Lecture 3: Large deviations bounds and applications

Lecturer: *Pravesh Kothari* Scribe:*Pravesh*

(Slightly modified version of last year's lecture notes scribed by Sanjeev.)

Today's topic is deviation bounds: what is the probability that a random variable deviates from its mean by a lot? Recall that a random variable X is a mapping from a probability space to $\mathbf{R}$. The *expectation* or *mean* is denoted $\mathbf{E}[X]$ (usually denoted by μ), the *variance* of X (usually denoted by σ^2) is defined as $\mathsf{Var}(X) = \mathbf{E}\,(X - \mathbf{E}[X])^2$ and the standard deviation of X is $\sqrt{\mathsf{Var} X}$.

In many settings we have a set of n random variables $X_1, X_2, X_3, \ldots, X_n$ defined on the same probability space. To give an example, the probability space could be that of all possible outcomes of n tosses of a fair coin, and X_i is the random variable that is 1 if the ith toss is a head, and is 0 otherwise, which means $E[X_i] = 1/2$. Recall the basic fact of **Linearity of Expectation** from the first lecture, viz.

$$\mathbf{E}[\sum_i X_i] = \sum_i \mathbf{E}[X_i]$$

It is important to realize that linearity holds *regardless* of the whether or not the random variables are independent. This lecture is about understanding how much a random variable X deviates from its mean. An important property of random variables that comes into play in this context is *independence*.

Can we say something about $\mathbf{E}[X_1 X_2]$? In general, nothing much but if X_1, X_2 are independent events (formally, this is the property that for all a, b $\mathbf{Pr}[X_1 = a, X_2 = b] = \mathbf{Pr}[X_1 = a]\,\mathbf{Pr}[X_2 = b]$) then $\mathbf{E}[X_1 X_2] = \mathbf{E}[X_1]\,\mathbf{E}[X_2]$.

Note that if the X_i's are pairwise independent (i.e., each pair are mutually independent) then this means that $\mathrm{var}[\sum_i X_i] = \sum_i \mathrm{var}[X_i]$.

1 Three progressively stronger tail bounds

Now we give three methods that give progressively stronger bounds.

1.1 Markov's Inequality (aka averaging)

The first of a number of inequalities presented today, **Markov's inequality** says that any *non-negative* random variable X satisfies

$$\mathbf{Pr}\,(X \geq k\,\mathbf{E}[X]) \leq \frac{1}{k}.$$

Note that this is just another way to write the trivial observation that $\mathbf{E}[X] \geq k \cdot \mathbf{Pr}[X \geq k]$.

Can we give any meaningful upperbound on $\mathbf{Pr}[X < c \cdot \mathbf{E}[X]]$ where $c < 1$, in other words the probability that X is a lot less than its expectation? In general we cannot.

However, if we know an upperbound on X then we can. For example, if $X \in [0, 1]$ and $\mathbf{E}[X] = \mu$ then for any $c < 1$ we have (simple exercise)

$$\mathbf{Pr}[X \leq c\mu] \leq \frac{1-\mu}{1-c\mu}.$$

Sometimes this is also called an averaging argument.

Example 1 Suppose you took a lot of exams, each scored from 1 to 100. If your average score was 90 then in at least half the exams you scored at least 80.

1.2 Chebyshev's Inequality

The *variance of a random variable* X is one measure (there are others too) of how "spread out"it is around its mean. It is defined as $E[(x-\mu)^2] = E[X^2] - \mu^2$.

A more powerful inequality, **Chebyshev's inequality**, says

$$\mathbf{Pr}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2},$$

where μ and σ^2 are the mean and variance of X. Recall that $\sigma^2 = \mathbf{E}[(X-\mu)^2] = \mathbf{E}[X^2] - \mu^2$. Actually, Chebyshev's inequality is just a special case of Markov's inequality: by definition,

$$\mathbf{E}\left[|X-\mu|^2\right] = \sigma^2,$$

and so,

$$\mathbf{Pr}\left[|X-\mu|^2 \geq k^2\sigma^2\right] \leq \frac{1}{k^2}.$$

Here is simple fact that's used a lot: *If $Y_1, Y_2, \ldots, Y_t$ are iid (which is jargon for independent and identically distributed) then the variance of their average $\frac{1}{k}\sum_i Y_t$ is exactly $1/t$ times the variance of one of them.* Using Chebyshev's inequality, this already implies that the average of iid variables converges sort-of strongly to the mean.

1.2.1 Example: Load balancing

Suppose we toss m balls into n bins. You can think of m jobs being randomly assigned to n processors. Let X = number of balls assigned to the first bin. Then $\mathbf{E}[X] = m/n$. What is the chance that $X > 2m/n$? Markov's inequality says this is less than $1/2$.

To use Chebyshev we need to compute the variance of X. For this let Y_i be the indicator random variable that is 1 iff the ith ball falls in the first bin. Then $X = \sum_i Y_i$. Hence

$$\mathbf{E}[X^2] = \mathbf{E}[\sum_i Y_i^2 + 2\sum_{i<j} Y_iY_j] = \sum_i \mathbf{E}[Y_i^2] + \sum_{i<j} \mathbf{E}[Y_iY_j].$$

Now for independent random variables $\mathbf{E}[Y_iY_j] = \mathbf{E}[Y_i]\,\mathbf{E}[Y_j]$ so $\mathbf{E}[X^2] = \frac{m}{n} + \frac{m(m-1)}{n^2}$. Hence the variance is very close to m/n, and thus Chebyshev implies that the probability that $\Pr[X > 2\frac{m}{n}] < \frac{n}{m}$. When $m > 3n$, say, this is stronger than Markov.

1.3 Large deviation bounds

When we toss a coin many times, the expected number of heads is half the number of tosses. How tightly is this distribution concentrated? Should we be very surprised if after 1000 tosses we have 625 heads?

Intuitively, such an event appears unlikely - we think that independent coin tosses cannot conspire to create a significant bias towards either heads or tails. Large deviation inequalities make this intuition precise. Specifically, we will see such inequalities for sums of independent random variables.

A general rule of thumb to employ in understanding sums of independent random variables is via the *central limit theorem* that says that the sum of n independent random variables (with bounded mean and variance) converges to the famous Gaussian distribution (popularly known as the *Bell Curve*). This is very useful in algorithm design: we maneuver to design algorithms so that the analysis boils down to estimating the sum of independent (or somewhat independent) random variables.

To do a back of the envelope calculation, if all n coin tosses are fair (Heads has probability 1/2) then the Gaussian approximation implies that the probability of seeing N heads where $|N - n/2| > a\sqrt{n}$ is at most $e^{-a^2/2}$. The chance of seeing at least 625 heads in 1000 tosses of an unbiased coin is less than 5.3×10^{-7}. These are pretty strong bounds!

This kind of back-of-the-envelope calculations will get most of the credit in homeworks.

Of course, for finite n the sum of n random variables need not be an exact Gaussian and that's where Chernoff bounds come in - these help us bound the probability of a random variable X that is a sum of independent random variables falling in the "tail" of its distribution - this is the region that can be roughly defined as a constant number of standard deviations far from the mean. (By the way these bounds are also known by other names in different fields since they have been independently discovered.)

Our proof of the Chernoff bound will be based on a simple application of the "moment method". At a high level, the idea is to relate the probability that the deviations $|X - \mu|$ takes a value that's (say) too large to the probability that $(X-\mu)^{2k}$ for some large enough k takes too large a value. One can upper bound this probability by an application of Markov's inequality to the non-negative random variable $(X - \mu)^{2k}$ - a degree $2k$ polynomial of the random variable X. Chebyshev's inequality is an example of this method with $k = 1$. Working with $(X - \mu)^{2k}$ allows us to use the independence structure in X - in particular, if we know that X is a sum of $2k$-wise independent random variables, then one can usually obtain a good upper bound on $(X - \mu)^{2k}$. In the proof of Chernoff bound, we will deal with X that is a sum of mutually independent random variables and to make full use of this assumption, we will use the "ultimate polynomial" - the exponential function.

First we give an inequality that works for general variables that are real-valued in $[-1, 1]$. (To apply it to more general bounded variables just scale them to $[-1, 1]$ first.)

Theorem 1 (Quantitative version of CLT due to H. Chernoff)
If $X_1, X_2, \ldots, X_n$ are independent random variables and each $X_i \in [-1, 1]$. Let $\mu_i = E[X_i]$ and $\sigma_i^2 = \mathrm{var}[X_i]$. Then $X = \sum_i X_i$ satisfies

$$\Pr[|X - \mu| > k\sigma] \leq 2\exp(-\frac{k^2}{4n}),$$

where $\mu = \sum_i \mu_i$ and $\sigma^2 = \sum_i \sigma_i^2$. Also, $k \leq 2\sigma$.

Instead of proving the above we prove a simpler theorem for binary valued variables which showcases the basic idea.

THEOREM 2

Let $X_1, X_2, \ldots, X_n$ be independent 0/1-valued random variables and let $p_i = \mathbf{E}[X_i]$, where $0 < p_i < 1$. Then the sum $X = \sum_{i=1}^n X_i$, which has mean $\mu = \sum_{i=1}^n p_i$, satisfies

$$\mathbf{Pr}[X \geq (1+\delta)\mu] \leq (c_\delta)^\mu$$

where c_δ is shorthand for $\left[\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right]$.

Remark: There is an analogous inequality that bounds the probability of deviation *below* the mean, whereby δ becomes negative and the $\geq$ in the probability becomes $\leq$ and the c_δ is very similar.

PROOF: Surprisingly, this inequality also is proved using the Markov inequality, albeit applied to a different random variable.

We introduce a positive dummy variable t and observe that

$$\mathbf{E}[\exp(tX)] = \mathbf{E}[\exp(t\sum_i X_i)] = \mathbf{E}[\prod_i \exp(tX_i)] = \prod_i \mathbf{E}[\exp(tX_i)], \tag{1}$$

where the last equality holds because the X_i r.v.s are independent. Now,

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

therefore,

$$\begin{aligned}\prod_i \mathbf{E}[\exp(tX_i)] = \prod_i [1 + p_i(e^t - 1)] &\leq \prod_i \exp(p_i(e^t-1)) \\ &= \exp(\sum_i p_i(e^t-1)) = \exp(\mu(e^t-1)),\end{aligned} \tag{2}$$

as $1 + x \leq e^x$. Finally, apply Markov's inequality to the random variable $\exp(tX)$, viz.

$$\mathbf{Pr}[X \geq (1+\delta)\mu] = \mathbf{Pr}[\exp(tX) \geq \exp(t(1+\delta)\mu)] \leq \frac{\mathbf{E}[\exp(tX)]}{\exp(t(1+\delta)\mu)} = \frac{\exp((e^t-1)\mu)}{\exp(t(1+\delta)\mu)},$$

using lines (1) and (2) and the fact that t is positive. Since t is a dummy variable, we can choose any positive value we like for it. The right hand size is minimized if $t = \ln(1+\delta)$—just differentiate—and this leads to the theorem statement. □

2 Application 1: Sampling/Polling

Opinion polls and statistical sampling rely on tail bounds. Suppose there are n arbitrary numbers in $[0,1]$ If we pick t of them randomly (with replacement!) then the sample mean is within $(1 \pm \epsilon])$ of the true mean with probability at least $1 - \delta$ if $t > \Omega(\frac{1}{\epsilon^2} \log 1/\delta)$. (Verify this calculation!)

In general, Chernoff bounds implies that taking k independent estimates and taking their mean ensures that the value is highly concentrated about their mean; large deviations happen with exponentially small probability.

3 Balls and Bins revisited: Load balancing

Suppose we toss m balls into n bins. You can think of m jobs being randomly assigned to n processors. Then the expected number of balls in each bin is m/n. When $m = n$ this expectation is 1 but we saw in Lecture 1 that the most overloaded bin has $\Omega(\log n/\log\log n)$ balls. However, if $m = cn\log n$ then the expected number of balls in each bin is $c\log n$. Thus Chernoff bounds imply that the chance of seeing less than $0.5c\log n$ or more than $1.5c\log n$ is less than $\gamma^{c\log n}$ for some constant γ (which depends on the $0.5, 1.5$ etc.) which can be made less than say $1/n^2$ by choosing c to be a large constant.

Moral: if an office boss is trying to allocate work fairly, he/she should first create more work and then do a random assignment.

4 What about the median?

Given n numbers in $[0, 1]$ can we approximate the median via sampling? This will be part of your homework.
Exercise: Show that it is impossible to estimate the *value* of the median within say 1.1 factor with $o(n)$ samples.

But what is possible is to produce a number that is an approximate median: it is greater than at least $n/2 - n/t$ numbers below it and less than at least $n/2 - n/t$ numbers. The idea is to take a random sample of a certain size and take the median of that sample. (Hint: Use balls and bins.)