A Theorist's Toolkit **(CMU 18-859T, Fall 2013)**

Lecture 03: Chernoff/Tail Bounds

September 16, 2013

Lecturer: Ryan O'Donnell *Scribe: Elara Willett*

1 INTRODUCTION

Define $H \sim \text{Binomial}(n, \frac{1}{2})$, i.e. H is the number of heads that occur when we flip a coin n times. Recall that the mean of H is $\mu = \frac{n}{2}$ and the standard deviation is $\sigma = \frac{\sqrt{n}}{2}$. Last time, we applied the Berry-Esseen Theorem to get

$$Pr\left[H \geq \frac{n}{2} + t\frac{\sqrt{n}}{2}\right] \approx \overline{\Phi}(t) \leq \frac{\phi(t)}{t} \leq e^{\frac{-t^2}{2}}$$

where $\overline{\Phi}(t)$ is the probability that the Gaussian distribution is at most t and $t \geq 1$. This tells us that the probability that we exceed the standard deviation decreases very rapidly.

As an example, consider $t = 10\sqrt{\ln n}$. Then we get $Pr\left[H \geq \frac{n}{2} + t\frac{\sqrt{n}}{2}\right] \leq e^{-50 \ln n} = \frac{1}{n^{50}}$. However, the error term of our above approximation is $\pm O(\frac{1}{\sqrt{n}})$, which is important in this example because the error is so much larger than the bound. We will see in the analysis to come that the bound does, indeed, hold in this example.

In this lecture, we will cover bounds on random variables given varying assumptions. For a comprehensive reference on probabilistic techniques used in approximation algorithms, including topics covered in this lecture and more, see Dubhashi and Panconesi's book [DP$^+$]. For another reference, see Mitzenmacher and Upfal's book [MU05], which is an introduction to probabilistic techniques used in algorithm analysis, including the topics covered in this lecture.

2 BOUNDS BASED ONLY ON MEAN & VARIANCE

Say we have a random variable X. Generally speaking, the more we know about X, the better bounds we can calculate for $P[X \geq t]$. So we'll begin by supposing we know only the expectation $E[X]$. Assume $X \geq 0$ always and X is not always 0.

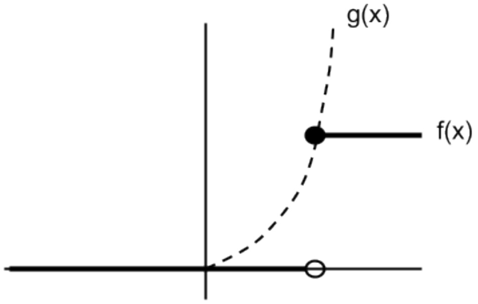
Theorem 2.1. *Markov Inequality. For all $t > 0$, $P[X \geq t \cdot E[X]] \leq \frac{1}{t}$.*

Proof. WOLOG assume $E[X] = 1$. (We can multiply X by a constant and we do not change the probabilities.) For contradiction, suppose $P[X \geq t] > \frac{1}{t}$. Then

$$E[X] \geq t \cdot P[X \geq t] + 0 \cdot P[X < t] > 1$$

which contradicts our assumption that $E[X] = 1$. □

Here is an alternative proof, using the intuition from the depicted graph.



Proof. Define $f(x)$ to be 0 for $x < t$ and 1 for $x \geq t$, so that $Pr[X \geq t] = E[f(X)]$. Define $g(x) = \frac{x}{t}$. Since $g(x) \geq f(x)$, we get

$$Pr[X \geq t] = E[f(X)] \leq E[g(X)] = E\left[\frac{X}{t}\right] = \frac{1}{t}$$

where the last equality holds because $E[X] = 1$. □

We can prove another similar fact based only on expectation using a similar argument, known as a 'Markov-type' or 'average-type' argument.

Fact 2.2. *Suppose* $E[X] = \epsilon$ *and* $0 \leq X \leq 1$, *then* $Pr[X \geq \frac{\epsilon}{2}] \geq \frac{\epsilon}{2}$.

Proof. For contradiction, suppose $Pr[X \geq \frac{\epsilon}{2}] < \frac{\epsilon}{2}$, then

$$E[X] \leq 1 \cdot Pr\left[X \geq \frac{\epsilon}{2}\right] + \frac{\epsilon}{2} \cdot Pr\left[X < \frac{\epsilon}{2}\right] < 1 \cdot \frac{\epsilon}{2} + \frac{\epsilon}{2} \cdot 1 = \epsilon$$

which contradicts our assumption that $E[X] = \epsilon$. □

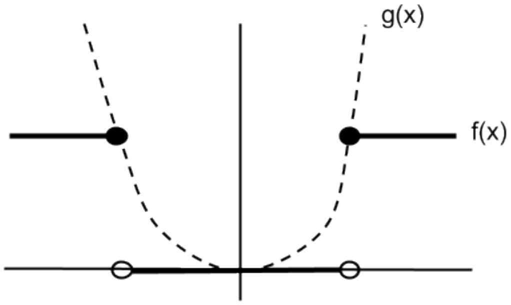
Now, we can move on to the scenerio where we know the expectation and the variance of X. In this case we get the following bounds.

Theorem 2.3. *Chebyshev's Inequality. Let* $E[X] = \mu$ *and* $\text{stddev}[X] = \sigma \neq 0$, *then for all* $t > 0$, $P[|X - \mu| \geq t\sigma] \leq \frac{1}{t^2}$.

Proof. WOLOG assume $\mu = 0$. (We can subtract μ from the mean without changing the standard deviation.) Also, WOLOG assume $\sigma = 1$. (We can multiply X by $\frac{1}{\sigma}$ without changing the probabilites.) Note that these assumptions imply $E[X^2] = 1$. Now, it is sufficient to show $P[|X| \geq t] \leq \frac{1}{t^2}$. Well, the event $|X| \geq t$ is equivalent to the event $X^2 \geq t^2$. Since X^2 is nonnegative and $t^2 = t^2 \cdot E[X^2]$, we can apply Markov's Inequality to bound $P[X^2 \geq t^2]$, which gives us the desired bound

$$P[|X| \geq t] = P[X^2 \geq t^2] \leq \frac{1}{t^2}$$ □

Of course, as in the proof of Markov's Inequality, we can alternatively prove this theorem with a picture.



Proof. Again WOLOG assume $\mu = 0$ and $\sigma = 1$. Now define $f(x)$ to be 0 for $x < t$ and 1 for $x \geq t$ and define $g(x) = \frac{x}{t^2}$. As before, since $g(x) \leq f(x)$,

$$Pr[|X| \geq t] = E[f(X)] \leq E[g(X)] = E\left[\frac{X^2}{t^2}\right] = \frac{1}{t^2} \qquad \square$$

3 BACK TO SUMS OF VARIABLES

Let $X = X_1 + X_2 + \cdots + X_n$, and assume (for now) that the X_i's are independent. Given this additional information about X, what bounds can we prove? In this section we'll explore Chernoff Bounds to answer this question.

3.1 Markov and Chebyshev on Sums

To start, let's again consider $H \sim \text{Binomial}(n, \frac{1}{2})$. Recall that we are interested in bounding $Pr[H \geq \frac{n}{2} + t\frac{\sqrt{n}}{2}]$, where $t = 10\sqrt{\ln n}$. We know that $\mu = \frac{n}{2}$ and $\sigma = \frac{\sqrt{n}}{2}$, so we can apply the inequalities form the previous section. Using only the mean, we can apply Markov's Inequality and get

$$Pr\left[H \geq \mu\left(1 + O\left(\frac{\sqrt{\ln n}}{\sqrt{n}}\right)\right)\right] \leq \frac{1}{1 + O(\frac{\sqrt{\ln n}}{\sqrt{n}})} = \frac{1}{1 + \widetilde{O}(\frac{1}{\sqrt{n}})}$$

Using the mean and the standard deviation, we can apply Chebyshev' Inequality and get

$$Pr[|X - \mu| \geq 10\sqrt{\ln n} \cdot \sigma] \leq \frac{1}{(10\sqrt{\ln n})^2} = \frac{1}{100 \ln n}$$

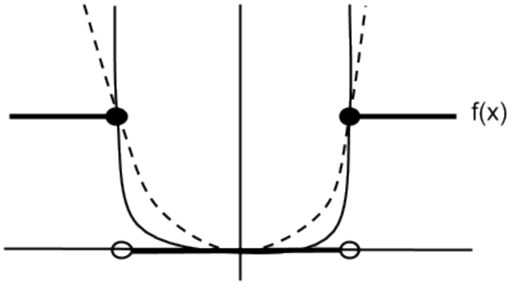
Chebyshev's bound will still hold if we only have pairwise independence of the W_i's. Pairwise independence means that for all pairs $i \neq j$, $E[X_i X_j] = E[X_i]E[X_j]$. The mean will remain the same as in the independence case because of linearity of independence. To see that the variance will also remain unchanged, notice that pairwise independence gives us

$$E\left[(\sum_{i=1}^{n} X_i)^2\right] = \sum_{i=1}^{n} E[X_i^2] + 2\sum_{i \neq j} E[X_i X_j] = \sum_{i=1}^{n} E[X_i^2] + 2\sum_{i \neq j} E[X_i]E[X_j]$$

By definition, $Var[X_1+X_2+\cdots+X_n] = E[(\sum_{i=1}^{n} X_i)^2] - (E[\sum_{i=1}^{n} X_i])^2$, and by our calculation above this value is the same as in the case of independent variables, so variance is unchanged. Since the mean and variance are the same in the cases of independence and pairwise independence, Chebychev's bound is the same for both. Therefore, we see that Chebychev's bound is strong in some cases and weak in others.

3.2 Fourth Moment Method

So far we have seen bounds based on the mean, e.g. the first moment, and the variance, e.g. the second moment, so now we will look at bounds based on higher moments. The graph below shows how a fourth-degree polynomial could more closely fit our function, f, than a quadratic polynomial. An odd-degree polynomial could not upper bound f, so we do not consider the third moment.



Let $X = X_1 + X_2 + \cdots + X_n$, where the X_i's are independent identical random variables and

$$X_i = \begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$$

We will use the Fourth Moment Method to upper bound $Pr[|X| \geq 10\sqrt{\ln n} \cdot \sqrt{n}]$.

Well, $Pr[|X| \geq t\sqrt{n}] = P[X^4 \geq t^4 n^2]$, and X^4 is a nonnegative variable, so we can apply Markov's Inequality, namely,

$$Pr[|X| \geq t\sqrt{n}] = P[X^4 \geq t^4 n^2] \leq \frac{E[X^4]}{t^4 n^2}$$

Now, let's take a close look at $X^4 = \sum_i X_i^4$. When we multiple the variables out, we get

$$X^4 = \sum_i X_i^4 + c\sum_{i\neq j} X_i^2 X_j^2 + c' \sum_{i\neq j} X_i^3 X_j + c'' \sum_{i\neq j\neq k} X_i^2 X_j X_k + c''' \sum_{i\neq j\neq k\neq l} X_i X_j X_k X_l$$

for some integers c, c', c'', c'''. Since $E[X_i] = 0$ and the variables are independent,

$$E[X_i X_j X_k X_l] = E[X_i]E[X_j]E[X_k]E[X_l] = 0 \cdot 0 \cdot 0 \cdot 0 = 0$$

Similarly, $E[X_i^3 X_j] = E[X_i^3]E[X_j] = E[X_i^3] \cdot 0 = 0$, and $E[X_i^2 X_j X_k] = 0$. Thus,

$$E[X^4] = \sum_i E[X_i^4] + c\sum_{i\neq j} E[X_i^2]E[X_j^2]$$

and if we have to calculate it, $c = \binom{n}{2}\binom{4}{2} = 3n^2 - 3n$. Thus,

$$Pr[|X| \geq t\sqrt{n}] \leq \frac{E[X^4]}{tn^2} = \frac{3n^2 - 3n}{t^4 n^2} \leq \frac{3}{t^4}$$

Furthermore, in our specific example,

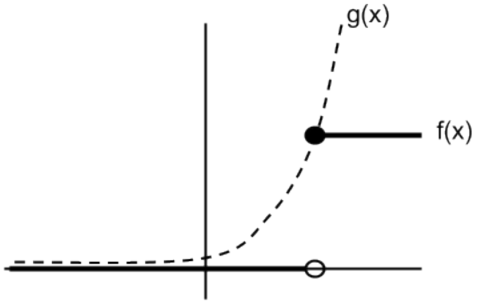
$$Pr[|X| \geq 10\sqrt{\ln n} \cdot \sqrt{n}] \leq \frac{3}{(10\sqrt{\ln n})^4} = \Theta\left(\frac{1}{\ln^2 n}\right)$$

We could continue in the same manner, applying the $2s$ Moment Method to get bounds of the form $Pr[|X| \geq t\sqrt{n}] \leq \frac{C_s}{t^{2s}}$ for some constant t. Notice that as s goes to infinity, c_s also goes to infinity. Thus, to see which of these methods gives the best bound, we would need to optimize over all possible s, geting $s(t)$ for a given t. This would be quite difficult, but it might give us a slightly better bound.

3.3 Chernoff Method

For this subsection, we assume X is the sum of independent identical variables.

Define $c = e^\lambda$ for some $\lambda > 0$. Let $g(x) = c^x/c^u$ for a given $u = t\sqrt{n}$. Looking at the diagram below, it is clear that g is strictly increasing and $g(u) = 1$.



This tells us that the event that $g \geq 1$ is equivalent to the event that $X \geq u$, which implies $P[X \geq u] = Pr[g \geq 1] = Pr[e^{\lambda X} \geq e^{\lambda u}]$. Since $e^{\lambda X}$ is a nonegative random variable, we can now apply Markov's Inequality to get,

$$P[X \geq u] = Pr[e^{\lambda X} \geq e^{\lambda u}] \leq \frac{E[e^{\lambda X}]}{e^{\lambda u}}$$

This bound is a function of λ and is called the moment generating function of X. By optimizing over $\lambda > 0$, we could determine the best bound achievable by this function for any given u.

In order to use this strategy, we first need to get a function for $E[e^{\lambda X}]$. Well,

$$E[\exp(\lambda X)] = E[\exp(\lambda X_1 + \lambda X_2 + \cdots + \lambda X_n)] = E[\exp(\lambda X_1)\exp(\lambda X_2)\cdots\exp(\lambda X_n)]$$

By independence of the X_i's, we get

$$E[\exp(\lambda X_1)\exp(\lambda X_2)\cdots\exp(\lambda X_n)] = E[\exp(\lambda X_1)] \cdot E[\exp(\lambda X_2)] \cdots E[\exp(\lambda X_n)]$$

and since the X_i's and identical, we get

$$E[\exp(\lambda X_1)] \cdot E[\exp(\lambda X_2)] \cdots E[\exp(\lambda X_n)] = (E[\exp(\lambda X_1)])^n$$

Therefore, $E[e^{\lambda X}] = (E[\exp(\lambda X_1)])^n$. Now, $E[\exp(\lambda X_1)] = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda}$, which is known as the hyperbolic cosine. We can simplify this function by applying the Taylor Expansion for e^x. Namely, $e^{\lambda} = (1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3!} + \cdots)$ and $e^{-\lambda} = (1 - \lambda + \frac{\lambda^2}{2} - \frac{\lambda^3}{3!} + \cdots)$. The odd terms cancel each other out, so we get $E[\exp(\lambda X_1)] = (1 + \frac{\lambda^2}{2} + \frac{\lambda^4}{4!} + \frac{\lambda^6}{6!} + \cdots)$. By Claim 3.1, this series is upper bounded by the Taylor Expansion of e^x for $x = \frac{\lambda^2}{2}$. Thus, we get

$$P[X \geq u] \leq \frac{E[e^{\lambda X}]}{e^{\lambda u}} \leq \frac{e^{\lambda^2 n/2}}{e^{\lambda u}} = e^{\lambda^2 n/2 - \lambda u}$$

Now, we have an upper bound for $P[X \geq u]$ that we could optimize over λ for a given u. For example, if we choose $\lambda = \frac{u}{n}$ then we get the bound:

$$P[X \geq u] \leq e^{-\frac{u^2}{2n}}$$

Furthermore, for $u = 10\sqrt{\ln n} \cdot \sqrt{n}$, we get $P[X \geq u] \leq e^{-50 \ln n} = \frac{1}{n^{50}}$. Recall from the Introduction that this is precisely the bound we desired. For this particular random variable and choice of t, we can deduce from the Berry-Esseen Theorem that this is about the best bound we can get. In more general cases, this bound is still about the best we can do. See [AS04] or [Slu77] for more details.

Claim 3.1. *Let* $x = 1 + \frac{\lambda^2}{2} + \frac{\lambda^4}{4!} + \frac{\lambda^6}{6!} + \cdots$ *and* $y = 1 + (\frac{\lambda^2}{2}) + \frac{1}{2}(\frac{\lambda^2}{2})^2 + \frac{1}{3!}(\frac{\lambda^2}{2})^2 + \cdots$*. then* $x \leq y$*.*

Proof. Well, $y = \sum_{i=0}^{\infty} \frac{(\lambda^2/2)^i}{i!} = \sum_{i=0}^{\infty} \frac{(\lambda)^{2i}}{2^i i!} \geq \sum_{i=0}^{\infty} \frac{(\lambda)^{2i}}{(2i)!}$=x, where the inequality follows from

$$\frac{(2i)!}{i!} \geq \frac{2i(2i-1)\cdots(i+1)}{i(i-1)\cdots 1} = \left(\frac{2i}{i}\right)\left(\frac{2i-1}{i-1}\right)\cdots\left(\frac{i+1}{1}\right) \geq 2^i$$

□

Note: Suppose our X_i's were not identical, but had some common structure, such as

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

As long as the X_i's are independent, we could do a similar, yet somewhat more complicated, analysis and get a good bound.

We will now formally state two variations on the bound we proved, the Hoeffding Bound and the Chernoff Bound. These bounds assume $X = X_1 + X_2 + \cdots + X_n$ with independent X_i's and $\mu = E[X] = \sum_i E[X]$. These bounds are extremely useful and **the Chernoff Bound should be memorized.**

Theorem 3.2. *Hoeffding Bound. Suppose $a_i \leq X_i \leq b_i$ always. Then for all $t > 0$,*

$$Pr[X \geq \mu + t] \leq \exp\left(-\frac{2t^2}{\sum_i (b_i - a_i)^2}\right)$$

and the exact same bound holds for $Pr[X \geq \mu - t]$.

The Chernoff Bound, below, is a little stronger than the Hoeffding Bound, which is also sometimes called the Chernoff Bound.

Theorem 3.3. *Chernoff Bound. Suppose $0 \leq X_i \leq 1$. Then for all $\epsilon > 0$,*

$$Pr[X \leq (1-\epsilon)\mu] \leq \exp\left(-\frac{\epsilon^2}{2}\mu\right), \textit{ and } Pr[X \geq (1+\epsilon)\mu] \leq \exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu\right)$$

The Chernoff Bound is even useful in the case where you only know some range within which μ falls. Suppose $\mu_L \leq \mu \leq \mu_H$, then as you might expect,

$$Pr[X \leq (1-\epsilon)\mu_L] \leq \exp\left(-\frac{\epsilon^2}{2}\mu_L\right), \text{ and } Pr[X \geq (1+\epsilon)\mu_H] \leq \exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu_H\right)$$

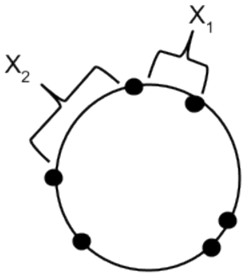
4 RELAXING INDEPENDENCE

So far we have mostly considered only sums of independent variables. In this section we will consider two special cases where we can prove good bounds about the sums of dependent variables: negatively associated random variables and martingales.

4.1 Negatively Associated Random Variables

For intuition's sake, negative association (NA) means that if one variable in a set of negatively associated random variables is large than the others will be small. The following are examples of negatively associated variables.

1. Uniformly at random, draw n points on the unit circle as depicted, and let $X_1, X_2, ...X_n$ be the arc lengths between these points. Notice that if you know that one arc length is large, then the others must be small.



2. Independently throw n balls into n bins, and let X_i be the number of balls in bin i. If there are a lot of balls in bin i, then there are few balls in the other bins.

3. Let $x_1, x_2, ..., x_n$ be any real numbers. Let $X_1, ..., X_n$ be a random permutation of these numbers. If we know $X_i = x_j$ for a high x_j, than the other variables cannot take on x_j, so they will probably be comparatively low. Note, this setup can be used to model sampling without replacement.

We will now formally define negative association.

Definition 4.1. Random variables $X_1, X_2, ..., X_n$ are randomly associated if

$$E[f(X_i : i \in A)g(X_j : j \in B)] \leq E[f(X_i : i \in A)]E[g(X_j : j \in B)]$$

for any disjoint sets $A, B \subseteq \{1, 2, ...n\}$ and any nondecreasing functions, f, g.

This definition implies that $E[X_i X_j] \leq E[X_i]E[X_j]$, or $\text{cov}[X_i, X_j] \leq 0$, but the converse is not true. In general, Negative association is stronger than this claim about covariance. See [JDP83] for more details.

By induction, one can show that negative association is enough to give,

$$E[\exp(\lambda X_1)\exp(\lambda X_2) \cdots \exp(\lambda X_n)] \leq E[\exp(\lambda X_1)] \cdot E[\exp(\lambda X_2)] \cdots E[\exp(\lambda X_n)]$$

In our discussion of the Chernoff Method, we only needed independence to get this one fact. Therefore, the Chernoff Bounds hold for negatively associated variables.

Here are a few interesting facts to ponder about negative association:

- Independence implies negative association.
- Negative association is closed under subsets and independent unions.
- Negative association is closed under sapplying nondecreasing functions to subsets of variables.

4.2 Martingales

Suppose you bought a lottery ticket with numbers 12 33 5 60. Before the winning numbers are revealed, you have an expected payoff of about 0, but after the first number is revealed to be 12, your expected payoff is much higher. Your expected payoff becomes further refined as more of the numbers are revealed. Finally, the last number is revealed and you know your payoff exactly. We could model your series of guesses as a martingale.

Definition 4.2. Let $X_1, X_2, ..., X_n$ be discrete random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$. Define $Y_i = E[f(X_1, ..., X_n)|X_1, ..., X_i]$. Then $Y_0, ..., Y_n$ is a **martingale** with respect to $X_1, ..., X_n$.

So we have $Y_0 = E[f(X_1, ..., X_n)] = \mu$, $Y_1 = E[f(X_1, ..., X_n)|X_1]$,..., $Y_n = f(X_1, ..., X_n)$. Given a property of the Y_i's, i.e. bounded diffference, we can get a bound on f.

Theorem 4.3. *Azuma's Inequality or Method of Bounded Difference. Let $X_1, ..., X_n$ be independent. If f satisfies*

$$f(X_1, ..., X_n) - f(X_1, ..., X_{i-1}, X_i', X_{i+1}, ..., X_n) \leq c_i$$

then

$$Pr[f(X_1, ..., X_n) \geq \mu + t] \leq \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right)$$

and the same upper bound holds for $Pr[f(X_1, ..., X_n) \geq \mu - t]$.

For more information on martingales see [McD89].

References

[AS04] N. Alon and J.H. Spencer. *The Probabilistic Method*, chapter Appendix A.2. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2004.

[DP$^+$] D Dubhashi, Alessandro Panconesi, et al. Concentration of measure for the analysis of randomised algorithms.

[JDP83] Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295, 1983.

[McD89] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[MU05] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[Slu77] Eric V Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, 5(3):404–412, 1977.