Indiana University Bloomington

Fall-2023
CSCI-B505 / INFO-I500
APPLIED ALGORITHMS
Examination – III

December 07, 2023, Thursday, 6:30 p.m. – 7:30 p.m.

Name & Surname	
University ID	
Signature	

Rules:

1. There are 25 questions in this examination, each question carries 4 marks.
2. Duration of the exam is 60 minutes.
3. Write your name and surname on every page at the designated positions.
4. Put your ID card on your desk so that the proctors can check your identity.
5. The use of lecture notes, books, and any other resources, calculators, computers, mobile phones, and any digital equipment is prohibited.
6. Every student taking this examination is subject to the university discipline code. Any act or attempt of cheating, including helping others, will be considered a violation of the code.

Name & Surname:

1. Which of the following best describes a randomized algorithm?

   (a) For multiple runs with different inputs, the time complexity or output may vary across runs

   (b) For multiple runs with the same input, the time complexity or output may vary across runs ✓

   (c) For multiple runs with the same input, the time complexity or output may remain the same across runs

   (d) For multiple runs with different inputs, the time complexity or output may remain the same across runs

Answer the next 3 questions based on the provided information on Approximate Median Randomized Algorithms

Pseudo-Code for the ApproxMedian1, ApproxMedian2, and ApproxMedian3 algorithms are provided below. Assume there is an array A with 249 numbers in it. The parameters for the below algorithms are $c = 4$ and $\delta = 0.2$. Answer the below questions using the information provided.

Algorithm *ApproxMedian1*(δ, A)

1. $\triangleright$ $A[1..n]$ is array of n distinct numbers.
2. $r \leftarrow Random(1, n)$
3. $x^* \leftarrow A[r]$; $k \leftarrow 1$
4. **for** $i \leftarrow 1$ **to** n
5. **do if** $A[i] < x^*$ **then** $k \leftarrow k + 1$
6. **if** $\lfloor(\frac{1}{2} - \delta)(n+1)\rfloor \leqslant k \leqslant \lceil(\frac{1}{2} + \delta)(n+1)\rceil$
7. **then return** x^*
8. **else return** "error"

Algorithm *ApproxMedian2*(δ, A)

1. $j \leftarrow 1$
2. **repeat** *result* $\leftarrow$*ApproxMedian1*(A, δ); $j \leftarrow j + 1$
3. **until** (*result* $\neq$ "error") **or** $(j = c + 1)$
4. **return** *result*

Algorithm *ApproxMedian3*(δ, A)

1. **repeat** *result* $\leftarrow$*ApproxMedian1*(A, δ)
2. **until** *result* $\neq$ "error"
3. **return** *result*

2. For the ApproxMedian1 algorithm, what are the lower and upper bound positions of array A that ensure the returned value is considered a valid approximate median?

   (a) 65 and 145

   (b) 75 and 175 ✓

   (c) 85 and 165

   (d) 70 and 165

3. What is the probability that ApproxMedian2 fails to find a valid answer after execution? (Choose the value closest to your calculated probability.)

   (a) 0.08

   (b) 0.22

   (c) 0.13 ✓

   (d) 0.36

4. What category of randomized algorithms do ApproximateMedian2 and ApproximateMedian3 fall into, respectively?

   (a) Las Vegas, Monte Carlo

   (b) Monte Carlo, Las Vegas ✓

   (c) Las Vegas, Las Vegas

   (d) Monte Carlo, Monte Carlo

5. Given a Las Vegas algorithm with an expected running time of $E[T(n)] = \log n$. If we converted the Las Vegas algorithm into a Monte Carlo algorithm that runs in $8 \cdot \log n$. What should be the expected success probability for the Monte Carlo algorithm?

   (a) $1/8$

   (b) $7/8$ ✓

   (c) $1/7$

   (d) $6/7$

6. When using randomization to evaluate a majority node with three children, what is the probability that the third child's value need not be visited after randomly visiting two children?

   (a) $1/3$ ✓

   (b) $1/2$

   (c) $2/3$

   (d) $1/4$

7. Freivald's algorithm for matrix multiplication verification has a time complexity of $O(k \cdot n^2)$. What does k represent in this context?

   (a) The number of matrices

   (b) The dimensions of the matrices

   (c) The number of times the verification is repeated ✓

   (d) The failure probability

8. We would like to verify whether $A \times B = C$, where all are $n \times n$ matrices. If we run the Freivalds algorithm three times, what is the probability that the Freivalds algorithm says that the $A \times B = C$ is true when in reality $A \times B = C$ is false?

   (a) 0.125 ✓

   (b) 0.25

   (c) 0.5

   (d) 1

9. You are provided with the load factor $\lambda = 0.6$ and Hash Table Capacity = 15. But you are curious about how many cells were occupied in the current hash table. How many cells are currently occupied in the hash table?

   (a) 9 ✓

   (b) 10

   (c) 6

   (d) 12

10. A Hash table uses hashing function h(k) = k mod 10 and linear probing. A few insertions are made to the table, as seen in the image below. What could have been the order in which these insertions were made?

Index	0	1	2	3	4	5	6	7	8	9
Number		51	22	21	44	25	93		18	

   (a) 18, 51, 44, 21, 22, 25, 93

   (b) 51, 21, 23, 44, 25, 93, 18

   (c) 25, 51, 22, 21, 44, 93, 18 ✓

   (d) 25, 51, 22, 21, 93, 18, 44

11. Consider a hash table of size 13 that uses the hashing function $h(k) = k \bmod 13$ and quadratic probing $f(i) = i^2$. A few insertions have already been made to the table as shown below. Now, we want to add a new element 19 to this hash table. At what index will this new object be added?

Index	0	1	2	3	4	5	6	7	8	9	10	11	12
Number	13		41	16			6	20			49		

(a) 8

(b) 9 ✓

(c) 1

(d) 4

Answer the next 2 questions based on the Bloom-Filters Concept

Assume that you are using the Bloom Filters to create a hash table for the data you are inserting. You are provided with two different hash functions to insert the following elements 15, 1, 3, 13, 16, 5, 2, 19, 4, and 7 in the given order.

Hash Functions:

$$h_1(x) = (2x+5) \mod 11$$
$$h_2(x) = (x+8) \mod 11$$

12. Select the option that is the correct hash table [0..10] for bloom filters after inserting all the values.

(a) $[1,1,0,1,1,0,1,1,1,1,1]$

(b) $[1,1,1,1,1,1,1,1,1,1,1]$

(c) $[1,1,1,0,1,1,0,1,1,1,1]$ ✓

(d) $[1,1,1,0,1,0,1,1,1,1,0]$

13. After inserting the given elements into a Bloom filter with the specified configuration, you need to check if a new random element is present. Which of the following options for the new data would result in a false positive, indicating its presence?

(a) 17

(b) 21

(c) 26 ✓

(d) 53

14. In Bloom Filters, how many hash functions are required if the required false alarm probability is (1/32), and how many bits per element are required respectively?

    (a) 5, 7.2 ✓
    (b) 6, 10
    (c) 5, 10
    (d) 6, 7.2

15. For perfect hashing using the hash function $h(k) = k \mod 5$ at level 1, what should be the size of level 2 for the given sequence of keys: 15, 2, 1, 5, 20, 31, 12, 13, 34?

    (a) 16
    (b) 17
    (c) 18
    (d) 19 ✓

16. Considering a streaming integer array S with duplicates, the objective is to estimate the number of distinct values in S using Minwise Hash with a constant amount of memory. Given a hash size of 50 and the minimum hash observed is 2, what is the estimated value of distinct numbers?

    (a) 24 ✓
    (b) 25
    (c) 48
    (d) 49

17. Select the most accurate statement that distinguishes a streaming algorithm from a traditional batch processing algorithm:

    (a) Streaming algorithms can only handle discrete data.
    (b) Streaming algorithms process data in a single pass only. ✓
    (c) Streaming algorithms always achieve optimal performance.
    (d) All of the above.

18. Select k elements RANDOMLY from the streaming sequence S, where k = 3. In the below table, S1, S2, and S3 are the first, second, and third sampled items respectively. R represents the random values in the range [1..100] for the probability generation such that if $R_i \leq 100 \cdot p_i$ then perform sampling. 'Replace' indicates which sampled item is to be replaced if the sampling condition is met.

    What are the final values of S1, S2, and S3, respectively, after the streaming data has been processed?

	1	2	3	4	5	6	7	8	9	10
S	9	3	26	42	33	12	49	10	36	1
R	-	-	-	35	70	75	22	5	52	17
Replace	-	-	-	3	2	3	1	2	1	2
S1										
S2										
S3										

    (a) 49, 1, 42 ✓

    (b) 49, 10, 42

    (c) 42, 10, 49

    (d) 9, 3, 42

    Answer the next 2 questions using the Misra-Gries algorithm

    Assume the input data stream S = $< 1,2,3,3,5,6,4,8,7,6,3,9,7,3,4,8,7,4,3,5 >$. Implement the Misra-Gries algorithm with the allowed error $\varepsilon = 0.2$

19. What will be the final state of the counters? (CX:1, CY:2 represents that there are 2 counters CX and CY, where the element X has count 1 and the element Y has count 2.)

    (a) {C3: 5, C7: 3, C9: 1, C4: 2, C5: 2}

    (b) {C3: 4, C7: 2, C9: 1, C4: 1, C5: 1}

    (c) {C3: 5, C7: 3, C8: 2, C4: 3, C5: 2}

    (d) {C3: 3, C7: 2, C8: 1, C4: 1, C5: 1} ✓

20. What are the 0.29 Heavy Hitters?

    (a) only 3

    (b) 3, 4, and 7 only

    (c) 3 and 7 only ✓

    (d) No 0.29 heavy hitters

21. Which of the following is true about the Misra-Gries algorithm?

    I) It returns all items that appeared more than $\phi \cdot n$

    II) It never returns an item that appeared less than $(\phi - \varepsilon) \cdot n$

    III) All the items that appeared more than $(\phi - \varepsilon) \cdot n$ but less than $\phi \cdot n$ will be reported

    IV) Some of the items that appeared more than $(\phi - \varepsilon) \cdot n$ but less than $\phi \cdot n$ may still be reported

    (a) I and II only

    (b) I and III only

    (c) I, II, and III only

    (d) I, II, and IV only ✓

22. Which of the following is true about the Count-Min Sketch algorithm?

    I) It returns all items that appeared more than $\phi \cdot n$

    II) It can return an item that appeared less than $(\phi - \varepsilon) \cdot n$

    III) It works well in a distributed computing environment.

    IV) It is a probabilistic data structure that maps input streaming data to a defined space using hash functions.

    (a) I and II only

    (b) I and III only

    (c) I, II, and III only

    (d) All of the above ✓

23. Consider a hash function that maps even integer values to +1 and odd integer values to -1. Given the input data stream $< 1,2,3,4,5,6,4,3,2,4 >$, Using the half-decent estimator algorithm, what is the estimated skewness value for this input stream?

    (a) 2

    (b) 4 ✓

    (c) -1

    (d) 1

Answer the next 2 questions using the Count-Min Sketch algorithm

The Count-Min Sketch algorithm is being implemented in two distributed environments. The same hash functions are used in both environments. The hash functions employed and the resultant matrices from running the Count-Min Sketch on the streaming data in both environments are provided below.

$$h1(x) = (2x+9)mod5$$
$$h2(x) = (3x+7)mod5$$
$$h3(x) = (7x+3)mod5$$

$$\text{Environment 1} = \begin{bmatrix} 1 & 1 & 0 & 2 & 1 \\ 1 & 1 & 1 & 2 & 0 \\ 1 & 0 & 2 & 1 & 1 \end{bmatrix}$$

$$\text{Environment 2} = \begin{bmatrix} 0 & 1 & 1 & 3 & 0 \\ 1 & 0 & 0 & 3 & 1 \\ 1 & 1 & 3 & 0 & 0 \end{bmatrix}$$

24. Based on the information provided above, what is the total number of elements when both environments are combined?

    (a) 5

    (b) 10 ✓

    (c) 15

    (d) 30

25. Based on the information provided above, Which of the following is a 0.4 heavy hitter when both environments are combined?

    (a) 1

    (b) 2 ✓

    (c) 3

    (d) 4