Lecture 3

1 Balls in bins.

Recall, the process of throwing n balls into n bins uniformly at random. You should have proved that with high probability (i.e., probability at least $1 - 1/n^c$) that the maximum load is $O(\log n/ \log \log n)$. (Indeed, we showed an even better bound of $O(\log \log n)$ for the process of "Pick the best of two bins".)

Just to review: You did this by bounding the number of ways that the maximum load could be higher than k on one bin, showing this is very small compared to the sample space when $k >> \log n$, and then using the union bound to show that the probability that the load on *any* bin exceeds k is small.

Recall that a setting where this is useful, is one of assigning jobs to servers. The result states that one can assign jobs at random and the max load is at most $\theta(\log n)$ in the case that n jobs are assigned to n servers. What is the "efficiency" of this approach?

Well, say each server could do a job in one time unit. With this approach, indeed you showed that it takes $O(\log n/ \log \log n)$ time to complete the jobs. Indeed, with this approach it takes $\Omega(\log n/ \log \log n)$ time to complete the jobs as is the result of the following exercise.

Question 1: Show that with constant probability that some bin has load $\Omega(\log n/ \log \log n)$.

Perhaps, one can get better efficiency m balls into n bins, when $m >> n$. Well, we can use the previous analysis to show that the max load is $\Theta((m/n) \log n/ \log \log n)$ just by dividing the balls into m/n groups of n balls.

Can we do better? Today, we discuss other techniques to do this. In many case, one could derive results either from counting, but sometimes it is easier with the theorems that we will describe.

Again, we will start with basic deviation bounds.

2 Deviation bounds.

Deviation bounds are bounds on the probability for a random variable being very different from its mean. In the balls in bins example, the mean is m/n, and we wish to bound the deviation from this mean.

The most basic deviation bound is the following.

Theorem 1
[Markov's inequality.] For a positive random variable X,

$$Pr[X > cE[X]] \leq 1/c.$$

Proof:

$$\begin{aligned} E[X] &= \sum_a aPr[X = a] \\ &= \sum_{a \le cE[X]} aPr[X = a] + \sum_{a > cE[x]} aPr[X = a] \\ &\ge \sum_{a > cE[X]} aPr[X = a] \\ &\ge cE[X] \sum_{a > cE[X]} Pr[X = 1] = cE[X]Pr[X > cE[X]] \end{aligned}$$

The first line is by definition of expectation. The next several are algebra. The last equality is the definition of probability of the event "X ¿ c E[X]".

We now have

$$E[X] \ge cE[X]Pr[X > cE[X]].$$

Dividing both sides by $cE[X]$ yields the theorerm. (Where do we use the fact that X is positive?)

□

Exercise: do not turn in. Show that for a random variable Y whose expectation is μ and whose maximum value is 2μ that the probability that the random variable is less than $\mu/4$ is at most $1/3$.

For our n balls into n bins, we get that the probability that there are more than k balls in bin 1 is at most $1/k$. This does not particularly help us prove an interesting upper bound in the *max* load, since we use the union bound on all the n bins.

A stronger deviation bound is called Chebyshev's inequality.

Theorem 2

[Chebyshev's inequality.] For any random variable X, and $\mu = E[X]$, and $\sigma^2 = E[(X-\mu)^2]$,

$$Pr[|X - E[X]| > t\sigma] \le 1/t^2.$$

Proof:

This inequality follows from Markov's inequality. Consider the positive random variable $Y = (X - E[X])^2$. Note that $E[Y] = \sigma^2$. We have

$$Pr[Y > t^2\sigma^2] \le 1/t^2,$$

by Markov's inequality.

But, $Pr[|X - E[X]| > t\sigma] = Pr[Y > t^2\sigma^2]$. Thus, the theorem holds.

□

The quantity $E[(X-E[X])^2]$ is called the variance of X, sometimes denoted by $Var[X]$. The square root of the variance is the standard deviation of X, usually denoted by the symbol σ. (By the way, the symbol μ often denotes $E[X]$.) One easy to verify inequality about the variance is

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2. \tag{1}$$

To use Chebyshev, we need to compute the varaiance of our random variable. For our balls in bins experiments, the load on bin 1 is $X = \sum_i X_i$ where X_i is $0 - 1$ indicator variable for ball i choosing bin 1. Recall that $E[X] = 1$. Now,

$$X^2 = \sum_{i,j} X_i X_j = \sum_i X_i^2 + \sum i \neq j X_i X_j$$

.

Noting that $X_i^2 = X_i$, and taking expectations, we get

$$E[X^2] = \sum_i E[X_i] + \sum i \neq j E[X_i X_j].$$

Now, $\sum_i E[X_i] = E[X]$, and $E[X_i X_j] = Pr[X_i = 1 \text{ and } X_j = 1] = (1/n)^2$. So, we have

$$E[X^2] = E[X] + n(n-1)/n^2,$$

and plugging into equation 1, we get

$$Var[X] = (n-1)/n.$$

Since this is upper bounded by 1. We can now use Chebyshev to bound the probability that the load is greater than $2\sqrt{n}$ by $1/4n$, which then allows us to say that the probability that any bin exceeds load $2\sqrt{n}$ is at most $1/4$.

This still is not so good.

3 Chernoff Bounds

Now, we examine a bound that gets reasonably tight answers. The following theorem is ascribed to Chernoff and/or Hoeffding. There are numerous forms, of which this is one.

Theorem 3

For a random variable, $X = \sum_i X_i$, where X_i are $0 - 1$ random variables, and with mean $\mu = E[X]$, for $\delta > 0$

$$Pr[X > (1+\delta)\mu] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\mu . \tag{2}$$

and for $1 > \delta > 0$,

$$Pr[X < (1-\delta)\mu] \leq \left(\frac{e^{-}\delta}{(1-\delta)^{1-\delta}} \right)^\mu . \tag{3}$$

Proof:

We only give the proof for bounding the probability of the "upper tail" or equation 2. Let t be an arbitrary positive constant.

$$
\begin{aligned}
Pr[X > (1+\delta)\mu] &\leq Pr[e^{tX} > e^{(1+\delta)t\mu}] \\
&\leq \frac{E[e^{tX}]}{e^{(1+\delta)t\mu}} \\
&\leq \frac{\Pi_i E[e^{tX_i}]}{e^{(1+\delta)t\mu}} \\
&\leq \frac{\Pi_i (p_i e^t + (1-p_i)e^0)}{e^{(1+\delta)t\mu}} \\
&\leq \frac{\Pi_i (1 + p_i(e^t - 1))}{e^{(1+\delta)t\mu}} \\
&\leq \frac{\Pi_i e^{p_i(e^t-1)}}{e^{(1+\delta)t\mu}} \\
&\leq \frac{e^{\sum_i p_i(e^t-1)}}{e^{(1+\delta)t\mu}} \\
&\leq \frac{e^{(e^t-1)\sum_i p_i}}{e^{(1+\delta)t\mu}} \\
&\leq \frac{e^{(e^t-1)\mu}}{e^{(1+\delta)t\mu}} = \left(\frac{e^{(e^t-1)}}{e^{(1+\delta)t}}\right)^{\mu}
\end{aligned}
$$

The third line follows from the fact that the X_i are independent. (That is, $E[XY] = E[X]E[Y]$ for independent random variables X and Y.) The sixth line follows from the fact that for positive x, $(1+x) < e^x$.

Choosing $t = \ln(1+\delta)$, we get equation 2.

$\square$

The Lemma above gives us an upper bound on the probability of the *upper and lower tails* of the distribution of X. Note, that the case of equation 2 can be expressed in a simpler form, which is more convenient for our needs:

$$
\Pr\big[X \geq (1+\delta)\mu\big] \leq \begin{cases} e^{-\delta^2\mu/3} & \text{if } \delta \leq 1, \\ e^{-\delta^2\mu/4} & \text{if } \delta \leq 2e-1, \\ 2^{-\delta\mu} & \text{if } \delta > 2e-1. \end{cases} \tag{4}
$$

Question 2: Prove these inequalities using the lemma above.

We can use equation 4 to show an upper bound on the maximum load with m balls in n bins, with high probability. As we would like to bound the probability by

$$
\Pr\big[X \geq (1+\delta)\left(\frac{m}{n}\right)\big] \leq \frac{1}{n^2},
$$

we need to determine the appropriate δ values for which this may hold. For the simple case where $m = n$, assuming $\delta \geq 2e$ we get

$$\Pr\bigl[X \geq (1+\delta)\bigr] \leq 2^{-\delta}$$

hence, for $\delta \geq 2\log_2 n$, it holds that

$$\Pr\bigl[X \geq (1+\delta)\bigr] \leq 2^{-2\log_2 n} = \frac{1}{n^2}$$

Obviously, a deviation of $2\log n$ is not too satisfying... However, we can achieve a better bound with high probability, for the general case and with bigger values of m: opting for the first case of equation 4, we require that

$$e^{-\delta^2\mu/3} \leq \frac{1}{n^2}$$

therefore

$$\begin{aligned} -\frac{\delta^2\mu}{3} &\leq -2\ln n \\ \delta^2 &\geq \frac{6\ln n}{\mu} = \frac{6n\ln n}{m} \\ \delta &\geq \sqrt{\frac{6n\ln n}{m}}\ . \end{aligned}$$

Since it is assumed that $\delta \leq 1$, we get that $m \geq 6n\ln n$. Then, denoting $c = \frac{m}{6n\ln n} \geq 1$, the maximum load is

$$\left(1+\frac{1}{\sqrt{c}}\right)\left(\frac{m}{n}\right)$$

with high probability.

Question 3: Say a poll gives support of 50% for a candidate and used 1000 samples. Use one or more of the simplified Chernoff inequalites above to show that the probability that her or his actual support is less than 40rough bound. (You may assume that their support greater than 0 since none of the bounds above give anything for $\mu = 0$.) Be careful, the expectation of the random variable is what the pollster is estimating.