CHAPTER 4

Tail Inequalities

In this chapter we present some general bounds on the tail of the distribution of the sum of independent random variables, with some extensions to the case of dependent or correlated random variables. These bounds are derived via the use of moment generating functions and result in "Chernoff-type" or "exponential" tail bounds. These Chernoff bounds are applied to the analysis of algorithms for global wiring in chips and routing in parallel communications networks. For applications in which the random variables of interest cannot be modeled as sums of independent random variables, martingales are a powerful probabilistic tool for bounding the divergence of a random variable from its expected value. We introduce the concept of conditional expectation as a random variable, and use this to develop a simplified definition of martingales. Using measure-theoretic ideas, we provide a more general description of martingales. Finally, we present an exponential tail bound for martingales and apply it to the analysis of an occupancy problem.

4.1. The Chernoff Bound

In Chapter 3 we initiated the study of techniques for bounding the probability that a random variable deviates far from its expectation. In this chapter we focus on techniques for obtaining considerably sharper bounds on such tail probabilities.

The random variables we will be most concerned with are sums of independent Bernoulli trials; for example, the outcomes of tosses of a coin. In designing and analyzing randomized algorithms in various settings, it is extremely useful to have an understanding of the behavior of this sum. Let $X_1, \ldots, X_n$ be independent Bernoulli trials such that, for $1 \leq i \leq n$, $\mathbf{Pr}[X_i = 1] = p$ and $\mathbf{Pr}[X_i = 0] = 1 - p$. Let $X = \sum_{i=1}^{n} X_i$; then X is said to have the *binomial distribution*. More generally, let $X_1, \ldots, X_n$ be independent coin tosses such that, for $1 \leq i \leq n$, $\mathbf{Pr}[X_i = 1] = p_i$ and $\mathbf{Pr}[X_i = 0] = 1 - p_i$. Such coin tosses are

referred to as *Poisson trials.* Our discussion below will focus on the random variable $X = \sum_{i=1}^{n} X_i$, where the X_i are Poisson trials. Of course, all our bounds apply to the special case when the X_i are Bernoulli trials with identical probabilities, so that X has the binomial distribution.

We consider two questions regarding the deviation of X from its expectation $\mu = \sum_{i=1}^{n} p_i$. For a real number $\delta > 0$, we might ask "what is the probability that X exceeds $(1 + \delta)\mu$?" We thus seek a bound on the *tail probability* of the sum of Poisson trials. An answer to this type of question is useful in *analyzing* an algorithm, showing that the chance it fails to achieve a certain performance is small. We face a different type of question in *designing* an algorithm: how large must δ be in order that the tail probability is less than a prescribed value ϵ?

Tight answers to such questions come from a technique known as the *Chernoff bound.* This technique proves to be extremely useful in designing and analyzing randomized algorithms. We focus on the Chernoff bound on the sum of independent Poisson trials.

For a random variable X, the quantity $\mathbf{E}[e^{tX}]$ is called the *moment generating function* of X. This is because $\mathbf{E}[e^{tX}]$ can be written as a power-series with terms of the form $t^k\mathbf{E}[X^k]/k!$, and $\mathbf{E}[X^k]$ is the kth *moment* of X for any positive integer k. The basic idea behind the Chernoff bound technique is to take the moment generating function of X and apply the Markov inequality to it. The sum of independent random variables appears in the exponent, and this turns into the product of random variables whose expectation we then bound.

Theorem 4.1: *Let $X_1, X_2, \ldots, X_n$ be independent Poisson trials such that, for $1 \le i \le n$, $\mathbf{Pr}[X_i = 1] = p_i$, where $0 < p_i < 1$. Then, for $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbf{E}[X] = \sum_{i=1}^{n} p_i$, and any $\delta > 0$,*

$$\mathbf{Pr}[X > (1+\delta)\mu] < \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}. \tag{4.1}$$

PROOF: For any positive real t,

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

Applying the Markov inequality to the right-hand side, we have

$$\mathbf{Pr}[X > (1+\delta)\mu] < \frac{\mathbf{E}[\exp(tX)]}{\exp(t(1+\delta)\mu)}. \tag{4.2}$$

Notice that the inequality is strict: this stems from our assumption that the p_i are not all identically 0 or 1, so that X assumes more than one value. The reader may wish to recall the proof of the Markov inequality to see this.

We bound the right-hand side by observing that

$$\mathbf{E}[\exp(tX)] = \mathbf{E}[\exp(t\sum_{i=1}^{n} X_i)] = \mathbf{E}[\prod_{i=1}^{n} \exp(tX_i)].$$

Since the X_i are independent, the random variables $\exp(tX_i)$ are also independent. It follows that $\mathbf{E}[\prod_{i=1}^{n} \exp(tX_i)] = \prod_{i=1}^{n} \mathbf{E}[\exp(tX_i)]$. Using these facts in (4.2) gives

$$\mathbf{Pr}[X > (1+\delta)\mu] < \frac{\prod_{i=1}^{n} \mathbf{E}[\exp(tX_i)]}{\exp(t(1+\delta)\mu)}. \tag{4.3}$$

The random variable e^{tX_i} assumes the value e^t with probability p_i, and the value 1 with probability $1 - p_i$. Computing $\mathbf{E}[e^{tX_i}]$ from these observations, we have that

$$\begin{aligned} \mathbf{Pr}[X > (1+\delta)\mu] &< \frac{\prod_{i=1}^{n} [p_i e^t + 1 - p_i]}{\exp(t(1+\delta)\mu)} \\ &= \frac{\prod_{i=1}^{n} [1 + p_i(e^t - 1)]}{\exp(t(1+\delta)\mu)}. \end{aligned} \tag{4.4}$$

Now we use the inequality $1 + x < e^x$ with $x = p_i(e^t - 1)$, to obtain

$$\begin{aligned} \mathbf{Pr}[X > (1+\delta)\mu] &< \frac{\prod_{i=1}^{n} \exp(p_i(e^t - 1))}{\exp(t(1+\delta)\mu)} \\ &= \frac{\exp(\sum_{i=1}^{n} p_i(e^t - 1))}{\exp(t(1+\delta)\mu)} \\ &= \frac{\exp((e^t - 1)\mu)}{\exp(t(1+\delta)\mu)}. \end{aligned} \tag{4.5}$$

Observe that all of the above has been proved for any positive real t; we are now free to choose a particular value for t that yields the best possible bound. For this, we differentiate the last expression with respect to t and set to zero; solving for t now yields $t = \ln(1+\delta)$, which is positive for $\delta > 0$. Substituting this value for t, we obtain our theorem. □

There were three main ingredients in the above proof:

1. We studied the random variable e^{tX} rather than X.
2. The expectation of the product of the e^{tX_i} turns into the product of their expectations owing to independence.
3. We pick a value of t to obtain the best possible upper bound – indeed, we choose a value of t that depends on the deviation δ.

These ingredients are generic and do not hinge on the particular case of the sum of Poisson trials. For example, Problem 4.4 is concerned with applying this technique to the sum of geometrically distributed random variables.

For succinctness in what follows, we define an upper tail bound function for the sum of Poisson trials.

► **Definition 4.1:** $F^+(\mu, \delta) \triangleq \left[e^{\delta}/(1+\delta)^{(1+\delta)}\right]^{\mu}$.

► **Example 4.1:** The Arkansas Aardvarks win each game they play with probability 1/3. Assuming that the outcomes of the games are independent, derive an upper

bound on the probability that they have a winning season in a season lasting n games.

Let X_i be 1 if the Aardvarks win the ith game and 0 otherwise; let $Y_n = \sum_{i=1}^{n} X_i$. Applying Theorem 4.1 to Y_n, we find that $\mathbf{Pr}[Y_n > n/2] < F^+(n/3, 1/2) < (0.965)^n$. Thus, the probability that the Aardvarks have a winning season in n games is exponentially small in n, suggesting that the longer they play the more likely it is that their true colors show through.

The reader should verify that the term within the brackets in $F^+(\mu, \delta)$ is always strictly less than 1. Since the power μ is always positive, we will always get an upper bound that is less than 1.

The right-hand side of (4.1) is difficult to interpret, especially since we will require answers to questions such as "how large need δ be in order that $\mathbf{Pr}[X > (1+\delta)\mu]$ is at most 0.01?" We will presently work on simplifying it. But first, we consider deviations of X *below* its expectation μ.

Theorem 4.2: *Let $X_1, X_2, \ldots, X_n$ be independent Poisson trials such that, for $1 \le i \le n$, $\mathbf{Pr}[X_i = 1] = p_i$, where $0 < p_i < 1$. Then, for $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbf{E}[X] = \sum_{i=1}^{n} p_i$, and $0 < \delta \le 1$,*

$$\mathbf{Pr}[X < (1-\delta)\mu] < \exp(-\mu\delta^2/2). \tag{4.6}$$

PROOF: The proof is very similar to the proof for the upper tail we saw in Theorem 4.1. As before,

$$\begin{aligned} \mathbf{Pr}[X < (1-\delta)\mu] &= \mathbf{Pr}[-X > -(1-\delta)\mu] \\ &= \mathbf{Pr}[\exp(-tX) > \exp(-t(1-\delta)\mu)], \end{aligned}$$

for any positive real t. Applying the Markov inequality and proceeding as in equations (4.2–4.3), we obtain that

$$\mathbf{Pr}[X < (1-\delta)\mu] < \frac{\prod_{i=1}^{n} \mathbf{E}[\exp(-tX_i)]}{\exp(-t(1-\delta)\mu)}.$$

Computing $\mathbf{E}[\exp(-tX_i)]$ and proceeding as in equations (4.4–4.5),

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

This time, we let $t = \ln(1/(1-\delta))$ to obtain that

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

We simplify this by noting that for $\delta \in (0, 1]$,

$$(1-\delta)^{1-\delta} > \exp(-\delta + \delta^2/2),$$

using the McLaurin expansion for $\ln(1-\delta)$. This yields the desired result. □

We define the lower tail bound function for the sum of Poisson trials as follows.

► **Definition 4.2:** $F^-(\mu, \delta) \triangleq \exp\left(\frac{-\mu\delta^2}{2}\right)$.

It is immediate that $F^-(\mu, \delta)$ is always less than 1 for positive μ and δ. Note two differences between the proofs of Theorems 4.1 and 4.2. First, we directly apply the basic Chernoff technique to the random variable $-X$ rather than apply Theorem 4.1 to $Y = n - X$ (a plausible option, which leads, however, to a slightly weaker bound than the one derived below). Second, the form of the McLaurin expansion for $\ln(1 - \delta)$ allows us to obtain a "cleaner" closed form here, whereas the McLaurin expansion for $\ln(1 + \delta)$ did not permit this in Theorem 4.1.

► **Example 4.2:** The Arkansas Aardvarks hire a new coach, and critics revise their estimates of the probability of their winning each game to 0.75. What is the probability that the Aardvarks suffer a losing season assuming the critics are right and the outcomes of their games are independent of one another?

Setting up the random variable Y_n as before, we find that $\mathbf{Pr}[Y_n < n/2] < F^-(0.75n, 1/3)$, which evaluates to $< (0.9592)^n$. Thus, this probability is also exponentially small in n.

The bounds in Theorems 4.1 and 4.2 do not depend on n, but only on μ and δ. These bounds do not distinguish, for instance, between 1000 trials each with $p_i = 0.02$ and 100 each with $p_i = 0.2$, even though the distributions of X are different in the two cases. Thus, even if the actual tail probabilities are different in these cases, our estimates are the same in both cases.

We make the following definitions to facilitate our second kind of question, i.e.,"how large need δ be for $\mathbf{Pr}[X > (1 + \delta)\mu]$ to be less than ϵ?"

► **Definition 4.3:** For any positive μ and ϵ, $\Delta^+(\mu, \epsilon)$ is that value of δ that satisfies

$$F^+(\mu, \Delta^+(\mu, \epsilon)) = \epsilon. \tag{4.7}$$

Similarly, $\Delta^-(\mu, \epsilon)$ is that value of δ that satisfies

$$F^-(\mu, \Delta^-(\mu, \epsilon)) = \epsilon. \tag{4.8}$$

In other words, a deviation of $\delta = \Delta^+(\mu, \epsilon)$ suffices to keep $\mathbf{Pr}[X > (1 + \delta)\mu]$ below ϵ, irrespective of the values of n and the p_i's.

A nice feature of the bound in Theorem 4.2 is the convenient form of the right-hand side: it is easy to derive $\Delta^-(\mu, \epsilon)$ explicitly. Equating the right-hand side of (4.6) to ϵ yields

$$\Delta^-(\mu, \epsilon) = \sqrt{\frac{2 \ln 1/\epsilon}{\mu}}. \tag{4.9}$$

► **Example 4.3:** Suppose that $p_i = 0.75$. How large must δ be so that $\mathbf{Pr}[X < (1-\delta)\mu]$ is less than n^{-5}? Using (4.9), we find that the value of δ that suffices for ϵ

to be less than n^{-5} is

$$\Delta^-(0.75n, n^{-5}) = \sqrt{\frac{10 \ln n}{0.75n}}.$$

Thus, to obtain a tail probability that is inversely polynomial in n, we need only go slightly away from the expectation - in this case out to $\delta = \sqrt{(13.333 \ln n)/n}$.

What if we wanted that $\mathbf{Pr}[X < (1-\delta)\mu]$ be less than $e^{-1.5n}$? Using (4.9), we find that for $\epsilon = e^{-1.5n}$,

$$\Delta^-(0.75n, e^{-1.5n}) = \sqrt{\frac{3n}{0.75n}} = 2,$$

which tells us nothing (for deviations below the expectation, values of δ bigger than 1 cannot occur).

We return to the simplification of (4.1) to obtain tractable estimates for $\Delta^+(\mu, \epsilon)$.

Exercise 4.1: Prove that

$$F^+(\mu, \delta) < [e/(1+\delta)]^{(1+\delta)\mu}. \tag{4.10}$$

Hence infer that if $\delta > 2e - 1$,

$$F^+(\mu, \delta) < 2^{-(1+\delta)\mu}.$$

Exercise 4.1 gives us a simple form for $F^+(\mu, \delta)$ when δ is "large." For such deviations, we have the bound

$$\Delta^+(\mu, \epsilon) < \frac{\log_2 1/\epsilon}{\mu} - 1. \tag{4.11}$$

We now present the following simplification of $F^+(\mu, \delta)$ for δ in a restricted range $(0, U]$. A pointer to the proof is given in the Notes section.

Theorem 4.3: *For* $0 < \delta \leq U$,

$$F^+(\mu, \delta) \leq \exp(-c(U)\mu\delta^2),$$

where $c(U) = [(1+U)\ln(1+U) - U]/U^2$.

For $U = 2e - 1$, this simplifies to $F^+(\mu, \delta) < \exp(-\mu\delta^2/4)$. Consequently, provided $\delta \leq 2e - 1$, we can use the estimate

$$\Delta^+(\mu, \epsilon) < \sqrt{\frac{4 \ln 1/\epsilon}{\mu}}. \tag{4.12}$$

Thus, between Theorem 4.3 and Exercise 4.1, we have bounds on $\Delta^+(\mu, \epsilon)$; however, we require some idea of the correct value of $\Delta^+(\mu, \epsilon)$ before deciding

which of these forms to use. Moreover, the result of Exercise 4.1 may be slack for some values of μ and ϵ, as in the following example. This example uses Chernoff bounds to approach the occupancy problem considered in Section 3.1.

► **Example 4.4:** Consider throwing n balls uniformly and independently into n bins. Let the random variable Y_1 denote the number of balls that fall into the first bin. We wish to determine a quantity m such that $\mathbf{Pr}[Y_1 > m] \leq 1/n^2$.

Consider the Bernoulli trials indicating whether or not the ith ball falls into the first bin. Each of the p_i's is thus $1/n$. It follows that $\mu = 1$; the number m we seek is $1 + \Delta^+(1, 1/n^2)$. Guessing that $\Delta^+(1, 1/n^2)$ is larger than $2e$, we use the result in (4.11) to obtain $\Delta^+(1, 1/n^2) < 2\log_2 n - 1$.

Unfortunately, this is not the tightest possible answer in this case. Returning to (4.1), we can apply it with $\delta \approx (1.5 \ln n)/\ln\ln n$ and simplify to obtain $F^+(\mu, \delta)$ less than n^{-2}, so that our original estimate of $2\log_2 n - 1$ was asymptotically an overestimate.

A good rule of thumb from examples like this is: for ϵ of the order of n^{-c} (a value arising often in algorithmic applications), estimates such as (4.11) and (4.12) are satisfactory provided μ is $\Omega(\log n)$; when μ is smaller, we must return to (4.1) in order to obtain the tightest possible estimate.

► **Example 4.5 (Set Balancing):** This problem is known variously as *set-balancing*, or *two-coloring a family of vectors*. Given an $n \times n$ matrix A all of whose entries are 0 or 1, find a column vector $\boldsymbol{b} \in \{-1, +1\}^n$ minimizing $||A\boldsymbol{b}||_\infty$.

Consider the following algorithm for choosing $\boldsymbol{b}$: each entry of $\boldsymbol{b}$ is independently and equiprobably chosen from $\{-1, +1\}$. Note that this choice ignores the given matrix A. Clearly the inner product of any row of A with our randomly chosen $\boldsymbol{b}$ has expectation 0. We now study the deviation of this inner product from 0.

Consider the ith row of A. Applying (4.9), the probability that the inner product of this row with $\boldsymbol{b}$ is bounded by $-4\sqrt{n \ln n}$ is less than n^{-2}. An identical argument shows that the probability that the inner product of this row with $\boldsymbol{b}$ exceeds $4\sqrt{n \ln n}$ is less than n^{-2}. Thus, the probability that the absolute value of the inner product exceeds $4\sqrt{n \ln n}$ is less than $2n^{-2}$.

Let us say that the ith *bad event* occurs if the absolute value of the inner product of the ith row of A with $\boldsymbol{b}$ exceeds $4\sqrt{n \ln n}$. There are n possible bad events, one for each row, and the argument of the previous paragraph shows that the probability that any of them occurs is at most $2n^{-2}$. The probability of the union of the bad events is no more than the sum of their probabilities, which is $2/n$. In other words, with probability at least $1 - 2/n$, we find a vector $\boldsymbol{b}$ for which $||A\boldsymbol{b}||_\infty \leq 4\sqrt{n \ln n}$.