Theoretical Computer Science III Term 2 Summative

wtxd25

March 11, 2020

1 Advanced Algorithms

Question 1

We have a hash table of size n, with arbitrary hash function $h : \mathscr{U} \rightarrow \{0, ..., n-1\}$. $|\mathscr{U}| \geq n \cdot m$ for some $m \geq 1$. Want to prove that $\mathscr{U}$ contains a subset $\mathscr{U}_0$ of size m such that $h(x_1) = h(x_2)$ **(1)**.
Proof By Induction:
Let m = 1: This implies that $|\mathscr{U}| \geq n$, and we need to find a subset $\mathscr{U}_0$, $|\mathscr{U}_0| = 1$. The above condition is trivially true for this case because there will only be one element in the subset which obviously hashes to the same value as itself.
Assume true for m = k: This means that $|\mathscr{U}| \geq k \cdot n$ and there exists a subset $\mathscr{U}_0$ of size k where each of the pairs of values hash to the same key.
Let m = k + 1: Assume each of the keys that the values can be hashed to are called buckets ($\{0, ..., n-1\}$). When $|\mathscr{U}| \geq (k+1) \cdot n$, the smallest maximum size of a bucket is when all of the values are evenly distributed over all the keys. In the case where $|\mathscr{U}| = (k+1) \cdot n$, this would mean that all buckets would have $k+1$ entries. If this were the case then any of the buckets could be the subset $\mathscr{U}_0$. For a bucket to have fewer than $k+1$ values hashed to it, another bucket would have greater than $k+1$ values hashed to it. As such, the latter of these two buckets could be $\mathscr{U}_0$; $\mathscr{U}_0$ equals some random selection of $k+1$ elements from this bucket.
By the principle of induction **(1)** is true for all $n \in \mathbb{Z}^+ \square$.
If you have to hash v values, and k keys, and for the first value of m for which $v \geq m \cdot k$, then there is a subset of size m values which hash to the same key. As such in hashing with chaining, the worst case time complexity to find a value is $\Omega(m)$.

Question 2

(a)

There are $n = 1000$ different buckets that values could be hashed into. This means that the probability that two values are hashed to two different buckets is $1 - \frac{1}{1000} = \frac{999}{1000}$. This means that the probability of two values hashing to the same bucket in k turns is $1-(\frac{999}{1000})^{\binom{k}{2}}$. As such, we need to solve the inequality $1-(\frac{999}{1000})^{\binom{k}{2}} \geq 0.8$ for k. Let $k=57$, then $1-(\frac{999}{1000})^{\binom{57}{2}} = 1-(\frac{999}{1000})^{1596} = 0.7975$. Now let $k=58$, then $1-(\frac{999}{1000})^{\binom{58}{2}} = 1-(\frac{999}{1000})^{1653} = 0.8087$. This means that on the 58th insertion the probability of collision exceeds 80% for the first time.

(b)

We are given that in the first $k-1$ insertions there have been no collisions and that on the k-th insertion the probability of collision must be strictly less than 20%. The first fact means that $k-1$ buckets are used in the hash table $\Rightarrow$ the probability of collision in the k-th insertion should be less than $\frac{k-1}{1000}$.

From the second fact provided, we know that $\frac{k-1}{1000} < 0.2 \Rightarrow k-1 < 200 \Rightarrow k < 201$. As such, the value of k when the size of the hash table should be increased and thus the first time that the collision probability is greater than or equal to 20% is on the **200th insertion**.

Question 3

Need to prove that the expected running time of an entire sequence of m operations, on a hash table of size n, is upper bounded by $m \cdot (1+\frac{m}{2n})$. H is a 2-universal family of hash functions from which we use hash functions. E_i denotes the expected running time of i-th operation, and K_i is the expected number of collisions on the i-th operation. Can assume that $E_i = 1 + K_i$.

There are m operations, therefore, using the notation above, the overall expected running time, E is $\sum_{i=1}^{m} E_i$. Let $E_{max} = max(E_1, ..., E_m)$. $\therefore E \leq m \cdot E_{max}$, and thus $E \leq m \cdot (1+K_{max})$. As the family of hash functions we are using is 2-universal we can say that $K_i = P(h(x_i) = h(x_j)) \leq \frac{1}{n}$ where h is an arbitrary hash function from H and x_j is some random value in the hash table. Let K be the expected number of collisions over all m operations, we can say that $K \leq \binom{m}{2} \cdot \frac{1}{n} \leq \frac{m^2}{2n}$. The expected number of collisions for one key is $K_i \leq \frac{1}{m} \cdot \frac{m^2}{2n} = \frac{m}{2n}$. Using the earlier relation ($E_i = 1 + K_i$), we get $E_i \leq 1 + \frac{m}{2n}$. There are m operations and so the total expected running time is upper bounded as follows: $E \leq m \cdot E_i \leq m \cdot (1+\frac{m}{2n})$ $\square$.

2 Information Theory

Question 4

The capacity of a channel is defined as follows: $C = \max_{p(x)} I(X;Y)$. $I(X;Y) = H(Y) - H(Y|X)$. This gives us $I(X;Y) = H(Y)$ because of the fact that the channel is deterministic; if $Y = f(X)$, then $H(Y|X) = 0$. $\therefore C = \max_{p(x)} H(Y) = \max_{p(x)} H(f(X))$. $p(x)$ is the probability distribution of, which maximises entropy when the distribution is uniform. $\therefore C = log|f(X)|$ when $p(x)$ is a uniform distribution $\square$.

Question 5

Property 2

Line 1: The sum of probabilities of all input sequences $x^n \in \mathscr{X}^n$ is equal to 1.
Line 2: The typical set $A_\varepsilon^{(n)}$ is a subset of $\mathscr{X}^n$ with particular properties. Therefore, it is obvious that the number of sequences in the typical set is not greater than the number of sequences in $\mathscr{X}^n$. Thus, the sum of the probabilities for this subset is ≤ 1.
Line 3: One of the properties of typical set is that $H(X) - \varepsilon \leq -\frac{1}{n}\log p(x^n) \leq H(X) + \varepsilon$. Will focus on the RHS of this inequality; $-\frac{1}{n}\log p(x^n) \leq H(X) + \varepsilon$. Rearranging this we get $p(x^n) \geq 2^{-n(H(X)+\varepsilon)}$. $\therefore$ we get the inequality shown.
Line 4: The value $2^{-n(H(X)+\varepsilon)}$ is constant so we can pull it out of the sum $\sum_{x^n \in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} = 2^{-n(H(X)+\varepsilon)} \sum_{x^n \in A_\varepsilon^{(n)}} 1 = 2^{-n(H(X)+\varepsilon)} |A_\varepsilon^{(n)}|$. To complete the proof simply multiply both sides of the inequality by $2^{n(H(X)+\varepsilon)}$. This leaves $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$, as required.

Property 3

Line 1: The first line makes use of Property 1 of the AEP theorem for typical sequences. It states that, for n large enough, the probability of typical set is nearly 1. Should be noted that $Pr\{A_\varepsilon^{(n)}\} = \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)$.
Line 2: One of the properties of typical set is that $H(X) - \varepsilon \leq -\frac{1}{n}\log p(x^n) \leq H(X) + \varepsilon$. Will now focus on the LHS of this inequality; $-\frac{1}{n}\log p(x^n) \geq H(X) - \varepsilon$. Rearranging this we get $p(x^n) \leq 2^{-n(H(X)-\varepsilon)}$. $\therefore$ we get the inequality shown.
Line 3: The value $2^{-n(H(X)-\varepsilon)}$ is constant so we can pull it out of the sum $\sum_{x^n \in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = 2^{-n(H(X)-\varepsilon)} \sum_{x^n \in A_\varepsilon^{(n)}} 1 = 2^{-n(H(X)-\varepsilon)} |A_\varepsilon^{(n)}|$. To complete the proof simply multiply both sides of the inequality by $2^{n(H(X)-\varepsilon)}$. This leaves $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)2^{n(H(X)+\varepsilon)}$, as required.

Question 6

(a)

$H(X_{S\cup T})$ is the joint entropy of all the discrete random variables that are members of sets X_S or X_T or both. We can see that $X_S, X_T = X_{s_1}, \ldots, X_{s_k}, X_{t_1}, \ldots, X_{t_m}$ and $X_{S\cup T} = X_S, X_T = X_{s_1}, \ldots, X_{s_k}, X_{t_1}, \ldots, X_{t_m} \Rightarrow H(X_{S\cup T}) = H(X_S, X_T)$ (def. 1).
$H(X_{S\cap T})$ is the joint entropy of all the discrete random variables that are members of both X_S and X_T. Intuitively, the mutual information, $I(X_S;X_T)$ is the amount of information in common about X_S and X_T $\therefore H(X_{S\cap T}) \leq I(X_S;X_T)$. We prove this below:

$$I(X_S;X_T) = H(X_S,X_T) - H(X_T|X_S) - H(X_S|X_T) \text{ - Using the Venn Diagram}$$
$$\geq H(X_S,X_T) = H(X_{S\cup T}) \text{ - Using def. 1 (def. 2)}$$

The number of members of the set $S\cap T$ is less than or equal to the number of members of $S\cup T$. This means that the number of random variables in the joint entropy of $H(X_{S\cap T})$ is less than or equal to the number in $H(X_{S\cup T})$. The joint entropy can be rewritten as:

$$H(X_1,...,X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1},...,X_1)$$

$\therefore$ as the number of variables in the joint entropy $H(X_{S\cap T})$ is less than or equal to the number in $H(X_{S\cup T})$, this means that the number of entries in the sum if also fewer for $H(X_{S\cap T})$. Hence, because conditional entropy is always greater than 0, the summation for $H(X_{S\cup T})$ is greater than or equal to the summation for $H(X_{S\cap T})$, and thus the following is true:

$$H(X_{S\cup T}) \geq H(X_{S\cap T})$$
$$\Rightarrow I(X_S;X_T) \geq H(X_S \cap X_T) \text{ - from def. 2 } \square.$$

$H(X_{S\cap T})+H(X_{S\cup T}) \leq H(X_S,X_T)+I(X_S;X_T) = H(X_S,X_T)+H(X_S)+H(X_T)-H(X_S,X_T) = H(X_S)+H(X_T)$ $\square$.

(b)

$$\sum_{i=1}^{m} H(X_{[m]/\{i\}}) = H(X_2,...,X_m)+H(X_1,X_3,...,X_m)+...+H(X_1,...,X_{m-1})$$
$$= \sum_{i=1}^{m} H(X_i|X_{i-1},...,X_2) - H(X_1) + \sum_{i=1}^{m} H(X_i|X_{i-1},...,X_3,X_1)$$
$$-H(X_2|X_1)+...+\sum_{i=1}^{m-1} H(X_i|X_{i-1},...,X_1)$$

Conditioning reduces entropy which means that adding the respective missing discrete random variable to RHS of each conditional entropy reduces the overall entropy.

$$\therefore \sum_{i=1}^{m} H(X_{[m]/\{i\}}) \geq m\sum_{i=1}^{m} H(X_i|X_{i-1},...,X_1) - H(X_1) - H(X_2|X_1) - ... - H(X_m|X_{m-1},...,X_1)$$
$$= m\sum_{i=1}^{m} H(X_i|X_{i-1},...,X_1) - \sum_{i=1}^{m} H(X_i|X_{i-1},...,X_1)$$
$$= (m-1)\sum_{i=1}^{m} H(X_i|X_{i-1},...,X_1)$$
$$= (m-1)H(X_1,...,X_m) = (m-1)H(X_{[m]}).$$

Thus, $\sum_{i=1}^{m} H(X_{[m]/\{i\}}) \leq (m-1)H(X_{[m]})$ $\square$.