6.042/18.062J Mathematics for Computer Science December 7, 2006
Tom Leighton and Ronitt Rubinfeld Lecture Notes

Large Deviations

Today we are going to talk more about the probability that a random variable deviates from its expectation. We have already seen examples where a random variable tends to be close to its expectation and where it tends to be far from its expectation. We also saw that the variance can be used to get some handle on a random variable. Today we will cover some new tools.

1 Markov and Chebyshev's Inequality

Markov's theorem say that if a random variable is never negative, then it is unlikely to greatly exceed its mean.

Theorem 1. *If R is a non-negative random variable, then for all $x > 0$, $\Pr(R \geq x) \leq \frac{\mathrm{Ex}(R)}{x}$.*

In other words, if R is never negative and $\mathrm{Ex}(R)$ is small, then R will also be small with probability near 1.

Proof. From the theorem of total expectation from recitation last week, we know

$$\begin{aligned}\mathrm{Ex}(R) &= \mathrm{Ex}(R \mid R \geq x)\Pr(R \geq x) + \mathrm{Ex}(R \mid R < x)\Pr(R < x)\\ &\geq x\Pr(R \geq x).\end{aligned}$$

Hence, $\Pr(R \geq x) \leq \frac{\mathrm{Ex}(R)}{x}$. □

Markov's theorem is often expressed in an alternate form, which is an easy corollary.

Corollary 2. *If R is a non-negative random variable, then $\forall c > 0$, $\Pr(R \geq c\,\mathrm{Ex}(R)) \leq \frac{1}{c}$.*

Proof. Set $x = c\,\mathrm{Ex}(R)$. Then we get $\frac{\mathrm{Ex}(R)}{x} = \frac{1}{c}$ in the bound above. □

For example, suppose R is the weight of a random person, and $\mathrm{Ex}(R) = 100$. Suppose we don't know the distribution of R. We can still compute useful information, like $\Pr(R \geq 200) \leq \frac{1}{2}$, by using Markov's bound. The interpretation here is that at most $1/2$ of the population have weight ≥ 200. If this were not the case, then the expectation would be larger than 100, a contradiction.

Now, when we say that at most $1/2$ of the population weighs over 200, this is a deterministic fact if the average weight is 100. Thus, Markov's theorem is about probabilities but it implies deterministic facts. This is because we have weighted all the sample points with a certain probability, in this case, the uniform distribution. Probabilities are really just weights, so it is natural to go back and forth. We turn N sample points into a random variable by assigning each probability $\frac{1}{N}$.

Let's revisit the Chinese Appetizer Problem and the Hat Check Problem. In the first problem there are N people at a Chinese restaurant that spin a Lazy Susan with appetizers. As argued in earlier lectures, the expected number of people to get the right appetizer is 1. Moreover, the Markov bound tells us that the probability N people get the right appetizer is at most $\frac{1}{N}$, and in fact in this case it is tight! On the other hand, for the Hat Check Problem, the probability all N people pick up their same hat is only $\frac{1}{N!}$, which is much smaller.

Okay, now what if the random variable R can be negative? In this case Markov's bound does not apply. Indeed, consider R with $\Pr(R = 1000) = \frac{1}{2}$ and $\Pr(R = -1000) = \frac{1}{2}$. Then $\mathrm{Ex}(R) = 0$, but $\Pr(R \geq 1000) \neq 0$.

Markov's bound gives us an upper bound on the probability that a random variable is large. It turns out, though, that there is a related result to get an upper bound on the probability that a random variable is small.

Corollary 3. *If $R \leq u$ for some $u \in \mathbb{R}$, then for all $x < u$,*

$$\Pr(R \leq x) \leq \frac{u - \mathrm{Ex}(R)}{u - x}.$$

Proof. Note that $\Pr(R \leq x) = \Pr(u - R \geq u - x)$. Now we can apply Markov's bound on the random variable $u - R$, which is non-negative.

$$\begin{aligned}\Pr(u - R \geq u - x) &\leq \frac{\mathrm{Ex}(u - R)}{u - x} \\ &= \frac{u - \mathrm{Ex}(R)}{u - x}\end{aligned}$$

□

As an example, let's look at quiz scores. Suppose R is the score of a random student, and the maximum score obtainable is 100. Suppose $\mathrm{Ex}(R) = 75$. Then

$$\Pr(R \leq 50) \leq \frac{100 - 75}{100 - 50} = \frac{1}{2},$$

by the previous corollary.

In some cases, like the Chinese Appetizer Problem, Markov's theorem is tight, but in other cases, like the Hat Check Problem, the bound is far from reality. Not surprisingly, if you know more about the distribution, you can get better bounds. For example, if you know the variance you can often get better bounds on the probability of deviating from the mean by using a result known as Chebyshev's Theorem.

Theorem 4. *For all $x > 0$ and for any random variable R,*

$$\Pr\left(|R - \mathrm{Ex}\left(R\right)| \geq x\right) \leq \frac{\mathrm{Var}\left[R\right]}{x^2}.$$

Chebyshev's Theorem is very similar to Markov's Theorem but uses the added info supplied by the variance to get a better upper bound. In fact, the proof uses Markov's Theorem.

Proof.

$$\Pr\left(|R - \mathrm{Ex}\left(R\right)| \geq x\right) = \Pr\left((R - \mathrm{Ex}\left(R\right))^2 \geq x^2\right) \leq \frac{\mathrm{Ex}\left((R - \mathrm{Ex}\left(R\right))^2\right)}{x^2} = \frac{\mathrm{Var}\left[R\right]}{x^2}.$$

□

A useful corollary is the following.

Corollary 5.

$$\Pr\left(|R - \mathrm{Ex}\left(R\right)| \geq c\sigma(R)\right) \leq \frac{1}{c^2}.$$

Proof. Just set $x = c\sigma(R)$ in the above. □

Here is an example. Suppose R is the IQ of a random person. We assume $R \geq 0$, although this may in fact be false (remember Bobo?). Also assume $\mathrm{Ex}\left(R\right) = 100$ and $\sigma(R) = 10$.

Let's compute $\Pr\left(R \geq 200\right)$. By Markov's Theorem, this is at most $1/2$. However, by Chebyshev's Theorem,

$$\begin{aligned}
\Pr\left(R \geq 200\right) &= \Pr\left(R - 100 \geq 100\right) \\
&= \Pr\left(R - \mathrm{Ex}\left(R\right)\right) \geq 10\sigma(R) \\
&\leq \Pr\left(|R - \mathrm{Ex}\left(R\right)| \geq 10\sigma(R)\right) \\
&\leq \frac{1}{10^2} = \frac{1}{100}.
\end{aligned}$$

Thus, we've used the standard deviation, some extra information about the distribution, to derive a much better bound. Notice in the proof that we used $\Pr\left(R - \mathrm{Ex}\left(R\right) \geq 10\sigma(R)\right) \leq \Pr\left(|R - \mathrm{Ex}\left(R\right)| \geq 10\sigma(R)\right)$. This always holds, though a common mistake is to assume that $\Pr\left(R - \mathrm{Ex}\left(R\right) \geq 10\sigma(R)\right) = \frac{1}{2}\Pr\left(|R - \mathrm{Ex}\left(R\right)| \geq 10\sigma(R)\right)$, which holds, e.g., if the distribution is symmetric about the mean. This is by no means true in general.

However, if as in the example above, you only care about 1-sided error, then you can do slightly better than Chebyshev's Theorem, but not a lot better.

Theorem 6. *For any random variable R,*

$$\Pr\left(R - \operatorname{Ex}(R) \geq c\sigma(R)\right) \leq \frac{1}{c^2+1},$$

and

$$\Pr\left(R - \operatorname{Ex}(R) \leq -c\sigma(R)\right) \leq \frac{1}{c^2+1}.$$

Proving this theorem is a bit trickier, and we won't do it in class. In general, this is the best you can say given only the expected value and standard deviation. Returning to the IQ example, we get $\Pr\left(R \geq 200\right) \leq \frac{1}{101}$, which is a slight improvement.

Here's another example. Say we give an exam. What fraction of the class can score more than 2 standard deviations away from average? If R is the score of a random student, the answer is

$$\Pr\left(\left|R - \operatorname{Ex}(R)\right| \geq 2\sigma(R)\right) \leq \frac{1}{4}.$$

For one-sided error, the fraction that could be 2 standard deviations or more high is at most $1/(2^2+1) = 1/5$. This holds no matter what the test scores were, and is again a deterministic fact derived using probabilistic tools.

2 Chernoff Bounds

The bounds we get using the Markov Theorem and the Chebyshev Theorem are sometimes very good, and sometimes very bad. Now we're going to turn to a special case of a random variable which arises in practice. The special case is when the random variable is the sum of lots of other random variables that are mutually independent, and the bound on the probability of deviating from the mean is known as the Chernoff bound. Chernoff was a professor here, and at the time of discovery did not put much importance on his bounds, which were much later applied by many others in many situations.

Theorem 7 (Chernoff Bounds). *Let $T_1, \ldots, T_n$ be any mutually independent random variables such that $\forall j$, $0 \leq T_j \leq 1$. Let $T = \sum_{j=1}^{N} T_j$. Then for any $c > 1$,*

$$\Pr\left(T \geq c \operatorname{Ex}(T)\right) \leq e^{-\alpha \operatorname{Ex}(T)},$$

where $\alpha = c \ln c + 1 - c > 0$.

In general this is a much better bound than you get from Markov or Chebyshev. The probability from Markov is $1/c$. The bound from Chebyshev is only slightly better. With Chernoff, the bound is exponentially small in $c \ln c$ times the expected value. This is a huge difference.

For example, using Chernoff Bounds, $\Pr\left(T \geq 2 \operatorname{Ex}(T)\right) \leq e^{-38}$ if $\operatorname{Ex}(T) = 100$. In this case Markov would only give $1/2$, and the one-sided extension of Chebyshev would only give $1/(2^2+1) = 1/5$.

Of course, Chernoff Bounds do not apply to all distributions. They only work when R is the sum of random variables defined on the interval $[0, 1]$. But this is a pretty broad class, and includes, for example, the binomial distribution. In fact, the T_j form a much broader class since the T_j need not have the same distribution, and their distribution can be arbitrary on the interval $[0, 1]$, rather than just Bernoulli.

Another nice thing about the Chernoff Bound is that the bound does not directly depend on the number of random variables being summed, and you can show that even small deviations are unlikely.

For example, suppose 10 million people play "Pick 4". This is a lottery game where you pick a 4-digit number and you win if you get an exact natch. Then the probability of winning is just $\Pr(\text{Win}) = \frac{1}{10000}$, and the expected number of winners is 10 million divided by a thousand, or 1000.

Suppose each of the 10 million players pick mutually independent random numbers. Then by Chernoff Bounds, we know $\Pr(\geq 2000 \text{ winners}) \leq e^{-.38 \cdot 1000} = e^{-380}$. Moreover, $\Pr(\geq 1100 \text{ winners}) \leq e^{-4.8} < .01$, so only a 1% chance that the number of winners is 10% over the expectation. Note that the Markov and Chebyshev Theorems are useless here.

So let's prove the theorem. Then we'll apply it to load balancing. The proof of the theorem is similar to the proof of Chebyshev's Theorem. As in the proof of Chebyshev, we'll use Markov's Theorem, but in this case, we exponentiate the deviation instead of squaring it before applying Markov's Theorem. The proof is clever and a bit tricky. We don't expect you to be able to derive such a proof on your own in this class.

Proof. Define $R_j = T_j - \text{Ex}(T_j)$ for $j = 1, 2, \ldots, N$. Then

$$-\text{Ex}(T_j) \leq R_j \leq 1 - \text{Ex}(T_j),$$

and $\text{Ex}(R_j) = 0$. We have,

$$\begin{aligned}
\Pr(T \geq c\,\text{Ex}(T)) &= \Pr(R \geq (c-1)\,\text{Ex}(T)) \\
&= \Pr\left(c^R \geq c^{(c-1)\,\text{Ex}(T)}\right) \\
&\leq \frac{\text{Ex}\left(c^R\right)}{c^{(c-1)\,\text{Ex}(T)}} \quad \text{Markov's Bound.}
\end{aligned}$$

Thus,

$$\begin{aligned}
\text{Ex}\left(c^R\right) &= \text{Ex}\left(c^{R_1+R_2+\cdots+R_N}\right) \\
&= \text{Ex}\left(\Pi_{j=1}^N c^{R_j}\right) \\
&= \Pi_{j=1}^N \text{Ex}\left(c^{R_j}\right) \quad \text{independence of the } T_j \text{ and thus } R_j.
\end{aligned}$$

We need the following facts.

Fact 1. *If $-m \leq z \leq 1-m$, then $c^z \leq c^{-m}(1 + m(c-1)) + z(c^{1-m} - c^{-m})$.*

This follows from the fact that the right-hand-side describes a line which intersects the curve c^z (here z is the variable) at $z = -m$ and $z = 1 - m$. Moreover, since c^z is convex, the curve is entirely below this line.

Fact 2. $1 + m(c-1) \leq e^{m(c-1)}$.

This follows from the Taylor expansion $1 + x \leq e^x$.

Resuming the proof, set $m = \text{Ex}\,(T_j)$ and $z = R_j$. Then

$$\begin{aligned} \text{Ex}\left(c^{R_j}\right) &\leq \text{Ex}\left(c^{-m}e^{m(c-1)} + (c^{1-m} - c^{-m})R_j\right) \\ &= c^{-m}e^{m(c-1)} \quad \text{since } \text{Ex}\,(R_j) = 0 \\ &= e^{m(c-1-\ln c)} \end{aligned}$$

Thus,

$$\begin{aligned} \Pi\, \text{Ex}\left(c^{R_j}\right) &\leq \Pi e^{\text{Ex}(T_j)(c-1-\ln c)} \\ &= e^{(c-1-\ln c)\sum \text{Ex}(T_j)} \\ &= e^{(c-1-\ln c)\,\text{Ex}(T)} \end{aligned}$$

Concluding,

$$\begin{aligned} \Pr\left(T \geq c\,\text{Ex}\,(T)\right) &\leq \frac{e^{(c-1-\ln c)\,\text{Ex}(T)}}{c^{(c-1)\,\text{Ex}(T)}} \\ &= e^{\text{Ex}(T)(c-1-\ln c - c\ln c + \ln c)} \\ &= e^{\text{Ex}(T)(-c\ln c - 1 + c)} \\ &= e^{-\alpha\,\text{Ex}(T)}, \end{aligned}$$

where $\alpha = c\ln c 1 - c$. □

3 Load Balancing

Suppose we need to build a load balancing device to assign a set of N jobs $B_1, B_2, \ldots, B_N$ to a set of m servers $S_1, S_2, \ldots, S_m$. If you are hosting a decent-sized website, N might be about $100K$ and m might be about 10. Suppose the ith job B_j takes L_j time, $0 \leq L_j \leq 1$ (say, in seconds). The goal is to assign the N jobs to the m servers so that the load is as balanced as possible (i.e., so that the busiest server finishes as quickly as possible).

Suppose each server works sequentially through the jobs that are assigned to it and finishes in time equal to the sum of job lengths assigned to the server. Let $L_{Tot} = \sum_{j=1}^{N} L_j$ be the total sum of job lengths. With perfect load balancing then, each server would take $L_{Bal} = \frac{L_{Tot}}{m}$ time. Now if you know the L_j, this is a variant of the knapsack problem. It is hard to get all the tasks perfectly balanced but good algorithms exist to get close.

But we are interested in the case when you don't know the job lengths until after you make the assignments, which is often the case in practice. At first it seems hopeless. The idea, however, is to assign the jobs randomly, i.e., to pick 1 of m processors uniformly at random for each job. This is a very useful technique in computer science when you don't have enough information to solve a problem or a deterministic solution is too hard to figure out.

Let's see how it works in this case. Normally this kind of technique isn't covered until grad school, but we're ready for it now! We'll start by seeing how much load gets assigned to the ith server S_i.

Let $R_{i,j}$ be the load on S_i from job B_j. Then $R_{i,j} = L_j$ if B_j is assigned to S_i, and is 0 otherwise. Note, $0 \leq R_{i,j} \leq 1$. Let R_i be the total load on S_i from all jobs. Then $R_i = \sum_{j=1}^{N} R_{i,j}$. So,

$$\begin{aligned}
\operatorname{Ex}(R_i) &= \sum_{j=1}^{N} \operatorname{Ex}(R_{i,j}) \\
&= \sum_{j=1}^{N} L_j/m \\
&= \frac{1}{m} L_{Tot} \\
&= L_{Bal}.
\end{aligned}$$

So the expected load on the ith server is what we would get if the load were perfectly balanced. Because of mutual independence and that $0 \leq R_{i,j} \leq 1$, we can apply Chernoff's Bound. Thus, $\Pr(R_i \geq cL_{Bal}) \leq e^{-\alpha L_{Bal}}$, where $\alpha = c \ln c + 1 - c$.

Now this holds for each i individually, but we need to make sure that none of the servers has too much load. To do this, we need to bound the probability that the worst server takes more than cL_{Bal} steps. This is just

$$\Pr(R_1 \geq cL_{Bal} \vee R_2 \geq cL_{Bal} \vee \cdots \vee R_m \geq cL_{Bal}).$$

We can upper bound this by summing the probabilities,

$$\leq \Pr(R_1 \geq cL_{Bal}) + \cdots + \Pr(R_m \geq cL_{Bal}) \leq me^{-\alpha L_{Bal}}.$$

For example, if we plug in $c = 1.1$ (10% above the optimal), $\alpha > .0048$. Let $N = 100K$ and $m = 10$. Say the average job is $1/4$ seconds. Then $L_{Tot} = 25K$, and $L_{Bal} = 2500$. Then, $\Pr(\exists \text{ server with } \geq 10\% \text{ extra load }) \leq 10e^{-.0048 \cdot 2500} < e^{-9}$.

Thus, our randomized algorithm got all 10 servers to within 10% of the optimal load with very high probability. This gives a very nice performance, especially compared to the worst case. The only way we do poorly is if we have a bad random assignment; that is, this result holds no matter what the job lenghts are. This is a really useful and powerful tool in computer science.