CS 361, Lecture 11

Jared Saia
University of New Mexico

Outline

- Birthday Paradox
- Analysis of Randomized Quicksort

Probability Definitions

(from Appendix C.3)

- A *random variable* is a variable that takes on one of several values, each with some probability. (Example: if X is the outcome of the role of a die, X is a random variable)
- The *expected value* of a random variable, X is defined as:

$$E(X) = \sum_x x * P(X = x)$$

(Example if X is the outcome of the role of a three sided die,

$$\begin{aligned} E(X) &= 1 * (1/3) + 2 * (1/3) + 3 * (1/3) \\ &= 2 \end{aligned}$$

Probability Definitions

- Two events A and B are *mutually exclusive* if $A \bigcap B$ is the empty set (Example: A is the event that the outcome of a die is 1 and B is the event that the outcome of a die is 2)
- Two random variables X and Y are *independent* if for all x and y, $P(X = x$ and $Y = y) = P(X = x)P(Y = y)$ (Example: let X be the outcome of the first role of a die, and Y be the outcome of the second role of the die. Then X and Y are independent.)

Probability Definitions

- An *Indicator Random Variable* associated with event A is defined as:
  - $I(A) = 1$ if A occurs
  - $I(A) = 0$ if A does not occur
- Example: Let A be the event that the role of a die comes up 2. Then $I(A)$ is 1 if the die comes up 2 and 0 otherwise.

Linearity of Expectation

- Let X and Y be two random variables
- Then $E(X+Y) = E(X) + E(Y)$
- (Holds even if X and Y are not independent.)

- More generally, let $X_1, X_2, \ldots, X_n$ be n random variables
- Then

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$$

Example

- For $1 \leq i \leq n$, let X_i be the outcome of the i-th role of three-sided die
- Then

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = 2n$$

Example

- Indicator Random Variables and Linearity of Expectation used together are a very powerful tool
- The "Birthday Paradox" illustrates this point
- To analyze the run time of quicksort, we will also use indicator r.v.'s and linearity of expectation (analysis will be similar to "birthday paradox" problem)

"Birthday Paradox"

- Assume there are k people in a room, and n days in a year
- Assume that each of these k people is born on a day chosen uniformly at random from the n days
- Q: What is the expected number of pairs of individuals that have the same birthday?
- We can use indicator random variables and linearity of expectation to compute this

Analysis

- For all $1 \leq i < j \leq k$, let $X_{i,j}$ be an indicator random variable defined such that:
  - $X_{i,j} = 1$ if person i and person j have the same birthday
  - $X_{i,j} = 0$ otherwise
- Note that for all i, j,

$$\begin{aligned} E(X_{i,j}) &= P(\text{person i and j have same birthday}) \\ &= 1/n \end{aligned}$$

Analysis

- Let X be a random variable giving the number of pairs of people with the same birthday
- We want $E(X)$
- Then $X = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} X_{i,j}$
- So $E(X) = E(\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} X_{i,j})$

Analysis

$$\begin{aligned} E(X) &= E(\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} X_{i,j}) \\ &= \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} E(X_{i,j}) \\ &= \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} 1/n \\ &= \binom{k}{2} 1/n \\ &= \frac{k(k-1)}{2n} \end{aligned}$$

The second step follows by Linearity of Expectation

Reality Check

- Thus, if $k(k-1) \geq 2n$, expected number of pairs of people with same birthday is at least 1
- Thus if have at least $\sqrt{2n}+1$ people in the room, can expect to have at least two with same birthday
- For $n = 365$, if $k = 28$, expected number of pairs with same birthday is 1.04

In-Class Exercise

- Assume there are k people in a room, and n days in a year
- Assume that each of these k people is born on a day chosen uniformly at random from the n days
- Let X be the number of groups of *three* people who all have the same birthday. What is $E(X)$?
- Let $X_{i,j,k}$ be an indicator r.v. which is 1 if people i,j, and k have the same birthday and 0 otherwise

In-Class Exercise

- Q1: Write the expected value of X as a function of the $X_{i,j,k}$ (use linearity of expectation)
- Q2: What is $E(X_{i,j,k})$?
- Q3: What is the total number of groups of three people out of k?
- Q4: What is $E(X)$?

Scratch Space

R-Partition

```
//PRE: A[p..r] is the array to be partitioned, p>=1 and r <= size
//     of A
//POST: Let A' be the array A after the function is run.  Then
//      A'[p..r] contains the same elements as A[p..r].  Further,
//      all elements in A'[p..res-1] are <= A[i], A'[res] = A[i],
//      and all elements in A'[res+1..r] are > A[i], where i is
//      a random number between $p$ and $r$.
R-Partition (A,p,r){
  i = Random(p,r);
  exchange A[r] and A[i];
  return Partition(A,p,r);
}
```


Randomized Quicksort

```
//PRE: A is the array to be sorted, p>=1,  and r is <= the size of A
//POST: A[p..r] is in sorted order
R-Quicksort (A,p,r){
  if (p<r){
    q = R-Partition (A,p,r);
    R-Quicksort (A,p,q-1);
    R-Quicksort (A,q+1,r);
}
```


Analysis

- R-Quicksort is a *randomized* algorithm
- The run time is a *random variable*
- We'd like to analyze the *expected* run time of R-Quicksort
- To do this, we first need to learn some basic probability theory.

Plan of Attack

"If you get hold of the head of a snake, the rest of it is mere rope" - Akan Proverb

- We will analyze the *total* number of comparisons made by quicksort
- We will let X be the total number of comparisons made by R-Quicksort
- We will write X as the sum of a bunch of indicator random variables
- We will use linearity of expectation to compute the expected value of X

Notation

- Let A be the array to be sorted
- Let z_i be the i-th smallest element in the array A
- Let $Z_{i,j} = \{z_i, z_{i+1}, \ldots, z_j\}$

Indicator Random Variables

- Let $X_{i,j}$ be 1 if z_i is compared with z_j and 0 otherwise
- Note that $X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{i,j}$
- Further note that

$$E(X) = E(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{i,j}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E(X_{i,j})$$

Questions

- Q1: So what is $E(X_{i,j})$?
- A1: It is $P(z_i$ is compared to $z_j)$
- Q2: What is $P(z_i$ is compared to $z_j)$?
- A2: It is:

  $P($either z_i or z_j are the first elems in $Z_{i,j}$ chosen as pivots)

- Why?
  - If no element in $Z_{i,j}$ has been chosen yet, no two elements in $Z_{i,j}$ have yet been compared, and all of $Z_{i,j}$ is in same list
  - If some element in $Z_{i,j}$ other than z_i or z_j is chosen first, z_i and z_j will be split into separate lists (and hence will never be compared)

More Questions

- Q: What is

  $P($either z_i or z_j are first elems in $Z_{i,j}$ chosen as pivots)

- A: $P(z_i$ chosen as first elem in $Z_{i,j}) +$ $P(z_j$ chosen as first elem in $Z_{i,j})$
- Further note that number of elems in $Z_{i,j}$ is $j - i + 1$, so

$$P(z_i \text{ chosen as first elem in } Z_{i,j}) = \frac{1}{j - i + 1}$$

and

$$P(z_j \text{ chosen as first elem in } Z_{i,j}) = \frac{1}{j - i + 1}$$

- Hence

$$P(z_i \text{ or } z_j \text{ are first elems in } Z_{i,j} \text{ chosen as pivots}) = \frac{2}{j - i + 1}$$

Conclusion

$$
\begin{aligned}
E(X_{i,j}) &= P(z_i \text{ is compared to } z_j) && (1)\\
&= \frac{2}{j-i+1} && (2)
\end{aligned}
$$

Putting it together

$$
\begin{aligned}
E(X) &= E\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{i,j}\right) && (3)\\
&= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E(X_{i,j}) && (4)\\
&= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{2}{j-i+1} && (5)\\
&= \sum_{i=1}^{n-1}\sum_{k=1}^{n-i} \frac{2}{k+1} && (6)\\
&< \sum_{i=1}^{n-1}\sum_{k=1}^{n} \frac{2}{k} && (7)\\
&= \sum_{i=1}^{n-1} O(\log n) && (8)\\
&= O(n \log n) && (9)
\end{aligned}
$$

Questions

- Q: Why is $\sum_{k=1}^{n} \frac{2}{k} = O(\log n)$?
- A:

$$
\begin{aligned}
\sum_{k=1}^{n} \frac{2}{k} &= 2\sum_{k=1}^{n} 1/k && (10)\\
&\leq 2(\ln n + 1) && (11)
\end{aligned}
$$

- Where the last step follows by an integral bound on the sum (p. 1067)

Take Away

- The expected number of comparisons for r-quicksort is $O(n \log n)$
- Competitive with mergesort and heapsort
- Randomized version is "better" than deterministic version

Todo

- Finish Chapter 7