1. 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.

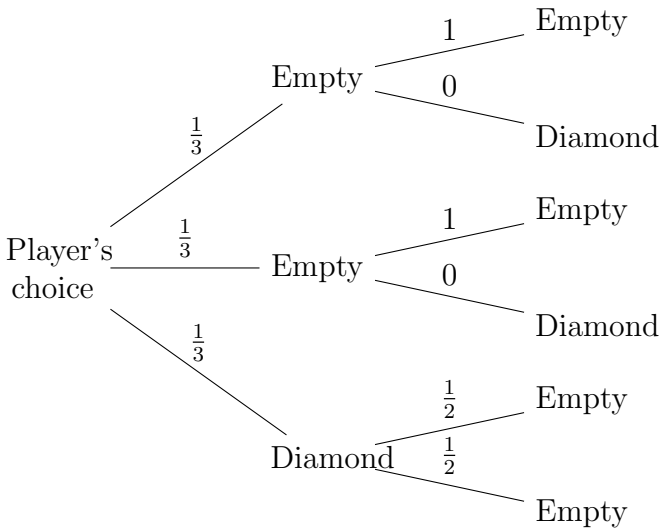
   Which option will you choose in the last step and why? (Hint: compute probability of winning in both cases)

   A. The player sticks to the box he chose earlier as his chance of winning is greater in this case.

   B. The player switches and choose the other box as his chance of winning is greater in this case.

   C. His choice in the last step does not influence his chance of winning.

Solution: B is the correct answer.
We can draw the following game tree where the first level denotes the player's choice and the second level denotes the Host's choice.



The above game tree shows that if the player chooses the diamond box in his first choice and sticks with it, the probability of his winning will be $\frac{1}{3}.\frac{1}{2} + \frac{1}{3}.\frac{1}{2} = \frac{1}{3}$.

But if he chooses an empty box(he will choose an empty box with $\frac{2}{3}$ probability) and then switches, his probability of winning will be $\frac{1}{3}.1 + \frac{1}{3}.1 = \frac{2}{3}$, as the host

will never reveal the diamond box. Therefore, the chance of winning is more if the player switches and chooses the other box.

2. A and B are two random variables which can take values 0 or 1. A joint probability distributions over A and B is provided in the table below. Given the table, which of the following statement is correct.

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

Table 1: P(A,B)

A. A and B are independent random variables.

B. A and B are dependent random variables.

Solution: Option B is the correct answer.
To show that A and B are independent, we have to show that

$$P(A)P(B) = P(A, B)$$

Using the Table 1, we can find the marginals of A and B which will be given by,

P(A) =

A=0	A=1
0.48	0.52

and P(B) =

B=0	B=1
0.38	0.62

We can clearly see,

P(A)P(B) $\neq$

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

which is $\neq$ P(A,B).

Therefore, the random variables A and B are not independent.

3. The joint probability distribution of two binary random variables, X and Y, is given below:

X	**Y**	**P(X,Y)**
0	0	0.1
0	1	0.7
1	0	0.15
1	1	0.05

Find the marginal distribution of X *i.e* $P(X) = ?$

A.

X	**P(X)**
0	0.5
1	0.5

B.

X	**P(X)**
0	0.2
1	0.8

C.

X	**P(X)**
0	0.70
1	0.25

D.

X	**P(X)**
0	0.8
1	0.2

Solution: Option D is the correct answer.

4. 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

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

Testosterone	**Neighbourhood**	**Aggression** 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

Given the above information, what is the probability that a male who grew up in a non-violent neighbourhood is highly aggressive ?

A. 0.178

B. 0.248

C. 0.314

Solution: Option A is the correct answer.
Let's define certain notations:

- Neighbourhood(N) : Violent(V) = 0, Non Violent(NV) = 1
- Testosterone Levels(T) : High(H) = 0, Low(L) = 1
- Aggression Levels(A) : High(H) = 0, Low(L) = 1

Therefore, we can find certain joint probabilities as follows:

- $P(T \cap N)$

  $= P(T|N)P(N)$

  $=$

	T=0	T=1
N=0	0.7	0.3
N=1	0.4	0.6

.

N=0	N=1
0.2	0.8

  $=$

	T=0	T=1
N=0	0.14	0.06
N=1	0.32	0.48

- Similarly, given $P(A|T,N)$, we can find $P(A \cap T \cap N)$ as follows:

  $P(A \cap T \cap N)$

  $= \quad P(A|T,N).P(T \cap N)$

  $=$

		A=0	A=1
T=0	N=0	0.105	0.035
T=0	N=1	0.0704	0.2496
T=1	N=0	0.036	0.024
T=1	N=1	0.072	0.408

- Marginalizing T over $P(A \cap T \cap N)$, we get

  $P(A \cap N)$

  $=$

	A=0	A=1
N=0	0.141	0.059
N=1	0.1424	0.6576

Therefore, $P(A=0|N=1) = \frac{P(A=0 \cap N=1)}{P(N=1)} = \frac{0.1424}{0.8} = 0.178$

5. Keeping the information given in Question 4 in mind, what is the probability that an arbitrarily chosen male who is highly aggressive, has high levels of testosterone and grew up in a non-violent neighbourhood?

   A. 0.178

   B. 0.248

   C. 0.314

Solution: Option B is the correct answer.

$P(T = 0, N = 1|A = 0) = \frac{P(T=0 \cap N=1 \cap A=0)}{P(A=0)} = \frac{.0704}{0.2834} = 0.248$

6. Consider the random variables X, Y, Z, W which take $3, 4, 4, 2$ values respectively. 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?

   A. 94

   B. 95

   C. 96

Solution: Option B is the correct answer.
The minimum number of parameters required to represent this distribution will be $3 * 4 * 4 * 2 - 1 = 95$ independent parameters

7. Consider the random variables X, Y, Z, W which take $3, 4, 4, 2$ values respectively. An insight into the variables now reveals the information that $(X \perp W|Z)$ and $(Y \perp X|Z, W)$. What is the minimum number of parameters needed to represent this distribution in this case?

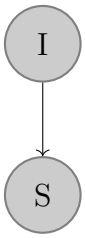
   A. 39

   B. 56

   C. 63

Solution: Option A is the correct answer Given this new information, we can represent,
$P(X, Y, W, Z) = P(X|Y, W, Z)P(Y|W, Z)P(W|Z)P(Z)$
$= P(X|Z)P(Y|W, Z)P(W|Z)P(Z)$. Therefore number of independent parameters $= 8 + 24 + 4 + 3 = 39$

1. Consider the following student Bayesian network.



where I represents the student's intelligence and S the student's SAT score. Let P(I,S) be the joint distribution representing the above network.

P(I,S) =

I	S	P(I,S)
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.05
i^1	s^1	0.25

What will be $P(s^1|i^1)$?

A. 0.167

B. 0.2

C. 0.8

D. 0.833

Solution: Option D is the correct answer.

Marginalizing P(I,S) over S, we get, P(I) =

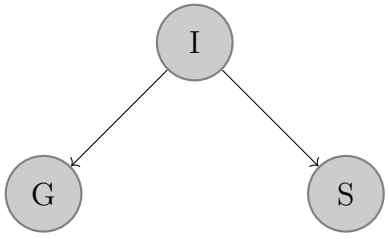
i^0	i^1
0.7	0.3

. From the given network, we can see S is dependent on I. therefore, $P(I,S) = P(I)P(S|I)$. Hence, $P(S|I)$ will be given by the table:

I	s^0	s^1
i^0	0.95	0.05
i^1	0.167	0.833

. Therefore, from the table, we can see $P(s^1|i^1) = 0.833$

2. Consider the following student Bayesian network.



where I represents the student's intelligence, G the student's grade and S the student's SAT score. Let P(I,S) be the joint distribution representing the relationship between random variables I and S.

P =

I	S	P(I,S)
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.05
i^1	s^1	0.25

Also, let $P(G|I)$ is given by,

I	g^1	g^2	g^3
i^0	0.2	0.34	0.46
i^1	0.76	0.15	0.09

What will be $P(i^1, s^1, g^2)$?

A. 0.0375

B. 0.0408

C. 0.0562

D. 0.0695

Solution: Option A is correct.

Marginalizing P(I,S) over S, we get, P(I) =

i^0	i^1
0.7	0.3

. From the given network, we can see S is dependent on I. therefore, $P(I, S) = P(I)P(S|I)$. Hence, $P(S|I)$ will be given by the table:

I	s^0	s^1
i^0	0.95	0.05
i^1	0.167	0.833

. From the given network,

$$P(I, S, G) = P(S, G|I)P(I)$$

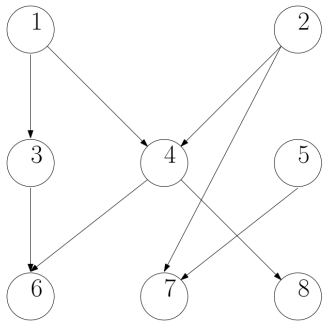
From the Bayesian network, notice that, $G \perp S|I$, i.e G is independent of S given I. Therefore,

$$P(I, S, G) = P(S|I)P(G|I)P(I)$$

Therefore, substituting values, we get,

$$\begin{aligned} P(i^1, s^1, g^2) &= P(s^1|i^1)P(g^2|i^1)P(i^1) \\ &= 0.833 \times 0.15 \times 0.3 \\ &= 0.0375 \end{aligned}$$

3. Consider the Bayesian network G given below:



Consider a distribution over 8 random variables $X_1, \ldots, X_8$ given by the Bayesian network to the left. What is the largest set of random variables that is independent of X_3.

A. $\{X_2, X_5, X_7\}$

B. $\{X_5, X_4, X_8\}$

C. $\{X_5, X_7, X_8\}$

Solution: Option A is the correct answer. $\{X_2, X_5, X_7\}$ is the largest independent set. Please refer Figure 1 for the detailed solution.

4. Consider the Bayesian network G given in question 3. What is the largest set of random variables that is independent of X_3, conditioned on X_1.

A. $\{X_1, X_2, X_4, X_6\}$

B. $\{X_1, X_2, X_4, X_5, X_7\}$

C. $\{X_1, X_2, X_4, X_5, X_7, X_8\}$

Solution: Option C is the correct answer. Please refer Figure 1 for the detailed solution.

5. Consider the Bayesian network G given in question 3. What is the largest set of random variables that is independent of X_3, conditioned on X_1 and X_4.

   A. $\{X_1, X_2, X_4, X_6\}$

   B. $\{X_1, X_2, X_4, X_5, X_7\}$

   C. $\{X_1, X_2, X_4, X_5, X_7, X_8\}$

Solution: Option C is the correct answer. Please refer Figure 1 for the detailed solution.

6. Consider the following two statements and choose the correct option.

   1. If G is an I-Map for P, then P factorizes according to G.
   2. If P factorizes according to G, then G is an I-Map for P,

   where G is a Bayesian network and P is the probability distribution.

   A. Statement 1 is correct but Statement 2 is not.

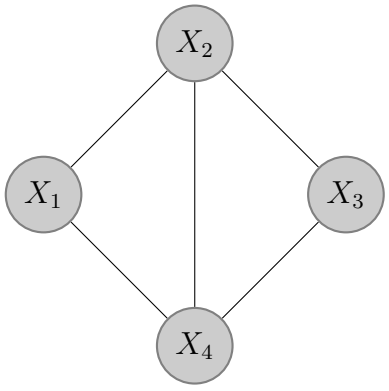
   B. Statement 2 is correct but Statement 1 is not.

   C. Both are correct.

   D. Both are incorrect

Solution: Option C is correct.

7. Consider the Markov Network H given below:



If Q is a probability distribution that factorised according to H, what can be said about the form of the distribution Q?

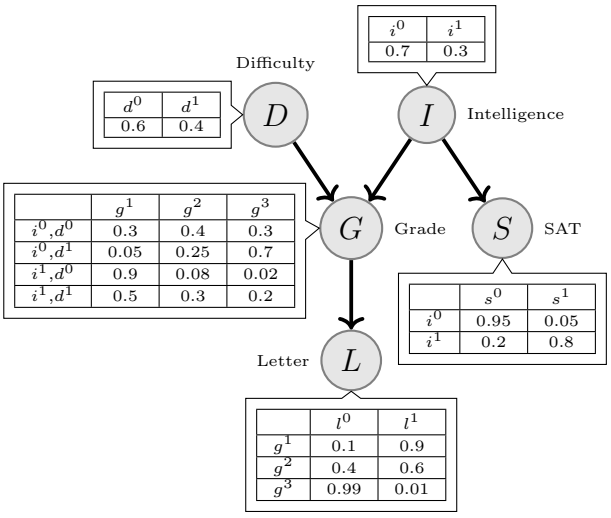
A. $Q(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = \frac{1}{Z}\psi_1(x_1, x_2, x_4)\psi_2(x_2, x_3, x_4)$

B. $Q(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = \frac{1}{Z}\psi_1(x_1, x_2, x_4)\psi_2(x_2, x_3, x_4)\psi_3(x_2, x_4)$

C. $Q(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = \frac{1}{Z}\psi_1(x_1, x_2)\psi_2(x_2, x_3)\psi_3(x_3, x_4)\psi_4(x_1, x_4)\psi_5(x_2, x_4)$

Solution: Option A is the correct option.

8. Consider the following bayesian network along with their respective conditional probability distribution.



i^0	i^1
0.7	0.3

d^0	d^1
0.6	0.4

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Find $P(i^0, d^1, g^1, s^1, l^0)$

A. 0.00003

B. 0.00005

C. 0.00007

D. 0.00010

Solution: Option C is correct.
The graph gives us a natural factorization for the joint distribution. In this case,

$$P(I, D, G, S, L) = P(I)P(D)P(G|I, D)P(S|I)P(L|G)$$

$$\begin{aligned} P(i^0, d^1, g^1, s^1, l^1) &= P(i^0)P(d^1)P(g^1|i^0, d^1)P(s^1|i^0)P(l^0|g^1) \\ &= 0.7 \times 0.4 \times 0.05 \times 0.05 \times 0.1 \\ &= 0.00007 \end{aligned}$$

The detailed solution of question 3,4 and 5 is given below:

Solution:

Trails from X3 to
X1 :
(i) $X3 \leftarrow X1$
(ii) $X3 \rightarrow X6 \leftarrow X4 \leftarrow X1$
X2 :
(iii) $X3 \leftarrow X1 \rightarrow X4 \leftarrow X2$
(iv) $X3 \rightarrow X6 \leftarrow X4 \leftarrow X2$
X4 :
(v) $X3 \leftarrow X1 \rightarrow X4$
(vi) $X3 \rightarrow X6 \leftarrow X4$
X5 :
(vii) $X3 \leftarrow X1 \rightarrow X4 \leftarrow X2 \rightarrow X7 \leftarrow X5$
(viii) $X3 \rightarrow X6 \leftarrow X4 \leftarrow X2 \rightarrow X7 \leftarrow X5$
X6 :
(ix) $X3 \rightarrow X6$
(x) $X3 \leftarrow X1 \rightarrow X4 \rightarrow X6$
X7 :
(xi) $X3 \leftarrow X1 \rightarrow X4 \leftarrow X2 \rightarrow X7$
(xii) $X3 \rightarrow X6 \leftarrow X4 \leftarrow X2 \rightarrow X7$
X8 :
(xiii) $X3 \leftarrow X1 \rightarrow X4 \rightarrow X8$
(xiv) $X3 \rightarrow X6 \leftarrow X4 \rightarrow X8$

(a) The largest set of random variables that is independent of X3:
Trail numbers (i), (v), (ix) and (xiii) are active. So X1, X4, X6 and X8 are dependent on X3. Therefore largest independent set is { X2, X5, X7 }.

(b) The largest set of random variables that is independent of X3 conditioned on X1:
Only active trail is (ix), which means X6 is dependent on X3. Therefore largest independent set is { X1, X2, X4, X5, X7, X8 }.

(c) The largest set of random variables that is independent of X3 conditioned on X1 and X4:
Here also, only active trail is (ix), which means X6 is dependent on X3. Therefore largest independent set is { X1, X2, X4, X5, X7, X8 }.

Figure 1: Solution

1. Consider you have M hidden units and N visible units. What would be the number of parameters for a Restricted Boltzmann machine and a Boltzmann machine?

   A. C_1^{MN}, C_2^{MN}

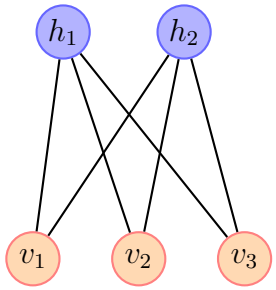
   B. C_1^{M+N}, C_2^{M+N}

   C. C_1^{MN}, C_2^{M+N}

   D. C_2^{M+N}, C_1^{MN}

Solution: Option C is the correct answer. The number of parameters for RBM is $C_1^{MN} = M \times N$ whereas for a Boltzmann machine, it is C_2^{M+N} which is equal to the number of edges in a undirected graph with $M \times N$ nodes.

2. Consider the Markov Network J given below:



where $|V| = 3$ (i.e., $V \in \{0,1\}^3$) and $|H| = 2$ (i.e., $H \in \{0,1\}^2$).

The joint probability distribution of J is given as follows:

$$P(V,H) = \frac{1}{Z}\prod_i\prod_j \phi_{ij}(v_i,h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)$$

and the corresponding factors are given below:

$\phi_{11}(v_1,h_1)$			$\phi_{12}(v_1,h_2)$			$\phi_{21}(v_2,h_1)$			$\phi_{22}(v_2,h_2)$			$\phi_{31}(v_3,h_1)$			$\phi_{32}(v_3,h_2)$		
0	0	20	0	0	6	0	0	3	0	0	2	0	0	6	0	0	3
0	1	3	0	1	20	0	1	3	0	1	1	0	1	3	0	1	1
1	0	5	1	0	10	1	0	2	1	0	10	1	0	5	1	0	10
1	1	10	1	1	2	1	1	10	1	1	10	1	1	10	1	1	10

$\psi_1(v_1)$		$\psi_2(v_2)$		$\psi_3(v_3)$		$\xi_1(h_1)$		$\xi_2(h_2)$	
0	30	0	100	0	1	0	100	0	10
1	1	1	1	1	100	1	1	1	10

Which of the following options correctly computes $P(V =< 0,1,0 >, H =< 0,1 >)$?

A.

$$\begin{aligned} &P(V =< 0,1,0 >, H =< 0,1 >) \\ =&\frac{1}{Z}\phi_{11}(0,1)\phi_{12}(0,1)\phi_{21}(0,1) \\ &\phi_{22}(0,1)\phi_{31}(0,1)\phi_{32}(0,1) \\ &\psi_1(0)\psi_2(0)\psi_3(0)\xi_1(1)\xi_2(1) \end{aligned}$$

B.

$$\begin{aligned}P(V =&< 0,1,0 >, H =< 0,1 >)\\ =&\frac{1}{Z}\phi_{11}(0,1)\phi_{12}(0,1)\phi_{21}(0,1)\\ &\phi_{22}(0,1)\phi_{31}(0,1)\phi_{32}(0,1)\\ &\psi_1(0)\psi_2(1)\psi_3(0)\xi_1(0)\xi_2(1)\end{aligned}$$

C.

$$\begin{aligned}P(V =&< 0,1,0 >, H =< 0,1 >)\\ =&\frac{1}{Z}\phi_{11}(0,0)\phi_{12}(0,1)\phi_{21}(1,0)\\ &\phi_{22}(1,1)\phi_{31}(0,0)\phi_{32}(0,1)\\ &\psi_1(0)\psi_2(0)\psi_3(0)\xi_1(1)\xi_2(1)\end{aligned}$$

D.

$$\begin{aligned}P(V =&< 0,1,0 >, H =< 0,1 >)\\ =&\frac{1}{Z}\phi_{11}(0,0)\phi_{12}(0,1)\phi_{21}(1,0)\\ &\phi_{22}(1,1)\phi_{31}(0,0)\phi_{32}(0,1)\\ &\psi_1(0)\psi_2(1)\psi_3(0)\xi_1(0)\xi_2(1)\end{aligned}$$

Solution: Option D is the correct answer.
For $(V =< 0,1,0 >, H =< 0,1 >)$, we need the following 3×2 values: $\phi_{11}(0,0), \phi_{12}(0,1), \phi_{21}(1,0), \phi_{22}(1,1), \phi_{31}(0,0), \phi_{32}(0,1), \psi_1(0), \psi_2(1), \psi_3(0)$, $\xi_1(0)$ and $\xi_2(1)$. Therefore,

$$\begin{aligned}P(V =&< 0,1,0 >, H =< 0,1 >)\\ =&\frac{1}{Z}\phi_{11}(0,0)\phi_{12}(0,1)\phi_{21}(1,0)\\ &\phi_{22}(1,1)\phi_{31}(0,0)\phi_{32}(0,1)\\ &\psi_1(0)\psi_2(1)\psi_3(0)\xi_1(0)\xi_2(1)\end{aligned}$$

3. Following question 3 above, which of the following options is the correct choice for the partition function?

A.

$$Z = \sum_{h_1=0}^{1} \sum_{h_2=0}^{1} P(V =< v_1, v_2, v_3 >, H =< h_1, h_2 >)$$

B.

$$Z = \sum_{v_1=0}^{1} \sum_{v_2=0}^{1} \sum_{v_3=0}^{1} P(V =< v_1, v_2, v_3 >, H =< h_1, h_2 >)$$

C.

$$Z = \sum_{v_1=0}^{1} \sum_{v_2=0}^{1} \sum_{v_3=0}^{1} \sum_{h_1=0}^{1} \sum_{h_2=0}^{1} P(V =< v_1, v_2, v_3 >, H =< h_1, h_2 >)$$

Solution: Option C is the correct answer.

4. Consider a Restricted Boltzmann Machine which has m visible variables V and n hidden variables H. Let us assume that the parametric form of the factors is given by the energy function, E, as defined below:

$$E(V,H) = -\sum_i \sum_j w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$

The joint probability distribution of V and H is given as the exponential of the energy function defined on V and H:

$$P(V,H) = \frac{e^{-E(V,H)}}{\sum_V \sum_H e^{-E(V,H)}}$$

Using the information given above, derive the formula for calculating $P(V)$

A. $P(V) = \sum_H e^{-E(V,H)}$

B. $P(V) = \sum_V e^{-E(V,H)}$

C. $P(V) = \frac{\sum_H e^{-E(V,H)}}{\sum_V \sum_H e^{-E(V,H)}}$

D. $P(V) = \frac{\sum_V e^{-E(V,H)}}{\sum_V \sum_H e^{-E(V,H)}}$

Solution: Option C is the correct answer. For calculating $P(V)$, we have to marginalize $P(V,H)$ over H.

5. Using the information given in question 5, derive a formula for calculating $P(v_l = 0|H)$ where v_l is the l^{th} visible unit.

A. $P(v_l = 0|H) = \sigma(-\sum_{i=1}^{n} w_{il}h_i - b_l)$

B. $P(v_l = 0|H) = \sigma(\sum_{i=1}^{n} w_{il}h_i + b_l)$

C. $P(v_l = 0|H) = 1 - \sigma(-\sum_{i=1}^{n} w_{il}h_i - b_l)$

D. $P(v_l = 0|H) = 1 + \sigma(-\sum_{i=1}^{n} w_{il}h_i - b_l)$

Solution: Option A is the correct answer.
Given in Slide 50 of lecture 18, $P(v_l = 1|H) = \sigma(\sum_{i=1}^{n} w_{il}h_i + b_l)$.
Therefore, $P(v_l = 0|H) = 1 - P(v_l = 1|H) = \sigma(-\sum_{i=1}^{n} w_{il}h_i - b_l)$
The derivation for the same is as follows:

$$
\begin{aligned}
P(v_l = 0|H) &= 1 - P(v_l = 1|H) \\
&= 1 - \sigma(\sum_{i=1}^{n} w_{il}h_i + b_l) \\
&= 1 - \sigma(x) && \text{(Let } x = \textstyle\sum_{i=1}^{n} w_{il}h_i + b_l\text{)} \\
&= 1 - \frac{1}{1+e^{-x}} \\
&= \frac{1+e^{-x}-1}{1+e^{-x}} \\
&= \frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{\frac{1}{e^{-x}} + \frac{e^{-x}}{e^{-x}}} \\
&= \frac{1}{e^{x}+1} \\
&= \frac{1}{1+e^{x}} \\
&= \sigma(-x) \\
&= \sigma(-\sum_{i=1}^{n} w_{il}h_i - b_l)
\end{aligned}
$$

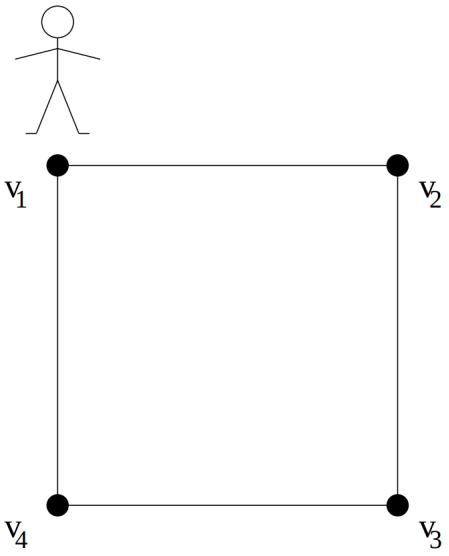
Note that in $x = \sum_{i=1}^{n} w_{il}h_i + b_l$, summation is over i.
So we can rewrite x as $x = (\sum_{i=1}^{n} w_{il}h_i) + b_l$

Therefore,

$$-x = -(\sum_{i=1}^{n} w_{il} h_i) - b_l = -\sum_{i=1}^{n} w_{il} h_i - b_l$$

Therefore, **Option A** is the correct answer.

1. Consider a very small town which has only four corners v_1, v_2, v_3, v_4 as shown in the figure below. Pappu is standing at corner v_1 at time step $t = 0$. At every time step, Pappu flips a fair coin and decides to take a step in clockwise direction if its heads or takes a step in an anti-clockwise direction if its tails. For example, at time step $t = 1$, Pappu flips a fair coin and moves to corner v_2 if it is heads or moves to corner v_4 if it is tails.



This process is repeated at time steps 1,2,3,4,...

Let X_t be a random variable which denotes the index of the street corner at which Pappu is standing at time step t. Hence $(X_0, X_1, X_2, ...)$ is a Markov Chain where $X_t \in \{1, 2, 3, 4\}$.

Now consider the following statements with respect to the above defined Markov Chain:

Statement 1 : It is a Time homogeneous Markov Chain.

Statement 2 : It is a Discrete Time Markov Chain.

Statement 3 : It is a Discrete Space Markov Chain.

Statement 4 : Its Transition matrix is a symmetric matrix.

Statement 5 : There exists a Stationary distribution for this Markov Chain.

Which of the following is the correct option with respect to the Markov Chain defined above.

A. Only statements (1), (2) and (3) are True.

B. Only statements (2), (3) and (4) are True.

C. Only statements (2), (3) and (5) are True.

D. Only statements (1), (2), (3) and (4) are True.

E. All statements are True.

Solution: Option D is the correct option.
This Markov Chain does not have a stationary distribution because Pappu will never be in the same corner in the consecutive time steps for any time step t and hence, $X_t, X_{t+1}, X_{t+2}, ...$ will never follow the same distribution.

2. Recall that initial distribution tells us how the Markov Chain starts *i.e*

$$\mu^0 = [P(X_0 = v_1) \;\; P(X_0 = v_2) \;\; P(X_0 = v_3) \;\; P(X_0 = v_4)]$$

What is the initial distribution μ^0 for the Markov Chain defined in question 1?

A. $\mu^0 = [0.25 \;\; 0.25 \;\; 0.25 \;\; 0.25]$
B. $\mu^0 = [1 \;\; 0 \;\; 0 \;\; 0]$
C. $\mu^0 = [0 \;\; 1 \;\; 0 \;\; 0]$
D. $\mu^0 = [0 \;\; 0 \;\; 0 \;\; 1]$
E. $\mu^0 = [0 \;\; 0 \;\; 1 \;\; 0]$

Solution: Option B is the correct answer.
As Pappu is initially at cornere v_0, $P(X_0 = v_1) = 1$, $P(X_0 = v_2) = 0$, $P(X_0 = v_3) = 0$, $P(X_0 = v_4) = 0$.

3. Recall that the ij^{th} entry of a transition matrix T tells us the probability of Pappu moving to j^{th} corner in t^{th} time step when he is standing at i^{th} corner in $(t-1)^{th}$ time step.
   For the Markov Chain defined in question 1, what will be its transition matrix T ?

   A. $\begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$

   B. $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

   C. $\begin{bmatrix} 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.4 \end{bmatrix}$

   D. $\begin{bmatrix} 0 & 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 & 0 \end{bmatrix}$

Solution: Option D is the correct answer.

$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{array} \begin{array}{c} \begin{array}{cccc} v_1 & v_2 & v_3 & v_4 \end{array} \\ \begin{pmatrix} 0 & 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 & 0 \end{pmatrix} \end{array}$$

4. What will be the transition matrix T of the Markov Chain defined in question 1 if instead of a fair coin, Pappu flips a biased coin which has 20% probability of getting heads.

A. $\begin{bmatrix} 0 & 0.8 & 0 & 0.2 \\ 0.2 & 0 & 0.8 & 0 \\ 0 & 0.2 & 0 & 0.8 \\ 0.8 & 0 & 0.2 & 0 \end{bmatrix}$

B. $\begin{bmatrix} 0.8 & 0 & 0.2 & 0 \\ 0 & 0.2 & 0 & 0.8 \\ 0.2 & 0 & 0.8 & 0 \\ 0 & 0.8 & 0 & 0.2 \end{bmatrix}$

C. $\begin{bmatrix} 0 & 0.2 & 0 & 0.8 \\ 0.8 & 0 & 0.2 & 0 \\ 0 & 0.8 & 0 & 0.2 \\ 0.2 & 0 & 0.8 & 0 \end{bmatrix}$

D. $\begin{bmatrix} 0.2 & 0 & 0.8 & 0 \\ 0 & 0.8 & 0 & 0.2 \\ 0.8 & 0 & 0.2 & 0 \\ 0 & 0.2 & 0 & 0.8 \end{bmatrix}$

Solution: Option C is the correct answer.

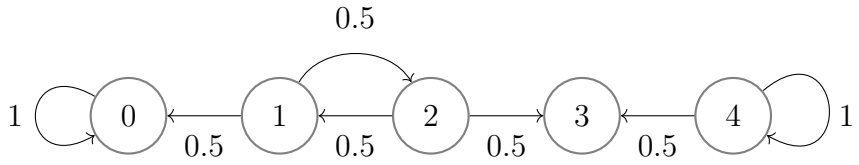
$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{array} \begin{array}{c} \begin{array}{cccc} v_1 & v_2 & v_3 & v_4 \end{array} \\ \begin{pmatrix} 0 & 0.2 & 0 & 0.8 \\ 0.8 & 0 & 0.2 & 0 \\ 0 & 0.8 & 0 & 0.2 \\ 0.2 & 0 & 0.8 & 0 \end{pmatrix} \end{array}$$

5. Consider a Markov Chain with 5 states, $0, 1, 2, 3, 4$, which has the transition probability matrix T as given below:

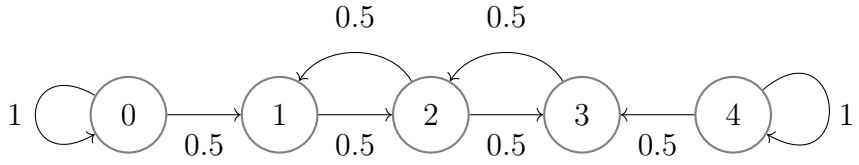
$$\begin{array}{c} \\ \mathbf{0} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Which of the following options is the correct transition diagram for the above defined Markov Chain?

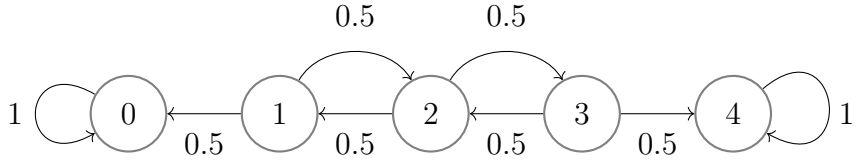
A.



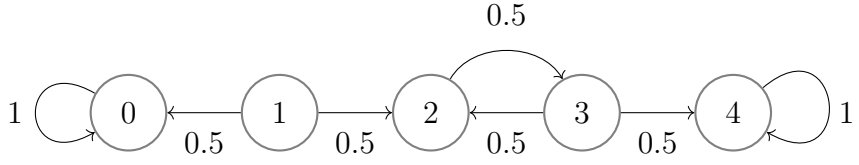
B.



C.



D.



Solution: Option C is the correct option.

6. In a Markov Chain, a state i is called an **Absorbing state** if it is impossible to leave that state *i.e* the probability of transition from state i to any other state j ($j \neq i$) is 0 and the probability of transition from state i to itself is 1.
Keeping this definition in mind, are there any absorbing states in the Markov Chain defined in question 4?

   A. Yes, there is a 1 absorbing state

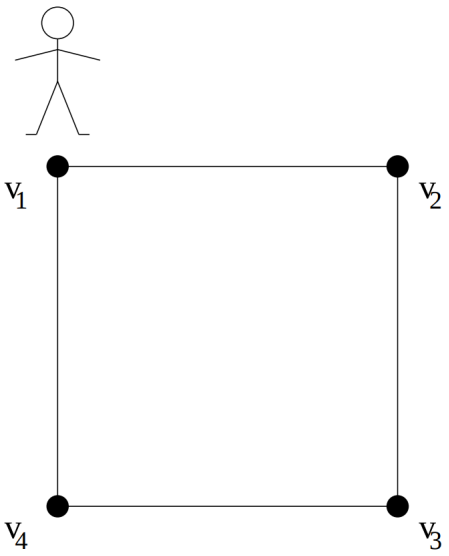
   B. Yes, there are 2 absorbing states

   C. Yes, there are 3 absorbing states

   D. No, there are no absorbing states

Solution: Option B is the correct answer.
State 0 and State 4 are the absorbing states as there is a self-loop with probability 1.

1. Consider a very small town which has only four corners v_1, v_2, v_3, v_4 as shown in the figure below. Pappu is standing at corner v_1 at time step $t = 0$. At every time step, Pappu flips a fair coin and decides to take a step in clockwise direction if its heads or takes a step in an anti-clockwise direction if its tails. For example, at time step $t = 1$, Pappu flips a fair coin and moves to corner v_2 if it is heads or moves to corner v_4 if it is tails.



This process is repeated at time steps 1,2,3,4,...
Let X_t be a random variable which denotes the index of the street corner at which Pappu is standing at time step t. Hence $(X_0, X_1, X_2, ...)$ is a Markov Chain.

Now consider the following statements with respect to the above defined Markov Chain:

Statement 1 : The chain is an irreducible Markov Chain.

Statement 2 : The chain is an aperiodic Markov Chain.

Which of the following is the correct option with respect to the Markov Chain defined above.

A. Only statement (1) is True.

B. Only statement (2) is True.

C. Both statements (1) and (2) are True.

D. Both statements (1) and (2) are False.

Solution: Option A is the correct option.
Irreducible because there is a positive probability to go from any corner to any other corner in finite number of steps.
The chain is not aperiodic because Pappu has to take even number of steps to get back to the same corner ($gcd\{2, 4, 6, ...\} \neq 1$) and hence the chain is periodic with time period 2.

2. Consider a chess board with a lonely white king making random moves, meaning that at each move, he picks one of the possible chess squares to move to, uniformly at random. Which of the following option is true with respect to the corresponding Markov Chain?

   A. Periodic and Reducible

   B. Periodic and Irreducible

   C. Aperiodic and Reducible

   D. Aperiodic and Irreducible

Solution: Option D is the correct option. Chain is irreducible and aperiodic.

3. Consider a Markov chain $(X^{(0)}, X^{(1)}, X^{(2)}, \ldots)$ where $X^{(t)}$ is the value that a random variable X takes at time step t. Further assume that $X = (X_1, X_2, \ldots, X_n)$ where each X_i can take values in $[0, c-1]$.
Let $T^{(t)}$ be a Transition matrix whose i^{th} row tells us the probability of X_i taking one of the c values at time step t keeping the values of all the other variables fixed.
Among the following options, what will be the right choice for the Transition kernel at time step t?

A. $T_i^t = P(X^t{}_i | X^t{}_j = x_j^t), \ \forall j \in \{1, 2, \ldots, n\}$

B. $T_i^t = P(X^t{}_i | X^t{}_j = x_j^t), \ \forall j \in \{1, 2, \ldots, n\} \ and \ j \neq i$

C. $T_i^t = P(X^t{}_i | X^t{}_j = x_j^{t-1}), \ \forall j \in \{1, 2, \ldots, n\}$

D. $T_i^t = P(X^t{}_i | X^t{}_j = x_j^{t-1}), \ \forall j \in \{1, 2, \ldots, n\} \ and \ j \neq i$

Solution: Option D is the correct option.

4. Continuing question 3, lets assume we have access to an unnormalised distribution $\widetilde{P}(X_1, X_2, \ldots, X_n)$, i.e. we have access to a function which takes in $X_1, \ldots, X_n$ as input and gives $\widetilde{P}(X)$ as output.
   Suppose we do *Gibbs sampling* by choosing to apply the transition kernels one after another sequentially.
   how many times do we have to access the unnormalized distribution $\widetilde{P}(X_1, X_2, \ldots, X_n)$ for one complete cycle of Gibbs sampling. One cycle means all the random variables have a chance to change its values. In particular, it means n transitions must happen.

   A. n

   B. n^c

   C. cn

   D. c^n

Solution: Option C is the correct answer.
A single transition of Gibbs sampling involves sampling from the distribution $P(X_i = x_i | X_{-i} = x_{i-1})$. Note that calculating the normalization constant Z is not a concern since we have access to $\widetilde{P}$. It can be normalized to give a valid distribution to sample from.
Let the current sample be $X = (x_i, x_{-i})$. When random variable X_i is being changed, all the samples $P(X_i = x_i | X_{-i} = x_{i-1}), x_i \in [0, c-1]$ are required. This means that for one complete cycle of Gibbs sampling, cn calls are required.

5. In question 3, suppose we use *block Gibbs sampling* instead of Gibbs sampling. How many times do we have to access the unnormalized distribution $\widetilde{P}(X_1, X_2, \ldots, X_n)$ for one complete cycle of *block Gibbs sampling*? The blocking occurs by partitioning the n random variables into $\frac{n}{k}$ partitions of k variables each, and grouping the variables in each partition to form $\frac{n}{k}$ super random variables. Gibbs sampling is then done on this set of super random variables. Note that the number of transitions in a cycle depends on k.
(You can assume that n is divisible by k)

A. $\frac{n}{k}$

B. $\frac{cn}{k}$

C. $\frac{nc^k}{k}$

D. $\frac{cn^k}{k}$

Solution: Option C is the correct answer.
Let $m = \frac{n}{k}$.
Now there are m "super random variables" $Y_1, Y_2, \ldots, Y_m$. Each Y_i consists of k regular variables, each regular variable can take c values and therefore each Y_i can take c^k values. Now Gibbs sampling proceeds as before but using $\frac{n}{k}$ super variables instead. Proceeding as before, a single transition will require sampling $\widetilde{P}(Y_i = y_i | Y_{-i} = y_{-i})$ for all y_i, which needs c^k calls. Since there are $m = \frac{n}{k}$ such random variables, $\frac{nc^k}{k}$ calls are required for a single cycle of Gibbs sampling.

6. Consider the following pseudocode for training RBMs with Block Gibbs Sampling.

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D
Output: Learned Parameters $\mathbf{W}, b, c$

(1) init $\mathbf{W}, b, c$
(2) **forall** $v \in D$ **do**
(3) Randomly initialize $v^{(0)}$
(4) **for** $t = 0, ..., k, k+1, ..., k+r$ **do**
(5) **for** $i = 1, ..., n$ **do**
(6) sample $h_i^{(t)} \sim p(h_i|v^{(t)})$
end
(7) **for** $j = 1, ..., m$ **do**
(8) sample $v_j^{(t+1)} \sim p(v_j|h^{(t)})$
end
end
(9) $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}v_d + c)v_d^T - \frac{1}{r}\sum_{t=k+1}^{k+r} \sigma(\mathbf{W}v^{(t)} + c)v^{(t)T}]$
end

All the notations carry their meaning as defined in the lecture.
Now suppose you want to train the RBM using k-contrastive divergence algorithm. Modify the above pseudocode such that it can be used to train RBM using k-contrastive divergence algorithm. In the following options, the number in the parenthesis preceding the statements indicates the line number in the pseudocode.

A. (3) Initialize $\mathbf{v^{(0)}} \leftarrow \mathbf{v}$
(4) **for** t = 0,. . . ,k
(9) $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}v_d + c)v_d^T + \frac{1}{r}\sum_{t=0}^{k} \sigma(\mathbf{W}\tilde{v} - c)\tilde{v}]$

B. (3) Initialize $\mathbf{v^{(0)}} \leftarrow \mathbf{v}$
(4) **for** t = 0,. . . ,k
(9) $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}v_d + c)v_d^T - \sigma(\mathbf{W}\tilde{v} + c)\tilde{v}]$

C. (3) Initialize $\mathbf{v^{(0)}} \leftarrow \mathbf{v}$
(4) **for** t = 0,. . . ,k
(9) $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}v_d + c)v_d^T - \frac{1}{r}\sum_{t=0}^{k} \sigma(\mathbf{W}\tilde{v} + c)\tilde{v}]$

D. (3) Initialize $\mathbf{v}^{(\mathbf{0})} \leftarrow \mathbf{v}$
(4) **for** t = 0,. . . ,k
(9) $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}v_d + c)v_d^T + \sigma(\mathbf{W}\tilde{v} - c)\tilde{v}]$

Solution: Option B is the correct option.

1. The decoder in a Variational Autoencoder is responsible for predicting a probability distribution over X, i.e $P(X|z)$. It is further assumed that the distribution $P(X|z)$ is a Gaussian distribution with unit variance. Based on the above, which of the following statements are correct?

    1. The covariance matrix of $P(X|z)$ is an identity matrix.
    2. In a Variational Autoencoder, given the latent variables, the visible variables (X) are considered to be independent of each other.

    A. Statement 1 is correct.

    B. Statement 2 is correct.

    C. Both the statements are correct.

    D. None of the statement is correct

Solution: Option C is correct.

2. Choose the correct option according to the statements given below with respect to the generative models - Restricted Boltzmann Machines (RBMs) and Variational Autoencoders (VAEs).

    1. Both of the models have to deal with an intractable probability distribution $P(X)$
    2. Both of the models deal with $P(X)$ by approximating it with Gibb's Sampling.
    3. Both of the models try to estimate the parameters of the posterior distribution $Q_\theta(z|X)$ instead of dealing with the intractable distribution $P(z|X)$.
    4. RBMs deal with $P(X)$ by approximating it with Gibb's Sampling whereas VAEs try to estimate the parameters of the posterior distribution $Q_\theta(z|X)$ instead of dealing with the intractable distribution $P(z|X)$.

    A. Statement 1 and 2 are correct

    B. Statement 1 ad 3 are correct

    C. Statement 1 and 4 are correct

    D. Only Statement 1 is correct.

Solution: Option C is correct.

3. Suppose a intractable distribution $p = p(x_1, x_2, \ldots, x_n; \theta)$, parameterized by θ is given. We can also write as p as:

$$p(x_1, x_2, \ldots, x_n; \theta) = \frac{\tilde{p}(x_1, x_2, \ldots, x_n; \theta)}{Z(\theta)},$$

where $Z(\theta)$ is the intractable partition function. The goal of Variational Inference is to try to solve an optimization problem over a class of tractable distributions Q in order to find a $q \in Q$ that is most similar to p. How would you model this goal of Variational inference as an optimization problem?

A. maximize $KL(p||q)$

B. minimize $KL(p||q)$

C. maximize $KL(q||p)$

D. minimize $KL(q||p)$

Solution: Option D is the correct answer.
We have to minimize the KL difference between q and p. Also, notice that, we have to minimize $KL(q||p)$ as it involves an expectation with respect to q whereas $KL(p||q)$ requires computing expectations with respect to p, which is typically intractable even to evaluate.

4. Continuing question 3, instead of working with p, we work with the unnormalized distribution $\tilde{p}$ and try to optimize the following function,

$$J(q) = \sum_x q(x) log \frac{q(x)}{\tilde{p}(x)}$$

What will be the lower bound on the log partition function, log $Z(\theta)$?

A. $J(q)$

B. $-J(q)$

C. log $J(q)$

D. log $\frac{1}{J(q)}$

Solution: Option B is the correct answer.

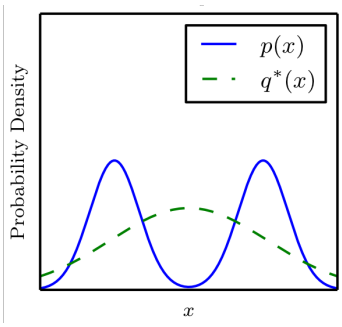
$$\begin{aligned} J(q) &= \sum_x q(x) log \frac{q(x)}{\tilde{p}(x)} \\ &= \sum_x q(x) log \frac{q(x)}{p(x)} - logZ(\theta) \\ &= KL(q||p) - logZ(\theta) \end{aligned}$$

Since, $KL(q||p) > 0$, by rearranging terms we get,

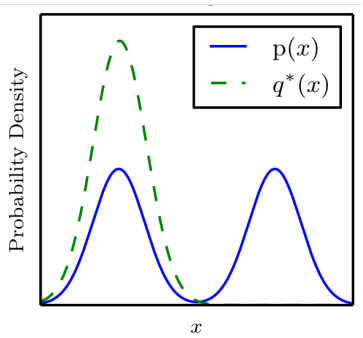
$$logZ(\theta) = KL(q||p) - J(q) >= -J(q)$$

Therefore, $-J(q)$ is the lower bound on log $Z(\theta)$.

5. Suppose the true data distribution p is a mixture of two Gaussians as shown in blue in the figures (here $x \in \mathbb{R}$). Of course, we do not know this true distribution and assume that the data comes from a single Gaussian distribution q. Given some training data you are trying to learn the parameters of this Gaussian distribution q and you experiment with two objective functions *minimize* $KL(p||q)$ and *minimize* $KL(q||p)$.
You are shown two figures below in which the green curve is the learnt distribution q. Identify which objective function each figure corresponds to.



(a) Figure 1



(b) Figure 2

A. Figure 1 corresponds to $KL(q||p)$ and Figure 2 corresponds to $KL(p||q)$

B. Figure 1 corresponds to $KL(p||q)$ and Figure 2 corresponds to $KL(q||p)$

C. Both the figures correspond to $KL(q||p)$

D. Both the figures correspond to $KL(p||q)$

Solution: Option B is the correct option.
Check this blog for details.

6. Given visible variables X and hidden variables h, we know that Restricted Boltzmann machines (RBMs) are capable of learning a probability distribution over the h-space, i.e the distribution $P(h|X)$. Suppose we have a 4-dimensional hidden representation and each h_i is binary, i.e it can take a value of either 0 or 1. From a RBM, we get the following probabilities:

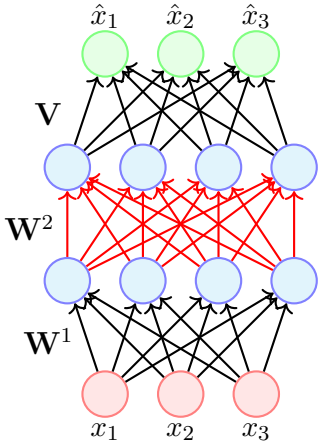
   - $P(h_1 = 1) = 0.1$
   - $P(h_2 = 1) = 0.3$
   - $P(h_3 = 1) = 0.55$
   - $P(h_4 = 1) = 0.8$

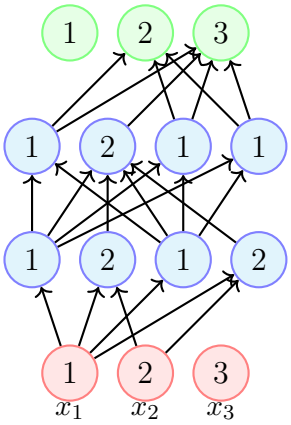
   Now, suppose in order to generate a particular hidden representation, we sample from the uniform distribution and appropriately set the value of h_i's according to the above distribution.
   The samples drawn from the uniform distribution are $0.3, 0.4, 0.2, 0.75$, for h_1, h_2, h_3, h_4 respectively. What will be the hidden representation $\begin{bmatrix} h_1 & h_2 & h_3 & h_4 \end{bmatrix}$ that can be constructed from the samples?

   A. $\begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$
   B. $\begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}$
   C. $\begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}$
   D. $\begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix}$

Solution: Option C is correct.

1. Consider a Masked Autoencoder Density Estimator (MADE) as given in the (left) figure below where the input $x \in \{0,1\}^3$, i.e, there are 3 input and 3 output nodes. The nodes in the hidden layer have been randomly assigned numbers from 1 to 2 as shown in the (right) figure. Given a weight matrix $\mathbf{W}^2$, what should be the mask matrix that must be applied to it in order to emulate the connections given in the (right) figure?





A. $\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$

B. $\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$

C. $\begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$

D. $\begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$

Solution: Option B is correct.

2. Which of the following operations helps us in providing sparse connectivity in a neural network?

   A. Masking

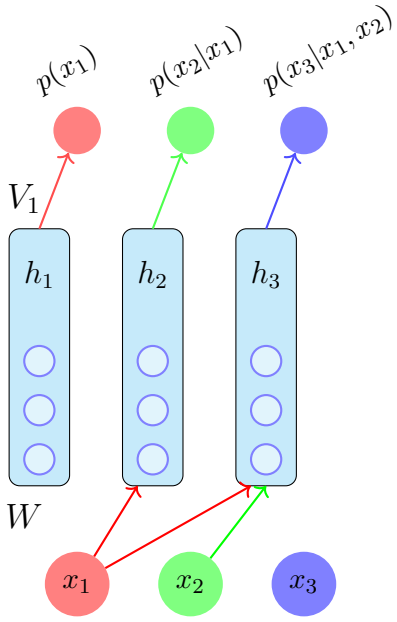
   B. Convolution operator

   C. Both A and B

   D. Neither A nor B

Solution: Option C is the correct answer.

3. Consider a Neural Autoregressive Density Estimator (NADE) as given in the figure below. Note that the dimensions of shared parameters (W and b) are $W \in \mathbb{R}^{3\times 3}$ and $b \in \mathbb{R}^{3\times 1}$. In addition, we have the $V_k \in \mathbb{R}^{3\times 1}$ and $c_k \in \mathbb{R}^1$ for each of the 3 factors. Given the above information and the equations of the NADE model, what will be the total number of parameters for this network?



A. 16

B. 19

C. 24

D. 27

Solution: Option D is the correct answer.
Number of parameters in NADE ($n = 3, d = 3$): $2nd + n + 2d = 2 * 3 * 3 + 3 + 6 = 27$

4. Which of the following statement/s is/are correct?

    1. The Auto-regressive models(NADE, MADE) do not make any independence assumptions.
    2. Both NADE, MADE are latent variable models and hence can do "abstraction"
    3. The procedure of "generation" in auto-regressive models is very slow.

    A. Statement 1 and 2 is correct.

    B. Statement 2 and 3 is correct.

    C. Statement 1 and 3 is correct.

    D. Only Statement 3 is correct.

Solution: Option C is correct.

5. Consider the NADE model given in question 3. As per the notation used in the slides, given $W = \begin{bmatrix} 0.1 & 0.25 & 0.2 \\ 0.2 & 0.4 & 0.3 \\ 0.5 & 0.5 & 0.6 \end{bmatrix}$, $b = \begin{bmatrix} 0.1 \\ 0.05 \\ 0.3 \end{bmatrix}$, $V_1 = \begin{bmatrix} 0.3 \\ 0.7 \\ 0.5 \end{bmatrix}$, $h_1 = \begin{bmatrix} 0.2 \\ 0.8 \\ 0.7 \end{bmatrix}$ and $c_1 = -0.02$, what will be the value of $p(x_1 = 1)$?

A. 0.05

B. 0.95

C. 0.28

D. 0.72

Solution: Option D is the correct answer.

$$\begin{aligned} p(x_1 = 1) &= \sigma(V_1 h_1 + c_1) \\ &= \sigma((0.3) * (0.2) + (0.7) * (0.8) + (0.5) * (0.7) - 0.02) \\ &= \sigma(0.95) \\ &= 0.72 \end{aligned}$$

6. Continuing question 5, suppose after sampling x_1 from a uniform distribution between 0 and 1, we get the value as 0.6. Further, given the values of $V_2 = \begin{bmatrix} 0.4 \\ 0.7 \\ 0.1 \end{bmatrix}$, and $c_2 = -0.5$, what will be the value of $p(x_2 = 1|x_1)$?

A. 0.55

B. 0.45

C. 0.18

D. 0.82

Solution: Option A is the correct answer.
Firstly, we have to sample x_1 from the Bernoulli distribution $p(x_1 = 1) = 0.72$. As given in the question, on sampling from the uniform distribution, we get the value 0.6 which is less than 0.72, therefore the sampled value of $x_1 = 1$. Now, to find h_2,

$$\begin{aligned} h_2 &= \sigma(W_{.,<2}x_{<2} + b) \\ &= \sigma(W. \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + b) \\ &= \sigma(\begin{bmatrix} 0.1 \\ 0.2 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.05 \\ 0.3 \end{bmatrix}) \\ &= \sigma(\begin{bmatrix} 0.2 \\ 0.25 \\ 0.8 \end{bmatrix}) \\ &= \begin{bmatrix} 0.55 \\ 0.56 \\ 0.69 \end{bmatrix} \end{aligned}$$

Given h_2, we can find $p(x_2 = 1|x_1)$ as,

$$\begin{aligned} p(x_2 = 1|x_1) &= \sigma(V_2 h_2 + c_2) \\ &= \sigma((0.4) * (0.55) + (0.7) * (0.56) + (0.1) * (0.69) - 0.5) \\ &= \sigma(0.18) \\ &= 0.55 \end{aligned}$$

1. Suppose,

$$\min_D \left[\frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(D(x)-b)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z))-a)^2]\right]$$

is the modified objective function for the discriminator where a and b are the class labels for the generated images and real images. Given a fixed generator G, which of the following options is the optimum discriminator D which will minimize the above objective function?

A.
$$D^*(x) = \frac{p_G(x) + p_{data}(x)}{ap_G(x) + bp_{data}(x)}$$

B.
$$D^*(x) = \frac{p_G(x) + p_{data}(x)}{bp_G(x) + ap_{data}(x)}$$

C.
$$D^*(x) = \frac{bp_G(x) + ap_{data}(x)}{p_G(x) + p_{data}(x)}$$

D.
$$D^*(x) = \frac{ap_G(x) + bp_{data}(x)}{p_G(x) + p_{data}(x)}$$

Solution: Option D is the correct answer. Refer this paper for more details.
We will first expand it to its integral form:

$$\min_D \int_x \left(\frac{1}{2}\left(p_{data}(x)(D(x)-b)^2\right)\right) + \left(\frac{1}{2}\left(p_z(z)(D(G(z))-a)^2\right)\right)$$

Using law of the unconscious statistician our revised objective is given by,

$$\min_D \int_x \left[\left(\frac{1}{2}\left(p_{data}(x)(D(x)-b)^2\right)\right) + \left(\frac{1}{2}\left(p_G(x)(D(x)-a)^2\right)\right)\right] dx$$

Given a generator G, we are interested in finding the optimum discriminator D which will minimize the above objective function.
To find the optima we will take the derivative of the term inside the integral w.r.t. D and set it to zero

$$\frac{d}{d(D(x))}\left(\frac{1}{2}\left(p_{data}(x)(D(x)-b)^2\right)\right) + \left(\frac{1}{2}\left(p_G(x)(D(x)-a)^2\right)\right) = 0$$

$$p_{data}(x)(D(x)-b) = -p_G(x)(D(x)-a)$$

$$D(x) = \frac{ap_G(x) + bp_{data}(x)}{p_G(x) + p_{data}(x)}$$

2. Consider an extension of GANs in which generator G will sample $z \sim N(0, I)$ and learns to make a series of complex transformations on z so that the output looks as if it came **from a particular class of our training distribution.** For example, suppose you are working with the MNIST dataset and you want the generator to generate the images of a specific class having class label y and the discriminator to classify the generated image as real or fake.

   Note that in vanilla GANs, generator can generate images of any class from the training data whereas in this variant, generator is restricted to generate the images of a particular class y.

   Keeping the above information in mind, what will be the modified objective function for this variant of GAN?

   A. $\min_G \max_D \left[\mathbb{E}_{x \sim p_{data}(x)}[log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z, y)))]\right]$

   B. $\min_G \max_D \left[\mathbb{E}_{x \sim p_{data}(x)}[log(D(x|y))] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))]\right]$

   C. $\min_G \max_D \left[\mathbb{E}_{x \sim p_{data}(x)}[log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z|y)))]\right]$

   D. $\min_G \max_D \left[\mathbb{E}_{x \sim p_{data}(x)}[log(D(x|y))] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z|y)))]\right]$

Solution: Option D is the correct option.
Refer this link for intuition.

3. Recall that GANs have a generator network G, and a discriminator network D.
θ and ϕ are the parameters of D and G respectively.
Now consider the following statements in the context of GANs.

I. While training GANs, the only signal that the generator receives is the probability that the discriminator assigns to a sample generated by the generator.

II. There is no explicit representation of $\mathbf{p_G}(\mathbf{x})$ and D must be well synchronized with G during training. ($\mathbf{p_G}(\mathbf{x})$ is the Generator Network's distribution.)

III. Markov chains are needed.

IV. GANs are capable of doing both, *Abstraction* and *Generation.*

In the context of GANs, which of the following options is correct?

A. Only statements I and II are False

B. Only statements III and IV are False

C. Only statements II, III and IV are False

D. All the statements are False

Solution: Option B is the correct answer.
Markov chains are never needed, only back-propagation algorithm is used to obtain gradients.
GANs are capable of doing *Generation* only.

4. Fill in the blank (i) in the following pseudo-code for training GANs (all notations carry their usual meaning),

---

1: **procedure** GAN Training
2: **for** number of training iterations **do**
3: **for** k steps **do**
4: • Sample minibatch of m noise samples $\{\mathbf{z}^{(1)}, .., \mathbf{z}^{(m)}\}$ from noise prior $p_g(\mathbf{z})$
5: • Sample minibatch of m examples $\{\mathbf{x}^{(1)}, .., \mathbf{x}^{(m)}\}$ from data generating distribution $p_{data}(\mathbf{x})$
6: • Update the discriminator by ascending its stochastic gradient:

$$(i)$$

7: **end for**
8: • Sample minibatch of m noise samples $\{\mathbf{z}^{(1)}, .., \mathbf{z}^{(m)}\}$ from noise prior $p_g(\mathbf{z})$
9: • Update the generator by ascending its stochastic gradient

$$(ii)$$

10: **end for**
11: **end procedure**

A. $\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} \left[\log G_\phi\left(x^{(i)}\right) + \log\left(1 - G_\phi\left(D_\theta\left(z^{(i)}\right)\right)\right)\right]$

B. $\nabla_\phi \frac{1}{m} \sum_{i=1}^{m} \left[\log\left(G_\phi\left(D_\theta\left(z^{(i)}\right)\right)\right)\right]$

C. $\nabla_\phi \frac{1}{m} \sum_{i=1}^{m} \left[\log\left(D_\theta\left(G_\phi\left(z^{(i)}\right)\right)\right)\right]$

D. $\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} \left[\log D_\theta\left(x^{(i)}\right) + \log\left(1 - D_\theta\left(G_\phi\left(z^{(i)}\right)\right)\right)\right]$

Solution: Option D is the correct answer.

5. Fill in the blank (ii) in the pseudo-code for training GANs given in the previous question.

A. $\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} \left[\log G_\phi\left(x^{(i)}\right) + \log\left(1 - G_\phi\left(D_\theta\left(z^{(i)}\right)\right)\right)\right]$

B. $\nabla_\phi \frac{1}{m} \sum_{i=1}^{m} \left[\log\left(G_\phi\left(D_\theta\left(z^{(i)}\right)\right)\right)\right]$

C. $\nabla_\phi \frac{1}{m} \sum_{i=1}^{m} \left[\log\left(D_\theta\left(G_\phi\left(z^{(i)}\right)\right)\right)\right]$

D. $\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} \left[\log D_\theta\left(x^{(i)}\right) + \log\left(1 - D_\theta\left(G_\phi\left(z^{(i)}\right)\right)\right)\right]$

Solution: Option C is the correct answer.

1: **procedure** GAN TRAINING
2: **for** number of training iterations **do**
3: **for** k steps **do**
4: • Sample minibatch of m noise samples $\{\mathbf{z}^{(1)}, .., \mathbf{z}^{(m)}\}$ from noise prior $p_g(\mathbf{z})$
5: • Sample minibatch of m examples $\{\mathbf{x}^{(1)}, .., \mathbf{x}^{(m)}\}$ from data generating distribution $p_{data}(\mathbf{x})$
6: • Update the discriminator by ascending its stochastic gradient:

$$\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} \left[\log D_\theta\left(x^{(i)}\right) + \log\left(1 - D_\theta\left(G_\phi\left(z^{(i)}\right)\right)\right)\right]$$

7: **end for**
8: • Sample minibatch of m noise samples $\{\mathbf{z}^{(1)}, .., \mathbf{z}^{(m)}\}$ from noise prior $p_g(\mathbf{z})$
9: • Update the generator by ascending its stochastic gradient

$$\nabla_\phi \frac{1}{m} \sum_{i=1}^{m} \left[\log\left(D_\theta\left(G_\phi\left(z^{(i)}\right)\right)\right)\right]$$

10: **end for**
11: **end procedure**

6. Answer this question in the context of GANs.
When the generator is optimal what is the minimum loss value that the discriminator can achieve?

A. $-log1$

B. $-log2$

C. $-log3$

D. $-log4$

Solution: Option D is the correct answer.

7. Consider the following statements:

   I. Kulback-Leibler (KL) Divergence is a symmetric measure.

   II. Jensen-Shannon Divergence (JSD) is a symmetric measure.

   Which of the following options is correct?

   A. Only statement I is True.

   B. Only statement II is True.

   C. Both statements I and II are True.

   D. Both statements I and II are False.

Solution: Option B is the correct answer.

8. Which of the following models are capable of doing *Abstraction* as well as *Generation*?

I. Restricted Boltzman Machines (RBMS)

II. Variational Autoencoders (VAEs)

III. Autoregressive models (AR models)

IV. Generative Adversial Networks (GANs)

A. Only I and II

B. Only I and III

C. Only II and III

D. Only III and IV

E. Only I

Solution: Option A is the correct answer.

	RBMs	VAEs	AR models	GANs
Abstraction	Yes	Yes	No	No
Generation	Yes	Yes	Yes	Yes
Compute P(X)	Intractable	Intractable	Tractable	No
Sampling	MCMC	Fast	Slow	Fast
Type of GM	Undirected	Directed	Directed	Directed
Loss	KL-divergence	KL-divergence	KL-divergence	JSD
Assumptions	X independent given z	X independent given z	None	N.A.
Samples	Bad	Ok	Good	Good (best)

Table: Comparison of Generative Models

9. For any given generator G, what is the optimum discriminator $D^*_\theta(x)$ which will maximize the following objective function for GANs,

$$\min_{\phi} \max_{\theta} \int_x \left(p_{data}(x) \log D_\theta(x) + p_G(x) \log(1 - D_\theta(x))\right) dx$$

All notations carry their usual meaning as defined in the lecture.

A.
$$D^*_\theta(x) = \frac{p_{data}(x)}{p_G(x)}$$

B.
$$D^*_\theta(x) = \frac{p_G(x)}{p_{data}(x)}$$

C.
$$D^*_\theta(x) = \frac{p_{data}(x)}{p_G(x) + p_{data}(x)}$$

D.
$$D^*_\theta(x) = \frac{p_G(x)}{p_G(x) + p_{data}(x)}$$

Solution: Option C is the correct answer.