

CSE3358 Problem Set 6 Solution

Problem 2: Yet another variation

Consider k lists that are not necessarily sorted containing a total of n elements. In this variation, all the elements in the first list are less than or equal to all the elements in the second list, and so on... One possible method for sorting the elements is to sort the individual lists independently and then concatenate the sorted results. This will take $\Theta(\sum_i n_i \log n_i)$ where n_i is the number of elements in list i .

Algorithm OBVIOUS

```
for  $i \leftarrow 1$  to  $k$ 
  do sort list  $i$            ▷ for example using merge sort
concatenate the lists
```

Although the above algorithm is OBVIOUS, nothing better can be done. Regardless of what algorithm we use for this variation of the sorting problem, show that $\Omega(\sum_i n_i \log n_i)$ time is needed. Note that it is not rigorous to sum the individual lower bounds because an algorithm does not necessarily work like the OBVIOUS algorithm.

Note: If all lists have the same size, this bound will be $\Omega(n \log n/k)$.

ANSWER: Well, let's see how many possible answers we can have for this sorting problem. Obviously, we can permute elements within a list, but not across lists. Therefore, any permutation consisting of permuting elements within lists could be the answer for the sorting problem. The number of these permutations is simply $n_1!n_2!\dots n_k! = \prod_i n_i!$.

In the decision tree representation of the sorting algorithm, the number of leaves has to be therefore greater or equal to $\prod_i n_i!$. Since the decision tree is a binary tree, $2^h \geq \text{number leaves}$. Therefore, $2^h \geq \prod_i n_i!$. Taking log on each side, we get

$$h \geq \log \prod_i n_i! = \sum_i \log n_i! = \sum_i \Theta(n_i \log n_i)$$

using the fact that $\log x! = \Theta(x \log x)$. Therefore $h \geq \Theta(\sum_i n_i \log n_i)$, which means $h = \Omega(\sum_i n_i \log n_i)$.

Since the running time of the sorting algorithm in the worst case is $\Theta(h)$, the running time is $\Omega(\sum_i n_i \log n_i)$.

Problem 9: Constructing a universal family of hash function

In this problem we will look at an easy way for constructing a universal family of hash function. Recall the definition of a universal hash family.

Universal Hash Family H : For any two keys k and l , let $S_{kl} = \{h \in H | h(k) = h(l)\}$, then $|S_{kl}| \leq \frac{|H|}{m}$, where m is the size of the hash table.

The important consequence of the above definition is that by picking a hash function at random from a universal family of hash functions, the probability that two keys will collide is $\leq \frac{1}{m}$. This makes the expected number of keys that collide with a particular key k , at most $\frac{n-1}{m}$ (each of the $n-1$ other

keys has a probability $\leq \frac{1}{m}$ of being in the list). Therefore, the expected length of the list in which k hashes to is at most $1 + \frac{n-1}{m} = O(1 + \alpha)$.

Without loss of generality, let the keys be decimal numbers with n digits. In other words, every key k has the form $k = x_1x_2\dots x_n$ where $x_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Let m be a prime number greater than 10 (we will see why later) and consider the following family of hash functions.

$$h(k = x_1x_2\dots x_n) = \left(\sum_{i=1}^n a_i x_i\right) \bmod m$$

for $a_i \in \{0, 1, 2, \dots, m-1\}$.

(a) What is the size of H ?

ANSWER: For any choice of a_1, a_2, \dots, a_n , we have a hash function. Therefore, since $a_i \in \{0, 1, 2, \dots, m-1\}$, we have m^n hash functions in H .

(b) What needs to be done to pick a hash function from H uniformly at random?

ANSWER: It is enough to pick every a_i from the set $\{0, 1, 2, \dots, m-1\}$ independently and uniformly at random. Therefore, for every $i = 1..n$, let a_i be any element of $\{0, 1, 2, \dots, m-1\}$ with probability $1/m$.

Now we want to see if H is universal. This means we need to verify whether $S_{kl} \leq \frac{|H|}{m}$ for every pair of keys k and l . Let's assume that $k = x_1\dots x_n$ and $l = y_1\dots y_n$ collide. This means

$$\sum_{i=1}^n a_i x_i \equiv \sum_{i=1}^n a_i y_i \pmod{m} \quad k \neq l$$

Since $k \neq l$ they must differ in at least one digit. Without loss of generality, let's say they differ in the first digit, i.e. $x_1 \neq y_1$. Therefore, we can rewrite the above as:

$$a_1(x_1 - y_1) \equiv \sum_{i=2}^n a_i(y_i - x_i) \pmod{m} \quad x_1 - y_1 \neq 0$$

This is where the choice of m being prime becomes important. From number theory, if m is prime, then every integer $z \in \{1, 2, \dots, m-1\}$ has a **unique inverse** $z^{-1} \in \{1, 2, \dots, m-1\}$ such that $zz^{-1} \equiv 1 \pmod{m}$.

(c) Argue that $x_1 - y_1$ has a unique inverse $(x_1 - y_1)^{-1}$ such that $(x_1 - y_1)(x_1 - y_1)^{-1} \equiv 1 \pmod{m}$.

First, m is prime. Second, $0 \leq x_i \leq 9$ and $0 \leq y_i \leq 9$. Without loss of generality we can assume that $x_i > y_i$; therefore, $x_i - y_i \leq 9$. Since $m > 10$, $x_i - y_i \in \{1, 2, \dots, m-1\}$. By the property stated above, $x_i - y_i$ has a unique inverse $(x_i - y_i)^{-1}$ such that $(x_i - y_i)(x_i - y_i)^{-1} \equiv 1 \pmod{m}$.

(d) By multiplying both sides of the equation above by $(x_1 - y_1)^{-1}$, show that a_1 is uniquely determined from a_2, a_3, \dots, a_n .

ANSWER:

$$\begin{aligned}a_1(x_1 - y_1) &\equiv \sum_{i=2}^n a_i(y_i - x_i) \pmod{m} \\a_1(x_1 - y_1)(x_1 - y_1)^{-1} &\equiv \left(\sum_{i=2}^n a_i(y_i - x_i)\right)(x_1 - y_1)^{-1} \pmod{m} \\a_1 \cdot 1 &\equiv \left(\sum_{i=2}^n a_i(y_i - x_i)\right)(x_1 - y_1)^{-1} \pmod{m} \\a_1 &\equiv \left(\sum_{i=2}^n a_i(y_i - x_i)\right)(x_1 - y_1)^{-1} \pmod{m}\end{aligned}$$

Thus a_1 is uniquely determined from a_2, a_3, \dots, a_n .

(e) Using part (d), find the number of hash functions $|S_{kl}|$ that can cause k and l to collide, and verify that H is universal.

The number of hash functions that cause k and l to collide is m^{n-1} because for every choice of a_2, a_3, \dots, a_n , a_1 is uniquely determined such that $h(k) = h(l)$, and we have m^{n-1} possible choices for a_2, a_3, \dots, a_n . Therefore, $|S_{kl}| = m^{n-1}$. This means that H is universal because $|S_{kl}| = m^{n-1} = \frac{m^n}{m} = \frac{|H|}{m}$.