Instructions:

- This assignment is meant to help you grok certain concepts we will use in the course. Please don't copy solutions from any sources.
- Avoid verbosity.
- The assignment needs to be typeset in latex using the attached tex file. The solution for each question should be written in the solution block in space already provided in the tex file. **Handwritten assignments will not be accepted.**
- Deadline for submission is **11:55PM 1/4/2018**.

1. **Independence of Random Variables**
   A and B are two random variables which can take values 0 or 1. Two joint probability distributions over A and B are provided in the tables below. For each case, argue whether A and B are independent.

Table 1: (a)

	A=0	A=1
B=0	0.12	0.18
B=1	0.28	0.42

Table 2: (b)

	A=0	A=1
B=0	0.20	0.18
B=1	0.28	0.34

Solution:

2. Ram is trying to study the causes of aggressive behaviour in males. For his initial experiments, he decides to take into account two parameters, namely, the basal level of testosterone in the male (high or low) and the kind of neighbourhood he grew up in (violent/non-violent). Based on a survey of males in a city that he conducted, he estimated that 80% of the males grew up in non-violent neighbourhoods. He also gathered the following posteriors

	Testosterone	
Neighbourhood	High	Low
Violent	0.7	0.3
Non-Violent	0.4	0.6

		Aggression	
Testosterone	**Neighbourhood**	High	Low
High	Violent	0.75	0.25
High	Non-Violent	0.22	0.78
Low	Violent	0.60	0.40
Low	Non-violent	0.15	0.85

What is the probability that

(a) A male who grew up in a non-violent neighbourhood is highly aggressive.

Solution:

(b) An arbitrarily chosen male who is highly aggressive, has high levels of testosterone and grew up in a non-violent neighbourhood.

Solution:

3. **Game of Diamonds**
You are playing a game in which you have an opportunity to win diamonds. You are shown three identical boxes, one of which contains diamonds and the other two boxes are empty. The game proceeds as follows:

- You choose one box, which you think might contain diamonds.
- Among the remaining boxes, either one or both are empty. The game host opens one such empty box.
- Now you have two options: stick to the choice you made earlier, or choose the other box. Depending on the option you choose, you win or lose.

Which option will you choose in the last step and why? (Hint: compute probability of winning in both cases)

Solution:

4. **Sampling from continuous distributions**
You are given a random number generator R, which gives a real number output sampled from the probability distribution $U_{0,1}$. $U_{a,b}$ is defined as:

$$U_{a,b}(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & otherwise \end{cases}$$

(a) Use R to sample from $U_{a,b}$ for $a, b \in \mathbb{R}$ $(a < b)$.

Solution:

(b) Given a probability distribution $P_X(x)$, whose PDF is given by $F_X(x)$ and CDF is given by $C_X(x)$, using a random number generated using R, how will you obtain a sample from the distribution $P_X(x)$?

Solution:

5. Consider the random variables X, Y, Z, W which take $3, 4, 4, 2$ values respectively.
   (a) Consider a joint distribution P_1 over these 4 variables. Without any information about the (in)dependencies between the variables, what is the minimum number of parameters you will need to represent this distribution?

   Solution:

   (b) An insight into the variables now reveals the information that $(X \perp W | Z)$. What is the minimum number of parameters needed to represent this distribution in this case?

   Solution:

   (c) An oracle further tells you that $(Y \perp X | Z, W)$. What is the minimum number of parameters needed to represent this distribution in this case?

   Solution:

6. The students of a college have the option of choosing between two mess caterers, namely, Fake Foods (FF) and Terrible Taste (TT). At the end of each month, each student needs to pick his/her caterer of choice. Based on past experience, we know that 80% of the students choosing FF opt to continue eating in FF for the next month, while 40% of the students eating in TT choose to opt for FF in the subsequent month.
   (a) If 50% of the students are assigned to FF in the first month and the rest for TT, what fraction of students are in FF at the start of the 4^{th} month?

   Solution:

   (b) Does the fraction of students eating in FF converge to a certain value? If yes, what is the value?

   Solution:

   (c) Repeat part (a) with 60% of students assigned to TT for the first month. What is the answer to part (b) in this case? If it converges, does it converge to the same value or is it different? Justify your answer.

   Solution:

7. A Markov Chain is a discrete time stochastic process. It consists of N states and is characterized a $N \times N$ transition probability matrix P, whose entries lie in the interval $[0, 1]$, entries in each row adding up to 1. The entry P_{ij} is the probability of the state in

the next time step being j, given that the state at the current time step is i.

Consider a Markov Chain with x states and the following transition probability matrix:

$$P = \begin{bmatrix} 0 & 0.7 & 0.3 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

(a) Define $p_{ij}^{(n)}$ as the probability of reaching state j in n steps starting from state i. Calculate $p_{12}^{(3)}$ and $p_{22}^{(3)}$.

Solution:

(b) The period d_i of a state i is defined as $d_i = \gcd\{n \geq 1 \ : p_{ii}^(n) > 0\}$. If $p_{ii}^{(n)} = 0$ $\forall n \geq 1$, $d_i = \inf$. Find the period of each state in the Markov Chain.

Solution:

(c) The state j is said to be accessible from state i if $p_{ij}^{(n)} > 0$ for some n. States i and j are said to communicate if they are accessible from each other. Show that communication is an equivalence relation.

Solution:

(d) A Markov Chain can be partitioned into classes based on the communication relation defined previously. For the given Markov Chain, find all the equivalence classes. Out of these, which classes are aperiodic (i.e. $d_i = 1$ $\forall i$ in the class)?

Solution:

8. $\mathbb{X} = \{X_1, ..., X_N\} \in \Lambda^N$ is a multivariate random variable, with $x_i \in \Lambda, \forall i \in \{1, 2, ..., N\}$. The possible values taken by the samples($\mathbf{x}$) of $\mathbb{X}$ can be thought of as a state of a Markov Chain (refer Question 7) with Λ^N states. Consider such a Markov Chain with transition probability between states defined as

$$p_{\mathbf{xy}} = \begin{cases} q(i)\pi(y_i|(x_v)_{i\in\{1,\ldots,N\}\setminus\{i\}}), \text{ if } \exists i \in \{1,...,N\} \text{ such that } \forall j \in 1,...,N \text{ with } j \neq i, x_j = y_j \\ 0, \text{ otherwise} \end{cases}$$

where q is a density function over the indices $\{1, ..., N\}$ and π is a joint distribution over $\{X_1, X_2, ..., X_N\}$ (you can think of π as the current state probabilities). Show that for this Markov Chain, the following condition (called the detailed balance condition) is satisfied

$$\pi(\mathbf{x})p_{\mathbf{xy}} = \pi(\mathbf{y})p_{\mathbf{yx}}$$

$\forall \mathbf{x}, \mathbf{y} \in \Lambda^N$.

Hint: Prove it separately for the cases where i) $\mathbf{x} = \mathbf{y}$, ii) $\mathbf{x}$ and iii) $\mathbf{y}$ differ in only one variable and ii) $\mathbf{x}$ and $\mathbf{y}$ differ in more than one variables.

Solution:

9. Consider binary random variables $\mathbf{V} = \{V_1, V_2, ..., V_m\}$, $\mathbf{H} = \{H_1, H_2, ..., H_n\}$ taking values $(\mathbf{v}, \mathbf{h}) \in \{0,1\}^{m+n}$. The joint probability distribution is given by

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})}$$

where E is an energy function defined as

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i$$

and Z is the normalizing constant.

$$Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$$

(a) Show that

$$P(V_l = 1|\mathbf{h}) = \sigma(\sum_{i=1}^{n} w_{il} h_i + b_l)$$

Solution:

(b) Show that

$$P(h_k = 1|\mathbf{v}) = \sigma(\sum_{j=1}^{n} w_{kj} v_j + c_k)$$

Solution:

(c) Show that the marginal is given as below

$$p(\mathbf{v}) = \frac{1}{Z} \prod_{j=1}^{m} e^{b_j v_j} \prod_{i=1}^{n} (1 + e^{c_i + \sum_{j=1}^{m} w_{ij} v_j})$$

Solution:

10. (a) Show that for random variable $x \in \mathbf{R^n}$ drawn from the distribution $\mathcal{N}(\mu, \Sigma)$, the random variable $y = Ax + b$ follows the distribution $\mathcal{N}(A\mu + b, A^T)$.

Solution:

(b) Using the result from part (a), give a method for sampling from an arbitrary normal distribution with mean μ and variance Σ, given that you have the means to sample from the standard normal distribution.

Solution:

11. We use the notion of "distance" to measure the difference between 2 quantities. One standard measure of distance is the Euclidean norm for points in d-dimensional space. There are problems where we would like to measure distances between different mathematical ojects such as probability distributions. Consider the situation where the model outputs a probability distribution (for example: softmax) and we want the loss function to capture the distance between the predicted output and the true output. One of the commonly used functions to measure distance between 2 probability distributions is the KL-divergence. The KL-divergence between 2 distributions $p(x)$ and $q(x)$ is defined as

$$KL(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

The summation is replaced by integration for continuous distributions.
Based on the above definition, answer the following questions

(a) Under what condition(s) is $KL(p||q) = 0$?

Solution:

(b) Is the function symmetric?

Solution:

(c) One necessary property of a distance function is that it needs to be ≥ 0. Prove that $KL(p||q) \geq 0$.

Solution:

12. In future classes we will encounter situations where we need to compute the KL-divergence between 2 gaussians. As a warm up, derive the expression for the KL-divergence between

2 univariate gaussians p and q.
$p - \mathcal{N}(\mu_1, \sigma_1)$ and $q - \mathcal{N}(\mu_2, \sigma_2)$.
Understand the effect of each of the terms to the value of the divergence and try to convince yourself of these findings.

Solution:

Bonus question Derive the expression for the general multivariate case.

Solution: