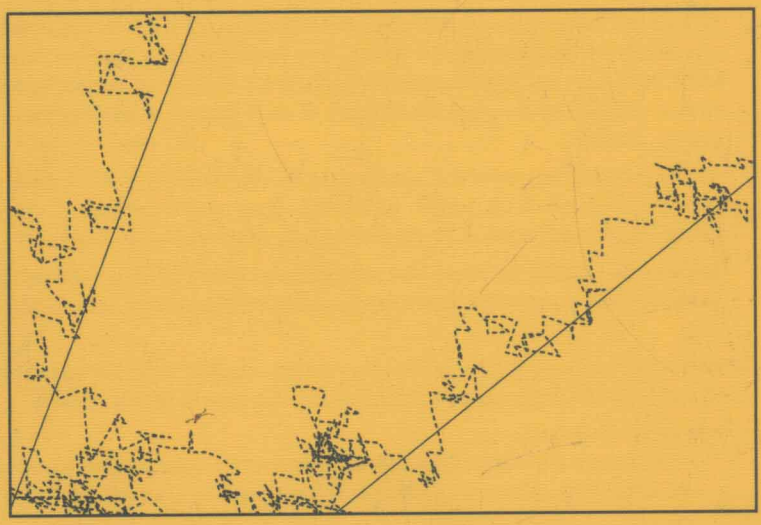
Ayalvadi Ganesh
Neil O'Connell
Damon Wischik

Big Queues

Lecture Notes in Mathematics
1838



Springer

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.

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

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

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

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

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 $(Q_t, t \ge 0)$ 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 $q \in \mathbb{N}$,

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

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

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

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 $\delta > 0$,

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

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

$$Q_t = \left[Q_{t-1} + A_t - C\right]_0^B$$

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 distribu-

tion 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

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

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,

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

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.