Appendix A
Bounding of Large Deviations

A.1 CHERNOFF BOUNDS

We give here some basic bounds on large deviations that are useful when employing the probabilistic method. Our treatment is self-contained. Most of the results may be found in, or immediately derived from, the seminal paper of Chernoff (1952). While we are guided by asymptotic considerations the inequalities are proved for all values of the parameters in the specified region. The first result, while specialized, contains basic ideas found throughout the appendix.

Theorem A.1.1 *Let $X_i, 1 \leq i \leq n$, be mutually independent random variables with*

$$\Pr[X_i = +1] = \Pr[X_i = -1] = \frac{1}{2}$$

and set, following the usual convention,

$$S_n = X_1 + \cdots + X_n .$$

Let $a > 0$. Then

$$\Pr[S_n > a] < e^{-a^2/2n} .$$

We require Markov's Inequality, which states: Suppose that Y is an arbitrary nonnegative random variable, $\alpha > 0$. Then

$$\Pr[Y > \alpha \mathrm{E}[Y]] < \frac{1}{\alpha}.$$

Proof. Fix n, a and let, for the moment, $\lambda > 0$ be arbitrary. For $1 \le i \le n$,

$$\mathrm{E}\left[e^{\lambda X_i}\right] = \frac{e^{\lambda} + e^{-\lambda}}{2} = \cosh(\lambda).$$

We require the inequality $\cosh(\lambda) \le e^{\lambda^2/2}$, valid for all $\lambda > 0$, the special case $\alpha = 0$ of Lemma A.1.5 below. (The inequality may be shown more easily by comparing the Taylor series of the two functions termwise.)

$$e^{\lambda S_n} = \prod_{i=1}^{n} e^{\lambda X_i}.$$

Since the X_i are mutually independent so are the $e^{\lambda X_i}$; expectations multiply and

$$\mathrm{E}\left[e^{\lambda S_n}\right] = \prod_{i=1}^{n} \mathrm{E}\left[e^{\lambda X_i}\right] = \cosh^n(\lambda) < e^{\lambda^2 n/2}.$$

We note that $S_n > a$ if and only if $e^{\lambda S_n} > e^{\lambda a}$ and apply Markov's Inequality so that

$$\Pr[S_n > a] = \Pr\left[e^{\lambda S_n} > e^{\lambda a}\right] < \mathrm{E}\left[e^{\lambda S_n}\right]/e^{\lambda a} \le e^{\lambda^2 n/2 - \lambda a}.$$

We set $\lambda = a/n$ to optimize the inequality, $\Pr[S_n > a] < e^{-a^2/2n}$ as claimed. ∎

By symmetry we immediately have the following.

Corollary A.1.2 *Under the assumptions of Theorem A.1.1,*

$$\Pr[|S_n| > a] < 2e^{-a^2/2n}.$$

The proof of Theorem A.1.1 illustrates the basic idea of the Chernoff bounds. We wish to bound $\Pr[X > a]$ for some random variable X. For any positive λ we bound

$$\Pr[X > a] = \Pr\left[e^{\lambda X} > e^{\lambda a}\right] \le \mathrm{E}\left[e^{\lambda X}\right] e^{-\lambda a}. \tag{A.1}$$

The core idea of the Chernoff bounds is to select that λ that minimizes $\mathrm{E}\left[e^{\lambda X}\right] e^{-\lambda a}$. The art to the Chernoff bounds is to select a λ that is reasonably close to optimal and easy to work with, yielding upper bounds on $\Pr[X > a]$ that are, one hopes, good enough for our purposes. Bounds on $\Pr[X < a]$ are similar. For any positive λ we bound

$$\Pr[X < a] = \Pr\left[e^{-\lambda X} > e^{-\lambda a}\right] \le \mathrm{E}\left[e^{-\lambda X}\right] e^{\lambda a}.$$

Chernoff bound arguments tend to be cleaner when $\mathrm{E}[X] = 0$. A simple translation, replacing X by $X - \mu$ where $\mu = \mathrm{E}[X]$, is often quite helpful.

It is instructive to examine the case when N is the standard normal distribution and a is positive. In this instance $\mathrm{E}\left[e^{\lambda N}\right] = e^{\lambda^2/2}$ and so

$$\Pr[N > a] = \Pr\left[e^{\lambda N} > e^{\lambda a}\right] \le \mathrm{E}\left[e^{\lambda N}\right] e^{-\lambda a} = e^{\lambda^2/2 - \lambda a}.$$

Elementary calculus leads to the optimal choice $\lambda = a$ so that

$$\Pr[N > a] < e^{-a^2/2}.$$

This compares well, as $a \to \infty$, with the actual asymptotics

$$\Pr[N > a] = \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{-t^2/2}\, dt \sim \frac{1}{\sqrt{2\pi} a} e^{-a^2/2}.$$

Results with N being normal with mean μ and variance σ^2 are similarly good. This explains, to some extent, the efficacy of the Chernoff bounds. When a random variable X is "roughly" normal the Chernoff bounds on $\Pr[X > a]$ should be quite close to the actual values for a large. In practice, however, precise calculations of $\mathrm{E}\left[e^{\lambda X}\right]$ can be difficult or impossible to achieve and there can be considerable art in finding approximations for $\mathrm{E}\left[e^{\lambda X}\right]$ that will allow for good bounds on $\Pr[X > a]$.

Many of our remaining results will deal with distributions X of the following prescribed type.

Assumptions A.1.3

- $p_1, \ldots, p_n \in [0, 1]$,
- $X_1, \ldots, X_n$ *are mutually independent with* $\Pr[X_i = 1 - p_i] = p_i$ *and* $\Pr[X_i = -p_i] = 1 - p_i$,
- $p = (p_1 + \cdots + p_n)/n$ *and* $X = X_1 + \cdots + X_n$.

Remark. Clearly $\mathrm{E}[X] = \mathrm{E}[X_i] = 0$. When all $p_i = 1/2$, X has distribution $S_n/2$. When all $p_i = p$, X has distribution $B(n,p) - np$, where $B(n,p)$ is the usual binomial distribution.

Theorem A.1.4 *Under Assumptions A.1.3 and with* $a > 0$,

$$\Pr[X > a] < e^{-2a^2/n}.$$

Lemma A.1.5 *For all reals* α, β *with* $|\alpha| \le 1$,

$$\cosh(\beta) + \alpha \sinh(\beta) \le e^{\beta^2/2 + \alpha\beta}.$$

Proof. This is immediate if $\alpha = \pm 1$ or $|\beta| \ge 100$. If the lemma were false the function $f(\alpha, \beta) = \cosh(\beta) + \alpha \sinh(\beta) - e^{\beta^2/2 + \alpha\beta}$ would assume a positive global maximum in the interior of the rectangle $R = \{(\alpha, \beta) : |\alpha| \le 1, |\beta| \le 100\}$.

Setting partial derivatives equal to zero we find

$$\begin{aligned}\sinh(\beta) + \alpha\cosh(\beta) &= (\alpha+\beta)e^{\beta^2/2+\alpha\beta}, \\ \sinh(\beta) &= \beta e^{\beta^2/2+\alpha\beta},\end{aligned}$$

and thus $\tanh(\beta) = \beta$, which implies $\beta = 0$. But $f(\alpha, 0) = 0$ for all α, a contradiction. ■

Lemma A.1.6 *For all $\theta \in [0,1]$ and all λ, $\theta e^{\lambda(1-\theta)} + (1-\theta)e^{-\lambda\theta} \le e^{\lambda^2/8}$.*

Proof. Setting $\theta = \frac{1}{2}(1+\alpha)$ and $\lambda = 2\beta$, this lemma reduces to Lemma A.1.5. ■

Proof [Theorem A.1.4]. Let, for the moment, $\lambda > 0$ be arbitrary.

$$\mathrm{E}\left[e^{\lambda X_i}\right] = p_i e^{\lambda(1-p_i)} + (1-p_i)e^{-\lambda p_i} \le e^{\lambda^2/8}$$

by Lemma A.1.6. Then

$$\mathrm{E}\left[e^{\lambda X}\right] = \prod_{i=1}^{n} \mathrm{E}\left[e^{\lambda X_i}\right] \le e^{\lambda^2 n/8}.$$

Applying Markov's Inequality,

$$\Pr[X > a] = \Pr\left[e^{\lambda X} > e^{\lambda a}\right] < \mathrm{E}\left[e^{\lambda X}\right]/e^{\lambda a} \le e^{\lambda^2 n/8 - \lambda a}.$$

We set $\lambda = 4a/n$ to optimize the inequality: $\Pr[X > a] < e^{-2a^2/n}$, as claimed. ■

Again by symmetry we immediately have the following.

Corollary A.1.7 *Under Assumptions A.1.3 and with $a > 0$,*

$$\Pr[|X| > a] < 2e^{-2a^2/n}.$$

Under Assumptions A.1.3 with λ arbitrary,

$$\begin{aligned}\mathrm{E}\left[e^{\lambda X}\right] &= \prod_{i=1}^{n} \mathrm{E}\left[e^{\lambda X_i}\right] = \prod_{i=1}^{n}\left(p_i e^{\lambda(1-p_i)} + (1-p_i)e^{-\lambda p_i}\right) \\ &= e^{-\lambda pn}\prod_{i=1}^{n}\left(p_i e^{\lambda} + (1-p_i)\right).\end{aligned}$$

With λ fixed, the function

$$f(x) = \ln(xe^{\lambda} + 1 - x) = \ln(Bx + 1) \text{ with } B = e^{\lambda} - 1$$

is concave and hence $\sum_{i=1}^{n} f(p_i) \le nf(p)$ (Jensen's Inequality). Exponentiating both sides,

$$\prod_{i=1}^{n}\left(p_i e^{\lambda} + (1-p_i)\right) \le \left(pe^{\lambda} + (1-p)\right)^n,$$

so that we have the following.

Lemma A.1.8 *Under the Assumptions A.1.3,*

$$\mathrm{E}\left[e^{\lambda X}\right] < e^{-\lambda pn}\left(pe^{\lambda} + (1-p)\right)^n.$$

Applying this lemma with inequality (A.1) yields the following.

Theorem A.1.9 *Under the Assumptions A.1.3 and with $a > 0$,*

$$\Pr[X > a] < e^{-\lambda pn}\left(pe^{\lambda} + (1-p)\right)^n e^{-\lambda a}$$

for all $\lambda > 0$.

Remark. For given p, n, a, an optimal assignment of λ in Theorem A.1.9 is found by elementary calculus to be

$$\lambda = \ln\left[\left(\frac{1-p}{p}\right)\left(\frac{a+np}{n-(a+np)}\right)\right].$$

This value is oftentimes too cumbersome to be useful. We employ suboptimal λ to achieve more convenient results.

Setting $\lambda = \ln(1+a/pn)$ and using the fact that $(1+a/n)^n \le e^a$, Theorem A.1.9 implies the following.

Corollary A.1.10 $\Pr[X > a] < e^{a - pn\ln(1+a/pn) - a\ln(1+a/pn)}$.

To simplify further, apply the inequality $\ln(1+u) \ge u - u^2/2$, valid for all $u \ge 0$, to Corollary A.1.10 with $u = a/pn$. This gives the following.

Theorem A.1.11 $\Pr[X > a] < e^{-a^2/2pn + a^3/2(pn)^2}$.

When all $p_i = p$, X has variance $np(1-p)$. With $p = o(1)$ and $a = o(pn)$ this bound reflects the approximation of X by a normal distribution with variance $\sim np$. The bound of Theorem A.1.11 hits a minimum at $a = 2pn/3$. For $a > 2pn/3$ we have the simple bound

$$\Pr[X > a] \le \Pr[X > 2pn/3] < e^{-2pn/27}.$$

This is improved by the following.

Theorem A.1.12 *For $\beta > 1$,*

$$\Pr[X > (\beta - 1)pn] < [e^{\beta-1}\beta^{-\beta}]^{pn}.$$

Proof. Direct "plug in" to Corollary A.1.10. ■

$X + pn$ may be interpreted as the number of successes in n independent trials when the probability of success in the ith trial is p_i.

Theorem A.1.13 *Under Assumptions A.1.3 and with $a > 0$,*

$$\Pr[X < -a] < e^{-a^2/2pn}.$$

Note that one cannot simply employ "symmetry" as then the roles of p and $1 - p$ are interchanged.

Proof. Let $\lambda > 0$ be, for the moment, arbitrary. Then by the argument preceding Lemma A.1.8,

$$\mathrm{E}\left[e^{-\lambda X}\right] \le e^{\lambda pn}\left(pe^{-\lambda} + (1-p)\right)^n.$$

Thus

$$\Pr[X < -a] = \Pr\left[e^{-\lambda X} > e^{\lambda a}\right] < e^{\lambda pn}\left(pe^{-\lambda} + (1-p)\right)^n e^{-\lambda a},$$

analogous to Theorem A.1.9. We employ the inequality $1 + u \le e^u$, valid for all u, so that

$$pe^{-\lambda} + (1-p) = 1 + (e^{-\lambda} - 1)p < e^{p(e^{-\lambda}-1)}$$

and

$$\Pr[X < -a] \le e^{\lambda pn + np(e^{-\lambda}-1) - \lambda a} = e^{np(e^{-\lambda}-1+\lambda)-\lambda a}.$$

We employ the inequality

$$e^{-\lambda} \le 1 - \lambda + \lambda^2/2,$$

valid for all $\lambda > 0$. (Note: The analogous inequality $e^{\lambda} \le 1 + \lambda + \lambda^2/2$ is *not* valid for $\lambda > 0$ and so this method, when applied to $\Pr[X > a]$, requires an "error" term as the one found in Theorem A.1.11.) Now

$$\Pr[X < -a] \le e^{np\lambda^2/2 - \lambda a}.$$

Set $\lambda = a/np$ to optimize the inequality: $\Pr[X < -a] < e^{-a^2/2pn}$, as claimed. ■

For clarity the following result is often useful.

Corollary A.1.14 *Let Y be the sum of mutually independent indicator random variables, $\mu = \mathrm{E}[Y]$. For all $\epsilon > 0$,*

$$\Pr[|Y - \mu| > \epsilon\mu] < 2e^{-c_\epsilon \mu},$$

where $c_\epsilon > 0$ depends only on ϵ.

Proof. Apply Theorems A.1.12 and A.1.13 with $Y = X + pn$ and

$$c_\epsilon = \min\{-\ln(e^{\epsilon}(1+\epsilon)^{-(1+\epsilon)}), \epsilon^2/2\}.$$

■

The asymmetry between $\Pr[X < a]$ and $\Pr[X > a]$ given by Theorems A.1.12 and A.1.13 is real. The estimation of X by a normal distribution with zero mean and variance np is roughly valid for estimating $\Pr[X < a]$ for any a and for estimating $\Pr[X > a]$ while $a = o(np)$. But when a and np are comparable or when $a \gg np$ the Poisson behavior "takes over" and $\Pr[X > a]$ *cannot* be accurately estimated by using the normal distribution.

We conclude with several large deviation results involving distributions other than sums of indicator random variables.

Theorem A.1.15 *Let P have Poisson distribution with mean μ. For $\epsilon > 0$*

$$\Pr[P \le \mu(1-\epsilon)] \le e^{-\epsilon^2\mu/2},$$
$$\Pr[P \ge \mu(1+\epsilon)] \le \left[e^{\epsilon}(1+\epsilon)^{-(1+\epsilon)}\right]^{\mu}.$$

Proof. For any s

$$\Pr[P = s] = \lim_{n\to\infty} \Pr\left[B\left(n, \frac{\mu}{n}\right) = s\right].$$

Apply Theorems A.1.12 and A.1.13. ■

Theorem A.1.16 *Let X_i, $1 \le i \le n$, be mutually independent with all $\mathrm{E}[X_i] = 0$ and all $|X_i| \le 1$. Set $S = X_1 + \cdots + X_n$. Then*

$$\Pr[S > a] < e^{-a^2/2n}.$$

Proof. Set, as in the proof of Theorem A.1.1, $\lambda = a/n$. Set

$$h(x) = \frac{e^{\lambda} + e^{-\lambda}}{2} + \frac{e^{\lambda} - e^{-\lambda}}{2}x.$$

For $x \in [-1, 1]$, $e^{\lambda x} \le h(x)$. [$y = h(x)$ is the chord through the points $x = \pm 1$ of the convex curve $y = e^{\lambda x}$.] Thus

$$\mathrm{E}\left[e^{\lambda X_i}\right] \le \mathrm{E}[h(X_i)] = h(\mathrm{E}[X_i]) = h(0) = \cosh\lambda.$$

The remainder of the proof follows as in Theorem A.1.1. ■

Theorem A.1.17 *Suppose* $\mathrm{E}[X]=0$ *and no two values of* X *are ever more than one apart. Then for all* $\lambda \geq 0$,

$$\mathrm{E}\left[e^{\lambda X}\right] \leq e^{\lambda^2/8}.$$

Proof. Fix $b \in [-\frac{1}{2}, \frac{1}{2}]$ with $X \in [\frac{1}{2}(-1+b), \frac{1}{2}(+1+b)]$. Let $y = h(x)$ be the straight line intersecting the curve $y = e^{\lambda x}$ at the points $\frac{1}{2}(\pm 1 + b)$. As $e^{\lambda x}$ is a convex function, $e^{\lambda x} \leq h(x)$ for all $x \in [\frac{1}{2}(-1+b), \frac{1}{2}(+1+b)]$. Thus

$$\mathrm{E}\left[e^{\lambda X}\right] \leq \mathrm{E}\left[h(X)\right] = h\left(\mathrm{E}\left[X\right]\right) = h(0).$$

We calculate $h(0) = e^{\lambda b/2}[\cosh(\lambda/2) - b\sinh(\lambda/2)]$, which is at most $e^{\lambda^2/8}$ by Lemma A.1.5. ■

Theorem A.1.18 *Let* X_i, $1 \leq i \leq n$, *be independent random variables with each* $\mathrm{E}[X_i] = 0$ *and no two values of any* X_i *ever more than one apart. (We allow, however, values of different* X_i, X_j *to be further apart.) Set* $S = X_1 + \cdots + X_n$. *Then*

$$\Pr\left[S > a\right] < e^{-2a^2/n}.$$

Proof. $\mathrm{E}\left[e^{\lambda S}\right] = \prod_{i=1}^{n} \mathrm{E}\left[e^{\lambda X_i}\right] \leq e^{n\lambda^2/8}$ by Theorem A.1.17. Then for $\lambda \geq 0$,

$$\Pr\left[S > a\right] = \Pr\left[e^{\lambda S} \geq e^{\lambda a}\right] \leq \exp\left[\frac{n\lambda^2}{8} - \lambda a\right]$$

and we set $\lambda = 4a/n$. ■

We have been roughly guided by the notion that if X has mean zero and variance σ^2 then $\Pr[X \geq a\sigma]$ should go like $e^{-a^2/2}$. There are times when this idea is very wrong. Consider Assumptions A.1.3 with all $p_i = 1/n$ so that $X = P_n - 1$, where P_n has the binomial distribution $B(n, 1/n)$, which is asymptotically P, the Poisson distribution with mean one. Then $\mathrm{E}[X] = 0$ and $\mathrm{Var}[X] \sim 1$. For a fixed $\Pr[X = a] \to 1/e(a+1)!$, which is far bigger than $e^{-a^2/2}$. With this cautionary preamble, we give a general situation for which the notion is asymptotically correct when a is not too large.

Theorem A.1.19 *For every* $C > 0$ *and* $\varepsilon > 0$ *there exists* $\delta > 0$ *so that the following holds: Let* $\{X_i\}_{i=1}^{n}$, n *arbitrary, be independent random variables with* $\mathrm{E}[X_i] = 0$, $|X_i| \leq C$ *and* $\mathrm{Var}[X_i] = \sigma_i^2$. *Set* $X = \sum_{i=1}^{n} X_i$ *and* $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$ *so that* $\mathrm{Var}[X] = \sigma^2$. *Then for* $0 < a \leq \delta\sigma$,

$$\Pr\left[X > a\sigma\right] < e^{-\frac{1}{2}a^2(1-\varepsilon)}.$$

Proof. We set $\lambda = a/\sigma$ so that $0 \leq \lambda \leq \delta$. Then

$$\mathrm{E}\left[e^{\lambda X_i}\right] = \sum_{k=0}^{\infty} \mathrm{E}\left[\frac{\lambda^k}{k!} X_i^k\right] = 1 + \frac{\lambda^2}{2}\sigma_i^2 + \sum_{k=3}^{\infty} \frac{\lambda^k}{k!} \mathrm{E}\left[X_i^k\right].$$

As $|X_i^k| \leq C^{k-2} X_i^2$ we bound

$$\mathrm{E}\left[X_i^k\right] \leq \mathrm{E}\left[|X_i^k|\right] \leq C^{k-2}\mathrm{E}\left[X_i^2\right] = C^{k-2}\sigma_i^2.$$

For $k \geq 3$ we bound $2/k! \leq 1/(k-2)!$ so that

$$\mathrm{E}\left[e^{\lambda X_i}\right] \leq 1 + \frac{\lambda^2}{2}\sigma_i^2 \left[1 + \sum_{k=3}^{\infty} \frac{(C\lambda)^{k-2}}{(k-2)!}\right] = 1 + \frac{\lambda^2}{2}\sigma_i^2 e^{\lambda C}.$$

We choose δ to satisfy $e^{C\delta} \leq 1 + \varepsilon$. As $\lambda \leq \delta$,

$$\mathrm{E}\left[e^{\lambda X_i}\right] \leq 1 + \frac{\lambda^2}{2}\sigma_i^2(1+\varepsilon) < \exp\left[\frac{\lambda^2}{2}\sigma_i^2(1+\varepsilon)\right].$$

This inequality has held for all X_i so

$$\mathrm{E}\left[e^{\lambda X}\right] = \prod_{i=1}^{n} \mathrm{E}\left[e^{\lambda X_i}\right] < \exp\left[\frac{\lambda^2}{2}\sigma^2(1+\varepsilon)\right]$$

and

$$\Pr\left[X > a\sigma\right] \leq \mathrm{E}\left[e^{\lambda X}\right] e^{-\lambda a\sigma} < e^{-\frac{1}{2}a^2(1-\varepsilon)}.$$

■

A.2 LOWER BOUNDS

The Chernoff bounds of the previous section give upper bounds for $\Pr[X > a]$ by examining one value (albeit, the right one!) of the Laplace transform $\mathrm{E}\left[e^{\lambda X}\right]$. Here we use three values of the Laplace transform to give lower bounds for $\Pr[X > a]$. We shall set

$$f(\lambda) = \mathrm{E}\left[e^{\lambda X}\right],$$
$$g_a(\lambda) = f(\lambda)e^{-\lambda a}.$$

With this notation $\Pr[X > a] \leq g_a(\lambda)$ and the Chernoff bound is achieved by taking that λ minimizing $g_a(\lambda)$. For any positive u and ε,

$$X \geq a + u \quad \Rightarrow \quad \lambda X \leq (\lambda + \varepsilon)X - \varepsilon a - \varepsilon u,$$
$$X \leq a - u \quad \Rightarrow \quad \lambda X \leq (\lambda - \varepsilon)X + \varepsilon a - \varepsilon u,$$

so that

$$\begin{aligned} \mathrm{E}\left[e^{\lambda X}\chi(X \geq a+u)\right] &\leq f(\lambda+\varepsilon)e^{-\varepsilon a}e^{-\varepsilon u}, \\ \mathrm{E}\left[e^{\lambda X}\chi(X \leq a-u)\right] &\leq f(\lambda-\varepsilon)e^{\varepsilon a}e^{-\varepsilon u}. \end{aligned}$$

Subtracting these from $\mathrm{E}\left[e^{\lambda X}\right]$ yields

$$\mathrm{E}\left[e^{\lambda X}\chi(|X-a|<u)\right] \geq f(\lambda) - e^{-\varepsilon u}[f(\lambda+\varepsilon)e^{-\varepsilon a} + f(\lambda-\varepsilon)e^{\varepsilon a}].$$

When $|X-a|<u$, $e^{\lambda X} \leq e^{\lambda u}e^{\lambda a}$ so

$$\Pr[|X-a|<u] \geq e^{-\lambda u}e^{-\lambda a}\mathrm{E}\left[e^{\lambda X}\chi(|X-a|<u)\right].$$

But $\Pr[X>a-u] \geq \Pr[|X-a|<u]$, giving our general result as follows.

Theorem A.2.1 *For any $a, u, \lambda, \varepsilon$ with $u, \lambda, \varepsilon, \lambda-\varepsilon$ all positive,*

$$\Pr[X>a-u] \geq e^{-\lambda u}\left[g_a(\lambda) - e^{-\varepsilon u}[g_a(\lambda+\varepsilon)+g_a(\lambda-\varepsilon)]\right].$$

We note that this bound has used only three values of the Laplace transform: $f(\lambda), f(\lambda-\varepsilon), f(\lambda+\varepsilon)$.

It is instructive to examine the case when N is the standard normal distribution. We assume a is positive and are interested in the asymptotics as $a \to +\infty$. We set $\lambda = a$ so that $g_a(\lambda) = e^{-a^2/2}$. Now

$$g_a(\lambda \pm \varepsilon) = e^{(\lambda\pm\varepsilon)^2/2 - a(\lambda\pm\varepsilon)} = g_a(\lambda)e^{\varepsilon^2/2}.$$

The cancellation of the linear (in ε) terms was not serendipity, but rather reflected the critical choice of λ to minimize $\ln(g_a(\lambda))$. Now

$$\Pr[N>a-u] \geq g_a(a)e^{-au}\left[1-2e^{-\varepsilon u}e^{\varepsilon^2/2}\right].$$

Suppose we take $\varepsilon = u = 2$. This gives

$$\Pr[N>a-2] \geq e^{-a^2/2}e^{-2a}\left[1-2e^{-2}\right].$$

Rescaling: $\Pr[N>a] = \Omega(e^{-a^2/2}e^{-4a})$. In contrast we have the upper bound $\Pr[N>a] \leq e^{-a^2/2}$.

In many applications one does not have the precise values of the Laplace transform $f(\lambda)$. Suppose, however, that we have reasonably good estimates *in both directions* on $f(\lambda)$. Then Theorem A.2.1 will give a lower bound for $\Pr[X>a-u]$ by using a lower bound for $g_a(\lambda)$ and upper bounds for $g_a(\lambda\pm\varepsilon)$. Our goal will be less ambitious than the estimate achieved for the standard normal N. We shall be content to find the asymptotics of the logarithm of $\Pr[X>a]$. In the next result, the X_n may be imagined to be near the normal distribution. The interval for λ could easily be replaced by $[(1-\gamma)a_n, (1+\gamma)a_n]$ for any fixed positive γ.

Theorem A.2.2 *Let X_n be a sequence of random variables and a_n a sequence of positive reals with $\lim_{n\to\infty} a_n = \infty$. Assume*

$$\mathrm{E}\left[e^{\lambda X_n}\right] = e^{\frac{1}{2}\lambda^2(1+o(1))}$$

uniformly for $\frac{1}{2}a_n \leq \lambda \leq \frac{3}{2}a_n$. Then

$$\ln\Pr[X_n > a_n] \sim -\frac{a_n^2}{2}.$$

Remark. For $X_n = S_n n^{-1/2}$, $\mathrm{E}\left[e^{\lambda X_n}\right] = \cosh^n(\lambda n^{-1/2})$. When $u \to 0$, $\ln\cosh(u) \sim \frac{1}{2}u^2$. The conditions of Theorem A.2.2 therefore hold when $a_n = o(\sqrt{n})$ and $a_n \to +\infty$. That is, $\ln\Pr[S_n > b_n] \sim -b_n^2/2n$ when $\sqrt{n} \ll b_n \ll n$.

Proof. The upper bound is the Chernoff bound with $\lambda = a_n$.

$$\Pr[X_n > a_n] \leq \mathrm{E}\left[e^{\lambda X_n}\right]e^{-a_n\lambda} = e^{-\frac{1}{2}a_n^2(1+o(1))}.$$

For the lower bound we first let $\delta \in (0, 0.01)$ be fixed. We set $\lambda = a = a_n(1+\delta)$, $u = a_n\delta$, $\varepsilon = \lambda\delta/10$. Applying Theorem A.2.1

$$\Pr[X > a_n] \geq e^{-\lambda u}B$$

with

$$B = g_a(a) - e^{-\varepsilon u}[g_a(a+\varepsilon) + g_a(a-\varepsilon)].$$

But $\ln[g_a(a)] \sim -\frac{1}{2}a^2$ and, analogous to our result for the standard normal,

$$\ln[g_a(a\pm\varepsilon)] \sim \frac{a^2}{2}\left(1 \pm \frac{\delta}{10}\right)^2 - a^2\left(1 \pm \frac{\delta}{10}\right) = \frac{a^2}{2}\left(-1 + \frac{\delta^2}{100}\right).$$

As $\varepsilon u = a^2\delta^2/10(1+\delta)$ we have $e^{-\varepsilon u}g_a(a\pm\varepsilon) \ll g_a(a)$. Now B is dominated by its initial term and

$$\Pr[X > a_n] \geq e^{-\lambda u}g_a(a)(1-o(1)).$$

Taking logarithms:

$$\ln[\Pr[X > a_n]] \geq -a_n^2\delta(1+\delta) - \frac{a_n^2}{2}(1+\delta)^2(1+o(1)) - o(1).$$

As this holds for any fixed $\delta \in (0, 0.01)$

$$\ln[\Pr[X > a_n]] \geq -\frac{a_n^2}{2}(1+o(1)).$$

■

We have seen that $\Pr[S_n > b_n]$ can be well approximated by $\Pr[\sqrt{n}N > b_n]$ as long as $\sqrt{n} \ll b_n \ll n$. For $b_n = \Theta(n)$ this approximation by the normal distribution is no longer valid. Still, we shall see that the Chernoff bounds continue to give the right asymptotic value for $\ln \Pr[S_n > b_n]$. We place this in a somewhat wider context. Ellis (1984) has given far more general results.

Theorem A.2.3 *Let Z_n be a sequence of random variables. Let a be a fixed positive real. Set*

$$F(\lambda) = \lim_{n\to\infty} \frac{1}{n} \ln \mathrm{E}\left[e^{\lambda Z_n}\right].$$

Suppose that there exists $\lambda > 0$ and an open interval I containing λ such that

1. *$F(s)$ exists and has a first and second derivative for all $s \in I$.*
2. *$F'(\lambda) = a$.*
3. *F' is a strictly increasing function in I.*
4. *There is a K so that $|F''(s)| \le K$ for all $s \in I$.*

Then

$$\lim_{n\to\infty} \frac{1}{n} \ln \Pr[Z_n > an] = F(\lambda) - a\lambda.$$

Remark. Let X be a random variable whose Laplace transform is well defined. Let Z_n denote the sum of n independent copies of X. Then $F(\lambda) = \ln \mathrm{E}\left[e^{\lambda X}\right]$. In particular, suppose $\Pr[X = 1] = \Pr[X = -1] = \frac{1}{2}$ so that $Z_n = S_n$. Then $F(\lambda) = \ln\cosh(\lambda)$. For any $a \in (0, 1)$ there is a positive λ for which $a = F'(\lambda) = \tanh(\lambda)$. The conditions of Theorem A.2.3 hold and give the asymptotics of $\ln(S_n > an)$.

Proof. The upper bound is the Chernoff bound as

$$\Pr[Z_n > an] \le \mathrm{E}\left[e^{\lambda Z_n}\right] e^{-a\lambda n} = e^{n(F(\lambda) - a\lambda + o(1))}.$$

For the lower bound we will apply Theorem A.2.1. First note that since F' is continuous and monotone over I it has a continuous inverse H defined over some interval J containing a. Note $H(a) = \lambda$. Let u be positive and sufficiently small so that $H(a+u) \pm u/k \in I$. As

$$\lim_{u\to 0} H(a+u) \pm \frac{u}{K} = H(a) = \lambda,$$

all sufficiently small u satisfy this criterion.

Set $a^* = a + u$ and $\lambda^* = H(a^*)$ so that $F'(\lambda^*) = a^*$. We define

$$g_n(s) = \mathrm{E}\left[e^{sZ_n}\right] e^{-sa^*}.$$

Theorem A.2.1 (noting that $an = a^*n - un$) states

$$\Pr[Z_n > an] \ge e^{-\lambda^* a^* n}\left[g_n(\lambda^*) - e^{-\varepsilon u n}[g_n(\lambda^* + \varepsilon) + g_n(\lambda^* - \varepsilon)]\right].$$

We select $\varepsilon = u/K$. Our selection of u assures us that $\lambda^* \pm \varepsilon$ belong to I. We have

$$\lim_{n\to\infty} \frac{1}{n} \ln\left(\frac{e^{-\varepsilon u n} g_n(\lambda^* + \varepsilon)}{g_n(\lambda^*)}\right) = -\varepsilon u + F(\lambda^* + \varepsilon) - F(\lambda^*) - \varepsilon a^*.$$

We have selected λ^* so that $F'(\lambda^*) = a^*$. Since $|F''(s)| \le K$ in the interval I Taylor series bounds

$$|F(\lambda^* + \varepsilon) - F(\lambda^*) - \varepsilon a^*| \le \frac{K}{2}\varepsilon^2.$$

Our choice of ε (chosen to minimize the quadratic though any sufficiently small ε would do) gives that

$$-\varepsilon u + F(\lambda^* + \varepsilon) - F(\lambda^*) - \varepsilon a^* \le -\frac{u^2}{2K}.$$

Thus $e^{-\varepsilon n} g_n(\lambda^* + \varepsilon)/g_n(\lambda^*)$ drops exponentially quickly. We only use that for n sufficiently large the ratio is less than 0.25. The same argument shows that for n sufficiently large $e^{-\varepsilon n} g_n(\lambda^* - \varepsilon)/g_n(\lambda^*) < 0.25$. For such n we then have

$$\Pr[Z_n > an] \ge \frac{1}{2} e^{-\lambda^* a^* n} g_n(\lambda^*).$$

This lower bound is $\exp[n(F(\lambda^*) - \lambda^* a^* + o(1))]$. Now consider $F(\lambda^*) - \lambda^* a^*$ as a function of u. As $u \to 0$, $\lambda^* = H(a+u) \to H(a) = \lambda$. As F is continuous $F(\lambda^*) \to F(\lambda)$. Clearly $a^* = a + u \to a$ and therefore $\lambda^* a^* \to \lambda a$. Thus

$$F(\lambda^*) - \lambda^* a^* \to F(\lambda) - \lambda a$$

so

$$\Pr[Z_n > an] \ge e^{n(F(\lambda) - \lambda a + o(1))}.$$

■

Remark. Let Z_n be a sequence of random variables with mean and variance μ_n and σ_n^2, respectively. The analysis of $\Pr[Z_n > \mu_n + \lambda_n \sigma_n]$ frequently (S_n being the premier example) splits into three parts:

1. *Small Deviations.* $\lambda_n \to \lambda$, a positive constant. One hopes to prove asymptotic normality so that $\Pr[Z_n > \mu_n + \lambda_n\sigma_n] \to \Pr[N > \lambda]$. There is a huge literature on asymptotic normality but, for the most part, asymptotic normality is not covered in this work.
2. *Large Deviations.* $\lambda_n \to +\infty$ and $\lambda_n = o(\sigma_n)$. One hopes to show that Z_n is approximately normal in the sense that $\ln \Pr[Z_n > \mu_n + \lambda_n\sigma_n] \sim -\lambda_n^2/2$.
3. *Very Large Deviations.* $\lambda_n \to +\infty$ and $\lambda_n = \Omega(\sigma_n)$. Here the approximation of Z_n by the normal distribution generally fails but one hopes that the asymptotics of $\ln \Pr[Z_n > \mu_n + \lambda_n\sigma_n]$ may still be found by the methods we have given.

A.3 EXERCISES

1. The *Hajós number* of a graph G is the maximum number k such that there are k vertices in G with a path between each pair so that all the $\binom{k}{2}$ paths are internally pairwise vertex disjoint (and no vertex is an internal vertex of a path and an endpoint of another). Is there a graph whose chromatic number exceeds twice its Hajós number ?

2. For two subsets A and B of the set $\mathbb{Z}_m$ of integers modulo m and for $g \in \mathbb{Z}_m$, denote

$$s(A, B, g) = |\{(a, b) : a \in A, b \in B, a + b = g\}| .$$

For a partition of $\mathbb{Z}_m$ into two disjoint sets $\mathbb{Z}_m = A \cup B$, $A \cap B = \emptyset$ denote

$$c(A, B) = \max_{x \in \mathbb{Z}_m} |s(A, A, x) + s(B, B, x) - 2s(A, B, x)| .$$

Prove that for every odd m there is a partition of $\mathbb{Z}_m$ into two disjoint sets A and B such that $c(A, B) = O(\sqrt{m \log m})$.

3. For $a \in (0, 1)$ apply Theorem A.2.3 to find $\lim_n (1/n) \ln \Pr[S_n > an]$ explicitly. Express $\Pr[S_n > an]$ combinatorially as 2^{-n} times the sum of binomial coefficients. Use Stirling's formula to asymptotically evaluate this sum and show that you get the same result for $\lim_n (1/n) \ln \Pr[S_n > an]$.

4. More generally, for $p \in (0, 1)$ fixed, apply Theorem A.2.3 to find the asymptotics of $\ln \Pr[B(n, p) > an]$ for $p < a < 1$ and of $\ln \Pr[B(n, p) < an]$ for $0 < a < p$. Show that an application of Stirling's formula gives the same answer.

5. Let $\{X_i\}_{i=1}^n$ be independent random variables, each chosen uniformly from $\{+1, +2, -3\}$. Set $Y_n = \sum_{i=1}^n X_i$. Let $f(n)$ be the minimal value so that $\Pr[Y_n > f(n)] < 1/n$. Find the asymptotics of $f(n)$. Redo with $1/n$ replaced by n^{-50}. (Note that it doesn't change the answer much!)

THE PROBABILISTIC LENS:

Triangle-Free Graphs Have Large Independence Numbers

Let $\alpha(G)$ denote the independence number of a graph G. It is easy and well known that for every graph G on n vertices with maximum degree d, $\alpha(G) \geq n/(d+1)$. Ajtai, Komlós and Szemerédi (1980) showed that in case G is triangle-free, this can be improved by a logarithmic factor and in fact $\alpha(G) \geq (cn \log d)/d$, where c is an absolute positive constant. Shearer (1983) simplified the proof and improved the constant factor to $c = 1 + o(1)$. Here is a very short proof, without any attempt to optimize c, which is based on a different technique of Shearer (1995) and its modification in Alon (1996).

Proposition 1 *Let $G = (V, E)$ be a triangle-free graph on n vertices with maximum degree at most $d \geq 1$. Then $\alpha(G) \geq (n \log d)/8d$, where the logarithm here and in what follows is in base 2.*

Proof. If, say, $d < 16$ the result follows from the trivial bound $\alpha(G) \geq n/(d+1)$ and hence we may and will assume that $d \geq 16$. Let W be a random independent set of vertices in G, chosen uniformly among all independent sets in G. For each vertex $v \in V$ define a random variable $X_v = d|\{v\} \cap W| + |N(v) \cap W|$, where $N(v)$ denotes the set of all neighbors of v. We claim that the expectation of X_v satisfies $\mathrm{E}[X_v] \geq \frac{1}{4} \log d$.

To prove this claim, let H denote the induced subgraph of G on $V - (N(v) \cup \{v\})$, fix an independent set S in H and let X denote the set of all non neighbors of S in the set $N(v)$, $|X| = x$. It suffices to show that the conditional expectation

$$\mathrm{E}\left[X_v \mid W \cap V(H) = S\right] \geq \frac{\log d}{4} \tag{1}$$

for each possible S. Conditioning on the intersection $W \cap V(H) = S$ there are precisely $2^x + 1$ possibilities for W: one in which $W = S \cup \{v\}$ and 2^x in which $v \notin W$ and W is the union of S with a subset of X. It follows that the conditional expectation considered in (1) is precisely

$$\frac{d}{2^x + 1} + \frac{x 2^{x-1}}{2^x + 1}.$$

To check that the last quantity is at least $\frac{1}{4}\log d$ observe that the assumption that this is false implies that $x \geq 1$ and $2^x(\log d - 2x) > 4d - \log d$, showing that $\log d > 2x \geq 2$ and hence $4d - \log d < \sqrt{d}(\log d - 2)$, which is false for all $d \geq 16$. Therefore

$$\mathrm{E}\left[X_v \mid W \cap V(H) = S\right] \geq \frac{\log d}{4},$$

establishing the claim.

By linearity of expectation we conclude that the expected value of the sum $\sum_{v \in V} X_v$ is at least $\frac{1}{4} n \log d$. On the other hand, this sum is clearly at most $2d|W|$, since each vertex $u \in W$ contributes d to the term X_u in this sum, and its degree in G, which is at most d, to the sum of all other terms X_v. It follows that the expected size of W is at least $(n \log d)/8d$, and hence there is an independent set of size at least this expectation, completing the proof. ■

The *Ramsey number* $R(3,k)$ is the minimum number r such that any graph with at least r vertices contains either a triangle or an independent set of size k. The asymptotic behavior of this function has been studied for over fifty years. It turns out that $R(3,k) = \Theta(k^2/\log k)$. The lower bound is a recent result of Kim (1995), based on a delicate probabilistic construction together with some thirty pages of computation. There is no known explicit construction of such a graph, and the largest known explicit triangle-free graph with no independent set of size k, described in Alon (1994), has only $\Theta(k^{3/2})$ vertices. The tight upper bound for $R(3,k)$, proved in Ajtai et al. (1980), is a very easy consequence of the above proposition.

Theorem 2 [Ajtai et al. (1980)] *There exists an absolute constant b such that $R(3,k) \leq bk^2/\log k$ for every $k > 1$.*

Proof. Let $G = (V, E)$ be a triangle-free graph on $8k^2/\log k$ vertices. If G has a vertex of degree at least k then its neighborhood contains an independent set of size k. Otherwise, by Proposition 1 above, G contains an independent set of size at least

$$\frac{8k^2}{\log k}\frac{\log k}{8k} = k.$$

Therefore, in any case $\alpha(G) \geq k$, completing the proof. ■