Large Deviations with Applications to Telecommunications

Neil O'Connell
Basic Research Institute in the Mathematical Sciences
HP Laboratories Bristol
HPL-BRIMS-2000-27
December 18th, 2000*

queueing networks

These are lecture notes for a course given at Uppsala University in November 1999. Acknowledgments are due to my fellow instructors, John Lewis, Raymond Russell and Fergal Toomey; the organisers, Ingemar Kaj and Tobias Ryden; and, of course the students. Thanks also to the Swedish Foundation for Strategic Research for funding the course.

* Internal Accession Date Only
Approved for External Publication

1 The single server queue

Let $(X_n,\ n \in \mathbb{Z})$ be a stationary ergodic sequence of random variables with $EX_0 < 0$, and consider the recursion

$$Q_n = (Q_{n-1} + X_n)^+. \tag{1}$$

This is known as Lindley's recursion. It arises in the analysis of first-come-first-served single server queues, in both continuous and discrete-time settings.

In continuous time, customers are labelled by the integers and X_n is the difference between the service time of customer n and the interarrival time between customers n and $n+1$; in this case Q_n is the *waiting time* of customer $n+1$ (that is, the time spent in the queue before commencing service).

In discrete time, X_n is the difference between the amount of work to arrive at the queue at time n and the amount of work which can be processed at that time; in this case, Q_n is the amount of work remaining in the queue.

We shall adopt the latter interpretation, but clearly most of the results presented here will have implications in the former context.

It was shown by Loynes [40] that, for any initial condition Q_0, the law of Q_n converges as $n \to \infty$ to a unique equilibrium distribution (independent of Q_0). Moreover, the sequence

$$Q_n = \left[\sup_{m \le n} \sum_{k=m}^{n} X_k \right]^+, \tag{2}$$

$n \in \mathbb{Z}$, defines a stationary ergodic solution.

Exercise 1.1 *Show that the sequence Q_n defined by (2) satisfies (1). Convince yourself that this is the unique solution by repeatedly applying the Lindley recursion (1).*

An nice example to keep in mind is the following. Suppose the X_n are iid with $P(X_0 = 1) = 1 - P(X = -1) = p < 1/2$. Then the process Q (defined by (2)) is a stationary birth and death Markov chain with equilibrium distribution

$$P(Q_0 \geq q) = \left(\frac{p}{1-p}\right)^q \tag{3}$$

for $q \in \mathbb{Z}_+$. This is a discrete-time analogue of the M/M/1 queue. We shall rewrite (3) as

$$\log P(Q_0 \geq q) = -\delta q \tag{4}$$

where $\delta = \log[(1-p)/p]$.

It is a remarkable fact that an approximate version of the formula (4) holds quite generally: for large q,

$$\log P(Q_0 \geq q) \sim -\delta q \tag{5}$$

for some $\delta > 0$. We will soon make this statement precise and give a proof using large deviation theory. But first, let us consider the implications.

If (5) holds we can (in principle) estimate the frequency with which large queues build up by empirically observing the queue-length distribution over a relatively short time period: plot the log-frequency with which each level q is exceeded against q, and linearly extrapolate. I have qualified this statement because actually this is a very challenging statistical problem. Nevertheless,

this ingeneous idea, which was first proposed in [11], has inspired major new developments in the application of large deviation theory to queueing networks and network management generally.

Before we give a formal statement and proof of (5) we present some background material on one-dimensional large deviation theory. For more detailed accounts, and more on the theory of large deviations in general, see [15, 14, 25, 39].

1.1 One-dimensional large deviation theory

1.1.1 Cramér's theorem

Let Y_k be a sequence of iid random variables and set $S_n = Y_1 + \cdots + Y_n$. The cumulant generating function associated with Y_1 is defined by

$$\Lambda(\theta) = \log E e^{\theta Y_1}.$$

This is a convex function on $\mathbb{R}$ taking values in the extended real numbers $\mathbb{R}^* == (-\infty, +\infty]$.

Exercise 1.2 *Prove that Λ is convex. (Hint: use Hölder's inequality.)*

The *convex dual*, or *Fenchel-Legendre transform*, of Λ is a non-negative function on $\mathbb{R}$ defined by

$$\Lambda^*(x) = \sup_{\theta \in \mathbb{R}} [\theta x - \Lambda(\theta)].$$

Theorem 1.1 *The sequence of random variables S_n/n satisfies the large deviation principle with rate function Λ^*: for all closed sets F,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P(S_n \in F) \leq -\inf_F \Lambda^*, \tag{6}$$

and for all open sets G,

$$\liminf_{n\to\infty} \frac{1}{n} \log P(S_n \in G) \geq -\inf_G \Lambda^*. \tag{7}$$

The inequality (6) is usually referred to as the large deviations upper bound, and (7) as the large deviations lower bound. If both hold we say that the sequence S_n/n satisfies the large deviation principle with rate function Λ^*.

1.1.2 A generalisation of Cramér's theorem

Cramér's theorem generalises far beyond the realm of sums of iid random variables. In the standard one-dimensional generalisation, S_n is any sequence of random variables, and the statement of Cramér's theorem can be shown to hold with Λ defined to be the limiting scaled cumulant generating function

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E e^{\theta S_n},$$

provided this limit exists and is well-behaved. It is easy to see that this is consistent with the case of sums of iid random variables.

To state this generalisation we need some definitions. Let f be a function on $\mathbb{R}$ which takes values in the extended real numbers. The *effective domain* of f is defined by $\mathcal{D}_f = \{\theta : \ f(\theta) < \infty\}$. The function f is *steep* if, for any sequence θ_n which converges to a boundary point of $\mathcal{D}_f$, $\lim_{n\to\infty} |f'(\theta_n)| =$

$+\infty$. Note that $-\infty$ and $+\infty$ are never considered to be boundary points. The function f is *essentially smooth* if it is steep, the interior of its effective domain is non-empty and it is differentiable there.

Let S_n be a sequence of random variables with respective cumulant generating functions

$$\Lambda_n(\theta) = \log E e^{\theta S_n}.$$

Theorem 1.2 *If the limiting scaled cumulant generating function*

$$\Lambda(\theta) = \lim_{n\to\infty} \Lambda_n(\theta)/n \tag{8}$$

exists for each $\theta \in \mathbb{R}$ as an extended real number and zero lies in the interior of its effective domain then, for all closed sets F,

$$\limsup_{n\to\infty} \frac{1}{n} \log P(S_n \in F) \leq -\inf_F \Lambda^*. \tag{9}$$

If, in addition, Λ is essentially smooth, the corresponding lower bound

$$\liminf_{n\to\infty} \frac{1}{n} \log P(S_n \in G) \geq -\inf_G \Lambda^* \tag{10}$$

holds for all open sets G.

If the upper and lower bounds of Theorem 1.2 hold, we say the sequence S_n/n satisfies the large deviation principle (LDP) with rate function Λ^*.

Let Y_k be a sequence of random variables and set

$$S_n = Y_1 + \cdots + Y_n.$$

Exercise 1.3 *Compute Λ and Λ^* for the following models.*

(1) Y_k iid Poisson with mean λ.

(2) Y_k iid exponential with mean μ.

(3) Y_k iid Gaussian with mean μ and variance σ^2.

(4) Y_k is a stationary auto-regressive process of degree 1. That is,

$$Y_0 = \sum_{k=0}^{\infty} \alpha^k \epsilon_{-k},$$

and

$$Y_k = \alpha Y_{k-1} + \epsilon_k$$

for all $k > 0$, where $-1 < \alpha < 1$ and the ϵ_k are iid Gaussian with zero mean and variance σ^2.

Theorem 1.2 has a converse. A function $I : \mathbb{R} \to [0, \infty]$ is a *good rate function* if the level set $\{x : \ I(x) \leq \alpha\}$ is compact for each $\alpha \geq 0$.

Theorem 1.3 *If the sequence S_n/n satisfies the LDP with good rate function I and*

$$\limsup_{n\to\infty} \frac{1}{n} \log E \exp((\theta + \epsilon) S_n) < \infty,$$

for some $\epsilon > 0$, then

$$\lim_{n\to\infty} \frac{1}{n} \log E e^{\theta S_n} = I^*(\theta).$$

This is a consequence of Laplace's method, and can also be regarded as a special case of Varadhan's lemma. Note that I needn't be convex, whereas Theorem 1.2 always leads to convex rate functions.

A related thing is the *principle of the largest term.* Let a_n and b_n be positive sequences of real numbers. If

$$\lim_{n\to\infty} \frac{1}{n} \log a_n = a$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log b_n = b,$$

then

$$\lim_{n\to\infty} \frac{1}{n} \log(a_n + b_n) = \max\{a, b\}.$$

This extends easily to finite sums.

1.1.3 Russell's time-change formula

The following is a useful trick for computing Λ^* in the case of 'on/off sources'. It follows from Russell's time-change formula (see [52] for more details). Let R_k and T_k be a pair of iid sequences, and set $Y_k = 1$ for $k \leq R_1$, $Y_k = 0$ for $R_1 < k \leq R_1 + T_1$, $Y_k = 1$ for $R_1 + T_1 < k \leq R_1 + T_1 + R_2$, and so on. As before, set $S_n = Y_1 + \cdots + Y_n$. If the cumulant generating functions $\Lambda_R(\theta) = \log Ee^{\theta R_1}$ and $\Lambda_T(\theta) = \log Ee^{\theta T_1}$ are finite in a neighbourhood of zero, the sequence $S_n/n = (Y_1 + \cdots + Y_n)/n$ satisfies the LDP with rate function given by

$$I(x) = \inf_{a>0} a[\Lambda_R^*(x/a) + \Lambda_T^*((1-x)/a)]. \tag{11}$$

Exercise 1.4 *Try to give a heuristic proof of this formula, using the principle of the largest term.*

Exercise 1.5 *Use (11) to compute the rate function associated with S_n/n when Y_k is a Markov chain on $\{0,1\}$. (Hint: use the fact that the times between transitions are independent and geometrically distributed.) If the transition probabilities are given by $p(0,1) = a$ and $p(1,0) = d$, show that*

$$\Lambda(\theta) = \log\left(c + \sqrt{c^2 - (1-a-d)e^{\theta}}\right),$$

where $2c = 1 - a + (1-d)e^{\theta}$. (Hint: use Theorem 1.3.)

1.2 Application to the single-server queue

The following is one of the fundamental theorems in the application of large deviation theory to queueing networks. It has been demonstrated by several authors, under similar conditions [6, 16, 20, 30]. The proof given here is similar to the proof given in [6].

Recall that the queue-length at time zero is given by

$$Q_0 = \sup_{n \geq 0} S_n^+$$

where

$$S_n = X_0 + X_{-1} + \cdots + X_{-n}.$$

Set

$$\Lambda_n(\theta) = \log E e^{\theta S_n}.$$

Theorem 1.4 *Suppose that the limiting scaled cumulant generating function*

$$\Lambda(\theta) = \lim_{n \to \infty} \Lambda_n(\theta)/n \tag{12}$$

exists for each $\theta \in \mathbb{R}$ *as an extended real number, and* $\Lambda(\theta) < 0$ *for some* $\theta > 0$.

If $\Lambda_n(\theta) < \infty$ *for all* $\theta > 0$ *such that* $\Lambda(\theta) < 0$, *then*

$$\limsup_{q\to\infty} \frac{1}{q} \log P(Q_0 \geq q) \leq -\delta, \tag{13}$$

where

$$\delta = \sup\{\theta > 0 : \ \Lambda(\theta) < 0\}. \tag{14}$$

If the large deviations lower bound

$$\liminf \frac{1}{n} \log P(S_n > xn) \geq \Lambda^*(x)$$

holds for all $x > 0$, *then*

$$\liminf_{q\to\infty} \frac{1}{q} \log P(Q_0 \geq q) \geq -\delta. \tag{15}$$

Combining this with Theorem 1.2 we have:

Corollary 1.5 *If the limiting scaled cumulant generating function*

$$\Lambda(\theta) = \lim_{n\to\infty} \Lambda_n(\theta)/n \tag{16}$$

exists for each $\theta \in \mathbb{R}$ *as an extended real number,* $\Lambda(\theta) < 0$ *for some* $\theta > 0$, $\Lambda_n(\theta) < \infty$ *for all* $\theta > 0$ *such that* $\Lambda(\theta) < 0$, *and* Λ *is essentially smooth, then*

$$\lim_{q\to\infty} \frac{1}{q} \log P(Q_0 \geq q) = -\delta.$$

As will become clear when we present the proof, one interpretation of Theorem 1.4 is that the naive approximation

$$P(\sup_n S_n \geq q) \simeq \sup_n P(S_n \geq q)$$

is justified on a logarithmic scale.

Proof of Theorem 1.4 — *Upper bound.* Fix $\theta > 0$ with $\Lambda(\theta) < 0$. By the inequalities of Boole and Markov, for $q > 0$,

$$\begin{aligned} P(Q_0 \geq q) &= P(\sup_{n \geq 0} S_n \geq q) \\ &\leq \sum_{n \geq 0} P(S_n \geq q) \\ &\leq e^{-\theta q} \sum_{n \geq 0} e^{\Lambda_n(\theta)}. \end{aligned}$$

Now choose $\epsilon > 0$ such that $\Lambda(\theta) + \epsilon < 0$. By hypothesis, $\Lambda_n(\theta)/n \leq \Lambda(\theta) + \epsilon$ for all n sufficiently large, so there exists a constant $C > 0$ such that

$$e^{\Lambda_n(\theta)} \leq C e^{[\Lambda(\theta)+\epsilon]n},$$

for all n. Thus,

$$\begin{aligned} P(Q_0 \geq q) &\leq C e^{-\theta q} \sum_{n \geq 0} e^{[\Lambda(\theta)+\epsilon]n} \\ &= C e^{-\theta q} e^{\Lambda(\theta)+\epsilon} (1 - e^{\Lambda(\theta)+\epsilon})^{-1}, \end{aligned}$$

and so

$$\limsup_{q \to \infty} \frac{1}{q} \log P(Q_0 \geq q) \leq -\theta.$$

Since this holds for any θ with $\Lambda(\theta) < 0$ we have established the upper bound.

Lower bound. We can assume that $\delta < \infty$, because otherwise there is nothing to prove. For any $\tau > 0$,

$$\begin{aligned}
\liminf_{q\to\infty} \frac{1}{q} \log P(Q_0 \geq q) &= \liminf_{q\to\infty} \frac{1}{q} \log P \bigcup_{n\geq 0} \{S_n \geq q\} \\
&\geq \liminf_{q\to\infty} \frac{1}{q} \log P\{S_{[\tau q]} \geq q\} \\
&= -\tau \Lambda^*(1/\tau).
\end{aligned}$$

Thus,

$$\liminf_{q\to\infty} \frac{1}{q} \log P(Q_0 \geq q) \geq -\inf_{\tau>0} \tau\Lambda^*(1/\tau).$$

To complete the proof:

$$\begin{aligned}
\sup\{\theta > 0 :\ \Lambda(\theta) < 0\} &= \sup\{\theta > 0 :\ \sup_{x\in\mathbb{R}} [x\theta - \Lambda^*(x)] < 0\} \\
&= \sup\{\theta > 0 :\ x\theta - \Lambda^*(x) < 0 \text{ for all } x \in \mathbb{R}\} \\
&= \sup\{\theta > 0 :\ \theta < \Lambda^*(x)/x \text{ for all } x > 0\} \\
&= \inf_{\tau>0} \tau\Lambda^*(1/\tau)
\end{aligned}$$

Here we have used the fact that $\Lambda^*(0) = -\inf \Lambda > 0$. □

Recall that we are interpreting X_k as the difference $A_k - C_k$ between the amount of work arriving at time k and the amount of work which can be processed at time k. If we assume that the sequence A is independent of C then $\Lambda(\theta) = \Lambda_A(\theta) + \Lambda_C(-\theta)$, where

$$\Lambda_A(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E \exp\left(\theta \sum_{k=1}^{n} A_k\right)$$

and Λ_C is defined similarly.

Exercise 1.6 *(1) Show that* $\Lambda^*(x) = \inf_y [\Lambda_A^*(y) + \Lambda_C^*(y - x)]$.

(2) If $C_k = c$ for all k, then $\Lambda_C(\theta) = c\theta$ and

$$\delta = \sup\{\theta > 0 : \ \Lambda_A(\theta) \leq c\theta\} = \inf_{x>0} \Lambda_A^*(x+c)/x.$$

Exercise 1.7 *Compute δ for the following examples.*

(1) A_k iid Poisson with mean λ and C_k iid Poisson with mean $\mu > \lambda$.

(2) A_k iid exponential with mean λ and C_k iid exponential with mean $\mu > \lambda$.

(3) X_k iid Gaussian with mean $\mu < 0$ and variance σ^2.

(4) A_k a Markov chain on $\{0, 1\}$ and $C_k = c$ for all k where c is bigger that the equilibrium probability of the chain being in state 1.

1.3 The many-sources asymptotic

There is also an asymptotic regime which considers what happens when a queue, or network of queues, is shared by a large number of independent traffic sources. This is interesting not just from a potentially practical point of view, but also because it demonstrates the benefits of statistical multiplexing.

Consider a single-server queue as before, with N sources and constant service capacity cN. Denote by A_k^i the amount of work which arrives from source i at time k. For each i, $(A_k^i, \ k \in \mathbb{Z})$ is a stationary and ergodic sequence of random variables and these sequences are assumed to be independent of each other and identically distributed. For stability we require $EX_0^1 < c$. To put this in a familiar context, set

$$X_k^N = A_k^1 + \cdots + A_k^N - cN.$$

Then the queue-length at time zero is given by

$$Q_0^N = \left[\sup_{n\geq 0} S_n^N\right]^+$$

where

$$S_n^N = X_0^N + X_{-1}^N + \cdots + X_{-n}^N.$$

We will consider the asymptotic behaviour of $P(Q_0^N \geq q)$ as the number of sources N becomes large. Using similar techniques as before we obtain the following result. Set

$$\Lambda_n(\theta) = \log E \exp\left(\theta \sum_{k=-n}^{0} A_k^1\right)$$

and, for $q \geq 0$,

$$I(q) = \inf_{n\geq 0} \Lambda_n^*(q + cn).$$

Theorem 1.6 *Fix $q > 0$. The lower bound*

$$\liminf_{N\to\infty} \frac{1}{N} \log P(Q_0^N > qN) \geq -I(q),$$

holds without any assumptions. If

$$\limsup_{m\to\infty} \limsup_{N\to\infty} \frac{1}{N} \log \sum_{n>m} e^{-\Lambda_n^*(q+cn)N} \leq -I(q), \tag{17}$$

then

$$\lim_{N\to\infty} \frac{1}{N} \log P(Q_0^N > qN) = -I(q).$$

Proof of Theorem 1.6 — *Lower bound.* For each n,

$$\begin{aligned}
\liminf_{N\to\infty} \frac{1}{N} \log P(Q_0^N > qN) &= \liminf_{N\to\infty} \frac{1}{N} \log P(\sup_{m\geq 0} S_m^N > qN) \\
&\geq \liminf_{N\to\infty} \frac{1}{N} \log P(S_n^N > qN) \\
&= \liminf_{N\to\infty} \frac{1}{N} \log P\left(\sum_{i=1}^{N} \sum_{k=-n}^{0} A_k^i > (q+cn)N\right) \\
&\geq -\Lambda_n^*(q+cn),
\end{aligned}$$

by Cramér's theorem. Now optimise this bound over n.

Upper bound. We can apply Boole's inequality and Markov's inequality as before, but this time with a sequence θ_n:

$$\begin{aligned}
P(Q_0^N > qN) &= P(\sup_{n\geq 0} S_n^N > qN) \\
&\leq \sum_{n\geq 0} P(S_n^N > qN) \\
&= \sum_{n\geq 0} P\left(\sum_{i=1}^{N} \sum_{k=-n}^{0} A_k^i > (q+cn)N\right) \\
&\leq \sum_{n\geq 0} e^{-[\theta_n(q+cn)-\Lambda_n(\theta_n)]N}.
\end{aligned}$$

For each n we can choose the optimal value of θ_n and this becomes

$$P(Q_0^N > qN) \leq \sum_{n\geq 0} e^{-\Lambda_n^*(q+cn)N}.$$

The hypothesis (17) was chosen specifically to control the terms in the 'tails' of this summation. By the principle of the largest term,

$$\limsup_{N\to\infty} \frac{1}{N} \log P(Q_0^N > qN) \leq \max\left\{-\inf_{0\leq n\leq m} \Lambda_n^*(q+cn), \epsilon(m)\right\},$$

where

$$\epsilon(m) = \limsup_{N\to\infty} \frac{1}{N} \log \sum_{n>m} e^{-\Lambda_n^*(q+cn)N}.$$

By hypothesis, $\limsup_{m\to\infty} \epsilon(m) \leq -I(q)$, and the result follows. □

The many sources asymptotic first appeared in [57]. Variants of Theorem 1.6 are given in [5] and [12]. For the last word, see [59].

1.4 Effective bandwidths

For a single-server queue with arrivals process A_k we can ask: how much service capacity do we need in order to ensure that

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

for large q and some *prespecified* value of δ? From Theorem [30] (see also Exercise 1.6) we see that the answer to this question, assuming the conditions are satisfied, is approximately $\Lambda_A(\delta)/\delta$, where

$$\Lambda_A(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E \exp\left(\theta \sum_{k=1}^{n} A_k\right).$$

This quantity is called the *effective bandwidth* of the source A.

The notion of effective bandwidths was introduced by Kelly [34]. See also [29, 35], where the two-parameter effective bandwidth

$$\alpha(\delta, n) = \frac{1}{\delta n} \log E \exp\left(\sum_{k=1}^{n} A_k\right)$$

is proposed as a more detailed traffic descriptor. Note that that this function also contains information which is relevant to the many sources asymptotic.

2 Queueing networks

The techniques outlined in the previous section are ideal for studying the single-server queue but soon become cumbersome when one tries to apply them to more complicated queueing networks. In this section we present a 'variational' approach which has many advantages.

2.1 The general framework

The following is a general scheme which can be applied to an endless variety of network problems where the goal is to establish probability approximations for aspects of a system (such as queue lengths) under very general ergodicity and mixing assumptions about the network inputs.

We will suppose that the inputs to a network can be represented by a sequence of random variables (X_k) in $\mathbb{R}^d$, and that the (sequence of) objects of interest, (O_n), can be expressed as a continuous function of the partial sums process corresponding to X. To make this more precise, for $t \geq 0$ set

$$S_n(t) = \frac{1}{n} \sum_{k=0}^{[nt]} X_{-k}. \tag{18}$$

Write $\tilde{S}_n$ for the polygonal approximation to S_n:

$$\tilde{S}_n(t) = S_n(t) + \left(t - \frac{[nt]}{n}\right) \left(S_n\left(\frac{[nt]+1}{n}\right) - S_n\left(\frac{[nt]}{n}\right)\right). \tag{19}$$

For $\mu \in \mathbb{R}^d$, denote by $\mathcal{A}_\mu$ the space of absolutely continuous paths $\phi : \mathbb{R}_+ \to \mathbb{R}^d$, with $\phi(0) = 0$ and limits $\lim_{t\to\infty} \phi(t)/t = \mu$, equipped with the topology

induced by the norm

$$\|\phi\|_u = \sup_t \left| \frac{\phi(t)}{1+t} \right|. \tag{20}$$

Our supposition is that there exists a continuous function $f : \mathcal{A}_\mu \to \mathcal{X}$, for some Hausdorff topological space $\mathcal{X}$, such that $O_n = f(\tilde{S}_n)$, for each n. (Note that we are also implicitly assuming that $\tilde{S}_n \in \mathcal{A}_\mu$, for each n.)

For example, consider the single-server queue. In this case $d = 1$ and X_k is the difference between the amount of work arriving at time k and the amount of work that can be processed at that time. If X is stationary and ergodic with $\mu = EX_0 < 0$, then

$$\lim_{t\to\infty} \tilde{S}_n(t)/t = \lim_{n\to\infty} \sum_{k=1}^{n} X_{-k}/n = \mu$$

almost surely and hence $\tilde{S}_n$ is almost surely in $\mathcal{A}_\mu$ for each n. Recall (Section 1) that the queue length at time zero (for the equilibrium system) is given by

$$Q_0 = \sup_{n\geq 0} \sum_{k=0}^{n} X_{-k}, \tag{21}$$

or, equivalently, $Q_0/n = f(\tilde{S}_n)$, where $f : \mathcal{A}_\mu \to \mathbb{R}_+$ is defined by

$$f(\phi) = \sup_{t>0} \phi(t). \tag{22}$$

Note that f is only well-defined when $\mu < 0$.

Exercise 2.1 *Check that f is a continuous function.*

Why is this a useful supposition? To answer this, we need to introduce some rather abstract large deviation theory.

2.2 Large deviations and the contraction principle

Let $\mathcal{X}$ be a Hausdorff topological space with Borel σ-algebra $\mathcal{B}$, and let μ_n be a sequence of probability measures on $(\mathcal{X}, \mathcal{B})$. We say that μ_n satisfies the *large deviation principle* (LDP) with rate function I, if $I : \mathcal{X} \to \mathbb{R}_+$ is lower semicontinuous and, for all $B \in \mathcal{B}$,

$$-\inf_{x \in B^\circ} I(x) \leq \liminf_n \frac{1}{n} \log \mu_n(B) \leq \limsup_n \frac{1}{n} \log \mu_n(B) \leq -\inf_{x \in \bar{B}} I(x); \quad (23)$$

if, for each n, Z_n is a realisation of μ_n, it is sometimes convenient to say that the sequence Z_n satisfies the LDP. A rate function I is *good* if its level sets $\{x : \ I(x) \leq \alpha\}$, $\alpha \geq 0$, are compact subsets of $\mathcal{X}$.

A useful tool in large deviation theory is the *contraction principle*. This states that if Z_n satisfies the LDP in a Hausdorff topological space $\mathcal{X}$ with good rate function I, and f is a continuous mapping from $\mathcal{X}$ into another Hausdorff topological space $\mathcal{Y}$, then the sequence $f(Z_n)$ satisfies the LDP in $\mathcal{Y}$ with good rate function given by $J(y) = \inf\{I(x) : \ f(x) = y\}$.

Consider the partial sums process $\tilde{S}_n$. The contraction principle tells us that, if the sequence $\tilde{S}_n$ satisfies the LDP in $\mathcal{A}_\mu$ with a good rate function I, for any continuous function f taking values in a Hausdorff topological space, the sequence $f(\tilde{S}_n)$ satisfies the LDP with good rate function given by $J(y) = \inf\{I(\phi) : \ f(\phi) = y\}$. In practise this will mean that once we have established (or simply assumed) the LDP for the partial sums process, we immediately have LDP's for the objects we are interested in provided we can write them as continuous functions of $\tilde{S}_n$. We only have to solve the variational problem to identify the rate function. This approach considerably

reduces the technical difficulties normally associated with proving LDP's in queueing networks (and in fact in many other applications of large deviation theory). In some sense it provides a mechanism for turning heuristics into theorems: solving the variational problem is equivalent to finding the most likely way in which the associated event occurs, and this, at least in many applications to queueing networks, is the heuristic which is used to predict the LDP one is trying to prove.

2.3 Large deviations for partial sums processes

Under quite general conditions, the sequence $\tilde{S}_n$ satisfies the LDP in $\mathcal{A}_\mu$ with good rate function given by

$$I(\phi) = \int_0^\infty \Lambda^*(\dot{\phi})ds,$$

where, as before, Λ^* is the convex dual of the limiting scaled cumulant generating function

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E e^{n\theta \cdot S_n(1)}, \tag{24}$$

which is assumed to exist for each $\theta \in \mathbb{R}^d$ as an extended real number, and also assumed to be differentiable at the origin with $\nabla\Lambda(0) = \mu$. In the applications we will consider next, we will regard this as a hypothesis. In this section, we will present a very quick summary of the relevant background.

Denote by $\tilde{S}_n^T$ the restriction of $\tilde{S}_n$ to the interval $[0, T]$ and by $\mathcal{A}^T$ the space of absolutely continuous functions on $[0, T]$ with $\phi(0) = 0$, eqquipped with the uniform topology. Dembo and Zajic [13] establish quite general conditions

for which $\tilde{S}_n^T$ satisfies the LDP in $\mathcal{A}^T$ with good convex rate function given by

$$I(\phi) = \int_0^T \Lambda^*(\dot{\phi})ds.$$

For the LDP to hold in the iid case, it is sufficient that the moment generating function $Ee^{\lambda \cdot X_1}$ exists and is finite everywhere; this is a classical result, due to Varadhan [56] and Mogulskii [43]. The proof given in Dembo and Zeitouni [14] only requires finiteness of Λ in a neighbourhood of the origin. This family of LDP's can immediately be extended to spaces of functions indexed by the entire half-line via the Dawson-Gärtner theorem for projective limits (see, for example, [14].) However, the projective limit topology (the topology of uniform convergence on compact intervals) is not strong enough for many applications; for example, the function f defined by (22) is not continuous in this topology on any supporting subspace, and so the contraction principle cannot be applied. This has motivated the consideration of stronger topologies by Dobrushin and Pechersky [18] and Ganesh and O'Connell [27]. In the latter it is proved that if the LDP holds for $\tilde{S}_n^1$ in $\mathcal{A}^1$ and Λ is differentiable at the origin with $\nabla\Lambda(0) = \mu$, then the LDP holds for $\tilde{S}_n$ in the space $\mathcal{A}_\mu$ with the topology induced by the norm (20), and with good convex rate function given by

$$I(\phi) = \int_0^\infty \Lambda^*(\dot{\phi})ds.$$

As we remarked earlier, the function f defined by (22) is continuous in this topology, provided $\mu < 0$.[1]

[1]The motivation for working with the norm (20) rather than the gauge topology introduced in [18] is the following. In the topology of [18], the mapping $\phi \mapsto \sup_t[\phi(t) - t]$ is

2.4 Solving the variational problem

Throughout the remainder of this section we will assume that the sequence $\tilde{S}_n$ satisfies the LDP in $\mathcal{A}_\mu$ with good convex rate function given by

$$I(\phi) = \int_0^\infty \Lambda^*(\dot{\phi})ds,$$

where $\mu = \nabla\Lambda(0)$ and

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E e^{n\theta \cdot S_n(1)}.$$

Recall now that, by the contttraction principle, if $\mathcal{Y}$ is Hausdorff and $f : \mathcal{A}_\mu \to \mathcal{Y}$ is continuous, then the sequence $f(\tilde{S}_n)$ satisfies the LDP in $\mathcal{Y}$ with good rate function

$$J(y) = \inf \left\{ \int_0^\infty \Lambda^*(\dot{\phi})ds : \ f(\phi) = y \right\}. \tag{25}$$

The basic tool used to simplify the variational problems which arise in queueing networks is Jensen's inequality. To illustrate this, consider the simplest possible example, where $d = 1$ and $f(\phi) = \phi(1)$. This mapping is certainly continuous, so we can apply the contraction principle to get that the sequence $f(\tilde{S}_n)$ satisfies the LDP in $\mathbb{R}$ with good rate function

$$J(y) = \inf \left\{ \int_0^\infty \Lambda^*(\dot{\phi})ds : \ \phi(1) = y \right\}.$$

continuous on the subspace of increasing paths in $\mathcal{A}_\mu$, for $\mu < 1$. This allows one to treat the single server queue with constant service rate. However, the mapping $\phi \mapsto \sup_t \phi(t)$ is not continuous in that topology, so the single-server queue with stochastic service rate is immediately out of reach. Another advantage with using the norm (20) is that the space $C(\mathbb{R}_+)$ of continuous functions $\phi : \mathbb{R}_+ \to \mathbb{R}^d$ with limits $\lim_{t\to\infty} \phi(t)/t$ is, when equipped with the norm (20), isomorphic to $C([0,1])$; in particular, it is Polish, which is a useful property to have in large deviation theory.

Now, for any path ϕ with $\phi(1) = y$, the path $\psi \in \mathcal{A}_\mu$ defined by $\dot{\psi} = y$ on $[0, 1)$ and $\dot{\psi} = \mu$ on $[1, \infty)$ also has $\psi(1) = y$ and, moreover,

$$\int_0^\infty \Lambda^*(\dot{\psi})ds \leq \int_0^\infty \Lambda^*(\dot{\phi})ds.$$

To see this, first note that

$$\int_1^\infty \Lambda^*(\dot{\psi})ds = 0 \leq \int_1^\infty \Lambda^*(\dot{\phi})ds.$$

On the interval $[0, 1)$ we have

$$\int_0^1 \Lambda^*(\dot{\psi})ds = \Lambda^*(y) = \Lambda^*\left(\int_0^1 \dot{\phi}ds\right) \leq \int_0^1 \Lambda^*(\dot{\phi})ds,$$

by Jensen's inequality. Thus, $J(y) = \Lambda^*(y)$. This should be compared with Theorem 1.2.

2.5 Interpreting the infimiser

What is the meaning of the path which achieves the infimum in the variational problem (25)? The answer to this is simple: it is the 'most likely' path among all paths ϕ with $f(\phi) = y$. In fact, one can show that, given $f(\tilde{S}_n) = y$, the probability that $\tilde{S}_n$ lies in any fixed neighbourhood of the infimiser tends to one as n goes to infinity. Thus, operating at the level of sample paths has the advantage of not just providing estimates of probabilities of rare events but also revealing precisely the manner in which they occur.

Consider the example of the previous section: $f(\phi) = \phi(1)$. We saw that the infimiser in this case is the path with constant gradient y on the interval $[0, 1)$ (and gradient μ thereafter). Thus, the most likely way to any point

is via a straight line and given that $S_n(1)$ takes an extreme value, with high probability it got there approximately along a staight line. This basic property, which we refer to as the 'linear geodesic property', considerably simplifies many network problems.

2.6 Examples and exercises

2.6.1 The single server queue

Now consider the single-server queue: $d = 1$ and X_k is the difference between the amount of work arriving at time k and the amount of work which can be processed at that time. Suppose also that $\mu = \Lambda'(0) < 0$. If X is stationary and ergodic with $EX_0 = \mu$ then the queue length at time zero is defined and given by

$$Q_0 = \sup_{n \geq 0} \sum_{k=0}^{n} X_{-k}, \tag{26}$$

or, equivalently, $Q_0/n = f(\tilde{S}_n)$, where $f : \mathcal{A}_\mu \to \mathbb{R}_+$ is defined by

$$f(\phi) = \sup_{t>0} \phi(t).$$

As this mapping is continuous (see Exercise 2.1) we can apply the contraction principle to get that the normalised queue length at time zero, Q_0/n, satisfies the LDP in $\mathbb{R}_+$ with good rate function

$$J(q) = \inf \left\{ \int_0^\infty \Lambda^*(\dot{\phi}) ds : \sup_{t>0} \phi(t) = q \right\}.$$

For any path $\phi \in \mathcal{A}_\mu$ with

$$\sup_{t>0} \phi(t) = q,$$

there must exist $\tau < \infty$ at which the supremum is achieved (here we are using that fact that $\lim_{t\to\infty} \phi(t)/t = \mu < 0$). The path $\psi \in \mathcal{A}_\mu$ defined by $\dot{\psi} = q/\tau$ on $(0, \tau]$ and $\dot{\psi} = \mu$ elsewhere also has $\sup_{t>0} \psi(t) = q$ and by Jensen's inequality (as above)

$$\int_0^\infty \Lambda^*(\dot{\psi})ds = \tau\Lambda^*(q/\tau) \leq \int_{-\infty}^\infty \Lambda^*(\dot{\phi})ds.$$

It follows that

$$J(q) = \inf_{\tau>0} \tau\Lambda^*(q/\tau) = \delta q,$$

where

$$\delta = \inf_{x>0} \Lambda^*(x)/x.$$

The LDP in this case states that, for any $q > 0$,

$$\lim_{q\to\infty} \frac{1}{q} \log P(Q_0 > q) = -\delta.$$

This should be compared with Theorem 1.4.

Recalling Section 2.5 we can ask: how to large queues build up? From the above we saw that the most likely path which achieves a large queue-length q is for the queue to start more or less empty and then grow at a constant rate q/τ^* for a time τ^*, where $\tau^* = \text{arginf}_{\tau>0} \tau\Lambda^*(c + q/\tau)$. (To convince yourself of this you will have to reverse the order of time.)

Exercise 2.2 *Compute τ^* as a function of c and σ when $\Lambda^*(x) = \sigma^2(x - c)^2/2$.*

2.6.2 Single-server queue with finite waiting space

Consider the following variation on the single-server queue. As before, $d = 1$ and X_k is the difference between the amount of work arriving at time k and the amount of work which can be processed at that time. However, there is a maximum allowable queue-length, which evolves as follows:

$$Q_n^b = \min\{(Q_{n-1}^b + X_k)^+, b\},$$

where $b > 0$ is the waiting space.

Equilibrium properties of such queues have been studied by Borovkov [3] and Toomey [55]. In the latter, the following characterisation is given for the equilibrium queue-length distribution, assuming it exists. Assume X is stationary and ergodic with $EX_0 = \mu = \nabla\Lambda(0) < 0$, and define, for $q \in \mathbb{R}$,

$$t_q = \inf\{t \geq 0 : \ n\tilde{S}_n(t) = q\}.$$

Note that, for $0 \leq q \leq b$, t_{q-b} is almost surely finite. It is (implicitly) shown in [55] that under these assumptions there is a stationary ergodic solution Q^b, where the queue-length at time zero is given by

$$Q_0^b = \inf\{q \in [0, b] : \ t_{q-b} < t_q\}.$$

For $a \in \mathbb{R}$ define mappings $T_a : \mathcal{A}_\mu \to [0, \infty]$ by

$$T_a(\phi) = \inf\{t \geq 0 : \ \phi(t) = a\},$$

and $f : \mathcal{A}_\mu \to [0, 1]$ by

$$f(\phi) = \inf\{y \in [0, 1] : \ T_{1-y}(\phi) < T_y(\phi)\}.$$

We can now write $Q_0^n/n = f(\tilde{S}_n)$.

Applying the contraction principle, if f is continuous we can deduce the LDP for Q_0^n/n in $[0,1]$ with good rate function given by

$$J(y) = \inf\left\{\int_0^\infty \Lambda^*(\dot{\phi})ds : \ f(\phi) = y\right\}.$$

Exercise 2.3 *Show that f is continuous, and that $J(y) = \delta y$ where $\delta = \inf_{x>0} \Lambda^*(x)/x$. Compare this with Theorem 1.4.*

This LDP was obtained in [55] using the essentially same approach. It justifies approximating the frequency of overflow in a queue with (large) finite waiting space by the frequency with which that level is exceeded in the corresponding queue with infinite waiting space.

2.6.3 Departures from a single server-queue

Consider the single-server queue of Section 2.6.1, except that now $d = 2$ and $X_k = (A_k, C_k)$, where A_k is the amount of work to arrive at time k and C_k is the amount of work that can be served. We shall assume that X is stationary and ergodic and that the sample path LDP holds with

$$\mu = \nabla\Lambda(0) = (\mu_1, \mu_2) = (EA_0, EC_0)$$

and $EA_0 < EC_0$. The departures at time k are defined by

$$D_k = A_k + Q_{k-1} - Q_k.$$

Consider the partial sums process associated with D: for $t \geq 0$ set

$$T_n(t) = \frac{1}{n}\sum_{k=0}^{[nt]} D_{-k},$$

and write $\tilde{T}_n$ for the polygonal approximation to T_n. Then

$$\tilde{T}_n(t) = \tilde{S}_n^1(t) + \sup_{s>t}[\tilde{S}^1(s) - \S^1(t) - (\tilde{S}^2(s) - \tilde{S}^2(t))] - \sup_{s>0}[\tilde{S}^1(s) - \tilde{S}^2(s)],$$

and so we can write $\tilde{T}_n = f(\tilde{S}_n)$ where $f : \mathcal{A}_\mu \to \mathcal{A}_{\mu_1}$ is defined by $f(\phi) = \psi$ where

$$\psi(t) = \phi^1(t) + \sup_{s>t}[\phi^1(s) - \phi^1(t) - (\phi^2(s) - \phi^2(t))] - \sup_{s>0}[\phi^1(s) - \phi^2(s)].$$

Exercise 2.4 *Show that f is continuous.*

We can therefore deduce an LDP for the sequence $\tilde{T}_n$ with rate function given by

$$J(\psi) = \inf\left\{\int_0^\infty \Lambda^*(\dot{\phi})ds : \ f(\phi) = \psi\right\}.$$

To solve this variational problem in general is quite hard.

Let us first concentrate on the sequence $\tilde{T}_n(1)$. By the same argument we have the LDP for this sequence with rate function

$$K(x) = \inf\left\{\int_0^\infty \Lambda^*(\dot{\phi})ds : \ [f(\phi)](1) = x\right\}.$$

If we also assume that $\Lambda^*(x, y) = \Lambda_A^*(x) + \Lambda_C^*(y)$ or, equivalently, $\Lambda(\theta_1, \theta_2) = \Lambda_A(\theta_1) + \Lambda_C(\theta_2)$, we can make some progress. This assumption is satisfied if the arrivals and service processes, A and C are independent.

Exercise 2.5 *Show that*

$$K(z) = \inf\{\delta q + \Lambda_A^*(x) + \Lambda_C^*(y) : q \geq 0, \ \min\{y, q + x\} = z\},$$

where $\delta = \sup\{\theta > 0 : \ \Lambda_A(\theta) \leq \Lambda_C(\theta)\}$.

Exercise 2.6 (Constant service rate) *If $\Lambda_C(\theta) = c\theta$, for some $c > \Lambda'_A(0)$, then $K = \Lambda^*_A$ on $[0, c]$ and $+\infty$ elsewhere.*

Exercise 2.7 *Assume that Λ_C is finite and differentiable. Show that if*

$$\Lambda_A(\theta) = \Lambda_C(\theta) - \Lambda'_C(\mu_1)(\theta - \mu_1)$$

then $K = \Lambda^*_A$. Find some examples where this property is satisfied.

The last exercise identifies an arrival process, for any given arrival rate $\mu_1 < \Lambda'_C(0)$, whose one-dimensional large deviations behaviour is left invariant by the queue with service process C. For more on this, see [28].

A natural question to ask at this point is: when is the linear geodesic property preserved? In other words, when do we have

$$J(\psi) = \int_0^\infty K(\dot{\psi})ds? \tag{27}$$

This is the subject of the paper [27]. It is shown there (see also [28]) that the rate function J is convex if, and only if, $\Lambda^*_A \leq \Lambda^*_C$ on $(-\infty, \mu_1]$. If this condition is satisfied then (27) holds.

2.6.4 Other examples

More complicated examples where this approach has been applied can be found in [45, 46, 47]. These include the first-come-first served single-server queue with multiple inputs, where the state of the system in equilibrium and the joint large deviations behaviour of the outputs are analysed, and queues

with dedicated buffers. Most of the problems we have discussed, including those studied in [45, 46, 47], along with variants and related problems, have been analysed using different methods in (for example) [1, 2, 4, 6, 7, 8, 10, 17, 18, 21, 22, 23, 24, 32, 51, 54].

2.7 Application: A problem in stochastic control

Consider the following queueing system. We have a stationary and ergodic *arrivals* process X_k, and the queue evolves as follows:

$$Q_n = \min\{(Q_{n-1} + X_n - c(Q_{n-1}))^+, B\}, \tag{28}$$

where c is some function which we are allowed to choose. The object is to keep the queue-length away from the boundaries 0 and b.

This problem arises in many contexts, from storage and inventory control to tape-speed control in backup drives. It can also be regarded as an abstraction of the file-transfer problem in communications networks (if you send the data too fast it causes congestion, if you send it too slowly you loose on throughput).

We will only consider the following control functions: for some threshold $T \in [0, B]$, $c = c_1$ on $[T, B]$ and $c = c_0$ on $[0, T)$, where $c_0 < EX_0 < c_1$. It is not hard to convince yourself that this class of control functions is in some sense an optimal subclass of all control functions which take values in the interval $[c_0, c_1]$. The problem is to determine, from among this subclass, an optimal choice of T. We will assume that the events of hitting either boundary are equally undesireable, so the object is simply to minimise the

frequency of time spent at either boundary.

If $\tilde{S}_n$ satisfies the usual LDP in $\mathcal{A}_\mu$ with $c_0 < \mu = \nabla\Lambda(0) < c_1$ and B is large, we can use large deviation theory to estimate the frequency of time spent at either boundary and hence choose a value of T which is optimal in an approximate sense.

This problem turns out to be a kind of two-sided variant of the single-server queue with finite waiting space. Define

$$t_q^1 = \inf\{t \geq 0 : \ n(\tilde{S}_n(t) - c_1 t) = q\},$$

and

$$t_q^0 = \inf\{t \geq 0 : \ n(c_0 t - \tilde{S}_n(t)) = q.$$

Now set

$$Q_0^+ = \inf\{q \in [0, B-T] : \ t_{q-B+T}^1 < t_q\},$$

and

$$Q_0^- = \inf\{q \in [0, T] : \ t_{q-T}^0 < t_q\}.$$

Finally, we define

$$Q_0 = T + Q_0^+ - Q_0^-.$$

Note that $\min\{Q_0^+, Q_0^-\} = 0$ (this follows from the definition).

Exercise 2.8 *Argue that this represents a stable equilibrium for the system defined by (28).*

Note that Q_0^+ and Q_0^- are queue-lengths in single-server queues with finite waiting space. We can thus analyse this system as in Section 2.6.2 to obtain:

Exercise 2.9 *Write Q_0^B to express the dependence on B, and let $T = aB$ for some $0 < a < 1$. Show that*

$$\lim_{B\to\infty} \frac{1}{B} \log P(Q_0^B = B) = -\delta_1(1-a)$$

and

$$\lim_{B\to\infty} \frac{1}{B} \log P(Q_0^B = 0) = -\delta_0 a,$$

where

$$\delta_1 = \inf_{x>0} \Lambda^*(x + c_1)/x = \sup\{\theta > 0 : \ \Lambda(\theta) \le c_1\theta\}$$

and

$$\delta_0 = \inf_{x>0} \Lambda^*(c_0 - x)/x = \sup\{\theta > 0 : \ \Lambda(-\theta) \le -c_0\theta\}.$$

By the principle of the largest term, the overall frequency of hitting either boundary is given approximately by

$$\exp(-\max\{\delta_0 a, \delta_1(1-a)\}B).$$

This is minimised, assuming δ_0 and δ_1 are finite, by choosing the threshold aB such that $\delta_0 a = \delta_1(1-a)$, or

$$a = \frac{\delta_1}{\delta_0 + \delta_1}.$$

2.8 Scaling properties of networks

Often the variational problems that arise in networks are too complicated to be of practical use. However, there are some elementary observations which can be made from the abstract theory which turn out to be surprisingly useful for providing heuristics, and these can really be used in practical network

management situations. Imagine a complicated network, and suppose we are interested in the contents of a buffer at a particular location. As before, we assume the inputs to the network, *and the service capacities in the network*[2], can be represented by a sequence of random variables (X_k) in $\mathbb{R}^d$ and $\tilde{S}_n$ is defined to be the corresponding partial sums process (19). We also approximate the network by assuming the buffers are infinite. Although the queue-length at this buffer is a complicated function of the inputs to network, it will generally be of the form $Q = f(n\tilde{S}_n)$. Moreover, since the queue-length is expressed in the same units as the inputs and service capacities, the function f will be *homogeneous*: $F(a\phi) = aF(\phi)$ for any $a > 0$. From this alone, assuming the mapping f is continuous, we can deduce that this queue-length has exponential tails! Indeed, the normalised queue-length, by homogeneity, is given by $Q/n = f(\tilde{S}_n)$. From this we deduce that the sequence Q/n satisfies the LDP with rate function given by

$$J(q) = \inf \left\{ \int_0^\infty \Lambda^*(\dot{\phi}) ds : \ f(\phi) = q \right\}.$$

It follows that

$$\lim_{n\to\infty} \frac{1}{n} \log P(Q > n) = -J(1),$$

and of course that $J(q) = J(1)q$.

To establish that $J(1)$ is strictly positive and finite, we need 'stability, continuity, and non-degeneracy'. By stability we mean the following: if ϕ is the path with constant gradient μ, then $f(\phi) = 0$. Continuity refers to the continuity of the mapping f. By non-degeneracy we mean there exists a

[2]In what follows, as will become clear when we introduce the homogeneity requirement below, it is important to include the service capacities in $\tilde{S}_n$, even if they are deterministic.

$\phi \in \mathcal{A}_\mu(\mathbb{R}_+)$ for which

$$\int_0^\infty \Lambda^*(\dot{\phi})ds < \infty$$

and $f(\phi) > 0$. Roughly speaking these conditions are equivalent to the requirement that there exists a non-trivial equilibrium, which is reasonable. For more on this, see [49].

2.9 Heavy traffic models

All of the above can be applied to heavy traffic models. The canonical model for a single server queue which is heavily loaded evolves as follows: for $t \in \mathbb{R}$,

$$Q_t = \sup_{s<t}[B_s - B_t - (t-s)]$$

where $(B_t,\ t \geq 0)$ and $(B_t,\ t \leq 0)$ are two independent standard Brownian motions started from 0. This arises as a diffusion approximation to the single-server queue described in Section 1, provided the sequence X is suffiently mixing (weakly dependent) with finite variance, and EX_0 is close to zero. Heavy-traffic approximation theorems are usually stated in terms of continuous-time queues, such as the $M/M/1$ queue, but they are equivalent. See [58] for a recent survey of heavy traffic queueing models. The queue-length process Q is in fact a reflected Brownian motion with negative drift. For this model, the equilibrium distribution is exactly exponential. For more complicated networks one can apply the techniques outlined in this section with the basic sample path LDP for partial sums replaced by *Schilder's theorem*, which states that if B is a standard Brownian motion started at zero,

the sequence $B/\sqrt{n}$ satisfies the LDP in $\mathcal{A}_0$ with good convex rate function

$$I(\phi) = \frac{1}{2}\int_0^\infty \dot{\phi}^2 ds. \tag{29}$$

(See, for example, [15].) See [1] and references therein for more details.

3 Long range dependence

In the early nineties, a collection of papers ([38] and references therein) published by researchers at AT&T caused quite a stir in the world of communications networking and traffic modelling. Based on a huge collection of traffic measurements taken from broadband networks, it was claimed that internet traffic exhibits long range dependence (LRD). Confusion and controversy ensued. Networking engineers, familiar with traditional Markovian queueing models (which do not exhibit LRD), were worried because the implications of this finding were unclear. The controversy arose naturally because of deep philosophical difficulties associated with fitting models to data which exhibits long range dependence. It was soon realised that this was not a new dilemma. For example, a similar controversy arose in the hydrology literature some twenty years earlier (see, for example, [37]).

In this section we explain the notion of LRD, its implications for networks, how it can arise and the philosophical issues associated with fitting LRD models to data.

3.1 What is long range dependence?

Let X_n be a stationary sequence of random variables, which we assume to be bounded for simplicity, and set $S_n = X_1 + \cdots X_n$. If the X_n's are independent, then $\mathrm{var} S_n = n \mathrm{var} X_1$. In particular, the variance of S_n grows linearly with n, a property which is not specific to the iid case: this property holds quite generally for Markov chains and other weakly dependent sequences.

There is no standard definition for long range dependence. Rather, it is a loosely used term to refer to the case where the variance grows non-linearly. The most common LRD models used for teletraffic have $\mathrm{var}S_n \approx ct^{2H}$, where $1/2 < H < 1$ (the *Hurst parameter*).

LRD sequences typically have fluctuations at every time scale.

In the case of unbounded variables, where variances may not even exist, one has to be a little more careful in defining LRD.

3.2 Implications for networks

In the large deviations framework of the last chapter, the existence of a sufficiently smooth limiting cumulant generating function

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log E e^{\lambda S_n},$$

implies (by Taylor expansion) that $\mathrm{var}S_n \simeq \Lambda''(0)n$. Thus, everything we have presented so far does not apply to LRD models.

There is an analogue of Theorem 1.2 (actually it follows from that theorem) in the case where the limit

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{a(n)} \log E e^{\theta a(n) S_n/n}$$

exists and is well-behaved, for sequences $a(n) \to \infty$. In this case the log-probabilities $\log P(S_n/n \in F)$ are normalised by $a(n)$. If the variance is growing non-linearly, we can expect (from the Taylor expansion again) to have

$$a(n) = n^2/\mathrm{var}S_n.$$

This new framework can be used to prove an analogue of Theorem 1.4 (see, for example, [20]). The bottom line, in the case

$$\mathrm{var} S_n \sim ct^{2H},$$

is that the tails of the queue-length distribution do not decay exponentially if $H > 1/2$. Instead, we have

$$\log P(Q_0 > q) \sim -\delta q^{\gamma},$$

where $0 < \gamma = 2(1-H) < 1$ and $\delta = \inf_{\tau>0} \tau^{\gamma} \Lambda^*(1/\theta)$. It is possible to state and prove this result using techniques similar to those presented earlier. (You may wish to regard this as a challenging exercise!) Instead, we will focus on the most popular and well-known LRD queueing model, which is based on fractional Brownian motion.

This is a continuous time model which has been widely adopted for its parsimonious structure, as it depends on just three parameters: mean, variance and Hurst parameter. The Hurst parameter reflects the degree of LRD.

A *Gaussian process* is any stochastic process whose finite-dimensional distributions are all multivariate normal.

Standard fBM can be characterised as the centered (zero mean) Gaussian process $(W_t,\ t \in \mathbb{R})$ with $W_0 = 0$, stationary increments, continuous paths and $EW_t^2 = |t|^{2H}$, for some $1/2 \leq H < 1$. The case $H = 1/2$ corresponds to standard Brownian motion.

In the corresponding queueing model, introduced by Norros [44], the queue-

length at time t is given by

$$Q_t = \sup_{s<t}[\sigma(W_s - W_t) - \mu(t-s)],$$

where σ and μ are strictly positive. This is a stationary ergodic process with equilibrium distribution characterised by

$$Q = \sup_{t>0}[\sigma W_t - \mu t].$$

For $H > 1/2$, the tails of the queue-length distribution for this model are not exponential. In general, we have:

Theorem 3.1

$$\lim_{q\to\infty} \frac{1}{q^\gamma} \log P(Q > q) = -\kappa^2/2\sigma^2$$

where $\gamma = 2(1-H)$ *and*

$$\kappa = \frac{\mu^H}{H^H(1-H)^{(1-H)}}.$$

Proof — We follow the proof given in [42]. By scaling we can set $\sigma = 1$. *Lower bound.* Set

$$\bar{\varphi}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-x^2/2},$$

and recall that W_t is Gaussian with mean zero and variance t^{2H}.

$$\begin{aligned} P(Q > q) &= P(\sup_{t>0}[W_t - \mu t] > q) \\ &= P\bigcup_{t>0}\{W_t - \mu t > q\} \\ &\geq \sup_{t>0} P(W_t - \mu t > q) \\ &= \sup_{t>0} \bar{\varphi}\left(\frac{q+\mu t}{t^H}\right) \\ &= \bar{\varphi}(q^{1-H}\kappa). \end{aligned}$$

At the last step we minimise the argument in $\bar{\varphi}$. The lower bound now follows from the fact that $\log \bar{\varphi}(x) \sim -x^2/2$ as $x \to \infty$.

Upper bound. To prove the upper bound we use *Borell's inequality* : if X_t is a Gaussian process (on any index set) and $v^2 = \sup_t \mathrm{var} X_t < \infty$, then $m = E \sup_t X_t < \infty$ and

$$P(\sup_t X_t > x) \leq 2 \exp\left(-\frac{x-m}{2v^2}\right),$$

for all $x > m$. Observe that

$$\begin{aligned} P(Q > q) &= P(\sup_{t>0}[W_t - \mu t] > q) \\ &= P \bigcup_{t>0} \{W_t - \mu t > q\} \\ &= P \bigcup_{t>0} \left\{ \frac{W_t}{q+\mu t} > 1 \right\} \\ &= P\left(\sup_{t>0} \frac{W_t}{q+\mu t} > 1 \right) \end{aligned}$$

Why did we rewrite the probability in this way? Because now we can apply Borell's inequality: the variance of $W_t - \mu t$ is unbounded, whereas

$$v^2 = \sup_t \mathrm{var} \frac{W_t}{q+\mu t} = \sup_t \frac{t^{2H}}{(q+\mu t)^2} = q^{-\gamma} \kappa^{-2}.$$

Moreover,

$$|m| = \left| E \sup_t \frac{W_t}{q+\mu t} \right| \leq E \sup_t \frac{|W_t|}{q+\mu t}$$

which goes to zero as $q \to \infty$ by monotone convergence. Applying Borell's inequality, for q sufficiently large,

$$P(Q > q) \leq 2 \exp\left(-\frac{1-m}{2v^2}\right) = 2 \exp(-(1-m)\kappa^2 q^\gamma/2),$$

and the result follows. □

The case $H = 1/2$ coresponds to the heavy traffic model discussed in Section 2.9; in this case, as expected, the tails do indeed decay exponentially.

3.2.1 Sample path large deviations for fBM

The statement of Theorem 3.1 tells us how the tails of the queue-length distribution decay, but it tells us nothing about *how large queue-lengths occur*. What is the most likely path? Is is linear?

There is an analogue of Schilder's theorem for fBM (see, for example, [15]) and the rate function is not of the form (29). In particular, for fBM, the most likely path between two points is not linear.

This sample path LDP can be used to compute most likely paths. An alternative approach is presented in [48], which uses a representation of fBM as a stochastic integral against standard Brownian motion, Schilders theorem and the contraction principle. See also [9].

There is an even easier way to compute geodesics, which was pointed out to me by Peter Glynn (personal communication). We use the following fact: if (X, Y) has a bivariate normal distribution with $EX = EY = 0$, $EX^2 = EY^2 = 1$ and $\mathrm{cov}(X, Y) = \rho$, then the conditional law of X given $Y = y$ is normal with mean ρy and variance $1 - \rho^2$.

Exercise 3.1 *Use this fact to show that, for* $\alpha \in [0, 1]$,

$$g(\alpha, x) := \lim_{t \to \infty} E(W_{\alpha t}/t |\ W_t/t = x) = [\alpha^{2H} + 1 - (1 - \alpha)^{2H}]x/2,$$

and

$$\lim_{t\to\infty} \sup_{\alpha} \text{var}\ (W_{\alpha t}/t|\ W_t/t = x) = 0.$$

Deduce, using Borell's inequality, that

$$\lim_{t\to\infty} P(\sup_{\alpha} |W_{\alpha t}/t - g(\alpha, x)| > \epsilon|\ W_t/t = x) = 0.$$

(You may assume that $\{(W_{\alpha t}/t|W_t/t = x),\ 0 \leq \alpha \leq 1\}$ *is a Gaussian process.) Use this to describe how large queues build up in the fBM queueing model.*

3.3 How does long range dependence arise in natural systems?

A concrete example from everyday life where LRD arises naturally is traffic patterns on country roads. Local interactions (cars cannot overtake each other) can give rise to long-range interactions (huge backlogs followed by long stretches without any cars at all).

Another example from everyday life is a magnet. Microscopic local interaction between molecules can lead to macroscopic organisation: this is LRD. In statistical physics, magnets are modelled as a Markov random field (higher dimensional analogue of a Markov chain) and it can be shown that, if the local interaction (dependence) is strong enough, the system will exhibit long range dependence. This does not occur in one dimension (recall Exercise 1.5).

In the context of teletraffic, the following observation is relevant. The aggregation of many independent traffic sources, each with heavy-tailed interarrival times, can be approximated by fractional Brownian motion with Hurst

parameter $H > 1/2$ (see [33] and references therein). To make this more precise, supppose we have N sources, each an independent copy of stationary renewal process with inter-renewal time distribution F, and F lies in the domain of attraction of a stable law with index $1+\beta$, $0 < \beta < 1$, that is

$$1 - F(t) \sim t^{-(1+\beta)} L(t)$$

where L is slowly varying at infinity:

$$\lim_{t\to\infty} L(xt)/L(t) = 1,$$

for any $x > 0$. Note that F has a mean μ but no variance. Denote the number of arrivals in the time interval $[0, t)$ by $A^N(t)$. A scaling sequence a_N can be chosen such that $a_N^\beta \sim L(a_N)m$. In [33] it is shown that the the sequence of processes Y^N, defined by $Y^N(s) = A^N(a_N s)/a_N - Ns/\mu$ for $s > 0$, converges in law to centered fractional Brownian motion with variance parameter

$$\sigma^2 = \frac{2}{\mu^3(1-\beta)(2-\beta)},$$

and Hurst parameter $H = 1 - \beta/2$.

3.4 Philosophical difficulties with LRD modelling

First let us suppose that we have observed a high *empirical* value for the Hurst paremeter associated with a particular time series. There are various schemes for estimating Hurst parameters, but whichever one has been adopted, a large empirical Hurst parameter indicates that there is a fluctuation at the time-scale over which the data is observed. To fit a LRD model to this data is to

regard this fluctuation as random. The alternative is to regard the data as non-stationary.

Without any further information about the data and where it came from, the fact that there is a fluctuation at the time-scale over which the data is observed makes prediction beyond the short term[3] a difficult task; to hope to say something useful about future fluctuations at the same time-scale is somewhat optimistic. It is a sample-size problem: with one sample (of fluctuations at this time scale) we don't have very much information (about fluctuations at this time scale).

In general, it is impossible to distinguish between LRD and non-stationarity. If there is a fluctuation at the time-scale over which the data is observed, then either proposition is consistent with the data. For all practical purposes they are equivalent. See [37] for an excellent discussion on this point.

Having said that, there is a fundamental difference at the philosophical level, similar in nature to the difference between frequentist and Bayesian statistics. I claim that, in this context, to take the LRD view and regard the single, unexplained fluctuation as random, is to be Bayesian. The alternative viewpoint is essentially frequentist.

Suppose, on the other hand, we have some reason to believe that the data is, in a truly statistical sense, long-range dependent in nature. For example, suppose we know that the data is an aggregate of many independent sources with heavy-tailed inter-arrival times, as discussed in the previous

[3]Ironically, as one of the students pointed out, short term prediction is often considerably easier with such data sets because of the presence of 'trends'.

section. Then things are somewhat different. Modelling and prediction will be difficult, but no more difficult that modelling heavy-tailed distributions, and as such one can hope to have some success. Robustness is now the key issue. Domains of attraction of heavy-tailed stable laws (or as we saw in the last section, fractional Brownian motion) are, in a sense which is difficult to formulate precisely but which is nevertheless meaningful, much smaller than the domain of attraction of the usual central limit theorem (and standard Brownian motion), and for that reason predictions based on the former are in practise more prone to error.

Note that these issues are a function of the data, not the approach. If a time series appears non-stationary, exhibits LRD or heavy tails, there will be difficulties with prediction no matter what approach is adopted. There are many instances in practise where it preferable to try something, even if confidence is limited, rather than throw our arms in the air and say "this is impossible!".

References

[1] Florin Avram, J.G. Dai, John J. Hasenbein (1999). Explicit solutions for variational problems in the quadrant. Preprint.

[2] Dimitris Bertsimas, Ioannis Ch. Paschalidis and John N. Tsitsiklis (1998). On the large deviations behaviour of acyclic networks of G/G/1 queues. *Ann. Appl. Prob.* 8, 1027–1069.

[3] A.A. Borovkov. *Asymptotic methods in queueing theory.* Wiley, 1984.

[4] A.A. Borovkov and A.A. Mogulskii (1995). Large deviations for stationary Markov chains in a quarter plane. In: *Probability Theory and Mathematical Statistics (Tokyo, 1995)*, 12–19. World Scientific Publishing, 1996.

[5] D.D. Botvich and N.G. Duffield (1995). Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems* 20, 293–320.

[6] Cheng-Shang Chang (1994). Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. on Automatic Control* 39:913–931.

[7] Cheng-Shang Chang. Approximations of ATM networks: effective bandwidths and traffic descriptors. ???

[8] Cheng-Shang Chang, Philip Heidelberger, Sandeep Juneja and Perwez Shahabuddin (1994). Effective Bandwidth and Fast Simulation of ATM Intree Networks. *Performance Evaluation* 20:45–66.

[9] C.-S. Chang, D.D. Yao, T. Zajic, Large Deviations, Moderate Deviations, and Queues with Long-Range Dependent Input. To appear in *J. Appl. Prob.*

[10] C.-S. Chang and T. Zajic (1995). Effective bandwidths of departure processes from queues with time varying capacities. *Infocom.*

[11] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand and R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. *IEEE Trans. Comm.* ???

[12] C. Courcoubetis and R. Weber (1996). Buffer overflow asymptotics for a buffer handling many traffic sources. *J. Appl. Prob.* 33, 886–903.

[13] Amir Dembo and Tim Zajic (1995). Large deviations: from empirical mean and measure to partial sums process. *Stoch. Proc. Appl.* 57:191–224.

[14] Amir Dembo and Ofer Zeitouni (1992). *Large Deviations Techniques and Applications.* Jones and Bartlett, London.

[15] Jean-Dominique Deuschel and Daniel W. Stroock. *Large deviations.* Academic Press, 1989.

[16] G. de Veciana, C. Courcoubetis and J. Walrand (1993). Decoupling bandwidths for networks: a decomposition approach to resource management. *Infocom*, 1994.

[17] G. de Veciana and G. Kesidis (1993). Bandwidth allocation for multiple qualities of service using generalised processor sharing. Preprint.

[18] R.L. Dobrushin and E.A. Pechersky (1995). Large deviations for random processes with independent increments on infinite intervals. Preprint.

[19] N.G. Duffield, J.T. Lewis, Neil O'Connell, Raymond Russell and Fergal Toomey (1995). Entropy of ATM traffic streams: a tool for estimating QoS parameters. *IEEE Journal of Selected Areas in Communications* 13(6):981–990.

[20] N.G. Duffield and Neil O'Connell (1995). Large deviations and overflow probabilities for the general single server queue, with applications. *Proc. Camb. Phil. Soc.* 118(1).

[21] N.G. Duffield and Neil O'Connell (1994). Large deviations for arrivals, departures, and overflow in some queues of interacting traffic. *Proceedings of the 11^{th} IEE Teletraffic Symposium,* Cambridge, March 1994.

[22] Paul Dupuis and Richard S. Ellis (1996). The large deviation principle for a general class of queueing systems, I. *Trans. Amer. Math. Soc.* 347, 2689–2751.

[23] Paul Dupuis and Richard S. Ellis (1994). Large deviation analysis of queueing systems. Proceedings of the IMA workshop "Stochastic Networks, February 28 - March 4, 1994", F. Kelly and R. Williams, eds. Springer-Verlag.

[24] Paul Dupuis and Kavita Ramanan (1998). A Skorohod problem formulation and large deviation analysis of a processor sharing model. *Queueing Systems* 28, 109–124.

[25] Richard S. Ellis. *Large Deviations and Statistical Mechanics.* Springer ???

[26] A. Ganesh and Neil O'Connell (1998). The linear geodesic property is not generally preserved by a FIFO queue. *Ann. Appl. Prob.* 8(1), 98–111.

[27] A. Ganesh and Neil O'Connell (1999). A large deviation principle with queueing appllications. *Stochastics and Stochastic Reports*, to appear.

[28] A. Ganesh, Neil O'Connell and Balaji Prabhakar (1999). Large deviations and fixed points for discrete-time queues. In preparation.

[29] R.J. Gibbens (1996). Traffic characterisation and effective bandwidths for broadband network traces. In: *Stochastic Networks*, ed. by Frank Kelly and Ilze Zeidens. Oxford University Press.

[30] Peter W. Glynn and Ward Whitt (1995). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.* ???

[31] M. Harrison (1985). *Brownian Motion and Stochastic Flow Systems.* Wiley.

[32] I.A. Ignatyuk, V. Malyshev and V.V. Scherbakov (1994). The boundary influence in the problems of large deviations. *Uspehi Matematicheskih Nauk* 49, 43–102.

[33] Ingemar Kaj (1999). Convergence of scaled renewal processes to fractional Brownian motion. Preprint.

[34] F. P. Kelly (1991). Effective bandwidths at multi-class queues. *Queueing Systems* 9, 5–15.

[35] F.P. Kelly (1996). Notes on effective bandwidths. In: *Stochastic Networks*, ed. by Frank Kelly and Ilze Zeidens. Oxford University Press.

[36] F.P. Kelly and P.B. Key (1994). Dimensioning playout buffers from an ATM network. *Proceedings of the 11th IEE Teletraffic Symposium,* Cambridge, March 1994.

[37] V. Klemes (1974). The Hurst phenomenum: a puzzle? *Water Resour. Res.* 10(4), 675–688.

[38] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson (1994). On the Self-Similar Nature of Ethernet Traffic (extended version). *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, pp. 1-15.

[39] J.T. Lewis, C. Pfister and W. Sullivan. Thermodynamic aspects of probability. *Markov Proc. Rel. Fields* 1. ???

[40] Loynes. ???

[41] B. Mandelbrot and J. Van Ness (1968). Fractional Brownian Motions, Fractional Gaussian Noises and Applications. *SIAM Review*, Vol. 10, No. 4, pp. 422-437.

[42] Laurent Massoulie and Alain Simonian. Large buffer asymptotics for the queue with fBM input. Preprint.

[43] A.A. Mogulskii. Large deviations for trajectories of multi-dimensional random walks. *Th. Prob. Appl.* 21, 300-315, 1976.

[44] I. Norros (1995). On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks. *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 6, pp. 953-962.

[45] Neil O'Connell (1997). Large deviations for departures from a shared buffer. *J. Appl. Prob.* 34, 753–766.

[46] Neil O'Connell (1998). Large deviations for queue lengths at a multi-buffered resource. *J. Appl. Prob.* 35, 240–245.

[47] Neil O'Connell (1996). Queue-lengths and departures at single-server resources. In: *Stochastic Networks*, ed. by Frank Kelly and Ilze Zeidens. Oxford University Press, 1996.

[48] Neil O'Connell and Gregorio Procissi. On the build-up of large queues in a queue with fBM input. BRIMS Technical Report HPL-BRIMS-97??

[49] Neil O'Connell and Fergal Toomey. Scaling and continuity in queueing networks. In preparation.

[50] Shyam Parekh and Jean Walrand (1989). A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. Aut. Contr.* 34:54–66, 1989.

[51] Kavita Ramanan and Paul Dupuis. Large deviation properties of data streams that share a buffer. *Ann. Appl. Prob.*

[52] Raymond Russell (1998). *The Large Deviations of Random Time-Changes*. PhD thesis, Trinity College Dublin.

[53] G. Samorodnitsky and M. S. Taqqu (1994). *Stable Non-Gaussian Random Processes.* Chapman and Hall.

[54] Adam Shwartz and Alan Weiss (1995). *Large Deviations for Performance Analysis.* Chapman and Hall.

[55] Fergal Toomey (1998). Bursty traffic and finite capacity queues. *Ann. Oper. Res.* 79, 45–62.

[56] S.R.S. Varadhan (1966). Asymptotic probabilities and differential equations. *Comm. Pure Appl. Math.* 19, 261–286.

[57] A. Weiss (1986). A new technique for analysing large traffic systems. *Adv. Appl. Prob.* 18, 506–532.

[58] R.J. Williams (1996). On the approximation of queueing networks in heavy traffic. In: *Stochastic Networks*, ed. by Frank Kelly and Ilze Zeidens. Oxford University Press, 1996.

[59] Damon Wischik (1999). *Large Deviations and Internet Congestion.* PhD thesis, University of Cambridge.