Tail Inequalities

William Hunt
Lane Department of Computer Science and Electrical Engineering,
West Virginia University,
Morgantown, WV
William.Hunt@mail.wvu.edu

1 Introduction

In this chapter, we are interested in obtaining significantly stronger and sharper ways for making the statement: "with high probability". Using Markov's Inequality just doesn't give us a strong enough statement.

Chernoff bounds are another kind of tail bound. They bound the total amount of probability of some random variable Y that is far from the mean.

The Chernoff bound applies to a class of random variable that are a sum of independent indicator random variables. Occupancy problems for balls in bins is such a class, so Chernoff bounds apply.

2 Poisson Trials

The Chernoff bound works for random variables that are a sum of indicator variables with the same distribution (Bernoulli trials).

Let $X_1, X_{2,} \cdots, X_n$ be independent Bernoulli trials such that $\mathbf{Pr}[X_i = 1] = p_i$ and $\mathbf{Pr}[X_i = 0] = (1 - p_i)$

The sums of independent Bernoulli trials are considered Poisson trials. Thus, $X = X_1 + X_2 + \cdots + X_n$ is a Poisson trial. All our bounds apply to the special case when X_i are Bernoulli trials with identical probabilities, so that X has the binomial distribution.

For any $\delta > 0$, we want to answer the following questions about how much deviation of the expected value occurs so that we can find a bound on the tail probability.

1. What is the probability that X exceeds $(1 + \delta)\mu$?
   $\mathbf{Pr}[X > (1 + \delta)\mu]$
2. How large should δ be so the probability is less than ε?
   $\mathbf{Pr}[X > (1 + \delta)\mu] < \varepsilon$

3 Chernoff Bound

To answer the type of question useful in analyzing an algorithm, we can use the Chernoff Bound technique. This is seeking a bound on the tail probabilities.

Definition: 3.1 *For a random variable X, $\mathbf{E}[e^{tX}]$ is the moment generating function.*

We are going to focus on the sum of independent random variables. In particular, we are going to look at bounds on the sum of independent Poisson trials. The idea for the Chernoff Bound technique is to take the moment generating function of X and apply the Markov inequality to it.

Theorem: 3.1 *Let $X_1, X_{2,} \cdots, X_n$ be independent Poisson trials such that, for $1 \leq i \leq n$, $\mathbf{Pr}[X_i = 1] = p_i$ with $0 < p_i < 1$. Then, for $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbf{E}[X] = \sum_{i=1}^{n} p_i$, and for any $\delta > 0$,*

$$\mathbf{Pr}[X > (1+\delta)\mu] < \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}$$

Proof:

For any positive real number t,

$$\mathbf{Pr}[X > (1+\delta)\mu] = \mathbf{Pr}[e^{tX} > e^{t(1+\delta)\mu}]$$

Applying the Markov inequality to the right hand side, we get

$$\mathbf{Pr}[e^{tX} > e^{t(1+\delta)\mu}] < \frac{\mathbf{E}[e^{tX}]}{e^{t(1+\delta)\mu}}$$

This is easy to calculate because the products in the exponents become sums in the expected value expression and since each X_i is independent, the expectation of the product becomes the product of the expectations. So, the numerator can be calculated the expectation as:

$$\mathbf{E}[e^{tX}] = \mathbf{E}[e^{t\sum_{i=1}^{n} X_i}] = \mathbf{E}[\prod_{i=1}^{n} e^{tX_i}] = \prod_{i=1}^{n} \mathbf{E}[e^{tX_i}]$$

The random variable e^{tX_i} has the value of e^t with probability p_i and the value 1 with probability $1-p_i$. Therefore, the expected value is $p_i(e^t) + (1-p_i)(1)$ making the numerator:

$$\prod_{i=1}^{n} (p_i e^t + 1 - p_i) = \prod_{i=1}^{n} \left(1 + p_i(e^t - 1)\right)$$

Using the inequality $1+x<e^x$ with $x=p_i(e^t-1)$, we get the right hand side to be:

$$\mathbf{Pr}[e^{tX} > e^{t(1+\delta)\mu}] < \frac{\prod_{i=1}^{n} e^{p_i(e^t-1)}}{e^{t(1+\delta)\mu}}$$

$$\mathbf{Pr}[e^{tX} > e^{t(1+\delta)\mu}] < \frac{e^{\sum_{i=1}^{n} p_i(e^t-1)}}{e^{t(1+\delta)\mu}}$$

$$\mathbf{Pr}[e^{tX} > e^{t(1+\delta)\mu}] < \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}}$$

Now, we choose a t such that the right hand side is maximized: $t = \ln(1+\delta)$

Substituting t in to the inequality gives us our theorem.

□

Definition: 3.2 *The upper tail bound function for the sum of Poisson trials is*

$$\mathbf{Pr}[x > (1+\delta)\mu] = F^{+}(\mu,\delta) \triangleq \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}$$

We can use a similar argument to find the lower tail bound function for the sum of Poisson trials.

Theorem: 3.2 *Let $X_1, X_2, \cdots, X_n$ be independent Poisson trials such that, for $1 \le i \le n$, $\mathbf{Pr}[X_i = 1] = p_i$ with $0 < p_i < 1$. Then, for $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbf{E}[X] = \sum_{i=1}^{n} p_i$, and for any $\delta > 0$,*

$$\mathbf{Pr}[X < (1-\delta)\mu] < e^{\frac{-\mu\delta^2}{2}}$$

Proof: *For any positive real number t,*

$$\mathbf{Pr}[X < (1-\delta)\mu] = \mathbf{Pr}[-X > -(1-\delta)\mu]$$

$$\mathbf{Pr}[X < (1-\delta)\mu] = \mathbf{Pr}[e^{-tX} > e^{-t(1-\delta)\mu}]$$

Applying the Markov inequality to the right hand side, we get

$$\mathbf{Pr}[e^{-tX} > e^{-t(1-\delta)\mu}] < \frac{\mathbf{E}[e^{-tX}]}{e^{-t(1-\delta)\mu}}$$

Following the same steps as done in the upper bound proof, the numerator becomes:

$$\mathbf{E}[e^{-tX}] = \mathbf{E}[e^{-t\sum_{i=1}^{n} X_i}] = \mathbf{E}[\prod_{i=1}^{n} e^{-tX_i}] = \prod_{i=1}^{n} \mathbf{E}[e^{-tX_i}]$$

The random variable e^{-tX_i} has the value of e^{-t} with probability p_i and the value 1 with probability $1-p_i$. Therefore, the expected value is $p_i(e^{-t})+(1-p_i)(1)$ making the numerator:

$$\prod_{i=1}^{n}(p_ie^{-t}+1-p_i)=\prod_{i=1}^{n}\left(1+p_i(e^{-t}-1)\right)$$

Using the inequality $1+x<e^{-x}$ with $x=p_i(e^{-t}-1)$, we get the right hand side to be:

$$\mathbf{Pr}[e^{-tX}>e^{-t(1-\delta)\mu}]<\frac{\prod_{i=1}^{n}e^{p_i(e^{-t}-1)}}{e^{-t(1-\delta)\mu}}$$

$$\mathbf{Pr}[e^{-tX}>e^{-t(1-\delta)\mu}]<\frac{e^{\sum_{i=1}^{n}p_i(e^{-t}-1)}}{e^{-t(1-\delta)\mu}}$$

$$\mathbf{Pr}[e^{-tX}>e^{-t(1-\delta)\mu}]<\frac{e^{(e^{-t}-1)\mu}}{e^{-t(1-\delta)\mu}}$$

Now, we choose t to be: $t=\ln\left(\frac{1}{1-\delta}\right)$

Substituting t in to the inequality gives us:

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

This can be simplified since for $\delta\in(0,1]$, we note that $(1-\delta)^{(1-\delta)}>e^{(-\delta+\frac{\delta^2}{2})}$. And by using the McLaurin expansion for $\ln(1-\delta)$, we get the result of the theorem.

□

Definition: 3.3
The lower tail bound function for the sum of Poisson trials is

$$\mathbf{Pr}[X<(1-\delta)\mu]=F^{-}(\mu,\delta)\stackrel{\Delta}{=}e^{\frac{-\mu\delta^2}{2}}$$

Example (1): Suppose that sports commentators say that WVU wins each game that they play with a probability of $\frac{1}{3}$. Assuming that the winning the games are independent of each other, what is the upper bound on the probability that they will have a winning season in a season that lasts n games?

We can characterize this event by the following: $\mathbf{Pr}\left[X>\frac{n}{2}\right]$, that is, what is the probability that more than half the games are won.

Using the fact that the probability of winning one game is $\frac{1}{3}$, we can calculate the expected value of winning n games to be

$$\mathbf{E}[X]=\frac{n}{3}$$

We now need to rewrite the probability in terms of δ and μ. In other words, we need to rewrite $X > \frac{n}{2}$ to be in the form of $X > (1+\delta)\mu$. So, we just do simple algebra:

$$\begin{aligned}&(1+\delta)\mu = \tfrac{n}{2}\\&(1+\delta)\tfrac{n}{3} = \tfrac{n}{2}\\&1+\delta = \tfrac{3}{2}\\&\delta = \tfrac{1}{2}\end{aligned}$$

So, now we have the parameters needed and can now plug into the upper bound formula:

$$\mathbf{Pr}[X > \frac{n}{2}] < F^{+}(\frac{n}{3}, \frac{1}{2}) < \left[\frac{e^{\frac{1}{2}}}{\left(1+\frac{1}{2}\right)^{\left(1+\frac{1}{2}\right)}}\right]^{\frac{n}{3}} < (0.965)^{n}$$

Is this good or bad? As n increases, this number drops off very rapidly. That means that the longer WVU plays the more likely they will have a winning season.

Example (2): Suppose WVU now has a probability of winning each game to be $\frac{3}{4}$. What is the probability that they suffer a losing season assuming that the outcomes of their games are independent of one another?

We can use the formula for finding the lower bound by again finding δ and μ.

$$\begin{aligned}&(1-\delta)\mu = \tfrac{n}{2}\\&(1-\delta)\tfrac{3n}{4} = \tfrac{n}{2}\\&1-\delta = \tfrac{2}{3}\\&\delta = \tfrac{1}{3}\end{aligned}$$

$$\mathbf{Pr}[X < \frac{n}{2}] < F^{-}(\frac{3n}{4}, \frac{1}{3}) < e^{\frac{-\left(\frac{3n}{4}\right)\left(\frac{1}{3}\right)^{2}}{2}} < (0.9592)^{n}$$

This probability is also exponentially small in n.

4 Deviation by δ

To answer the 2^{nd} type of question, for example, "how large does δ need to be such that $\mathbf{Pr}[X > (1+\delta)\mu] \leq \frac{1}{\varepsilon}$?", we can use the following definitions:

Definition: 4.1 *For any positive μ and ε, $\Delta^{+}(\mu, \varepsilon)$ is the value of δ that satisfies*

$$F^{+}(\mu, \Delta^{+}(\mu, \varepsilon)) = \varepsilon$$

Definition: 4.2 *For any positive μ and ε, $\Delta^-(\mu, \varepsilon)$ is the value of δ that satisfies*

$$F^-(\mu, \Delta^-(\mu, \varepsilon)) = \varepsilon$$

$$\Delta^-(\mu, \varepsilon) = \sqrt{\frac{2\ln\frac{1}{\varepsilon}}{\mu}}$$

These definitions show that a deviation from the upper and lower bounds are irrespective of the values of n and the p_i's. It shows that the deviation is kept below ε.

Example (3): Suppose that we want to find a δ such that with probability $p_i = \frac{3}{4}$, the probability $\mathbf{Pr}[X < (1-\delta)\mu] \leq n^{-5}$.

We can plug into the $\Delta^-(\mu, \varepsilon)$ formula for the result.

$$\Delta^-(\frac{3n}{4}, \frac{1}{n^5}) = \sqrt{\frac{10\ln n}{\frac{3n}{4}}} = \sqrt{\frac{13.333\ln n}{n}}$$

This is very small, so it means that we don't need to go very far away from the expectation.

Example (4): Suppose that we want to find a δ such that with probability $p_i = \frac{3}{4}$, the probability $\mathbf{Pr}[X < (1-\delta)\mu] \leq e^{-1.5n}$.

We can plug into the $\Delta^-(\mu, \varepsilon)$ formula for the result.

$$\Delta^-(\frac{3n}{4}, \frac{1}{e^{1.5n}}) = \sqrt{\frac{3n}{\frac{3n}{4}}} = 2$$

This means nothing since for deviations below the expectation, values of δ bigger than 1 cannot occur.

Example (5):
Suppose we have n balls and n bins. What is the number m such that the first bin has m balls in it? Each ball is thrown into a bin independently and at random.

We can express this by letting Y be a random variable denoting the number of balls in the 1^{st} bin.

We want to find m such that $\mathbf{Pr}[Y > m] \leq \frac{1}{n^2}$.

If we consider the Bernoulli trails of whether or not the i^{th} ball falls into the first bin, we have each of these events occurring with $p_i = 1$. Thus, the expected value is the sum of each of the probabilities, so, $\mu = 1$.

We can take the event and express m in terms of $m = \Delta^+$ as follows:

$$\begin{aligned} m &= (1+\delta)\mu \\ &= (1 + \big(\Delta^+(1, \tfrac{1}{n^2})\big)) \end{aligned}$$

5 Set Balancing

The Set Balancing problem can be characterized by the following:

Given an n x n matrix A, all of whose entries are either 0 or 1, find a common vector, $\vec{b}$, all of whose entries are –1 or 1, such that $\left\|A\vec{b}\right\|_\infty$ is minimized.

5.1 Norm of a vector

Definition: 5.1 *The norm of a vector,$\|\vec{x}\|$, is a non-negative real number, such that the following conditions are satisfied:*

$$\|\vec{x}\| \geq 0, \quad \|\vec{x}\| = 0 \ \textit{iff}\ \vec{x} = \vec{0}$$

$$\|c\vec{x}\| = |c| \cdot \|\vec{x}\|$$

$$\|\vec{x} + \vec{y}\| = \|\vec{x}\| + \|\vec{y}\|$$

$$\|\vec{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |\vec{x}_i|^p}$$

If $p = 2$, the norm is the Euclidean norm (our well known formula for distance).

If $p = \infty$, the norm is the $\max_{1 \leq i \leq n} |\vec{x}_i|$.

5.2 Set Balancing Solution

We can analyze the set balancing solution with the following:

Suppose we choose a vector $\vec{b}$ uniformly and at random with probability $= \frac{1}{2}$ from {-1,+1}.
This choice completely ignores the given matrix A.

If the above statement is our algorithm, we can now ask a few questions:

1. What is the expected value of picking any row of A with our randomly chosen $\vec{b}$?
   $\mathbf{E}[a_i b] = \sum_{j=1}^{n} \vec{a}_{ij}\vec{b}_j = 0$
2. How much away from 0 does the expectation deviate?
   We can show that $\mathbf{Pr}[\vec{a}_i\vec{b} < -4\sqrt{n \ln n}] \leq \frac{1}{n^2}$, by using F^- to get the right hand side. Similarly, we can show that $\mathbf{Pr}[\vec{a}_i\vec{b} > 4\sqrt{n \ln n}] \leq \frac{1}{n^2}$.

3. Suppose we say that a bad event occurs if the absolute value of the inner product of the i^{th} row of A with $\vec{b}$ exceeds $4\sqrt{n \ln n}$. What is the probability that a bad event occurs for the i^{th} row?

   Since both events are disjoint and mutually exclusive, we can sum the probabilities to show that $\mathbf{Pr}[\left|\vec{a}_i\vec{b}\right| > 4\sqrt{n \ln n}] \leq \frac{2}{n^2}$.

4. What is the probability that a bad event occurs?
   There are n possible bad events for each row in A. Since any one of them can occur at most $\frac{2}{n^2}$. The union of these bad events is no more than the sum of their probabilities, which is $\frac{2}{n}$.

Result: 5.1 *The probability that we find a vector, $\vec{b}$, for which $\left\|A\vec{b}\right\|_\infty$ is minimized is* $1 - \frac{2}{n^2}$