COMPSCI 330: Design and Analysis of Algorithms — September 21, 2017

Lecture 8 : Randomized Algorithms

Lecturer: Rong Ge — *Scribe: Will Wang*

1 Basic Probability

Randomized algorithms are algorithms that make use of use of random decisions. Before we delve into them, we first review of basic probability.

Random Variable: Variable whose value depend on a random phenomenon. For example, tossing a fair coin, we have 0.5 probability to get either a head or a tail.

Joint Probability: The probability of event Y occurring at the same time event X occurs. $P(X = i, Y = j)$

Independence: The probability that one event occurs in no way affects the probability of the other event occurring. $P(X = i, Y = j) = P(X = i)P(Y = j)$

Conditional Probability: Probability of event X given we already know occurance of event Y. $P(X = i|Y = j) = \frac{P(X=i,Y=j)}{P(Y=j)}$

Expectation: Weighted average of the possible values that X can take, each value being weighted according to the probability of that event occurring. $E(X) = \sum X * P(X)$

Conditional Expectation: Expectations of conditioned random variable $E(X|Y = j) = \sum X * P(X|Y = j)$

Law of Total Expectation: Expectations of conditioned random variable $E(X) = \sum E(X|Y = j) * P(Y = j)$. Here the $E(X)$ can be interpreted as expected running time of ALG, when $E(X|Y = j)$ is runtime of ALG after fixing the first decision and $P(Y = j)$ is first random decision in ALG.

1.1 Relationship Between Joint Probabilities and Conditional Probabilities

Here we show the relationship between joint probabilities and conditional probabilities:

$$\begin{aligned} P(X = i, Y = j) &= P(X = i)P(Y = j|X = i) \\ &= P(Y = j)P(X = i|Y = j) \\ P(Y = j|X = i) &= \frac{P(Y = j)P(X = i|Y = j)}{P(X = i)} \qquad \text{Bayes Theorem} \end{aligned}$$

1.2 Prove Linearity of Expectation

Here we prove the linearity of expectation $E(X+Y) = E(X) + E(Y)$:

$$\begin{aligned}
P(X+Y=k) &= \sum_{i+j=k} P(X=i, Y=j) \\
E(X+Y) &= \sum_{k} P(X+Y=k)k \\
&= \sum_{k} \sum_{i+j=k} P(X=i, Y=j)(i+j) \\
&= \sum_{k} \sum_{i+j=k} P(X=i, Y=j)i + \sum_{k} \sum_{i+j=k} P(X=i, Y=j)j \\
&= \sum_{i,j} P(X=i, Y=j)i + \sum_{i,j} P(X=i, Y=j)j \\
&= \sum_{i} (\sum_{j} P(X=i, Y=j)i) + \sum_{j} (\sum_{i} P(X=i, Y=j)j) \\
&= \sum_{i} P(X=i)i + \sum_{j} P(Y=j)j \\
&= E(X) + E(Y)
\end{aligned}$$

2 Two Types of Randomized Algorithms

Randomized algorithms can be classified into two types:

Las Vegas Algorithm:

Always outputs the correct answer but running time is random.
Analysis: Compute expected running time.

Monte Carlo Algorithm:

Always run in a fixed amount of time but result may be incorrect.
Requirement: Result is correct with probability at least 2/3.

3 Quick Sort

We are interested in the average/expected running time of QuickSort which can be categorized into Las Vegas Alogrithm. Recall the algorithm: it first divides a large array into two smaller sub-arrays: the low elements and the high elements. Quicksort can then recursively sort the sub-arrays.

Example: to sort a list of numbers a[] = 4, 2, 8, 6, 3, 1, 7, 5. We first pick a random pivot number (say 3). Then we partition the array into numbers smaller and larger than the pivot (2, 1, 4, 8, 6, 7, 5). And then we recursively sort the two parts.

Worst Case: Suppose we always pick the first number. Then we can easily see the the running time is $\Theta(n^2)$.

Randomness Helps: The goal here is to analyze the expected running time of quicksort. Our intuition is that the running time looks like $O(nlogn)$ on average. We prove this by using induction in the following:

Let X_n be running time of quick sort with n numbers. Then $E(X_n)$ is the average running time we are trying to analyze.

Induction Hypothesis: $E(X_n) \leq Cn \log_2 n$

Recursion:

$$\begin{aligned}
E(X_n) &= \sum_{i=1}^{n} P(\text{Pivot number is ith smallest}) E(X_n|\text{pivot number is ith smallest}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(E(X_{i-1} + E(X_{n-1})) + An)
\end{aligned}$$

Base Case: $E(X_0) = 0$, $E(X_1) = 0$
Induction Step:

$$\begin{aligned}
E(X_n) &= \frac{1}{n}\sum_{i=1}^{n}(E(X_{i-1}) + E(X_{n-1})) + An \\
&= \frac{2}{n}\sum_{i=1}^{n} E(X_{i-1}) + An \\
&\leq \frac{2}{n}C\sum_{i=1}^{n}(i-1)\log_2(i-1) + An \\
&\leq \frac{2}{n}C[\sum_{i=1}^{\frac{n}{2}}(i-1)\log_2(\frac{n}{2}) + \sum_{i=\frac{n}{2}+1}^{n}(i-1)\log_2 n] + An \\
&= \frac{2}{n}C[\sum_{i=1}^{\frac{n}{2}}(i-1)(\log_2 n - 1) + \sum_{i=\frac{n}{2}+1}^{n}(i-1)\log_2 n] + An \\
&= \frac{2}{n}C[\sum_{i=1}^{n}(i-1)\log_2 n - \sum_{i=1}^{\frac{n}{2}}(i-1)] + An \\
&= \frac{2}{n}C[\frac{n(n-1)}{2}\log_2 n - \frac{\frac{n}{2}(\frac{n}{2}-1)}{2}] + An \\
&\leq C * n * \log_2 n - C * \frac{n}{4} + An \\
&\leq C * n * log_2 n \qquad \text{when } C \geq 4A
\end{aligned}$$

Therefore the running time is bounded by $4 * A * n * log_2 n$