CS 70 SPRING 2008 — DISCUSSION #13

LUQMAN HODGKINSON, AARON KLEINMAN, MIN XU

1. Distributions

Exercise 1. A drawer contains 10 socks. 6 of them have holes, and 4 of them do not. What's the probability that after pulling out two random socks at the same time 5 times (and putting back each pair after you look at it), you pull out a pair with no holes precisely 4 out of 5 times?

Exercise 2. Min is trying to hitchhike along a deserted road. The probability that a car drives by during any given minute is 5%.

(a) What's the probability that no cars will appear in a particular span of 20 minutes, assuming that the cars travelling along the road are independent?

(b) Now suppose Min has waited 20 minutes and no cars have come. What's the probability that no cars will arrive during the next 20 minutes?

Exercise 3. Suppose a box has 6 brown balls and 4 purple balls. A random sample of size n is selected with replacement. Let $X=$ "number of brown balls selected". Write the distribution of X. Can this be closely approximated with a Poisson distribution? If so, write the approximate distribution. If not, explain why not. If so, but only under certain conditions, explain these conditions and write the approximate distribution.

Exercise 4. Now suppose the box has 6 brown balls and 4,000 purple balls. As in the previous exercise a random sample of size n is selected with replacement and $X=$ "number of brown balls selected". Write the distribution of X. Can this be closely approximated with a Poisson distribution? If so, write the approximate distribution. If not, explain why not. If so, but only under certain conditions, explain these conditions and write the approximate distribution.

Exercise 5. Suppose that in the scenario from Quiz 13, Professor Wagner buys two light bulbs from a web merchant who sells bulbs that have a probability p of burning out on any given day. He installs the first bulb in his kitchen. Let the random variable X denote the number of days until it burns out. When it burns out, he replaces it with the second bulb. Let the random variable Y denote the number of days until the second bulb burns out. That is, X and Y are independent geometrically distributed random variables with parameter p. What is the distribution of the total number of days until the two bulbs both burn out?

2. Perfect Hashing

Years after CS70, Aaron is now working as the Chief Technology Officer of AT&T. The CEO, jealous of Aaron's talents, gives Aaron a seemingly impossible task: the CEO wants Aaron to design an semi-efficient (i.e. $O(\sqrt{n})$ look-up time) hashing scheme that maps a set of n 10-digit phone numbers (call this set S) to a hash table of size n, where n is a prime number and much smaller than 10^{10}. The tricky part of this challenge is that Aaron's hashing scheme must work semi-efficiently for any subset of n phone numbers and Aaron will be fired if his hash function is inefficient for any subset S.

Aaron at first despairs since the pigeon hole principle says that if n is much smaller than 10^{10}, then no matter how clever Aaron's hash function is, there will always exist a subset S of all phone numbers where S creates lots of collisions in the hash table.

Exercise 6. Just as Aaron is cleaning up his desk, he suddenly remembered his CS70 and designed the following hashing scheme: Instead of using a single hash function, Aaron's scheme randomly chooses a hash function from a set of hash functions H. To be more precise, Aaron will first uniformly at random choose $a_1, a_2, a_3, a_4, a_5 \in \{0, ...n-1\}$ and use the result to create a hash function to hash all the phone numbers.

Suppose Aaron is given phone number x, then let x_1, x_2, x_3, x_4, x_5 represent first, second, third, etc 2-digits of the number respectively. The hash function will then be

Date: April 29, 2008.

$$h(x) = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 \pmod{n}$$

Prove that the probability that such a randomly chosen hash function will map 2 different phone numbers to the same index is $1/n$. You should assume that $n > 10^2$.

Exercise 7. Let X be the number of collisions that Aaron's hashing scheme creates after it hashes every single member of S. What is $\mathbf{E}[X]$? (Note, hashing 4 different phone numbers to the same index creates $\binom{4}{2} = 6$ collisions, not 4)

Exercise 8. Let Y be the number of elements in the hash table index with the largest number of elements (i.e. the most crowded index). Prove that $\mathbf{E}[Y] \leq 2\sqrt{n}$. (Hint: think about the amount of collisions created in the most crowded index and use the previous problem. You may also find the inequality $\mathbf{E}[Y]^2 \leq \mathbf{E}[Y^2]$ helpful).

Exercise 9. Based on the previous problem, we know that on average, Aaron's hashing scheme is pretty good; the most crowded index has on average, at most $\sqrt{2n}$ elements so look-up is at worst $O(\sqrt{2n})$. Give an upper bound on the probability that the most crowded index contains more than $2\sqrt{2n}$ element.

Exercise 10. Although Aaron's scheme is semi-efficient with a high probability, propose a simple way to enhance Aaron's hashing algorithm such that he can guarantee that his hashing scheme is always semi-efficient. Show that your modification is practical in the sense that it doesn't affect the look-up time and allows for reasonable creation time on average.