** variables $\mathbf{V} = \{V_{1,1}, \ldots, V_{m,n}\}$. Each voter $V_{i,j}$ is contacted exactly once by volunteers D_i and R_j. The value of $V_{i,j}$ reflects which party they voted for ($V_{i,j} = 1$ if they voted Democrat, and $V_{i,j} = 0$ for Republican).

We assume that this Bayesian network is fully specified with table CPDs (and that Bob knows these CPDs).

(a) **[4 points]** Since Bob has spent his whole life growing up in this community, he has built priors over the quality of all the $m+n$ volunteers for both parties, i.e., he knows $P(D_1), \ldots, P(D_m)$ and $P(R_1), \ldots, P(R_n)$. He wants to use this information to predict which way the upcoming election will turn out. Let the sum of votes for the Democrat party be $S = \sum_{i=1}^{m} \sum_{j=1}^{n} V_{i,j}$. One way to measure the predicted election outcome is to compute the expected value of S, $\boldsymbol{E}[S]$. Describe in detail how to compute this efficiently. What is the best time complexity (in big O notation) that you can achieve, in terms of m and n?

Answers Given $\boldsymbol{D}$ and $\boldsymbol{R}$, the V's are all independent. Since $P(V_{i,j}) = \sum_{D_i, R_j} P(D_i, R_j, V_{i,j}) = \sum_{D_i, R_j} P(D_i)P(R_j)P(V_{i,j}|D_i, R_j)$, and both D_i and R_j are binary variables, we can compute this in $O(1)$ time. Using linearity of expectations, the total time complexity of calculating $\boldsymbol{E}[S] = \sum_{i=1}^{m} \sum_{j=1}^{n} P(V_{i,j} = 1)$ is therefore $O(mn)$, since there are mn voters.

(-4) Major error / blank solution

(-2) Attempt to use linearity of expectations, but wrong.

(b) **[5 points]** Suppose that after the election, Bob is able to get access to the individual votes of all the electors $\mathbf{V} = \overline{\mathbf{v}}$. Bob wants to use this information to better estimate the quality of the volunteers $\mathbf{D}, \mathbf{R}$ using $P(\mathbf{D}, \mathbf{R}|\mathbf{V})$. Which of the following is true

- $D_i \perp D_{i'} \mid \mathbf{V}$
- $R_j \perp R_{j'} \mid \mathbf{V}$
- $D_i \perp R_j \mid \mathbf{V}$
- $D_i \perp D_{i'} \mid \mathbf{V}, \mathbf{R}$
- $R_j \perp R_{j'} \mid \mathbf{V}, \mathbf{D}$

for $i \neq i'$, $j \neq j'$?

Answers: No, No, No, Yes, Yes

(c) **[4 points]**

Draw or otherwise describe a Markov Random Field $\mathcal{Q}$, which has variables $\mathbf{D} \bigcup \mathbf{R}$. $\mathcal{Q}$ should be a *perfect* map for the distribution $P(\mathbf{D}, \mathbf{R}|\mathbf{V})$, and should be described using factors that correctly

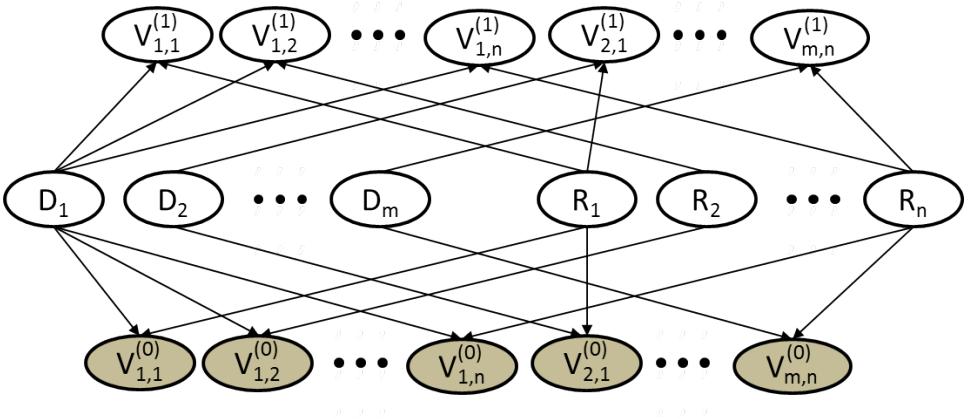
represent $P(\mathbf{D}, \mathbf{R}|\mathbf{V})$. Describe the set of factors associated with your formulation of $\mathcal{Q}$ in terms of the original CPDs in $\mathcal{G}$.

Answers: $\mathcal{Q}$ is a fully connected bipartite Markov network between the two sets of variables $\boldsymbol{D}$ and $\boldsymbol{R}$. There are $\phi_{i,j}(D_i, R_j) = P(V_{i,j} = v_{i,j}|D_i, R_j)$ for each i, j, $\phi_i(D_i) = P(D_i)$ for each i and $\phi_j(R_j) = P(R_j)$ for each j.

(-4) Major error / blank solution

(-2) Correct graph, but factors absent / wrong; or correct factors, but wrong graph.

(d) **[6 points]** Bob constructs a new Bayesian network $\mathcal{G}'$ to incorporate the previous year's votes $\boldsymbol{V}^{(0)}$:



Once again, Bob knows all the CPDs in this network. He wants to use this to repeat the calculation that you made in (a). Let $S = \sum_{i=1}^{m} \sum_{j=1}^{n} V_{i,j}^{(1)}$. Describe in detail how to compute $\boldsymbol{E}[S|\boldsymbol{V}^{(0)}]$ in this network. What is the best time complexity (in big O notation) that you can achieve, in terms of m and n? You may assume that $m < n$. Hint: consider the structure of $P(\mathbf{D}, \mathbf{R}|\mathbf{V})$ you derived in the previous part.

Answers: We're in trouble now, because $\boldsymbol{D}$ and $\boldsymbol{R}$ are activated by the V-structures. We have that $P(V_{i,j}^{(1)}|\boldsymbol{V}^{(0)}) = \sum_{D_i,R_j} P(D_i, R_j, V_{i,j}^{(1)}|\boldsymbol{V}^{(0)}) = \sum_{D_i,R_j} P(D_i, R_j|\boldsymbol{V}^{(0)}) P(V_{i,j}^{(1)}|D_i, R_j)$. We can compute the second term by looking at the CPD, but in order to compute the first term, we have to run inference in $\mathcal{Q}$. It's a fully connected bipartite graph, and as we saw in class, the minimum sepset size is m.

We can construct a chain clique tree for $\mathcal{Q}$ with cliques $\boldsymbol{C}_j = \{D_1, \ldots, D_m, R_j\}$ for $j = 1, 2, \ldots, n$. Doing inference in this clique tree takes $O(n \cdot 2^m)$ time. Once we've calibrated the clique tree, we can extract the approrpriate marginals from each clique. The straightforward method of computing this for the mn pairs takes $O(mn \cdot 2^m)$ time. Full credit was given for this solution.

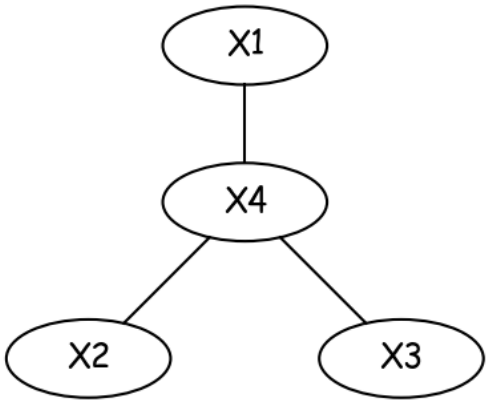
However, we can reuse computation in this step. For example, given a clique over $\{D_1, D_2, \ldots, D_8, R_1\}$, we can first compute the marginals over $\{D_1, D_2, \ldots, D_4, R_1\}$ and $\{D_5, D_6, \ldots, D_8, R_1\}$, then $\{D_1, D_2, R_1\}$ etc. Each marginal computation step involving d variables takes $O(2^d)$, so the running time for a clique is $(2 \cdot 2^m) + (2^2 \cdot 2^{\frac{m}{2}}) + \ldots)$, which is dominated by the top level step and hence $O(2^m)$. We have n cliques, our total time complexity is $O(n \cdot 2^m)$.

(-4) The exp is wrong or missing.

(-2) The multiplier is wrong or missing.

5. **[17 points] Message Passing**

Suppose we have the following Markov network on 4 binary variables X_1, X_2, X_3, X_4:



The joint probabilty can be represented as:

$$P(X_1, X_2, X_3, X_4) = \frac{1}{Z}\phi_1(X_1, X_4)\phi_1(X_2, X_4)\phi_1(X_3, X_4)\phi_2(X_4),$$

where Z is the partition function, and

$\phi_1(X, Y) =$

XY	0	1
0	1	2
1	1	1

that is, $\phi_1(0,0) = 1$, $\phi_1(0,1) = 2$, $\phi_1(1,0) = 1$, $\phi_1(1,1) = 1$, and

$\phi_2(X) =$

X	0	1
	1	2

that is, $\phi_2(0) = 1$ and $\phi_2(1) = 2$.

Note that the order of the arguments matters.

Suppose you run belief propagation (sum product) on this network.

(a) **[3 points]** What's the message $M_{1\to4}$? Give a symbolic and a numeric answer.
Answers
symbolic: $\sum_{X_1} \phi_1(X_1, X_4)$
Numeric:[2, 3]

(b) **[5 points]** What's the message $M_{4\to2}$? Give a symbolic and a numeric answer.
Answers
symbolic:

$$M_{4\to2}(X_2) = \sum_{X_4} \phi_1(X_2, X_4)\phi_2(X_4)M_{1\to4}(X_4)M_{3\to4}(X_4) \quad (1)$$

$$= [40, 22] \quad (2)$$

$$(3)$$

(c) **[2 points]** What's the marginal probability $P(X_2 = 1)$?

Answers

$$M_{4\to 2}(1)/(M_{4\to 2}(1) + M_{4\to 2}(0))$$

(d) **[2 points]** Consider the following procedure. Sample $x_1 \sim P(X_1)$, $x_2 \sim P(X_2)$, $x_3 \sim P(X_3)$. Sample $x_4 \sim P(X_4|X_1 = x_1, X_2 = x_2, X_3 = x_3)$. Is (x_1, x_2, x_3, x_4) a sample from P?

Answers No

(e) **[2 points]** Consider the following procedure. Sample $x_2 \sim P(X_2)$, $x_4 \sim P(X_4|X_2 = x_2)$, $x_1 \sim P(X_1 \mid X_4 = x_4)$. Sample $x_3 \sim P(X_3 \mid X_4 = x_4)$. Is (x_1, x_2, x_3, x_4) a sample from P?

Answers Yes

(f) **[3 points]** Consider the following procedure. Initialize $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = x_4^{(0)} = 0$, $t = 0$. Sample $x_1^{(t+1)} \sim P(X_1|X_4 = x_4^{(t)})$. Sample $x_2^{(t+1)} \sim P(X_2|X_4 = x_4^{(t)})$. Sample $x_3^{(t+1)} \sim P(X_3|X_4 = x_4^{(t)})$. Sample $x_4^{(t+1)} \sim P(X_4|X_1 = x_1^{(t+1)}, X_2 = x_2^{(t+1)}, X_3 = x_3^{(t+1)})$. As $t \to \infty$, is $(x_1^{(t)}, x_2^{(t)}, x_3^{(t)}, x_4^{(t)})$ a sample from P?

Answers Yes, this is a gibbs sampler

6. **[11 points] Score-Based Structure Learning**

In score-based approaches of structure learning, we first define a score function $score(\mathcal{G}; \mathcal{D})$ that can score each candidate structure $\mathcal{G}$ with respect to the training data $\mathcal{D}$. After the definition of score function, we search in the space of directed acyclic graphs (DAGs) to find the graph structure $\mathcal{G}$ that maximizes the score $score(\mathcal{G}; \mathcal{D})$.

(a) **[3 points]** If the score function $score(\mathcal{G}; \mathcal{D})$ is defined as the log-likelihood $LL(\mathcal{D} \mid \mathcal{G})$, we have seen that a fully-connected graph $\mathcal{G}^{full}$ always maximizes the score. Let G' be another Bayes Net structure. Let $\widehat{p}$ denote the empirical data distribution corresponding to $\mathcal{D}$. Under what conditions on $\widehat{p}$ is $LL(\mathcal{D} \mid \mathcal{G}^{full}) = LL(\mathcal{D} \mid \mathcal{G}')$? You may assume that the topological order of G' and G^{full} is the same. Hint: Your answer should consist of a list of conditional independencies.

Answer: For every variable, $X_i \perp \{X_1, \cdots, X_{i-1}\} \setminus \mathrm{Pa}_{X_i} \mid \mathrm{Pa}_{X_i}$ according to $\widehat{p}$

(b) **[2 points]** An additional term is usually added to the score function,

$$score(\mathcal{G}; \mathcal{D}) = LL(\mathcal{D} \mid \mathcal{G}) - \psi(M)\|G\|,$$

where $\|\mathcal{G}\|$ is the number of parameters in $\mathcal{G}$ and M is the number of data samples. When $\psi(M) = 1$, it is called the AIC score; when $\psi(M) = \frac{1}{2}\log_e M$, it is called the BIC score. Briefly explain why this additional term will prevent us from obtaining a fully-connected graph. Would you expect to obtain a simpler model $\mathcal{G}$ using AIC or BIC?

Answer: Fully connected graph will be penalized because it requires many parameters to be specified. BIC, because it penalized complexity more.

(c) **[2 points]** Given a score function, a local search algorithm could help us find a locally optimal graph structure. Concretely, a local search method is an iterative procedure that starts with an initial guess for the best structure. Then, at each step we will consider all the neighbors of the current best graph. If none of them has a higher score than the current guess, we terminate the search. Otherwise, we pick the neighbor with the highest score as our new best guess. The neighbors of the graph are defined as all valid Bayesian networks obtained by a single operation of edge addition, deletion or reversal.

Which of the following are neighbors of the graph in Figure 2?

(a) removal of edge $B \to A$

(b) reversal of edge $D \to A$

(c) addition of edge $E \to C$

Answer: a,c

(d) **[4 points]** If we consider to modify the structure in Figure 2 by reversing the edge $D \to B$, obtaining a new graph $\mathcal{G}''$, express the change in score

$$score(\mathcal{G}; \mathcal{D}) - score(\mathcal{G}''; \mathcal{D})$$

in terms of mutual information terms $MI(\mathbf{X}_i, \mathbf{Z}_i)$ (with respect to the data distribution), entropy terms $H(X_i)$, and changes in number of parameters in the CPDs of $\mathcal{G}$ and $\mathcal{G}''$.

Hint: use the fact that the score can be decomposed as a sum of local terms over each variable and its parents (a "family score")

Answer: Let

$$FamScore(X_i; pa(X_i)|\mathcal{D}) = M\left(MI(X_i, \mathbf{X}_{pa(X_i)}) - H(X_i)\right) - \psi(M)||\theta_{X_i|pa(X_i)}||$$

where $||\theta_{X_i|pa(X_i)}||$ is the number of independent parameters of the CPD $P(X_i|pa(X_i))$ The change in score is

$$FamScore(B; D|\mathcal{D}) + FamScore(D; E|\mathcal{D}) - FamScore(B; \emptyset|\mathcal{D}) - FamScore(D; B, E|\mathcal{D})$$

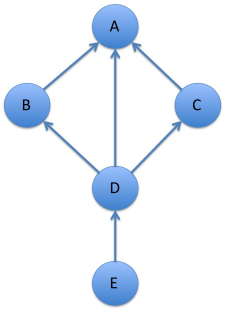


Figure 2: Part (c)

7. **[14 points] Parameter estimation**

Suppose $X = (X_1, \cdots, X_p)$ consists of p binary variables taking value one or zero. If the variables X_j are independent, and each X_j takes value one with probability π_j , then the joint distribution of X is given by

$$P[X = x] = \prod_{j=1}^{p} \pi_j^{x_j}(1 - \pi_j)^{1-x_j}$$

where $x = (x_1, \cdots, x_p)$, and $x_j \in \{0, 1\}$.

We now consider a mixture of distributions of the preceding type. With M mixture components, X has the mixture distribution

$$f(x; \theta) = \sum_{m=1}^{M} \phi_m f_m(x; \theta_m)$$

$$f_m(x; \theta_m) = \prod_{j=1}^{p} \pi_{j,m}^{x_j}(1 - \pi_{j,m})^{1-x_j}$$

where $\theta = (\phi_1, \cdots, \phi_M, \theta_1, \cdots, \theta_M)$ and each $\theta_m = (\pi_{1,m}, \cdots, \pi_{p,m})$ is itself a vector of p parameters.

We can interpret the mixture distribution using a latent variable. Suppose there is a latent variable Z which takes values $1, \cdots, M$ with probabilities $\phi_1, \cdots, \phi_M$ respectively. Assume given $Z = m$, the conditional distribution of X is given by $f_m(x; \theta_m)$. In this way we have specified a joint distribution of (X, Z). If you compute the marginal distribution of X from this joint distribution, you will obtain the mixture distribution.

(a) **[3 points]**
Suppose we have a training set $(x^{(1)}, ..., x^{(N)})$, from which we want to estimate the parameter set θ. Each training point $x^{(i)} = (x_1^{(i)}, \cdots, x_p^{(i)})$ is a p-dimensional vector with binary entries.
Imagine that for each training point $x^{(i)}$, there is an associated latent variable $z^{(i)} \in \{1, \cdots, M\}$, which we did not observe. Assuming that $z^{(i)}$ are observed, write down the log likelihood function given by the complete data $(x^{(1)}, z^{(1)}), \cdots, (x^{(N)}, z^{(N)})$. Find out the MLE for θ.
Answer:

$$\sum_{i=1}^{N} \sum_{m=1}^{M} 1[z^{(i)} = m](\log \phi_m + \sum_{j=1}^{p} x_j^{(i)} \log(\pi_{j,m}) + (1 - x_j^{(i)}) \log(1 - \pi_{j,m}))$$

MLE estimates

$$\phi_m^* = \frac{\sum_{i=1}^{N} 1[z^{(i)} = m]}{N}$$

$$\pi_{j,m}^* = \frac{\sum_{i=1}^{N} x_j^{(i)} 1[z^{(i)} = m]}{\sum_{i=1}^{N} 1[z^{(i)} = m]}$$

(b) **[5 points]** Now let's assume the latent variables $z^{(i)}$ are not observed, and we use EM to estimate θ. Provide the E-step of the algorithm.
Answer:

$$\gamma[z^{(i)}, m] = P[Z = m | X = x^{(i)}] = \frac{\phi_m \prod_{j=1}^{p} \pi_{j,m}^{x_j^{(i)}}(1 - \pi_{j,m})^{1-x_j^{(i)}}}{\sum_m \phi_m \prod_{j=1}^{p} \pi_{j,m}^{x_j^{(i)}}(1 - \pi_{j,m})^{1-x_j^{(i)}}}$$

(c) **[5 points]** Provide the M-step of the algorithm.
Answer:

$$\phi_m^* = \frac{\sum_{i=1}^{N} \gamma[z^{(i)}, m]}{N}$$

$$\pi_{j,m}^* = \frac{\sum_{i=1}^{N} x_j^{(i)} \gamma[z^{(i)}, m]}{\sum_{i=1}^{N} \gamma[z^{(i)}, m]}$$

(d) **[1 points]** Provide a valid initialization for the EM algorithm.
Answer: Anything within the bounds works.