Linear Algebra and its Applications
Final Examination

(**Note**: You may use a calculator, but not your phone)

1. For each of the following statements, give a very short proof (not more than 10 to 12 words) if it is true or give a counter example to show it is false. [10 points]

   (a) Let A, B be two matrices such that their product AB is defined. Then
   $$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}.$$

   (b) Let A be a real matrix. Then
   $$\ker(A^T A) = \ker(A).$$

   (c) Let λ be an eigenvalue of a square matrix A. Then
   $$|\lambda| \leq \sigma_1,$$
   where σ_1 is the largest singular value of A.

   (d) Let S be a symmetric positive definite matrix. Then the positive square root of an eigenvalue of S is a singular value of S.

   (e) The rank of a matrix is the number of nonzero eigenvalues.

2. Find the (approximate) dominant eigenvector of the matrix [10 points]
   $$A = \begin{bmatrix} 4 & 5 \\ 6 & 5 \end{bmatrix}$$
   using power iteration. Start with the vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and perform 4 iterations, with scaling. Use this approximate eigenvector to calculate the dominant eigenvalue.

3. Consider the matrix [30 points]
   $$A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

   (a) Find QR decomposition of A using the Gram-Schmidt algorithm.

   (b) Find the singular value decomposition of A (i.e., $A = U_{3\times3}\Sigma V_{2\times2}^T$).

   (c) Write down the rank 1 and rank 2 approximations of A using singular values of A.

   (d) Write down an orthonormal basis for the following four fundamental spaces: row space of A, column space of A, null space of A and null space of A^T.

   (e) What is the pseudoinverse of A?

4. Let A be an $m \times n$ tall matrix with linearly independent columns. Use the reduced QR decomposition of A to define an expression for $A^\dagger$, the pseudoinverse of A. Explicitly prove that the Moore-Penrose criterion are satisfied by your formula. [10 points]

   - $(AA^\dagger)A = A$.
   - $(A^\dagger A)A^\dagger = A^\dagger$.
   - $(AA^\dagger)^T = AA^\dagger$.
   - $(A^\dagger A)^T = A^\dagger A$.

   For A in Problem 3 above find $A^\dagger$ using your formula.

ALGO

End-semester Exam

April 25, 2023

- This exam has 3 questions for a total of 100 marks. All answers will be evaluated.
- Read each question *carefully* before attempting it. You will not get *any* credit for writing stuff that is vaguely related to the question but does not match the requirements stated in the question, *even if* the stuff that you wrote is correct.
- Use a pen to write your answers; do *not* write your answers with a pencil.
- Write *legibly*. If you strike something off, strike it off clearly. Do *not* overwrite text that you wish me to read. (Parts of) Answers that are not easy to read, or can be interpreted in more than one way, will **not** be evaluated.
- Warning: CMI's academic policy regarding cheating applies to this quiz.

Unstated assumptions and lack of clarity in solutions can and will be used against you during evaluation. Please ask the invigilator if you have questions about the questions.

1. Let A be an array of integers with an odd number of elements $|A|$. The *median* of A is defined as the element that would appear at position $(|A|+1)/2$, counting from 1, in a *sorted version* of A. In the next two questions A is an array of integers with an odd number of elements. Array A is *not necessarily* sorted, and it may—or may not—contain repeated elements. You may invoke $\mathrm{len}(A)$ to get the number $|A|$, if you wish. Assume that one call to $\mathrm{len}(A)$ takes $\mathcal{O}(1)$ time. You may assume that accessing any one element of A, and each operation involving a constant number of integers, take $\mathcal{O}(1)$ time. Array indexing starts with 1.

   (a) Write the pseudocode for a *deterministic* procedure $\textsc{IsMedian}(A, x)$ that takes an array A and an integer x as arguments, runs in time $O(|A|)$, and checks whether x is the median element of A. Note that A, x are the *only* arguments to the procedure; pseudocode that assumes any other argument(s) will not get you credit. Explain why your procedure is correct, and why it runs within the required time bound. [10]

   You will get the credit for this answer only if all three parts—code, explanation, and running time analysis—are correct.

   (b) Write the pseudocode for a *randomized* procedure $\textsc{GetBigElement}(A, k)$ that takes an array A and an integer k as arguments, runs in time $O(k)$, and returns an element x of A which is *at least as large* as the median element of A. Your procedure should succeed with a probability at least $f(k)$ which (i) is a function of k alone (and is independent of $|A|$), and (ii) becomes closer to 1 as k grows larger. [20]

What is the function $f(k)$? Explain why your procedure succeeds with probability at least $f(k)$, and why it runs within the required time bound.

Note that A, k are the *only* arguments to the procedure; pseudocode that assumes any other argument(s) will not get you credit. You may assume that picking a random number takes $\mathcal{O}(1)$ time.

You will get the credit for this answer only if all three parts—code, analysis of success probability, and running time analysis—are correct.

2. A *cut* of an undirected graph $G = (V, E)$ is any partition of V into two non-empty parts, say S and $V \setminus S$. This cut is denoted $(S, V \setminus S)$. The *cutset* corresponding to the cut $(S, V \setminus S)$ is the set of all edges with exactly one end-point in S. In the following, let G be a *connected* graph with at least two vertices.

   (a) Prove that deleting all the edges in *any one* cutset of G, makes G disconnected. [10]

   (b) Let $F \subseteq E$ be a set of edges such that: [20]

   - Deleting all the edges in F from G makes G disconnected, and,
   - There is *some* proper subset $F' \subsetneq F$ such that deleting all the edges in F' from G also makes G disconnected.

   Is it true that such an F is always a cutset corresponding to some cut of G? Justify your answer. If your answer is YES, the justification should be a proof. If your answer is NO, it should be a counter-example. **You will get the credit for this answer only if your justification is also correct.**

3. For a fixed positive integer k, an instance of the MAX kSAT problem is a Boolean formula in conjunctive normal form (CNF) with n variables and m clauses, where each clause contains exactly k literals. The task is to find an assignment of truth values—$\{0, 1\}$—to the variables that satisfies as many clauses as possible.

   Consider an algorithm RANDMAXKSAT that takes as input an instance of MAX kSAT with n variables and m clauses where $m \geq 2^k$, and assigns a value 0 or 1 to each variable uniformly and independently at random. Thus each variable has probability exactly $\frac{1}{2}$ of getting the value 1. An assignment of truth values to the variables is said to be *heavy* if it satisfies at least $\lfloor (m+1)(1 - \frac{1}{2^k}) \rfloor$ clauses.

   Kuch kaam nahi aaya

   (a) Show that the probability that a given run of this algorithm will find a heavy assignment, is at least as large as $\frac{1}{m+1}$. [20]

   (b) Describe a randomized algorithm that, given an input an instance of MAX kSAT with n variables and $m \geq 2^k$ clauses, finds a heavy assignment for this instance with probability at least half. Explain how your algorithm achieves this probability of success. [20]

Data Mining and Machine Learning 2023

Final Exam

Chennai Mathematical Institute

26 April 2023, 09:30–12:30 (3 hours) Marks: 50, Weightage: 50%

1. Our task is to build a classifier to predict if a person is at risk from diabetes. The training set has patient data with age, gender, presence or absence of 1000 genes in their DNA, and whether or not they have diabetes.

   (a) Describe how to build and use a Naive Bayes classifier for this dataset. Mention any assumptions that you have made. *(5 marks)*

   (b) Suppose we are given additional patient data which does not include information about whether each individual has diabetes or not. Can we use this additional data to improve your classifier? If so, how? *(5 marks)*

2. We are studying a species of fish found in the Indian Ocean. This species has evolved into sub-species based on the region of the ocean that they live in. We have a dataset of latitudes and longitudes where this species of fish has been observed. How would we estimate the number of sub-species of this fish species? *(5 marks)*

3. We have a dataset of customers who have taken personal loans from a bank over the past 30 years. The data about each customer consists of their age, loan amount, address and whether or not they have paid back the loan in full. Additionally, for the past 5 years, we have data about whether the customer was a salaried employee or a businessman.

   The bank wishes to estimate the probability of full repayment of the loan for salaried employees. How can we use expectation-maximization to estimate this probability based on the entire data over 30 years? *(5 marks)*

4. We have a dataset of 10000 rice farms growing the same variety of rice. For each farm, we have data about how much fertilizer was used per acre and the yield per acre (in kgs). In addition, for 100 equally spaced days between planting and harvesting, there is data about the water-level (in cm), the average temperature and the average height of the rice plants.

   (a) Explain how to build a model that predicts the yield of rice, given all the other data about a farm. *(2 marks)*

   (b) We wish to run our ML model on a mobile phone (in an app). How can we reduce the computational cost without compromising too much on accuracy? *(3 marks)*

5. Suppose we have a neural network with four input features x_1, x_2, x_3, x_4 and a single output y. As usual, we assume that each pair of adjacent layers is completely connected and there is a single output layer. How many parameters do we have to estimate in the following situations?

   (a) A shallow network with 1 hidden layer consisting of 18 nodes.

   (b) A deep network with 3 hidden layers, where the first two layers have 3 nodes each and the third layer has 2 nodes. *(5 marks)*

✓ 6. We made the following assumptions about the loss (cost) function C for neural networks.

- For each input x, $C(x)$ is a function of only the output layer activation.
- The total cost across the training set is the average of the individual input costs.

Explain why these assumptions are important for effective learning of the parameters.

(5 marks)

✓

7. (a) For $z = wx + b$, how does the shape of the sigmoid function $\sigma(z) = (1 + e^{-z})^{-1}$ vary with w and b?

   (b) Given two input features x_1, x_2, explain how to construct a neural network to approximate a "rectangular box" function $g(x_1, x_2)$ with height h for $\ell_1 \leq x_1 \leq r_1$ and $\ell_2 \leq x_2 \leq r_2$. In other words, the function to be approximated is the following.

$$g(x_1, x_2) = \begin{cases} h & \text{if } \ell_i \leq x_i \leq r_i, i \in \{1, 2\} \\ 0 & \text{otherwise} \end{cases}$$

(5 marks)

✓ 8. Consider a neural network that is layered and completely connected. Suppose we initialize two nodes n_1 and n_2 from the same layer with the same biases and same weights on incoming and outgoing edges. What can we say about the final weights and biases that will be learned for n_1 and n_2 through backpropagation? What can we conclude about initialization strategies for such networks?

(5 marks)

✓ 9. In a nuclear power station, an alarm is triggered when a temperature gauge exceeds a given threshold. The gauge measures the temperature of the core of the reactor. Consider the boolean variables A (alarm sounds), F_A (alarm is faulty), and F_G (gauge is faulty) along with multivalued variables G (gauge reading) and T (actual core temperature).

   (a) Draw a Bayesian network for this scenario, given that the gauge is more likely to fail when the core temperature gets too high. Explain the structure of your network.

   (b) Suppose G and T each take just two values, normal and high. Assume that the gauge gives the correct temperature with probability x when it is working and with probability y when it is faulty. Describe the conditional probability table for G.

(5 marks)

moving computation is cheaper than moving data

n=3, K=3 Mod

CMI/22/4/2023/Final Exam

Apr 2023

200 quadrillion cal
= 200 petaflops.

(Note:)

Chennai Mathematical Institute

Distributed Computing and Big Data

sequence files (4)

DURATION: 3 HOURS

MAX MARKS: 35.

→ Inodes ch1 concept.
(le 3 file systems)

Instructions

- Please remember to mention your name and roll number in your answer sheet.
- This is an individual task. Do not discuss with anyone.
- This is a closed book exam. You are not allowed to carry books or cheatsheets.
- No electronic devices (calculators, laptops, etc) are allowed in the exam hall. Wherever heavy calculation is involved, you need not evaluate it to the final number unless it is explicitly asked for. For example, it is acceptable to leave the answer as $\frac{1}{1+\frac{5}{32}}$. You need not evaluate it to 0.865.
- First section has negative marks. No negative marks for the rest of the sections.

Partitioner
UML exactly?

Section 1: All questions carry one mark each. -0.5 for wrong answers. Answer in True/False.

draw the obj diagram for scenario

To be studied

Question 1. The name notwithstanding, there are most definitely servers in serverless computing. 'Serverless' describes the developer's experience with those servers—they are are invisible to the developer, who doesn't see them, manage them, or interact with them in any way. T

How will you implement these with map reduce ?

Question 2. Poorly maintained data lakes are often called Data Swamps. T

Question 3. Apache Kafka is an open-source distributed event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, data integration, and mission-critical applications. T

how will you compute the median? by map reduce

Question 4. For the services that do not expose metrics, we can use the ladder based scaling strategy. Scaling ladders can be defined per million concurrent users on the platform (1M, 2M ..). This works well for predictable workloads. T

partitioner example

Question 5. A Content Delivery Network (CDN) is a distributed network of servers that are geographically distributed across the globe. T

Question 6. Load balancer is a device or software that distributes incoming network traffic across multiple servers. T T

Question 7. Pods are the smallest deployable units of computing that you can create and manage in Kubernetes. T

Question 8. Neo4j is not ACID compliant. F

min/max count why ?

map reduce code again and again

kab kon sa delete krna h ye kaise pta chelga neo4j me

write query by pen in neo4j

→ CP

T

✓**Question 9.** Redis is a key-value store and MongoDB is a document Store. T

✓**Question 10.** On the CAP triangle, MongoDB falls on the AP side and Cassandra falls on the CP side. F

→ falls in AP

F

✓**Question 11.** General-purpose computing on graphics processing units is the use of a GPU, which typically handles computation only for computer graphics, to perform computation in applications traditionally handled by the CPU. F (GPGPU)

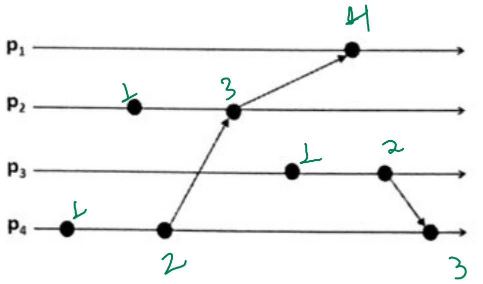
$= \frac{1}{\frac{90}{100} + \frac{10}{10\times10}}$

$= \frac{100}{100} = 1X$

Section 2: All questions carry two marks each.

✓**Question 12.** If only 10% computation can be executed in parallel, and if we have 10 processors, what is the best speed up achievable as per Amdahl's law? 1X

✓**Question 13.** Annotate the following space-time diagram with scalar time.



✓**Question 14.** For the same space-time diagram as in the previous question, annotate p_2 events with matrix time.

✓**Question 15.** What will be the output of the following pig script if the input file, file1, contains a single line "1,V Rao,40,Chennai"?

```
A = LOAD 'file1' USING PigStorage(',')
    AS (id:int, name:chararray, age:int,  city:chararray);
B = FOREACH A GENERATE TOKENIZE(name);
DUMP B;
```

({(V), (Rao)})

✓**Question 16.** What will be the output of the following pig script if the input file, file2, contains numbers 1 to 10, each in one line (i.e., 1 in first line, 2 in second line, and so on)?

```
A = Load 'file2' using PigStorage(',') as (num:int);
B = Foreach A generate 1 as gid, num;
C = Group B by gid;
D = Foreach C generate SUM(B.gid);
Dump D;
```

A:

num	gid	num
1	1	1
2	1	2
3	1	3
4	⋮	⋮
⋮		
10	1	10

B → gid num
1 1
1 2

C → group B:{gid:int, num:int}
1 {(1,1), (1,2) ... (1,10)}

1
2
3
4
5
6

C has ony 1 group se, this process for ony 1 group.
if there were more elements Ans would be (10)
(10)
(10)

(1, {(1,1), (1,2), ... (1,10)}) → C

group results in (group, values) pairs where values contain all the occurencies of group value x rpm .

Read Time (T) = for 1 sector

Rot. Delay (R) = ½ a rotation

$1 \rightarrow \frac{60}{x}$ sec

$\frac{1}{2} \rightarrow \frac{60}{2x}$ sec $= 3\times10^{-3}$

$\Rightarrow \frac{\cancel{60}\,\cancel{30}\,10}{\cancel{2}\times\cancel{3}\times10^{-3}} = x$

$\boxed{10^{4} = x \text{ RPM}}$

$\frac{1}{20} \rightarrow \frac{\cancel{60}\,15}{\cancel{2}\times10^{\cancel{4}5}\times\cancel{2}} = \boxed{15\times10^{-5}\text{ sec}}$

CMI/22/4/2023/Final Exam Apr 2023

~~**Question 17.**~~ Draw a sequence diagram to capture the CMI's admission process.

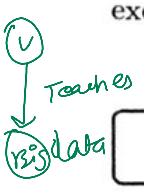
→ R = 3ms

Question 18. Ramesh bought a hard disk with rotational delay of 3ms. With what RPM does the disk spin? If it had 20 sectors per track, what is its read time?

Question 19. How many nodes are created when the following statement is executed by Neo4j?

2

```
create (p:Person {name:'Venkatesh'})-[:Teaches]->(c:Course {name:'BigData'})
```



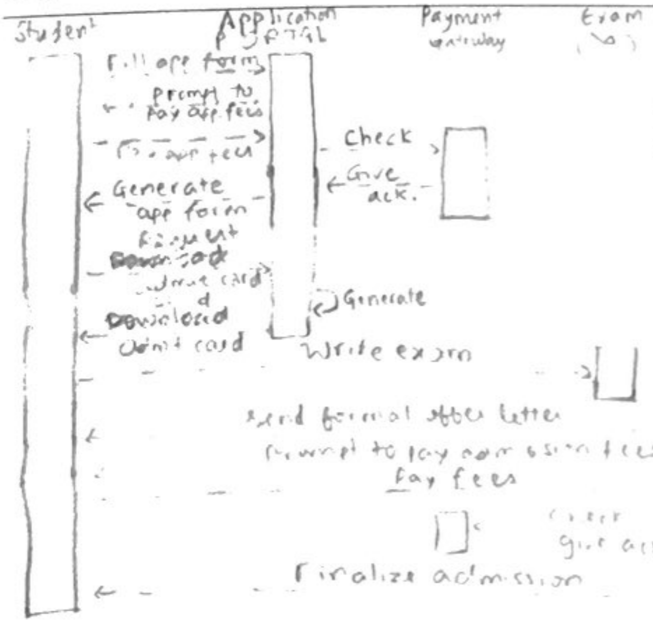
Section 3: All questions carry 4 marks each.

Question 20. The Chennai Public School wants to automate its system for grading students. Specifically, this system will allow creation, modification and deletion of exam marks and student grades from class V to class X. Design a RESTful web service for this scenario. You may scope your answer to three identified resources. Your answer must cover at least one idempotent method assignment and one non-idempotent method assignment.

Question 21. From a very large text file, we need to find the least five frequent words that contain only alphabets (i.e., no digits, no punctuations, etc). Describe a map-reduce design pattern to achieve the same.

Use a filter in the mapper

Bottom 5 from every mapper



① Identify resources

- students
- classes & sections
- grades

② Create endpoints/URIs

/students/{id}

/classes/{class}/sections/{secid}

/grades/{grade}

③ Resource representation (XML, JSON)

④ Relevant HTTP methods

1. create student
2. Put in class & section
3. Enter grades
4. ↳ display student by ... by class & section