Computer Science and Engineering, UCSD
Tail Inequalities

October 12, 1999
Author: Bellare

Tail Inequalities

1 The problem

We have a collection $X_1, \ldots, X_n$ of random variables each ranging between 0 and 1. We let $p_i = \mathbf{E}[X_i]$ for $i = 1, \ldots, n$ and we let $X = X_1 + \cdots + X_n$. We let $\mu = \mathbf{E}[X]$. Linearity of expectation tells us that $\mu = p_1 + \cdots + p_n$. We fix some parameter $A > 0$ and are interested in the probability that $X - \mu \geq A$, namely that X exceeds its expectation by some amount A.

A particular form in which we wish to study this probability is the following. We let $A = x\mu$ for some $x > 0$. Then $\Pr[\, X - \mu \geq A \,] = \Pr[\, X \geq (1+x)\mu \,]$. We are interested in how this behaves as a function of x, with all other quantities being fixed. Mostly we want good upper bounds.

This situation arises extremely often. Typically, something is known about the "amount of independence" of the random variables $X_1, \ldots, X_n$. The simplest case is that they are actually independent. Another case common in computer science is that they satisfy some limited form of independence, for example pairwise independence, or, more generally, t-wise independence where $t \geq 2$ is some integer. (When $t = n$ we have independence.) Alternatively, they may satisfy some form of "almost independence". Tail inequalities deal with these situations.

In mathematics courses on introductory probability theory, these problems are typically treated via the "laws of large numbers" and the "central limit theorem". These provide qualitative understanding of how the probabilities in question behave as a function of n. Tail inequalities are the quantitative analogue.

We will begin with some background and then go to the most common case, the one where the random variables are (fully) independent. Then we address limited independence.

2 Basic inequalities

The most basic inequality is Markov's.

Proposition 1 [Markov's Inequality] For any non-negative random variable X and any real number $a > 0$ we have

$$\Pr[\, X \geq a \,] \leq \frac{\mathbf{E}[X]}{a} \;. \quad \blacksquare$$

As an example let $a = 2 \cdot \mathbf{E}[X]$. Then the above says $\Pr[\, X \geq 2 \cdot \mathbf{E}[X] \,] \leq 1/2$. Namely, if you move out to twice the expectation, you can have only half the area under the curve to your right. This is quite intuitive.

Proof of Proposition 1: This is a simple computation:

$$\begin{aligned} a \cdot \Pr[X \geq a] &= a \cdot \sum_{x\,:\,x \geq a} \Pr[X = x] \\ &\leq \sum_{x\,:\,x \geq a} x \cdot \Pr[X = x] \\ &\leq \sum_{x} x \cdot \Pr[X = x] \\ &= \mathbf{E}[X] . \end{aligned}$$

Where in this proof did we use that $X \geq 0$? Third line above. ▮

Markov's inequality is rather weak. The curious thing is that nonetheless it is in the end the root of the most powerful tail inequalities around. You just have to find the right way to use it.

One step up from Markov's inequality is Chebyschev's inequality. To state it we first recall that if X is a random variable then its variance is $\mathbf{Var}[X] = \mathbf{E}[(X-\mu)^2] = \mathbf{E}[X^2] - \mu^2$ where $\mu = \mathbf{E}[X]$ is the expectation of X.

Proposition 2 [Chebychev's inequality] Let X be a random variable, and let $A > 0$. Then

$$\Pr[\,|X-\mu| \geq A\,] \;\leq\; \frac{\mathbf{Var}[X]}{A^2} . \;▮$$

Proof of Proposition 2: Let $\mu = \mathbf{E}[X]$. Let Y be the random variable defined by $Y = (X-\mu)^2 = X^2 - 2\mu \cdot X + \mu^2$. Then

$$\mathbf{E}[Y] = \mathbf{E}\left[X^2\right] - 2\mu \cdot \mathbf{E}[X] + \mu^2 = \mathbf{E}\left[X^2\right] - 2\mu^2 + \mu^2 = \mathbf{Var}[X] .$$

Since $Y \geq 0$ we can apply Proposition 1 to it. We have

$$\begin{aligned} \Pr[\,|X-\mu| \geq A\,] &= \Pr\left[Y \geq A^2\right] \\ &\leq \frac{\mathbf{E}[Y]}{A^2} \\ &= \frac{\mathbf{Var}[X]}{A^2} \end{aligned}$$

as desired. ▮

This will come in useful for tail inequalities on pairwise independent random variables.

3 Tail inequalities for independent random variables

We have *independent* random variables $X_1, \ldots, X_n$ ranging between 0 and 1. For simplicity let's assume they are actually boolean, meaning assume only the values 0 and 1. (This turns out to be the worse case for the situations we deal with.) In that case, $p_i \overset{\text{def}}{=} \mathbf{E}[X_i] = \Pr[X_i = 1]$ for

$i = 1, \ldots, n$. As usual set $X = X_1 + \cdots + X_n$ and $\mu = \mathbf{E}[X]$. We are interested in upper bounding $\Pr[X - \mu > A]$ where $A > 0$ is some given real number.

The law of large numbers says that if we fix A, p then

$$\lim_{n\to\infty} \Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - p_i)\right| > A\right] \; = \; 0 \; .$$

That is, the probability that X deviates from its expectation gets smaller and smaller as the number of samples n grows. In computer science we want more precise information: our interest is in how this probability tails of as a function of n.

Nomenclature in this area is not uniform, but the bounds we will now discuss sometimes go under the name of Chernoff-type bounds. A good reference is [1].

The first bound we specify is the simplest, yet good enough in many of the applications.

Proposition 3 Let $X_1, \ldots, X_n$ be independent, 0/1 valued random variables, and let $p_i = \mathbf{E}[X_i]$ for $i = 1, \ldots, n$. Let $X = X_1 + \cdots + X_n$ and let $\mu = \mathbf{E}[X]$. Let $A > 0$ be a real number. Then

$$\Pr[X - \mu > A] \;\leq\; e^{-A^2/2n} \; . \; \blacksquare$$

We won't prove this because below we will prove something stronger. But let's discuss it. To get an understanding of the bound we consider the case where $p_1 = p_2 = \cdots = p_n$.

Corollary 4 Let $X_1, \ldots, X_n$ be independent, 0/1 valued random variables all having the same expectation p. Let $X = X_1 + \cdots + X_n$ and let $\mu = \mathbf{E}[X]$. Let $x > 0$ be a real number. Then

$$\begin{aligned} \Pr[X > (1+x)\mu] &\leq e^{-x^2p^2n/2} \\ &= e^{-x^2\mu p/2} \; . \; \blacksquare \end{aligned}$$

Proof: Set $A = x\mu$ in Proposition 3 and use the fact that $\mu = pn$. $\blacksquare$

This shows us the punch-line: roughly, $\Pr[X > (1+x)\mu]$ decreases exponentially with n for fixed x, p. In other words, the probability that the sum of independent random variables deviates significantly from its mean (expectation) drops very quickly as the number of random variables grows. For example if you toss many fair coins, the probability of getting significantly more than 50% heads is very small.

However you have to be careful with the above bound. The intuition that $\Pr[X > (1+x)\mu]$ decreases exponentially with n is quite sensitive to the values of x, p, and there are many common situations in which the bounds obtained from the above are not good enough. One way to see why is to consider the second line in the bound of Corollary 4, which we got just by substituting μ for np in the first line. It shows us that if p is small, the bound is worse even for a fixed value of the expectation μ of the sum X. This is actually a weakness in the bound, not necessarily a reflection of reality.

Here now is a stronger Chernoff-type bound. This is pretty much "tight", meaning as good as you can get. It may at first be hard to interpret, but we'll elucidate it later.

Theorem 5 Let $X_1, \ldots, X_n$ be independent, 0/1 valued random variables, and let $p_i = \mathbf{E}[X_i]$ for $i = 1, \ldots, n$. Let $X = X_1 + \cdots + X_n$ and let $\mu = \mathbf{E}[X]$. Let $\beta > 1$ be a real number. Then

$$\Pr[X > \beta \cdot \mu] \quad \leq \quad e^{-g(\beta)\cdot\mu}$$

where we define the function $g(\cdot)$ by $g(\beta) = \beta \ln(\beta) + 1 - \beta$. ▌

To visualize this it is again useful to set $\beta = 1 + x$ for $x > 0$ and see what happens to the bound viewed as a function of x.

Corollary 6 Let $X_1, \ldots, X_n$ be independent, 0/1 valued random variables, and let $p_i = \mathbf{E}[X_i]$ for $i = 1, \ldots, n$. Let $X = X_1 + \cdots + X_n$ and let $\mu = \mathbf{E}[X]$. Let $0 < x \leq 2$ be a real number. Then

$$\Pr[X > (1+x) \cdot \mu] \quad \leq \quad e^{-3x^2\mu/10} \ . \ ▌$$

Notice the improvement over Corollary 4: the factor of p in the exponent has vanished. That is quite a change; the new bound is much better.

Be careful when you apply this bound to note that it only holds for $x \leq 2$. If x is larger, our intuition is that the probability in question should be even lower, yet the bound above does not apply. If you want a bound that works for "large" x you will need to go back to the proofs of Theorem 5 and Corollary 6 and try to extend them. It would be nice to do this and get a clean bound for larger x, actually. We might explore these issues later, but right now I want to look more at the proofs. Let's begin by seeing why Corollary 6 follows from Theorem 5.

Proof of Corollary 6: Theorem 5 tells us that

$$\Pr[X > (1+x) \cdot \mu] \quad \leq \quad e^{-g(1+x)\cdot\mu}$$

where $g(\cdot)$ is the function defined in the statement of Theorem 5. So it suffices to show that $g(1+x) \geq 3x^2/10$ for $0 < x \leq 2$. We do this using Taylor series approximations. We have

$$\begin{aligned}
g(1+x) &= (1+x)\ln(1+x) + 1 - (1+x) \\
&= (1+x)\ln(1+x) - x \\
&= -x + (1+x) \cdot \sum_{i\geq 1} (-1)^{i-1} \cdot \frac{x^i}{i} \\
&= -x + \sum_{i\geq 1} (-1)^{i-1} \cdot \frac{x^i}{i} + \sum_{i\geq 2} (-1)^{i} \cdot \frac{x^i}{i-1} \\
&= \sum_{i\geq 2} (-1)^{i-1} \cdot \frac{x^i}{i} + (-1)^{i} \cdot \frac{x^i}{i-1} \\
&= \sum_{i\geq 2} (-1)^{i} \cdot \frac{x^i}{i(i-1)} \\
&= \frac{x^2}{2} - \sum_{i\geq 3} (-1)^{i-1} \cdot \frac{x^i}{i(i-1)} \\
&\geq \frac{x^2}{2} - \left(\frac{x^3}{6} - \frac{x^4}{12} + \frac{x^5}{20}\right) .
\end{aligned}$$

We now want to upper bound the expression in parentheses in the last line above. We consider the function $f(x) = 1/6 - x/12 + x^2/20$. It attains its minimum at $x = 5/6$ so for $0 < x \leq 2$ the maximum value of f is attained at $x = 2$. Thus the above is

$$\begin{aligned} &\geq \frac{x^2}{2} - x^2 \cdot 2 \cdot f(2) \\ &= x^2 \cdot \left(\frac{1}{2} - \frac{1}{5}\right) \\ &= \frac{3x^2}{10} \ . \end{aligned}$$

That concludes the proof. ▮

Now we come to the interesting part, namely the proof of Theorem 5. It introduces the idea of *exponential generating functions*. It is quite neat, illustrating many simple but powerful techniques.

Proof of Theorem 5: We introduce a parameter $\lambda > 0$ whose value will be set later. Recall that $\mu = p_1 + \cdots + p_n = \mathbf{E}[X]$. We use the monotonicity of the exponential function and then apply Markov's inequality to get

$$\begin{aligned} \Pr[\, X > \beta \cdot \mu \,] &= \Pr\left[e^{\lambda X} > e^{\lambda \beta \cdot \mu} \right] \\ &\leq \frac{\mathbf{E}\left[e^{\lambda X}\right]}{e^{\lambda\beta\cdot\mu}} \ . \end{aligned} \tag{1}$$

This is the exponential generating function trick. We do something that looks really trivial. We note that the probability is unchanged if we exponentiate the terms involved, and then we use, of all things, the weakest of the inequalities around, namely Markov's inequality. Yet as we will now see, rather strong bounds emerge.

The next thing we do is bound the expectation in Equation (1). We start with the following–

$$\begin{aligned} \mathbf{E}\left[e^{\lambda X}\right] &= \mathbf{E}\left[e^{\lambda(X_1+X_2+\cdots+X_n)}\right] \\ &= \mathbf{E}\left[e^{\lambda X_1} \cdot e^{\lambda X_2} \cdot \ldots \cdot e^{\lambda X_n}\right] . \end{aligned}$$

The independence of $X_1, \ldots, X_n$ (this is the one and only place we use this) implies that the above equals

$$\mathbf{E}\left[e^{\lambda X_1}\right] \cdot \mathbf{E}\left[e^{\lambda X_2}\right] \cdot \ldots \cdot \mathbf{E}\left[e^{\lambda X_n}\right] .$$

Now we compute these individual expectations: For any $i = 1, \ldots, n$ we have

$$\begin{aligned} \mathbf{E}\left[e^{\lambda X_i}\right] &= 1 \cdot \Pr[\, X_i = 0 \,] + e^{\lambda} \cdot \Pr[\, X_i = 1 \,] \\ &= (1 - p_i) + e^{\lambda} \cdot p_i \\ &= 1 + (e^{\lambda} - 1) \cdot p_i \ . \end{aligned}$$

So at this point we have

$$\mathbf{E}\left[e^{\lambda X}\right] = [1 + (e^{\lambda} - 1) \cdot p_1] \cdot [1 + (e^{\lambda} - 1) \cdot p_2] \cdot \ldots \cdot [1 + (e^{\lambda} - 1) \cdot p_n] \ .$$

Notice that so far we have done no bounding; we have equalities.

At this point, what can we do? We are looking at a complex bound, product of many terms. We will start seeing terms involving products of the values $p_1, \ldots, p_n$, which is not something we know much about. What we do know something about is the sum of $p_1, \ldots, p_n$, because this is exactly μ, the expectation of X. We'd like to work this in. This is done by applying a very common and useful little inequality, namely that $1 + y \le e^y$ for any real number y. Set $y_i = (e^\lambda - 1)p_i$ and we get

$$\begin{aligned}
\mathbf{E}\left[e^{\lambda X}\right] &\le (1+y_1)\cdot(1+y_2)\cdot\ldots\cdot(1+y_n) \\
&\le e^{y_1}\cdot e^{y_2}\cdot\ldots\cdot e^{y_n} \\
&= e^{y_1+\cdots+y_n} \\
&= e^{(e^\lambda-1)(p_1+\cdots+p_n)} \\
&= e^{(e^\lambda-1)\mu} \; .
\end{aligned}$$

Let's now put this back together with Equation (1). That gives us

$$\begin{aligned}
\Pr\left[\, X > \beta\cdot\mu \,\right] &\le \frac{e^{(e^\lambda-1)\mu}}{e^{\lambda\beta\cdot\mu}} \\
&= e^{(e^\lambda-1)\mu-\lambda\beta\cdot\mu} \\
&= e^{-f(\lambda)\cdot\mu}
\end{aligned}$$

where

$$f(\lambda) \;=\; \lambda\beta - (e^\lambda - 1) \; .$$

Now we want to analyze the function $f(\cdot)$ and choose the value of $\lambda > 0$ that makes f as large as possible. Since all the above is true for any value of $\lambda > 0$ we can plug in this special value and that will be our bound. To analyze $f(\cdot)$ we use high-school level calculus. We compute the derivate: $f'(\lambda) = \beta - e^\lambda$. The function f' is positive for $\lambda < \ln(\beta)$, zero at $\lambda = \ln(\beta)$, and then negative for $\lambda > \ln(\beta)$. This tells us that f is increasing for $0 < \lambda < \ln(\beta)$ and decreasing for $\lambda > \ln(\beta)$. So the maximum is at $\lambda = \ln(\beta)$. Now we note that

$$f(\ln(\beta)) \;=\; \ln(\beta)\cdot\beta - \beta + 1 \; .$$

This is exactly what we called $g(\beta)$, so the proof is complete. ▌

The technique of this proof is the one used in all proofs of Chernoff-type bounds, with minor variations. It is useful to know it so that you can derive your own bounds if necessary.

4 Tail inequalities for pairwise independent random variables

Pairwise independence means that we cannot infer anything extra about X_i given X_j where $j \ne i$, even though we might be able to infer something or even everything about X_i if we were given X_j and X_k where i, j, k are distinct.

Definition 7 We say that $X_1, \ldots, X_n$ are *pairwise independent* random variables if for every $1 \leq i < j \leq n$ and every $a, b \in \mathbf{R}$ we have

$$\Pr[\, X_i = a \text{ and } X_j = b \,] \;=\; \Pr[\, X_i = a \,] \cdot \Pr[\, X_j = b \,] \ . \ \blacksquare$$

The tail inequality for such random variables makes use of the fact that the variance of a sum of pairwise independent random variables behaves exactly like the variance of a sum of independent random variables: it is the sum of the individual variances.

Lemma 8 Let $X_1, \ldots, X_n$ be pairwise independent random variables. Then

$$\mathbf{Var}\left[X_1 + \cdots + X_n\right] = \mathbf{Var}\left[X_1\right] + \cdots + \mathbf{Var}\left[X_n\right] \ . \ \blacksquare$$

Proof of Lemma 8: Use the formula for the variance and the linearity of expectation to get

$$\begin{aligned}
\mathbf{Var}\left[X_1 + \cdots + X_n\right] &= \mathbf{E}\left[(X_1 + \cdots + X_n)^2\right] - \mathbf{E}\left[X_1 + \cdots + X_n\right]^2 \\
&= \mathbf{E}\left[(X_1 + \cdots + X_n)(X_1 + \cdots + X_n)\right] - \left(\mathbf{E}\left[X_1\right] + \cdots + \mathbf{E}\left[X_n\right]\right)^2 \\
&= \mathbf{E}\left[\textstyle\sum_{i,j} X_i X_j\right] - \sum_{i,j} \mathbf{E}\left[X_i\right] \cdot \mathbf{E}\left[X_j\right] \\
&= \sum_{i,j} \mathbf{E}\left[X_i X_j\right] - \sum_{i,j} \mathbf{E}\left[X_i\right] \cdot \mathbf{E}\left[X_j\right] \\
&= \sum_{i} \mathbf{E}\left[X_i^2\right] + \sum_{i \neq j} \mathbf{E}\left[X_i X_j\right] - \sum_{i} \mathbf{E}\left[X_i\right]^2 - \sum_{i \neq j} \mathbf{E}\left[X_i\right] \cdot \mathbf{E}\left[X_j\right] \\
&= \sum_{i} \left(\mathbf{E}\left[X_i^2\right] - \mathbf{E}\left[X_i\right]^2\right) + \sum_{i \neq j} \left(\mathbf{E}\left[X_i X_j\right] - \mathbf{E}\left[X_i\right] \cdot \mathbf{E}\left[X_j\right]\right) \\
&= \sum_{i} \mathbf{Var}\left[X_i\right] + \sum_{i \neq j} \left(\mathbf{E}\left[X_i X_j\right] - \mathbf{E}\left[X_i\right] \cdot \mathbf{E}\left[X_j\right]\right) \ .
\end{aligned}$$

The pairwise independence means that $\mathbf{E}\left[X_i X_j\right] = \mathbf{E}\left[X_i\right] \cdot \mathbf{E}\left[X_j\right]$ whenever $i \neq j$. Thus the second sum above is zero, and we are done. $\blacksquare$

We can now obtain the tail inequality by applying Chebyshev's inequality.

Lemma 9 Let $X_1, \ldots, X_n$ be pairwise independent random variables, let $X = X_1 + \cdots + X_n$, let $A > 0$ be a real number, and let $\mu = \mathbf{E}\left[X\right]$. Then

$$\Pr[\, |X - \mu| > A \,] \;\leq\; \frac{\mathbf{Var}\left[X_1\right] + \cdots + \mathbf{Var}\left[X_n\right]}{A^2} \ . \ \blacksquare$$

Proof of Lemma 9: Proposition 2 tells us that

$$\Pr[\, |X - \mu| > A \,] \;\leq\; \frac{\mathbf{Var}\left[X\right]}{A^2} \ .$$

Now apply Lemma 8. $\blacksquare$

References

[1] R. Motwani and P. Raghavan, *Randomized algorithms*, Cambridge University Press, 1995.