Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2020

QUIZ 1 ANSWER BOOKLET

Thursday, October 15, 2020
9:00 am – 11:00 am / 7:00 pm – 9:00 pm

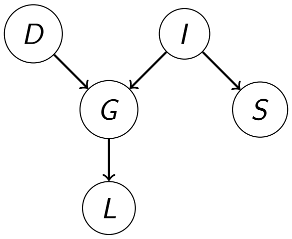
NAME: Andy Markov

- Don't forget to put your name on all sheets.
- Remember that only this answer booklet will be considered in the grading of your exam.
- Be sure to **show all relevant work and reasoning.**
- Please be neat! You may want to first work things through on scratch paper and then neatly transfer to this answer booklet the work you would like us to look at.

Problem	Your score
1	20
2	20
3	20
Total	60

Problem 1

(a)(i) Draw the DAG here:



(a)(ii) Determine if the CI relations are true or false:

- $D \perp\!\!\!\perp I$: True
- $D \perp\!\!\!\perp S \mid L$: False
- $S \perp\!\!\!\perp L \mid I$: True

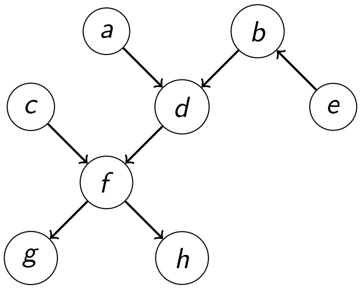
Reasoning/Work to be looked at for Problem 1(a):

(a)(i) The DAG can be obtained from the causal relationships described in the problem.

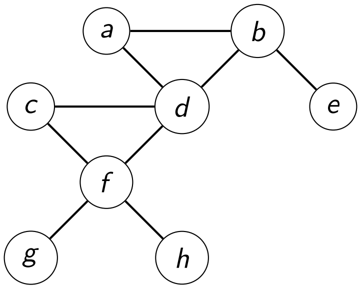
(a)(ii) The random variables D and I are independent of each other because both G and I are not observed, so the path between D and I is blocked by G. The random variables D and S are dependent given L, because I is not observed (hence doesn't block the path), and L, which is a descendant of G, is observed (hence G doesn't block the path). The random variables S and L are independent given I because the path is blocked by I.

One common mistake is that the students did not comment about I not blocking the path for $D \perp\!\!\!\perp S \mid L$ to be false. Also, quite a few students did not have the correct explanation as to why $D \perp\!\!\!\perp I$ is True.

(b)(i) Draw the I-equivalent DAG here:



(b)(ii) Draw the minimal undirected I-map here:



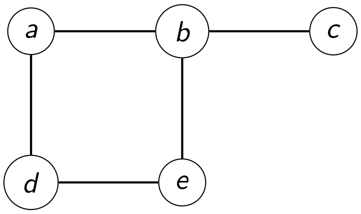
Reasoning/Work to be looked at for Problem 1(b):

(b)(i) As seen in Problem Set 1, two DAGs are I-equivalent if and only if they have the same skeleton and the same set of immoralities. The DAG drawn in the answer is the only DAG that is I-equivalent to $\mathcal{G}$, because flipping any other edges will introduce/remove immoralities in the graph.

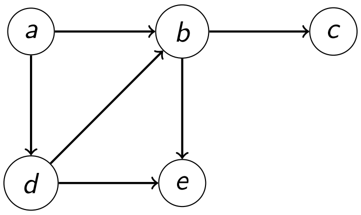
A few students failed to explain why there is only one DAG that is I-equivalent to $\mathcal{G}$.

(b)(ii) An undirected graph is the *moralization* of a directed graph if the undirected graph has an (undirected) edge between every pair of nodes that the directed graph does, and, in addition, (undirected) edges between every pair nodes in the set of parents π_i, for each node $i \in \mathcal{V}$.

(c)(i) Draw the undirected graph here:



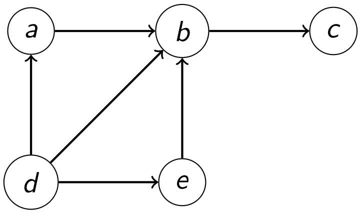
(c)(ii) Draw the DAG here:



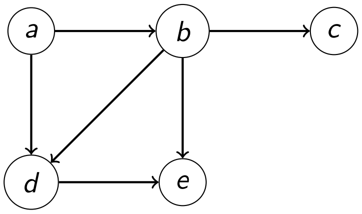
Reasoning/Work to be looked at for Problem 1(c):

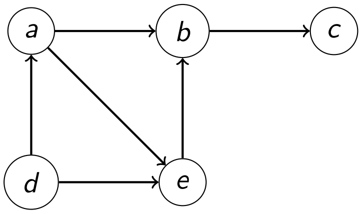
(c)(i) Using the list CI relations and the pairwise Markov property, we can start from a complete graph, delete edges, and arrive at the solution. Also, using the separation of the nodes in the graph, we can verify that the graph is indeed a P-map for the distribution.

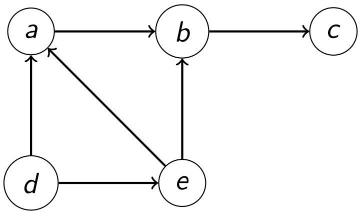
(c)(ii) As seen in the lecture materials, a non-chordal undirected graph must be chordalized to be converted into a DAG. However, just chordalizing the graph and assigning arbitrary directions to the edges are *not* enough. We have to be careful about the direction of edges. For example, the following DAG is *not* a minimal I-map, because it introduces $a \perp\!\!\!\perp e \mid d$ which is not on the list of CI relations.

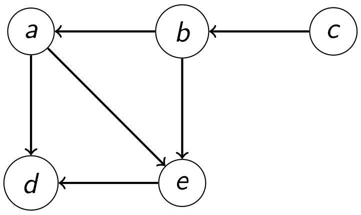


The DAG drawn in the solution is just one example of correct chordalizations. The DAG successfully represents all the CI relations, except $b \perp\!\!\!\perp d \mid a, e$ and $b \perp\!\!\!\perp d \mid a, c, e$. There are many correct answers to this problem. Other equivalently correct answers include:


a
b
c
d
e


a
b
c
d
e


a
b
c
d
e


a
b
c
d
e

Problem 2

(a)

$$\begin{aligned}
f(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)};\beta_{12}) &= \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_2^{(i)}-\beta_{12}x_1^{(i)})^2\right)\\
g(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)};\beta_{13}) &= \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_3^{(i)}-\beta_{13}x_1^{(i)})^2\right)\\
h(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)};\beta_{24},\beta_{34}) &= \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_4^{(i)}-\beta_{24}x_2^{(i)}-\beta_{34}x_3^{(i)})^2\right)
\end{aligned}$$

or any valid scaling of the three.

Reasoning/Work to be looked at for Problem 2(a):

We first factorize based on the DAG and the independence of the samples:

$$p(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)}) = \prod_{i=1}^{N} p(x_1^{(i)})p(x_2^{(i)} \mid x_1^{(i)})p(x_3^{(i)} \mid x_1^{(i)})p(x_4^{(i)} \mid x_2^{(i)}, x_3^{(i)})$$

$p(x_1)$ doesn't depend on any of the parameters. We see that

$$\prod_{i=1}^{N} p(x_2^{(i)} \mid x_1^{(i)}) \propto \prod_{i=1}^{N} \exp\left(-\frac{(x_2^{(i)}-\beta_{12}x_1^{(i)})^2}{2}\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_2^{(i)}-\beta_{12}x_1^{(i)})^2\right)$$

$$\prod_{i=1}^{N} p(x_3^{(i)} \mid x_1^{(i)}) \propto \prod_{i=1}^{N} \exp\left(-\frac{(x_3^{(i)}-\beta_{13}x_1^{(i)})^2}{2}\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_3^{(i)}-\beta_{13}x_1^{(i)})^2\right)$$

$$\prod_{i=1}^{N} p(x_4^{(i)} \mid x_2^{(i)}, x_3^{(i)}) \propto \prod_{i=1}^{N} \exp\left(-\frac{(x_4^{(i)}-\beta_{24}x_2^{(i)}-\beta_{34}x_3^{(i)})^2}{2}\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_4^{(i)}-\beta_{24}x_2^{(i)}-\beta_{34}x_3^{(i)})^2\right)$$

One common mistake is that students did not explicitly give the pdfs for the conditional distributions. However, full credit was still given if this formula appeared in parts (b) or (c).

(b)

$$\hat{\beta}_{12} = \frac{\sum x_1^{(i)} x_2^{(i)}}{\sum {x_1^{(i)}}^2}$$

$$\hat{\beta}_{13} = \frac{\sum x_1^{(i)} x_3^{(i)}}{\sum {x_1^{(i)}}^2}$$

$$\hat{\beta}_{24} = \frac{(\sum x_2^{(i)} x_4^{(i)})(\sum {x_3^{(i)}}^2) - (\sum x_3^{(i)} x_4^{(i)})(\sum x_2^{(i)} x_3^{(i)})}{(\sum {x_2^{(i)}}^2)(\sum {x_3^{(i)}}^2) - (\sum x_2^{(i)} x_3^{(i)})^2}$$

$$\hat{\beta}_{34} = \frac{(\sum x_3^{(i)} x_4^{(i)})(\sum {x_2^{(i)}}^2) - (\sum x_2^{(i)} x_4^{(i)})(\sum x_2^{(i)} x_3^{(i)})}{(\sum {x_2^{(i)}}^2)(\sum {x_3^{(i)}}^2) - (\sum x_2^{(i)} x_3^{(i)})^2}$$

Reasoning/Work to be looked at for Problem 2(b):

We can maximize each of the factors independently.

The log likelihood for $x_2 \mid x_1$ is:

$$\log p(x_2 \mid x_1) = -\frac{1}{2} \sum (x_2^{(i)} - \beta_{12} x_1^{(i)})^2 + C$$

The gradient with respect to β_{12} is

$$\frac{\partial \log p(x_2 \mid x_1)}{\partial \beta_{12}} = \sum (x_2^{(i)} - \beta_{12} x_1^{(i)}) x_1^{(i)} = \sum x_1^{(i)} x_2^{(i)} - \beta_{12} \sum {x_1^{(i)}}^2$$

Therefore

$$\hat{\beta}_{12} = \frac{\sum x_1^{(i)} x_2^{(i)}}{\sum {x_1^{(i)}}^2}$$

Similarly,

$$\hat{\beta}_{13} = \frac{\sum x_1^{(i)} x_3^{(i)}}{\sum {x_1^{(i)}}^2}$$

Finally, the log-likelihood for $x_4 \mid x_2, x_3$ is

$$\log p(x_4 \mid x_2, x_3) = -\frac{1}{2} \sum (x_4^{(i)} - \beta_{24} x_2^{(i)} - \beta_{34} x_3^{(i)})^2 + C$$

The gradient with respect to β_{24} is

$$\frac{\partial \log p(x_4 \mid x_2, x_3)}{\partial \beta_{24}} = \sum x_2^{(i)} x_4^{(i)} - \beta_{24} \sum {x_2^{(i)}}^2 - \beta_{34} \sum x_2^{(i)} x_3^{(i)}$$

Likewise

$$\frac{\partial \log p(x_4 \mid x_2, x_3)}{\partial \beta_{34}} = \sum x_3^{(i)} x_4^{(i)} - \beta_{24} \sum x_2^{(i)} x_3^{(i)} - \beta_{34} \sum x_3^{(i)2}$$

Solving the system of equations, we obtain

$$\hat{\beta}_{24} = \frac{(\sum x_2^{(i)} x_4^{(i)})(\sum {x_3^{(i)}}^2) - (\sum x_3^{(i)} x_4^{(i)})(\sum x_2^{(i)} x_3^{(i)})}{(\sum {x_2^{(i)}}^2)(\sum {x_3^{(i)}}^2) - (\sum x_2^{(i)} x_3^{(i)})^2}$$
$$\hat{\beta}_{34} = \frac{(\sum x_3^{(i)} x_4^{(i)})(\sum {x_2^{(i)}}^2) - (\sum x_2^{(i)} x_4^{(i)})(\sum x_2^{(i)} x_3^{(i)})}{(\sum {x_2^{(i)}}^2)(\sum {x_3^{(i)}}^2) - (\sum x_2^{(i)} x_3^{(i)})^2}.$$

Alternatively, you could recognize that each individual optimization is equivalent to solving a least-squares linear regression problem, in which case the MLE can be computed by using the pseudoinverse.

(c) Sufficient Statistics:

$$x_1^2, x_2^2, x_3^2, x_4^2, x_1x_2, x_1x_3, x_2x_3, x_2x_4, x_3x_4$$

Natural Parameters:

$$-\frac{1}{2}(1+\beta_{12}^2+\beta_{13}^2), -\frac{1}{2}(1+\beta_{24}^2), -\frac{1}{2}(1+\beta_{34}^2), -\frac{1}{2}, \beta_{12}, \beta_{13}, -\beta_{24}\beta_{34}, \beta_{24}, \beta_{34}$$

Reasoning/Work to be looked at for Problem 2(c):

We see that

$$p(x_1) \propto \exp\left(-\frac{1}{2}x_1^2\right)$$

$$\begin{aligned} p(x_2 \mid x_1) &\propto \exp\left(-\frac{(x_2-\beta_{12}x_1)^2}{2}\right) \\ &= \exp\left(-\frac{1}{2}x_2^2 + \beta_{12}x_1x_2 - \frac{\beta_{12}^2}{2}x_1^2\right) \end{aligned}$$

$$\begin{aligned} p(x_3 \mid x_1) &\propto \exp\left(-\frac{(x_3-\beta_{13}x_1)^2}{2}\right) \\ &= \exp\left(-\frac{1}{2}x_3^2 + \beta_{13}x_1x_3 - \frac{\beta_{13}^2}{2}x_1^2\right) \end{aligned}$$

$$\begin{aligned} p(x_4 \mid x_2, x_3) &\propto \exp\left(-\frac{(x_4-\beta_{24}x_2-\beta_{34}x_3)^2}{2}\right) \\ &= \exp\left(-\frac{1}{2}x_4^2 + \beta_{24}x_2x_4 + \beta_{34}x_3x_4 - \beta_{24}\beta_{34}x_2x_3 - \frac{\beta_{24}^2}{2}x_2^2 - \frac{\beta_{34}^2}{2}x_3^2\right) \end{aligned}$$

Therefore this is an exponential family with sufficient statistics

$$x_1^2, x_2^2, x_3^2, x_4^2, x_1x_2, x_1x_3, x_2x_3, x_2x_4, x_3x_4$$

and corresponding natural parameters:

$$-\frac{1}{2}(1+\beta_{12}^2+\beta_{13}^2), -\frac{1}{2}(1+\beta_{24}^2), -\frac{1}{2}(1+\beta_{34}^2), -\frac{1}{2}, \beta_{12}, \beta_{13}, -\beta_{24}\beta_{34}, \beta_{24}, \beta_{34}$$

A common mistake was writing the generic sufficient statistics and natural parameters for the exponential family of a joint Gaussian, in terms of the information matrix J. Remember to only use terms that are defined in the problem.

(d) **Reasoning/Work to be looked at for Problem 2(d):**

Each conditional distribution is of the form:

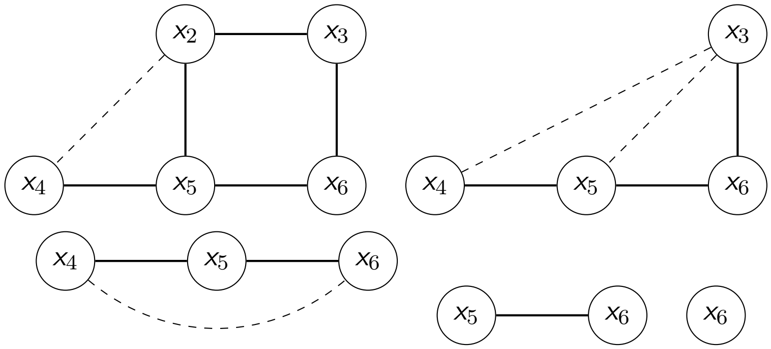
$$p(x_i \mid x_{\pi_i}) \propto \exp\left(-\frac{1}{2}(x_i - \sum_{j\in\pi_i} \beta_{ij}x_j)^2\right)$$

Expanding, we observe that there's an $x_j x_k$ term if (1) $j = i$ and $k \in \pi_i$ or (2) $j, k \in \pi_i$. Therefore $x_i x_j$ is a sufficient statistic iff either i and j are connected by an edge, or i, j have a common child.

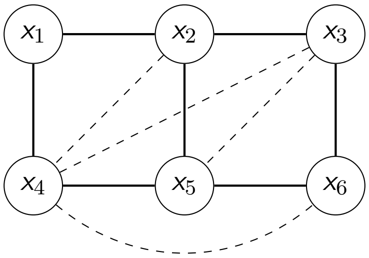
Consider any I-Map of $\mathcal{G}$. Clearly if there is a directed edge $i \rightarrow j$, then i and j are connected in the I-Map. Similarly, if $i \rightarrow k \leftarrow j$ is a V-structure, then i and j are not conditionally independent given all other vertices as the path $i - j - k$ is unblocked. Therefore i and j must be connected in any I-Map. Finally, these edges (the original edges and moralizing edges) are the only edges that must be in every I-Map, since the moralizing graph (consisting of only the original edges and moralizing edges) is an I-Map itself. Therefore $x_i x_j$ is a sufficient statistic if and only if the edge (i, j) is in every I-map. One common mistake was mistaking undirected I-maps for minimal undirected I-maps, and stating that a DAG and an I-map have the same skeleton.

Problem 3

(a)(i) Draw the sequence of graphs:



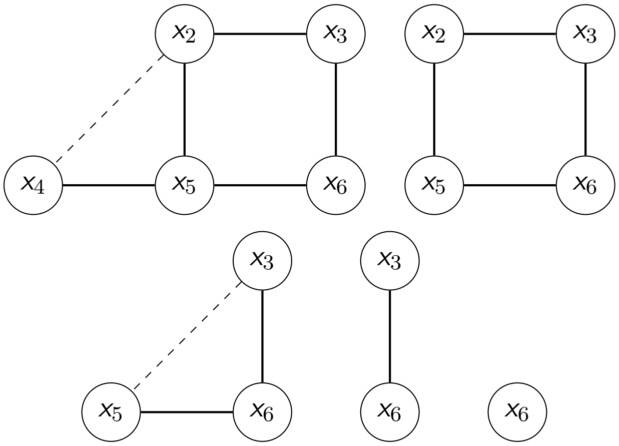
Draw the reconstituted graph:



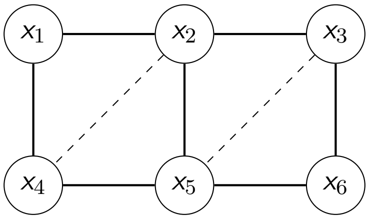
(a)(ii) Choose your elimination order: $(1, 4, 2, 5, 3, 6)$.

With this elimination order, the maximum clique size in the reconstituted graph is reduced from 4 to 3. There are other elimination orders that result in the same maximum clique size. However, note that the goal of this elimination is to compute the marginal distribution of x_6, so x_6 has to be the *last* element in the order.

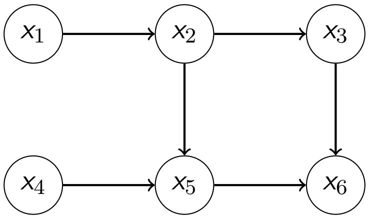
Draw the sequence of graphs:



Draw the reconstituted graph:



(b)(i) Draw the interventional graph:



(b)(ii) Write the distribution $p(x_1, x_2, x_3, x_5, x_6 \mid \mathrm{do}(\mathsf{x}_4 = x_4))$ in terms of factors:

$$p(x_1, x_2, x_3, x_5, x_6 \mid \mathrm{do}(\mathsf{x}_4 = x_4)) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_5 \mid x_2, x_4)p(x_6 \mid x_3, x_5)$$

(b)(iii) **Reasoning/Work to be looked at for Problem 3(b)(iii):**

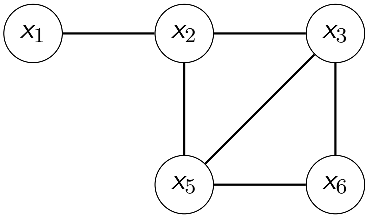
Fix x_4. Then, the interventional distribution factorizes as

$$p(x_1, x_2, x_3, x_5, x_6 \mid \mathrm{do}(\mathsf{x}_4 = x_4)) = \psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{25}(x_2, x_5)\psi_{356}(x_3, x_5, x_6)$$

where

$$\begin{aligned} \psi_{12}(x_1, x_2) &:= p(x_1)p(x_2 \mid x_1) \\ \psi_{23}(x_2, x_3) &:= p(x_3 \mid x_2) \\ \psi_{25}(x_2, x_5) &:= p(x_5 \mid x_2, x_4) \\ \psi_{356}(x_3, x_5, x_6) &:= p(x_6 \mid x_3, x_5) \end{aligned}$$

Thus we're working with the following undirected graphical model:



One can see that if we run the elimination algorithm with say $(1, 2, 3, 5, 6)$ as the ordering, the reconstituted graph is the same as the original graph and as the maximum clique size is 3, runs in $O(|\mathcal{X}|^3)$ time. Repeating this for each value of $x_4 \in \mathcal{X}$ leads to an algorithm which runs in $O(|\mathcal{X}|^4)$ time.