Lecture Notes in Mathematics 1838

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Ayalvadi Ganesh
Neil O'Connell
Damon Wischik

Big Queues

Springer

Authors

Ayalvadi Ganesh

Microsoft Research
7 J.J. Thomson Avenue
Cambridge CB3 0FB, UK

e-mail: ajg@microsoft.com

Neil O'Connell

Mathematics Institute
University of Warwick
Coventry CV4 7AL, UK

e-mail: noc@maths.warwick.ac.uk

Damon Wischik

Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WB, UK

e-mail: D.J.Wischik@statslab.cam.ac.uk

Cataloging-in-Publication Data available

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at http://dnb.ddb.de

Mathematics Subject Classification (2000):
60K25, 60K30, 90B15, 90B18, 90B20, 90B22, 60F10

ISSN 0075-8434
ISBN 3-540-20912-3 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag is a part of Springer Science+Business Media
springeronline.com

Typesetting: Camera-ready TEX output by the authors

SPIN: 10983423 41/3142/DU - 543210 - Printed on acid-free paper

To our parents:
Jagannathan and Lalita,
Michael and Kath O'Connell,
Claude and Irene Wischik.

Preface

Aims and scope

Big Queues aims to give a simple and elegant account of how large deviations theory can be applied to queueing problems. Large deviations theory is a collection of powerful results and general techniques for studying rare events, and has been applied to queueing problems in a variety of ways.

The strengths of large deviations theory are these: it is powerful enough that one can answer many questions which are hard to answer otherwise, and it is general enough that one can draw broad conclusions without relying on special case calculations. This latter strength has been hidden by the rather piecemeal development of the subject so far, and we feel it is now time for an account which shows that (in this case at least) abstraction can serve to simplify rather than to obscure.

We are not aiming to write an encyclopaedia on the subject, nor is this an attempt to survey the vast literature (including books by Shwartz and Weiss [91] and Chang [13]) which has evolved on this and related topics. Instead we present a certain point of view regarding the application of large deviations theory to queueing problems. Specifically, we will use the 'continuous mapping' approach, which has several benefits.

First, it suggests a style of simple heuristic argument which is easy to make rigorous.

Second, by basing our results on one key concept, the presentation is made much simpler. The continuous mapping approach lets us use exactly the same framework to describe three important scaling regimes: the large buffer regime; the regime for describing long-range dependence, which has attracted a good deal of attention in Internet traffic modelling; and the many-flows regime, which often gives better numerical approximations.

Third, this approach allows us to make very general statements about how various quantities of interest scale as the system scales, without needing to make any explicit calculations. In designing networks, it is commonly more important to understand such scaling behaviour than it is to obtain explicit answers. With the help of the continuous mapping approach, we aim to give an elementary introduction to rare-event scaling phenomena in queueing theory.

Intended readership

Big Queues targets graduate students in probability and mathematically-inclined graduate students in engineering, especially those interested in applications to communications networks. Much of the material is drawn from lecture courses given by the authors at Uppsala, Cambridge and Bangalore.

The introductory chapters and Chapter 10 on heuristics might also be of interest to the wider network-engineering research community.

Online material

The website for this book is `www.bigqueues.com`. It contains corrections, as well as an 'active bibliography' containing links to online versions of the papers cited (where available) and references to more recent articles.

Acknowledgements

Some of the material evolved from a lecture course given by N.O'C. at Uppsala University in November 1999. Acknowledgments are due to the organisers, Ingemar Kaj and Tobias Ryden, to fellow-instructors Raymond Russell and Fergal Toomey, and to the students.

Some of the material draws on a lecture course given by Stephen Turner at Cambridge. Thanks to him for making his notes available to D.J.W.

The presentation benefited from the helpful comments of Professor Anurag Kumar who invited A.J.G. to give a course on this material at the Indian Institute of Science at Bangalore.

Thanks to proof readers, including Devavrat Shah, Xiaofei Ji, Silke Meiner, and Chris Purcell.

Finally we would like to thank Venkat Anantharam, John Lewis and Frank Kelly for introducing us to large deviations theory.

Ayalvadi Ganesh,
Neil O'Connell,
Damon Wischik.

Cambridge, November 2003.

Contents

Chapter 1

The Single Server Queue

The study of queueing models is an appealing part of applied mathematics because queues are familiar and intuitive—we face queues nearly every day—and because they can be used to model many different systems.

The simplest queue is a line of customers, in which the customer at the head of the line receives service from a single server and then departs, and arriving customers join the tail of the line. Given the interarrival times and service requirements, we may wish to know how often the server is idle, what the average waiting time is, how often the number in the queue exceeds some level, and so on.

Queues can also be used to model problems in insurance. Suppose an insurance broker starts with a certain amount of capital. Every day a certain amount of money is paid out in claims (the 'service' that day), and a certain amount of money is paid in in premiums (the 'arrivals' that day), and the capital at the end of the day is the starting capital plus arrivals minus service. We may wish to know how likely it is that there is insufficient capital to meet the claims on a given day.

Another application is to packet-based data networks. Data is parcelled up into packets and these are sent over wires. At points where several wires meet, incoming packets are queued up, inspected, and sent out over the appropriate wire. When the total number of packets in the queue (the 'amount of work' in the queue) reaches a certain threshold (the 'buffer size'), incoming packets are discarded. We may wish to know the frequency of packet discard, to know how large to make the buffer.

There are many more applications, and many extensions—multiple servers, different service disciplines, networks of queues, etc. etc.

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 1–21, 2004.

Consider now the recursion

$$Q_t = (Q_{t-1} + A_t - C_t)^+,$$

where $t \in \mathbb{N}$ (or $\mathbb{Z}$) and Q_t, A_t and $C_t \in \mathbb{R}^+$, and x^+ denotes the positive part of x, i.e. $\max(x, 0)$. This is known as Lindley's recursion. It can be used to describe customers waiting in a line. Interpret Q_t as the time that the $(t+1)$th customer spends waiting before his service starts, A_t as the service time of the tth customer, and C_t as the interarrival time between customers t and $t+1$.

It can also be used to describe the insurance model. Interpret Q_{t-1} as the amount of capital at the start of day t, and A_t as the arrivals and C_t as the service that day.

For the packet data model, consider the modified recursion

$$Q_t = \big[Q_{t-1} + A_t - C_t\big]_0^B$$

where $[x]_0^B = \max(\min(x, B), 0)$. Interpret Q_t as the amount of work in the queue just after time $t \in \mathbb{Z}$, A_t as the number of packets that arrive in the interval $(t-1, t)$, C_t as the number of packets served at time t, and B as the buffer size.

For these simple models, the goal of queueing theory is to understand the qualitative behaviour of the queueing system, when the input sequence A and the service sequence C are random.

If they are both sequences of i.i.d. random variables, then Q_t is a random walk constrained to stay positive, and one can obtain certain results using the theory of random walks and renewal processes. If in addition either A or C is a sequence of exponential random variables, one can obtain further results by considering certain embedded Markov chains. In the latter setting, even for more complicated queueing models, there is beautiful mathematical theory which has been remarkably successful as a basis for many applications. See, for example, the introductory texts [3, 11, 49, 52].

However, in recent years, there has been increasing demand for a theory which is tractable and yet allows one to consider input and service sequences which exhibit highly non-Markovian characteristics. This is especially important for modelling internet traffic. In general, this is a tall order—many of the important structures of classical queueing theory break down completely—but not so tall as it may seem if one restricts one's attention to rare events.

For example, in the packet data example, we may want to make the buffer size sufficiently large that packet discard is a rare event. To quantify how large the buffer size needs to be, we need to estimate the probability of the rare event that the buffer overflows and packets are discarded.

The basic tool for studying rare events is large deviations theory. In this book we will describe one approach to large deviations theory for queues. The strength of the theory (and particularly of this approach) is that one can draw broad conclusions, for systems which are otherwise hard to analyse, without relying on special-case calculations.

In the remainder of this chapter we focus on the simplest single-server queueing model and describe how one can apply some elementary large deviations theory in this context.

1.1 The Single-Server Queueing Model

Consider Lindley's recursion

$$Q_t = (Q_{t-1} + A_t - C_t)^+, \tag{1.1}$$

where $t \in \mathbb{N}$ (or $\mathbb{Z}$), Q_t, A_t and $C_t \in \mathbb{R}^+$, and x^+ denotes the positive part of x.

> *Note.* Throughout this book we will adopt the interpretation that Q_t is the amount of work in a queue just after time t, A_t is the amount of work that arrives in $(t-1,t)$, and C_t is the amount of work that the server can process at time t.
>
> As we have noted, the recursion can also be interpreted as describing customers waiting in a line. Most of our results can be interpreted in this context.

It is of course unnecessary to keep track of both A_t and C_t. We could just define $X_t = A_t - C_t$, $A_t \in \mathbb{R}$, and look at the recursion $Q_{t+1} = (Q_t + X_t)^+$. Nonetheless, we shall (for the moment) persist in keeping separate account of service, because it is helpful in building intuition. So we will deal with the recursion

$$Q_t = (Q_{t-1} + A_t - C)^+, \tag{1.2}$$

where C is a fixed constant, and allow $A_t \in \mathbb{R}$.

This recursion may have many solutions. One way to get around this is to impose boundary conditions. For example, suppose we are interested in Q_0. If we impose the boundary condition $Q_{-T} = 0$, for some $T > 0$, then the recursion specifies a unique value for Q_0—call it Q_0^{-T} to emphasize the rôle of the boundary condition. Now Q_0^{-T} has a simpler form:

Lemma 1.1 *Let S_t, $t \geq 1$, be the cumulative arrival process: $S_t = A_{-t+1} + \cdots + A_0$. By convention, let $S_0 = 0$. Then*

$$Q_0^{-T} = \max_{0 \leq s \leq T} S_s - Cs$$

To prove this, simply apply Lindley's recursion T times, to Q_0 then to Q_{-1} and so on to Q_{-T+1}.

One particularly important solution to Lindley's recursion can be obtained by letting $T \to \infty$. The above lemma implies that Q_0^{-T} is increasing in T, which means that the limit

$$Q_0^{-\infty} = \lim_{T \to \infty} Q_0^{-T}$$

exists (though it may be infinite). The lemma also gives a convenient form:

$$Q_0^{-\infty} = \sup_{s \geq 0} S_s - Cs.$$

Of course, there is nothing special about time 0, so we can just as well define

$$Q_{-t}^{-\infty} = \sup_{s \geq t} S[t, s) - C(s - t) \tag{1.3}$$

where $S[t, s) = A_{-t} + \cdots + A_{-s+1}$ and $S[t, t) = 0$. Think of $Q_{-t}^{-\infty}$ intuitively as the queue size at time $-t$, subject to the boundary condition that the queue was empty at time $-\infty$.

This boundary condition is so useful that from now on we will drop the superscript and write Q_{-t} for $Q_{-t}^{-\infty}$, where the context is clear.

If the arrival process is stationary, i.e. if $(A_{-t}, \ldots, A_0)$ has the same distribution as $(A_{-t-u}, \ldots, A_{-u})$ for every t and u, then Q_0 has the same distribution as Q_{-t} for every t, and this distribution is called the *steady state* distribution of queue size.

> *Note.* Why is this boundary condition interesting? Exercise 1.2 shows that if we impose the boundary condition $Q_{-T} = r$ and let $T \to \infty$ we get the same answer, for any r, as long as the mean arrival rate is less than the service rate.
>
> This construction was used by Loynes [60]. He showed that if $(A_t,\ t \in \mathbb{Z})$ is a stationary ergodic sequence of random variables with $EA_0 < C$, then for any initial condition Q_0 the sequence Q_t, as defined by the recursion (1.2), converges in distribution as $t \to \infty$ to a limit which does not depend on Q_0. (It is easy to see that $Q_0^{-\infty}$ has this distribution.) Moreover, the sequence $(Q_t^{-\infty},\ t \in \mathbb{Z})$ defines a stationary ergodic solution to (1.2)

Exercise 1.1
Show that (1.3) satisfies (1.2). ◇

Exercise 1.2
Let $R_0^{-T}(r)$ be the queue size at time 0, subject to the boundary condition that $Q_{-T} = r$. Show that

$$R_0^{-T}(r) = \max_{0\leq s\leq T}\Big[S_s - Cs\Big] \vee (r + S_T - CT).$$

Deduce that, if $S_t/t \to \mu$ almost surely as $t \to \infty$ for some $\mu < C$, then almost surely

$$\lim_{T\to\infty} R_0^{-T}(r) = Q_0^{-\infty} \quad \text{for all } r.$$

This shows that we could just as well take any value for the 'queue size at time $-\infty$'—it makes no difference to the queue size at time 0. ◇

A nice example to keep in mind is the following, a discrete-time analog of the $M/M/1$ queue.

Example 1.3
Let $C = 1$ and let the A_t be independent and identically distributed: $A_t = 2$ with probability p and $A_t = 0$ with probability $1-p$, $p < 1/2$. Fix Q_0. Then the process $(Q_t, t \geq 0)$ defined by Lindley's recursion is a birth-and-death Markov chain, and it is easy to work out the distribution of the equilibrium queue length Q: for $q \in \mathbb{N}$,

$$P(Q \geq q) = \Big(\frac{p}{1-p}\Big)^q. \tag{1.4}$$

The distribution of the Markov chain converges to this equilibrium distribution, whatever the value of Q_0. Thus, the distribution of Q_0^{-T} converges to it also as $T \to \infty$. So the distribution of Q_0 (i.e. of $Q_0^{-\infty}$) is the equilibrium distribution of queue size.

We will rewrite (1.4) as

$$\log P(Q_0 \geq q) = -\delta q \tag{1.5}$$

where $\delta = \log\big((1-p)/p\big)$. ◇

It is a remarkable fact that an approximate version of (1.5) holds quite generally: for some $\delta > 0$,

$$\log P(Q_0 \geq q) \sim -\delta q \quad \text{for large } q. \tag{1.6}$$

We will frequently refer to the event $\{Q_0 \geq q\}$ by saying that *'the queue size at time* 0 *overflows a buffer level q'*; then the statement (1.6) is that the probability of overflow decays exponentially. The rest of this book is about making such statements precise.

> *Note.* So far we have assumed that the queue size can grow arbitrarily large. Similar results also apply when the queue size cannot grow beyond a maximum value, known as the *buffer size*, as suggested by the following example.
>
> *Exercise 1.4*
> Suppose the queue has a finite buffer of size B, and we use the modified version of Lindley's equation
>
> $$Q_t = \big[Q_{t-1} + A_t - C\big]_0^B$$
>
> where $[x]_0^B = \max(\min(x, B), 0)$. Find the equilibrium distribution of queue size for the Markov model of Example 1.3.
>
> It is now possible that incoming work is discarded because the buffer is full. In this model, if $Q_{t-1} = B$ and $Q_t = B$ then one unit of work was dropped at time t. Let the steady-state probability of this event be $p(B)$. Calculate $p(B)$, and show that
>
> $$\lim_{B\to\infty} \frac{1}{B} \log p(B) = -\delta$$
>
> where again $\delta = \log\big((1-p)/p\big)$. ◇

Before we go on to make (1.6) precise, let us consider one application. If the approximation 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. We have qualified this statement because actually this is a very challenging statistical problem. Nevertheless, this ingenious idea, which was first proposed in [19], has inspired major new developments in the application of large deviation theory to queueing networks.

We will make (1.6) precise using large deviations theory. In this chapter we will give an explicit proof in a simple setting, and in later chapters we will draw on more powerful large deviations techniques to prove more general results. First, we need to introduce some basic large deviations theory.

1.2 One-Dimensional Large Deviations

Let X be a random variable, and let $(X_n, n \in \mathbb{N})$ be a sequence of independent, identically distributed random variables, each with the same distribution as X, and let $S_n = X_1 + \cdots + X_n$. If EX is finite, then the strong law of large numbers says that

$$\frac{S_n}{n} \to EX \quad \text{almost surely}$$

as $n \to \infty$.

What about fluctuations of S_n/n around EX? If X has finite variance, then the central limit theorem says that the sequence of random variables

$$\sqrt{n}\Big(\frac{S_n}{n} - EX\Big)$$

converges in law to a normal distribution. The central limit theorem thus deals with fluctuations of S_n/n from EX of size $O(1/\sqrt{n})$. The probability of such a fluctuation is $O(1)$.

The theory of large deviations deals with larger fluctuations. In this book, we will primarily be interested in fluctuations that are $O(1)$ in size; the probability of such large fluctuations typically decays exponentially in n.

Example 1.5
Suppose X is exponential with mean $1/\lambda$. Then for $x > 1/\lambda$,

$$\frac{1}{n}\log P\Big(\frac{S_n}{n} \geq x\Big) \to -\big(\lambda x - \log(\lambda x) - 1\big) \tag{1.7}$$

(which is strictly negative). ◇

It is not straightforward to prove (1.7). Happily, it is easy to find it as an upper bound, even for general X. Define

$$\Lambda(\theta) = \log Ee^{\theta X}.$$

This is known as the *cumulant* or *log moment generating function* of X. It is a function defined on $\theta \in \mathbb{R}$, and taking values in the extended real numbers $\mathbb{R}^* = \mathbb{R} \cup \{+\infty\}$. Closely related to it is

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

This is known as the *convex conjugate* or *Fenchel-Legendre* transform of Λ. It is a function defined on $x \in \mathbb{R}$, and taking values in $\mathbb{R}^*$.

Lemma 1.2 *Let X_n and S_n be as above, and let $\Lambda(\theta)$ be the log moment generating function of X. Then*

$$\frac{1}{n}\log P\Big(\frac{S_n}{n} \geq x\Big) \leq -\sup_{\theta\geq 0} \theta x - \Lambda(\theta). \tag{1.8}$$

Proof. For any $\theta \geq 0$,

$$\begin{aligned} P(S_n/n \geq x) &= E\big(1[S_n - nx \geq 0]\big) \\ &\leq E\big(e^{\theta(S_n - nx)}\big) = e^{-\theta nx} E e^{\theta S_n}. \end{aligned}$$

This inequality is known as the Chernoff bound. Since the X_n are independent and identically distributed,

$$Ee^{\theta S_n} = \big(Ee^{\theta X}\big)^n = e^{n\Lambda(\theta)}.$$

Taking logarithms and dividing by n,

$$\frac{1}{n}\log P(S_n \geq nx) \leq -\big(\theta x - \Lambda(\theta)\big).$$

Optimising this bound over θ yields the result. □

When $x > EX$, we show in Lemma 2.6 that taking the supremum over $\theta \in \mathbb{R}$ in $\Lambda^*(x)$ is the same as taking the supremum over $\theta \geq 0$, and so the right hand side in (1.8) is $-\Lambda^*(x)$. (A similar bound applies to $P(S_n \leq nx)$ for $x < EX$ by considering $\theta \leq 0$.)

Exercise 1.6
Calculate $\Lambda^*(x)$ in the case where X is exponential with mean $1/\lambda$. Check that your answer agrees with Example 1.5. ◇

It turns out that Chernoff's bound is tight, in the sense that it gives the correct exponential rate of decay of the probability $P(S_n/n \geq x)$. This is the content of Cramér's theorem.

Cramér's Theorem

As before, let X be a random variable and let $(X_n, n \in \mathbb{N})$ be independent, identically distributed random variables each distributed like X, and let $S_n = X_1 + \cdots + X_n$. Let $\Lambda(\theta)$ be the log moment generating function of X, and let $\Lambda^*(x)$ be its convex conjugate.

Theorem 1.3 (Cramér's theorem) *For any measurable set $B \subset \mathbb{R}$,*

$$
\begin{aligned}
-\inf_{x\in B^\circ} \Lambda^*(x) &\le \liminf_{n\to\infty} \frac{1}{n} \log P(S_n/n \in B) && (1.9)\\
&\le \limsup_{n\to\infty} \frac{1}{n} \log P(S_n/n \in B) \le -\inf_{x\in \bar{B}} \Lambda^*(x), && (1.10)
\end{aligned}
$$

where B° denotes the interior of B, and $\bar{B}$ its closure.

Cramér's theorem is an example of a large deviations principle (LDP), a general statement of which can be found in Chapter 4. The inequality (1.9) is called the large deviations lower bound, and (1.10) is called the large deviations upper bound. If both hold, we say that the sequence S_n/n satisfies the large deviations principle with rate function Λ^*.

We collect the proof of Cramér's theorem, together with some important properties of Λ and Λ^*, in Chapter 2.

> *Note.* The reader is strongly encouraged to skim through Chapter 2 now. The results look technical, but it does help to develop a feel for Λ and Λ^*.

Here are some properties which we will need shortly. The function Λ^* is convex, which means that the region where it is finite is an interval, and that it is continuous on the interior of this region. Moreover, supposing EX is finite, Λ^* attains its minimum at EX, and for $x > EX$

$$\limsup_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \ge x\Big) \le -\inf_{y\ge x} \Lambda^*(y) = -\Lambda^*(x), \tag{1.11}$$

and also

$$\liminf_{n\to\infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} > x\Big) \ge -\inf_{y> x} \Lambda^*(y) = -\Lambda^*(x+) \tag{1.12}$$

where $\Lambda^*(x+) = \lim_{y\downarrow x} \Lambda^*(y)$ (and this limit is guaranteed to exist as an extended real number).

1.3 Application to Queues with Large Buffers

The following theorem is one of the fundamental results in the application of large deviation theory to queues.

Recall the queueing model from the beginning of this chapter. Let A be a random variable and let $(A_t, t \in \mathbb{Z})$ be a collection of independent random variables each distributed like A. Interpret A_t as the amount of work arriving at the queue at (or rather just before) time t. Let the queue have constant service rate C. Let Q be the queue size at time 0, given by

$$Q = \sup_{t \geq 0} S_t - Ct,$$

where $S_t = A_0 + \cdots + A_{-t+1}$ and $S_0 = 0$.

Let $\Lambda(\theta)$ be the log moment generating function of A, and assume that it is finite for all θ. (This is not necessary, but it does simplify the proof slightly.)

Theorem 1.4 (LDP for queue size) *Assume that $EA < C$. Then for sequences of $l \in \mathbb{R}$, and $q > 0$,*

$$\lim_{l \to \infty} \frac{1}{l} \log P(Q/l > q) = -I(q)$$

where

$$I(q) = \inf_{t \in \mathbb{R}^+} t\Lambda^*(C + q/t).$$

Before the proof, some remarks. There will be more remarks and examples after the proofs.

i. Equivalently, $q^{-1} \log P(Q > q) \to -I(1)$. We have written out the fuller form to make it look more like a large deviations principle.

ii. This is a restricted sort of large deviations principle. The theorem only concerns intervals $[q, \infty)$, whereas a full large deviations principle would deal with general events $\{Q/l \in B\}$.

iii. The theorem proves a limit, whereas large deviations principles typically give upper and lower bounds. In fact we will prove large deviations upper and lower bounds, but in this case they happen to agree.

iv. The assumption that the A_t are independent is overly restrictive. We will give a more sophisticated version of this theorem in Chapter 3, drawing on more advanced large deviations theory than Cramér's theorem, with less restrictive assumptions.

v. For this type of theorem to be meaningful, it is necessary that Λ be finite in a neighbourhood of the origin. When that is not the case, a different scaling regime is appropriate, sometimes referred to as the *heavy tail* or *subexponential* regime. See for example Asmussen [3], Whitt [98], and Zwart [103].

Proof. The following lemmas prove: a lim sup result with rate function I_1, then a lim inf result with rate function I_2, and finally that $I_1 = I_2 = I$. □

Lemma 1.5 *In the setting of Theorem 1.4,*

$$\limsup_{l\to\infty}\frac{1}{l}\log P(Q/l>q)\le -q\sup\{\theta>0:\Lambda(\theta)<\theta C\}.$$

Lemma 1.6 *In the setting of Theorem 1.4,*

$$\liminf_{l\to\infty}\frac{1}{l}\log P(Q/l>q)\ge -\inf_{t\in\mathbb{R}^+}t\Lambda^*(C+q/t).$$

Lemma 1.7 *In the setting of Theorem 1.4,*

$$\begin{aligned} I(q) &= \inf_{t\in\mathbb{R}^+} t\Lambda^*(C+q/t) && (1.13)\\ &= \inf_{t\in\mathbb{R}^+}\sup_{\theta\ge 0}\theta(q+Ct)-t\Lambda(\theta) && (1.14)\\ &= q\sup\{\theta>0:\Lambda(\theta)<\theta C\}. && (1.15)\end{aligned}$$

Proof of Lemma 1.5 As in the upper bound for Cramér's Theorem, Lemma 1.2, we will use Chernoff's bound. First, write down the probability we are trying to estimate. From the definition of Q,

$$P(Q>lq)=P(\sup_{t\ge 0}S_t-Ct>lq)\le\sum_{t\ge 0}P(S_t-Ct\ge lq)$$

and so by Chernoff's bound,

$$\le e^{-\theta lq}\sum_{t\ge 0}e^{t\left(\Lambda(\theta)-C\theta\right)}.$$

for any $\theta>0$. Restrict attention to those θ for which $\Lambda(\theta)<\theta C$. This makes the sum finite:

$$\le e^{-\theta lq}\frac{e^{\Lambda(\theta)-C\theta}}{1-e^{\Lambda(\theta)-C\theta}},$$

and so

$$\limsup_{l\to\infty}\frac{1}{l}\log P(Q>lq)\le-\theta q.$$

Taking the supremum over all such θ,

$$\limsup_{l\to\infty}\frac{1}{l}\log P(Q>lq)\le -q\sup\{\theta>0:\Lambda(\theta)<\theta C\}.\qquad\square$$

If there were no such θ, the bound would be trivial. In fact this is never the case, by our assumption that $EA < C$. We assumed that $\Lambda(\theta)$ is finite and, in particular, finite in a neighbourhood of the origin. Hence it is differentiable in a neighbourhood of the origin (Lemma 2.3), and furthermore $EA = \Lambda'(0)$. Since we have assumed $EA < C$, there exists a $\theta > 0$ such that $\Lambda(\theta) < \theta C$.

Proof of Lemma 1.6 We will prove the lower bound by estimating the probability that the queue overflows over some fixed timescale. It is a common method, in proving large deviations lower bounds, to bound the probability of a rare event by finding the probability that it occurs in a specific way. Fix $t > 0$, $t \in \mathbb{R}$. Then, from the definition of Q,

$$P(Q > lq) = P(\exists u : S_u - Cu > lq) \geq P(S_{\lceil lt \rceil} - C\lceil lt \rceil > lq),$$

where $\lceil x \rceil \in \mathbb{Z}$ is the smallest integer greater than or equal to x, i.e. $\lceil x \rceil - 1 < x \leq \lceil x \rceil$. Hence

$$\liminf_{l\to\infty} \frac{1}{l} \log P(Q \geq lq) \geq \liminf_{l\to\infty} \frac{1}{l} \log P(S_{\lceil lt \rceil} - C\lceil lt \rceil > lq)$$

so using the fact that $l \leq \lceil lt \rceil / t$,

$$\geq \liminf_{l\to\infty} \frac{t}{\lceil lt \rceil} \log P\Big(S_{\lceil lt \rceil} - C\lceil lt \rceil > \frac{\lceil lt \rceil}{t} q\Big).$$

Defining $n = \lceil lt \rceil$,

$$\begin{aligned} &= t \liminf_{n\to\infty} \frac{1}{n} \log P(S_n - Cn > nq/t) \\ &= t \liminf_{n\to\infty} \frac{1}{n} \log P(S_n/n > C + q/t). \end{aligned}$$

By the lower bound in Cramér's theorem, (1.12), this is

$$\geq -t\Lambda^*\big((C + q/t)+\big).$$

Since $t > 0$ was arbitrary

$$\liminf_{l\to\infty} \frac{1}{l} \log P(Q > lq) \geq -\inf_{t>0} t\Lambda^*\big((C + q/t)+\big)$$

and it is easy to see from the properties of Λ^* that this is

$$= -\inf_{t>0} t\Lambda^*(C + q/t). \qquad \square$$

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

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

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 $t \in \mathbb{R}^+$,

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

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):

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

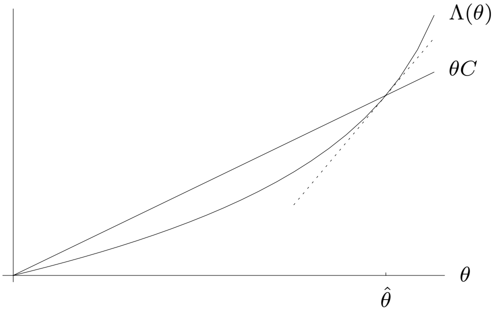

Figure 1.1: Illustration of $\hat\theta = \sup\{\theta : \Lambda(\theta) < \theta C\}$.

Performing the optimization over θ,

$$= \inf_{t>0} \begin{cases} \infty & \text{if } t < q/(\Lambda'(\theta^*) - C) \\ \theta^* t(\Lambda'(\theta^*) - C) & \text{if } t \geq q/(\Lambda'(\theta^*) - C) \end{cases}$$
$$= \theta^* q = (1.15).$$

This completes the proof. □

Some remarks on the proofs.

i. One interpretation of Theorem 1.4 is that the approximation

$$P(\sup_t S_t - Ct \geq q) \approx \sup_t P(S_t - Ct \geq q)$$

is justified (for large q) on a logarithmic scale. In other words, to estimate the probability that the queue size is large, we need to find the timescale t over which it is most likely that the queue fills up to level q, and then estimate the probability that it fills up over that timescale.

ii. From the proof of Lemma 1.6, the most likely time for the queue to fill up to some high level lq is lt, where t is the optimizing parameter in $I(q)$.

iii. Another interpretation of Theorem 1.4 is that, on a macroscopic scale, the process $S_t - Ct$ is effectively a simple random walk with negative drift. The theorem implies that the approximation

$$P(\sup_t S_t - Ct \geq q_1 + q_2) \approx P(\sup_t S_t - Ct \geq q_1) P(\sup_t S_t - Ct \geq q_2) \quad (1.16)$$

is valid (for large q_1 and q_2). If S_t were a simple random walk, we would have equality in (1.16), by the strong Markov property. Thus the effects of 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

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

which agrees with our earlier conclusion. ◇

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

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

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 $C_t = C > \mu$. ◇

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).

Consider a single-server queue as before, with N sources and constant service rate CN. Let $A_t^{(i)}$ be the amount of work arriving from source i at time t. Assume that for each i, $(A_t^{(i)}, t \in \mathbb{Z})$ is a stationary sequence of random variables, and that these sequences are independent of each other, and identically distributed.

To put this in a familiar context, set $A_t^N = A_t^{(1)} + \cdots + A_t^{(N)}$ and $S_t^N = A_0^N + \cdots + A_{-t+1}^N$ (and $S_0^N = 0$). So S_t^N is the total amount of work arriving at the queue in the interval $(-t, 0]$. Then the queue length at time 0 is given by

$$Q^N = \sup_{t \geq 0} S_t^N - NCt.$$

We will consider the behaviour of $P(Q^N \geq Nq)$ as the number of sources N becomes large.

We will prove a simple sort of large deviations principle for Q^N/N, using the same techniques as before: Chernoff's bound for the upper bound, Cramér's theorem for the lower bound. Define

$$\Lambda_t(\theta) = \log E e^{\theta S_t^1}.$$

Assume that, for all t, $\Lambda_t(\theta)$ is finite for θ in a neighbourhood of the origin. Assume that the limit

$$\Lambda(\theta) = \lim_{t \to \infty} \frac{1}{t} \Lambda_t(\theta) \tag{1.17}$$

exists and is finite and differentiable, for θ in a neighbourhood of the origin.

Theorem 1.8 *Under these two assumptions, and the stability assumption that $ES_1^1 < C$,*

$$\begin{aligned} -I(q+) &\leq \liminf_{N \to \infty} \frac{1}{N} \log P(Q^N > Nq) \\ &\leq \limsup_{N \to \infty} \frac{1}{N} \log P(Q^N > Nq) \leq -I(q) \end{aligned}$$

where

$$I(q) = \inf_{t \in \mathbb{N}} \Lambda_t^*(q + Ct).$$

Some remarks.

i. Note that the rate function $I(q)$ has nearly the same form as that for the large-buffer limit, (1.14).

ii. As does Cramér's theorem, this result involves both an upper and a lower bound. If Λ_t^* is continuous for each t, then the two bounds agree and we obtain a straightforward limit.

iii. We have assumed that $t^{-1}\Lambda_t(\theta)$ converges to a limit $\Lambda(\theta)$, and yet $\Lambda(\theta)$ does not appear anywhere in the result. The assumption is just a way to control the tails—to say that there are no surprises in the distribution of S_t^1 for large t. See Exercise 1.3 for an example. There are different sorts of assumption that we can use to control the tail; see Section 3.2 for more.

The many sources limit was first introduced by Weiss [97]. Versions of Theorem 1.8 were proved independently by Courcoubetis and Weber [21] (in discrete time, for queues with finite buffers), by Botvich and Duffield [9] (in discrete and continuous time, for queues with infinite buffers) and by Simonian and Guibert [92] (in continuous time, for queues with infinite buffers fed by on-off sources). For more, see Section 3.2 and Chapter 7.

The proof is in two parts, a lower bound and an upper bound, presented in the following two lemmas.

Lemma 1.9 *Under the assumptions of Theorem 1.8,*

$$\liminf_{N\to\infty} \frac{1}{N} \log P\big(Q^N > Nq\big) \geq -\lim_{r\downarrow q} \inf_{t\in\mathbb{N}} \Lambda_t^*(q+Ct).$$

Lemma 1.10 *Under the assumptions of Theorem 1.8,*

$$\limsup_{N\to\infty} \frac{1}{N} \log P\big(Q^N > Nq\big) \leq -\inf_{t\in\mathbb{N}} \Lambda_t^*(q+Ct).$$

Proof of Lemma 1.9 First, write out the probability we are estimating:

$$P(Q^N > Nq) = P(\sup_{t\geq 0} S_t^N - NCt > Nq).$$

Fix t. This probability is then

$$\geq P(S_t^N/N > q+Ct).$$

(It seems we are throwing away a lot of probability mass. It turns out in the large deviations limit that we aren't: as is common in large deviations lower bounds, we need only consider the probability that the rare event occurs in a specific way.) Now, apply the lower bound part of Cramér's theorem, (1.12):

$$\liminf_{N\to\infty} \frac{1}{N} \log P(Q^N > Nq) \geq \Lambda_t^*\big((q+Ct)+\big).$$

Taking the supremum over all $t > 0$, and using the properties of Λ^* mentioned in Section 1.2,

$$\begin{aligned}\liminf_{N\to\infty} \frac{1}{N} \log P(Q^N > Nq) &\geq -\inf_{t>0} \inf_{r>q} \Lambda_t^*(r+Ct)\\ &= -\inf_{r>q} \inf_{t>0} \Lambda_t^*(r+Ct)\\ &= -\lim_{r\downarrow q} \inf_{t>0} \Lambda_t^*(r+Ct),\end{aligned}$$

where the last equality follows from the fact that, for each t, $\Lambda_t^*(r+Ct)$ is increasing in r. □

Proof of Lemma 1.10 First, write out the probability we are estimating:

$$\begin{aligned}P(Q^N > Nq) &= P(\sup_{t\geq 0} S_t^N - NCt > Nq)\\ &\leq \sum_{t\geq 0} P(S_t^N \geq NCt + Nq).\end{aligned}$$

The result we are trying to prove suggests that all the action is happening at finite timescales. This motivates us to break up the sum, into some finite-timescale parts and an infinite-timescale tail:

$$\begin{aligned}&= P(S_0^N \geq Nq) + \cdots + P(S_{t_0}^N \geq NCt_0 + Nq)\\ &\quad + \sum_{t>t_0} P(S_t^N \geq NCt + Nq).\end{aligned}$$

Now, look at this probability on the large deviations scale: take logarithms, divide by N, take the lim sup. Using the fact that for sequences a_n and b_n in $\mathbb{R}^+$

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

we obtain

$$\begin{aligned}\limsup_{N\to\infty} \frac{1}{N} \log P(Q^N > Nq) \leq \max_{0\leq t\leq t_0} &\limsup_{N\to\infty} \frac{1}{N} \log P(S_t^N/N \geq q + Ct)\\ &\vee \limsup_{N\to\infty} \frac{1}{N} \log \sum_{t>t_0} P(S_t^N/N \geq q + Ct).\end{aligned}$$

(This is an instance of the *principle of the largest term*, a key idea in large deviations theory. Here, we are only using a simple form of the principle, a form which is discussed in more detail in Section 2.2.)

Each of the finite-timescale parts is easy to deal with on its own. By the upper bound part of Cramér's theorem, (1.11),

$$\limsup_{N\to\infty} \frac{1}{N} \log P(S_t^N/N \geq q+Ct) \leq -\Lambda_t^*(q+Ct).$$

The following lemma controls the infinite-timescale part: it says that

$$\limsup_{N\to\infty} \frac{1}{N} \log \sum_{t>t_0} P(S_t^N/N \geq q+Ct) \to -\infty \text{ as } t_0 \to \infty.$$

Combining these two, we have the result. □

Lemma 1.11 *Under the assumptions of Theorem 1.8,*

$$\limsup_{N\to\infty} \frac{1}{N} \log \sum_{t>t_0} P(S_t^N/N \geq q+Ct) \to -\infty \text{ as } t_0 \to \infty.$$

Proof. We will use Chernoff's bound, and then the tail-controlling assumption. First, using Chernoff's bound, for any $\theta > 0$

$$\sum_{t>t_0} P(S_t^N/N \geq q+Ct) \leq \sum_{t>t_0} e^{-N\theta(q+Ct)} E e^{\theta S_t^N}.$$

Using the fact that the input flows are all independent and identically distributed, this is

$$= \sum_{t>t_0} e^{-N\left(\theta(q+Ct)-\Lambda_t(\theta)\right)}. \tag{1.18}$$

We want to choose a θ such that $\theta(q+Ct) - \Lambda_t(\theta)$ is strictly negative, uniformly in t. To do this, use the tail-controlling assumption, as follows.

First, note that $\Lambda(\cdot)$ is the limit of convex functions $t^{-1}\Lambda_t(\cdot)$. We have assumed that for each t, $\Lambda_t(\cdot)$ is finite in a neighbourhood of the origin; hence, by Lemma 2.3, it is differentiable at the origin, with $\Lambda_t'(0) = ES_t^1$. By the stationarity assumption, $t^{-1}\Lambda_t'(0) = \mu$ where $\mu = ES_1^1$. By the following lemma, Lemma 1.12, it follows that $\Lambda'(0) = \mu$. Also, by the stability assumption, $\mu < C$. So there exists some $\theta > 0$ for which

$$\Lambda(\theta) < \theta(C - 2\delta),$$

for some $\delta > 0$.

Since $\Lambda_t(\theta)/t \to \Lambda(\theta)$, there exists a t_0 such that for $t > t_0$,

$$\Lambda_t(\theta) < t\big(\Lambda(\theta) + \delta\theta\big)$$

and so

$$\theta(q + Ct) - \Lambda_t(\theta) \geq \theta Ct - \Lambda_t(\theta) > \theta Ct - t\big(\Lambda(\theta) + \delta\theta\big).$$

By our choice of θ, this is

$$> \theta Ct - t\big(\theta(C - 2\delta) + \delta\theta\big) = \theta\delta t.$$

Now we can estimate the probability we want: for t_0 sufficiently large,

$$(1.18) \leq \sum_{t>t_0} e^{-N\theta\delta t} = \frac{e^{-N\theta\delta(t_0+1)}}{1 - e^{-N\theta\delta}}$$

and so

$$\limsup_{N\to\infty} \frac{1}{N} \log \sum_{t>t_0} P(S_t^N/t \geq q + Ct) \leq -\theta\delta(t_0 + 1).$$

Taking the limit as $t_0 \to \infty$ completes the proof. □

Lemma 1.12 *Let $(f_n,\ n \in \mathbb{N})$ be a sequence of convex functions $f_n : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, converging pointwise to a function f (which is necessarily convex). Suppose that for each n, f_n is finite and differentiable in a neighbourhood of the origin, and that $f_n'(0) = \mu_n$. If f is differentiable at the origin and if $\mu_n \to \mu$ then $f'(0) = \mu$.*

Proof. To see that f is convex, note that for any $0 \leq \alpha \leq 1$

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &= \lim_{n\to\infty} f_n(\alpha x + (1-\alpha)y) \\ &\leq \lim_{n\to\infty} \alpha f_n(x) + (1-\alpha) f_n(y) = \alpha f(x) + (1-\alpha) f(y). \end{aligned}$$

Now, by convexity and differentiability of f_n, for all θ

$$f_n(\theta) \geq \theta f_n'(0) = \theta\mu_n.$$

Hence

$$f(\theta) = \lim_{n\to\infty} f_n(\theta) = \liminf_{n\to\infty} f_n(\theta) \geq \liminf_{n\to\infty} \theta\mu_n = \theta\mu.$$

Since f is differentiable at the origin, it must be that $f'(0) = \mu$. □

Some remarks.

i. The proofs made heavy use of the principle of the largest term: loosely, the idea that to estimate the probability of a rare event, we only need concern ourselves with the most likely way in which this event can occur. In proving the lower bound, we estimated the probability that the queue overflows over a fixed time t, and took the most likely time t^*. In proving the upper bound, we showed that the probability of overflow over time $t \neq t^*$ is negligible, on a logarithmic scale.

ii. In the large-buffer limit, Theorem 1.4, the optimizing τ^* relates the queue size q to the most likely time to overflow $q\tau^*$. Thus the most likely rate for the queue to build up is $1/\tau^*$, and this does not depend on q. In the many-sources limit, the optimizing t^* is simply the most likely time to overflow. It typically depends on q in a non-linear way.

iii. Compare the forms of the rate functions for the large-buffer limit

$$I(q) = \inf_{t \in \mathbb{R}^+} \sup_{\theta \in \mathbb{R}} \theta(q + Ct) - t\Lambda(\theta)$$

and the many-flows limit

$$I(q) = \inf_{t \in \mathbb{N}_0} \sup_{\theta \in \mathbb{R}} \theta(q + Ct) - \Lambda_t(\theta).$$

If in the many-flows case the amount of work arriving from each flow at each timestep is independent, then $\Lambda_t(\theta) = t\Lambda(\theta)$, and the two expressions are nearly identical.

iv. The tail-controlling assumption was needed to prove the upper bound, Lemma 1.10, but not the lower bound, Lemma 1.9.

Exercise 1.9
Let $(A_t^{(i)}, t \in \mathbb{Z})$ be a two-state Markov chain representing a source which produces an amount of work h in each timestep while in the on state and no work in the off state, and which flips from on to off with probability p and from off to on with probability q. Show that

$$\Lambda_t(\theta) = \log\Bigl(\frac{q}{q+p} E_t + \frac{p}{q+p} F_t\Bigr)$$

where

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

Explain why $\Lambda_t(\theta)/t \to \Lambda(\theta)$, and show that $\Lambda(\theta)$ is everywhere differentiable. ◇

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

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

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

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

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 23–45, 2004.

Thus

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

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 $q < p$. ◇

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 $S_n = Y_1 + \ldots + Y_n$. The sample mean S_n/n is also normally distributed, with mean zero and variance $1/n$. Thus, for any $x > 0$,

$$\begin{aligned} P\Big(\frac{S_n}{n} > x\Big) &= \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-nz^2/2} dz \\ &\le \frac{1}{x\sqrt{2\pi}} \int_x^\infty z e^{-nz^2/2} dz = \frac{1}{nx\sqrt{2\pi}} e^{-nx^2/2}. \end{aligned}$$

Also

$$\begin{aligned} P\Big(\frac{S_n}{n} > x\Big) &\ge \frac{1}{\sqrt{2\pi}} \int_x^{x+n^{-1}} e^{-nz^2/2} dz \\ &\ge \frac{1}{n\sqrt{2\pi}} \exp\Big(-\frac{n}{2}\Big(x+\frac{1}{n}\Big)^2\Big) = \frac{1}{n\sqrt{2\pi}} e^{-x-(1/2n)} e^{-nx^2/2}. \end{aligned}$$

Thus, ignoring terms that are subexponential in n, we get

$$P\Big(\frac{S_n}{n} > x\Big) \approx \exp\Big(-n\frac{x^2}{2}\Big).$$

By symmetry, a similar expression holds for $P(S_n/n < x)$ when $x < 0$. ◇

A natural question to ask at this point is whether a similar approximation holds for sample means of i.i.d. random variables with arbitrary distribution. The answer is provided by Cramér's theorem, and is affirmative. This is one instance of what is called a large deviation principle.

In these examples, the random variables took values in $\mathbb{R}$. Large deviation principles can be stated much more generally, for random variables taking values in abstract spaces. In later chapters we shall have occasion to use the LDP for random variables taking values in the space of continuous functions $\mathbb{R}_0^+ \to \mathbb{R}$. For a general statement of the large deviations principle, see Chapter 4. In this chapter, we will stick to $\mathbb{R}$ and $\mathbb{R}^d$.

2.2 Principle of the Largest Term

In both the previous examples, the probability estimates were governed by a single point: $n^{-1}L_n = q$ in the first, $n^{-1}S_n = x$ in the second. This is to do with the *principle of the largest term*. We shall have a great deal more to say about this principle when we introduce abstract large deviations, and when we apply it to queues. For now, here is a simple concrete explanation.

Let A_n and B_n be sequences of events, with A_n and B_n disjoint. Suppose that

$$\frac{1}{n}\log P(A_n) \to -a \quad \text{and} \quad \frac{1}{n}\log P(B_n) \to -b.$$

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

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

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

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

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*

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

and

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

(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 deviations principle with rate function Λ^*. In fact this definition can also be applied to sequences of random variables which do not arise as sums of i.i.d. random variables, as the examples in Section 4.2 show.

Write $\mathbb{R}^*$ for the *extended real numbers*, $\mathbb{R} \cup \{+\infty\}$.

Definition 2.1 *A function $I : \mathbb{R}^d \to \mathbb{R}^*$ is a* rate function *if*

- *$I(x) \geq 0$ for all $x \in \mathbb{R}^d$;*
- *I is lower semicontinuous, i.e. the level sets $\{x : I(x) \leq \alpha\}$ are all closed, for $\alpha \in \mathbb{R}$.*

It is called a good rate function *if in addition*

- *the level sets are all compact.*

Definition 2.2 (Large deviations principle) *Let $(X_n,\ n \in \mathbb{N})$ be a sequence of random variables taking values in $\mathbb{R}^d$. Say that X_n satisfies a large deviations principle in $\mathbb{R}^d$ with rate function I if I is a rate function, and if for any measurable set $B \subset \mathbb{R}^d$*

$$-\inf_{x \in B^\circ} I(x) \leq \liminf_{n\to\infty} \frac{1}{n} \log P(X_n \in B) \tag{2.2}$$

$$\leq \limsup_{n\to\infty} \frac{1}{n} \log P(X_n \in B) \leq -\inf_{x \in \bar{B}} I(x), \tag{2.3}$$

where B° denotes the interior of B, and $\bar{B}$ its closure.

We will now go on to prove Cramér's theorem, and some extensions. First we need a good deal of technical work, to understand some properties of Λ and Λ^*.

2.4 Cumulant Generating Functions

Definition 2.3 *The* cumulant generating function *or* logarithmic moment generating function *of a real-valued random variable X is a function $\Lambda : \mathbb{R} \to \mathbb{R}^*$, defined by*

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

Definition 2.4 *The* effective domain *of a function $f : \mathcal{X} \to \mathbb{R}^*$ is the set $\{x \in \mathcal{X} : f(x) < \infty\}$.*

The following lemma summarises some of the key properties of the cumulant generating function.

Lemma 2.3 *The cumulant generating function Λ is convex and lower semicontinuous, and $\Lambda(0) = 0$. It is differentiable in the interior of its effective domain, with derivative*

$$\Lambda'(\theta) = E(Xe^{\theta X})/e^{\Lambda(\theta)}.$$

Proof. Using Hölder's inequality,

$$\begin{aligned} E\big(e^{(\alpha\theta_1+(1-\alpha)\theta_2)X}\big) &= E\big((e^{\theta_1 X})^{\alpha}(e^{\theta_2 X})^{1-\alpha}\big) \\ &\le \big(E(e^{\theta_1 X})\big)^{\alpha}\big(E(e^{\theta_2 X})\big)^{1-\alpha}, \end{aligned}$$

for $\alpha \in [0,1]$. Taking logarithms, the convexity of Λ is immediate. Next, fix $\theta \in \mathbb{R}$ and let θ_n be any sequence converging to θ. Then, by Fatou's lemma,

$$E(e^{\theta X}) \le \liminf_{n\to\infty} E(e^{\theta_n X}).$$

Taking logarithms yields that Λ is lower semicontinuous. It is also straightforward to see that $\Lambda(0) = \log E(1) = 0$.

To verify differentiability, let θ be in the interior of the effective domain of Λ and observe that

$$\lim_{\delta\to 0}\frac{1}{\delta}\Big(E(e^{(\theta+\delta)X}) - E(e^{\theta X})\Big) = \lim_{\delta\to 0} E\Big(\frac{e^{(\theta+\delta)X} - e^{\theta X}}{\delta}\Big), \tag{2.4}$$

by the linearity of expectations. Now, $(e^{(\theta+\delta)X} - e^{\theta X})/\delta$ converges pointwise to $Xe^{\theta X}$ as δ goes to zero, and is dominated by $Z := e^{\theta X}(e^{\epsilon|X|} - 1)/\epsilon$, for all $\delta \in (-\epsilon, \epsilon)$. (This can be readily verified using the convexity of $z \mapsto e^{zX}$ for every X.) By the assumption that θ is in the interior of the effective domain of Λ, so are $\theta + \epsilon$ and $\theta - \epsilon$, for small enough ϵ. Hence EZ is finite. Thus, by the dominated convergence theorem,

$$E\Big(\frac{e^{(\theta+\delta)X} - e^{\theta X}}{\delta}\Big) \to E(Xe^{\theta X}) \quad \text{as } \delta \to 0. \tag{2.5}$$

Define $M(\theta) = E(e^{\theta X})$. It follows from (2.4) and (2.5) that $M'(\theta) = E(Xe^{\theta X})$. Since $\Lambda(\theta) = \log M(\theta)$, the last claim of the lemma follows from the chain rule. □

Exercise 2.4

Let X and Y be independent random variables with cumulant generating functions Λ_X and Λ_Y, and let a be non-zero. Find the cumulant generating functions of $X + Y$ and aX. ◇

Exercise 2.5
Suppose N takes values in $\mathbb{N}$, has cumulant generating Λ_N, and is independent of $X_1, X_2, \ldots$, which are i.i.d. with cumulant generating function Λ_X. Find the cumulant generating function of $X_1 + \cdots + X_N$. $\diamond$

Exercise 2.6
Calculate the cumulant generating functions of the following.

i. $X \sim$ Bernoulli(p),
ii. $X \sim$ Binomial(n, p),
iii. $X \sim$ Poisson(λ),
iv. $X \sim$ Exponential(λ),
v. $X \sim$ Geometric(ρ),
vi. $X \sim$ Normal(μ, σ^2),
vii. $X \sim$ Cauchy, with density $f(x) = \pi^{-1}(1 + x^2)^{-1}$, $x \in \mathbb{R}$. $\diamond$

Exercise 2.7
Let X be an $\mathbb{R}^d$-valued random variable. Its cumulant generating function $\Lambda : \mathbb{R}^d \to \mathbb{R}^*$ is defined by

$$\Lambda(\theta) = \log E\big(\exp(\theta \cdot X)\big),$$

where $\theta \cdot x$ denotes the inner product of θ and x in $\mathbb{R}^d$. Show that Lemma 2.3 still applies, i.e. Λ is convex, $\Lambda(0) = 0$, and Λ is differentiable in the interior of its effective domain, with $\nabla\Lambda(\theta) = E(Xe^{\theta \cdot X})/e^{\Lambda(\theta)}$. $\diamond$

2.5 Convex Duality

Let f be a function on $\mathbb{R}^d$ taking values in the extended real numbers, $\mathbb{R}^*$, and suppose f isn't identically infinite.

Definition 2.5 *The* convex conjugate *or* Fenchel-Legendre transform *of f, denoted f^*, is another extended real-valued function on $\mathbb{R}^d$, defined by*

$$f^*(\theta) = \sup_{x \in \mathbb{R}^d} \big(\theta \cdot x - f(x)\big), \tag{2.6}$$

where $\theta \cdot x$ denotes the inner product of θ and x.

The function f^* doesn't take the value $-\infty$ as there is at least one x for which $f(x)$ is finite. Geometrically, $f^*(\theta)$ is the smallest amount by which the hyperplane $y = \theta \cdot x$ has to pushed down (or the negative of the largest amount it can be pushed up) so as to lie below the graph of the function f. This is easiest to visualise for $d = 1$, where $y = \theta x$ is a straight line with slope θ.

Lemma 2.4 *The function f^* is convex and lower semicontinuous. If f is a convex function, differentiable at $x \in \mathbb{R}^d$ with $\nabla f(x) = \eta$, then $f^*(\eta) = \eta \cdot x - f(x)$.*

Proof. Recall that a function $g : \mathbb{R}^d \to \mathbb{R}^*$ is said to be convex if its *epigraph*, $\{(x, y) \in \mathbb{R}^{d+1} : y \geq g(x)\}$, is a convex set. The supremum of convex functions is convex since its epigraph is the intersection of the epigraphs of the functions over which the supremum is taken. Since f^* is the supremum of the convex (in fact, linear) functions, $g_x(\theta) = \theta \cdot x - f(x)$, it is convex. Next, observe that each g_x is continuous, and so the level sets $\{\theta : g_x(\theta) \leq \alpha\}$ are closed for all $\alpha \in \mathbb{R}$. The level sets of $f^* = \sup_x g_x$ are the intersection of the corresponding level sets of the g_x, and hence they are closed. Therefore, f^* is lower semicontinuous.

Suppose next that f is convex, and that $\nabla f(x) = \eta$. Then, for all $y \in \mathbb{R}^d$, $f(y) \geq f(x) + \eta \cdot (y - x)$. To see this, note that

$$\frac{f\big((1-\delta)x + \delta y\big) - f(x)}{\delta} \leq f(y) - f(x) \quad \text{for all } y \in \mathbb{R}^d \text{ and } \delta \in (0, 1],$$

and that the left hand side converges to $\eta \cdot (y - x)$ as δ decreases to zero. Hence

$$f^*(\eta) = \sup_{y \in \mathbb{R}^d} \big(\eta \cdot y - f(y)\big) \leq \eta \cdot x - f(x).$$

In fact, equality holds above, as can be seen by taking $y = x$. This completes the proof of the lemma. □

A function is said to be *closed convex* if its epigraph is a closed convex set. A convex function is closed if it is lower semicontinuous. We have shown that, for any extended real-valued function f, the convex conjugate f^* is closed and convex. The following lemma says that, if f is itself a closed convex function, then it is the conjugate of f^*, i.e., the conjugacy relation is a duality. For a proof see Rockafellar [88].

Lemma 2.5 (Duality of convex conjugate) *If f is a closed convex function on $\mathbb{R}^d$, then $(f^*)^* \equiv f$.*

Exercise 2.8 (Weak duality)
Show that, for all $f : \mathbb{R}^d \to \mathbb{R}^*$ (i.e. not just for convex functions), $(f^*)^*(x) \leq f(x)$ for all $x \in \mathbb{R}^d$. ◇

Exercise 2.9
Compute the convex conjugate of each of the cumulant generating functions in Exercise 2.6. ◇

Exercise 2.10
Compute f^* and $(f^*)^*$ in each of the following cases:
i. $f(x) = |x|$.
ii. $f(x) = \sin(x)$.
iii. $f(x) = 1/(1 - x^2)$ if $|x| < 1$ and $f(x) = +\infty$ otherwise.
iv. $f(x) = x^2$ if $|x| \leq 1$ and $f(x) = +\infty$ otherwise.
v. $f(x) = x^2$ if $|x| < 1$ and $f(x) = +\infty$ otherwise. ◇

We conclude this section with a discussion of some properties of the convex conjugates of cumulant generating functions. These will be useful for the proof of Cramér's theorem in the next section.

Lemma 2.6 *Let Λ be the cumulant generating function of a real-valued random variable X, and let Λ^* be its convex conjugate.*
i. Λ^* is non-negative, convex and lower semi-continuous.
ii. $(\Lambda^*)^* = \Lambda$.
If $\Lambda(\theta)$ is finite in a neighbourhood of zero, then
iii. $\mu = EX$ is finite and equal to $\Lambda'(0)$,
iv. $\Lambda^*(\mu) = 0$,
v. Λ^* is decreasing on $(-\infty, \mu]$ and increasing on $[\mu, \infty)$ (though in neither case is it necessarily strictly so), and

$$\Lambda^*(x) = \sup_{\theta \geq 0} \theta x - \Lambda(\theta) \quad \text{for } x \geq \mu \tag{2.7}$$

$$\Lambda^*(x) = \sup_{\theta \leq 0} \theta x - \Lambda(\theta) \quad \text{for } x \leq \mu. \tag{2.8}$$

Some of the properties apply not only to cumulant generating functions but also to the sort of generalized cumulant generating function which appears in the generalized version of Cramér's theorem, Theorem 2.11.

Lemma 2.7 *Let Λ be a convex real-valued function, taking value zero at the origin, and let Λ^* be its convex conjugate. Then Λ^* satisfies (i) of Lemma 2.6. If Λ is differentiable at the origin with $\mu = \Lambda'(0)$ then it also satisfies (iv) and (v).*

Proof of Lemma 2.6 Since Λ is a cumulant generating function, it is convex (by Lemma 2.3) and $\Lambda(0) = 0$. If it is finite in a neighbourhood of the origin then, by the same lemma it is differentiable at the origin with $\Lambda'(0) = EX$, yielding (iii). So the conditions of Lemma 2.7 are satisfied, and (i), (iv) and (v) follow.

For (ii), note that Λ is convex and lower-semicontinuous by Lemma 2.3, hence closed convex, hence is equal to $(\Lambda^*)^*$ by Lemma 2.5. □

Proof of Lemma 2.7 The convexity and lower-semicontinuity of Λ^* follow from Lemma 2.4. Its non-negativity follows from the fact that $\Lambda(0) = 0$.

If it is differentiable at the origin with $\Lambda'(0) = \mu$ then, by Lemma 2.4, $\Lambda^*(\mu) = 0$. Since Λ^* is convex and non-negative, and $\Lambda^*(\mu) = 0$, it must be decreasing on $(-\infty, \mu]$ and increasing on $[\mu, \infty)$.

Finally, if $x \geq \mu$ and $\theta < 0$,

$$\theta x - \Lambda(\theta) \leq \theta\mu - \Lambda(\theta) \leq \Lambda^*(\mu) = 0,$$

and so it is sufficient to take the supremum over $\theta \geq 0$ in (2.7). Likewise, (2.8) follows by considering $x \leq \mu$ and $\theta > 0$. □

Note. There is a close connection between exponentially-tilted random variables and convex conjugation, a connection which is important in large deviations theory.

Let X be a random variable with cumulant generating function Λ, and probability law μ. Suppose that θ is in the interior of the effective domain of Λ, and consider the *exponentially tilted* probability law $\tilde{\mu}$ defined by

$$\frac{d\tilde{\mu}}{d\mu}(x) = \exp\bigl(\theta x - \Lambda(\theta)\bigr).$$

Let $\tilde{X}$ be a random variable drawn from $\tilde{\mu}$. The tilted mean is

$$E\tilde{X} = EXe^{\theta X - \Lambda(\theta)} = \Lambda'(\theta)$$

by Lemma 2.3. Cramér's theorem, which follows, concerns a well-chosen exponential tilting.

Since θ is in the interior of the effective domain of Λ, Λ is infinitely differentiable at θ. We also know that Λ is convex. Suppose now that it is strictly convex, i.e. that $\Lambda''(\theta) > 0$. One can check that Λ^* is differentiable at x, with derivative θ.

So we have dual ways of looking at exponential tilts. If X is tilted to have mean x then the tilt parameter θ satisfies $\Lambda'(\theta) = x$; if X is tilted by parameter θ then the tilted mean satisfies $(\Lambda^*)'(x) = \theta$.

2.6 Cramér's Theorem

At last we are ready to prove Cramér's theorem. In general, our approach in this book is to take as given the standard results from large deviation theory, and to focus on their application to queues. We make an exception for Cramér's theorem and present a proof, for two reasons. First, it helps to 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,

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

Taking logarithms, for $x \geq EX$,

$$\begin{aligned}\frac{1}{n}\log P\Big(\frac{S_n}{n} \in [x,\infty)\Big) &\leq -\sup_{\theta \geq 0} \theta x - \Lambda(\theta)\\ &= -\Lambda^*(x) \quad \text{by (2.7)}\\ &= -\inf_{y \in [x,\infty)} \Lambda^*(y)\end{aligned}$$

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

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

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,

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

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

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

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

$$\liminf_{n \to \infty} \frac{1}{n} \log P\Big(\frac{S_n}{n} \in G\Big) \ge -\Lambda^*(x). \tag{2.9}$$

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:

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

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

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

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 $g(\theta) \to \infty$ as $|\theta| \to \infty$, 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

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

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.)

We can now estimate the probability of interest, using the fact that (since G is open), the set $(x-\delta, x+\delta)$ is contained in G for sufficiently small δ. Let $\tilde{S}_n$ be the sum of n i.i.d. copies of $\tilde{X}$. Then

$$\begin{aligned}
&P\Big(\Big|\frac{S_n}{n}-x\Big|<\delta\Big)\\
&=\int\cdots\int_{|x_1+\cdots+x_n-nx|<n\delta}\mu(dx_1)\cdots\mu(dx_n)\\
&=\int\cdots\int_{|x_1+\cdots+x_n-nx|<n\delta}e^{-n\hat{\theta}(x_1+\cdots+x_n)+n\Lambda(\hat{\theta})}\tilde{\mu}(dx_1)\cdots\tilde{\mu}(dx_n)\\
&=E\Big(e^{-n\hat{\theta}\tilde{S}_n+n\Lambda(\hat{\theta})}1_{|\tilde{S}_n/n-x|<\delta}\Big)\\
&\geq E\Big(e^{-n(\hat{\theta}x-\Lambda(\hat{\theta})+|\hat{\theta}|\delta)}1_{|\tilde{S}_n/n-x|<\delta}\Big)\\
&=e^{-n(\hat{\theta}x-\Lambda(\hat{\theta})+|\hat{\theta}|\delta)}P\Big(\Big|\frac{\tilde{S}_n}{n}-x\Big|<\delta\Big).
\end{aligned}$$

By the weak law of large numbers, and the fact that our tilted distribution has mean x, the term $P(\cdot)$ tends to 1 as $n\to\infty$. Taking logarithms and then lim inf,

$$\begin{aligned}
\liminf_{n\to\infty}\frac{1}{n}\log P\Big(\frac{S_n}{n}\in G\Big)&\geq\liminf_{n\to\infty}\frac{1}{n}\log P\Big(\Big|\frac{S_n}{n}-x\Big|<\delta\Big)\\
&\geq-\big(\hat{\theta}x-\Lambda(\hat{\theta})+|\hat{\theta}|\delta\big).
\end{aligned}$$

But δ was arbitrarily small, so

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

The case when the infimum of Λ is attained at a boundary point of its effective domain involves some additional technicalities. It can be handled by considering the truncated random variables $X\wedge n$, the effective domain of whose cumulant generating function is the entire real line, and letting n go to infinity; see Dembo and Zeitouni [25] for details.

Goodness and convexity of rate function. It only remains to verify that Λ^* is a good convex rate function. We have already established that it is a convex rate function in Lemma 2.6. Choosing any $\theta>0$ such that $\Lambda(\theta)<\infty$ we see that $\Lambda^*(x)\geq\theta x-\Lambda(\theta)$ and so $\Lambda^*(x)\to\infty$ as $x\to\infty$. Similarly, by choosing a $\theta<0$ for which Λ is finite, $\Lambda^*(x)\to\infty$ as $x\to-\infty$. Therefore any level set $\{x:\Lambda^*(x)\leq\alpha\}$ is bounded; by lower-semicontinuity it is closed; hence it is compact. So Λ^* is good. □

Some remarks on the proof.

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

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

holds for all closed intervals $F\subseteq\mathbb{R}$ and all n, not just on a logarithmic scale in the limit as $n\to\infty$. 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 $n\to\infty$.

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 $(X^{\lfloor\alpha N\rfloor}/N,\ N\in\mathbb{N})$ satisfies a large deviations principle in $\mathbb{R}$ with rate function $J(x)=\alpha I(x/\alpha)$.

Hint: prove the upper bound for closed half-spaces, then extend to general closed sets; prove the lower bound for small open balls, then extend to general open sets. ◇

2.7 Sanov's Theorem for Finite Alphabets

We shall now use Cramér's theorem in $\mathbb{R}^d$ to derive an LDP for empirical distributions.

Let $(X_n,\ n \in \mathbb{N})$ be an i.i.d. sequence of random variables, taking values in a finite set A consisting of d elements. Let μ denote the probability law of X_1. We assume without loss of generality that $\mu(a) = P(X_1 = a) > 0$ for all $a \in A$, by restricting A to be the set of x for which this is true. The empirical distribution L_n of $X_1, \dots, X_n$ is a probability measure on A defined by

$$L_n(a) = \frac{1}{n} \sum_{i=1}^{n} 1[X_i = a], \quad a \in A.$$

Let $M_1(A)$ denote the space of probability measures on A. Observe that L_n is essentially the sample mean of n i.i.d. random variables in $\mathbb{R}^d$, and so Cramér's theorem can tell us about it, as follows. (The following result is a simple version of Sanov's theorem. For a more general version, see Chapter 4.)

Theorem 2.9 *The sequence of random variables* $(L_n,\ n \in \mathbb{N})$*, satisfies an LDP in* $\mathbb{R}^d$*, with the good convex rate function*

$$I(\nu) = \begin{cases} H(\nu|\mu) & \textit{if } \nu \in M_1(A), \\ +\infty & \textit{otherwise} \end{cases}$$

where $H(\nu|\mu)$*, the Kullback-Leibler divergence of* ν *with respect to* μ*, is given by*

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

with the convention that $0 \log 0 = 0$.

(We have stated the LDP for L_n in $\mathbb{R}^d$ as that is what we get by applying Cramér's theorem. It makes more sense to think of L_n as lying in $M_1(A)$, as we will note in Lemma 2.10.)

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

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

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

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

From (2.10), Λ is differentiable, with gradient

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

(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 $\nu^k \to \nu$ with $\nu^k(a) > 0$ for all $a \in A$. By lower-semicontinuity of Λ^*,

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

(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

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

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 $\to \infty$ as $k \to \infty$.

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

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

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

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

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

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., $P\big(L_n \in M_1(A)\big) = 1$ 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. □

Example 2.12

Let A be a finite subset of $\mathbb{R}$, and as usual let $S_n = X_1 + \cdots + X_n$, where the X_i are i.i.d. random variables taking values in A. What is the most likely value of L_n conditional on $S_n/n = \bar{x}$? In other words, what is the most likely way for $S_n \approx n\bar{x}$ to happen?

Fix $\delta > 0$ and let M denote the set of all probability distributions on A whose mean is in $[\bar{x} - \delta, \bar{x} + \delta]$. Since $H(\cdot|\mu)$ is continuous on $M_1(A)$, its infimum on the interior and closure of M are the same, and so

$$\lim_{n\to\infty} \frac{1}{n} \log P(L_n \in M) = - \inf_{\lambda \in M} H(\lambda|\mu).$$

Now M is a closed convex set, so by convexity of $H(\cdot|\mu)$ there is a unique $\nu \in M$ at which $H(\cdot|\mu)$ is minimised. Let $B(\nu, \epsilon)$ denote the open ball in $M_1(A)$ with centre ν and radius ϵ. Its complement $B(\nu, \epsilon)^c$ is a closed set. Let the infimum of $H(\cdot|\mu)$ on $M \cap B(\nu, \epsilon)^c$ be attained at ν'. By strict convexity, $H(\nu'|\mu) > H(\nu|\mu)$, and using the large deviation upper bound,

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log P(L_n \notin B(\nu, \epsilon)|L_n \in M) \\
&\qquad = \limsup_{n\to\infty} \frac{1}{n} \log P(L_n \in M \cap B(\nu, \epsilon)^c) - \lim_{n\to\infty} \frac{1}{n} \log P(L_n \in M) \\
&\qquad \leq -\big(H(\nu'|\mu) - H(\nu|\mu)\big) < 0.
\end{aligned}$$

Thus, conditional on $L_n \in M$ (i.e. on $|(S_n/n) - x| \leq \delta$), the probability that L_n is outside an arbitrarily small neighbourhood of ν decays to zero at some strictly positive exponential rate. In other words, if the sample mean is close to x, then the empirical distribution is close to the probability measure ν that has minimum relative entropy with respect to μ among all distributions whose mean is x. ◇

Exercise 2.13
Suppose we roll a fair die ten thousand times, and observe that the mean value of the outcome is 3.8. How many sixes did we roll? ◇

Exercise 2.14
Let μ be a probability measure on a finite subset A of $\mathbb{R}$. Show that the distribution ν that minimizes $H(\nu|\mu)$ subject to the mean, $\sum_{a\in A} a\nu(a)$, being equal to $\bar{a}$ corresponds to an exponential tilting of μ. ◇

2.8 A Generalisation of Cramér's Theorem

Cramér's theorem generalizes far beyond the realm of sums of independent identically distributed random variables. In Theorem 2.8, S_n was the sum of n independent identically distributed random variables, with common cumulant generating function $\Lambda(\theta)$, and so

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

In the standard generalisation, S_n is any sequence of random variables, and we consider

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E e^{\theta S_n}. \tag{2.12}$$

If this limit exists and is well-behaved, then the statement of the theorem still holds. We will use this result in Chapter 3, to prove an LDP for a queue with a weakly dependent input flow. To state the result properly, we need a definition.

Definition 2.6 (Essential smoothness) *Let f be a function on $\mathbb{R}^d$ taking values in the extended reals $\mathbb{R}^*$. The function f is* essentially smooth *if the interior of its effective domain is non-empty, f is differentiable in the interior of its effective domain, and f is* steep, *namely, for any sequence θ_n which converges to a boundary point of the effective domain of f, $\lim_{n\to\infty} |\nabla f(\theta_n)| = +\infty$.*

Theorem 2.11 (Generalized Cramér's Theorem) *If the limit* (2.12) *exists for each $\theta \in \mathbb{R}^d$ as an extended real number, and if $\Lambda(\theta)$ is finite in a neighbourhood of $\theta = 0$ and essentially smooth and lower-semicontinuous, then the sequence of random variables S_n/n satisfies the LDP in $\mathbb{R}^d$ with good convex rate function Λ^*.*

Some remarks.

i. This particular generalization of Cramér's theorem is presented in the book of Dembo and Zeitouni [25], and referred to as the Gärtner-Ellis theorem.

ii. The existence of a non-trivial limit $\Lambda(\theta)$ is essentially a mixing condition, saying that autocorrelations decay sufficiently fast.

iii. The essential smoothness condition is stronger than necessary. For example, in one dimension, a sufficient condition is that the effective domain of Λ^* be contained in $\Lambda'(\mathbb{R})$.

iv. Λ^* is non-negative and, by Lemma 2.4, lower-semicontinuous; hence it is a rate function. By the same lemma, it is also convex.

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 $X_n = f(\xi_n)$, $S_n = X_1 + \ldots + X_n$. 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

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

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,

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

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 $a_1, \ldots, a_r$ be given constants, and consider the recursion

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

where the ε_t are independent standard normal random variables. If all roots of the characteristic equation, $1 - \sum_{k=1}^{r} a_k z^{-k} = 0$, lie strictly within the unit circle in the complex plane, then the recursion is stable and has a unique stationary solution. The solution $(X_t, t \in \mathbb{Z})$ is a stationary zero mean Gaussian process whose covariance structure is most easily described through its Fourier transform, which is called the power spectral density of the process. The power spectrum is defined as

$$\mathcal{S}_X(\omega) = \sum_{t=-\infty}^{\infty} \mathrm{Cov}(X_0, X_t) e^{i\omega t} \quad \text{for } \omega \in \mathbb{R}.$$

As the X_t have mean zero, $\mathrm{Cov}(X_0, X_t) = E(X_0 X_t)$. Let $A(\omega) = 1 - \sum_{k=1}^{r} a_k e^{i\omega k}$. It can be shown that $\mathcal{S}_X(\omega) = |A(\omega)|^2 = A(\omega)A(-\omega)$.

Let $S_n = X_1 + \ldots + X_n$. Define the process

$$R_m = \sum_{t=m+1}^{m+n} X_t \quad \text{for } m \in \mathbb{Z},$$

so that $S_n = R_0$. Now, the sequence $R_m, m \in \mathbb{Z}$ is obtained by the convolution of the sequence X_m and the sequence h_m defined as $h_m = 1$ for $m \in \{-n, \ldots, -1\}$, and $h_m = 0$ otherwise. Hence, for fixed n, R_m is a stationary, zero mean Gaussian process with power spectral density

$$\mathcal{S}_R(\omega) = \mathcal{S}_X(\omega) \frac{\sin^2(n\omega/2)}{\sin^2(\omega/2)}.$$

By Parseval's theorem, the variance of $S_n = R_0$ is given by

$$\begin{aligned}
\mathrm{Var}(S_n) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{S}_R(\omega) d\omega \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{S}_X(\omega) \frac{\sin^2(n\omega/2)}{\sin^2(\omega/2)} d\omega \\
&= n\mathcal{S}_X(0) + \int_{-\pi}^{\pi} \frac{A(\omega)A(-\omega) - A(0)^2}{2\sin^2(\omega/2)} \Big(1 - \cos(n\omega)\Big) d\omega.
\end{aligned}$$

We have used the fact that $\int_{-\pi}^{\pi} \sin^2(n\omega/2)/\sin^2(\omega/2) d\omega = 2\pi n$ to obtain the last equality above. Now, the function

$$f(\omega) = \begin{cases} \big(A(\omega)A(-\omega) - A(0)^2\big)/\big(2\sin^2(\omega/2)\big) & \text{if } \omega \neq 0 \\ 2A'(0)^2 & \text{if } \omega = 0 \end{cases}$$

is continuous, hence bounded, on the compact interval $[-\pi, \pi]$. Consequently, $\int_{-\pi}^{\pi} f(\omega)(1-\cos(n\omega))d\omega$ is bounded in absolute value by a constant that does not depend on n. Therefore

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{Var}(S_n) = \mathcal{S}_X(0) = A(0)^2. \tag{2.13}$$

Since S_n is Gaussian with zero mean, we also have

$$\log E(e^{\theta S_n}) = \tfrac{1}{2}\theta^2 \operatorname{Var}(S_n).$$

Hence, by (2.13),

$$\lim_{n\to\infty} \frac{1}{n} \log E(e^{\theta S_n}) = \tfrac{1}{2}\theta^2 A(0)^2.$$

This is a quadratic function of θ, so it is finite and differentiable on all of $\mathbb{R}$. Hence, S_n/n satisfies an LDP by Theorem 2.11, with rate function $I(x) = \frac{1}{2}x^2/A(0)^2$. ◇

Observe that different Gaussian processes having the same power spectral density at zero have the same limiting cumulant generating function and the same rate function.

The essential property we required of the Gaussian process above was that its power spectral density be finite and differentiable on $[-\pi, \pi]$. This basically requires that the correlations decay sufficiently fast. In Chapter 8, we shall encounter examples of Gaussian processes for which this isn't true, and the spectrum has a singularity at zero. It is still possibly to use large deviation theory, but it requires a different scaling in n.

Exercise 2.17

Let $(Y_n,\ n \in \mathbb{N})$ be an irreducible Markov chain on a finite state space E, and suppose that, conditional on $Y_n = i$, X_n is Poisson with mean λ_i, where λ_i, $i \in E$, are given constants. For $S_n = X_1 + \ldots + X_n$, show that S_n/n satisfies an LDP in $\mathbb{R}$, and compute the rate function. ◇

Chapter 3

More on the Single Server Queue

In this chapter we will take further the style of argument of the first chapter, for example to queues with long-range dependent input, and give more examples. We will need some of the more advanced large deviations theory of Chapter 2 to prove these results.

> *Note.* It is easy to get lost in the details of the calculations: this is because our present style of argument is crude and direct. In the following chapters we will come to a more elegant approach, using a tool from large deviations called the contraction principle. That approach is however more abstract, and it is good to see how far we can go with direct methods.

3.1 Queues with Correlated Inputs

This section generalizes the results in Section 1.3. Recall the setup: consider a queue with constant service rate C and arrival process $(A_t,\ t \in \mathbb{Z})$, A_t being the amount of work arriving at time t. The queue size at time 0 is

$$Q = \sup_{t \geq 0} S_t - Ct$$

where $S_t = A_0 + \cdots + A_{-t+1}$ and $S_0 = 0$. In that section we assumed that the A_t were independent and identically distributed. Now, we shall weaken this assumption, using the generalized version of Cramér's theorem, Theorem 2.11.

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 47–55, 2004.

Theorem 3.1 *Let $(A_t, t \in \mathbb{Z})$ be a stationary random process, with $EA_0 < C$. Let*

$$\Lambda_t(\theta) = \log Ee^{\theta S_t}.$$

Suppose that the limit

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{1}{t}\Lambda_t(\theta)$$

exists in $\mathbb{R}^*$ for each $\theta \in \mathbb{R}$, and that it is essentially smooth, and finite in a neighbourhood of $\theta = 0$; and that $\Lambda_t(\theta)$ is finite for all t whenever $\Lambda(\theta) < \theta C$. Then for $q > 0$,

$$\lim_{l\to\infty} \frac{1}{l} \log P(Q > lq) = -\inf_{t\in\mathbb{R}^+} t\Lambda^*(C + q/t).$$

Some remarks.

i. This theorem says that $S_t - Ct$ is effectively a simple random walk with negative drift, in that

$$P(\sup_t S_t - Ct > q_1 + q_2) \approx P(\sup_t S_t - Ct > q_1)P(\sup_t S_t - Ct > q_2)$$

for large q_1 and q_2. In other words, the (weak) dependence of the A_t is invisible at the macroscopic scale (though it does contribute to the value of $I(q)$).

ii. In stating the theorem, we have implied that $\Lambda(\theta) < \theta C$ for some $\theta > 0$. This is a consequence of Lemma 3.2 and the stability assumption that $EA_0 < C$.

iii. This theorem has appeared in the literature in various more or less equivalent forms. See for example Báartfai [5], Glynn and Whitt [48] and Chang [12].

Proof. The proof is very much like that of Theorem 1.4. The lower bound is proved in exactly the same way as Lemma 1.6: simply replace the appeal to Cramér's theorem with an appeal to the generalized version. The upper bound (the analogue of Lemma 1.5) and the result about rate functions (the analogue of Lemma 1.7) both need a little more work. They are proved in lemmas 3.3 and 3.4. □

Lemma 3.2 *Under the assumptions of Theorem 3.1, $\Lambda'(0) = EA_0$.*

Lemma 3.3 *Under the assumptions of Theorem 3.1,*

$$\limsup_{l\to\infty} \frac{1}{l} \log P(Q > lq) \le -q \sup\{\theta : \Lambda(\theta) < \theta C\}.$$

Lemma 3.4 *Under the assumptions of Theorem 3.1,*

$$I(q) = \inf_{t\in\mathbb{R}^+} t\Lambda^*(C + q/t) \tag{3.1}$$

$$= \inf_{t\in\mathbb{R}^+} \sup_{\theta\geq 0} \theta(q + Ct) - t\Lambda(\theta) \tag{3.2}$$

$$= q \sup\{\theta > 0 : \Lambda(\theta) < \theta C\}. \tag{3.3}$$

Proof of Lemma 3.2 Since $\Lambda(\cdot)$ is finite in a neighbourhood of the origin, so are all $t^{-1}\Lambda_t(\cdot)$ for t sufficiently large. Hence by Lemma 2.3 they are convex, and differentiable in a neighbourhood of the origin with $t^{-1}\Lambda_t'(0) = \mu$ where $\mu = EA_0$. Now by Lemma 1.12 the pointwise limit $\Lambda(\cdot)$ is convex; and, since it is assumed to be differentiable at the origin, $\Lambda'(0) = \mu$. □

Proof of Lemma 3.3 By expanding the definition of Q and using Chernoff's bound,

$$P(Q > lq) \leq e^{-\theta lq} \sum_{t\geq 0} e^{\Lambda_t(\theta)-\theta Ct} \tag{3.4}$$

for any $\theta > 0$. Pick some $\theta > 0$ such that $\Lambda(\theta) < \theta C$. (There exists such a θ—since by Lemma 3.2 $\Lambda'(0) = \mu$ and by the stability assumption $\mu < C$.) Choose $\varepsilon > 0$ such that $\Lambda(\theta) < \theta(C - 2\varepsilon)$. Since $\Lambda_t(\theta)/t \to \Lambda(\theta)$, there exists t_0 such that for $t > t_0$

$$\Lambda_t(\theta) < t\big(\Lambda(\theta) + \varepsilon\theta\big)$$

and hence

$$(3.4) < e^{-\theta lq}\Big(\sum_{t\leq t_0} e^{\Lambda_t(\theta)-\theta Ct} + \sum_{t>t_0} e^{-\varepsilon\theta t}\Big).$$

We have assumed that $\Lambda_t(\theta)$ is finite, so the first sum is finite; the second sum is clearly finite. Hence

$$\limsup_{l\to\infty} \frac{1}{l} \log P(Q > lq) \leq -\theta q.$$

Take the infimum over $\theta > 0$ such that $\Lambda(\theta) < \theta C$ to obtain the result. □

Proof of Lemma 3.4 The proof that (3.1) = (3.2) is similar to that in Lemma 1.7. The only part that needs to change is the appeal to Lemma 2.6, concerning properties of Λ^* when Λ is cumulant generating function. It should be replaced by an appeal to Lemma 2.7, which concerns the properties of Λ^* when Λ is merely akin to a cumulant generating function. That lemma requires Λ to be convex, to take value 0 at the origin, and to be differentiable at the origin with gradient less than C. The first follow from the fact that Λ is the limit of scaled cumulant generating functions, the last follows from Lemma 3.2.

The proof that (3.2) $\geq$ (3.3) is exactly the same as in Lemma 1.7. It remains to show that (3.2) $\leq$ (3.3). Let $\theta^* = \sup\{\theta > 0 : \Lambda(\theta) < \theta C\}$. If $\theta^* = \infty$, there is nothing to prove. If θ^* is in the interior of the effective domain of Λ, the proof of Lemma 1.7 still holds. (The proof uses the fact that $\Lambda(\theta)$ is convex, which is true since it is the limit of convex functions; and also the fact that $\Lambda(\theta)$ is differentiable at θ^*, which is true because θ^* is in the interior of the effective domain.)

It only remains to consider the case where θ^* is on the boundary of the effective domain. Let $\theta_n \uparrow \theta^*$, with θ_n in $(0, \theta^*)$. Considering the tangents at θ_n we see that

$$\Lambda(\theta) \geq \Lambda(\theta_n) + (\theta - \theta_n)\Lambda'(\theta_n).$$

Using the same argument as in Lemma 1.7,

$$(3.2) \leq q\theta_n \frac{\Lambda'(\theta_n) - \Lambda(\theta_n)/\theta_n}{\Lambda'(\theta_n) - C}.$$

As $n \to \infty$, $\theta_n \to \theta^*$. Also, by convexity of $\Lambda(\cdot)$, $\Lambda(\theta_n) \geq \theta_n \Lambda'(0)$ and so $\Lambda(\theta_n)/\theta_n$ is bounded below. Moreover, using the convexity of Λ once more,

$$\Lambda(\theta_n) \leq \Bigl(1 - \frac{\theta_n}{\theta^*}\Bigr)\Lambda(0) + \frac{\theta_n}{\theta^*}\Lambda(\theta^*) \leq \theta_n C$$

and so $\Lambda(\theta_n)/\theta_n$ is bounded above. Finally, by the assumption of steepness, $\Lambda'(\theta_n) \to \infty$. Thus, taking the limit as $n \to \infty$ we get

$$(3.2) \leq \theta^* q = (3.3)$$

as required. □

Example 3.1
Let $(A_t,\ t \in \mathbb{Z})$ be a stationary autoregressive process of degree 1. That is, $A_t = \mu + X_t$ where

$$X_t = aX_{t-1} + \sqrt{1 - a^2}\sigma\varepsilon_t$$

and $|a| < 1$ and the ε_t are i.i.d. standard normal random variables (i.e. with mean 0 and variance 1). With these parameters, $EA_t = \mu$ and $\operatorname{Var} A_t = \sigma^2$.

Note. We have given a recursion without specifying the initial condition. It is clear what is meant; to be precise we could let $X_t = \sqrt{1-a^2}\sum_{i\geq 0} a^t \sigma \varepsilon_{t-i}$. This satisfies the recursion and is stationary. However, in the calculations that follow, we will only be concerned with $X(-t,0]$, which by stationarity has the same distribution as $X(0,t]$, so we could just as well set $X_0 \sim \text{Normal}(0,\sigma^2)$ and define $X|_{(0,t]}$ from this.

What is $\Lambda(\theta)$? Clearly $ES_t = \mu t$, and $\text{Cov}(A_0, A_t) = a^t\sigma^2$, giving

$$\text{Var}\, S_t = \sigma^2 \sum_{1\leq i,j\leq t} a^{|i-j|} = \frac{\sigma^2}{(1-a)^2}\Big(t(1-a^2) - 2a(1-a^t)\Big).$$

And since S_t is normal,

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

where $\sigma_t^2 = \text{Var}\, S_t$. Dividing by t and taking the limit,

$$\Lambda(\theta) = \theta\mu + \frac{\theta^2}{2}\Big(\frac{1+a}{1-a}\Big).$$

The rate function for queue size is

$$\begin{aligned} I(q) &= q \sup\{\theta > 0 : \Lambda(\theta) < \theta C\} \\ &= 2q(\mu - C)\frac{1-a}{1+a}. \end{aligned}$$

Note that when $a \to 1$ this becomes very small, meaning that the probability of large queues is high; and when $a \to -1$ this becomes very large, meaning that the probability of large queues is small. ◇

The following is a useful trick for computing rate functions for on/off sources. It follows from the time-change formula of Russell [89].

Exercise 3.2
Let $(R_k,\ k\in\mathbb{N})$ and $(T_k,\ k\in\mathbb{N})$, be a pair of i.i.d. sequences taking values in $\mathbb{N}$. Define another sequence $(X_n,\ n\in\mathbb{N})$ by setting $X_n = 1$ for $n \leq R_1$, $X_n = 0$ for $R_1 < n \leq R_1 + T_1$, $X_n = 1$ for $R_1 + T_1 < n \leq R_1 + T_1 + R_2$, and so on. As before, set $S_n = X_1 + \cdots + X_n$.

Let the cumulant moment generating functions be $\Lambda_R(\theta) = \log Ee^{\theta R_1}$ and $\Lambda_T(\theta) = \log Ee^{\theta T_1}$. If these are both finite in a neighbourhood of the origin, it can be shown that S_n/n satisfies an LDP with rate function

$$I(x) = \inf_{a>0} a\Lambda_R^*\Big(\frac{x}{a}\Big) + a\Lambda_T^*\Big(\frac{1-x}{a}\Big). \tag{3.5}$$

Give a heuristic explanation of why this is the right rate function.

Use (3.5) to find the rate function associated with S_n/n when X_n is a Markov chain on $\{0, h\}$ (cf. Exercise 1.9.) Use this to calculate the rate function for queue size.

See Exercise 4.10 for an extension of this example. ◇

3.2 Queues with Many Sources and Power-Law Source Scalings

Recall the setup of Section 1.4: let Q^N be the queue size in a queue serving N sources at constant service rate NC. Let $A_t^{(i)}$ be the amount of work arriving from source i at time t. Assume that for each i, $(A_t^{(i)},\ t \in \mathbb{Z})$ is a stationary sequence of random variables, and that these sequences are independent of each other, and identically distributed. Let S_t^1 be the amount of work produced by a typical source over time period t.

The theorem in Section 1.4 already allowed the $(A_t^{(i)},\ t \in \mathbb{Z})$ to have correlations in time. (That was why we needed our tail-controlling assumption.)

One way to extend the theorem would be to use the generalized version of Cramér's theorem, to account for queues where the many sources have some correlations between them. We will not go down that route. It will be taken care of automatically when we come to the more abstract formulation in Chapter 7.

Instead, here is a different sort of generalization. Our tail-controlling assumption was a restriction on the correlation structure of a single source: if $\Lambda_t(\theta) = \log Ee^{\theta S_t^1}$, we assumed that the limit

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{1}{t}\Lambda_t(\theta)$$

exists, and is finite and differentiable in a neighbourhood of the origin. This does not allow for sources with certain interesting scalings, notably fractional Brownian motion described below in Example 3.3. For such flows we need to introduce a scaling function v_t, described in the following theorem, which depends on how correlations decay in time. For the sources we have described so far $v_t = t$; for sources like fractional Brownian motion, v_t is a power of t, and so we say they have *power-law scalings* (or that they exhibit *long-range dependence*).

Theorem 3.5 *Suppose that for some sequence* $(v_t,\ t \in \mathbb{N})$ *taking values in* $\mathbb{R}^+$, *with* $v_t/\log t \to \infty$, *the limit*

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{\Lambda_t(\theta v_t/t)}{v_t}$$

exists and is finite and differentiable in a neighbourhood of the origin. Then the conclusion of Theorem 1.8 still holds, i.e. if $ES_1^1 < C$ *then*

$$\begin{aligned} -I(q+) &\leq \liminf_{N\to\infty} \frac{1}{N} \log P(Q^N > Nq) \\ &\leq \limsup_{N\to\infty} \frac{1}{N} \log P(Q^N > Nq) \leq -I(q) \end{aligned}$$

where

$$I(q) = \inf_{t\in\mathbb{N}} \Lambda_t^*(q + Ct).$$

Various extensions of this result have been proved. Likhanov and Mazumdar [59] prove a tighter bound under weaker conditions. Botvich and Duffield [10] prove it in continuous time, and Mandjes and Kim [69] prove it in continuous time under weaker conditions.

Proof. The only part of the proof that needs to be changed is the proof of Lemma 1.11. Choose θ as in that lemma. Let $\theta_t = \theta v_t/t$, and use θ_t in applying Chernoff's bound. Thus

$$\sum_{t>t_0} P(S_t^N/N > q + Ct) \leq \sum_{t>t_0} e^{-N\left(\theta_t(q+Ct)-\Lambda_t(\theta_t)\right)}$$

and using $\Lambda_t(\theta v_t/t)/v_t \to \Lambda(\theta)$ in the same way as before,

$$\leq \sum_{t>t_0} e^{-N\theta\delta v_t} \quad \text{for some } \delta > 0.$$

Since $v_t/\log t \to \infty$, given K we can choose a t_0 such that for $t > t_0$, $v_t > K\log t$. This makes

$$\begin{aligned} \limsup_{N\to\infty} \frac{1}{N} \log \sum_{t>t_0} e^{-N\theta\delta v_t} &< \limsup_{N\to\infty} \frac{1}{N} \log \sum_{t>t_0} t^{-KN\theta\delta} \\ &= K\theta\delta \limsup_{M\to\infty} \frac{1}{M} \log \sum_{t>t_0} t^{-M}. \end{aligned} \tag{3.6}$$

To bound the sum, note that (for $M \geq 2$)

$$\begin{aligned}\sum_{t>t_0} t^{-M} &= \frac{1}{(t_0+1)^M}\left(1+\Big(\frac{t_0+1}{t_0+2}\Big)^M+\Big(\frac{t_0+1}{t_0+2}\Big)^M+\cdots\right)\\ &\leq \frac{1}{(t_0+1)^M}\left(1+\Big(\frac{t_0+1}{t_0+2}\Big)^2+\Big(\frac{t_0+1}{t_0+3}\Big)^2+\cdots\right)\\ &\leq \frac{\pi^2/6}{(t_0+1)^{M-2}}\end{aligned}$$

so that

$$\limsup_{N\to\infty}\frac{1}{N}\log\sum_{t>t_0} t^{-\alpha M} \leq -\alpha\log(t_0+1) \quad \text{for any } \alpha>0, \tag{3.7}$$

and hence that (for $t_0 \geq 2$)

$$(3.6) \leq -K\theta\delta\log(t_0+1) < -K\theta\delta.$$

In other words, given an arbitrarily large K there exists a t_0 such that

$$\limsup_{N\to\infty}\frac{1}{N}\log\sum_{t>t_0} P(S_t^N/N \geq q+Ct) < -K\theta\delta,$$

and hence

$$\limsup_{N\to\infty}\frac{1}{N}\log\sum_{t>t_0} P(S_t^N/N \geq q+Ct) \to -\infty \ \text{ as } t_0\to\infty$$

as required. □

Exercise 3.3
Let $(S_t^1,\ t\in\mathbb{N})$ be a discrete-time Gaussian process with marginal distributions given by

$$S_t^1 \sim \text{Normal}(\mu t, \sigma^2 t^{2H})$$

for some $H \in [\frac{1}{2}, 1)$.

> *Note.* Such a process exists. In fact there is a continuous-time process with the same marginal distributions, with stationary increments (i.e. $S_{t+u}^1 - S_t^1$ has the same distribution regardless of the value of $u \geq 0$), and with continuous sample paths, which is known as fractional Brownian motion. For more on this see Chapter 8.

Show that $\Lambda_t(\theta)/t$ does not converge as $t \to \infty$. Show that nonetheless this traffic source does satisfy Theorem 3.5, and show that

$$I(q) = \inf_t \frac{\big(q + (C-\mu)t\big)^2}{2\sigma^2 t^{2H}} \approx \frac{q^{2(1-H)}(C-\mu)^{2H}}{H^{2H}(1-H)^{2(1-H)}},$$

and that the optimal t is

$$t \approx \frac{q}{C-\mu}\frac{H}{1-H}.$$

(Hint: Use scaling function $v_t = t^{2(1-H)}$.) ◇

3.3 Queues with Large Buffers and Power-Law Source Scalings

The last section showed an indifference result: it doesn't really make much difference whether the sources have power-law scalings, as far as the many-sources limit is concerned—the probability of large queues still decays exponentially in the number of sources.

A natural question to ask is: can we blend the techniques in Sections 3.1 and 3.2? In a queue with a single traffic source, does long-range dependence of the source make any difference to the queue size (as far as the large-buffer limit is concerned)? The answers are yes and yes. In cases where the limit

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{1}{v_t} \log E e^{\theta S_t v_t / t}$$

exists and is well-behaved, the log-probabilities $\log P(S_t/t)$ are normalized by v_t. This leads to an analogue of Theorem 3.1. For example, if $v_t = t^{2(1-H)}$ for some $H \in [\frac{1}{2}, 1)$ then under suitable conditions

$$\frac{1}{q^{2(1-H)}} \log P(Q > q) \approx -I.$$

It is possible to state and prove this result using the techniques in this chapter. But since this sort of scaling has a number of interesting consequences, we will postpone a discussion of it to Chapter 8.

Chapter 4

Introduction to Abstract Large Deviations

We are now ready to explain what is meant by a large deviations principle, in a more abstract setting. To develop intuition, think back to the large deviations results for $\mathbb{R}^d$ in Chapter 2. One space we will be interested in in later chapters is the space of continuous functions $\mathbb{R}_0^+ \to \mathbb{R}$, representing the space of input flows at a queue.

This chapter will also give some general large deviations results, paying particular attention to the *contraction principle*, a result which will be using heavily in later chapters.

4.1 Topology and Metric Spaces

It takes a hefty amount of basic topology to formulate large deviations in an abstract setting, so we gather together here some definitions as a handy reference.

Let $\mathcal{X}$ be a set.

A family of subsets τ of $\mathcal{X}$ is called a *topology* if

- $\emptyset \in \tau$ and $\mathcal{X} \in \tau$
- The union of any family of sets in τ, is in τ
- The intersection of a finite number of sets in τ, is in τ

The elements of τ are called the *open sets*; their complements are the *closed sets*. The pair $(\mathcal{X}, \tau)$ is called a *topological space*. We often write it $\mathcal{X}$, when the topology is clear from the context.

A function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *metric* if for all $x, y, z \in \mathcal{X}$,

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 57–76, 2004.

- $d(x, y) \geq 0$ with equality iff $x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$

This induces a natural topology, the smallest topology containing all the balls $B(x, \varepsilon) = \{y : d(x, y) < \varepsilon\}$.

A subset σ of τ is called a *basis* of the topology τ if every set in τ is the union of sets in σ. A topological space is *separable* if it has a countable basis of open sets.

The topology *induced* on any subset $\mathcal{Y} \subset \mathcal{X}$ is $\{A \cap \mathcal{Y} : A \in \tau\}$.

A topology σ is *finer* than a topology τ on the same set $\mathcal{X}$, if σ contains all the sets in τ; then τ is *coarser* than σ.

Let A be a subset of $\mathcal{X}$. A is an *open neighbourhood* of x if A is open and $x \in A$. A is a *neighbourhood* of x if it contains an open neighbourhood of x. The *interior* of A, A°, is the union of all open subsets of A. It is the largest open set contained in A. The *closure* of A, $\bar{A}$, is the intersection of all closed supersets of A. It is the smallest closed set containing A.

A set A is open iff for all $x \in A$ there exists an open neighbourhood B of x with $B \subset A$.

A set $A \subset \mathcal{X}$ is *dense* if its closure $\bar{A}$ is equal to $\mathcal{X}$.

A metric space is separable if it contains a countable dense set.

A topological space $\mathcal{X}$ is *Hausdorff* if for every $x \in \mathcal{X}$ $\{x\}$ is closed, and for all $x, y \in \mathcal{X}$ there exist open neighbourhoods B_x and B_y of x and y that are disjoint. It is *regular* if in addition for every closed set $\bar{A} \subset \tau$ and point $y \in \tau$ there exist open neighbourhoods $B_{\bar{A}}$ and B_y of $\bar{A}$ and y that are disjoint.

Every metric space is regular.

A sequence $x_1, x_2, \ldots$ *converges* to x if for all neighbourhoods B of x, $x_n \in B$ eventually.

In a metric space, $x_n \to x$ iff $\lim_{n\to\infty} d(x_n, x) = 0$.

A sequence $x_1, x_2, \ldots$ in a metric space is a *Cauchy sequence* if $d(x_m, x_n) \to 0$ as $m \wedge n \to \infty$. A metric space $\mathcal{X}$ is *complete* if every Cauchy sequence converges. A complete separable metric space is called *Polish.*

Let A be a subset of a topological space. An *open cover* is a collection of open sets whose union contains A. A is *compact* if every open cover has a finite subcover.

A closed subset of a compact set is compact. The intersection of a closed set and a compact set is compact. In a Hausdorff space, every compact set is closed.

A subset A of a topological space is *sequentially compact* if every sequence of points in A contains a subsequence which converges to a point in A.

If A is compact then A is sequentially compact. If $\mathcal{X}$ is a metric space, and A is closed and sequentially compact, then A is compact.

Let A be a subset of a metric space. A is closed iff whenever x_n is a sequence in A and $x_n \to x$ then $x \in A$.

Let $\mathcal{X}$ and $\mathcal{Y}$ be topological spaces, and let $f : \mathcal{X} \to \mathcal{Y}$. For $Y \subset \mathcal{Y}$, let $f^{-1}(Y) = \{x \in \mathcal{X} : f(x) \in Y\}$. We say f is *continuous* if whenever Y is open in $\mathcal{Y}$, $f^{-1}(Y)$ is open in $\mathcal{X}$.

If $\mathcal{X}$ and $\mathcal{Y}$ are metric spaces, a function f is continuous iff whenever $x_n \to x$, $f(x_n) \to f(x)$.

Let $f : \mathcal{X} \to \mathcal{Y}$ be continuous, and let $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$. If Y is open then $f^{-1}(Y)$ is open. If Y is closed then $f^{-1}(Y)$ is closed. If X is compact then $f(X)$ is compact.

Let $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$. The *level sets* of f are the sets of the form $\{x : f(x) \leq \alpha\}$ for $\alpha \in \mathbb{R}$. We say f is *lower-semicontinuous* if all level sets are closed.

Let $\mathcal{X}$ be a metric space and let $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$. Then f is lower-semicontinuous iff whenever $x_n \to x$, $\liminf_{n\to\infty} f(x_n) \geq f(x)$.

4.2 Definition of LDP

Let $\mathcal{X}$ be a Hausdorff space, throughout the rest of this chapter.

Definition 4.1 (Rate function) *A function $I : \mathcal{X} \to \mathbb{R}^*$ is called a* rate function *if it is non-negative and lower semicontinuous (defined in Section 4.1). A rate function is said to be* good *if, in addition, its level sets are compact.*

Let $(\mu_n,\ n \in \mathbb{N})$ be a sequence of Borel probability measures on $\mathcal{X}$, and let $\mathcal{B}$ be the Borel σ-algebra.

Definition 4.2 (Large deviations principle) *We say that μ_n satisfies the* large deviation principle *(LDP) on $\mathcal{X}$ with rate function I if I is a* *rate function and, for all $B \in \mathcal{B}$,*

$$\begin{aligned} -\inf_{x \in B^\circ} I(x) &\leq \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(B) \leq -\inf_{\bar{B}} I(x). \end{aligned} \tag{4.1}$$

Some remarks.

i. If X_n is a sequence of random variables with distribution μ_n, we may equivalently say that the sequence X_n satisfies the LDP.

ii. If $\mathcal{X}$ is a space of functions indexed by $\mathbb{R}$ or $\mathbb{N}$, we call the LDP a *sample path LDP.*

iii. We could allow n to take values in $\mathbb{R}^+$ instead of $\mathbb{N}$.

iv. Note that $\inf_{x\in\mathcal{X}} I(x) = 0$ since $\mu_n(\mathcal{X}) = 1$ for all n.

v. The above formulation of the LDP makes sense even if $\mathcal{B}$ is not the Borel σ-algebra on $\mathcal{X}$. The interior B° and the closure $\bar{B}$, and also the lower-semicontinuity of I, are specified by the topology on $\mathcal{X}$, which may be unrelated to the σ-algebra for the probability measures μ_n. To avoid technicalities, in this book we shall always work with Borel probability measures corresponding to the topology in which we are interested in establishing the LDP. In that case, the LDP can equivalently be stated as

$$-\inf_{x\in F} I(x) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(F) \quad \text{for all open sets } F\text{, and} \tag{4.2}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(G) \leq -\inf_{x\in G} I(x) \quad \text{for all closed sets } G. \tag{4.3}$$

vi. (4.2) is called the large deviations lower bound; (4.3) the upper.

vii. The statement of the LDP above may appear rather complicated. Why can't we just require $\lim n^{-1} \log \mu_n(A) = \inf_{x\in A} I(x)$ for all measurable sets A? Because if we did, we would not be able to use the LDP to describe continuous random variables. To see this, let the sequence μ_n consist of measures which are non-atomic, i.e. $\mu_n(\{x\}) = 0$ for all n and x; then it must be that $I(x) = \infty$ for all x; yet this contradicts the requirement that $\inf_{x\in\mathcal{X}} I(x) = 0$. So some topological restrictions are needed, and the LDP is a particularly convenient way of stating them—tight enough to be useful, loose enough to be true. However, this consideration motivates the following definition...

Definition 4.3 *A set $A \subset \mathcal{X}$ is called an I-*continuity set *if*

$$\inf_{x\in A^\circ} I(x) = \inf_{x\in \bar{A}} I(x).$$

For such a set, if it is measurable then $n^{-1}\log\mu_n(A) \to \inf_{x\in A} I(x)$. If $\mathcal{X} = \mathbb{R}$ and I is continuous, then all intervals are I-continuity sets.

> *Note.* Note the formal similarity of the LDP with the definition of weak convergence of probability measures on $\mathbb{R}$: we say that a sequence of probability measures μ_n converges weakly to a probability measure μ if $\limsup_n \mu_n(F) \le \mu(F)$ for all closed sets F and $\liminf_n \mu_n(G) \ge \mu(G)$ for all opens sets G. If a set A is a μ-continuity set, i.e. $\mu(A^\circ) = \mu(\bar{A})$, then $\lim_n \mu_n(A)$ exists and is equal to $\mu(A)$. Likewise in the large deviation setting, if A is a measurable I-continuity set, then $\lim_n n^{-1}\log\mu_n(A)$ exists and is equal to $-\inf_{x\in A} I(x)$. For more on the analogy with weak convergence, see the book of Dupuis and Ellis [37].

Lemma 4.1 *For any rate function I, if A is a compact set then the infimum $\inf_{x\in A} I(x)$ is attained at some $\hat{x} \in A$. If I is a good rate function then the infimum is attained on any closed set.*

The proof is trivial, though it must take account of the fact that I can take value $+\infty$.

The interpretation is that $\hat{x}$ is the most likely way for an event A to occur. It is the largest term (i.e. has the largest value of $I(x)$), and it dominates in $P(X_n \in A)$; the smaller terms do not contribute. This is another instance of the principle of the largest term, which we can make precise as follows.

Lemma 4.2 (Rare events occur in the most likely way) *Suppose X_n satisfies an LDP with good rate function I, and C is a closed set with $\inf_{x\in C} I(x) = k < \infty$. This infimum must be attained; suppose it is attained in C°. Let B be a neighbourhood of $\{x \in C : I(x) = k\}$. Then*

$$P(X_n \notin B | X_n \in C) \to 0.$$

Proof. We can simply estimate

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n}\log P(X_n \notin B | X_n \in C) \\
&= \limsup_{n\to\infty} \frac{1}{n}\big(\log P(X_n \in C \setminus B) - \log P(X_n \in C)\big) \\
&\le \limsup_{n\to\infty} \frac{1}{n}\log P(X_n \in C \setminus B) - \liminf_{n\to\infty} \frac{1}{n}\log P(X_n \in C) \\
&\le -\Big(\inf_{x\in\overline{C\setminus B}} I(x) - \inf_{x\in C^\circ} I(x)\Big).
\end{aligned}$$

The second term is equal to k, since the minimum is attained in C°. If the first term is equal to ∞, we are done. Otherwise, we might as well take B to be an open neighbourhood, so that $C \setminus B$ is closed, implying that the infimum is attained, say at $\hat{x}$. Since $\hat{x} \in C$, $I(x) \geq k$; since $\hat{x} \notin B$, $I(\hat{x}) \neq k$. So we are done. □

Example 4.1
Let X be an exponential random variable with mean λ^{-1}, and let $X_n = n^{-1}X$. We will now show that X_n satisfies an LDP in $\mathbb{R}_0^+$ with good rate function $I(x) = \lambda x$.

First, is I a good rate function? Clearly $I(x) \geq 0$. The level sets $\{x \in \mathbb{R}_0^+ : I(x) \leq \alpha\}$ are just the intervals $[0, \alpha/\lambda]$ which are certainly compact.

Does X_n satisfy the large deviations upper bound (4.3)? If the closed set G is a semi-infinite interval $[x, \infty]$ then $P(X_n \in G)$ is $\exp(-n\lambda x)$ so the upper bound is exact, even without taking limits. If G is a general closed set, let $x = \inf G$. Then $P(X_n \in G)$ is bounded above by $P(X_n \geq x)$, and the upper bound holds, even without taking limits.

Does X_n satisfy the large deviations lower bound (4.2)? Consider first neighbourhoods of points $x \in \mathbb{R}_0^+$. If $x > 0$, let F be the open set $(x-\delta, x+\delta)$ for some $0 < \delta < x$. Then

$$\begin{aligned} P(X_n \in F) &= P\big(X > n(x-\delta)\big) - P(X \geq n(x+\delta)\big) \\ &= e^{-\lambda n(x-\delta)} - e^{-\lambda n(x+\delta)}. \end{aligned}$$

By the principle of the largest term, Lemma 2.2,

$$\liminf_{n\to\infty} \frac{1}{n} \log P(X_n \in F) = -\lambda(x-\delta) \geq -\lambda x.$$

If $x = 0$, let F be the open interval $[0, \delta)$ in $\mathbb{R}_0^+$.

$$P(X_n \in F) = 1 - e^{-n\lambda\delta}$$

and so

$$\liminf_{n\to\infty} \frac{1}{n} \log P(X_n \in F) = 0.$$

Now, if F is any open set, for any point $x \in F$ there is an open interval containing x, and by the above

$$\liminf_{n\to\infty} \frac{1}{n} \log P(X_n \in F) \geq -I(x).$$

Taking the supremum over $x \in F$ gives the result. ◇

Exercise 4.2
Let $(\varepsilon_n,\ n \in \mathbb{N})$ be a sequence of real numbers converging to zero, and define the sequence of 'random' variables $X_n = \varepsilon_n$. Show that X_n satisfies an LDP in $\mathbb{R}$ with good rate function

$$I(x) = \begin{cases} 0 & \text{if } x = 0 \\ \infty & \text{otherwise.} \end{cases}$$

In other words, verify the inequalities (4.3) and (4.2), and verify that I is a good rate function. ◇

Example 4.3 (Schilder's theorem)
Let $(W_t, t \geq 0)$ be a Brownian motion. It can be shown that $(n^{-1/2}W_t,\ 0 \leq t \leq 1)$ satisfies an LDP in the space of continuous functions on $[0,1]$ with good rate function

$$I(f) = \begin{cases} \frac{1}{2}\int_0^1 \dot{f}(t)^2 dt & \text{if } f \text{ is absolutely continuous and } f(0) = 0 \\ \infty & \text{otherwise.} \end{cases}$$

For a proof see Dembo and Zeitouni [25, Theorem 5.2.3]. ◇

In general, it is very tedious to establish an LDP directly by verifying the large deviation lower and upper bounds for all open and closed sets. Fortunately, a number of tools are available that can be applied to a wide variety of problems. To establish the lower bound, it is enough to show that for any $x \in \mathcal{X}$, and any open neighbourhood A of x,

$$\liminf_{n\to\infty} \frac{1}{n} \log P(X_n \in A) \geq -I(x).$$

This is known as 'proving the bound locally'. We saw the technique used in Example 4.1. In establishing the upper bound for compact set, it turns out the principle of the largest term is often very useful. Extending the upper bound from compact sets to all closed sets can be difficult, and relies on problem-specific techniques; see also Exercise 4.8.

In this book, we typically do not establish the LDP from scratch for the random variables that we are interested in, but rely on more indirect methods in general, which we now explain.

4.3 The Contraction Principle

Once we have an LDP for one sequence of random variables, we can effortlessly obtain LDPs for a whole other class of random sequences, namely those that are obtained via continuous transformations. The tool for doing this is the contraction principle.

For example, in queueing applications, we might start with an LDP for the entire arrival process in the space of continuous functions $\mathbb{R}_0^+ \to \mathbb{R}$, show that the queue size function is continuous, and deduce an LDP for queue size.

The term 'effortlessly' is somewhat misleading! It can be quite difficult to establish the continuity of a given function, and to compute the rate function for the resulting LDP. Nevertheless, the contraction principle is a very general and powerful tool, and is the primary technique we use in this book to establish LDPs for quantities of interest in queueing models.

Let $\mathcal{X}$ be a Hausdorff topological space. All measures and functions will be assumed to be Borel-measurable.

Theorem 4.3 (Contraction principle) *Suppose that X_n satisfies an LDP in $\mathcal{X}$ with good rate function I, and that $f : \mathcal{X} \to \mathcal{Y}$ is a continuous map to another Hausdorff space $\mathcal{Y}$. Then $f(X_n)$ satisfies a large deviations principle in $\mathcal{Y}$, with good rate function*

$$J(y) = \inf_{x \in \mathcal{X} : f(x) = y} I(x).$$

Proof. It is simple to show that the large deviations upper and lower bounds hold for $f(X_n)$. Take the upper bound. Let $B \subset \mathcal{Y}$ be closed. Then

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log P\big(f(X_n) \in B\big) \\
&\qquad = \limsup_{n\to\infty} \frac{1}{n} \log P(X_n \in f^{-1}B) \\
&\qquad \le - \inf_{x \in f^{-1}B} I(x) \quad \text{since } f^{-1}B \text{ is closed} \\
&\qquad = - \inf_{y \in B} \inf_{x : f(x) = y} I(x) = \inf_{y \in B} J(y).
\end{aligned}$$

Similarly for B open.

What is harder is to show that J is a good rate function. It's trivial that $J(y) \ge 0$; we need to show that J has compact level sets. Consider a level set:

$$\begin{aligned}\{y : J(y) \le \alpha\} &= \{y : \inf_{x:f(x)=y} I(x) \le \alpha\} \\ &= \{y : \exists x : f(x) = y, I(x) \le \alpha\} \quad \text{using Lemma 4.1} \\ &= f(\{x : I(x) \le \alpha\}) \\ &= f(\text{compact set}) \quad \text{as } I \text{ is good} \\ &= \text{compact set}, \quad \text{as } f \text{ is continuous.}\end{aligned}$$ □

Example 4.4
Let X_n be the average of n independent Normal(μ, σ^2) random variables. By Cramér's theorem, X_n satisfies an LDP with good rate function $I(x) = (2\sigma^2)^{-1}(x-\mu)^2$. So X_n^2 satisfies an LDP with good rate function

$$J(y) = \inf_{x:x^2=y} I(x) = \begin{cases} 0 & \text{if } y < 0 \\ (2\sigma^2)^{-1}(\mu - \sqrt{y})^2 & \text{if } y \ge 0,\ \mu \ge 0 \\ (2\sigma^2)^{-1}(\mu + \sqrt{y})^2 & \text{if } y \ge 0,\ \mu \le 0. \end{cases}$$ ◇

We have to assume the rate function is good, for otherwise the function J may not even be a rate function:

Example 4.5
Let X_n be the average of n independent Cauchy random variables. By Cramér's theorem, X_n satisfies a large deviations principle with rate function $I(x) = 0$. Let $Y_n = e^{X_n}$. Then $J(y) = 0$ if $y > 0$ and ∞ if $y \le 0$. This is not lower-semicontinuous, hence not a rate function. ◇

The contraction principle leads to another instance of the idea that rare events occur in the most likely way. The following lemma follows immediately from Lemma 4.2.

Lemma 4.4 *Under the assumptions of Theorem 4.3: Let D be a closed set in $\mathcal{Y}$. Then, supposing it to be finite, the infimum $\inf_{y\in D} J(y)$ is attained. Suppose it is attained only at $\hat{x}$, $f(x) \in D^\circ$. Then for any neighbourhood B of $\{x = \hat{x}\}$,*

$$P\big(X_n \notin B \;\big|\; f(X_n) \in D\big) \to 0.$$

> *Note.* The standard contraction principle is not always sufficient for our applications, and the next two results describe certain extensions. Their use only becomes clear in the context of the application, so the rest of this section should be omitted on first reading. Perhaps it should even be omitted on subsequent readings!

For the applications in Chapter 9, the quantity we are interested in is 'nearly a continuous function but not quite', in the sense that Y_n, the quantity of interest, is exponentially equivalent (defined in Section 4.4) to $f(X_n)$ for some continuous function f. Then Theorems 4.3 and 4.8 imply the following.

Corollary 4.5 (Approximate contraction principle) *Suppose that X_n satisfies an LDP in $\mathcal{X}$ with good rate function I, and that $f : \mathcal{X} \to \mathcal{Y}$ is a continuous map to a metric space $\mathcal{Y}$. Suppose that $f(X_n)$ is exponentially equivalent to Y_n. Then Y_n satisfies an LDP in $\mathcal{Y}$ with good rate function*

$$J(y) = \inf_{x \in \mathcal{X} : f(x) = y} I(x).$$

Exercise 4.6
Suppose that, in the context of Corollary 4.5, the metric on $\mathcal{Y}$ is d. If $d\big(f(X_n), Y_n\big) \to 0$ uniformly then Y_n is exponentially equivalent to $f(X_n)$, and so the result holds. This is used in Section 5.8.

Prove this directly, without recourse to exponential equivalence. ◇

For the applications in Chapters 7 and 9, the function f is not continuous on the whole space, but only on the subspace where the rate function is finite. The contraction principle can easily be modified:

Theorem 4.6 (Extended contraction principle) *Suppose X_n satisfies an LDP in $\mathcal{X}$ with good rate function I, and that $f : \mathcal{X} \to \mathcal{Y}$ is a map to another Hausdorff space $\mathcal{Y}$. Suppose there exists an open neighbourhood A of the effective domain of I, such that f is continuous on $\bar{A}$. Then $f(X_n)$ satisfies an LDP in $\mathcal{Y}$ with good rate function*

$$J(y) = \inf_{x : f(x) = y} I(x) = \inf_{x \in A : f(x) = y} I(x) = \inf_{x \in \bar{A} : f(x) = y} I(x).$$

Proof. First, the large deviations upper bound. Let $B \subset \mathcal{Y}$ be closed. Then

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log P\big(f(X_n) \in B\big) \\
&\quad = \limsup_{n\to\infty} \frac{1}{n} \Big[P\big(f(X_n) \in B, X_n \in \bar{A}\big) + P\big(f(X_n) \in B, X_n \notin \bar{A}\big) \Big] \\
&\quad = \limsup_{n\to\infty} \frac{1}{n} \log P\big(X_n \in f^{-1}(B) \cap \bar{A}\big) \\
&\qquad \vee \limsup_{n\to\infty} \frac{1}{n} \log P\big(X_n \in f^{-1}(B), X_n \notin \bar{A}\big)
\end{aligned}$$

where the last equality is by the principle of the largest term. Since f is continuous on $\bar{A}$, the set $f^{-1}(B) \cap \bar{A}$ is closed, and so by the LDP for X_n the first term is

$$\leq - \inf_{x \in f^{-1}(B) \cap \bar{A}} I(x) = - \inf_{y \in B} \inf_{x \in \bar{A} : f(x) = y} I(x).$$

The second term is

$$\leq \limsup_{n \to \infty} \frac{1}{n} \log P\big(X_n \notin A\big) \leq - \inf_{x \notin A} I(x) = -\infty.$$

Putting these together, we get the upper bound

$$\limsup_{n \to \infty} \frac{1}{n} \log P\big(f(X_n) \in B\big) \leq - \inf_{y \in B} \inf_{x \in \bar{A} : f(x) = y} I(x).$$

The large deviations lower bound for $f(X_n)$ works similarly: we find that for open B

$$\liminf_{n \to \infty} \frac{1}{n} \log P\big(f(X_n) \in B\big) \geq - \inf_{y \in B} \inf_{x \in A : f(x) = y} I(x).$$

The three expressions for $J(y)$ are all equal. To see this, consider a sequence $x^n \to x$, with $x^n \notin A$. Since A is open, $x \notin A$, and so $I(x) = \infty$. Thus, in seeking

$$\inf_{x : f(x) = y} I(x) \quad \text{or} \quad \inf_{x \in \bar{A} : f(x) = y} I(x)$$

we can restrict attention to $x \in A$.

Finally we need to prove that J is a good rate function. Trivially, $J(y) \geq 0$. So we need to show that the level sets $\{y : J(y) \leq \alpha\}$ are compact for $\alpha \in \mathbb{R}^+$. But this level set is just

$$f\big(\{x \in \bar{A} : I(x) \leq \alpha\}\big).$$

Since I is good it has compact level sets. Also $\bar{A}$ is closed. So this expression is the continuous image of a compact set, so it is compact. □

4.4 Other Useful LDP Results

First, uniqueness of the rate function.

Lemma 4.7 *Let $\mathcal{X}$ be a regular Hausdorff space. If X_n satisfies an LDP in $\mathcal{X}$ with rate function I, and with rate function J, then $I = J$.*

We will only prove it when $\mathcal{X}$ is a metric space.

Proof. Suppose there exists $\hat{x}$ such that $J(\hat{x}) < I(\hat{x})$. Intuitively I says $\hat{x}$ is less likely than does J. Consider a neighbourhood of $\hat{x}$: since I is lower-semicontinuous, for all $\delta > 0$ there exists $\varepsilon > 0$ such that if $d(\ \hat{x}) \leq \varepsilon$ then

$$\begin{aligned} f(x) &\geq \begin{cases} f(\hat{x} - \delta) \text{ if } f(\hat{x}) < \infty \\ 1/\delta \text{ if } f(\hat{x}) = \infty \end{cases} \\ &\geq \bigl(f(\hat{x}) - \delta\bigr) \wedge \frac{1}{\delta}. \end{aligned}$$

Now

$$\begin{aligned} -J(\hat{x}) &\leq - \inf_{x:d(x,\hat{x})<\varepsilon} J(x) \quad \text{since } \hat{x} \text{ lies in that set} \\ &\leq \liminf_{n\to\infty} \frac{1}{n} \log P\bigl(d(X_n, \hat{x}) < \varepsilon\bigr) \quad \text{by the LD lower bound} \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log P\bigl(d(X_n, \hat{x}) \leq \varepsilon\bigr) \quad \text{as this is a bigger event} \\ &\leq - \inf_{x:d(x,\hat{x})\leq\varepsilon} I(x) \quad \text{by the LD upper bound} \\ &\leq -\bigl(I(\hat{x}) - \delta\bigr) \wedge \frac{1}{\delta} \quad \text{as } I \text{ is lower-semicontinuous.} \end{aligned}$$

So for all $\delta > 0$, $J(\hat{x}) \geq (I(\hat{x}) - \delta) \wedge \delta^{-1}$, contradicting $J(\hat{x}) < I(\hat{x})$. □

The next definition and theorem are also concerned with uniqueness. If X_n and Y_n are close (in a suitable sense) and X_n satisfies an LDP with good rate function I, then so does Y_n.

Definition 4.4 (Exponential equivalence) *Let $\mathcal{X}$ be a metric space, with metric d. Let X_n and Y_n be sequences of random variables on $\mathcal{X}$. They are* exponentially equivalent *if for all $\delta > 0$*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\bigl(d(X_n, Y_n) > \delta\bigr) = -\infty.$$

If the probability they differ even by δ decays superexponentially, and large deviations can only pick up exponentially-decaying probabilities, then X_n and Y_n should be indistinguishable in their large deviations properties. This is made precise in the following theorem, a proof of which is given by Dembo and Zeitouni [25].

Theorem 4.8 *If X_n and Y_n are exponentially equivalent, and X_n satisfies an LDP with good rate function I, then so does Y_n.*

Sometimes we apply standard results about large deviations, and find an LDP in a space which is larger than the one we are actually interested in. (We already saw an example of this with regard to Sanov's theorem in Section 2.7. The use of Cramer's theorem naturally yielded an LDP on $\mathbb{R}^{|A|}$, though it was natural to expect that the LDP hold on the smaller space $M_1(A)$ as that is the space in which all the random variables live.) This motivates the following.

Lemma 4.9 *Let $\mathcal{E}$ be a measurable subset of $\mathcal{X}$ such that $P(X_n \in \mathcal{E}) = 1$ for all $n \in \mathbb{N}$. Equip $\mathcal{E}$ with the topology induced by $\mathcal{X}$, and suppose $\mathcal{E}$ is closed.*

- *If $(X_n,\ n \in \mathbb{N})$ satisfies an LDP in $\mathcal{E}$ with rate function I, then it satisfies an LDP in $\mathcal{X}$ with rate function*

$$I'(x) = \begin{cases} I(x) & \textit{if } x \in \mathcal{E} \\ \infty & \textit{otherwise.} \end{cases}$$

- *If $(X_n,\ n \in \mathbb{N})$ satisfies an LDP in $\mathcal{X}$ with rate function I then it satisfies an LDP in $\mathcal{E}$ with the same rate function I.*

Exercise 4.7
Prove Lemma 4.9. Hint: for the second part, note that since $P(X_n \notin \mathcal{E}) = 0$, by using the large deviations lower bound, $\inf_{x \in \mathcal{X} \setminus \mathcal{E}} I(x) = \infty$. For a proof, see Dembo and Zeitouni [25, Lemma 4.1.5]. ◇

The following definition concerns another sort of restriction.

Definition 4.5 (Exponential tightness) *Let X_n be a sequence of random variables in $\mathcal{X}$. It is* exponentially tight *if for all $\alpha \in \mathbb{R}^+$ there exist compact sets $K_\alpha \subset \mathcal{X}$ such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log P(X_n \notin K_\alpha) < -\alpha.$$

In other words, X_n is exponentially tight if exponentially much of the probability mass is found in compact sets.

Exponential tightness is used in proving the following theorem, a partial converse to the contraction principle. It says that if X_n is a sequence of random variables in $\mathcal{X}$, and f is a continuous bijection $\mathcal{X} \to \mathcal{Y}$, and $f(X_n)$

satisfies an LDP, then (under certain conditions) so does X_n. We will most often use this where f is the identity map and $\mathcal{Y}$ is $\mathcal{X}$ but with a coarser topology, and we will refer to this use as *strengthening the LDP*. For a proof of the theorem see Dembo and Zeitouni [25].

Theorem 4.10 (Inverse contraction principle) *Let f be a continuous bijection from $\mathcal{X}$ to another Hausdorff space $\mathcal{Y}$, and suppose $f(X_n)$ satisfied an LDP in $\mathcal{Y}$ with rate function J. If X_n is exponentially tight in $\mathcal{X}$ then X_n satisfies an LDP in $\mathcal{X}$ with good rate function $I(x) = J(f(x))$.*

Exercise 4.8
Suppose that X_n satisfies a *weak LDP* in $\mathcal{X}$ with rate function I, i.e.

$$-\inf_{x \in B^\circ} I(x) \le \liminf_{n\to\infty} \frac{1}{n} \log P(X_n \in B) \quad \text{for all sets } B \text{, and}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log P(X_n \in B) \le -\inf_{x \in B} I(x) \quad \text{for compact sets } B.$$

(This is not to be confused with the stronger/weaker LDPs referred to above.) Suppose that X_n is exponentially tight. Show that X_n satisfies an LDP with good rate function I. Hint: For any closed set B, divide the event $\{X_n \in B\}$ into $\{X_n \in B \cap K_\alpha\}$, which is compact, and $\{X_n \in B \setminus K_\alpha\}$, the probability of which vanishes as $\alpha \to \infty$. For goodness, show that $\{x : I(x) \le \alpha\}$ is contained in K_α. ◇

In Cramér's theorem and its generalisation, we saw that if a sequence of random variables has finite exponential moments which are smooth in a neighbourhood of the origin, then the random variables satisfy an LDP. Conversely, if a sequence of random variables satisfies an LDP, do smooth functions of them have finite exponential moments? The answer is provided by Varadhan's lemma. This lemma can be useful in applications of large deviation theory. Here we state two special cases; for more general versions see Dembo and Zeitouni [25, Theorem 4.3.1].

Lemma 4.11 (Varadhan's lemma) *Let X_n satisfy an LDP in $\mathcal{X}$ with rate function I, and let $f : \mathcal{X} \to \mathbb{R}$ be a bounded and continuous function. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log E e^{n f(X_n)} = \sup_{x \in \mathcal{X}} f(x) - I(x). \tag{4.4}$$

Exercise 4.9
Prove Lemma 4.11. Hint: For the upper bound, divide the range of f into a finite number of closed sets, and thence divide $\mathcal{X}$ into a finite number

of closed sets F_i such that f varies little on each F_i. Use the principle of the largest term, Lemma 2.2, to find the F_i that contributes most to the expectation. For the lower bound, pick any $x \in \mathcal{X}$ and estimate the moment generating function only on points near x. ◇

Lemma 4.12 (Varadhan's lemma) *Let X_n satisfy an LDP in $\mathcal{X}$ with good rate function I and let $f : \mathcal{X} \to \mathbb{R}$ be a continuous function. Assume that for some $\gamma > 1$*

$$\limsup_{n\to\infty} \frac{1}{n} \log E e^{\gamma n f(X_n)} < \infty.$$

Then (4.4) *holds.*

Exercise 4.10
Let X_n be a Markov chain on $\{0, h\}$ with transition probabilities p and q for the jumps $h \to 0$ and $0 \to h$ respectively. Let $S_n = X_1 + \cdots + X_n$. Using your answer to Exercise 3.2, and the second version of Varadhan's lemma with $f(x) = \theta x$, show that

$$\lim_{n\to\infty} \frac{1}{n} \log E e^{\theta S_n} = \log\Big(\xi + \sqrt{\xi^2 - (1-p-q)e^{\theta h}}\Big)$$

where $$\xi = \frac{1}{2}\big(1 - q + (1-p)e^{\theta}\big).$$ ◇

Sanov's theorem, which we first came across in Section 2.7, does not apply just to empirical measures on finite sets. We include the general statement here for completeness. A proof of the theorem in a general setting can be found in [25]. As is often the case in large deviations, the idea is simple but getting the topology right can be difficult.

Given a set $\mathcal{X}$ let $M_1(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$ equipped with the weak topology, namely that generated by the open balls

$$B(\phi, x, \delta) = \big\{\nu \in M_1(\mathcal{X}) : \big|E_\nu \phi(X) - x\big| < \delta\big\}$$

where ϕ is a bounded continuous real-valued function on $\mathcal{X}$, X is a random variable drawn from ν, $x \in \mathbb{R}$, and $\delta > 0$.

Theorem 4.13 (Sanov's theorem) *Let $(X_i, i \in \mathbb{N})$ be a sequence of i.i.d. random variables taking values in a Polish space $\mathcal{X}$, with distribution μ. The sequence of empirical measures*

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1[X_i \in A] \quad \textit{for } A \subset \mathcal{X}$$

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

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

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

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

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, 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

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

where C_n and D_n are closed. The set $C \times D$ is closed in $\mathcal{X} \times \mathcal{Y}$ if C and D are closed in $\mathcal{X}$ and $\mathcal{Y}$.

Goodness of K. Clearly $K(x,y) \geq 0$. A typical level set is

$$\begin{aligned}
\big\{(x,y) : K(x,y) \leq \alpha\big\} &= \big\{(x,y) : I(x) + J(y) \leq \alpha\big\} \\
&= \bigcap_{n \in \mathbb{N}} \bigcup_{m \leq n+1} \{(x,y) : I(x) \leq \frac{m}{n}\alpha, J(y) \leq \frac{n+1-m}{n}\alpha\} \\
&= \bigcap_{n \in \mathbb{N}} \bigcup_{m \leq n+1} \{x : I(x) \leq \frac{m}{n}\alpha\} \times \{y : J(y) \leq \frac{n+1-m}{n}\alpha\} \\
&= \bigcap_{n \in \mathbb{N}} \bigcup_{m \leq n+1} \text{compact} \times \text{compact} \\
&= \text{compact}.
\end{aligned}$$

Thus K is a good rate function.

LD lower bound. Let B be an open set in $\mathcal{X} \times \mathcal{Y}$, and $(x,y) \in B$. By the basis we have described, B is the union of sets of the form $O \times P$ where O and P are open in $\mathcal{X}$ and $\mathcal{Y}$; and $(x,y) \in O \times P$ for some O and P. Now

$$\begin{aligned}
P\big((X_n, Y_n) \in B\big) &\geq P\big((X_n, Y_n) \in O \times P\big) \\
&= P(X_n \in O)P(Y_n) \in P \quad \text{by independence.}
\end{aligned}$$

To turn this into a large deviations lower bound,

$$\begin{aligned}
&\liminf_{n \to \infty} \frac{1}{n} \log P\big((X_n, Y_n) \in B\big) \\
&\qquad = \liminf_{n \to \infty} \frac{1}{n} \Big[\log P(X_n \in O) + \log P(Y_n \in P)\Big] \\
&\qquad \geq \liminf_{n \to \infty} \frac{1}{n} \log P(X_n \in O) + \liminf_{n \to \infty} \frac{1}{n} \log P(Y_n \in P) \\
&\qquad \geq - \inf_{x' \in O} I(x') - \inf_{y' \in P} J(y') \\
&\qquad \geq -\big(I(x) + J(y)\big) \\
&\qquad = -K(x,y).
\end{aligned}$$

LD upper bound for cylinders. Consider a closed set of the form

$$B_N = \bigcap_{n \leq N} (C_n \times \mathcal{Y}) \cup (\mathcal{X} \times D_n).$$

For such a set,

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

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

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

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 $B=\lim_{N\to\infty} B_N$.) Thus

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

Hence the lim sup is

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

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

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

Certainly, LHS≤RHS. Suppose the inequality is strict. Then the left hand side is bounded, and hence finite. Since each B_N is closed, and K is good, for each N the infimum is attained at some z_N. Furthermore, since K is good and the left hand side is bounded, the z_N all live in some compact space, hence there is a convergent subsequence with limit z^*. Since the B_N form a decreasing sequence, $z_N, z_{N+1}, \ldots$ all lie in B_N, which is closed, so $z^* \in B_N$ for all N, so $z^* \in B$. Furthermore

$$\begin{aligned}\lim_{N\to\infty} \inf_{z\in B_N} K(z) &= \liminf_{z_N \to z} K(z_N)\\ &\geq K(z^*) \quad \text{by lower-semicontinuity}\\ &\geq \inf_{z\in B} K(z).\end{aligned}$$

This completes the proof. □

The next result is the basic tool used to prove sample path LDPs, that is, LDPs for processes. To state the result properly requires the concept of projective limit spaces, which is somewhat abstract, and which we will not define. Instead we will give three examples of projective limit spaces, the three examples used in this book, and refer the interested reader to Dembo and Zeitouni [25, Theorem 4.6.1] for a proper definition and a proof of the following.

Theorem 4.15 (Dawson-Gärtner theorem for projective limits) *Let $\mathcal{X}$ be a projective limit space, with projections $(p_j,\ j \in J)$. Let X_n be a sequence of random variables in $\mathcal{X}$. Suppose that for every j, $p_j(X_n)$ satisfies an LDP in $p_j(\mathcal{X})$ with good rate function I_j. Then X_n satisfies an LDP in $\mathcal{X}$ equipped with the projective limit topology, with good rate function*

$$I(x) = \sup_{j\in J} I_j(p_j(x)).$$

Example 4.11

Let $\mathcal{X}$ be the space of functions $x : \mathbb{N} \to \mathbb{R}$, equipped with projections $p_j : \mathcal{X} \to \mathbb{R}^j$ given by

$$p_j(x) = (x_1, \ldots, x_j) \quad \text{for } j \in \mathbb{N}.$$

Let $p_j(\mathcal{X})$ be equipped with the standard topology on $\mathbb{R}^j$. The projective limit topology is the weakest topology which makes every p_j continuous, namely the topology of pointwise convergence. ◇

Example 4.12
Let $\mathcal{X}$ be the space of continuous functions $x : \mathbb{R}_0^+ \to \mathbb{R}$, and $\mathcal{X}_j$ the space of continuous functions from $x : [0, j] \to \mathbb{R}$, and consider the projections $p_j : \mathcal{X} \to \mathcal{X}_j$ given by

$$p_j(x) = x|_{[0,j]} \quad \text{for } j \in \mathbb{R}_0^+.$$

Let $p_j(\mathcal{X})$ be equipped with the topology of uniform convergence. The projective limit topology is the weakest topology which makes every p_j continuous, namely the topology of uniform convergence on compact sets. ◇

Example 4.13
Let $\mathcal{X}$ be the space of continuous functions $x : [0, 1] \to \mathbb{R}$. Let J be the set of partitions

$$J = \Big\{0 \leq t_1 < t_2 < \cdots < t_n \leq 1\Big\}$$

and for $j = (t_1 < \cdots < t_n)$ consider the projection

$$p_j(x) = \big(x(t_1), \ldots, x(t_n)\big).$$

Let $p_j(\mathcal{X})$ be equipped with the standard topology on $\mathbb{R}^n$. The projective limit topology is the weakest topology which makes every p_j continuous, namely the topology of pointwise convergence. ◇

Chapter 5

Continuous Queueing Maps

The techniques outlined in Chapter 1 are fine for studying the single-server queue, but they quickly become cumbersome when one tries to apply them to more complicated queueing networks. Here we introduce a different approach which uses abstract large deviations theory and which has many advantages. In this chapter we will describe the general approach; and in the next two chapters we will apply it to the two scaling regimes described earlier, the large-buffer regime and the many-flows regime.

5.1 Introduction

Recall the simple queueing model of Chapter 1. Consider a queue operating in discrete time, with constant service rate C. Let Y_t be the amount of work arriving at time $t \in \mathbb{Z}$, and for $t > 0$ define the cumulative sum process $A(t) = Y_{-t+1} + \cdots + Y_0$, with $A(0) = 0$. We found an expression for the queue size at time 0:

$$Q_0 = \sup_{t \geq 0} A(t) - Ct.$$

Write this more suggestively as

$$Q_0 = f(A)$$

where A is the entire process $A = (A(t), t \geq 0)$.

> *Note.* More generally, consider a network of queues fed by a collection of arrival processes. Write $\underline{A}$ for the vector of arrival processes. Many quantities of interest (such as the queue size at any queue in the network, or the departure process from some queue) can be written as functions $f(\underline{A})$. We

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 77–104, 2004.

> may wish to include in $\underline{A}$ any other random influences on the network. For example, if the service process at a queue is random, that should be included in $\underline{A}$.

The big idea of this chapter (and the rest of the book) is that we can use the contraction principle to investigate $f(A)$ (or rather $f(A^L)$ for various interesting sequences of processes A^L):

> *Note.* Look back through Chapter 4 now if you are not familiar with the contraction principle.

if the sequence A^L satisfies a large deviations principle, and if the function f is continuous, then the contraction principle tells us that $f(A^L)$ satisfies a large deviations principle, and it even tells us the rate function.

> *Note.* This chapter, and indeed the rest of the book, deals with queues in steady state; that is, the functions we are interested in are functions of the entire history of the process $(A(t),\ t \in \mathbb{R}^+)$. The techniques we will describe can also be applied to non-steady-state scenarios. For example, we could suppose that the queue starts empty at time $-T$, and only consider the evolution of the system over $[-T, 0]$, which is a function of $(A(t),\ t \in [0, T])$. This considerably simplifies the problem.

5.2 An Example: Queues with Large Buffers

To explain how we might choose A^L, and to show the steps involved, here is an outline of how we can use the contraction principle to derive the results for queues with large buffers in Section 1.3.

We want to let A^L be a version of A which is speeded up in time and scaled down in space:

$$A^L(t) = \frac{1}{L}A(Lt).$$

Actually, we need to be a little more careful than this. It turns out to be more helpful to work in continuous time. To achieve this, we make the following definition:

Given a discrete-time process $(X(t), t \in \mathbb{Z})$, define its polygonalization $(\tilde{X}(t), t \in \mathbb{R})$ by

$$\tilde{X}(t) = (\lfloor t+1 \rfloor - t)X(\lfloor t \rfloor) + (t - \lfloor t \rfloor)X(\lfloor t+1 \rfloor).$$

Now, we can properly define the scaled processes: if $\tilde{A}$ is the polygonalized version of A, then for $t \in \mathbb{R}^+$ define

$$\tilde{A}^L(t) = \frac{1}{L}\tilde{A}(Lt).$$

Define also the continuous-time version $\tilde{f}$ of the queue size function f:

$$\tilde{f}(\tilde{A}) = \sup_{t\in\mathbb{R}^+} \tilde{A}(t) - Ct.$$

Consider this function applied to $\tilde{A}^L$:

$$\begin{aligned}\tilde{f}(\tilde{A}^L) &= \sup_{t\in\mathbb{R}^+} \tilde{A}^L(t) - Ct\\ &= L^{-1} \sup_{t\in\mathbb{R}^+} \tilde{A}(t) - Ct\\ &= L^{-1} \sup_{t\in\mathbb{N}_0} A(t) - Ct\\ &= L^{-1} f(A) = L^{-1} Q_0.\end{aligned} \tag{5.1}$$

This means that

$$P(\tilde{f}(\tilde{A}^L) > b) = P(Q_0 > Lb).$$

The contraction principle gives us estimates for the first quantity, and hence for the second. (We will drop the $\sim$s after this section!)

Specifically, $\tilde{A}^L$ will typically satisfy a large deviations principle of the form

$$\frac{1}{L}\log P(\tilde{A}^L \in B) \approx -\inf_{a\in B} I(a)$$

for some good rate function I, in some topological space. (The approximation is in the large deviations sense, (4.1).) If the queue size function $\tilde{f}$ is continuous on that space, then

$$\frac{1}{L}\log P(L^{-1}Q_0 > b) \approx -J(b)$$

where

$$J(b) = \inf_{a:\tilde{f}(a)>b} I(a). \tag{5.2}$$

This is exactly the sort of estimate we obtained in Section 1.3.

We glibly define $J(b)$ thus, but to simplify it into a useful form can be a lot of work. But in some sense, this is a difficulty inherent in queueing theory. Often, when one seeks to estimate the probability of a rare event, a first heuristic is to consider the most likely way for that event to occur. Around this heuristic one can build various probability estimates. The rate function J expresses exactly this heuristic, and the contraction principle does all the work of building probability estimates around it.

The most likely way for the rare event $\{Q_0 > Lb\}$ to occur is when the input process $\tilde{A}^L$ is close to the optimizing a in (5.2). The sense of this can be made precise, using the idea that rare events occur only in the most likely way (Lemma 4.2). Since this a is a function indexed by time, we call it the most likely path to overflow.

> *Note.* We can now see why it was necessary to work with continuous-time polygonalizations: the second equality in (5.1) only works if we are taking the supremum over $\mathbb{R}_0^+$, not $\mathbb{N}_0$. It is worth checking that this translation does not mess up the queue size function:
>
> *Exercise 5.1*
> Explain why the third equality in (5.1) is true. ◇

5.3 The Continuous Mapping Approach

The general approach is this. Consider a queueing network, or rather a sequence of queueing networks indexed by L, in which the Lth network has a vector of inputs $\underline{A}^L$. Include in $\underline{A}^L$ all relevant information, such as input processes and service processes, so that we can express any quantity of interest as a function $f(\underline{A}^L)$.

Prove a large deviations principle for $\underline{A}^L$, in some topological space. Show that f is continuous on that space. Use the contraction principle to derive a large deviations principle for $f(\underline{A}^L)$, and simplify the resulting rate function.

What sort of functions f are suitable? In the last section, $f(A)$ was the function giving queue size at time 0 in a single-server queue with cumulative arrival process A. There are many other possible functions: for example, the queue size in a finite-buffer queue, or the departure process.

What is an interesting sequence of inputs $\underline{A}^L$? In the last section, we used a particular definition of A^L, and this led to a large deviations principle for queues with large buffers (Section 1.3). If we had chosen a different A^L, we could have reproduced the large deviations results for queues with many input flows (Section 1.4) or for queues with power-law scalings (Section 3.3). These scalings, and others, will be elaborated in the following chapters.

The power of this general technique is threefold:

- *We only need to prove a sample path LDP for $\underline{A}^L$ once, and we can get LDPs for many different functions $f(\underline{A}^L)$;*
- *We only need to prove continuity of f once, and we get LDPs for many different scalings $\underline{A}^L$;*
- *We can not only estimate the probability of a rare event $\{f(\underline{A}^L) \in B\}$, but also find the most likely path $\underline{\hat{a}}$ to lead to that event.*

The work of finding a large deviations principle for A^L, and of simplifying the rate function, depends very much on how A^L is defined. The work of proving that f is continuous is much the same, though, and the rest of this chapter looks at a variety of interesting functions on appropriately chosen spaces.

A continuous mapping approach is useful for stochastic process limits at scales other than the large deviations scale. See for example Whitt [98] and references therein.

5.4 Continuous Functions

The first job is to decide on a topological space in which A^L satisfies a large deviations principle and on which f is continuous. This involves a tradeoff: if the topology is too fine, it will be hard to prove a large deviations principle; if the topology is too coarse, it will be hard to prove continuity. And the space will of course depend on the application. Happily there are common themes. The following defines a suitable space for the single-server queue, and for many other systems.

Definition 5.1 *Define the space $\mathcal{C}_\mu$ to be the set of continuous functions $x : \mathbb{R}^+ \to \mathbb{R}$ for which $x(0) = 0$ and*

$$\lim_{t\to\infty} \frac{x(t)}{t+1} = \mu \tag{5.3}$$

equipped with the topology induced by the scaled uniform norm

$$\|x\| = \sup_{t\in\mathbb{R}^+} \left|\frac{x(t)}{t+1}\right|. \tag{5.4}$$

(We will use the same notation x for continuous-time and discrete-time processes; it should be clear from the context which is meant.)

We can think of x as the (polygonalized) cumulative arrival process to a queue. The μ in (5.3) is called the *mean arrival rate.* Or we can think of x as the (polygonalized) cumulative service process, in which case μ is called the *mean service rate.*

> *Note.* The role of the scaled uniform norm (5.4) will become clearer after reading Example 5.2.
>
> A useful property is that it induces a topology which makes the space $\mathcal{C}_\mu$ isomorphic to the space of functions $\{x : [0,1] \to \mathbb{R} : x(1) = \mu\}$ equipped with the uniform norm. This is a Polish space, i.e. a complete separable metric space. Completeness makes it easier to check continuity; separability is important in finding LDPs for product spaces, as we saw in Theorem 4.14.

A simple but very important property of the scaled uniform norm is the following:

Lemma 5.1 *Suppose $x^n \to x$ in $\mathcal{C}_\mu$. Then $x^n \to x$ uniformly on compact intervals. In other words, for all $T > 0$,*

$$\sup_{0\le t\le T} |x^n(t) - x(t)| \to 0.$$

The left hand side of this expression is called the supremum norm over the interval $[0,T]$.

The proof is trivial. The lemma is nonetheless worth stating, because the topology of uniform convergence on compact intervals plays an important role in functions related to queue size, which often involve supremums and infimums, because of the following trivial result:

Lemma 5.2 *The function $x \mapsto \sup_{0\le t\le T} x(t)$ is continuous with respect to the topology of uniform convergence on compact intervals.*

For some of our applications we will need to work in the smaller space $\mathcal{A}_\mu$, which is $\mathcal{C}_\mu$ restricted to the set of absolutely continuous functions. (Absolute continuity implies continuity.) We will point out when we need to assume absolute continuity. A key feature is that if x is absolutely continuous then its derivative exists almost everywhere and

$$x(t) = \int_{u=0}^{t} \dot{x}(u)\,du.$$

> *Note. Absolute continuity* is defined as follows. Consider a sequence (finite or infinite) of non-overlapping intervals $[u_i, v_i]$ in $\mathbb{R}^+$. A function $x : \mathbb{R}^+ \to \mathbb{R}$ is absolutely continuous if
>
> $$\sum |x(v_i) - x(u_i)| \to 0$$
>
> as
>
> $$\sum (v_i - u_i) \to 0.$$

Sometimes we will have to deal with truncations. Define $\mathcal{C}^T$ to be the space of continuous functions $x : [0,T] \to \mathbb{R}$ equipped with the topology of uniform convergence, and define $\mathcal{A}^T$ similarly.

5.5 Some Convenient Notation

Because of the many transformations of the arrival process which we will use in the rest of this chapter, and because it is bothersome to work in 'reverse time', we will first give some notation.

In discrete time, suppose that $X(t)$ is the cumulative sums process corresponding to a sequence $(Y_t,\ t \in \mathbb{N})$, i.e. $X(t) = Y_0 + \cdots + Y_{t-1}$, with $X(0) = 0$. Define

$$X(-u,-v] = X(u) - X(v) \quad \text{for } u,v \in \mathbb{N}_0.$$

If X is an arrival process, with Y_t the amount of work arriving at time $-t$, then $X(-u,-v]$ is the amount arriving in the interval $(-u,-v]$. In particular,

$$X(-u,0] = X(u)$$

and we will use the $(\cdot,0]$ notation to try to avoid confusion.

In continuous time, for $x \in \mathcal{C}_\mu$, we similarly define

$$x(-u,-v] = x(u) - x(v) \quad \text{for } u,v \in \mathbb{R}_0^+.$$

If $x \in \mathcal{C}_\mu$ is an arrival process, then $x(-u,-v]$ is the amount of work arriving in the interval $(-u,-v]$. Since we will usually work with continuous functions, there is no point in insisting on $(\cdot,0]$ rather than $[\cdot,0]$ or $[\cdot,0)$. But we will stick with this notation because it better parallels the discrete-time notation we have been using: with these definitions, if x is the polygonalization of X then $x(-u,-v] = X(-u,-v]$ for $u,v \in \mathbb{N}_0$, whereas $x[-u,-v] \neq X[-u,-v]$.

Sometimes we will need to deal with truncations. Define $\mathcal{C}^T$ to be the space of continuous functions $x : [0,T] \to \mathbb{R}$ equipped with the topology of uniform convergence. Write

$$x|_{(-t,0]}$$

for the restriction of $x \in \mathcal{C}_\mu$ to $\mathcal{C}^T$. We will use the same notation for truncations of discrete-time processes.

Finally, if x is absolutely continuous, then for $t \geq 0$ write

$$\dot{x}_{-t} \ \text{ for } \ \frac{d}{dt}x(t) \quad \text{whenever the derivative exists,}$$

so that for fixed $u > 0$ and $t \in [-u, 0]$

$$\dot{x}_t = \frac{d}{dt} x(-u, t] \quad \text{whenever the derivate exists.}$$

If x is an absolutely continuous arrival process, $\dot{x}_{-t}$ is the instantaneous arrival rate at time $-t$.

If X is a discrete-time process it doesn't make sense to talk about derivatives, but we will nevertheless use the notation

$$\dot{X}_{-t} = X(t+1) - X(t).$$

If X is an arrival process, $\dot{X}_{-t}$ is the amount of work arriving at time $-t$.

5.6 Queues with Infinite Buffers

In Section 5.2, we explained how to use the contraction principle to estimate the queue length distribution, assuming that the queue size function is continuous. Here we will prove continuity, along with several other useful properties.

Theorem 5.3 *Suppose $\mu < \nu$. Then the queue size function*

$$f(a, c) = \sup_{t \in \mathbb{R}_0^+} a(-t, 0] - c(-t, 0] \tag{5.5}$$

is continuous on $\mathcal{C}_\mu \times \mathcal{C}_\nu$. Furthermore, the supremum is attained (and is finite).

Proof. Since $\mathcal{C}_\mu$ and $\mathcal{C}_\nu$ are metric spaces, we can check continuity using sequences. Let $(a^n, c^n) \to (a, c)$ in $\mathcal{C}_\mu \times \mathcal{C}_\nu$. We want to show that for any $\varepsilon > 0$ there exists N such that for all $n > N$, $|f(a^n, c^n) - f(a, c)| < \varepsilon$.

Let $\delta > 0$. First, since $a^n \to a$ in $\mathcal{C}_\mu$, there exists N such that for all $n > N$

$$\sup_{t \in \mathbb{R}^+} \left| \frac{a^n(-t, 0] - a(-t, 0]}{t+1} \right| < \delta$$

and hence, for $n > N$ and $t \geq 0$

$$a^n(-t, 0] < a(-t, 0] + \delta(t+1).$$

Second, since $a \in \mathcal{C}_\mu$, it has mean rate μ, and hence by (5.3) there exists T such that for $t > T$

$$\left| \frac{a(-t, 0]}{t+1} - \mu \right| < \delta,$$

and hence

$$a(-t,0] < (\mu+\delta)(t+1).$$

Putting these together: for $n > N$ and $t > T$,

$$a^n(-t,0] < (\mu+2\delta)(t+1).$$

We have assumed that $\mu < \nu$. Fix σ with $\mu < \sigma < \nu$. By choosing δ sufficiently small, we find that for $n > N$ and $t > T'$,

$$a^n(-t,0] < \sigma t,$$

and (by a similar argument)

$$c^n(-t,0] > \sigma t.$$

Thus

$$\begin{aligned} f(a^n,c^n) &= \sup_{t\in\mathbb{R}^+} a^n(-t,0] - c^n(-t,0] \\ &= \sup_{0\le t\le T'} a^n(-t,0] - c^n(-t,0]. \end{aligned} \tag{5.6}$$

The same is true for $f(a,c)$, with (without loss of generality) the same T'. In other words, there exists a time interval $[0,T']$ such that (for n sufficiently large) we can ignore what happens outside that interval. Also, as we have already remarked, $(a^n,c^n) \to (a,c)$ in $\mathcal{C}_\mu \times \mathcal{C}_\nu$ implies that there is uniform convergence over that interval. Now Lemma 5.2 implies that (5.6) is continuous, and so

$$f(a^n,c^n) \to f(a,c)$$

as required.

Since $(a,c) \in \mathcal{C}_\mu \times \mathcal{C}_\nu$, $a-c$ is continuous. Thus the supremum in $f(a,c) = \sup_{0\le t\le T'} a(-t,0] - c(-t,0]$ is attained. Since both a and c are real-valued functions, the supremum is finite. □

Example 5.2

The function f is not continuous on the set $\mathcal{C}_\mu \times \mathcal{C}_\nu$ equipped instead with the topology of uniform convergence on compact intervals. To see this, for $t \in \mathbb{R}^+$ let

$$a^n(-t,0] = \begin{cases} \mu t & \text{if } t < n \\ \mu t + (t-n)\big((n+1)(\nu-\mu)+1\big) & \text{if } n \le t < n+1 \\ \mu t + (n+1)(\nu-\mu)+1 & \text{if } t \ge n+1. \end{cases}$$

Each arrival process a^n has mean rate μ; the arrival rate is μ except in the interval $(-n-1,-n]$ when it is larger. Also, $a^n \to \mu e$ uniformly on compact intervals, where by μe we mean the constant process of rate μ. Let each c^n be the service process of constant rate ν, $c^n(-t,0] = \nu t$, i.e. $c^n = \nu e$. The arrival process was chosen so as to make $f(a^n,c^n) = 1$; however $f(\mu e,\nu e) = 0$. Thus f is not continuous with respect to the topology of uniform convergence on compact intervals. ◇

In the course of proving Theorem 5.3, we implicitly established a fact which merits attention in its own right. To make it explicit, we need another definition. We have already defined $f(a,c)$, the function giving queue size at time 0, which we will now call $q_0(a,c)$. Now define

$$q_{-t}(a,c) = \sup_{u\geq t} a(-u,-t] - c(-u,-t]. \tag{5.7}$$

Interpret $q_{-t}(a,c)$ as the queue size at time $-t$. Note that this is consistent with $q_0(s)$.

> *Note.* Strictly speaking, what we should do is the following. For discrete-time arrival process a and service process c, let $q_{-t}(a,c)$ be the queue size at time $-u$, as defined by (1.3):
>
> $$q_{-t}(s) = \sup_{-u\in\mathbb{Z},-u\leq -t} a(-u,-t] - c(-u,-t].$$
>
> Then $q_{-t}(a,c) = \tilde{q}_{-t}(\tilde{a},\tilde{c})$, where $\tilde{q}_{-t}$ is our equation (5.7). (You should check this.) In other words, (5.7) is a sensible way to define queue size for continuous-time input processes, consistent with the discrete-time definition.

Now we can make the claim.

Lemma 5.4 *Suppose a and c are continuous.*
i. If

$$q_0(a,c) = a(-t,0] - c(-t,0]$$

then the queue is empty at time $-t$.
ii. The queue is empty at some point in $[-T,0]$ if and only if

$$q_0(a,c) = \sup_{-T\leq -t\leq 0} a(-t,0] - c(-t,0].$$

This helps us to understand the role of the supremum in (5.5). If $(a,c) \in \mathcal{C}_\mu \times \mathcal{C}_\nu$ and $\mu < \nu$, then the optimal t^* in $q_0(s)$ is attained. By continuity of a and c, there is a smallest such t^*. For this t^*: by (i), the queue is empty at $-t^*$; by (ii), the queue is non-empty (and hence the server is busy) in $(-t^*,0]$.

Proof of Lemma 5.4 First, note that for all $-t \leq 0$,

$$\begin{aligned} q_0(a,c) &= \sup_{-u\leq 0} a(-u,0] - c(-u,0] \\ &\geq \sup_{-u\leq -t} a(-u,0] - c(-u,0] \\ &= a(-t,0] - c(-t,0] + \sup_{-u\leq -t} a(-u,-t] - c(-u,-t] \\ &= a(-t,0] - c(-t,0] + q_{-t}(a,c). \end{aligned} \tag{5.8}$$

(This last expression is what the queue size would be at time 0 if it didn't idle in $(-t,0]$.) Now to prove the claims of the lemma.

i. Suppose not: i.e. $q_{-t}(a,c) > 0$. By (5.8),

$$q_0(a,c) \geq a(-t,0] - c(-t,0] + q_{-t}(a,c) > a(-t,0] - c(-t,0]$$

which contradicts the assumption.

ii. *The if part.* We need to show that $q_{-t} = 0$ for some $-t \in [-T,0]$. Suppose not, i.e. $q_{-t} > 0$ for all $-t \in [-T,0]$. By (5.8),

$$q_0(a,c) > a(-t,0] - c(-t,0] \quad \text{for all } -t \in [-T,0].$$

But by assumption

$$q_0(a,c) = \sup_{-T\leq t\leq 0} a(-t,0] - c(-t,0]. \tag{5.9}$$

We have assumed that a and c are continuous. Thus the supremum in (5.9) is attained at some $-t \in [-T,0]$. Hence a contradiction.

The only if part. Suppose that

$$q_0(a,c) = \sup_{-t\leq 0} a(-t,0] - c(-t,0] \neq \sup_{-T\leq -t\leq 0} a(-t,0] - c(-t,0].$$

Then for any $-t \in [-T,0]$,

$$q_0(a,c) = \sup_{-u<-t} a(-u,0] - c(-u,0], \tag{5.10}$$

and

$$q_0(a,c) > a(-t,0] - c(-t,0]. \tag{5.11}$$

By (5.10),

$$\begin{aligned} q_0(a,c) &= a(-t,0] - c(-t,0] + \sup_{-u<-t} a(-u,-t] - c(-u,-t] \\ &= a(-t,0] - c(-t,0] + q_{-t}, \end{aligned}$$

and, using (5.11), $q_{-t} > 0$. Hence the queue is never empty in $[-T,0]$. □

Also, a corollary which makes explicit another fact that we proved in the course of Theorem 5.3.

Lemma 5.5 *Let $a^n \to a$ in $\mathcal{C}_\mu$ and $c^n \to c$ in $\mathcal{C}_\nu$, and $\mu < \nu$. Then there exists a T such that the queue $q_t(a,c)$, and every queue $q_t(a^n,c^n)$, are each empty at some point in $[-T,0]$ (though not necessarily the same point).*

Finally, a useful way to rewrite the queue size function for time 0, assuming we know the queue size at some time $-t$.

Exercise 5.3
Write q_0 for $q_0(a,c)$ and q_{-1} for $q_{-1}(a,c)$. Show that

$$q_0 = \Big(q_{-t} + a(-t,0] - c(-t,0]\Big) \vee \Big(\sup_{0\le u\le 1} a(-u,0] - c(-u,0]\Big) \qquad \diamond$$

5.7 Queues with Finite Buffers

There are plenty of other interesting continuous functions apart from queue size in a queue with an infinite buffer. Now we turn our attention to queues with finite buffers. This section introduces some new machinery: a procedure for turning a finite-horizon function into an infinite-horizon one. (Another term for this is turning a transient problem into an steady-state problem).

The plan of the section is: introduce the queue size function for discrete-time queues with finite buffers over finite horizons, find a consistent continuous-time version, extend it to an infinite-horizon version.

For a similar approach to finite-buffer queues, see Toomey [94], who finds a different form of the queue size function.

Consider a queue with finite buffer B. Let $A(-t,0]$ be the amount of work arriving in $(-t,0]$, let $C(-t,0]$ be the amount of service offered in that time, and let Q_t be the queue size at time t. As usual, let $A(-t,0] = \dot{A}_{-t+1} + \cdots + \dot{A}_0$, and similarly for C. The evolution of Q_t is described by the recursion

$$Q_t = [Q_{t-1} + \dot{A}_t - \dot{C}_t]_0^B \tag{5.12}$$

where $[x]_b^a = (x \vee b) \wedge a$. As was the case with infinite buffers, this recursion may have many solutions, so we will impose boundary conditions. First we will work over a finite horizon, imposing the boundary condition $Q_{-T} = 0$

and studying Q_{-t}, $-T \leq -t \leq 0$, which we will call Q_{-t}^{-T} to emphasize the boundary condition.

We want to find an equivalent description in continuous time. We will start by finding a more explicit form of the queue size function.

Lemma 5.6

$$Q_0^{-T} = \Big[X(-T,0]\Big]_{\max_{u\in[0,T]} N_u \wedge (M_u+B)}^{\min_{u\in[0,T]} N_u \vee (M_u+B)} \tag{5.13}$$

where

$$\begin{aligned} M_u &= \min_{v\in[0,u]} X(-v,0] \\ N_u &= \max_{v\in[0,u]} X(-v,0] \\ X(-v,0] &= A(-v,0] - C(-v,0]. \end{aligned}$$

Sketch proof. This expression comes from expanding the recursion (5.12). It could be proved formally by induction on T; here instead is a sketch proof which is certainly less formal but we hope more revealing. Let $\dot{X}_t = \dot{A}_t - \dot{C}_t$, so that $X(-v,0] = \dot{X}_{-v+1} + \cdots + \dot{X}_0$. Then

$$\begin{aligned} Q_0 &= \big[Q_{-1} + \dot{X}_0\big]_0^B \\ &= \big[\big[Q_{-2} + \dot{X}_{-1}\big]_0^B + \dot{X}_0\big]_0^B \\ &= \big[\big[Q_{-2} + \dot{X}_{-1} + \dot{X}_0\big]_{\dot{X}_0}^{\dot{X}_0+B}\big]_0^B \\ &= \big[\big[Q_{-2} + X(-2,0]\big]_{X(-1,0]}^{X(-1,0]+B}\big]_{X(0,0]}^{X(0,0]+B} \\ &\;\;\vdots \\ &= \big[\cdots\big[Q_{-T} + X(-T,0]\big]_{X(-T+1,0]}^{X(-T+1,0]+B}\cdots\big]_{X(0,0]}^{X(0,0]+B} \\ &= \big[\cdots\big[X(-T,0]\big]_{X(-T,0]}^{X(-T,0]+B}\cdots\big]_{X(0,0]}^{X(0,0]+B}. \end{aligned}$$

The last equality is by the assumption that $Q_{-T} = 0$, and because $x = [x]_x^{x+B}$.

Convince yourself that this is equal to (5.13), by drawing a picture of the successive truncations of $X(-T,0]$, along the lines of Figure 5.1. □

This expression (5.13) suggests how we should define the queue size function in continuous time. Let a and c be continuous functions, and

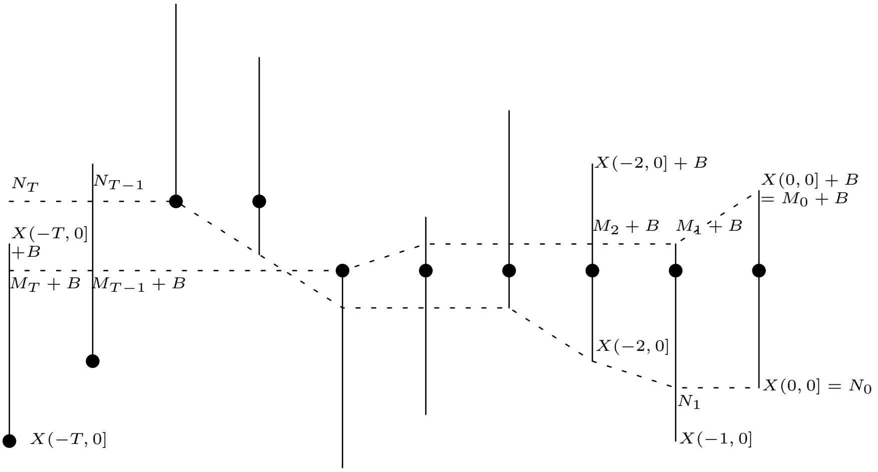

Figure 5.1: A plot of X, $M+B$ and N. The dots show the successive boundings of $X(-T,0]$. At each timestep, the position of the dot within its vertical line indicates how full the buffer is.

simply use the same expression as before, but allowing the time variables to range over real numbers rather than just integers. Convince yourself that if a and c are polygonalized versions of discrete-time processes A and C, then the two versions of the equation agree. In other words, the continuous-time definition of queue size is consistent with the discrete-time definition. We can of course extend the definition to other times: for $-T \le -t \le 0$, let

$$q_{-t}^{-T}(a,c) = \Big[x(-T,-t]\Big]_{\sup_{u\in[t,T]} n_u \wedge (m_u+B)}^{\inf_{u\in[t,T]} n_u \vee (m_u+B)} \tag{5.14}$$

where

$$\begin{aligned} m_u &= \inf_{v\in[t,u]} x(-v,-t], \\ n_u &= \sup_{v\in[t,u]} x(-v,-t], \\ x(-v,-t] &= a(-v,-t] - c(-v,-t]. \end{aligned}$$

The following remark is trivial.

Lemma 5.7 *The function $q_{-t}^{-T}(a,c)$ is continuous on $\mathcal{C}^T \times \mathcal{C}^T$.*

The next step is to extend the queue size function to deal with an infinite time horizon—to use the more intuitive boundary condition, that 'the queue was empty at time $-\infty$'. We want to define

$$q_{-t}(a,c) = \lim_{T\to\infty} q_{-t}^{-T}(a,c). \tag{5.15}$$

We need to verify that the limit exists, and then we will show that the function is continuous on $\mathcal{C}_\mu \times \mathcal{C}_\nu$ for $\mu < \nu$.

The following lemma implies that the limit exists.

Lemma 5.8 *The function $q_0^{-T}(a,c)$ is increasing (though perhaps not strictly) in T.*

Sketch proof. This is hard to see from the formula (5.14), but easy to see from a picture. Plot the two curves

$$m_T + B = \inf_{v\in[0,T]} x(-v,0]$$

and

$$n_T = \sup_{v\in[0,T]} x(-v,0].$$

The former is decreasing, the latter increasing, both are continuous, and $m_0 + B > n_0$.

Suppose the two curves cross at U. Then for any $T \geq U$, q_0^{-T} is constant (with value $n_U = m_U + B$), and thus increasing.

Suppose alternatively that either $T < U$ or the curves do not cross. Then

$$q_0^{-T} = \big[x(-T,0]\big]_{n_T}^{m_T+B} = n_T$$

which is certainly increasing. □

This is sufficient to define $q_0(a,c)$, but it doesn't reveal all the structure. To see more: Observe that q_{-t}^{-T} is increasing in buffer size B, and that if we let r_{-t}^{-T} be q_{-t}^{-T} with B set to ∞ the expression simplifies to

$$r_{-t}^{-T}(a,c) = \sup_{v\in[t,T]} x(-v,-t],$$

reassuringly the same as the infinite-buffer equations in Section 5.6. Now, we know from that section that a stable queue empties from time to time. To be precise, if $(a,c) \in \mathcal{C}_\mu \times \mathcal{C}_\nu$ and $\mu < \nu$, then there must be some time $-t$ such that $r_{-t}^{-\infty} = 0$, and hence $r_{-t}^{-T} = 0$ for all $T \geq t$. If the infinite-buffer queue is empty at $-t$, the finite-buffer queue must be empty too (we have just seen that queue size is increasing in buffer size), and so $q_{-t}^{-T} = 0$ for all $T \geq t$. Now, by rewriting (5.14),

$$q_0^{-T}(a,c) = \Big[q_{-t}^{-T} + x(-t,0] \Big]_{\sup_{u\in[0,t]} n_u \wedge (m_u+B)}^{\inf_{u\in[0,t]} n_u \vee (m_u+B)}.$$

So

$$q_0^{-T} = q_0^{-t} \quad \text{for all } T \geq t. \tag{5.16}$$

This is another justification for why the limit (5.15) exists.

The last task is to show that the queue size function is continuous. Let $(a^n, c^n) \to (a,c)$ in $\mathcal{C}_\mu \times \mathcal{C}_\nu$. First, let us restate (5.16). We have seen that if $r_{-t} = 0$ then $q_0^{-U} = q_0^{-t}$ for any $U \geq t$. In particular, if there exists a T such that $r_{-t} = 0$ for some $t \leq T$, then $q_0^{-U} = q_0^{-T}$ for any $U \geq T$. From Lemma 5.5, this time horizon T can be chosen uniformly in (a,c) and (a^n,c^n). So there exists T such that $q_0(a,c) = q_0^{-T}(a,c)$ and $q_0(a^n,c^n) = q_0^{-T}(a^n,c^n)$. Lemma 5.7 says that q_0^{-T} is continuous with respect to the topology of uniform convergence on compact intervals, so we are done. We have proved

Theorem 5.9 *The finite-buffer queue size function $q_{-t} : \mathcal{C}_\mu \times \mathcal{C}_\nu \to \mathbb{R}$ is continuous, if $\mu < \nu$.*

Exercise 5.4
In fact, for stable queues, the queue size definition (5.15) can be written more simply. Show that if $a \in \mathcal{C}_\mu$ and $c \in \mathcal{C}_\nu$ and $\mu < \nu$ then

$$\inf_{u \geq t} n_u \vee (m_u + B) = \sup_{u\geq t} n_u \wedge (m_u + B).$$

Deduce that $q_{-t}(a,c)$ is equal to this common value, and in particular that

$$q_0(a,c) = \sup_{t\geq 0} \Big(\sup_{0\leq s\leq t} x(-s,0] \Big) \wedge \Big(B + \inf_{0\leq s\leq t} x(-s,0] \Big)$$

where $x(-s,0] = a(-s,0] - c(-s,0]$. Conclude that the queue size in a finite-buffer queue is no larger than that in an infinite-buffer queue with the same service rate. ◇

5.8 Queueing Delay

How long does it take to 'drain' the queue? Let Q_0 be the queue size at time 0, and $C(0,t]$ the amount of service available in the interval $(0,t]$. The time to drain the queue is defined to be

$$W = \inf\big\{ t \in \mathbb{N}_0 : C(0,t] \geq Q_0 \big\}.$$

To apply the contraction principle, we want to express this in continuous time. The natural guess is the function

$$w(a,c) = \inf\{t \in \mathbb{R}^+ : c(0,t] \geq q_0(a,c)\}.$$

There are several problems with this. First, it involves $c(0,t]$ for $t > 0$, but we have so far only settled on a topology for the history of a process prior to time 0, not its future. Second, the function w is not even continuous. Third, because of discretization effects, the two expressions do not agree when a and c are polygonalized versions of the arrival and service processes. We will deal with these problems in turn.

The space $\mathcal{C}_\mu$ we defined in Section 5.4 only contains enough information to describe the past of a process. It is simple enough to add in the future, though: let $c_{\text{pre}}(\cdot)$ and $c_{\text{post}}(\cdot)$ be two processes in $\mathcal{C}_\mu$, and extend our convenient notation accordingly:

$$c(u,v] = c_{\text{pre}}([-u]^+) - c_{\text{pre}}([-v]^+) + c_{\text{post}}(v^+) - c_{\text{post}}(u^+).$$

Treat c as living in the space $\mathcal{C}_\mu^2$. (In fact, we can make do with a coarser topology for the future half of c; but if the past satisfies an LDP in $\mathcal{C}_\mu$ it is natural to suppose that the future does too, so we will not complicate matters by working with different topologies.)

It turns out that there are insurmountable difficulties unless the service rate is bounded below, so let us assume that $c(u,v] \geq c_0(v-u)$ for all $u \leq v$, for some $c_0 > 0$. Then the function w is continuous:

Lemma 5.10 *The function w is continuous on $\mathcal{C}_\mu \times \mathcal{X}_\nu$ for $\mu < \nu$, where $\mathcal{X}_\nu$ is the restriction of $\mathcal{C}_\nu^2$ to the set of service processes whose service rate is bounded below by $c_0 > 0$.*

Proof. Let $(a^n, c^n) \to (a,c)$ in $\mathcal{C}_\mu \times \mathcal{X}_\nu$. By continuity of the queue size function, and since $\mu < \nu$, $q_0(a^n,c^n) \to q_0(a,c)$ and $q_0(a,c) < \infty$. Since the service rate is bounded below, $w(a,c)$ is finite.

Pick any $\varepsilon > 0$ and let $u = w(a,c)$. Since $c \in \mathcal{X}_\nu$, c is continuous, so $c(0,u] = q$. Using the fact that the service rate is bounded below,

$$\begin{aligned} q_0(a,c) - c(0,u-\varepsilon] &= q_0(a,c) + \Big(c(0,u] - c(0,u-\varepsilon]\Big) - c(0,u] \\ &= c(u-\varepsilon,u] \geq c_0\varepsilon. \end{aligned}$$

Now q_0 is continuous, and $c^n \to c$ uniformly over finite intervals, so

$$q_0(a^n,c^n) - c^n(0,u-\varepsilon] \to q_0(a,c] - c(0,u-\varepsilon].$$

For n sufficiently large, the left hand side must be strictly positive, and so $w(a^n, c^n) > u - \varepsilon$. Similarly one can show that $w(a^n, c^n) < u + \varepsilon$. But ε was arbitrary, hence $w(a^n, c^n) \to w(a, c)$. □

If the service rate is not bounded below, w may not even be finite. Even when $w(a, c)$ and every $w(a^n, c^n)$ lie in some finite interval, it may still be that w is discontinuous. (Demonstrating this is left as an exercise.)

The final problem is that W is not equal to $w(\tilde{A}, \tilde{C})$ where $\tilde{A}$ and $\tilde{C}$ are the polygonalized versions of A and C. This is not in fact a serious problem. We will later study queueing delay in the regime described in Section 5.2, which considers the sequence of scaled arrival process

$$\tilde{A}^N(-t, 0] = \frac{1}{N}\tilde{A}(-Nt, 0]$$

and similarly for $\tilde{C}^N$, as $N \to \infty$. It is clear that

$$\left| w(\tilde{A}^N, \tilde{C}^N) - W/N \right| \le \frac{1}{N}$$

Therefore we will be able to use the version of the approximate contraction principle described in Example 4.6 to obtain an LDP for W/N.

To study queueing delay for discrete-time processes, of the sort described in Chapters 7 and 9, one should redefine W to be an infimum over $t \in \mathbb{R}_0^+$. Otherwise neither this approximation technique nor even Lemma 5.10 works.

5.9 Priority Queues

This is our first example of a queue which shares its capacity between several flows. Happily, the proof of continuity is trivial, much simpler than the processor sharing queue in the next section.

Define a priority queue, operating in discrete time, as follows. Let A_t^1 be the amount of high-priority work arriving at time t, let A_t^2 be the amount of low priority work, and let C_t be the amount of service available. Let Q_t^1 and Q_t^2 be the high-priority and low-priority queue sizes at time t. Their evolution is described by the recursion

$$\begin{aligned} Q_t^1 &= [Q_{t-1}^1 + A_t^1 - C_t]^+ \qquad (5.17) \\ Q_t^2 &= [Q_{t-1}^2 + A_t^2 - (C_t - Q_{t-1}^1 - A_t^1)^+]^+. \end{aligned}$$

The second equation can be rewritten

$$Q_t^1 + Q_t^2 = [Q_{t-1}^1 + Q_{t-1}^2 + A_t^1 + A_t^2 - C_t]^+. \qquad (5.18)$$

These equations are the standard Lindley equations: (5.17) for a queue fed by A^1, (5.18) for a queue fed by $A^1 + A^2$. So it is not hard to see that a consistent version of the queue-size equations in continuous time is

$$\begin{aligned} q^1_{-t} &= q_{-t}(a^1, c) \\ q^2_{-t} &= q_{-t}(a^1 + a^2, c) - q_{-t}(a^1, c) \end{aligned}$$

where $q_{-t}(a, c)$ is the normal queue size function $q_{-t}(a, c) = \sup_{u \geq t} a(-u, -t] - c(-u, -t]$.

Assume the high-priority input a^1 is an arrival process in $\mathcal{C}_{\mu^1}$, the low-priority input a^2 is an arrival process in $\mathcal{C}_{\mu^2}$, and c is a service process in $\mathcal{C}_\nu$, for some $\nu > \mu^1 + \mu^2$. Since $(a^1, c) \mapsto q_{-t}(a^1, c)$ is continuous, and $(a^1, a^2, c) \mapsto q_{-t}(a^1 + a^2, c)$ is continuous, the function $(a^1, a^2, c) \mapsto (q^1_{-t}, q^2_{-t})$ is also continuous.

5.10 Processor Sharing

Loosely speaking, a processor sharing queue divides its service between several inputs according to a weighted priority scheme. This can be used to model queueing systems in which the capacity is divided fairly between concurrent users. In this section we describe the queue size function, and also outline some new machinery—a way to show that the queue size function is continuous without having to find an explicit formula. (In fact in this case an explicit formula is known; but it is worth mentioning the new machinery nonetheless.)

> *Note.* We will not develop LDPs for processor sharing models in this book. Such results can be found in [4, 8, 22, 38, 50, 72], and also in [78] which follows our general approach. The cover illustration is of a processor sharing queue; it plots the amount of work in the queue from each input flow, and shows paths to overflow, both a simulated path and the most likely path predicted by large-buffer large deviations theory.

To explain the model precisely: a processor sharing queue (with two inputs, and constant service rate, for simplicity) evolves in discrete time according to

$$Q^i_t = \left[Q^i_{t-1} + \dot{A}^i_t - \begin{cases} p^i C & \text{if } Q^j_{t-1} + \dot{A}^j_t > p^j C \\ C - Q^j_{t-1} - \dot{A}^j_t & \text{if } Q^j_{t-1} + \dot{A}^j_t \leq p^j C \end{cases} \right]^+ \tag{5.19}$$

where the two arrival streams are A^1 and A^2, the weights are p^1 and p^2, and the service rate is C, and the equation holds for $i \neq j \in \{1,2\}$. (Recall our notation for discrete-time processes: $\dot{A}_t$ is the amount of work that arrives at time t.)

In words, each queue i is offered an amount of service p^iC. If one queue does not have enough work to make use of all this offered service, the unused service is offered to the other queue.

Some more notation: use a superscript $(\cdot)^\Sigma$ to denote sums and omit the superscript to denote pairs, so that $p^\Sigma = p^1 + p^2$ and $p = (p^1, p^2)$ etc.

Toomey [93] has found an explicit solution to this recursion, a solution which extends to continuous time. If the continuous-time arrival processes are $(a^1, a^2) \in \mathcal{C}^T \times \mathcal{C}^T$, we set $x^i(s,t] = a^i(s,t] - p^iC(t-s)$, and we impose the boundary condition that the queues are empty at time $-T$, then

$$q_0^i = \inf_{0 \le u \le T} \max \left\{ \begin{array}{l} \sup\limits_{0 \le s \le u} x^i(-s,0], \\ \sup\limits_{u \le s \le T} x^\Sigma(-s,-u] + x^i(-u,0] \end{array} \right\} \tag{5.20}$$

where the equation holds for $i \neq j$. Of course, we can also define q_{-t}^i for $0 \le t \le T$ in this way. It is tedious but not difficult to show that the continuous-time solution does indeed agree with the discrete-time recursion, i.e. that if we take a discrete-time input process, polygonalize it, and apply the continuous-time queue size functions, we get the same answer as we do from the discrete-time recursion. This is left as an exercise. (Hint: write down a recursion based on (5.20).)

It is also not difficult, and left as an exercise, to show that these functions are continuous on $\mathcal{C}^T \times \mathcal{C}^T$.

Thus, given the boundary condition $q_{-T}^i = 0$, we have defined the queue size functions q_0^i. Write this latter quantity $q_0^i(a|_{(-T,0]})$, to emphasize the boundary condition. The last step is to extend the definition to complete input sample paths in $\mathcal{C}_{\mu^1} \times \mathcal{C}_{\mu^2}$, for $\mu^1 + \mu^2 < C$, and to check that the resulting functions are continuous. As before, we will use the intuitive boundary condition that 'the queue was empty at time $-\infty$'. In other words, we would like to define

$$q_0^i(a) = \lim_{T \to \infty} q_0^i(a|_{(-T,0]}). \tag{5.21}$$

Lemma 5.11 *For $a \in \mathcal{C}_{\mu^1} \times \mathcal{C}_{\mu^2}$, and $\mu^\Sigma < C$, this limit exists, and furthermore, the functions q_0^i are continuous on $\mathcal{C}_{\mu^1} \times \mathcal{C}_{\mu^2}$.*

Sketch proof. Consider the single-server queue size function $q_{-t}(a^\Sigma)$. Since $\mu^\Sigma < C$, the single-server queue fed by a^Σ is stable, and so there exists some t such that $q_{-t}(a^\Sigma) = 0$. Fix some $T \geq t$; then $q_{-t}(a^\Sigma|_{(-U,0]}) = 0$ for $U \geq T$. (The purpose of this circumlocution about T will become clear later.)

It is easy to verify from (5.20) that if $q_0(a^\Sigma|_{(-U,0]}) = 0$ then $q_0^i(a|_{(-U,0]}) = 0$ for each i. (In fact, we would expect that the sum of the two q_0^i should equal $q_0(a^\Sigma)$, but this is harder to check.) Applying this to time $-t$, $q_{-t}^i(a|_{(-U,0]}) = 0$ for all $U \geq T$.

Now, one can rewrite $q_0^i(a|_{(-U,0]})$ as a function of the queue sizes at time $-t$, $q_{-t}^i(a|_{(-U,0]})$, and of the arrival process thereafter, $a|_{(-t,0]}$. Since the former quantity does not depend on U (for $U \geq T$), $q_0^i(a|_{(-U,0]})$ does not depend on U, and in particular

$$q_0^i(a|_{(-U,0]}) = q_0^i(a|_{(-T,0]}) \quad \text{for all } U \geq T.$$

This proves the existence of the limit (5.21).

Furthermore, it establishes continuity, by virtue of Lemma 5.5, which says that for a convergent sequence of arrival processes, one can find a uniform horizon bound T; the continuity of the finite-horizon queue size functions has already been remarked upon. □

Fluid Equations

One does not always have the ingenuity to take a discrete-time recursion and write down an explicit continuous-time solution. (For example, no analogue of (5.20) is known when there are more than two buffers.) For such occasions, there is a more powerful approach, which we now outline. First, rewrite the discrete-time equation in terms of $X^i(s,t] = A^i(s,t] - p^iC(t-s)$:

$$Q_t^i - Q_{t-1}^i = \left[\dot{X}_t^i + \begin{cases} 0 & \text{if } Q_{t-1}^j + \dot{X}_t^j > 0 \\ Q_{t-1}^j + \dot{X}_t^j & \text{if } Q_{t-1}^j + \dot{X}_t^j \leq 0 \end{cases}\right] \vee \left(-Q_{t-1}^i\right).$$

This helps to motivate the following system of integral equations (also called fluid equations): for $i \neq j$ and $t \in [-T, 0]$,

$$q_{-T}^i = 0 \quad \text{and} \quad q_t^i = \int_{s=-T}^{t} \left[\dot{x}_s^i + \begin{cases} 0 & \text{if } q_s^j > 0 \\ \dot{x}_s^j & \text{if } q_s^j = 0 \end{cases}\right]^{+(q_s^i=0)} ds \qquad (5.22)$$

where the notation $[x]^{+(y=0)}$ means

$$[x]^{+(y=0)} = \begin{cases} x^+ & \text{if } y = 0 \\ x & \text{if } y > 0. \end{cases}$$

When the arrival processes are polygonalized versions of discrete-time arrival processes, it is easy to check that this system of equations has a unique solution, which one can construct recursively on intervals, and that this solution agrees with the discrete-time recursion (5.19).

For the equations to make sense for general arrival processes, we have to assume that $(a^1, a^2) \in \mathcal{A}^T \times \mathcal{A}^T$, i.e. that the arrival processes are absolutely continuous, meaning that the instantaneous rates $\dot{x}^1_s$ and $\dot{x}^2_s$ exist almost everywhere. (Note that the slicker explicit definition (5.20) works even when the arrival processes are not absolutely continuous.) Given this assumption, Dupuis and Ramanan [38] show that the system of equations (5.22) has a unique solution, and that the solution is continuous with respect to the topology of uniform convergence.

It remains to extend the queue size functions from $\mathcal{A}^T \times \mathcal{A}^T$ to $\mathcal{A}_{\mu^1} \times \mathcal{A}_{\mu^2}$. This is very similar to Lemma 5.11—even easier, in fact, since it is easy to see from (5.22) that

$$q^\Sigma_{-T} = 0 \quad \text{and} \quad q^\Sigma_t = \int_{s=-T}^{t} \left[\dot{x}^\Sigma_s\right]^{+(q^\Sigma_s=0)}$$

and thence to relate the queue size functions for the processor sharing model to that for a single server queue.

Note. In (5.22), the process x is called the free process. The queue size process q tries to follow x, but is constrained so that $q \geq 0$. It is only constrained when it is at the boundary of its permissible region; the manner of the constraint is implicit in the equation. This is a rather general way of looking at many problems in queueing theory, referred to as Skorokhod problems. Dupuis and Ramanan [38] develop this approach.

5.11 Departures from a Queue

A rather harder problem is to show that the function which maps an input process to the corresponding output process is continuous. If we can prove this, we can use the contraction principle to deduce an LDP for the output process; and, indeed, for the traffic flow at any point in a feedforward network.

The first step is to define the output process. Consider a queue with an infinite buffer and variable service rate. Let A_t be the amount of work arriving at time t, and let C_t be the amount of service offered at time t. Define the amount of work departing the queue at time t to be

$$D_t = A_t + Q_{t-1} - Q_t. \tag{5.23}$$

As before, Q_t denotes the queue size at time t.

To express what we want to show, we need some more precise notation. As usual, write $A(-t,0]$ for the cumulative arrival process, $A(-t,0] = A_{-t+1} + \cdots + A_0$, with $A(0,0] = 0$, and $C(-t,0]$ similarly. Let a and c be the polygonalizations of A and C. As usual, the queue size at time $-t$ is

$$q_{-t} = \sup_{-u \leq -t} a(-u,-t] - c(-u,-t].$$

To make sure that the queue is well-behaved, we will assume from now on that the arrival process a lives in $\mathcal{C}_\mu$ and the service process c lives in $\mathcal{C}_\nu$, and that $\mu < \nu$.

Now we can define the departure process properly. The polygonalized departure process is simply

$$d(-t,0] = a(-t,0] + q_{-t} - q_0. \tag{5.24}$$

Persuade yourself that this is consistent with (5.23).

The work of this section is to show that $d \in \mathcal{C}_\mu$, and that the map from (a,c) to d is continuous.

Lemma 5.12 *If $a \in \mathcal{C}_\mu$ and $c \in \mathcal{C}_\nu$ and $\mu < \nu$, then $d \in \mathcal{C}_\mu$.*

Proof. We need to show that $d(t)$ is continuous in t, that $d(0) = 0$, that $\sup_t |d(t)/(t+1)|$ is finite, and that $d(t)/(t+1) \to \mu$.

For continuity: we want to show that $d(t)$, or equivalently $d(-t,0]$, is continuous in t. Rewrite it by expanding q_{-t} to give

$$d(-t,0] = c(-t,0] + \sup_{-u \leq -t} \Big\{ a(-u,0] - c(-u,0] \Big\} + \sup_{-u \leq 0} \Big\{ a(-u,0] - c(-u,0] \Big\}. \tag{5.25}$$

The first term $c(-t,0]$ is continuous. The third term does not involve t. And it is not hard to check that the second term is continuous also. So d is continuous.

It's also obvious from (5.24) that $d(0,0] = 0$.

For the last part, note that $a(-t,0]/(t+1) \to \mu$ (since $a \in \mathcal{C}_\mu$), and that q_0 is finite (by Theorem 5.3) so that $q_0/(t+1) \to 0$. It will therefore suffice to show that $q_{-t}/(t+1) \to 0$, which we do now.

Start with the expression for q_{-t}:

$$q_{-t} = \sup_{-u \leq -t} a(-u,-t] - c(-u,-t].$$

Given any $\varepsilon > 0$, there exists T such that for $t \geq T$, $|a(t) - \mu t| < \varepsilon$, and so

$$|a(-u,-t] - \mu(u-t)| < \varepsilon(t+u).$$

Similarly for c. This implies that, for $t \geq T$,

$$\begin{aligned} q_{-t} &\leq \sup_{-u \leq -t} -(\nu-\mu)(u-t) + 2\varepsilon(u+t) \\ &= \sup_{-u \leq -t} -(\nu-\mu-2\varepsilon)u + (\nu-\mu+2\varepsilon)t. \end{aligned}$$

We may assume $2\varepsilon < \nu - \mu$. Then, taking the supremum over $-u \leq -t$,

$$q_{-t} \leq -(\nu-\mu-2\varepsilon)t + (\nu-\mu+2\varepsilon)t = 4\varepsilon t.$$

Thus $q_{-t} \leq 4\varepsilon t$ for $t \geq T$. Since ε was arbitrary, $q_{-t}/(t+1) \to 0$. □

Note. If we are dealing with absolutely continuous processes, a similar corresponding result holds. We just need to argue that if a and c are absolutely continuous, then so is d.

To see this, rewrite $d(-t,0]$ as in (5.25). The first term $c(-t,0]$ is absolutely continuous. The third term does not involve t. So we just need to check that the second term is absolutely continuous. Write $x(u) = a(-u,0] - c(-u,0]$; this is absolutely continuous. Write $y(t) = \sup_{u \geq t} x(u)$. Note that for $s > t$,

$$\begin{aligned} y(s) - y(t) &= \sup_{u \geq s} x(s) - \sup_{u \geq t} x(t) \\ &= \Big[\sup_{t \leq u \leq s} x(u) - \sup_{u \geq t} x(u) \Big]^{+} \\ &\leq \Big| \sup_{t \leq u \leq s} x(u) - x(s) \Big| \\ &= |x(t') - x(s)| \end{aligned}$$

for some $t' \in [t,s]$. If $[t_i, u_i]$ is a partition of $\mathbb{R}_+$, then $[t'_i, u_i]$ is a smaller partition, in the sense of (5.4). Since x is absolutely continuous, it must be that y is absolutely continuous. Hence d is absolutely continuous.

Theorem 5.13 (Continuity of departure process) *The function* $(a,c) \mapsto d$*, defined in* (5.24)*, is a continuous function* $\mathcal{C}_\mu \times \mathcal{C}_\nu \to \mathcal{C}_\mu$*, for* $\mu < \nu$*.*

Proof. Let $a^n \to a$ and $c^n \to c$. We want to show that $d^n \to d$. Write it out:

$$\begin{aligned}\|d^n - d\| &= \sup_{t\geq 0}\Big|\frac{d(-t,0] - d^n(-t,0]}{t+1}\Big| \\ &\leq \sup_{t\geq 0}\Big|\frac{a(-t,0] - a^n(-t,0]}{t+1}\Big| + \sup_{t\geq 0}\Big|\frac{q_0 - q_0^n}{t+1}\Big| + \sup_{t\geq 0}\Big|\frac{q_{-t} - q_{-t}^n}{t+1}\Big|.\end{aligned}$$

The first term $\to 0$ since $a^n \to a$. The supremum in the second term is attained at $t = 0$, and the second term $\to 0$ since q_0 is a continuous function of (a, c). It remains to show that the third term $\to 0$.

We will adopt the strategy of breaking up the supremum into three parts: for any T,

$$\sup_{t\geq 0}\Big|\frac{q_{-t} - q_{-t}^n}{t+1}\Big| \leq \sup_{0\leq t<T}\Big|\frac{q_{-t} - q_{-t}^n}{t+1}\Big| + \sup_{t\geq T}\Big|\frac{q_{-t}}{t+1}\Big| + \sup_{t\geq T}\Big|\frac{q_{-t}^n}{t+1}\Big|.$$

The second two term first. As in Lemma 5.12, given $\varepsilon > 0$ (and assuming $2\varepsilon < \nu - \mu$), there exists a T (depending on s and c) such that for $t \geq T$,

$$q_{-t} < 4\varepsilon t.$$

Take in addition N such that for $n \geq N$, $\|a^n - a\| < \varepsilon$ and $\|c^n - c\| < \varepsilon$, so that

$$|a(-u,-t] - a^n(-u,-t]| \leq \varepsilon(t + u + 2) \quad \text{for all } t,u \tag{5.26}$$

and similarly for c. By a similar argument (and assuming $4\varepsilon < \nu - \mu$), for the same T as before, for $t \geq T$ and $n \geq N$,

$$q_{-t}^n < 8\varepsilon(t+1).$$

So given $\varepsilon > 0$ there exists a T and an N such that for $t \geq T$ and $n \geq N$

$$\sup_{t\geq T}\Big|\frac{q_{-t}}{t+1}\Big| + \sup_{t\geq T}\Big|\frac{q_{-t}^n}{t+1}\Big| < 12\varepsilon.$$

Now for the first term. With T and N as above,

$$\begin{aligned}a(-u,-t] - c(-u,-t] &\leq -(\nu-\mu)(u-t) + 2\varepsilon(t+u) \\ &= -(\nu - \mu - 2\varepsilon)u + (\nu - \mu + 2\varepsilon)t, \quad \text{and} \\ a^n(-t,-u] - c^n(-u,-t] &\leq -(\nu - \mu - 4\varepsilon)u + (\nu - \mu + 4\varepsilon)t + 2\varepsilon.\end{aligned}$$

Assume that $4\varepsilon < \nu - \mu$. Then there is a T' (depending on ε and T) such that for $-u \leq -T'$,

$$a(-u,-t] - c(-u,-t] < 0 \quad \text{and} \quad a^n(-u,-t] - c^n(-u,-t] < 0.$$

So the supremums over $-u \leq -t$ in q_{-t} and q^n_{-t} can be replaced by supremums over $-T' \leq -u \leq -t$. Now, for $-t \geq -T$ and $-T' \leq -u \leq -t$, again using (5.26),

$$\Big|\big(a(-u,-t]-c(-u,-t]\big)-\big(a^n(-u,-t]-c^n(-u,-t]\big)\Big| \leq 2\varepsilon(t+u+2) \leq 2\varepsilon(T+T'+2),$$

and putting this into the quantity we want to estimate,

$$\begin{aligned}|q_{-t} - q^n_{-t}| &= \Big|\big(\sup_{t\leq u\leq T'} a(-u,-t] - c(-u,-t]\big) - \big(\sup_{t\leq u\leq T'} a^n(-u,-t] - c^n(-u,-t]\big)\Big| \\ &\leq 2\varepsilon(T+T'+2).\end{aligned}$$

This bound does not depend on the value of $t \in [0,T]$. Thus

$$\sup_{t\leq T}\left|\frac{q_{-t} - q^n_{-t}}{t+1}\right| \leq 2\varepsilon(T+T'+2).$$

This completes the proof. □

An extension is to the case where there are several separate input flows, sharing the buffer say in a first-come–first-served fashion. A limitation is that our results only apply to queues with infinite buffers:

Example 5.5
Consider a bufferless resource with constant service rate c. Consider the arrival process a which has constant arrival rate $\mu > c/2$, so that $a(t) = \mu t$, and the collection of arrival processes a^n defined by their time-derivatives

$$\dot{a}^n_s = \begin{cases} \mu & \text{if } s < n \\ \mu + (c - \mu + \varepsilon) & \text{if } s > n \text{ and } \lfloor s \rfloor \text{ is odd} \\ \mu - (c - \mu + \varepsilon) & \text{if } s > n \text{ and } \lfloor s \rfloor \text{ is even.} \end{cases}$$

for some $0 < \varepsilon < 2\mu - c$. Observe that

$$\sup_{t\geq 0}\left|\frac{a^n(t) - a(t)}{t+1}\right| = \frac{c - \mu + \varepsilon}{n+2}$$

which implies that $\|a^n - a\| \to 0$ in $\mathcal{C}_\mu$.

We haven't formally defined the departure process for queues with finite buffers, or even for bufferless resources like this one. We will simply assert that the natural choice of departure process is

$$\dot{d}_s = \dot{a}_s \wedge c$$

and similarly for $\dot{d}^n$. So $d(t) = a(t) = \mu t$ and $d(t)/(t+1) \to \mu$, whereas

$$\frac{d^n(t)}{t+1} \to \mu - \tfrac{1}{2}\varepsilon$$

and so

$$\frac{d^n(t) - d(t)}{t+1} \to -\tfrac{1}{2}\varepsilon.$$

Thus $\|d^n - d\| \not\to 0$. ◇

We are not aware of a suitable topology for dealing with departures from queues with finite buffers.

5.12 Conclusion

We have explained the second ingredient for the contraction principle—a continuous function $A^L \mapsto f(A^L)$. We have introduced several general pieces of machinery for producing continuous functions. The general procedure we have followed is this:

i. Start with a discrete-time recursion for the queueing system of interest. If we are able, write down an explicit solution for the state of the system at time 0, given a boundary condition of the form 'the system was empty at time $-T$'. This is known as a finite-horizon solution.

ii. Produce a continuous-time definition of the quantity of interest, subject to the same finite-horizon boundary condition. If we have an explicit discrete-time solution, this part is easy. If we do not have an explicit solution, we may be able to make do by working with a system of differential equations, referred to as the *fluid equations*, which is the analogue of the discrete-time recursion. Verify that this continuous-time definition is consistent with the discrete-time solution, i.e. that it agrees on piecewise-linear inputs.

iii. Prove that the finite-horizon function is continuous with respect to the topology of uniform convergence over finite time horizons.

iv. Extend the finite-horizon definition to the infinite-horizon boundary condition, which says that ‘the queue was empty at time $-\infty$’. This can often be done by comparing the system of interest to a simple queue with an infinite buffer, noting that this simple queue must be empty at some time $-T$, deducing that the system of interest was also empty at time $-T$, and using the finite-horizon definition to work out the state of the system at time 0.

v. Prove that the infinite-horizon function found in (iv) is continuous with respect to the topology of interest. This can often be done by finding a uniform horizon $-T$, such that the system of interest must be empty at some time in $[-T, 0]$.

It may be that we are only interested in the behaviour of the system over a finite horizon, in which case steps (iv) and (v) are unnecessary. This is also known as studying the *transients* of the system, as opposed to the *steady state* behaviour. It’s clear from the above list that to understand the steady state behaviour one must first understand the transients.

Item (ii) may not always be necessary: it depends whether we want our function $f(A^L)$ to be defined on continuous-time processes A^L, or on discrete-time processes A^L. In Chapter 6 on the large-buffer scaling regime and Chapter 8 on long-range dependence, it is necessary to work in continuous time; in Chapter 7 on the many-flows scaling regime and in Chapter 9 on moderate deviations we could work in either, but we choose to work in discrete time. In this chapter we have worked in continuous time, because this yields discrete-time results as trivial corollaries.

Item (ii) refers to the case where we are not able to find an explicit solution to the discrete-time recursion, and must make do with the fluid equations. It is well known that for more complicated systems, the fluid equations do not have a unique solution; if that is so, they clearly cannot give us a definition for the quantity of interest, so we must either work harder to find an explicit solution, or content ourselves with working solely with discrete-time processes.

It remains to explain the first ingredient for the contraction principle, a large deviations principle for A^L. We do this in several different ways in the following chapters.

Chapter 6

Large-Buffer Scalings

In this chapter, we describe how the continuous mapping approach, presented in the previous chapter, can be used to obtain large-buffer asymptotics for queueing networks. The key steps are as follows.

i. We try to express the quantities of interest (e.g. queue lengths) as continuous functions of the suitably scaled inputs (e.g. sample paths of the arrival and service processes) in a suitable topology. Suitable functions have been given in Chapter 5; part of the work of this chapter is to understand how those functions relate to large-buffer scalings.

ii. Assuming that the inputs satisfy an appropriate LDP, we obtain an LDP for the quantity of interest, via the contraction principle. In Section 6.2 we give conditions under which the inputs satisfy an appropriate LDP.

iii. Typically, the rate function for this LDP will be given as the solution to a variational problem. We will show in a number of cases how to solve this variational problem, by exploiting convexity. More complicated examples where this approach has been applied, including the first-in first-out single-server queue with multiple inputs, have been studied by Majewski [63] and also in [76, 77, 78]. Nonetheless, the fact remains that explicit solutions can only be obtained in a few special cases.

> *Note.* You should ensure you are familiar with the direct proof of the large-buffer limiting result in Section 1.3 and with the introduction to Chapter 5 before proceeding.

6.1 The Space of Input Processes

We will suppose that the raw data for the network can be represented as a sequence of $\mathbb{R}^d$-valued random variables $(X(t),\ t \in \mathbb{N})$, for some fixed d. For

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 105–150, 2004.

example, in modelling a simple queue we might let $A(t)$ be the amount of work arriving in the interval $(-t,0]$ and $C(t)$ the amount of service offered in that same interval, and let $X(t) = \big(A(t), C(t)\big)$.

We will suppose that we are interested in a sequence of quantities which can be expressed as a continuous function of a sequence of scaled versions of X. To make this precise, for fixed $N \in \mathbb{N}$ let $(\tilde{X}^N(t),\ t \in \mathbb{R}_0^+)$ be the piecewise linear approximation to a scaled version of X: for $t \in \mathbb{R}_0^+$,

$$\tilde{X}(t) = \big(\lfloor t+1 \rfloor - t\big)X(\lfloor t \rfloor) + \big(t - \lfloor t \rfloor\big)X(\lfloor t+1 \rfloor)$$

and

$$\tilde{X}^N(t) = \frac{1}{N}\tilde{X}(Nt), \tag{6.1}$$

where $\lfloor t \rfloor$ denotes the integer part of t. The function $\tilde{X}$ is called the *polygonalization* of X. We will suppose we are interested in the sequence of quantities $f(\tilde{X}^N)$, for some continuous function f.

In order to talk about the continuity of f, we need to specify a topological space for the processes $\tilde{X}^N$. When $d = 1$ we will work with $\mathcal{C}_\mu$ (for some $\mu \in \mathbb{R}$) as described in Section 5.4. This is the space of continuous functions $x : \mathbb{R}_0^+ \to \mathbb{R}$ for which $x(0) = 0$ and $\lim_{t\to\infty} x(t)/(t+1) = \mu$, equipped with the topology induced by the scaled uniform norm

$$\|x\| = \sup_{t \in \mathbb{R}_0^+} \left| \frac{x(t)}{t+1} \right|. \tag{6.2}$$

We will also work with $\mathcal{A}_\mu$, the subspace of $\mathcal{C}_\mu$ consisting of absolutely continuous functions. We will also need to refer to $\mathcal{C}^T$ and $\mathcal{A}^T$, the spaces of continuous and absolutely continuous functions $x : [0,T] \to \mathbb{R}$ with $x(0) = 0$, equipped with the topology of uniform convergence. When $d > 1$ and $\mu \in \mathbb{R}^d$ we will refer to $\mathcal{C}_\mu = \mathcal{C}_{\mu_1} \times \ldots \times \mathcal{C}_{\mu_d}$ etc. in the obvious way.

If the function f is continuous, and if the sequence of processes $\tilde{X}^N$ satisfies a large deviations principle in the appropriate space, then $f(\tilde{X}^N)$ will satisfy a large deviations principle

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

$x(-t,0]$	for $x(t)$
$x(-t,-u]$	for $x(t)-x(u)$, when $t \geq u$
$x\vert_{(-t,0]}$	for the restriction of x to $[0,t]$
$\dot{x}_{-t}$	for $-dx/dt$, when $x(t)$ is indexed by $t \in \mathbb{R}$
$\dot{X}_{-t}$	for $X(t+1)-X(t)$, when $X(t)$ is indexed by $t \in \mathbb{N}$

6.2 Large Deviations for Partial Sums Processes

In order to apply the general method outlined above, we need the sequence $\tilde{X}^N$ of scaled polygonalized sample paths to satisfy an LDP in $\mathcal{C}_\mu$. We want to find conditions on the input process $(X(t),\ t \in \mathbb{N})$ under which the $\tilde{X}^N$ can be expected to satisfy such an LDP. In the rest of this section we give such conditions. For the applications in this chapter all that will matter is the main result, which we now state as a definition.

Definition 6.1 *Say that the sequence of processes* $(\tilde{X}^N,\ N \in \mathbb{N})$ *satisfies a* sample path LDP with linear geodesics*, with instantaneous rate function h, if the following hold:*

i. h is a convex rate function and $h(\mu)=0$ *for some* μ*, referred to as the* mean rate*;*

ii. $(\tilde{X}^N,\ t \in \mathbb{N})$ *satisfies an LDP in the topological* $\mathcal{C}_\mu$ *with good rate function*

$$I(x) = \begin{cases} \int_0^\infty h\big(\dot{x}(s)\big) & \textit{if } x \in \mathcal{A}_\mu \\ +\infty & \textit{otherwise.} \end{cases} \tag{6.3}$$

Note that I is convex because Λ^* is, and that Λ^* is good because I is. The meaning of the term 'linear geodesics' will be made clear in Section 6.3.

A standard example of such a process is obtained from $X(t) = Y_1 + \cdots + Y_t$ where $(Y_t,\ t \in \mathbb{N})$ is a sequence of i.i.d. random variables, and where the $\tilde{X}^N$ are the scaled polygonalized versions of X given by (6.1). Let $\Lambda(\theta) = \log E e^{\theta Y_1}$, and suppose it is finite in a neighbourhood of the origin. Then the $\tilde{X}^N$ satisfy a sample path LDP with linear geodesics, with instantaneous rate function Λ^*, and by Lemma 2.6 the mean rate is EY_1. We will now go on to detail weaker conditions under which conclusion holds.

Note. The rest of this section is rather technical. The details are interesting from the point of view of general large deviations theory, but less interesting from the point of view of applications to queueing. We suggest that it be omitted on first reading.

LDP over Finite Horizon

To start with, let $(Y_t,\ t \in \mathbb{N})$ be a sequence of i.i.d. random variables taking values in $\mathbb{R}$, and let

$$X(t) = Y_1 + \cdots + Y_t.$$

Let Λ be the cumulant generating function of Y_0 and Λ^* its convex conjugate,

$$\Lambda(\theta) = \log Ee^{\theta Y_1} \ \text{ and } \ \Lambda^*(x) = \sup_{\theta\in\mathbb{R}} \theta x - \Lambda(\theta).$$

We saw in Chapter 2 that, by Cramér's Theorem, the sequence $X(t)/t$ satisfies a large deviations principle in $\mathbb{R}$ with rate function Λ^*.

Now fix positive constants T and $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_k = T$, and consider the sequence of random variables

$$Z^N = \frac{1}{N}\Big(X\big(\lfloor \alpha_1 N\rfloor\big), X\big(\lfloor \alpha_2 N\rfloor\big) - X\big(\lfloor \alpha_1 N\rfloor\big), \\ \ldots, X\big(\lfloor \alpha_k N\rfloor\big) - X\big(\lfloor \alpha_{k-1} N\rfloor\big)\Big).$$

One can show that, for each $j \in \{1, \ldots, k\}$, the sequence

$$\frac{X\big(\lfloor \alpha_j N\rfloor\big) - X\big(\lfloor \alpha_{j-1} N\rfloor\big)}{N}, \quad N \in \mathbb{N},$$

satisfies an LDP in $\mathbb{R}$ with rate function

$$(\alpha_j - \alpha_{j-1})\Lambda^*\Big(\frac{x}{\alpha_j - \alpha_{j-1}}\Big).$$

(See Exercise 2.11.) Since these random variables for different values of j are independent of each other, it is natural to expect that $(Z^N,\ N \in \mathbb{N})$ satisfies an LDP in $\mathbb{R}^k$ with rate function

$$I\Big((x_k, \ldots, x_1)\Big) = \sum_{j=1}^{k} (\alpha_j - \alpha_{j-1})\Lambda^*\Big(\frac{x_j}{\alpha_j - \alpha_{j-1}}\Big), \tag{6.4}$$

and this is indeed the case.

Loosely speaking, if we look at the partial means of the sequence Y_t over disjoint intervals, then any finite collection of such partial means satisfies an LDP. Not only does the sample mean $X(NT)/NT$ concentrate around EX_1, but also the *process* $(X(Nt)/N,\ 0 \le t \le T)$ concentrates around the straight line of slope EX_1. And just as we can ask how likely the sample mean is to be 'close to' any $x \in \mathbb{R}$, so can we ask how likely the sample path is to lie 'close to' an arbitrary curve $x : [0, T] \to \mathbb{R}$.

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 $x : [0,T] \to \mathbb{R}$ 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 $X(t) = Y_1 + \cdots + Y_t$, 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*

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

> *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 $x : \mathbb{R}_0^+ \to \mathbb{R}$ for which $x(0) = 0$, equipped with the topology of uniform convergence on compact intervals, with good rate function

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

where I_T is the rate function for $\tilde{X}^N|_{[0,T]}$. By non-negativity of Λ^*, the supremum is

$$I(x) = \begin{cases} \int_0^\infty \Lambda^*\bigl(\dot{x}(s)\bigr)\,ds & \text{if } x \text{ is absolutely continuous} \\ \infty & \text{otherwise.} \end{cases}$$

Strengthening the Topology

However, this topology is not fine enough for many queueing applications, since even the queue size function for a single-server queue with an infinite buffer is not continuous with respect to it (as we noted in Chapter 5, Example 5.2). This has prompted the consideration of finer topologies, for example by Dobrushin and Pechersky [30] and Ganesh and O'Connell [41]. The next theorem follows the latter.

Let $(Y_t,\ t \in \mathbb{N})$ be a stationary sequence of $\mathbb{R}^d$-valued random variables, and let $X(t) = Y_1 + \cdots + Y_t$. Let $\tilde{X}^N$ be the scaled polygonalized version of X, as in (6.1). Suppose that for each $\theta \in \mathbb{R}^d$ the limit

$$\Lambda(\theta) = \lim_{N\to\infty} \frac{1}{N} \log E \exp\bigl(N\theta \cdot \tilde{X}^N(1)\bigr) \tag{6.6}$$

exists as an extended real number, and that, for each $T > 0$, the sequence $(\tilde{X}^N|_{[0,T]},\ N \in \mathbb{N})$ satisfies an LDP in $\mathcal{C}^T$ with good rate function

$$I_t(x) = \begin{cases} \int_0^T \Lambda^*\bigl(\dot{x}(s)\bigr)\,ds & \text{if } x \in \mathcal{A}^T \\ \infty & \text{otherwise} \end{cases} \tag{6.7}$$

where Λ^* is the convex dual of Λ.

Theorem 6.2 *In the above setting, if Λ is differentiable in a neighbourhood of the origin, then $\tilde{X}^N$ satisfies a sample path LDP with linear geodesics, with mean rate $\mu = \nabla\Lambda(0)$ and instantaneous rate function Λ^*, that is, it satisfies the LDP in the topological space $\mathcal{C}_\mu$ with good rate function*

$$I(x) = \begin{cases} \int_0^\infty \Lambda^*\bigl(\dot{x}(s)\bigr)\,ds & \textit{if } x \in \mathcal{A}_\mu \\ \infty & \textit{otherwise.} \end{cases} \tag{6.8}$$

Proof. From the assumptions of the theorem, and using the Dawson-Gärtner theorem for projective limits, the sequence $\tilde{X}^N$ satisfies an LDP in the space of continuous functions $x : \mathbb{R}_0^+ \to \mathbb{R}^d$ for which $x(0) = 0$, equipped with the topology of uniform convergence on compacts, with good rate function $I(x)$ given by (6.8).

Let $\mu = \nabla\Lambda(0)$. Lemma 6.3 shows that the effective domain of I is contained in $\mathcal{C}_\mu$, and that $P(\tilde{X}^N \in \mathcal{C}_\mu) = 1$; and so by Lemma 4.9, $\tilde{X}^N$ satisfies an LDP in the set $\mathcal{C}_\mu$ equipped with the topology of uniform convergence on compacts, with good rate function I. Denote this space by $(\mathcal{C}_\mu, \tau_p)$ (where τ_p stands for projective limit topology).

It remains to strengthen the topology to the scaled uniform topology, i.e. the topology induced by the norm $\|\cdot\|$ given by (6.2). The space $\mathcal{C}_\mu$ is defined to have this topology, but we will emphasize the fact by writing it as $(\mathcal{C}_\mu, \|\cdot\|)$. We will strengthen the LDP from $(\mathcal{C}_\mu, \tau_p)$ to $(\mathcal{C}_\mu, \|\cdot\|)$ by appealing to the inverse contraction principle. We need to find, for each $\alpha \in \mathbb{R}^+$, a compact set K_α in $(\mathcal{C}_\mu, \|\cdot\|)$ for which

$$\lim_{\alpha\to\infty} \limsup_{N\to\infty} \frac{1}{N} \log P(\tilde{X}^N \notin K_\alpha) = -\infty. \tag{6.9}$$

We construct these sets as follows. We know that $\tilde{X}^N|_{[0,1]}$ satisfies an LDP in $\mathcal{C}^1$ with good rate function I_1, and that $\mathcal{C}^1$ is a Polish space. It follows that $\tilde{X}^N|_{[0,1]}$ is exponentially tight, i.e. that there exists a family of compact sets $L'_\alpha \subset \mathcal{C}^1$ for which

$$\limsup_{N\to\infty} \frac{1}{N} \log P\big(\tilde{X}^N|_{[0,1]} \notin L'_\alpha\big) \le -\alpha.$$

From this, define the set L_α by

$$L_\alpha = \bigcap_{n\in\mathbb{N}} L'_\alpha(n) \quad \text{where} \quad L'_\alpha(n) = \Big\{x : \frac{1}{n} x^{\circlearrowright n} \in L'_\alpha\Big\} \subset \mathcal{C}^n$$

and where the speeded-up version $x^{\circlearrowright n}$ is defined by $x^{\circlearrowright n}(t) = x(nt)$. (The sets $L'_\alpha(n)$ are stretched out versions of L'_α, and they are compact in $\mathcal{C}^n$.) Next take some process $(\delta_t,\ t \in \mathbb{R})$ and define

$$M_\alpha = \Big\{x : \Big|\frac{x(t)}{t} - \mu\Big| \le \alpha\delta_t \text{ for all } t \ge 1\Big\}$$

Finally define

$$K_\alpha = \mathcal{C}_\mu \cap L_\alpha \cap M_\alpha.$$

Lemma 6.4 shows that if $\delta_t \to 0$ as $t \to \infty$ then K_α is compact in $(\mathcal{C}_\mu, \|\cdot\|)$. Lemma 6.5 presents such a process δ for which (6.9) holds. Thus $\tilde{X}^N$ is exponentially tight in $(\mathcal{C}_\mu, \|\cdot\|)$. This completes the proof of the theorem.□

Lemma 6.3 *In the setting of Theorem 6.2, if $\mu = \nabla\Lambda(0)$ and I is given by* (6.8)*, then the effective domain of I is contained in $\mathcal{C}_\mu$ and also $P(\tilde{X}^N \in \mathcal{C}_\mu) = 1$.*

Proof. Let x belong to the effective domain of I. Then x is absolutely continuous. Since Λ^* is non-negative and convex, by Jensen's inequality

$$t\Lambda^*\Big(\frac{x(t)}{t}\Big) \le I(x).$$

This holds for all $t > 0$, and so $\Lambda^*\big(x(t)/t\big) \to 0$ as $t \to \infty$. Now, by the assumption that Λ is differentiable at the origin, Λ^* has a unique zero at $\mu = \nabla\Lambda(0)$. Hence $x(t)/t \to \mu$ as $t \to \infty$, and so $x \in \mathcal{C}_\mu$.

Next, by the assumption that $\tilde{X}^N|_{[0,1]}$ satisfies an LDP in $\mathcal{C}^1$, and using the contraction principle, $\tilde{X}^N(1)$ satisfies an LDP in $\mathbb{R}^d$ with rate function Λ^*. We have already seen that Λ^* has a unique zero at $\mu = \nabla\Lambda(0)$. Hence, for any $\varepsilon > 0$ there is a $\delta > 0$ such that

$$P\big(|\tilde{X}^N(1) - \mu| > \varepsilon\big) < e^{-N\delta}$$

for all N sufficiently large. Thus, by the Borel-Cantelli lemma, $\tilde{X}^N = X(N)/N$ converges to μ almost surely as $N \to \infty$. It is then immediate that

$$\lim_{t\to\infty} \frac{\tilde{X}^N(t)}{1+t} = \mu \quad \text{almost surely}$$

i.e. that $P(\tilde{X}^N \in \mathcal{C}_\mu) = 1$. □

Lemma 6.4 *In the setting of Theorem 6.2, with the sets K_α as defined there, if $\delta(t) \to 0$ as $t \to \infty$ then K_α is compact in $\mathcal{C}_\mu$.*

Proof. By Tychonoff's theorem, the set L_α is compact in the projective limit topology, and so $L_\alpha \cap \mathcal{C}_\mu$ is compact in the space $(\mathcal{C}_\mu, \tau_p)$. Hence, given any sequence x^n in $K_\alpha \subset L_\alpha \cap \mathcal{C}_\mu$, we can find a subsequence x^j which converges to some x, in this topology. We will show that $x \in K_\alpha$, and that $x^j \to x$ in $(\mathcal{C}_\mu, \|\cdot\|)$. In other words, we will have shown that K_α is sequentially compact. Since K_α is a metric space, this proves compactness.

Since $x^j \to x$ uniformly on compact intervals,

$$\lim_{j\to\infty} \sup_{t\in[0,T]} \left| \frac{x^j(t)}{1+t} - \frac{x(t)}{1+t} \right| = 0 \quad \text{for every } T > 0.$$

Also, since $x^j \in K_\alpha$,

$$\Big|\frac{x^j(t)}{t} - \mu\Big| \leq \alpha\delta_t \quad \text{for all } t \geq 1$$

Therefore

$$\Big|\frac{x(t)}{t} - \mu\Big| \leq \alpha\delta_t \quad \text{for all } t \geq 1$$

as well, i.e. $x \in K_\alpha$.

Moreover, for any $\varepsilon > 0$ we can choose $T > 0$ such that $\alpha\delta_t < \varepsilon$ for all $t > T$. Hence, for j sufficiently large,

$$\begin{aligned}\|x^j - x\| &\leq \sup_{t \leq T} \Big|\frac{x^j(t)}{1+t} - \frac{x(t)}{1+t}\Big| + \sup_{t>T}\Big|\frac{x^j(t)}{1+t} - \frac{x(t)}{1+t}\Big| \\ &\leq \varepsilon + 2\alpha\delta_T < 3\varepsilon.\end{aligned}$$

So $x^j \to x$ in $(\mathcal{C}_\mu, \|\cdot\|)$. □

Lemma 6.5 *In the setting of Theorem 6.2, with the sets K_α as defined there, there is a function δ such that $\delta(t) \to 0$ as $t \to \infty$, and also*

$$\lim_{\alpha\to\infty} \limsup_{N\to\infty} \frac{1}{N} \log P(\tilde{X}^N \notin K_\alpha) = -\infty. \tag{6.10}$$

Proof. We will proceed in four stages. First we will show that the sets L_α are suitably big, i.e. that the limit (6.10) holds for the family of sets L_α. Then, working in one dimension, we will explain how to choose δ_t, and prove that (6.10) holds for the M_α. Since

$$K_\alpha = \mathcal{C}_\mu \cap L_\alpha \cap M_\alpha$$

it follows that (6.10) holds as it is stated, for the K_α. Finally we extend the argument to $d > 1$ dimensions.

The L_α are suitably big. By choice of L'_α,

$$\limsup_{N\to\infty} \frac{1}{N} \log P\big(\tilde{X}^N|_{[0,1]} \notin L'_\alpha\big) \leq -\alpha.$$

Thus, given $\varepsilon > 0$, there is an N_0 such that for $N > N_0$

$$P\big(\tilde{X}^N|_{[0,1]} \notin L'_\alpha\big) \leq e^{-N(\alpha-\varepsilon)}.$$

Thus

$$\begin{aligned}
&\limsup_{N\to\infty}\frac{1}{N}\log P\big(\tilde{X}^N\notin L_\alpha\big)\\
&\qquad=\limsup_{N\to\infty}\frac{1}{N}\log P\big(\tilde{X}^N|_{[0,n]}\notin L'_\alpha(n)\text{ for some }n\big)\\
&\qquad\le\limsup_{N\to\infty}\frac{1}{N}\log\sum_{n=1}^{\infty}P\big(\tilde{X}^N|_{[0,n]}\notin L'_\alpha(n)\big)\\
&\qquad=\limsup_{N\to\infty}\frac{1}{N}\log\sum_{n=1}^{\infty}P\big(\tilde{X}^{Nn}|_{[0,1]}\notin L'_\alpha\big)\\
&\qquad\qquad\qquad\text{(by definition of }L'_\alpha(n))\\
&\qquad\le\limsup_{N\to\infty}\frac{1}{N}\log\sum_{n=1}^{\infty}e^{-Nn(\alpha-\varepsilon)}\\
&\qquad=\limsup_{N\to\infty}\frac{1}{N}\log\frac{1}{e^{N(\alpha-\varepsilon)}-1}\\
&\qquad\le-(\alpha-\varepsilon).
\end{aligned}$$

Since ε was arbitrary, this gives

$$\limsup_{N\to\infty}\frac{1}{N}\log P\big(\tilde{X}^N\notin L_\alpha\big)\le-\alpha.$$

Choice of δ in one dimension. Assume for now that X is a real-valued process, i.e. that we are working in dimension $d=1$. Define

$$\Lambda_N(\theta)=\frac{1}{N}\log E\exp\big(N\theta\tilde{X}^N(1)\big)\quad\text{for }\theta\in\mathbb{R}.$$

By assumption, the extended real valued functions Λ_N converge pointwise to Λ, which is differentiable, hence finite and continuous, in a neighbourhood of the origin. Let $\phi>0$ be such that Λ is differentiable on $|\theta|\le\phi$. Clearly Λ_N is also finite at $\theta=\pm\phi$ for N sufficiently large. By Lemma 2.3 the scaled cumulant generating function Λ_N is convex and continuous on the interior of its effective domain. Thus there exists a positive $\phi'<\phi$ such that Λ_N is continuous on $|\theta|\le\phi'$, for N sufficiently large. By continuity, the pointwise convergence $\Lambda_N\to\Lambda$ must be uniform on $|\theta|\le\phi'$. In other words, if we set

$$\varepsilon_n=\sup_{m\ge n}\sup_{|\theta|\le\phi'}\big|\Lambda_m(\theta)-\Lambda(\theta)\big|$$

then $\varepsilon_n \downarrow 0$ as $n \to \infty$. Now define a sequence θ_n by

$$\theta_n = \bigl(\sqrt{\varepsilon_n} + \nu_n\bigr) \wedge \phi' \quad \text{where} \quad \nu_n = \sqrt{\frac{1+\log n}{n}}.$$

Clearly $\theta_n \downarrow 0$ as $n \to \infty$. From these we can finally define

$$\begin{aligned}\delta_t &= \frac{\Lambda(\theta_n)}{\theta_n} + \frac{\Lambda(-\theta_n)}{\theta_n} + \frac{\varepsilon_n}{\theta_n} + \nu_n \quad \text{where} \quad n = \lfloor t \rfloor \\ &= \frac{\Lambda(\theta_n) - \mu\theta_n}{\theta_n} + \frac{\Lambda(-\theta_n) + \theta_n\mu}{\theta_n} + \frac{\varepsilon_n}{\theta_n} + \nu_n.\end{aligned}$$

In this last expression, the first two terms both decrease to 0 as $n \to \infty$, by convexity and differentiability of Λ; the third term decreases to 0, as one can see by substituting in the definition of θ_n; and the last term clearly decreases to 0. Thus $\delta_t \downarrow 0$ as $t \to \infty$.

The M_α are suitably big, in one dimension. We want to show

$$\limsup_{N\to\infty} \frac{1}{N} \log P\Bigl\{\Bigl|\frac{\tilde{X}^N(t)}{t} - \mu\Bigr| > \alpha\delta_t \text{ for some } t \geq 1\Bigr\} \qquad (6.11)$$
$$\to -\infty \quad \text{as } \alpha \to \infty.$$

If such a limit holds for both the events

$$\frac{\tilde{X}^N(t)}{t} - \mu > \alpha\delta_t \quad \text{and} \quad \frac{\tilde{X}^N(t)}{t} - \mu < \alpha\delta_t \qquad (6.12)$$

then (6.11) is true, by the principle of the largest term. We will give the proof for the first of these; the proof for the second is similar. First write down the probability we wish to estimate:

$$\begin{aligned}&P\Bigl(\frac{\tilde{X}^N(t)}{t} - \mu > \alpha\delta_t \text{ for some } t \geq 1\Bigr) \qquad (6.13)\\ &\qquad\qquad \leq P\Bigl(\tilde{X}^N(t) > \alpha t\delta_t + \mu t \text{ for some } t \geq 1\Bigr)\end{aligned}$$

Because of the scaling used in its definition, the polygonalized process $\tilde{X}^N(t)$ is linear over intervals of length $1/N$; also δ_t is constant over intervals of length 1. Thus it suffices to check the condition at the endpoints of intervals of length $1/N$, that is, at points $t = k + i/N$ where $i, k \in \mathbb{N}$ and $0 \leq i < N$. This gives

$$(6.13) \leq \sum_{k=1}^{\infty} \sum_{i=0}^{N-1} P\Bigl(X^N(t) > \alpha t\delta_t + \mu t\Bigr)$$

$$\text{(where by writing } t \text{ we mean } k+i/N)$$

$$\leq \sum_{k=1}^{\infty}\sum_{i=0}^{N-1} P\Big(X(Nk+i) > (Nk+i)\Big[\alpha\delta_k + \mu\Big]\Big)$$

$$\text{(using the fact that } \delta_{k+i/N} = \delta_k)$$

$$\leq \sum_{k=1}^{\infty}\sum_{i=0}^{N-1} \exp\Big(-(Nk+i)\Big[\theta_k\alpha\delta_k - \Lambda_{Nk+i}(\theta_k) + \mu\theta_k\Big]\Big)$$

$$\text{(by Chernoff's bound.)}$$

(To estimate the probability associated with the lower bound part of (6.12), we would use $-\theta_k$ rather than θ_k in Chernoff's bound.) A typical term in brackets $[\cdot]$ in this expression is

$$\begin{aligned}
&\theta_k\alpha\delta_k - \Lambda_{Nk+i}(\theta_k) + \mu\theta_k \\
&\quad \geq \theta_k\alpha\delta_k - \Lambda(\theta_k) + \mu\theta_k - \varepsilon_k \quad \text{(by definition of } \varepsilon_k) \\
&\quad = \alpha\Big[\Lambda(\theta_k) - \theta_k\mu + \Lambda(-\theta_k) + \theta_k\mu + \varepsilon_k + \theta_k\nu_k\Big] \\
&\qquad\qquad - \Big[\Lambda(\theta_k) - \theta_k\mu\Big] - \varepsilon_k \quad \text{(using definition of } \delta_k) \\
&\quad = (\alpha-1)\Big[\Lambda(\theta_k) - \theta_k\mu + \varepsilon_k\Big] + \alpha\Big[\Lambda(-\theta_k) + \theta_k\mu + \theta_k\nu_k\Big] \\
&\quad \geq \alpha\theta_k\nu_k \quad \text{(for } \alpha \geq 1\text{, since } \Lambda(\theta) - \theta\mu \geq 0 \text{ by convexity)} \\
&\quad \geq \alpha\nu_k^2 \quad \text{(for } k \text{ sufficiently large that } \theta_k < \phi').
\end{aligned}$$

Let k_0 be such that this last condition holds for $k > k_0$. Now we can bound the terms in the sum we derived from (6.13), to find that

$$\begin{aligned}
(6.13) &\leq \sum_{k=1}^{k_0}\sum_{i=0}^{N-1} e^{-(Nk+i)\alpha\theta_k\nu_k} + \sum_{k>k_0}\sum_{i=0}^{N-1} e^{-(Nk+i)\alpha\nu_k^2} \\
&\leq N\sum_{k=1}^{k_0} e^{-Nk\alpha\theta_k\nu_k} + N\sum_{k>k_0} e^{-Nk\alpha\nu_k^2} \\
&= N\sum_{k=1}^{k_0} e^{-Nk\alpha\theta_k\nu_k} + Ne^{-N\alpha}\sum_{k>k_0} k^{-N\alpha} \quad \text{(by definition of } \nu_k).
\end{aligned}$$

By the principle of the largest term, and using (3.7),

$$\limsup_{N\to\infty}\frac{1}{N}\log(6.13) \leq -\Big(\Big[\inf_{k\leq k_0}\alpha k\theta_k\nu_k\Big] \wedge \Big[\alpha + \alpha\log(k_0+1)\Big]\Big)$$

$$\to -\infty \quad \text{as } \alpha \to \infty.$$

This proves that the sets M_α satisfy (6.10).

Extension to $d > 1$ dimensions. To prove that the M_α satisfy (6.10) in one dimension, we needed to estimate the probability

$$P\Big(\Big|\frac{\tilde{X}^N(t)}{t} - \mu\Big| > \alpha\delta_t\Big) \tag{6.14}$$

We did this by splitting it into two parts, corresponding to

$$\frac{\tilde{X}^N(t)}{t} - \mu > \alpha\delta_t \quad \text{and} \quad \frac{\tilde{X}^N(t)}{t} - \mu < \alpha\delta_t$$

and proving that the probability of each of these two events is suitably big, in the sense that (6.10) holds.

In $\mathbb{R}^d$, any natural norm is equivalent to the componentwise supremum norm $|x| = \max_{i \leq d} x_i$. Under this norm we can split (6.14) into $2d$ parts, two parts for each component of $\tilde{X}^N(t)$. The one-dimensional proof given above applies without modification to each of these parts. We end up constructing a different function $\delta_t(i)$ for each component, and thus a different set $M_\alpha(i)$. By taking $\delta_t = \max_i \delta_t(i)$, and defining M_α from it, we ensure that the M_α satisfy (6.10). □

6.3 Linear Geodesics

All the quantities that we study in this chapter can be expressed as functions of the sequence $(\tilde{X}^N,\ N \in \mathbb{N})$ of scaled input processes, where each $\tilde{X}^N$ is a function $\mathbb{R} \to \mathbb{R}^d$. We will assume that this sequence satisfies a large deviations principle with linear geodesics, with instantaneous rate function $\Lambda^* : \mathbb{R}^d \to \mathbb{R}$. Let μ be the mean rate. The rate function for $\tilde{X}^N$, which is good and convex, is as specified by (6.3):

$$I(x) = \begin{cases} \int_0^\infty \Lambda^*\big(\dot{x}(s)\big) & \text{if } x \in \mathcal{A}_\mu \\ +\infty & \text{otherwise.} \end{cases}$$

By the contraction principle, if $\mathcal{Y}$ is a Hausdorff topological space and if $f : \mathcal{C}_\mu \to \mathcal{Y}$ is continuous, then $f(\tilde{X}^N)$ satisfies an LDP in $\mathcal{Y}$ with good rate function

$$J(\xi) = \inf_{\substack{x \in \mathcal{A}_\mu: \\ f(x) = \xi}} \int_0^\infty \Lambda^*\big(\dot{x}(s)\big)\, ds. \tag{6.15}$$

In general, it is very hard to find an explicit solution to this variational problem.

The basic tool which we use (in this chapter) to simplify the rate function is Jensen's inequality. To illustrate this, consider the simple case where we are interested in $f(x) = x(T)$ and $d = 1$. This function is certainly continuous, so the contraction principle applies. Now, take any path $x \in \mathcal{A}_\mu$ with $x(T) = \xi$. Construct a new path $x' \in \mathcal{A}_\mu$, the *straightened* version of x, by

$$\dot{x}'(t) = \begin{cases} \xi/T & \text{if } 0 \le t < T \\ \mu & \text{otherwise} \end{cases}$$

It is the case that $I(x) \ge I(x')$. To see this, note that

$$\begin{aligned} \int_0^T \Lambda^*\big(\dot{x}'(t)\big)\,dt &= T\Lambda^*(\xi/T) \\ &= T\Lambda^*\left(\frac{1}{T}\int_0^T \dot{x}(t)\,dt\right) \le T\frac{1}{T}\int_0^T \Lambda^*\big(\dot{x}(t)\big)\,dt \end{aligned}$$

using Jensen's inequality; and also that

$$\int_T^\infty \Lambda^*\big(\dot{y}(t)\big)\,dt = 0 \le \int_T^\infty \Lambda^*\big(\dot{x}(t)\big)\,dt.$$

Thus $J(\xi) \ge I(x') = T\Lambda^*(\xi/T)$. On the other hand, x' satisfies $x'(T) = \xi$, and so $J(\xi) \le I(x')$. Therefore $J(\xi) = T\Lambda^*(\xi/T)$.

> *Note.* In particular, if $\tilde{X}^N$ is the scaled polygonalized version of X, where $X(t) = Y_1 + \cdots + Y_t$, and $f(x) = x(1)$, then $f(\tilde{X}^N) = X(N)/N$; and by the above it satisfies an LDP with rate function $\Lambda^*(\xi)$. Compare this result to Cramér's theorem, Theorem 2.8.

The path x' which achieves the minimum in the above variational problem has constant gradient ξ/T on the interval $(0, T)$ (and gradient μ thereafter). This path can be interpreted as the *most likely path* to satisfy the constraint $f(x) = \xi$, in the sense that, conditional on $f(\tilde{X}^N) = \xi$, $\tilde{X}^N$ lies in a small neighbourhood of y with high probability. (This idea, that rare events occur in the most likely way, is made precise in Lemma 4.2.)

Thus, given that $\tilde{X}^N$ takes an extreme value, it is likely to have got there along a straight line. This basic property, known as the *linear geodesics property*, considerably simplifies many network problems.

Example 6.1

What is the range of a Brownian bridge? A Brownian bridge is a Brownian motion over the interval $[0, 1]$, conditioned to be 0 at the right endpoint

$t = 1$. An easy way to construct a Brownian bridge is to take a standard Brownian motion $B(t)$ and set $X(t) = B(t) - tB(1)$. Then X is a Brownian bridge. Its vertical span is

$$R = \max_{t \in [0,1]} X(t) - \min_{t \in [0,1]} X(t).$$

(A Brownian motion has continuous sample paths, almost surely, and so the maximum and minimum are attained.) How can we use large deviations to estimate $P(R > r)$?

The starting point is a large deviations principle for Brownian motion. Schilder's theorem says that if $B(t)$ is a Brownian motion and we set $B^N(t) = B(t)/\sqrt{N}$, then the sequence $(B^N,\ N \in \mathbb{N})$ satisfies a large deviations principle in $\mathcal{C}^1$ with good rate function

$$I(b) = \begin{cases} \frac{1}{2}\int_0^1 \dot{b}_t^2\,dt & \text{if } b \in \mathcal{A}^1 \\ \infty & \text{otherwise.} \end{cases}$$

See Dembo and Zeitouni [25, Theorem 5.2.3] for a proof. Now R is clearly a continuous function of X, and X is a continuous function of B, so R is a continuous function f of B. By the contraction principle, we immediately obtain an LDP for $R/\sqrt{N}$ with good rate function

$$J(r) = \inf\Big\{ I(b) : b \in \mathcal{C}^1, f(b) = r \Big\} \tag{6.16}$$

We will now solve this variational problem, using the linear geodesics property.

Suppose we have a path $b \in \mathcal{A}^1$ which satisfies the constraint $f(b) = r$. Let $\beta = b(1)$ and $m = \min_t \big(b(t) - t\beta\big)$ and $M = \max_t \big(b(t) - t\beta\big)$. Suppose that the minimum is attained at u and the maximum at v, and suppose for now that $0 < u < v < 1$. Construct a straightened path b' by

$$\dot{b}'(t) = \begin{cases} m/u + \beta & \text{for } t < u \\ (M-m)/(v-u) + \beta & \text{for } u < t < v \\ -M/(1-v) + \beta & \text{for } v < t \end{cases}$$

Now $I(b) \geq I(b')$ by Jensen's inequality. If $0 < v < u < 1$, or if u or v were extreme, we would construct b' slightly differently, but in all cases we can construct a path b' for which $I(b) \geq I(b')$. Thus we obtain a lower bound on $J(r)$.

On the other hand, suppose we are given β, m, M, u and v, such that $r = M - m$. The same construction yields a path b' for which $f(b') = r$, and thus we obtain an upper bound on $J(r)$ which agrees with the lower bound.

Thus we can reduce the general variational problem (6.16) to the simpler optimization problem

$$\begin{aligned}
J(r) &= \tfrac{1}{2} \inf_{\substack{m,M\in\mathbb{R}:\\ r=M-m}} K(m,M) \wedge L(m,M) \\
K(m,M) &= \inf_{0<u<v<1} u\big(\beta + \frac{m}{u}\big)^2 + (v-u)\big(\beta + \frac{M-m}{v-u}\big)^2 \\
&\qquad\qquad + (1-v)\big(\beta - \frac{M}{1-v}\big)^2 \\
L(m,M) &= \inf_{0<v<u<1} v\big(\beta + \frac{M}{v}\big)^2 + (u-v)\big(\beta - \frac{M-m}{u-v}\big)^2 \\
&\qquad\qquad + (1-u)\big(\beta - \frac{m}{1-u}\big)^2
\end{aligned}$$

This is simple to solve. The optimum has $\beta = 0$, and there are many possible optimal values of the other variables. The optimum value is simply $J(r) = 2r^2$.

Now we can answer the original question. From the LDP for $R/\sqrt{N}$,

$$\lim_{N\to\infty} \frac{1}{N} \log P(R/\sqrt{N} > r) = -2r^2.$$

(The LDP specified lower and upper bounds, and for this event the bounds agree.) Rewriting,

$$\lim_{r\to\infty} \frac{1}{r^2} \log P(R > r) = -2.$$

Note that, because there are many paths which are optimal for (6.16), there is no single “most likely path”. ◇

6.4 Queues with Infinite Buffers

Consider the single-server queue. Let $A(-t,0]$ be the amount of work arriving in the interval $(-t,0]$, and let $C(-t,0]$ be the amount of service offered, for $t \in \mathbb{N}$. (We've switched to the extended notation, described in Section 6.1, since it is more suggestive.) Let $X(-t,0] = A(-t,0] - C(-t,0]$. Let $\tilde{X}$ be the polygonalized version of X, and let

$$\tilde{X}^N(-t,0] = \frac{1}{N}\tilde{X}(-Nt,0] \quad \text{for } t \in \mathbb{R}_0^+.$$

In Chapter 1, we saw that the queue length at time 0 is given by

$$Q_0 = \sup_{t \in \mathbb{N}_0} X(-t, 0]. \tag{6.17}$$

Equivalently

$$Q_0 = \sup_{t \in \mathbb{R}_0^+} \tilde{X}(-t, 0]$$

and so

$$Q_0/N = f(\tilde{X}^N) \quad \text{where} \quad f(x) = \sup_{t \in \mathbb{R}_0^+} x(-t, 0]. \tag{6.18}$$

Assume that the sequence $(\tilde{X}^N,\ N \in \mathbb{N})$ satisfies the sample path large deviations principles with linear geodesics, with some instantaneous rate function Λ^* and mean rate σ. Suppose that $\sigma < 0$. Note that $\tilde{X}^N$ satisfies an LDP in $\mathcal{C}_\sigma$, hence must almost surely lie in $\mathcal{C}_\sigma$, and hence that $f(\tilde{X}^N)$ is almost surely finite, by Theorem 5.3.

> *Note.* The typical setting is this. Let $X(-t, 0] = Y_{-t+1} + \cdots + Y_0$ for some i.i.d. sequence $(Y_t,\ t \in \mathbb{Z})$. Let Λ be the cumulant generating function for Y_1. As mentioned in Section 6.2, $\tilde{X}^N$ satisfies a sample path LDP with linear geodesics, with instantaneous rate function Λ^*, and with mean rate EY_1. However, the results in this section apply to much more general sequences Y_t.
>
> As we remarked in Chapter 1, if Y_t is stationary and ergodic, and $\sigma < 0$, then Q_0 as defined above has the steady state queue size distribution.

In Theorem 5.3 we showed that f is continuous on $\mathcal{C}_\sigma$, for $\sigma < 0$. Thus we can apply the contraction principle to obtain a large deviations principle for Q_0/N in $\mathbb{R}_0^+$ with good rate function

$$J(q) = \inf\Big\{ I(x) : x \in \mathcal{C}_\sigma,\ f(x) = q \Big\} \tag{6.19}$$

for the standard rate function

$$I(x) = \begin{cases} \int_{-\infty}^0 \Lambda^*(\dot{x}_s)\, ds & \text{if } x \in \mathcal{A}_\sigma \\ +\infty & \text{otherwise.} \end{cases}$$

We now proceed to simplify this rate function.

Theorem 6.6 *If $\tilde{X}^N$ satisfies a sample path LDP with linear geodesics, with instantaneous rate function Λ^* and mean rate $\sigma < 0$, then Q_0/N satisfies an LDP in $\mathbb{R}^+$ with good rate function*

$$J(q) = \delta q \quad \textit{where} \quad \delta = \inf_{\xi > 0} \Lambda^*(\xi)/\xi. \tag{6.20}$$

Proof. Let $J(q)$ be defined by (6.19). Note first that we might as well take the infimum in $J(q)$ over $\{x \in \mathcal{A}_\sigma : f(x) = q\}$, since $I(x) = \infty$ for $x \in \mathcal{C}_\sigma \setminus \mathcal{A}_\sigma$.

Suppose $q = 0$. The path $x \in \mathcal{A}_\sigma$ of constant rate $\dot{x}_s = \sigma$ has rate function $I(x) = 0$, and satisfies $f(x) = q$, thus $J(0) = 0$.

Suppose now $q > 0$. For any path $x \in \mathcal{C}_\sigma$ with $f(x) = q$, there must exist $0 < t < \infty$ such that $x(-t, 0] = q$, by Theorem 5.3, and so

$$J(q) \geq \inf_{t>0} \inf_{x:x(-t,0]=q} I(x) \geq \inf_{t>0} t\Lambda^*(q/t)$$

where the last inequality follows by Jensen's inequality, as in Section 6.3. Conversely, fix $t > 0$ and consider the path x defined by

$$\dot{x}_s = \begin{cases} \sigma & \text{for } s \leq -t \\ q/t & \text{for } -t < s \leq 0 \end{cases}.$$

This path satisfies $f(x) = q$ and has rate function $I(x) = t\Lambda^*(q/t)$. Thus

$$J(q) \leq t\Lambda^*(q/t) \text{ for all } t > 0.$$

We conclude that

$$J(q) = \inf_{t>0} t\Lambda^*(q/t) \tag{6.21}$$

which may be rewritten to give the result. □

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

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

(The large deviations upper and lower bounds agree, since $J(q)$ is continuous, as long as $q > 0$. If $q = 0$ they may not agree, in the case $\delta = \infty$.) The LDP may be rewritten

$$\lim_{q\to\infty} \frac{1}{q} \log P(Q_0 > q) = -\delta, \tag{6.22}$$

which means that the queue length distribution has an exponential tail.

This result should be compared with Theorem 1.4 and Theorem 3.1. In the latter theorem, we assumed that Λ^* was the convex conjugate of some function Λ and found, under certain conditions on Λ, the following alternative characterization of δ:

$$\delta = \sup\{\theta : \Lambda(\theta) \leq 0\}. \tag{6.23}$$

Recalling Section 6.3 we can ask: how do large queues build up? From the above we see that, given a large queue length q, the most likely path satisfies $\dot{x}_{-s} = q/t^*$ where t^* is the optimizing parameter in (6.21), or equivalently $\dot{x}_{-s} = \xi^*$ where ξ^* is the optimizing parameter in (6.20). (Note that $t^* = q/\xi^*$.) So the most likely way for the queue to reach q is for it to start more or less empty and then to grow at constant rate ξ^* for a time t^*.

> *Note.* To justify calling it *the* most likely path, we should be more careful. There is not, in general, a unique most likely path. However...
>
> *Exercise 6.2*
> Suppose that Λ^* is strictly convex. Show that the most likely path is unique.◇

One can also investigate the empirical distribution of the $\dot{X}_{-s}$ over the interval in which the queue builds up from empty to scaled level q. This has been done by Anantharam [2].

Exercise 6.3
Suppose that $(\dot{X}_{-t},\ t \in \mathbb{N})$ is a sequence of i.i.d. Gaussian random variables with mean $-c$ and variance σ^2, for some $c > 0$. Show that $\Lambda^*(x) = (x + c)^2/2\sigma^2$. Compute ξ^* as a function of c and σ. ◇

Exercise 6.4
Suppose that $X(-t, 0] = A(-t, 0] - C(-t, 0]$, where $A(-t, 0]$ is the amount of work arriving in $(-t, 0]$ and $C(-t, 0]$ is the service capacity. Suppose that $(\dot{A}_{-t},\ t \in \mathbb{Z})$ and $(\dot{C}_{-t},\ t \in \mathbb{Z})$ are i.i.d. sequences and independent of each other. Compute ξ^* in the following cases.

i. Arrivals and service capacity in each time slot are Bernoulli random variables, i.e. $P(\dot{A}_1 = 1) = p$, $P(\dot{A}_1) = 0) = 1 - p$, $P(\dot{C}_1 = 1) = q$, and $P(\dot{C}_1) = 0) = 1 - q$, and the queue is stable, i.e. $p < q$.
ii. Arrivals and service capacity in each time slot are geometric random variables, i.e. $P(\dot{A}_1 = j) = (1 - \mu)\mu^j$, $P(C_1 = j) = (1 - \nu)\nu^j$ and $0 < \mu < \nu < 1$. ◇

Exercise 6.5
Suppose X is as in the previous exercise, that arrivals in each time slot are exponential random variables with mean μ, i.e. $P(\dot{A}_1 > x) = e^{-x/\mu}$, and that service capacity in each time slot is exponential with mean $\nu > \mu$. Compute ξ^*, and write down the LDP for queue size.

As noted in Chapter 1, the solution Q_0 of Lindley's recursion can be interpreted either as the amount of work in the buffer in a discrete time queueing model or the waiting time in a continuous time model. With the latter interpretation, this model corresponds to an $M/M/1$ queue with arrival rate μ^{-1} and service rate ν^{-1}; it is known that the waiting time distribution for this queue is a mixture of an atom at zero and an exponential distribution with parameter $\nu^{-1} - \mu^{-1}$. How does this relate to the LDP you have found? ◇

Example 6.6 (Heavy traffic models)
In heavy traffic theory, an object of interest is the queue size in a queue fed by a Brownian motion. Let $B(t)$ be a standard Brownian motion, and let $A(-t, 0] = B(t)$ and $X(-t, 0] = A(-t, 0] - t$.

> *Note.* This model arises as a diffusion approximation to a single-server queue with general arrivals and service, provided the arrival and service processes are both sufficiently mixing, and have finite variance, and the mean arrival rate is close to the mean service rate. Heavy traffic approximation theorems are normally stated in terms of continuous-time queues, such as the $M/M/1$ queue, but also apply to discrete-time queues. Whitt [98] gives a survey of heavy traffic models.

As usual, define the scaled version

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

(Since B is indexed by $t \in \mathbb{R}$, we do not need to polygonalize A.) Schilder's theorem, recounted in Example 6.1, gives an LDP for the scaled process $B/\sqrt{N}$ on the interval $[-1, 0]$, and because of the scaling properties of Brownian motion this gives us an LDP for $A^N|_{(-1,0]}$. In fact, this LDP can be extended to give an LDP for A^N in $\mathcal{A}_0$ with good rate function

$$I(a) = \begin{cases} \frac{1}{2}\int_{-\infty}^{0} \dot{a}_t^2\, dt & \text{if } a \in \mathcal{A}_0 \\ \infty & \text{otherwise.} \end{cases} \tag{6.24}$$

So A^N satisfies a sample path LDP with linear geodesics, with instantaneous rate function $\Lambda_A^*(\alpha) = \alpha^2/2$ and mean rate 0; and so the sequence of similarly scaled processes X^N does too but with instantaneous rate function $\Lambda^*(\xi) = (\xi + 1)^2/2$ and mean rate -1.

Thus we obtain an LDP for Q_0/N with good rate function $J(q) = 2q$. In fact, in this model, the process $(Q_t,\ t \in \mathbb{R})$ is a reflected Brownian motion with negative drift, and the steady state distribution Q_0 is exactly exponential. In more complicated networks, the steady distribution is not so simple, and the large deviations techniques described in later sections can be useful. ◇

6.5 Queues with Finite Buffers

Consider the single-server queue with a finite buffer. Let $A(-t,0]$ be the amount of work arriving in the interval $(-t,0]$, and let $C(-t,0]$ be the service offered, for $t \in \mathbb{N}$. Let $X(-t,0] = A(-t,0] - C(-t,0]$. The queue length, for a queue with finite buffer B, evolves as follows:

$$Q_t^B = \left[Q_{t-1}^B + \dot{X}_t\right]_0^B$$

where $[q]_0^B = (q \vee 0) \wedge B$.

Let $\tilde{X}$ be the polygonalized version of X, and let $\tilde{X}^N$ be the scaled version of $\tilde{X}$:

$$\tilde{X}^N(-t,0] = \frac{1}{N}\tilde{X}(-Nt,0].$$

Suppose that $(\tilde{X}^N,\ N \in \mathbb{N})$ satisfies the sample path LDP with linear geodesics, with some mean rate $\sigma < 0$. In particular, $\tilde{X}^N \in \mathcal{C}_\sigma$ almost surely for each N. Let I be the rate function for the $\tilde{X}^N$.

We saw in Section 5.7 that if $\tilde{X} \in \mathcal{C}_\sigma$ then $Q_0^B = f(\tilde{X})$ where

$$f(x) = \sup_{t\geq 0}\Big(\sup_{0\leq s\leq t} x(-s,0]\Big) \wedge \Big(B + \inf_{0\leq s\leq t} x(-s,0]\Big) \tag{6.25}$$

Thus

$$Q_0^{NB}/N = f(\tilde{X}^N).$$

We also saw that f is continuous on $\mathcal{C}_\sigma$. Thus, using the contraction principle, we can find an LDP for Q_0^{NB}/N.

It remains to identify the rate function. We could in principle do this directly from the rate function that the contraction principle gives us. This is left as an exercise for the bored reader. We shall instead use a more indirect approach. Let

$$J(q) = \inf_{\substack{x\in\mathcal{C}_\sigma:\\ f(x)=q}} I(x). \tag{6.26}$$

Lemma 6.7 *Suppose $\sigma < 0$. For $0 \leq q \leq B$, $J(q)$ is equal to the rate function for the infinite-buffer queue, as given by* (6.21)*. If $q > B$ then $J(q) = \infty$.*

Proof. The cases $q = 0$ and $q > B$ are trivial. So assume $0 < q \leq B$.

Let f be the queue size function for the finite-buffer queue (6.25), and let g be the queue size function for the infinite-buffer queue (6.18). Let J be the rate function for $f(\tilde{X}^N)$ and K the rate function for $g(\tilde{X}^N)$.

We saw in Section 5.7 that the queue length in the infinite buffer queue dominates that in the finite buffer queue, i.e. $f(x) \leq g(x)$ for all $x \in \mathcal{C}_\sigma$. Thus $\{f(x) = q\} \subset \{g(x) \geq q\}$. It follows immediately that

$$\begin{aligned} J(q) &= \inf_{x:f(x)=q} I(x) \\ &\geq \inf_{x:g(x)\geq q} I(x) = \inf_{r \geq q} K(r). \end{aligned}$$

Now we have found that $K(r) = \delta r$ where $\delta \geq 0$, so the infimum is attained at $r = q$, and we get $J(q) \geq K(q)$.

To obtain the reverse inequality, recall that the sample path x that achieves the infimum in $K(q)$ gives rise to a queue size that increases linearly from 0 to q over some time period $[-t, 0]$. Thus, as $q \leq B$, the queue size in the infinite-buffer queue never exceeds B. Hence the finite-buffer queue evolves in the same way as the infinite-buffer queue, and $f(x) = g(x) = q$. Or, more rigorously, simply observe from (6.25) that $f(x) = q$. Thus

$$J(q) \leq I(x) = K(q). \qquad \square$$

The same LDP was obtained by Toomey [94]. It justifies approximating the frequency of overflow in a queue with (large) finite buffer B by the frequency with which queue level B is exceeded in the corresponding queue with infinite buffer.

6.6 Queueing Delay

Consider the single-server queue with an infinite buffer. Let $A(u, v]$ be the amount of work arriving in the interval $(u, v]$, and let $C(u, v]$ be the amount of service offered, for $u, v \in \mathbb{Z}$. Let $\tilde{A}$ be the polygonalized version of A, and define the scaled version

$$\tilde{A}^N(u, v] = \frac{1}{N} \tilde{A}(Nu, Nv] \quad \text{for } u, v \in \mathbb{R}$$

and similarly for $\tilde{C}^N$. Let Q_0 be the queue size at time 0. In Section 5.8 we defined the queueing delay

$$W = \inf\bigl\{t \in \mathbb{N}_0 : C(0, t] \geq Q_0\bigr\}$$

and saw that

$$W/N \approx f(\tilde{A}^N, \tilde{C}^N) \quad \text{where} \quad f(a, c) = \inf\bigl\{t \in \mathbb{R}^+ : c(0, t] \geq q_0(a, c)\bigr\} \tag{6.27}$$

To make these definitions we need to describe the behaviour of the service process after time 0, and we explained in Section 5.8 how to do this by extending the space $\mathcal{C}_\mu$ to $\mathcal{C}_\mu^2$.

Assume that the sequence $\big((\tilde{A}^N, \tilde{C}^N),\ N \in \mathbb{N}\big)$ satisfies a sample path LDP in $\mathcal{C}_\mu^2 \times \mathcal{C}_\nu^2$ with instantaneous rate function Λ^*. It should be clear what is meant by this: we mean that the rate function is

$$I(a,c) = \begin{cases} \int_{-\infty}^{\infty} \Lambda^*(\dot{a}_t, \dot{c}_t)\, dt & \text{if } (a,c) \in \mathcal{A}_\mu^2 \times \mathcal{A}_\nu^2 \\ \infty & \text{otherwise.} \end{cases}$$

If the service rate is bounded below by $c_0 > 0$ almost surely, we can use Lemma 4.9 to restrict the LDP to $\mathcal{C}_\mu^2 \times \mathcal{X}_\nu$, where $\mathcal{X}_\nu$ is the restriction of $\mathcal{C}_\nu^2$ to the set of service processes whose service rate is bounded below by c_0. Suppose also that $\mu < \nu$. Then, as described in Section 5.8, the function f is continuous, and the sense of the approximation in (6.27) is such as to allow us to apply the approximate contraction principle, as it is described in Exercise 4.6, to obtain an LDP for W/N with good rate function

$$J(w) = \inf\big\{I(a,c) : f(a,c) = w\big\}$$

We now proceed to simplify this rate function.

Theorem 6.8 *If $(\tilde{A}^N, \tilde{C}^N)$ satisfies a sample path LDP in $\mathcal{C}_\mu^2 \times \mathcal{C}_\nu^2$ with instantaneous rate function Λ^*, and the service rate is bounded below by $c_0 > 0$ almost surely, and $\mu < \nu$, then W/N satisfies an LDP with good rate function*

$$J(w) = \gamma w \quad \textit{where} \quad \gamma = \inf_{\sigma \geq c_0} \delta\sigma + \Lambda^*(\mu, \sigma) \tag{6.28}$$

and δ is as in Theorem 6.6.

Proof. First introduce a dummy variable q into the variational problem (6.28), representing the queue size at time 0. The problem becomes

$$\begin{aligned} &\text{minimize} \quad I(a,c) \\ &\text{over} \quad a \in \mathcal{A}_\mu^2,\ c \in \mathcal{A}_\nu^2,\ q \geq 0 \\ &\text{such that} \quad q_0(a,c) = q,\ f(a,c) = w. \end{aligned}$$

The reason we introduce q is that, given q, the constraints split into two parts, one part over $(-\infty, 0]$ and the other over $(0, \infty)$. The objective function also splits into two parts. Therefore the optimization problem splits into two subproblems, each parameterized by q, and the solution to the overall problem is the infimum over $q \geq 0$ of the sum of the solutions to these two subproblems.

The first subproblem is

$$\begin{aligned}&\text{minimize} \quad \int_{-\infty}^{0} \Lambda^*(\dot{a}_t, \dot{c}_t)\,dt \\ &\text{over} \quad a \in \mathcal{A}_\mu,\ c \in \mathcal{A}_\nu \\ &\text{such that} \quad q_0(a,c) = q.\end{aligned}$$

We have already solved this, in Section 6.4. The solution is δq.

The second subproblem is

$$\begin{aligned}&\text{minimize} \quad \int_{-\infty}^{0} \Lambda^*(\dot{a}_t, \dot{c}_t)\,dt \\ &\text{over} \quad a \in \mathcal{A}_\mu,\ c \in \mathcal{A}_\nu \\ &\text{such that} \quad f(a,c) = w \quad \text{given} \quad q_0(a,c) = q.\end{aligned}$$

Since the service rate is bounded below by c_0 almost surely, the rate function is infinite for $c \notin \mathcal{X}_\nu$. Furthermore the optimization problem is over absolutely continuous paths; thus we can restrict attention to paths c for which $\dot{c}_t \geq c_0$ for all t. This means that the constraint $f(a,c) = w$ simplifies to $c(0,w] = q$. We described in Section 6.3 how to solve such optimization problems. By the linear geodesics property, one can show that the solution to this problem is $w\Lambda^*(\mu, q/w)$.

The solution to the overall optimization problem is thus

$$\inf_{q \geq 0} \delta q + w\Lambda^*(\mu, q/w)$$

which by reparameterizing in terms of $\sigma = q/w$ becomes

$$w \inf_{\sigma \geq 0} \delta\sigma + \Lambda^*(\mu, \sigma)$$

As we have noted, the rate function $I(a,c)$ is infinite if $c \notin \mathcal{X}_\nu$, which means that $\Lambda^*(\mu,\sigma) = \infty$ for $\sigma < c_0$, so we can restrict the infimum to $\sigma \geq c_0$. This gives the desired rate function. □

6.7 Departure Process

In this section we turn to the departure process. The LDPs we have obtained so far have been for queue length, which is a real-valued function of the input process; the departure process, on the other hand, takes values in the space of continuous functions on the real line. Even though this is a much more complicated example in some sense, the techniques are essentially the same.

Consider the single-server queue with an infinite buffer, described in Section 6.4. Let $A(-t,0]$ be the amount of work arriving in $(-t,0]$, and let $C(-t,0]$ be the amount of work that can be served, for $t \in \mathbb{N}$. Let $Z = (A,C)$. As usual, define the scaled polygonalized version $\tilde{Z}^N$, and assume that $(\tilde{Z}^N,\ N \in \mathbb{N})$ satisfies the sample path LDP with linear geodesics. Let the mean rate be (μ,ν). The departure process was discussed in Section 5.11. It is defined by

$$D(-t,0] = A(-t,0] + Q_{-t} - Q_0.$$

Let $X(-t,0] = A(-t,0] - C(-t,0]$, and define the polygonalized versions $\tilde{A}$ etc. Note that $\tilde{A} \in \mathcal{C}_\mu$ and $\tilde{C} \in \mathcal{C}_\nu$ almost surely. Define for $t \in \mathbb{R}_0^+$ the queue size function $q_{-t} : \mathcal{C}_\mu \times \mathcal{C}_\nu \to \mathbb{R}$ by

$$q_{-t}(a,c) = \sup_{-s \le -t} a(-s,-t] - c(-s,-t]. \tag{6.29}$$

Also define the process-valued function f by

$$f(a,c)(-t,0] = a(-t,0] + q_{-t}(a,c) - q_0(a,c) \quad \text{for } t \in \mathbb{R}_0^+. \tag{6.30}$$

It is easy to see that $D(-t,0] = f(\tilde{A},\tilde{C})(-t,0]$ for $t \in \mathbb{N}$. It is somewhat harder to see, but still the case, that $\tilde{D} = f(\tilde{A},\tilde{C})$. It is challenging to see that $f : \mathcal{C}_\mu \times \mathcal{C}_\nu \to \mathcal{C}_\mu$ and that it is continuous; this is proved in Section 5.11.

Finally, define the scaled polygonalized versions $\tilde{A}^N$ etc. It is easy to see that

$$\tilde{D}^N = f(\tilde{A}^N, \tilde{C}^N).$$

Using the contraction principle, we can immediately deduce an LDP for $\tilde{D}^N$, with rate function

$$J(d) = \inf_{\substack{a \in \mathcal{C}_\mu,\, c \in \mathcal{C}_\nu: \\ f(a,c) = d}} I(a,c) \tag{6.31}$$

where the input rate function is of the usual form

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

To solve this variational problem in general is quite hard.

A natural question to ask at this point is: does the departure process satisfy a sample path LDP with linear geodesics? If so, then we could in principle recursively use the approach outlined above to analyse quite complicated networks of queues. Unfortunately the answer to the question is: in general, no. Counterexamples are exhibited in [41]. It is shown there that in the case where $\Lambda^*(\dot{a}_t, \dot{c}_t) = \Lambda^*_A(\dot{a}_t) + \Lambda^*_C(\dot{c}_t)$, a necessary condition for the departure process to satisfy a sample path LDP with linear geodesics is that $\Lambda^*_A \le \Lambda^*_C$ on $[0, \mu]$. It seems to be difficult to obtain sufficient conditions.

There is one simple case where it is possible give an affirmative answer:

Exercise 6.7
Suppose that the service rate is c, a constant, and greater than the mean arrival rate μ. Show that the departure process $\tilde{D}^N$ satisfies a sample path large deviations principle with linear geodesics, with mean rate μ and instantaneous rate function

$$\mathrm{M}^*(x) = \begin{cases} \Lambda^*(x) & \text{if } x \le c \\ \infty & \text{otherwise.} \end{cases}$$

Hint. Let J be the actual rate function for the departure process, which we know to be good, and let I be the sample path rate function corresponding to the instantaneous rate function M^*. We want to show that $I(d) = J(d)$. Show that $I(d) \le J(d)$, by assuming $J(d)$ is finite, taking an optimal arrival process a, and if $a(-t, 0] \ne d(-t, 0]$ for some t then straightening a to make it equal. Show that $I(d) \ge J(d)$, by assuming $I(d)$ is finite, and showing that the queue fed by arrival process d produces departure process d. ◇

Large deviations for departure processes have also been studied by Bertsimas et al. [7].

6.8 Mean Rate of Departures

Since we cannot in general find the rate function for the departure process, let us ask a simpler question. Can we find an LDP for the mean departure rate over an interval? In the setup of the preceding section, the mean rate of departures in the interval $(-N, 0]$ is $\tilde{D}^N(-1, 0]$. By the contraction principle we have an LDP for the sequence $(\tilde{D}^N(-1, 0],\ N \in \mathbb{N})$ with good rate function

$$J(\xi) = \inf\Big\{ I(a, c) : (a, c) \in \mathcal{C}_\mu \times \mathcal{C}_\nu,\ f(a, c)(-1, 0] = \xi \Big\} \tag{6.32}$$

where f is given by (6.30).

Lemma 6.9

$$J(\xi) = \inf_{q\geq 0} \delta q + K(\xi,q) \wedge L(\xi,q)$$

where

$$\begin{aligned}
\delta &= \inf_{\alpha>\sigma} \frac{\Lambda^*(\alpha,\sigma)}{\alpha-\sigma} \\
K(\xi,q) &= \inf_{C_1} \Lambda^*(\alpha,\xi) \\
&\quad \textit{and } C_1 = \{\alpha : q+\alpha \geq \xi\} \\
L(\xi,q) &= \inf_{C_2} (1-t)\Lambda^*(\alpha_1,\sigma_1) + t\Lambda^*(\alpha_2,\sigma_2) \\
&\quad \textit{and } C_2 = \Big\{ (\alpha_1,\alpha_2,\sigma_1,\sigma_2,t) : t\in[0,1],\ \alpha_2\geq\sigma_2, \\
&\qquad\qquad q+(1-t)\alpha_1 \leq (1-t)\sigma_1, \\
&\qquad\qquad\quad q+(1-t)\alpha_1 + t\sigma_2 = \xi \Big\}
\end{aligned}$$

Before going on to follow the proof, think what the equations suggest. They suggest that the queue size at time -1 is equal to q, for some q, and that thereafter either the queue never empties in $(-1,0]$, or the queue becomes empty at $-t$ and is then busy in $(-t,0]$.

Proof. Write $x(t,u]$ for $a(t,u]-c(t,u]$. Let d be the departure process $f(a,c)$. Recall that

$$d(-1,0] = a(-1,0] + q_{-1} - q_0 \tag{6.33}$$

where the two quantities appearing in this equation are

$$q_{-1} = \sup_{-s\leq -1} x(-s,-1]$$

and

$$q_0 = \sup_{-s\leq 0} x(-s,0] \quad .$$

We will adopt the strategy of introducing a dummy variable q for q_{-1}, and (in a large deviations sense) conditioning on its value. Indeed, let us rewrite the optimization problem (6.32) as

$$\begin{aligned}
&\text{minimize} \quad \int_{-\infty}^{-1} \Lambda^*(\dot a_t,\dot c_t)\,dt + \int_{-1}^{0} \Lambda^*(\dot a_t,\dot c_t)\,dt \\
&\text{over} \quad (a,c)\in\mathcal{A}_\mu\times\mathcal{A}_\nu \text{ and } q\geq 0 \\
&\text{subject to} \quad q_{-1}=q \text{ and } d(-1,0]=\xi.
\end{aligned}$$

The reason we introduce q is that, given that $q_{-1} = q$, we can write $d(-1,0]$ purely as a function of q and $a|_{(-1,0]}$ and $c|_{(-1,0]}$ (cf. Exercise 5.3). Thus the constraints are split into two parts, one part over $(-\infty,-1]$ and the other over $(-1,0]$. The objective function is also split into two parts. Therefore the optimization problem splits into two subproblems, each parameterized by q. The first (after translating the time coordinates) is

$$\begin{aligned}&\text{minimize} \quad \int_{-\infty}^{0} \Lambda^*(\dot{a}_t, \dot{c}_t)\, dt\\ &\text{over} \quad (a,c) \in \mathcal{A}_\mu \times \mathcal{A}_\nu,\\ &\text{subject to} \quad q_0 = q.\end{aligned}$$

The second problem is

$$\begin{aligned}&\text{minimize} \quad I^1(a,c) \quad \text{where } I^1(a,c) = \int_{-1}^{0} \Lambda^*(\dot{a}_t, \dot{c}_t)\, dt\\ &\text{over} \quad (a,c) \in \mathcal{A}^1 \times \mathcal{A}^1,\\ &\text{subject to} \quad d(-1,0] = \xi \quad \text{given} \quad q_{-1} = q.\end{aligned}$$

The solution to the overall optimization problem (6.32) is the infimum over $q \geq 0$ of the sum of the solutions to these two subproblems.

The first subproblem we have already solved, in Section 6.4. The value of the infimum is δq where (in our present notation)

$$\delta = \inf_{\alpha > \sigma} \frac{\Lambda^*(\alpha, \sigma)}{\alpha - \sigma}$$

It remains to solve the second subproblem. Let $M(\xi, q)$ be the solution. We will show that

$$M(\xi, q) = K(\xi, q) \wedge L(\xi, q). \tag{6.34}$$

Then, we will have completed the proof. First we will rewrite the constraint in a more useful way. According to Exercise 5.3,

$$q_0 = q_0^{-1} \vee \Big(q + a(-1,0] - c(-1,0]\Big) \quad \text{where} \quad q_0^{-1} = \sup_{-1 \leq -s \leq 0} x(-s, 0].$$

Interpret q_0^{-1} as the queue size at time 0 if the system were started empty at -1. This yields

$$d(-1,0] = \Big(a(-1,0] + q - q_0^{-1}\Big) \wedge c(-1,0].$$

We will call this function $g(a,c)$.

LHS≥RHS in (6.34). First, take any path (a,c) which satisfies the constraint. Then ξ is equal to the minimum of $c(-1,0]$ and $a(-1,0]+q-q_0^{-1}$. Consider first the case where ξ is equal to the former. Let $\alpha = a(-1,0]$; and define a new path (a',c'), a straightening of (a,c), by

$$\begin{aligned} \dot a'_{-t} &= \alpha \\ \dot c'_{-t} &= \xi. \end{aligned} \tag{6.35}$$

Note that $a'(-1,0] = a(-1,0]$ and $c'(-1,0] = c(-1,0]$ and so $I^1(a,c) \geq I^1(a',c')$ by Jensen's inequality. Furthermore, $c(-1,0] \leq a(-1,0]+q-q_0^{-1}$ and so $\xi \leq \alpha + q$, which implies that α lies in C_2 and $I^1(a',c') \geq K(\xi,q)$.

Consider next the case where ξ is equal to $a(-1,0]+q-q_0^{-1}$, and suppose that the supremum in q_0^{-1} is attained at $-t \in [-1,0]$. Let $\alpha_1 = a(-1,-t]/(1-t)$, $\alpha_2 = a(-t,0]/t$, $\sigma_1 = c(-1,-t]/(1-t)$ and $\sigma_2 = c(-t,0]/t$; and define a new path (a',c'), a straightening of (a,c), by

$$\begin{aligned} \dot a'_t &= \alpha_1 \text{ for } -1 \leq -t, \text{ and } \alpha_2 \text{ otherwise} \\ \dot c'_t &= \sigma_1 \text{ for } -1 \leq -t, \text{ and } \sigma_2 \text{ otherwise.} \end{aligned} \tag{6.36}$$

Note that $a'(-1,-t] = a(-1,-t]$ and $a'(-t,0] = c(-t,0]$, and similarly for c and c', and so $I^1(a,c) \geq I^1(a',c')$ by Jensen's inequality. Furthermore, $\alpha_2 \geq \sigma_2$ since the supremum in q_0^{-1} is attained at $t > 0$ (the special case $t=0$ is left as an exercise); also $q + (1-t)\alpha_1 \leq (1-t)\sigma_1$ for the same reason (the special case $t=1$ is left as an exercise); finally the constraint says that $q + (1-t)\alpha_1 + t\sigma_2 = \xi$; thus $(\alpha_1,\alpha_2,\sigma_1,\sigma_2,t)$ lies in C_2 and so $I^1(a,c) \geq L(\xi,q)$.

We have shown that for any path (a,c) satisfying the constraints of the $M(\xi,q)$ problem, either $I^1(a,c) \geq K(\xi,q)$ or $I^1(a,c) \geq L(\xi,q)$ hence $I^1(a,c) \geq K(\xi,q) \wedge L(\xi,q)$, hence this is a lower bound for $M(\xi,q)$ also.

LHS≤RHS in (6.34). We prove the reverse inequality in two separate cases. For the first case, fix any α in C_1, and define the path

$$\begin{aligned} \dot a_{-t} &= \alpha \\ \dot c_{-t} &= \xi. \end{aligned}$$

Observe that $g(a,c) = \xi$. Thus $M(\xi,q) \leq I^1(a,c)$, and taking the minimum over such (a,c) we obtain $M(\xi,q) \leq K(\xi,q)$.

For the second case, fix $(\alpha_1,\alpha_2,\sigma_1,\sigma_2,t)$ in C_2, and define the path

$$\begin{aligned} \dot a_{-t} &= \alpha_1 \text{ for } -1 \leq -t, \text{ and } \alpha_2 \text{ otherwise} \\ \dot c_{-t} &= \sigma_1 \text{ for } -1 \leq -t, \text{ and } \sigma_2 \text{ otherwise.} \end{aligned}$$

For this path, $c(-1,0] = (1-t)\sigma_1 + t\sigma_2$, $a(-1,0] = (1-t)\alpha_1 + t\alpha_2$, and $q_0^{-1} = t(\alpha_2 = \sigma_2)$; by the constraints in C_2, $g(a,c) = \xi$. Taking the minimum over such (a,c) we obtain $M(\xi,q) \leq L(\xi,q)$.

Putting these two cases together, we have proved that $M(\xi,q) \leq K(\xi,q) \wedge L(\xi,q)$. This completes the proof. □

The above problem can be solved numerically; it also admits closed-form solutions in special cases. Note that both $K(\xi,q)$ and $L(\xi,q)$ are the solutions to finite-dimensional optimization problems.

The strategy of the proof was as follows. We bounded the rate function below, by considering certain constraints that an optimal path must satisfy. We bounded it above, by exhibiting certain paths. If we choose the constraints and the paths well, the upper and lower bounds agree. Since the inputs have linear geodesics, the paths we exhibit will be piecewise linear, and the constraints will reflect this.

Similar reductions can be obtained in dealing with more complex acyclic queueing networks; however the number of possible segments in the piecewise linear paths grows very quickly, rendering this approach impractical for large networks.

The exercises below provide some instances where it is possibly to derive closed-form expressions for the rate function governing the mean departure rate. While such instances are somewhat rare, the situation is not so different from classical queueing theory, where closed-form solutions for steady state distributions are available only in a few special cases.

Exercise 6.8

Compute J, the rate function for the mean rate of departures, in each of the following cases. Assuming that the arrivals $\{\dot{A}_t, t \in \mathbb{Z}\}$ and service capacities $\{\dot{C}_t,\ t \in \mathbb{Z}\}$ are i.i.d. sequences and independent of each other.

i. Arrivals and service capacity in each time slot are Bernoulli random variables, i.e. $P(\dot{A}_1 = 1) = p$, $P(\dot{A}_1 = 0) = 1-p$, $P(\dot{C}_1 = 1) = q$ and $P(\dot{C}_1 = 0) = 1-q$, and the queue is stable, i.e. $p < q$.

ii. Arrivals and service capacity in a time slot are geometric random variables, i.e. $P(\dot{A}_1 = j) = (1-\mu)\mu^j$ and $P(\dot{C}_1 = j) = (1-\nu)\nu^j$, and $0 < \mu < \nu < 1$.

iii. The amount of work arriving in a time slot is exponential with mean μ, i.e. $P(\dot{A}_1 > x) = e^{-x/\mu}$, and the service capacity is exponential with mean ν, and $\mu < \nu$. ◇

A special case of considerable practical importance is that of queues with deterministic service capacity. The next exercise shows that we can explicitly compute the rate function for the mean rate of departures from such a queue, for arbitrary arrival processes.

Exercise 6.9 (Mean departure rate with constant service rate)
Suppose that the arrival process satisfies a sample path LDP with linear geodesics, with instantaneous rate function Λ^*_A and mean rate μ, and that the service process is $C(-t,0] = ct$ for some $c > \mu$. Show that

$$J(\xi) = \begin{cases} \Lambda^*(\xi) & \text{for } 0 \le \xi \le c \\ +\infty & \text{otherwise.} \end{cases} \qquad \diamond$$

We now proceed to derive a joint LDP for the mean rate of departures over an interval and the work in queue at the end of this interval. This will turn out to be useful in Section 6.9 below. Specifically, suppose we are interested in the pair

$$\big(D(-N,0]/N, Q_0/N\big).$$

As we noted at the beginning of this section, the first component is $\tilde{D}^N(-1,0]$ which is a continuous function of $(\tilde{A}^N, \tilde{C}^N)$; the second component is also a continuous function $q_0(\tilde{A}^N, \tilde{C}^N)$. Thus we can use the contraction principle to obtain an LDP for the pair, with rate function

$$J(\xi, r) = \inf\Big\{ I(a,c) : (a,c) \in \mathcal{C}_\mu \times \mathcal{C}_\nu, \\ f(a,c)(-1,0] = \xi, q_0(a,c) = r \Big\} \qquad (6.37)$$

where f is the departure map, given by (6.30).

Lemma 6.10 *The rate function $J(\xi, r)$ has the same form as $J(\xi)$ in Lemma 6.9, except that the constraint sets have each been narrowed by the addition of an equation involving r:*

$$\begin{aligned} C_1 &= \{\alpha : q + \alpha \ge \xi,\ r = q + \alpha - \xi\} \\ C_2 &= \Big\{(\alpha_1, \alpha_2, \sigma_1, \sigma_2, t) : t \in [0,1],\ \alpha_2 \ge \sigma_2, \\ &\qquad q + (1-t)\alpha_1 \le (1-t)\sigma_1, \\ &\qquad q + (1-t)\alpha_1 + t\sigma_2 = \xi,\ r = t(\alpha_2 - \sigma_2)\Big\} \end{aligned}$$

Sketch proof. The proof is very much like that of Lemma 6.9: we first of all introduce a dummy variable q to represent the queue size at time -1; then we split the optimization problem into two parts, one over $(-\infty, -1]$ and the other over $(-1, 0]$, and to each problem we add the constraint $q_{-1}(a,c) = q$. We solve the first part using the result about queue size in Section 6.4, and we solve the second part by considering well-chosen straightenings of optimal paths.

In Lemma 6.9, the constraint sets reflected the properties of the optimal path. At time -1, the queue size has reached some level $q = q_{-1}$. Let q_0 be the queue size at time 0, and q_0^{-1} the queue size at time 0 if the queue were started empty at time -1. We identified two cases: either $q_0 > q_0^{-1}$, in which case the mean rate of departures is given by

$$d(-1,0] = c(-1,0]$$

or $q_0 = q_0^{-1}$, in which case it is given by

$$d(-1,0] = a(-1,0] + q_{-1} - q_0^{-1}.$$

In the first case, we study piecewise linear paths of the form (6.35), and in the second we study piecewise linear paths of the form (6.36)

Now, the rate function (6.37) that we want to calculate here specifies an additional constraint: that $q_0 = r$. So in the linear sample paths that we construct must satisfy this constraint. In the first case, $q_0 = q + \alpha - \xi$; in the second case, $q_0 = t(\alpha_2 - \sigma_2)$. This explains the additional constraints in C_1 and C_2 specified in the statement of the lemma.

(You should work through the proof of Lemma 6.10, and verify that adding these constraints is the correct thing to do in proving both the lower and upper bounds on $M(\xi, r, q)$.) □

We can in fact find a slightly more explicit form for $J(\xi, r)$, by simplifying the constraint sets C_1 and C_2 above: by setting $\zeta_1 = q + (1-t)\alpha_1$ and $\zeta_2 = (1-t)\sigma_1$, we find that

$$\begin{aligned} J(\xi,r) &= \inf_{q\geq 0} \delta q + K(\xi,r,q) \wedge L(\xi,r,q) \\ &\quad \delta \text{ is as in Lemma 6.9} \\ K(\xi,r,q) &= \Lambda^*(r-q+\xi,\xi) \qquad (6.38) \\ L(\xi,r,q) &= \inf_{C_2} (1-t)\Lambda^*\Big(\frac{\zeta_1 - q}{1-t}, \frac{\zeta_2}{1-t}\Big) + t\Lambda^*\Big(\frac{r+\xi-\zeta_1}{t}, \frac{\xi-\zeta_1}{t}\Big) \qquad (6.39) \\ &\quad \text{and } C_2 = \big\{(\zeta_1,\zeta_2,t) : t\in[0,1], \zeta_1 \leq \zeta_2\big\}. \qquad (6.40) \end{aligned}$$

Exercise 6.10
Use Lemma 6.10 and the contraction principle to derive the rate function $J(\xi)$ governing the LDP for $\big(D(-N,0]/N,\ N \in \mathbb{N}\big)$. Verify that this is the same as the rate function given by Lemma 6.9. ◇

Exercise 6.11
Compute the rate function $J(\xi, q)$ governing the joint LDP of the pair

$$\big((D(-N,0]/N, Q_0/N),\ N \in \mathbb{N}\big)$$

for the arrival and service processes in Exercise 6.8. ◇

This problem has also been studied by Ramanan and Dupuis [87] and Chang and Zajic [15], and by Puhalskii and Whitt [84] in the continuous-time setting.

6.9 Quasi-Reversibility

A continuous-time queueing system is called *quasi-reversible* if its 'state' is a stationary Markov process, with the property that the state at time t is independent of both the arrival process after time t, and the departure process prior to time t. It can be shown that the arrival and departure processes are both Poisson (with the same rate). It can also be shown that, in a network of quasi-reversible queues, the joint distribution of queue states is product-form. This makes such networks analytically tractable, and has contributed to the popularity of quasi-reversible queueing models in performance analysis. For more details, see Muntz [74], Kelly [52] and Walrand [95].

In this section we will explore the large deviations analogue of quasi-reversibility. Consider a single-server queue whose service process satisfies a sample path LDP with linear geodesics. We will exhibit a rate function such that, if the arrival process satisfies a sample path LDP with linear geodesics and our specified rate function, then the joint rate function for current queue size, future arrivals, and past departures, is simply the sum of their individual rate functions. (This is the large-deviations analogue of independence: see Theorem 4.14). Furthermore, this rate function is invariant, i.e. the departure process has exactly the same rate function as the arrival process.

In fact, one can say more: if c is the mean service rate, then for any mean arrival rate $\mu < c$ there is a unique rate function for the sample paths of the arrival process which both is invariant and satisfies the large-deviations analogue of quasi-reversibility. We will not prove this here; the details can be found in [44].

The setup for the rest of this section is as follows. Consider a single-server queue with an infinite buffer. Let $C(-t, 0]$ be the amount of service offered in the interval $(-t, 0]$. As usual, let $\tilde{C}$ be the polygonalized version, and define the scaled polygonalized version

$$\tilde{C}^N(-t, 0] = \frac{1}{N}\tilde{C}(-Nt, 0].$$

Assume that the sequence $(\tilde{C}^N,\ N \in \mathbb{N})$ satisfies a sample path LDP with linear geodesics, with instantaneous rate function Λ_C^* and mean rate c. Let $A(-t, 0]$ be the amount of work arriving in the interval $(-t, 0]$. Let D be the departure process, defined by (6.29). Define the scaled polygonalized versions $\tilde{A}^N$ and $\tilde{D}^N$ accordingly.

Theorem 6.11 *Assume that Λ_C^* is strictly convex, and finite and differentiable on $(0, c]$.*

For $\mu \in (0, c)$, let $\theta_\mu = (\Lambda_C^*)'(\mu)$, and let

$$\Lambda_A^*(x) = \Lambda_C^*(x) - \Lambda_C^*(\mu) - \theta_\mu(x - \mu). \tag{6.41}$$

Then Λ_A^* is a convex rate function, and $\Lambda_A^*(\mu) = 0$. Suppose the sequence $(\tilde{A}^N,\ N \in \mathbb{N})$ satisfies a sample path LDP with linear geodesics, with instantaneous rate function Λ_A^*, and that $\tilde{A}^N$ is independent of $\tilde{C}^N$. Then

i. The sequence $(\tilde{D}^N,\ N \in \mathbb{N})$ satisfies a sample path LDP with linear geodesics, with instantaneous rate function Λ_A^* and mean rate μ;

ii. The sequence $(Q_0/N,\ N \in \mathbb{N})$ satisfies an LDP in $\mathbb{R}_0^+$ with good rate function $I(q) = \delta q$, with

$$\delta = \inf_{\alpha > \sigma} \frac{\Lambda_A^*(\alpha) + \Lambda_C^*(\sigma)}{\alpha - \sigma}$$

iii. The sequence $\big((\tilde{D}^N, Q_0/N),\ N \in \mathbb{N}\big)$ satisfies a large deviations principle, and the rate function is the sum of the two individual rate functions.

After making some remarks about interpretation, the rest of this section is given over to proving Theorem 6.11.

i. The theorem does not claim that there exists an arrival process $\tilde{A}^N$ with the specified rate function. For some results on existence see [44]. Nor does the theorem claim that the departure process has the same distribution as the arrival process—only that the distribution is the same on a large deviations scale.

ii. Since Λ_C^* is strictly convex, so is Λ_A^*. It is easy to check that $\Lambda_A^*(\mu) = 0$, and this must thus be the unique zero, and hence the mean rate of A. The theorem says that for each $\mu \in (0, c)$ there is (*pace* the above remark) an arrival process with mean rate μ, whose large deviations rate function is the same as that of the corresponding departure process. Clearly, there is no such rate function if $\mu > c$ since the mean rate of departures cannot exceed the mean service rate.

iii. When the service process has i.i.d. increments, there is a concrete interpretation of (6.41). Let L denote the probability law of $\dot{C}_0$, the service offered in a single timeslot, and define the exponentially tilted law $\tilde{L}$ by

$$\frac{d\tilde{L}}{dL}(\xi) = \exp\bigl(\theta_\mu \xi - \Lambda_C(\theta_\mu)\bigr), \quad \xi \in \mathbb{R}.$$

(This tilting was seen in the note at the end of Section 2.5 and in the proof of Cramér's theorem.) Note that by the self-duality property of Λ_C (see Lemmas 2.5 and 2.6)

$$\Lambda_C(\theta_\mu) = \theta_\mu \mu - \Lambda_C^*(\mu)$$

and in particular that $\Lambda_C(\theta_\mu)$ is finite, so the exponential tilt is well-defined. Now, let X be a random variable drawn from $\tilde{L}$, and consider its log moment generating function, and its convex dual:

$$\begin{aligned}
\Lambda_X(\theta) &= \log E e^{\theta X} e^{\theta_\mu X - \Lambda_C(\theta_\mu)} = \Lambda_C(\theta + \theta_\mu) - \Lambda_C(\theta_\mu) \qquad (6.42)\\
\Lambda_X^*(\xi) &= \sup_\theta \theta\xi - \bigl(\Lambda_C(\theta+\theta_\mu) - \Lambda_C(\theta_\mu)\bigr)\\
&= -\theta_\mu \xi + \Lambda_C(\theta_\mu) + \sup_\theta (\theta+\theta_\mu)\xi - \Lambda_C(\theta+\theta_\mu)\\
&= -\theta_\mu \xi + \Lambda_C(\theta_\mu) + \Lambda_C^*(\xi)\\
&= \Lambda_C^*(\xi) - \Lambda_C^*(\mu) - \theta_\mu(\xi - \mu) \quad \text{using (6.42)}
\end{aligned}$$

This is exactly as specified by (6.41). Therefore if we let $(\dots, \dot{A}_{-1}, \dot{A}_0)$ be a sequence of i.i.d. random variables drawn from $\tilde{L}$, then the sequence of scaled arrival processes $\tilde{A}^N$ satisfies the conditions of Theorem 6.11.

The following lemmas are key steps in the proof of Theorem 6.11.

Lemma 6.12 *Under the assumptions of Theorem 6.11, $\delta = -\theta_\mu$.*

Proof. On one hand, by choosing the 'flipped' values $\alpha = c$ and $\sigma = \mu$,

$$\delta = \inf_{\alpha > \sigma} \frac{\Lambda_A^*(\alpha) + \Lambda_C^*(\sigma)}{\alpha - \sigma} \le \frac{\Lambda_A^*(c) - \Lambda_C^*(\mu)}{c - \mu} = -\theta_\mu.$$

(The intuition behind this choice is as follows. The parameter δ represents an exponential tilt of $\dot{A}_0$ and $\dot{C}_0$. We claim that the appropriate tilt of $\dot{A}_0$

is $-\theta_\mu$. Bearing in mind the note at the end of Section 2.5, the tilted mean of $\dot{A}_0$ must satisfy $(\Lambda_A^*)'(\alpha) = -\theta_\mu$. The tilted mean is thus c. Similarly for $\dot{C}_0$.)

On the other hand, by using the convexity bound

$$\Lambda_C^*(\sigma) \geq \Lambda_C^*(\mu) + \theta_\mu(\sigma - \mu)$$

and by substituting the equation (6.41) for Λ_A^* into δ,

$$\delta = \inf_{\alpha > \sigma} \frac{\Lambda_A^*(\alpha) + \Lambda_C^*(\sigma)}{\alpha - \sigma} \geq \inf_{\alpha > \sigma} \frac{\Lambda_C^*(\alpha) - \theta_\mu(\alpha - \sigma)}{\alpha - \sigma} \geq -\theta_\mu. \qquad \square$$

Lemma 6.13 *Under the assumptions of Theorem 6.11, and using the notation of Lemma 6.10,*

$$\delta q + K(\xi, r, q) \geq \delta r + \Lambda_A^*(\xi) \tag{6.43}$$

and

$$\delta q + L(\xi, r, q) \geq \delta r + \Lambda_A^*(\xi). \tag{6.44}$$

Proof. First (6.43):

$$\begin{aligned} \delta q + K(\xi, r, q) &= \delta q + \Lambda_A^*(r - q + \xi) + \Lambda_C^*(\xi) \quad \text{using (6.38)} \\ &= \delta q + \Lambda_C^*(r - q + \xi) + \Lambda_A^*(\xi) - \delta q + \delta r \qquad (6.45) \\ &\qquad \text{using the definition of } \Lambda_A^* \text{ twice} \\ &\geq \delta r + \Lambda_A^*(\xi) \quad \text{since } \Lambda_C^* \text{ is non-negative.} \end{aligned}$$

Now for (6.44). This involves the infimum over $\zeta_1 \leq \zeta_2$ of

$$\begin{aligned} \delta q + \ (1-t)\Big[\Lambda_A^*\Big(\frac{\zeta_1 - q}{1-t}\Big) + \Lambda_C^*\Big(\frac{\zeta_2}{1-t}\Big)\Big] \\ +t\Big[\Lambda_A^*\Big(\frac{r + \xi - \zeta_1}{t}\Big) + \Lambda_C^*\Big(\frac{\xi - \zeta_1}{t}\Big)\Big] \end{aligned} \tag{6.46}$$

By noting that Λ_C^* is convex, and attains its minimum at c, we can restrict the infimum to

$$\big\{\zeta_1 = \zeta_2\big\} \cup \big\{\zeta_1 \leq \zeta_2 \leq c(1-t)\big\}.$$

In the first case, when $\zeta_1 = \zeta_2$,

$$\begin{aligned} (6.46) &\geq \delta q + \Lambda_A^*(\zeta_1 - q + r + \xi - \zeta_1) + \Lambda_C^*(\zeta_2 + \xi - \zeta_1) \\ &\qquad \text{by convexity of } \Lambda_A^* \text{ and } \Lambda_C^* \\ &= \delta q + \Lambda_A^*(r - q + \xi) + \Lambda_C^*(\xi) \quad \text{since } \zeta_1 = \zeta_2 \\ &\geq \delta r + \Lambda_A^*(\xi) \quad \text{as it was for } \delta q + K(\xi, r, q). \end{aligned}$$

Now consider the alternative, that $\zeta_1 \leq \zeta_2 \leq c(1-t)$. In this case

$$
\begin{aligned}
(6.46) &\geq \delta q + (1-t)\Lambda_A^*\Big(\frac{\zeta_1 - q}{1-t}\Big) + t\Big[\Lambda_A^*\Big(\frac{r+\xi-\zeta_1}{t}\Big) + \Lambda_C^*\Big(\frac{\xi-\zeta_1}{t}\Big)\Big] \\
&\qquad \text{by non-negativity of } \Lambda_C^* \\
&\geq \delta q + (1-t)\Big[\Lambda_C^*\Big(\frac{\zeta_1 - q}{1-t}\Big) - \Lambda_C^*(\mu) + \delta\Big(\frac{\zeta_1 - q}{1-t} - \mu\Big)\Big] + t\Big[\cdots\Big] \\
&\qquad \text{using the definition of } \Lambda_A^* \\
&\geq (1-t)\Big[\Lambda_C^*\Big(\frac{\zeta_1}{1-t}\Big) - \Lambda_C^*(\mu) + \delta\Big(\frac{\zeta_1}{1-t} - \mu\Big)\Big] + t\Big[\cdots\Big] \\
&\qquad \text{since } \Lambda_C^* \text{ is decreasing in } (-\infty, c] \\
&= (1-t)\Lambda_A^*\Big(\frac{\zeta_1}{1-t}\Big) + t\Big[\Lambda_A^*\Big(\frac{r+\xi-\zeta_1}{t}\Big) + \Lambda_C^*\Big(\frac{\xi-\zeta_1}{t}\Big)\Big] \\
&\qquad \text{using the definition of } \Lambda_A^* \\
&= \delta r + (1-t)\Lambda_A^*\Big(\frac{\zeta_1}{1-t}\Big) + t\Big[\Lambda_A^*\Big(\frac{\xi-\zeta_1}{t}\Big) + \Lambda_C^*\Big(\frac{\xi-\zeta_1}{t}\Big)\Big] \\
&\qquad \text{using the definition of } \Lambda_A^* \text{ twice} \\
&\geq \delta r + \Lambda_A^*(\xi) + t\Lambda_C^*\Big(\frac{r+\xi-\zeta_1}{t}\Big) \quad \text{by convexity of } \Lambda_A^* \\
&\geq \delta r + \Lambda_A^*(\xi) \quad \text{by non-negativity of } \Lambda_C^*
\end{aligned}
$$

□

At several points in the proof, we used the definition of Λ_A^* (and the characterization of δ as $-\theta_\mu$) in order to rewrite the expression $\Lambda_A^*(\alpha) + \Lambda_C^*(\beta)$ as $\Lambda_A^*(\beta) + \Lambda_A^*(\alpha) + \delta(\alpha - \beta)$. This idea, of flipping arrival and service rates, was also used in proving $\delta = -\theta_\mu$, and recurs in the following result.

Lemma 6.14 *Under the assumptions of Theorem 6.11, and using the notation of Lemma 6.10,*

$$J(\xi, r) = \delta r + \Lambda_A^*(\xi).$$

Proof. Lemma 6.10 gives the expression

$$J(\xi, r) = \inf_{q\geq 0} \delta q + K(\xi, r, q) \wedge L(\xi, r, q).$$

Lemma 6.13 shows that $J(\xi, r) \geq \delta r + \Lambda_A^*(\xi)$. We will now show that

$$\inf_{q\geq 0}\ \delta q + K(\xi, r, q) \leq \delta r + \Lambda_A^*(\xi) \quad \text{if } \xi/c + r \geq 1$$

and

$$\inf_{q\geq 0}\ \delta q + L(\xi, r, q) \leq \delta r + \Lambda_A^*(\xi) \quad \text{if } \xi/c + r < 1.$$

This will complete the proof.

First, suppose $\xi/c + r \geq 1$. Pick $q = r + \xi/c - 1$. As we saw in (6.45) in the proof of Lemma 6.13,

$$\begin{aligned}\delta q + K(\xi, r, q) &= \delta r + \Lambda_A^*(\xi) + \Lambda_C^*(r - q + \xi)\\ &= \delta r + \Lambda_A^*(\xi) \quad \text{by choice of } q\end{aligned}$$

Now suppose $\xi/c + r < 1$. Pick $q = 0$, $t = r/(c - \xi)$, $\zeta_1 = \xi(1-t)$ and $\zeta_2 = c(1-t)$. Then $0 < t < 1$ (the trivial case $t = 0$ should be dealt with separately), and $\zeta_1 < \zeta_2$, so these parameters lie in C_2. With this choice of parameters,

$$\begin{aligned}\delta q + L(\xi, r, q) &\leq \delta q + \ (1-t)\Big[\Lambda_A^*\Big(\frac{\zeta_1 - q}{1-t}\Big) + \Lambda_C^*\Big(\frac{\zeta_2}{1-t}\Big)\Big]\\ &\qquad + t\Big[\Lambda_A^*\Big(\frac{r + \xi - \zeta_1}{t}\Big) + \Lambda_C^*\Big(\frac{\xi - \zeta_1}{t}\Big)\Big]\\ &= (1-t)\big[\Lambda_A^*(\xi) + \Lambda_C^*(c)\big] + t\big[\Lambda_A^*(c) + \Lambda_C^*(\xi)\big]\\ &= (1-t)\Lambda_A^*(\xi) + t\big[\Lambda_A^*(c) + \Lambda_C^*(\xi)\big] \quad \text{since } \Lambda_C^*(c) = 0\\ &= (1-t)\Lambda_A^*(\xi) + t\big[\Lambda_A^*(\xi) + \Lambda_C^*(c) + \delta(C - \xi)\big]\\ &\qquad\qquad \text{by flipping arrival and service rates}\\ &= \delta r + \Lambda_A^*(\xi) \quad \text{since } \Lambda_C^*(c) = 0.\end{aligned}$$

This completes the proof. □

Lemma 6.15 *Under the assumptions of Theorem 6.11, for any $k \in \mathbb{N}$ and $0 = t_0 < t_1 < \cdots < t_k$, the random vector*

$$\big(\tilde{D}^N(-t_k, -t_{k-1}], \ldots, \tilde{D}^N(-t_1, 0],\, Q_0/N\big)$$

satisfies an LDP in $\mathbb{R}^k \times \mathbb{R}^+$ with good rate function

$$J(\xi_j, \ldots, \xi_1,\, r) = \delta r + \sum_{i=1}^{k} (t_i - t_{i-1}) \Lambda_A^*\Big(\frac{\xi_i}{t_i - t_{i-1}}\Big)$$

Sketch proof. The proof is by induction on k. We will in fact only argue cases $k = 1$ and $k = 2$; then the induction argument should be clear.

When $k = 1$, this is just a slightly more general version of Lemma 6.14. That lemma described the case when $t_1 = 1$, but it is easy to extend it to general t_1 by simply rescaling the most likely path leading to $d(-1, 0] = \xi$ and $q_0 = r$ identified there.

Now consider the case $k = 2$. We already know (from Section 6.7) that the entire departure process $\tilde{D}^N$ is a continuous function of the scaled arrival and service processes, as is Q_0/N. This means that the vector we are interested in is also a continuous function, which means we can use the contraction principle to compute the rate function J. The variational problem we have to solve is

$$\begin{aligned}&\text{minimize} \quad I(a,c)\\ &\text{over} \quad (a,c) \in \mathcal{A}_\mu \times \mathcal{A}_\nu\\ &\text{subject to} \quad d(-t_2,-t_1] = \xi_2,\ d(-t_1,0] = \xi_1,\ q_0 = r.\end{aligned}$$

We will adopt the same strategy as in Lemma 6.9, of introducing a dummy variable q to represent the queue size at time t_{-1}, giving the expanded problem

$$\begin{aligned}&\text{minimize} \quad I(a,c)\\ &\text{over} \quad (a,c) \in \mathcal{A}_\mu \times \mathcal{A}_\nu,\ q \geq 0\\ &\text{subject to} \quad d(-t_2,-t_1] = \xi_2,\ q_{-t_1} = q,\ d(-t_1,0] = \xi_1,\ q_0 = r.\end{aligned}$$

Given q, this problem splits into two parts, one over $(-\infty,-t_1]$ and the other over $(-t_1,0]$. We know the solution to the first part, by the $k = 1$ case. It results in the simplified problem

$$\begin{aligned}&\text{minimize} \quad \delta q + (t_2 - t_1)\Lambda_A^*\Big(\frac{\xi_2}{t_2 - t_1}\Big) + \int_{-t_1}^0 \Lambda_A^*(\dot{a}_t) + \Lambda_C^*(\dot{c}_t)\,dt\\ &\text{over} \quad (a,c) \in \mathcal{A}^{t_1} \times \mathcal{A}^{t_1},\ q \geq 0\\ &\text{subject to} \quad q_{-t_1} = q,\ d(-t_1,0] = \xi_1,\ q_0 = r.\end{aligned}$$

But this is, apart from the addition of $\big((t_2 - t_1)\Lambda_A^*(\xi_2/(t_2 - t_1)\big)$, exactly the optimization problem that was solved in Lemma 6.10; and by Lemma 6.14 the solution is just

$$(t_2 - t_1)\Lambda_A^*\Big(\frac{\xi_2}{t_2 - t_1}\Big) + t_1\Lambda_A^*\Big(\frac{\xi_2}{t_1}\Big) + \delta r.$$

This completes the proof. □

We are finally ready to complete the proof of the main result.

Sketch proof of Theorem 6.11 We need to go from a discrete-time rate function like that in 6.15 to a full rate function for $(\tilde{D}^N, Q_0/N)$. The argument is similar to that of Mogulskii's theorem, described in Section 6.2.

First, consider $(\tilde{D}^N|_{(-t,0]}, Q_0/N)$. This is the projective limit of systems of the form appearing in Lemma 6.15, and by the Dawson-Gärtner theorem it satisfies an LDP in $\mathcal{C}^t \times \mathbb{R}^+$. It can be shown that the rate function is

$$J_t(d,r) = \begin{cases} \int_{-t}^0 \Lambda_A^*(\dot{d}_t)\,dt + \delta r & \text{if } d \in \mathcal{A}^t \\ \infty & \text{otherwise.} \end{cases}$$

The topology on $\mathcal{C}^t$ is that of pointwise convergence, which is what the Dawson-Gärtner theorem gives us, and let us write $(\mathcal{C}^t, \tau_p)$ to emphasize this. We would like to establish the LDP in $(\mathcal{C}^t, \tau_u) \times \mathbb{R}^+$, where by the first term we mean $\mathcal{C}^t$ equipped with the topology of uniform convergence. We can do this using the contraction principle. We know from Section 6.7 that $(\tilde{D}^N, Q_0/N)$ satisfies an LDP in $\mathcal{C}_\mu \times \mathbb{R}^+$, and so by the contraction principle $(\tilde{D}^N|_{[-t,0]}, Q_0/N)$ satisfies an LDP in $(\mathcal{C}^t, \tau_u) \times \mathbb{R}^+$ with some good rate function K_t. Applying the contraction principle again, this time to the identity map, $(\tilde{D}^N|_{[-t,0]}, Q_0/N)$ satisfies an LDP in $(\mathcal{C}^t, \tau_p) \times \mathbb{R}^+$ with the same good rate function K_t. But, by uniqueness of the rate function, $K_t = J_t$. Thus we obtain an LDP in the finer topology with good rate function J_t.

By taking projective limits again, we obtain an LDP for $(\tilde{D}^N, Q_0/N)$ in $\mathcal{X} \times \mathbb{R}^+$ with rate function

$$J(d,r) = \begin{cases} \int_{-\infty}^0 \Lambda_A^*(\dot{d}_t)\,dt + \delta r & \text{if } d \in \mathcal{A} \\ \infty & \text{otherwise} \end{cases}$$

where $\mathcal{X}$ is the space of continuous functions $\mathbb{R}^+ \to \mathbb{R}$ equipped with the topology of uniform convergence on compacts, which is what the Dawson-Gärtner theorem gives us. By the same argument we used before, we can strengthen this to an LDP in $\mathcal{C}_\mu \times \mathbb{R}^+$. □

Exercise 6.12
We have studied quasi-reversibility in the large deviations sense for the simple single-server queue. Kelly [52] describes a broad class of queueing network models which are quasi-reversible in the conventional sense. Are they also quasi-reversible in the large deviations sense? ◇

6.10 Scaling Properties of Networks

Often the variational problems that arise in networks are too complicated to be of practical use. However, there are some simple observations which 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 $\big(X(-t,0],\ t\in\mathbb{N}\big)$, 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.

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 $x^{\circlearrowright\kappa}$ by $x^{\circlearrowright\kappa}(-t,0]=x(-\kappa t,0]$ then $f(x^{\circlearrowright\kappa})=f(x)$. 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 $N^{-1}\tilde{X}^{\circlearrowright N}$ satisfies an LDP, from which we obtain an LDP for Q/N, from which we deduced (6.22), namely that

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

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

Theorem 6.16 *Suppose that the sequence of inputs $N^{-1}\tilde{X}^{\circlearrowright N}$ 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)*.*

Proof. By the contraction principle, Q/N satisfies an LDP with good rate function

$$J(q) = \inf\Big\{ I(x) : x \in \mathcal{A}_\mu, f(x) = q \Big\}$$

We will first show that $J(q) = qJ(1)$. Pick any $r > 0$ and consider $J(r)$. Since the rate function is good, the optimal path in $J(r)$ is attained; let it be y. Now consider the scaled path $z = (q/r)\, y^{\circlearrowright r/q}$. By (i) $f(z) = q$; and by a change of variables $I(z) = (q/r)\, I(y)$. Thus $J(q) \leq (q/r)\, J(r)$. So $J(q) \leq qJ(1)$, and $J(1) \leq J(q)/q$ for $q > 0$, and so $J(q) = qJ(1)$.

Second, we will show that $J(1) > 0$. Suppose $J(1) = 0$, and let y be the optimal path. By the form of the rate function, y is absolutely continuous and $\Lambda^*(\dot{y}_t) = 0$ for almost all t. Since the rate function is strictly convex, $\dot{y}_t = \mu$ for almost all t. By (ii), $f(y) = 0$, which contradicts its optimality for $J(1)$.

Third, we will show that $J(1) < \infty$. Let x be the path in (iii). By the above bound, $J(1) \leq I(x)/f(x) < \infty$.

Now consider the event $\{Q/N > q\}$. As $f(q)$ is linear in q, the large deviations upper and lower bounds agree, and so

$$\lim_{N\to\infty} \frac{1}{N} \log P(Q/N > q) = -J(1)q.$$

If we write q' for N and let $q = 1$ we obtain the desired result. □

6.11 Statistical Inference for the Tail-Behaviour of Queues

A key observation we made in Section 6.4 is that the queue length distribution has an exponential tail, under very general assumptions about the arrival and service processes. On the basis of the large deviations limit result (6.22) we make the approximation

$$P(Q_0 > q) \approx \exp(-\delta q)$$

In that section we explained how the parameter δ is related to the distributions of the arrival and service processes. However, in practice, those distributions are rarely known. This motivates us to consider how to estimate δ from observations, either of the arrival and service processes or the queue length process. In this section we will consider this problem, in both frequentist and Bayesian settings.

Note. Exactly the same estimation problem arises in risk theory, in which context δ is called the risk adjustment coefficient.

A Frequentist Approach

Recall the discussion of the single-server queue in Section 1.3. Let $A(-t,0]$ be the amount of work arriving at the queue in interval $(-t,0]$, let $C(-t,0]$ be the service offered, and let $X(-t,0] = A(-t,0] - C(-t,0]$. We showed in Theorem 3.1 that

$$\delta = \sup\bigl\{\theta > 0 : \Lambda(\theta) < 0\bigr\} \quad \text{where} \quad \Lambda(\theta) = \lim_{t\to\infty} \log Ee^{\theta X(-t,0]} \tag{6.48}$$

Suppose first that $(\dot{X}_t,\ t \in \mathbb{Z})$ are i.i.d.. One approach to estimating δ is to specify a parametric model for the distribution of $\dot{X}_t$, estimate the parameters from observations of $\dot{X}_1, \dots, \dot{X}_n$, and use the fitted model to compute δ. This is usually appropriate if the distribution of $\dot{X}_t$ can be described by a simple model.

In more complicated situations, there is typically far more information in the model than is needed to quantify its impact on the build-up of queues. The formula (6.48) suggests that one can, alternatively, adopt a non-parametric approach and estimate δ directly, without reference to a particular model. Thus, if we use the empirical distribution of $\dot{X}_1, \dots, \dot{X}_n$ as an estimate of the distribution of $\dot{X}_1$, we obtain the estimate

$$\hat{\Lambda}(\theta) = \log\left(\frac{1}{n}\sum_{i=1}^{n} \exp\theta\dot{X}_i\right)$$

for its cumulant generating function. Substituting $\hat{\Lambda}$ into the equation for δ yields an estimate $\hat{\delta}$ for δ. This approach, for estimating δ and finding confidence intervals for it, has been studied by Pitts et al. [83] and also by [32].

More generally, when $(\dot{X}_t,\ t \in \mathbb{Z})$ are not independent, the formula (6.48) suggests the following estimator for the limiting scaled cumulant generating function:

$$\hat{\Lambda}(\theta) = \frac{1}{t}\log\left(\frac{1}{n-t+1}\sum_{i=0}^{n-t} \exp\theta X(i,i+t]\right)$$

Thus, we divide up the observations into overlapping blocks of size t. We estimate the moment generating function corresponding to a block of size t by averaging over blocks. We use overlapping blocks instead of non-overlapping blocks with the intention of reducing the variance of the estimator. The block size t needs to be chosen carefully. We want t to be large, since $\Lambda(\theta)$

is given by a limit as $t \to \infty$ and so by increasing t we decrease the bias of the estimator. On the other hand, we want t to be small, since this gives more independent blocks of size t in the data, and hence a lower variance for the estimator. Statistical properties of this estimator are discussed in [32, 42, 86].

These estimators rely on observations of the input process. An alternative approach is based directly on observations of the queue length; in the light of the scaling result of Section 6.10 this can of great practical value. Such an approach was first suggested by Courcoubetis et al. [19], and has subsequently been studied by many others. The basic idea is to take a sample of the queue length process, estimate the distribution function $P(Q > q)$ by the empirical distribution $\hat{P}(Q > q)$, plot $\log \hat{P}(Q > q)$ against q, draw a straight line through the points, and estimate δ by the slope of the line.

There are not many rigorous results concerning direct estimators. One case where rigorous results are known is as follows: Let M_t be the maximum queue size attained in the interval $[0, t]$ and let $\hat{\delta}_t = -\log M_t/t$. Glynn and Zeevi [47] have shown that, under very general conditions, this estimator is strongly consistent i.e. $\hat{\delta}_t \to \delta$ almost surely as $t \to \infty$. However, convergence is slow; the error is of order $1/\log t$ in probability.

A Bayesian Approach

We now consider a Bayesian approach to estimating the probability that the queue size is large. Assume again that the $(\dot{X}_t,\ t \in \mathbb{Z})$ are i.i.d., and let L be their common probability law. We will suppose we start with a prior distribution $\pi(L)$ for L, make a sequence of observations $X|_{[1,n]} = (\dot{X}_1, \dots, \dot{X}_n)$, use Bayes's rule to find the posterior distribution $\pi'_n(L) = \pi\big(L \mid X|_{[1,n]}\big)$, and we want to compute the expected posterior loss probability

$$E_{\pi'_n}\Big[P_L(Q > q)\Big].$$

In fact it is difficult to compute this exactly in all but a few special cases. Instead, we will use large deviations techniques to find

$$\lim_{n\to\infty} \frac{1}{n} \log E_{\pi'_n}\Big[P_L(Q > nq)\Big]. \tag{6.49}$$

(Note that we are scaling both the number of observations and the queue size threshold.) We know that $P_L(Q > q) \approx e^{-q\delta(L)}$ where $\delta(L)$ is the standard large deviations rate; we would therefore expect (6.49) to be involve $-q\delta(M)$

for some 'most likely' posterior law M. The following theorem makes this rigorous.

For technical reasons, we have already assumed that the $(\dot{X}_t,\ t \in \mathbb{Z})$ are i.i.d. We will also assume that they take values in a finite set A. Let $M_1(A)$ denote the set of probability measures on A, and for $L \in M_1(A)$ write $L(a)$ for $L(\{a\})$.

> *Note.* The following result involves large deviations for empirical distributions. You may like to look back to Sanov's theorem in Section 2.7 before proceeding. Sanov's theorem involves the relative entropy $H(L|M)$ as a function of L; it not a typographical error that the result below involves the relative entropy as a function of M.

Theorem 6.17 *Under the above assumptions, for L-almost every X,*

$$(6.49) = -\inf_{M \in \operatorname{supp} \pi} q\delta(M) + H(L|M)$$

where

$$H(L|M) = \sum_{a \in A} L(a) \log \frac{L(a)}{M(a)}$$

is the relative entropy *of L with respect to M, and* $\operatorname{supp}\pi$ *is the support of π.*

Sketch proof. We will use Varadhan's lemma. This involves a continuous function and an LDP. The continuous function we will use will be an estimate of $P_L(Q > nq)$, and the LDP will be for a random measure drawn from the posterior distribution π'_n.

It is a standard result, a refinement of Wald's approximation for hitting probabilities of random walks, that

$$c_1 e^{-q\delta(L)q} \le P_L(Q > q) \le c_2 e^{-q\delta(L)}$$

for some positive constants c_1 and c_2. From this it follows that (6.49) is equal to

$$\lim_{n \to \infty} \frac{1}{n} \log E_{\pi'_n} e^{-nq\,\delta(L)} \tag{6.50}$$

provided this limit exists. Now, $\delta(L)$ is a continuous bounded function on $M_1(A)$. To see this, consider the characterization of $\delta(L)$ given in (6.48). In the present case it simplifies, since

$$\Lambda(\theta) = \log\Bigl(\sum_{a \in A} L(a) e^{\theta a}\Bigr).$$

From this it is easy to verify the necessary properties of δ.

Now for the posterior distribution. In fact, by our assumption that the $\dot{X}_t$ are i.i.d., the posterior distribution $\pi(L|X_{[1,n]})$ simplifies to $\pi(L|L_n)$ where L_n is the empirical distribution

$$L_n(a) = \frac{1}{n}\sum_{i=1}^{n} 1[\dot{X}_i = a].$$

Let L^n be a random measure on A drawn from the posterior distribution $\pi(\cdot|L_n)$. It can be shown that L^n satisfies an LDP in $M_1(A)$ with good rate function

$$I(M) = \begin{cases} H(L|M) & \text{if } M \in \operatorname{supp}\pi \\ \infty & \text{otherwise} \end{cases}$$

where L is the distribution of the $\dot{X}_t$. A proof of this can be found in [40]; it is also implicit in [29], and is a corollary of a more general result for posteriors of Markov transition matrices established in [80].

We can now apply Varadhan's lemma, to conclude that

$$(6.50) = \sup_{M \in M_1(A)} -q\delta(M) - I(M).$$

This completes the proof. □

While we are still faced with a variational problem, this is much simpler than exact computation of the posterior expectation, and can be solved numerically. It can also be simplified in special cases. Under additional conditions on the prior, the result can be extended to i.i.d. $\dot{X}_t$ taking values in a compact set [43, 61]. It can also be extended to dependent sequences; the Markovian case is treated in [39, 81].

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:

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

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

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 151–181, 2004.

$x(-t,0]$	for $x(t)$
$x(-t,-u]$	for $x(t)-x(u)$, when $t \geq u$
$x\vert_{(-t,0]}$	for the restriction of x to $\{-(t+1),\dots,0\}$
$\dot{x}_{-t}$	for $x(t+1)-x(t)$

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

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

Any arrival process A^N lies in this set; we do not need any polygonalization tricks, and we don't need to switch to continuous time. This simplifies working with queue-size functions.

For example, consider the processor-sharing queue in Section 5.10. In that section we started with a pair of discrete-time equations (5.19) which describe the system. With a finite-horizon boundary condition, they yield a finite-horizon queue-size function $q(\cdot|_{(-T,0]})$. We then proposed a pair of continuous-time equations (5.22) with the same finite-horizon boundary condition, and established three things:

- that the continuous-time queue-size equations have a unique solution $\tilde{q}^{-T}$ over any finite time horizon;
- that they are consistent with the discrete-time queue-size functions (i.e. $q(x) = \tilde{q}(\sim(x))$, where $\sim$ is the polygonalization operator);
- that the continuous-time finite-horizon functions $\tilde{q}^{-T}$ are continuous on $\mathcal{C}^T$, that is, with respect to the topology of uniform convergence on compact intervals.

For our work on the many-flows scaling it is sufficient to work in discrete time. This makes things simpler:

- the discrete-time equations (5.19), together with the finite-horizon boundary condition, clearly have a unique solution q^{-T};
- we haven't defined any new functions, so we don't need to check consistency;
- the discrete-time finite-horizon functions q^{-T} are clearly continuous on $\mathcal{D}$ with respect to the topology of uniform convergence on compact intervals, i.e. with respect to pointwise convergence.

So, it is simpler to describe the finite-horizon behaviour of the queueing system in discrete time than in continuous time. We still need to extend to the infinite-horizon boundary condition, that 'the queue was empty at time $-\infty$', and to prove that the infinite-horizon queue-size function is continuous. We have already done all this work in continuous time:

- we defined the infinite-horizon queue-size function $\tilde{q}(x) = \lim_{T\to\infty} \tilde{q}^{-T}(x)$;
- we proved that $\tilde{q}$ was continuous on $\mathcal{C}_\mu$.

We have done this work, so we will simply

- define the infinite-horizon queue-size function $q(x) = \lim_{T\to\infty} q^{-T}(x)$;
- note that $q(x) = \tilde{q}(\sim(x))$, and that the map $\sim$ is continuous (with respect to a topology on $\mathcal{X}$ which we will define in a moment), and conclude that if $\tilde{q}$ is continuous then so is q.

It is possible to prove continuity directly, but given the work we have done already, there is no need.

Mean arrival rate. The part that is harder is dealing with mean arrival rates. The trouble is that a typical arrival process A^N may not have the right mean rate. For example, let $A^{(i)}$ be independent copies of a constant-rate random process, each rate drawn independently from Normal(μ, σ^2). Then the limit

$$\lim_{t\to\infty} \frac{A^N(t)}{t+1}$$

is not necessarily equal to μ—in fact, the limit is a normal random variable with mean μ and variance $N^{-1}\sigma^2$. We can define a space $\mathcal{D}_\mu$ analogously to $\mathcal{C}_\mu$, but A^N almost surely does not lie in $\mathcal{D}_\mu$, and so we cannot speak about a large deviations principle for A^N in $\mathcal{D}_\mu$.

The way around this problem is to work in a larger space and use the extended contraction principle, Theorem 4.6. It turns out that the rate function of any sample path not in $\mathcal{D}_\mu$ is infinite, which essentially means we can ignore those sample paths that are not in a neighbourhood of $\mathcal{D}_\mu$. To make this precise, we will define some more topological spaces. We will see how to use them in Section 7.5.

For convenience define the lower and upper mean rates of an arrival process:

$$\begin{aligned} \underline{x} &= \liminf_{t\to\infty} \frac{x(t)}{t+1} \\ \bar{x} &= \limsup_{t\to\infty} \frac{x(t)}{t+1}. \end{aligned}$$

Equip $\mathcal{D}$ with the extended scaled uniform topology defined below; we will obtain an LDP for A^N in this space.

Definition 7.1 *Define the* extended scaled uniform topology *on $\mathcal{D}$ as follows. If $\underline{x} = \infty$ or $\bar{x} = -\infty$ then let x be an isolated point. Equip the remainder of $\mathcal{D}$ with the topology obtained from the normal scaled uniform norm* (5.4)*: recall its definition*

$$\|x\| = \sup_{t\geq 0} \left| \frac{x(t)}{t+1} \right|.$$

Next, define two subspaces of $\mathcal{D}$:

$$\mathcal{D}_{(\mu,\nu)} = \{x \in \mathcal{D} : \mu < \underline{x} \leq \bar{x} < \nu\}$$

and

$$\mathcal{D}_{[\mu,\nu]} = \{x \in \mathcal{D} : \mu \leq \underline{x} \leq \bar{x} \leq \nu\}.$$

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 $\mathcal{D}_\mu$, 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 $\mathcal{D}_\mu$, which is Polish and hence separable.

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*

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

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 $v : \mathbb{N} \to \mathbb{R}$ for which $v_t / \log t \to \infty$. 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*

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

where

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

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 $X^N|_{(-t,0]}$ 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 $(\mathcal{D}, \tau_p)$, 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 $(\mathcal{D}, \tau_p)$ 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 $(\mathcal{D}, \|\cdot\|) \to (\mathcal{D}, \tau_p)$ is continuous, which is trivial; an LDP for X^N in $(\mathcal{D}, \tau_p)$, which we have just found; and exponentially 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 $\Lambda_t^*(x|_{(-t,0]})$ is increasing in t. □

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,

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

Now suppose that $\bar{x} > \mu$. Then, for any $\varepsilon > 0$, there exist infinitely many t such that $x(-t, 0]/t > \bar{x} - \varepsilon$. By convergence of the $\tilde{\Lambda}_t$, given $\theta > 0$ sufficiently small and $\varepsilon > 0$, t it is the case that $\tilde{\Lambda}_t(\phi)/\phi < \tilde{\Lambda}(\phi)/\phi + \varepsilon$ for sufficiently large t. Putting these together, for $\phi > 0$ sufficiently small and for infinitely many t,

$$I(x) \geq \phi v_t \big[\bar{x} - \tilde{\Lambda}(\phi)/\phi - 2\varepsilon\big].$$

We will show shortly that $\tilde{\Lambda}'(0) = \mu$. Thus there exists a $\phi_0 > 0$ such that $\tilde{\Lambda}(\phi_0) < \phi_0(\mu + \varepsilon)$, and so, for some sufficiently small $\phi < \phi_0$ and for infinitely many t,

$$I(x) \geq \phi v_t \big[\bar{x} - \mu - 3\varepsilon\big].$$

Letting $t \to \infty$, recalling from the definition of a scaling function that $v_t \to \infty$, and choosing ε sufficiently small,

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

Similarly, if $\bar{x} < \mu$ we can choose $\phi < 0$, with the same conclusion.

It remains to show that $\tilde{\Lambda}'(0) = \mu$. We have assumed that $\tilde{\Lambda}$ is differentiable at 0. Furthermore, it is the pointwise limit of convex functions which are all differentiable at 0, with common derivative μ. By Lemma 1.12, $\tilde{\Lambda}'(0) = \mu$. □

Exponential Tightness

We now prove claim about exponential tightness in Theorem 7.1. The proof is rather technical, and should be omitted on first reading.

Lemma 7.3 *If X^N is regular over finite and infinite horizons, then it is exponentially tight in $(\mathcal{D}, \|\cdot\|)$. In other words, there exist compact sets K_α in $(\mathcal{D}, \|\cdot\|)$ for which*

$$\lim_{\alpha\to\infty} \limsup_{N\to\infty} \frac{1}{N} \log P(X^N \notin K_\alpha) = -\infty.$$

A suitable choice is

$$K_\alpha = \Big\{ x \in \mathcal{D} : \frac{x(-t,0]}{t} \in \big[\mu - \alpha\delta_t, \mu + \alpha\delta_t\big] \text{ for all } t \Big\}$$

for well-chosen δ_t, where $\mu = \tilde{\Lambda}'(0)$. The proof of compactness is given in Lemma 7.4; it requires that δ_t be sufficiently small. The proof of the limit is given in Lemma 7.5; it requires that δ_t be sufficiently large. The correct choice of δ_t hinges on the scaling function v_t in the definition of infinite-horizon regularity.

Lemma 7.4 *If $(\delta_t,\ t \in \mathbb{N})$ is chosen so that $\delta_t \to 0$, then the sets K_α are compact in $\mathcal{D}$.*

Proof. Note first that $K_\alpha = \bigcap_T K_\alpha(T)$, where

$$K_\alpha(T) = \Big\{ x \in \mathcal{D} : \frac{x(-t,0]}{t} \in \big[\mu - \alpha\delta_t, \mu + \alpha\delta_t\big] \quad \text{for } t \le T \Big\}$$

Because we are working in a metric space, it suffices to show that the sets K_α are sequentially compact. So, let x^k be a sequence of processes. Since the T-dimensional truncation of $K_\alpha(T)$ is compact in $\mathbb{R}^T$ for each T, the intersection K_α is compact under the projective limit topology. That is, there is a subsequence $x^{j(k)}$ which converges pointwise, say to x. It remains to show that $x^j \to x$ under the scaled uniform topology on $\mathcal{D}$.

Given any $\varepsilon > 0$, since $\delta_t \to 0$ as $t \to \infty$, we can find t_0 such that for $t \ge t_0$, $2\delta_t\alpha < \varepsilon$. And since x and all the x^j are in K_α,

$$\sup_{t \ge t_0} \Big| \frac{x^j(-t,0]}{t} - \frac{x(-t,0]}{t} \Big| < \varepsilon.$$

Also, since the x^j converge pointwise, there exists a j_0 such that for $j \ge j_0$

$$\sup_{t < t_0} \Big| \frac{x^j(-t,0]}{t} - \frac{x(-t,0]}{t} \Big| < \varepsilon.$$

Putting these two together gives the result. □

Lemma 7.5 *There is a choice of $(\delta_t,\ t \in \mathbb{N})$ for which $\delta_t \to 0$ as $t \to \infty$, and also*

$$\lim_{\alpha\to\infty} \limsup_{N\to\infty} \frac{1}{N} \log P(X^N \notin K_\alpha) = -\infty. \tag{7.3}$$

Proof. We will split up the set K_α into several parts. First note that if $K_\alpha = L_\alpha \cap M_\alpha$, and that if both L_α and M_α satisfy conditions of the form (7.3) then so does K_α.

The first way we will split K_α is by

$$P(X^N \notin K_\alpha) \leq \sum_{t\in\mathbb{N}} P\Big(\frac{X^N(-t,0]}{t} > \mu+\alpha\delta_t\Big)+\sum_{t\in\mathbb{N}} P\Big(\frac{X^N(-t,0]}{t} < \mu+\alpha\delta_t\Big). \tag{7.4}$$

We will only prove the condition (7.3) for the first term, as the second term can be dealt with similarly.

We will split the probability further, breaking the infinite sum up into several parts: several finite-timescale parts, and a long-timescale infinite part. This strategy is at the core of the proof of our first many-flows limit theorem, Theorem 1.8.

First, fix t and consider a single term in the sum.

$$\begin{aligned}\limsup_{N\to\infty}\frac{1}{N}\log P\Big(\frac{X^N(-t,0]}{t} > \mu+\alpha\delta_t\Big)&\\ \leq -\Big[\theta(\mu t+\alpha t\delta_t)-\Lambda_t(\theta e)\Big] \quad &\text{for all } \theta>0\end{aligned}$$

by Chernoff's bound. Choosing any θ for which $\Lambda_t(\theta e)$ is finite, we see that this quantity → $-\infty$ as $\alpha\to\infty$.

Now for the remaining terms in the sum. We will show below that

$$\limsup_{N\to\infty}\frac{1}{N}\log\sum_{t>t_0} P\Big(\frac{X^N(-t,0]}{t} > \mu+\alpha\delta_t\Big) \leq -\alpha\log(t_0+1) \tag{7.5}$$

for t_0 sufficiently large, not depending on α, and $\alpha \geq 1$. This will complete the proof. To show (7.5) we will first explain how to choose δ_t, and then go on to establish the limit.

Choice of δ_t. By assumption, the functions $\tilde\Lambda_t$ converge pointwise to $\tilde\Lambda$, which is differentiable, hence finite and continuous, in a neighbourhood of the origin. Let $\phi > 0$ be such that $\tilde\Lambda$ is differentiable on $|\theta| \leq \phi$. Clearly, $\tilde\Lambda_t$ is also finite at $\theta = \pm\phi$ for t sufficiently large. By Lemma 2.3 the scaled cumulant generating function $\tilde\Lambda_t$ is convex and continuous on the interior of its effective domain. Thus there exists a positive $\phi' < \phi$ such that $\tilde\Lambda_t$ is continuous on $|\theta| \leq \phi'$, for t sufficiently large. By convexity, the pointwise convergence $\tilde\Lambda_t \to \tilde\Lambda$ must be uniform. In other words, if we set

$$\varepsilon_t = \sup_{u>t}\sup_{|\theta|\leq\phi'}\big|\tilde\Lambda_t(\theta)-\tilde\Lambda(\theta)\big|$$

then $\varepsilon_t \downarrow 0$ as $t\to\infty$. Now define a sequence θ_t by

$$\theta_t = \big(\sqrt{\varepsilon_t}+\nu_t\big)\wedge\phi' \quad\text{where}\quad \nu_t=\sqrt{\frac{\log t}{v_t}}$$

By definition of scaling function, $\nu_t \to 0$; thus $\theta_t \to 0$ as $t \to \infty$. 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 $t \to \infty$, 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 $\nu_t \to 0$. Thus $\delta_t \to 0$ as $t \to \infty$.

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,

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

$$\leq \limsup_{N\to\infty} \frac{1}{N} \log \sum_{t>t_0} \exp\Big(-N\psi_t(\mu t + \alpha t\delta_t) + N\Lambda_t(\psi_t)\Big)$$

$$\text{(for any choice of } \psi_t > 0\text{, by Chernoff's bound)} \tag{7.7}$$

$$\leq \limsup_{N\to\infty} \frac{1}{N} \log \sum_{t>t_0} \exp\Big(-Nv_t\Big[\theta_t(\mu + \alpha\delta_t) - \tilde{\Lambda}_t(\theta_t)\Big]\Big)$$

$$\text{(by choosing } \psi_t = \theta_t v_t/t) \tag{7.8}$$

(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

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

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

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

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

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. □

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 $(X_t, t \in \mathbb{Z})$ is a stationary Gaussian process characterized by its mean and covariance structure:

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

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

> *Note.* Instead of Σ_t, we could specify the autocorrelation structure
>
> $$\rho_t = \text{Cov}(X_{-t+1}, X_0)$$
>
> 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,

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

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

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

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

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

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

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

There is no simple form for the rate function

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

unless Σ^t is invertible in which case

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

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 $\theta \in \mathbb{R}^t$

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

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

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

(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

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

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 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

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

and

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

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

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

Is this regular over the infinite horizon?

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

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

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 $\theta = 0$. ◇

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.

The rate function is

$$I(x) = \sum_{t=-\infty}^{0} \Lambda_1^*(x_t).$$

This is qualitatively just like the rate function in Chapter 6, for large-buffer scalings; and so the basic method of argument used there, namely straightening sample paths into piecewise linear paths, carries through. ◇

7.5 Applying the Contraction Principle

We now have the ingredients to apply the extended contraction principle, Theorem 4.6, which gives us

Corollary 7.6 *Suppose X^N is the average of N independent copies of a process X, and X^N satisfies the sample path LDP and has mean rate μ. Suppose f is a function which is continuous on $\mathcal{D}_{[\mu-\varepsilon,\mu+\varepsilon]}$ for some $\varepsilon > 0$. Then $Y^N = f(X^N)$ satisfies an LDP with good rate function*

$$J(y) = \inf_{x\in\mathcal{D}:f(x)=y} I(x) = \inf_{x\in\mathcal{D}_\mu:f(x)=y} I(x)$$

where $I(x)$ is given by (7.2).

Proof. We have the ingredients for the extended contraction principle: by Theorem 7.2, the rate function is infinite outside $\mathcal{D}_\mu$; it is easy to see that $\mathcal{D}_{(\mu-\varepsilon,\mu+\varepsilon)}$ is an open neighbourhood of $\mathcal{D}_\mu$; and it is also easy to see that $\mathcal{D}_{[\mu-\varepsilon,\mu+\varepsilon]}$ is closed. This gives us: Y^N satisfies an LDP with good rate function

$$J(y) = \inf_{x\in\mathcal{D}:f(x)=y} I(x) = \inf_{x\in\mathcal{D}_{[\mu-\varepsilon,\mu+\varepsilon]}:f(x)=y} I(x).$$

Since the rate function is infinite outside $\mathcal{D}_\mu$, we can restrict the infimum to $\mathcal{D}_\mu$. This completes the proof. □

The trouble with the rate function $I(x)$ is that it can be very general. This makes it hard to say anything about $J(y)$. Only in some special cases can it be simplified to something useful. In the following sections we will look at some of these special cases.

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*

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

First an example.

Example 7.5 (Fractional Brownian motion)
Let A be a fractional Brownian motion input as in Example 7.2. This has

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

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

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

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. ◇

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. □

Lemma 7.8 *If A^N is regular over both finite and infinite horizons, and has mean rate μ, then*

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

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. □

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*

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

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

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

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 $a = e\mu$, 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

$$K_t(r) = \inf_{\substack{a \in \mathbb{R}^t: \\ a(-t,0]=r}} \Lambda_t^*(a).$$

By the contraction principle, K_t is a rate function, and in particular K_t is non-negative. Since Λ_t^* is convex, so is K_t. Since $\Lambda_t^*(e\mu) = 0$, $K_t(\mu t) = 0$, 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

$$\begin{aligned} I(\hat{a}) &= \sup_t \Lambda_t^*(\hat{a}|_{(-t,0]}) \\ &\geq \Lambda_{\hat{t}}^*(\hat{a}|_{(-\hat{t},0]}) \geq (7.11). \end{aligned}$$

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 Λ_t^*. And an optimal $\hat{t}$ is also attained. For suppose not, and take a sequence $t_n \to \infty$ and $a^n|_{(-t_n,0]}$ with $a^n(-t_n,0] = q + Ct_n$ and $\Lambda_t^*(a^n|_{(-t_n,0]})$ 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 $a^k \to a$, also with $I(a) < K$. But then $a(-t_k,0]/t_k \to C$ 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 $I(\hat{a}) = \Lambda_t^*(\hat{a}|_{(-\hat{t},0]})$. 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

$$\begin{aligned}\inf_t \inf_{\substack{a\in\mathbb{R}^t:\\a(-t,0]=q+Ct}} \Lambda_t^*(a|_{(-t,0]}) &\geq \inf_{s\neq\hat{t}} \inf_{\substack{a\in\mathbb{R}^s:\\a(-s,0]=q'+Ct}} \Lambda_s^*(a|_{(-s,0]})\\ &\geq \inf_{s\neq\hat{t}} \inf_{\substack{a\in\mathbb{R}^s:\\a(-s,0]=q+Ct}} \Lambda_s^*(a|_{(-s,0]})\end{aligned}$$

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

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

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

$$\begin{aligned}\sup_{\theta\in\mathbb{R}}\theta x-\Lambda_t(e\theta) &= \sup_{\theta\in\mathbb{R}}\theta x-\Big[\sup_{y\in\mathbb{R}}\sup_{\substack{a\in\mathbb{R}^t:\\a(-t,0]=y}} e\theta\cdot a-\Lambda_t^*(a)\Big]\\ &= \sup_{\theta\in\mathbb{R}}\inf_{y\in\mathbb{R}}\theta(x-y)+K_t(y)\\ &= K_t(x)+\sup_{\theta\in\mathbb{R}}\inf_{y\in\mathbb{R}}K_t(y)-\big(K_t(x)+\theta(y-x)\big)\\ &\geq K_t(x),\end{aligned}$$

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

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

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

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 $\theta > 0$.

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. □

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 $J(q) = \bar{J}(q)$. □

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*

$$J^+(0) = \sup_\theta \theta C - \Lambda_1(\theta e)$$

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

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

This reduces to

$$\inf_{t>0} \sup_\theta \theta Ct - \Lambda_t(\theta e)$$

as in Theorem 7.7. By convexity,

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

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 $\varepsilon > 0$, let

$$a^\varepsilon = (\dots, a_{-2}, a_{-1}, \varepsilon + a_0).$$

Then $q(a^\varepsilon) > 0$ so $a^\varepsilon \in F$. Also $a^\varepsilon \to a$ as $\varepsilon \to 0$, 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.

LHS≥RHS in (7.13). Let $a \in \bar{F}$. Then there exist $a^n \to a$ in F, and $f(a^n) \to f(a)$ by Theorem 5.3. If $f(a) > 0$ then

$$I(a) \geq \inf_{q>0} J(q) \geq \inf_{t>0} \sup_{\theta} \theta Ct - \theta Ct - \Lambda_t(\theta e)$$

because the optimal $\hat{t}$ in (7.12) must be strictly positive for $q > 0$. So suppose $q(a^n) \to 0$. As in Theorem 5.3, there exist an n_0 and t_0 such that, for $n \geq n_0$,

$$q(a^n) = \sup_{t \leq t_0} a^n(-t, 0] - Ct.$$

And because $q(a^n) > 0$, the supremum must be attained at $t > 0$. Some t must be repeated infinitely often as $n \to \infty$; for that t, $a(-t, 0] = Ct$. Taking the infimum over such a gives the result. □

The same technique can be used to estimate the probability that a queue with a finite buffer is non-empty, or that work is lost. Let $\bar{f}$ be the queue size function for a queue with finite buffer B.

Corollary 7.12 *If A^N satisfies the conditions of Theorem 7.11, and $B > 0$, then $\{\bar{f}(A^N) > 0\}$ has the same large deviations bounds as $\{f(A^N) > 0\}$.*

Proof. If $\bar{f}(a) > 0$ then $f(a) > 0$ also, so the same upper bound works. As for $\{q > 0\}$, the lower bound is straightforward. □

The technique for estimating the probability of lost work is similar, so the following is left as an exercise.

Exercise 7.6
Show that the event that incoming work is lost has large deviations lower bound $-J(B^+)$ and upper bound $-J(B)$ (or $-J^+(0)$ if $B = 0$). ◇

7.9 Paths to Overflow

The expressions for the rate function in Section 7.6 tell us more than just the probability that the queue reaches a certain level: they tell us *how* the queue reaches that level. Because the rate function I is good, the infimum in

$$J(q) = \inf_{a \in \mathcal{D}: q(a) = q} I(a) \tag{7.14}$$

is attained, as long as $J(q) < \infty$. Furthermore, the sample path LDP tells us the probability of any deviation from this path.

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 $t \geq \hat{t}$,*

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

where s is the step function

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

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 $\hat{a}|_{(-t,0]} = \nabla\Lambda_t(\hat{\theta}s)$. (These definitions, one for each t, are clearly all consistent.) This path has the right rate function: using Lemma 2.4, $\Lambda_t^*(\hat{a}|_{(-t,0]})$ is equal to (7.9). And it also causes the queue to reach at least the right level: from differentiating $\theta(q + C\hat{t}) - \Lambda_{\hat{t}}(e\theta)$ 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. □

Example 7.7 (Gaussian sources)
Let A be a Gaussian, as in Example 7.1. It is easy to work out the optimal path:

$$\nabla\Lambda_t(\theta s|_{(-t,0]}) = \mu e + \theta\Sigma_t s.$$

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

Consider the case of fractional Brownian motion, 7.2, where

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

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

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

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. ◇

Exercise 7.8
Let A be a single-step autoregressive process:

$$A_t = \mu + a(A_{t-1} - \mu) + \sqrt{1 - \alpha^2}\sigma\varepsilon_t$$

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

$$a_{-t} = \mu + \hat{\theta}\sigma^2\Big(1 + \frac{1 - a^{t+1}}{1 - a} + \frac{1 - a^{\hat{t}-t}}{1 - a}\Big).$$ ◇

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

$$a_{-t} = \big(\nabla\Lambda_t(\hat{\theta}s)\big)_{-t} = \frac{E(A_{-t}e^{\theta A(-\hat{t},0]})}{E(e^{\theta A(-\hat{t},0]})}.$$

We can now calculate

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

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

$$a_{-t} = \frac{qhe^{\theta h}F_{\hat{t}-t-1}F_t}{qF_{\hat{t}} + pG_{\hat{t}}}.$$

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.) ◇

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

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

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

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

Proof. It is easy to prove lower bounds for rate functions: we just need to make some simple observations on properties of optimal paths.

If $J(\cdot, r)$ is finite, then optimal paths (a, b) must be attained, since the rate function is good. For such paths, there must be a last time $-t$ at which both queues were empty; and after time $-t$, $a(-t, 0] + b(-t, 0] \geq r + Ct$. Now apply the contraction principle to the sample path LDP for (A^N, B^N) to find that the rate function for $x = a(-t, 0] + b(-t, 0]$ is

$$\sup_\theta \theta x - \Lambda_t(\theta e) - \mathrm{M}_t(\theta e).$$

As in Theorem 7.7, this is increasing in x for $x \geq C$. Taking the infimum over t yields the result. □

The expression on the right hand side in (7.15) looks just like the rate function for overflow $J(r)$, given in (7.9), for a single server queue fed by B^N, except that the constant service rate C has been replaced by an *effective service rate* $\tilde{C}(\theta, t) = C - (\theta t)^{-1}\Lambda_t(\theta e)$. In other words, the low priority flow sees an effective service rate of at least $\tilde{C}(\theta, t)$, the total service rate less the effective bandwidth of the high priority flow. (The sense of 'effective service rate' and 'effective bandwidth' is described later in Section 10.1.)

Berger and Whitt [6] have observed a similar result in the large-buffer scaling, and they stress the point that approximating $J(\cdot, r)$ by the expression in (7.15) leads to simple control decisions. On the other hand, [101] gives an example showing when the inequality is strict. Priority queues have also been studied by Mandjes and van Uitert [70], Shakkottai and Srikant [90] and Delas et al. [23].

7.11 Departures from a Queue

In Chapter 5 we saw that the function which maps arrival process to departure process is continuous, and in Chapter 6 we used this result and the contraction principle to derive an LDP for the departure process.

Exactly the same procedure works here. The aggregate departure process from a queue with many flows satisfies an LDP. The problem is that the rate function is so complicated that it gives no insight at all. Even for the large-buffer limit, when the rate function for arrivals has a simple form, the rate function for departures is typically intractable; in the many-flows limit the rate function for arrivals is much more general, so what hope is there of simplifying the rate function for aggregate departures?

It is an open question whether the sample path rate function for departures even has the same form as that for arrivals. It is simple to see that it satisfies most points of Lemma 7.8, but it is not clear whether it satisfies the last two. So we doubt that the rate function for downstream queues can be simplified to the form in Theorem 7.7. If this is the case, it is akin to the negative result for the large buffer scaling, described in Section 6.7.

Instead, here is a totally different approach. We are considering queues fed by many input flows. What if, instead of asking about aggregate departures, we want to investigate how a *single flow* is affected by passing through the queue?

Let A be a typical input flow in a queue fed by N independent identically distributed flows, and served at rate NC, and let $D^{(N)}$ be the corresponding departure flow. (We will be more precise about how this is defined later.) We will investigate the characteristics of $D^{(N)}$ as $L \to \infty$.

Now, A on its own does not satisfy a large deviations principle, so we cannot hope that $D^{(N)}$ does either. Instead, A is described via a large deviations principle for A^N, the average of N independent copies of A. So it is natural to try to describe $D^{(N)}$ by seeking a large deviations principle for D^N, the average of N independent copies of $D^{(N)}$.

The surprising result of this section is that D^N satisfies the same large deviations principle as A^N. In other words, in this large deviations sense, the characteristics of a flow of traffic are *not changed* as it passes through a queue.

> *Note.* It's worth emphasizing that we are *not* attempting to find a large deviations principle for the aggregate departures. Think carefully about what it is that we are attempting to describe here with an LDP.

Let us now be precise about the setup. Suppose the queue has service rate NC and finite buffer NB. (It will turn out to be important for our proof technique that the buffer be finite.) Let it be fed by the aggregate of N independent copies of A, assumed to be regular over finite and infinite horizons. Let $D^{(N)}$ be a typical departure flow. It doesn't matter exactly what the service discipline is, as long as all work that arrived at time $t-1$ is served before any of the work that arrives at time t. Assume that $\dot{A}_t \geq 0$ almost surely (otherwise it is hard to interpret the departure process). Assume that the mean rate of A is strictly less than C.

We want to find an LDP for D^N, the average of N i.i.d. copies of $D^{(N)}$. Following Section 7.3, we will ask whether it is regular over finite and infinite horizons.

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

$$\mathrm{M}_t(\theta) = \lim_{N\to\infty} \frac{1}{N} \log E \exp(N\theta \cdot D^N|_{(-t,0]})$$

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.

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

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

is uniformly integrable, since

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

For any $-t < s \leq 0$, $P(D^{(N)}_{-s} \neq A_{-s})$ 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

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

Thus

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

and taking logarithms gives the result. □

At the core of the proof is the rather simple idea, that in this limiting regime the queue is very often empty, and so most of the time the work passes through unsmoothed.

It's interesting to know how large N needs to be for this to be accurate. Corollary 7.12 gives us an estimate for the probability that the queue is non-empty: if J is the rate function for that event, then for any $\varepsilon > 0$ there exists an N_0 such that for $N \geq N_0$, $P(\text{queue non-empty}) \leq e^{-N(J-\varepsilon)}$. Therefore

$$P\big(A(-t,0] \neq D^{(N)}(-t,0]\big) \leq (t+1)e^{-N(J-\varepsilon)}.$$

For fixed θ and t, the difference in log moment generating functions can be bounded similarly. So the error decays exponentially in NJ at least.

Infinite-Horizon Regularity of D^L

Proving regularity over the infinite-horizon is much harder: in fact, it is impossible. It turns out that this is not a problem. Infinite-horizon regularity was just a technical condition to control the tail behaviour of A, and there are other ways to achieve this. In fact, if we are interested in transient behaviour rather than steady state behaviour, it is not even necessary to worry about infinite-horizon regularity.

Specifically, with weaker conditions on the tail, we can prove a sample path LDP for D^N with is weaker, but still useful. It is weaker because it uses a weaker topology, the weak queue topology.

Definition 7.3 (Weak queue topology) *Let q be the queue-size function for a queue with service rate C and finite buffer B. Define the* weak queue topology *$wq(C,B)$ on $\mathcal{X}$ by the metric*

$$d(x,y) = |q(x) - q(y)| + \sum_{t=0}^{\infty} \frac{1 \vee |\dot{x}_{-t} - \dot{y}_{-t}|}{2^t}.$$

The first term in $d(x,y)$ measures the distance between $q(x)$ and $q(y)$; the second measures the distance between x and y in the topology of pointwise convergence. We know from Chapter 5 that q is continuous on $\mathcal{D}_\mu$ equipped with the scaled uniform norm topology $\|\cdot\|$, for $\mu < C$. Thus $\|\cdot\|$ is finer than wq which is finer than pointwise convergence; yet wq is still fine enough that q is continuous on $(\mathcal{D}_\mu, wq)$. So if D^N satisfies an LDP in $(\mathcal{D}, wq)$ and has mean rate $\mu < C$, we can still use the contraction principle to derive an LDP for $q(D^N)$.

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

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

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?

Chapter 8

Long Range Dependence

In the early nineties, a collection of papers ([58] 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, [57]).

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

8.1 What Is Long Range Dependence?

Let $(Y_n,\ n \in \mathbb{N})$ be a stationary sequence of random variables, which we assume to be bounded for simplicity, and set $S(n) = Y_1 + \cdots + Y_n$. If the Y_n are independent, then $\operatorname{Var} S(n) = n \operatorname{Var} Y_1$. In particular, the variance of $S(n)$ grows linearly with n. This property holds quite generally, for Markov chains and other weakly dependent sequences.

Long range dependence refers to when the variance grows non-linearly.

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 183–198, 2004.

The most common LRD models used for teletraffic have

$$\operatorname{Var} S(n) \approx \sigma^2 n^{2H} \quad \text{for large } n \tag{8.1}$$

where $H \in [\frac{1}{2}, 1)$ is the Hurst parameter. (In the case of unbounded variables, for which variances may be infinite, one must be more careful.) There is no standard definition of long range dependence; rather the term is loosely used to cover a range of related phenomena.

One related phenomenon is self-similarity. Let $\tilde{S}$ be the polgonalized version of S (see Section 5.2 for a definition of polygonalization) and define the speeded-up version $\tilde{S}^{\circlearrowright N}$ for $N \in \mathbb{R}^+$ by

$$\tilde{S}^{\circlearrowright N}(t) = \tilde{S}(Nt). \tag{8.2}$$

If the sequence of scaled processes

$$\frac{1}{N^H} S^{\circlearrowright N}$$

converges in distribution to some non-trivial limit as $N \to \infty$, then S is said to be asymptotically self-similar with Hurst parameter H. If the limit process has finite variance then S satisfies (8.1). Self-similar processes have fluctuations at every timescale, and the Hurst parameter relates the size of fluctuations to their timescale.

The most popular and well-known process which satisfies the above is fractional Brownian motion. This is a continuous-time process which has been widely adopted for its parsimonious structure—it depends on just three parameters, drift, variance parameter, and Hurst parameter. In this chapter we will focus on fractional Brownian motion, rather than working with general long range dependent processes.

A standard fractional Brownian motion $(Z(t),\ t \in \mathbb{R})$ with Hurst parameter H is characterized by the following properties:

- Z is Gaussian, i.e. its finite-dimensional distributions are multivariate normal;
- $Z(t)$ is normal with mean 0 and variance $|t|^{2H}$;
- Z has stationary increments, i.e. $Z(u+t) - Z(u) \sim Z(t)$;
- $Z(0) = 0$;
- Z has continuous sample paths.

If $H = \frac{1}{2}$ then this describes standard Brownian motion. A fractional Brownian motion with drift μ and variance parameter σ^2 can be written as $Z'(t) = \mu t + \sigma Z(t)$. Fractional Brownian motion satisfies (8.1) exactly, and is exactly self-similar, i.e.

$$\frac{1}{a^H} Z^{\circlearrowleft a} \sim Z \quad \text{for all } a > 0. \tag{8.3}$$

8.2 Implications for Queues

Consider a queue with service rate C, fed by a long-range dependent source X, and write $X(-t, 0]$ for the amount of work arriving in time interval $(-t, 0]$. In Section 3.1 we found a large deviations principle for queue size

$$\lim_{q\to\infty} \frac{1}{q} \log P(Q > q) = -\delta \quad \text{where} \quad \delta = \inf_{t>0} t\Lambda^*(C + 1/t) \tag{8.4}$$

where we assumed the existence of a sufficiently smooth limiting cumulant generating function

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{1}{t} \log Ee^{\theta X(-t,0]}.$$

If this limit exists then usually (by a Taylor expansion) $\operatorname{Var} X(-t, 0] \sim t\Lambda''(0)$. Thus, nothing we have presented so far applies to LRD models.

As we remarked at the end of Section 3.1, there is a variant of (8.4) which holds when the limit

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{1}{v_t} \log Ee^{\theta X(-t,0] v_t/t}$$

exists and is well-behaved, for sequences $v_t \to \infty$. If the variance grows non-linearly, we can expect (from the Taylor expansion again) that $\operatorname{Var} X(-t, 0] \sim \Lambda''(0)t^2/v_t$. This can be used to prove an analogue of Theorem 3.1. (See for example [33], but note that the proof given there contains a technical gap, which has since been addressed in [34].) The bottom line, in the case $\operatorname{Var} X(-t, 0] \sim \sigma^2 t^{2H}$, is that the queue size does not decay exponentially. Instead,

$$\lim_{q\to\infty} \frac{1}{q^{2(1-H)}} \log P(Q > q) = -\delta \quad \text{where} \quad \delta = \inf_{t>0} t^{2(1-H)} \Lambda^*(C + 1/t).$$

It is possible to state and prove this result using techniques similar to those presented in Chapter 3. We will not present the proof here, but instead focus on a specific traffic model: fractional Brownian motion.

Queues fed by fractional Brownian motion

Queues fed by fractional Brownian motion were first studied by Norros [75]. Let $(X(t),\ t \in \mathbb{R})$ be a fractional Brownian motion with drift μ, variance parameter σ^2, and Hurst parameter $H \in [\frac{1}{2}, 1)$. Let Q_t be the queue size at time t in an infinite-buffer queue with service rate $C > \mu$ fed by X:

$$Q_t = \sup_{s \le t} X(s,t] - C(t-s).$$

As usual, we are using our extended notation to describe traffic processes, in which we write $X(s,t]$ for $X(t) - X(s)$; see Section 5.5 for the rest of the extended notation. The process Q_t is stationary, and can be shown to be ergodic, with marginal distribution given by that of Q_0.

The following result shows that (for $H > \frac{1}{2}$) the queue length distribution does not have an exponential tail. When $H = \frac{1}{2}$ the problem reduces to the heavy traffic model described in Example 6.6, and the tail is exponential.

Theorem 8.1

$$\lim_{q \to \infty} \frac{1}{q^{2(1-H)}} \log P(Q_0 > q) = -\gamma^2/2$$

where

$$\gamma = \frac{(C-\mu)^H}{\sigma} \kappa \quad \textit{and} \quad \kappa = \frac{1}{H^H (1-H)^{1-H}}.$$

Proof. We follow the proof given by Massoulie and Simonian [73].

Lower bound. Using the fact that $X(-t,0] \sim \text{Normal}(\mu t, \sigma^2 t^{2H})$,

$$\begin{aligned} P(Q_0 > q) &= P\big(\sup_{t \ge 0} X(-t,0] - Ct > q\big) \\ &\ge \sup_{t \ge 0} P\big(X(-t,0] > q + Ct\big) \\ &= \sup_{t \ge 0} P\Big(N(0,1) > \frac{q + (C-\mu)t}{\sigma t^H}\Big) \end{aligned} \tag{8.5}$$

which is maximized at $t = H(1-H)^{-1}\, q/(C-\mu)$, giving

$$= P\Big(N(0,1) > q^{1-H}\gamma\Big).$$

Using the fact that

$$\frac{1}{x^2} \log P(N(0,1) > x) \to \frac{1}{2} \quad \text{as } x \to \infty$$

we obtain a lower bound on $\liminf P(Q_0 > q)$. (Note that the most likely time to overflow, the optimizing parameter in (8.5), is proportional to q.)

Upper bound. To prove the upper bound we will use Borell's inequality, which says the following. Let $(Y_t,\ t \in I)$ be a Gaussian process on any index set $I \subset [0,\infty)$. Suppose Y is centred, i.e. $EY_t = 0$ for all $t \in I$, and that the sample paths of Y are (almost surely) bounded. If $\rho^2 = \sup_{t\in I} \operatorname{Var} Y_t$ is finite, then $m = E\sup_{t\in I} Y_t$ is finite and

$$P\bigl(\sup_{t\in I} Y_t > x\bigr) \le 2e^{-(x-m)^2/2\rho^2} \quad \text{for all } x > m.$$

Now we can estimate the probability of a large queue size:

$$\begin{aligned} P(Q_0 > q) &= P\bigl(\sup_{t\ge 0} X(-t,0] - Ct > q\bigr) \\ &= P\Bigl(\sup_{t\ge 0} \frac{X(-t,0] - \mu t}{q + (C-\mu)t} > 1\Bigr) \end{aligned}$$

Why did we rewrite the probability in this way? Because now we can apply Borell's inequality: the variance of $X(-t,0] - Ct$ is unbounded, whereas

$$\rho^2 = \sup_{t\ge 0} \operatorname{Var} \frac{X(-t,0] - \mu t}{q + (C-\mu)t} = \sup_{t\ge 0} \frac{\sigma^2 t^{2H}}{(q + (C-\mu)t)^2} = q^{-2(1-H)}\gamma^{-2}.$$

Moreover, the process appearing here is centred, and since $C > \mu$ it almost surely has bounded sample paths. Thus

$$m_q = E \sup_{t\ge 0} \frac{X(-t,0] - \mu t}{q + (C-\mu)t}$$

is finite; and it tends to zero as $q \to \infty$ by monotone convergence. Applying Borell's inequality, for q sufficiently large,

$$P(Q_0 > q) \le 2e^{-(1-m_q)/2\rho^2} \le 2\exp\Bigl(-\frac{q^{2(1-H)}(1-m_q)\gamma^2}{2}\Bigr).$$

Taking logarithms and the lim sup as $q \to \infty$ gives the upper bound. □

8.3 Sample Path LDP for Fractional Brownian Motion

Theorem 8.1 tells us how the tail of the queue length distribution decays, but it doesn't tell us about how large queue lengths occur. What is the most likely path? Is it linear?

In Example 6.1 we used Schilder's theorem to analyse most likely paths of a standard Brownian motion. That theorem says that, if $B(t)$ is a standard Brownian motion and we set $B^N(t) = B(t)/\sqrt{N}$ then $(B^N|_{[0,1]},\ N \in \mathbb{R}^+)$ satisfies a large deviations principle in $\mathcal{C}^1$ with good rate function

$$I(x) = \begin{cases} \frac{1}{2}\int_0^1 \dot{x}_t^2\,dt & \text{if } x \in \mathcal{A}^1 \\ \infty & \text{otherwise.} \end{cases}.$$

Here $\mathcal{C}^1$ is the space of continuous functions $x : [0,1] \to \mathbb{R}$ with $x(0) = 0$, equipped with the topology of uniform convergence.

There is a similar sample path LDP for fractional Brownian motion. If Z is a standard fractional Brownian motion with Hurst parameter H, and we set $Z^N(t) = Z(t)/\sqrt{N}$ then the sequence $(Z^N,\ N \in \mathbb{R}^+)$ satisfies an LDP in $\mathcal{C}_0$ with good rate function

$$I(x) = \begin{cases} \frac{1}{2}\|x\|_R^2 & \text{if } z \in R \text{ (defined below)} \\ \infty & \text{otherwise.} \end{cases}$$

Here $\mathcal{C}_0$ is the the natural extension of the space we defined in Section 5.4 to functions $x : \mathbb{R} \to \mathbb{R}$, namely, the space of functions for which $x(0) = 0$ and

$$\lim_{t\to\infty} \frac{x(t)}{1+|t|} = \lim_{t\to-\infty} \frac{x(t)}{1+|t|} = 0$$

equipped with the norm

$$\|x\| = \sup_{t\in\mathbb{R}} \frac{x(t)}{1+|t|}.$$

Note. In fact, this result applies to a broad class of Gaussian processes; see Addie et al. [1]. It is an extension of a result known as the generalized form of Schilder's theorem; see Deuschel and Stroock [28, Theorem 3.4.5].

We need to define R, the *reproducing Hilbert space* for Z, and the norm $\|\cdot\|_R$. First, let

$$\Gamma(s,t) = \operatorname{Cov}\bigl(Z(s), Z(t)\bigr) = \frac{1}{2}\bigl(V_s + V_t - V_{|t-s|}\bigr)$$

where $V_t = \operatorname{Var} Z(t)$. Write $\Gamma_s(\cdot)$ for $\Gamma(s,\cdot)$. Consider the set of functions $\{\Gamma_s : s \in \mathbb{R}\}$, equipped with the inner product

$$\langle \Gamma_s, \Gamma_t \rangle = \Gamma(s,t).$$

The space R is obtained by closing this space of functions with respect to linear combinations, and completing with respect to the norm

$$\|x\|_R^2 = \langle x, x \rangle.$$

The sample paths in R, being composed of smooth functions Γ_s, are typically smoother than the sample paths of Z.

The self-similarity property (8.3) of Z lets us write down an LDP which is more useful for queueing applications. Recalling the definition of speed-up, and letting $N = L^{2(1-H)}$, the self-similarity property says that

$$\frac{1}{L} Z^{\circlearrowleft L} \sim \frac{1}{\sqrt{N}} Z$$

and hence $L^{-1} Z^{\circlearrowleft L}$ satisfies an LDP of the form

$$\frac{1}{L^{2(1-H)}} \log P(L^{-1} Z^{\circlearrowleft L} \in B) \approx -\inf_{z \in B} I(z).$$

The sense of the approximation is that the appropriate large deviations lower and upper bounds apply for open and closed sets B. The denominator $L^{2(1-H)}$ is called the *speed* of this large deviations principle. If X is a fractional Brownian motion with drift μ, then $L^{-1} X^{\circlearrowleft L}$ also satisfies an LDP at this speed, in the space $\mathcal{C}_\mu$.

All our results based on the contraction principle carry through. We can find LDPs for queue size in queues with finite and infinite buffers, priority queues, etc. etc. This approach to fractional Brownian motion has also been taken by Majewski [64].

The answer to our initial question—do queues build up linearly?—is no. The sample path LDP can be used to find the most likely path, which is the approach taken by Addie et al. [1]. Another approach is to use a representation of fractional Brownian motion as a stochastic integral against a standard Brownian motion, then use Schilder's theorem and the contraction principle [79]. Yet another approach is taken by Chang et al. [14]. Also see Example 7.7. In fact, as was pointed out to us by Peter Glynn, a direct calculation is possible in this case:

Example 8.1
This calculation relies on the following observation. If (X, Y) has a bivariate normal distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \gamma \\ \gamma & \rho^2 \end{pmatrix} \right]$$

then conditional on $Y = y$, X has the distribution

$$X \mid (Y = y) \;\sim\; \text{Normal}\Big(\mu + \frac{\gamma(y-\nu)}{\rho^2}, \sigma^2 - \frac{\gamma^2}{\rho^2}\Big).$$

Furthermore, if $(X_t,\ t \in [0,T])$ is a Gaussian process then so is $(X_t,\ t \in [0,T]) \mid X_s = \xi$.

Let A be a fractional Brownian motion with drift μ, variance parameter σ^2 and Hurst parameter H. Define the scaled version

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

As we have remarked, this satisfies a sample path LDP with speed $N^{2(1-H)}$. Conditional on $A^N(-T,0] = \xi$, what is the distribution of A?

It is easy to check that

$$E\big(A^N(-t,0] \;\big|\; A^N(-T,0] = \xi\big) = \mu t + \frac{\Gamma(t,T)}{V_T}(\xi - \mu T).$$

Call this quantity $a(t)$. It is also easy to check that

$$\text{Var}\big(A^N(-t,0] \;\big|\; A^N(-T,0] = \xi\big) = \frac{\sigma^2}{N^{2(1-H)}}\Big(t^{2H} - \frac{\Gamma(t,T)^2}{T^{2H}}\Big).$$

With a little work we can apply Borell's inequality to deduce that

$$\lim_{N\to\infty} P\Big(\sup_{t\in[0,T]} |A^N(-t,0] - a(t)| > \varepsilon \;\Big|\; A^N(-T,0] = \xi\Big) = 0.$$

◇

8.4 Scaling Properties

A great deal can be said about scaling properties of networks, without having to solve variational problems arising from the contraction principle. We discussed this approach in Section 6.10, in the context of the standard large-buffer scaling. What about long range dependence?

A good word to use here is *Hurstiness*, a term introduced to us by John Lewis. We will use it to mean the following. Given a process X, define the scaled version

$$X^N(-t,0] = \frac{1}{N} X(-Nt,0].$$

Say that X has Hurstiness H if the sequence $(X^N,\ N \in \mathbb{N})$ satisfies a sample path LDP in $\mathcal{C}_\mu$ (for some μ, called the mean rate) of the form

$$\frac{1}{N^{2(1-H)}} \log P(X^N \in B) \approx -\inf_{x \in B} I(x), \tag{8.6}$$

where I is a good rate function, and where the only path x for which $I(x) = 0$ is the path with constant rate μ, and where there is some path x for which $0 < I(x) < \infty$. When we write the approximation, we mean that the standard large deviations lower and upper bounds hold for open and closed sets, as $N \to \infty$ in $\mathbb{R}^+$. A fractional Brownian motion with Hurst parameter H and drift μ has Hurstiness H and mean rate μ. We will see that

- if a flow has a Hurstiness, it has a unique Hurstiness;
- the Hurstiness of the arrival process determines the shape of the tail of the queue length distribution;
- the Hurstiness of an aggregate is equal to the maximum Hurstiness of the constituent parts;
- the Hurstiness of the departure process from a queue is equal to the Hurstiness of the arrival process.

The second of these relies on a key scaling property of the rate function I, implied by (8.6), made precise in the following lemma. (Recall from (8.2) the meaning of $x^{\circlearrowright 1/\kappa}$.)

Lemma 8.2 *If X^N satisfies an LDP of the form* (8.6), *then the rate function I must satisfy*

$$I(\kappa x^{\circlearrowright 1/\kappa}) = \kappa^{2(1-H)} I(x). \tag{8.7}$$

Proof. Define $Y^N = X^{\kappa N}$, that is,

$$Y^N = f(X^N) \quad \text{where} \quad f(x) = \frac{1}{\kappa} x^{\circlearrowright \kappa}.$$

Each of these two representations will lead us to an LDP for Y^N; we will then use the uniqueness of the rate function (Lemma 4.7) to obtain the desired equality (8.7).

First, we will use the contraction principle. It is easy to verify that $\|f(x)\| \leq (\kappa^{-1} \vee 1)\|x\|$, and to deduce that f is continuous. Hence, by the contraction principle, Y^N satisfies an LDP of the form (8.6) with good rate function

$$J(y) = \inf_{x : f(x) = y} I(x) = I(\kappa y^{\circlearrowright 1/\kappa}).$$

We can also obtain an LDP directly by using the scaling in (8.6):

$$\begin{aligned}
&\frac{1}{N^{2(1-H)}} \log P(Y^N \in B) \\
&= \frac{1}{N^{2(1-H)}} \log P(X^{\kappa N} \in B) \\
&= \kappa^{2(1-H)} \frac{1}{(N\kappa)^{2(1-H)}} \log P(X^{\kappa N} \in B) \\
&\approx -\kappa^{2(1-H)} \inf_{x \in B} I(x) = -\inf_{x \in B} \kappa^{2(1-H)} I(x).
\end{aligned}$$

where we write $\approx$ as before to denote the lower and upper LD bounds for open and closed sets.

Comparing these two expressions for the rate function completes the proof. □

Uniqueness. Suppose X has Hurstiness H, so that it satisfies an LDP of the form

$$\frac{1}{N^{2(1-H)}} \log P(X^N \in B) \approx -\inf_{x \in B} I(x).$$

Can it have Hurstiness $G > H$? We will show that for any $G > H$, X satisfies an LDP at speed G but with a trivial rate function. By uniqueness of the rate function (Lemma 4.7), it cannot satisfy an LDP at speed G with a non-trivial rate function. Therefore it cannot have Hurstiness G. This establishes the uniqueness of Hurstiness.

We will argue that the LDP at speed G is of the form

$$\frac{1}{N^{2(1-G)}} \log P(X^N \in B) \approx -\inf_{x \in B} I'(x) \tag{8.8}$$

where

$$I'(x) = \begin{cases} 0 & \text{if } I(x) = 0 \\ \infty & \text{otherwise.} \end{cases}$$

Note that I' is trivially a good rate function. For the large deviations upper bound, suppose that $\inf_{x \in B} I(x) > 0$ for some closed set, and consider

$$\frac{1}{N^{2(1-G)}} \log P(X^N \in B) = N^{2(G-H)} \frac{1}{N^{2(1-H)}} \log P(X^N \in B).$$

The $\limsup$ of the second term is < 0, and the first term $\to \infty$, so

$$\limsup_{N \to \infty} \frac{1}{N^{2(1-G)}} \log P(X^N \in B) = -\infty.$$

If $\inf_{x\in B} I(x) = 0$, the bound is trivial. The lower bound for open sets B is trivial unless $\inf_{x\in B} I(x) = 0$, in which case it can be obtained by considering the complement of B and using the upper bound.

Queues. Consider an infinite-buffer queue fed by a source with Hurstiness H. As usual, we look at the queue size $Q = q(X)$ where

$$q(x) = \sup_{t\in\mathbb{R}^+} x(-t,0] - Ct$$

and C is the service rate. If the service rate is larger than the mean arrival rate then q is continuous, and so the contraction principle applies, giving an LDP for $Q/N = q(X^N)$ of the form

$$\frac{1}{N^{2(1-H)}} \log P(Q/N \in B) \approx -\inf_{q\in B} J(q)$$

where J is the good rate function

$$J(q) = \inf\bigl\{I(x) : q(x) = q\bigr\}.$$

It must be that $J(q) > 0$ for $q > 0$. For suppose $J(q) = 0$. By goodness of the rate function I, the optimum is attained, say at $\hat{x}$; by the definition of Hurstiness, $\hat{x}$ has constant rate equal to the mean arrival rate; yet this cannot satisfy $q(\hat{x}) = q > 0$.

In fact, we can say a great deal more about the form of J, using only the form of the rate function (8.6). Suppose that that $J(q_0) < \infty$ for some q_0, i.e. that overflow is 'plausible'. Let $\hat{x}$ be optimal in $J(q_0)$, so that $q(\hat{x}) = q_0$ and $J(q_0) = I(\hat{x})$. Now let us seek to evaluate $J(q)$. Consider the path y defined by

$$y = \frac{q}{q_0}\hat{x}^{\circlearrowright q_0/q}.$$

This path satisfies $q(y) = q(\hat{x})q/q_0 = q$, and so $J(q) \le I(y)$. Furthermore, using Lemma 8.2,

$$J(q) \le I(y) = \Bigl(\frac{q}{q_0}\Bigr)^{2(1-H)} I(\hat{x}) = \Bigl(\frac{q}{q_0}\Bigr)^{2(1-H)} J(q_0).$$

We can obtain a reverse inequality similarly. The conclusion is that

$$J(q) = q^{2(1-H)}\delta \quad \text{where} \quad \delta = \frac{J(q_0)}{q_0^{2(1-H)}} \quad \text{and} \quad 0 < \delta < \infty.$$

Since the rate function J is continuous, the LD lower and upper bounds for the event $\{Q/N > q\}$ agree, and so

$$\lim_{N\to\infty} \frac{1}{N^{2(1-H)}} \log P(Q/N > q) = -\delta q^{2(1-H)}.$$

Let $q = 1$ and relabel N as q to see that

$$\lim_{N\to\infty} \frac{1}{q^{2(1-H)}} \log P(Q > q) = -\delta q^{2(1-H)}.$$

Thus (under the condition that overflow is plausible) the Hurstiness of the arrival process determines the shape of the tail of the queue length distribution.

Aggregates. Consider two independent flows X and Y, one with Hurstiness H and rate function I, the other with Hurstiness G, and rate function J, and suppose that $H < G$. What is the Hurstiness of the aggregate $Z = X + Y$?

As we remarked above, X satisfies a trivial LDP of the form (8.8). Now, consider the aggregate $Z = X+Y$. By the contraction principle, Z^N satisfies a large deviations principle of the form

$$\frac{1}{N^{2(1-G)}} \log P(Z^N \in B) \approx -\inf_{z\in B} K(z)$$

where K is the good rate function

$$K(z) = \inf_{(x,y)\,:\,x+y=z} I'(x) + J(y).$$

Since I' is trivial,

$$K(z) = J(z - \mu)$$

where by μ we mean the path with constant rate equal to the mean rate of X. Since J is assumed to be non-trivial (from the definition of Hurstiness), K is non-trivial. Thus the Hurstiness of the aggregate is equal to the greater of the Hurstinesses of the two component flows.

Departures. Consider an infinite-buffer queue fed by a source with Hurstiness H and mean rate μ, and service rate greater than the mean rate. The departure process D is a continuous function of the arrival process, so by the contraction principle D^N satisfies an LDP with speed $N^{2(1-H)}$.

If the arrival process has constant rate μ, then the departure process does too; and so this path has rate function 0. It is not hard to see that this is the only path with rate function 0. Also, an arrival process with a finite value of the rate function yields some departure process with a finite value of the rate function; thus the rate function for the departure process is non-trivial. So the departure process has Hurstiness H.

Networks. It seems likely that similar results apply to networks of queues. Indeed, the argument given above for queues applies to queues anywhere inside a network. However, it can be difficult to verify the two properties we needed, namely

- stability, i.e. $q(\mu) = 0$;
- non-degeneracy, i.e. there exists a path x with $I(x) < \infty$ and $q(x) > 0$.

One simple setting where these hold is the tandem queue. Suppose we have two queues in tandem, and we are interested in the tail of the queue size distribution for the downstream queue. The stability condition is satisfied if the two service rates are greater than the mean arrival rate; and the non-degeneracy condition is satisfied assuming that the upstream service rate is greater than the downstream service rate and that it is plausible for the upstream queue to overflow.

8.5 How Does Long Range Dependence Arise?

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, i.e. LRD. In statistical physics, magnets are modelled as a Markov random field, a higher-dimensional analogue of a Markov chain; and it can be shown that, if the local interaction is strong enough, the system will exhibit long range dependence. (This does not occur in one dimension, as suggested by Exercise 3.2.)

What is the cause of LRD in teletraffic? A possible cause is the fact that file sizes typically have heavy-tailed distributions. A random variable T is said to be heavy-tailed if

$$\lim_{x \to \infty} P\big(T > t + x \bigm| T > t\big) = 1 \quad \text{for all } x > 0.$$

There are many possible traffic models based on heavy-tailed file sizes; the general conclusion is that the aggregate of many independent flows, each with heavy-tailed renewal times, converges to a fractional Brownian motion (after rescaling appropriately). See for example Willinger et al. [99].

Another construction is the following, due to Kaj [51]. Suppose there are N sources, each an independent copy of a stationary renewal process with inter-renewal time distributed like T. Let $H \in (\frac{1}{2}, 1)$ and suppose that

$$P(T > t) \sim t^{-(3-2H)} L(t)$$

where L is slowly varying at infinity, i.e.

$$\lim_{t\to\infty} \frac{L(xt)}{t} = 1 \quad \text{for all } x > 0.$$

Note that T has a finite mean but no variance. Let $A^N(-t, 0]$ be the total number of arrivals (i.e. renewal events) in time $(-t, 0]$. One can choose a scaling sequence v_N such that $v_N^{2(1-H)} \sim N\,L(v_N)$; for such a sequence, define the scaled arrival process

$$\tilde{A}^N(-t, 0] = \frac{A^N(-v_N t, 0]}{v_N} - \frac{Nt}{ET}.$$

This process converges in distribution to a fractional Brownian motion with Hurst parameter H, drift μ, and variance parameter

$$\sigma^2 = \frac{1}{H(2H-1)(ET)^3}.$$

In the context of large deviations, here is a result due to Mandjes [67]. It concerns a queue fed by the aggregate of N independent sources. Let each source be $M/G/\infty$, that is, jobs arrive as a Poisson process of rate λ, and stay active for a holding time which has distribution T, and holding times are independent. Suppose the job generates work at unit rate while active. Take T to be Pareto, that is,

$$P(T > t) = (t+1)^{-(3-2H)}$$

for some $H \in (0, 1)$. Now let A be a Gaussian approximation to this source. A has mean arrival rate λET, and variance

$$\operatorname{Var} A(-t, 0] = \frac{\lambda}{H(2H-1)(2H-2)} \Big(1 - (t+1)^{2H} + \frac{2H}{t}\Big).$$

Consider a queue fed by N independent copies of A, with service rate NC where C is greater than the mean arrival rate of a source. Let Q^N be the queue size. Then Q^N satisfies a large deviations principle:

$$\lim_{N\to\infty} \frac{1}{N} \log P(Q^N > Nb) = -I(b).$$

The rate function $I(b)$ has two different forms, depending on the value of H. If $H < \frac{1}{2}$ then $I(b) \sim \kappa b$ for some constant κ; if $H > \frac{1}{2}$ then $I(b) = O(b^{2(1-H)})$. This result links the fractional Brownian motion limit above, with the queue size result for fractional Brownian motion in Section 8.2.

8.6 Philosophical Difficulties with LRD Modelling

First let us suppose that we have observed a high *empirical* value for the Hurst parameter 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 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. (Ironically, short term prediction is often considerably easier with such data sets because of the presence of 'trends'.)

In general, it is impossible to distinguish between long range dependence 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 [57] 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 points of view. There is a sense in which to take the LRD view and regard the single, unexplained fluctuation as random, is to be effectively Bayesian; and the alternative viewpoint is frequentist.

Suppose that we do have further information about the data—suppose 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 interarrival times, as discussed in the previous section. Then things are somewhat different. Modelling and prediction will be difficult, but no more difficult than 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 practice 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 practice where it preferable to try something, even if confidence is limited, rather than throw our arms in the air and say ‘this is impossible!’

Chapter 9

Moderate Deviations Scalings

In our discussion of long range dependence, we introduced a different style of working with large deviations problems: by looking at the *speed* of the LDP rather than calculating the exact value of a rate function, we were able to draw robust conclusions without too much work. In this chapter we will look at another class of LDP where such arguments are useful, the class of moderate deviations principles. This is a large deviations analogue of heavy traffic theory.

We will tackle moderate deviations problems using the contraction principle. Moderate deviations have also been studied using the tools of heavy traffic theory, by Puhalskii [85] and Majewski [62].

9.1 Motivation

Moderate deviations concerns a collection of scales between large deviations and the central limit theorem. To be concrete, let $(Y_n, n \in \mathbb{N})$ be a collection of i.i.d. random real-valued random variables all distributed like Y and let

$$S_n = \sum_{i=1}^{n} Y_i.$$

We have worked extensively with the large deviations limit

$$\frac{1}{n} \log P\Big(\frac{S_n}{n} \in B\Big) \approx - \inf_{x \in B} I(x)$$

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 199–209, 2004.

(where by $\approx$ we mean that the large deviations bounds hold: an upper bound on the lim sup for closed sets, and a lower bound on the lim inf for open sets). The central limit theorem says that, if $\mu = EY$ and $\sigma^2 = \operatorname{Var} Y$,

$$P\Big(n^{1/2}\Big(\frac{S_n}{n} - \mu\Big) \in B\Big) \approx P\big(\mathrm{Normal}(0,\sigma^2) \in B\big).$$

(where by $\approx$ we mean there is convergence in distribution of the random variables). Moderate deviations concerns a range of scales in between these. Specifically, for $\beta \in (0,1)$,

$$\frac{1}{n^\beta}\log P\Big(n^{(1-\beta)/2}\Big(\frac{S_n}{n} - \mu\Big) \in B\Big) \approx -\inf_{x\in B} \tfrac{1}{2}x^2/\sigma^2 \tag{9.1}$$

(where by $\approx$ we mean that the large deviations upper and lower bounds hold, and we are assuming $\sigma^2 > 0$.) If this is so, we say that $n^{-1}S_n$ satisfies a moderate deviations principle with mean μ at scale β. A moderate deviations principle is just a large deviations principle with a particular scaling.

Theorem 9.1 *If the log moment generating function of Y,*

$$\Lambda(\theta) = \log Ee^{\theta Y},$$

is finite for θ in a neighbourhood of the origin then (9.1) *holds.*

Proof. If $\Lambda(\theta)$ is finite in a neighbourhood of the origin then it is infinitely differentiable at the origin, and has a power series expansion

$$\Lambda(\theta) = \theta\mu + \tfrac{1}{2}\sigma^2\theta^2 + O(\theta^3).$$

Now we will simply use the generalized Cramér's theorem. Let

$$T_n = n^{(1-\beta)/2}(n^{-1}S_n - \mu).$$

This has log moment generating function

$$\mathrm{M}_n(\theta) = n\Lambda(\theta n^{-(1+\beta)/2}) - \theta n^{(1-\beta)/2}\mu,$$

and so

$$\frac{1}{n^\beta}\mathrm{M}_n(\theta n^\beta) = \frac{\Lambda(\theta\delta) - \theta\delta\mu}{\delta^2}$$

where $\delta = n^{-(1-\beta)/2}$. Note that

$$n^{-\beta}\mathrm{M}_n(\theta n^\beta) \to \tfrac{1}{2}\sigma^2\theta^2 \quad \text{as } n \to \infty$$

by standard results for power series. So we can apply the generalized Cramér's theorem, Theorem 2.11, to deduce that T^n satisfies a large deviations principle of the form

$$\frac{1}{n^\beta}\log P(T_n \in B) \approx -\inf_{x\in B} I(x)$$

where $\approx$ is in the usual sense of meaning upper and lower bounds, and the rate function is

$$I(x) = \sup_\theta \theta x - \tfrac{1}{2}\sigma^2\theta^2 = \tfrac{1}{2}x^2/\sigma^2.$$

(The statement of Cramér's theorem in Chapter 2 only used probability scalings of the form $n^{-1}\log P(\cdot)$, but all the results apply equally to this scaling.) □

Some remarks.

i. The moderate deviations estimate is like a mixture between the large deviations limit and the central limit. If β is close to 0 it deals with variations which are close to those studied by the central limit theorem; if β is close to 1 it deals with variations which are close to those studied by large deviations theory.

ii. As with large deviations, the moderate deviations estimate is governed by the principle of the largest term: only the most likely $\hat{x} \in B$ contributes to the estimate.

iii. As with the central limit theorem, the moderate deviations estimate depends only on the variance σ^2. Higher moments are not involved.

iv. Let Y' be a normal random variable with mean μ and variance σ^2. Let S'_n be the sum of n i.i.d. copies of Y'. Then the large deviations estimate for S'_n is

$$\lim_{n\to\infty}\frac{1}{n}\log P\Big(\frac{S'_n}{n} - \mu \in B\Big) = -\inf_{x\in B}\tfrac{1}{2}x^2/\sigma^2.$$

The rate function is just the same as for the moderate deviations estimate. Crudely, moderate deviations is like large deviations but ignoring moments that are higher-order than the mean and variance.

v. We assumed $\Lambda(\theta)$ was finite in a neighbourhood of the origin. In fact, it is not even necessary for all moments to be finite (though the larger β is, the more control we need on the moments). See Deo and Babu [27] for tighter conditions.

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

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

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

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:

i. If X^N is the average of N independent copies of some traffic process X, then item (ii) just says that the rate function I is what it would be for a Gaussian approximation to X.

ii. If X^N is a time-rescaled version of a process X, then the covariance function is typically trivial, $\gamma_t = 0$ for $t > 0$.

iii. Item (iii) is a condition on the long-timescale regularity of X^N. It is used to strengthen the topology on $\mathcal{D}$ from that of pointwise convergence to that of convergence in the scaled uniform norm topology. See Example 7.1 for how this works for Gaussian processes in the many-flows limit.

iv. Item (iv) is in fact a consequence of (iii), and we only include it in the definition for convenience.

It can be seen that the theory is all very similar to Chapter 6 and especially to Chapter 7. The only distinctive feature is in the scaling factor N^β. What does it mean for queueing systems?

9.3 Queue Scalings

We can straightforwardly apply the extended contraction principle. If the sequence of arrival processes A^N satisfies a sample path moderate deviations principle, then, with some abuse of language, we can say that the queue size $q(A^N)$ satisfies a moderate deviations principle. The form is this:

$$\frac{1}{N^\beta} \log P\Big[q\Big(N^{(1-\beta)/2}(A^N - \mu e)\Big) \in B\Big] \approx - \inf_{a \in \mathcal{D}: q(a) \in B} I(a).$$

The rate function can be significantly simplified. From Section 7.6, the form of $J(q) = \inf\{I(a) : a \in \mathcal{D}, q(a) = q\}$ is

$$J(q) = \inf_{t \geq 0} \sup_{\theta \in \mathbb{R}} \theta(q + Ct) - \tfrac{1}{2}\theta^2 V_t = \inf_{t \geq 0} \frac{(q + Ct)^2}{2V_t}. \tag{9.3}$$

All the other applications of the contraction principle in Chapter 7 carry through too. (Though results must be interpreted with care, bearing in mind the scaling involved in the moderate deviations principle.)

To try to understand what these results mean, suppose for instance that A^N is the average of N independent copies of some arrival process A. Write $q(A^N, C)$ for the queue size function with service rate C, because the scaling of service rate will turn out to be important. Then

$$\begin{aligned}
&q\big(N^{(1-\beta)/2}(A^N-\mu e),C\big)\\
&\qquad=N^{-(1+\beta)/2}q\big(NA^N-N\mu e,N^{(1+\beta)/2}C\big)\\
&\qquad=N^{-(1+\beta)/2}q\big(NA^N,N\mu+N^{(1+\beta)/2}C\big).
\end{aligned}$$

So this moderate deviations result is describing a sequence of queues, in which the Nth queue serves N flows, and the excess service rate and queue size scale as $N^{(1+\beta)/2}$. The traffic intensity at the Nth queue is

$$\rho^N=\frac{N\mu}{N\mu+N^{(1+\beta)/2}C}$$

so $\rho^N\to 1$ and

$$1-\rho^N\sim N^{-(1-\beta)/2}C/\mu.$$

It may be helpful to give a more intuitive account of the parameter β. Recall the statement of the moderate deviations principle, (9.2), which we will informally rewrite as

$$P(A^N\approx\mu+N^{-(1-\beta)/2}x)\approx\exp(-N^\beta I(x)).$$

This expresses a relationship between the *size* of a deviation, of scale $N^{-(1-\beta)/2}$ relative to the mean, and its *frequency* $\exp(-N^\beta)$. A random arrival process A^N will typically satisfy moderate deviations principles at *all* scales of size and frequency, that is, for all $\beta\in(0,1)$. When the arrival process NA^N is fed into a queue with traffic intensity of scale $\rho\approx 1-N^{-(1-\gamma)/2}$, there are deviations in the queue size of scale $N^{(1+\gamma)/2}$, caused by deviations in the traffic of the same scale, and these deviations have frequency $\exp(-N^\gamma)$. (There will of course be deviations at many scales, but large deviations tools only tell us non-trivial things about this particular scale.)

Interpretation

The formal procedure we have described here—applying the contraction principle and solving a variational problem to find the rate function—is exactly the same for moderate deviations as for large deviations (be it large-buffer, many-flows or long-range dependence). However, it will become apparent in the next two sections that the results must be interpreted carefully, because of the distinctive 'moderately heavy traffic' scaling. In particular, quantities like the departure process have a very different interpretation.

If one studies large deviations in queues fed by Gaussian traffic processes, intending this to be a heavy traffic approximation of some non-Gaussian sytem (rather than just a convenient and tractable mathematical object), it is important to consider the scaling implications of the heavy traffic limit, and to interpret any large deviations results in the light of those implications. Crudely speaking, one can apply the large deviations techniques in Chapters 6 and 7 to heavy traffic approximations in order to study total queue size and most likely paths to overflow, but not (unless one has good reasons for doing so) to queueing delay, to shared buffers, or to departures. The right way to interpret such quantities is given below.

9.4 Shared Buffers

Consider a single queue fed by two input flows A^N and B^N, and assume that each, normalized, satisfies a sample path moderate deviations principle. Suppose the work is served with a first-come–first-served discipline. How much of the work in the queue comes from each of the two flows?

To be precise, suppose that work $N\dot{A}^N_{-t}$ and $N\dot{B}^N_{-t}$ arrives uniformly spread throughout the interval $(-t-1,-t]$. Let the queue have service rate $N\mu + N^{(1+\beta)/2}C$, where μ is the mean rate of $A^N + B^N$. Let the total queue size be $N^{(1+\beta)/2}Q^N_{-t}$. Let the amount of work due to A^N and B^N be $N^{(1+\beta)/2}R^N_{-t}$ and $N^{(1+\beta)/2}S^N_{-t}$, defined with the usual boundary condition, that 'the queue was empty at time $-\infty$'. We know that Q^N_{-t} satisfies a moderate deviations principle. We seek a moderate deviations principle for (R^N_{-t}, S^N_{-t}).

Theorem 9.2 *Let ν be the mean rate of A^N, and let*

$$\tilde{R}^N_{-t} = \frac{\nu}{\mu} Q^N_{-t}.$$

Then R^N_{-t} is exponentially equivalent to $\tilde{R}^N_{-t}$ at scale β, in that

$$\limsup_{N\to\infty} \frac{1}{N^\beta} \log P\big(|R^N_{-t} - \tilde{R}^N_{-t}| > \delta\big) = -\infty \quad \textit{for all } \delta > 0.$$

Thus R^N_{-t} satisfies a moderate deviations principle with good rate function $J(q\mu/\nu)$, where $J(q)$ is the rate function for Q^N_{-t}. In Chapter 4 we called this the approximate contraction principle.

An immediate consequence (easily proved from the definition of exponential equivalence) is that (R^N_{-t}, S^N_{-t}) is exponentially equivalent to $(\nu/\mu, 1 - \nu/\mu)Q^N_{-t}$.

The idea of the proof is to unwrap the scaling that defines R^N_{-t}. At time $-t-1$ there is an (unscaled) amount of work $N^{(1+\beta)/2}Q^N_{-t-1}$ in the system. Over time $(-t-1,-t]$, some extra work arrives: that extra work is $N\mu + N^{(1+\beta)/2}\dot{X}^N_{-t}$, where X^N is of the same scale as Q^N_{-t}. Compare the scales N and $N^{(1+\beta)/2}$ and observe that the vast majority of work in the queue is due to the $N\mu$ term. So the ratio of work due to the two flows is dominated by their means, and so (after serving an amount $N\mu + N^{(1+\beta)/2}C$ of work) the queue at time $-t$ is largely made up of the two flows in proportion to their mean rates.

Proof of Theorem 9.2 Consider how R^N_{-t} comes about. At time $-t-1$ there was a certain amount of work $N^{(1+\beta)/2}Q^N_{-t-1}$ in the queue, with work from the two flows distributed somehow. Then $N\dot{A}^N_{-t} + N\dot{B}^N_{-t}$ arrives, and work from the two flows is distributed evenly. Of the total work, $N\mu + N^{(1+\beta)/2}$ is served, the original work $N^{(1+\beta)/2}Q^N_{-t-1}$ coming first.

So either $N^{(1+\beta)/2}Q^N_{-t-1}$ is served completely, in which case

$$R^N_{-t} = Q^N_{-t}\frac{\dot{A}^N_{-t}}{\dot{A}^N_{-t}+\dot{B}^N_{-t}},$$

or it is not, which requires

$$N^{(1+\beta)/2}Q^N_{-t-1} > N\mu + N^{(1+\beta)/2}C.$$

Thus

$$\begin{aligned} &P\big(|R^N_{-t} - \tilde{R}^N_{-t}| > \delta\big) \\ &\quad \leq P\Big(Q^N_{-t-1} > N^{(1-\beta)/2}\mu + C\Big) + P\Big(\Big|\frac{\dot{A}^N_{-t}}{\dot{A}^N_{-t}+\dot{B}^N_{-t}} - \frac{\nu}{\mu}\Big|Q^N_{-t} > \delta\Big). \end{aligned} \tag{9.4}$$

By the principle of the largest term, it is sufficient to show that

$$\limsup_{N\to\infty} N^{-\beta}\log P(\cdot) = -\infty$$

for each of these parts.

Consider the first term in (9.4). We know that Q^N_{-t-1} satisfies a moderate deviations principle, say with rate function $J(q)$ as in (9.3). Thus

$$\limsup_{N\to\infty}\frac{1}{N^\beta}\log P\big(Q^N_{-t-1} > N^{(1-\beta)/2}\mu + C\big) \leq -J(q) \tag{9.5}$$

for every $q > 0$. (This assumes $\mu > 0$.) But $J(q)$ is unbounded as $q \to \infty$. (This is from the assumption that $V_t = o(t^2/\log t)$, where V_t is the marginal variance of $A^N + B^N$.) So the lim sup is equal to $-\infty$.

Now for the second term in (9.4). Let $\delta_1 = |\dot{A}^N_{-t} - \nu|$ and $\delta_2 = |\dot{A}^N_{-t} + \dot{B}^N_{-t} - \mu|$. If $\delta_2 < \mu$ then

$$\Big|\frac{\dot{A}^N_{-t}}{\dot{A}^N_{-t} + \dot{B}^N_{-t}} - \frac{\nu}{\mu}\Big| \leq \frac{\mu\delta_1 + \nu\delta_2}{\mu(\mu - \delta_2)}.$$

This lets us break up the second term into separate parts:

$$P\Big(\Big|\frac{\dot{A}^N_{-t}}{\dot{A}^N_{-t} + \dot{B}^N_{-t}} - \frac{\nu}{\mu}\Big| Q^N_{-t} > \delta\Big) \\ \leq P(\delta_1 Q^N_{-t} > \delta\mu) + P(\delta_2 Q^N_{-t} > \delta\mu/\nu) + P(\delta_2 > \mu/2),$$

for each of which, as with (9.5), $\limsup_N N^{-\beta} \log P(\cdot) = -\infty$. □

This leaves us with the following picture of the evolution of the queue. At the level of fluctuations described by moderate deviations, a total amount of work $\dot{a}_{-t} + \dot{b}_{-t}$ arrives at the queue at time $-t$. The queue size fluctuates according to the standard Lindley recursion, $q_{-t} = (q_{-t-1} + \dot{a}_{-t} + \dot{b}_{-t} - c)^+$. All of q_{-t-1} is served at time $-t$, and the amount of work left over in the queue from a and b is $q_{-t}\nu/\mu$ and $q_{-t}(1 - \nu/\mu)$.

Observe that the distinctive moderate deviations scaling has led us to this result, which is completely unlike results for either large-buffer or many-flows scaling.

Exercise 9.1
Let D^N be the departure process for traffic coming from A^N:

$$ND^N(-t, 0] = NA^N(-t, 0] + N^{(1+\beta)/2} R^N_{-t} - N^{(1+\beta)/2} R^N_0.$$

Find a moderate deviations principle for $D^N|_{(-t,0]}$ suitably normalized. ◇

Exercise 9.2
Prove that the *process* $(R^N_{-t}, t \geq 0)$ is exponentially equivalent to $(Q^N_{-t}\nu/\mu, t \geq 0)$. *(Challenging.)* ◇

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 $N^{(1+\beta)/2}Q^N_{-t}$ be the queue size at time $-t$. Define the departure process in the usual way:

$$ND^N(-t,0] = NA^N(-t,0] + N^{(1+\beta)/2}Q^N_{-t} - N^{(1+\beta)/2}Q^N_0.$$

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 $\delta > 0$,

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

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

$$\limsup_{N\to\infty} \frac{1}{N^{\beta'}} \log P\Big(\sup_{t>0} N^{(1+\beta)/2}\Big|\frac{Q^N_0}{t+1} - \frac{Q^N_{-t}}{t+1}\Big| > \delta N^{(1+\beta')/2}\Big) = -\infty.$$

Now,

$$\sup_{t>0}\Big|\frac{Q^N_0}{t+1}-\frac{Q^N_{-t}}{t+1}\Big|\leq Q^N_0+\sup_{t>0}\frac{Q^N_{-t}}{t+1}.$$

The following lemma proves that

$$\limsup_{N\to\infty}\frac{1}{N^{\beta'}}\log P\big(N^{(1+\beta)/2}Q^N_0>N^{(1+\beta')/2}\delta\big)=-\infty.$$

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

Lemma 9.3 *If the arrival process A^N, normalized, satisfies the sample path moderate deviations principle at scale β'; and if $Q^N_0=q(LA^N,N\mu+N^{(1+\beta)/2}C)$ for $\beta'<\beta$ and $\mu<C$, then*

$$\limsup_{N\to\infty}\frac{1}{N^{\beta'}}\log P\big(N^{(1+\beta)/2}Q^N_0>N^{(1+\beta')/2}\delta\big)=-\infty. \tag{9.7}$$

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

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

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:

$$J(q,C)=\inf_{t\geq0}\frac{(q+Ct)^2}{2V_t}\geq C^2\inf_{t\geq0}\frac{t^2}{2V_t}.$$

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

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. ◇

Chapter 10

Interpretations

Readers of a practical bent will by now be bursting with the question: This large deviations theory is all very well, but how do I actually *use* it?

We start in Section 10.1 by describing the theory in a more tangible way, in terms of effective bandwidths.

One of the purposes of a limit theorem—but by no means the most important—is to give a numerical approximation to a quantity which is hard to calculate exactly. How numerically accurate are the different large deviations estimates? We investigate this in Section 10.2.

Another purpose of an estimate—perhaps more important than simply giving numerical approximations—is to tell us about the general structure of the solution. The different large deviations principles can tell us different things about the scaling properties of a queueing system. We address this in Sections 10.3 and 10.4. In particular, we introduce the *global approximation*, a heuristic formula which predicts the proper scaling to use for all the large deviations results in this book.

Finally, in Section 10.5 we describe large deviations results for some standard traffic models.

10.1 Effective Bandwidths

There is a certain transformation of the log moment generating function which has become popular in the literature on communications networks. It is called the *effective bandwidth* of a traffic flow, and it is a convenient and intuitive descriptor of the flow's stochastic properties, at least as far as large deviations queueing behaviour is concerned. Effective bandwidths of

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 211–238, 2004.

the sort that arise from large deviations theory were introduced by Kelly [53]; for further details see Kelly [55], Gibbens [45] and Kesidis et al. [56].

10.1.1 Effective Bandwidths for the Large-Buffer Scaling

Recall our very first result about large deviations for queues, Theorem 1.4. Let $A(s,t]$ be the amount of work arriving at a queue in the interval $(s,t]$, $s<t\in\mathbb{N}$, and suppose that the increments $\dot{A}_t$ are i.i.d. (We are using the extended notation for arrival processes described in Section 5.5. We mean that $A(s,t]=\dot{A}_{s+1}+\cdots+\dot{A}_t$.) Let the queue have service rate C. Let Q be the queue size at time 0 (or equivalently, the stationary queue size). Then, if $E\dot{A}_0<C$,

$$\lim_{l\to\infty}\frac{1}{l}\log P(Q/l>q)=-q\sup\{\theta>0:\Lambda(\theta)<\theta C\}$$

where $\Lambda(\theta)$ is the log moment generating function

$$\Lambda(\theta)=\log Ee^{\theta\dot{A}_0}.$$

Interpret this another way. For some given tail decay parameter $\gamma>0$, what service rate C do we need to ensure

$$P(Q>q)<e^{-\gamma q}\,?$$

(In the context of telecommunications, one may wish to provide a guarantee of this form. Such guarantees are called 'quality of service' guarantees, and so γ is sometimes called a 'quality of service' parameter.) The answer is approximately

$$C=\gamma^{-1}\Lambda(\gamma)$$

and for this reason, the function $\alpha(\theta)=\theta^{-1}\Lambda(\theta)$ is known as the *effective bandwidth function* of the arrival process. Note that effective bandwidth is additive for independent sources.

The effective bandwidth function has the following property: $\alpha(\theta)$ is increasing in θ, and lies between the mean rate $E\dot{A}_0$ and the peak rate, which is $\operatorname{ess\,sup}\dot{A}_0=\inf\{x:P(\dot{A}_0\le x)=1\}$.

Exercise 10.1
Show that $\lim_{\theta\to 0}\alpha(\theta)=E\dot{A}_0$. Show that $\alpha(\theta)$ is increasing in θ. (Hint: Use Hölder's inequality.) Show that $\lim_{\theta\to\infty}\alpha(\theta)=\operatorname{ess\,sup}\dot{A}_0$. ◇

Of course, by Theorem 3.1, the same interpretation holds when the increments $(\dot{A}_t, t \in \mathbb{Z})$ are weakly dependent, though the interpretation of peak rate is slightly different. In this case the effective bandwidth is

$$\alpha(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E e^{\theta A(-t,0]}.$$

Effective bandwidths can also be used for 'admission control', in the following sense. Suppose there are m independent copies of arrival process A, with effective bandwidth $\alpha(\theta)$, and n independent copies of arrival process B, with effective bandwidth $\beta(\theta)$. What values of m and n can we admit while maintaining quality of service γ? The answer is simple and linear:

$$\big\{m, n : m\alpha(\gamma) + n\beta(\gamma) < C\big\}. \tag{10.1}$$

The effective bandwidth functions thus measure the tradeoff between flows. Alternatively it measures the tradeoff between mean rate and burstiness of flows:

Exercise 10.2
Let $\alpha(\theta)$ be the effective bandwidth of an arrival process where the arrivals at different timesteps are independent normal random variables with mean μ and variance σ^2. Let $\beta(\theta)$ be the same, but with mean ν and variance ρ^2. Sketch the admission control region (10.1) for different values of γ. ◇

Having studied sample path large deviations for queues with large buffers in Chapter 6, we are in a better position to understand effective bandwidths in networks. The reason the effective bandwidth is important is because the sequence of scaled polygonalized arrival processes $\tilde{A}^N$ (see Section 5.2 to remind yourself of the scaling) satisfies a sample path LDP in the space of continuous processes with good rate function

$$I(a) = \int_{-\infty}^{0} \Lambda^*(\dot{a}_t)\, dt \tag{10.2}$$

where $\dot{a}_t$ is the instantaneous arrival rate at time t and Λ^* is the convex conjugate of $\theta\alpha(\theta)$,

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

Consider the output of a queue with fixed service rate C. We saw in Exercise 6.7 that the sequence of scaled departure processes $\tilde{D}^N$ satisfies a sample path LDP with good rate function

$$J(d) = \int_{-\infty}^{0} \mathrm{M}^*(\dot{d}_t)\, dt$$

where the rate function M^* given by

$$\mathrm{M}^*(x) = \begin{cases} \Lambda^*(x) & \text{if } x \le C \\ \infty & \text{otherwise.} \end{cases}$$

In fact, the departure process has an effective bandwidth. Since Λ^* is convex and lower semicontinuous, so is M^*. Now by duality of convex conjugates (Lemma 2.5) M^* satisfies

$$\mathrm{M}^*(x) = \sup_\theta \theta x - \theta\delta(\theta)$$

where $\delta(\theta)$ is defined by

$$\delta(\theta) = \frac{1}{\theta}\Big(\sup_x \theta x - \mathrm{M}^*(x)\Big),$$

which is simply

$$= \frac{1}{\theta}\Big(\sup_{x \le C} \theta x - \Lambda^*(x)\Big).$$

We may call this the effective bandwidth function of the departure process, since it plays the same role that $\alpha(\theta)$ does for the arrival process, that is, it determines the integrand in the rate function. Note that by duality of convex conjugates

$$\alpha(\theta) = \frac{1}{\theta}\Big(\sup_x \theta x - \Lambda^*(x)\Big),$$

This implies that the effective bandwidth function for the departure process is smaller (pointwise) than the effective bandwidth function for the arrival process.

> *Note.* The above is a formal way to obtain an effective bandwidth function for the departure process. A concrete approach is as follows. By the contraction principle, as in Section 6.8, $\tilde{D}^N(-t,0]$ satisfies an LDP with good rate function $I(x) = x\mathrm{M}^*(x/t)$. By Varadhan's Lemma (Lemma 4.11), using the the bounded continuous function $f(x) = \theta(x \wedge Ct)$,
>
> $$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \log E e^{N\theta\tilde{D}^N(-t,0]} &= \lim_{N\to\infty} \frac{1}{N} \log E e^{N\theta(\tilde{D}^N(-t,0]\wedge Ct)} \\ &= \sup_{x\in\mathbb{R}} \theta(x \wedge Ct) - I(x) \\ &= \sup_{x \le Ct} \theta x - t\Lambda^*(x/t). \end{aligned}$$

In particular, the scaled limiting log moment generating function (i.e. the effective bandwidth) $\delta(\theta)$ of the departure process is

$$\delta(\theta) = \frac{1}{\theta} \lim_{t\to\infty} \frac{1}{t} \log E e^{\theta D(-t,0]} = \sup_{x \leq C} \theta x - \Lambda^*(x).$$

However, in order to apply the results of Chapter 6, it is not sufficient to have an LDP for $\tilde{D}^N(-t,0]$—we need a full sample path LDP with linear geodesics. Exercise 6.7 shows that such an LDP holds, if the service rate is constant.

Unfortunately there is little more positive that can be said about departure processes. If the service is stochastic, or if there are several input flows, then as noted in Section 6.7 the output process may not have a rate function of the form (10.2). Therefore, while it may be possible to find the scaled cumulant moment generating function, it will not have the same interpretation and it is not appropriate to call it an effective bandwidth.

10.1.2 Effective Bandwidths for the Many-Flows Scaling

Effective bandwidths of a different sort make sense in the many-flows scaling. Recall the basic result, Theorem 1.8. Consider a single-server queue with N identical independent arrival flows and constant service rate CN. Let $A(-t,0]$ be the amount of work arriving in time interval $(-t,0]$ from one of the sources, and assume the mean arrival rate is less than CN. Let Q^N be the queue size at time 0 (or equivalently, the stationary queue size). Then

$$\lim_{N\to\infty} \frac{1}{N} \log P(Q^N > Nq) = -\inf_{t>0} \sup_{\theta>0} \theta(q+Ct) - \Lambda_t(\theta) \tag{10.3}$$

where

$$\Lambda_t(\theta) = \log E e^{\theta A(-t,0]}.$$

(Lemma 7.9 gives a rate function of this form. That lemma has $t \geq 0$ and $\theta \geq 0$. We can clearly replace the latter by $\theta > 0$, and we can replace the former by $t > 0$ if $q > 0$. The rate function is increasing, though not necessarily continuous; we assume for convenience that it is continuous at q, in order to obtain a limit.)

Suppose the optimum in (10.3) is attained, and the optimizing parameters are $\hat{\theta}$ and $\hat{t}$, both strictly positive. (These are referred to as the critical spacescale and timescale.) Suppose we replace a relatively small number

$\lfloor \delta N \rfloor$ of these flows by constant-rate flows of rate $(\hat{\theta}\hat{t})^{-1}\Lambda(\hat{\theta},\hat{t})$. The rate function is now

$$J_\delta(q) = \inf_{t>0} \sup_{\theta>0} \theta(q+Ct) - \big((1-\delta)\Lambda(\theta,t) + \delta\Lambda(\hat{\theta},\hat{t})\big).$$

Locally, at $(\hat{\theta},\hat{t})$, these new flows have the same log moment generating function as the flows they replace, so the rate function $J_\delta(q)$ is the same as $J(q)$ (to first order, under appropriate smoothness conditions). In other words, a flow with log moment generating function Λ_t has the same effect, at operating point $(\hat{\theta},\hat{t})$, as a flow of constant rate $\alpha(\hat{\theta},\hat{t})$, where

$$\alpha(\theta,t) = \frac{1}{\theta t}\Lambda_t(\theta).$$

This quantity is called the *effective bandwidth* of the flow. Note that effective bandwidth is additive for independent sources.

Alternatively, think of effective bandwidths in terms of admission regions. Suppose there are $\lfloor mN \rfloor$ flows with effective bandwidth $\alpha(\theta,t)$, and $\lfloor nN \rfloor$ flows with effective bandwidth $\beta(\theta,t)$. For what values of m and n does the system meet the quality of service constraint

$$P(Q^N > Nq) < e^{-\gamma N}\,?$$

The admissible region is

$$\bigcap_{t>0} \Big\{ m,n : \exists\, \theta>0 : m\alpha(\theta,t) + n\beta(\theta,t) < C + \frac{q}{t} - \frac{\gamma}{\theta t} \Big\}.$$

So the effective bandwidth gives the tradeoff between flows of different types. Note that the queue looks like an infinite collection of logical resources, one for each t, and the pair (m,n) is admissible if it is admissible at each of the logical resources.

Rabinovitch [86] has investigated statistical estimation of effective bandwidth, and Courcoubetis et al. [20] have investigated ways to estimate overflow probability and admission control regions without explicitly estimating the effective bandwidth.

Effective bandwidths characterize the queue size distribution, in a large deviations sense. Note however that they do not characterize the sample path large deviations of a flow, as the following example (prompted by Frank Kelly) shows.

Example 10.3
We will construct two arrival processes X and Y with stationary increments, with identical distributions for $X(-t,0]$ and $Y(-t,0]$.

Let U be a random variable with the uniform distribution on $\{0,1,2\}$. Let $X(-t,0] = \dot{X}_{-t+1} + \cdots + \dot{X}_0$ where the $\dot{X}_t$ are independent and distributed like U.

Let Y be

$$(\ldots,\dot{Y}_{-3},\dot{Y}_{-2},\dot{Y}_{-1},\dot{Y}_0) = \begin{cases} (\ldots,P_1,Q_1,P_0,Q_0) & \text{with probability } \frac{1}{2} \\ (\ldots,Q_2,P_1,Q_1,P_0) & \text{with probability } \frac{1}{2} \end{cases}$$

where the pairs (P_i,Q_i) are independent and identically distributed, P_i is distributed like U, and $Q_i = f(P_i,Z_i)$, where the Z_i are independent and distributed like U, and

$$f(i,j) = \begin{cases} j+1 & \text{if } j = i+1 \\ j & \text{otherwise.} \end{cases}$$

Then Q_i is also distributed like U, and, by simple counting of probabilities, $P_i + Q_i$ has the same distribution as the sum of two independent copies of U. The regenerative structure of Y ensures that $Y(-t,0]$ has the same distribution as $X(-t,0]$ for all t. Thus the two processes have the same effective bandwidth function, even though they have different distributions (and their typical paths to overflow are also different). ◇

What about the effective bandwidth of the departure process? Let D^N be the aggregate departure process. As in the large buffer case, we know that D^N satisfies a large deviations principle, and in particular that $D^N(-t,0]$ satisfies a large deviations principle, which lets us compute a log moment generating function $\mathrm{M}_t(\theta)$ for departures over $(-t,0]$. However, it is not known whether the rate function for the departure process is convex, which means it is not known whether Lemma 7.8 is satisfied, which means that Theorem 7.7 might not apply, which means it is not known whether the rate function for overflow at a downstream queue has the form (10.3) (though that equation will still hold as a conservative bound). Thus we do not know if the scaled cumulant moment generating function of the departure process deserves to be called an effective bandwidth.

Section 7.11 describes another way of thinking about networks in which flows follow diverse routes. In that setup, it make sense to talk about the effective bandwidth of a single departure flow, and that effective bandwidth is exactly the same as that of the corresponding arrival flow.

10.2 Numerical Estimates

The large deviations results in the preceding chapters concern limiting results; in this section we turn them into numerical estimates. We have looked at several different scaling regimes. Which is most useful? The answer is trite: it depends on what question we are trying to answer. We seek, then, to give some guidance about the strengths and weaknesses of the different approaches.

10.2.1 From Limits to Estimates

Consider, for example, the limit theorem for the many-flows scaling. Let Q^N be the steady-state queue size in a single-server queue with N identical independent arrival flows, constant service rate CN and finite buffer Nq. Let $A(-t,0]$ be the amount of work arriving in time interval $(-t,0]$ from one of the sources, and assume the total mean arrival rate is less than CN. Then

$$\lim_{N\to\infty} \frac{1}{N} \log P(Q^N = Nq) = -I(q)$$

where

$$\begin{aligned} I(q) &= \inf_{t>0} \sup_{\theta>0} \theta(q+Ct) - \Lambda_t(\theta) \\ \Lambda_t(\theta) &= \log E e^{\theta A(-t,0]}. \end{aligned}$$

Note that this event corresponds to overflow—if $\{Q_t^N = Nq\}$ then an amount of work $Q_{t-1}^N + \dot{A}_t^N - NC - Nq$ is lost at time t.

(Technically, this is a large deviations upper bound. If the rate function for overflow $I(q)$ is continuous at q, then the limit holds. The same limit holds for the probability that the queue size in an infinite buffer exceeds Nq.)

Take this as a naive estimate: $\log P(\text{overflow}) \approx -NI$: and rewrite it in terms of the actual parameters of the system

$$\begin{aligned} \tilde{Q} &= Q^N \\ \tilde{C} &= NC \\ \tilde{q} &= Nq \\ \tilde{\Lambda}_t(\theta) &= \log E e^{\theta N A^N(-t,0]}, \\ \tilde{I} &= \inf_{t>0} \sup_{\theta>0} \theta(\tilde{q} + \tilde{C}t) - \tilde{\Lambda}_t(\theta), \end{aligned}$$

where $NA^N(-t,0]$ is the total amount of work arriving at the queue in interval $(-t,0]$, to get

$$\log P(\tilde{Q}=\tilde{q}) \approx -\tilde{I}.$$

Notice that N has cancelled out. In other words, to apply the many-flows estimate, we don't actually need to know how many flows are present! Notice also that (formally) this estimate can be obtained by simply setting $N=1$. The same happens in all the other scalings we have studied in this book.

10.2.2 Common Form of Estimates

In fact, all the estimates from the different scaling regimes have a common form. Consider a queue with service rate C and buffer size q. (We have dropped the $\sim$ notation, but we still mean that these are the actual service rate and buffer size.) This common form is

$$\log P(\text{overflow}) \approx -I(q)$$

where

$$I(q) = \inf_{t>0} \sup_{\theta>0} \theta(q+Ct) - \mathrm{M}_t(\theta). \tag{10.4}$$

The different scaling regimes correspond to different expressions for $\mathrm{M}_t(\theta)$, given in Table 10.1. These expressions are given in terms of $\Lambda_t(\theta)$,

$$\Lambda_t(\theta) = \log E e^{\theta A(-t,0]},$$

where $A(-t,0]$ is the total amount of work arriving in an interval of length t. The table refers to the large-buffer scaling (LDLB), the many-flows scaling (LDMF), the large-buffer scaling for traffic with long range dependence (LDLBH), and also moderate deviations versions of the large-buffer (MDLB) and many-flows (MDMF) scalings.

In LDLB, the rate function simplifies to $I(q)=qI(1)$, as shown in Lemma 3.4. In the moderate deviations scalings, $\mathrm{M}_t(\theta)$ is quadratic in θ, so $I(q)$ simplifies. In MDLB, $\mathrm{M}_t(\theta)$ is linear in t, so $I(q)$ is very simple indeed.

10.2.3 Refined Estimates

Some of these estimates can be improved upon, although this requires further assumptions. In the many-flows limit, Likhanov and Mazumdar [59] have used the Bahadur-Rao theorem, details of which are given by Dembo and

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

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

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

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

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

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,

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

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 $\mathrm{M}_t(\theta)$ 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,

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

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

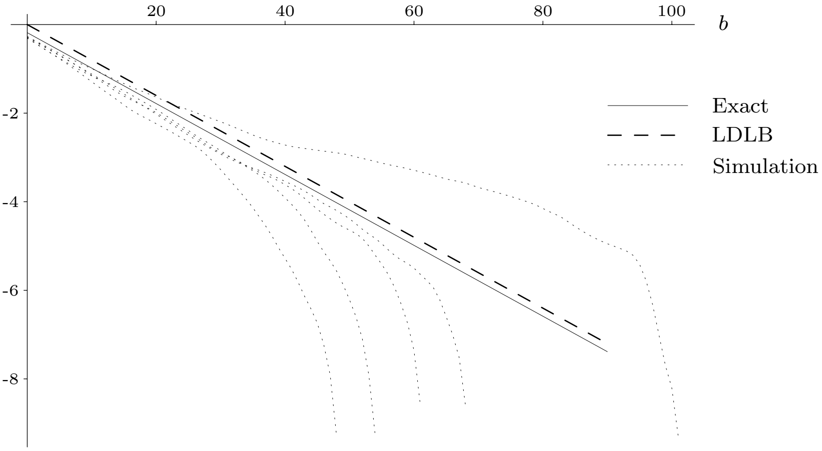

Figure 10.1: Watermark plot of $\log P(Q > b)$ against b, for a Markov on/off source. The dotted lines are from simulation, the black line is the true value (which in this case we can calculate exactly), and the dashed line is the LDLB estimate. This estimate clearly has the right slope.

watermark curves tail off. A small number of simulation runs reach $\{Q > b\}$, and on reaching $\{Q > b\}$ they are reasonably likely to go on a spree and reach higher values, leading to a small number of watermark curves with inflated estimates for the probability.

The next example illustrates the difference between the LDLB and LDMF estimates. In Figure 10.2 the traffic is generated by a Gaussian autoregressive source:

$$A_t = \mu + a(A_{t-1} - \mu) + \sqrt{1-a^2}\sigma\varepsilon_t$$

where the ε_t are independent standard normal random variables. With this choice of parameters, $A_t \sim \text{Normal}(\mu, \sigma^2)$. In (a) $a = 0.8$ and in (b) $a = -0.6$. The other parameters are the same: the mean rate is $\mu = 0.8$ and the marginal variance is $\sigma^2 = 0.2$; and the service rate of the queue is 1. In (b), over the range of buffer sizes plotted, the most likely number of timesteps until overflow is small (and odd); so the LDLB estimate, which is only concerned with long-timescale correlations, is not accurate. For larger buffer sizes it takes a long time for the buffer to fill, and LDLB is relatively more accurate. Note the different scales of buffer size in the two plots—queues fed by negatively-correlated traffic are much less likely to overflow than queues fed by positively correlated traffic.

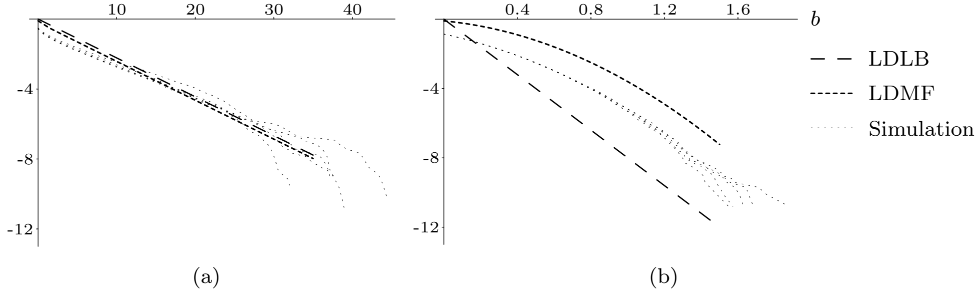

Figure 10.2: Watermark plots of $\log P(Q > b)$ against b, for a Gaussian autoregressive source which is (a) positively correlated and (b) negatively correlated. In (a), the estimates LDLB and LDMF are in good agreement; in (b) they are not. Both have the correct asymptotic slope.

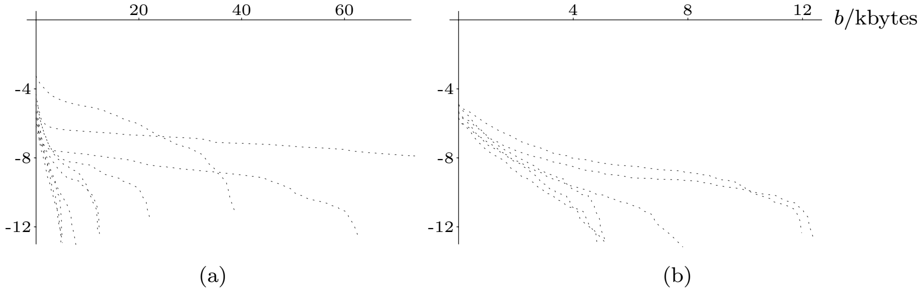

Figure 10.3: Watermark plots for real data. The data is an hour-long trace of TCP traffic. Figure (a) shows ten watermark plots for ten intervals of 42 seconds each, taken throughout the hour. Figure (b) shows the watermark plots for the six of these intervals that occur between 22 and 52 minutes. The scales are very different!

Figure 10.3 shows what a watermark plot looks like for real data. The data comes from Lawrence Berkeley Laboratory, and consists of an hour's worth of TCP packets between the laboratory and the rest of the world [82]. This traffic trace was fed into a simulated queue with service rate 400 kbytes/second (bearing in mind a 40 byte TCP/IP overhead for each packet, running the queue in continuous time, sampling the queue size at period intervals, measuring the total number of bytes in the queue rather than the number of packets, and supposing that bytes are drained from the queue at a constant rate). Rather than creating a single watermark plot of the queue length distribution over the entire hour, we took 10 intervals of 42 seconds each, and plotted 10 different watermark plots. The mean arrival rates over those 10 slices were

$$83.5 \quad 31.8 \quad 37.9 \quad 44.2^b \quad 47.6^b \quad 34.8^b \quad 45.5^b \quad 41.5^b \quad 36.1^b \quad 27.4$$

kbytes/second. The slices marked with a superscript are plotted on their own in Figure 10.3(b). For these parts of the hour, the queue size seems to have a reasonably well-behaved exponential tail, whereas for the other parts it seems worse than exponential. Is this a consequence of long range dependence, or does it indicate that traffic over the course of the hour was non-stationary? This is a philosophical issue, discussed in Section 8.6.

The next example looks at the many flows limit. In Figure 10.4 we let the queue be fed by N copies of an autoregressive Gaussian source (each with mean rate $\mu = 0.9$, autocorrelation parameter $a = 0.6$, and marginal variance $\sigma^2 = 0.5$), and let the queue have service rate N and buffer threshold N. (For Gaussian sources it is reasonable to imagine N to be fractional.) Both the crude and refined LDMF estimates leave something to be desired; and the LDLB estimate doesn't even have the right slope.

Our last example looks at moderate deviations estimates. Figure 10.5 plots overflow probability as a function of traffic intensity ρ. The queue has buffer size 1 and service rate 1. In (a) the traffic has independent increments: $A_t \sim \mathrm{Bin}(2, \rho/2)$. In (b) the traffic has a correlation structure: it is a Markov on/off source with peak rate 2, which jumps from off to on with probability $0.1\rho/(2-\rho)$ and from on to off with probability 0.1.

In (a), the traffic has independent increments, so the LDLB and LDMF estimates agree, and are in fact exact. (The figure plots the LDMF refined estimate, which is actually worse!) In (b) the traffic has significant correlations, and the LDLB estimate is so poor we haven't plotted it—it estimates $\log P(Q^\rho > b) > -0.1$ for the full range of ρ plotted.

In (a) we plot another estimate, PLT. This estimate comes from the

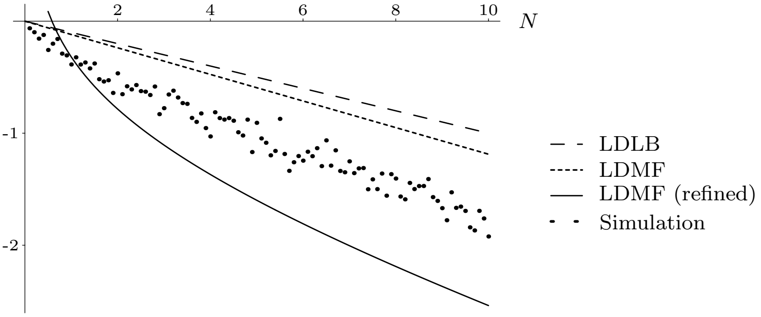

Figure 10.4: Plot of $\log P(Q^N > Nb)$ against N, where Q^N is fed by N independent flows and has service rate NC. The dots are simulated values. Theory says that $\log P(Q^N > Nb)$ is linear in N, and that both LDMF estimates have the correct limiting slope, though it is hard to see this from the plot!

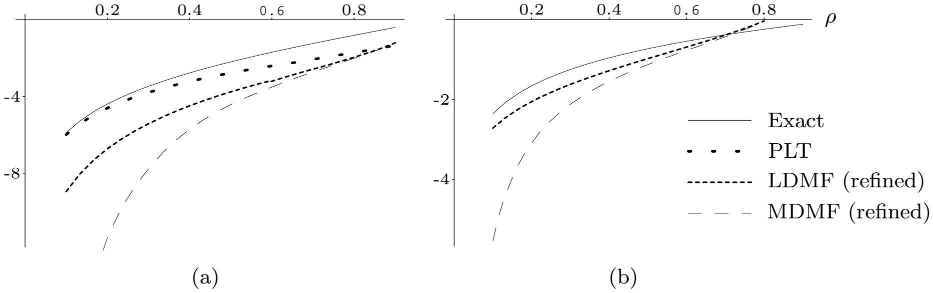

Figure 10.5: $\log P(Q^\rho > b)$ as a function of traffic intensity ρ. The queue parameters are fixed, and the parameters of the traffic process are changed to give intensity ρ. In (a) the traffic has independent increments; in (b) there are significant correlations. The moderate deviations estimate is reasonable at high traffic intensities, when overflow is governed by first and second moments; it is poor at low traffic intensities, when higher order moments are more important.

principle of the largest term, which says

$$P(Q > b) = P(\sup_t A(-t,0] > b + Ct) \approx \sup_t P(A(-t,0] > b + Ct).$$

The plot shows that this estimate is good for small ρ, and worse for large ρ. The LDMF estimate is an approximation to PLT—it further approximates

$$\log P\bigl(A(-t,0] > b + Ct\bigr) \approx -\sup_\theta \theta(b + Ct) - \log Ee^{\theta A(-t,0]}.$$

So the inaccuracy of LDMF arises in two ways, and we can see from the plot that the relative contribution of the two errors varies as ρ increases.

The moderate deviations estimate MDMF makes an additional approximation:

$$\log Ee^{A(-t,0]} \approx \theta EA(-t,0] + \tfrac{1}{2}\theta^2 \operatorname{Var}(A(-t,0]),$$

i.e. it uses a normal approximation to $A(-t,0]$. This is reasonable at high traffic intensities, when the optimizing θ is small and the LDMF estimate is governed by first and second moments; it is poor at low traffic intensities, when the optimizing θ is large, meaning that higher order moments are more important.

10.3 A Global Approximation

The last section addressed the question: how accurate an estimate can we find? A more interesting question is: how much information can we throw away, and still have a reasonable estimate? By choosing a scaling regime we are, in effect, choosing what sort of information to throw away.

Consider as usual a single-server queue with service rate C and an infinite buffer, to which an amount of work $A(-t,0]$ arrives in the time interval $(-t,0]$. Consider the following approximation (forgetting its relationship to the many-flows rate function):

$$\begin{aligned}
&\log P(Q > q)\\
&\quad = \log P\bigl(\sup_{t\ge 0} A(-t,0] - Ct > q\bigr)\\
&\quad \approx \sup_{t\ge 0} \log P\bigl(A(-t,0] > q + Ct\bigr) \quad \text{(principle of the largest term)}\\
&\quad \approx \sup_{t\ge 0} \inf_{\theta\ge 0} \log Ee^{\theta A(-t,0] - \theta(q+Ct)} \quad \text{(Chernoff estimate)}\\
&\quad = -\inf_{t\ge 0} \sup_{\theta\ge 0} \theta(q + Ct) - \Lambda_t(\theta) \qquad (10.7)
\end{aligned}$$

where

$$\Lambda_t(\theta) = \log Ee^{\theta A(-t,0]}.$$

The two approximation steps are likely to be valid in nearly any large deviations limit.

This turns out to be a *global approximation*, in the sense that Ward and Glynn [96] use the term, namely that it can be used to motivate all the different limit results we have come across.

Large buffers. Take the large-buffer scaling. Let N be large, let $t = Ns$, and rewrite (10.7):

$$\log P(Q > Nq) \approx -\inf_{s\geq 0}\sup_{\theta\geq 0}\theta(Nq + CNs) - Ns\frac{1}{Ns}\Lambda_{Ns}(\theta).$$

If the scaled log moment generating functions converge, say $t^{-1}\Lambda_t(\theta) \to \Lambda_\infty(\theta)$, then approximating further

$$\frac{1}{N}\log P(Q/N > q) \approx -\inf_{s\geq 0}\sup_{\theta\geq 0}\theta(q + Cs) - s\Lambda_\infty(\theta), \tag{10.8}$$

precisely the form of the large deviations principle in Chapter 6.

Long range dependence. Or take the large-buffer scaling for flows with long-range dependence, where

$$\frac{1}{t^{2(1-H)}}\Lambda_t(t^{-(2H-1)}\theta) \to \Lambda_{\infty(H)}(\theta).$$

Note that then

$$\frac{1}{N^{2(1-H)}}\Lambda_{Ns}(N^{1-2H}\theta) \to \mathrm{M}_t(\theta) = t^{2(1-H)}\Lambda_{\infty(H)}(\theta t^{2H-1}). \tag{10.9}$$

Again, by reparameterizing carefully, this time in terms of $t = Ns$ and $\theta = N^{1-2H}\phi$,

$$\log P(Q > Nq) \approx$$
$$-\inf_{s\geq 0}\sup_{\phi\geq 0}\phi N^{1-2H}(Nq + CNs) - N^{2(1-H)}\frac{1}{N^{2(1-H)}}\Lambda_{Ns}(N^{1-2H}\phi),$$

thus
$$\frac{1}{N^{2(1-H)}} \log P(Q/N > q) \approx -\inf_{s\geq 0} \sup_{\phi \geq 0} \phi(q + Cs) - \mathrm{M}_s(\phi)$$
which corresponds to the large deviations principle in Chapter 8. The parameter scalings (Nq, Ns and $N^{1-2H}\phi$) are as they are in order to fit in with (10.9).

Many flows. Or take the many-flows scaling. Let Q^N be a queue with service rate NC, fed by A^N, the aggregate of N independent copies of A. Let $\Lambda_t(\theta)$ be the log moment generating function for a single copy of A, so that the log moment generating function for the aggregate input is
$$\log Ee^{\theta A^N(-t,0]} = N\Lambda_t(\theta).$$

Then
$$\log P(Q^N > Nq) \approx -\inf_{t\geq 0} \sup_{\theta\geq 0} \theta(Nq + NCt) - N\Lambda_t(\theta)$$

and
$$\frac{1}{N} \log P(Q^N/N > q) \approx -\inf_{t\geq 0} \sup_{\theta\geq 0} \theta(q + Ct) - \Lambda_t(\theta),$$

the form of the large deviations principle in Chapter 7.

Moderate deviations. The global approximation also suggests the form of the moderate deviations principle. This is a more involved calculation. If $\Lambda_t(\theta)$ is finite in a neighbourhood of the origin, it has a power series expansion:
$$\Lambda_t(\theta) = \theta\mu t + \tfrac{1}{2}\theta^2\sigma_t^2 + O(\theta^3)$$
where $\mu t = EA(-t, 0]$ and $\sigma_t^2 = \operatorname{Var} A(-t, 0]$. Thus
$$N^{2\gamma}\big(\Lambda_t(\phi N^{-\gamma}) - \phi\mu t N^{-\gamma}\big) = \tfrac{1}{2}\sigma_t^2\phi^2 + O(N^{-\gamma}) \tag{10.10}$$

as $N \to \infty$, for $\gamma > 0$. Let
$$\mathrm{M}_t(\theta) = \tfrac{1}{2}\sigma_t^2\theta^2 \; :$$

this is the log moment generating function of a normal random variable with mean 0 and variance σ_t^2. Let Q^N be the steady state queue size in a queue fed by N independent copies of A, and let $\Lambda_t(\theta)$ be the log moment generating function for a single copy. How should we scale the buffer size and service rate to bring out the approximation (10.10)? It isn't immediately clear, so just let the buffer size be q^N and the service rate C^N. Then

$$\begin{aligned}
&\log P(Q^N > q^N)\\
&\quad\approx -\inf_{t\geq 0}\sup_{\theta\geq 0}\ \theta(q^N + C^N t) - N\Lambda_t(\theta)\\
&\quad\approx -\inf_{t\geq 0}\sup_{\phi\geq 0}\ \phi N^{-\gamma}(q^N + C^N t)\\
&\qquad\qquad\qquad - N\Big(\phi N^{-\gamma}\mu t + \tfrac{1}{2}\phi^2 N^{-2\gamma}\sigma_t^2 + O(N^{-3\gamma})\Big)\\
&\quad\approx -\inf_{t\geq 0}\sup_{\phi\geq 0}\ \phi\Big(N^{-\gamma}q^N + (N^{-\gamma}C^N - N^{1-\gamma}\mu)t\Big)\\
&\qquad\qquad\qquad - N^{1-2\gamma}\mathrm{M}_t(\phi) + O(N^{1-3\gamma})
\end{aligned}$$

so

$$\begin{aligned}
&\frac{1}{N^{1-2\gamma}}\log P(Q^N > q^N)\\
&\quad\approx -\inf_{t\geq 0}\sup_{\phi\geq 0}\ \phi\Big(N^{-(1-\gamma)}q^N + (N^{-(1-\gamma)}C^N - N^{\gamma}\mu)t\Big)\\
&\qquad\qquad\qquad - \mathrm{M}_t(\phi) + O(N^{-\gamma}).
\end{aligned}$$

If we scale the system according to $q = N^{-(1-\gamma)}q^N$ and $C = N^{-(1-\gamma)}C^N - N^{\gamma}\mu$, then

$$\frac{1}{N^{1-2\gamma}}\log P(Q^N > N^{1-\gamma}q) \approx -\inf_{t\geq 0}\sup_{\phi\geq 0}\phi(q + Ct) - \mathrm{M}_t(\phi).$$

We can equivalently describe Q^N as the steady state queue size in a queue fed by N independent copies of A and served at rate $N\mu + N^{1-\gamma}C$. We prefer to write it in terms of $\beta = 1 - 2\gamma$:

$$\frac{1}{N^{\beta}}\log P(Q^N > N^{(1+\beta)/2}q) \approx -\inf_{t\geq 0}\sup_{\phi\geq 0}\phi(q + Ct) - \mathrm{M}_t(\phi) \qquad (10.11)$$

where Q^N is the steady-state queue size in a queue fed by N independent copies of A and served at rate $N\mu{+}N^{(1+\beta)/2}C$, and thus with traffic intensity roughly $1 - N^{-(1-\beta)/2}C/\mu$. We need $\beta < 1$ for the approximation of Λ_t to work. And we need $\beta > 0$ for this to be a rare event, a suitable limit to study using large deviations techniques.

Conclusion. So the different scaling regimes are, in effect, taking the global approximation (10.7) and approximating the log moment generating function in different ways. They approximate $\Lambda_t(\theta) \approx \mathrm{M}_t(\theta)$, where $\mathrm{M}_t(\theta)$ is as in the table on page 220.

Take, for example, the LDLB approximation $\Lambda_t(\theta) \approx t\mathrm{M}(\theta)$. The large-buffer estimate of the probability of overflow (10.8) will be good if this approximation is valid at the timescale of interest. In this case the timescale of interest is $t = \hat{s}N$, where $\hat{s}$ is the optimizing parameter in (10.8) and N is the scale factor for the queue size Nq.

Or we can say the same thing the other way round. If the LDLB approximation to Λ_t is good for t in a certain range, we can find the range of system parameters (buffer size and service rate) for which the large-buffer probability estimate is good: namely those system parameters which lead to $t = \hat{s}N$ in that good range.

Exercise 10.4
For the moderate deviations large-buffer scaling (MDLB), we approximate

$$\Lambda_t(\theta) \approx \theta\mu t + \tfrac{1}{2}\theta^2\sigma^2 t.$$

Find the parameter scaling under which this is a good approximation. ◇

Exercise 10.5
Let Q be the queue size in a queue fed by a single source. The LDLB limit concerns a large deviations principle for Q/N as $N \to \infty$. Use the global approximation to guess an LDP for Q/N^γ, where $\gamma \in (0,1)$. ◇

10.4 Scaling Laws

So far in this chapter, and indeed in most of this book, we have been missing the wood for the trees—the wood being the scaling law, the trees being the rate function. In all our results, there are four quantities that are scaled: the number of flows, the buffer threshold, the service rate, and the probability that the queue size exceeds the threshold. Each limiting regime describes a different relationship between these quantities.

For example, in a large-buffer large-deviations statement like

$$\log P(\text{queue length exceeds } q) \approx -q \sup\{\theta : \Lambda(\theta) < \theta C\} \quad \text{for large } q$$

we hold the number of flows and the service rate fixed, and consider the relationship between threshold q and probability of exceeding that threshold.

The most important part of the statement is that the queue length has an exponential tail; the value of the decay rate is subsidiary to this.

We remarked in Section 6.10 on the wide applicability of this idea. If the queue size at any point in the network can be written as a continuous function of the collection of all input and service processes, then that queue size should have an exponential tail. Even though the rate function is unworkably complicated, we can still say something useful about the scaling law.

This sort of result applies to the other scalings we have considered. The LRD result says that for a queue fed by a single source with Hurst parameter H, the probability of overflow decays like $\exp(-q^{2(1-H)}I)$; if all inputs to a network have the same Hurst parameter, then all queues in the network should have the same decay rate. (In Chapter 8 we saw that if the inputs have different Hurst parameters then it is the largest Hurst parameter that dominates.)

The many-flows result says that if buffer size and service rate are scaled up in proportion to the number of independent arrival processes N, then the probability of overflow decays like $\exp(-NI)$. This is also true in a network, as long as the buffer sizes and service rates and levels of multiplexing are scaled up in proportion everywhere.

Putting these together, a very rough heuristic is that the probability of overflow decays like $\exp(-Nq^{2(1-H)}I)$ (where q is now the buffer per flow).

Another way to use scaling laws is in the setting of moderate deviations. A typical traffic process has fluctuations of many sizes, and larger fluctuations have lower frequency. In moderate deviations theory, the relationship between size and frequency is codified by a parameter $\beta \in (0,1)$. In Chapter 9 we considered systems whose various parameters were scaled by different β, and this led to strikingly simple results.

Estimation. A further application of scaling laws was mentioned in Section 1.1. If the queue length distribution has an exponential tail, we can plot the distribution for small values of q and exponentially extrapolate to find the probability of large values of q. Or we could run a 'shadow' system with a fraction α of the input flows, service rate and buffer size, measure the shadow probability of overflow, and exponentially extrapolate to find the true (hopefully small) probability of overflow.

Here is a more involved example, worked out in detail.

Example 10.6
In the moderate-deviations many-flows scaling, the basic large deviations

result is that

$$\frac{1}{N^\beta}\log P(Q^N > N^{(1+\beta)/2}q) \to -I$$

where Q^N is the queue size in a queue fed by N independent identically distributed sources, with total mean arrival rate $N\mu$ and service rate $N\mu + N^{(1+\beta)/2}C$. The value of I will depend on the parameters q and C (and of course on the distribution of the arrival process). Let p be the probability in question.

Now consider a rescaled system, in which N is replaced by kN: the number of sources is multiplied by k, and the excess service rate by $k^{(1+\beta)/2}$; and we are interested in p', the probability that the queue length exceeds a level which is $k^{(1+\beta)/2}$ times larger than before. Then

$$p' \approx p^{-k^\beta}.$$

This could be used as a basis for estimating the probability of overflow in a larger system. Note that this calculation does not involve the form of the rate function I. ◇

10.5 Types of Traffic

We will now describe some popular traffic models that have been studied using large deviations theory, paying particular attention to the form of the global approximation

$$I(q) = \inf_{t\geq 0}\sup_{\theta\geq 0}\ \theta(q+Ct) - \log Ee^{\theta A(-t,0]}.$$

There is a vast literature on traffic modelling, and this is only a tiny smattering.

10.5.1 Markov Jump Processes

The simplest Markov jump process is a Poisson process. If A is a Poisson process of rate λ, and A^N is defined by

$$A^N(-t,0] = \frac{1}{N}A(-Nt,0]$$

(or equivalently A^N is the average of N independent copies of A), then A^N satisfies a sample path LDP with linear geodesics, with good rate function

$$I(a) = \begin{cases} \int_{t=-\infty}^{0} \Lambda^*(\dot a_t)\,dt & \text{if } a \text{ is absolutely continuous} \\ \infty & \text{otherwise} \end{cases}$$

where the instantaneous rate function Λ^* is

$$\Lambda^*(x) = \sup_{\theta \in \mathbb{R}} \theta x - \lambda(e^\theta - 1).$$

The global approximation is

$$\begin{aligned} I(q) &= q \ \sup\{\theta > 0 : \lambda(e^\theta - 1) < \theta C\} \\ &= q\bigl(-\rho - \operatorname{plog}(-\rho e^{-\rho})\bigr) \quad \text{where } \rho = \lambda/C. \end{aligned}$$

The first expression comes from a simplification of I which holds whenever A has independent increments, and which was proved in Lemma 3.4. The second expression involves the product logarithm function $\operatorname{plog}(x)$, the solution to $\xi \exp(\xi) = x$. (In fact, if $x < -1/e$ there are no solutions; if $-1/e < x < 0$ there are two solutions, of which we want the smaller; if $x > 0$ there is one solution.)

In higher dimensions, one may want to work with

$$A(-t,0] = \sum_{i=1}^n N_i(-t,0] r_i$$

where the r_i are vectors in $\mathbb{Z}^d$ and the N_i are Poisson processes with rates λ_i. The instantaneous rate function is now

$$\Lambda^*(x) = \sup_{\theta \in \mathbb{R}^d} \theta \cdot x - \sum_{i=1}^n \lambda_i (e^{\theta \cdot r_i} - 1).$$

A natural generalization is to allow the rates λ_i to depend on the state of the system. For example, q_t might represent the current queue size, and the service process might be a Poisson process with a rate which depends on q_t. The instantaneous rate function is then

$$\Lambda^*(x;q) = \sup_{\theta \in \mathbb{R}^d} \theta \cdot x - \sum_{i=1}^n \lambda_i(q) (e^{\theta \cdot r_i} - 1)$$

and the sample path rate function (for absolutely continuous a) is

$$I(a) = \int_{t=-\infty}^0 \Lambda^*\bigl(\dot{a}_t; q_t(a)\bigr)\, dt.$$

There are very many interesting systems which can be modelled in this way, such as circuit-switched networks. The book of Shwartz and Weiss [91] gives many applications, and this is also the setting of work by Dupuis and Ellis [35, 36].

10.5.2 Markov Additive Sources

Let X_t be an irreducible aperiodic Markov chain on a finite state space (with $t \in \mathbb{Z}$). Let $Y_t = f(X_t)$ for some function f, and let $A(-t, 0] = Y_1 + \cdots + Y_t$. This is a Markov additive source. The continuous-time analogue is called a Markov modulated fluid source.

Botvich and Duffield [9] show that, in general and not just for Markov additive sources, if the limit

$$\alpha = - \lim_{t \to \infty} \Lambda_t(\beta) - \beta C t$$

exists and is finite, and certain technical conditions are also satisfied, then the rate function for a queue with service rate C has the form, for large q,

$$I(q) = \alpha + \beta q + o(1).$$

They show that Markov additive sources satisfy these conditions.

Of course the limit α can only exist if

$$\Lambda(\theta) = \lim_{t \to \infty} \frac{1}{t} \Lambda_t(\theta).$$

Then it can be shown that β must be equal to

$$\inf_{t > 0} t \Lambda^*(c + 1/t).$$

Compare this result to the large-buffer limit in Theorem 3.1.

10.5.3 On-off Sources

An on/off source alternates between the active state and the silent state. While active, it produces work at constant rate p. While silent, it produces no work. It remains active for a duration distributed like T and silent for a duration distributed like U, and all durations are independent. The mean arrival rate is

$$\rho = p \frac{ET}{ET + EU}$$

In order that $I(q)$ not be trivial, suppose the service rate C lies in (ρ, p).

Mandjes and Kim [69] have found the form of $I(q)$ for small q. Assume that ET and EU are finite, and that $T + U$ is non-lattice. They show that

$$I(q) = \alpha + \beta \sqrt{q} + O(q) \tag{10.12}$$

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*,

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

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

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

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

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

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 constant rate p. This is closely related to the on-off model. Indeed, Mandjes and Kim [69] show that (10.12) is still true, though the constants are of course different; and Duffield [31] shows that (10.13) is true.

See also the example in Section 8.5 of a Gaussian approximation to the $M/G/\infty$ system, due to Mandjes [67].

10.5.5 Cell-Burst Sources

A periodic source emits one unit of work every time unit, so that the amount of work generated in an interval $[0,t]$ is

$$\lfloor t\rfloor + 1_{t-\lfloor t\rfloor > P}$$

where P is the phase, a uniform random variable in $[0,1]$. A cell-burst source is like an on-off source, except that while in the active state it behaves like a periodic source (the phase being constant from one active period to another). Let T and U be as before, except that now we will take them to be integer. Assume that they have finite expectation, and let $\rho = ET/(ET+EU)$. This has been used as a model for digital voice transmission: while a person is speaking (the burst), his voice is digitized and sent in periodic packets (cells).

Mandjes and Kim [65] have investigated the form of $I(q)$ for this source model. They show that, for service rate $c > \rho$, there exists a critical queue size q_{crit} such that for $q < q_{\text{crit}}$, $I(q)$ is exactly equal to the rate function for queue size in a queue fed by a pure periodic source and served at rate c/ρ. Furthermore, the most likely time to overflow is less than $\frac{1}{2}$. They show that for small q,

$$I(q) = \alpha q + O(q^2)$$

where α depends only on c and ρ.

They also find asymptotics for large q, in the case where T and U are geometric. Then

$$I(q) = \alpha + \beta q + o(1)$$

for certain constants α and β.

10.5.6 Gaussian Sources

If the arrival process is Gaussian with mean $EA(-t,0] = \mu t$ and variance $\operatorname{Var} A(-t,0] = V_t$ then the global approximation $I(q)$ simplifies:

$$I(q) = \inf_t \frac{(q+(C-\mu)t)^2}{2V_t}.$$

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

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

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

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

where $Z^{\circlearrowright a}$ denotes the speeded-up process $Z_t^{\circlearrowright a} = Z_{at}$. 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.

Brownian bridge. The Brownian bridge (defined in Example 6.1) is useful in constructing a Gaussian approximation to a periodic source. A periodic source generates an amount of work

$$A(0,t] = \lfloor t \rfloor + 1_{t-\lfloor t \rfloor > P}$$

in the time interval $(0,t]$, where P is the phase, a random variable uniformly distributed on $[0,1]$. This has $EA(0,t] = t$ and variance $\operatorname{Var} A(0,t] = (t - \lfloor t \rfloor)(\lceil t \rceil - t)$. This is exactly the same mean and variance as for the Gaussian process $Z(t) = t + X(t)$ where X is a Brownian bridge. For the source Z,

$$I(q) = 2q(q+c-1),$$

the infimum in the definition of $I(q)$ being attained at

$$t = \frac{q}{2q+c-1} < \frac{1}{2}.$$

Addie et al. [1] point out that the approximation this leads to is in fact exact.

A scaling relationship. Let A^N be the average of N copies of A. This has variance function $\operatorname{Var} A^N(-t,0] = V_t/N$. Equivalently, the process A^ε with variance $\operatorname{Var} A^\varepsilon(-t,0] = \varepsilon V_t$ can be represented as the average of $1/\varepsilon$ copies of A. This means that the many-flows limit results can equally be seen as small-variance limit results.

Other Gaussian models. Further examples are given by Botvich and Duffield [9], Addie et al. [1] and Mannersalo and Norros [71]. Mandjes [67] gives more examples, and shows that (under certain conditions) the rate function $I(q)$ is convex at q if and only if there are negative correlations at the critical timescale for q. See also Example 3.1, a Gaussian autoregressive source.

Bibliography

[1] Ron Addie, Petteri Mannersalo, and Ilkka Norros. Most probable paths and performance formulae for buffers with Gaussian input traffic. *European transactions on telecommunications*, 13, 2002. (Pages 188, 189, 237 and 238.)

[2] V. Anantharam. How large delays build up in a GI/G/1 queue. *Queueing Systems*, 5:345–368, 1989. (Page 123.)

[3] Søren Asmussen. *Applied Probability and Queues.* Springer, 2nd edition, 2003. (Pages 2 and 10.)

[4] Florin Avram, J. G. Dai, and John J. Hasenbein. Explicit solutions for variational problems in the quadrant. *Queueing Systems*, 37:261–291, 2001. (Page 95.)

[5] P. Báartfai. Large deviations in the queueing theory. *Periodica Mathematica Hungarica*, 2:165–172, 1972. (Page 48.)

[6] Arthur W. Berger and Ward Whitt. Effective bandwidths with priorities. *IEEE/ACM Transactions on Networking*, 6(4), August 1998. (Page 177.)

[7] Dimitris Bertsimas, Ioannis Ch. Paschalidis, and John N. Tsitsiklis. On the large deviations behaviour of acyclic networks of G/G/1 queues. *Annals of Applied Probability*, 8:1027–1069, 1998. (Page 130.)

[8] A. A. Borovkov and A. A. Mogulskii. Large deviations for stationary Markov chains in a quarter plane. In S. Watanabe, M. Fukushima, Yu. V. Prohorov, and A. N. Shiryaev, editors, *Probability Theory and Mathematical Statistics*, pages 12–19. World Scientific Publishing, Tokyo, 1995. (Page 95.)

A. Ganesh, N. O'Connell, and D. Wischik: LNM 1838, pp. 239–248, 2004.

[9] D. D. Botvich and N. G. Duffield. Large deviations, economies of scale, and the shape of the loss curve in large multiplexers. *Queueing systems*, 20, 1995. (Pages 17, 234 and 238.)

[10] D.D. Botvich and N.G. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems*, 20:293–320, 1995. (Page 53.)

[11] Pierre Bremaud and Francois Baccelli. *Elements of Queueing Theory*. Springer, 2nd edition, 2003. (Page 2.)

[12] Cheng-Shang Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39:913–931, 1994. (Page 48.)

[13] Cheng-Shang Chang. *Performance Guarantees in Communication Networks*. Springer, 2000. (Page VII.)

[14] Cheng-Shang Chang, David D. Yao, and Tim Zajic. Large deviations, moderate deviations, and queues with long-range dependent input. *Advances in Applied Probability*, 31(1):254–278, 1999. (Page 189.)

[15] Cheng-Shang Chang and Tim Zajic. Effective bandwidths of departure processes from queues with time varying capacities. In *Proceedings of IEEE Infocom*, pages 1001–1009, 1995. (Page 137.)

[16] Jinwoo Choe and Ness B. Shroff. A central limit theorem based approach for analyzing queue behaviour in ATM networks. *IEEE/ACM Transactions on Networking*, 6(5):659–671, 1998. (Page 221.)

[17] Jinwoo Choe and Ness B. Shroff. Use of the supremum distribution of Gaussian processes in queueing analysis with long-range dependence and self-similarity. *Stochastic Models*, 16:209–231, 2000. (Page 221.)

[18] Gagan L. Choudhury, David M. Lucantoni, and Ward Whitt. On the effectiveness of effective bandwidths for admission control in ATM networks. In *Proceedings of the 14th International Teletraffic Congress — ITC 14*, pages 411–420. Elsevier Science, 1994. (Page 221.)

[19] 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 Transactions on Communications*, 43:1778–1784, 1995. (Pages 6 and 148.)

[20] Costas Courcoubetis, Vasilios A. Siris, and George D. Stamoulis. Application of the many sources asymptotic and effective bandwidths to traffic engineering. *Telecommunication Systems*, 12:167–191, 1999. (Page 216.)

[21] Costas Courcoubetis and Richard Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33:886–903, 1996. (Page 17.)

[22] G. de Veciana and G. Kesidis. Bandwidths allocation for multiple qualities of service using generalised processor sharing. *IEEE Transactions on Information Theory*, 42:268–272, 1996. (Page 95.)

[23] Sylvain Delas, Ravi Mazumdar, and Catherine Rosenberg. Tail asymptotics for HOL priority queues handling a large number of independent stationary sources. *Queueing Systems*, 40:183–204, 2002. (Page 177.)

[24] Amir Dembo and Tim Zajic. Large deviations: From empirical mean and measure to partial sums process. *Stochastic Processes and their Applications*, 57:191–224, 1995. (Page 109.)

[25] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Springer, 2nd edition, 1998. (Pages 23, 36, 37, 42, 63, 68, 69, 70, 71, 75, 119 and 220.)

[26] F. den Hollander. *Large Deviations.* Fields Institute Monographs. American Mathematical Society, 2000. (Page 23.)

[27] Chandrakant M. Deo and Gutti Jogesh Babu. Probabilities of moderate deviations in Banach space. *Proceedings of the AMS*, 1981. (Page 201.)

[28] Jean-Dominique Deuschel and Daniel W. Stroock. *Large deviations.* Academic Press, 1989. (Pages 23 and 188.)

[29] Persi Diaconis and David Freedman. On the uniform consistency of Bayes estimates for multinomial probabilities. *Annals of Statistics*, 18: 1317–1327, 1990. (Page 150.)

[30] R. L. Dobrushin and E. A. Pechersky. Large deviations for random processes with independent increments on infinite intervals. In *Probability theory and mathematical statistics (St. Petersburg, 1993)*, pages 41–74. Gordon and Breach, Amsterdam, 1996. (Page 110.)

[31] N. G. Duffield. Queueing at large resources driven by long-tailed $M/G/\infty$-modulated processes. *Queueing Systems*, 28, 1998. (Page 236.)

[32] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russell, and F. Toomey. Entropy of ATM traffic streams: a tool for estimating QoS parameters. *IEEE Journal of Selected Areas in Communications*, 13:981–990, 1995. (Pages 147 and 148.)

[33] N. G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Mathematical Proceedings of the Cambridge Philosophical Society*, 118:363–374, 1995. (Page 185.)

[34] Ken Duffy, John T. Lewis, and Wayne G. Sullivan. Logarithmic asymptotics for the supremum of a stochastic process. *Annals of Applied Probability*, 13:430–445, 2003. (Page 185.)

[35] Paul Dupuis and Richard S. Ellis. Large deviation analysis of queueing systems. In F. P. Kelly and R. Williams, editors, *Proceedings of the IMA Workshop*. Springer-Verlag, February 1994. (Page 233.)

[36] Paul Dupuis and Richard S. Ellis. The large deviation principle for a general class of queueing systems, I. *Transactions of the American Mathematical Society*, 347:2689–2751, 1995. (Page 233.)

[37] Paul Dupuis and Richard S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley, New York, 1997. (Page 61.)

[38] Paul Dupuis and Kavita Ramanan. A Skorokhod problem formulation and large deviation analysis of a processor sharing model. *Queueing Systems*, 28:109–124, 1998. (Pages 95 and 98.)

[39] Peter Eichelsbacher and Ayalvadi Ganesh. Bayesian inference for Markov chains. *Journal of Applied Probability*, 39:91–99, 2002. (Page 150.)

[40] A. Ganesh and N. O'Connell. An inverse of Sanov's theorem. *Statistics and Probability Letters*, 42:201–206, 1999. (Page 150.)

[41] A. Ganesh and Neil O'Connell. A large deviation principle with queueing applications. *Stochastics and Stochastic Reports*, 73(1–2):25–35, 2002. (Pages 110 and 130.)

[42] A. J. Ganesh. Bias correction in effective bandwidth estimation. *Performance Evaluation*, 27–28:319–330, 1996. (Page 148.)

[43] A. J. Ganesh and N. O'Connell. A large deviation principle for Dirichlet posteriors. *Bernoulli*, 6:1021–1034, 2000. (Page 150.)

[44] Ayalvadi Ganesh, Neil O'Connell, and Balaji Prabhakar. Invariant rate functions for discrete-time queues. *Annals of Applied Probability*, 13:446–474, 2003. (Pages 137 and 138.)

[45] R. J. Gibbens. Traffic characterisation and effective bandwidths for broadband network traces. In Kelly et al. [54]. (Page 212.)

[46] R. J. Gibbens and Y. C. Teh. Critical time and space scales for statistical multiplexing in multiservice networks. In P. Key and D. Smith, editors, *Proceedings of the 16th International Teletraffic Congress*. Elsevier Science, 1999. (Page 167.)

[47] Peter Glynn and Assaf Zeevi. Estimating tail probabilities in queues via extremal statistics. In *Analysis of communication networks: call centres, traffic and performance*, volume 28 of *Fields Institute Communications*. American Mathematical Society, 2000. (Page 148.)

[48] Peter W. Glynn and Ward Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability*, 31A:131–156, 1994. Special edition. (Page 48.)

[49] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and Random Processes*. Oxford, 3rd edition, 2001. (Page 2.)

[50] I. A. Ignatyuk, V. Malyshev, and V. V. Scherbakov. The influence of boundaries in problems on large deviations. *Uspekhi Matematicheskikh Nauk*, 49:43–102, 1994. (Page 95.)

[51] Ingemar Kaj. Convergence of scaled renewal processes to fractional Brownian motion. Preprint, 1999. (Page 196.)

[52] F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, 1979. (Pages 2, 137 and 144.)

[53] F. P. Kelly. Effective bandwidths at multi-class queues. *Queueing Systems*, 9:5–15, 1991. (Page 212.)

[54] F. P. Kelly, S. Zachary, and I. Ziedins, editors. *Stochastic Networks: Theory and Applications*. Royal Statistical Society Lecture Note Series. Oxford, 1996. (Pages 243, 244 and 245.)

[55] Frank Kelly. Notes on effective bandwidths. In Kelly et al. [54], chapter 8, pages 141—168. (Pages 212 and 220.)

[56] George Kesidis, Jean Walrand, and Cheng-Shang Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, 1:424–428, 1993. (Page 212.)

[57] V. Klemes. The Hurst phenomenon: a puzzle? *Water Resources Research*, 10:675–688, 1974. (Pages 183 and 197.)

[58] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, 1994. (Page 183.)

[59] N. Likhanov and R. R. Mazumdar. Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *Journal of Applied Probability*, 36:86–96, 1999. (Pages 53 and 219.)

[60] R. M. Loynes. The stability of a queue with non-independent interarrival and service times. *Proceedings of the Cambridge Philosophical Society*, 58, 1962. (Page 4.)

[61] Claudio Macci. Large deviations for posterior distributions concerning exponential statistical models. Preprint, 2002. (Page 150.)

[62] Kurt Majewski. Heavy traffic approximations of large deviations of feedforward queueing networks. *Queueing Systems*, 1998. (Page 199.)

[63] Kurt Majewski. Large deviations of the steady-state distribution of reflected processes with applications to queueing systems. *Queueing Systems*, 29, 1998. (Page 105.)

[64] Kurt Majewski. Large deviations for multi-dimensional reflected fractional Brownian motion. *Stochastics and stochastic reports*, 75(4): 233–257, 2003. (Page 189.)

[65] M. Mandjes and J. H. Kim. An analysis of the phase transition phenomenon in packet networks. *Advances in applied probability*, 33, 2001. (Page 236.)

[66] M. R. H. Mandjes and S. C. Borst. Overflow behavior in queues with many long-tailed inputs. CWI Report PNA-R9911, CWI, 1999. (Page 245.)

[67] Michel Mandjes. A note on the benefits of buffering. *Stochastic Models*, 2003. To appear. (Pages 196, 236 and 238.)

[68] Michel Mandjes and Sem Borst. Overflow behavior in queues with many long-tailed inputs. *Advances in applied probability*, 32, 2000. Preliminary version online in [66]. (Page 235.)

[69] Michel Mandjes and Jeong Han Kim. Large deviations for small buffers: An insensitivity result. *Queueing systems*, 37, 2001. (Pages 53, 234 and 236.)

[70] Michel Mandjes and Miranda van Uitert. Sample-path large deviations for tandem and priority queues with Gaussian inputs. Technical Report PNA-R0221, CWI, 2002. (Page 177.)

[71] Petteri Mannersalo and Ilkka Norros. A most probable path approach to queueing systems with general Gaussian input. *Computer networks*, 40, 2002. (Page 238.)

[72] Laurent Massoulie. Large deviations estimates for polling and weighted fair queueing service systems. *Advances in Performance Analysis*, 2, 1999. (Page 95.)

[73] Laurent Massoulie and Alain Simonian. Large buffer asymptotics for the queue with fractional Brownian input. *Journal of Applied Probability*, 36(3):894–906, 1999. (Page 186.)

[74] R. R. Muntz. Poisson departure processes and queueing networks. *IBM Research Report RC 4145*, 1972. (Page 137.)

[75] I. Norros. On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications*, 13(6):953–962, 1995. (Page 186.)

[76] Neil O'Connell. Queue lengths and departures at single-server resources. In Kelly et al. [54], chapter 5. (Page 105.)

[77] Neil O'Connell. Large deviations for departures from a shared buffer. *Journal of Applied Probability*, 34:753–766, 1997. (Page 105.)

[78] Neil O'Connell. Large deviations for queue lengths at a multi-buffered resource. *Journal of Applied Probability*, 35:240–245, 1998. (Pages 95 and 105.)

[79] Neil O'Connell and Gregorio Procissi. On the build-up of large queues in a queue with fBm input. Technical Report HPL-BRIMS-9818, BRIMS, Hewlett-Packard Laboratories, 1998. (Page 189.)

[80] F. Papangelou. Large deviations and the Bayesian estimation of higher-order markov transition functions. *Journal of Applied Probability*, 33:18–27, 1996. (Page 150.)

[81] I. Ch. Paschalidis and S. Vassilaras. On the estimation of buffer overflow probabilities from measurements. *IEEE Transactions on Information Theory*, 47:178–191, 2001. (Page 150.)

[82] V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 1995. (Page 224.)

[83] S. M. Pitts, R. Grübel, and P. Embrechts. Confidence bounds for the adjustment coefficient. *Advances in Applied Probability*, 28:802–827, 1996. (Page 147.)

[84] A. A. Puhalskii and W. Whitt. Functional large deviation principles for waiting and departure processes. *Probability in the Engineering and Informational Sciences*, pages 479–7507, 1998. (Page 137.)

[85] Anatolii A. Puhalskii. Moderate deviations for queues in critical loading. *Queueing Systems*, 31:359–392, 1999. (Page 199.)

[86] Peter Rabinovitch. Statistical estimation of effective bandwidth. Master's thesis, Carleton University, 2000. (Pages 148 and 216.)

[87] Kavita Ramanan and Paul Dupuis. Large deviation properties of data streams that share a buffer. *Annals of Applied Probability*, 8(4), 1998. (Page 137.)

[88] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970. (Page 30.)

[89] Raymond Russell. *The Large Deviations of Random Time-Changes*. PhD thesis, Trinity College, Dublin, 1998. (Page 51.)

[90] S. Shakkottai and R. Srikant. Many-sources delay asymptotics with applications to priority queues. *Queueing Systems*, 39:183–200, 2001. (Page 177.)

[91] Adam Shwartz and Alan Weiss. *Large Deviations for Performance Analysis.* Chapman and Hall, 1995. (Pages VII and 233.)

[92] Alain Simonian and Jacky Guibert. Large deviations approximation for fluid queues fed by a large number of on/off sources. *IEEE Journal on selected areas in communication*, 13, 1995. (Page 17.)

[93] Fergal Toomey. Private communication, 1995. (Page 96.)

[94] Fergal Toomey. Bursty traffic and finite capacity queues. *Annals of Operations Research*, 79:45–62, 1998. (Pages 88 and 126.)

[95] J. Walrand. *An Introduction to Queueing Networks.* Prentice-Hall, New Jersey, 1988. (Page 137.)

[96] Amy R. Ward and Peter W. Glynn. A diffusion approximation for a Markovian queue with reneging. *Queueing systems*, 2002. (Page 227.)

[97] A. Weiss. A new technique for analyzing large traffic systems. *Advances in Applied Probability*, 18:506–532, 1986. (Page 17.)

[98] Ward Whitt. *Stochastic-process limits.* Springer, 2001. (Pages 10, 81 and 124.)

[99] W. Willinger, V. Paxson, and M. S. Taqqu. Self-similarity and heavy tails: structural modelling of network traffic. In R. Adler, R. Feldman, and M. S. Taqqu, editors, *A practical guide to heavy tails: statistical techniques for analysing heavy tailed distributions.* Birkhauser, 1998. (Page 196.)

[100] Damon Wischik. The output of a switch, or, effective bandwidths for networks. *Queueing Systems*, 32:383–396, 1999. (Page 181.)

[101] Damon Wischik. Sample path large deviations for queues with many inputs. *Annals of Applied Probability*, 11:379–404, 2001. (Pages 155 and 177.)

[102] Damon Wischik. Moderate deviations in queueing theory. Preprint, 2004. (Pages 202 and 208.)

[103] A. P. Zwart. *Queueing Systems with Heavy Tails.* PhD thesis, Eindhoven University of Technology, 2001. (Page 10.)

See also the active bibliography at `www.bigqueues.com`

Index of Notation

$a \wedge b$	$\min(a,b)$
$a \vee b$	$\max(a,b)$
$[a]^+$	$\max(a,0)$
$[a]_b^c$	$\min(\max(a,b),c)$
B°	interior of B
$\bar{B}$	closure of B
$\mathbb{N}$	$\{1,2,\ldots\}$
$\mathbb{N}_0$	$\{0,1,2,\ldots\}$
$\mathbb{R}_0^+$	$\{x \in \mathbb{R} : x \geq 0\}$
$\mathbb{R}^+$	$\{x \in \mathbb{R} : x > 0\}$
$\mathbb{R}^*$	$\mathbb{R} \cup \{+\infty\}$

Index