An Introduction to Large Deviations for Teletraffic Engineers

John T. Lewis[1] Raymond Russell[1]

November 1, 1997

1 What Large Deviation Theory is About

Roughly speaking, Large Deviations is a theory of rare events. It is probably the most active field in probability theory at present, one which has many surprising ramifications. One of its applications is to the analysis of the tails of probability distributions and, in recent years, this aspect of the theory has been widely used in queuing theory. The aim of this tutorial is to introduce the reader to the ideas underlying the theory, to explain why it is called "Large Deviations", and to outline the main theorems and some important applications, in particular the application to queuing theory. Here is a summary of what you will learn from this tutorial:

- What Large Deviation Theory is About
- Coin Tossing: Exploring Large Deviations Using Your PC
- Cramér's Theorem: Introducing Rate-Functions
- Why "Large" in Large Deviations?
- Chernoff's Formula: Calculating the Rate-Function
- The Connection with Shannon Entropy
- Varadhan's Theorem and the Scaled CGF
- The Contraction Principle: Moving Large Deviations About

[1]Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland; e-mail:lewis@stp.dias.ie and russell@stp.dias.ie

- Large Deviations in Queuing Networks: Effective Bandwidths
- Bypassing Modelling: Estimating the Scaled CGF

2 The Basic Ideas Underlying Large Deviations

If you have a little skill in programming, you can very quickly get a good feel for the basic ideas of Large Deviation theory by carrying out on your PC the experiments we are going to describe. Even if you can't program – and you can't rope anyone into programming for you – you will find it useful to read this section: consider what we are going to describe as a thought-experiment, and we will supply the results.

Coin Tossing: Exploring Large Deviations Using Your PC

Imagine a coin-tossing experiment, where we toss a coin n times and record each result. There are 2 possible outcomes for each toss, giving 2^n possible outcomes in all. What can we say about the total number of heads? Firstly there are $n+1$ possible values for the total, ranging from 0 heads to n heads; secondly, of the 2^n possible outcomes, ${}^n\mathrm{C}_r$ result in r heads (${}^n\mathrm{C}_r$ is the binomial coefficient $n!/r!(n-r)!$). If the coin is fair, every outcome is equally likely, and so the probability of getting r heads is ${}^n\mathrm{C}_r/2^n$. Thus the average number of heads per toss has $n+1$ possible values, $0, 1/n, 2/n, \ldots 1$ and the value r/n has weight ${}^n\mathrm{C}_r/2^n$. To calculate the probability of the average number of heads per toss lying in a particular range, we add up the weight of each of those possible values which fall inside that range. If we let M_n be the average number of heads in n tosses, then

$$\mathrm{P}(\, x < M_n < y \,) = \sum_{\left\{\, r\,:\, x<\frac{r}{n}<y \right\}} \binom{n}{r} \frac{1}{2^n}.$$

Exercise 1 *Write a function/procedure to take an integer n and two floating-point numbers x and y and return the value of the expression above. Use this function/procedure to write a program to produce histograms of the distribution of M_n for selected values of n.*

We have done this for n=16, n=32, n=64 and n=128 and the results are shown in Figure 1. We can see clearly the Law of Large Numbers at work: as n increases, the distribution becomes more and more sharply peaked about the mean, 1/2, and the tails become smaller and smaller.

Exercise 2 *Pick some point x greater than 1/2 and write a program to calculate, for a range of values of n, the logarithm of the probability of M_n exceeding x.*

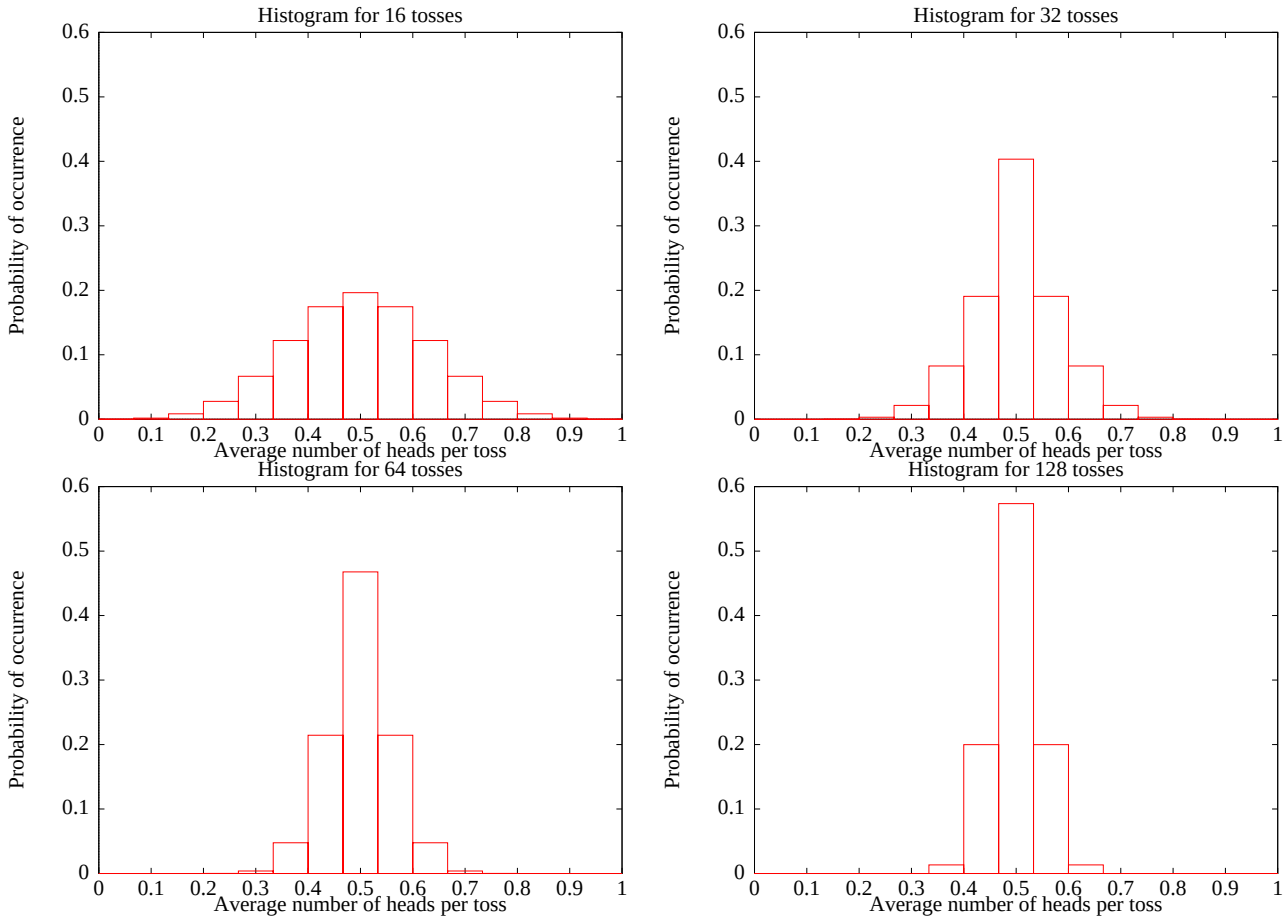

Figure 1: Histograms of the distributions of the average number of heads per toss for increasing numbers of tosses.

We have chosen x=0.6 and produced a plot of $\ln \mathrm{P}(\, M_n > x \,)$ against n for n up to 100, shown in Figure 2. It is clear that, although things are a little jumpy initially, the plot becomes linear for large n. Repeat the experiment for a different value of x and you will see that the same thing happens: no matter what value of x greater than $1/2$ you take, the plot will always be linear for n large. We can see this from Figure 3 which shows $\ln \mathrm{P}(\, M_n > x \,)$ against n for several values of x. How quickly it becomes linear, and what the asymptotic slope is, depends on the value of x, but the graph of $\ln \mathrm{P}(\, M_n > x \,)$ against n is always linear for large n. Let's call this asymptotic slope $-I(x)$.

Exercise 3 *Repeat the experiment for a range of values of x from 1/2 to 1, measure the asymptotic slope in each case, and plot the values of $I(x)$ you get against x. Do the same thing for $\ln \mathrm{P}(\, M_n < x \,)$ for a range of values of*

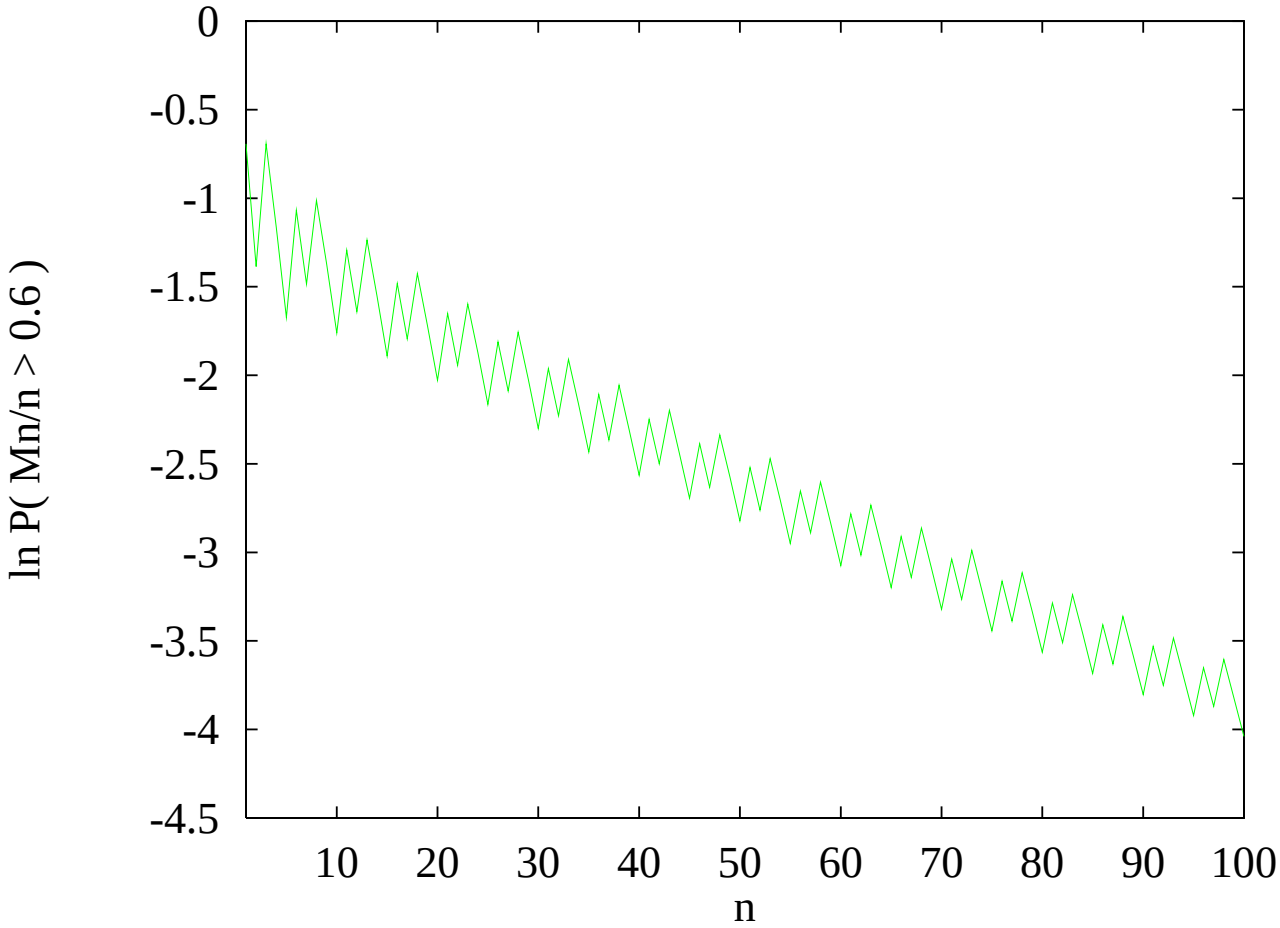

Figure 2: $\ln \mathrm{P}(\, M_n > 0.6\,)$ against n.

x from 0 to 1/2.

Depending on how accurately you measured the slope, you should get results similar to those shown in Figure 4.
You have made a discovery:

> THE TAIL OF THE DISTRIBUTION OF THE AVERAGE NUMBER OF HEADS IN n TOSSES DECAYS EXPONENTIALLY AS n INCREASES

The plot you have made tells you the local rate at which a tail decays as a function of the point from which the tail starts: you have built up a picture of the *rate-function* $I(x)$.

Exercise 4 *Plot the graph of the function* $x \ln x + (1-x)\ln(1-x) + \ln 2$ *against* x *and compare it with your previous plot.*

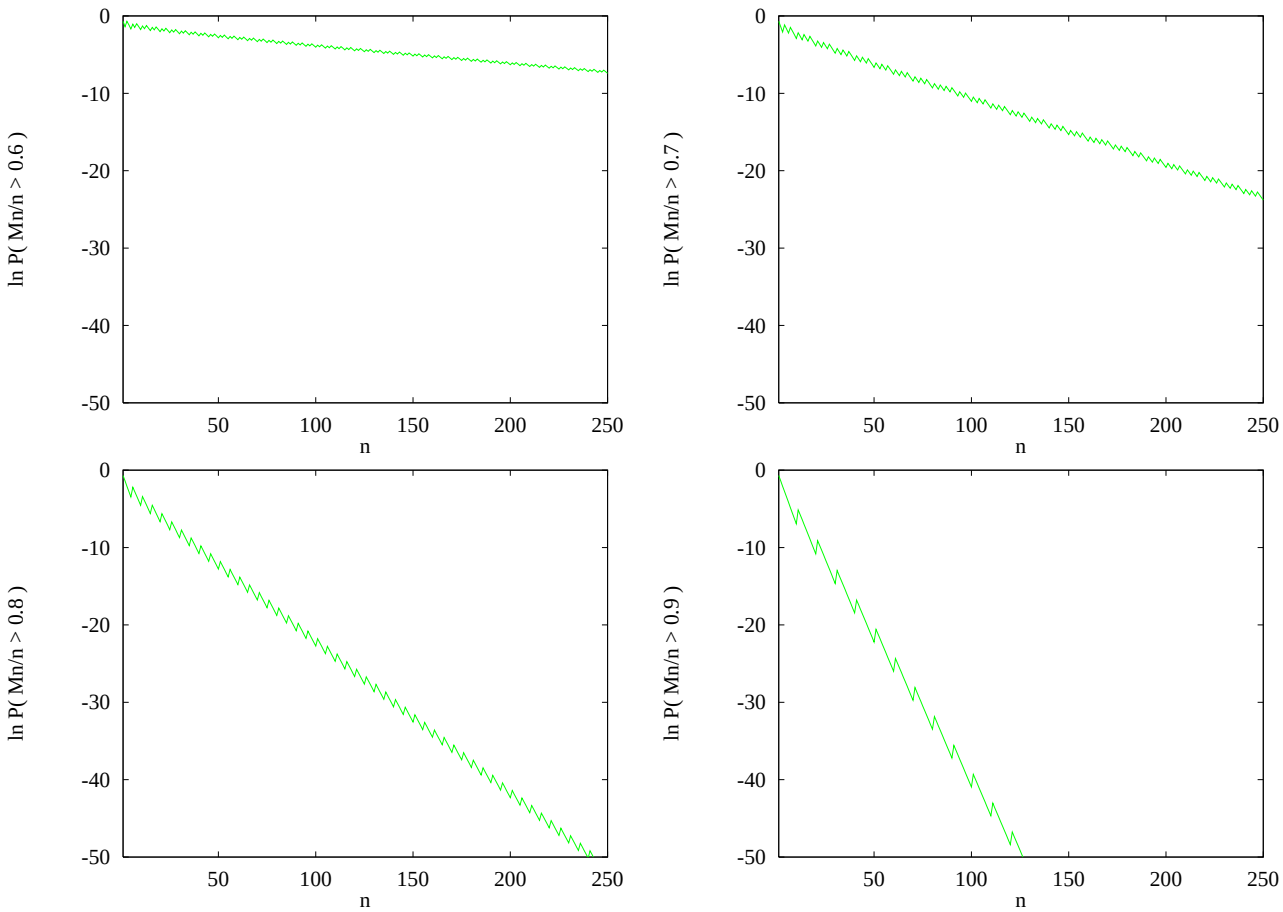

Figure 3: $\ln \mathrm{P}(\, M_n > x \,)$ against n for several values of x.

We have added a plot of this function to the plot in Figure 4 to get Figure 5. We see that the two plots fit: we have guessed a formula for $I(x)$, the rate-function for coin-tossing.

One of the goals of Large Deviation theory is to provide a systematic way of calculating the rate-function; we will show you later one way of achieving this.

To summarise: we have found that, for coin tossing, the tails of the distribution of M_n, the average number of heads in n tosses, decay exponentially fast:

$$\begin{aligned} \mathrm{P}(\, M_n > x \,) &\asymp e^{-nI(x)} \qquad \text{for} \quad x > 1/2, \\ \mathrm{P}(\, M_n < x \,) &\asymp e^{-nI(x)} \qquad \text{for} \quad x < 1/2, \end{aligned}$$

as n becomes large; in fact, as you can see from Figure 2, the approximation is quite good for surprisingly small values of n. The combinatorial approach

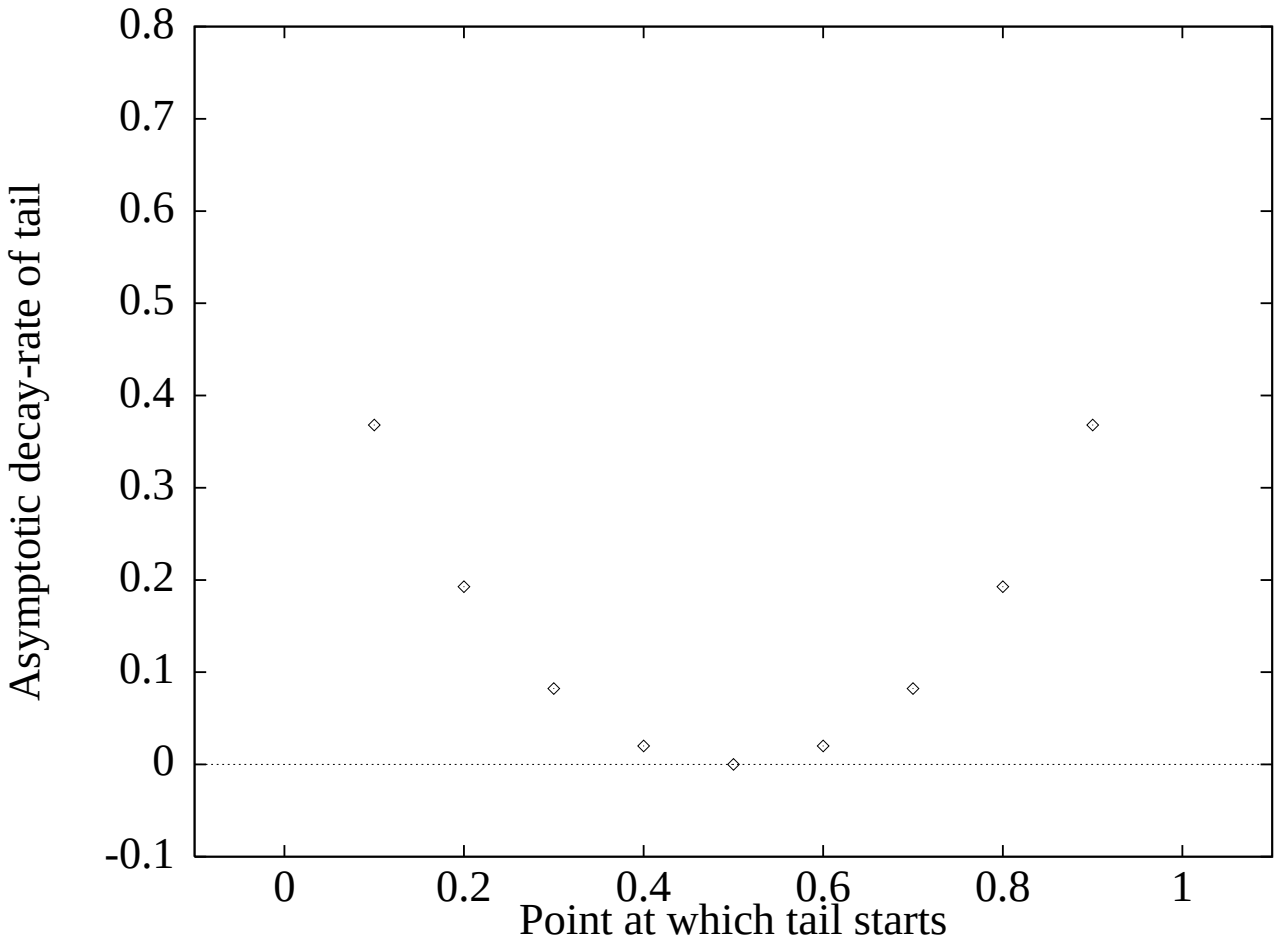

Figure 4: Measured decay-rates of tails against starting point of tail.

we have investigated numerically can be made, using Stirling's Formula, to yield a proof of this result; this is sketched in Appendix A.

The Weak Law of Large Numbers Regained

Notice that a consequence of this result is the Weak Law of Large Numbers for coin-tossing. It states that, as n increases, the distribution of M_n becomes more and more sharply peaked about the mean; in symbols:

$$\lim_{n\to\infty} \mathrm{P}(\mid M_n - 1/2 \mid < \epsilon) = 1$$

for each positive number ϵ. This is the same thing as saying that, as n increases, the tails become smaller and smaller; in symbols:

$$\lim_{n\to\infty} \mathrm{P}(\mid M_n - 1/2 \mid > \epsilon) = 0$$

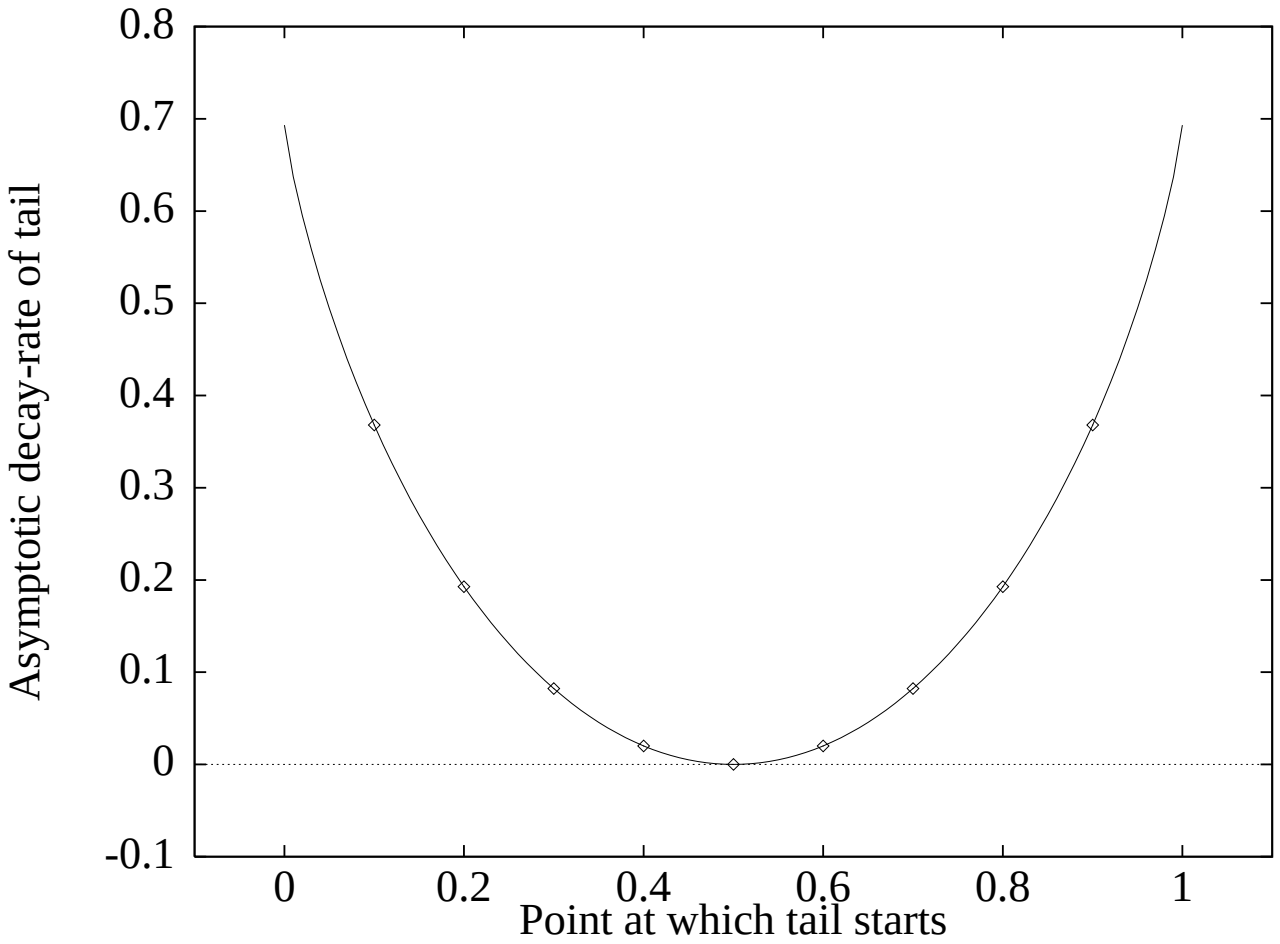

Figure 5: Measured decay-rates with the rate-function superimposed.

for each positive number ϵ. How do we set about proving this? First, let us write out $\mathrm{P}(\mid M_n - 1/2 \mid > \epsilon)$ in detail and see if we can approximate it or get a bound on it:

$$\mathrm{P}(\mid M_n - 1/2 \mid > \epsilon) = \mathrm{P}(M_n < 1/2 - \epsilon) + \mathrm{P}(M_n > 1/2 + \epsilon).$$

Now we have found that, for coin tossing, the tails of the distribution of M_n decay exponentially fast:

$$\begin{aligned} \mathrm{P}(M_n > x) &\asymp e^{-nI(x)} \qquad \text{for} \quad x > 1/2, \\ \mathrm{P}(M_n < x) &\asymp e^{-nI(x)} \qquad \text{for} \quad x < 1/2, \end{aligned}$$

as n becomes large, with $I(x) > 0$; it follows that both terms on the right-hand side of the equation decay to zero so that

$$\lim_{n\to\infty} \mathrm{P}(\mid M_n - 1/2 \mid > \epsilon) = 0$$

for each positive number ϵ. This shows that the Weak Law of Large Numbers is a consequence of the Large Deviation Principle.

3 Cramér's Theorem: Introducing Rate-Functions

Harald Cramér was a Swedish mathematician who served as a consultant actuary for an insurance company; this led him to discover the first result in Large Deviation theory. The Central Limit Theorem gives information about the behaviour of a probability distribution near its mean while the risk theory of insurance is concerned with rare events out on the tail of a probability distribution. Cramér was looking for a refinement of the Central Limit Theorem; what he proved was this:

Cramér's Theorem *Let $X_1, X_2, X_3, \ldots$ be a sequence of bounded, independent and identically distributed random variables each with mean m, and let*

$$M_n = \frac{1}{n}(X_1 + \ldots + X_n)$$

denote the empirical mean; then the tails of the probability distribution of M_n decay exponentially with increasing n at a rate given by a convex rate-function $I(x)$:

$$\begin{aligned} \mathrm{P}(\, M_n > x \,) &\asymp e^{-nI(x)} \qquad \text{for} \quad x > m, \\ \mathrm{P}(\, M_n < x \,) &\asymp e^{-nI(x)} \qquad \text{for} \quad x < m. \end{aligned}$$

Historically, Cramér used complex variable methods to prove his theorem and gave $I(x)$ as a power-series; later in this section we will explain why this is a natural approach to the problem. However, there is another approach which has the advantage that it can be generalised in several directions and which has proved to be of great value in the development of the theory. We illustrate this approach in Appendix B by sketching a proof of Cramér's Theorem which uses it; but first let us see how the theorem can be used in risk-theory.

An Application to Risk-Theory

Large Deviation theory has been applied to sophisticated models in risk theory; to get a flavour of how this is done, consider the following simple model. Assume that an insurance company settles a fixed number of claims in a fixed period of time, say one a day; assume also that it receives a steady income from premium payments, say an amount p each day. The sizes of the claims are random and there is therefore the risk that, at the end of some planning period of length T, the total amount paid in settlement of claims will exceed the total income from premium payments over the period. This risk is inevitable, but the company will want to ensure that it is small (in the

interest of its shareholders, or because it is required by its reinsurers or some regulatory agency). So we are interested in the small probabilities concerning the sum of a large number of random variables: this problem lies squarely in the scope of Large Deviations.

If the sizes X_t of claims are independent and identically distributed, then we can apply Cramér's Theorem to approximate the probability of ruin, the probability that the amount $\sum_{t=1}^{T} X_t$ paid out during the planning period T exceeds the premium income pT received in that period:

$$\mathrm{P}\left(\sum_{t=1}^{T} X_t > pT \right) \asymp e^{-TI(p)}.$$

So, if we require that the risk of ruin be small, say e^{-r} for some large positive number r, then we can use the rate-function I to choose an appropriate value of p:

$$\begin{aligned} \mathrm{P}\left(\frac{1}{T} \sum_{t=1}^{T} X_t > p \right) &\approx e^{-r} \\ e^{-TI(p)} &\approx e^{-r} \\ I(p) &\approx r/T \end{aligned}$$

Since $I(x)$ is convex, it is monotonically increasing for x greater than the mean of X_t and so the equation

$$I(p) = r/T$$

has a unique solution for p.

Of course, to solve this equation, we must know what $I(x)$ is and that means knowing the statistics of the sizes of the claims. For example, if the size of each claim is normally distributed with mean m and variance σ^2, then the rate-function is

$$I(x) = \frac{1}{2} \left(\frac{x-m}{\sigma} \right)^2.$$

It is easy to find the solution to the equation for p in this case: it is $p = m + \sigma\sqrt{2r/T}$; thus the premium should be set so that the daily income is the mean claim size plus an additional amount to cover the risk. The ratio $(p-m)/m$ is called the *safety loading*; in this case, it is given by $(\sigma/m)\sqrt{2r/T}$. Notice that σ/m is a measure of the size of the fluctuations in claim-size, while $\sqrt{T/2r}$ is fixed by the regulator.

Why "Large" in Large Deviations?

Recall what the Central Limit Theorem tells us: if $X_1, X_2, X_3, \ldots$ is a sequence of independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$, then the average of the first n of them, $M_n = \frac{1}{n}(X_1 + \ldots + X_n)$ is approximately normal with mean μ and variance σ^2/n. That is, its probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{n}{2}\frac{(x-\mu)^2}{\sigma^2}},$$

and the approximation is only valid for x within about $\sigma/\sqrt{n}$ of μ. If we ignore the prefactor in f and compare the exponential term with the approximation that Cramér's Theorem gives us, we see that the terms $(x-\mu)^2/2\sigma^2$ occupy a position analogous to that of the rate function. Let us look again at the coin tossing experiments: for x close to $1/2$, we can expand our rate-function in a Taylor series:

$$x \ln x + (1-x)\ln(1-x) + \ln 2 = \frac{(x - \frac{1}{2})^2}{2 \times \frac{1}{4}} + \ldots$$

The mean of each toss of a coin is $1/2$, and the variance of each toss is $1/4$; thus the rate-function for coin tossing gives us the Central Limit Theorem. In general, whenever the rate-function can be approximated near its maximum by a quadratic form, we can expect the Central Limit Theorem to hold.

So much for the similarities between the CLT and Large Deviations; the name "Large Deviations" arises from the contrast between them. The CLT governs random fluctuations only near the mean – deviations from the mean of the order of $\sigma/\sqrt{n}$. Fluctuations which are of the order of σ are, relative to typical fluctuations, much bigger: they are *large deviations* from the mean. They happen only rarely, and so Large Deviation theory is often described as *the theory of rare events* – events which take place away from the mean, out in the tails of the distribution; thus Large Deviation theory can also be described as a theory which studies the tails of distributions.

Chernoff's Formula: Calculating the Rate-Function

One way of calculating the rate-function for coin-tossing is to apply Stirling's Formula in conjunction with the combinatorial arguments we used earlier; this is sketched in Appendix A. There is, however, an easier and more general method for calculating the rate-function for the independent case; it is known as *Chernoff's Formula*. To understand the idea behind it, we look at a way of getting an upper bound on a tail probability which is often used in probability theory.

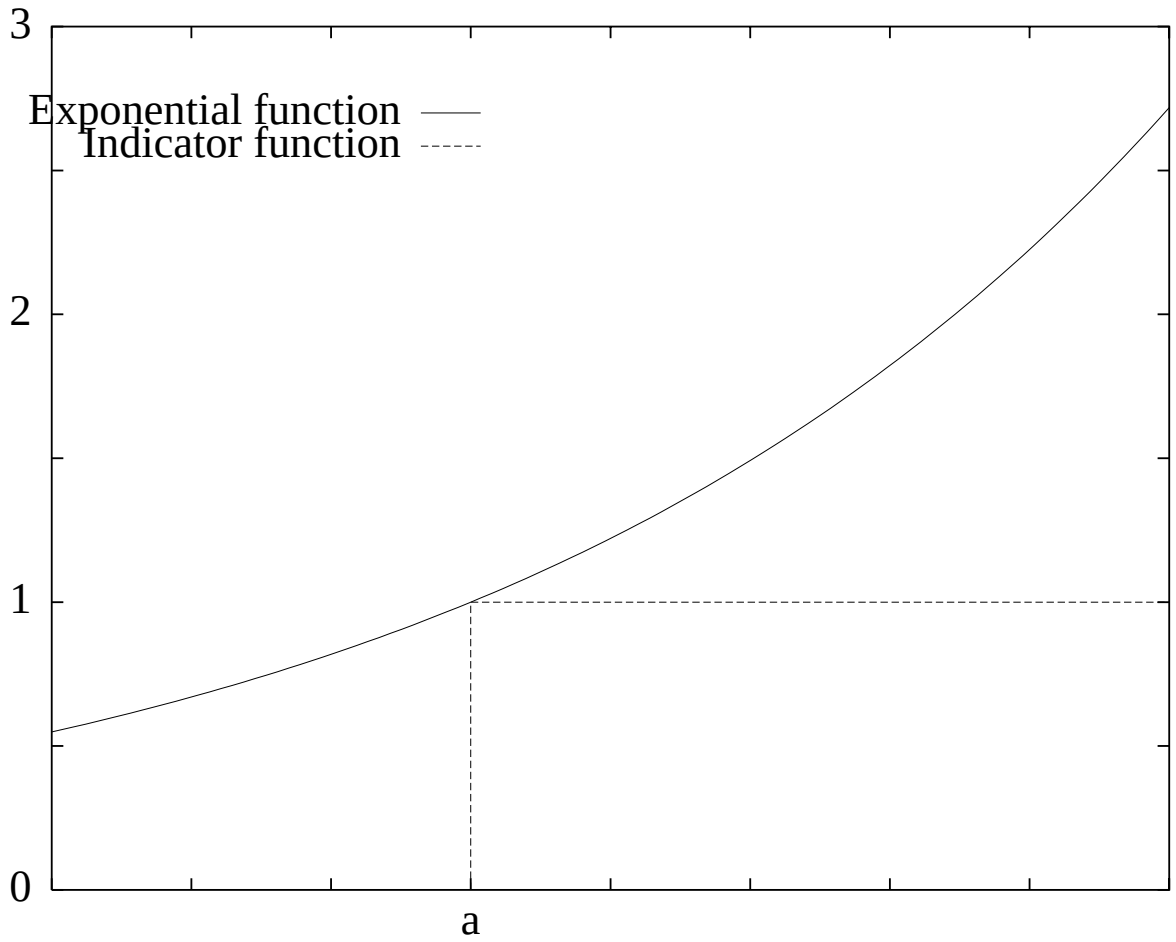

Figure 6: $\mathbb{I}_{(a,\infty)}x \leq e^{\theta x}/e^{\theta a}$ for each a and each $\theta > 0$.

Chernoff's Bound

First let us look at a simple bound on $\mathrm{P}(\, M_n > a \,)$ known as *Chernoff's bound*. We rewrite $\mathrm{P}(\, M_n > a \,)$ as an expectation using indicator functions. The indicator function of a set $A \subset \mathbb{R}$ is the function $\mathbb{I}_A\cdot$ defined by

$$\mathbb{I}_A x := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 6 shows graphically that, for each number a and each positive number θ, $\mathbb{I}_{(a,\infty)}x \leq e^{\theta x}/e^{\theta a}$. Now note that $\mathrm{E}\,\mathbb{I}_{(na,\infty)}nM_n = \mathrm{P}(\, nM_n > na \,)$ and so

$$\begin{aligned} \mathrm{P}(\, M_n > a \,) &= \mathrm{P}(\, nM_n > na \,) \\ &= \mathrm{E}\,\mathbb{I}_{(na,\infty)}nM_n \\ &\leq \mathrm{E}\, e^{\theta n M_n}/e^{\theta na} \end{aligned}$$

$$
\begin{aligned}
&= e^{-\theta na}\mathrm{E}\, e^{\theta(X_1+\ldots+X_n)} \\
&= e^{-\theta na}\left(\mathrm{E}\, e^{\theta X_i}\right)^n;
\end{aligned}
$$

we take this last step 1 the X_i's are independent and identically distributed. By defining $\lambda(\theta) = \ln \mathrm{E}\, e^{\theta X_1}$ we can write $\mathrm{P}(\, M_n > a\,) \leq e^{-n\{\theta a - \lambda(\theta)\}}$; since this holds for each θ positive, we can optimise over θ to get

$$
\mathrm{P}(\, M_n > a\,) \leq \min_{\theta>0} e^{-n\{\theta a-\lambda(\theta)\}} = e^{-n\max_{\theta>0}\{\theta a-\lambda(\theta)\}}.
$$

If a is greater than the mean m, we can obtain a lower bound which gives

$$
\mathrm{P}(\, M_n > a\,) \asymp e^{-n\max_{\theta}\{\theta a-\lambda(\theta)\}};
$$

This is just a statement of the Large Deviation principle for M_n, and is the basis for

Chernoff's Formula: *the rate-function can be calculated from* λ*, the* cumulant generating function*:*

$$
I(x) = \max_{\theta}\{x\theta - \lambda(\theta)\},
$$

where λ *is defined by*

$$
\lambda(\theta) := \ln \mathrm{E}\, e^{\theta X_j}.
$$

Coin-Tossing Revisited

We explored the large-deviations of the coin-tossing experiment with a fair coin in some detail and painstakingly built up a picture of the rate-function from the asymptotic slopes of our log-probability plots. We could do the same again for a biased coin, but Cramér's Theorem tells us immediately that we will again have a Large Deviation principle. Since the outcomes of the tosses are still bounded (either 0 or 1) and independent and identically distributed, Cramér's Theorem applies and we can calculate the rate-function using Chernoff's Formula: let p be the probability of getting heads, so that the cumulant generating function λ is

$$
\lambda(\theta) = \ln \mathrm{E}\, e^{\theta X_1} = \ln(pe^{\theta} + 1 - p).
$$

By Chernoff's Formula, the rate-function is given by

$$
I(x) = \max_{\theta}\left\{x\theta - \lambda(\theta)\right\}.
$$

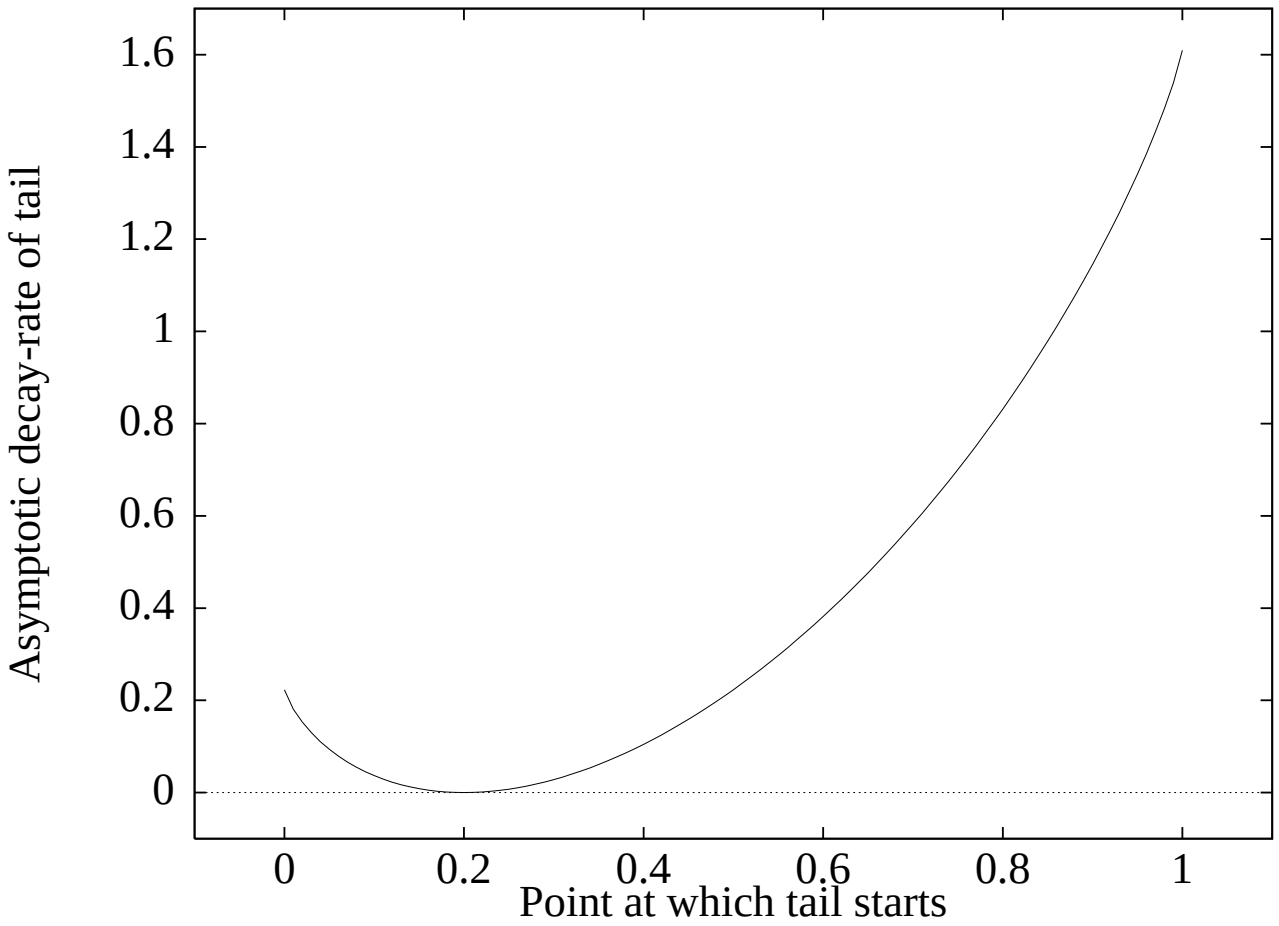

Figure 7: The rate-function for coin tossing with a biased coin (p=0.2).

Exercise 5 *Use differential calculus to show that the value of θ which maximises the expression on the right-hand side is $\theta_x = \ln x/p - \ln(1-x)/(1-p)$ and hence that the rate-function is*

$$I(x) = x \ln \frac{x}{p} + (1-x) \ln \frac{1-x}{1-p}.$$

Figure 7 shows a plot of $I(x)$ against x for $p = 0.25$.

Why Does Chernoff's Formula Work?

The cumulants of a distribution are closely related to the moments. The first cumulant is simply the mean, the first moment; the second cumulant is the variance, the second moment less the square of the first moment.

The relationship between the higher cumulants and the moments is more complicated, but in general the k^{th} cumulant can be written in terms of the first k moments. The relationship between the moments and the cumulants is more clearly seen from their respective generating functions. The function $\phi(\theta) = \mathrm{E}\, e^{\theta X_1}$ is the moment generating function for the X's: the k^{th} moment of the X's is the k^{th} derivative of ϕ evaluated at $\theta = 0$:

$$\begin{aligned} \frac{d^k}{d\theta^k}\phi(\theta) &= \mathrm{E}\, X_1^k e^{\theta X_1} \\ \left.\frac{d^k\phi}{d\theta^k}\right|_{\theta=0} &= \mathrm{E}\, X_1^k = k^{\text{th}} \text{ moment} \end{aligned}$$

The cumulant generating function (CGF) is defined to be the logarithm of the moment generating function, $\lambda(t) := \ln \phi(t)$, and the cumulants are then just the derivatives of λ:

$$\begin{aligned} \left.\frac{d}{d\theta}\lambda(\theta)\right|_{\theta=0} &= m, \\ \left.\frac{d^2}{d\theta^2}\lambda(\theta)\right|_{\theta=0} &= \sigma^2, \ldots \end{aligned}$$

So, why does Chernoff's Formula work? In order to calculate the Central Limit Theorem approximation for the distribution for M_n, we must calculate the mean and variance of the X's: essentially we use the first two cumulants to get the first two terms in a Taylor expansion of the rate-function to give us a quadratic approximation. It is easy to see that, if we want to get the full functional form of the rate-function, we must use all the terms in a Taylor series — in other words, we must use all the cumulants. The CGF packages all the cumulants together, and Chernoff's Formula shows us how to extract the rate-function from it.

Proving Cramér's Theorem

Cramér's proof of his theorem was based essentially on an argument using moment generating functions and gave the rate-function as a power series. After seeing the connection with the Central Limit Theorem and Chernoff's Formula, we can see how this is a natural approach to take; indeed, one of the standard proofs of the Central Limit Theorem is based on the moment generating function. This method of proof has the drawback that it is not easy to see how to adapt it to more general situations. There is a more elegant argument which establishes the theorem and which can easily be modified to apply to random variables which are vector-valued and to random variables

which are not necessarily independent. This argument shows at the same time that the rate-function is convex; we sketch this proof in appendix B.

4 The Connection with Shannon Entropy

We are often asked "How does Shannon entropy fit into Large Deviation theory?" To answer this question fully would sidetrack us into an exposition of the ideas involved in Information Theory. However, we will use Cramér's Theorem to give a proof of the Asymptotic Equipartition Property, one of the basic results of Information Theory; in the course of this, we will see how Shannon entropy emerges from a Large Deviation rate-function.

Let $A = \{a_1, \dots a_r\}$ be a finite alphabet. We describe a language written using this alphabet by assigning probabilities to each of the letters; these represent the relative frequency of occurrence of the letters in texts written in the language. We write p_k for the relative frequency of the letter a_k. As a first approximation, we ignore correlations between letters and take the probability of a text to be the product of the probabilities of each of the component letters. This corresponds to the assumption that the sequence of letters in any text is an independent and identically distributed sequence of random letters. Let Ω_n be the set of all texts containing exactly n letters; then the probability of the text $\omega = (\omega_1, \omega_2, \dots \omega_n)$ (where each of the ω_i's represents a letter) is just $p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}$, where n_k is the number of occurrences of the letter a_k in the text ω. We write $\alpha[\omega] = p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}$ to explicitly denote the dependence of this probability on the text.

Claude Shannon made a remarkable discovery: if some letters are more probable (occur more frequently), then there is a subset Γ_n of Ω_n consisting of "typical texts" with the following properties:

- texts not in Γ_n occur very rarely, so that $\alpha[\Gamma_n] = \sum_{\omega \in \Gamma_n} \alpha[\omega]$ is close to 1;
- for n large, Γ_n is much smaller than Ω_n:

$$\frac{\#\Gamma_n}{\#\Omega_n} \asymp e^{-n\delta} \qquad \text{for some } \delta > 0;$$

- all texts in Γ_n have roughly the same probability.

This result is known as the *Asymptotic Equipartition Property* (AEP). Shannon introduced a quantity which measures the non-uniformity of a probability measure; it is known as the *Shannon entropy*. The Shannon entropy is defined by

$$h(\alpha) := -p_1 \ln p_1 - p_2 \ln p_2 \dots - p_r \ln p_r.$$

We see that $0 \leq h(\alpha) \leq \ln r$; the maximum value $\ln r$ occurs when all the p_k are equal. Because of the applications of information theory to computer

science and computer engineering, a binary alphabet is often considered, so that $r = 2$, $a_1 = 0$ and $a_2 = 1$, and Shannon entropy is often defined using base 2 logarithms instead of natural logs; in that case, the maximum entropy is 1. The number of elements in Ω_n is $\#\Omega_n = r^n = e^{n \ln r}$; this grows exponentially in n. The number of elements in Γ_n also grows exponentially in n, and the Shannon entropy gives the growth-rate: $\#\Gamma_n \asymp e^{nh(\alpha)}$; thus $\#\Gamma_n / \#\Omega_n \asymp e^{-n(h(\alpha) - \ln r)}$. This is where the second of the three parts of the AEP comes from: the constant δ which appears there is the difference in the growth rates $\delta = \ln r - h(\alpha)$. If α is far from uniform, then $h(\alpha) \ll \ln r$ and Γ_n is substantially smaller than Ω_n even for low values of n.

To prove the AEP, we must first set up some notation. Let X_i be the random variable defined on Ω_n which picks out the i^{th} letter in the text: if $\omega = (\omega_1, \omega_2, \ldots \omega_n)$, then $X_i(\omega) = \omega_i$. For each a in A, let $\delta_a[\cdot]$ be the Dirac measure defined on the subsets of A by

$$\delta_a[B] = \begin{cases} 1, & \text{if } a \in B, \\ 0, & \text{otherwise.} \end{cases}$$

Now put

$$L_n(\omega, B) = \frac{1}{n} \left(\delta_{X_1(\omega)}[B] + \ldots + \delta_{X_n(\omega)}[B] \right);$$

L_n is called the *empirical distribution* because $L_n(\omega, \{a_k\}) = n_k/n$, where n_k is the number of occurrences of the letter a_k in the text ω. The random variables $X_1, \ldots X_n$ are independent and identically distributed with respect to the probability measure α; it is not difficult to see that the same is true of the variables $\delta_{X_1}, \ldots \delta_{X_n}$. If follows that we can apply Cramér's Theorem to $\{L_n\}$ to get the result known as Sanov's Theorem:

Sanov's Theorem: *There exists a convex function I on the space $\mathcal{M}(A)$ of probability measures on A such that*

$$\alpha[L_n \approx \mu] \asymp e^{-nI(\mu)}.$$

Once we know that the rate-function I exists and is convex, we can use Chernoff's Formula to compute it. Since I is a function of a (probability) vector, the cumulant generating function is also a function of a vector $\mathbf{t} = (t_1, \ldots t_r)$:

$$\begin{aligned} \lambda(\mathbf{t}) &= \ln \mathrm{E}\, e^{t_1 \delta_{X_1}[a_1] + \ldots + t_r \delta_{X_1}[a_r]} \\ &= \ln \left(p_1 e^{t_1} + \ldots + p_r e^{t_r} \right) \end{aligned}$$

To compute I, we must calculate the Legendre transform of λ:

$$I(\mu) = \max_{\mathbf{t}} \{\, t_1 \mu[\{a_1\}] + \ldots + t_r \mu[\{a_r\}] - \lambda(\mathbf{t}) \,\}.$$

Exercise 6 *Use differential calculus to show that* $\mathbf{t}$ *must satisfy*

$$\frac{\partial \lambda}{\partial t_k} = \frac{p_k e^{t_k}}{p_1 e^{t_1} + \ldots + p_r e^{t_r}} = \mu[\{a_k\}],$$

and so

$$t_k = \ln \frac{\mu[\{a_k\}]}{p_k} + \lambda(\mathbf{t}).$$

Substitute this into the expression to be minimised to show that

$$I(\mu) = m_1 \ln \frac{m_1}{p_1} + \ldots + m_r \ln \frac{m_r}{p_r},$$

where $m_k = \mu[\{a_k\}]$.

The expression $m_1 \ln m_1/p_1 + \ldots + m_r \ln m_r/p_r$ is written more simply as $D(\mu\|\alpha)$ and is known as the *informational divergence*.

Let us look again at the statement of Cramér's Theorem: we see that the distribution of M_n is concentrated near the mean m, the place where the rate-function vanishes.

Exercise 7 *Show that it follows from Cramér's Theorem that, as* n *increases,*

$$\lim_{n\to\infty} \mathrm{P}(\, m - \delta < M_n < m + \delta \,) = 1,$$

for any $\delta > 0$.

In Sanov's Theorem, the rate-function is $D(\mu\|\alpha)$; this vanishes if and only if $\mu = \alpha$. So we have that

$$\lim_{n\to\infty} \alpha[L_n \approx \alpha] = 1,$$

and it follows that, if we choose Γ_n to be those texts in which the relative frequencies of the letters are close to those specified by α, then $\alpha[\Gamma_n]$ will converge to 1 as n increases. Thus Γ_n consists of the most probable texts and texts which are not in Γ_n occur only very rarely. To estimate the size of Γ_n, we apply Sanov's Theorem a second time. Let β be the uniform probability measure which assigns probability $1/r$ to each of the r letters in A; then $\beta[\Gamma_n] = \#\Gamma_n/\#\Omega_n$. Now, Γ_n is the set of texts ω for which $L_n(\omega)$ is close to α, and so we can apply Sanov's Theorem to the distribution of L_n with respect to β:

$$\beta[\Gamma_n] = \beta[L_n \approx \alpha] \asymp e^{-nD(\alpha\|\beta)}.$$

But

$$\begin{aligned} D(\alpha\|\beta) &= p_1(\ln p_1 - \ln 1/r) + \ldots + p_r(\ln p_r - \ln 1/r) \\ &= \ln r - h(\alpha) \\ &= \delta, \end{aligned}$$

and so $\#\Gamma_n/\#\Omega_n \asymp e^{-n\delta}$.

5 Some General Principles

Varadhan's Theorem

Varadhan's Theorem is one of the most famous results in Large Deviation theory; it is a landmark in the development of the subject. It concerns the asymptotic behaviour of sequences of integrals. Consider the integral

$$G_n = \int_0^\infty e^{ng(x)} \, d\mathrm{P}(\, M_n \approx x \,) \, ;$$

if M_n is such that $\mathrm{P}(\, M_n \approx x \,) \asymp e^{-nI(x)}$, we might guess that

$$\begin{aligned} G_n &\asymp \int_0^\infty e^{ng(x)} e^{-nI(x)} \, dx \\ &= \int_0^\infty e^{n\{g(x)-I(x)\}} \, dx \\ &\asymp e^{n \max_x \{g(x)-I(x)\}} \end{aligned}$$

so that

$$\lim_n \frac{1}{n} \ln \int_0^\infty e^{ng(x)} \, d\mathrm{P}(\, M_n \approx x \,) = \max_x \{g(x) - I(x)\}.$$

Varadhan wrote down a list of four hypotheses which he used in his proof that this asymptotic formula holds whenever g is a bounded continuous function. We list them in appendix C (they are too technical to state here); they give precise meaning to our suggestive notation

$$\mathrm{P}(\, M_n \approx x \,) \asymp e^{-nI(x)}.$$

When they hold, we say that the sequence $\{\mathrm{P}(\, M_n \approx x \,)\}$ (or, more loosely, the sequence $\{M_n\}$) *satisfies a Large Deviation principle with rate-function* I.

Generalisations of Cramér's Theorem

So far, we have talked only of the case in which the random variables are independent and identically distributed (I.I.D.). Can we still prove that $\mathrm{P}(\, M_n \approx x \,) \asymp e^{-nI(x)}$ when these conditions are relaxed? If so, can we calculate $I(x)$? There are two main approaches to these problems: the first, which stems from the work of Ruelle and Lanford on the foundations of

statistical thermodynamics, yields a direct proof of the *existence* of a rate-function satisfying the requirements listed in appendix C and uses Varadhan's Theorem to calculate it; the second is known as the Gärtner-Ellis Theorem — we will examine it after we have described the first approach.

The Ruelle-Lanford Approach

We illustrate the Ruelle-Lanford approach by outlining the proof of Cramér's Theorem for I.I.D. random variables which is given in appendix B. We then indicate how the argument goes when we relax the condition of independence.

It is well known that if a sequence $\{a_n\}$ of numbers is such that

$$a_n \geq a_m \qquad \text{whenever } n > m \tag{1}$$

then its limit exists and is equal to its least upper bound:

$$\lim_{n\to\infty} a_n = \sup_{n>0} a_n \tag{2}$$

Let $s_n(x) := \ln \mathrm{P}(\, M_n > x\,)$; if we could prove that the sequence $\{s_n/n\}$ satisfies Equation 1, then the existence of the limit would follow. We can't quite manage that. However we can use the I.I.D. properties of the random variables $\{X_n\}$ to prove that

$$s_{m+n}(x) \geq s_m(x) + s_n(x) \tag{3}$$

It then follows from a theorem in analysis that

$$\lim_{n\to\infty} \frac{s_n}{n} = \sup_{n>0} \frac{s_n}{n}.$$

This proves the existence of the rate-function

$$I(x) = -\lim_{n\to\infty} \frac{1}{n} \ln \mathrm{P}(\, M_n > x\,).$$

If we give names to these conditions, we can write slogans which are easy to remember. If Equation 1 holds, we say that $\{a_n\}$ is *monotone increasing*; if Equation 2 holds, we say that $\{a_n\}$ is *approximately monotone increasing*; if Equation 3 holds, we say that $\{s_n\}$ is *super-additive*. This is illustrated in Figure 8 using the coin-tossing data; the super-additive sequence $\{\ln \mathrm{P}(\, M_n > 0.6\,)\}$ is shown on top, and the approximately monotone increasing sequence $\{\ln \mathrm{P}(\, M_n > 0.6\,)/n\}$ is shown on the bottom converging to its maximum value.

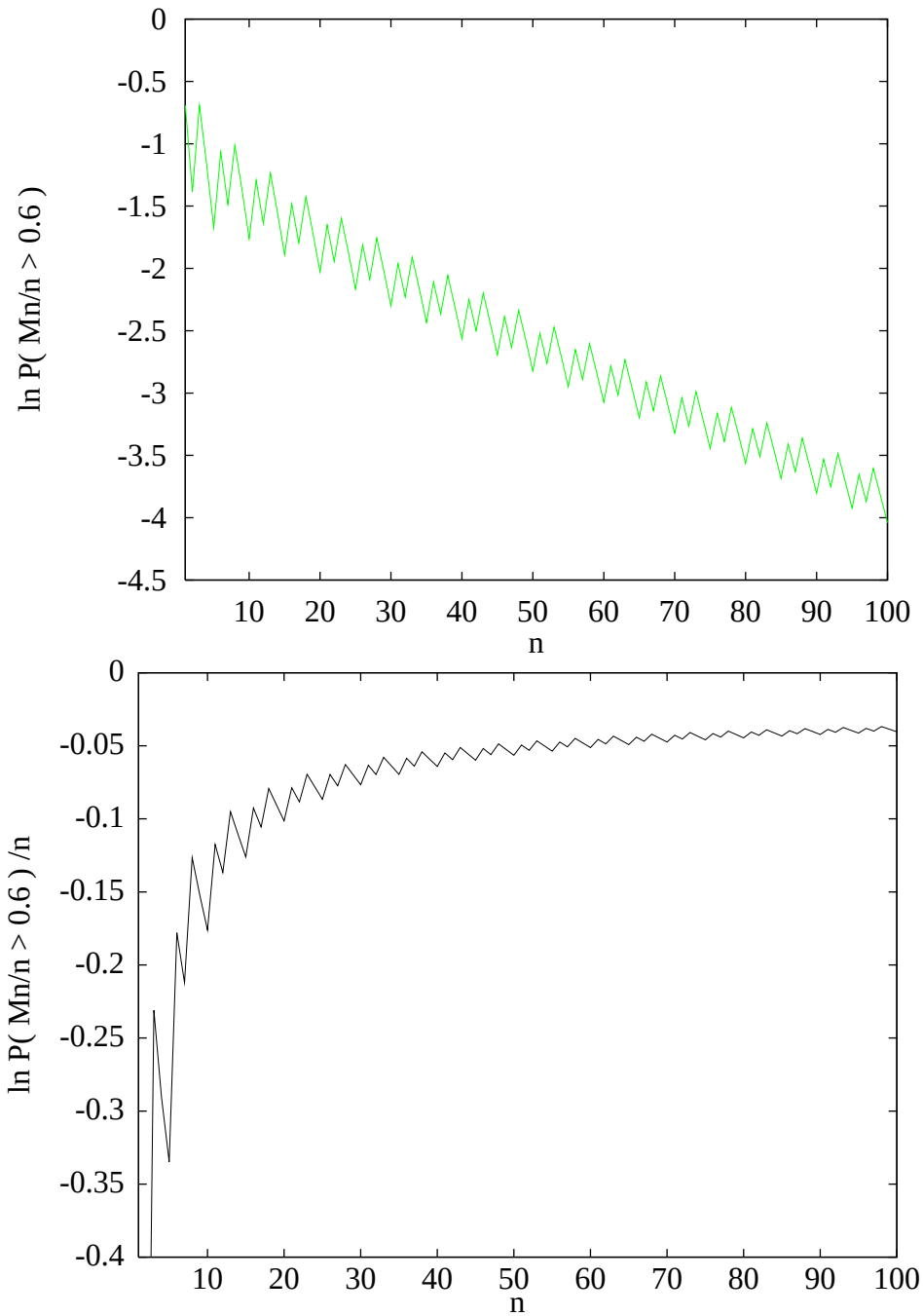

Figure 8: The distribution of the mean for coin-tossing: $\ln \mathrm{P}(\, M_n > 0.6\,)$ (top) is super-additive and so $\ln \mathrm{P}(\, M_n > 0.6\,)\,/n$ (bottom) is approximately monotone; as n increases, $\ln \mathrm{P}(\, M_n > 0.6\,)\,/n$ converges to its maximum value.

The proof of Cramér's Theorem goes like this:

$$\begin{aligned} \{X_n\} \text{ I.I.D.} &\Rightarrow \{s_n(x)\} \text{ super-additive} \\ &\Rightarrow \{s_n(x)/n\} \text{ approximately monotone increasing} \\ &\Rightarrow I(x) \text{ exists.} \end{aligned}$$

The independence condition can be replaced by a condition of *weak dependence* (the definition is technical, so we refrain from giving it here — enough to say that it is satisfied by Markov chains, for example) and the identical distribution condition by *stationarity*; under these conditions, the existence of the rate-function can still be proved. The chain of implications now looks like this:

$$\begin{aligned} &\{X_n\} \text{ stationary and weakly dependent} \\ &\Rightarrow \{s_n(x)\} \text{ approximately super-additive} \\ &\Rightarrow \{s_n(x)/n\} \text{ approximately monotone increasing} \\ &\Rightarrow I(x) \text{ exists.} \end{aligned}$$

A modification of this argument proves that I is a convex function; this is the key to calculating it.

The Scaled CGF: Calculating the Rate-Function

Once we have established the existence of the rate-function, we can apply Varadhan's Theorem. Choosing the function g to be linear,

$$g(x) = \theta x \qquad \text{for some number } \theta,$$

we have

$$\begin{aligned} \lambda(\theta) &:= \lim_{n\to\infty} \frac{1}{n} \ln \mathrm{E}\, e^{n\theta M_n} \\ &= \lim_{n\to\infty} \frac{1}{n} \ln \int_0^\infty e^{n\theta x}\, d\mathrm{P}(\, M_n \approx x\,) \\ &= \max_x \{\theta x - I(x)\}. \end{aligned}$$

The last expression is known as the *Legendre transform* of I. We call λ the *scaled cumulant generating function* (the use of this name should cause no confusion; when $\{X_n\}$ is an I.I.D. sequence, λ reduces to the cumulant generating function — see exercise 8). This application of Varadhan's Theorem proves that when the rate-function I exists, so does the scaled cumulant generating function λ and that λ is the Legendre transform I^* of I. The

Legendre transform is like the Fourier transform in that, for an appropriate class of functions, the transformed function contains exactly the same information as the original function and so the transform is invertible. The Legendre transform is invertible on the class of convex functions and is inverted by repeating it; if I is convex, then its double transform I^{**} is just I itself. Thus, if I is convex, we have

$$I(x) = \lambda^*(x) = \max_\theta \{\theta x - \lambda(\theta)\}.$$

This remark is useful because it frequently happens that the scaled CGF can be computed directly from its defining expression and because the super-additivity argument used to prove the existence of the rate-function proves also that it is convex.

Exercise 8 *Show that, if the X_n's are independent, then the scaled CGF reduces to the CGF of X_1. (Hint: Start by using the fact that the X's are I.I.D. to show that*

$$\mathrm{E}\, e^{n\theta M_n} = \left(\mathrm{E}\, e^{\theta X_1}\right)^n .)$$

Thus, in this case, Varadhan's Theorem yields Chernoff's Formula.

The Gärtner-Ellis Theorem

The second approach, which is exemplified by the Gärtner-Ellis Theorem, assumes the existence of the scaled CGF $\lambda(\theta)$. The upper bound then follows using an extension of the argument which we used to get Chernoff's bound. To get the lower bound, we assume in addition that $\lambda(\theta)$ is differentiable. This approach is very popular — the conditions are easily stated and often not difficult to check, provided we have an explicit expression for the scaled CGF; sometimes, however, they are unnecessarily restrictive.

The Contraction Principle: Moving Large Deviations Around

In many applications of probability theory, we model a process with a sequence $\{X_n\}$ of random variables but, ultimately, we are only interested in some subset of the properties of the process. We may only be interested in a given function f of the value of X_n, and so it is natural to ask about the Large Deviation behaviour of the sequence $\{f(X_n)\}$, given that of the original $\{X_n\}$. The answer is given by the

Contraction principle: *If $\{X_n\}$ satisfies a Large Deviation principle with*

rate-function I and f is a continuous function, then $\{f(X_n)\}$ satisfies a Large Deviation principle with rate-function J, where J is given by

$$J(y) = \min\{\, I(x) \;:\; f(x) = y \,\}.$$

It is called the "contraction" principle because typically f is a contraction in that it throws detail away by giving the same value $f(x)$ to many different values of x.

Why does the contracted rate-function J have this form? Consider $\mathrm{P}(\, f(X_n) \approx y\,)$: there may be many values of x for which $f(x) = y$, say $x_1, x_2, \ldots, x_m$, and so

$$\begin{aligned}
\mathrm{P}(\, f(X_n) \approx y\,) &= \mathrm{P}(\, X_n \approx x_1 \;\text{ or }\; X_n \approx x_2 \ldots \;\text{ or }\; X_n \approx x_m\,) \\
&= \mathrm{P}(\, X_n \approx x_1\,) + \mathrm{P}(\, X_n \approx x_2\,) + \ldots + \mathrm{P}(\, X_n \approx x_m\,) \\
&\asymp e^{-nI(x_1)} + e^{-nI(x_2)} + \ldots + e^{-nI(x_m)}.
\end{aligned}$$

As n grows large, the term which dominates all the others is the one for which $I(x_k)$ is smallest, and so

$$\mathrm{P}(\, f(X_n) \approx y\,) \asymp e^{-n \min\{\, I(x) \,:\, f(x) = y\,\}} = e^{-nJ(y)}.$$

To see the Contraction Principle in action, let us return to the AEP: we saw that the empirical distribution $L_n[B] = (\delta_{X_1}[B] + \ldots + \delta_{X_n}[B])/n$ satisfies a Large Deviation principle in that

$$\mathrm{P}(\, \mathbb{L}_n \approx \mu\,) \asymp e^{-nI(\mu)}$$

Suppose we take the letters $a_1, \ldots a_r$ to be real numbers and that, instead of investigating the distribution of the empirical distribution L_n, we decide to investigate the distribution of the empirical mean $M_n = a_1 n_1/n + \ldots + a_r n_r/n$. Do we have to go and work out the Large Deviation Principle for M_n from scratch? No, because M_n is a continuous function of L_n; it is a very simple function

$$M_n = f(L_n) = a_1 L_n[\{a_1\}] + \ldots + a_r L_n[\{a_r\}].$$

The contraction principle applies, allowing us to calculate the rate-function J for $\{M_n\}$ in terms of the rate-function I for $\{L_n\}$. We have that

$$J(x) = \min_{\mu} D(\mu \| \alpha) \qquad \text{subject to} \quad a_1 m_1 + \ldots + a_r m_r = x,$$

where $m_k = \mu[\{a_k\}]$. This is a simple constrained optimisation problem: we can solve it using Lagrange multipliers.

Exercise 9 *Show that the value of μ which achieves the minimum is given by*

$$\mu[\{a_k\}] = \frac{e^{\beta a_k} p_k}{e^{\beta a_1} p_1 + \ldots + e^{\beta a_r} p_r},$$

where β is the Lagrange multiplier whose value can be determined from the constraint. (Hint: Note that there are two constraints on $\mathbf{m}$ – the requirement that $a_1 m_1 + \ldots + a_r m_r$ be x and the fact that μ is a probability vector – and so you will need two Lagrange multipliers!)

6 Large Deviations in Queuing Systems

A queuing network consists of a network through which customers flow, requiring some kind of processing by servers at the nodes. Typically the servers have limited capacity and the customers must queue while waiting for service. The basic problems in queuing theory is to analyse the behaviour of a single element of a network: the single-server queue.

For simplicity, we will consider a single-server queue in discrete time and, ignoring any considerations of multiple classes of customers or priority, we take the discipline to be FIFO (First In, First Out). Let us set up some notation: let X_i be the amount of work brought by customers arriving at time i, let Y_i be the amount of work the server can do at time i, and let Q_i be the queue-length (i.e. the amount of work waiting to be done) at time i. The dynamics of the system are very simple: the length of the queue at time 0, say, is the sum of the length of the queue at time -1 and the work which arrives at time -1 less the work which can be done at time -1. The service cannot be stored and so, if the service available is greater than the sum of the work in the queue and the work which arrives, then the new queue-length is not negative but zero. This can all be expressed neatly in the formula

$$Q_0 = \max\{0, X_{-1} - Y_{-1} + Q_{-1}\},$$

known as Lindley's equation. We can iterate it, substituting for Q_{-1}; to simplify notation, we set $Z_i = X_{-i} - Y_{-i}$ for all i:

$$\begin{aligned} Q_0 &= \max\{0, Z_1 + Q_{-1}\} \\ &= \max\{0, Z_1 + \max\{0, Z_2 + Q_{-2}\}\} \\ &= \max\{0, Z_1, Z_1 + Z_2 + Q_{-2}\}. \end{aligned}$$

Applying it again gives us

$$\begin{aligned} Q_0 &= \max\{0, Z_1, Z_1 + Z_2 + \max\{0, Z_3 + Q_{-3}\}\} \\ &= \max\{0, Z_1, Z_1 + Z_2, Z_1 + Z_2 + Z_3 + Q_{-3}\}. \end{aligned}$$

It is clear that, by defining $W_t = Z_1 + \ldots + Z_t$, we can write

$$Q_0 = \max\{W_0, W_1, \ldots W_{t-1}, W_t + Q_{-t}\};$$

W_t is called the *workload process.* If we started off at some finite time $-T$ in the past with an empty queue, then

$$Q_0 = \max\{W_0, W_1, \ldots W_{T-1}, W_T\}.$$

However, we may be interested in the equilibrium behaviour of the queue — what the queue-length is when the system has been running for a very long time, when the initial queue-length has no influence. For an equilibrium distribution to exist, the work-process and the service process must satisfy some requirements, the most obvious of which is *stationarity*: the probability distributions of $\{X_i\}$ and $\{Y_i\}$ must be independent of the time. In that case, the equilibrium queue-length Q is given by

$$Q = \max_{t \geq 0} W_t,$$

provided the stability condition

$$\mathrm{E}\, X_i < \mathrm{E}\, Y_i$$

is satisfied. For many purposes, we are interested in the behaviour of the tail of the probability distribution: how big is $\mathrm{P}(\, Q > q\,)$ when q is large? The best way to answer this question is to look at a picture of a typical queue-length distribution. Consider $\mathrm{P}(\, Q > q\,)$, which we can think of as the fraction of time for which the queue-length Q is greater than q. If the queue is stable, then this probability typically decays very rapidly with increasing q, and so in Figure 9 we plot it against q on a logarithmic scale. While there is some detail to the plot for small values of q, the most striking feature is that, for q greater than about 40, it is linear in q. This means that $\mathrm{P}(\, Q > q\,)$ decays exponentially with q for large values of q:

$$\mathrm{P}(\, Q > q\,) \asymp e^{-\delta q}$$

where $-\delta$ is the asymptotic slope of Figure 9. Can this behaviour be explained using Large Deviations? The answer is "Yes": if the arrivals process and the service process are stationary and satisfy the stability condition and the workload process satisfies a Large Deviation principle

$$\mathrm{P}(\, W_t/t \approx x\,) \asymp e^{-tI(x)},$$

with rate-function I, then

$$\mathrm{P}(\, Q > q\,) \asymp e^{-\delta q}$$

and the decay-constant δ can be calculated from the rate-function for the workload process:

$$\delta = \min_x \frac{I(x)}{x}.$$

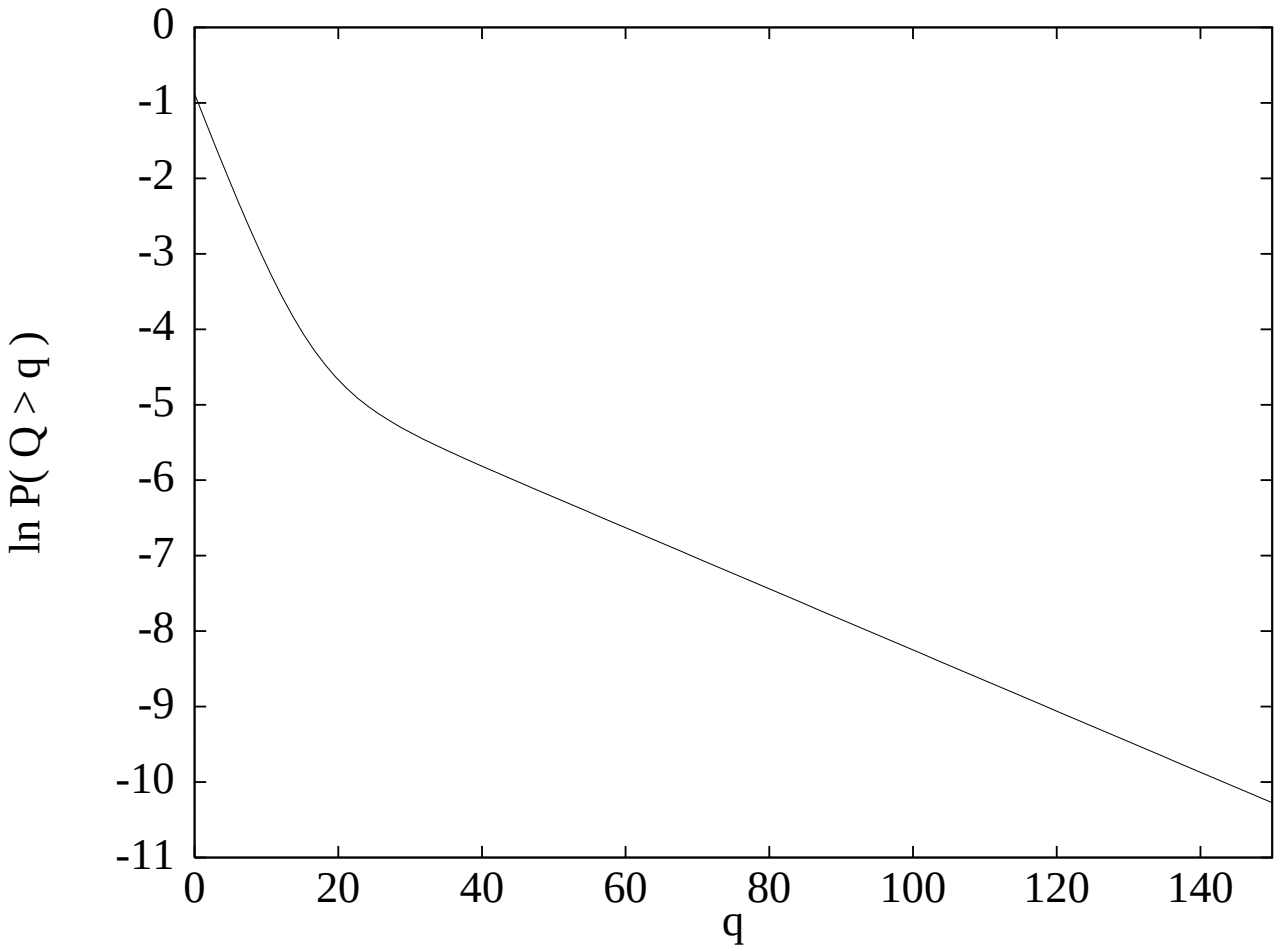

Figure 9: The logarithm of a typical queue-length distribution

We can get a feeling for why this is true from the following crude argument (a proof is given in Appendix E); first notice that

$$\begin{aligned} \mathrm{P}(\, Q > q \,) &= \mathrm{P}\left(\max_{t \geq 0} W_t > q \right) \\ &= \mathrm{P}(\, \cup_{t \geq 0} \{ W_t > q \} \,) \\ &\leq \sum_{t \geq 0} \mathrm{P}(\, W_t > q \,) . \end{aligned}$$

Since $\mathrm{P}(\, W_t/t > x \,) \asymp e^{-tI(x)}$, we have

$$\mathrm{P}(\, W_t > q \,) = \mathrm{P}(\, W_t/t > q/t \,) \asymp e^{-tI(q/t)} = e^{-q\frac{I(q/t)}{q/t}}$$

so that

$$\mathrm{P}(\, Q > q \,) \asymp e^{-q\frac{I(q)}{q}} + e^{-q\frac{I(q/2)}{q/2}} + \ldots + e^{-q\frac{I(q/t)}{q/t}} + \ldots$$

and, just as in the discussion of the Contraction Principle, the term which dominates when q is large is the one for which $I(q/t)/(q/t)$ is smallest, that is the one for which $I(x)/x$ is a minimum:

$$\mathrm{P}(\, Q > q \,) \asymp e^{-q \min_x \frac{I(x)}{x}} = e^{-q\delta}.$$

Note that we can also characterise δ in terms of the scaled CGF:

$$\begin{aligned} \theta \leq \min_x I(x)/x \quad & \text{if and only if} \quad \theta \leq I(x)/x \qquad \text{for all } x \\ & \text{if and only if} \quad \theta x - I(x) \leq 0 \qquad \text{for all } x \\ & \text{if and only if} \quad \max_x \{\theta x - I(x)\} \leq 0; \end{aligned}$$

thus $\theta \leq \delta$ if and only if $\lambda(\theta) \leq 0$ and so

$$\delta = \max \{\, \theta \,:\, \lambda(\theta) \leq 0 \,\}.$$

Using the Scaled CGF for Resource Allocation

Buffer Dimensioning

If the queue has only a finite waiting space, then δ gives us an estimate of what that buffer-size must be in order to achieve a given probability of overflow. Look again at Figure 9: it is clear that, if we require the probability of overflow to be smaller than e^{-9} (a little over 10^{-4}), we need a buffer larger than 120. We can use δ to give an estimate $\hat{P}(b)$ of the probability of a buffer of size b overflowing:

$$\hat{P}(b) = e^{-\delta b}.$$

Effective Bandwidths

Note that in the case in which the service capacity is a constant, s per unit time, the workload process is $W_t = X_{-1} + \ldots + X_{-t} - st$ and the scaled CGF is

$$\begin{aligned}
\lambda(\theta) &= \lim_{t\to\infty} \frac{1}{t} \ln \mathrm{E}\, e^{\theta W_t} \\
&= \lim_{t\to\infty} \frac{1}{t} \ln \mathrm{E}\, e^{\theta(X_{-1}+\ldots+X_{-t}-st)} \\
&= \lim_{t\to\infty} \frac{1}{t} \ln \mathrm{E}\, e^{\theta(X_{-1}+\ldots+X_{-t})} - s\theta \\
&= \lambda_A(\theta) - s\theta
\end{aligned}$$

where λ_A is the scaled CGF of the arrivals process. Thus, given the arrivals scaled CGF, we can calculate δ as a function of s:

$$\delta(s) = \max\{\, \theta \,:\, \lambda_A(\theta) \le s\theta \,\}.$$

It may happen, as in ATM networks, that the buffer-size is fixed but the capacity of the server can be varied; in that case, a natural question to ask is "Given the arrivals process A_t, what is the minimum service capacity s_p required to guarantee that the probability of buffer-overflow is less than some given level p?". We can use the approximation $\mathrm{P}(\, Q > b \,) \approx e^{-\delta(s)b}$ to give an estimate of s_p:

$$s_p = \min\{\, s \,:\, e^{-\delta(s)b} \le p \,\}.$$

Exercise 10 *Show that* $s_p = \lambda_A(\theta_p)/\theta_p$ *where* $\theta_p = (-\ln p)/b$.

s_p is, in general, larger than the mean bandwidth of the arrivals; and is known as the *effective bandwidth* of the arrivals. The function $\lambda_A(\theta)/\theta$ is known as the *effective bandwidth function* and the approximation $\mathrm{P}(\, Q > b \,) \approx e^{-\delta(s)b}$ is known as the *effective bandwidth approximation*.

Effective Bandwidths in Risk Theory

Recall the simple model of risk theory we discussed in Section 3: an insurance company settles a fixed number of claims in a fixed period of time, say one a day, and receives a steady income from premium payments, say an amount p each day. Since the sizes of the claims are random, there is the risk that, at the end of the planning period T, the total amount paid in settlement of claims will exceed the total assets of the company. When we discussed this model before, we assumed that the only asset the company has to cover

the claims is the income from premium payments; it is, however, likely that a company would have some other assets, say of fixed value u. We might now want to evaluate the risk of the amount $\sum_{t=1}^{T} X_t$ paid out over the planning period T exceeding the total assets $pT + u$ of the company. This risk-theory model is similar to a single-server queue: the claims are like customers, the premium income is like service capacity and the initial assets u are like a buffer, guarding temporarily against large claims which exceed the premium income; thus ruin (exhaustion of all assets) in the risk-theory model corresponds to buffer-overflow in the queuing system.

We assume, as before, that the sizes X_t of claims are independent and identically distributed so that we can apply Cramér's Theorem to approximate the probability of ruin.

$$\mathrm{P}\left(\frac{1}{T} \sum_{t=1}^{T} X_t > x \right) \asymp e^{-TI(x)},$$

where $x = p + u/T$. We require that the risk of ruin be e^{-r} for some large positive number r and we use the rate-function I to choose an appropriate value of x:

$$\begin{aligned} \mathrm{P}\left(\frac{1}{T} \sum_{t=1}^{T} X_t > x \right) &\approx e^{-r} \\ e^{-TI(x)} &\approx e^{-r} \\ I(x) &\approx r/T \end{aligned}$$

Since $I(x)$ is convex, it is monotonically increasing for x greater than the mean of X_t and so the equation

$$I(x) = r/T$$

has a unique solution for x; we call the unique solution x^*. We are free to choose any values of p and u such that $p + u/T = x^*$ but there is one choice which can be interpreted in terms of queuing theory. Recall that the rate-function I of the claims process can be expressed as the Legendre transform of λ, the CGF of the claim-size,

$$I(x) = \max_{\theta} \{\theta x - \lambda(\theta)\},$$

and so the requirement that $I(p + u/T) = r/T$ can be rewritten as

$$\max_{\theta} \{\theta p + \theta u/T - \lambda(\theta)\} = r/T.$$

If we denote by θ^* the value of θ which maximises this expression, then we get that

$$\theta^* p + \theta^* u/T - \lambda(\theta^*) = r/T$$

hence, by taking $p = \lambda(\theta^*)/\theta^*$ and $u = r/\theta^*$ we get a solution in which the premium rate p is the effective bandwidth of the claims process at a value θ^* of θ which gives a probability $e^{-\theta^* u} = e^{-r}$ of ruin in a buffer of size u.

Bypassing Modelling: Estimating the Scaled CGF

All we need for the scaled CGF of the arrivals to exist is for the arrivals to be stationary and mixing. If these two conditions are satisfied, we can use the scaled CGF to make predictions about the behaviour of the queue. One way to get the scaled CGF for the arrivals is to make a suitable statistical model, fit the parameters of the model to the data and then calculate the scaled CGF from the model. There are a number of problems with this approach. Firstly, real traffic streams cannot be accurately represented by simple models; any realistic model would have to be quite complex, with many parameters to be fitted to the data. Secondly, the calculation of the scaled CGF for any but the simplest model is not easy. Thirdly, even if you could find a realistic model, fit it to your data and calculate the scaled CGF, this would be a wasteful exercise: the scaled CGF is a Large Deviation object, and it does not depend on the details of the model, only on its "bulk properties". Hence all the effort you put into fitting your sophisticated model to the data is, to a large extent, lost. Our approach is to ask "Why not measure directly what you are looking for?" There are many good precedents for this approach. When engineers design a steam turbine they need to know the thermodynamic properties of steam. To find this out, they do not make a sophisticated statistical mechanical model of water and calculate the entropy from that; instead, they measure the entropy directly in a calorimetric experiment, or (more likely) they use steam tables – based on somebody else's measurements of the entropy. Now, we make the observation that *thermodynamic entropy is nothing but a rate-function*. So if you want the thermodynamics (that is, the Large Deviations) of your queuing system, why not measure directly the entropy (that is, the rate-function) of the workload process? How do we measure the rate-function – or, equivalently, the scaled CGF – of the workload process? Consider the case in which the service capacity is a constant, s per unit time; the workload process is then $W_t = X_{-1} + \ldots + X_{-t} - st$ and the scaled CGF is

$$\lambda(\theta) = \lambda_A(\theta) - s$$

where λ_A is the scaled CGF of the arrivals process. If the arrivals X_i are weakly dependent, we can approximate the scaled CGF by a finite-time cumulant generating function:

$$\lambda_A(\theta) \approx \lambda_A^{(T)}(\theta) = \frac{1}{T} \ln \mathrm{E}\, e^{\theta A_T},$$

for T sufficiently large. We can now estimate the value of the expectation by breaking our data into blocks of length T and averaging over them:

$$\hat{\lambda}_A(\theta) := \frac{1}{T} \ln \frac{1}{K} \sum_{k=1}^{k=K} e^{\theta \tilde{X}_k},$$

where the $\tilde{X}$'s are the block sums

$$\tilde{X}_1 := X_1 + \ldots + X_T, \qquad \tilde{X}_2 := X_{T+1} + \ldots + X_{2T}, \qquad \text{etc.}$$

This yields estimates of both the effective bandwidth $\hat{\lambda}(\theta)/\theta$ and, given the value of s, the asymptotic decay-rate $\hat{\delta}$ of the queue-length distribution through

$$\hat{\delta} := \max \left\{ \theta \,:\, \hat{\lambda}(\theta) \leq 0 \right\}.$$

Other Estimators

Is this the only way to estimate λ? Not at all; there are many ways to estimate it and, indeed, different methods may be appropriate for different arrivals processes. Consider what exactly we are doing when we use the estimator

$$\hat{\lambda}(\theta) = \frac{1}{T} \ln \frac{1}{K} \sum_{k=1}^{k=K} e^{\theta \tilde{X}_k}.$$

If the $\tilde{X}_k$'s are independent, then the scaled CGF of the X_i's is just the cumulant generating function of the $\tilde{X}_k$'s, and the above estimator is exactly what we need to measure the latter. Thus, when we say that we needed T to be large enough for the finite time cumulant generating function $\lambda^{(T)}$ to be a good approximation to the asymptotic one λ, we really mean that we need T to be large enough for the block-sums $\tilde{X}_k$ to be independent. Not only can we easily calculate the scaled CGF for independent sequences, but we can also do so for Markov sequences. Thus, if we use an estimator based on a Markov structure, we only need T to be large enough that the $\tilde{X}_k$'s are approximately Markov; such an estimator is tailor-made for the case of Markovian arrivals.

There is a whole class of estimators, similar to the simple I.I.D. estimator $\hat{\lambda}$ shown above, but which, instead of using one fixed size T for each block, use variable block-sizes. To see how they work, we need to consider the Large Deviations of random time-changes.

The Large Deviations of Random Time-Changes

Suppose we have a Large Deviation principle (LDP) for some process $\{S_t\}$, in that

$$\mathrm{P}(\, S_t \approx s \,) \asymp e^{-tI(s)};$$

if we take an increasing sequence $\{t_n\}$ of times such that $t_n \to \infty$, then obviously we also have a LDP for the sequence $\{S_{t_n}\}$:

$$\mathrm{P}(\, S_{t_n} \approx s \,) \asymp e^{-t_n I(s)};$$

If $t_n/n \to \tau$, then we can also write $\mathrm{P}(\, S_{t_n} \approx s \,) \asymp e^{-n\tau I(s)}$. What happens if, instead of a deterministic sequence $\{t_n\}$, we take a sequence of random times $\{T_n\}$? If $\{T_n\}$ satisfies the Weak Law of Large Numbers, so that $\lim_n \mathrm{P}(\, |T_n/n - \tau| > \epsilon \,) = 0$ for any positive number ϵ, then we might expect that

$$\mathrm{P}(\, S_{T_n} \approx s \,) \asymp e^{-n\tau I(s)}.$$

as before. What if $\{T_n\}$ satisfies a LDP? Obviously the Large Deviations of $\{T_n\}$ get mixed in with those of $\{S_t\}$. We could ask a more specific question: if we have a joint LDP for S_{T_n} and T_n, so that

$$\mathrm{P}(\, S_{T_n} \approx s,\ T_n/n \approx \tau \,) \asymp e^{-nJ(s,\tau)},$$

then how are $I(s)$ and $J(s,\tau)$ related?

To answer that, we need to look at the joint Large Deviation principle for S_t and N_t, where N_t is the *counting process* associated with T_n:

$$N_t = \sup\{n : T_n \leq t\};$$

N_t is the number of random times which have occurred up to time t. A better question is the following: if S_t and N_t satisfy a LDP jointly

$$\mathrm{P}(\, S_t \approx s,\ N_t/t \approx \nu \,) \asymp e^{-tI(s,\nu)},$$

then is it true that

$$\mathrm{P}(\, S_{T_n} \approx s,\ T_n/n \approx \tau \,) \asymp e^{-nJ(s,\tau)}$$

and, if so, what is the relationship between $I(s,\nu)$ and $J(s,\tau)$?

The answer is simple: yes (under appropriate hypotheses on the processes S_t and N_t and the rate-function I), and the relationship between the rate-functions is

$$J(s,\tau) = \frac{I(s,1/\tau)}{\tau}.$$

Consider the following scaled cumulant generating function

$$\mu(\theta) := \lim_{n\to\infty} \frac{1}{n} \ln \mathrm{E}\, e^{\theta T_n S_{T_n}};$$

an application of Varadhan's Theorem allows us to calculate μ in terms of J:

$$\begin{aligned} \mathrm{E}\, e^{\theta T_n S_{T_n}} &= \int e^{\theta n\tau s}\, d\mathrm{P}(\, S_{T_n} \approx s,\, T_n/n \approx \tau\,) \\ &\asymp \int e^{\theta n\tau s} e^{-nJ(s,\tau)}\, ds\, d\tau \\ &\asymp e^{n \max_{s,\tau}\{\theta\tau s - J(s,\tau)\}}. \end{aligned}$$

Thus $\mu(\theta) = \max_{s,\tau}\{\theta\tau s - J(s,\tau)\}$ and, rewriting J in terms of I, we have that

$$\mu(\theta) = \max_{s,\tau} \tau\{\theta s - I(s,1/\tau)\}.$$

Consider also the scaled CGF of S_t

$$\lambda(\theta) = \lim_{t\to\infty} \frac{1}{t} \ln \mathrm{E}\, e^{\theta t S_t}$$

Exercise 11 *Use Varadhan's Theorem as we did above to show that*

$$\lambda(\theta) = \max_{s,\nu}\{\theta s - I(s,\nu)\}.$$

(Hint: Apply Varadhan's Theorem to $\mathrm{E}\, e^{\theta t S_t + \alpha N_t}$ *and then set* $\alpha = 0$*.)*

Is there any connection between λ and μ? Yes, they are related as follows:

$$\begin{aligned} \max_{s,\nu}\{\theta s - I(s,\nu)\} \le 0 \quad &\text{if and only if} \quad \{\theta s - I(s,\nu)\} \le 0 \qquad \text{for all } s \text{ and all } \nu > 0 \\ &\text{if and only if} \quad \{\theta s - I(s,1/\tau)\} \le 0 \qquad \text{for all } s \text{ and all } \tau > 0 \\ &\text{if and only if} \quad \tau\{\theta s - I(s,1/\tau)\} \le 0 \qquad \text{for all } s \text{ and all } \tau > 0 \\ &\text{if and only if} \quad \max_{s,\tau} \tau\{\theta s - I(s,1/\tau)\} \le 0 \end{aligned}$$

and so $\lambda(\theta) \le 0$ if and only if $\mu(\theta) \le 0$ and hence

$$\{\, \theta \,:\, \lambda(\theta) \le 0 \,\} = \{\, \theta \,:\, \mu(\theta) \le 0 \,\}.$$

This means that, if the process S_t is the average workload W_t/t of a queue up to time t, then we can characterise the asymptotic decay rate δ of the queue-length distribution as either

$$\delta = \max\{\,\theta\,:\,\lambda(\theta) \le 0\,\} \qquad \text{or} \qquad \delta = \max\{\,\theta\,:\,\mu(\theta) \le 0\,\}$$

Let us look more closely at μ: it is defined as

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathrm{E}\, e^{\theta W_{T_n}} = \lim_{n\to\infty} \frac{1}{n} \ln \mathrm{E}\, e^{\theta(X_{-1}+\ldots+X_{-T_n}-sT_n)};$$

where s is the service rate of the queue and T_n is any increasing sequence of random times. This tells us that, when estimating δ, we are not restricted to using a sequence of fixed block-sizes to aggregate the observations of the arrivals but are free to use any sequence $\{T_n\}$ of large block-sizes. We can therefore choose blocks which reflect the structure of the arrivals process to give us estimators which have lower bias and/or variance.

A The Large Deviations of Coin-Tossing: A Bare-Hands Calculation

Another way to see Large Deviations at work is to use Stirling's Formula in conjunction with the combinatorial expressions for the tail-probabilities. Thus consider $\mathrm{P}(\,M_n < a\,)$:

$$\mathrm{P}(\,M_n < a\,) = \sum_{k=0}^{\lceil a\rceil - 1} \binom{n}{k} \frac{1}{2^n}.$$

If $a < 1/2$, then each term in the sum is bounded by ${}^n\mathrm{C}_{\lceil na\rceil}$ and so

$$\mathrm{P}(\,M_n < a\,) \le \lceil na\rceil \binom{n}{\lceil na\rceil} \frac{1}{2^n} =: A_n.$$

Now consider $\ln A_n$: we can rewrite $\ln {}^n\mathrm{C}_{\lceil na\rceil}$ as

$$-\frac{\lceil na\rceil}{n}\left(\frac{1}{\lceil na\rceil}\ln\lceil na\rceil! - \ln\lceil na\rceil\right) - \frac{\lfloor n(1-a)\rfloor}{n}\left(\frac{1}{\lfloor n(1-a)\rfloor}\ln\lfloor n(1-a)\rfloor! - \ln\lfloor n(1-a)\rfloor\right)$$
$$+\left(\frac{1}{n}\ln n! - \ln n\right) - \frac{\lceil na\rceil}{n}\ln\frac{\lceil na\rceil}{n} - \frac{\lfloor n(1-a)\rfloor}{n}\ln\frac{\lfloor n(1-a)\rfloor}{n}$$

and use Stirling's Formula

$$\lim_{n\to\infty}\left(\frac{1}{n}\ln n! - \ln n\right) = -1$$

to get

$$\lim_{n\to\infty} \frac{1}{n} \ln A_n = -a \ln a - (1-a)\ln(1-a) - \ln 2.$$

Let us look again at $\mathrm{P}(\, M_n < a\,)$: not only can we bound it above by something which decays exponentially, but we can also bound it below by something which decays exponentially at the same rate. So long as $a > 0$,

$$\mathrm{P}(\, M_n < a\,) \geq \binom{n}{\lceil na \rceil - 1} \frac{1}{2^n}.$$

Exercise 12 *Show that*

$$\lim_{n\to\infty} \frac{1}{n} \ln \binom{n}{\lceil na \rceil - 1} \frac{1}{2^n} = -a \ln a - (1-a)\ln(1-a) - \ln 2.$$

So we have, for $0 < a < 1/2$,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \ln \mathrm{P}(\, M_n < a\,) &= -a \ln a - (1-a)\ln(1-a) - \ln 2 \\ &= -I(a) \\ &= -\inf_{x<a} I(x), \end{aligned}$$

our first Large Deviation principle, established using only a little combinatorics.

B A Proof of Cramér's Theorem

Cramér's proof of his theorem was based essentially on an argument using moment generating functions and power-series expansions. After seeing the connection with the Central Limit Theorem and Chernoff's Formula, we can see how this is a natural approach to take; indeed, one of the standard proofs of the Central Limit Theorem is based on the moment generating function. However there is a more elegant argument which establishes the theorem and shows at the same time that the rate-function is convex; we present it here.

Suppose $\{\, X_n \,\}$ is a sequence of independent and identically distributed random variables; define $M_n^{(m)}$ by

$$M_n^{(m)} := \frac{1}{n} \left(X_{m+1} + \ldots + X_{m+n} \right).$$

Now

$$M_{m+n}^{(0)} = \frac{m}{m+n} M_m^{(0)} + \frac{n}{m+n} M_n^{(m)} \tag{4}$$

and so, if $M_m^{(0)} > x$ and $M_n^{(m)} > x$, then $M_{m+n}^{(0)} > x$. Hence

$$\left\{ M_m^{(0)} > x \right\} \cap \left\{ M_n^{(m)} > x \right\} \subset \left\{ M_{m+n}^{(0)} > x \right\},$$

so that

$$\mathrm{P}\Big(\left\{ M_m^{(0)} > x \right\} \cap \left\{ M_n^{(m)} > x \right\} \Big) \leq \mathrm{P}\Big(M_{m+n}^{(0)} > x \Big);$$

but, since the X_n's are independent,

$$\mathrm{P}\Big(\left\{ M_m^{(0)} > x \right\} \cap \left\{ M_n^{(m)} > x \right\} \Big) = \mathrm{P}\big(M_m^{(0)} > x \big)\, \mathrm{P}\big(M_n^{(m)} > x \big)$$

and, since the X_n's are identically distributed,

$$\mathrm{P}\big(M_n^{(m)} > x \big) = \mathrm{P}\big(M_n^{(0)} > x \big).$$

If we define $s_n(x) := \ln \mathrm{P}\Big(M_n^{(0)} > x \Big)$, then this implies that

$$s_m(x) + s_n(x) \leq s_{m+n}(x)$$

and we say that the sequence $\{ s_n \}_n$ is *super-additive*.

It can be shown that, if a sequence $\{ a_n \}_n$ is super-additive, then $\lim_n a_n/n$ exists and is equal to $\sup_n a_n/n$. (We say that $\{a_n/n\}$ is *almost monotone increasing*). Thus $\lim_n s_n(x)$ exists; let us define $I(x) = -\lim_n s_n(x)$. We have proved that there exists a function I such that

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathrm{P}(\, M_n > x \,) = -I(x),$$

where $M_n = M_n^{(0)}$. Returning to Equation 4, we see that, if $M_n^{(0)} > x$ and $M_n^{(n)} > y$, then $M_{2n}^{(0)} > (x+y)/2$; hence

$$\left\{ M_n^{(0)} > x \right\} \cap \left\{ M_n^{(n)} > y \right\} \subset \left\{ M_{2n}^{(0)} > \frac{x+y}{2} \right\},$$

so that

$$\mathrm{P}\Big(\left\{ M_n^{(0)} > x \right\} \cap \left\{ M_n^{(n)} > y \right\} \Big) \leq \mathrm{P}\left(M_{2n}^{(0)} > \frac{x+y}{2} \right).$$

Using the fact that the X_n's are independent and identically distributed, we have

$$s_n(x) + s_n(y) \leq s_{2n}\left(\frac{x+y}{2} \right)$$

so that

$$\frac{1}{2}\left\{\frac{1}{n}s_n(x)+\frac{1}{n}s_n(y)\right\}\le\frac{1}{2n}s_{2n}\left(\frac{x+y}{2}\right).$$

Since $I(x)=-\lim_n s_n(x)$ exists for all x, we have

$$\frac{1}{2}\left\{I(x)+I(y)\right\}\ge I\left(\frac{x+y}{2}\right);$$

this implies that I is convex.

C A Precise Statement of the Large Deviation Principle

We use the suggestive notation

$$\mathrm{P}(\,M_n\approx x\,)\asymp e^{-nI(x)}$$

to indicate that the sequence $\mathrm{P}(\,M_n>x\,)$ of probability distributions *satisfies a Large Deviation principle with rate-function* I; that is

(LD1) the function I is lower-semicontinuous

(LD2) for each real number a, the level set $\{\,x\in\mathbb{R}\,:\,I(x)\le a\,\}$ is compact

(LD3) for each closed subset F of $\mathbb{R}$,

$$\limsup_{n\to\infty}\frac{1}{n}\ln\mathrm{P}(\,M_n\in F\,)\le-\inf_{x\in F}I(x)$$

(LD4) for each open subset G of $\mathbb{R}$,

$$\liminf_{n\to\infty}\frac{1}{n}\ln\mathrm{P}(\,M_n\in G\,)\ge-\inf_{x\in G}I(x)$$

D The Gärtner-Ellis Theorem

We state the Gärtner-Ellis Theorem for a sequence $\{M_n\}$ of real-valued random variables; the theorem can be extended to vector-valued random variables without too much difficulty. Define

$$\lambda_n(\theta):=\frac{1}{n}\ln\mathrm{E}\,e^{n\theta M_n}$$

for $\theta\in\mathbb{R}$ and assume :

1. $\lambda_n(\theta)$ is finite for all θ;
2. $\lambda(\theta) := \lim_{n\to\infty} \lambda_n(\theta)$ exists and is finite for all θ;

then the upper bound (LD3) holds with rate-function

$$I(x) = \sup_{\theta\in\mathbb{R}}\{x\theta - \lambda(\theta)\}.$$

If, in addition, $\lambda(\theta)$ is differentiable for all $\theta \in \mathbb{R}$, then the lower bound (LD4) holds.

E Deriving the Asymptotics of the Queue-Length from the Large Deviations of the Workload

The proof of the lower bound is easy, so we give it first. The queue-length Q is related to the workload W_n by $Q = \sup_n W_n$ and so the event $\{Q > b\}$ can be expressed as

$$\{Q > b\} = \bigcup_{n\geq 0} \{W_n > b\}\,.$$

Thus, for each $n \geq 0$,

$$\{Q > b\} \supset \{W_n > b\}$$

and so

$$\mathrm{P}(\, Q > b\,) \geq \mathrm{P}(\, W_n > b\,)$$

for all $n \geq 0$. Fix $c > 0$; since $b/c \leq \lceil b/c \rceil$, we have that $c \geq \frac{b}{\lceil b/c\rceil}$ so that

$$\left\{W_{\lceil b/c\rceil} > b\right\} = \left\{\frac{1}{\lceil b/c\rceil} W_{\lceil b/c\rceil} > \frac{b}{\lceil b/c\rceil}\right\} \supset \left\{\frac{1}{\lceil b/c\rceil} W_{\lceil b/c\rceil} > c\right\}$$

and hence

$$\mathrm{P}(\, Q > b\,) \geq \mathrm{P}\big(\, W_{\lceil b/c\rceil} > b\,\big) \geq \mathrm{P}\left(\, \frac{1}{\lceil b/c\rceil} W_{\lceil b/c\rceil} > c\,\right).$$

Since $b/c \geq \lfloor b/c \rfloor$, we have $\frac{1}{b} \leq \frac{1}{c}\cdot\frac{1}{\lfloor b/c\rfloor}$ so that

$$\frac{1}{b}\ln \mathrm{P}(\, Q > b\,) \geq \frac{1}{c}\cdot\frac{1}{\lfloor b/c\rfloor}\ln \mathrm{P}\left(\, \frac{1}{\lceil b/c\rceil} W_{\lceil b/c\rceil} > c\,\right).$$

(Remember that, for any event A, $\mathrm{P}(\,A\,) \le 1$ so that $\ln \mathrm{P}(\,A\,) \le 0$!) It follows that, for each $c > 0$, we have

$$\begin{aligned}
\liminf_{b\to\infty} \frac{1}{b} \ln \mathrm{P}(\,Q > b\,) &\ge \frac{1}{c} \liminf_{b\to\infty} \frac{1}{\lfloor b/c \rfloor} \ln \mathrm{P}\left(\frac{1}{\lceil b/c \rceil} W_{\lceil b/c \rceil} > c \right) \\
&= \frac{1}{c} \liminf_{n\to\infty} \frac{1}{n} \ln \mathrm{P}\left(\frac{1}{n} W_n > c \right) \\
&\ge -\frac{1}{c} I(c),
\end{aligned}$$

where in the last step we used (LD4). But this holds for all $c > 0$, so that we have the lower bound

$$\begin{aligned}
\liminf_{b\to\infty} \frac{1}{b} \ln \mathrm{P}(\,Q > b\,) &\ge \sup_{c>0} (-\frac{1}{c} I(c)) \\
&= -\inf_{c>0} \frac{1}{c} I(c).
\end{aligned}$$

The get the upper bound from a general result which follows from putting together the Chernoff bound and the Principle of the Largest Term. Let $\{\, W_n \,\}_{n\ge 0}$ be a sequence of random variables such that, for each real number θ, $\mathrm{E}\, e^{\theta W_n}$ is finite for all n and $\lambda(\theta) := \lim_n \lambda_n(\theta)$ exists and is finite, where

$$\lambda_n(\theta) := \frac{1}{n} \ln \mathrm{E}\, e^{\theta W_n}.$$

Define $\delta := \sup \{\, \theta > 0 \,:\, \lambda(\theta) < 0 \,\}$; then

$$\limsup_{b\to\infty} \frac{1}{b} \ln \mathrm{P}\left(\sup_{n\ge 0} W_n > b \right) \le -\delta.$$

Proof: If the set $\{\, \theta > 0 \,:\, \lambda(\theta) < 0 \,\}$ is empty, then $\delta = -\infty$ and there is nothing to prove. Otherwise, choose $\bar\theta$ such that $0 < \bar\theta < \delta$ and $\lambda(\bar\theta) < 0$. Then, using the Chernoff Bound, we have for each integer n

$$\mathrm{P}(\,W_n > b\,) \le e^{-\bar\theta b} \mathrm{E}\, e^{\bar\theta W_n};$$

since $\mathrm{E}\, e^{\bar\theta W_n} < \infty$, we have

$$\limsup_{b\to\infty} \frac{1}{b} \ln \mathrm{P}(\,W_n > b\,) \le \bar\theta.$$

Thus, for each integer N, we have

$$\mathrm{P}\left(\sup_{n\le N} W_n > b \right) \le \sum_{n\le N} \mathrm{P}(\,W_n > b\,) \le N \sup_{n\le N} \mathrm{P}(\,W_n > b\,)$$

so that

$$\limsup_{b\to\infty} \frac{1}{b} \ln \mathrm{P}\left(\sup_{n\le N} W_n > b \right) \le \sup_{n\le N} \limsup_{b\to\infty} \frac{1}{b} \ln \mathrm{P}(\, W_n > b\,) \le \bar{\theta}.$$

On the other hand, we have

$$\mathrm{P}\left(\sup_{n>N} W_n > b \right) \le \sum_{n>N} \mathrm{P}(\, W_n > b\,) \le e^{-\bar{\theta}b} \sum_{n>N} \mathrm{E}\, e^{\bar{\theta}W_n} = e^{-\bar{\theta}b} \sum_{n>N} e^{n\lambda_n(\bar{\theta})}.$$

Since $\lim_n \lambda_n(\bar{\theta}) = \lambda(\bar{\theta}) < 0$, there exists a positive number ϵ such that $\epsilon < -\lambda(\bar{\theta})$ and an integer $N_{\bar{\theta}}$ such that $\lambda_n(\bar{\theta}) < -\epsilon$ for every $n > N_{\bar{\theta}}$. It follows that

$$\mathrm{P}\left(\sup_{n>N_{\bar{\theta}}} \right) \le e^{-\bar{\theta}b} \sum_{n>N_{\bar{\theta}}} e^{-n\epsilon} < e^{-\bar{\theta}b} \cdot \frac{1}{1-e^{-\epsilon}}.$$

Applying the Principle of the Largest Term once more, we have

$$\limsup_{b\to\infty} \frac{1}{b} \ln \mathrm{P}\left(\sup_{n\ge 0} W_n > b \right) =$$
$$\max\left\{ \limsup_{b\to\infty} \frac{1}{b} \ln \mathrm{P}\left(\sup_{n\le N_{\bar{\theta}}} W_n > b \right), \limsup_{b\to\infty} \frac{1}{b} \ln \mathrm{P}\left(\sup_{n>N_{\bar{\theta}}} W_n > b \right) \right\}$$

so that

$$\limsup_{b\to\infty} \frac{1}{b} \ln \mathrm{P}\left(\sup_{n\ge 0} W_n > b \right) \le -\bar{\theta};$$

but this is true for all $\bar{\theta} < \delta$ and so the result follows:

$$\limsup_{b\to\infty} \frac{1}{b} \ln \mathrm{P}\left(\sup_{n\ge 0} W_n > b \right) \le -\delta.$$

□