CPSC 536N: Randomized Algorithms 2011-12 Term 2

Lecture 2

Prof. Nick Harvey *University of British Columbia*

1 Concentration Inequalities

One of the most important tasks in analyzing randomized algorithms is to understand what random variables arise and how well they are **concentrated**. A variable with good concentration is one that is close to its mean with good probability. A "concentration inequality" is a theorem proving that a random variable has good concentration. Such theorems are also known as "tail bounds".

1.1 Markov's inequality

This is the simplest concentration inequality. The downside is that it only gives very weak bounds, but the upside is that needs almost no assumptions about the random variable. It is often useful in scenarios where not much concentration is needed, or where the random variable is too complicated to be analyzed by more powerful inequalities.

Theorem 1 *Let Y be a random variable that assumes only nonnegative values. Then, for all $a > 0$,*

$$\Pr[Y \geq a] \;\leq\; \frac{\mathrm{E}[Y]}{a}.$$

Note that if $a \leq \mathrm{E}[Y]$ then the right-hand side of the inequality is at least 1, and so the statement is trivial. So Markov's inequality is only useful if $a > \mathrm{E}[Y]$. Typically we use Markov's inequality to prove that Y has only a constant probability of exceeding its mean by a constant factor, e.g., $\Pr[Y \geq 2 \cdot \mathrm{E}[Y]] \leq 1/2$.

Proof: Let X be the indicator random variable that is 1 if $Y \geq a$. Since Y is non-negative, we have

$$X \;\leq\; Y/a. \tag{1}$$

Then

$$\Pr[Y \geq a] \;=\; \Pr[X \geq 1] \;=\; \mathrm{E}[X] \;\leq\; \mathrm{E}[Y/a] \;=\; \frac{\mathrm{E}[Y]}{a},$$

where the inequality comes from taking the expectation of both sides of (1). □

Note that Markov's inequality only bounds the **right tail** of Y, i.e., the probability that Y is much greater than its mean.

1.2 The Reverse Markov inequality

In some scenarios, we would also like to bound the probability that Y is much smaller than its mean. Markov's inequality can be used for this purpose if we know an upper-bound on Y. The following result is an immediate corollary of Theorem 1.

Corollary 2 *Let Y be a random variable that is never larger than B. Then, for all $a < B$,*

$$\Pr[Y \leq a] \;\leq\; \frac{\mathrm{E}[B-Y]}{B-a}.$$

1.3 Application to Max Cut

Recall the Max Cut problem from the last lecture. We are given a graph $G = (V, E)$ and wish to approximately solve $\max\{|\delta(U)| \,:\, U \subseteq V\}$. Recall our algorithm simply chooses U to be a uniformly random subset of V.

Let's now analyze the probability that this algorithm gives a large cut. Let $Y = |\delta(U)|$ and $B = |E|$, and note that Y is never larger than B. Fix any $\epsilon \in [0, 1/2]$ and set $a = (\frac{1}{2} - \epsilon)|E|$. By the Reverse Markov inequality,

$$\begin{aligned}
\Pr[|\delta(U)| \leq (1/2-\epsilon)|E|] \;&\leq\; \frac{\mathrm{E}[|E| - |\delta(U)|]}{|E| - (1/2-\epsilon)|E|} \\
&=\; \frac{|E| - \mathrm{E}[|\delta(U)|]}{(1/2+\epsilon)|E|} \qquad \text{(linearity of expectation)} \\
&=\; \frac{1}{1+2\epsilon} \qquad \text{(since } \mathrm{E}[|\delta(U)|] = |E|/2\text{)} \\
&\leq\; 1-\epsilon,
\end{aligned}$$

where we have used Inequality 2 from the Notes on Convexity Inequalities.

This shows that, with probability at least ϵ, the algorithm outputs a set U satisfying

$$|\delta(U)| \;>\; (1/2-\epsilon)\, OPT.$$

This statement is quite unsatisfying. If we want a 0.499 approximation, then we have only shown that the algorithm has probability 0.001 of succeeding. Next we will show how to increase the probability of success.

2 Amplification by Independent Trials

In many cases where the probability of success is positive but small, we can "amplify" that probability by perfoming several independent trials and taking the best outcome.

For Max Cut, consider the following algorithm. First it picks several sets $U_1, \ldots, U_k \subseteq V$, independently and uniformly at random. Let j be the index for which $|\delta(U_j)|$ is largest. The algorithm simply outputs the set U_j. (Note that our Max Cut algorithm in the last lecture did not even look at the edges of the graph, but this new algorithm must look at the edges to compute $|\delta(U_j)|$.)

To analyze this algorithm, we wish to argue that $|\delta(U_j)|$ is large. Well, $|\delta(U_j)|$ is small only if *all* $|\delta(U_i)|$ are small, and this is rather unlikely.

$$\begin{aligned}
\Pr[\text{every } |\delta(U_i)| \leq (1/2-\epsilon)|E|] \;&=\; \prod_{i=1}^{k} \Pr[|\delta(U_i)| \leq (1/2-\epsilon)|E|] \qquad \text{(by independence)} \\
&\leq\; (1-\epsilon)^k \qquad \text{(by our analysis above)} \\
&\leq\; e^{-\epsilon k}.
\end{aligned}$$

Here we have used the standard trick $1 + x \leq e^x$, which is Inequality 1 from the Notes on Convexity Inequalities.

Thus, setting $k = \log(1/\delta)/\epsilon$, we obtain that

$$\Pr[\text{every } |\delta(U_i)| \leq (1/2 - \epsilon)|E|] \;\leq\; \delta.$$

In summary, our new Max Cut algorithm picks $k = \log(1/\delta)/\epsilon$ sets $U_1, \ldots, U_k$, finds the j which maximizes $|\delta(U_j)|$, then outputs U_j. With probability at least $1 - \delta$ we have

$$\Pr[|\delta(U_j)| \geq (1/2 - \epsilon)|E|].$$

In this analysis we used the following general fact, which is elementary but often very useful.

Claim 3 *Consider a random trial in which the probability of success is p. If we perform k trials, then*

$$\Pr[\textit{all failures}] \;=\; (1-p)^k \;\leq\; e^{-pk}$$

and therefore

$$\Pr[\textit{at least one success}] \;=\; 1 - (1-p)^k \;\geq\; 1 - e^{-pk}.$$

2.1 Example: Flipping Coins

Consider flipping a fair coin. If the coin turns up heads, call this a "success". So the probability of success is $p = 1/2$. By Claim 3,

$$\Pr[\text{at least one heads after } k \text{ trials}] \;=\; 1 - 1/2^k.$$

This is a true statement, but not all that interesting: it is quite obvious that after k coin flips, the probability of seeing at least one head is very close to 1. Note that the expected number of heads is $k/2$. Can't we say that there are probably close to $k/2$ heads? This takes us beyond the subject of amplification and into a discussion of the binomial distribution.

3 The binomial distribution

Again, let us flip a fair coin k times and let X be the number of heads seen. Then X has the **binomial distribution**. So

$$\Pr[\text{exactly } i \text{ heads}] \;=\; \binom{k}{i} 2^{-k}.$$

How can we show that X is probably close to its expected value (which is $k/2$). Well, the probability of X being "small" is:

$$\Pr[\text{at most } i \text{ heads}] \;=\; \sum_{0 \leq j \leq i} \binom{k}{j} 2^{-k}. \tag{2}$$

Here the meaning of "small" depends on the choice of i. For what values of i is this sum small? Unfortunately the sum is a bit too complicated to get a feel for its magnitude. Can we simplify the expression so it's easier to see what's going on?

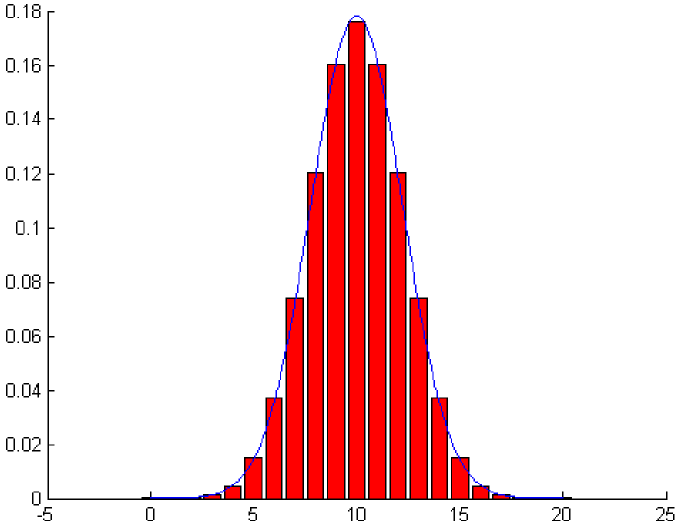
As far as I know, this sum does *not* have a simple closed-form expression. Instead, can we usefully approximate that sum? It would be great to have a provable upper bound for the sum that is simple, useful and asymptotically nearly tight. The **Chernoff bound**, which we discuss in Section 4, is a very general and powerful tool for analyzing tails of probability distributions, and it gives a fantastic bound on (2). But first, to motivate the bound that we will obtain, let us describe some interesting behavior of binomial distributions.

3.1 Different behavior in different regimes

In most introductory courses on probability theory, one often encounters statements that describe the limiting behavior of the binomial distribution. For example:

Fact 4 *Let X be a binomially distributed random variable with parameters n and p. (That is, X gives the number of successes in n independent trials where each trial succeeds with probability p.) As n gets large while p remains fixed, the distribution of X is well approximated by the normal distribution with mean np and variance $np(1-p)$.*

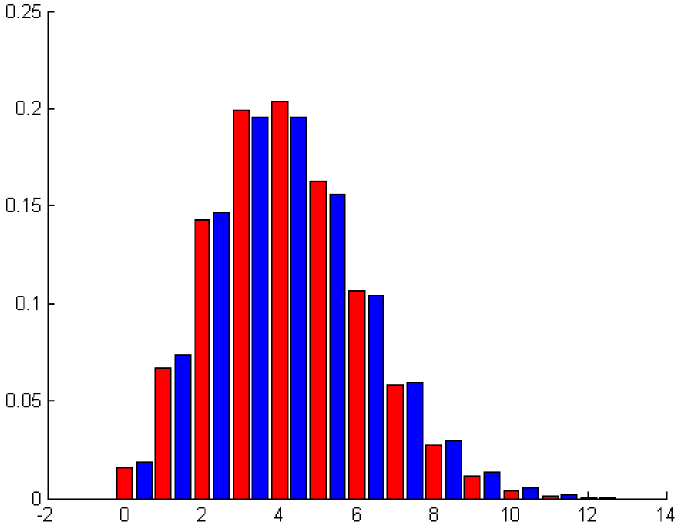
The following plot shows the binomial distribution for $n = 20$ and $p = 1/2$ in red and the corresponding normal approximation in blue.



However the binomial distribution has qualitatively different behavior when p is very small.

Fact 5 *Let X be a binomially distributed random variable with parameters n and p. As n gets large while np remains fixed, the distribution of X is well approximated by the Poisson distribution with parameter $\lambda = np$.*

The following plot shows the binomial distribution for $n = 50$ and $p = 4/n$ in red and the corresponding Poisson approximation in blue. Note that the red plot is quite asymmetric, so we would not expect a normal distribution (which is symmetric) to approximate it well.



Ideally would like an upper bound on (2) which works well for all ranges of parameters, and captures the phenomena described in Fact 4 and Fact 5. Remarkably, the Chernoff bound is able to capture both of these phenomena.

4 The Chernoff Bound

The Chernoff bound is used to bound the tails of the distribution for a sum of independent random variables, under a few mild assumptions. Since binomial random variables are sums of independent Bernoulli random variables, it can be used to bound (2). Not only is the Chernoff bound itself very useful, but its proof techniques are very general and can be applied in a wide variety of settings.

The Chernoff bound is by far the most useful tool in randomized algorithms. Numerous papers on randomized algorithms have only three ingredients: Chernoff bounds, union bounds and cleverness. Of course, there is an art in being clever and finding the right way to assemble these ingredients!

4.1 Formal Statement

Let $X_1, \ldots, X_m$ be independent random variables such that X_i always lies in the interval $[0,1]$. Define $X = \sum_i X_i$ and $\mu = \mathrm{E}[X]$. Let $p_i = \mathrm{E}[X_i]$ and note that $\mu = \sum_i p_i$.

Theorem 6 (Chernoff Upper Tail) *For any $\delta > 0$,*

$$\Pr[X \geq (1+\delta)\mu] \;\leq\; \exp\Big(-\mu\big((1+\delta)\ln(1+\delta) - \delta\big)\Big). \tag{3}$$

For any $\delta \in [0,1]$,

$$\Pr[X \geq (1+\delta)\mu] \;\leq\; \exp(-\mu\delta^2/3). \tag{4}$$

For any $\delta \geq 1$,

$$\Pr[X \geq (1+\delta)\mu] \;\leq\; \exp(-\mu\delta/3). \tag{5}$$

Remarks. All of these bounds are useful in different scenarios. The statement in (3) is the most general bound and it implies the statements in (4) and (5). The difference between (4) and (5) is due to the different phenomena illustrated in Facts 4 and 5.

Useful fact. This theorem does not actually require that $\mu = \mathrm{E}[X]$. It suffices to have $\mu \geq \mathrm{E}[X]$.

Theorem 7 (Chernoff Lower Tail) *For any $\delta \in [0,1]$,*

$$\Pr[X \leq (1-\delta)\mu] \;\leq\; \exp(-\mu\delta^2/3).$$

Useful fact. This theorem does not actually require that $\mu = \mathrm{E}[X]$. It suffices to have $\mu \leq \mathrm{E}[X]$.

Historical remark. Chernoff actually only considered the case in which the X_i random variables are identically distributed. The more general form stated here is due to Hoeffding.

4.2 Proof

The Chernoff bounds would not be true without the assumption that the X_is are independent. What special properties do independent random variables have? One basic property is that

$$\mathrm{E}[XY] \;=\; \mathrm{E}[X] \cdot \mathrm{E}[Y] \tag{6}$$

for any independent random variables X and Y.

But the Chernoff bound has nothing to do with *products* of random variables, it is about *sums* of random variables. So one trick we could try is to convert sums into products using the exponential function. Fix some parameter $t > 0$ whose value we will choose later. Define

$$\begin{aligned} Y_i &= \exp(tX_i) \\ Y &= \exp(tX) = \exp(t\sum_i X_i) = \prod_i \exp(tX_i) = \prod_i Y_i. \end{aligned}$$

It is easy to check that, since the X_is are independent, the Y_is are also independent. Therefore, by (6),

$$\mathrm{E}[Y] \;=\; \prod_i \mathrm{E}[Y_i]. \tag{7}$$

This is quite useful because we can now analyze $\mathrm{E}[Y]$ by separately analyzing the $\mathrm{E}[Y_i]$ terms. Furthermore, Y_i is closely related to X_i. Is there also a relationship between $\mathrm{E}[Y_i]$ and $\mathrm{E}[X_i]$?

Claim 8 $\mathrm{E}[Y_i] \leq \exp((e^t - 1)p_i)$.

Proof: We would like to relate $\mathrm{E}[e^{tX_i}]$ and $\mathrm{E}[X_i]$. If e^{tx} were a linear function of x we could just use linearity of expectation, but unfortunately it is not. However, since we know that $X_i \in [0,1]$, we can instead use a linear approximation of e^{tx} on that interval.

By convexity (see Inequality 3 in the Notes on Convexity Inequalities), we have

$$e^{tx} \;\leq\; 1 + (e^t - 1)x.$$

Thus,

$$\mathrm{E}[e^{tX_i}] \leq \mathrm{E}[1 + (e^t - 1)X_i] = 1 + (e^t - 1)p_i \leq \exp((e^t - 1)p_i),$$

where again we have used our favorite inequality $1 + x \leq e^x$. □

Now we are ready to prove the Chernoff bound. In fact, we will only prove (3); all the other inequalities follow using the same ideas and additional calculations.

$$\begin{aligned} \Pr[X \geq (1+\delta)\mu] &= \Pr\big[\exp(tX) \geq \exp\big(t(1+\delta)\mu\big)\big] \quad \text{(by monotonicity)} \\ &\leq \frac{\mathrm{E}[\exp(tX)]}{\exp(t(1+\delta)\mu)} \quad \text{(by Markov's inequality)} \\ &\leq \frac{\prod_i \exp((e^t-1)p_i)}{\exp(t(1+\delta)\mu)}, \end{aligned}$$

by (7) and Claim 8. Gathering everything inside one exponential we get

$$\Pr[X \geq (1+\delta)\mu] \;\leq\; \exp\big((e^t-1)\sum_i p_i - t(1+\delta)\mu\big).$$

Finally, substituting $t = \ln(1+\delta)$ and using $\mu = \sum_i p_i$ proves the desired inequality.