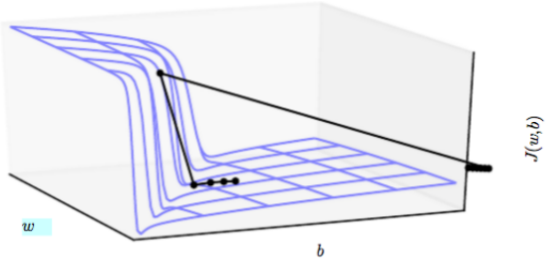
Instructions:

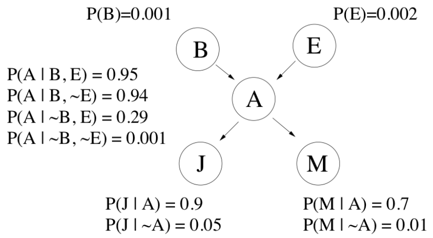
- Questions 1 to 8 are mandatory. [**5 Marks**]
- Answer any five out of the questions 9-16. [**10 Marks**]
- Answer any one question from 17, 18. [**3 Marks**]
- Question 19 is mandatory. [**5 Marks**]

1. **(0.5 Mark)** Assuming $x \in \mathbb{R}$, for what value of x will the logistic function $\frac{1}{1+e^{-(wx+b)}}$ output the value 0.5.
2. **(0.5 Mark)** Write down the update rules for RMSProp ?
3. **(0.5 Mark)** Consider the following corpus: "human machine interface for computer applications . user opinion of computer system response time . user interface management system . system engineering for improved response time". What is the size of the vocabulary of the above corpus ?
4. **(0.5 Mark)** What situation is being depicted in the following diagram (Hint: think gradient descent) ? Assume $J(w, b)$ is the loss function and w, b are parameters.



5. **(0.5 Mark)** Consider the joint distribution of two binary random variables A and B such that $P(A = a, B = b) = 0.25\ \forall(a, b) \in \{(00), (01), (10), (11)\}$. Write down all the independence relations (if any) which exist in this joint distribution.
6. **(1 Mark)** Consider a joint distribution over n binary random variables $X_1, X_2, ..., X_n$. There are several joint distributions possible such that each of them would entail a different set of independence conditions (for example, in some joint distributions X_1 and X_2 may be independent and in some they may not be). Suggest a Bayesian Network which will be an I-Map for any distribution over these n random variables ?
7. **(1 Mark)** Prove that the JSD between two distributions p and q is 0 if and only if $p = q$.
8. **(0.5 Marks)** Consider the Bayesian Network shown in the figure involving A, B, C, D, E which are binary random variables.

   Note that $P(B) = 0.001$ means $P(B = 1) = 0.001$. Similarly $P(A|B, \sim E) = 0.94$ means that $P(A = 1|B = 1, E = 0) = 0.94$. Compute $P(A = 1, B = 1, E = 0, J = 1, M = 0)$.



9. **(2 Marks)** When using gradient descent for training RBMs we encounter the following expression for the gradient w.r.t. a parameter w_{ij} :

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]$$

In contrastive divergence, we approximate the first expectation by a training instance and the second expectation by a sampled instance. Can you argue mathematically that this is the same as maximizing the likelihood of the training instance and minimizing the likelihood of the sampled instance.

10. **(2 Marks)** While deriving the expression for bias correction in Adam, we assumed that the gradients come from a stationary distribution (*i.e.*, $\mathbb{E}[g_t] = \mathbb{E}[g] \;\; \forall t$). Argue why or why isn't this assumption reasonable in practice.

11. **(2 Marks)** Consider the task of generating descriptions from structured data (similar to what you did in the RNN assignment). Another instance of this problem is to generate the description of a product given a structured table about the product. For example, Flipkart may have several products such as phones, laptops, refrigerators, TVs, *etc.* with tabular data about each product. When Pappu trained a vanilla encode-attend-decode model for this task he observed that in many cases the description generated by the model does not cover all the fields mentioned in the product table (brand, OS, RAM size, storage, battery life, etc). Further, in some cases the model repeats a phrase in the description. For example, the model generates the following description "the mobile has 2GB RAM, a 6 inch screen with dual camera and a 2GB RAM". Can you suggest a refinement to the vanilla encode-attend-decode framework to explicitly ensure that maximum number of fields from the input are covered in the description and a given field is not repeated twice in the description. Mention data, model, parameters, objective function and learning algorithm.

12. Consider a time homogeneous Markov Chain involving 2 binary random variables and hence 4 states. At every time step, there is a transition in the state as determined by the transition matrix T.

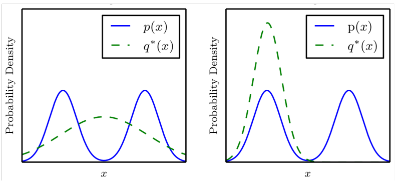
    (a) **(0.5 Mark)** Suggest a transition matrix for which the chain will **not be irreducible**.

(b) **(0.5 Mark)** Suggest a transition matrix for which the chain will **not be aperiodic**.

(c) **(1 Mark)** Suggest a transition matrix for which the chain will **be irreducible but not aperiodic**.

For each of the above cases if you feel such a transition matrix does not exist then explain why does it not exist.

13. **(2 Marks)** Suppose the true data distribution p is a mixture of two Gaussians as shown in blue in the figure (here $x \in \mathbb{R}$). Of course, you 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-Divergence*$(p||q)$ and *minimize KL-Divergence*$(q||p)$. You are shown two figures below in which the green curve is the learnt distribution q. Looks like you get two different solutions based on the objective function that you choose. Identify the objective function each figure corresponds to and justify your answer.



14. We studied NADE in the context of binary random variables. How would you adapt the NADE architecture if instead of binary variables you have

(a) **(1 Mark)** k-ary random variables (*i.e.*, each X_i can take k values)

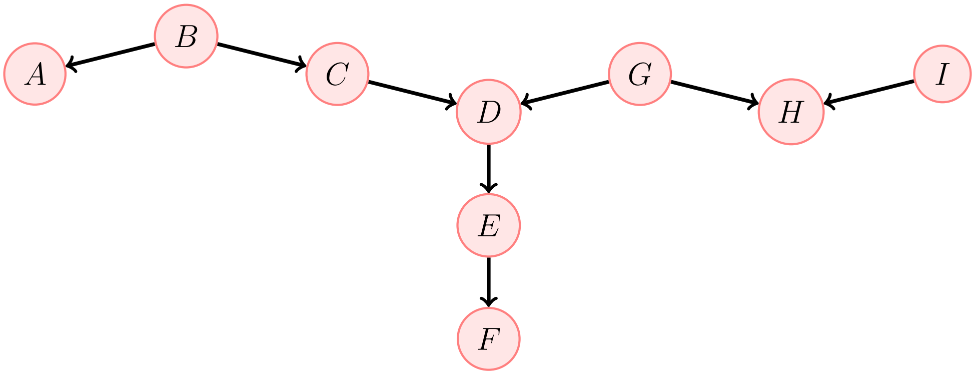
(b) **(1 Mark)** continuous random variables (*i.e.*, each $X_i \in \mathbf{R}$)

Feel free to make certain assumptions but explain your assumptions clearly.

15. **(2 Marks)** While training GANs, the only signal that the generator receives is the probability that the discriminator assigns to a sample generated by the generator. Suppose that the generator and the discriminator have the same number of hidden layers. Of course the output layers of the two networks would still be different because the generator needs to produce an n-dimensional output whereas the discriminator only needs to produce a single output (probability). Given this situation that the generator and discriminator have the same number of hidden layers, suggest an additional signal that can be fed to the generator while training. (Hint: We are asking you to modify the loss function).

16. **(2 Marks)** Answer the following question in the context of GANs. What if we completely ditch the discriminator and have a generator which does the following: (i) Pick any image from your training data, say x_i (ii) Sample a value for z (iii) Pass this z through the generator network and generate an image $\hat{x}_i$ (iv) Minimize the squared error loss between $\hat{x}_i$ and x_i (no discriminator). Justify or argue against the above generator.

17. **(3 Marks)** Consider the Bayesian Network shown in the figure below.



Let $\mathbf{O} \subseteq \{A, B, C, D, E, F, G, H, I\}$ be the set of all observed variables, *i.e.*, the set of all variables whose values are known. Now consider the path between A and I. This path is said to be active if every consecutive triple of variables X, Y, Z on this path satisfies one of the following conditions:

- The path between X, Y, Z is of the form $X \rightarrow Y \rightarrow Z$ and Y is **unobserved** $(Y \notin O)$
- The path between X, Y, Z is of the form $X \leftarrow Y \leftarrow Z$ and Y is **unobserved** $(Y \notin O)$
- The path between X, Y, Z is of the form $X \leftarrow Y \rightarrow Z$ and Y is **unobserved** $(Y \notin O)$
- The path between X, Y, Z is of the form $X \rightarrow Y \leftarrow Z$ and Y or any of its descendants is **observed**

Given a set of observed variables $\mathbf{O}$, A is said to be independent of I if there is no active path between A and I. Write down the condition under which A is independent of I. (Note: When we say observed variables we mean the variables which are given. For example, when we say X is independent of Y given Z we mean Z is observed.)

18. In VAEs, we assumed that z comes from a unimodal Gaussian distribution $\mathbb{N}(\mu, \Sigma)$

    (a) **(1 Mark)** Give an example where such an assumption would be unrealistic

    (b) **(2 Marks)** How would you modify VAE to deal with situations where z comes from a mixture of Gaussian distributions ? (you need to think about how to sample a z to pass to the decoder and how to implement the reparameterization trick.)

19. **(5 Marks)** Suggest a problem of social importance in the Indian context where AI/Deep Learning can help. For example, we can use a CNN based image classifier to predict the diseases that have affected a plant by looking at a real time image of the plant (say, taken by a farmer and uploaded on your app). Suggest other such applications of AI for social good outside the agriculture domain. Marks will be given based on (i) novelty of the problem (ii) availability of data or novel ideas for collecting data for training the model, (iii) potential social impact (iv) details of the solution and (v) feasibility of deployment.