

Proof of Lemma 1.7 First, we show (1.14)=(1.13). The latter is

$$\begin{aligned} (1.13) &= \inf_{t \in \mathbb{R}^+} t \left(\sup_{\theta \in \mathbb{R}} \theta \left(C + \frac{q}{t} \right) - \Lambda(\theta) \right) \\ &= \inf_{t > 0} \sup_{\theta \in \mathbb{R}} \theta(q + Ct) - t\Lambda(\theta). \end{aligned}$$

Since $EA < C$, $EA < (q + Ct)/t$ and so by Lemma 2.6 we can restrict the supremum to be over $\theta \geq 0$, yielding (1.14)=(1.13).

Now we show (1.14) \geq (1.15). For any $\theta > 0$ such that $\Lambda(\theta) < \theta C$, and $t \in \mathbb{R}^+$,

$$\theta(q + Ct) - t\Lambda(\theta) = \theta q + t(\theta C - \Lambda(\theta)) \geq \theta q.$$

Taking the supremum over such θ ,

$$\sup_{\theta > 0: \Lambda(\theta) < \theta C} \theta(q + Ct) - t\Lambda(\theta) \geq q \sup_{\theta > 0: \Lambda(\theta) < \theta C} \theta$$

and so by relaxing the left hand side

$$\sup_{\theta \geq 0} \theta(q + Ct) - t\Lambda(\theta) \geq q \sup\{\theta > 0 : \Lambda(\theta) < \theta C\}.$$

Since the right hand side does not depend on t , taking the infimum over t yields the result.

Finally, we show (1.14) \leq (1.15). Let $\theta^* = \sup\{\theta > 0 : \Lambda(\theta) < \theta C\}$. If $\theta^* = \infty$ there is nothing to prove. So assume $\theta^* < \infty$. We will see in Lemma Lemma 2.3 that $\Lambda(\theta)$ is convex, and also that (from our assumption that it is finite everywhere) it is continuous and differentiable everywhere. It must then be that $\Lambda(\theta^*) = \theta^*C$ and $\Lambda'(\theta^*) > C$ (see the sketch in Figure 1.1 to convince yourself of this).

Since $\Lambda(\theta)$ is convex, it is bounded below by the tangent at θ^* :

$$\Lambda(\theta) \geq \theta^*C + \Lambda'(\theta^*)(\theta - \theta^*).$$

We will use this to bound (1.14):

$$\begin{aligned} (1.14) &= \inf_{t > 0} \sup_{\theta \geq 0} \theta(q + Ct) - t\Lambda(\theta) \\ &\leq \inf_{t > 0} \sup_{\theta \geq 0} \theta(q + Ct) - t \left(\theta^*C + \Lambda'(\theta^*)(\theta - \theta^*) \right) \\ &= \inf_{t > 0} \sup_{\theta \geq 0} \theta(q - t(\Lambda'(\theta^*) - C)) + \theta^*t(\Lambda'(\theta^*) - C). \end{aligned}$$

Figure 1.1 and caption should say θ^* not $\hat{\theta}$

path ‘discontinuities’ (sometimes referred to as overshoot) are invisible at the macroscopic scale. (Note however that these effects contribute to the value of $I(q)$.)

Example 1.7

Consider again the system in Example 1.3. This has $C = 1$, and log moment generating function for A given by

$$\Lambda(\theta) = \log(1 - p + pe^{2\theta}).$$

This gives

$$\begin{aligned} I(q) &= q \sup\{\theta > 0 : \log(1 - p + pe^{2\theta}) < \theta\} \\ &= \log\left(\frac{1 + \sqrt{1 - 4p(1-p)}}{2p}\right) \\ &= \log\frac{1-p}{p} \end{aligned}$$

which agrees with our earlier conclusion. \diamond

Exercise 1.8

If the service is a random variable, say C_t , we can apply the theorem to the random variable $A_t - C_t$ (rather than to A_t) and set $C = 0$. Then

$$\Lambda(\theta) = \Lambda_A(\theta) + \Lambda_C(-\theta)$$

$$\Lambda(\theta) = \Lambda_A(\theta) - \Lambda_C(\theta)$$

where Λ_A and Λ_C are the log moment generating functions for the A_t and C_t . Show that $\Lambda^*(x) = \inf_y \Lambda_A^*(y) + \Lambda_C^*(y - x)$. Compute $I(q)$ for the following examples:

- i. A_t are Poisson random variables with mean λ and C_t are independent Poisson random variables with mean $\mu > \lambda$,
- ii. A_t are exponential random variables with mean λ and C_t are independent exponential random variables with mean $\mu > \lambda$,
- iii. A_t are Gaussian random variables with mean μ and variance σ^2 , and $C_t = C > \mu$. \diamond

1.4 Application to queues with many sources

There is another limiting regime, which considers what happens when a queue is shared by a large number of independent traffic flows (also called sources).

Chapter 2

Large deviations in Euclidean spaces

In Section 1.2 we alluded to Cramér's Theorem, a result about large deviations for averages of random variables in \mathbb{R} . In this chapter we will give a proof, and a generalisation, and explore some consequences. This chapter does not mention queues! The presentation given here and in Chapter 4 owes much to the book of Dembo and Zeitouni [25]. Other good sources include the books of Deuschel and Stroock [28] and den Hollander [26].

2.1 Some examples

First, a couple of examples.

Example 2.1

Let L_n , $n \in \mathbb{N}$, denote the proportion of heads in n independent tosses of a biased coin, which has probability p of coming up heads. Say n is large and that we are interested in the probability that L_n exceeds q , for some $q > p$. For notational convenience, suppose that qn is an integer. Since nL_n has a binomial distribution, we see that

$$P(L_n > q) = \sum_{k=qn}^n \binom{n}{k} p^k (1-p)^{n-k}. \quad (2.1)$$

It is straightforward to check that the largest term in the above sum corresponds to $k = qn$. Indeed, for any $j > qn > pn$,

$$\binom{n}{j+1} p^{j+1} (1-p)^{n-(j+1)} \Big/ \binom{n}{j} p^j (1-p)^{n-j} = \frac{n-j}{j} \frac{p}{1-p} < 1.$$

The probability we're after is $P(L_n \geq q)$, not $P(L_n > q)$. The right hand side of the bound on the ratio of successive terms should be

$$\frac{n-j}{j+1} \frac{p}{1-p}$$

which is less than

$$\frac{n-j}{j} \frac{p}{1-p}.$$

Thus

$$\binom{n}{qn} p^{qn} (1-p)^{(1-q)n} \leq P(L_n > q) \leq (1-q)n \binom{n}{qn} p^{qn} (1-p)^{(1-q)n}.$$

We can use Stirling's formula to simplify the above expression. Ignoring terms that are in subexponential in n , we get

$$P(L_n > q) \approx \exp(-nH(q; p)),$$

where $H(q; p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}$. This quantity is called the relative entropy, or Kullback-Leibler divergence, of the probability distribution $(q, 1-q)$ with respect to the probability distribution $(p, 1-p)$. A similar expression can be obtained for $P(L_n < q)$ when $q < p$. \diamond

There are two key points to note from this derivation.

i. For all sets A in some class (here $A \in \{(q, 1], [0, q)\}$), and a sequence of random variables L_n , we have $P(L_n \in A) \approx \exp(-nI(A))$, where $I(\cdot)$ is some set function. Large deviation theory deals with probability approximations of precisely this form: given a parametrised family of random variables or their probability laws, these are approximated by a term that is exponential in the parameter. In our example, the parameter space was the natural numbers, but it is equally easy to deal with an uncountable parameter set, such as the positive reals.

ii. A single term in the sum in (2.1), namely the term with $k = qn$, is sufficient to determine the correct exponential decay rate in n of this sum. Since it is only this decay rate that we are interested in, we can replace the sum by the largest term. It turns out this feature is characteristic of many situations where the theory of large deviations is applicable. See Section 2.2 for more.

The random variable L_n considered earlier is nothing but the average of n i.i.d. Bernoulli random variables X_i , with $P(X_1 = 1) = p = 1 - P(X_1 = 0)$, and this made the calculation easy. Here is another example where the calculation is also easy: the average is of normal random variables.

Example 2.2

Let Y_i be an i.i.d. sequence of normal random variables with zero mean and unit variance, and let $S_n = Y_1 + \dots + Y_n$. The sample mean S_n/n is also normally distributed, with mean zero and variance $1/n$. Thus, for any $x > 0$,

Replace 2π by $2\pi/n$ in the estimates for S_n/n .

and $a > 0$ and $b > 0$. By the following elementary lemma,

$$\frac{1}{n} \log P(A_n \cup B_n) \rightarrow -(a \wedge b).$$

Lemma 2.1 (Principle of the largest term) *Let a_n and b_n be sequences in \mathbb{R}^+ . If $n^{-1} \log a_n \rightarrow a$ and $n^{-1} \log b_n \rightarrow b$ then $n^{-1} \log(a_n + b_n) \rightarrow a \vee b$. (This extends easily to finite sums.)*

The principle of the largest term is often expressed in the probability context by the phrase *rare events occur in the most likely way*. The event $A_n \cup B_n$ is rare, in that $a \wedge b > 0$, and

$$P(A_n | A_n \cup B_n) = \frac{P(A_n)}{P(A_n) + P(B_n)} \rightarrow \begin{cases} 1 & \text{if } a < b \\ 0 & \text{if } a > b \end{cases}$$

An extension of the principle of the largest term, which we will need for various estimates in later chapters, is this.

Lemma 2.2 *Let a_n and b_n be sequences in \mathbb{R}^+ . Then*

In fact, there is equality in the lim sup case.

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log(a_n + b_n) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log(a_n) \vee \limsup_{n \rightarrow \infty} \frac{1}{n} \log(b_n),$$

and

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log(a_n + b_n) \geq \liminf_{n \rightarrow \infty} \frac{1}{n} \log(a_n) \vee \liminf_{n \rightarrow \infty} \frac{1}{n} \log(b_n).$$

(This extends easily to finite sums.)

If the gods of probability are being kind, as they are in Lemma 1.10, this can extend to infinite sums.

Exercise 2.3

Prove Lemma 2.2. ◇

2.3 Large deviations principle

Now it is time to state what we mean by a large deviations principle in \mathbb{R}^d . Look back at theorem 1.3, Cramér's theorem, which we used in Chapter 1 to derive expressions for the tail of the queue length distribution.

Cramér's theorem can be rephrased in terms of the following definition—in the notation of Section 1.2, the theorem says that S_n/n satisfies a large

demystify the theory and develop some feel for it by seeing at least one proof worked out in detail. Second, the techniques used here for deriving upper and lower bounds are of wider applicability and can be used to derive bounds fairly easily even in situations where it may be quite hard to establish an LDP.

Theorem 2.8 *Let $(X_n, n \in \mathbb{N})$ be a sequence of independent random variables each distributed like X , and let $S_n = X_1 + \cdots + X_n$. Let $\Lambda(\theta) = \log Ee^{\theta X}$, and let Λ^* be the convex conjugate of Λ . Suppose that Λ is finite in a neighbourhood of zero. Then the sequence of random variables $(S_n/n, n \in \mathbb{N})$, satisfies an LDP in \mathbb{R} with good convex rate function Λ^* .*

Proof. We first establish the large deviations upper bound (2.3) for closed half-spaces, i.e. sets of the form $[x, \infty)$ and $(-\infty, x]$. We then extend it to all closed sets. We then establish the large deviations lower bound (2.2). Finally we show that Λ^* is a good convex rate function.

Upper bound for closed half-spaces. Applying Chernoff's bound,

$$P\left(\frac{S_n}{n} \in [x, \infty)\right) \leq e^{-n\theta x} Ee^{\theta S_n} = e^{-n\theta x} (Ee^{\theta X})^n \quad \text{for all } \theta \geq 0.$$

Taking logarithms, for $x \geq EX$,

$$\begin{aligned} \log P\left(\frac{S_n}{n} \in [x, \infty)\right) &\leq -\sup_{\theta \geq 0} \theta x - \Lambda(\theta) \\ &= -\Lambda^*(x) \quad \text{by (2.7)} \\ &= -\inf_{y \in [x, \infty)} \Lambda^*(y) \end{aligned}$$

Should be $n^{-1} \log P(\cdot)$
not plain $\log P(\cdot)$.

where the last equality is by the monotonicity of Λ^* on $[EX, \infty)$, shown in Lemma 2.6. On the other hand, if $x < EX$, then trivially

$$\frac{1}{n} \log P\left(\frac{S_n}{n} \in [x, \infty)\right) \leq 0 = -\Lambda^*(EX) = -\inf_{y \in [x, \infty)} \Lambda^*(y).$$

The proof of the LD upper bound for sets of the form $(-\infty, x]$ follows by considering the random variable $-X$.

LD upper bound for general closed sets. Let F be an arbitrary closed set. If F contains EX , then the LD upper bound holds trivially since

$$\inf_{x \in F} \Lambda^*(x) = \Lambda^*(EX) = 0.$$

Otherwise, F can be written as the union $F = F_1 \cup F_2$ where F_1 and F_2 are closed and

$$F_1 \subseteq [EX, \infty) \text{ and } F_2 \subseteq (-\infty, EX).$$

Suppose F_1 is non-empty, and let x be the infimum of F_1 . By closure, $x \in F_1$. Now,

$$\begin{aligned} \frac{1}{n} \log P\left(\frac{S_n}{n} \in F_1\right) &\leq \frac{1}{n} \log P\left(\frac{S_n}{n} \in [x, \infty)\right) \\ &\leq -\Lambda^*(x) \quad \text{by the upper bound for closed half-spaces} \\ &= -\inf_{y \in F_1} \Lambda^*(y) \end{aligned}$$

where the last equality is by monotonicity of Λ^* on $[EX, \infty)$, in which F_1 is contained. Similarly, by considering the LD upper bound for $(-\infty, x]$, where x is the supremum of F_2 , we obtain

$$\frac{1}{n} \log P\left(\frac{S_n}{n} \in F_2\right) \leq -\inf_{y \in F_2} \Lambda^*(y).$$

In other words, the LD upper bound holds for both of F_1 and F_2 . Hence, by the principle of the largest term, it holds for $F = F_1 \cup F_2$.

LD lower bound. Let G be any open set, and let $x \in G$. We will show that

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P\left(\frac{S_n}{n} \in G\right) \geq -\Lambda^*(x). \quad (2.9)$$

Taking the supremum over $x \in G$ will then yield the large deviations lower bound. We will proceed by calculating the value of $\Lambda^*(x)$. We will do this in two cases: first the case when $P(X < x) = 0$ or $P(X > x) = 0$, second the case when neither holds.

Suppose that $P(X < x) = 0$. We can calculate Λ^* explicitly as follows:

$$\begin{aligned} &= \sup_{\theta \in \mathbb{R}} \theta x - \Lambda(\theta) \\ &= -\inf_{\theta \in \mathbb{R}} \log E e^{\theta(X-x)} \\ &= -\lim_{\theta \rightarrow -\infty} \log E e^{\theta(X-x)} \\ &= -\log E 1_{X=x} \\ &= -\log P(X = x). \end{aligned}$$

$$\begin{aligned} \Lambda^*(x) &= \sup_{\theta \in \mathbb{R}} \theta x - \Lambda(\theta) \\ &= -\inf_{\theta \in \mathbb{R}} \log E e^{\theta(X-x)} \\ &= -\lim_{\theta \rightarrow -\infty} \log E e^{\theta(X-x)} \quad \text{since } X \geq x \text{ almost surely} \\ &= -\log E 1_{X=x} \quad \text{by monotone convergence} \\ &= -\log P(X = x). \end{aligned}$$

If $P(X = x) = 0$, then the lower bound in (2.9) is trivial. If $P(X = x) = p > 0$ then

$$\begin{aligned} \frac{1}{n} \log P\left(\frac{S_n}{n} \in (x - \delta, x + \delta)\right) &\geq \frac{1}{n} \log P(X_1 = \cdots = X_n = x) \\ &= \frac{1}{n} \log p^n = \log p \end{aligned}$$

and so (2.9) is also satisfied. If $P(X > x) = 0$, a similar argument shows that the large deviations lower bound holds.

Assume now that $P(X > x) > 0$ and $P(X < x) > 0$. Again, we investigate the value of the lower bound:

$$\begin{aligned} \Lambda^*(x) &= \sup_{\theta \in \mathbb{R}} \theta x - \Lambda(\theta) \\ &= - \inf_{\theta \in \mathbb{R}} \Lambda(\theta) - \theta x = - \inf_{\theta \in \mathbb{R}} \log E e^{\theta(X-x)}. \end{aligned}$$

Now, the function $g(\theta) = \Lambda(\theta) - \theta x$ satisfies $g(\theta) \rightarrow \infty$ as $|\theta| \rightarrow \infty$, by the assumption that there is probability mass both above and below x ; and it inherits lower-semicontinuity from Λ . Any set of the form $\{g(\theta) \leq \alpha\}$ is thus bounded as well as closed, hence compact, and so g attains its infimum, say

$$\Lambda^*(x) = \hat{\theta}x - \Lambda(\hat{\theta}).$$

We will use $\hat{\theta}$ to estimate the probability in question.

We will do this using a *tilted distribution*. Let μ be the measure of X , and define a tilted measure $\tilde{\mu}$ by

$$\frac{d\tilde{\mu}}{d\mu}(x) = e^{\hat{\theta}x - \Lambda(\hat{\theta})}.$$

Let \tilde{X} be a random variable drawn from $\tilde{\mu}$. Observe that

$$\begin{aligned} E\tilde{X} &= \int x \tilde{\mu}(dx) = \int x e^{\hat{\theta}x - \Lambda(\hat{\theta})} \mu(dx) \\ &= EX e^{\hat{\theta}X - \Lambda(\hat{\theta})} = \Lambda'(\hat{\theta}) \end{aligned}$$

where the last equality comes from Lemma 2.3, making the assumption that Λ is differentiable at $\hat{\theta}$. (We will leave the case where it is not differentiable to later.) Note also that, by optimality of $\hat{\theta}$ in $\Lambda^*(x)$, $\Lambda'(\hat{\theta}) = x$. Thus $E\tilde{X} = x$. (This tilted random variables captures the idea of being close in distribution to X , conditional on having a value close to x .)

Clearer to write $S_n/n \in G$ rather than $S_n/n \in (x - \delta, x + \delta)$.

Some remarks on the proof.

- i. It is clear from the proof that the upper bound

$$\frac{1}{n} \log P\left(\frac{S_n}{n} \in F\right) \leq - \inf_{x \in F} \Lambda^*(x)$$

holds for all closed intervals $F \subseteq \mathbb{R}$ and all n , not just on a logarithmic scale in the limit as $n \rightarrow \infty$. This follows from the corresponding upper bound for half-spaces, which is known as Chernoff's bound. Since Chernoff's bound also holds for half-spaces in \mathbb{R}^d , it holds for all convex subsets of \mathbb{R}^d , as these are the intersection of half-spaces.

ii. The lower bound is local (the bound for open balls implies the bound for all open sets) and its proof uses a change of measure argument. Both these ideas are applicable in more abstract settings, not just in \mathbb{R} . They often yield easy lower bounds, even if these aren't tight or can't easily be turned into a full large deviation principle.

iii. The measure $\tilde{\mu}$ is called an exponential tilting of the measure μ , with tilt parameter $\hat{\theta}$. In order to derive a bound on the probability that the sample mean lies in $(x - \delta, x + \delta)$, we seek a tilt parameter $\hat{\theta}$ that makes the mean of the tilted distribution equal to x . If $\hat{\theta}$ lies at the boundary of the effective domain of Λ , then the tilted distribution may not have a mean, so this method is not applicable. The tilted measure $\tilde{\mu}$ is not just a convenient tool for a proof. It also tells us the *most likely way* by which the mean of a large sample turns out to be close to x . More precisely, conditional on the sample mean S_n/n being in $(x - \delta, x + \delta)$, the empirical distribution of X_1, \dots, X_n approaches $\tilde{\mu}$ as $n \rightarrow \infty$.

Cramér's theorem is applicable to random variables for which the origin may not be in the interior of the effective domain of Λ , with the modification that the rate function need not be good. The theorem also holds for \mathbb{R}^d -valued random variables, with the modification that Λ is defined on \mathbb{R}^d as $\Lambda(\theta) = \log E(e^{\theta \cdot X})$. Proofs of these results can be found in the book of Dembo and Zeitouni [25].

The following exercise will test whether you have understood the proof of Cramér's theorem. The result is also handy in understanding Chapter 6 on large-buffer scalings.

Exercise 2.11

Let $(X^N/N, N \in \mathbb{N})$ satisfy a large deviations principle in \mathbb{R} with convex rate function I . Let α be a positive real number. Show that $(X^{\lfloor \alpha N \rfloor}, N \in \mathbb{N})$ satisfies a large deviations principle in \mathbb{R} with rate function $J(x) = \alpha I(x/\alpha)$.

Should be $X^{\lfloor \alpha N \rfloor}/N$,
not $X^{\lfloor \alpha N \rfloor}$.

Proof. Let $A = \{a_1, \dots, a_d\}$. Observe that L_n is the sample mean of Z_1, \dots, Z_n , where $Z_i = (1[X_i = a_1], \dots, 1[X_i = a_d])$ is an \mathbb{R}^d -valued random variable. Moreover, $(Z_i, i \in \mathbb{N})$ are i.i.d., and the cumulant generating function of Z_1 is given, for $\theta \in \mathbb{R}^d$, by

$$\Lambda(\theta) = \log E e^{\theta \cdot Z_1} = \log \sum_{i=1}^d \mu(a_i) e^{\theta_i}, \quad (2.10)$$

which is finite for all θ . Hence, by Cramér's theorem, L_n satisfies the LDP in \mathbb{R}^d with the good convex rate function Λ^* , which is the convex conjugate of Λ .

The rest of the proof is just finding an explicit form of Λ^* . We will first show that

$$\Lambda^*(\nu) = \sum_{a \in A} \nu(a) \log \frac{\nu(a)}{\mu(a)} \quad \text{for } \nu \in M_1(A) \text{ with } \nu(a) > 0 \text{ for all } a \in A. \quad (2.11)$$

From (2.10), Λ is differentiable, with gradient

$$(\nabla \Lambda(\theta))_i = \mu(a_i) e^{\theta_i - \Lambda(\theta)}.$$

(Thus $\nabla \Lambda(\theta)$ is a probability distribution on A , and in fact corresponds to an exponential tilting of μ .) Pick $\nu \in M_1(A)$ and suppose first that $\nu(a) > 0$ for all $a \in A$. We can find $\theta \in \mathbb{R}^d$ such that

$$\nu = \nabla \Lambda(\theta) :$$

just take $\theta_i = \log \nu(a_i) / \mu(a_i)$. Then, by Lemma 2.4,

$$\Lambda^*(\nu) = \theta \cdot \nu - \Lambda(\theta) = \sum_{a \in A} \nu(a) \log \frac{\nu(a)}{\mu(a)}.$$

Next we deal with the case where $\nu(a) = 0$ for some $a \in A$. Let $\nu^k \rightarrow \nu$ with $\nu^k(a) > 0$ for all $a \in A$. By lower-semicontinuity of Λ^* ,

$$\Lambda^*(\nu) \leq \liminf_{k \rightarrow \infty} \Lambda^*(\nu^k) = \sum_{a \in A} \nu(a) \frac{\nu(a)}{\mu(a)}$$

(using the convention that $0 \log 0 = 0$). For the reverse inequality, choose θ^k such that $\theta_i^k = \log \nu(a_i) / \mu(a_i)$ if $\nu(a_i) > 0$ and $\theta_i^k = -k$ otherwise. Then

$$\begin{aligned} \Lambda^*(\nu) &= \sup_{\theta} \theta \cdot \nu - \Lambda(\theta) \geq \limsup_{k \rightarrow \infty} \theta^k \cdot \nu - \Lambda(\theta^k) \\ &= \sum_{a \in A} \nu(a) \frac{\nu(a)}{\mu(a)}. \end{aligned}$$

In both bounds for $\Lambda^*(\nu)$ we should have

$$\sum_{a \in A} \nu(a) \log \frac{\nu(a)}{\mu(a)}$$

Next, suppose $\nu \notin M_1(A)$ and that $\nu(a_i) < 0$ for some i . Choose θ^k by taking $\theta_i^k = -k$ and $\theta_j = 0$ for $j \neq i$. Then it is easy to check that $\Lambda(\theta^k) \leq 0$, and hence deduce

$$\Lambda^*(\nu) \geq \theta^k \cdot \nu - \Lambda(\theta^k) \geq -k\nu(a_i)$$

which $\rightarrow \infty$ as $k \rightarrow \infty$.

Finally, suppose $\nu \notin M_1(A)$ and that $\nu(a) \geq 0$ for all $a \in A$. Then $\sum_a \nu(a) \neq 1$. Choose $\theta^{(k,c)}$ by

Should have a_i not a in the equation for $\theta_i^{(k,c)}$.

$$\theta_i^{(k,c)} = \begin{cases} c + \log \nu(a) / \mu(a) & \text{if } \nu(a) > 0 \\ -k & \text{if } \nu(a) = 0 \end{cases}$$

for some constant k and c , to be specified. Then

Should be $\{a : \nu(a) = 0\}$ not $\nu(\{a : \nu(a) = 0\})$.

$$\begin{aligned} \Lambda^*(\nu) &\geq \theta^{(k,c)} \cdot \nu - \Lambda(\theta^{(k,c)}) \\ &= \sum_{a \in A} \nu(a) \log \frac{\nu(a)}{\mu(a)} + c\nu(A) - \log \left(e^c \nu(A) + e^{-k} \nu(\{a : \nu(a) = 0\}) \right). \end{aligned}$$

If $\nu(A) = 0$ then taking $k \rightarrow \infty$ we see that $\Lambda^*(\nu) = \infty$. If $\nu(A) > 1$ then taking $c \rightarrow \infty$ while keeping k fixed we see that $\Lambda^*(\nu) = \infty$. If $0 < \nu(A) < 1$ then letting $n = -2c$ and taking $c \rightarrow \infty$ we see that $\Lambda^*(\nu) = \infty$. \square

This is an LDP for L_n in \mathbb{R}^d , because that is what we get by applying Cramér's theorem. Since the L_n live in $M_1(A)$, i.e., $P(L_n \in M_1(A)) = 1$ for all n , it follows by the large deviation lower bound that the infimum of I on the open set $\mathbb{R}^d \setminus M_1(A)$ must be infinite, as verified by the theorem. It is natural to expect that $(L_n, n \in \mathbb{N})$ also satisfies the LDP in $M_1(A)$, with the same rate function $H(\cdot|\nu)$. This is indeed the case.

Lemma 2.10 *The sequence of random variables $(L_n, n \in \mathbb{N})$, satisfies the LDP on $M_1(A)$ with rate function $H(\cdot|\mu)$, which is continuous on $M_1(A)$ and strictly convex.*

Sketch proof. This is because $M_1(A)$ is a closed subset of \mathbb{R}^d and the rate function $I(\nu)$ is infinite outside $M_1(A)$. This is simple to prove, by writing out the large deviations bounds; alternatively see the abstract result Lemma 4.9. It is straightforward to verify continuity and strict convexity. \square

Example 2.12

Let A be a finite subset of \mathbb{R} , and as usual let $S_n = X_1 + \dots + X_n$, where the X_i are i.i.d. random variables taking values in A . What is the most likely

v. If the limit (2.12) exists then Λ , being the pointwise limit of convex functions, is itself convex. However, although it is the pointwise limit of functions which are lower-semicontinuous and differentiable in the interior of their effective domain, it does not necessarily satisfy these two conditions, and they must be part of the assumption of the theorem.

We illustrate the theorem with a couple of examples.

Example 2.15 (Additive functionals of Markov chains)

Let $(\xi_n, n \in \mathbb{N})$ be an irreducible Markov chain, taking values in a finite set E , with transition matrix P and invariant distribution π . Let f be a function from E to \mathbb{R} and define $X_n = f(\xi_n)$, $S_n = X_1 + \dots + X_n$. We will show that the sequence S_n/n satisfies an LDP and compute the rate function. For $i \in E$, define $v_n(i) = E[e^{\theta S_n} | \xi_1 = i]$. We have

$$\begin{aligned} v_n(i) &= e^{\theta f(i)} E[e^{\theta(X_2 + \dots + X_n)} | \xi_1 = i] \\ &= e^{\theta f(i)} \sum_{j \in E} p_{ij} E[e^{\theta(X_2 + \dots + X_n)} | \xi_2 = j]. \end{aligned}$$

Let $Q(\theta)$ denote the $E \times E$ matrix whose ij^{th} entry is $e^{\theta f(i)} p_{ij}$, and let v_n be the column vector whose i^{th} entry is $v_n(i)$. We can now rewrite the equation above as $v_n = Q(\theta)v_{n-1}$. Hence, $v_n = Q(\theta)^n v_0$, where v_0 is the $|E|$ -dimensional vector of ones. Let $\rho(\theta)$ denote the spectral radius of the non-negative irreducible matrix $Q(\theta)$. By the Perron-Frobenius theorem, $\rho(\theta)^{-n} v_n$ converges to (a scaled version) of the eigenvector of $Q(\theta)$ corresponding to the eigenvalue $\rho(\theta)$. Since this eigenvector has strictly positive entries,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log E[e^{\theta S_n}] = \log \rho(\theta),$$

Should be $\log \rho(\theta)$ not $\rho(\theta)$.

for any initial condition ξ_1 . Hence, $\Lambda(\theta) = \log \rho(\theta)$, and Λ is finite for all $\theta \in \mathbb{R}$. Thus, steepness is not an issue and, in order to apply Theorem 2.11, we need only to verify that Λ is differentiable everywhere. This follows from standard results in linear algebra and the fact that $\rho(\theta)$ is an isolated eigenvalue of $Q(\theta)$, which is a consequence of the Perron-Frobenius theorem. \diamond

Example 2.16 (Gaussian autoregressive processes)

Let a_1, \dots, a_r be given constants, and consider the recursion

$$X_t = \sum_{k=1}^r a_k X_{t-k} + \varepsilon_t \quad \text{for } t \in \mathbb{Z},$$

satisfies an LDP in $M_1(\mathcal{X})$ with good convex rate function $H(\cdot|\mu)$ given by

$$H(\nu|\mu) = \begin{cases} \int_{\mathcal{X}} \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu} d\mu & \text{if } \nu \text{ is absolutely continuous with respect to } \mu \\ \infty & \text{otherwise.} \end{cases}$$

Here, $d\nu/d\mu$ denotes the density, or Radon-Nikodym derivative, of ν with respect to μ . If ν and μ have densities p and q then $d\nu/d\mu(x) = p(x)/q(x)$.

The next result concerns independent random variables. It is perhaps the second most useful result, after the contraction principle, in applying large deviations theory to queues. Intuitively speaking, if

$$P(X_n \approx x) \approx e^{-nI(x)} \quad \text{and} \quad P(Y_n \approx y) \approx e^{-nJ(y)}$$

then by independence

$$P((X_n, Y_n) \approx (x, y)) \approx e^{-n[I(x)+J(y)]}$$

Theorem 4.14 *Let X_n satisfy an LDP in \mathcal{X} with good rate function I , let Y_n satisfy an LDP in \mathcal{Y} with good rate function J , and suppose that X_n is independent of Y_n , for each n . Assume that \mathcal{X} and \mathcal{Y} are separable. Then the pair (X_n, Y_n) satisfies an LDP in $\mathcal{X} \times \mathcal{Y}$ with good rate function $K(x, y) = I(x) + J(y)$.*

In fact the theorem holds if $\mathcal{X} \times \mathcal{Y}$ is regular (which is the case if e.g. both \mathcal{X} and \mathcal{Y} are regular). See the note in the proof below.

In fact, if I and J are infinite outside separable subsets of \mathcal{X} and \mathcal{Y} , the result still holds. This turns out to be useful in Chapter 7.

Proof. First we will recall some basic properties of the product topology on $\mathcal{X} \times \mathcal{Y}$. Then the proof proceeds in three steps: a proof that that K is a good rate function; a proof of the large deviations lower bound for open sets, proved locally; a proof of the large deviations upper bound for closed cylinder sets, then for general closed sets.

Topology on $\mathcal{X} \times \mathcal{Y}$. If σ and τ are bases for \mathcal{X} and \mathcal{Y} then $\{O \times P : O \in \sigma, P \in \tau\}$ is a basis for $\mathcal{X} \times \mathcal{Y}$. Since \mathcal{X} and \mathcal{Y} are separable, they have countable bases, and so $\mathcal{X} \times \mathcal{Y}$ has a countable basis of sets of the form $\{O_m \times P_n, m, n \in \mathbb{N}\}$ where each O_m and P_n is open in \mathcal{X} or \mathcal{Y} . Open sets in $\mathcal{X} \times \mathcal{Y}$ are of the form

$$\bigcup_{n \in \mathbb{N}} O_n \times P_n,$$

where O_n and P_n are open; and closed sets are of the form

$$\bigcap_{n \in \mathbb{N}} (C_n \times \mathcal{Y}) \cup (\mathcal{X} \times D_n)$$

For such a set,

$$\begin{aligned} & P((X_n, Y_n) \in B_N) \\ &= P\left((X_n, Y_n) \in \bigcup_{\substack{(i_1, \dots, i_N) \\ \in \{0,1\}^N}} \bigcap_{n \leq N} \begin{cases} i_n = 0: & C_n \times \mathcal{Y} \\ i_n = 1: & \mathcal{X} \times D_n \end{cases}\right) \\ &= P\left((X_n, Y_n) \in \bigcup_{(i_1, \dots, i_N)} \left(\bigcap_{n: i_n=0} C_n \right) \times \left(\bigcap_{n: i_n=1} D_n \right)\right). \end{aligned}$$

and so, by the principle of the largest term and independence,

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \frac{1}{n} \log P((X_n, Y_n) \in B_N) \\ & \leq - \inf_{i_1, \dots, i_N} \left(\inf_{x \in \bigcap_{i_n=0} C_n} I(x) + \inf_{y \in \bigcap_{i_n=1} D_n} J(y) \right) \\ & = - \inf_{(x,y) \in B_N} I(x) + J(y). \end{aligned}$$

LD upper bound for closed sets. By our remarks on topology, any closed set B is of the form

$$B = \bigcap_{N \in \mathbb{N}} B_N$$

(and in fact the sets B_N are decreasing, so $B = \lim_{N \rightarrow \infty} B_N$.) Thus

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \frac{1}{n} \log P((X_n, Y_n) \in B) \\ & \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log P((X_n, Y_n) \in B_N) \quad \text{for all } N \\ & \leq - \inf_{(x,y) \in B_N} K(x, y). \end{aligned}$$

Hence the lim sup is

$$\leq - \lim_{N \rightarrow \infty} \inf_{(x,y) \in B_N} K(x, y).$$

(This is an increasing limit, as the sets B_N are decreasing.) We will now show that

$$\lim_{N \rightarrow \infty} \inf_{z \in B_N} K(z) = \inf_{z \in B} K(z).$$

$$B \subset \left(\bigcup_{n \leq N} O_{z_n} \times P_{z_n} \right)^c.$$

Using the earlier result for cylinders, $\limsup \leq -(\alpha - \varepsilon)$. Since ε was arbitrary, we have the result.

There is a much cleaner and better proof. Let B be closed, and let $\alpha = \inf_{z \in B} K(z)$. If $\alpha = 0$ we are done. Otherwise pick $\varepsilon > 0$ such that $\alpha - \varepsilon > 0$. Now consider $L_{\alpha - \varepsilon}$. By the assumption that $\mathcal{X} \times \mathcal{Y}$ is regular, for every point $z \in L_{\alpha - \varepsilon}$ we can choose an open set which contains z but does not intersect B . Without loss of generality, this open set is of the form $O_z \times P_z$. By goodness of the rate function, this cover has a finite subcover, $O_{z_n} \times P_{z_n}$ for $n \leq N$. So

The following result makes this rough idea precise. It is due to Varadhan and Mogulskii. Let \mathcal{C}^T denote the space of continuous functions $x : [0, T] \rightarrow \mathbb{R}$ for which $x(0) = 0$, equipped with the topology of uniform convergence, and let \mathcal{A}^T denote the subspace consisting of absolutely continuous functions. (The result is often stated for $T = 1$. It is simple to generalise it to arbitrary $T > 0$.)

Theorem 6.1 (Sample path LDP for the partial sums process) *Let $(Y_t, t \in \mathbb{N})$ be a sequence of i.i.d. random variables, and let Λ be the cumulant generating function for Y_1 . Assume that $\Lambda(\theta)$ is finite for all $\theta \in \mathbb{R}$. Let $X(t)$ be the partial sums process $X(t) = Y_1 + \dots + Y_t$, and let $\tilde{X}^N \in \mathcal{C}^T$ be the scaled polygonalized partial sums process as in (6.1), restricted to the interval $[0, T]$. Then the sequence $(\tilde{X}^N, N \in \mathbb{N})$ satisfies an LDP in \mathcal{C}^T with rate function*

$$I_T(x) = \begin{cases} \int_0^T \Lambda^*(\dot{x}(s)) ds & \text{if } x \in \mathcal{A}^T \\ \infty & \text{otherwise.} \end{cases} \quad (6.5)$$

Note. Observe that this rate function is consistent with (6.4). Intuitively, (6.4) specifies the rate function for piecewise linear x . Since any sufficiently smooth x can be approximated by piecewise linear functions, the rate function in (6.5) is as we would expect.

In fact, the LDP for \tilde{X}^N holds even if Λ is finite only in a neighbourhood of zero, and it holds even if the random variables $(Y_t, t \in \mathbb{N})$ are weakly dependent. Dembo and Zajic [24] describe rather general conditions under which the LDP can be proved.

LDP over infinite horizon

This family of results (an LDP for $\tilde{X}^N|_{[0, T]}$ for each T) can immediately be extended to an LDP for the entire process \tilde{X}^N using a standard result known as the Dawson-Gärtner theorem for projective limits. This establishes that $(\tilde{X}^N, N \in \mathbb{N})$ satisfies an LDP in the space of continuous functions $x : \mathbb{R}_0^+ \rightarrow \mathbb{R}$ for which $x(0) = 0$, equipped with the topology of uniform convergence on compact intervals, with good rate function

$$I(x) = \sup_{T \in \mathbb{R}_0^+} I_T(x|_{[0, T]})$$

Not true! Suppose $Y_t \sim \exp(1)$ and that $(Y_t, t \in \mathbb{N})$ are independent. Then

$$I(x) = \int_0^1 \dot{x}_t - 1 - \log \dot{x}_t dt.$$

This is not a rate function. Consider a sequence indexed by n with $\dot{x}^n(t) = n$ for $t \leq T_n$ and $\dot{x}^n(t) = 1$ for $t > T_n$. By choosing T_n suitably, $I(x^n) = 1$. As $n \rightarrow \infty$ the only possible limit is $x(0) = 0$, $x(t) = 1 + t$ for $t > 0$. But this process does not lie in \mathcal{C}^1 since it is not continuous. See paper by Ganesh, Macci, Torrisi, Elec. J. Prob. vol 10, pp. 1026–1043, 2005.

can be made about scaling behaviour, which are perhaps more important than knowing the exact value of a rate function.

Imagine a complicated queueing network, and suppose we are interested in the queue size at some queue. As usual, we assume that the amount of work arriving to the network, and the service capacities, can be represented by a sequence of \mathbb{R}^d -valued random variables $(X(-t, 0], t \in \mathbb{N})$, for some fixed d . Let \tilde{X} be the polygonalized version of X . The queue size at the queue we are interested in will generally be of the form $Q = f(\tilde{X})$, although the function f can be quite complicated.

For consistency with Section 5.1, we should write \underline{X} not X , throughout this section.

This function f generally satisfies two basic properties. First, since the queue length is expressed in the same units as the inputs and service capacities, and taking all buffer sizes to be infinite, the function f is *linear in space*: $f(\kappa x) = \kappa f(x)$ for any $\kappa > 0$. Second, f is *homogeneous in time*: if we define the speeded-up input process $x^{\circ\kappa}$ by $x^{\circ\kappa}(-t, 0] = x(-\kappa t, 0]$ then $f(x^{\circ\kappa}) = f(x)$. These two properties, together with continuity of f , are sufficient for us to deduce that the queue size Q has exponential tails!

To illustrate: when we studied the single-server queue with an infinite buffer in Section 6.4, the function was $f(x) = \sup_t x(-t, 0]$, which is linear in space and homogeneous in time. The sequence of scaled input processes $N^{-1}\tilde{X}^{\circ N}$ satisfies an LDP, from which we obtain an LDP for Q/N , from which we deduced (6.22), namely that

$$\lim_{q \rightarrow \infty} \frac{1}{q} \log P(Q > q) = -\delta \quad \text{for some } \delta > 0 \quad (6.47)$$

(under quite general conditions on the LDP for $N^{-1}\tilde{X}^{\circ N}$).

Theorem 6.16 *Suppose that the sequence of inputs $N^{-1}\tilde{X}^{\circ N}$ satisfies a sample path LDP with linear geodesics, with mean rate μ and instantaneous rate function Λ^* which is strictly convex. Write $I(x)$ for the rate function*

$$I(x) = \begin{cases} \int_{-\infty}^0 \Lambda^*(\dot{x}_t) dt & \text{if } x \in \mathcal{A}_\mu \\ \infty & \text{otherwise.} \end{cases}$$

Let $Q = f(\tilde{X})$, and suppose

- i. f is continuous, and linear in space and homogeneous in time;
- ii. the system is stable, i.e. for the path x with constant gradient $\dot{x}_t = \mu$, $f(x) = 0$;
- iii. the problem is non-degenerate, i.e. there exists some $x \in \mathcal{A}_\mu$ for which $I(x) < \infty$ and $f(x) > 0$.

Then Q has exponential tails, i.e. it satisfies (6.47).

Chapter 7

Many-flows scalings

In this chapter we will systematize the result of Section 1.4: a large deviations principle for queues with many input flows. We will develop the theory using the general framework outlined in Chapter 5: decide on the scaling (Section 7.1), find a suitable topological space to work in (Section 7.2), establish an LDP for traffic processes (Section 7.3), then apply the contraction principle to deduce LDPs for various functions of interest (Sections 7.6–7.10).

We will then go on (Section 7.11) to describe some results concerning networks, which do not fit into this general framework.

7.1 Traffic scaling

Consider a queue fed by many input flows. Let $A^{(i)}(t)$ be the amount of work arriving to the queue in the interval $(-t, 0]$, $t \in \mathbb{Z}$, from input flow i . Suppose that each flow is a random process, and that the different flows are independent and identically distributed. (These assumptions are neither necessary nor sufficient for what we will do later. In Section 7.3 we will be precise; for now this will do.)

Let A^N be the average of N input flows:

$$A^N(t) = \frac{1}{N} \left(A^{(1)}(t) + \cdots + A^{(N)}(t) \right).$$

Some convenient notation. For talking about abstract processes, we will use the notation $A(t)$. When we come to study queues, it will be more convenient to use the extended notation which we described in Section 5.5. Write

$$\begin{array}{ll}
x(-t, 0] & \text{for } x(t) \\
x(-t, -u] & \text{for } x(t) - x(u), \text{ when } t \geq u \\
x|_{[-t, 0]} & \text{for the restriction of } x(\cdot) \text{ to } \{0, \dots, t\} \\
\dot{x}_{-t} & \text{for } x(t+1) - x(t)
\end{array}$$

and also $\dot{x}|_{(-t, 0]}$ for $\dot{x}_{-t+1}, \dots, \dot{x}_0$.

Throughout this chapter, many occurrences of $Ee^{\theta \cdot X|_{(-t, 0]}}$ should be replaced by $Ee^{\theta \cdot \dot{X}|_{(-t, 0]}}$. These changes are made in red, but not annotated in the margin.

Queue scaling. Consider the queue size function (with constant service rate, for simplicity) applied to A^N :

$$\begin{aligned}
q(A^N, C) &= \sup_{t \geq 0} A^N(-t, 0] - Ct \\
&= N^{-1} \sup_t \sum_{i=1}^N A^{(i)}(-t, 0] - NCt \\
&= N^{-1} R_0^N
\end{aligned}$$

where R_0^N is the queue size at time 0 in a queue fed by N flows $A^{(1)}, \dots, A^{(N)}$ and served at rate NC .

Now suppose that A^N satisfies a large deviations principle with good rate function I and that q is continuous. Applying the contraction principle, we obtain a large deviations principle of the form

$$\frac{1}{N} \log P(q(A^N, C) \geq b) \approx -J(b)$$

and hence

$$\frac{1}{N} \log P(R_0^N \geq Nb) \approx -J(b),$$

the usual form of the many-flows estimate (as in Theorem 1.8).

7.2 Topology for sample paths

In some ways it is easier to study continuity of queue-size functions in the many-flows limit than in the large-buffer limit, in some ways harder. Easier because we only need to work in discrete time; harder because it is harder to deal with the mean rate of an arrival process.

Discrete-time sample paths. We start with the set of sample paths

$$\mathcal{D} = \{x : \mathbb{N}_0 \rightarrow \mathbb{R}, x(0) = 0\}. \quad (7.1)$$

Recall also

$$\mathcal{D}_\mu = \{x \in \mathcal{D} : \underline{x} = \bar{x} = \mu\}.$$

It is not hard to check that the results in Chapter 5, which show that various queue-size functions are continuous on \mathcal{D}_μ , carry through to $\mathcal{D}_{[\mu-\varepsilon, \mu+\varepsilon]}$ for ε sufficiently small.

Note. These isolated points will not be at all important. As we have already mentioned, we can effectively ignore all sample paths outside $\mathcal{X}_{[\mu-\varepsilon, \mu+\varepsilon]}$, including all these isolated points. If we had made the extra assumption that X is ergodic, it would not even have been necessary to include them, since if A^L is ergodic then $\bar{A}^N = \underline{A}^N$ and this is finite.

The space $\mathcal{D}_{[\mu-\varepsilon, \mu+\varepsilon]}$ is not Polish; it is not even separable. Separability is needed for certain LDP results such as Theorem 4.14 for product spaces. However, as noted after that theorem, it is sufficient if there is a separable subspace which contains the effective domain of the rate function. It will turn out that the rate function is infinite outside \mathcal{D}_μ , which is Polish and hence separable.

As noted beside that theorem, it is sufficient for the space to be regular—which it is.

7.3 The sample path LDP

We will give here a simplified version of the large deviations principle. There are some extra subtleties which are needed to describe networks, which we will give in Section 7.11, and some generalizations, which can be found in [101].

Let X^N be the average of N independent arrival processes with common distribution X . We will prove an LDP for X^N .

Note. We could simply state the LDP as an assumption, as we did for the large-buffer scaling, and then apply the contraction principle; and this would be the most elegant way to proceed. However, the LDP is somewhat convoluted and abstract, and it is perhaps more helpful to work with a concrete theorem.

Definition 7.2 For $t \in \mathbb{N}$ and $\theta \in \mathbb{R}^t$, define the log moment generating function

$$\Lambda_t(\theta) = \log E \exp(\theta \cdot \dot{X}|_{(-t, 0]}).$$

Say that X is regular over finite horizons if each Λ_t is finite in a neighbourhood of 0, and essentially smooth (i.e. differentiable in the interior of its effective domain, and steep).

A scaling function is a function $v : \mathbb{N} \rightarrow \mathbb{R}$ for which $v_t / \log t \rightarrow \infty$. Given a scaling function, define for $\theta \in \mathbb{R}$ the scaled log moment generating function

$$\tilde{\Lambda}_t(\theta) = \frac{1}{v_t} \Lambda_t(e\theta v_t/t).$$

where e is the vector of 1s. Say that X is regular over the infinite horizon if the functions $\tilde{\Lambda}_t$ converge pointwise to a limit $\tilde{\Lambda}$ which is differentiable in a neighbourhood of the origin.

Theorem 7.1 (Many-flows sample path LDP) *Let X^N be the average of N independent identically distributed copies of some process X . If X is regular over finite horizons, then X^N satisfies a sample path large deviations principle with good rate function*

$$I(x) = \sup_{t \in \mathbb{N}} \Lambda_t^*(\dot{x}|_{(-t,0]}) = \lim_{t \rightarrow \infty} \Lambda_t^*(\dot{x}|_{(-t,0]}), \quad (7.2)$$

where

$$\Lambda_t^*(y) = \sup_{\theta \in \mathbb{R}^t} \theta \cdot y - \Lambda_t(\theta) \quad \text{for } y \in \mathbb{R}^t,$$

in the space \mathcal{D} defined by (7.1) and equipped with the topology of pointwise convergence.

If in addition X^N is regular over the infinite horizon then it is exponentially tight in \mathcal{D} equipped with the extended scaled uniform topology, and satisfies an LDP in that space with the same good rate function.

Note. The concept of regularity over the infinite horizon is needed to establish tightness, but neither the scaling function v_t nor the limiting scaled log moment generating function $\tilde{\Lambda}$ appear in the LDP above. Their only purpose is to control the tail behaviour of X^N . For most processes, the scaling function $v_t = t$ is appropriate; though it is useful to allow the more general v_t to cope with processes with non-standard tail behaviour, such as fractional Brownian motion.

In the rest of this chapter, we will say that X^N satisfies a sample path LDP if it is regular over finite and infinite horizons, and if it has stationary increments, i.e. $X^N(-t+u, u]$ has the same distribution as $X^N(-t, 0]$ for all $u \leq 0$.

Proof. We will first appeal to the generalized Cramér's theorem (Theorem 2.11) to establish an LDP for $X^N|_{(-t,0]}$ in \mathbb{R}^t . (Since X^N is the average of i.i.d. random variables, the generalized version of the theorem is in fact

overkill.) By assumption, Λ is essentially smooth and finite in a neighbourhood of the origin; by Lemma 2.3 it is lower-semicontinuous. Thus the conditions of the theorem are satisfied, and so $\dot{X}^N|_{(-t,0]}$ satisfies an LDP in \mathbb{R}^t for each t , with good rate function Λ_t^* .

Second, by the Dawson-Gärtner theorem we can extend this collection of LDPs to an LDP for X^N in (\mathcal{D}, τ_p) , by which we mean the set \mathcal{D} equipped with the topology of pointwise convergence (which is the projective limit topology for discrete sequences), with good rate function $I(x)$.

Third, we want to turn the LDP in (\mathcal{D}, τ_p) into an LDP in $(\mathcal{D}, \|\cdot\|)$, by which we mean \mathcal{D} equipped with our extended scaled uniform norm topology. (When we write \mathcal{D} this topology is to be understood; we are just spelling it out here for emphasis.) This can be done using the inverse contraction principle (Theorem 4.10). The ingredients are as follows: to show that the identity map $(\mathcal{D}, \|\cdot\|) \rightarrow (\mathcal{D}, \tau_p)$ is continuous, which is trivial; an LDP for X^N in (\mathcal{D}, τ_p) , which we have just found; and exponential tightness of X^N in $(\mathcal{D}, \|\cdot\|)$. The proof of exponential tightness is very technical, and is left to Lemma 7.3 at the end of this section.

To see that the two expressions for the rate function are equal, simply note that $\Lambda_t^*(\dot{x}|_{(-t,0]})$ is increasing in t . \square

We have used the term *mean rate* in connection with sample paths, in Section 5.4. The following theorem establishes equality between the mean rate and $t^{-1}EX(-t, 0]$. (Since X has stationary increments, this quantity does not depend on t .)

Theorem 7.2 *Under the conditions of Theorem 7.1, if $\mu = t^{-1}EX(-t, 0]$, then for $x \notin \mathcal{D}_\mu$*

$$I(x) = \infty.$$

Proof. Let $\mu = t^{-1}EX(-t, 0]$. We want to show that $I(x) = \infty$ if $x \notin \mathcal{D}_\mu$. Now,

$$\begin{aligned} I(x) &= \sup_{t \in \mathbb{N}} \Lambda_t^*(\dot{x}|_{(-t,0]}) \\ &= \sup_{t \in \mathbb{N}} \sup_{\theta \in \mathbb{R}^t} \theta \cdot \dot{x}|_{(-t,0]} - \Lambda_t(\theta) \\ &\geq \sup_{t \in \mathbb{N}} \sup_{\phi \in \mathbb{R}} \frac{\phi v_t}{t} x(-t, 0] - \Lambda_t\left(\frac{\phi v_t}{t} e\right) \quad (\text{by choosing } \theta = e\phi v_t/t) \\ &= \sup_{t \in \mathbb{N}} \sup_{\phi \in \mathbb{R}} \phi v_t \left[\frac{x(-t, 0]}{t} - \frac{\tilde{\Lambda}_t(\phi)}{\phi} \right]. \end{aligned}$$

By definition of scaling function, $\nu_t \rightarrow 0$; thus $\theta_t \rightarrow 0$ as $t \rightarrow \infty$. From these we can finally define

$$\delta_t = \frac{\tilde{\Lambda}(\theta_t) - \mu\theta_t}{\theta_t} + \frac{\tilde{\Lambda}(-\theta_t) + \mu\theta_t}{\theta_t} + \frac{\varepsilon_t}{\theta_t} + \nu_t.$$

The first two terms both decrease to 0 as $t \rightarrow \infty$, since $\tilde{\Lambda}$ is convex and differentiable at 0 with derivative $\mu = \tilde{\Lambda}'(0)$; the third term decreases to 0, as one can see by substituting in the definition of θ_t ; and we have already said why $\nu_t \rightarrow 0$. Thus $\delta_t \rightarrow 0$ as $t \rightarrow \infty$.

Establishing the limit. Pick t_0 sufficiently large that the convergence discussed above of $\tilde{\Lambda}_t$ to $\tilde{\Lambda}$ is uniform on $|\theta| \leq \phi'$. Now,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log \sum_{t > t_0} P\left(\frac{X^N(-t, 0]}{t} > \mu + \alpha\delta_t\right) \quad (7.6)$$

$$\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \log \sum_{t > t_0} \exp\left(-N\psi_t(\mu t + \alpha t\delta_t) + N\Lambda_t(\psi_t e)\right) \\ \text{(for any choice of } \psi_t > 0, \text{ by Chernoff's bound)} \quad (7.7)$$

$$\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \log \sum_{t > t_0} \exp\left(-Nv_t\left[\theta_t(\mu + \alpha\delta_t) - \tilde{\Lambda}_t(\theta_t)\right]\right) \\ \text{(by choosing } \psi_t = \theta_t v_t/t) \quad (7.8)$$

(To estimate the probability associated with the lower bound part of (7.4), we would use $-\psi_t$ rather than ψ_t in Chernoff's bound.) A typical term in brackets $[\cdot]$ in this expression is

$$\begin{aligned} & \theta_t(\mu + \alpha\delta_t) - \tilde{\Lambda}_t(\theta_t) \\ & \geq \theta_t(\mu + \alpha\delta_t) - \tilde{\Lambda}(\theta_t) - \varepsilon_t \quad \text{(by definition of } \varepsilon_t) \\ & = \alpha\theta_t\delta_t - (\tilde{\Lambda}(\theta_t) - \mu\theta_t) - \varepsilon_t \\ & = \alpha\left[\tilde{\Lambda}(\theta_t) - \mu\theta_t + \tilde{\Lambda}(-\theta_t) + \mu\theta_t + \varepsilon_t + \theta_t\nu_t\right] - (\tilde{\Lambda}(\theta_t) - \mu\theta_t) - \varepsilon_t \\ & = (\alpha - 1)\left[\tilde{\Lambda}(\theta_t) - \mu\theta_t + \varepsilon_t\right] + \alpha\left[\tilde{\Lambda}(-\theta_t) + \mu\theta_t + \theta_t\nu_t\right] \\ & \geq \alpha\theta_t\nu_t \quad \text{(assuming } \alpha \geq 1, \text{ and since } \tilde{\Lambda}(\theta) - \theta\mu \geq 0 \text{ by convexity)} \\ & \geq \alpha\nu_t^2 \quad \text{(for } t \text{ sufficiently large that } \theta_t < \phi') \\ & = \alpha \log t/v_t. \end{aligned}$$

We can use this to bound the sum we derived from (7.6), to find that

$$(7.6) \leq \limsup_{N \rightarrow \infty} \frac{1}{N} \log \sum_{t > t_0} e^{-N\alpha \log t}$$

$$\begin{aligned}
&= \limsup_{N \rightarrow \infty} \frac{1}{N} \log \sum_{t > t_0} t^{-\alpha N} \\
&= \alpha \limsup_{M \rightarrow \infty} \frac{1}{M} \log \sum_{t > t_0} t^{-M} \\
&\leq -\alpha M \log(t_0 + 1) \quad (\text{by (3.7)}).
\end{aligned}$$

Note that this holds for t_0 sufficiently large, and that the choice of t_0 does not depend on α . This completes the proof. \square

7.4 Example sample path LDPs

Example 7.1 (Gaussian)

Let X^N be the average of N independent arrival processes each distributed like X , where $(\dot{X}_t, t \in \mathbb{Z})$ is a stationary Gaussian process characterized by its mean and covariance structure:

$$\dot{X}|_{(-t,0]} \sim \text{Normal}(\mu e, \Sigma_t)$$

where Σ_t is the $t \times t$ matrix $(\Sigma_t)_{ij} = \text{Cov}(\dot{X}_{-t+i}, \dot{X}_{-t+j})$.

Note. Instead of Σ_t , we could specify the autocorrelation structure

$$\rho_t = \text{Cov}(\dot{X}_{-t+1}, \dot{X}_0)$$

or even the marginal variances $V_t = \text{Var} X(-t, 0]$, by using the relations

$$\begin{aligned}
V_1 &= \rho_0, \\
V_{t+1} &= V_t + 2(\rho_1 + \cdots + \rho_t) + \rho_0.
\end{aligned}$$

For such a process,

$$\Lambda_t(\theta) = \mu\theta \cdot e + \frac{1}{2}\theta \cdot \Sigma_t\theta$$

which is everywhere continuous, so X is regular over finite horizons. The scaled log moment generating function is

$$\tilde{\Lambda}_t(\theta) = \theta\mu + \frac{1}{2}\theta^2 v_t / t^2 V_t.$$

The natural choice of scaling function is $v_t = t^2 / V_t$, which gives

$$\tilde{\Lambda}(\theta) = \tilde{\Lambda}_t(\theta) = \theta\mu + \frac{1}{2}\theta^2.$$

Whether or not X^N is regular over the infinite horizon depends on the speed with which $v_t \rightarrow \infty$. It is regular if $V_t = o(t^2/\log t)$, i.e. if

$$\frac{V_t}{t^2/\log t} \rightarrow \infty \quad \text{as } t \rightarrow \infty.$$

There is no simple form for the rate function

$$\Lambda_t^*(y) = \sup_{\theta \in \mathbb{R}^t} \theta \cdot y - \mu\theta \cdot e - \frac{1}{2}\theta \cdot \Sigma_t \theta,$$

unless Σ^t is invertible in which case

$$\Lambda_t^*(y) = \frac{1}{2}(y - \mu e) \cdot \Sigma_t^{-1}(y - \mu e). \quad \diamond$$

Example 7.2 (Fractional Brownian motion)

Let X^N be the average of N independent copies of the process X , defined by

$$X(-t, 0] = \mu t + \sigma Z_t$$

where Z_t is a fractional Brownian motion with Hurst parameter H . Then for $\theta \in \mathbb{R}^t$

$$\Lambda_t(\theta) = \mu\theta \cdot e + \frac{1}{2}\sigma^2\theta \cdot S_t\theta$$

where the $t \times t$ matrix S_t is given by

$$(S_t)_{ij} = (|j - i - 1|^{2H} + |j - i + 1|^{2H} - 2|j - i|^{2H}).$$

(This gives the marginal variances $V_t = \sigma^2 t^{2H}$.) To show regularity over infinite horizons, choose the scaling function

$$v(t) = t^{2(1-H)},$$

so that

$$\tilde{\Lambda}_t(\theta) = \mu\theta + \frac{1}{2}\sigma^2\theta^2.$$

This does not depend on t , so it is equal to $\tilde{\Lambda}(\theta)$, and X^N is regular over the infinite horizon. \diamond

Example 7.3 (Markov-modulated fluid)

Let X^N be the average of N independent sources distributed like X , where \dot{X} is a Markov chain which produces an amount of work h each timestep while in the on state and no work while in the off state, and which flips from on to off with probability p and from off to on with probability q .

Since $X(-t, 0]$ can only take a finite number of values, it is clear that X^N is regular over finite horizons.

We can calculate $\Lambda_t(\theta e)$. First define

$$F_t = E(e^{\theta X(-t, 0]} | \dot{X}_{-t} = \text{on})$$

and

$$G_t = E(e^{\theta X(-t, 0]} | \dot{X}_{-t} = \text{off}).$$

We can find expressions for F_t and G_t by conditioning on \dot{X}_{-t+1} :

$$\begin{pmatrix} F_t \\ G_t \end{pmatrix} = \begin{pmatrix} (1-p)e^{\theta h} & p \\ qe^{\theta h} & 1-q \end{pmatrix}^t \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

And now

$$\Lambda_t(\theta e) = \log\left(\frac{p}{p+q}F_t + \frac{q}{p+q}G_t\right).$$

Is this regular over the infinite horizon?

We can rewrite $\Lambda_t(\theta e)$ as

$$\Lambda_t(\theta e) = \log\left(\kappa_1(\theta)e^{t \log \lambda_1(\theta)} + \kappa_2(\theta)e^{t \log \lambda_2(\theta)}\right)$$

where λ_1 and λ_2 are the two eigenvalues of the matrix above. This suggests the scaling function $v_t = t$, leading to the limiting scaled log moment generating function $\tilde{\Lambda}(\theta) = \log \lambda_1(\theta) \vee \log \lambda_2(\theta)$. This is differentiable in θ near $\theta = 0$. \diamond

Example 7.4 (Sources with independent increments)

Let X^N be the average of N independent sources distributed like X , and suppose that the $(\dot{X}_t, t \in \mathbb{Z})$ are independent. Suppose that

$$\Lambda_1(\theta) = \log E e^{\theta \dot{X}_1}$$

is finite in a neighbourhood of the origin, and essentially smooth. Then

$$\Lambda_t(\theta) = \sum_i \Lambda_1(\theta_i)$$

which is also finite in a neighbourhood of the origin and essentially smooth. So X^N is regular over finite horizons. With the scaling function $v_t = t$,

$$\tilde{\Lambda}_t(\theta) = \Lambda_1(\theta)$$

so X^N is regular over the infinite horizon.

Should be $\log E e^{\theta \dot{X}_1}$

7.6 Queues with infinite buffers

Consider the single server queue with an infinite buffer. Let $A^{(i)}(-t, 0]$ be the amount of work arriving in the interval $(-t, 0]$ from a single flow i , and suppose the queue is fed by N i.i.d. flows. Let

$$A^N(-t, 0] = \frac{1}{N} \sum_{i=1}^N A^{(i)}(-t, 0].$$

Suppose that the service rate is scaled in proportion to be NC . In Chapter 1, we saw that the queue length at time 0 is given by

$$Q_0^N = \sup_{t \in \mathbb{N}_0} NA^N(-t, 0] - NCt$$

and so

$$Q_0^N/N = f(A^N) \quad \text{where} \quad f(a) = \sup_{t \in \mathbb{N}_0} a(-t, 0] - Ct.$$

Assume that A^N satisfies the sample path LDP and has mean rate μ . Theorem 5.3 shows that f is continuous on $\mathcal{D}_{[\mu-\varepsilon, \mu+\varepsilon]}$, for $\mu + \varepsilon < C$. So, by the extended contraction principle, Q_0^N/N satisfies a large deviations principle with good rate function

$$J(q) = \inf_{a \in \mathcal{D}: f(a)=q} I(a).$$

The following theorem is rather heavy work, but it does tell us a lot about $J(q)$.

Theorem 7.7 *If A^N is regular over both finite and infinite horizons, and $\mu < C$, then $J(q)$ is given by*

$$J(q) = \inf_{t \geq 0} \sup_{\theta \geq 0} \theta(q + Ct) - \Lambda_t(e\theta). \quad (7.9)$$

First an example.

Example 7.5 (Fractional Brownian motion)

Let A be a fractional Brownian motion input as in Example 7.2. This has

$$\Lambda_t(\theta e) = \mu\theta t + \frac{1}{2}\theta^2\sigma^2 t^{2H}$$

$$\Lambda_t(\theta e) = \mu\theta + \frac{1}{2}\sigma^2 t^{2H}.$$

We can calculate the optimizing parameters in (7.9) explicitly. They are known as the critical spacescale and the critical timescale, and they are respectively

$$\hat{\theta} = \frac{q + (C - \mu)\hat{t}}{\sigma^2 \hat{t}^{2H}} \quad \text{and} \quad \hat{t} = \frac{q}{C - \mu} \frac{H}{1 - H}$$

(or rather, \hat{t} is an integer close to this value; but we will ignore this minor complication). This gives rate function

$$J(q) = \frac{1}{2\sigma^2} q^{2(1-H)} (C - \mu)^{2H} \left(\frac{H}{1 - H} \right)^{2(1-H)} \frac{1}{H^2}.$$

Gibbens and Teh [46] estimate the rate function corresponding to certain Internet traffic traces, and investigate how well it can be approximated by this analytical rate function for fractional Brownian motion. \diamond

Proof of Theorem 7.7 We will give an oblique proof of this theorem, breaking it into two lemmas which we will later refer to separately. The first lemma makes explicit which properties of the rate function we are using; the second lemma proves the rate function from them. \square

Lemma 7.8 *If A^N is regular over both finite and infinite horizons, and has mean rate μ , then, with $M_t(x) = \Lambda_t^*(x)$,*

- i. For all t , $\dot{A}^N|_{(-t,0]}$ satisfies an LDP in \mathbb{R}^t with good rate function $M_t(\cdot)$.*
- ii. $I(a) = \sup_t M_t(\dot{a}|_{(-t,0]})$, and I is good.*
- iii. $I(a) = \infty$ if $a \notin \mathcal{D}_\mu$.*
- iv. $M_t(e\mu) = 0$.*
- v. M_t is convex.*
- vi. $\Lambda_t(\theta) = M_t^*(\theta)$*

Proof. Items (i) and (ii) come from Theorem 7.1. Item (iii) comes from Theorem 7.2. Item (iv) is by Exercise 2.7. Items (v) and (vi) are from Lemma 2.6. \square

Lemma 7.9 *If A^N satisfies an LDP with rate function I , and satisfies the conclusions of Lemma 7.8, and the mean rate μ is less than C , then $J(q)$ is increasing and*

$$J(q) = \inf_{a \in \mathcal{D}: f(a)=q} I(a) \tag{7.10}$$

$$= \inf_{t \geq 0} \inf_{\substack{a|_{(-t,0]} \in \mathbb{R}^t: \\ a(-t,0]=q+Ct}} M_t(\dot{a}|_{(-t,0]}) \tag{7.11}$$

$$= \inf_{t \geq 0} \sup_{\theta \geq 0} \theta(q + Ct) - \Lambda_t(e\theta). \tag{7.12}$$

This was confusingly written. The point of this lemma is that all we need to know is that $\dot{A}^N|_{(-t,0]}$ satisfies an LDP with *some* rate function M_t , and that this rate function satisfies certain properties. We don't need to know that the rate function is any type of convex conjugate. The properties have been rewritten to make this clearer.

If $q > 0$ then the infimum can be taken over $t > 0$; if additionally $\Lambda_t(\theta)$ is differentiable at $\theta = 0$ it can be taken over $\theta > 0$.

Proof. If $q = 0$, then (7.10) takes the value 0 on the path $\dot{a} = e\mu$, and (7.11) and (7.12) take the value 0 at $t = 0$. So restrict attention to the case $q > 0$.

First, $J(q)$ is increasing. To see this, let

$$K_t(r) = \inf_{\substack{a|_{[-t,0]} \in \mathbb{R}^t: \\ a(-t,0)=r}} \mathbf{M}_t(a).$$

By the contraction principle, K_t is a rate function, and in particular K_t is non-negative. Since \mathbf{M}_t is convex, so is K_t . Since $\mathbf{M}_t(e\mu) = 0$, $K_t(\mu t) = 0$, and by convexity $K_t(r)$ is increasing for $r \geq \mu t$, and in particular $K_t(q + Ct)$ is increasing for $q \geq 0$. Now,

$$J(q) = \inf_{t \geq 0} K_t(q + Ct)$$

and the infimum of increasing functions is increasing, so J is increasing.

Next, (7.10) \geq (7.11). Suppose (7.10) is finite (otherwise the inequality is trivial). The sample path rate function I is good, so an optimal path \hat{a} is attained. And $I(\hat{a}) < \infty$, so $\hat{a} \in \mathcal{D}_\mu$. Now $q(\hat{a}) = \sup_t \hat{a}(-t, 0] - Ct = q$, and by Theorem 5.3 this supremum is attained, say at \hat{t} . Thus

$$\begin{aligned} I(\hat{a}) &= \sup_t \mathbf{M}_t(\dot{\hat{a}}|_{[-t,0]}) \\ &\geq \mathbf{M}_{\hat{t}}(\dot{\hat{a}}|_{[-\hat{t},0]}) \geq (7.11). \end{aligned}$$

i.e. an optimal path in (7.11)

Next, (7.10) \leq (7.11). Suppose (7.11) is finite (otherwise the inequality is trivial). For given t , an optimal path $\hat{a}|_{[-t,0]}$ is attained, by goodness of the rate function \mathbf{M}_t . And an optimal \hat{t} is also attained. For suppose not, and take a sequence $t_n \rightarrow \infty$ and $a^n|_{[-t_n,0]}$ with $a^n(-t_n, 0] = q + Ct_n$ and $\mathbf{M}_t(\dot{a}^n|_{[-t_n,0]})$ bounded above by K say. By the contraction principle and the goodness of the rate function I , we can extend $a^n|_{[-t_n,0]} \in \mathbb{R}^{t_n}$ to $a^n \in \mathcal{D}$, with $I(a^n) < K$. Since I is good it has compact level sets, so the a^n have a convergent subsequence, say $a^k \rightarrow a$, also with $I(a) < K$. But then $a(-t_k, 0]/t_k \rightarrow C$ so $a \notin \mathcal{D}_\mu$ so $I(a) = \infty$, a contradiction.

By the contraction principle and the goodness of the rate function I , we can extend $\hat{a}|_{[-\hat{t},0]} \in \mathbb{R}^{\hat{t}}$ to $\hat{a} \in \mathcal{X}$, with $I(\hat{a}) = \mathbf{M}_{\hat{t}}(\dot{\hat{a}}|_{[-\hat{t},0]})$. Since the rate function is finite, $\hat{a} \in \mathcal{D}_\mu$. If $q(\hat{a}) = q$ the inequality is proved. So suppose, for some \hat{t} , that $q(\hat{a}) = q' \neq q$ for all such extensions \hat{a} of all optimal $\hat{a}|_{[-\hat{t},0]}$.

Since $\hat{a}(-\hat{t}, 0] = q + C\hat{t}$, $q' > q$. Then there is some $s \neq \hat{t}$ with $\hat{a}(-s, 0] = q'$. But then

$$\begin{aligned} \inf_t \inf_{\substack{a|_{[-t,0]} \in \mathbb{R}^t: \\ a(-t,0]=q+Ct}} \mathbf{M}_t(\dot{a}|_{(-t,0]}) &\geq \inf_{s \neq \hat{t}} \inf_{\substack{a|_{[-s,0]} \in \mathbb{R}^s: \\ a(-s,0]=q'+Ct}} \mathbf{M}_s(\dot{a}|_{(-s,0]}) \\ &\geq \inf_{s \neq \hat{t}} \inf_{\substack{a|_{[-s,0]} \in \mathbb{R}^s: \\ a(-s,0]=q+Ct}} \mathbf{M}_s(\dot{a}|_{(-s,0]}) \end{aligned}$$

where the last inequality is because $K_t(q + Ct)$ is increasing in q . The inequalities must then both be inequalities. Repeat this procedure until we find some \hat{a} for which $q(\hat{a}) = q$. We will eventually find some such \hat{a} , for otherwise there are arbitrarily large optimal \hat{t} , and as in the previous paragraph this yields a contradiction.

Next, (7.11) = (7.12). We will first show

$$K_t(x) = \sup_{\theta \in \mathbb{R}} \theta x - \Lambda_t(e\theta).$$

Note that \mathbf{M}_t is closed convex (it is a rate function, hence lower semicontinuous, and we assume it to be convex). By Lemma 2.4, $\mathbf{M}_t = \Lambda_t^*$, where Λ_t is given by (vi) in Lemma 7.8. For the upper bound on $K_t(x)$,

$$\begin{aligned} K_t(x) &= \inf_{\substack{a|_{[-t,0]} \in \mathbb{R}^t: \\ a(-t,0]=x}} \sup_{\theta \in \mathbb{R}^t} \theta \cdot \dot{a} - \Lambda_t(\theta) \\ &\geq \inf_{\substack{a|_{[-t,0]} \in \mathbb{R}^t: \\ a(-t,0]=x}} \sup_{\theta \in \mathbb{R}} \theta x - \Lambda_t(e\theta) \\ &= \sup_{\theta \in \mathbb{R}} \theta x - \Lambda_t(e\theta). \end{aligned}$$

For the lower bound on $K_t(x)$,

$$\begin{aligned} \sup_{\theta \in \mathbb{R}} \theta x - \Lambda_t(e\theta) &= \sup_{\theta \in \mathbb{R}} \theta x - \left[\sup_{y \in \mathbb{R}} \sup_{\substack{a|_{[-t,0]}: \\ a(-t,0]=y}} e\theta \cdot a - \mathbf{M}_t(a) \right] \\ &= \sup_{\theta \in \mathbb{R}} \inf_{y \in \mathbb{R}} \theta(x - y) + K_t(y) \\ &= K_t(x) + \sup_{\theta \in \mathbb{R}} \inf_{y \in \mathbb{R}} \left[K_t(y) - (K_t(x) + \theta(y - x)) \right] \\ &\geq K_t(x), \end{aligned}$$

where the last equality comes from taking a supporting plane to the convex function $K_t(y)$ at x .

To complete the proof that (7.11) = (7.12): should be $\theta(q + Ct)$ not θx

For $J(q)$, we are interested in $K_t(q + Ct)$. As we noted before, $K_t(x)$ is increasing for $x \geq \mu t$, so the supporting plane has $\theta \geq 0$. It is clear we can also restrict attention to $\theta \geq 0$ in the upper bound for $K_t(q + Ct)$. Hence

$$K_t(q + Ct) = \sup_{\theta \geq 0} \theta x - \Lambda_t(e\theta).$$

Last clause of Lemma 7.9:

Finally, the case $q > 0$. The rate function at $t = 0$ is infinite, so we can restrict attention to $t > 0$. As we have just seen, the supremum can be taken over $\theta \geq 0$. The upper bound for $K_t(x)$ still works if we restrict attention to $\theta > 0$. For the lower bound, except in the pathological case $K_t(x) = 0$ for all $x \geq \mu t$, it can similarly be shown that, for $x \geq \mu t$,

$$\sup_{\theta > 0} \theta x - \Lambda_t(e\theta) \geq K_t(x) - \varepsilon$$

where ε can be arbitrarily small, so we restrict attention to $\theta > 0$.

This requires that the derivative is equal to μt , which is true by Lemma 2.6, and which should have been mentioned in Lemma 7.9.

The pathological case cannot happen if Λ_t is differentiable at the origin. For then $d/d\theta \Lambda(e\theta) = \mu t$ at $\theta = 0$, so there is some $\theta > 0$ for which $\Lambda(e\theta) < Ct$, and the lower bound for $K_t(q + Ct)$ is strictly positive. \square

7.7 Queues with finite buffers

Consider now the single-server queue with a finite buffer. As in the previous section, let $NA^N(-t, 0]$ be the total amount of work arriving in the interval $(-t, 0]$ from N i.i.d. flows, and suppose that A^N satisfies the sample path LDP. Suppose the service rate is scaled in proportion to be NC , and the buffer size is scaled in proportion to be NB . Let Q_0^N be the queue size at time 0.

By rescaling the units in which work is expressed, it is clear that $Q_0^N/N = \bar{f}(A^N)$, where \bar{f} is the finite-buffer queue size function described in Section 5.7. In that section we proved that \bar{f} is continuous on $\mathcal{D}_{[\mu-\varepsilon, \mu+\varepsilon]}$ for $\mu + \varepsilon < C$, so the extended contraction principle again tells us that $\bar{q}(X^N)$ satisfies a large deviations principle with good rate function

$$\bar{J}(q) = \inf_{x \in \mathcal{D}: \bar{f}(x)=q} I(x).$$

What is $\bar{J}(q)$? The following theorem relates $\bar{J}(q)$ to the rate function $J(q)$ for the infinite-buffer queue from the preceding section.

Theorem 7.10 For $q \leq B$, $\bar{J}(q) = J(q)$; and for $q > B$, $\bar{J}(q) = \infty$.

Proof. The last clause is obvious: the queue size can never be greater than B . So suppose $q \leq B$. We wish to show $\bar{J}(q) = J(q)$. This hints that the most likely path might be the same in each case. To relate the queue sizes for a given path, note that $f(a) \geq \bar{f}(a)$. This was discussed in Section 5.7.

Suppose $\bar{J}(q)$ is finite. Then there is an optimal path \hat{a} for which $\bar{f}(\hat{a}) = q$, so $f(\hat{a}) \geq q$. Since $J(q)$ is increasing, $J(q)$ must be finite. In other words, if $J(q)$ is infinite, then $\bar{J}(q)$ is infinite also.

So suppose $J(q)$ is finite. Let \hat{a} be an optimizing path in Theorem 7.7. Consider the queue size for the infinite-buffer queue under this path. The queue is empty at $-\hat{t}$ by Lemma 5.4. The queue then builds up. Suppose it first reaches level $q' \geq q$ at time $-s$. Consider the truncated process $b = \hat{a}|_{(-\infty, -s]}$. Suppose we feed b into the finite-buffer queue. Since finite-buffer queue size is no larger than infinite-buffer queue size, the finite-buffer queue must be empty at time $-(\hat{t} - s)$. By construction, the finite-buffer queue will not reach level q before time 0, so $\bar{f}(b) = f(b)$, so $\bar{J}(q) \leq I(b)$.

What is this rate function? By stationarity,

$$I(\hat{a}) \geq I(b).$$

Also $f(b) \geq q$. Since $J(q)$ is increasing, $J(q) \leq I(b)$. By optimality, $J(q) = I(\hat{a})$. So $I(b) = I(\hat{a})$, and thus $\bar{J}(q) \leq J(q)$.

Consider the optimal path \hat{b} in $\bar{J}(q)$. It causes $\bar{q}(\hat{b}) = q$, and so $q(\hat{b}) \geq q$. Since $J(q)$ is increasing, $J(q) \leq \bar{J}(q)$.

Hence $J(q) = \bar{J}(q)$. \square

7.8 Overflow and underflow

Before leaving the simple single-server queue, there are some more large deviations results which are interesting, and which are, at first sight, easily confused with those of Sections 7.6 and 7.7.

The first gives the probability that a queue with an infinite buffer is non-empty. At first sight, we can find this from the LDP in Section 7.6: just consider the event that $q > 0$. But the large deviations upper bound we get is useless, because it involves the closure of this set—which is $q \geq 0$, the entire space. So for a better bound, we can go back to the sample path LDP and look at the closure of the set of sample paths for which $f(a) > 0$ (where f is the infinite-buffer queue size function), now not the entire space.

Theorem 7.11 *If A^N is regular over finite and infinite horizons, and has mean rate $\mu < C$, then the event $\{f(A^N) > 0\}$ has large deviations lower*

bound $-J(0^+)$ and upper bound $-J^+(0)$, where

$$J^+(0) = \sup_{\theta \in \mathbb{R}} \theta C - \Lambda_1(\theta)$$

and

$$J(q^+) = \lim_{r \downarrow q} J(r).$$

Proof. Lower bound. Let F be the event $\{f(a) > 0\}$. The large deviations lower bound is

$$\inf_{a \in F} I(a).$$

Since $F = \cup_{q>0} \{f(a) = q\}$,

$$\inf_{a \in F} I(a) = \inf_{q>0} J(q).$$

But since $J(q)$ is increasing, this is just

$$\lim_{q \downarrow 0} J(q).$$

Upper bound. We will prove that

$$\inf_{a \in \bar{F}} I(a) = \inf_{t>0} \inf_{a: a(-t,0] = Ct} I(a). \quad (7.13)$$

This reduces to

$$\inf_{t>0} \sup_{\theta \in \mathbb{R}} \theta Ct - \Lambda_t(\theta)$$

as in Theorem 7.7. By convexity,

$$\Lambda_t(\theta e) \leq t \Lambda_1(\theta e),$$

Should read

$$\Lambda_t(\theta e) \leq \Lambda_1(\theta t e).$$

Proof:

$$\log E e^{\theta(X_1 + \dots + X_t)/t}$$

$$\leq \log E \left(\frac{e^{\theta X_1} + \dots + e^{\theta X_t}}{t} \right)$$

$$= \log E e^{\theta X_1}.$$

so the optimum is attained at $t = 1$ and we are left with $J^+(0)$.

LHS ≤ RHS in (7.13). Suppose $a(-t, 0] = Ct$ for some $t > 0$. For $\varepsilon > 0$,

$$\dot{a}^\varepsilon = (\dots, \dot{a}_{-2}, \dot{a}_{-1}, \varepsilon + \dot{a}_0).$$

Then $q(a^\varepsilon) > 0$ so $a^\varepsilon \in F$. Also $a^\varepsilon \rightarrow a$ as $\varepsilon \rightarrow 0$, so $a \in \bar{F}$. Thus

$$\{a : \exists t > 0, a(-t, 0] = Ct\} \subset \bar{F}.$$

Taking the infimum of I over these sets gives the result.

For the large-buffer scaling, we showed in Section 6.4 the most likely path to overflow was linear, a consequence of the linear geodesic property. In the many-flows scaling, the most likely path is not in general linear. Nonetheless, we can still find its form explicitly, at least for the case of a single-server queue.

Theorem 7.13 *If $J(q)$ is finite then the optimal timescale \hat{t} and the optimizing path \hat{a} are both attained. If additionally the optimizing parameter $\hat{\theta}$ is attained, and the rate function $J(q)$ is strictly increasing at q , then an optimal path is given by, for $t \geq \hat{t}$,*

Bad notation. Prefer

$$\dot{a}|_{(-t,0]} = \nabla \Lambda_t(\theta)$$

$$\dot{a}|_{(-t,0]} = \nabla \Lambda_t(\hat{\theta}s|_{(-t,0]}),$$

where s is the step function

$$s = e|_{(-\hat{t},0]} + 0e.$$

evaluated at $\theta = \hat{\theta}s$, where $s \in \mathbb{R}^t$ consists of \hat{t} ones followed by $t - \hat{t}$ zeros.

Proof. We explained why the optimal timescale is attained, in the proof of Theorem 7.7. Suppose that the optimal parameter $\hat{\theta}$ is attained. Since $J(q)$ is finite, $\Lambda_{\hat{t}}(\hat{\theta}e)$ is finite. This is equal to $\Lambda_t(\hat{\theta}s)$ for $t \geq \hat{t}$, which is thus also finite. By essential smoothness (a consequence of being regular over finite horizon t) Λ_t must be differentiable at $\hat{\theta}s$. Define \hat{a} by $\dot{a}|_{(-t,0]} = \nabla \Lambda_t(\hat{\theta}s)$. (These definitions, one for each t , are clearly all consistent.) This path has the right rate function: using Lemma 2.4, $\Lambda_t^*(\dot{a}|_{(-t,0]})$ is equal to (7.9). And it also causes the queue to reach at least the right level: from differentiating $\theta(q + Ct) - \Lambda_{\hat{t}}(e\theta)$ with respect to θ at $\theta = \hat{\theta}$, $\hat{a}(-\hat{t}, 0] = q + Ct$. If $q(\hat{a}) = q$ then we are done. If $q(\hat{a}) = q' > q$ then $J(q') \leq I(\hat{a})$. But $J(q') > J(q) = I(\hat{a})$, a contradiction. \square

since we've assumed J is strictly increasing

Example 7.7 (Gaussian sources)

Let A be a Gaussian, as in Example 7.1. It is easy to work out the optimal path:

$$\nabla \Lambda_t(\theta s) = \mu e + \theta \Sigma_t s.$$

where $(\Sigma_t)_{ij} = \rho_{|i-j|}$.

Consider the case of fractional Brownian motion, 7.2, where

$$\rho_t = \frac{1}{2}\sigma^2 \left((t-1)^{2H} - 2t^{2H} + (t+1)^{2H} \right) \quad \text{and} \quad \rho_0 = \sigma^2.$$

The most likely path to overflow can be computed to be, for $-\hat{t} < -t \leq 0$,

$$\dot{a}_{-t} = \mu + \frac{1}{2}\hat{\theta}\sigma^2 \left((t+1)^{2H} - t^{2H} + (\hat{t}-t-2)^{2H} - (\hat{t}-t-1)^{2H} \right).$$

If $H > \frac{1}{2}$, the source exhibits long-range dependence, and the most likely input path $t \mapsto \dot{a}_t$ leading to overflow is concave; whereas if $H < \frac{1}{2}$, the path to overflow is convex. \diamond

Exercise 7.8

Let A be a single-step autoregressive process:

$$\dot{A}_t = \mu + a(\dot{A}_{t-1} - \mu) + \sqrt{1 - \alpha^2} \sigma \varepsilon_t$$

where the ε_t are independent $\text{Normal}(0, 1)$ and $|a| < 1$. Then $\rho_t = \sigma^2 a^t$. Show that the most likely path to overflow is, for $-\hat{t} < -s \leq 0$,

$$\dot{a}_{-t} = \mu + \hat{\theta} \sigma^2 \left(1 + \frac{1 - a^{t+1}}{1 - a} + \frac{1 - a^{\hat{t}-t}}{1 - a} \right). \quad \diamond$$

In the brackets, replace a by α

Example 7.9 (Markov-modulated on-off source)

Let A be an on-off Markov fluid flow, as in Example 7.3. To calculate the most likely path to overflow, note

$$\dot{a}_{-t} = (\nabla \Lambda_t(\hat{\theta}s))_{-t} = \frac{E(\dot{A}_{-t} e^{\theta A(-\hat{t}, 0]})}{E(e^{\theta A(-\hat{t}, 0]})}$$

We can now calculate

$$\begin{aligned} E(\dot{A}_{-t} e^{\theta A(-\hat{t}, 0]}) &= E\left[\dot{A}_{-t} E(e^{\theta A(-\hat{t}, -t-1]} | \dot{A}_{-t}) e^{\theta \dot{A}_{-t}} E(e^{\theta A(-t, 0]} | \dot{A}_{-t})\right] \\ &= \frac{q}{p+q} h F_{t-1} e^{\theta h} F_{\hat{t}-t}. \end{aligned}$$

The first equality follows from the Markov property, and the second equality follows from reversibility. This gives

$$\dot{a}_{-t} = \frac{q h e^{\theta h} F_{\hat{t}-t-1} F_t}{q F_{\hat{t}} + p G_{\hat{t}}}$$

If $p + q < 1$ the path to overflow $t \mapsto \dot{a}_t$ is concave over $t \in (-\hat{t}, 0]$: the sources start slowly, then conspire to produce lots of work in the middle of the critical timeperiod, then slow down again at the end. (If $p + q > 1$ it is convex.) \diamond

7.10 Priority queues

In the examples so far, the rate function has simplified enough that we can draw fairly detailed conclusions. That is the exception: under the many-flows scaling, very often, all we can write down is that J is the solution to a complicated optimization problem, and leave it at that.

A priority queue is such a case. But even though we cannot work out the rate function exactly, we can still interpret the result and give some interesting bounds.

Consider a priority queue fed by two flows: the high priority flow A^N , the average of N independent copies of some stationary process A , and the low priority flow B^N , the average of L independent copies of some stationary process B . Let μ and ν be the mean rates of A and B . Suppose A and B are regular over finite and infinite horizons. Let the queue be served at constant service rate $C > \lambda + \mu$, and let it have an infinite buffer.

Let Q^N be the amount of high priority work in the queue, and R^N the amount of low priority work. As we discussed in Section 5.9, the easiest way to define these is

$$\begin{aligned} Q^N &= q(A^N) \\ R^N &= r(A^N, B^N) = q(A^N + B^N) - q(A^N), \end{aligned}$$

where q is the queue size function for the single-server queue with infinite buffer. (We have described how to interpret the scaling of similar quantities in Sections 7.6 and 7.7, and we will not repeat it here.)

The function $(a, b) \mapsto (q(a), r(a, b))$ is continuous on $\mathcal{D}_{[\lambda-\varepsilon, \lambda+\varepsilon]} \times \mathcal{D}_{[\mu-\varepsilon, \mu+\varepsilon]}$ for ε sufficiently small. By the extended contraction principle, (Q^N, R^N) satisfies an LDP with good rate function

$$J(q, r) = \inf_{\substack{a \in \mathcal{D}, b \in \mathcal{D}: \\ q(a) = q, q(a+b) = q+r}} \sup_t \Lambda_t^*(\dot{a}|_{(-t,0]}) + \sup_t M_t^*(\dot{b}|_{(-t,0]}),$$

where Λ_t and M_t are the log moment generating functions of A and B . By the contraction principle, R^N satisfies an LDP with good rate function

$$J(\cdot, r) = \inf_{q \geq 0} J(q, r).$$

The following lemma gives a bound on the rate function.

Lemma 7.14

$$J(\cdot, r) \geq \inf_t \sup_{\theta} \theta(r + Ct) - \Lambda_t(\theta e) - M_t(\theta e). \quad (7.15)$$

Finite-horizon regularity of D^N

Our earlier definition of regularity over finite horizons is not entirely appropriate here. In Section 7.3 we dealt with a process X^N which was the average of L independent copies of X ; here D^N is the average of N independent copies of $D^{(N)}$, which depends on N . This is not a significant obstacle: to obtain the sample path LDP, it is sufficient that the limit

$$M_t(\theta) = \lim_{N \rightarrow \infty} \frac{1}{N} \log E \exp(N\theta \cdot \dot{D}^N|_{(-t,0]})$$

exists, and that M_t satisfies the usual conditions: for each t the origin belongs to the interior of the effective domain of M_t and M_t is essentially smooth.

The following theorem tells us that $D^{(N)}$ is regular over finite horizons (with this enhanced definition), and that furthermore its statistical characteristics are essentially the same as those of X .

Should be A not X

Write Λ_t for the log moment generating function associated with A , and I for its rate function.

Theorem 7.15 M_t exists, and is equal to Λ_t , for θ in the interior of the effective domain of Λ_t .

Proof. Let A be the arrival process which becomes $D^{(N)}$. First note that

$$D^{(N)}(-t, 0] \leq A(-t - \lfloor B/C \rfloor, 0]$$

since any work arriving before $-t - \lfloor B/C \rfloor$, even if it finds the queue full, must have left by time $-t$. In what follows we drop the $\lfloor \cdot \rfloor$ notation.

For fixed t , the collection

$$\{\exp(\theta \cdot \dot{D}^{(N)}|_{(-t,0]})\}$$

is uniformly integrable, since

$$0 \leq \theta \cdot \dot{D}^{(N)}|_{(-t,0]} \leq \max_i |\theta_i| A(-t - B/C, 0].$$

For any $-t < s \leq 0$, $P(\dot{D}^{(N)}_{-s} \neq \dot{A}_{-s})$ is bounded by the probability that the queue is non-empty at either $-s - 1$ or $-s$. By Corollary 7.12 this tends to 0. So

$$\exp(\theta \cdot \dot{D}^{(N)}|_{(-t,0]}) - \exp(\theta \cdot \dot{A}|_{(-t,0]}) \rightarrow 0 \quad \text{in probability.}$$

Thus

$$E \exp(\theta \cdot \dot{D}^{(N)}|_{(-t,0]}) - E \exp(\theta \cdot \dot{X}|_{(-t,0]}) \rightarrow 0$$

Should be \dot{A} not \dot{X}

and taking logarithms gives the result. □

The way in which this topology is used is rather technical; full details can be found in [100]. We will restrict ourselves to stating the conclusion: D^N satisfies an LDP in $(\mathcal{D}_\mu, wq(C, B))$ with exactly the same rate function as A^N , for any $C > \mu$ and B .

Decoupling and other extensions

Consider now a queue fed by many independent flows of different types: L flows like A and L flows like B . Let $D^{(N)}$ and $E^{(N)}$ be typical outputs, and let D^N and E^N be as before. If the total mean arrival rate is less than the service rate, then the above proofs still work with minor modifications, and we conclude that $D^N \sim A^N$ and $E^N \sim B^N$ (in a large deviations sense), which tells us that $D^{(N)} \sim A$ and $E^{(N)} \sim B$ (in a heuristic sense). So the marginal distributions of the flows are essentially unchanged. What about their joint distributions? The basic fact, that the queue is frequently empty, is still true. By considering now the log moment generating function

$$\log E \exp(\theta \cdot \dot{D}^{(N)} + \phi \cdot \dot{E}^{(N)})$$

one can show that $D^{(N)}$ and $E^{(N)}$ are essentially independent (in a heuristic sense).

It might be expected that traffic flows would influence each other. For example, if A is very bursty and B is smooth, one might expect $D^{(N)}$ to be less bursty than A and $E^{(N)}$ to be less smooth than B , and indeed this can happen when the router only has a small number of inputs. But we have seen that in the many flows scaling regime it is not the case. In other words, $D^{(N)}$ and $E^{(N)}$ do not depend on the traffic mix at the router (so long as the total mean input rate is less than the service rate). This is known as *decoupling*.

We have only described the output of a single queue. Obviously it would be nice to describe networks. The results can be extended to flows which have passed through several queues (each queue empties often, so there is a high probability that the flow passes through each queue unchanged) but it is hard to interpret them, since it is not clear even how to formulate sensible network limits in the many-flows regime. This regime describes systems with many independent flows; as the number of independent flows increases, should the network topology be scaled up too, and if so how?

9.2 Traffic processes

We will not attempt to prove here a moderate deviations principle for traffic processes, but simply state it as an assumption. For details see [102]. We will work in discrete time, as we did in Chapter 7—look back at Section 7.2 to remind yourself of the space \mathcal{D} of real-valued integer-indexed processes. Let $(X^N, N \in \mathbb{N})$ be a sequence of processes in \mathcal{D} .

Note. The process X^N could be the average of N independent copies of a process X , in which case this definition describes a many-flows result. Or it could be a speeded-up version of a process, $X^N(-t, 0] = N^{-1}X(-t, 0]$, in which case this definition describes a large-buffer result (although for large-buffer results it is more natural to work with polygonalized processes in continuous time). The theory in this chapter applies equally.

Definition 9.1 *Say that X^N , normalized, satisfies the sample path moderate deviations principle with mean $\mu > 0$ and covariance structure $(\gamma_t)_{t \geq 0}$ if the following four conditions hold:*

i. For each $\beta \in (0, 1)$, X^N satisfies a large deviations principle of the form

$$\frac{1}{N^\beta} \log P(N^{(1-\beta)/2}(X^N - \mu e) \in B) \approx - \inf_{x \in B} I(x) \quad (9.2)$$

with good rate function I , in the space \mathcal{D} given in Definition 7.1, where e is the vector of 1s.

ii. The rate function I has the form

$$I(x) = \sup_{t \in \mathbb{N}} \sup_{\theta \in \mathbb{R}^t} \theta \cdot x(-t, 0] - \Lambda_t(\theta)$$

where $\Lambda_t(\theta) = \frac{1}{2}\theta \cdot \Sigma_t \theta$ and Σ_t is the $t \times t$ matrix $(\Sigma_t)_{ij} = \gamma_{|i-j|}$ for some function γ , called the covariance function.

iii. Let $V_t = e \cdot \Sigma_t e$ be the variance function corresponding to γ . Require that $V_t = o(t^2 / \log t)$.

iv. $I(x) = 0$ if $x \notin \mathcal{D}_\mu$.

There are many clauses to this definition, and it may not be immediately apparent where they come from. If so, write down a moderate deviations principle for $X^N(-t, 0]$ (where X^N may come from either a many-flows scaling or a large-buffer scaling), as described in Theorem 9.1, and see that this is the natural extension of that result from real-valued random variables to processes. Some more remarks on the definition:

Strictly speaking, e is the vector of 1s. What we want here instead of e is the constant-rate arrival process f defined by $f(-t, 0] = t$, i.e. $\dot{f} = e$. With some abuse of notation, we will write e to mean f , in this chapter.

9.5 Mixed limits

Consider a queue with an infinite buffer, fed by an input flow A^N . What are statistical characteristics of the departure process? We will make this question precise, in a novel way, using the scaling parameter β .

Let the aggregate input process be NA^N , where A^N , normalized, satisfies a sample path moderate deviations principle at all scales $\beta \in (0, 1)$; and suppose the service rate is scaled accordingly to be $N\mu + N^{(1+\beta)/2}C$. Let $N^{(1+\beta)/2}Q_{-t}^N$ be the queue size at time $-t$. Define the departure process in the usual way:

$$ND^N(-t, 0] = NA^N(-t, 0] + N^{(1+\beta)/2}Q_{-t}^N - N^{(1+\beta)/2}Q_0^N.$$

By the contraction principle, the departure process, normalized, satisfies some moderate deviations principle at scale β . Does it satisfy a moderate deviations principle at *other scales* $\beta' \neq \beta$? (In the large buffer scaling and the many flows scaling, it is not even possible to ask this question. But it does relate very closely to the question of Hurstiness in Chapter 8.)

First, suppose $\beta' < \beta$. It turns out that, at this scale, A^N and D^N are exponentially equivalent: that is, for any $\delta > 0$,

Should be $\mu e \in \mathcal{D}$ not plain $\mu \in \mathbb{R}$

$$\limsup_{N \rightarrow \infty} \frac{1}{N^{\beta'}} \log P(\|N^{(1-\beta')/2}(A^N - \mu) - N^{(1-\beta')/2}(D^N - \mu)\| > \delta) = -\infty. \quad (9.6)$$

Thus at scale $\beta' < \beta$, D^N satisfies exactly the same moderate deviations principle as does A^N . In other words, the burstiness of the traffic at scales $\beta' < \beta$ has not been affected at all. (We will not give a full proof; see [102] for that. In a moment is a sketch proof, and a full proof of an important step.)

What about scales $\beta' > \beta$? This is harder to say. One statement though is trivial. The queue cannot emit work at a rate greater than its service rate; so $ND^N(-t, 0] \leq N\mu + N^{(1+\beta)/2}C$; so if ND^N were fed into a downstream queue with service rate $N\mu + N^{(1+\beta')/2}C'$, that downstream queue would never overflow (for N sufficiently large).

Sketch proof of (9.6). Let $\beta' < \beta$. Substituting into (9.6) the definition of D^N , and rescaling, we need to show

$$\limsup_{N \rightarrow \infty} \frac{1}{N^{\beta'}} \log P\left(\sup_{t>0} N^{(1+\beta)/2} \left| \frac{Q_0^N}{t+1} - \frac{Q_{-t}^N}{t+1} \right| > \delta N^{(1+\beta')/2}\right) = -\infty.$$

Now,

$$\sup_{t>0} \left| \frac{Q_0^N}{t+1} - \frac{Q_{-t}^N}{t+1} \right| \leq Q_0^N + \sup_{t>0} \frac{Q_{-t}^N}{t+1}.$$

The following lemma proves that

$$\limsup_{N \rightarrow \infty} \frac{1}{N^{\beta'}} \log P(N^{(1+\beta)/2} Q_0^N > N^{(1+\beta')/2} \delta) = -\infty.$$

With some harder work, using essentially the same technique, we can prove a similar result for $\sup_t Q_{-t}^N/(t+1)$. Hence the result. \square

Lemma 9.3 *If the arrival process A^N , normalized, satisfies the sample path moderate deviations principle at scale β' ; and if $Q_0^N = q(LA^N, N\mu + N^{(1+\beta)/2}C)$ for $\beta' < \beta$ and $\mu < C$, then*

$$\limsup_{N \rightarrow \infty} \frac{1}{N^{\beta'}} \log P(N^{(1+\beta)/2} Q_0^N > N^{(1+\beta')/2} \delta) = -\infty. \tag{9.7}$$

Proof. The proof consists mostly in changing the scales of the equation, as follows.

$$\begin{aligned} (9.7) &= \limsup_{N \rightarrow \infty} \frac{1}{N^{\beta'}} \log P(q(NA^N, N\mu + N^{-(1-\beta)/2}C) > \delta N^{(1+\beta')/2}) \\ &= \limsup_{N \rightarrow \infty} \frac{1}{N^{\beta'}} \log P(q(N^{(1-\beta')/2}(A^N - \mu), N^{(\beta-\beta')/2}C) > \delta) \\ &\leq J(\delta, C') \quad \text{for all } C' > 0 \end{aligned}$$

where $J(\cdot, C')$ is the rate function for queue size in a queue with service rate C' . But we have a formula for J :

$$J(q, C) = \inf_{t \geq 0} \frac{(q + Ct)^2}{2V_t} \geq C^2 \inf_{t \geq 0} \frac{t^2}{2V_t}.$$

Since we have assumed $V_t = o(t^2/\log t)$, $J(q, C) \rightarrow \infty$ as $C \rightarrow \infty$. Hence the result. \square

Exercise 9.3

Consider a priority queue. Suppose the high priority input is NA^N , the sum of N independent copies of a traffic flow A , and the low priority input is NB^N , the sum of N independent copies of a traffic flow B . Let the buffer be infinite, and let the service rate be $N\mu + N^{(1+\beta)/2}$ where μ is the mean rate of $A^N + B^N$. Let Q^N be the total queue size at time 0, R^N the high priority queue size and S^N the low priority queue size. Write down a moderate deviations principle for $N^{-(1+\beta)/2}Q^N$. Show that $N^{-(1+\beta)/2}(R^N, S^N)$ is exponentially equivalent to $N^{-(1+\beta)/2}(0, Q^N)$. Recalling Chapter 7, find a large deviations principle for $N^{-1}R^N$ with speed N . \diamond

Should be $\mu \in \mathcal{D}$ not plain $\mu \in \mathbb{R}$

Scaling	$M_t(\theta)$
LDLB §3.1	$M_t(\theta) = t\Lambda_\infty(\theta),$ where $\Lambda_\infty(\theta) = \lim_{t \rightarrow \infty} t^{-1}\Lambda_t(\theta)$
LDMF §1.4	$M_t(\theta) = \Lambda_t(\theta)$
LDLBH §8.2	$M_t(\theta) = t^{2(1-H)}\Lambda_{\infty(H)}(\theta t^{2H-1}),$ where $\Lambda_{\infty(H)}(\theta) = \lim_{t \rightarrow \infty} t^{-2(1-H)}\Lambda_t(t^{-(2H-1)}\theta)$
MDMF §9.3	$M_t(\theta) = \theta\mu t + \frac{1}{2}\theta^2\sigma_t^2$ where $\mu t = EA(-t, 0] = \Lambda'_t(0)$ and $\sigma_t^2 = \text{Var } A(-t, 0] = \Lambda''_t(0)$
MDLB §9.3	$M_t(\theta) = \theta\mu t + \frac{1}{2}\theta^2 t\sigma^2$ where $\sigma^2 = \lim_{t \rightarrow \infty} t^{-1}\sigma_t^2 = \Lambda''_\infty(0)$

Table 10.1: The log moment generating function $M_t(\theta)$ appearing in the rate function (10.4), for various different scaling regimes.

Zeitouni [25, Theorem 3.7.4], to find conditions under which they can obtain a tighter limit on the probability of overflow in a queue

$$P(Q^N = Nq) = \frac{1}{\hat{\theta}\sqrt{2\pi N\sigma^2(\hat{t}, \hat{\theta})}} e^{-NI(q)} \left(1 + O(N^{-1})\right) \quad (10.5)$$

and on L^N , the expected amount of work that is lost each timestep,

$$L^N = \frac{1}{\hat{\theta}^2\sqrt{2\pi N\sigma^2(\hat{t}, \hat{\theta})}} e^{-NI(q)} \left(1 + O(N^{-1})\right).$$

Note that this work deals with a finite-buffer queue (otherwise it wouldn't make sense to measure L^N). The queue size can't exceed the buffer size Nq , which is why we measure $P(Q^N = Nq)$ rather than $P(Q^N \geq Nq)$.

Here, \hat{t} and $\hat{\theta}$ are the optimizing parameters in $I(q)$, μ is the mean arrival rate of a single flow, and $\sigma^2(\hat{t}, \hat{\theta})$ is the 'tilted variance'

$$\sigma^2(t, \theta) = \frac{d^2}{d\theta^2} \Lambda_t(\theta).$$

Again, the same expression (10.5) holds for the probability that the queue size exceeds Nq in a queue with an infinite buffer.

To turn these limit theorems into estimates, simply set $N = 1$ and write $I(q)$ in terms of the actual parameters, namely the total buffer size and service rate, and aggregate arrival process. Kelly [55] recounts similar results for bufferless resources shared by many flows.

In the large-buffer scaling, Choudhury et al. [18] explain that it is often possible to show that, for a queue with service rate C and buffer size q ,

$$P(\text{overflow}) \sim ae^{-I(q)} \quad \text{as } q \rightarrow \infty \quad (10.6)$$

for some constant a , where $I(q)$ is given by (10.4) using the large-buffer version of M_t . They remark that a is often close to 1 when the queue is fed by a single source. When the queue is fed by many sources, a can be far from 1, and so large-buffer approximation is unsuitable.

We conjecture that these refined approximations also apply to the moderate-deviations scalings, with $\Lambda_t(\theta)$ replaced by $M_t(\theta)$ as specified in Table 10.1.

When the input traffic is Gaussian, one can say more. Choe and Shroff [16] have found a tight upper bound for the constant a in (10.6), when the input is not long-range dependent. They go on to show in [17] that for a wider class of Gaussian processes, including long-range dependent processes,

$$P(Q > q) \leq e^{-I(q)} q^{K+o(1)} \quad \text{as } q \rightarrow \infty$$

for some constant K , where $I(q)$ is given by (10.4) with the *many-flows* version of M_t . This result is intriguing, because it involves the many-flows rate function yet describes a large-buffer limit.

10.2.4 Numerical comparison

Now we are ready to numerically compare these different estimates. Figure 10.1 is a *watermark plot* of queue length. What this means is that we run a simulation of the queue with an infinite buffer, and count the proportion of time $P(b)$ that it spends with buffer size greater than b ; we then plot $\log P(b)$ against b . This sort of plot emphasizes the behaviour of the queue for large buffer sizes: the limiting slope of $-\log P(b)$ is the LDLB rate function $I(b)$, and the vertical intercept tells us about the prefactor a in (10.6).

The parameters for Figure 10.1 are as follows. The queue has service rate 1. The traffic is generated by a Markov on/off source with peak rate 2, which jumps from off to on with probability 2/15 and from off to on with probability 3/15, so that the mean arrival rate is 4/5. Theory says that the LDLB estimate has the right limiting slope; the plot shows that it has nearly the right prefactor.

The simulated watermark curves are typical. When b is small, the probability of $\{Q > b\}$ is reasonably large, so $-\log P(Q > b)$ is easy to estimate and the watermark curves are close to the truth. When b is large the probability is small, and many simulation runs do not even reach $\{Q > b\}$, so the

off to on with probability 2/15 and on to off with probability 3/15

where α and β depend only on the peak rate, the service rate, and ET and EU , but *not* on the distribution of T or U .

They also show that this result extends to source models which have more than two states, where the durations in each state are independent and generally distributed, and the transitions are Markov.

Mandjes and Borst [68] have found the form of $I(q)$ for large q . Assume that EU and $ET^{1+\varepsilon} > 0$ are finite, for some $\varepsilon > 0$, and that $T + U$ is non-lattice. Let T^* be the *residual active-time*,

$$P(T^* > t) = \frac{1}{ET} \int_{u=t}^{\infty} P(T > u) du$$

$$\int_{u=t}^{\infty} P(T > u) du$$

and let $v_t = -\log P(T^* > t)$. If T^* is subexponential and subexponentially varying of index $h \in [0, 1)$ (see the reference for a definition of these two terms; also note that this class includes Pareto, lognormal and Weibull distributions) they show

$$I(q) = \alpha v_q (1 + o(1)) \tag{10.13}$$

where α depends only on the peak rate, the service rate, ET and EU , and h , but not otherwise on the distribution of T or U .

The scaling function v_t is important in proving this result. They use the transformation

$$I(q) = \inf_t v_t \tilde{\Lambda}_t^*(q/t + c)$$

where $\tilde{\Lambda}_t^*$ is the convex conjugate of $\tilde{\Lambda}_t$,

$$\tilde{\Lambda}_t(\theta) = \frac{1}{v_t} \log E e^{\theta A(-t, 0] v_t / t}$$

(Compare to Theorem 3.5 and Definition 7.2.) Their proof hinges on a limit result expressed by the approximation

$$\begin{aligned} E e^{\theta A(-t, 0] v_t / t} &\approx P(A^* > t) e^{\theta r v_t} + (1 - P(A^* > t)) e^{\theta \rho v_t} \\ &\approx \exp[v_t (\theta \rho \vee (\theta p - 1))] \end{aligned}$$

which has the interpretation that, over an interval of length t , either the source is always on, or it is sending at the mean rate.

10.5.4 $M/G/\infty$ models

In the $M/G/\infty$ model, calls arrive as a Poisson process, remain active for some duration T , and then depart. While active they produce work at

This makes it very simple to estimate the probability of overflow. (There are, however, pitfalls in using large deviations to study other quantities—like the departure process—in queues fed by Gaussian processes, if these are intended as heavy traffic approximations to non-Gaussian processes. See the note at the end of Section 9.3.)

Fractional Brownian motion. The archetypal Gaussian source is fractional Brownian motion. Let $A(-t, 0] = \mu t + \sigma Z_t$, where Z_t is a standard fractional Brownian motion with Hurst parameter $H \in (0, 1)$. Then

$$A(-t, 0] \sim \text{Normal}(\mu t, \sigma^2 t^{2H})$$

and

$$I(q) = \frac{(C - \mu)^{2H} q^{2(1-H)}}{2\sigma^2} \frac{1}{H^{2H} (1-H)^{2(1-H)}}$$

and the most likely time to overflow is

$$t = \frac{q}{c - \mu} \frac{H}{1 - H}.$$

Compare to Theorem 8.1. As Addie et al. [1] point out, the approximation this leads to is exact for Brownian motion ($H = \frac{1}{2}$) and reasonably good otherwise.

For fractional Brownian motion traffic, the many-flows limit and the large-buffer limit are related. This is because of the self-similarity relationship

$$Z \sim \frac{1}{a^{2H}} Z^{\circ a}$$

where $Z^{\circ a}$ denotes the speeded-up process $Z_t^{\circ a} = Z_{at}$. Thus if A^N is the average of N independent copies of A ,

$$A^N|_{(-t,0]} \sim \frac{1}{N^{1/(2-2H)}} A|_{(-N^{1/(2-2H)}t,0]},$$

Purely from the definition of the queue size function

$$Q(A) = \sup_{t \geq 0} A(-t, 0] - Ct$$

we obtain

$$P(Q(A^N) > q) = P(Q(A) > N^{1/(2-2H)} q).$$

In this way, the many-flows LDP is equivalent to the large-buffer LDP in the case of self-similar traffic.

punctuation: . not ,

typography: use bigger brackets