An Introduction to Randomized Algorithms

The focus of this lecture is to study a randomized algorithm for quick sort, analyze it using probabilistic recurrence relations, and also provide more general tools for analysis of randomized algorithms. For a quick overview of probability and terms associated with it, the reader is advised to see the Appendix.

1 Randomized QuickSort

In this section, we present the classic quick sort algorithm and compute the expected running time of the algorithm. We assume that the elements of the set are all distinct. Below is the randomized quick sort algorithm.

```
 Algorithm RandQuickSort(S)
Choose a pivot element x_i u.a.r from S = {x_1, x_2, ..., x_n}
Split the set S into two subsets S_1 = {x_j | x_j < x_i}
and S_2 = {x_j | x_j > x_i} by comparing each x_j with the chosen x_i
Recurse on sets S_1 and S_2
Output the sorted set S_1 then x_i and then sorted S_2.
end Algorithm
```

Analysis

Let $T(n)$ be the number of steps taken by the RandQuickSort algorithm on a set of size n. Note that the maximum value of $T(n)$ occurs when the pivot element x_i is the largest/smallest element of the remaining set during each recursive call of the algorithm. In this case, $T(n) = n + (n-1) + \cdots + 1 = O(n^2)$. This value of $T(n)$ is reached with a very low probability of $\frac{1}{n} \cdot \frac{1}{n-1} \cdot \cdots \frac{1}{2} = \frac{1}{n!}$. Also, the best case occurs when the pivot element splits the set S into two equal sized subsets and then $T(n) = O(n \ln n)$. This implies that $T(n)$ has a distribution between $O(n \ln n)$ and $O(n^2)$. Now we derive the expected value of $T(n)$. Note that if the i^{th} smallest element is chosen as the pivot element then S_1 and S_2 will be of sizes $i-1$ and $n-i-1$ respectively and this choice has a probability of $\frac{1}{n}$. The recurrence relation for $T(n)$ is:

$$T(n) = n + T(X) + T(n-1-X) \tag{1}$$

where, $Pr[X = i] = \frac{1}{n}$ for $i \in \{0, 1, \cdots, n-1\}$. Hence, $Pr[n-1-X = i] = Pr[X = n-1-i] = \frac{1}{n}$. However, it is incorrect to write $T(n-1-X) = T(X)$.

Taking expectations on both sides of (1),

$$\begin{aligned} E[T(n)] &= n + \tfrac{1}{n}\textstyle\sum_{i=1}^{n-1} E[T(i)] + \tfrac{1}{n}\sum_{j=1}^{n-1} E[T(j)] \\ &= n + \tfrac{2}{n}\textstyle\sum_{i=1}^{n-1} E[T(i)]. \end{aligned}$$

Let $f(i) = E[T(i)]$. Then, $f(n) = n + \frac{2}{n}\sum_{i=1}^{n-1} f(i)$. Simplifying,

$$nf(n) = n^2 + 2(f(1) + f(2) + \cdots + f(n-1)) \tag{2}$$

Substituting $n-1$ for n in (2),

$$(n-1)f(n-1) = (n-1)^2 + 2(f(1) + f(2) + \cdots + f(n-2)) \tag{3}$$

Subtracting (3) from (2), we get $nf(n) - (n-1)f(n-1) = (2n-1) + 2f(n-1)$ or $f(n) = \frac{n+1}{n} f(n-1) + \frac{2n-1}{n}$. We prove by induction that $f(n) \leq 2n \ln n$.

1.0.1 Claim: $f(n) \leq 2n \ln n$

Proof: Induction basis: $n = 1$, the claim holds.

Let the claim hold for all values up to $n-1$. Then,

$$\begin{aligned} f(n) &= \tfrac{n+1}{n} f(n-1) + \tfrac{2n-1}{n} \\ &\leq \tfrac{n+1}{n} 2(n-1)\ln(n-1) + \tfrac{2n-1}{n} \text{ by induction hypothesis} \\ &= \tfrac{2(n^2-1)}{n} \ln(n-1) + \tfrac{2n-1}{n} \\ &= \tfrac{2(n^2-1)}{n} (\ln n + \ln(1 - \tfrac{1}{n})) + \tfrac{2n-1}{n} \end{aligned}$$

We make use of the standard inequality stated below.

Fact 1.1 *$1 + x \leq e^x$ for $x \in R$. For example, $1 + 0.2 \leq e^{0.2}$ and $1 - 0.2 \leq e^{-0.2}$.*

$$\begin{aligned} f(n) &\leq \tfrac{2(n^2-1)}{n} (\ln n - \tfrac{1}{n}) + \tfrac{2n-1}{n} \\ &= 2n \ln n - \tfrac{2}{n} \ln n - 2 + \tfrac{2}{n^2} + 2 - \tfrac{1}{n} \\ &\leq 2n \ln n, \text{ establishing the inductive step.} \end{aligned}$$

Hence, the expected running time of the randomized quick sort algorithm is $O(n \ln n)$. But one of the limitations of the recurrence relation approach is that we do not how the running time of the algorithm is spread around its expected value. Can this analysis be extended to answer questions such as, with what probability does the algorithm RandQuickSort needs more than $12n \ln n$ time steps? Later on, we apply a different technique and establish that this probability is very small. Similarly, for the case of the Maximum algorithm, by solving the recurrence relation for the expected running time, we will not be able to answer questions of the form, what is the probability that the run time is greater than $3n$?

Answers to such questions give how close the random variable is spread around its expected value or how varied the distribution is. To be able to answer such queries, we study Tail inequalities in the next section.

2 Tail Inequalities

In this section, we study three ways to estimate the tail probabilities of random variables. It will be noted that, the more information we know about the random variable the better the estimate we can derive about a given tail probability.

2.1 Markov Inequality

Theorem 2.1 Markov Inequality
If X is a non-negative valued random variable with an expectation of μ, then $Pr[X \geq c\mu] \leq \frac{1}{c}$.

Proof. By definition,

$$\begin{aligned} \mu &= \textstyle\sum_a a Pr[X = a] \\ &= \textstyle\sum_{a < c\mu} a Pr[X = a] + \sum_{a \geq c\mu} a Pr[X = a] \\ &\geq 0 + \textstyle\sum_{a \geq c\mu} c\mu Pr[X = a] \text{ as } X \text{ is non-negative valued} \\ &= c\mu \textstyle\sum_{a \geq c\mu} Pr[X = a] \\ &= c\mu Pr[X \geq c\mu] \end{aligned}$$

Hence, $Pr[X \geq c\mu] \leq \frac{\mu}{c\mu} = \frac{1}{c}$. □

The knowledge of the standard deviation of the random variable X would most often give a better bound.

2.2 Chebychev Inequality

We first define the terms standard deviation and variance of a random variable X.

Definition 2.2 *Let X be a random variable with an expectation of μ. The variance of X, denoted by var(X), is defined as var$(X) = E[(X - \mu)^2]$. The standard deviation of X, denoted by σ_X, is defined as $\sigma_X = \sqrt{var(X)}$.*

Note that by definition, var$(x) = E[(X - \mu)^2] = E[X^2 - 2X\mu + \mu^2] = E[X^2] - \mu^2$. The second equality follows from the linearity of expectations.

Theorem 2.3 Chebychev Inequality
Let X be a random variable with expectation μ_X and standard deviation σ_X. Then, $Pr[|X - \mu_X| \geq c\sigma_X] \leq \frac{1}{c^2}$.

Proof. Let random variable $Y = (X - \mu_X)^2$. Then, $E[Y] = E[(X - \mu_X)^2] = \sigma_X^2$ by definition and also Y is a non-negative valued random variable.

Now, $Pr[|X - \mu_X| \geq c\sigma_X] = Pr[(X - \mu_X)^2 \geq c^2\sigma_X^2] = Pr[Y \geq c^2\sigma_X^2]$.

Applying Markov Inequality to the random variable Y, $Pr[Y \geq c^2\sigma_X^2] = Pr[Y \geq c^2\mu_Y] \leq \frac{1}{c^2}$. Hence the theorem. □

2.3 Chernoff Bounds

The tail estimates given by Theorem 2.1, Theorem 2.3 work for random variables in general. But if the random variable X can be expressed as a sum of n independent random variables each of which is $0, 1-$ valued, then we can obtain very tight bounds on the tail estimates. This is expressed in the following theorem and the bounds are commonly called as Chernoff Bounds.

Theorem 2.4 Chernoff Bounds:
Let X be a random variable defined as $X = X_1 + X_2 + \cdots + X_n$ where each $X_i, 1 \leq i \leq n$, is a $0, 1-$ valued random variable and all X_i's are independent. Also, let $E[X] = \mu$ and $Pr[X_i = 1] = p_i, 1 \leq i \leq n$. Then for any $\delta > 0$, $Pr[X \geq \mu(1 + \delta)] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$.

Proof. The normal strategy employed to prove tail estimates of sums of independent random variables is to make use of exponential moments. While proving Chebychev inequality (Theorem 2.3), we made use of second-order moment. It can be observed that using higher order moments would generally improve the bound on the tail inequality. But using exponential moments would result in a vast improvement in the bound as we shall see later in some examples.

Observe that,

$$\mu = \sum_{i=1}^{n} p_i \tag{4}$$

by definition of X and using linearity of expectations. Let $Y_i = e^{tX_i}$ for some parameter t to be chosen later. Note that Y_i's are also independent as X_i's are. Define the random variable $Y = Y_1 Y_2 \cdots Y_n$. Then,

$$E[Y_i] = E[e^{tX_i}] = p_i e^t + (1 - p_i)e^0 = 1 - p_i + p_i e^t \tag{5}$$

$$E[Y] = E[Y_1 Y_2 \cdots Y_n] = \Pi_{i=1}^n E[Y_i] = \Pi_{i=1}^n (1 - p_i + p_i e^t) \tag{6}$$

where the second equality follows from independence of Y_i's and the last equality follows from (5). Now,

$$Pr[X \geq \mu(1 + \delta)] = Pr[Y \geq e^{t\mu(1+\delta)}] \leq \frac{E[Y]}{e^{t\mu(1+\delta)}} = \frac{\Pi_{i=1}^n (1 - p_i + p_i e^t)}{e^{t\mu(1+\delta)}} \tag{7}$$

by using Markov inequality from Theorem 2.1 and equation (6).

We now make use of the inequality from Fact 1.1 in equation (7) which then reduces to,

$$Pr[X \geq \mu(1 + \delta)] = \frac{\Pi_{i=1}^n e^{-p_i(1-e^t)}}{e^{t\mu(1+\delta)}} = \frac{e^{-\mu(1-e^t)}}{e^{t\mu(1+\delta)}} = e^{\mu[(e^t-1)-t(1+\delta)]} \tag{8}$$

where the second equality follows from equation (4). In (8), we can choose a value of t that minimizes the probability estimate. To find the minimum let $f(t) = \ln e^{\mu[(e^t-1)-t(1+\delta)]} = -\mu(1 - e^t) - t\mu(1 + \delta)$. Differentiating $f(t)$ with respect to t and equating it to zero gives us $\mu e^t - \mu(1 + \delta) = 0$ or $t = \ln(1 + \delta)$. Using this value of t in (8),

$$Pr[X \geq \mu(1 + \delta)] \leq \frac{e^{-\mu(1-(1+\delta))}}{(1 + \delta)^{\mu(1+\delta)}} = \frac{e^{\mu\delta}}{(1 + \delta)^{(1+\delta)}} = \left(\frac{e^\delta}{(1 + \delta)^{(1+\delta)}}\right)^\mu. \tag{9}$$

That completes the proof. □

But the form of the inequality in equation (9) is not very convenient to handle. In addition this form is hard to invert, i.e. given the probability bound, choose an appropriate δ. Instead, we use the following form most often.

Theorem 2.5 *Let X be defined as in Theorem 2.4. Then,*

$$Pr[X \geq (1+\delta)\mu] \leq \begin{cases} e^{-\mu\delta^2/4} & \text{if } \delta \leq 1 \\ e^{-\mu\delta \ln \delta} & \text{if } \delta > 1 \end{cases}$$

2.4 Application of Tail Inequalities

We consider the use of tail inequalities to two problems.

2.4.1 n Balls and n Bins

Consider throwing n balls independently and uniformly at random into n bins. We are interested in the probability that bin 1 contains more than 4 balls. Define n $0,1-$ valued random variables $X_i, 1 \leq i \leq n$, defined as $X_i = 1$ if ball i falls into bin 1 and 0 otherwise. By uniformity, $Pr[X_i = 1] = \frac{1}{n}$. Define the random variable $X = X_1 + X_2 + \cdots X_n$. Thus X denotes the number of balls that fall in bin 1. By linearity of expectations, $E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = n\frac{1}{n} = 1$.

Using Markov inequality from Theorem 2.1, we get

$$Pr[X \geq 4] \leq \frac{1}{4} \tag{10}$$

Before using Chebychev inequality (Theorem 2.3), we first compute the standard deviation of X as follows.

$$\text{var}(X_i) = E[X_i^2] - E[X_i]^2 = \frac{1}{n} - \frac{1}{n^2} \tag{11}$$

$$\text{var}(X) = n\text{var}(X_1) = n(\frac{1}{n} - \frac{1}{n^2}) = 1 - \frac{1}{n} \tag{12}$$

where the first equality in (12) follows from the independence of $X_i's$. Hence,

$$\sigma_X = \sqrt{1 - \frac{1}{n}} \tag{13}$$

Applying Chebychev inequality to the random variable X,

$$Pr[X \geq 4] = Pr[X - 1 \geq 3] \leq Pr[|X-1| \geq 3] \leq \frac{1}{\left(\frac{3}{\sqrt{1-\frac{1}{n}}}\right)^2} = \frac{1-\frac{1}{n}}{9} \leq \frac{1}{9} \tag{14}$$

Using Chernoff bounds from Theorem 2.4,

$$Pr[X \geq 4] = Pr[X \geq (1+3)1] \leq e^{-1 \cdot 3 \ln 3} = \frac{1}{27} \tag{15}$$

where in we used the simpler form of Chernoff bounds as stated in Theorem 2.5.

Now comparing equations (10), (14) and (15) we can see that using Chebychev inequality gives a better bound than that of Markov inequality and a much better bound is obtained using Chernoff bounds. As another example, let us consider the probability that bin 1 has more than $1 + 10\ln n$ balls. Using Markov inequality, we get $Pr[X \geq 1 + 10\ln n] \leq \frac{1}{1+10\ln n}$. Using Chebychev inequality we get that $Pr[X \geq 1 + 10\ln n] \leq Pr[|X-1| \geq 10\ln n] \leq \frac{1-\frac{1}{n}}{100\ln^2 n} \leq \frac{1}{100\ln^2 n}$. Whereas, using Chernoff bounds, we can get $Pr[X \geq 1 + 10\ln n] \leq e^{-10\ln n \ln(10\ln n)} = n^{-10\ln 10\ln n} \leq \frac{1}{n^{10}}$. With a tighter analysis using Chernoff bounds it can be shown that for the case of n balls and n bins, the probability that bin 1 has more than $\frac{c\ln n}{\ln\ln n}$ for some constant c is very small, say $\frac{1}{n^2}$.

2.4.2 $n \ln n$ balls and n bins

In this scenario we have $n \ln n$ balls and n bins where balls are thrown into the bins independently and uniformly at random. Define the random variables $X_i, 1 \leq i \leq n \ln n$ and X as above. We have $Pr[X_i = 1] = 1/n$ and $E[X] = \ln n$. Let us estimate the probability that bin 1 has more than $10 \ln n$ balls. Using Markov inequality, we get that $Pr[X \geq 10 \ln n] \leq 1/10$. Instead, using Chernoff bounds, we get $Pr[X \geq 10 \ln n] = Pr[X \geq (1+9) \ln n] \leq e^{-9 \ln n \ln 9} \leq 1/n^{20}$ which is exponentially small.

In general, it can be observed that if the expectation of the random variable is small then we pay a higher penalty to derive a w.h.p bound.

A Probability Theory

We start by defining probability and then introduce some well-known inequalities that we often use.

Let Ω be an arbitrary set, called the sample space. We start by defining a σ–field, also sometimes called a σ–algebra.

Definition A.1 (σ–field) *A collection $\mathcal{F}$ of subsets of Ω is called a σ–field if it satisfies:*

1. $\Omega \in \mathcal{F}$
2. *$A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$, and*
3. *For any countable sequence $A_1, A_2, \ldots$, if $A_1, A_2, \ldots \in \mathcal{F}$ then $A_1 \cup A_2 \cup \ldots \in \mathcal{F}$.*

Definition A.2 *A set function $\Pr$ on a σ–field $\mathcal{F}$ of subsets of Ω such that $\Pr : \mathcal{F} \to [0,1]$ is called a probability measure if it satisfies:*

1. $0 \leq \Pr(A) \leq 1, \ \forall A \in \mathcal{F}$.
2. $\Pr(\Omega) = 1$, *and*
3. *If $A_1, A_2, \ldots$ is a disjoint sequence of sets in $\mathcal{F}$ then*

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$$

The triad $(\Omega, \mathcal{F}, \Pr)$ is often called a probability space. For equivalent and alternate definitions, examples, and a more complete introduction, we refer the reader to standard textbooks on probability [**?**]. In the following, if no probability space is mentioned then any space $(\Omega, \mathcal{F}, \Pr)$ can be taken.

We often use the following inequality called "Boole's inequality" which is part of a general Boole-Bonferroni inequalities [**?**] and this is also sometimes referred to as the "union bound" as it provides a bound on the probability of a union of events. This inequality is also referred to as the (finite) sub-additivity property of the probability measure.

Proposition A.3 (Boole's inequality) *For any arbitrary events $A_1, A_2, \ldots A_n$,*

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \Pr(A_i)$$

The notion of independence is an important concept in the study of probability.

Definition A.4 (Independence) *A collection of events $\{A_i : i \in I\}$ is said to be* independent *if for all $S \subseteq I$, $\Pr(\cap_{i \in S} A_i) = \Pi_{i \in S} \Pr(A_i)$.*

We now define random variable, which is any *measurable* function from Ω to $\mathbb{R}$. Let $\mathcal{R}$ denote the standard Borel σ–field associated with $\mathbb{R}$, which is the σ–field generated by left-open intervals of $\mathbb{R}$ [**?**].

Definition A.5 (Random Variable) *Given a probability space $(\Omega, \mathcal{F}, \Pr)$, a mapping $X : \Omega \to \mathbb{R}$ is called a random variable if it satisfies the condition that $X^{-1}(R) \in \mathcal{F}$ for every $R \in \mathcal{R}$.*

We represent as $\{X \leq x\}$ as the set $\{\omega \in \Omega | X(\omega) \leq x\}$ for $x \in \mathbb{R}$ and also write $\Pr(X \leq x)$ as the probability of the above event. Similar definition can be made for representing the set $\{\omega \in \Omega | X(\omega = x\}$ as $\{X = x\}$.

The notion of independence also extends to random variables. Two random variables X and Y are said to be *independent* if the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for $x, y \in R$. The definition extends to multiple random variables just as in Definition A.4.

Associated with any random variable is a distribution function defined as follows.

Definition A.6 (Distribution function) *The distribution function $F : \mathbb{R} \rightarrow [0, 1]$ for a random variable X is defined as $F_X(x) = \Pr(X \leq x)$.*

A random variable X is said to be a *discrete* random variable if the range of X is a finite or countably infinite subset of $\mathbb{R}$. For discrete random variables, the following definition can be provided for the *density* of a random variable.

Definition A.7 (Density) *Given a random variable X, the density function $f_X : \mathbb{R} \rightarrow [0, 1]$ of X is defined as $f_X(x) = \Pr(X = x)$.*

The above definition can be extended to all types of random variables also with proper care. In the rest of this section, we focus on discrete random variables only and hence the definitions are made for the case of discrete random variables. With proper care, the definitions however can be extended [**?**].

An important quantity of interest of a random variable is its expectation.

Definition A.8 (Expectation) *Given a probability space $(\Omega, \mathcal{F}, \Pr)$ and a random variable X, the expectation of X, denoted $E[X]$, is defined as*

$$E[X] = \sum_{x \in \mathbb{R}} x \Pr[X = x]$$

with the convention that $0 \cdot \infty = \infty \cdot 0 = 0$.