Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur
CSL7590 - Deep Learning

April 7, 2025

Instructions:
Read the questions carefully. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 40 min | *Quiz-3* | Maximum Marks: 10

1. Choose the correct statement(s): [1]

   a) GAN is an implicit distribution estimation model
   b) The generator and the discriminator plays a cooperating game while training
   c) The generator and the discriminator are trained in an alternating fashion
   d) The GAN will fail to learn if few of the labels of the generated images and the real images are flipped while training the discriminator

2. In GANs, discriminator ($D(x)$) is generally trained while assigning label 0 to generated samples ($G(z)$) and label 1 to real samples. Considering this, answer the following. [1]

   [True/False] GAN training completes as soon as $D(G(z))$ becomes close to 1.

3. Which problem of GAN is solved by the non-saturating loss. Why does the problem occur and how does the non-saturating loss solve it? [4]

4. Show that, at the convergence of a standard GAN, the discriminator accuracy would be equal to 50%. [4]

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

CSL7590 - Deep Learning

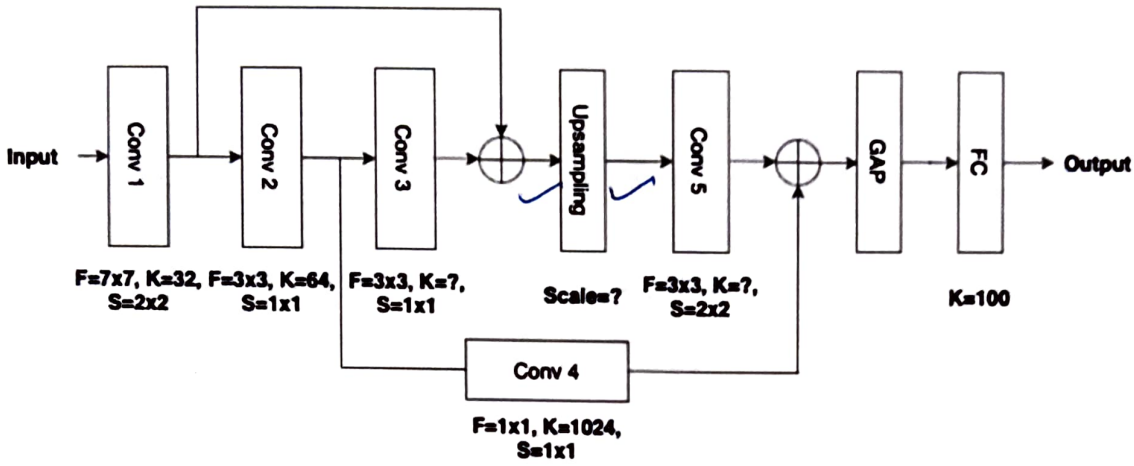
February 20, 2025

Instructions:
1. Read the questions carefully.
2. All questions are mandatory.
3. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 2 hour | *Minor* | Maximum Marks: 30

1. [True/False] Unsupervised pretraining helps in weight initialization. Justify your answer. [2]
2. [True/False] Applying ReLU activation before or after max pooling has no difference. Justify your answer. [2]
3. What is the limitation of AdaGrad? How is it addressed by RMSProp? [2]
4. How does LSTM solve the vanishing gradient problem of RNN?. [3]
5. Explain max-pooling layer and its operation during backpropogation. Describe a scenario where convolutional layer with max-pooling is better as compared to convolutional layer with stride. [4]
6. Recall the self-attention module discussed in the class and consider the combination of weight matrices of different sizes. Which of the following combinations is/are not possible choice/s for realizing a self-attention module and why? [4]
   1. $W_q \in \mathbb{R}^{d_1 \times d_2}$, $W_k \in \mathbb{R}^{d_1 \times d_2}$, $W_v \in \mathbb{R}^{d_1 \times d_2}$
   2. $W_q \in \mathbb{R}^{d_1 \times d_2}$, $W_k \in \mathbb{R}^{d_1 \times d_3}$, $W_v \in \mathbb{R}^{d_1 \times d_3}$
   3. $W_q \in \mathbb{R}^{d_1 \times d_2}$, $W_k \in \mathbb{R}^{d_1 \times d_2}$, $W_v \in \mathbb{R}^{d_3 \times d_2}$
   4. $W_q \in \mathbb{R}^{d_1 \times d_2}$, $W_k \in \mathbb{R}^{d_1 \times d_2}$, $W_v \in \mathbb{R}^{d_1 \times d_3}$
   5. $W_q \in \mathbb{R}^{d_1}$, $W_k \in \mathbb{R}^{d_1}$, $W_v \in \mathbb{R}^{d_1 \times d_2}$
7. Consider a LSTM cell at time step t with $h_{t-1} = [0\ 1]^T$ and $c_{t-1} = [-1\ 0.5]^T$. Weight matrices contain following values. $W_{fx} = \begin{bmatrix} 0 & -1 & 0 \\ -2 & 0 & -2 \end{bmatrix}$, $W_{fh} = \begin{bmatrix} 0 & -1 \\ -2 & 0 \end{bmatrix}$, $W_{ix} = \begin{bmatrix} 2 & 0 & 3 \\ 0 & 1 & 0 \end{bmatrix}$, $W_{ix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $W_{ox} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$, $W_{oh} = \begin{bmatrix} 0 & -1 \\ -2 & 2 \end{bmatrix}$, $W_{gx} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & -1 \end{bmatrix}$, and $W_{gh} = \begin{bmatrix} 0 & -1 \\ 0 & -2 \end{bmatrix}$. All bias vectors are zero. Given $x_t = [1\ 0\ 1]^T$ and $x_{t+1} = [1\ 0\ 2]^T$, find h_t, h_{t+1}, c_t, and c_{t+1}. You can round off all intermediate and final results to 1 decimal place. [6]

8. Consider a CNN shown on the right where Conv and FC represent convolutional and fully connected layers, respectively. GAP is the global average pooling layer. $\oplus$ is element-wise addition operation. F, K, and S represent filter size, number of filters/neurons, and stride, respectively. Each Conv layer uses zero padding of appropriate size. Answer the following for the input size (W×H×C) = 64×64×3 while showing calculation steps. [7]



1. Calculate the size of feature (activation) map before and after the Upsampling layer.
2. Calculate the value of K for Conv 3 and Conv-5 layer.
3. Calculate the total number of trainable parameters while assuming that there are no parameters in the Upsampling layer.
4. If GAP is replaced with a FC (K=2048) layer, what will be the total number of parameters now?

[A GAP layer refers to a "Global Average Pooling" layer, which calculates the average value of feature maps across the spatial dimensions, producing a single feature value per channel.]

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

CSL7590 - Deep Learning

April 25, 2025

Instructions:
1. Read the questions carefully.
2. If you find anything unclear/incorrect, make a reasonable assumption and proceed.

Time: 3 hour | *Major* | Maximum Marks: 50

1. Which among the following is computationally the most inefficient method for model compression? [1]

a) Knowledge Distillation b) Weight Pruning
c) Neural Architecture Search d) Quantization

2. Considering the traditional data-free knowledge distillation, which of the following is/are true? [1]

a) Teacher is released with weights b) Original data is available during distillation
c) Synthetic data is available during distillation d) Teacher is trained on synthetic data

3. [True/False] Unsupervised training guides the learning toward basins of attraction of minima that support better generalization from the training set. [1]

4. [True/False] An autoregressive model predicts the next component in a sequence by taking measurements from previous components in the sequence. [1]

5. [True/False] Co-occurrence matrix based word embedding is an example of distributed representation. [1]

6. How does ProxylessNas solve the memory problem of DARTS? [2]

7. Let us consider an autoencoder where $\boldsymbol{z}$, which represents the output of the encoder, is quantized into a corresponding vector ($\hat{\boldsymbol{z}}$) of 0's and 1's as [2]

$$\hat{\boldsymbol{z}}(i) = \begin{cases} 1 & \text{if } \boldsymbol{z}(i) > \tau \\ 0 & \text{if } \boldsymbol{z}(i) \leq \tau \end{cases}$$

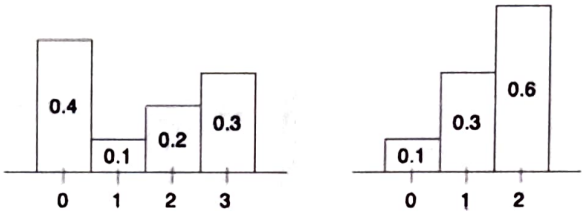
where $\boldsymbol{z}(i)$ is the i^{th} element of the vector $\boldsymbol{z}$ and τ is a user-defined threshold. $\hat{\boldsymbol{z}}$ is fed into the decoder for reconstruction. Do you observe any challenge in training of this autoencoder? Suggest a way to overcome the challenge.

8. How does ResNet handle the problem of vanishing gradient? Can the ResNet architecture be considered an ensemble? Justify your answer. [4]

9. Given three feature vectors $x_1 = [1\ 0\ 1\ 0\ 0]^T$, $x_2 = [0\ 0\ 2\ 0\ 1]^T$, and $x_3 = [2\ 1\ 0\ 0\ 0]^T$, find the updated feature vectors obtained after applying self-attention. Weight matrices contain following values. $W_Q = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \end{bmatrix}$, $W_K = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}$, and [6]

$W_V = \begin{bmatrix} 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 & 10 \\ 10 & 0 & 0 & 10 & 0 \end{bmatrix}$. You can round off all intermediate and final results to 1 decimal place. Also assume that $\exp(x) = 3^x$ for simplicity.

10. Consider the distributions of two random variables r and s shown below. [4]



Given the two transport plans (joint distributions) below, γ_1 and γ_2, find the corresponding transportation costs of transforming one distribution into the other.

γ_1

r \ s			
	0.1	0.3	0
	0	0	0.1
	0	0	0.2
	0	0	0.3

γ_2

r \ s			
	0.1	0	0.3
	0	0.1	0
	0	0.2	0
	0	0	0.3

11. Consider following sample types for an autoencoder and suggest possible choices of output activation and loss functions with justification. [6]

    1. Binary images
    2. Non-negative real-valued vectors
    3. Samples containing real values scaled in the range of -1 to +1.

12. Recall the ELBO maximization with respect to the variational posterior ($q(z|x)$) and the parameters of the decoder (θ). Show that the optimal variational posterior is same as true posterior *i.e.* $q^*(z|x) = p_\theta(z|x)$. [5]

13. Consider the problem of unsupervised domain adaptation where our goal is to train a feature extractor (f_θ) such that the distribution of target features ($\mathbb{P}_T$) matches with the source features distribution ($\mathbb{P}_S$). If we use a domain classifier for this purpose, model the problem as f-divergence minimization and derive the corresponding optimization objective. [6]

14. Recall your course project and answer the followings. [10]

    1. Write the problem statement of your project clearly. [1.5]
    2. Explain the methodology of your work. [4]
    3. Describe the novelty of your work. [1.5]
    4. Which components of your approach have critical dependency on deep learning and why? [3]

IIT Jodhpur	CSL-7480	Cryptography
Minor Exam	Semester 2, 2024-25	Instructor: Somitra Sanadhya
Time: 2 hours	2–4 pm, 21.02.25, LHC, Room 110	Max. Marks: 15

Note:

1. No queries will be entertained during the test.
2. If you do not understand any question then read it again.
3. Marks for each question/part are mentioned in the rightside margin. The total marks of each question are mentioned at the end of the question. A table showing the marks of all the question is provided at the end of the paper.

Notation: For n-bit strings x and y,

- $x \| y$ denotes concatenation of the two strings,
- $x \oplus y$ denotes bit-wise XOR of the two strings,
- $x + y$ means adding x and y and then taking remainder modulo 2^n (to ensure that the sum can be represented in n bits).

---

(1) 1. Data compression is often used in data storage or transmission. Suppose you want to use data compression in conjunction with encryption. Does it make more sense to

(a) Compress the data and then encrypt the result, or

(b) Encrypt the data and then compress the result.

Justify your answer. Total for Question 1: 1

(1) 2. Your friend is learning Cryptography from you. You taught her the perfect security of the OTP scheme. She realizes that when the secret key is all-zeros, the encryption of the message m is m itself (since $m \oplus 0^n = m$, where m and c are n-bit long). She thinks that this is a bad idea. Hence, she makes a slight modification to the scheme, by discarding the all-zero key, but otherwise retaining the same scheme. More precisely, her modified OTP scheme is:

$$E(k, m) = \begin{cases} m \oplus k & \text{if } k \neq 0^n \\ m \oplus 1^n & \text{if } k = 0^n \end{cases}$$

Discuss whether her thinking is right. If yes, then why? If not, then reconcile it with the fact that $c = m$ when the secret key is all-zeros. Total for Question 2: 1

(1) 3. (a) Assume a user of a login system chooses a password which is either *abcd* or *bedg*. Suppose the password is kept in an encrypted form on the server using the shift cipher, and an attacker gets to see the resulting ciphertext. Show how the attacker can determine the user's password, or explain why this is not possible.

(1) (b) State Kerckhoff's principle. Why is it the right model to be used in Cryptography.

(1) (c) Formally define perfect security. What is the primary limitation of perfect security?

Total for Question 3: 3

4. Explain how will the decryption be affected if the 3rd bit in the 2nd block of the ciphertext gets corrupted during transmission. Consider the following modes of operations:

(1) (a) CBC mode

(1) (b) Counter mode

Total for Question 4: 2

5. Let G_1 and G_2 be any secure PRG's from $\{0,1\}^n$ to $\{0,1\}^m$ where $m > n$. Are the following constructions secure PRG's? Why or why not? Justify your answer.

(1) (a) $G'(s) = G_1(s) \oplus G_2(s)$

(1) (b) $G'(s) = G_1(s) \,||\, G_2(s \oplus 1)$

Total for Question 5: 2

6. Let P_k be a family of pseudorandom permutations indexed by the key k. That is, fixing a specific key k^* is equivalent to choosing a specific permutation $\pi_{k^*} \in P_k$ which is hard to distinguish from a random permutation.

For each of the following encryption schemes which encrypt $2n$ bit messages, state with justification whether the scheme is CPA secure? That is, either show an attack or prove that the scheme is secure. The $||$ symbol stands for concatenating two strings.

(1) (a) $E(k, m_1||m_2) = \pi_k(m_1) \,||\, \pi_k(m_2)$

(1) (b) $E(k, m_1||m_2) = \text{IV} \,||\, (\pi_k(\text{IV} + 1) \oplus m_1) \,||\, (\pi_k(\text{IV}+2) \oplus m_2)$
where IV is a random n bit string (and $\oplus$ and $+$ are as defined in the notation earlier).

Total for Question 6: 2

(1) 7. (a) Prove that a 2-round Feistel construction is not a PRP. (i.e. show an attack which breaks the PRP property).

(2) (b) Assume that you have a block cipher E which supports an n-bit key and has t-bit block size. To increase the key-size to $2n$ bits, your friend suggests using the following encryption algorithm:

$$E'(k_1||k_2, m) = E(k_1, E(k_2, m)).$$

Keys k_1 and k_2 are n bits long, and the message m is of length t bits. Show an attack against the encryption scheme above. Mention the time and space complexity of your attack in terms of n and t.

Total for Question 7: 3

(1) 8. Derive the "birthday bound" for Cryptographic hash functions. What is its impact on the design of such functions?

Total for Question 8: 1

Table of marks:

Question:	1	2	3	4	5	6	7	8	Total
Points:	1	1	3	2	2	2	3	1	15
Score:									

Write briefly. Best wishes.

CSL7480 Cryptography End-semester exam Instructor: Somitra Sanadhya	26 April 2025, 2-5 pm	IIT Jodhpur LHC-110 Max. Marks: 40

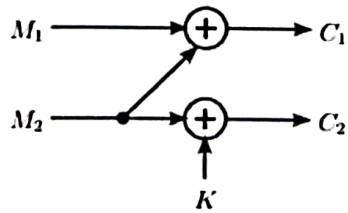
Roll number and Name: _____

Note:

1. The test has 2 parts. Both the parts are for 20 marks each.
2. Part A is meant to have short answers. I do not expect you to write more than one small paragraph in any question. Each question in this part carries 2 marks.
3. Part B has 5 questions of 4 marks each.
4. Write briefly and precisely. Do not waste your and my time with long-winded answers.
5. All answers should have valid reasoning. Writing the final answer, even when correct, does not lead to any marks.

Part A

1. Prove that (a) the shift cipher, and (b) the substitution cipher can be easily broken using a known-plaintext attack. How much known plaintext is needed to completely recover the key for each of these ciphers? (Assume that the underlying language is English with only lowercap letters being used.)
2. Fill in the blanks:
   (a) ..... is a way to create a stream cipher out of a block cipher.
   (b) A symmetric-key algorithm for ensuring that a message has not been tampered with is called .....
   (c) ..... is a widely used, standardized and secure hash function.
   (d) A value used in symmetric key cryptography to ensure that a new session that transmits the same text as a previous session does not result in identical ciphertext is called .....
3. Let the encryption of a $2n$-bit message $M = (M_1, M_2)$ be defined by the following figure.



   Is this encryption function perfectly secure? Why or why not?
4. If the second block of a long ciphertext gets corrupted and exactly two bits in this block are sent erroneously, what effect will it have on the decryption side if the CBC mode of operation is being used with AES?

5. State (no need to prove) the key complementation property of DES. What is the implication of this property on the security of DES?

6. An RSA cryptosystem has public key $n = 35$ and $e = 7$. Find the decryption exponent.

7. A bank manager uses Shamir's (2,3) secret sharing scheme, working modulo a prime 101. The shares created by the manager are (1,13), (3,12) and (2,*) where * represents unreadable data. What should be the value of * so that the scheme works correctly.

8. Given $n = 323737$ and $\phi(n) = 322596$, factorize n.

9. Let $n = p_1 \times p_2 \times p_3$ where p_i are prime numbers. How many square roots of 1 exist modulo n? Why?

10. State the encryption and decryption functions of the Goldwasser-Micali cryptosystem for the modulus $n = 35 = 5 \times 7$. Show the encryption of a bit b=0 for this modulus, and show how to perform the decryption of the generated ciphertext.

Part B

1. A b-bit block cipher with key K, denoted by $E_K(.)$, is used to construct a hash function $h(.)$ in the following way:

   First make the length of the input message m a multiple of b, by appending a single 1 bit and then enough 0's (i.e. the 10^* padding). Let the resultant message have n blocks and be denoted by $m_0||m_1||\ldots||m_{(n-1)}$ where each m_i is b bits.

   Run the following algorithm:

   $c = E_{m_0}(m_0)$
   for($i = 1$ to $n-1$){
   . $d = E_{m_0}(m_i)$
   . $c = c \oplus d$
   }
   Return $h(m) = c$

   Show that this is not a secure cryptographic hash function (by exhibiting an attack).

2. The ElGamal signature scheme is defined by the following 3 algorithms:

Algorithm 1 Key Generation

1: Choose a large prime p randomly.
2: Choose a generator α of Z_p^* randomly.
3: Choose a random integer $d \in \{2, 3, \ldots, (p-2)\}$.
4: Compute $\beta = \alpha^d \mod p$.
5: **Public key** $= (p, \alpha, \beta)$. **Secret key** $= d$.

Algorithm 2 Signature generation on message m

1: Choose a random number $k \in \{0, 1, \ldots, (p-2)\}$ such that $\gcd(k, p-1)$=1.
2: Compute parameters (r, s) as follows:
3: $r = \alpha^k \mod p$
4: $s = (m - d \cdot r)k^{-1} \mod (p-1)$.
5: **Return** $\mathrm{Sig}(m) = (r, s)$.

Algorithm 3 Signature Verification

1: **Input** message=m, signature $= (r, s)$. (Public key from Algorithm 1 is also known).
2: Compute $t = \beta^r \cdot r^s \mod p$.
3: The verification follows from:

$$t \begin{cases} = \alpha^m \bmod p & \Rightarrow \text{Valid signature} \\ \neq \alpha^m \bmod p & \Rightarrow \text{Invalid signature} \end{cases}$$

(a) Prove that the Signature scheme works correctly. That is, if a signature is produced by following Algorithms 1 and 2, then it will be verified by Algorithm 3.

(b) An "existential forgery" against a signature sceheme means the ability to create a valid signature on a random message, without knowing the secret key. Prove that one can create an Existential forgery against the ElGamal signature scheme.

3. Let $h : \{0,1\}^* \to \{0,1\}^n$ be a hash function which is preimage resistant, second preimage resistant, and collision resistant. Let $h' : \{0,1\}^* \to \{0,1\}^{n+1}$ be the hash function defined as follows:

$$h'(x) = \begin{cases} 0 \,||\, x & \text{if } x \in \{0,1\}^n, \\ 1 \,||\, h(x) & \text{otherwise.} \end{cases}$$

What can you say about the following properties of h' ?

(a) Preimage resistance

(a) Second preimage resistance

(a) Collision resistance

Note: $x \,||\, y$ denotes concatenation of strings x and y.

4. Let $E : y^2 = x^3 + Ax + B$ be an elliptic curve over a field K.

(a) Show that any point on E with y-coordinate equal to zero has order 2.

(b) Suppose $A = 0$. Show that any point on E with x-coordinate equal to zero has order 3.

Note: An element $a \in$ a group G with additive identity e is said to have order m if

$$a + a + \ldots (m \text{ times}) = e.$$

5. Consider the following protocol:

Input: RSA modulus n and a value $x \in QNR_n$.

The prover P knows the factorization of n and wants to prove that x is in QNR_n. (Recall that QNR_n is the set of elements which are quadratic non-residue modulo n. That is, there does not exist a z such that $z^2 \equiv x \pmod{n}$.

1. The verifier V forms a challenge value as follows: choose $c \in \{0,1\}$, choose random $r \in Z_n^*$. V sends $y = r^2 x^c$ to P.
2. The prover uses the factorization of n to determine whether $y \in QR_n$. If so, let $c' = 0$, otherwise let $c' = 1$. Send c' to V.
3. V checks that $c' = c$. If so, V accepts, otherwise V rejects.

Prove the following regarding this protocol:

(a) The protocol is sound. (Recall the definition of "soundness" for a zero-knowledge proof").

(b) The protocol is not zero-knowledge. That is, a cheating verifier can learn "something" by interacting with an honest prover. (Decide what is "something" by thinking along the definition of a zero-knowledge proof. If the proof is ZK then the entire process of the proof could be simulated by someone not having access to the secret information with the prover. If the verifier can learn anything more than what a simulator can, then this is "something").

*** End of the test paper ***

CSL 7030: Algorithms for Big Data Minor Exam

Date: 23/04/2025 2 to 5 PM Maximum points: 45

Formal mathematical proofs are expected; examples are not proofs. Partial credit may be given for ideas along the right lines. If you are using something proved in the class (such as an inequality, property etc.), then you should clearly mention the thing that you are using, and the reason why it is applicable.

1. Multiple Choice Questions: Each MCQ is worth 2 points. No points shall be awarded without a **short** valid explanation.

(i) Statement A: Picking a hash function from 2-universal hash family guarantees pairwise independence.
Statement B: Chernoff bound only requires pairwise independence.
Select the correct option.
(a) Statement A is true, statement B is false.
(b) Statement B is true, statement A is false.
(c) Both statements are true.
(d) Both statements are false.

(ii) What is the space complexity (in bits) is required for approximately counting up to m events in the streaming setting? (assume ϵ and δ are constant)
(a) $O(\log \log m)$ (b) $O(\log m)$ (c) $O(\frac{\log m}{\log \log m})$ (d) $O(\sqrt{m})$.

(iii) Statement A: Connectedness testing algorithm works in dynamic graph streaming model.
Statement B: Bipartiteness testing algorithm works in dynamic graph streaming model.
Select the correct option.
(a) Statement A is true, statement B is false.
(b) Statement B is true, statement A is false.
(c) Both statements are true.
(d) Both statements are false.

(iv) Out of the following options, select a valid combination of **number of machines** and **memory at each machine** in the MPC model. Here, N denotes the total input size.

(a) $\log N$, $\sqrt{N}$
(b) $\sqrt{N}$, $2^{\sqrt{\log N}}$ expo
(c) $N^{1/3}$, $\sqrt{N}$ small
(d) $N^{2/3}$, $\sqrt{N} \log N$. outside

(v) What is the space complexity (in bits) for finding approximate quantiles in the streaming setting? Here n is the size of universe and m is the stream length.
(a) $O(\frac{\sqrt{m+n}}{\epsilon^2 \delta})$
(b) $O(\frac{\log n + \log m}{\epsilon^2} \log(\frac{1}{\delta}))$
(c) $O(\frac{m \log n}{\epsilon} \log(\frac{1}{\delta}))$
(d) $O(\frac{\epsilon n \log m}{\delta})$.

Q2 to Q6 are descriptive questions, and each question is worth 7 points in total.

2. Let A, B, C be $n \times n$ matrices with 0/1 entries. We want to check whether $AB = C$ or not.

(a) Design a randomized algorithm for the problem that outputs the correct answer with probability at least $1 - \frac{1}{n}$. Analyze the correctness, running time, and probability. (5 points)

(b) Is the 0/1 assumption necessary? What changes (if any) are required in the algorithm so that it also works for matrices with real numbers? Give a proper explanation. (2 points)

3. L is a 0/1 array of length n. We want to check whether L is sorted in the **property testing model**. We are given two parameters $0 < \epsilon, \delta < 1$.

   (a) Design a sublinear algorithm for solving this problem with the following properties: (i) if L is sorted, then algorithm always outputs that L is sorted, and (ii) if L is ϵ-far from being sorted, then it should output "not sorted" with probability at least $1 - \delta$. What is the query complexity of the algorithm? (5 points)

   (b) Is the 0/1 assumption necessary? Why or why not? Give a proper explanation. (2 points)

4. Design an algorithm for finding a minimum spanning tree (MST) in the MPC model, using partitioning framework. Specify the values of different parameters, and prove the correctness of the algorithm.

Scenario for Q5 and Q6. You have joined the analytics team at QuantForge, a high-frequency trading firm processing millions of stock trades per second. Each trade is associated with the name of the company, a unique identifier, and other data that can be used for analysis, but is irrelevant for this question. You can assume that each stock trade can be stored using $O(\log n)$ bits of memory, where n is the total number of companies traded on the stock exchange.

Here is the situation: The number m of total trades throughout the day is **unknown** in advance and is **unbounded**. The monitoring system only has a fixed, limited amount of memory. Treat Q5 and Q6 as separate questions (don't use information from one question in another).

5. You want to find the list of *highly traded* companies for gaining insights into market trends. Specifically, you are given a parameter k, and you want to find those companies whose shares were traded more $\frac{m}{k}$ times.

   (a) Design a one-pass algorithm for finding a list containing all such companies. What is the space complexity of the algorithm? What is the property of the output of the algorithm, and what are the drawbacks? (4 points)

   (b) Suppose with the help of the central stock exchange, it is possible to make another pass over the trades happened throughout the day. Design an algorithm that makes the second pass over the stream, and overcomes the drawbacks of part (a). (3 points)

6. One day, a suspicious pattern among the trades is detected, suggesting that a malicious bot may be manipulating trades by occasionally injecting fraudulent transactions. To investigate, you need a uniform random sample of 1,000 **distinct** transactions at the end of the day, to check for fraud later.

   (a) Why is it infeasible to store all trades and take a sample at the end of the day? (1 point)

   (b) Design an algorithm that maintains a sample on-the-fly, such that each trade observed so far has an equal chance of being in the sample. (3 points)

   (c) Prove that the algorithm maintains a uniform sample of trades. Analyze time and space complexity of the algorithm. (3 points)

Hint: Misra-Gries for Q5, Reservoir sampling for Q6

Name: _____ Roll no.

CSL 7030: Algorithms for Big Data Minor Exam

Date: 23/02/2025 10:30 to 12:30 PM Instructor: Tanmay Inamdar Maximum points: 25

Formal mathematical proofs are expected; examples are not proofs. Partial credit may be given for ideas along the right lines. If you are using something proved in the class (such as an inequality, property etc.), then you should clearly mention the thing that you are using, and the reason why it is applicable.

1. 3 points
Consider the problem of counting the number of 1s in a 0/1 string of length n.
Prove that there cannot exist a *deterministic* algorithm that makes strictly less than n queries, and *always* outputs the correct answer. Be as mathematically formal as you can.

2. Total 6 points
Suppose there is an election involving 10 crore (10^8) voters. Each voter either supports Party A or Party B. We want to find an estimate $\widehat{n_A}$ of the true value n_A of the number of voters supporting Party A. For this, we will conduct a poll of sample size ℓ. The requirement is that, with at least 99% probability, $\widehat{n_A}$ should be within ± 1 lakh (10^5) from the true value of n_A.

   (a) Clearly state all the assumptions you will make in order to model this scenario mathematically, for solving part (b). For each assumption, explain why it is needed, and how realistic it is in the real world. (2 points)

   (b) Find the smallest value ℓ for the number of voters polled, which would ensure that your estimate satisfies the above property. (4 points)

3. Total 6 points
This question is about the sublinear algorithm for estimating the number of connected components in a graph $G = (V, E)$ where *every vertex has degree at most d*. Recall that n_v denotes the number of vertices in the connected component that contains the vertex v.

   (a) Let $u_1, u_2, \ldots, u_t$ be vertices chosen uniformly and independently at random, and let

$$S = \frac{n}{t} \sum_{i=1}^{t} \frac{1}{n_{u_i}}$$

   Show that the **expected value of S** is equal to the actual number of connected components in the graph G. (3 points)

   (b) Now, let $\widehat{n}_v = \min(n_v, q)$, where $q \geq 1$ is a threshold. Design an $O(d \cdot q)$ time algorithm which takes an input a vertex v and threshold q, and computes $\widehat{n}_v$ using neighbor queries. (3 points)

Turn over the page for Q4.

4. Total 10 points

Clustering: For a set of n points P and distances d satisfying *metric properties*, let κ denote the smallest radius, such that the entire set P can be clustered within a ball of radius κ centered at some point $p \in P$. In other words, κ is the smallest value such that, from some point $p \in P$, every point $q \in P$ is at a distance at most κ.

We are in the model where all $\binom{n}{2}$ distances are stored in the memory. We are allowed to ask queries of the form: "What is $d(p, q)$?" for any pair of points p and q.

(a) (5 points)

- Design an algorithm of query complexity $O(n)$ that outputs an estimate ℓ, such that $\kappa \leq \ell \leq 2 \cdot \kappa$.
- Prove why your algorithm's estimate satisfies the above property.
- Argue why it is a sublinear time algorithm.

(b) (5 points)

Now, suppose we have an additional information that input point set P actually comes from **three dimensional Euclidean space**. We are also given an ϵ such that $0 < \epsilon < 1$.

- Design an algorithm that outputs an estimate ℓ', such that $\kappa \leq \ell' \leq (1 + \epsilon)\kappa$.
- Prove why your algorithm's estimate satisfies the above property.
- Prove that the query complexity of the algorithm is $O(n/\epsilon^6)$.

Minor Examination

Course Name: Natural Language Understanding **Code:** CSL 7640

Total Scores-40 **Time:** 2 hours

Make reasonable assumptions as and whenever necessary. You can answer the questions in any sequence. However, answers of all the parts to any particular question should appear together.

Q1 (a). Explain how ambiguity in natural language is different/similar to the ambiguity in formal language. Give examples of lexical, and syntactic ambiguities (*scope, prepositional phrase attachment*), and state how these ambiguities could pose challenges in further processing. Explain how semantic information and context could help in resolving such ambiguities. **[3+ 5 + 3]**

Q1(b). Explain with examples and intuition how discourse level tasks introduce more complexities in comparison to semantics and syntactic level tasks. **[4]**

Q2(a). Consider following four tags for the Part-of-Speech (PoS) tagging problem.

N-Noun; A-Adjectives; V-Verb; O-Others

Historically, AI was used to understand and recommend information. Now, Generative AI can also help us create new content. Generative AI builds on existing technologies, like Large Language Models (LLMs) which are trained on large amounts of text and learn to predict the next word in a sentence. For example, "peanut butter and ___" is more likely to be followed by "jelly" than "shoelace"

Assign appropriate PoS tags to every token in the above sentences. **[5]**

Q2(b). Consider a first order HMM, and compute the transition and emission probabilities from the tagged sequence of Q2(a) (***treating this as a training set***).

Show how using Viterbi decoding, the sequence of tags will be assigned to the following sentence (***make appropriate assumption as and when required***):

AI was used to understand and recommend information.. **[8+3]**

Q3 (a). Considering proper nouns as the potential named entities, mark the named entities for the example paragraph of Q2 (a). Use the BIO (*where B, I and O denote the beginning, Intermediate and outside of NE*) notation for marking the boundaries of the NEs.

Q3(b). Explain with examples how contextual information influences Named Entity Recognition.

Q3 (c). Give an example where context could help in NE classification.

[4+3+2]

Best of Luck

Major Examination

Course Name: Natural Language Understanding
Code: CSL 7040

Full Marks-60 **Time:** 3 hours

Make reasonable assumptions as and whenever necessary. You can answer the questions in any sequence. However, answers of all the components to any particular question should appear together.

Q1(a). *Attention in NMT is equivalent to Alignment in SMT.* Explain this with proper intuition.

Q1 (b). Consider the following pair of parallel sentences in English-Hindi

India is a diverse country→भारत एक विविधतापूर्ण देश है (transliterated form: *bhaarat ek vividhataapoorn desh hai*)

Construct the word alignment table, and subsequently the phrase table from the word alignment information.

Q1(c). Consider a Pivot based NMT system (with three languages). What kind of transfer learning strategy will be adapted when the source is a more distant language to pivot than the target; and the target is a more distant language to pivot than the source? (*explain with appropriate architectural diagrams and necessary steps*). **(3+8 +6)**

Q2 (a). With an example of product review sentence (two aspects, each having different sentiment classes, *viz.*positive and negative), describe how a bi-directional LSTM by feeding target specific encoded representations at every time step help improve the performance over a model that does not make target specific representations (*for aspect based sentiment analysis*)?

Q2 (b). Show an example scenario, where a single word exhibits two different sentiments.

(6+2)

(Q3). (a). For semantics, why is the count based method (e.g. tf-idf) not as robust as the prediction based method (e.g.. word2vec)?

(Q3). (b). Explain with necessary architectural diagrams and optimization criteria how a Skip-gram word embedding model be implemented using CBOW based word embedding model. **(3+7)**

Q4(a). What is the intuition behind treating every task of NLU as a next word prediction problem in LLM?

Q4 (b). *Temperature sampling and top-k sampling for LLM generation are equivalent-* Justify in favor or against this claim with proper intuitions and explanations.

Q4(c). How caching of attention scores done in LLM and how does it benefit? How parameter efficient training is carried out in LLM? **(3+4+6)**

Q5 (a). Determine the coreference chain from the following text:

The development of AI has created new opportunities to improve the lives of people around the world, from business to healthcare to education. It has also raised new questions about the best way to build fairness, interpretability, privacy, and safety into these systems.

Q5 (b). Create the instances (*positive and negative*) for training of a machine learning based anaphora/coreference resolution model for the example of Q5 (a) (*consider noun phrases as the possible markable*).

Q5(c). For every instance of (b) above, extract the features to satisfy the constraints (e.g. number agreement, gender agreement etc). **(3+5+4)**

All the best

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur
CSL7590 - Deep Learning

March 19, 2025

Instructions:
Read the questions carefully. If you find anything unclear/incorrect in any question, make a reasonable assumption and proceed.

Time: 40 min | *Quiz-2* | Maximum Marks: 10

1. For a transfer learning task, which layers are generally transferred to another task? [1]

a) Higher layers (close to o/p) b) Lower layers (close to i/p)
c) Task Specific d) None

2. A sentiment predictor is trained on customer reviews of media content such as books, videos and music, it is then used to analyze comments about consumer electronics such as televisions or smartphones, this scenario describes? [1]

a) Concept drift b) Domain adaptation
c) Multi task learning d) None

3. Which of the following is an effective way of network compression? [1]

a) Knowledge distillation b) Low-precision arithmetic, like converting a float to an int.
c) Weight pruning d) Pruning of model activations

4. Which of the following can be chosen as a pretext image task for SSL? [1]

a) Rotation b) Jigsaw puzzles
c) Noise removal d) Reconstruction

5. How does self-supervised pre-training benefit downstream tasks like image classification? [1]

a) By making the model's predictions more uncertain b) By adding noise to the training data
c) By providing a better initialization point for the model d) By reducing the model's capacity

6. Recall the rotation based pretext task for self-supervised pretraining on image data. Suggest a way to realize a similar rotation based pretext task for audio data. [1]

7. What is 'dark knowledge' in context of knowledge distillation? How is it transferred from teacher model to student model? [2]

8. Dynamic Network Surgery uses two thresholds to control pruning, a_k and $b_k = a_k + t$ where t is a pre-defined margin. Discuss the impact of large and small values of a_k and t. [2]