PRACTICE MIDTERM EXAM

STA 414/2104 - Winter 2024

University of Toronto

Exam duration: **100 minutes**

Note: The midterm will have 7 questions and so it will be shorter than this midterm practice. No calculators will be allowed during the midterm exam.

Read the following instructions carefully:

1. Exam is closed book and internet. You can use an optional handwritten aid sheet - A4 double-sided.
2. If a question asks you to do some calculations, you must **show your work** for full credit.
3. Conceptual questions do not require long answers.
4. You will write your answers to each question in the space provided on the exam sheet. If you require additional paper, simply raise your hand.
5. After solving each question, you should write your answers immediately. Do not wait last minute to write them all at once.
6. Do not share the exam with anyone or in any platform!
7. Lastly, enjoy the problems!!!

1. Exponential families - 8pts. Probability mass function of a random variable X distributed as geometric distribution with parameter γ, with $0 < \gamma < 1$, is given as

$$\mathbb{P}(X = k) = \gamma(1-\gamma)^{k-1} \quad \text{for} \quad k = 1, 2, ...$$

(a) Show that this is a probability mass function. Hint: for $0 < p < 1$, $\sum_{k=0}^{\infty} p^k = 1/(1-p)$.
(b) Write the above distribution as an exponential family, and identify its sufficient statistics, natural parameter, and log-partition function.
(c) Assume that we observed $X_1, X_2, ..., X_n$ i.i.d. random variables from geometric distribution with an unknown parameter γ. Find the MLE for γ.

(a) $\mathbb{P}(X = k) \geq 0$ and $\sum_{k=0}^{\infty} \mathbb{P}(X = k) = \gamma \sum_{k=0}^{\infty} (1-\gamma)^{k-1} = 1$ by the hint.
(b) We can write $\mathbb{P}(X = x) = \exp\{\log(\gamma(1-\gamma)^{x-1})\} = \exp\{\log(\gamma) + (x-1)\log(1-\gamma))\}$. The sufficient statistics is $T(x) = x - 1$, natural parameter is $\eta = \log(1-\gamma)$; equivalently, $\gamma = 1 - e^{\eta}$. The normalizing constant (log-partition function) is $\log(1 - e^{\eta})$.
(c) The log-likelihood can be written as

$$\log p(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} \log p(X_i) = \sum_i \log(\gamma(1-\gamma)^{X_i - 1}) = n \log(\tfrac{\gamma}{1-\gamma}) + \log(1-\gamma) \sum_i X_i.$$

Taking derivatives w.r.t. γ, we solve for

$$\frac{n}{\gamma(1-\gamma)} - \frac{1}{1-\gamma} \sum_i X_i = 0$$

thus $\hat{\gamma} = 1/\bar{X}$ where $\bar{X} = \sum_i X_i / n$.

2. Maximum likelihood estimation and unnormalised models - 10 pts. Consider a model for three binary random variables (x_1, x_2, x_3),

$$p_\theta(x_1, x_2, x_3) \propto \exp\{\theta x_1 x_2 + \theta x_2 x_3\}, \qquad x_i \in \{0, 1\}.$$

1. What is the sufficient statistics of this exponential family?
   The sufficient statistics is $x_1 x_2 + x_2 x_3$.
2. Compute the partition function $Z(\theta)$ and the derivative of $A(\theta) = \log Z(\theta)$.
   There are 8 possible states $(0,0,0), \ldots, (1,1,1)$. In 5 cases the sufficient statistic is 0, in 2 cases it is 1, and in 1 case it is 2. Thus,

   $$Z(\theta) \;=\; 5 + 2e^{\theta} + e^{2\theta}.$$

   It follows that

   $$A'(\theta) \;=\; \frac{2e^{\theta} + 2e^{2\theta}}{5 + 2e^{\theta} + e^{2\theta}}.$$

3. Verify that for the sample $\{(1,1,1), (1,1,1), (1,1,0), (0,1,1), (0,1,0)\}$ the maximum likelihood is $\hat{\theta} = \ln(3)$. You will not need a calculator for this computation.
   For the given sample the sample average of the sufficient statistics is $\frac{1}{5}(2+2+1+1+0) = \frac{6}{5}$. We need to verify that

   $$A'(\ln(3)) \;=\; \tfrac{6}{5}$$

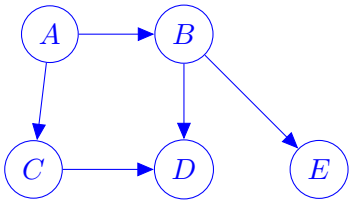
   which can be easily checked.
4. Compute the joint distribution $p_{\hat{\theta}}(x_1, x_2, x_3)$ corresponding to this MLE.
   Since $\exp\{\hat{\theta} x_1 x_2 + \hat{\theta} x_2 x_3\} = 3^{x_1 x_2 + x_2 x_3}$ and $Z(\hat{\theta}) = 20$ we have

(x_1, x_2, x_3)	000	001	010	011	100	101	110	111
$p(x_1, x_2, x_3)$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{9}{20}$

3. Graphical models - 16 pts. *No explanation needed, just your answers.*

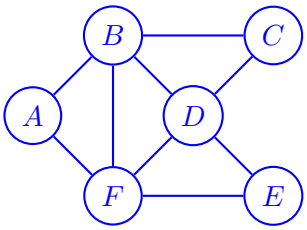
(a) (4 pts) Draw the DAG corresponding to the following factorization of a joint distribution:

$$p(A, B, C, D, E) = P(A)P(B|A)P(C|A)P(D|B, C)P(E|B)$$



(b) (4 pts) Draw the Markov Random Field that corresponds to the following factorization.

$$p(A, B, C, D, E, F) \propto \phi_{A,B,F}(A, B, F)\phi_{B,C,D}(B, C, D)\phi_{D,E,F}(D, E, F)\phi_{B,D,F}(B, D, F)$$



(c) (4 pts) Write the variables that belong to the Markov blanket of node 3 in the Figure 1.

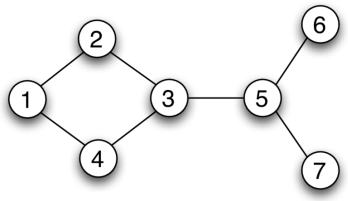


Fig 1: Simple MRF

Nodes 2,4,5.

(d) (4 pts) Belief propagation algorithm is run on a tree graph to compute the marginal of a node x.

- How many passes in which direction is sufficient to compute the marginal of x, given that we choose x to be the root?
- How many passes in which direction is sufficient to compute the marginal of x, given that we choose a root that is not the node x?

Here, the direction is either from leaves to root or from root to leaves, and a single pass refers to passing all messages pointing to one direction (either from root to leaves or from leaves to root).

i) from leaves to root, 1 pass. ii) 2 passes, from leaves to root and reverse.

4. Decision Theory - 5 pts. Imagine we are running a nuclear power plant that is undergoing a malfunction. We have two options: A) Vent the core, and B) do nothing.

Our current beliefs are that the amount of radiation in the core is uniform between 10 and 20 units, i.e.

$$\text{variant 1:} \quad R|\text{vent} \sim U(10, 20)$$

If we do nothing, there is a $X\%$ chance that no radiation will be released, and a $(1 - X)\%$ that 100 units of radiation will be released.

For what range of probabilities X would venting the core release less radiation in expectation?

The expected radiation if we vent is 15, and if we don't vent is $100 - 100X$. So if $X < 0.85$ we should vent the core.

5. Simple Monte Carlo - 12 pts. Imagine we have a rain prediction model that outputs samples of

$$P(R_1, R_2, \ldots, R_T | \text{measurements})$$

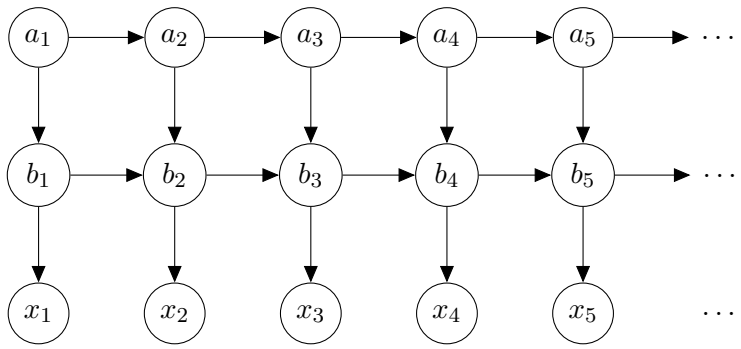
where each R_i is the predicted probability of rain i days ahead.

Given a set of N i.i.d. samples from this joint predictive distribution:

$$\begin{aligned} r_1^{(1)}, r_2^{(1)}, \ldots, r_T^{(1)} &\sim P(R_1, R_2, \ldots, R_T | \text{measurements}) \\ r_1^{(2)}, r_2^{(2)}, \ldots, r_T^{(2)} &\sim P(R_1, R_2, \ldots, R_T | \text{measurements}) \\ &\vdots \\ r_1^{(N)}, r_2^{(N)}, \ldots, r_T^{(N)} &\sim P(R_1, R_2, \ldots, R_T | \text{measurements}) \end{aligned} \tag{5.1}$$

1. **[3 points]** Write an unbiased estimator for the probability that it rains every day for the next T days. You might want to use the notation $I(\text{statement})$ which takes value 1 if the statement is true, and 0 if it is false.
   $\frac{1}{N}\sum_{i=1}^{N} I(r_1^{(i)} = 1, r_2^{(i)} = 1, ..., r_N^{(i)} = 1)$
2. **[3 points]** What is the variance of this estimator as a function of N?
   For $P(R_1 = 1, R_2 = 1, \ldots, R_T = 1 | \text{measurements}) = p$, variance is $p(1-p)/N$
3. **[3 points]** Write an unbiased estimator for the probability that it rains on day 3.
   $\frac{1}{N}\sum_{i=1}^{N} I(r_3^{(i)} = 1)$
4. **[3 points]** Write an unbiased estimator for the probability that it rains on day 3 given that it rained on day 4.
   $\sum_{i=1}^{N} I(r_3^{(i)} = 1, r_4^{(i)} = 1) / \sum_{i=1}^{N} I(r_4^{(i)} = 1)$

6. HMM Question - 12 pts. Given the following DAG:



1. **[2 points]** Write the factorized joint distribution implied by this DAG. Don't be afraid to add extra brackets or parentheses to avoid ambiguity.

$$p(a_1, a_2, \ldots, a_T, b_1, b_2, \ldots, b_T, x_1, x_2, \ldots, x_T) =$$

Answer:

$$p(a_1)\left[\prod_{i=2}^{T} p(a_i|a_{i-1})\right] p(b_1|a_1)\left[\prod_{i=2}^{T} p(b_i|a_i, b_{i-1})\right]\left[\prod_{i=1}^{T} p(x_i|b_i)\right]$$

2. If each variable a_i can take one of K_a states, each variable b_i can take one of K_b states, and each variable x_i can take one of K_a states:
   - **[2 points]** How many states can this set of variables take on?
     $(K_a K_b K_x)^T$, where T is the length of the chain.
   - **[2 points]** How many parameters are required to parameterize the joint distribution $p(a_1, a_2, \ldots, a_T, b_1, b_2, \ldots, b_T, x_1, x_2, \ldots, x_T)$, again assuming the factorization given by the DAG above? Note that this factorization does not imply that the factors at each time share any parameters. Also recall that for a categorical variable with K settings, only $K - 1$ parameters are required.
     Following the terms of the above factorization we get the parameter count:

$$\begin{aligned} & K_a - 1 \\ + \; & (T-1)(K_a \cdot (K_a - 1)) \\ + \; & K_a \cdot (K_b - 1) \\ + \; & (T-1)(K_a \cdot K_b \cdot (K_b - 1)) \\ + \; & T(K_b(K_x - 1)) \end{aligned}$$

3. **[1 point]** Is $x_1 \perp x_2$? Answer: No
4. **[1 point]** Is $x_1 \perp x_2 | b_1$? Answer: Yes
5. **[1 point]** Is $x_1 \perp x_2 | b_2$? Answer: Yes
6. **[1 point]** Is $a_1 \perp a_3 | a_2$? Answer: Yes
7. **[1 point]** Is $b_1 \perp b_3 | b_2$? Answer: No
8. **[1 point]** Is $b_1 \perp b_3 | a_2, b_2$? Answer: Yes

7. Markov chains and their stationary distributions - 15 pts. Consider a simple two-state Markov chain $x_0, x_1, x_2, \ldots$ with $x_t \in \{1, 2\}$ given by transition matrix

$$A = \begin{bmatrix} 2/3 & 1/3 \\ 1/2 & 1/2 \end{bmatrix}.$$

1. Find the stationary distribution $\pi = (\pi_1, 1 - \pi_1)$ of this Markov chain. The stationary distribution is given as the solution to the vector equation $A^\top \pi = \pi$.
   Answer: We solve for $A^\top \pi = \pi$. Standard linear algebra gives that $\pi = (3/5, 2/5)$.
2. Denote $p_t = \mathbb{P}(x_t = 1)$. Find the expression for p_{t+1} in terms of p_t.
   Answer: We have

$$p_{t+1} = \mathbb{P}(x_{t+1} = 1) = p_t A_{11} + (1 - p_t) A_{21} = \tfrac{1}{2} + \tfrac{1}{6} p_t$$

3. Show that p_t converges to π_1 as $t \to \infty$. You may want to use the fact that for $|q| < 1$ it holds that $\sum_{i=0}^{t-1} q^i = \frac{1-q^t}{1-q}$.
   We have

$$p_t = \tfrac{1}{2} + \tfrac{1}{6} p_{t-1} = \tfrac{1}{2}(1 + \tfrac{1}{6}) + \tfrac{1}{6^2} p_{t-2} = \cdots = \tfrac{1}{2}(1 + \tfrac{1}{6} + \cdots + \tfrac{1}{6^{t-1}}) + \tfrac{1}{6^t} p_0 = \frac{3}{5} + \tfrac{1}{6^t}(p_0 - \tfrac{3}{5}),$$

   which clearly converges to $\pi_1 = 3/5$ as $t \to \infty$.
4. Find the exact expression for the distance $|\pi_1 - p_t|$ in terms of t and p_0 to get a quantification of how quickly the Markov chain will converge to its stationary distribution.
   Answer: The calculations in 3 show that

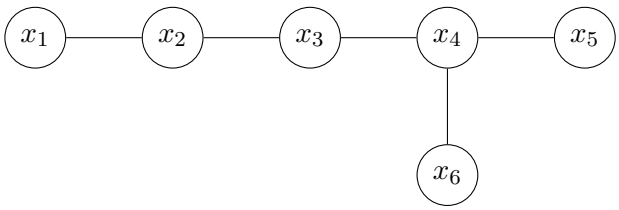
$$|\pi_1 - p_t| = \frac{1}{6^t} |p_0 - \tfrac{3}{5}| \leq \tfrac{3}{5} \tfrac{1}{6^t}.$$

   This is completely negligible already for small values of t.
5. Use the Metropolis-Hastings algorithm that uses this Markov chain to generate draws from the uniform distribution on $\{1, 2\}$.
   Answer: Using the Metropolis-Hastings construction we want to complement this auxiliary chain with a correction that assures that the limiting distribution of the so constructed Markov chain is $(1/2, 1/2)$. Suppose x_t is the current state. We first generate the proposed move x' from the Markov chain A. The acceptance probability is

$$\min\left\{1, \frac{\frac{1}{2} A_{x' x_t}}{\frac{1}{2} A_{x_t x'}}\right\}.$$

   Thus, moves $1 \to 1$, $2 \to 2$ and $1 \to 2$ are always accepted. The move $2 \to 1$ is accepted with probability $2/3$.

8. Belief propagation - 20 pts. Given the following graph of binary variables:



With x_4 being selected as root, having observed $\bar{x}_6 = 1$, and given the following potentials:

$\psi_{\text{even}}(x_i) = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ the node potential for all x_i where i is even

$\psi_{\text{odd}}(x_i) = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ The node potential for all x_i where i is odd

$\psi_{i,j}(x_i, x_j) = \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}$ for all i, j

1. (7 points) Calculate the message from 6 to 4: $m_{6\to4}(x_4)$

$$m_{6\to4} = \begin{pmatrix} \psi_6(1)\psi_{4,6}(0,1) \\ \psi_6(1)\psi_{4,6}(1,1) \end{pmatrix} = \begin{pmatrix} 3 \\ 15 \end{pmatrix}$$

2. (7 points) Given the normalized message $m_{3\to4}(x_3) = \begin{pmatrix} 0.55 \\ 0.45 \end{pmatrix}$ calculate $m_{4\to5}(x_5)$

$$\begin{aligned} m_{4\to5} &= \begin{pmatrix} \sum_{x_4} m_{3\to4}(x_4)m_{6\to4}(x_4)\psi_4(x_4)\psi_{5,4}(0,x_4) \\ \sum_{x_4} m_{3\to4}(x_4)m_{6\to4}(x_4)\psi_4(x_4)\psi_{5,4}(1,x_4) \end{pmatrix} \\ &= \begin{pmatrix} 0.55 \cdot 3 \cdot 1 \cdot 5 + 0.45 \cdot 15 \cdot 3 \cdot 1 \\ 0.55 \cdot 3 \cdot 1 \cdot 1 + 0.45 \cdot 15 \cdot 3 \cdot 5 \end{pmatrix} \\ &= \begin{pmatrix} 28.5 \\ 102.9 \end{pmatrix} \end{aligned}$$

3. (6 points) Calculate $p(x_5|\bar{x}_6)$

$$\begin{aligned} p(x_5|\bar{x}_6) &\propto \begin{pmatrix} \psi_5(0)m_{4\to5}(0) \\ \psi_5(1)m_{4\to5}(1) \end{pmatrix} \\ &= \begin{pmatrix} 4 \cdot 28.5 \\ 2 \cdot 102.9 \end{pmatrix} \\ &= \begin{pmatrix} 114 \\ 205.8 \end{pmatrix} \end{aligned}$$

$$p(x_5|\bar{x}_6) = \begin{pmatrix} 0.356 \\ 0.644 \end{pmatrix}$$

Note: In the midterm exam the numbers will be nicer and so no calculator will be needed.

9. Miscellaneous - 10 pts.

(a) (2 pts) Describe the Markov blanket of a set of variables A. Write the variables that belong to the Markov blanket of node 2 in the Figure 2.

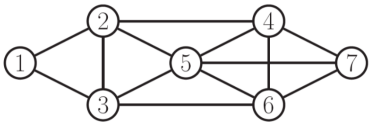


Fig 2: Simple MRF

Answer: The Markov blanket for 2 is $\{1, 3, 4, 5\}$.

(b) (2 pts) Identify all maximal and maximum cliques in the Figure 2.
Answer: The maximal cliques are: $\{1, 2, 3\}$, $\{2, 3, 5\}$, $\{2, 4, 5\}$, $\{3, 5, 6\}$, $\{4, 5, 6, 7\}$. The maximum clique is the largest maximal clique and so $\{4, 5, 6, 7\}$.

(c) (2 pts) Describe the connection between belief propagation and variable elimination on trees.
Belief propagation reduces to variable elimination algorithm on trees.

(d) (2 pts) Compare the methods Metropolis-Hasting algorithm vs rejection sampling in terms of i) the proposal densities used ii) dependencies among the samples produced.
Unlike MH, rejection sampling works only if the proposal density q(x) is similar to p(x) and it produces samples that are independent.

(e) (2 pts) In a classification problem over two classes $\mathcal{C}_1$ and $\mathcal{C}_2$, we are minimizing the misclassification error. Figure 3 shows the joint distributions. What is the decision rule that minimizes misclassification error (no derivation needed). Point at which red and blue curves

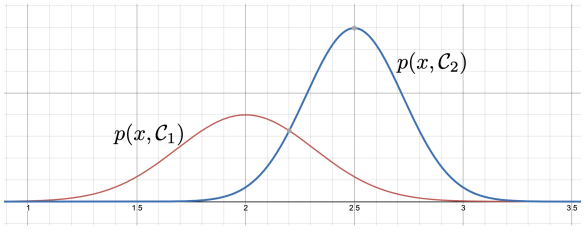


Fig 3: Decision theory

intersect.