中国科学技术大学计算机科学与技术学院

2022 年春季学期期中考试试卷

课程名称：大数据算法　　课程编号：011186

开课院系：计算机科学与技术学院　　考试形式：闭卷

姓　名：＿＿＿＿＿＿　　学　号：＿＿＿＿＿＿　　专　业：＿＿＿＿＿＿

题 号	1	2	3	4	5	6	7	8	总 分
得 分									

（以下为试卷正文）

（装订线内不要答题）

题目 1 *判断题 20 分*

(a) All the streaming algorithms with polylogarithmic space complexity are randomized (or probabilistic) algorithms

所有的空间复杂度为 $\mathrm{poly}\log(n)$ 的数据流算法都是随机（或概率）算法。

(b) The expected value of a random variable is always at most its variance.

一个随机变量的期望总是不超过它的方差。

(c) In the strict turnstile model, the CountMin Sketch algorithm has the following guarantee: for any query(i) such that $1 \leq i \leq n$, the algorithm outputs $\tilde{x}_i$ satisfying $\tilde{x}_i \geq x_i$, where x_i is the number of occurrences of i at the query moment.

在严格的 Turnstile 模型下面，CountMin Sketch 算法具有下面的保证：对于任意一个查询 query(i)（这里 $1 \leq i \leq n$），算法输出的 $\tilde{x}_i$ 满足 $\tilde{x}_i \geq x_i$，这里的 x_i 是在当前查询时刻 i 所出现的次数。

(d) Let i be an integer with $1 \leq i \leq n$. Assume X_i is a random variable that is 1 if it rains on the i-th day and 0 otherwise. Assume further that the probability for rain on any day is 0.45. Let $X = \sum_{i=1}^{n} X_i$. Then Chebyshev's inequality can be used to show that $\Pr[X > 0.9n] < 0.68^{0.45n}$.

令 i 为一个整数且满足 $1 \leq i \leq n$。假设 X_i 是一个随机变量，它满足当第 i 天下雨时 $X_i = 1$，当第 i 天不下雨时 $X_i = 0$。进一步假设每一天下雨的概率为 0.45。令 $X = \sum_{i=1}^{n} X_i$。那么，我们可以使用 Chebyshev 不等式证明 $\Pr[X > 0.9n] < 0.68^{0.45n}$。

(e) Let $A \in \mathbb{R}^{n \times d}$ be a matrix with a Singular Value Decomposition (SVD) such that $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, for $r \geq 2$. Then it holds that the left singular vectors $u_1, \cdots, u_r$ are pairwise orthogonal.

令 $A \in \mathbb{R}^{n \times d}$ 为一个矩阵，且它的奇异值分解（SVD）为 $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$，这里 $r \geq 2$。那么它的左奇异向量 $u_1, \cdots, u_r$ 是两两正交的。

上述陈述中对的有： _____

错的有： _____

题目 2 单选题 5 分

Let $A \in \mathbb{R}^{n\times d}$. Suppose we apply the random projection f as specified in the Johnson–Lindenstrauss (JL) lemma to reduce the dimension of each row a_i of A from d to k. We want to guarantee that with probability at least $1 - \frac{1}{n}$, for all $i, j \in [n]$,

$$(1 - \frac{1}{\log n})\|a_i - a_j\| \leq \|f(a_i) - f(a_j)\| \leq (1 + \frac{1}{\log n})\|a_i - a_j\|.$$

According to the JL lemma, what is the smallest k that we need to choose?

(a) $\Omega(\log n)$, (b) $\Omega(\log^2 n)$, (c) $\Omega(\log\log n)$, (d) $\Omega(\log^3 n)$

令 $A \in \mathbb{R}^{n\times d}$ 为一个矩阵。假设我们使用 Johnson–Lindenstrauss（JL）引理中的随机投影 f 将 A 中的每一行 a_i 的维度从 d 降到 k。为了保证以至少 $1 - \frac{1}{n}$ 的概率，对于任意 $i, j \in [n]$,

$$(1 - \frac{1}{\log n})\|a_i - a_j\| \leq \|f(a_i) - f(a_j)\| \leq (1 + \frac{1}{\log n})\|a_i - a_j\|$$

根据 JL 引理，我们可以选择的最小的 k 为多少?

(a) $\Omega(\log n)$, (b) $\Omega(\log^2 n)$, (c) $\Omega(\log\log n)$, (d) $\Omega(\log^3 n)$

你的选择： _____

题目 3 单选题 5 分

Consider the locality sensitive hashing (LSH) based algorithm for solving the (c, r)-ANN problem, where the input is a set $\mathcal{P}$ of n points from $\{0, 1\}^d$. Suppose we choose $c = 5$. If we use $\tilde{O}(f)$ to denote $O(f \operatorname{poly}\log(f))$ for any function f, which of the following is correct about the corresponding data structure?

(a) The space complexity is $\tilde{O}(n^{\frac{6}{5}} + nd)$.
(b) The query time is $\tilde{O}(n^{1/5})$.
(c) The space complexity is $\tilde{O}(n^{\frac{5}{6}} + nd)$.
(d) The query time is $\tilde{O}(n^{1/6})$.

考虑基于敏感哈希函数 (LSH) 的解决 (c, r)-ANN 的算法。假设其输入 $\mathcal{P}$ 是一个具有 n 个点的集合，其中每个点都来自 $\{0, 1\}^d$。假设 $c = 5$。如果对于任意函数 f，我们使用 $\tilde{O}(f)$ 来表示 $O(f \cdot \operatorname{poly}\log(f))$。关于该数据结构，以下哪个陈述是正确的?

(a) 空间复杂度为 $\tilde{O}(n^{\frac{6}{5}} + nd)$.
(b) 查询时间为 $\tilde{O}(n^{1/5})$.
(c) 空间复杂度为 $\tilde{O}(n^{\frac{5}{6}} + nd)$.
(d) 查询时间 $\tilde{O}(n^{1/6})$.

你的选择： _____

题目 4 *10 分*

Let $S = \langle 2, 6, 5, 5, 3, 5, 6, 3, 3, 2 \rangle$ be a stream of numbers. Suppose that we run Misra-Gries Algorithm with $k = 2$ for finding the first 2 most frequent items in a data stream. Recall that during the execution you will use an array A to maintain 2 items and their counters.

Write down the contents of the array A after seeing each number in the stream.

令 $S = \langle 2, 6, 5, 5, 3, 5, 6, 3, 3, 2 \rangle$ 为一个元素为整数的数据流。假设我们运行 $k = 2$ 情形的 Misra-Gries 算法，即查找出现最频繁的 2 个元素的算法。注意到算法在运行中将使用一个数组 A 来维护 2 组元素及其相应的计数器。

写下算法在看到数据流中的每个元素后，数组 A 中的相应内容。

将你的答案写在下表中:

time / A	0	1	2	3	4	5	6	7	8	9	10	11
Item 1												
Counter 1												
Item 2												
Counter 2												

题目 5 *10 分*

Let $A \in \mathbb{R}^{n \times d}$ be a matrix with a SVD such that $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, for $r \geq 2$. Let $w \in \mathbb{R}^d$ be a vector such that $w = \alpha v_1 + \beta v_2$.

(a) Write Aw as a linear combination of $u_1, \cdots, u_r$.

(b) Write $\|Aw\|_2$ as a function of α, β and $\sigma_1, \cdots, \sigma_r$.

令 $A \in \mathbb{R}^{n \times d}$ 为一个矩阵，且它的奇异值分解（SVD）为 $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$，这里 $r \geq 2$。令 $w \in \mathbb{R}^d$ 为一个向量且满足 $w = \alpha v_1 + \beta v_2$。

(a) 将 Aw 写成关于 $u_1, \cdots, u_r$ 的线性组合的形式。

(b) 将 $\|Aw\|_2$ 写成关于 α, β 及 $\sigma_1, \cdots, \sigma_r$ 的函数。

题目 6 *15 分*

Let $x \in \mathbb{R}^d$ be a random (column) vector such that each entry x_i is independently generated from the Gaussian distribution $N(0,1)$.

(a) Let $y_i = x_i^2 - 1$, for each $i \in [d] = \{1, \ldots, d\}$. What is $\mathrm{E}[\sum_{i=1}^d y_i]$? Give the details of your calculation.

(b) Let $a \in \mathbb{R}^d$ be an arbitrary row vector. Let $T = a \cdot x = \sum_{i=1}^d a_i \cdot x_i$ be the inner product of a and x. What is $\mathrm{E}[T]$ and $\mathrm{Var}[T]$? Give the details of your calculation.

令 $x \in \mathbb{R}^d$ 是一个随机的列向量，其每个元素 x_i 是独立地从 Gaussian 分布 $N(0,1)$ 中产生的。

(a) 令 $y_i = x_i^2 - 1$, 这里 $i \in [d] = \{1, \ldots, d\}$。计算 $\mathrm{E}[\sum_{i=1}^d y_i]$。给出计算步骤。

(b) 令 $a \in \mathbb{R}^d$ 是任意一个行向量。令 $T = a \cdot x = \sum_{i=1}^d a_i \cdot x_i$ 为 a 和 x 的内积。计算 $\mathrm{E}[T]$ 和 $\mathrm{Var}[T]$。给出计算步骤。

题目 7 *15 分*

Apply the following algorithm to a stream S of (possibly infinite many) items $\langle e_1, w_1\rangle, \langle e_2, w_2\rangle, \cdots$,, where each item $\langle e_i, w_i\rangle$ contains an identifier e_i and a weight $w_i > 0$.

(a) Initialize: set $x = 0$
(b) Process $\langle e_j, w_j\rangle$: with probability $w_j / \sum_{i=1}^{j} w_j$, set $x = e_j$; otherwise, do nothing.
(c) Output: x.

Let X_m be the output of the above algorithm at time m, or equivalently, X_m is the output after processing the stream $S_m = \{\langle e_1, w_1\rangle, \cdots, \langle e_m, w_m\rangle\}$. Prove that for each $\ell \in [m] = \{1, \cdots, m\}$, the identifier e_ℓ is sampled with probability proportional to its weight, i.e., $\Pr[X_m = e_\ell] = \frac{w_\ell}{\sum_{i=1}^{m} w_i}$.

将下面的数据流算法作用于输入为 $S = \langle e_1, w_1\rangle, \langle e_2, w_2\rangle, \cdots$, 的数据流上，这里 S 可能含有无穷多的元素，而每个元素 $\langle e_i, w_i\rangle$ 包含一个标识符 e_i 和一个权重 $w_i > 0$。

(a) 初始化：令 $x = 0$
(b) 处理 $\langle e_j, w_j\rangle$：以概率 $w_j / \sum_{i=1}^{j} w_j$，更新 $x = e_j$；否则，什么都不做
(c) 输出：x。

令 X_m 为时刻 m 时算法的输出，即 X_m 是算法在处理了数据流 $S_m = \{\langle e_1, w_1\rangle, \cdots, \langle e_m, w_m\rangle\}$ 之后的输出。证明对于任意 $\ell \in [m]$，标识符 e_ℓ 被抽样出的概率正比于其权重，即 $\Pr[X_m = e_\ell] = \frac{w_\ell}{\sum_{i=1}^{m} w_i}$。

题目 8 *20 分*

Consider a stream S of elements $a_1, \cdots, a_m$ such that each element a_i is an integer from the set $[n] = \{1, \cdots, n\}$, where n, m are sufficiently large integers. Let t be the number of distinct elements in the stream. Suppose that $t > k$ for some integer $k \geq 10$.

Choose a hash function $h : [n] \to [M]$ from a 2-wise independent hash family, where $M = n^3$. Then use h to hash all the elements in S to $[M]$.

(a) For each $i \in [t]$, what is the probability that the i-th distinct element is hashed to a value below $\frac{k \cdot M}{2t}$? It suffices to give a good upper bound on this probability.

(b) Let Y be the total number of distinct elements that are hashed to a value below $\frac{k \cdot M}{2t}$.

  (a) Give a good upper bound of $\mathrm{Var}[Y]$.

  (b) Prove that $\Pr[Y \geq k] \leq \frac{1}{6}$.

(c) Suppose we have an algorithm $\mathcal{A}$ that outputs an estimate $\tilde{t}$ such that $\frac{t}{4} \leq \tilde{t} \leq 4t$ with probability at least 0.6. How could you improve the probability that $\frac{t}{4} \leq \tilde{t} \leq 4t$ to $1 - \frac{1}{n^2}$? Describe your strategy and explain why.

令 $S = a_1, \cdots, a_m$ 是一个数据流，其中每个元素 a_i 都来自于集合 $[n] = \{1, \cdots, n\}$。假设 n, m 都是足够大的整数。令 t 为数据流中不同元素的个数。假设 $t > k$，这里的 $k \geq 10$ 且 k 为整数。

从一个 2-wise 独立的哈希函数类中选择一个函数 $h : [n] \to [M]$，这里 $M = n^3$。然后利用 h 将 S 中的元素哈希到集合 $[M]$ 中。

(a) 对于任意 $i \in [t]$，第 i 个不同的元素被哈希到一个小于 $\frac{k \cdot M}{2t}$ 的值上的概率是多少？你只需给出一个尽可能好的上界即可。

(b) 令 Y 是所有被哈希到一个小于 $\frac{k \cdot M}{2t}$ 的值上的不同元素的个数。

  (a) 给出 $\mathrm{Var}[Y]$ 的一个 (尽可能好的) 上界。

  (b) 证明 $\Pr[Y \geq k] \leq \frac{1}{6}$。

(c) 假设我们有一个算法 $\mathcal{A}$ 其输出估计值 $\tilde{t}$，满足：以至少 0.6 个概率，$\frac{t}{4} \leq \tilde{t} \leq 4t$。问：如何将事件 $\frac{t}{4} \leq \tilde{t} \leq 4t$ 的概率提高到 $1 - \frac{1}{n^2}$？描述你的策略并给出解释。