PRACTICE MIDTERM EXAM

CSC 412/2506 - Fall 2026

University of Toronto

Exam duration: **100 minutes**

Note: To help you prepare, this practice exam has considerably more questions and is longer than the actual midterm exam. No calculators will be allowed (or needed) during the actual exam.

Read the following instructions carefully:

1. Exam is closed book. 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. The correct answer to each question is short, a few lines at most.
4. Write your answers in the space provided on the exam sheet. If you require additional paper, simply raise your hand.

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

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

1. What is the sufficient statistic of this exponential family?
   $x_1 x_2$.

2. Compute the log-partition function $A(\theta)$ and its derivative $A'(\theta)$
   (x_1, x_2) can take four possible values, for three of them $x_1 x_2 = 0$, and for one of them $x_1 x_2 = 1$. To ensure the probabilities sum to 1, it follows that

   $$A(\theta) = \log(3 + e^\theta).$$

   Thus

   $$A'(\theta) = \frac{e^\theta}{3 + e^\theta}.$$

3. Using the expression for $A'(\theta)$, verify that $\hat{\theta} = \ln(3)$ is the maximum likelihood estimate for the sample

   $$\{(1, 1), (1, 1), (1, 0), (0, 1)\}.$$

   The sample average of the sufficient statistic is $\frac{1}{4}(1+1+0+0) = \frac{1}{2}$. The maximum likelihood estimate should satisfy $A'(\hat{\theta}) = \frac{1}{2} \quad \Rightarrow \quad \frac{e^{\ln(3)}}{3+e^{\ln(3)}} = \frac{1}{2}$, which indeed holds.

4. Compute the joint distribution $p_{\hat{\theta}}(x_1, x_2)$ corresponding to this MLE.
Since $\hat{\theta} = \ln(3)$,

$$\exp\{\hat{\theta} x_1 x_2\} = 3^{x_1 x_2},$$

so the distribution is

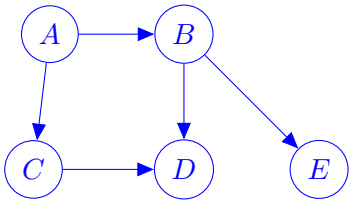
(x_1, x_2)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$p_{\hat{\theta}}(x_1, x_2)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$

.

2. Graphical models - 16 pts.

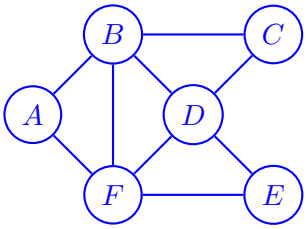
(a) (4 pts) Draw the DAG corresponding to the following factorization of a joint distribution:

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



(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.

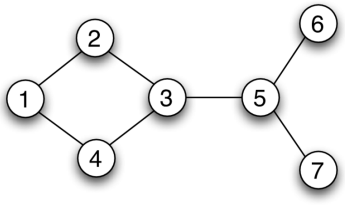


Fig 1: Simple MRF

Nodes 2,4,5.

(d) (4 pts) The 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.

3. Decision Theory - 5 pts. Imagine we are running a nuclear power plant that is undergoing a malfunction. We have two options: A) Shut it down, and B) do nothing.

Suppose that

- If we shut down, the amount of radiation that will be released is uniform between 10 and 20 units.
- If we do nothing, the amount of radiation released is 0 with probability p, and 100 with probability $1 - p$.

For what values of p would shutting down the plant release less radiation (in expectation) than doing nothing?

The expected radiation if we shut is 15, and if we do nothing it's $100(1-p)$. So $15 < 100(1-p)$ whenever $p < .85$.

4. 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. Suppose we sample N i.i.d. samples from this joint distribution, so each sample is a vector of T binary values:

$$(r_1^{(1)}, r_2^{(1)}, \ldots, r_T^{(1)}) \quad , \quad (r_1^{(2)}, r_2^{(2)}, \ldots, r_T^{(2)}), \quad \ldots \quad (r_1^{(N)}, r_2^{(N)}, \ldots, r_T^{(N)})$$

1. **[3 points]** Write a simple Monte Carlo estimator for the probability that it rains every day for the next T days. You might want to use the indicator function $\mathbb{1}(\text{E})$ which takes value 1 if E is true, and 0 if it is false.

   $\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(r_1^{(i)} = 1, r_2^{(i)} = 1, ..., r_T^{(i)} = 1)$

2. **[3 points]** Suppose that according to the model, $P(R_1 = 1, R_2 = 1, \ldots, R_T = 1 | \text{measurements}) = p$ for some number p. What is the variance of the estimator from the previous question?

   $p(1-p)/N$

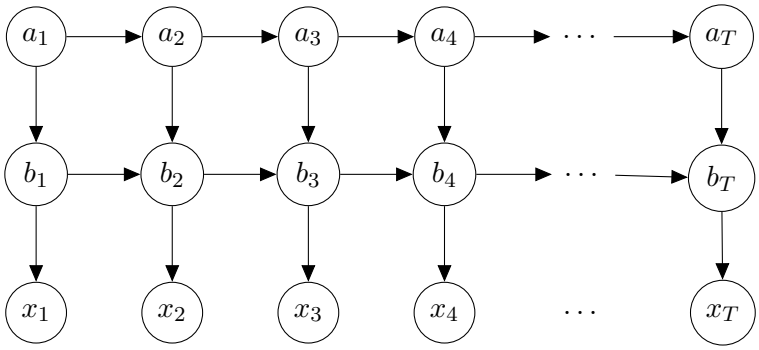
3. **[3 points]** Write a simple Monte Carlo estimator for the probability that it rains on day 3.

   $\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(r_3^{(i)} = 1)$

4. **[3 points]** Write an estimator for the probability that it rains on day 3 given that it rained on day 4 (hint: write this as the ratio between two estimators)

   $\sum_{i=1}^{N} \mathbb{1}(r_3^{(i)} = 1, r_4^{(i)} = 1) / \sum_{i=1}^{N} \mathbb{1}(r_4^{(i)} = 1)$

5. Conditional Independence - 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.

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

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

2. **[1 point]** Is $x_1 \perp x_2$? No

3. **[1 point]** Is $x_1 \perp x_2|b_1$? Yes

4. **[1 point]** Is $x_1 \perp x_2|b_2$? Yes

5. **[1 point]** Is $a_1 \perp a_3|a_2$? Yes

6. **[1 point]** Is $b_1 \perp b_3|b_2$? No

7. **[1 point]** Is $b_1 \perp b_3|a_2, b_2$? Yes

6. 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 $A^\top \pi = \pi$.
   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.

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

3. By comparing $|p_t - \pi_1|$ to $|p_{t+1} - \pi_1|$, show that $|p_t - \pi_1|$ converges to 0 as $t \to \infty$.
   Easy to check that $|p_{t+1} - \pi_1| = |\frac{1}{2} + \frac{1}{6} p_t - \frac{3}{5}| = \frac{1}{6} |p_t - \frac{3}{5}| = \frac{1}{6} |p_t - \pi_1|$. So by induction $|p_t - \pi_1| = \frac{1}{6^t} |p_0 - \frac{3}{5}|$ which clearly goes to 0 with t.

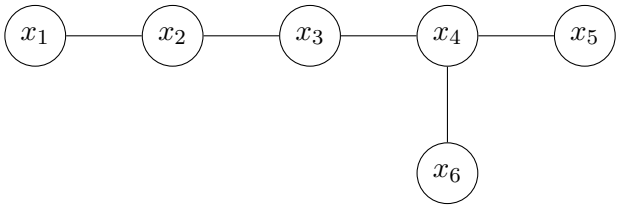
4. Use the Metropolis-Hastings algorithm that uses this Markov chain to generate draws from the uniform distribution on $\{1, 2\}$.
Suppose x_t is the current state. We first generate the proposed move x' from the Markov chain A. The acceptance probability is

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

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

7. Belief propagation - 20 pts. (**Note**: In the midterm exam the numbers will be much nicer than below, and no calculator will be needed)

Consider the following graph of binary variables (with values in $\{0, 1\}$):



Suppose we observed $\bar{x}_6 = 1$, and the potential functions are $\psi_{\text{even}}(x_i) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ for $i = 2, 4, 6$, $\psi_{\text{odd}}(x_i) = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ for $i = 1, 3, 5$, and $\psi_{i,j}(x_i, x_j) = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ for all (i, j).

1. (7 points) Calculate the message from 6 to 4: $m_{6\to4}(x_4)$.
   Since $\bar{x}_6 = 1$,
   $$m_{6\to4}(x_4) = \begin{pmatrix} \psi_6(1)\psi_{4,6}(0,1) \\ \psi_6(1)\psi_{4,6}(1,1) \end{pmatrix} = \begin{pmatrix} 2 \cdot 1 \\ 2 \cdot 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}.$$

2. (7 points) Given the normalized message $m_{3\to4}(x_4) = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$, calculate $m_{4\to5}(x_5)$.
   $$m_{4\to5}(x_5) = \sum_{x_4} m_{3\to4}(x_4) m_{6\to4}(x_4) \psi_4(x_4) \psi_{4,5}(x_4, x_5).$$
   Thus,
   $$m_{4\to5}(0) = \frac{1}{2}(2)(1)(2) + \frac{1}{2}(4)(2)(1) = 6,$$
   $$m_{4\to5}(1) = \frac{1}{2}(2)(1)(1) + \frac{1}{2}(4)(2)(2) = 9.$$
   Therefore $m_{4\to5}(x_5) = \begin{pmatrix} 6 \\ 9 \end{pmatrix}$.

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

$$p(x_5 \mid \bar{x}_6) \propto \begin{pmatrix} \psi_5(0) m_{4\to 5}(0) \\ \psi_5(1) m_{4\to 5}(1) \end{pmatrix} = \begin{pmatrix} 2 \cdot 6 \\ 1 \cdot 9 \end{pmatrix} = \begin{pmatrix} 12 \\ 9 \end{pmatrix}.$$

Normalizing, $p(x_5 \mid \bar{x}_6) = \begin{pmatrix} 12/21 \\ 9/21 \end{pmatrix} = \begin{pmatrix} 4/7 \\ 3/7 \end{pmatrix}$.

8. 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.

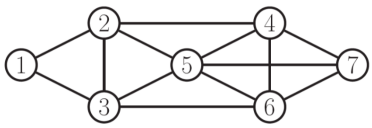


Fig 2: Simple MRF

$\{1, 3, 4, 5\}$.

(b) (2 pts) Identify all maximal cliques in the Figure 2.
$\{1, 2, 3\}$, $\{2, 3, 5\}$, $\{2, 4, 5\}$, $\{3, 5, 6\}$, $\{4, 5, 6, 7\}$.

(c) (2 pts) For trees, does belief propagation reduce to variable elimination?
Yes.

(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).

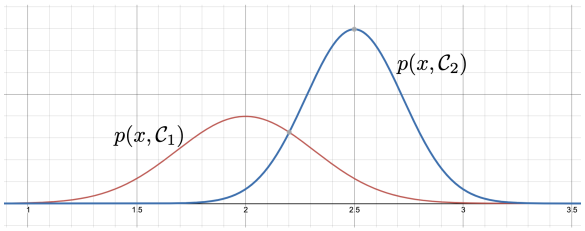


Fig 3: Decision theory

$\mathcal{C}_1$ if to the left of the point at which red and blue curves intersect. Otherwise $\mathcal{C}_2$.