COMPSCI 532: Design and Analysis of Algorithms	September 23, 2015
Lecture 8	
Lecturer: Debmalya Panigrahi	*Scribe: Allen Xiao*

1 Overview

We begin the discussion of randomized algorithms. Because the outcome of these algorithms depends on randomization, we analyze them in terms of their expected outcome and bound the probability they deviate far from this expectation. This lecture will focus on presenting tools for performing this analysis.

2 Probability Review

This section uses the presentation of probability spaces and random variables from the textbook by Mitzenmacher and Upfal [MU05]. The purpose of this section is to provide explicit definitions to the objects we use in this lecture. For that reason, there are several fundamental ideas from probability – like independence, Bayes' Rule, continuous random variables – that we will not define below.

Consider any random process (e.g. flipping a coin, rolling a dice). We model such a process using the notion of a *probability space*.

Definition 1. *A **probability space** consists of:*

1. *A **sample space** Ω of all possible outcomes of the random process.*
2. *Sets $\mathscr{F} \subseteq \Omega$ (possibly empty) called **events**. Singleton sets (i.e. individual elements of Ω) are called "simple" events.*
3. *A **probability function**, $\Pr : \mathscr{F} \to \mathbb{R}$, satisfying:*
   - *(i) For all events E, $0 \leq \Pr(E) \leq 1$.*
   - *(ii) $\Pr(\Omega) = 1$*
   - *(iii) For any finite or countably infinite sequence of mutually disjoint events $E_1, E_2, \ldots$*

$$\Pr\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} \Pr\left(E_i\right)$$

*If Ω is finite or countably infinite and $\mathscr{F}$ is the family of subsets of Ω, then we call this probability space **discrete**.*

Instead of dealing with events directly, we will often use *random variables* to give events some sort of value. For instance, when counting the number of heads in 10 coin flips, we could assign a value to each 10-flip outcome equal to the number of heads. This is a convenient representation for calculating the number of heads we "expect" to see from performing a 10-flip experiment.

Definition 2. *A random variable X is a function from events to $\mathbb{R}$.*

$$X : \mathscr{F} \to \mathbb{R}$$

A random variable is discrete if it only takes on a finite or countably infinite number of values. We will often argue on the probability that a random variable takes on a certain value:

$$\Pr(X = a) = \sum_{s \in \Omega : X(s) = a} \Pr(s)$$

Definition 3. *Let X be a discrete random variable. The **expectation** of X is defined as:*

$$\mathbb{E}[X] = \sum_{a \in \mathbb{R}} a \cdot \Pr(X = a)$$

3 Linearity of Expectation

Theorem 1 (Linearity of expectation)**.** *Let $X_1, X_2, \ldots$ be a finite or countably infinite set of random variables (not necessarily independent).*

$$\mathbb{E}\left[\sum_i X_i\right] = \sum_i \mathbb{E}[X_i]$$

Proof. This can be proved by inducting on the sum of two random variables. In other words, showing that:

$$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$$

One can prove this using the definition of expectation directly. □

Example 1. Suppose we have a coin which, when flipped, lands on heads with probability p (otherwise, it lands on tails). We flip this coin repeatedly until we get a heads. How many flips will we perform, in expectation? Consider any sequence of flips:

$$T \quad T \quad T \quad H \quad H \quad T \quad \cdots$$

What is the probability that it takes i trials to see the first head? For that occur, we must *fail* the first $(i-1)$ trials, and *succeed* exactly on the ith. Let X be a random variable equal to the first i with a heads. Since the trials are independent:

$$\Pr(X = i) = \Pr(\text{First } H \text{ is on trial } i) = (1-p)^{i-1}(p)$$

We call variables distributed in this way (independent trials until success) *geometric* random variables. We say that X is *drawn* from a geometric distribution with parameter p.

$$X \sim \text{Geo}(p) = \text{Geo}(1-p)$$

Using the definitions, one can show that the expectation of a geometric random variable is:

$$\mathbb{E}[X] = \frac{1}{p}$$

The expected number of flips before we see a heads is $1/p$.

Example 2. Suppose there are n different types of coupons, and each day we acquire a single coupon uniformly at random from the n types. The *coupon collector problem* asks: "How many days before we collect at least one of each type?"

We will count the time before seeing each new coupon type. We use random variables X_i:

$$X_i = \#\text{ days to see new type after seeing } i\text{th coupon type}$$

Adding up $\sum_{i=0}^{n-1} X_i$ gives us the total number of days before we see all n types.

The probability that it took us a days to see the $(i+1)$st from the ith is:

$$\Pr(X_i = a) = \left(\frac{i}{n}\right)^{a-1}\left(\frac{n-i}{n}\right)$$

It seems that X_i is drawn from geometric distribution with $p = (n-i)/n$.

$$X_i \sim \text{Geo}\left(\frac{n-i}{n}\right) = \text{Geo}\left(1-\frac{i}{n}\right)$$

When $n = 2$, this exactly the same as the coin tossing example (which fits our expectations). It follows that the expectation is:

$$\mathbb{E}[X_i] = \frac{n}{n-i}$$

Applying linearity of expectation gives us the expected number of days.

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=0}^{n-1} X_i\right] &= \sum_{i=0}^{n-1} \mathbb{E}[X_i] \\
&= \sum_{i=0}^{n-1} \frac{n}{n-i} \\
&= n + \frac{n}{2} + \frac{n}{3} + \ldots + \frac{n}{n} \\
&= nH_n \\
&= \Theta(n \log n)
\end{aligned}$$

Here, H_n is the nth harmonic number, and $H_n = \Theta(\log n)$. Intuitively, H_n is a discrete approximation for the logarithm integral:

$$\begin{aligned}
H_n &= \sum_{i=1}^{n} \frac{1}{i} \\
\ln n &= \int_1^n \frac{1}{x} dx
\end{aligned}$$

4 Union Bound and Tail Bounds

Theorem 2 (Union bound)**.** *Let $E_1, E_2, \ldots$ be a finite or countably infinite set of events, not necessarily disjoint.*

$$\Pr\left(\bigcup_{i\geq 1} E_i\right) \leq \sum_{i\geq 1} \Pr(E_i)$$

Proof. One can prove this inequality directly from the axioms of probability we introduced at the beginning. Additionally, equality holds when the events are independent. □

Example 3. Suppose we have n balls and m bins, and we toss the balls (uniformly at random) into the bins. This is called a *balls and bins* process, and is a popular model in practice. For example, hashing algorithms will often use a balls and bins process for analyzing collisions.

Let X_i be a random variable for the number of balls in the ith bin. We ask:

1. How large should n be before every bin is filled in expectation ($\mathbb{E}[X_i] \geq 1$)?
2. What is $\mathbb{E}[X_i]$?
3. When $m = n$, what is the maximum load over all the bins? If $X = \max_i X_i$, what is $\mathbb{E}[X]$?

(1) is actually a restatement of the coupon collector problem, so the answer is $n = \Theta(m \log m)$. For (2), notice that the expectation is symmetric across i.

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \cdots = \mathbb{E}[X_n]$$

Now, since we throw n balls,

$$\sum_{i=1}^{m} \mathbb{E}[X_i] = n$$

It follows that:

$$\mathbb{E}[X_i] = \frac{n}{m}$$

Finally, for (3), the probability a fixed bin has more than k balls is at least the probability that, for k of the n balls, they all landed in this bin. These trials are independent, so we can say:

$$\Pr(X_i \geq k) \leq \binom{n}{k} \left(\frac{1}{n}\right)^k$$

We can apply an inequality known as *Stirling's approximation* to give a tight bound on $\binom{n}{k}$.

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$$

This gives us:

$$\Pr(X_i \geq k) \leq \left(\frac{en}{k}\right)^k \left(\frac{1}{n}\right)^k = \left(\frac{e}{k}\right)^k$$

The definition of X gives us:

$$\Pr(X \geq k) = \Pr(\exists i \mid X_i \geq k) = \Pr\left(\bigcup_{i=1}^{n} (X_i \geq k)\right)$$

Applying the union bound:

$$\begin{aligned} \Pr(X \geq k) &\leq \sum_{i=1}^{n} \Pr(X_i \geq k) \\ &\leq \sum_{i=1}^{n} \left(\frac{e}{k}\right)^k \\ &= n\left(\frac{e}{k}\right)^k \end{aligned}$$

We know that $X \leq n$. We can divide the outcomes of X into cases where no $X_i \geq k$, and when at least one $X_i \geq k$. Taking the expectation of X:

$$\begin{aligned}\mathbb{E}[X] &\leq k\Pr(X < k) + n\Pr(X \geq k) \\ &\leq n \cdot n\left(\frac{e}{k}\right)^k + k\end{aligned}$$

We will choose k in order to remove the first half of the sum. Asymptotically, this will be the same as if we had tried to optimize for k.

$$\begin{aligned} & n^2\left(\frac{e}{k}\right)^k \leq 1 \\ \implies \quad & \left(\frac{e}{k}\right)^k \leq \frac{1}{n^2} \\ \implies \quad & \left(\frac{k^k}{e^k}\right) \geq n^2 \\ \implies \quad & k^k \sim \text{poly}(n) \\ \implies \quad & k\log k = O(\log n)\end{aligned}$$

How can we solve for k now? It turns out one natural solution to this expression is $k = O(\log n/\log\log n)$.

$$\begin{aligned}k\log k &= \frac{\log n}{\log\log n} \cdot \log\left(\frac{\log n}{\log\log n}\right) \\ &= \frac{\log n}{\log\log n} \cdot (\log\log n - \log\log\log n)\end{aligned}$$

The rightmost term is vanishingly small, so:

$$\begin{aligned}k\log k &= \Theta\left(\frac{\log n}{\log\log n} \cdot \log\log n\right) \\ &= \Theta(\log n)\end{aligned}$$

Similar to how $\log n$ solves $2^k = \text{poly}(n)$, $k^k = \text{poly}(n)$ is solved by $k = O(\log n/\log\log n)$. Using that choice of k:

$$\Pr\left(X_i \geq \frac{e\log n}{\log\log n}\right) \leq \left(\frac{\log\log n}{\log n}\right)^{\frac{e\log n}{\log\log n}} \leq \frac{1}{n^2}$$

This makes $\mathbb{E}[X]$:

$$\mathbb{E}[X] = n \cdot \frac{n}{n^2} + k = O\left(\frac{\log n}{\log\log n}\right)$$

In fact, this is tight; there is a lower bound example which matches this. Additionally, notice for X:

$$\Pr\left(X \geq \frac{e\log n}{\log\log n}\right) \leq \frac{1}{n}$$

When some event fails with probability $\Omega(1/n)$, we typically say it succeeds *with high probability*. Here, we say that with high probability $X = O((\log n)/(\log\log n))$. These types of statements are *tail bounds*, and provide a formal notion of controlling a distribution. Here, we used a tail bound to obtain a bound on the expectation. Often, we want to do the reverse: use expectation to obtain a tail bound ("value stays close to the mean with high probability").

4.1 Markov's Inequality

Theorem 3 (Markov's Inequality)**.** *Let X be a random variable such that $X \geq 0$. For $k > 0$:*

$$\Pr(X \geq k) \leq \frac{\mathbb{E}[X]}{k}$$

Proof. By nonnegativity of X:

$$\begin{aligned} \mathbb{E}[X] &\geq k \cdot \Pr(X \geq k) + 0 \cdot \Pr(X < k) \\ &= k \cdot \Pr(X \geq k) \end{aligned}$$

Afterwards, just rearrange. □

Markov's Inequality is one such tail bound constructed from expectation. However, it is often too weak on its own, and the nonnegativity constraint on X may be too restrictive. If we had applied Markov's in the balls and bins example, we would have gotten the bound:

$$\Pr(X_i \geq k) = \frac{1}{k}$$

This is exponentially weaker than the bound we proved. Markov turns out to be useful as a starting point for proving stronger inequalities.

Corollary 4. *Let X be a random variable, and $f(\cdot)$ be a function where $f(X) \geq 0$. For $k > 0$:*

$$\Pr(f(X) \geq k) \leq \frac{\mathbb{E}[f(X)]}{k}$$

Note that there is no restriction on the sign of X, only on the sign of $f(X)$. This follows directly from application of Markov's Inequality on the image of X in $f(\cdot)$.

4.2 Chebyshev's Inequality

Theorem 5 (Chebyshev's Inequality)**.** *Let X be a random variable, with mean $\mu = \mathbb{E}[X]$ and standard deviation $\sigma = (\mathbb{E}[X^2] - (\mathbb{E}[X])^2)^{1/2}$ For $k > 0$:*

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Note that X is not constrained in sign.

Proof. Chebyshev's inequality follows from a straightforward application of Corollary 4 using:

$$f(X) = (X - \mu)^2$$

As a quadratic function, $f(X) \geq 0$. The event in Chebyshev is equivalent to a similar event with this $f(X)$.

$$|X - \mu| \geq k\sigma \iff (X - \mu)^2 \geq k^2\sigma^2$$

It follows that their probabilities are equal.

$$\Pr(|X - \mu| \geq k\sigma) = \Pr((X - \mu)^2 \geq k^2\sigma^2)$$

Applying the corollary to $f(X)$ gives:

$$\begin{aligned}\Pr(|X-\mu| \geq k\sigma) &= \Pr\left((X-\mu)^2 \geq k^2\sigma^2\right) \\ &\leq \frac{\mathbb{E}\left[(X-\mu)^2\right]}{k^2\sigma^2} \\ &= \frac{\sigma^2}{k^2\sigma^2} \\ &= \frac{1}{k^2}\end{aligned}$$

□

4.3 Chernoff-Hoeffding Bounds

One way of interpreting the Chebyshev Inequality is that $\Pr(X=a)$ *drops quadratically* for each standard deviation we move from the mean (Markov's was linear). Both these bounds assumed nothing on the dependence of the random variables. As it turns out, this quadratic tail bound (i.e. Chebyshev's) is the best one can do without controlling the dependence of the random variables. The next class of bounds, due to Hoeffding and later Chernoff, use independence of random variables to create an *exponential* tail bound. Again, we use the corollary of Markov's Inequality and carefully choose $f(X)$.

Theorem 6 (Chernoff-Hoeffding Bounds)**.** *For independent $X_1, \ldots, X_n$ where:*

$$X_i = \begin{cases} 1 & \text{w.p. } p_i \\ 0 & \text{w.p. } (1-p_i) \end{cases}$$

Let $X = \sum_i X_i$, with mean $\mu = \mathbb{E}[X] = \sum_i p_i$. Then for $\varepsilon > 0$:

$$\Pr(X > (1+\varepsilon)\mu) < \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{(1+\varepsilon)}}\right)^{\mu}$$

$$\Pr(X < (1-\varepsilon)\mu) < \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{(1-\varepsilon)}}\right)^{\mu}$$

Proof. We will only prove the positive direction; the negative one is symmetric. Just like for Chebyshev, we apply Corollary 4. For all $t > 0$, let:

$$f(X) = e^{tX}$$

Again, $f(X) \geq 0$. $f(X)$ is also monotonically increasing, so:

$$\begin{aligned}\Pr(X > (1+\varepsilon)\mu) &= \Pr\left(e^{tX} > \exp(t(1+\varepsilon)\mu)\right) \\ &\leq \frac{\mathbb{E}\left[e^{tX}\right]}{\exp(t(1+\varepsilon)\mu)}\end{aligned}$$

We can use the independence of X_i to dissect the numerator. The expectation of a product of independent random variables is the product of their expectations.

$$\begin{aligned}
\mathbb{E}\left[e^{tX}\right] &= \prod_{i=1}^{n} \mathbb{E}\left[e^{tX_i}\right] \\
&= \prod_{i=1}^{n} \left(e^0(1-p_i) + e^t(p_i)\right) \\
&= \prod_{i=1}^{n} \left(1 + p_i\left(e^t - 1\right)\right) \\
&\leq \prod_{i=1}^{n} \exp\left(p_i\left(e^t - 1\right)\right) \\
&= \exp\left(\sum_{i=1}^{n} p_i\left(e^t - 1\right)\right) \\
&= \exp\left(\left(e^t - 1\right)\mu\right)
\end{aligned}$$

We want to choose t which minimizes the probability, which is:

$$\Pr(X > (1+\varepsilon)\mu) \leq \frac{\exp\left(\left(e^t - 1\right)\mu\right)}{\exp\left(t(1+\varepsilon)\mu\right)}$$

To do this, we can try to minimize the following function (minimize numerator, maximize denominator), after stripping the exponentiations and multiplicative constants.

$$g(t) = \left(e^t - 1\right) - \left(t(1+\varepsilon)\right)$$

Taking the derivative:

$$\frac{d}{dt}g(t) = \frac{d}{dt}\left(e^t - 1\right) - \frac{d}{dt}\left(t(1+\varepsilon)\right) = 0$$

Simplifying:

$$\begin{aligned}
0 &= e^t - (1+\varepsilon) \\
t &= \ln(1+\varepsilon)
\end{aligned}$$

Let $t = \log(1+\varepsilon)$, then the original expression gives us:

$$\begin{aligned}
\Pr\left(X > (1+\varepsilon)\mu\right) &\leq \frac{\exp\left(\left(e^t - 1\right)\mu\right)}{\exp\left(t(1+\varepsilon)\mu\right)} \\
&= \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{(1+\varepsilon)}}\right)^{\mu}
\end{aligned}$$

□

Additionally, we have a simpler form for when ε is small.

Corollary 7. *For $0 < \varepsilon < 1$, the Chernoff bound can be manipulated into a simpler form:*

$$\Pr(X > (1+\varepsilon)\mu) \leq \exp\left(-\varepsilon^2\mu/3\right)$$

$$\Pr(X < (1-\varepsilon)\mu) \leq \exp\left(-\varepsilon^2\mu/2\right)$$

Combining these, we can also write:

$$\Pr(|X-\mu| > \varepsilon\mu) \leq \exp\left(-\varepsilon^2\mu/3\right)$$

Proof. We will only give the proof for the first version; the second is similar. Rewriting the original statement of the bound:

$$\left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{(1+\varepsilon)}}\right)^{\mu} = \exp(\mu(\varepsilon - (1+\varepsilon)\ln(1+\varepsilon)))$$

We will focus on the inner term involving ε:

$$\varepsilon - (1+\varepsilon)\ln(1+\varepsilon)$$

We expand $\ln(1+\varepsilon)$ using the Taylor series.

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

Plugging into the expression:

$$\begin{aligned}
\varepsilon - (1+\varepsilon)\ln(1+\varepsilon) &= \varepsilon - (1+\varepsilon)\left(\varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3} - \cdots\right) \\
&= \varepsilon - \left(\varepsilon + \varepsilon^2\left(1-\frac{1}{2}\right) - \varepsilon^3\left(\frac{1}{2}-\frac{1}{3}\right) + \varepsilon^4\left(\frac{1}{3}-\frac{1}{4}\right) + \cdots\right) \\
&= -\varepsilon^2\left(1-\frac{1}{2}\right) + \varepsilon^3\left(\frac{1}{2}-\frac{1}{3}\right) - \varepsilon^4\left(\frac{1}{3}+\frac{1}{4}\right) + \cdots \\
&\leq -\varepsilon^2\left(1-\frac{1}{2}\right) + \varepsilon^3\left(\frac{1}{2}-\frac{1}{3}\right) \\
&= -\frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{6} \\
&\leq -\frac{\varepsilon^2}{3}
\end{aligned}$$

Because $\varepsilon < 1$, each successive term is smaller and the sum after the ε^3 term is negative. Therefore, we can truncate after the ε^3 term to get an upper bound. Returning to the original probability, we get:

$$\Pr(X > (1+\varepsilon)\mu) \leq \exp\left(-\varepsilon^2\mu/3\right)$$

□

It is often the case that this holds (we are within a multiplicative factor of 1 of the mean). For the rest of the class, we will tend to use the form in Corollary 7. Both the *law of large numbers* and the *Gaussian distribution* use similar proofs that have these exponential bounds in $(-\varepsilon^2\mu)$.

5 Summary

In this lecture, we gave a brief overview of probability spaces and random variables, and some useful bounds for reasoning about stochastic processes. These bounds were the union bound and tail bounds. The last two tail bounds we derived from careful applications of Markov's Inequality.

References

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.