1. Let A be an $m \times n$ real matrix. Show that the subordinate 1-norm A is the maximum of absolute column sums, i.e., prove [5 marks]

$$\|A\|_1 = \max_j \{\sum_{i=1}^{m} |a_{ij}| \mid 1 \le j \le n\}.$$

You may want to use the definition of the subordinate matrix norm first to show the inequality. Then prove the equality explicitly by finding a vector x such that vector 1-norm of Ax has the exact value.

2. Let $A = \begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix}$ be a 2×2 matrix such that ϵ is a very small real number. Answer the following questions: [15 marks]

   (a) Find the condition number $\kappa_\infty(A)$ by explicitly calculating the inverse.

   (b) Find the LU decomposition of A using GE *without pivoting*.

   (c) What are the ∞-condition numbers of the factors L, U?

   (d) Find the LU decomposition after permuting rows of A.

   (e) In which of the two scenarios above, the problem of solving $Ax = b$ (for any b and using GE, forward, backward substitutions etc.) well-conditioned? Explain.

3. Consider the following 4×4 system of linear equations: [15 marks]

$$\begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix}.$$

   (a) Solve the above system using GE/ Cholesky.

   (b) Let δb be the column vector $(0.1, -0.1, 0.1, -0.1)^t$. Find the solution of the system $A(x + \delta x) = b + \delta b$.

   (c) Compare $\|\delta x\|_1/\|x\|_1$ and $\|\delta b\|_1/\|b\|_1$ and deduce appropriate conclusions about the conditioning of the problem. Find a (good) bound on the condition number without computing the inverse explicitly.

4. Let A be a symmetric positive definite, tridiagonal matrix. [15 marks]

   (a) Write a pseudocode/ formula for the LDL^T factorization of A: Here L is unit lower triangular and D is a diagonal matrix. What is the exact operation count? Is it the least possible operation count?

   (b) Write a pseudocode/ formula for solving $Ax = b$? Your algorithm may either use the LDL^T factorization obtained above or directly overwrite b with the solution. In either case, find the operation count? Is it the least possible?

   (c) Consider the matrix

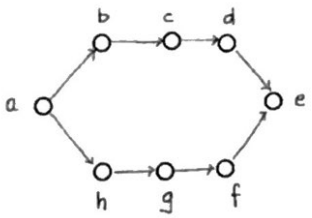
$$C = \begin{bmatrix} 4 & 2 & 0 & 0 \\ 2 & 5 & 2 & 0 \\ 0 & 2 & 5 & 2 \\ 0 & 0 & 2 & 5 \end{bmatrix};$$

   find the LDL^T factorization using your algorithm.

Algorithm Design Techniques 2023 Midsem Exam 9:30 AM - 12:30 PM

- This exam has 6 questions each worth 20 points. You may answer any subset of questions or parts of questions. All answers will be evaluated. You can score at most 100 points in total.
- You may use algorithms taught in class as subroutines without writing the pseudocode, unless explicitly asked not to do so.
- Write down your name and CMI ID at the top right corner of every page of your solutions sheet.
- Unstated assumptions and lack of clarity in solutions can lead to loss of credit. Please feel free to ask the invigilators if you need any clarification about the questions.
- CMI's academic honesty policy regarding cheating and plagiarism applies to this exam.

1. Consider the directed acyclic graph in the figure below.



(a) Write down three distinct topological orderings of vertices in the graph in the figure. **6 points**

(b) For general (odd) integral $n \geq 3$, describe a graph on n vertices with exactly $2^{\frac{n-1}{2}}$ distinct topological orderings. In other words, you must describe an infinite family consisting of n-vertex graphs satisfying the aforementioned property for all integers $n \geq 3$.

14 points

2. Consider a recursive algorithm A that, on an input of length n, recursively calls itself on two inputs of length $n/4$ each, combines the solutions returned by the recursive calls in $\sqrt{n}$ time, and then returns its solution. The algorithm is not recursive on inputs of length at most 4 and runs in time $O(1)$ on such inputs.

(a) Write a recurrence relation for the running time of the algorithm. Do not forget to include the boundary conditions. **6 points**

(b) Draw the recursion tree corresponding to the recurrence. Reason about the number of levels that the tree has and derive an expression for the total cost of the tree at the i-th level from the root. Assume that the level numbers begin from 0 corresponding to the root level. **8 points**

(c) Based on the above, solve the recurrence. **6 points**

3. In the following, n is a multiple of $2k$. A value in an array A of length n is called a k-heavy hitter if it occurs at least n/k times in the array.

(a) Given array A of length n, which is not necessarily sorted, let a_i for $i \in \{1, 2, \ldots, n\}$ denote the value occurring at index i upon sorting A in nondecreasing order, where we assume that arrays are indexed from 1 to n.
Prove that the set $\{a_j \mid j = t \cdot \frac{n}{2k}$ for $t \in \{1, 2, \ldots 2k\}\}$ contains all the k-heavy hitters in A.

10 points

(b) Using ideas from part (a), write the pseudocode for an algorithm that runs in time $O(kn)$, and outputs all the k-heavy hitters in an array A of length n. You will get full credit if your pseudocode is correct and has the desired running time. **10 points**

4. Consider a connected undirected graph $G = (V, E)$ on n vertices.

   (a) Prove that the DFS procedure, when run on G, examines at most $n - 1$ edges in the graph before encountering the first back edge, if any.

   You may use the fact that any undirected graph on n vertices that does not contain a cycle has at most $n - 1$ edges. **6 points**

   (b) Describe the pseudocode for a $O(n)$-time algorithm that takes a connected undirected graph $G = (V, E)$ as input and outputs whether G contains a cycle. If you want to use any of the algorithms taught in class as subroutines, write out their pseudocodes as well. **8 points**

   (c) Argue why your algorithm has running time $O(n)$. **6 points**

5. Let MERGE(A, B) for sorted arrays A and B denote a procedure that returns a sorted array C obtained by merging the arrays A and B. In particular, if A and B have lengths p and q, respectively, C has length $p + q$ and the MERGE procedure takes $O(p + q)$ time.

   (a) Consider the following algorithm that outputs a sorted array by merging the elements in k sorted arrays $A_1, A_2, \ldots A_k$, where each A_i is assumed to have length n. Carefully analyze its running time. **6 points**

Input: k sorted arrays $A_1, A_2, \ldots A_k$ each of length n

```
1 C ← A_1;
2 for i from 2 to k do
3 |  C ←MERGE(C, A_i);
4 end
5 return C
```

   (b) Write the pseudocode for a more efficient divide and conquer algorithm for the same problem as in part (a). **6 points**

   (c) Prove that your algorithm from part (b) has an asymptotically better running time than the algorithm described in part (a). **8 points**

6. Consider a directed acyclic graph $G = (V, E)$ in which each vertex $u \in V$ has an associated price p_u which is a positive integer. Define the function cost as follows: for each $u \in V$,

   cost(u) = price of the cheapest vertex reachable from u (including u itself).

   (a) Write the pseudocode for an $O(|V| + |E|)$-time algorithm that takes G as input and returns an array C such that $C[u] = \text{cost}(u)$ for all vertices $u \in V$. **10 points**

   (b) Prove that your algorithm computes the cost values correctly. **10 points**

   Note: You may be able to prove it by induction.

Data Mining and Machine Learning 2023

Mid-Semester Exam

Chennai Mathematical Institute

23 February 2023, 09:30–11:30 (2 hours) Marks: 30, Weightage: 20%

The following questions carry 5 marks each.

1. A research paper contains the data of about 1000 university students with the following attributes: (1) the time they went the sleep previous night, (2) the time they woke up, (3) the meals they had during the day and (4) their concentration level during the day. Can you use this data to build a ML model that takes the first three attributes and predicts the concentration level with good accuracy for an *average person*? Explain your answer.

2. Consider the class of *decision stumps*, which are decision trees of height 1.

   (a) Consider the ensemble model A build by Bagging 100 decision stumps. Can A be represented as a decision stump?

   (b) Consider another ensemble model B built by AdaBoosting 100 decision stumps. Can B be represented as a decision stump?

   Justify your answers.

3. You are building a decision tree on tabular data with attributes $\{A_1, A_2, \ldots, A_7\}$ where $\{A_3, A_5\}$ are numeric and the other five attributes are categorical. Attribute A_3 takes integer values in the range $[-100, 100]$ and attribute A_5 takes integer values in the range $[1, 10000]$. There are 2000 items in the training set. You adopt a pre-pruning strategy to build the tree where you do not split any node with fewer than 50 items. Across all possible decision trees that can be built on this training set, what is the maximum height of the resulting tree? Explain your answer.

4. An airport security system consists of a full body scanner that all passengers pass through, followed by manual frisking. Passengers who pass through the scanner without setting off the alarm are let through without frisking. Passengers who set off the alarm are manually frisked. If nothing suspicious is found during manual frisking they are let through, otherwise they are detained.

   Think of the full body scanner and manual frisking as two classifiers to detect suspicious passengers. A false positive is an innocent passenger who is classified as suspicious and a false negative is a suspicious passenger who is classified as innocent.

   In terms of the confusion matrix, explain what metrics the two stages in the classifier should aim to optimize so that, to the extent possible, all suspicious passengers are detained, and all innocent passengers are let through with minimum hassle.

5. You often hear airlines announce that a departure is delayed due to the late arrival of the incoming flight. Since these delays cascade, it would be useful to know, for instance, if a delay in a morning flight from Amritsar to Delhi would have an impact on your afternoon flight from Hyderabad to Cochin. Suppose the airline provides you with daily data about delayed flights for the past year. Each day's data has the list of flights that were delayed on that day. Explain how you could use this data to answer questions of the form "Does a delay in flight A imply a delay in flight B?" Be as precise as possible about how you model this problem and the analysis that you need to perform on the model.

6. Suppose your team computes the solution to a linear regression problem on n attributes as $\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$. Your partner argues that the relative importance of the attributes to the final answer can be computed from the coefficients. The most significant attribute has the largest coefficient (in magnitude), the second most significant attribute has the second largest coefficient, and so on. Explain whether your partner's claim is justified.

CMI/2023/Mid-Semester Exam 24/02/2023

DISTRIBUTED COMPUTING AND BIG DATA

DURATION: 90 MINS. Chennai Mathematical Institute MAX: 25 MARKS.

Instructions

- You are allowed to carry a single A4 size paper with hand written contents. You should not exchange these cheat sheets with other students.
- This is an individual task. Do not discuss with anyone.
- Calculators (that do no access internet) are allowed. But no other electronic devices are allowed. 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.
- Clearly mention your name and roll number in your answer sheet.
- Please submit your cheat sheet along with your answer sheet.

Section 1: Correct answers carry 1 mark each. Wrong answers carry -0.5 marks each.

Question 1. In 1928, IBM introduced a new version of the punched card with rectangular holes and 80 columns. It turned out to be one of IBM's most important technological innovations, propelling the company to the forefront of data processing. We classify such computers that used punched cards for storing programs as Von Neumann machines? True/False?

Question 2. IBM Summit, the fastest super computer of 2018, was capable of computing 200,000 trillion calculations per second. Oak Ridge National Laboratory has a super computer that could do 2000 teraFLOPS. IBM Summit is faster (purely based on the given data alone). True/False?

Question 3. The inode (index node) is a data structure in a Unix-style file system that describes a file-system object such as a file or a directory. True/False?

Question 4. A key benefit to STaaS is that you are offloading the cost and effort to manage data storage infrastructure and technology to a third-party cloud service provider. True/False?

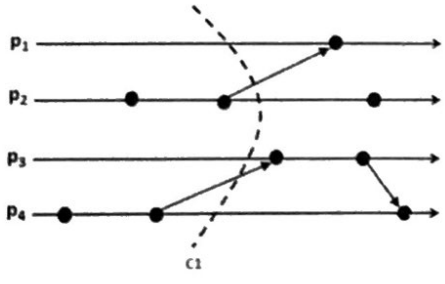
Question 5. Your company needs a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions. In such situations, your company can use a data lake. True/False?

Question 6. Ram bought a hard disk that has 500 GB of hard disk space and 5000 RPM rotation rate. Shyam bought a hard disk that has only 100 GB of hard disk space and it also has the same rotation rate. Ram's disk will have lesser rotational delay when compared to Shyam's disk. True/False?

Question 7. There is always one reducer in every map-reduce program. True / False?

Question 8. Map code is executed by the namenode in the hadoop cluster. True/False?

Question 9. The cut C1 is consistent. True/False?



Question 10. A solution to the General's Paradox is sending large number of messengers in each direction to guarantee a messenger gets through to the other side. True/False?

Question 11. Some technologists have estimated that all the words ever spoken by mankind would be equal to five Exabytes (an extraordinarily large unit of digital data). How much Gigabytes that (5 Exabytes) would be?

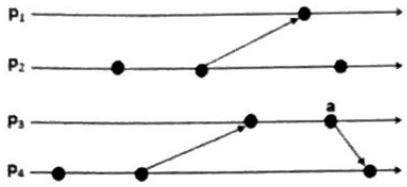
Question 12. The basic model of a distributed system, as discussed in the class, has a set of disconnected processes with no shared memory and no global clock. True/False?

Section 2: Correct answers carry 2 marks each. No negative marks.

Question 13. As per Amdahl's law, What is the best achievable speed up if only 20% can be parallelized, and we have 8 processors?

Question 14. Draw a space time execution diagram that provides an example of an inconsistent cut with exactly three processes, four events in the past and four events in the future.

Question 15. If we were to annotate the following space-time execution diagram with vector time stamps, how would we annotate the event marked as 'a'?



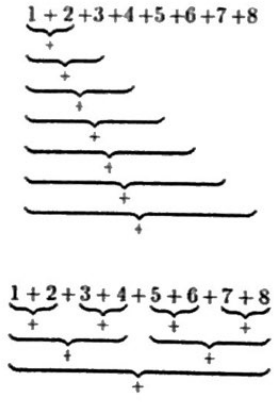
Question 16. In the muddy children puzzle, as discussed in the class, what would the children say during the first and second rounds if n=4 and k=2? i.e., there are four children, and two of them have muddy forehead and they are told that at least one of them have muddy forehead. Assume that each child can only say 'yes', 'no' or 'dont know'. Use the following format to provide your answer.

```
Round1: <1st child response>,<2nd ...>,<3rd ...>,<4th ...>
Round2: <1st child response>,<2nd ...>,<3rd ...>,<4th ...>
```

Question 17. Assume disk size = 256 GB, block size = 16 KB. How much space (in MB) will we need to store the free space bitmap?

Section 3: Question carries 3 marks. No negative marks.

Question 18. Our ability to write parallelizable programs decides the speed up we can achieve through scaling. Consider two programs to add large list of numbers. The first program adds the first two numbers, remembers the result, and adds that result to the next number. It continues doing this until the end of the list is reached. The second program adds two numbers at a time in parallel. It recursively does so until the final results are arrived at. The following figure explains their logic with an example of eight numbers.



Assume that the list is large and the numbers may be unordered. As per Gustafson's law, assuming each addition is an operation, how much speedup (approximately) can these programs achieve if there are four processors?