2/3 hrs	10 / Nov / 2023	Venue	EXAMS OFFICE USE ONLY

University of the Witwatersrand, Johannesburg

Course or topic No(s)	COMS4047A/7053A
Course or topic name(s) Paper Number & title	Probabilistic Graphical Models
Examination to be held during the month(s) of	November 2023
Year of study	1
Degrees/Diplomas for which this course is prescribed	BScHons (CS / BDA / CAM), MSc (AI / DS / CS / Robotics/ e-Science)
Faculties presenting candidates	Science
Internal examiner(s)	Prof. Ritesh Ajoodha x-76188
External examiner(s)	Prof. External Name (Ext Univ)
Special materials	Formula sheet and non-programmable calculator permitted
Time allowance	COMS7053A: 3 hours COMS4047A: 2 hours
Instructions to candidates	COMS7053A students must answer 4 of 5 questions for a total of 160 marks (100%). COMS4047A students can choose 3 out of 5 questions, worth 120 marks (100%). If a student answers multiple questions, then only the first three questions will be marked. This is a closed book exam with 30 pages.

Question 1 Particle-based Approximate Inference [40 Marks]

1.1. For each of the following MCQ questions, circle the correct answer label.

1.1.1. What is the main limitation of particle-based inference? [1]

(a) It only works on small datasets

(b) It cannot handle nonlinear relationships between variables.

(c) The estimate quality depends on the number of particles.

(d) It is only applicable to simple models with few variables

1.1.2. Which of the following statements is true regarding forward sampling? [1]

(a) Forward sampling is a process of generating samples from a given probability distribution.

(b) Forward sampling is a technique used to estimate counterfactual outcomes through random sampling.

(c) Forward sampling is a statistical method for sampling from imbalanced datasets.

(d) Forward sampling is a technique used to measure the representativeness of a sample in a survey.

1.1.3. In particle-based methods, the number of samples used has an impact on: [1]

(a) The accuracy of the approximation.

(b) The computational complexity of the algorithm.

(c) The convergence rate of the estimation.

(d) The dimensionality of the problem space.

1.1.4. Gibbs sampling is primarily used for: [1]

(a) Sampling from high-dimensional probability distributions.

(b) Estimating causal effects in observational studies.

(c) Training deep neural networks

(d) Solving optimization problems

1.1.5. Markov Chain Monte Carlo (MCMC) methods are commonly used for: [1]

(a) Sampling from complex probability distributions.

(b) Solving linear systems of equations

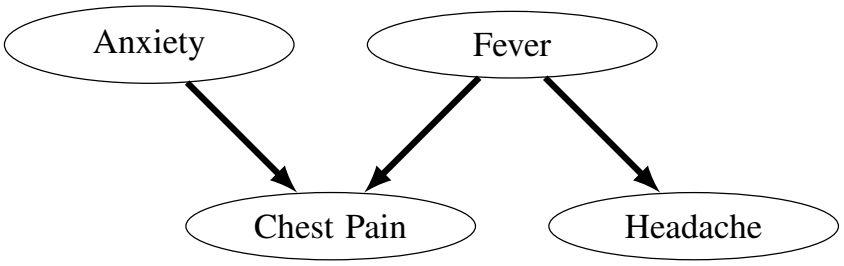
(c) Optimizing objective functions

(d) Conducting hypothesis testing

1.2 Use the below Bayesian network which models the relationships between anxiety, fever, chest pain, and headache to answer to answer the following questions.

Anxiety	
a^0	a^1
0.6	0.4

Fever	
f^0	f^1
0.7	0.3



	Chest Pain		
a,f	c^0	c^1	c^2
a^0, f^0	0.3	0.4	0.3
a^0, f^1	0.05	0.25	0.7
a^1, f^0	0.9	0.08	0.02
a^1, f^1	0.5	0.3	0.2

	Headache	
f	h^0	h^1
f^0	0.95	0.05
f^1	0.2	0.8

Figure 1: A Bayesian network, $\mathcal{B}$, for medical diagnostics.

(a) Suppose you would like to sample 100 instances of the distribution $P(a^0, f^1 \mid c^0, h^1)$ from $\mathcal{B}$. Explain how you could do this using rejection sampling. [2]

(b) Suppose that you would like to use likelihood weightings to to estimate $P(a^0, f^1 \mid c^0, h^1)$, and you have generated the following 10 samples for this purpose using inference given the evidence. Next to each particle below write down the likelihood weighting for that particle given the evidence. [2]

$\xi[1] = \{a^0, f^1, c^0, h^1\}$ ----------
$\xi[2] = \{a^1, f^0, c^0, h^1\}$ ----------
$\xi[3] = \{a^0, f^0, c^0, h^1\}$ ----------
$\xi[4] = \{a^1, f^1, c^0, h^1\}$ ----------
$\xi[5] = \{a^0, f^0, c^0, h^1\}$ ----------

$\xi[6] = \{a^1, f^0, c^0, h^1\}$ _____
$\xi[7] = \{a^0, f^1, c^0, h^1\}$ _____
$\xi[8] = \{a^1, f^1, c^0, h^1\}$ _____
$\xi[9] = \{a^0, f^0, c^0, h^1\}$ _____
$\xi[10] = \{a^1, f^1, c^0, h^1\}$ _____

(c) Finally, calculate the empirical probability $\hat{P}_{\mathcal{D}}(a^0, f^1 \mid c^0, h^1)$. Show your working and round off your answer two decimal places. [3]

1.3 Use the following Markov chain below which describes the distribution between 3 states x^1, x^2, and x^3, where x^3 is the initial state, to answer the following questions.

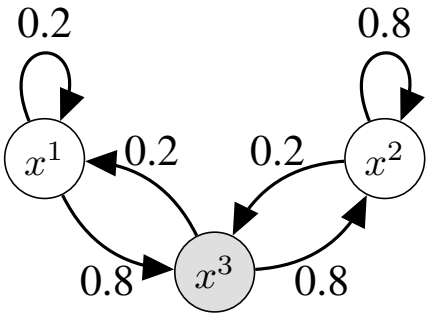


Figure 2: A Markov chain with 3 states.

(a) In the below table, complete the next 3 iterations of the Markov Chain Monte Carlo (MCMC) algorithm. The first iteration has been completed for you. Round of your answers to two decimal places. [10]

Iteration	x_1	x_3	x_2
1	0	1	0
2			
3			
4			

(b) Write the system of linear equations that governs the stationary distribution for the given system. [4]

(c) Use Gauss-Jordan elimination to calculate the stationary distribution: $\pi(x^1)$, $\pi(x^2)$, and $\pi(x^3)$. [10]

(d) Demonstrate that your stationary distribution provided in the previous question satisfies detailed balance equation between x^2 and x^3. [2]

1.4. Describe one way to determine if a Markov chain has mixed. [2]

Question 2 Parameter Estimation [40 Marks]

2.1. For each of the following MCQ questions, circle the correct answer label.

2.1.1. Which of the following statements about maximum likelihood estimation (MLE) is correct? [1]

(a) It aims to minimize the sum of squared errors between the observed data and the predicted values.

(b) It assumes that the errors are normally distributed.

(c) It guarantees unbiased estimates of the parameters.

(d) It maximizes the likelihood function to find the most likely parameter values for the observed data.

2.1.2. Bayesian estimation is based on: [1]

(a) Prior knowledge or beliefs about the parameters.

(b) Maximizing the likelihood of the observed data.

(c) Minimizing the sum of squared errors.

(d) Randomly sampling from the parameter space.

2.1.3. Which of the following is a limitation of the Expectation-Maximization (EM) algorithm? [1]

(a) The algorithm guarantees convergence to the global optimum.

(b) It can get stuck in local optima, leading to suboptimal solutions.

(c) The EM algorithm is only applicable to linear models.

(d) It requires prior knowledge of the true underlying probability distribution.

2.1.4. Both the K-means clustering algorithm and the Expectation-Maximization (EM) algorithm: [1]

(a) Are iterative algorithms used for unsupervised learning tasks.

(b) Involve the estimation of model parameters based on observed data.

(c) Utilize the concept of maximizing the likelihood of the data.

(d) Rely on an initial assignment of data points to clusters.

2.1.5. The convergence of the Expectation-Maximization (EM) algorithm is typically determined by: [1]

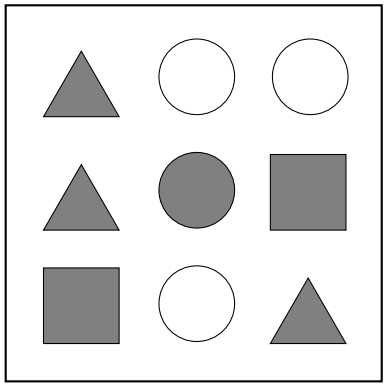
(a) Monitoring the change in the log-likelihood function between consecutive iterations.

(b) Setting a fixed number of iterations in advance.

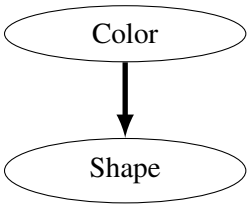
(c) Checking if the estimated parameters match the true underlying parameter values.

(d) Evaluating the convergence of the latent variable assignments.

2.2. Use the following dataset of shapes and colours to answer the following questions. The dataset contains three different shapes: circles (C), triangles (T), and squares (S). Each shape can take on two possible colours: gray (G), or white (W). (Note that "S" stands for square, not for shape.)



(a) A dataset consisting of shapes and colours.



(b) The Bayesian network.

Figure 3: A data set (a) and Bayesian network (b) which is related to different shapes of colours.

(a) Compute the maximum likelihood estimate: $P(C)$, $P(S \mid G)$, $P(C \mid W)$, $P(T \mid G)$, and $P(S)$. Show your working and round off your answer two decimal places. [5]

(b) Using a the Dirichlet prior, where $\alpha = 2$, compute the Bayesian estimate $P(C)$, $P(S \mid G)$, $P(C \mid W)$, $P(T \mid G)$, and $P(S)$. Show your working and round off your answer two decimal places. [5]

(c) Using the answers you provided above for $P(C \mid W)$, compare the estimates under the maximum likelihood estimation (MLE) and Bayesian estimation (BE) paradigms. Is there a difference? If so, can you justify why there is a difference? [3]

2.3. What is the purpose of the expectation maximisation (EM) algorithm? Explain the difference between the E-step and M-step in the EM algorithm. [3]

2.4. Use the below Bayesian network and the following incomplete data instances to answer the following questions: $\mathcal{D} = \{\langle a^1, ?, ?, d^0\rangle, \langle ?, b^1, ?, d^1\rangle\}$. You may assume that $P(a^1, d^0) = 0.12$ and $P(b^1, d^1) = 0.04$.

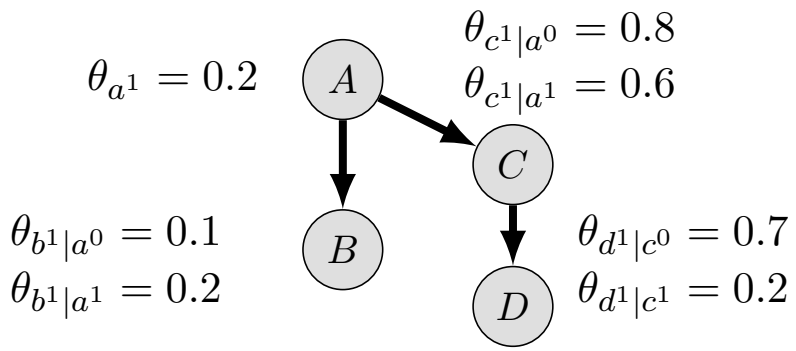


Figure 4: A Bayesian network that models four variables with full parameterization.

(a) Suppose that you are in the E-Step of the EM algorithm. For the data set, $\mathcal{D} = \{\langle a^1, ?, ?, d^0\rangle, \langle ?, b^1, ?, d^1\rangle\}$, calculate the likelihood of each possible completion of the data given the parameters. [10]

(b) Use your analysis from above, what is the most likely assignment for the dataset $\mathcal{D} = \{\langle a^1, ?, ?, d^0\rangle, \langle ?, b^1, ?, d^1\rangle\}$. [2]

(c) Now suppose that you are in the M-Step with $\mathcal{D}^+$ which contains the augmented data set with the likelihood weightings. Compute the sufficient statistics $\bar{M}_\theta[d^1, c^0]$ and $\bar{M}_\theta[c^0]$. Round off your answer two decimal places. [4]

(d) Finally, compute $\tilde{\theta}_{d^1|c^0}$ using the expected sufficient statistics with soft assignments from the previous question. Round off your answer two decimal places. [3]

Question 3 Structure Learning [40 Marks]

3.1. For each of the following MCQ questions, circle the correct answer label.

3.1.1. In structure learning, which of the following approaches is commonly used to estimate the dependencies between variables in a dataset? [1]

(a) Regression analysis

(b) Principal Component Analysis (PCA)

(c) Clustering algorithms

(d) Conditional Independence tests

(e) All of the options above.

3.1.2. Structure learning in knowledge discovery refers to: [1]

(a) Identifying the most relevant features in a dataset.

(b) Discovering the underlying causal relationships between variables.

(c) Evaluating the performance of a machine learning model.

(d) Generating synthetic data for training purposes.

3.1.3. Score-based approaches in structure learning for involve: [1]

(a) Assigning a numerical score to each data point in a dataset.

(b) Evaluating the performance of a machine learning model using a scoring metric.

(c) Selecting the model structure that maximizes a scoring function based on the observed data.

(d) Iteratively refining the scoring function based on the model structure.

3.1.4. Bayesian model averaging is a technique used for: [1]

(a) Combine the predictions of multiple Bayesian networks.

(b) Assess the uncertainty in model parameter estimates.

(c) Select the best model based on a predefined criterion.

(d) Estimate the probability distribution of the target variable.

3.1.5. The complexity of structure learning in graphical models typically depends on: [1]

(a) The number of variables in the model.

(b) The size of the dataset used for learning.

(c) The search space of possible model structures.

(d) The algorithm used for structure learning.

3.2. What is structure learning in machine learning, and why is it important? [3]

3.3. What does the faithfulness assumption imply in the context of structure learning? [3]

3.4. Consider three variables, A, B, and C, and the given set of samples in Table 1. You aim to learn the structure of the Bayesian network using a constraint-based approach. In this approach, you assume the null hypothesis (H_0) as $P^*(X, Y) = \hat{P}(X)\hat{P}(Y)$.

Using empirical mutual information, denoted as $d_{\mathbb{I}}(\mathcal{D})$, for independence tests, a p-value$(t) = 0.01$, and the ordering A, B, and C, determine the conditional independence assumptions between the variables. Show all working out. [15]

$\mathcal{D}$	$\langle A, B, C \rangle$
$\xi[1]$	$\langle a^1, b^0, c^1 \rangle$
$\xi[2]$	$\langle a^0, b^1, c^0 \rangle$
$\xi[3]$	$\langle a^1, b^1, c^1 \rangle$
$\xi[4]$	$\langle a^1, b^1, c^0 \rangle$
$\xi[5]$	$\langle a^0, b^0, c^0 \rangle$

Table 1: A set of instances for variables A, B, and C.

3.5. Suppose you are given the following set of shapes as depicted in Figure 5. Each shape can either be a square or a circle. Calculate the marginal likelihood, $P(x[1], \ldots, x[6])$, of the set of shapes with a Dirichlet$[\alpha_1, \alpha_2]$ prior, where $\alpha_1 = \alpha_2 = 1$. Round off your answer two decimal places. [6]

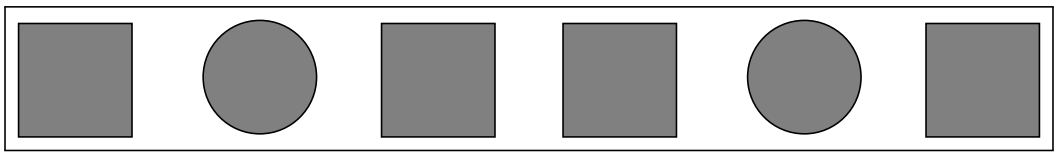


Figure 5: A dataset of shapes.

3.6. What is the Bayesian Information Criterion (BIC) score? Describe the intuition of how it works. [3]

3.7. Suppose you are using a score-based approach for structure learning, and your algorithm consistently gets stuck at local optima. Describe two techniques that can help you overcome this problem. [5]

Question 4 Causal Graphical Models [40 Marks]

4.1. For each of the following MCQ questions, circle the correct answer label.

4.1.1. Which of the following statements about causal models is correct? [1]

(a) They focus on correlational relationships rather than causal relationships.
(b) They assume that all variables are independent of each other.
(c) They can not be used to predict future events based on historical data.
(d) They aim to identify the underlying mechanisms that generate observed data.
(e) All of the options above.

4.1.2. Which of the following statements is true regarding Bayesian networks? [1]

(a) Bayesian networks with a causal structure tend to be denser and less natural.
(b) Bayesian networks with a causal structure tend to be sparser and more natural.
(c) Bayesian networks do not have any impact on sparsity or naturalness.
(d) The sparsity and naturalness of Bayesian networks depend on the size of the network.
(e) All of the options above.

4.1.3. Which of the following statements is not a causal query? [1]

(a) What is the effect of smoking on the likelihood of developing lung cancer?
(b) How does increasing education level impact job prospects?
(c) What is the average temperature in Johannesburg during the month of July?
(d) What is the influence of student study hours on the performance of students' in a course?
(e) How does parental involvement affect student academic performance?

4.1.4. Which of the following statements is true regarding causal identifiability? [1]

(a) Causal identifiability refers to the ability to determine cause-and-effect relationships between variables.
(b) Causal identifiability is not relevant in research studies.
(c) Causal identifiability only applies to observational studies, not experimental studies.
(d) Causal identifiability is determined solely by the sample size of a study.

4.1.5. Which of the following statements is true regarding Twinned Counterfactual Network (TCN)? [1]

(a) TCN is a statistical method used for estimating causal effects in observational studies.
(b) TCN is a type of deep learning architecture used for causal inference in image recognition tasks.
(c) TCN is a network topology commonly used in computer networks for modeling causal relationships.
(d) TCN is a programming language specifically designed for building causal web applications.

4.2. Suppose you are studying the impact of exercise on weight loss. You collect data from 200 individuals who participated in a weight loss program. The dataset includes the following variables:

- Exercise hours per week (X): Ranging from 0 to 10 hours
- Weight loss (Y): Measured in kilograms (kg)
- Diet plan (Z): 1 if the individual followed a specific diet plan, 0 otherwise.

Use this information to answer the following questions:

(a) Assume that 50 individuals exercised for 6 hours per week, and 30 of them lost more than 5kg. What is the probability that a randomly selected individual who exercised for 6 hours per week lost more than 5kg? [2]

(b) Assume that out of the 200 individuals, 70 exercised for 8 hours per week without following the diet plan, and their average weight loss was 7kg. Additionally, 60 individuals exercised for 8 hours per week and followed the diet plan, and their average weight loss was 10kg. What is the causal effect of following the diet plan on weight loss for individuals who exercised for 8 hours per week? [2]

4.3 Non-causal correlation is a statistical relationship between two variables not driven by direct cause-and-effect. Instead, various factors can induce non-causal correlations, leading to spurious or coincidental associations. Describe a factor that can contribute to non-causal correlations between variables? [3]

4.4 Consider the causal model is illustrated by Figure 6 which illustrates the causal effect of mental health on academic achievement. Use the causal model to answer the questions that follow.

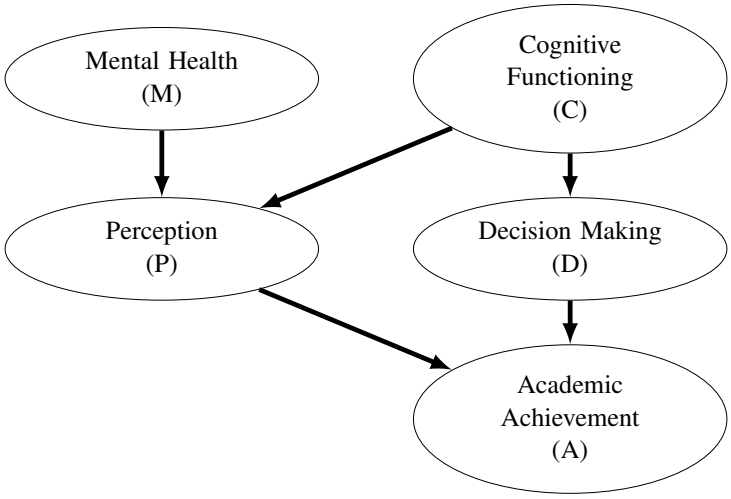


Figure 6: A causal model, $\mathcal{C}$, which illustrates the causal effect of mental health on academic achievement.

(a) A study examines the relationship between undergraduate students' perceived academic achievement and decision-making skills in two majors, A and B. Two groups of 200 students each, Group A and Group B, participate in the study. Perceived academic achievement is rated on a 1-10 scale, while decision-making skills are measured on a 1-100 scale. The results are as follows:

Group A (Major A):

- Perceived academic achievement score ≤ 6: Avg. decision-making skill score = 65 (n = 100)
- Perceived academic achievement score ≥ 7: Avg. decision-making skill score = 60 (n = 100)
- Overall average decision-making skill score: 62.5

Group B (Major B):

- Perceived academic achievement score ≤ 6: Avg. decision-making skill score = 55 (n = 150)
- Perceived academic achievement score ≥ 7: Avg. decision-making skill score = 70 (n = 50)
- Overall average decision-making skill score: 57.5

Compute the overall average decision-making skill score if we combine both groups and analyse the data without considering the perceived academic achievement as a factor. [4]

(b) Is Simpson's paradox is present in this scenario? Explain your answer. [2]

(c) Using the causal model $\mathcal{C}$, draw the mutilated causal network for $\mathcal{C}_{\mathbf{P}=\mathbf{p}}$. [4]

(d) Suppose that you have the following intervention query:

$$P(A \mid do(P := p), D, C, M)$$

Simplify the interventional query using the graph-dependent intervention rule. [2]

(e) Using the causal model $\mathcal{C}$, draw the augmented causal model to answer the following intervention query:

$$P(A \mid do(P := p), do(C := c), M, D)$$

. [4]

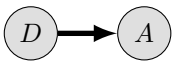
(f) Simplify the following interventional query using the information conservation rule.

$$P(A \mid do(P := p), do(C := c), M, D)$$

. [2]

(g) From the provided causal model C above, if there is a confounding variable influencing cognitive functioning (C) and academic achievement (A) when is $P(A \mid do(C := c))$ identifiable? [1]

4.5 Suppose a set of samples was collected to study evaluating the effectiveness of decision making (D) on academic achievement (A) of participants as indicated by $\mathcal{C}$. Suppose that we only use these two variables to generate interventional data, then the the samples are shown in Table 2. Use this information to answer the following questions.



Intervention	d^1, a^1	d^1, a^0	d^0, a^1	d^0, a^0
$\emptyset$	5	2	6	2
$do(D := d^1)$	2	4	0	0
$do(A := a^1)$	3	0	7	0

Table 2: Intervention Data for evaluating the effectiveness of decision making (D) on academic achievement (A) of participants.

(a) Calculate the causal sufficient statistics $M[d^1]$, $M[d^0]$, $M[a^1, d^1]$, $M[a^0, d^1]$, $M[a^1, d^0]$, and $M[a^0, d^0]$. [5]

(b) Calculate θ_{d^1}, $\theta_{a^1|d^1}$, and finally, $\theta_{a^1|d^0}$. Round off two decimal places. [4]

Question 5 Structured Decision Problems [40 Marks]

5.1. For each of the following MCQ questions, circle the correct answer label.

5.1.1. In decision theory, lotteries refer to: [1]

(a) Random drawings for prizes or rewards.
(b) Decision-making under uncertainty using probabilistic outcomes.
(c) Methods for selecting a course of action based on expected utility.
(d) Techniques for maximizing financial gains in games of chance.

5.1.2. Influence diagrams are graphical representations used for: [1]

(a) Analyzing the impact of social media on consumer behavior.
(b) Modeling the flow of information in computer networks.
(c) Visualizing decision-making problems under uncertainty.
(d) Studying the influence of genes on human traits.

5.1.3. In an influence diagram, decision nodes represent: [1]

(a) Uncertain events or outcomes.
(b) Actions or choices available to the decision-maker.
(c) Probability distributions of random variables.
(d) Utility functions representing the decision-maker's preferences.

5.1.4. A decision rule is a function that: [1]

(a) Assigns probabilities to possible outcomes.
(b) Determines the optimal action to take based on available information.
(c) Measures the uncertainty of a decision.
(d) Evaluates the performance of a decision-making model.

5.1.5. The Value of Perfect Information (VPI) represents: [1]

(a) The additional cost of acquiring perfect information for decision-making.
(b) The monetary value associated with having access to all available information.
(c) The potential improvement in decision outcomes with perfect information.
(d) The cost of uncertainty and risk in the decision-making process.

5.2. You are playing a game where you roll a fair six-sided die. If the result is an even number (2, 4, or 6), you win R10. If the result is an odd number (1, 3, or 5), you lose R5. Calculate the expected payoff for this game. [2]

5.3. Suppose you have a utility function defined over two goods, books (B) and hamburgers (H). Your utility function is given by

$$U(B, H) = 3 \times B^{0.6} \times H^{0.4}.$$

Use this information to answer the following questions.

(a) Suppose that you currently have 5 books and 4 hamburgers. You are offered a trade to give up 2 books in exchange for 3 hamburgers. Should you accept the trade? Justify your answer based on the change in utility. [3]

(b) Determine whether the utility function exhibits constant, increasing, or decreasing marginal utility for each good (books and hamburgers). [4]

5.4. Imagine you have an influence diagram, $\mathcal{I}$, that shows the connection between the weather and the traffic on the road when you are heading to work. This diagram is designed to

assist you in making a decision about whether you should work in the office or work online. Use this influence diagram to answer the questions that follow.

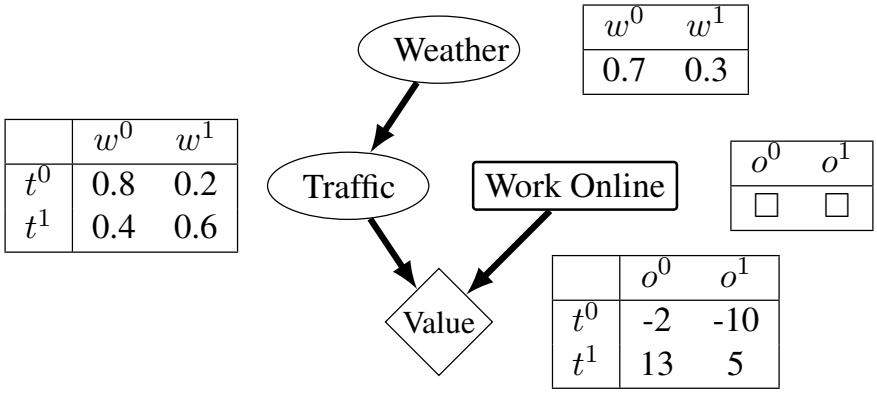


Figure 7: An influence diagram to help you decide to work from home.

(a) In decision theory, describe the concept of maximum expected utility. [3]

(b) Compute the expected utility associated with making a decision regarding whether to work online, denoted as $\mathrm{EU}[D[\delta_O]]$. Show your working out. [8]

(c) Which action will maximise $\mathrm{EU}[D[\delta_O]]$? Justify your answer. [2]

(d) Now suppose that we added an edge from the chance variable "Weather" to the action variable "Work Online". What is the expected utility with this perfect information, denoted $\mathrm{EU}[D_{X \to A}[\delta_O]]$? Show all of your working. [10]

(e) What is the best strategy to maximise the overall utility. Indicate the best strategy and the utility which it yields. [1]

(f) Calculate the value of perfect information of including the edge between the chance variable W and the decision variable O? [2]

END OF EXAM

Working out

Working out

Probabilistic Graphical Models Formula Sheet

Probability Theory

Chain Rule for Probabilities:
$P(X_1, \ldots, X_n) = P(X_1) \cdots P(X_n | X_1, X_{n-1})$

Bayes Rule:
$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)}{P(\alpha)P(\beta)}$

Probability Density Function: $p : \mathbb{R} \to \mathbb{R}$ is a probability density function (PDF) for $\mathcal{X}$ if it is a non-negative integrable function such that:
$\int_{Val(X)} p(x)\, dx = 1$.

Uniform Distribution: $X \sim \text{Unif}[a, b]$ if it has the PDF:
$$p(x) = \begin{cases} \frac{1}{b-a} & b \geq x \geq a \\ 0 & \text{otherwise.} \end{cases}$$

Gaussian Distribution: X has a Gaussian distribution: $X \sim \mathcal{N}(\mu; \sigma^2)$ if it has the PDF:
$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Joint Density Function: Let P be a joint distribution over $X_1, \ldots, X_n$. A function $p(x_1, \ldots, x_n)$ is a joint density function of $X_1, \ldots, X_n$ if:

1. $p(x_1, \ldots, x_n) \geq 0 \; \forall x_1, \ldots, x_n \in X_1, \ldots, X_n$.
2. p is integratable.
3. For any choice of $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$:

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n)$$
$$= \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} p(x_1, \ldots, x_n)\, dx_1 \ldots\, dx_n$$

Conditional Density Function: Suppose you would like to condition over the event: $x - \epsilon \leq X \leq x + \epsilon$. Then $P(Y \mid x) = \lim_{\epsilon \to 0} P(Y \mid x - \epsilon \leq X \leq x + \epsilon)$. If there is a continuous joint density function p(x,y) then
$= P(a \leq Y \leq b \mid x - \epsilon \leq X \leq x + \epsilon)$
$= \frac{P(a \leq Y \leq b, x - \epsilon \leq X \leq x + \epsilon)}{P(x - \epsilon \leq X \leq x + \epsilon)} = \frac{\int_a^b \int_{x-\epsilon}^{x+\epsilon} p(x', y)\, dy dx'}{\int_{x-\epsilon}^{x+\epsilon} p(x')\, dx'}$

Expectation of X under P is:
$\mathbb{E}_P[X] = \sum_x x.P(x)$.

Expectation if X is Continuous:
$\mathbb{E}_P[X] = \int x.p(x)\, dx$.

Linearity of Expectation:
$\mathbb{E}_P[X + Y] = \mathbb{E}_P[X] + \mathbb{E}_P[Y]$.

Conditional Expectation:
$\mathbb{E}_P[X \mid \mathbf{y}] = \sum_x x.P(x \mid \mathbf{y})$.

Variance of X:
$\mathbb{V}ar_P[X] = \mathbb{E}_P[(X - \mathbb{E}_P[X])^2]$.

Standard Deviation:
$\sigma_X = \sqrt{\mathbb{V}ar_P[X]}$.

Expectation and Variance of Gaussian distribution $X \sim \mathcal{N}(\mu; \sigma^2)$, then $\mathbb{E}[X] = \mu$ and $\mathbb{V}ar[X] = \sigma^2$.

Graph Theory

A **Graph** is a data structure $\mathcal{K} = (\mathcal{X}, \mathcal{E})$ consisting of a set of nodes, denoted $\mathcal{X} = X_1, \ldots, X_n$, and edges, denoted $\mathcal{E}$.

Induced Subgraph: Let $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, and $\mathbf{X} \in \mathcal{X}$, then an induced subgraph, denoted $\mathcal{K}[\mathbf{X}]$ is a graph $(\mathbf{X}, \mathcal{E}')$ where $\mathcal{E}'$ are all the edges $X \leftrightharpoons Y \in \mathcal{E}'$ such that $X, Y \in \mathbf{X}$.

Complete Graph (Clique): A subgraph over $\mathbf{X}$ is complete if every two nodes in $\mathbf{X}$ are connected by some edge. The set $\mathbf{X}$ is called a clique. A clique $\mathbf{X}$ is maximal if for any superset of nodes $\mathbf{Y} \supset \mathbf{X}$, $\mathbf{Y}$ is not a clique.

Upward Closure: A subset of nodes $\mathbf{X} \in \mathcal{X}$ is upwardly closed in $\mathcal{K}$ if, for any $\mathbf{X} \in \mathcal{X}$, we have that the $\text{Boundary}_X \subset \mathbf{X}$. We define upward closure of $\mathbf{X}$ to be the minimally upward closed subset $\mathbf{Y}$ that contains $\mathbf{X}$.

Topological ordering: An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \to X_j) \in \mathcal{E}$, then $i < j$.

Chordal Graph: Let $X_1 - X_2 - \ldots - X_k - X_1$ be a loop in a graph. A chord in a loop is an edge connecting X_i and X_j for two nonconsecutive nodes X_i, X_j. An undirected graph $\mathcal{H}$ is said to be chordal if and loop $X_1 - X_2 - \ldots - X_k - X_1$ for $k > 4$ has a chord. A directed graph $\mathcal{K}$ is said to be chordal if its underlying undirected graph is chordal.

Bayesian Networks

Naïve Bayes:
$P(C, X_1, \ldots, X_n) = P(C) \prod_{i=0}^n P(X_i \mid C)$

Bayesian Network:
$P(X_1, \ldots, X_n) = \prod_{i=1}^n P(X_i \mid \texttt{Pa}_{X_i}^{\mathcal{G}})$

Deterministic CPD: $f : Val(Pa_X) \mapsto Val(X)$ s.t.:
$$P(x \mid pa_x) = \begin{cases} 1 & \text{if } x = f(pa_X) \\ 0 & \text{if } x \texttt{ otherwise} \end{cases}$$

Time Granularity Assumption:
$P(\mathcal{X}^{(0:T)}) = P(\mathcal{X}^{(0)}) \prod_{t=0}^{T-1} P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(0:t)})$

Markov Assumption:
$P(\mathcal{X}^{(0:T)}) = P(\mathcal{X}^{(0)}) \prod_{t=0}^{T-1} P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(t)})$

Time Invariance Assumption:
$P(\mathcal{X}^{(t+1)} = \xi' \mid \mathcal{X}^{(t)} = \xi) = P(\mathcal{X}' = \xi' \mid \mathcal{X} = \xi)$

Two-TBN:
$P(\mathcal{X}' \mid \mathcal{X}) = P(\mathcal{X}' \mid \mathcal{X}_I) = \prod_{i=1}^n P(X_i' \mid Pa_{X_i'})$

Linear Dynamical Systems:
$P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}) = \mathcal{N}(A\mathbf{X}^{(t-1)}; Q)$
$P(O^{(t)} \mid \mathbf{X}^{(t)}) = \mathcal{N}(H\mathbf{X}^{(t)}; R)$

Gibbs Distribution: A distribution P_Φ is a Gibbs distribution parameterised by a set of factors $\Phi = \{\phi_1(\mathbf{D_1}), \ldots, \phi_K(\mathbf{D_k})\}$ if it is defined as:
$P_\Phi(X_1, \ldots, X_n) = \frac{1}{Z} \tilde{P}_\Phi(X_1, \ldots, X_n)$

Inference

Inference:
$P(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \frac{P(\mathbf{Y}, \mathbf{e})}{P(\mathbf{e})} = \frac{\sum_w P(\mathbf{y}, \mathbf{e}, \mathbf{w})}{\sum_{y,w} P(\mathbf{e})}$

Sum-Product Message Passing:
$\delta_{i \to j} = \sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} (\psi_i \times \prod_{k \in (Nb_i - \{j\})} \delta_{k \to i})$

Tree Calibration:
$\sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i(\mathbf{C}_i) = \sum_{\mathbf{C}_j - \mathbf{S}_{i,j}} \beta_j(\mathbf{C}_j)$

Graph Calibration:
$\sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i = \sum_{\mathbf{C}_j - \mathbf{S}_{i,j}} \beta_j$

MAP:
$\text{MAP}(\mathbf{Y} = \mathbf{y} \mid \mathbf{E} = \mathbf{e})$
$= \text{argmax}_y P(\mathbf{Y} = \mathbf{y} \mid \mathbf{E} = \mathbf{e})$

Convergence Bound:
$\hat{\mathbb{E}}_{\mathcal{D}}(f) = \frac{1}{M} \sum_{m=1}^M f(\xi[m])$.

Hoeffding Bound:
$P_{\mathcal{D}}(\hat{P}(\mathbf{y}) \notin [P(\mathbf{y}) - \epsilon, P(\mathbf{y}) + \epsilon]) \leq 2e^{-2M\epsilon^2}$

Chernoff Bound:
$P_{\mathcal{D}}(\hat{P}(\mathbf{y}) \notin [P(\mathbf{y})(\pm\epsilon)]) \leq 2e^{-MP(\mathbf{y})\epsilon^2/3}$
$M \geq 3 \frac{\ln(2/\delta)}{P(\mathbf{y})\epsilon^2}$.

Likelihood Weighting:
$\hat{P}_{\mathcal{D}}(\mathbf{y} \mid \mathbf{e}) = \frac{\sum_{m=1}^{M} w[m] \mathbb{1}\{\mathbf{y}[m]=\mathbf{y}\}}{\sum_{m=1}^{M} w[m]}$.

MCMC Sampling:
$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) =$
$\sum_{\mathbf{x} \in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x}) \mathcal{T}(\mathbf{x} \to \mathbf{x'})$

Stationary Distribution:
$\pi(\mathbf{X} = \mathbf{x'}) = \sum_{\mathbf{x} \in Val(\mathbf{X})} \pi(\mathbf{X} = \mathbf{x}) \mathcal{T}(\mathbf{x} \to \mathbf{x'})$

Detailed Balance Equation:
$\pi(x)\mathcal{T}(x \to x') = \pi(x')\mathcal{T}(x' \to x)$

Acceptance Probability:
$\mathcal{A}(x \to x') = \min[1, \frac{\pi(x')\mathcal{T}^{Q}(x' \to x)}{\pi(x)\mathcal{T}^{Q}(x \to x')}]$.

Metropolis-Hastings Acceptance Probability:
$\mathcal{A}(x_{-i}, x_i \to x_{-i}, x_i') =$
$\min[1, \frac{P_{\Phi}(x_i', x_{-i}) \mathcal{T}_i^{Q}(x_{-i}, x_i' \to x_{-i}, x_i)}{P_{\Phi}(x_i, x_{-i}) \mathcal{T}_i^{Q}(x_{-i}, x_i \to x_{-i}, x_i')}]$.

Learning

Relative Entropy:
$\mathbb{D}(P^* \mid\mid \tilde{P}) = \mathbb{E}_{\xi \sim P^*}[\log(\frac{P^*(\xi)}{\tilde{P}(\xi)})]$,

Negative Empirical Log-loss:
$\log P(\mathcal{D} : \mathcal{M}) = \sum_{m=1}^{M} \log P(\xi[m] : \mathcal{M})$.

Bayesian Parameter Estimation:
$P(\theta \mid x[1], \ldots, x[M]) = \frac{P(x[1], \ldots, x[M] \mid \theta) P(\theta)}{P(x[1], \ldots, x[M])}$

Expected Sufficient Statistics:
$\bar{M}_{\theta}[\mathbf{y}] =$
$\sum_{m=1}^{M} \sum_{\mathbf{h}[m] \in Val(\mathbf{H}[m])} Q(\mathbf{h}[m]) \mathbb{1}\{\xi[m]\langle \mathbf{Y} \rangle = \mathbf{y}\}$

Maximisation of Expected Parameter:
$\tilde{\theta}_{d^1 \mid c^0} = \frac{\bar{M}_{\theta}[d^1, c^0]}{\bar{M}_{\theta}[c^0]}$

Bayesian Clustering:
$\bar{M}_{\theta}[c] = \frac{\bar{M}_{\theta}[c]}{M}$
$\bar{M}_{\theta}[x_i \mid c] = \frac{\bar{M}_{\theta}[x_i, c]}{\bar{M}_{\theta}[c]}$

K-means Clustering:
$c[m] = \operatorname{argmax}_c P(c \mid x[m], \theta^t)$

Hypothesis Testing:
$d_{\mathbb{I}}(\mathcal{D}) = \sum_{x,y} \frac{M[x,y]}{M} \log \frac{M[x,y]/M}{M[x]/M \cdot M[y]/M}$
$\mathrm{R}_{d,t}(\mathcal{D}) \begin{cases} \text{Accept} & \text{if } d(\mathcal{D}) \leq t \\ \text{Reject} & \text{if } d(\mathcal{D}) > t \end{cases}$
$\text{p-value}(t) = P(\{\mathcal{D} : d(\mathcal{D}) > t\} \mid H_0, M)$

Likelihood:
$\mathbb{I}_{\hat{P}_{\mathcal{D}}}(X_i; Pa_{X_i}^{\mathcal{G}})$
$= \sum_{\mathbf{u}_i} \sum_{\mathbf{x}_i} \hat{P}(x_i, \mathbf{u}_i) \log \frac{\hat{P}(x_i, \mathbf{u}_i)}{\hat{P}(x_i)\hat{P}(\mathbf{u}_i)}$

Entropy:
$\mathbb{H}_{\hat{P}_{\mathcal{D}}}(X_i) = \sum_{x_i} \hat{P}(x_i) \log \frac{1}{\hat{P}(x_i)}$

Bayesian Structure Learning:
$P(\mathcal{G} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{G}) P(\mathcal{G})}{P(\mathcal{D})}$
$\text{score}_B(\mathcal{G} : \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$
$P(\mathcal{D} \mid \mathcal{G}) = \int_{\Theta_{\mathcal{G}}} P(\mathcal{D} \mid \boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G}) P(\boldsymbol{\theta}_{\mathcal{G}} \mid \mathcal{G}) d\boldsymbol{\theta}_{\mathcal{G}}$

Marginal Likelihood for Binomials:
$P(x[1], \ldots, x[M])$
$= P(x[1]) \cdot \ldots \cdot P(x[m] \mid x[1], \ldots, x[M-1])$

Marginal Likelihood for Multinomials:
$P(x[1], \ldots, x[M]) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+M)} \cdot \prod_{i=1}^{k} \frac{\Gamma(\alpha_i + M[x^i])}{\Gamma(\alpha_i)}$

Bayesian Score:
$P(\mathcal{D} \mid \mathcal{G}) = \prod_i \prod_{\mathbf{u}_i \in Val(Pa_{X_i}^{\mathcal{G}})} \frac{\Gamma(\alpha_{X_i \mid \mathbf{u}_i}^{\mathcal{G}})}{\Gamma(\alpha_{X_i \mid \mathbf{u}_i}^{\mathcal{G}} + M[\mathbf{u}_i])} \cdot$
$\prod_{\mathbf{x}_i^j \in Val(X_i)} \left[\frac{\Gamma(\alpha_{x_i^j \mid \mathbf{u}_i}^{\mathcal{G}} + M[x_i^j, \mathbf{u}_i])}{\Gamma(\alpha_{x_i^j \mid \mathbf{u}_i}^{\mathcal{G}})} \right]$

BIC Score:
$\text{score}_{BIC}(\mathcal{G} : \mathcal{D}) =$
$M \sum_{i=1}^{n} \mathbb{I}_{\hat{P}_{\mathcal{D}}}(X_i; Pa_{X_i}^{\mathcal{G}}) - \frac{\log M}{2} \dim[\mathcal{G}]$

Decomposibility:
$\text{score}(\mathcal{G} : \mathcal{D}) = \sum_i \text{FamScore}(X_i \mid Pa_{X_i}^{\mathcal{G}} : \mathcal{D})$

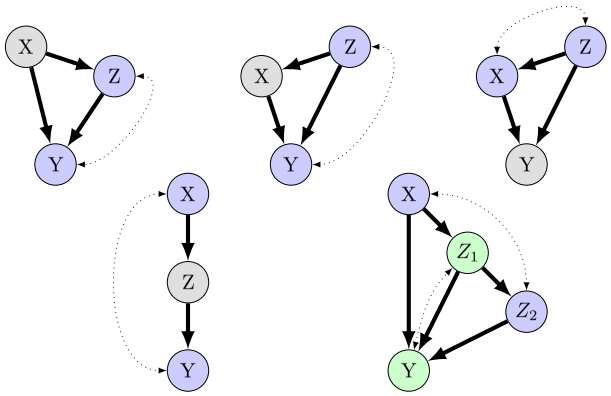
Tree weight:
$w_{i \to j} = \text{FamScore}(X_i \mid X_j : \mathcal{D}) -$
$\text{FamScore}(X_i : \mathcal{D})$

Learning Graphs:
$\mathcal{G}^* = \operatorname{argmax}_{\mathcal{G} \in \boldsymbol{\mathcal{G}}} \text{score}(\mathcal{G} : \mathcal{D})$

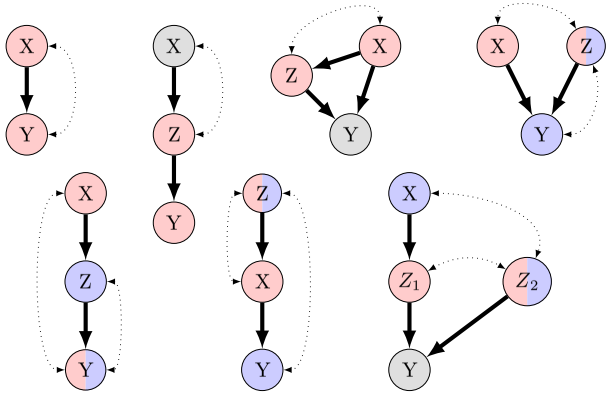
Causality

Intervention Query:
$P_{\mathcal{C}}(\mathbf{Y} \mid do(z), \mathbf{x}) = P_{\mathcal{C}_{\mathbf{Z}=\mathbf{z}}}(\mathbf{Y} \mid \mathbf{x})$

Identifiable when $P(Y \mid do(X))$:



Not Identifiable when $P(Y \mid do(X))$:



Learning with Intervention Data:
$P(\xi \mid do(\mathbf{Z} := \mathbf{z}), \mathcal{C}) = \prod_{X_i \notin \mathbf{Z}} P(x_i \mid \mathbf{u}_i)$

Sufficient Statistics (Intervention Data):
$M[x_i; \mathbf{u}_i] =$
$\sum_{m : X_i \notin \mathbf{Z}[m]} \mathbb{1}\{X_i[m] = x_i, Pa_{X_i}[m] = \mathbf{u}_i\}$

Likelihood of Data (Intervention):
$L(\mathcal{C} : \mathcal{D}) = \prod_{i=1}^{n} \prod_{x_i \in Val(X_i), \mathbf{u}_i \in Val(Pa_{X_i})} \theta_{x_i \mid \mathbf{u}_i}^{M[x_i; \mathbf{u}_i]}$

Decision Theory

Expected Utility:
$\text{EU}[D[a]] = \sum_{\mathbf{x}} P(\mathbf{x} \mid a) U(\mathbf{x}, a)$

Maximum Expected Utility:
$a^* = \operatorname{argmax}_a \text{EU}[D[a]]$
$= \operatorname{argmax}_a \sum_{\mathbf{x}} P(\mathbf{x} \mid a) U(\mathbf{x}, a)$

Expected Utility with Information:
$\text{EU}[D[\delta_A]] = \sum_{\mathbf{x}, a} P_{\delta_A}(\mathbf{x}, a) U(\mathbf{x}, a)$

Maximal Expected Utility (MEU) Strategy:
$\operatorname{argmax}_{\delta_{D_1}, \ldots, \delta_{D_k}} \text{EU}[\mathcal{I}[\delta_{D_1}, \ldots, \delta_{D_k}]]$

Value of Information:
$\text{VPI}(A \mid X) := \text{MEU}(D_{X \to A}) - \text{MEU}(D)$