PRACTICE MIDTERM EXAM - SOLUTIONS

STA 414/2104 WINTER 2026
University of Toronto

Exam duration: 120 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* $0 < p < 1$, $\sum_{k=0}^{\infty} p^k = 1/(1-p)$.

Solution: First, $\mathbb{P}(X = k) > 0$. Then, we must show that $\sum_{k=1}^{\infty} \mathbb{P}(X = k) = 1$.

$$\sum_{k=1}^{\infty} \gamma(1-\gamma)^{k-1} = \gamma \sum_{j=0}^{\infty} (1-\gamma)^j$$

Let $p = 1 - \gamma$. Since $0 < \gamma < 1$, we have $0 < p < 1$. Using the Hint:

$$\gamma \cdot \frac{1}{1-(1-\gamma)} = \gamma \cdot \frac{1}{\gamma} = 1.$$

(b) Write the above distribution as an exponential family, and identify its sufficient statistics, natural parameter, and log-partition function.

Solution: We rewrite the PMF in the exponential family form:

$$\begin{aligned}\mathbb{P}(X = x) &= \exp\left\{\log(\gamma(1-\gamma)^{x-1})\right\} \\ &= \exp\left\{\log\gamma + (x-1)\log(1-\gamma)\right\}\end{aligned}$$

- **Sufficient Statistic:** $T(x) = x - 1$
- **Natural Parameter:** $\eta = \log(1-\gamma)$ (which implies $\gamma = 1 - e^{\eta}$)
- **Log-Partition Function:** $A(\eta) = -\log\gamma = -\log(1-e^{\eta})$

(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.

Solution: The log-likelihood is:

$$\ell(\gamma) = \sum_{i=1}^{n} \left(\log\gamma + (X_i - 1)\log(1-\gamma)\right) = n\log\gamma + \log(1-\gamma)\sum_{i=1}^{n}(X_i - 1)$$

Taking the derivative w.r.t γ and setting to 0:

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

$$n(1-\gamma) = \gamma(n\bar{X} - n) \implies n - n\gamma = n\gamma\bar{X} - n\gamma$$

$$n = n\gamma\bar{X} \implies \hat{\gamma} = \frac{1}{\bar{X}}$$

where $\bar{X} = n^{-1}\sum X_i$.

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?

   Solution: $T(x) = x_1x_2 + x_2x_3 + x_1x_3$.

2. Compute the partition function $Z(\theta)$ and the derivative of $A(\theta) = \log Z(\theta)$.

   Solution: We enumerate the $2^3 = 8$ states and compute $S = x_1x_2 + x_2x_3 + x_1x_3$ for each:

   - Sum=0: (0,0,0), (1,0,0), (0,1,0), (0,0,1) [4 states]
   - Sum=1: (1,1,0), (0,1,1), (1,0,1) [3 states]
   - Sum=3: (1,1,1) [1 state]

   $$Z(\theta) = 4e^{0\cdot\theta} + 3e^{1\cdot\theta} + 1e^{3\cdot\theta} = 4 + 3e^\theta + e^{3\theta}$$

   $$A'(\theta) = \frac{d}{d\theta}\log Z(\theta) = \frac{Z'(\theta)}{Z(\theta)} = \frac{3e^\theta + 3e^{3\theta}}{4 + 3e^\theta + e^{3\theta}}$$

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.

   Solution: Calculate the observed average sufficient statistic:

   - (1,1,1): $T(x) = 3$ (occurs 2 times)
   - (0,1,1): $T(x) = 1$ (occurs 2 times)
   - (1,0,1): $T(x) = 1$ (occurs 2 times)

   $$\bar{T} = \frac{2*3 + 2*1 + 2*1}{6} = \frac{10}{6} = \frac{5}{3}$$

   At MLE, we have: $A'(\hat{\theta}) = \bar{T} = \frac{5}{3}$. Plug in $\hat{\theta} = \ln(2)$ (so $e^{\hat{\theta}} = 2$):

   $$A'(\ln 2) = \frac{3*2 + 3*2^3}{4 + 3*2 + 2^3} = \frac{6 + 24}{4 + 6 + 8} = \frac{30}{18} = \frac{5}{3}$$

   The values match, so $\hat{\theta} = \ln(2)$ is correct.

4. Compute the joint distribution $p_{\hat{\theta}}(x_1, x_2, x_3)$ corresponding to this MLE.

   Solution: We calculate $Z(\ln 2) = 4 + 6 + 8 = 18$. Probabilities are $\frac{1}{18}e^{\hat{\theta}\cdot T(x)}$.

State (x_1, x_2, x_3)	**Sufficient Statistic** $T(x)$	**Unnormalized** $e^{\hat{\theta}T(x)}$	**Probability** $p_{\hat{\theta}}(x)$
(0,0,0)	0	$2^0 = 1$	1/18
(1,0,0)	0	$2^0 = 1$	1/18
(0,1,0)	0	$2^0 = 1$	1/18
(0,0,1)	0	$2^0 = 1$	1/18
(1,1,0)	1	$2^1 = 2$	$2/18 = 1/9$
(0,1,1)	1	$2^1 = 2$	$2/18 = 1/9$
(1,0,1)	1	$2^1 = 2$	$2/18 = 1/9$
(1,1,1)	3	$2^3 = 8$	$8/18 = 4/9$

Joint distribution for $\hat{\theta} = \ln 2$ (where $Z(\hat{\theta}) = 18$).

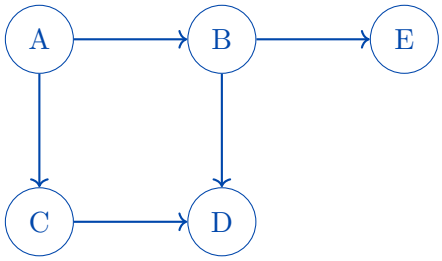
3. Graphical models (14 pts)

No explanation needed, just your answers.

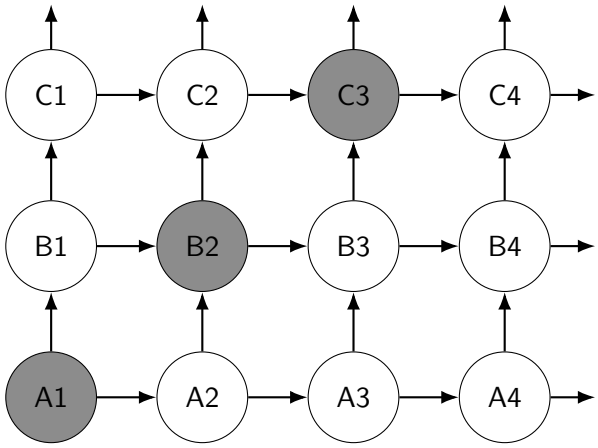
(a) (4 pts) Draw the DAG corresponding to the following factorization of a joint distribution:

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

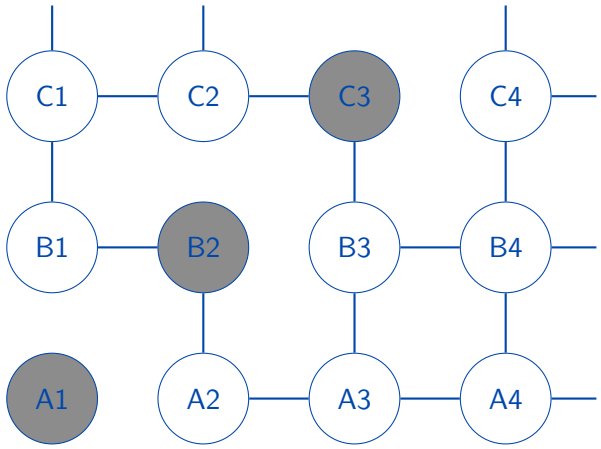
Solution:



(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.



Solution: Answer: The empty set (no nodes).
Justification: Applying the pruning algorithm, we find that there are no leaf node in this graph (all the nodes have two outgoing edges). Deleting the edges from the conditioning nodes and removing the arrows direction, we obtain the following graph



Any node in the upper diagonal region remains connected to C_2, as no edges are removed in that section. Similarly, in the lower diagonal region, all nodes connect to A_2, which in turn connects to C_2. Therefore, the set of nodes independent of C_2 is empty.

(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?

  Solution: 1 pass (from leaves to 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?

  Solution: 2 passes (leaves to root, then root to leaves).

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?

Solution:

- Expected radiation if Venting: $\mathbb{E}[R_{\text{vent}}] = \frac{10+20}{2} = 15$.
- Expected radiation if Nothing: $\mathbb{E}[R_{\text{nothing}}] = X * 0 + (1 - X) * 100 = 100(1 - X)$.

We vent if $15 < 100(1 - X)$:

$$0.15 < 1 - X \implies X < 0.85$$

Answer: $X < 85\%$.

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 $R_i \in \{0, 1\}$ is the predicted occurence 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.

   Solution:

   $$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(r_1^{(i)} = 1, r_2^{(i)} = 1, \ldots, r_T^{(i)} = 1)$$

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

   Solution: If $p = P(R_1 = 1, R_2 = 1, \ldots, R_T = 1 \mid \text{measurements})$ is the true probability, the variance is $\frac{p(1-p)}{N}$.

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

   Solution:

   $$\hat{p}_{day3} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(r_3^{(i)} = 1)$$

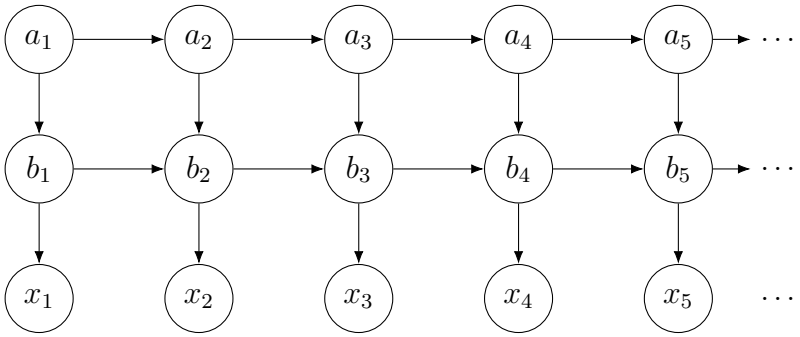
4. [3 points] Write an unbiased estimator for the probability that it rains on day 3 given that it rained on day 4.

   Solution:

   $$\hat{p}_{3|4} = \frac{\sum_{i=1}^{N} \mathbb{I}(r_3^{(i)} = 1, r_4^{(i)} = 1)}{\sum_{i=1}^{N} \mathbb{I}(r_4^{(i)} = 1)}$$

6. HMM Question (12 pts)

Given the following DAG:



1. [2 points] Write the factorized joint distribution implied by this DAG.

   Solution:

   $$p(\mathbf{a}, \mathbf{b}, \mathbf{x}) = p(a_1)\left[\prod_{t=2}^{T} p(a_t|a_{t-1})\right] p(b_1|a_1)\left[\prod_{t=2}^{T} p(b_t|a_t, b_{t-1})\right]\left[\prod_{t=1}^{T} p(x_t|b_t)\right]$$

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_x states:

   - [2 points] How many states can this set of variables take on?

     Solution: $(K_a K_b K_x)^T$, where T is the length of the chain

   - [2 points] How many parameters are required to parameterize the joint distribution?

     Solution:

     - a_1: $K_a - 1$
     - $a_t|a_{t-1}$: $(T-1) \times K_a(K_a - 1)$
     - $b_1|a_1$: $K_a(K_b - 1)$
     - $b_t|a_t, b_{t-1}$: $(T-1) \times K_a K_b(K_b - 1)$
     - $x_t|b_t$: $T \times K_b(K_x - 1)$

     This gives

     $$(K_a-1)+(T-1)\times K_a(K_a-1)+K_a(K_b-1)+(T-1)\times K_a K_b(K_b-1)+T\times K_b(K_x-1)$$

3.
- Is $x_1 \perp x_2$?
  Solution: No.
- Is $x_1 \perp x_2 \mid b_1$?
  Solution: Yes.
- Is $x_1 \perp x_2 \mid b_2$?
  Solution: Yes.
- Is $a_1 \perp a_3 \mid a_2$?
  Solution: Yes.
- Is $b_1 \perp b_3 \mid b_2$?
  Solution: No.
- Is $b_1 \perp b_3 \mid a_2, b_2$?
  Solution: 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 $\pi = (\pi_1, 1 - \pi_1)$.

   Solution: Solve $\pi A = \pi$. $\pi_1 = \frac{2}{3}\pi_1 + \frac{1}{2}(1 - \pi_1) \implies \pi_1 = \frac{2}{3}\pi_1 + \frac{1}{2} - \frac{1}{2}\pi_1 \implies (1 - \frac{2}{3} + \frac{1}{2})\pi_1 = \frac{1}{2} \implies \frac{5}{6}\pi_1 = \frac{1}{2} \implies \pi_1 = \frac{3}{5}$. Answer: $\pi = (3/5, 2/5)$.

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

   Solution: $p_{t+1} = p_t(2/3) + (1 - p_t)(1/2) = \frac{2}{3}p_t + \frac{1}{2} - \frac{1}{2}p_t = \frac{1}{2} + \frac{1}{6}p_t$.

3. Show that p_t converges to π_1 as $t \to \infty$.

   Solution: This is a linear recurrence relation giving $p_t = \frac{1}{2}\sum_{i=0}^{t-1}(1/6)^i + (1/6)^t p_0$. The geometric sum converges to $\frac{1}{1-1/6} = \frac{6}{5}$. Limit: $\frac{1}{2}(\frac{6}{5}) = \frac{3}{5} = \pi_1$.

4. Find the exact expression for the distance $|\pi_1 - p_t|$.

   Solution: Since $p_t = \frac{1}{2}\frac{1-(1/6)^5}{1-1/6} + (1/6)^t p_0 = \frac{3}{5} + (1/6)^t(p_0 - 3/5)$ and $\pi_1 = 3/5$, we can write that $|\pi_1 - p_t| = \left(\frac{1}{6}\right)^t |p_0 - 3/5| = \left(\frac{1}{6}\right)^t \left|\frac{3}{5} - p_0\right| \leqslant \left(\frac{1}{6}\right)^t \frac{3}{5}$.

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

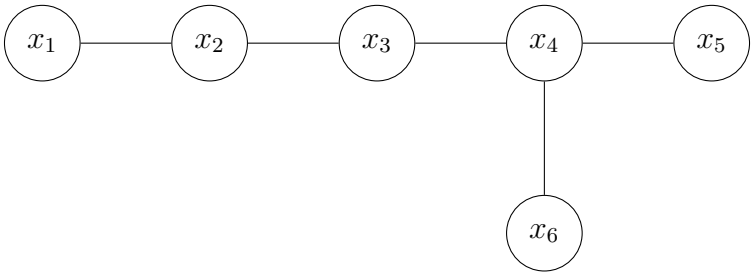
   Solution: Using the Metropolis-Hastings construction we want to complement this auxiliary chain with a correction that ensures that the limiting distribution of the constructed Markov chain is $\pi^* = (1/2, 1/2)$. Suppose x_t is the current state. We first generate the proposed move x' from the Markov chain A. The acceptance probability is

   $$\alpha = \min\left(1, \frac{\pi^*(x')q(x|x')}{\pi^*(x)q(x'|x)}\right) = \min\left(1, \frac{A_{x'x}}{A_{xx'}}\right).$$

   For $2 \to 1$: $\alpha = \min(1, \frac{1/3}{1/2}) = 2/3$. All other moves accepted with probability 1.

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: $m_{6\to4}(x_4)$.

   Solution:

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

2. (6 points) Given $m_{3\to4}(x_3) = \binom{0.55}{0.45}$, calculate $m_{4\to5}(x_5)$.

   Solution:

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

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

Solution:

$$
\begin{aligned}
p(x_5 \mid \bar{x}_6) &\propto \begin{pmatrix} \psi_5(0) m_{4\to 5}(0) \\ \psi_5(1) m_{4\to 5}(1) \end{pmatrix} \\
&= \begin{pmatrix} 4 \cdot 28.5 \\ 2 \cdot 102.9 \end{pmatrix} \\
&= \begin{pmatrix} 114 \\ 205.8 \end{pmatrix}
\end{aligned}
$$

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

9. Miscellaneous (6 pts)

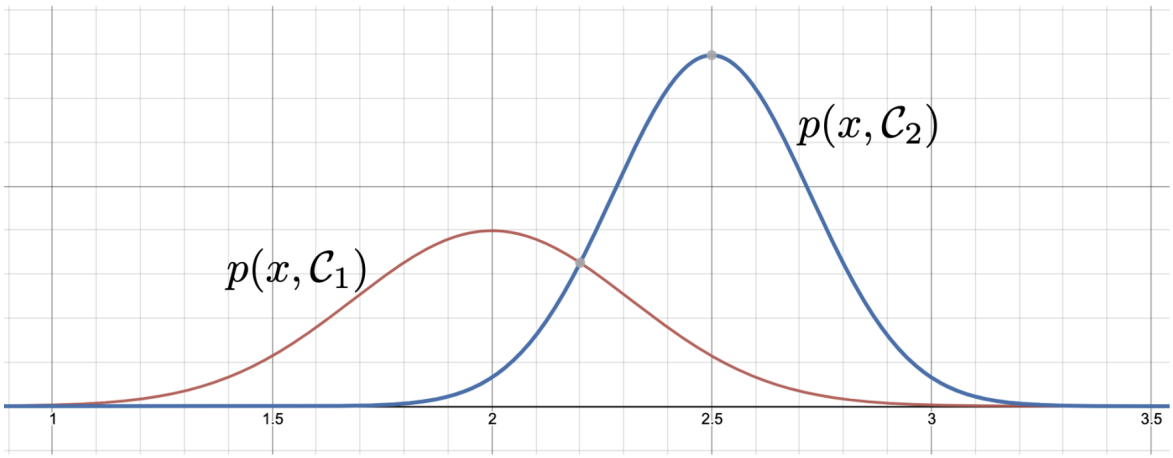
(a) (2 pts) Describe the connection between belief propagation and variable elimination on trees.

Solution: Belief propagation reduces to the variable elimination algorithm on trees. Specifically, the process of passing messages from the leaves to a root node in belief propagation is mathematically equivalent to eliminating variables one by one (summing them out) to compute the marginal distribution at that root.

(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.

Solution: Rejection sampling produces independent samples but requires a proposal distribution $q(x)$ that upper bounds $p(x)$ everywhere up to a multiplicative factor ($Mq(x) \geqslant p(x)$). Metropolis-Hastings produces correlated (dependent) samples (Markov Chain) but only requires $q(x'|x)$ to be defined and does not require an enveloping constant M.

(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).



Solution: The decision boundary should be placed at the point where the probability densities of the classes intersect ($p(x, \mathcal{C}_1) = p(x, \mathcal{C}_2)$). On the left of this point, which is $x \leqslant 2.2$, we select $\mathcal{C}_1$ ($\mathcal{C}_2$ otherwise).