PRACTICE MIDTERM EXAM

STA 414/2104 WINTER 2026
University of Toronto

Exam duration: 100 minutes
Note: The midterm will have 8 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 (8 pts)

The probability mass function of a random variable X distributed as a geometric distribution with parameter γ is given by

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

(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 $X_1, X_2, \ldots, X_n$ i.i.d. random variables from a geometric distribution with an unknown parameter γ. Find the MLE for γ.

2. Maximum likelihood estimation and unnormalised models (10 pts)

Consider a model for three binary random variables (x_1, x_2, x_3) where $x_i \in \{0, 1\}$.

$$p_\theta(x_1, x_2, x_3) \propto \exp\{\theta(x_1x_2 + x_2x_3 + x_1x_3)\}$$

1. What are the sufficient statistics of this exponential family?
2. Compute the partition function $Z(\theta)$ and the derivative of $A(\theta) = \log Z(\theta)$.
3. Verify that for the sample $\mathcal{D} = \{(1,1,1), (1,1,1), (0,1,1), (0,1,1), (1,0,1), (1,0,1)\}$ the maximum likelihood estimate is $\hat{\theta} = \ln(2)$. You will not need a calculator for this computation.
4. Compute the joint distribution $p_{\hat{\theta}}(x_1, x_2, x_3)$ corresponding to this MLE.

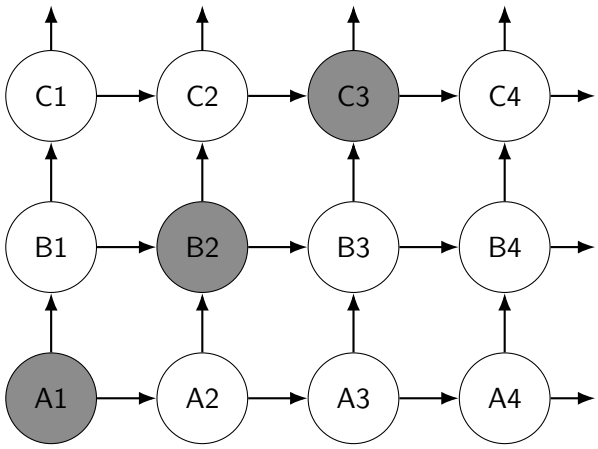
3. Graphical models (14 pts)

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) (6pts) Consider the following lattice structure with the diagonal nodes shaded. You may assume that it extends arbitrarily far upwards and also to the right. Conditioned on the shaded nodes, what are the set of all nodes independent of C_2? Justify your answer.



(c) (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 z, given that we choose a root that is not the node z?

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).

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.

$$R_{\text{vent}} \sim U(10, 20)$$

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

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

5. Simple Monte Carlo (12 pts)

Imagine we have a rain prediction model that outputs samples of

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

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:

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

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 $\mathbb{I}$(statement) which takes value 1 if the statement is true, and 0 if it is false.

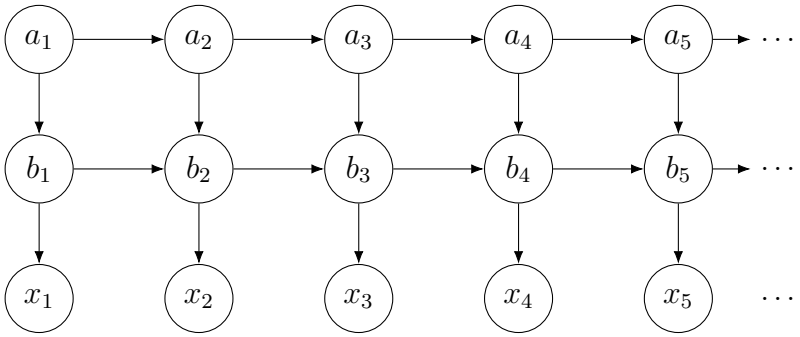
2. [3 points] What is the variance of this estimator as a function of N?

3. [3 points] Write an unbiased estimator for the probability that it rains on day 3.

4. [3 points] Write an unbiased estimator for the probability that it rains on day 3 given that it rained on day 4.

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.

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

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_x states:

   - [2 points] How many states can this set of variables take on?
   - [2 points] How many parameters are required to parameterize the joint distribution, 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.

3. [1 point] Is $x_1 \perp x_2$?

4. [1 point] Is $x_1 \perp x_2 \mid b_1$?

5. [1 point] Is $x_1 \perp x_2 \mid b_2$?

6. [1 point] Is $a_1 \perp a_3 \mid a_2$?

7. [1 point] Is $b_1 \perp b_3 \mid b_2$?

8. [1 point] Is $b_1 \perp b_3 \mid a_2, b_2$?

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

2. Denote $p_t = \mathbb{P}(x_t = 1)$. Find the expression for p_{t+1} in terms of p_t.

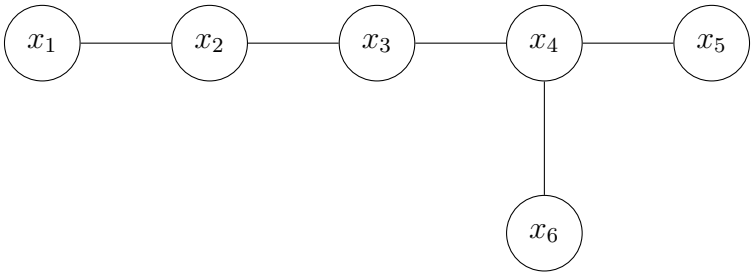
3. Show that p_t converges to π_1 as $t \to \infty$. You may want to use the fact that for $|q| < 1$ it holds that $\sum_{i=0}^{t-1} q^i = \frac{1-q^t}{1-q}$.

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.

5. Use the Metropolis-Hastings algorithm that uses this Markov chain to generate draws from the uniform distribution on $\{1, 2\}$.

8. Belief propagation (18 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:

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

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

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

1. (6 points) Calculate the message from 6 to 4: $m_{6\to4}(x_4)$.
2. (6 points) Given the normalized message $m_{3\to4}(x_3) = \binom{0.55}{0.45}$, calculate $m_{4\to5}(x_5)$.
3. (6 points) Calculate $p(x_5 \mid \bar{x}_6)$.

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

9. Miscellaneous (6 pts)

(a) (2 pts) Describe the connection between belief propagation and variable elimination on trees.

(b) (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.

(c) (2 pts) In a classification problem over two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ we are minimizing the misclassification error. Figure below shows the joint distributions. What is the decision rule that minimizes misclassification error (no derivation needed).