Logarithmic asymptotics in Queueing Theory and Risk Theory

by

Ken Duffy

A thesis submitted to
the University of Dublin
for the degree of

Doctor of Philosophy

School of Mathematics,
University of Dublin, Trinity College

September, 2000

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Ken Duffy.

Signature of Author ......Ken R. Duffy......

Ken Duffy
4 September, 2000

Summary

This thesis addresses four distinct, but related, problems. All four involve large deviation theory.

The first problem is to relate the logarithmic asymptotics of the single server queue length distribution to the long time-scale large deviations of the stochastic driving force. Sufficient conditions are presented under which this relationship has a simple form, even in the presence of nonlinear scales. Under the hypothesis that the service rate is constant, the logarithmic asymptotics of the queue length distribution are related to the logarithmic asymptotics of the waiting time distribution. Statistical multiplexing is investigated in the presence of nonlinear scales, and many-scale behavior is demonstrated: the scale on which large deviations are observed depends upon the size of the deviation.

The second problem relates the logarithmic asymptotics of the probability of exhaustion of a resource to the large deviations of a growing number of sources using the resource. We consider only a finite time interval: this enables us to tackle sources which are non-stationary. Nonlinear scales are covered in this treatment and, again, many-scale behavior is demonstrated.

The third problem deals with constructing a stochastic model which exhibits long range dependence. The proposals in the literature involve Gaussian processes. A class of two state processes exhibiting long range dependence is constructed. A relationship is presented between the large deviations of their sojourn times and the large deviations of the sources themselves. An explicit form of the rate-function is found in the case where the sojourn times have a semi-exponential distribution.

The fourth problem is to construct distribution-free confidence intervals for measurement of effective bandwidths. This treatment includes use of a first order auto-regressive filter to cope with non-stationarity.

Acknowledgements

It is a pleasure to thank my supervisor, John T. Lewis. John has given me the greatest gift I have received as a mathematician: confidence, I would not have done the work in this thesis without it. John works tirelessly on behalf of the Dublin Applied Probability Group, and for this too I owe him many thanks. As my good friend, Fergal Toomey, once said:

> *"My most ambitious hope for this document is that the mathematics contained in it might justify, to some small degree, all that John has done on my behalf."*

I must also thank the rest of my collaborators: Fergal Toomey with whom I have worked since the day I arrived in DIAS, Wayne Sullivan for unerring accuracy, Laslo Györfi, Andras Rácz, and the members of the DAPG : Séan Coffey, Mark Corluy, Ian Dowse, Mark Dukes, Eliza Heron, Brian McGurk, Raymond Russell and Cormac Walsh, for many stimulating discussions.

I would like to thank the Trinity College Dublin school of mathematics (David Simms and Siddhartha Sen in particular) for providing me with an education I only now appreciate. My undergraduate years would have been much more troublesome, educationally, if it were not for my friendship with David Malone. I would also like to thank Cíaran O'Sullivan for proving to me that teachers can be there because they love their subject. I would never have been involved in maths at all if it were not for my brother, Gavin, who is a far better mathematician than me.

I have too many friends, I cannot thank them all individually. However, I should mention two: Alister Kidd has been my friend for longer than I care to remember, I do hope it stays that way, and Josephine Hughes, to whom mathematics comes a sorry second.

I will never be able to properly express my respect and love for my parents, and my best friends, Ken and Mary Duffy.

My work in this thesis was supported by a grant from Microsoft Research, Cambridge.

Contents

Preface

The bulk of the material presented here is my own work and is, to the best of my knowledge, novel. I have, of course, presented some background material by way of introduction and in general I use standard terminology and notation. In order to distinguish my own work from the background material, I have used the convention that, unless I specify otherwise, any definitions, lemmas, theorems and proofs that are formally presented are original. In a few cases, I have given my own proof of a standard result or have modified an existing proof to suit my needs; I have clearly identified those cases where they occur in the text.

This thesis is for Ken and Mary.

Chapter 1

Introduction

Tempus Omnia Revelat.

1.1 On This Thesis

This thesis tackles four problems using large deviation theory, and probabilistic estimates in general. Although the problems are related, they differ in many ways. Hence it is natural to present each problem on its own, in its own chapter. This is what we have done. Chapters two and three use an example whose construction can be found in chapter four. We provide a general introduction, and motivation, to the subject matter of this thesis: Queueing Theory and Risk Theory.

The work in chapter two was done in collaboration with J. T. Lewis (DIAS) and W. Sullivan (UCD and DIAS), and is to be submitted for publication in the near future. The work in chapter three was done in collaboration with F. Toomey (DIAS), and has been submitted to the Journal of Applied Probability. The work in chapter four was done on my own, and is to be submitted for publication in the near future. The work in chapter five was done in collaboration with L. Györfi (Technical University of Budapest), J. T. Lewis, A. Rácz (Technical University of Budapest) and F. Toomey, and is to appear in the Journal of Applied Probability **37**:1 (2000).

We shall discuss the subject matter of this thesis in terms of its interpretation in Queueing Theory; in the section on Risk Theory, we shall point out the relation between questions about queueing behavior and questions about the Cramér-Lundberg model of Risk Theory.

1.2 Queueing Theory

1.2.1 An Historical Perspective

Queueing theory has been of interest to mathematicians and engineers since the turn of the twentieth century. Up until the middle of the last century, the focus of work had primarily been on application rather than theory. This all changed in 1952 when Lindley published his general treatment of the single server queue [46].

Lindley was the first to publish the recursion relation governing the evolution of the waiting times seen by customers at a single server queue. Let σ_n be the time taken to serve customer n, and let θ_n be the inter-arrival time between customer n and $n+1$. Then w_n, the time customer n spends waiting in the queue, evolves according to

$$w_{n+1} = (w_n + \sigma_n - \theta_n) \vee 0 \qquad \text{for } n \geq 0, \tag{1.1}$$

where $a \vee b$ is the maximum of a and b. Lindley considered i.i.d. service times and independent, i.i.d, inter-arrival times, and proved that a necessary and sufficient condition for the waiting time distribution to have a non-degenerate limit is that either $\mathbb{E}[\sigma_1 - \theta_1] < 0$, or $\sigma_n = \theta_n$ for all $n \geq 0$.

In 1962 Loynes investigated Lindley's equation (1.1) under very general hypotheses. He assumed the stochastic driving forces, σ_n and θ_n, to be stationary, almost surely finite sequences and proved two remarkable theorems. The first concerns the existence of a minimal solution to the Lindley's recursion (which happens to be stationary) and gives a form for the stationary distribution; the second concerns the stability of the queue and the importance of the stationary solution found in Theorem 1.1. Proofs of Theorems 1.1 and 1.2 can be found in Appendix A.

Theorem 1.1 *If $\{\sigma_n - \theta_n\}$ is a stationary sequence, then there exists a stationary sequence of random variables, Υ_n, which satisfy Lindley's equation (1.1). Each Υ_n is equal in distribution to the random variable Θ, defined by*

$$\Theta := \sup_{n \geq 0} \left(\sum_{i=-n}^{0} (\sigma_i - \theta_i) \right) \vee 0. \tag{1.2}$$

Theorem 1.2 *If the stability condition,*

$$\mathbb{P}\left[\limsup_{n\to\infty}\left(\sum_{i=-n}^{0}(\sigma_i-\theta_i)\right)<\infty\right]=1, \tag{1.3}$$

holds, then every solution of Lindley's equation (1.1) couples to the one defined in Theorem 1.1, after an almost surely finite time.

The Loynes stability condition (1.3) ensures that, no matter what the starting condition may be, the waiting time becomes zero after an almost surely finite number of customers have passed through the queue. If two solutions of Lindley's recursion have different starting conditions, then coupling occurs once the greater of the two solutions is zero. Hence, under the Loynes hypotheses, this will happen after an almost surely finite number of customers have passed. Loynes noted that if, along with his stationarity assumptions, $\sigma_i - \theta_i$ is ergodic, then the queue is stable if $\mathbb{E}[\sigma_i - \theta_i] < 0$. This has the following interpretation: under the assumptions of stationarity and ergodicity, the queue is stable if, on average, customers arrive at longer intervals than the time it takes them to be served.

There have been several attempts to justify (1.2) as the solution to the evolution of a single server queue where the parameter space is $\mathbb{R}^+$ rather than $\mathbb{Z}^+$. For a general reference, see the book of Harrison [34], and for an older reference see the paper of Kingman [39].

These days, most work in queueing theory revolves around the queue length rather than waiting times. Clearly the evolution of the queue length at a single server queue, in discrete time, can be defined in an identical fashion to that of the waiting times. Let S_t be the number of customers the queue can serve at time t, and let A_t be the number of customers which arrive at time t. Then q_t, the queue length at time t, evolves according to

$$q_{t+1} = (q_t + A_t - S_t) \vee 0 \qquad \text{for } t \geq 0.$$

With the workload process W_t defined by

$$W_t := \sum_{i=-t}^{-1}(A_i - S_i) \qquad \text{for } t \geq 1,$$

and $W_0 := 0$, the Loynes solution of Lindley's equation is equal in distribution to the random variable Q defined by

$$Q := \sup_{t \geq 0} W_t. \tag{1.4}$$

1.2.2 Long Time-Scale Asymptotic

A question that has become increasingly important with the introduction of high speed communication networks is the behavior of the tail of the distribution of Q. If the tail of Q is large, then buffers in the network will overflow and quality of service cannot be guaranteed. Connection admission control is a system whereby quality of service is guaranteed by restricting admission to the network.

The most successful approach, to date, has been the use of Large Deviation Theory to calculate the exponential rate of decay of the probability of the event $\{Q > q\}$ as q becomes large. In Glynn and Whitt's pioneering paper [28], they showed that if a technical set of conditions, including the Gartner-Ellis conditions [11], are satisfied, then

$$\mathbb{P}[Q > q] \asymp e^{-q\delta}$$

and δ can be calculated from the large deviation rate-function of the stochastic driving force $A_t - S_t$. [We use the suggestive notation

$$\mathbb{P}[Q > q] \asymp e^{-q\delta} \tag{1.5}$$

as shorthand for the statement that the limit

$$\lim_{q \to \infty} \frac{1}{q} \log \mathbb{P}[Q > q]$$

exists and is equal to $-\delta$; it captures the notion of asymptotically exponential decay.]

We have the following heuristic for Glynn and Whitt's result: assume that $Z_t := A_t - S_t$ is stationary and ergodic, and $\mathbb{E}[Z_t] < 0$ so that the Loynes stability condition is satisfied. Furthermore, assume that W_t satisfies a large deviation principle. [Roughly speaking this means that there exists a non-negative function, the rate-function, $I(\cdot) : \mathbb{R} \to [0, \infty]$, such that for all $x \geq \mathbb{E}[Z_t]$,

$$\mathbb{P}[W_t > xt] \asymp \exp\left(-tI(x)\right).] \tag{1.6}$$

From the definition (1.4) of Q we have that

$$\{Q > q\} = \{\sup_{t \geq 0} W_t > q\} = \{\sup_{c > 0} W_{cq} > q\} = \bigcup_{c>0} \{W_{cq} > q\}. \tag{1.7}$$

At time zero, $Q > q$ if and only if for some cq the workload W_{cq}, produced between time $-cq$ and time zero, exceeds q.

Consider fixed $c > 0$; what can we say about the probability that W_{cq} is greater than q? As $c > 0$ and $\mathbb{E}[Z_t] < 0$, equation (1.6) tells us that

$$\mathbb{P}[W_{cq} > q] = \mathbb{P}\left[W_{cq} > cq\left(\frac{1}{c}\right)\right] \asymp \exp\left(-cqI\left(\frac{1}{c}\right)\right). \tag{1.8}$$

Part of the special character of Large Deviation Theory is the principle of the largest term (see Lemma 2.3 of [43]); this principle says that the large deviations rate of decay of the probability of a finite union of not necessarily disjoint sets, is the smallest rate of decay of each of the individual sets. If we could justify using the principle of the largest term for a countable union of sets, then equation (1.8) would yield

$$\mathbb{P}[\bigcup_{c>0}\{W_{cq} > q\}] \asymp \exp\left(-q \inf_{c>0} cI\left(\frac{1}{c}\right)\right). \tag{1.9}$$

Using equations (1.7) and (1.9), we have Glynn and Whitt's result,

$$\mathbb{P}[Q > q] \asymp \exp\left(-q \inf_{c>0} cI\left(\frac{1}{c}\right)\right). \tag{1.10}$$

The biggest gap in turning this heuristic into a proof is the justification of use of the principle of the largest term in equation (1.9). The principle of the largest term applies only to finite unions; the following argument explains how we can use (1.8) to get (1.10).

Let c_* be the c at which $\inf_{c>0} cI\left(\frac{1}{c}\right)$ is attained. c_*q is the most likely scaled time in the past at which the queue was last empty. Given $0 < \overline{c} < c_* < \underline{c} < \infty$, as $\{\sup_{t\geq 0} W_t > q\} = \cup_{t\geq 0}\{W_t > q\}$,

$$\{Q > q\} = \bigcup_{t\geq 0}\{W_t > q\} = \bigcup_{t<\overline{c}q}\{W_t > q\} \bigcup_{\underline{c}q\geq t\geq \overline{c}q}\{W_t > q\} \bigcup_{t>\underline{c}q}\{W_t > q\}. \tag{1.11}$$

The principle of the largest term tells us that the asymptotic rate of decay of $\mathbb{P}[Q > q]$ equals the minimum rate of decay of the probability of

$$\bigcup_{t<\overline{c}q}\{W_t > q\}\ ,\quad \bigcup_{\underline{c}q\geq t\geq \overline{c}q}\{W_t > q\}, \text{ and } \bigcup_{t>\underline{c}q}\{W_t > q\}. \tag{1.12}$$

It will be shown in Chapter 2 that for any $0 < \overline{c} < \underline{c} < \infty$ the rate of decay of the probability of the middle term in (1.12) is governed by the large deviation principle,

$$\mathbb{P}\left[\bigcup_{q\underline{c}\geq t\geq \overline{c}q}\{W_t > q\}\right] = \mathbb{P}\left[\bigcup_{\underline{c}\geq c\geq \overline{c}}\{W_{cq} > q\}\right] \asymp \exp\left(-q\inf_{c>0} cI\left(\frac{1}{c}\right)\right).$$

It will also be shown in Chapter 2 that there exists a $0 < \underline{c} < \infty$ such that the rate of decay of the final term, $\mathbb{P}[\bigcup_{t>\underline{c}q}\{W_t > q\}]$, is at least as great as that of the middle term: $\mathbb{P}[\bigcup_{\underline{c}\geq c\geq \overline{c}}\{W_{cq} > q\}]$. To get Glynn and Whitt's result, we must impose conditions that ensure that the contribution from the first term does not dominate. The first and third terms in (1.12) have physical interpretations. Define the random variables

$$\tau_1(q) := \inf\{t : W_t > q\} \qquad \text{and} \qquad \tau_2(q) := \sup\{t : W_t > q\}; \tag{1.13}$$

$\tau_1(q)$ is the first time in the past at which the workload amounted to a level greater than q and $\tau_2(q)$ is the last time in the past at which the workload amounted to a level greater than q. Note that, given $\overline{c}$ and $\underline{c}$,

$$\{\tau_1(q) < \overline{c}q\} = \bigcup_{0\leq t<\overline{c}q}\{W_t > q\}, \qquad \text{and} \qquad \{\tau_2(q) > \underline{c}q\} = \bigcup_{t>\underline{c}q}\{W_t > q\}. \tag{1.14}$$

Thus equation (1.11) is equal to

$$\{Q > q\} = \{\tau_1(q) < q\overline{c}\} \bigcup_{\underline{c}\geq c\geq \overline{c}}\{W_{cq} > q\} \bigcup \{\tau_2(q) > q\underline{c}\}.$$

Hence the first term fails to dominate if the probability that the workload is greater than q does not grow too quickly.

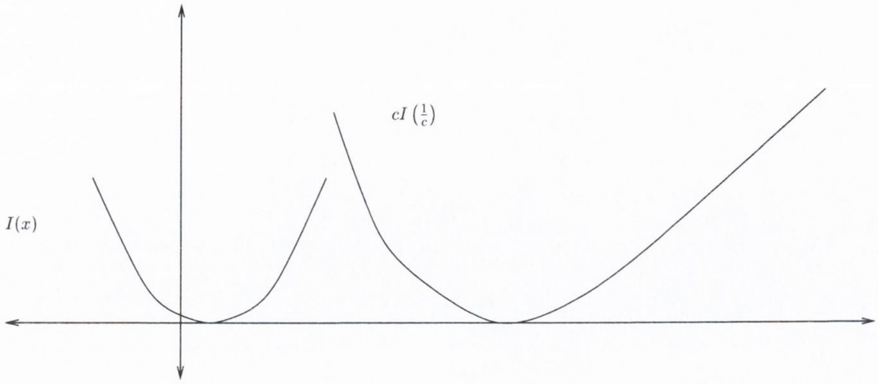

Figure 1-1: Rate-function transformation when the queue is unstable.

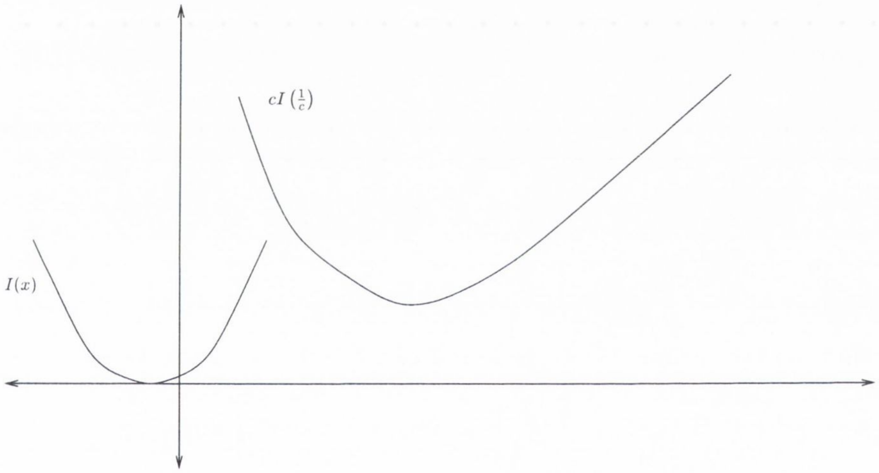

Figure 1-2: Rate-function transformation when the queue is stable.

If the queue is not stable, that is $\mathbb{E}[Z_1] > 0$, then $cI\left(\frac{1}{c}\right) = 0$ when $c = 1/\mathbb{E}[Z_1]$. Hence, if the queue is not stable, the tail of the queue length distribution does not decay. See figure 1-1 for a graphical representation of this rate-function transformation.

If the queue is stable, that is $\mathbb{E}[Z_1] < 0$, and $I(x)$ is strictly convex (which the Gartner-Ellis conditions ensure), then, as $I(x) = 0$ for $x = \mathbb{E}[Z_1]$, $cI\left(\frac{1}{c}\right)$ is positive for all $c > 0$. Thus the rate of decay of $\mathbb{P}[Q > q]$ is positive. Moreover, as $I(x)$ is convex, $cI\left(\frac{1}{c}\right)$ is convex for $c > 0$. See figure 1-2 for a graphical representation of this transformation.

1.2.3 Large Space-Scale asymptotic

Another question which has become increasingly important to queueing theorists, since the introduction of large capacity networks, is statistical multiplexing. This is where superpositions of bursty traffic streams lead to a smoother, less bursty, multiplex. As the multiplex is smoother it is less likely to cause a large deviation when queued in a single server queue, hence there are economies of scale. Calculating the exponential rate at which this smoothness occurs in the queue length, as the number of sources and service rate tend to infinity, is the subject matter of the large number of lines asymptotic.

This asymptotic was suggested independently by Botvich and Duffield [4], and Courcoubetis and Weber [8], who consider homogeneous and heterogeneous superpositions of sources using a single server queue whose service-rate is growing at the rate at which sources are added. A large deviation principle is assumed, with linear scaling, for the multiplex as the number of sources increases and the sources themselves are assumed to have linear long time-scale asymptotics. This connects their work to the earlier work of Glynn and Whitt.

They prove, in the case where there is fixed service rate per source, that

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbb{P}[Q^L > Lq] = -J(q),$$

where Q^L is the stationary queue length distribution associated to L sources, where $J(q)$ is a function which can be evaluated if one knows the finite time cumulant generating functions of the input traffic.

Duffield [13] then extended this work to include sources which obey long time-scale large deviation principles with power-law scaling functions. In particular, he found in this case that the economies of scale can be much greater than in the linear scaling case.

1.2.4 Long Range Dependent Models

In queueing networks service rates at switches are often fixed. Most of the modeling effort focuses on developing stochastic models which exhibit phenomena found in data-sets. One such phenomena is long range dependence, this is where correlations within a traffic source decay in time, t, more slowly than $\exp(-tC)$, for some positive constant C. Long range dependence has been of interest to teletraffic engineers since claims have been made that it is found in certain data-sets (see Leland *et. al* [42], Crovella and Bestavros [9], and Beran *et al.* [2]), leading to the suggestion of fractional Brownian motion, which exhibits long range dependence, as a model for Internet traffic. Fractional Brownian motion suffers from two drawbacks as a model of Internet data, it is unbounded, and it takes negative values. This is unrealistic given hardware constrains traffic to be bounded, and negative arrivals are as physically unsatisfying as the existence of tachyons.

This motivated the work found in chapter four of this thesis, where two state sources are constructed which exhibit long range dependence. The basic construction involves describing a two state source by the sojourn times it spends in either state. If these sojourn times are distributed via a heavy tail, namely a distribution that decays slower than exponentially, then the source will exhibit long range dependence.

Using results of Russell, which can be found in [62], we relate the large deviations of the sojourn times to the large deviations of the source itself.

1.2.5 Effective Bandwidths

Large deviation rate-functions are notoriously difficult to calculate, even for simple models. If a rate-function is convex then an alternative representation which contains the same information is its Legendre-Fenchel transform, the scaled cumulant generating function, which is almost always easier to calculate.

From Glynn and Whitt's result, we know how the tail of the queue length decays in terms of the workload rate-function, but how would we use this knowledge in practice? We could watch the traffic going through a switch, fit a stochastic model to the measured traffic, calculate the models rate-function, and hence evaluate δ found in equation (1.5). This process could be acceptable for off-line investigation but is too time-consuming and inaccurate to do 'on-the-fly'. If one wished to do connection admission control one needs to be able to calculate 'on-the-fly'.

In a radical departure from existing suggestions, Duffield *et al.* [15] proposed taking a cue from chemical engineers and measuring the scaled cumulant generating function directly. They suggested a simple estimator which is sufficiently computational efficient that it can be calculated in real-time, on the switch. The work found in chapter five of this thesis is the calculation of distribution-free confidence intervals for the estimator that they proposed.

1.3 Risk Theory

The subject matter of Risk Theory is the stochastic modeling of insurance companies. See the book by Grandell for an introduction [29]. The basic model, and the one that receives the most attention, is called the Cramér-Lundberg model. This is where we model an insurance company to have initial capital u, have income s per unit time, and for the value of claims at time t to be governed by the stochastic process $\{X_t\}$. An important issue is the calculation of the probability of bankruptcy. Bankruptcy, or ruin, occurs if total claims at any time, t, exceed the initial capital plus the income earned up to time t. The ruin event is,

$$\sup_{t\geq 0}\left\{\sum_{i=0}^{t}(X_t - s) > u\right\}.$$

Identifying $s = S_t$, and $X_t = A_t$, this event is mathematically identical to the event $\{Q > u\}$. Hence work done on the asymptotic tail of the queue length distribution is automatically true for the Cramér-Lundberg model. Use of large deviation theory, in a seminal form, to calculate asymptotics for the ruin event was proposed in 1986 by Martin-Löf [53].

Chapter 2

Logarithmic Asymptotics for a Stable Single Server Queue and Applications to Statistical Multiplexing

"Even the Royal Mail can't deliver us from what we've got into."

Gomez

2.1 Introduction

Let $\{W_t\}$ be a stochastic process. Define $Q := \sup_{t\geq 0} W_t$. Under the assumption that W_t satisfies a large deviation principle on regularly varying scales, we deduce tail asymptotics for $\mathbb{P}[Q > q]$ as q becomes large. We discuss the problem viewing Q as the stationary queue length distribution at a single server queue. While this interpretation helps our physical intuition, clearly it is not necessary.

In [17] Duffield and O'Connell study the tail of the distribution of Q. The basic aim of [17] is to compute

$$\lim_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q], \tag{2.1}$$

where $h(q)$ is an appropriately chosen scaling function. Glynn and Whitt [28] treat the case where $h(q) = q$ which is appropriate when a large deviation principle (LDP) is satisfied by the pair $(W_t/t, t)$. Duffield and O'Connell generalise to the case in which one has a LDP with a different scaling: $(W_t/a(t), v(t))$, where $a(t)$ and $v(t)$ are non-decreasing. Yet a third scaling function is employed for (2.1).

Fundamental to the discussion in [17] is the existence of a function $h(t)$ such that the limit $g(c)$ defined by

$$g(c) := \lim_{t\to\infty} \frac{v(a^{-1}(t/c)\,)}{h(t)} \tag{2.2}$$

exists for all $c > 0$. When such an h exists, one can define $\widetilde{h} := v \circ a^{-1}$. Then

$$\lim_{t\to\infty} \frac{h(t)}{\widetilde{h}(t)} = g(1).$$

The use of $\widetilde{h}$ instead of h affects the asymptotic properties of (2.1) only by the multiplicative constant $g(1)$, so there is no loss in defining

$$h(t) := v\Big(a^{-1}(t)\Big).$$

Then the limit (2.2) becomes

$$\lim_{t\to\infty} \frac{h(t/c)}{h(t)} = g(c),$$

which, if $a(t)$ and $v(t)$ are non-decreasing, suffices for h to be *regularly varying* (see Bingham *et al.* [3]). In this work we restrict to the case where both $a(t)$ and $v(t)$ are *regularly varying.* Specifically this means that there exists $V \geq 0$, and $A \geq 0$, such that

$$\lim_{t\to\infty}\frac{v(ct)}{v(t)} = c^V \qquad \text{and} \qquad \lim_{t\to\infty}\frac{a(ct)}{a(t)} = c^A, \tag{2.3}$$

for all $c > 0$. We shall assume that both V and A are greater than zero and that $a(t)$ is both continuous and strictly increasing. Note that, assuming $a(t)$ is continuous and increasing,

$$\liminf_{q\to\infty}\frac{1}{v(a^{-1}(q))}\log\mathbb{P}[Q > q] = \liminf_{q\to\infty}\frac{1}{v(q)}\log\mathbb{P}[Q > a(q)] \tag{2.4}$$

and

$$\limsup_{q\to\infty}\frac{1}{v(a^{-1}(q))}\log\mathbb{P}[Q > q] = \limsup_{q\to\infty}\frac{1}{v(q)}\log\mathbb{P}[Q > a(q)]. \tag{2.5}$$

This suggests that we are running the logarithmic asymptotics of Q on the same scales as W_t, which is essentially the case.

As Glynn and Whitt did, Duffield and O'Connell employ the Gartner-Ellis conditions to ensure that an LDP is satisfied with a convex rate-function; they employ methods based on the scaled cumulant generating function. These conditions are unnecessarily restrictive and hinder intuition.

Our approach makes evident a lacuna in Theorem 2.2 of [17]; their equation (34) is not valid in general. We assume that an LDP is satisfied by the pair $(W_t/a(t), v(t))$, but do not make assumptions about how it has been deduced. We place our estimates directly on the probabilities, making clear their interpretation, and we provide a simple form for the resulting asymptotic rate of decay.

With these assumptions in mind it is not difficult to prove the lower bound (details are presented in Proposition 2.1),

$$\liminf_{q\to\infty}\frac{1}{v(a^{-1}(q))}\log\mathbb{P}[Q > q] \geq -\inf_{c>0} c^V I\left(\frac{1}{c^A}\right). \tag{2.6}$$

More care, however, must be taken with the corresponding upper bound. This begins by splitting the set $\{Q > q\}$ into the union of three, not necessarily disjoint, subsets,

$$\{Q > q\} = \{\sup_{t:a(t)<q\bar{c}} W_t > q\} \cup \{\sup_{t:q\bar{c}\leq a(t)\leq q\underline{c}} W_t > q\} \cup \{\sup_{t:a(t)>q\underline{c}} W_t > q\}, \tag{2.7}$$

where $0 < \overline{c} < \underline{c} < \infty$. The principle of the largest term ensures that the rate of decay of the probability of $\{Q > q\}$ is less than or equal to the slowest rate of decay of these three sets. It will be shown in Proposition 2.3 that the rate of decay of the probability of the middle term in (2.7), for any $0 < \overline{c} < \underline{c} < \infty$, is bounded above by the rate found in the lower bound (2.6). It is shown in Proposition 2.2 that there exists a $\underline{c} < \infty$ such that the rate of decay of the probability of the final term in (2.7) is bounded above by the rate found in the lower bound (2.6).

The LDP hypothesis refers to limiting behavior of $\log \mathbb{P}[W_t > ca(t)]$, not values for specific t. The asymptotics of a single W_t could dominate those of $\log \mathbb{P}[Q > b]$. We need to impose an additional condition which excludes this possibility, ensuring that there exists a $\overline{c} > 0$ such that the rate of decay of the probability of the first term in (2.7) is as quick as the rate found in the lower bound (2.6). We then have that

$$\lim_{q\to\infty} \frac{1}{v(a^{-1}(q))} \log \mathbb{P}[Q > q] = -\inf_{c>0} c^V I\left(\frac{1}{c^A}\right). \tag{2.8}$$

In section 2.2 we set up our basic formulation and assumptions, stating the main result and providing sufficient conditions for it to hold. In section 2.3 we present examples to illustrate these results. In section 2.4 we relate the rate of decay of the stationary queue length distribution to the stationary waiting time distribution. In section 2.5 we discuss statistical multiplexing (which is of key interest to teletraffic engineers) in the presence of non-linear scales. In section 2.6 we present some examples of statistical multiplexing on non-linear scales. We illustrate how many-scale behavior can occur: the scale on which large deviations are observed depends upon the size of the deviation.

2.2 Main Results

We consider a family of random variables $\{W_t : t \in T\}$ where T is an unbounded subset of $\mathbb{R}_+$. In this work we shall be primarily interested in $T = \mathbb{Z}_+$ but provide an additional hypothesis under which the work extends to $T = \mathbb{R}_+$. We define $Q := \sup_{t\geq 0} W_t$.

Scaling hypothesis: The function $a : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous and strictly increasing, and the function $v : \mathbb{R}_+ \to \mathbb{R}_+$ is non-decreasing with $\lim_{t\to\infty} v(t) = \lim_{t\to\infty} a(t) = \infty$. For each $c > 0$

$$\lim_{t\to\infty} \frac{a(ct)}{a(t)} = c^A \qquad \text{and} \qquad \lim_{t\to\infty} \frac{v(ct)}{v(t)} = c^V, \tag{2.9}$$

where $A > 0$ and $V > 0$ are constants.

By Theorem 1.10.2 of [3] $a(t)$ and $v(t)$ are regularly varying.

Note that in Loynes original work [47], where the distribution $Q := \sup_{t\geq 0} W_t$ was first defined, $\{W_t : t \in \mathbb{Z}_+\}$ is defined to be a partial sums process associated with a stationary process $\{Z_t\}$, that is $W_t := \sum_{i=-t}^{-1} Z_i$, and $W_0 := 0$. Z_t is the arrivals less service at time t. In this setting, the individual ergodic theorem (see page 18 of Halmos [33]) proves that

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} Z_{-i} = \lim_{t\to\infty} \frac{W_t}{t} = \mathbb{E}[Z_1|\mathcal{F}],$$

where $\mathcal{F}$ is the invariant $\sigma-$algebra. Hence, in this situation, it seems likely that $a(t)$ would be set to be t; if $a'(t)$ is any other scale such that $\lim a'(t)/t \in \{0, \infty\}$, then the information about the mean behavior of the arrivals less service is lost on the scale $a'(t)$.

If $a(t)$ is regularly varying with constant A, then Theorem 1.5.12 of [3] proves that $a^{-1}(t)$ is regularly varying with constant $1/A$.

Definition 2.1 *We define the queue scaling function $h : \mathbb{R}_+ \to \mathbb{R}_+$ by*

$$h(q) := v(a^{-1}(q)). \tag{2.10}$$

From chapter 1 of [3], in particular Theorems 1.4.1 and 1.5.6, we deduce the following:

Lemma 2.1 *Under the scaling hypothesis the function h defined by (2.10) satisfies for each $c > 0$*

$$\lim_{t\to\infty} \frac{h(ct)}{h(t)} = c^{V/A}.$$

As $t \to \infty$ the ratio $h(ct)/h(t)$ converges to $c^{V/A}$ uniformly as a function of c on compact subsets of $\mathbb{R}_+$. Similarly, the ratios $v(ct)/v(t)$ and $a(ct)/a(t)$ converge uniformly to c^V and c^A as functions of c on compact subsets of $\mathbb{R}_+$. For each $\delta > 0$ and $C > 1$ there exists $b_{\delta,C}$ so that $q, t \geq b_{\delta,C}$ implies

$$h(t)/h(q) \leq C \max\{(t/q)^{V/A+\delta}, (t/q)^{V/A-\delta}\}.$$

Similarly for $v(t)/v(q)$ and $a(t)/a(q)$.

For each $x \in \mathbb{R}$, define $\lceil x \rceil$ to be the least integer greater than x and $\lfloor x \rfloor$ to be the greatest integer which is smaller than x.

Lemma 2.2 *Under the scaling hypothesis, for all $c > 0$,*

$$\lim_{t\to\infty} \frac{v(\lceil a^{-1}(ct) \rceil)}{v(a^{-1}(t))} = c^{V/A}.$$

PROOF By definition of the ceiling function, $\lceil x \rceil$,

$$a^{-1}(ct) \leq \lceil a^{-1}(ct) \rceil \leq a^{-1}(ct) + 1.$$

As, by Theorem 1.5.12 of [3], $a^{-1}(t)$ is regularly varying with constant $1/A$ and, by assumption, is increasing,

$$c^{1/A} = \lim_{t\to\infty} \frac{a^{-1}(ct)}{a^{-1}(t)} \leq \lim_{t\to\infty} \frac{\lceil a^{-1}(ct) \rceil}{a^{-1}(t)} \leq \lim_{t\to\infty} \left(\frac{a^{-1}(ct)}{a^{-1}(t)} + \frac{1}{a^{-1}(t)} \right) = c^{1/A}.$$

Hence, given $0 < \varepsilon < c$, there exists N_ε such that

$$\frac{v((c-\varepsilon)^{1/A} a^{-1}(t))}{v(a^{-1}(t))} \leq \frac{v(\lceil a^{-1}(ct) \rceil)}{v(a^{-1}(t))} \leq \frac{v((c+\varepsilon)^{1/A} a^{-1}(t))}{v(a^{-1}(t))},$$

for all $t > N_\varepsilon$. Hence, as $v(t)$ is regularly varying,

$$(c-\varepsilon)^{V/A} \leq \lim_{t\to\infty} \frac{v(\lceil a^{-1}(ct) \rceil)}{v(a^{-1}(t))} \leq (c+\varepsilon)^{V/A},$$

for all $0 < \varepsilon < c$; hence the result.

Lemma 2.3 *Under the scaling hypothesis, for each $\gamma > 0$,*

$$\lim_{n\to\infty} \frac{1}{-\gamma v(n)} \log \sum_{k=n}^{\infty} e^{-\gamma v(k)} = 1. \tag{2.11}$$

PROOF We need only show that the left has side in (2.11) is not less than 1. Fix α satisfying $0 < \alpha < V$. By Lemma 2.1, for $C > 1$ there exists b_C so that for $n \geq b_C$ $v(k) \geq v(n)(k/n)^\alpha/C$ when $k > n$. By considering Riemann sums we deduce

$$\sum_{k=n+1}^{\infty} \exp -\gamma v(k) \leq \sum_{k=n+1}^{\infty} \exp -\frac{\gamma v(n)(k/n)^\alpha}{C} \leq n \int_1^\infty \exp -\frac{\gamma v(n)\, x^\alpha}{C}\, dx.$$

For all sufficiently large n, $\gamma v(n)/C > 1$, so

$$\sum_{k=n}^{\infty} \exp -\gamma v(k) \leq \exp -\gamma v(n) + n \left(\int_1^\infty \exp -(x^\alpha - 1)\, dx \right) \exp -\frac{\gamma v(n)}{C}\,.$$

Since $(\log n)/v(n) \to 0$, the desired inequality follows by taking $C \downarrow 1$.

LDP hypothesis: $(W_t/a(t), v(t))$ satisfies a large deviation principle with rate-function $I(x)$. That is, there exists a lower semi-continuous function $I : \mathbb{R} \to [0, \infty]$ such that for all F closed

$$\limsup_{t\to\infty} \frac{1}{v(t)} \log \mathbb{P}\left[\frac{W_t}{a(t)} \in F\right] \leq -\inf_{x \in F} I(x),$$

and for all G open

$$\liminf_{t\to\infty} \frac{1}{v(t)} \log \mathbb{P}\left[\frac{W_t}{a(t)} \in G\right] \geq -\inf_{x \in G} I(x).$$

We do not assume that $I(x)$ has compact level sets (ie. is a good rate-function), as some authors do, as they are not necessary for this part of the treatment. We will, however, need compact level sets in the section on statistical multiplexing, Section 2.5, in order to use the simplest form of the contraction principle.

Stability and Continuity hypothesis: $I(x)$ is non-decreasing for $x > 0$, $I(0) > 0$ and there is some $x > 0$ such that $I(x) < \infty$. Moreover, $I(x)$ is assumed to be continuous on the interior of the set upon which it is finite, which we denote $\mathcal{I}$.

We say that a set A is a concentration set, with respect to the rate-function $I(x)$, if

$$\inf_{x \in A} I(x) = 0.$$

That is, A is a concentration set if the probability of A does not decay exponentially with respect to the scales upon which $I(x)$ is defined.

If $I(x) = 0$ for some $x > 0$, then the queue will be unstable: having more arrivals than service will be concentration set. If $I(x) = \infty$ for all $x > 0$, then $\mathbb{P}[Q > q]$ will be asymptotically zero with rate ∞.

Note that under the LDP, stability and continuity hypotheses, for all $c > 0$,

$$\begin{aligned} -I(c) &= -\inf_{x>c} I(x) \\ &\leq \liminf_{t\to\infty} \tfrac{1}{v(t)} \log \mathbb{P}\left[\tfrac{W_t}{a(t)} > c\right] \\ &\leq \limsup_{t\to\infty} \tfrac{1}{v(t)} \log \mathbb{P}\left[\tfrac{W_t}{a(t)} > c\right] \\ &\leq \limsup_{t\to\infty} \tfrac{1}{v(t)} \log \mathbb{P}\left[\tfrac{W_t}{a(t)} \geq c\right] \\ &\leq -\inf_{x\geq c} I(x) \\ &= -I(c). \end{aligned}$$

Lemma 2.4 *Under the LDP, stability and continuity hypotheses*

$$\lim_{t\to\infty} \frac{-1}{v(t)} \log \mathbb{P}\left[\frac{W_t}{a(t)} > c\right] = I(c)$$

converges uniformly as a function of c on compact intervals contained in, $\mathcal{I}$, the interior of the set upon which $I(x)$ is finite.

PROOF For each $t \in T$,

$$I_t(c) := \frac{-1}{v(t)} \log \mathbb{P}\left[\frac{W_t}{a(t)} > c\right]$$

is a non-decreasing function of c. Fix a compact interval $[\overline{c}, \underline{c}] \subset \mathcal{I}$ and let $\varepsilon > 0$ be given. As $I(c)$ is finite and continuous on $\mathcal{I}$, we can select $\overline{c} = x_0 < x_1 < \cdots < x_m = \underline{c}$ such that $I(x_i) - I(x_{i-1}) < \varepsilon$ for $i = 1, \ldots, m$. As $I_t(x)$ converges pointwise to $I(x)$, we can find N_ε such that $|I_t(x_i) - I(x_i)| < \varepsilon$ for all $t > N_\varepsilon$ and all $i \in \{0, \ldots, m\}$. Then, as $I_t(c)$ is non-decreasing, for any $c \in [\overline{c}, \underline{c}]$ and all $t > N_\varepsilon$, $|I_t(c) - I(c)| < 2\,\varepsilon$.

The LDP hypothesis refers to limiting behavior of $\log \mathbb{P}[W_t > ca(t)]$, not values for specific t. The asymptotics of a single W_t could dominate those of $\log \mathbb{P}[Q > b]$. The condition below excludes this possibility.

Uniform individual decay rate hypothesis: There exist constants $F > V/A$, $K > 0$ so that for all t and all $c > K$,

$$\frac{1}{v(t)} \log \mathbb{P}[W_t > ca(t)] \leq -c^F. \tag{2.12}$$

Note that if $a(t) = t$, then a sufficient condition for the uniform individual decay rate hypothesis to be satisfied is for the arrival rate to be almost surely bounded by a fixed K.

Extension hypothesis:

$$\lim_{q \to \infty} \sup_{k \in \mathbb{Z}_+} \frac{\log \mathbb{P}[\sup_{k<t\leq k+1} W_t > q]}{h(q)} = \lim_{q \to \infty} \sup_{k \in \mathbb{Z}_+} \frac{\log \mathbb{P}[W_k > q]}{h(q)}.$$

This hypothesis is trivially satisfied when $T = \mathbb{Z}_+$. For $T = \mathbb{R}_+$ additional information about $\{W_t\}$ is needed to assure that the supremum over $k < t \leq k+1$ does not differ significantly from W_{k+1}. Though this hypothesis may be difficult to prove for specific models, in actual queues it should be quite clear whether there is a significant difference between the maximum over all t and the maximum over $t \in \mathbb{Z}_+$. A sufficient condition is that the service rate be almost surely bounded above by a fixed K,

$$\sup_{t \in T: k<t\leq k+1} W_t \leq W_{k+1} + K,$$

for all $k \in \mathbb{Z}_+$.

The asymptotic lower bound for queue length probability is a direct consequence of the LDP and scaling hypotheses for $(W_t/a(t), v(t))$.

Proposition 2.1 *Under the scaling and LDP hypotheses*

$$\liminf_{q \to \infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] \geq -\inf_{c>0} c^V I\left(\frac{1}{c^A}\right). \tag{2.13}$$

PROOF Fix $c > 0$, as $Q = \sup_{t \geq 0} W_t$ we know that

$$\frac{1}{h(q)} \log \mathbb{P}[Q > q] \geq \frac{1}{h(q)} \log \mathbb{P}[W_{\lceil a^{-1}(cq) \rceil} > q]. \tag{2.14}$$

Using equation (2.10) we have that the right hand side of (2.14) is equal to

$$\frac{v(\lceil a^{-1}(cq) \rceil)}{v(a^{-1}(q))} \frac{1}{v(\lceil a^{-1}(cq) \rceil)} \log \mathbb{P}\left[W_{\lceil a^{-1}(cq) \rceil} > q\right]. \tag{2.15}$$

By Lemma 2.2,

$$\lim_{q \to \infty} \frac{v(\lceil a^{-1}(cq) \rceil)}{v(a^{-1}(q))} = c^{V/A}. \tag{2.16}$$

Let $b = a^{-1}(cq)$, then $q = a(b)/c$. So that,

$$\frac{1}{v(\lceil a^{-1}(cq) \rceil)} \log \mathbb{P}\left[W_{\lceil a^{-1}(cq) \rceil} > q\right] = \frac{1}{v(\lceil b \rceil)} \log \mathbb{P}\left[W_{\lceil b \rceil} > \frac{a(b)}{c}\right].$$

As $a(t)$ is increasing, we know that $a(\lceil b \rceil) \geq a(b)$, thus

$$\mathbb{P}\left[W_{\lceil b \rceil} > \frac{a(b)}{c}\right] \geq \mathbb{P}\left[W_{\lceil b \rceil} > \frac{a(\lceil b \rceil)}{c}\right].$$

The LDP hypothesis ensures that

$$\lim_{\lceil b \rceil \to \infty} \frac{1}{v(\lceil b \rceil)} \log \mathbb{P}\left[W_{\lceil b \rceil} > a(\lceil b \rceil)/c\right] = -I\left(\frac{1}{c}\right). \tag{2.17}$$

Thus using (2.14), (2.15), (2.16), and (2.17),

$$\liminf_{q \to \infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] \geq -c^{V/A} I\left(\frac{1}{c}\right).$$

As this is true for all $c > 0$,

$$\liminf_{q \to \infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] \geq -\inf_{c > 0} c^{V/A} I\left(\frac{1}{c}\right).$$

Substituting $c' = c^{1/A}$ in for c, we get the result (2.13).

The upper bound is treated by splitting the event $\{Q > q\}$ into three parts, each of which is dealt with separately.

Theorem 2.5 *For all* $\underline{c} > \underline{c} > 0$,

$$\limsup \frac{1}{h(q)} \log \mathbb{P}[Q > q] = \begin{cases} \limsup \frac{1}{h(q)} \log \mathbb{P}[\sup_{t:a(t)<q\bar{c}} W_t > q] \\ \limsup \frac{1}{h(q)} \log \mathbb{P}[\sup_{t:q\bar{c} \leq a(t) \leq q\underline{c}} W_t > q] \\ \limsup \frac{1}{h(q)} \log \mathbb{P}[\sup_{t:a(t)>q\underline{c}} W_t > q]. \end{cases} \tag{2.18}$$

PROOF Using (2.7), a direct application of the principle of the largest term (see Lemma 2.3 of Lewis and Pfister [43]) suffices.

First we treat the last term in equation (2.18).

Proposition 2.2 *Under the scaling, LDP, stability and continuity hypotheses, there exists* $0 < \underline{c} < \infty$ *such that*

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[\sup_{t:a(t)>q\underline{c}} W_t > q] < -\inf_{c>0} c^V I\left(\frac{1}{c^A}\right). \tag{2.19}$$

PROOF For $\underline{c} > 0$ we have

$$\mathbb{P}[\sup_{k:a(k)>q\underline{c}} W_k > q] = \mathbb{P}[\bigcup_{k:a(k)>q\underline{c}} W_k > q] \leq \sum_{k:a(k)>q\underline{c}} \mathbb{P}[W_k > q].$$

The LDP and stability hypotheses imply

$$\lim_{t\to\infty} \frac{-1}{v(t)} \log \mathbb{P}\left[\frac{W_t}{a(t)} > 0\right] = I(0) > 0.$$

Select γ, $0 < \gamma < I(0)$. Then for all sufficiently large k

$$\mathbb{P}[W_k > q] \leq \mathbb{P}[W_k > 0] \leq e^{-\gamma v(k)},$$

and for all sufficiently large q,

$$\sum_{k:a(k)>q\underline{c}} \mathbb{P}[W_k > q] \leq \sum_{k:a(k)>q\underline{c}} e^{-\gamma v(k)}.$$

By Lemma 2.3, there exists $\delta > 0$ so that

$$-\log \sum_{k=n}^{\infty} e^{-\gamma v(k)} > \delta \gamma v(n)$$

for all sufficiently large n. Then, letting $n = a^{-1}(q\underline{c})$,

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log \sum_{k:a(k)>q\underline{c}} \mathbb{P}[W_k > q] \leq \limsup_{q\to\infty} -\delta\gamma \frac{v(a^{-1}(q\underline{c}))}{h(q)} = -\delta\gamma \underline{c}^{V/A}.$$

Since $c^{V/A} \to \infty$ as $c \to \infty$, we may choose $\underline{c} > 0$ so that

$$-\delta\,\gamma\,\underline{c}^{V/A} < -\inf_{c>0} c^V I\left(\frac{1}{c^A}\right).$$

Now for the middle term in equation (2.18).

Proposition 2.3 *Under the scaling, LDP, stability and continuity hypotheses, for all* $0 < \bar{c} < \underline{c} < \infty$,

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[\sup_{t:q\bar{c}\le a(t)\le q\underline{c}} W_t > q] \le -\inf_{c>0} c^V I\left(\frac{1}{c^A}\right). \tag{2.20}$$

PROOF We have that

$$\mathbb{P}[\sup_{k:q\bar{c}\le a(k)\le q\underline{c}} W_k > q] \le a^{-1}(q\underline{c}) \max_{k:q\bar{c}\le a(k)\le q\underline{c}} \mathbb{P}\left[W_k > q\right]. \tag{2.21}$$

As $V > 0$

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log a^{-1}(q\underline{c}) = 0.$$

Thus, using equation (2.21) and the fact that log is order-preserving, we have that the left hand side of equation (2.20) is less than or equal to

$$\limsup_{q\to\infty} \max_{k:q\bar{c}\le a(k)\le q\underline{c}} \frac{1}{h(q)} \log \mathbb{P}\left[W_k > q\right]. \tag{2.22}$$

Let

$$I_k(c) := \frac{-\log P\left[W_k > c\,a(k)\right]}{v(k)}. \tag{2.23}$$

Select $\varepsilon > 0$, $\varepsilon < 1/\underline{c}$. If $I(1/\bar{c}) < \infty$, let $c^* := 1/\bar{c}$. Otherwise select c^* so that $I(c^*) < \infty$, $c^* + \varepsilon \le 1/\bar{c}$ and $I(c^* + \varepsilon) = +\infty$. Now $\lim_k I_k(c) = I(c)$ for each $c \in [1/\underline{c}, c^*]$ by the LDP hypothesis. By Lemma 2.4 we have uniform convergence on $[1/\underline{c}, c^*]$. Note $I(c) \ge I(0) > 0$ for $c > 0$. Then there exists N_ε so that $n \ge N_\varepsilon$ implies

$$I_n(c) > I(c)(1-\varepsilon) \text{ for } c \in [1/\underline{c}, c^*] \text{ and } I_n(c^*+\varepsilon) > \frac{1}{\varepsilon} \text{ if } I(1/\bar{c}) = +\infty. \tag{2.24}$$

Note that $I_n(c) \ge I_n(c^*)$ for $c \in [c^*, c^* + \varepsilon]$ and $I_n(c) \ge I_n(c^* + \varepsilon)$ for $c \in [c^* + \varepsilon, 1/\bar{c}]$. For $q > a(N_\varepsilon \bar{c})$ define c_k for each k, $a^{-1}(q/\bar{c}) \le k \le a^{-1}(q/\underline{c})$, by $c_k := q/a(k)$ so that we have $v(k) = h(q/c_k)$ and from (2.23) and (2.24),

$$-\log P[W_k > q] > \begin{cases} h(q/c_k)I(c_k)(1-\varepsilon) & \text{if } \; 1/\underline{c} < c_k \le c^* \\ h(q/c_k)I(c^*)(1-\varepsilon) & \text{if } \; c^* < c_k < c^* + \varepsilon \\ h(q/c_k)/\varepsilon & \text{if } \; c^* + \varepsilon \le c_k \le 1/\bar{c}. \end{cases} \tag{2.25}$$

The value of k at which the max in (2.22) occurs corresponds to the minimal term in (2.25). For sufficiently small ε the minimum does not occur at the $h(q/c_k)/\varepsilon$ term. Then this minimal term divided by $h(q)$ is not less than

$$\inf_{c\in[1/\underline{c},1/\overline{c}]} \frac{h(q/c)}{h(q)} I(c-\varepsilon)(1-\varepsilon) \geq \inf_{c>0} \frac{h(q/c)}{h(q)} I(c-\varepsilon)(1-\varepsilon).$$

Taking $q \to \infty$ and then the limit $\varepsilon \to 0$, and substituting $c' = c^{-1/A}$, yields (2.20).

Finally, the first term in equation (2.18).

Proposition 2.4 *Let the sequence $\{W_t\}$ satisfy the scaling, LDP, stability and continuity hypotheses, and the uniform individual decay rate hypothesis, then there exists $\overline{c} > 0$ such that*

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[\sup_{k:a(k)<q\overline{c}} W_k > q] < \inf_{c>0} c^V I\left(\frac{1}{c^A}\right). \tag{2.26}$$

PROOF Note that

$$\mathbb{P}[\sup_{k:a(k)<q\overline{c}} W_k > q] \leq a^{-1}(q\overline{c}) \max_{k:a(k)<q\overline{c}} \mathbb{P}[W_k > q],$$

and, as $V > 0$,

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log a^{-1}(q\overline{c}) = 0.$$

Therefore

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[\sup_{k:a(k)<q\overline{c}} W_k > q] = \limsup_{q\to\infty} \max_{k:a(k)<q\overline{c}} \frac{1}{h(q)} \log \mathbb{P}[W_k > q]. \tag{2.27}$$

Define c_k for each integer k, $0 < k \leq a^{-1}(q\overline{c})$, by $c_k := q/a(k)$, so that we have $v(k) = h(q/c_k)$ and from (2.12) for each $c_k > K$,

$$\frac{1}{h(q)} \log \mathbb{P}[W_k > c_k a(k)] \leq -\frac{h(a(k))}{h(q)} \left(\frac{q}{a(k)}\right)^F = -\left(\frac{h(q/c_k)}{h(q)}\right) c_k^F. \tag{2.28}$$

Take $C > 1$ and δ, $0 < \delta < F - V/A$. Letting $t = a(k)$, by Lemma 2.1 there exists $b_{\delta,C}$ so that for $q > t > b_{\delta,C}$,

$$\frac{h(t)}{h(q)} \geq \frac{1}{C} \left(\frac{t}{q}\right)^{V/A+\delta},$$

and, by (2.28), for $c_k := q/a(k) > \max\{K, 1\}$ and $a(k) > b_{\delta,C}$,

$$\frac{1}{h(q)} \log \mathbb{P}[W_k > c_k a(k)] \leq -\frac{1}{C}\left(\frac{1}{c_k}\right)^{V/A+\delta} c_k^F.$$

Choose $1/\bar{c} > \max\{K, 1\}$ so that

$$-\frac{1}{C}\left(\frac{1}{\bar{c}}\right)^{(F-V/A-\delta)} < -\inf_{c>0} c^V I\left(\frac{1}{c^A}\right).$$

Then for each $q > b_{\delta,C}$ and any k which satisfies $b_{\delta,C} < a(k) \leq q\bar{c}$ we have

$$\frac{1}{h(q)} \log \mathbb{P}[W_k > c_k\, a(k)] < -\inf_{c>0} c^V I\left(\frac{1}{c^A}\right). \tag{2.29}$$

Inequality (2.28) implies that for each fixed k, $\log \mathbb{P}[W_k > c_k a(k)]/h(q) \to -\infty$ as $q \to \infty$. Since there are only finitely many k with $a(k) \leq b_{\delta,C}$, (2.27) and (2.29) together imply (2.26).

From Proposition 2.1, Theorem 2.5 and Propositions 2.2, 2.3, 2.4, we deduce the following:

Theorem 2.6 *If the sequence $\{W_t\}$ satisfies the scaling, LDP, stability and continuity hypotheses, the uniform individual decay rate hypothesis and the extension hypothesis, then*

$$\lim_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = -\inf_{c>0} c^V I\left(\frac{1}{c^A}\right). \tag{2.30}$$

2.2.1 The Scaled Cumulant Generating Function

The Cumulant Generating Function (sCGF) of W_t, scaled by $(a(t), v(t))$, is defined by

$$\lambda_t(\theta) := \frac{1}{v(t)} \log \mathbb{E}\left[\exp\left(\theta v(t) W_t/a(t)\right)\right]. \tag{2.31}$$

The analysis in [28] and [17] is based on the sCGF. The hypotheses stated so far have simple expressions in terms of the sCGF, when it exists. The conditions we specify here for the sCGF case are intended for easy applicability, rather than maximum generality. Under these assumptions, the large deviation rate-function is convex. The sCGF technique is not applicable to models which have non-convex rate-functions.

LDP Hypothesis, sCGF case: For all $\theta \in \mathbb{R}$ and all t the scaled cumulant generating function given by (2.31) exists as a finite real value. For each $\theta \in \mathbb{R}$ the following finite limit exists and defines $\lambda(\theta)$:

$$\lambda(\theta) := \lim_{t\to\infty} \lambda_t(\theta).$$

Furthermore $\lambda(\theta)$ is assumed to be continuously differentiable. These assumptions imply the following (see [11]).

Proposition 2.5 *Under the above assumptions the pair* $\big(W_t/a(t), v(t)\big)$ *satisfies a large deviation principle with rate-function* $I(x)$ *given by the Legendre-Fenchel transform of* $\lambda(\theta)$,

$$I(x) := \sup_{\theta}\{\theta x - \lambda(\theta)\}. \tag{2.32}$$

This implies that $I(x)$ is a convex function and continuous on the interior of the set where it is finite.

Stability Hypothesis, sCGF case: There exists $\overline{\theta} > 0$ so that $\lambda(\overline{\theta}) < 0$.

The above implies that $I(0) \geq -\lambda(\overline{\theta})$.

The uniform individual decay rate hypothesis can be readily expressed in terms of the sCGF.

Proposition 2.6 *If there exists constants* F', M *such that* $F' > \max\{V/A, 1\}$ *and*

$$\lambda_t(\theta) \leq M\theta^{F'/(F'-1)}$$

for all $\theta > 0$ *and all* t*, then for each* F, $\max\{V/A, 1\} < F < F'$, *there exists* K_F *so that for all* $c > K_F$ *and all* t,

$$\frac{1}{v(t)} \log \mathbb{P}[W_t > c\,a(t)] \leq -c^F$$

and the uniform individual decay rate hypothesis is satisfied.

PROOF An elementary consequence of (2.31) is Chernoff's inequality,

$$\log \mathbb{P}[W_t > c\,a(t)] \leq -v(t)\left(c\theta - \lambda_t(\theta)\right).$$

It then follows that

$$\log \mathbb{P}[W_t > c\,a(t)] \leq -v(t)\left(c\theta - M\theta^{F'/(F'-1)}\right).$$

Choosing $\theta = (c(F'-1)/(MF'))^{F'-1}$ we have

$$\log \mathbb{P}[W_t > c\,a(t)] \leq -v(t)c^{F'}\left(M^{(1-F')}F'^{-F'}(F'-1)^{F'-1}\right). \tag{2.33}$$

Since M and F' are constants, for each F, $F' > F > \max\{V/A, 1\}$, there exists K_F such that, for all $c > K_F$, the right hand side of (2.33) will be less than $-v(t)c^F$.

2.3 Decay Rate Examples

We present two sets of examples, one set is based on Gaussian processes and the other on heavy tailed processes. Fractional Brownian motion (which is covered in the Gaussian processes section) has been proposed by Leland *et al.* [42] as a model for multiplexes of Ethernet data. Heavy tailed distributions are often studied in risk theory (see Mikosch and Nagaev [54] and references therein) and more recently in queueing theory (see Asmussen and Collamore [1]). For these models it is possible to get much more information than is available on a logarithmic scale. Interesting features are still displayed even after taking logs.

2.3.1 Application to Gaussian Processes

Perhaps the simplest examples to which the theory can be applied are those in which

$$W_t := X_t - \mu t,$$

where $\mu > 0$ is constant and $\{X_t\}$ are mean zero Gaussian random variables with covariance function

$$\Gamma(s,t) := \mathbb{E}[X_s X_t], \qquad \sigma_t^2 := \Gamma(t,t).$$

Define

$$a(t) := t \qquad \text{and} \qquad v(t) \equiv h(t) := t^2/\sigma_t^2.$$

$\lambda_t(\theta)$ is given, by

$$\lambda_t(\theta) = \frac{1}{2}\theta^2 - \mu\theta,$$

for all $t \geq 0$. Thus, using (2.32),

$$I(x) = \frac{1}{2}(x+\mu)^2.$$

Note that $I(x)$ is continuous and that it is non-decreasing for $x \geq 0$. Thus it suffices to assume that, for all $c > 0$,

$$\lim_{t\to\infty} \frac{\sigma_{ct}^2}{\sigma_t^2} = \lim_{t\to\infty} \frac{\Gamma(ct,ct)}{\Gamma(t,t)} = c^S, \qquad 0 < S < 2.$$

Then $v(t)$ is regularly varying with

$$\lim_{t\to\infty} \frac{v(ct)}{v(t)} = c^V, \qquad V = 2 - S.$$

Thus

$$\inf_{c>0} c^V I\left(\frac{1}{c}\right) = \frac{\mu^{2-V}}{V-2}\left(\frac{V-2}{V}\right)^{V/2}. \tag{2.34}$$

In particular, fractional Brownian motion has

$$2\Gamma(s,t) = s^{2H} + t^{2H} - |s-t|^{2H}, \qquad \sigma_t^2 = t^{2H},$$

where $0 < H < 1$. Here $a(t) := t$, $v(t) := h(t) := t^{2H}$. Note that the condition in Proposition 2.6 is satisfied with $F' = 2$ and $M = 1$. Thus (2.34) gives the exponential rate of decay of the tail of the stationary queue length distribution:

$$\lim_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = \lim_{q\to\infty} \frac{1}{v(q)} \log \mathbb{P}[Q > q] = \frac{\mu^{2-V}}{V-2}\left(\frac{V-2}{V}\right)^{V/2}.$$

Similarly, the Ornstein-Uhlenbeck position process has

$$\sigma_t^2 = 2\left(\frac{\mu}{\nu}\right)^2 (t + e^{-t} - 1),$$

where μ and ν are positive constants. Clearly σ_t^2 is regularly varying with $S = 1$. See [17] for details.

2.3.2 Application to Heavy Tailed Processes

In [62] Russell lays down a prescription to calculate the large deviations rate-function for the partial sums of a two state source which can be described in terms of the sojourn times it spends in the 'on' and 'off' states. In [19] Duffy uses this prescription to calculate the large deviations rate-function for a two state source whose sojourn times are distributed using a heavy tailed heavy distribution H,

$$\mathbb{P}[H \geq x] = d(x)e^{-v(x)},$$

where $d(x)$ is slowly varying (see Bingham *et al.* [3]), and $v(x)$ is regularly varying with constant $0 < V < 1$.

We define $\{Y_t\}$ to be a stationary sequence of two state random variables taking values in $\{0, 1\}$ whose sojourn times spent in the 0 and 1 states are distributed by an i.i.d. sequence with distribution H. Note that the mean of Y_t is $1/2$ as its sojourn times spent in the 'on' and 'off' states have finite, and equal, expectation. Define

$$Z_t := Y_t - \mu,$$

where $\mu \in (1/2, 1)$ so that the stability hypothesis will be satisfied. Define the workload process by

$$W_t := \int_0^t Z_s \, ds.$$

It is shown in [19] that $(W_t/t, v(t))$ satisfies a large deviation principle with rate-function, $I(x)$, given by

$$I(x) = \begin{cases} (1 - 2(x + \mu))^V & \text{if } x \in [-\mu, 1/2 - \mu] \\ (2(x + \mu) - 1)^V & \text{if } x \in [1/2 - \mu, 1 - \mu] \\ +\infty & \text{otherwise.} \end{cases}$$

For example, with $\mu := 3/4$, $d(x) := 1$ and $V := 1/2$, see figure 2-1 for a graph of $I(x)$. It is not surprising that the Gartner-Ellis conditions do not hold as $I(x)$ is not convex. Note that $I(x)$ is non-decreasing for $x \geq 0$; moreover, $I(x)$ is continuous where it is finite. We have

$$c^V I\left(\frac{1}{c}\right) = \begin{cases} +\infty & \text{if } c < \frac{1}{1-\mu} \\ (2 + c\,(2\mu - 1))^V & \text{if } c \geq \frac{1}{1-\mu}, \end{cases}$$

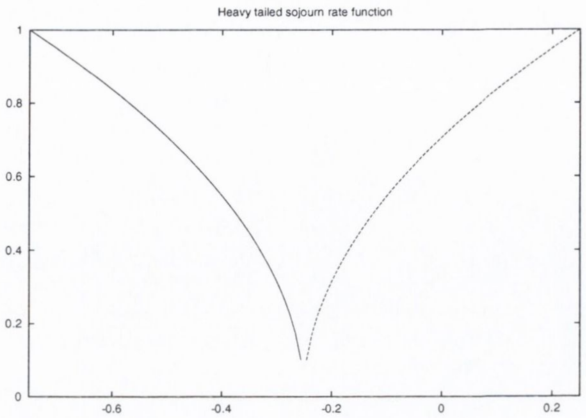

Figure 2-1: $I(x)$ vs. x for heavy tailed sojourn times.

and

$$\inf_{c>0} c^V I\left(\frac{1}{c}\right) = \left(2 + \frac{2\mu - 1}{1 - \mu}\right)^V. \tag{2.35}$$

As Z_t is bounded above by $1-\mu$, Proposition 2.4 applies. Hence (2.35) gives the exponential rate of decay of the tail of the stationary queue-length distribution:

$$\lim_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = \lim_{q\to\infty} \frac{1}{v(q)} \log \mathbb{P}[Q > q] = \left(2 + \frac{2\mu - 1}{1 - \mu}\right)^V.$$

2.4 Adjoint Processes and Queues with Fixed-Service Rate

We restrict to the case of a single server queue with fixed service-rate per unit time. We assume that the queue can serve a cell every $1/S$ clock ticks, and hence S cells can be served per clock tick. Note that the *Extension Hypothesis* is satisfied.

Although the results so far have been described in terms of the steady state queue length distribution, the original works of Lindley [46] and Loynes [47] were stated in terms of waiting times seen at the queue. Let U_n be the inter-arrival time between cell $n-1$ and n at the queue, and define $T_n := U_1 + \cdots + U_n$ to be the time of arrival of cell n. The waiting time w_n of cell n at the queue evolves via Lindley's recursion,

$$w_n = (w_{n-1} + 1/S - U_n) \vee 0 \qquad \text{for } n \geq 0. \tag{2.36}$$

If U_n is stationary then Lemma 1 of [47] proves that minimal solution to (2.36) is stationary, and that each element of this stationary solution is equal in distribution to the random variable W defined by

$$W := \sup_{n \geq 0} (n/S - T_n) . \tag{2.37}$$

All results from the previous section for Q hold for W. It is natural to ask whether there is a simple relationship between Q and W. In order to answer that question one needs the notion of inversely related processes.

The process which is inversely related to T_n is defined by $N_t := \sup\{n : T_n \leq t\}$. N_t is the number of cells that have arrived to the queue by time t. T_n and N_t satisfy the relation

$$\{T_n \leq t\} = \{N_t \geq n\}. \tag{2.38}$$

The steady state queue length distribution Q is given in this case by

$$Q := \sup_{t \geq 0}(N_t - St). \tag{2.39}$$

The first attempts to relate the large deviations of a counting process and its inverse were made by Glynn and Whitt [27], and by Russell [61], on the scales $v(t) := a(t) := t$. Glynn and Whitt then used this relation in [28] in order to characterise the decay rate of the tail of the queue length in terms of the rate of decay of the waiting times.

In [18], Duffield and Whitt extended these results to encompass scaling functions that are regularly varying. As an application they consider a single server queue with fixed service-rate under the hypotheses of Duffield and O'Connell in [17], and relate the rate of decay of the tail of Q to the rate of decay of the tail of W.

The following theorem proves a simple relationship for Q and W, and shows how the rate of decay of their tails are related.

Theorem 2.7 *Q is equal in distribution to SW, where W is defined in (2.37). Furthermore*

$$\lim_{q \to \infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = \frac{1}{S^{V/A}} \lim_{w \to \infty} \frac{1}{h(w)} \log \mathbb{P}[W > w],$$

if the limits exist.

PROOF To prove the first part it suffices to show that $\{W \geq w\} = \{Q \geq Sw\}$, for all $w \geq 0$. From (2.37) and (2.38), and as $\{T_n < 0\} = \emptyset$,

$$\begin{aligned} \{W \geq w\} &= \bigcup_{n \geq 0} \left\{T_n \leq \tfrac{n}{S} - w\right\} \\ &= \bigcup_{n \geq Sw} \left\{T_n \leq \tfrac{n}{S} - w\right\} \\ &= \bigcup_{n \geq Sw} \left\{N_{(n/S-w)} \geq n\right\}. \end{aligned}$$

Let $t = n/S - w$. Hence $n = S(t + w)$. $n \geq wS$ implies that $t \geq 0$, hence

$$\begin{aligned} \bigcup_{n \geq Sw} \left\{N_{n/S-w} \geq n\right\} &= \bigcup_{t \geq 0} \{N_t - St > Sw\} \\ &= \{Q \geq Sw\}. \end{aligned}$$

Therefore, in distribution, $Q = SW$. The second part requires only simple manipulation and the use of the scaling hypothesis. As $Q = SW$ in distribution

$$\lim_{q \to \infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = \lim_{q \to \infty} \frac{1}{h(q)} \log \mathbb{P}\left[W > \frac{q}{S}\right].$$

Multiplying by 1 we have that

$$\lim_{q \to \infty} \frac{1}{h(q)} \log \mathbb{P}\left[W > \frac{q}{S}\right] = \lim_{q \to \infty} \frac{h\left(\frac{q}{S}\right)}{h(q)} \frac{1}{h\left(\frac{q}{S}\right)} \log \mathbb{P}\left[W > \frac{q}{S}\right].$$

Thus, using the scaling hypothesis, we have

$$\lim_{q \to \infty} \frac{h\left(\frac{q}{S}\right)}{h(q)} \frac{1}{h\left(\frac{q}{S}\right)} \log \mathbb{P}\left[W > \frac{q}{S}\right] = \frac{1}{S^{V/A}} \lim_{q \to \infty} \frac{1}{h\left(\frac{q}{S}\right)} \log \mathbb{P}\left[W > \frac{q}{S}\right].$$

Hence, setting $w = q/S$, we get the result.

2.5 Statistical Multiplexing

Statistical multiplexing gain is of key interest to teletraffic engineers. This is where input processes are multiplexed before being fed through a single server queue. Economies of scale are then observed as the number of input processes is increased. For a general reference see Kelly [37].

One way to approach this issue is via the large number of lines asymptotic (see [8, 4, 13]). Here one considers homogeneous and heterogeneous superpositions of input processes which display long time-scale large deviation behavior. The service-rate, and buffer space, per input line are kept constant and the limit is taken as the number of inputs grows to infinity.

We shall consider statistical multiplexing with a finite number of lines. This allows us to consider sources which obey large deviation principles on different external scales and still deduce large deviation results.

Fix $L \in \mathbb{N}$ and consider a finite collection of sources

$$\{Z_t^i : i \in \{1, \dots, L\}, t \in T\},$$

where Z_t^i is the arrivals less service from source i at time t. For each $i \in \{1, \dots, L\}$, and each $t \in T$, we define the workload from source i up to time t to be

$$W_t^i := \sum_{j=-t}^{0} Z_j^i.$$

We also define the total workload up to time t to be

$$W_t := \sum_{i=1}^{L} W_t^i.$$

It is essential to this work that each source satisfies a LDP with the same internal scaling $a(t)$. If this is not the case, then it is not be possible to relate the large deviations of the multiplex to the large deviations of each of the sources.

LDP multiplexing hypothesis: There exists a scale $v(t)$ such that the sources satisfy a joint large deviation principle with good rate-function $J(\vec{x})$. That is, there exists a lower semi-continuous function $J : \mathbb{R}^L \to [0, \infty]$, whose level sets are compact, such that for all F closed in $\mathbb{R}^L$

$$\limsup_{t\to\infty} \frac{1}{v(t)} \log \mathbb{P}\left[\frac{(W_t^1, \dots, W_t^L)}{a(t)} \in F\right] \leq -\inf_{\vec{x}\in F} J(\vec{x}),$$

and for all G open in $\mathbb{R}^L$

$$\liminf_{t\to\infty} \frac{1}{v(t)} \log \mathbb{P}\left[\frac{(W_t^1, \dots, W_t^L)}{a(t)} \in G\right] \geq -\inf_{\vec{x}\in G} J(\vec{x}).$$

Theorem 2.8 *Under the LDP multiplexing hypothesis, $(W_t/a(t), v(t))$ satisfies a large deviation principle with good rate-function, $I(x)$, given by*

$$I(x) = \inf\{J(x_1, \dots, x_L) : x_1 + \dots + x_L = x\}$$

PROOF The function $\phi : \mathbb{R}^L \to \mathbb{R}$ defined by

$$\phi\left(W_t^1, \dots, W_t^L\right) := W_t^1 + \dots + W_t^L$$

is continuous as, for all G open in $\mathbb{R}$,

$$\phi^{-1}(G) := \{(x_1, \dots, x_L) : x_1 + \dots + x_L \in G\}$$

is open in $\mathbb{R}^L$. Therefore we can apply the contraction principle (see Theorem 6.4 of [43]), and the result follows immediately.

We shall investigate the case where the lines are independent, proving that they satisfy a joint large deviation principle and giving a form for the rate-function.

Source LDP hypothesis: For each $i \in \{1, \dots, L\}$, $v^i : \mathbb{R}_+ \to R_+$ is a non-decreasing function with $\lim_{t\to\infty} v^i(t) = \infty$. $(W_t^i/a(t), v^i(t))$ satisfies a large deviation principle with rate-function $I_i(x)$, whose level sets are compact.

We now demonstrate what happens to rate-function for W_t^i when the external scale is changed from $v^i(t)$ to $v^j(t)$.

Lemma 2.9 *Let $i \neq j$, if*

$$\lim_{t\to\infty} \frac{v^i(t)}{v^j(t)} =: \alpha$$

exists, then $(W_t^i/a(t), v^j(t))$ satisfies a large deviation principle with rate-function $\alpha I_i(x)$.

PROOF Under the *Source LDP hypothesis*, $(W_t^i/a(t), v^i(t))$ satisfies a large deviation principle. Hence for all F closed

$$\limsup_{t\to\infty} \frac{1}{v^i(t)} \log \mathbb{P}\left[\frac{W_t^i}{a(t)} \in F\right] \leq -\inf_{x\in F} I_i(x).$$

Therefore, multiplying by 1,

$$\limsup_{t\to\infty} \frac{1}{v^j(t)} \log \mathbb{P}\left[\frac{W_t^i}{a(t)} \in F\right] = \limsup_{t\to\infty} \frac{v^i(t)}{v^j(t)} \frac{1}{v^j(t)} \log \mathbb{P}\left[\frac{W_t^i}{a(t)} \in F\right].$$

Thus, using the existence of α and the assumed LDP,

$$\limsup_{t\to\infty} \frac{1}{v^j(t)} \log \mathbb{P}\left[\frac{W_t^i}{a(t)} \in F\right] \leq -\inf_{x\in F} \alpha I_i(x).$$

Similarly for all G open. $\alpha I_i(x)$ is lower semi-continuous as $I_i(x)$ is, and $\alpha I_i(x)$ has compact level sets as $I_i(x)$ does.

The following Lemma gives a form for the rate-function for $(W_t/a(t), v^j(t))$.

Lemma 2.10 *Under the source LDP hypothesis, let $\{W_t^i\}$ and $\{W_t^j\}$ be independent for all $i \neq j$. If for some $j \in \{1, 2, ..., M\}$, and for all $i \in \{1, 2, ..., M\}$, we have that*

$$\lim_{t\to\infty} \frac{v^i(t)}{v^j(t)} := \alpha^i,$$

then $(W_t/a(t), v^j(t))$ satisfies a large deviation principle with good rate-function, $I(x)$, given by

$$I(x) = \inf\left\{\alpha^1 I_1(x_1) + \cdots + \alpha^L I_L(x_L) : x_1 + \cdots + x_L = x\right\}.$$

PROOF By Lemma 2.9, $(W_t^i/a(t), v^j(t))$ satisfies a large deviation principle with good rate-function $\alpha^i I_i(x)$.

As $\{W_t^i\}$ and $\{W_t^j\}$ are independent for all $i \neq j$ it follows that $((W_t^1, \ldots, W_t^L)/a(t), v^j(t))$ satisfies a joint large deviation principle with good rate-function, $J(x)$, given by

$$J(x_1, \ldots, x_L) = \alpha^1 I_1(x_1) + \cdots + \alpha^L I_L(x_L).$$

The result follows applying Theorem 2.8.

2.6 Statistical Multiplexing Examples

2.6.1 Many-Scale Behavior

Consider the multiplex of two independent sources

$$\{Z_t^i : i \in \{1,2\}, t \in \mathbb{Z}\}.$$

Define Z_t^1 to be the heavy tailed sojourn sequence Z_t found in section 2.3.2. $(W_t^1/t, b(t)t^r)$ satisfies a large deviation principle with rate-function, $I_1(x)$, given by

$$I_1(x) = \begin{cases} (1-2(x+\mu))^r & \text{if } x \in [-\mu, 1/2-\mu] \\ (2(x+\mu)-1)^r & \text{if } x \in [1/2-\mu, 1-\mu] \\ +\infty & \text{otherwise.} \end{cases}$$

Define Y_t^2 to be Bernoulli random variables taking the values $\{0, B\}$ with $\mathbb{P}[Y_t^2 = B] = p$ and $\mathbb{P}[Y_t^2 = 0] = 1-p$. Let $\nu \in [Bp, B]$, and define Z_t^2 by

$$Z_t^2 := Y_t^2 - \nu.$$

Using the sCGF it is straight forward to show that $(W_t^2/t, t)$ satisfies a large deviation principle with rate-function, $I_2(x)$, given by

$$I_2(x) = \begin{cases} \left(\frac{B-x-\nu}{B}\right)\log\left(\frac{B-x-\nu}{1-p}\right) + \frac{x+\nu}{B}\log\frac{x+\nu}{p} - \log B & \text{if } y \in [-\nu, B-\nu] \\ \infty & \text{otherwise.} \end{cases}$$

See figure 2-2 for a graph of $I_2(x)$ with $B = 10$, $\nu = 5$, $p = 1/2$.

We now demonstrate how many-scale behavior can occur. The scale on which large deviations of the multiplex are observed depends upon the size of the deviation.

On the scale $v(t) = b(t)t^r$, by Lemma 2.10, we have that $(W_t/t, b(t)t^r)$ satisfies a large deviation principle with rate-function, $I(x)$, given by

$$I(x) = \inf\left\{I_1(x_1) + \infty I_2(x_2) : x_1 + x_2 = x\right\}.$$

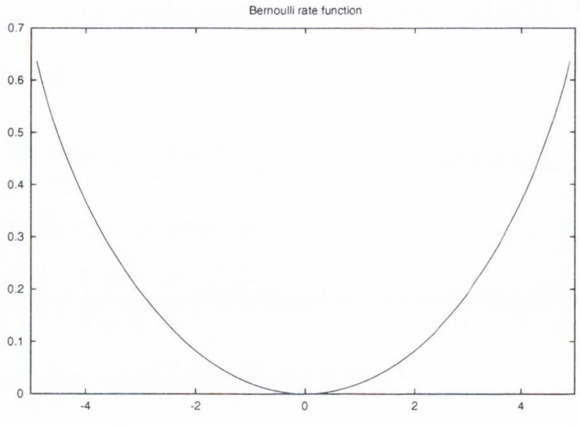

Figure 2-2: $I_2(x)$ vs. x for Bernoulli random variables.

Thus as $I_2(Bp-\nu)=0$, and with $\infty.0 := 0$, we see that

$$I(x) = I_1(x-(Bp-\nu)).$$

However, on the scale $v(t) = t$, by Lemma 2.10, we have that $(W_t/t, t)$ satisfies a large deviation principle with rate-function, $I(x)$, given by

$$I(x) = \inf\left\{0.I_1(x_1) + I_2(x_2) : x_1 + x_2 = x\right\}.$$

Thus we see that

$$I(x) = \begin{cases} I_2(x+\mu) & \text{if } x \in [-\nu-\mu, -\mu] \\ 0 & \text{if } x \in [-\mu, 1-\mu] \\ I_2(x-(1-\mu)) & \text{if } x \in [1-\mu, B-\nu-\mu] \\ \infty & \text{otherwise.} \end{cases}$$

Hence, on the scale $v(t) = t$, W_t^1 introduces a flat piece from $-\mu$ to $1-\mu$ into the middle of $I_2(x)$.

Thus if the deviation is in $[-\mu, 1-\mu]$ the scale on which large deviations are observed is $v(t) = b(t)t^r$. If the deviation is outside this interval, the scale on which large deviations are observed is $v(t) = t$.

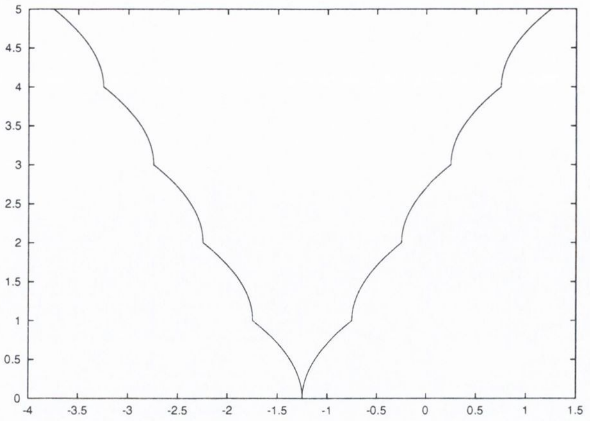

Figure 2-3: $I(x)$ vs. x for the multiplex of heavy tailed sojourn sources.

2.6.2 Concavity vs. Convexity

Consider L independent copies of Z_t^1 as defined in the previous section. By Lemma 2.10, on the scale $v(t) = b(t)t^r$,

$$I(x) = \inf \{I_1(x_1) + \cdots + I_1(x_L) : x_1 + \cdots + x_L = x\}.$$

Due to the concave nature of I_1 the way this infimum occurs is for one source to contribute as much as it can to the deviation. A second source contributes to the deviation only if the first source cannot create the entire deviation. For example, with $b(t) = 1$, $r = 1/2$, and $L = 5$, see figure 2-3 for a graph of $I(x)$ vs. x.

This is in stark contrast to the case where the rate-functions of the individual sources are convex. For example, if we have L independent copies of Z_t^2 as defined in the previous section then, by Lemma 2.10,

$$I(x) = LI_2\left(\frac{x}{L}\right).$$

Each of the sources contributes, in part, to cause the deviation.

Chapter 3

Large Deviations and Transient Multiplexing at a Buffered Resource

"I believe reflections bleed the sorrows of our souls."
Robb Flynn

3.1 Introduction

Large deviation asymptotics have been applied in many areas to calculate probabilities of exhaustion of a system resource as the size of the system becomes large. This body of work includes large buffer asymptotics in queueing theory with linear scalings such as Glynn and Whitt [27], and with nonlinear scalings such as Duffield and O'Connell [17]; large initial capital asymptotics in risk theory such as Martin-Lof [53] and Nyrhinen [57] and references therein; and also large number of lines asymptotics in queueing theory with linear number of lines/linear time scaling such as Botvitch and Duffield [4], Courcoubetis and Weber [8] and with nonlinear time scalings in Duffield [13].

The work on large number of lines asymptotics aims to capture the effect of statistical multiplexing, see Kelly [37] for a general reference. This is where superpositions of bursty traffic lead to a smoother, less bursty, multiplex and hence to economies of scale. The approach taken by other authors in their work on large number of lines asymptotics is to work with homogeneous and heterogeneous superpositions of traffic whose workload processes display long time-scale large deviations behavior. This connects their work with the work on large buffer asymptotics. Specifically Duffield [13] finds that in the case where the large deviation behavior of the underlying sources is on a nonlinear time scale, economies of scale can be much greater than in the linear case.

We will not approach the problem in this manner. We assume large deviation behavior on nonlinear scalings in the number of lines and consider only compact time intervals. This allows us to consider processes which are highly non-stationary in time and still deduce large deviation results. We have two applications in mind, one is the single server queue with many sources each of whom has an associated arrival and service process; the resource in question being the buffer space at the queue. The other application is to a version of the Cramer-Lundberg model in risk theory (for a general introduction to risk theory see Grandell [29]), here the resource is an insurance companies capital and the sources are the clients who make claims and pay premiums.

In section 3.2 we introduce the setup under which the results will be developed. In section 3.3 the main results are presented. In section 3.4 the notion of stability which was introduced in the early queueing literature (see Loynes [47]) is revisited in the current setting. In section 3.5 the results are considered in the presence of convex structure and a connection is made with work on effective bandwidths (see Kelly [37]). In section 3.6 references to the literature are given to conditions under which the main assumption of this work is true. In section 3.7 we present a range of examples to highlight the features of the approach adopted in this paper.

The first example illustrates how even in a simple i.i.d. setting the appropriate large deviation scale may be nonlinear. The second example displays many-scale behavior; the scale on which the large deviations are seen depend upon the size of the deviation. The third example is designed to show how the asymptotics of the two models, the single server queue and the Cramer-Lundberg model, can differ. It too displays many-scale behavior.

3.2 Setup

We consider L sources using a resource in the discrete time interval $[1, T]$ where T is finite. Ultimately we will take the limit as the number of sources using the resource becomes large. Each source i has an associated vector $X(i)$ in $\mathbb{R}^T$ which is almost surely finite. $X_t(i)$ describes the input imparted to the resource at time t by source i. We define the vector $Z(L) := X(1) + \cdots + X(L) \in \mathbb{R}^T$. $Z_t(L)$ is the total input imparted to the resource at time t from the first L sources. We consider two cases.

(I) The resource cannot grow with time. This is the case with queues where $X_t(i)$ represents the arrivals less service from source i at time t and the resource is the buffer space at the queue.

(II) The resource can grow with time. This is the case with the Cramér-Lundberg model of risk theory where $X_t(i)$ represents the claims made less premium received from customer i by an insurance company at time t. Here the resource is the company's capital which acts as a buffer to bankruptcy.

We define a scale to be a non-decreasing sequence of real numbers diverging to infinity.

We assume that the size of the resource, $\{B(L)\}$, grows as the number of sources attached to the resource grows. That is, there exists a scale $a(L)$ such that the size of the resource is given for $b > 0$ by $B(L) = a(L)b$. We keep b free as we shall be interested in what effect b, the resource space-scale, plays in the multiplexing.

We only consider discrete time although it is possible to deal with continuous time if a growth condition is placed on $X_t(\cdot)$ as a function of t such that for any compact interval $[1, T]$ it is sufficient to know the processes at some finite number of times within the interval, this sort of approach is taken by Duffield and O'Connell in [17] for large buffer asymptotics.

In case (I) the resource experiences overflow in the interval $[1, T]$ if the total arrivals less total service imparted to the resource by the sources exceeds $a(L)b$ in any interval $[r, s] \subseteq [1, T]$. That is if

$$\max_{1 \leq r \leq s \leq T} \sum_{t=r}^{s} Z_t(L) > a(L)b.$$

We define the maximum queue length operator

$$Q(\vec{y}) := \max_{1 \leq r \leq s \leq T} \sum_{t=r}^{s} y_t, \tag{3.1}$$

for $\vec{y} = (y_1, \ldots, y_T) \in \mathbb{R}^T$, so that overflow occurs if and only if

$$Q\left(\frac{Z(L)}{a(L)}\right) > b.$$

In case (II) the resource is exhausted in the interval $[1, T]$ if the total claims less total premium imparted to the resource by the sources exceeds $a(L)b$ in any interval $[1, s] \subseteq [1, T]$. That is if

$$\max_{1 \leq s \leq T} \sum_{t=1}^{s} Z_t(L) > a(L)b.$$

We define the worst negative bank balance operator

$$C(\vec{y}) := \max_{1\leq s\leq T} \sum_{t=1}^{s} y_t,$$

for $\vec{y} = (y_1, \dots, y_T) \in \mathbb{R}^T$, so that bankruptcy occurs if and only if

$$C\left(\frac{Z(L)}{a(L)}\right) > b.$$

We will calculate large deviation probabilities for these events in terms of an assumed underlying large deviations principle for $\{Z(L)\}$ as the number of sources using the resource tends towards infinity.

3.3 Main Results

The proofs for the results for cases (I) and (II) are almost identical so, although the results are stated in terms of both, they are only proved in case (I).

Assumption 3.1 *$\{Z(L)/a(L)\}$ satisfies a large deviation principle on the exterior scale $\{V(L)\}$ with good rate-function $I(\cdot)$. That is, there exists a lower semi-continuous function $I : \mathbb{R}^T \to [0, \infty]$, whose level sets are compact, such that for all F closed in $\mathbb{R}^T$*

$$\limsup_{L\to\infty} \frac{1}{V(L)} \log \mathbb{P}\left[\frac{Z(L)}{a(L)} \in F\right] \leq - \inf_{\vec{x}\in F} I(\vec{x}),$$

and for all G open in $\mathbb{R}^T$

$$\liminf_{L\to\infty} \frac{1}{V(L)} \log \mathbb{P}\left[\frac{Z(L)}{a(L)} \in G\right] \geq - \inf_{\vec{x}\in G} I(\vec{x}).$$

References to the literature giving sufficient conditions for this assumption to hold shall be presented in a later section. For a superb review of large deviation theory see Lewis and Pfister [43].

Lemma 3.1 *For each pair $[r,s]$ such that $1 \le r \le s \le T$ the sequence $\{S_{r,s}(L)/a(L)\}$, with $S_{r,s}(L) := \sum_{t=r}^{s} Z_t(L)$, satisfies a large deviation principle in $\mathbb{R}$ on the exterior scale $\{V(L)\}$ with good rate-function, $J_{r,s}(y)$, defined by*

$$J_{r,s}(y) := \inf\left\{ I(\vec{x}) : \sum_{t=r}^{s} x_t = y \right\}. \tag{3.2}$$

PROOF As the function $\phi(x_r, \ldots, x_s) = x_r + \cdots + x_s$ is continuous it follows directly from the contraction principle (see Theorem 6.4 of Lewis and Pfister[43]), and assumption 3.1, that the image measures $\mathbb{M}_L'[B] := \mathbb{P}\left[\frac{Z(L)}{a(L)} \in \phi^{-1}(B)\right] = \mathbb{P}\left[\frac{S_{r,s}(L)}{a(L)} \in B\right]$ satisfy a large deviation principle in $\mathbb{R}$ on the external scale $V(L)$ with good rate-function

$$J_{r,s}(y) := \inf\left\{ I(\vec{x}) : \sum_{t=r}^{s} x_t = y \right\}.$$

Theorem 3.2 *$\{Q(Z(L)/a(L))\}$ and $\{C(Z(L)/a(L))\}$ satisfy large deviation principles in $\mathbb{R}$ on the exterior scale $\{V(L)\}$ with good rate-functions $I_Q(y)$ and $I_C(y)$, respectively, which are defined by*

$$\begin{aligned} I_Q(x) &:= \inf\{I(\vec{y}) : Q(\vec{y}) = x,\ \vec{y} \in \mathbb{R}^T\} \\ I_C(x) &:= \inf\{I(\vec{y}) : C(\vec{y}) = x,\ \vec{y} \in \mathbb{R}^T\}. \end{aligned} \tag{3.3}$$

Moreover,

$$\inf_{x>b} I_Q(x) = \inf_{y>b} \min_{\{1\le r\le s\le T\}} J_{r,s}(y) \qquad \textit{and} \qquad \inf_{x>b} I_C(x) = \inf_{y>b} \min_{\{1\le s\le T\}} J_{1,s}(y).$$

PROOF The map $\vec{y} \to Q(\vec{y})$, defined in equation (3.1), is continuous. Hence we can apply the contraction principle (Theorem 6.4 of [43]) directly. Thus, $\{Q(Z(L)/a(L))\}$ satisfies an LDP with the rate function, $I_Q(x)$, given in equation (3.3). For the second part, note that

$$\begin{aligned} \inf\nolimits_{x>b} I_Q(x) &= \inf\{I(\vec{y}) : Q(\vec{y}) > b\} \\ &= \inf\{I(\vec{y}) : \max\nolimits_{1\le r\le s\le T} \textstyle\sum_{t=r}^{s} y_t > b\} \\ &= \min\nolimits_{\{1\le r\le s\le T\}} \inf\{I(\vec{y}) : \textstyle\sum_{t=r}^{s} y_t > b\} \\ &= \min\nolimits_{\{1\le r\le s\le T\}} \inf\nolimits_{y>b} J_{r,s}(y) \\ &= \inf\nolimits_{y>b} \min\nolimits_{\{1\le r\le s\le T\}} J_{r,s}(y), \end{aligned}$$

where $J_{r,s}(y)$ is defined in equation (3.2).

Corollary 3.3 *If $I_Q(\cdot)$ is continuous, the rate of decay of the probability of overflow satisfies,*

$$\lim_{L\to\infty}\frac{1}{V(L)}\log\mathbb{P}\left[Q\left(\frac{Z(L)}{a(L)}\right)>b\right]=-\inf_{x>b}I_Q(x).$$

If $I_C(\cdot)$ is continuous, the rate of decay of the probability of bankruptcy satisfies,

$$\lim_{L\to\infty}\frac{1}{V(L)}\log\mathbb{P}\left[C\left(\frac{Z(L)}{a(L)}\right)>b\right]=-\inf_{x>b}I_C(x).$$

PROOF Using the large deviation principle and the fact that for all a, $\mathbb{P}[Q(\vec{y})>a]\leq\mathbb{P}[Q(\vec{y})\geq a]$, we have that

$$\begin{aligned}-\inf_{x>b}I_Q(x) &\leq \liminf_{L\to\infty}\tfrac{1}{V(L)}\log\mathbb{P}\left[Q\left(\tfrac{Z(L)}{a(L)}\right)>b\right]\\ &\leq \limsup_{L\to\infty}\tfrac{1}{V(L)}\log\mathbb{P}\left[Q\left(\tfrac{Z(L)}{a(L)}\right)>b\right]\\ &\leq \limsup_{L\to\infty}\tfrac{1}{V(L)}\log\mathbb{P}\left[Q\left(\tfrac{Z(L)}{a(L)}\right)\geq b\right]\\ &\leq -\inf_{x\geq b}I_Q(x)\\ &= -\inf_{x>b}I_Q(x),\end{aligned}$$

where the last line uses the continuity of $I_Q(\cdot)$. Hence

$$\lim_{L\to\infty}\frac{1}{V(L)}\log\mathbb{P}\left[Q\left(\frac{Z(L)}{a(L)}\right)>b\right]=-\inf_{x>b}I_Q(x).$$

3.4 Stability

Traditionally with time stationary results for queues (see Loynes [47]) a queue is said to be stable if the mean arrival rate is less than the mean service rate. In this case there exists a minimum stationary sequence of random variables that satisfy the queueing recursion (Lindley's equation) and under a regularity condition every other solution of Lindley's equation couples to the stationary one in almost surely finite time.

Assumption 3.2

$$\lim_{L\to\infty}\frac{Z(L)}{a(L)}=Z\in\mathbb{R}^T,$$

in probability.

Let $Z = (z_1, \ldots, z_T)$ and, for all $1 \leq r \leq s \leq T$, define $M_{r,s} := \sum_{t=r}^{s} z_t$. Assumption 3.2 implies that, for all $1 \leq r \leq s \leq T$,

$$\lim_{L\to\infty} \frac{S_{r,s}(L)}{a(L)} = M_{r,s} \in \mathbb{R},$$

in probability. If $X(i)$ is stationary, then with $a(L) = L$, assumption 3.2 corresponds to a weak law of large numbers.

Lemma 3.4 *If $M_{r,s} > b$ for any $[r,s] \subseteq [1,T]$, then the probability of overflow does not decay exponentially on the scale $V(L)$. If $M_{1,s} > b$ for any $s \in [1,T]$, then the probability of bankruptcy does not decay exponentially on the scale $V(L)$.*

PROOF As $M_{r,s} > b$ for some $[r,s] \subseteq [1,T]$, we have that

$$\inf_{x>b} J_{r,s}(x) = J_{r,s}(M_{r,s}) = 0.$$

We know that

$$\liminf_{L\to\infty} \frac{1}{V(L)} \log \mathbb{P}\left[Q\left(\frac{Z(L)}{a(L)}\right) > b\right] \geq -\inf_{x>b} I_Q(x),$$

so that

$$\liminf_{L\to\infty} \frac{1}{V(L)} \log \mathbb{P}\left[Q\left(\frac{Z(L)}{a(L)}\right) > b\right] \geq -\inf_{x>b}\left(\min_{\{1\leq r\leq s\leq T\}} J_{r,s}(x)\right) = 0.$$

Similarly for the rate of decay of the probability of bankruptcy.

Assumption 3.3 *For each $[r,s] \subseteq [1,T]$ we have that $J_{r,s}(x)$ is non-decreasing to the right of $M_{r,s}$. Furthermore we assume there exists $\overline{m}_{r,s} \geq M_{r,s}$ such that for all $x > \overline{m}_{r,s}$ we know that $J_{r,s}(x) > 0$. We call this the monotone property.*

We note that if $I(\cdot)$ is strictly convex then $J_{r,s}(\cdot)$ is also strictly convex and hence satisfies the *monotone property* with $\overline{m}_{r,s} = M_{r,s}$.

If one proves the joint LDP in assumption 3.1 via the Gartner [24]–Ellis [20] theorem, then $I(\cdot)$ is automatically strictly convex. Other simple examples of where the *monotone* *property* holds but the underlying rate-function is not strictly convex come from models in statistical mechanics where flat spots in the rate-function around the mean correspond to phase transitions. We will provide an example adapted from Gantert [22] where the underlying rate-function is concave but still satisfies the *monotone property* with $\overline{m}_{r,s} = M_{r,s}$.

Lemma 3.5 *If $\overline{m}_{r,s} < b$ for all $[r,s] \subseteq [1,T]$, then the probability of overflow decays at least exponentially on the scale $V(L)$. If $\overline{m}_{1,s} < b$ for all $s \in [1,T]$, then the probability of bankruptcy decays at least exponentially on the scale $V(L)$.*

PROOF In this case we know that, for all $[r,s] \subseteq [1,T]$,

$$\inf_{x \geq b} J_{r,s}(x) = J_{r,s}(b) > 0.$$

Hence

$$\limsup_{L\to\infty} \frac{1}{V(L)} \log \mathbb{P}\left[Q\left(\frac{Z(L)}{a(L)}\right) \geq b\right] \leq -\inf_{x\geq b}\left(\min_{\{1\leq r\leq s\leq T\}} J_{r,s}(x)\right) < 0.$$

We note that if $X_t(i)$ is stationary in both t and i, then

$$M_{r,s} = (s-r+1)\mathbb{E}[X_1(1)],$$

and hence taking $r=1, s=T$ we see $M_{1,T} = T\mathbb{E}[X_1(1)]$. Letting T be large we end up with the well known Loynes [47] stability condition for a stable queue to exist, $\mathbb{E}[X_1(1)] < 0$; that is, that there are, on average, less arrivals than service.

3.5 Convexity and Effective Bandwidths

In the presence of convex structure large deviation theory becomes substantially more powerful. It becomes possible to deal with the Legendre-Fenchel transform of the rate-function, the scaled cumulant generating function (sCGF) $\lambda(\theta)$,

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

For a general reference to convex functions see Rockafellar [60] (specifically on convex conjugates section 12). A function which takes values in $[-\infty, +\infty]$ is *proper* if it never takes the value $-\infty$ and is not identically $+\infty$. Often large deviation principles are proved under conditions on the sCGF, the Gartner [24]–Ellis [20] conditions, which ensure that not only does the rate-function exist but also that it is strictly convex. There is however another way to use the convex structure. If we know a large deviation principle holds with a convex (but not necessarily strictly convex) rate-function and that the sCGF exists as a proper lower semi-continuous function, then we know that they are dual to each other (see Lemma 7.2 [43]).

Assumption 3.4 *The rate-function $I(\cdot)$ is convex,*

$$\lambda\left(\vec{\theta}\right) := \lim_{L\to\infty} \frac{1}{V(L)} \log \mathbb{E}\left[\exp\left(\frac{V(L)}{a(L)}\langle\vec{\theta}, Z(L)\rangle\right)\right],$$

exists as an extended real number for all $\vec{\theta} \in \mathbb{R}^T$ and is a proper lower-semi continuous function.

By Lemma 7.2 [43] $I(\cdot)$ and $\lambda(\cdot)$ are convex duals.

For each pair $[r, s] \subseteq [1, T]$ define $\lambda_{r,s}(\theta)$ for $\theta \in \mathbb{R}$ by

$$\lambda_{r,s}(\theta) := \lim_{L\to\infty} \frac{1}{V(L)} \log \mathbb{E}\left[\exp\left(\frac{V(L)}{a(L)}\theta S_{r,s}(L)\right)\right].$$

Lemma 3.6 *For all $[r, s] \subseteq [1, T]$, $J_{r,s}(\cdot)$ and $\lambda_{r,s}(\cdot)$ are convex dual to each other. Moreover $I_Q(\cdot)$ and $I_C(\cdot)$ satisfy*

$$\inf_{x>b} I_Q(x) = \inf_{x>b} \min_{\{1\leq r\leq s\leq T\}} \sup_{\theta}\{x\theta - \lambda_{r,s}(\theta)\}$$

and

$$\inf_{x>b} I_C(x) = \inf_{x>b} \min_{\{1\leq s\leq T\}} \sup_{\theta}\{x\theta - \lambda_{1,s}(\theta)\}.$$

PROOF $\lambda_{r,s}(\theta)$ is a proper lower semi-continuous function as $S_{r,s}(L)$ is a linear function of $Z(L)$. $I(\cdot)$ being convex implies that $J_{r,s}(\cdot)$ is convex thus, using Lemma 7.2 of [43], we see that $J_{r,s}(\cdot)$ and $\lambda_{r,s}(\cdot)$ are convex dual to each other. Thus, for each $[r, s] \subseteq [1, T]$,

$$J_{r,s}(x) = \sup_{\theta}\{x\theta - \lambda_{r,s}(\theta)\}.$$

Hence $I_Q(\cdot)$ satisfies

$$\inf_{x>b} I_Q(x) = \inf_{x>b} \min_{\{1 \le r \le s \le T\}} \sup_{\theta}\{x\theta - \lambda_{r,s}(\theta)\}.$$

Now for the connection to Effective Bandwidths. For a thorough review of effective bandwidth functions see Kelly [37]. The notion of Effective Bandwidth has become widely accepted as a measure of the resource requirements of traffic in a queueing network. Only linear scalings are dealt with in this context.

Consider a queue with fixed buffer-space per source, and fixed service-rate per source, being driven by independent sources. The stochastic behavior of the queue length arises from randomness within the sources. The objective is to find a service rate per source per unit time which ensures that the probability of overflow is below some prescribed threshold.

For the rest of this section $V(L) = L$, $a(L) = L$ and $X_t(i) := Y_t(i) - c$, where the $Y_.(\cdot)$ are almost surely non-negative, and $c > 0$ is the capacity of the resource per source per unit time. For each $[r, s] \subseteq [1, T]$, define $S^Y_{r,s}(L) := \sum_{t=r}^{s} Y_t(i)$.

Lemma 3.7 *If $\{Y_t(i)\}$ is stationary in t and if, for each $i \neq j$, $Y(i)$ and $Y(j)$ are mutually independent and identically distributed, then*

$$\lambda_{r,s}(\theta) = \log \mathbb{E}\left[\exp\left(\theta S^Y_{1,s-r+1}(1)\right)\right] - (s - r + 1)\theta c.$$

PROOF

$$\begin{aligned}
\lambda_{r,s}(\theta) &= \lim_{L\to\infty}\tfrac{1}{L}\log\mathbb{E}[\exp(\theta S_{r,s}(L))] \\
&= \lim_{L\to\infty}\tfrac{1}{L}\log\mathbb{E}\left[\exp\left(\theta S^Y_{r,s}(L)\right)\right] - (s-r+1)\theta c \\
&= \lim_{L\to\infty}\tfrac{1}{L}\log\mathbb{E}\left[\exp\left(\theta S^Y_{1,s-r+1}(1)\right)\right]^L - (s-r+1)\theta c \\
&= \log\mathbb{E}\left[\exp\left(\theta S^Y_{1,s-r+1}(1)\right)\right] - (s-r+1)\theta c.
\end{aligned}$$

The function

$$\alpha(\theta,s) := \frac{1}{\theta s}\log\mathbb{E}[\exp(\theta S^Y_{1,s}(1))]$$

is the *effective bandwidth* of a source (see Kelly [37]). Then, in this setting,

$$\lambda_{r,s}(\theta) = (s-r+1)\theta\alpha(\theta,s-r+1) - (s-r+1)\theta c.$$

Therefore in this scenario we have that the rate-functions for the probability of overflow, and the probability of bankruptcy, both satisfy

$$\inf_{x>b} I_Q(x) = \inf_{x>b} I_C(x) = \inf_{x>b}\min_{1\le s\le T}\sup_\theta\{\theta b - s\theta\alpha(\theta,s) + s\theta c\}.$$

For an article on measuring effective bandwidths and it's uses see Györfi *et al.* [31] and references therein.

3.6 Joint Large Deviation Principles

The main underlying assumption (assumption 3.1) has been the existence of a joint large deviation principle for the underlying sources, that is, a large deviation principle for their vectors. For a general reference to large deviation techniques see Dembo and Zeitouni [11]. Here we will try to illustrate the conditions under which this holds. Essentially we expect that it will hold when, for each $t \in [1,T]$, the partial sums of $X_t(\cdot)$ satisfy large deviation principles.

It is difficult to pin down a simple set of general mixing conditions for which the assumption is true. Ellis [20] proved a large deviation principle for random vectors under what would ultimately be called the Gartner-Ellis conditions. These conditions essentially amount to assuming the sCGF exists, is finite in a neighborhood of the origin and has steepness properties. There are plenty of examples where large deviation principles hold but these conditions are not true; see section 2.3 of Dembo and Zeitouni [11] for ones motivated from applied probability. In statistical mechanics any Ising model which allows a phase transition fails the steepness property. In heavy tailed distributions of risk theory (see for example T. Mikosch and A. V. Nagaev [54]) the sCGFs are not finite in the neighborhood of the origin.

There are other general mixing conditions under which large deviation principles can be deduced. See Bryc and Dembo [6] for conditions motivated by mixing conditions under which central limit theorems are proved and Lewis *et al.* [44] for a condition motivated by Gibbs measures of statistical mechanics.

It is certainly trivially true that if, for each $t \in [1,T]$ the partial sums of $X_t(\cdot)$ satisfy large deviation principles on the scales $a(\cdot)$ and $V(\cdot)$ with rate-functions $I_t(\cdot)$, and if $X_r(\cdot)$ and $X_s(\cdot)$ are independent for all $r \neq s$, then the partial sums of the $X(\cdot)$ satisfy a large deviation principle on the scales $a(\cdot)$ and $V(\cdot)$ with rate-function $I(\vec{x}) := I_1(x_1) + \cdots + I_T(x_T)$.

3.7 Examples

3.7.1 Example 1: Non-Linear Scale

This example presents a simple set of independent sources consisting of i.i.d. random variables which satisfy a large deviation principle on a nonlinear scale.

We consider fixed capacity (or premium) $c > 0$ per source per unit time and model the arrivals (or claims) process.

The behavior is described by a heavy tailed i.i.d. sequence adapted from Gantert [22]. Gantert has extended these results to include dependent random variables satisfying a mixing condition along the lines of that found in Bryc and Dembo [6]. Heavy tailed distributions are often studied in risk theory (see Mikosch and Nagaev [54] and references therein) and more recently in queueing theory (see Asmussen and Collamore [1]). For these models it is possible to get much more information than is available on a logarithmic scale. Interesting features are still displayed even after taking logs.

For each $t \in [1, T]$ we define the sequence $\{X_t(L)\}$ to be an independently and identically distributed sequence of random variables each with distribution equal to that of $Y_1(1) - c$, where,

$$\mathbb{P}[Y_1(1) \geq x] := d(x) \exp(-E(x)x^z),$$

$z \in (0, 1)$ and $d(\cdot)$ and $E(\cdot)$ are slowly varying, that is, for all $\eta \in (0, \infty)$,

$$\lim_{x \to \infty} \frac{d(\eta x)}{d(x)} = \lim_{x \to \infty} \frac{E(\eta x)}{E(x)} = 1,$$

(see Bingham *et al.* [3]). Define $M_{t,t} := \mathbb{E}[Y_t(1)] - c$. This example belongs to the class of semi-exponential distributions where all moments are finite but the cumulant generating function $\lambda(\theta)$ is infinite for all $\theta > 0$.

The sCGF for $X_t(\cdot)$ is not finite in a neighborhood of the origin so the Gartner-Ellis conditions are not satisfied. It is possible to show using a sub-additivity argument that a large deviation principle is satisfied on the internal scale $a(L) = L$ and external scale $V(L) = L$. The rate-function however is trivial: it is zero above it's mean and infinite below it. The scales on which the rate-function is non-trivial are $a(L) = L$ and $V(L) = E(L)L^z$. On these scales,

$$J_{t,t}(x) = \begin{cases} \infty & x < M_{t,t} \\ (x - M_{t,t})^z & x \geq M_{t,t}. \end{cases}$$

Note as $z \in (0, 1)$ that $J_{t,t}(\cdot)$ is concave, hence it is not surprising it fails the Gartner-Ellis conditions. We note also that in this case $\overline{m}_{t,t}$ as defined in section 3.4 can be set equal to $M_{t,t}$. For example, if $E(L) = d(L) = 1$, $c = 3$ and $z = 1/2$, then $M_{t,t} = -1$. See figure 3-1 for a graph of $J_{t,t}(\cdot)$.

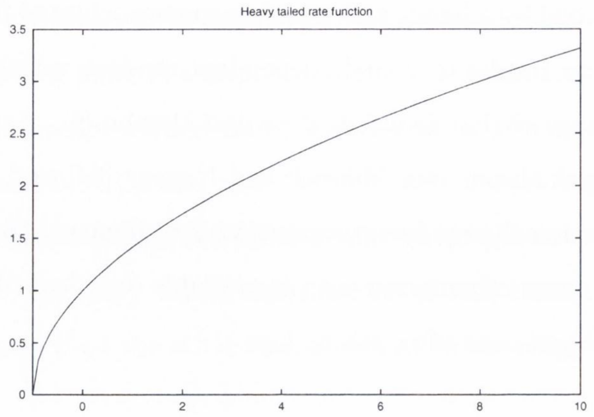

Figure 3-1: $J_{t,t}(y)$ vs. y for a heavy tailed random variables on the scale $V(L) = \sqrt{L}$.

By the comments at the end of the last section, on the scale $V(L) = E(L)L^z$ the partial sums of $X(\cdot)$ satisfy a large deviation principle with rate-function, $I(\cdot)$, given by

$$I(\vec{x}) = \begin{cases} \infty & \text{if } \ x_t < M_{t,t} \quad \text{for any } t \in [1,T] \\ \sum_{t=1}^{T}(x_t - M_{t,t})^z & \text{if } \ x_t \geq M_{t,t} \quad \text{for all } t \in [1,T]. \end{cases}$$

For each $[r,s] \subseteq [1,T]$, by Lemma 3.1, using the concavity of $J_{t,t}(x)$ for $x > 0$ and the fact that $J_{t,t}(x) = 0$ when $x = M_{t,t}$,

$$\begin{aligned} J_{r,s}(y) &= \inf\{I(\vec{x}) : \textstyle\sum_{t=r}^{s} x_t = y\} \\ &= \inf\{\textstyle\sum_{t=r}^{s} J_{t,t}(x_t) : \sum_{t=r}^{s} x_t = y\}, \\ &= J_{t,t}\left(y - (s-r)M_{t,t}\right). \end{aligned}$$

Assume that $M_{t,t} < 0$. As $J_{t,t}(y)$ is increasing for $y > 0$, $J_{t,t}(y) \leq J_{t,t}\left(y - (s-r)M_{t,t}\right)$ for $y > 0$ and all $1 \leq r \leq s \leq T$. Hence, assuming $M_{t,t} < 0$, on the scale $V(L) = E(L)L^z$,

$$\inf_{y>b} I_Q(y) = \inf_{y>b} \min_{\{1 \leq r \leq s \leq T\}} J_{r,s}(y) = \inf_{y>b} J_{t,t}(y) = (b - M_{t,t})^z,$$

for $b > 0$, and

$$\inf_{y>b} I_C(y) = \inf_{y>b} \min_{\{1 \leq s \leq T\}} J_{1,s}(y) = \inf_{y>b} J_{1,1}(y) = (b - M_{1,1})^z,$$

for $b > 0$.

The way the large deviations of the exhaustion of the resource occurs in this example is for one of the sources, at a single time, to place too great a demand on the resource. Hence in the queueing case, case (I), $\inf_{y>b} I_Q(y)$ is independent of T and, as $M_{r,r} = M_{s,s}$ for all $1 \leq r \leq s \leq T$, is independent of t.

In the risk theory case, case (II), $\inf_{y>b} I_C(y)$ is independent of T, but, as $M_{t,t} < 0$, it becomes more difficult as time progresses for a deviation to occur. Hence it is most likely to occur at time 1. The rate, however, is the same rate found in the queueing case, case (I).

3.7.2 Example 2: Many-Scale Behavior

The purpose of this example is to illustrate how many-scale behavior can arise. The scale on which large deviations are observed depends upon the size of the deviation.

Again we consider constant service rate (or premium), $c > 0$, per source per unit time. The stochastic driving force is a sequence of almost surely non-negative random vectors $Y(L)$ that describe the arrivals (or claims) process, that is $X_t(L) := Y_t(L) - c$. Consider time stationary sources which are made up of two independent parts. A large part, $U_t(L)$, which has no correlation across the sources, and a small part, $W_t(L)$, which is highly correlated across the sources. $Y_t(L) := U_t(L) + W_t(L)$.

We model $U_t(L)$ by i.i.d. Bernoulli random variables which take the values $\{0, A\}$ with $\mathbb{P}[U_1(1) = A] = p$ and $\mathbb{P}[U_1(1) = 0] = 1 - p$. The mean of $U_t(L)$ is $M^u_{t,t} := Ap$. The partial sums of $U_t(L)$ satisfy a large deviation principle on the scale $V(L) = L$ with rate-function $J^u_{t,t}(\cdot)$. This rate-function is simple to calculate by means of the sCGF,

$$J^u_{t,t}(y) = \begin{cases} \left(1 - \frac{x}{A}\right) \log\left(\frac{A-x}{1-p}\right) + \frac{x}{A} \log \frac{x}{p} - \log A & \text{if } y \in [0, A] \\ \infty & \text{otherwise.} \end{cases}$$

For example, with $A = 10$ and $p = 1/2$ see figure 3-2 for a graph of $J^u_{t,t}(\cdot)$. On the scale $V(L) = \sqrt{L}$ the rate-function for $U_t(\cdot)$ is trivial:

$$K^u_{t,t}(x) = \begin{cases} 0 & \text{if } x = Ap \\ \infty & \text{otherwise.} \end{cases}$$

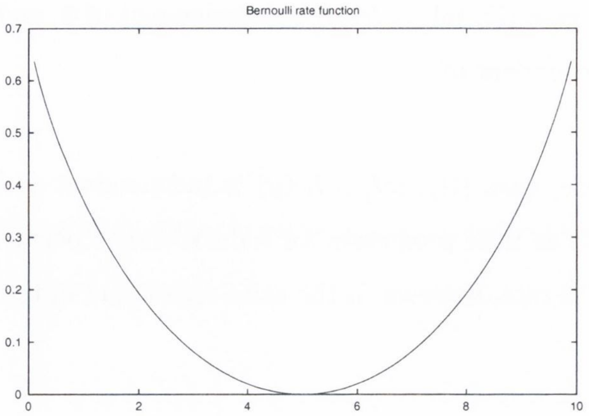

Figure 3-2: $J^u_{t,t}(y)$ vs. y for a Bernoulli random variables on the scale $V(L) = L$.

We model $W_t(L)$ as follows. Define the discrete heavy tailed distribution H by

$$\mathbb{P}[H \geq m] := e^{-\sqrt{m}},$$

for $m \in \mathbb{Z}^+$, and define $\{W_t(\cdot)\}$ to be a stationary sequence of two state random variables taking values in $\{0, B\}$, whose sojourn 'times' spent in the 0 and B states are distributed by an i.i.d. sequence with distribution H. On the scale $V(L) = L$ the partial sums of $W_t(\cdot)$ satisfy a large deviation principle with a rate-function, $J^w_{t,t}(\cdot)$, which is trivial,

$$J^w_{t,t}(x) = \begin{cases} 0 & \text{if } x \in [0, B] \\ \infty & \text{otherwise.} \end{cases}$$

In [62] Russell lays down a prescription to calculate the large deviations rate-function for the partial sums of a two state source which can be described in terms of the sojourn times it spends in the 'on' and 'off' states. Under technical conditions, Russell proves that the large deviations of a randomly sampled partial sums process is a simple functional of the large deviations of partial sums process itself, and of the large deviations of the random sampling. Hence setting random sampling to occur at the end of sojourn times, one can calculate the rate-function for the two state process by way of a functional of the rate-function for the sojourn times.

In [19] Duffy uses this prescription to calculate the large deviations rate-function, $K^w_{t,t}(\cdot)$,

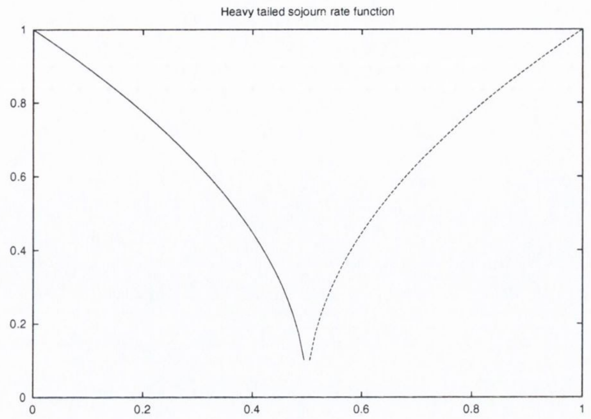

Figure 3-3: $K^w_{t,t}(y)$ vs. y for heavy tailed sojourn times on the scale $V(L) = \sqrt{L}$.

for the partial sums of $\{W_t(\cdot)\}$ on the scale $V(L) = \sqrt{L}$. $K^w_{t,t}(\cdot)$ is defined by

$$K^w_{t,t}(y) = \begin{cases} \left(1 - \frac{2y}{B}\right)^{1/2} & \text{if} \quad 0 \le y \le B/2 \\ \left(\frac{2y}{B} - 1\right)^{1/2} & \text{if} \quad B/2 \le y \le 1 \\ +\infty & \text{otherwise.} \end{cases}$$

Note that $\{W_t(\cdot)\}$ has mean $M^w_{t,t} := B/2$, as it's sojourn times spent in the 'on' and 'off' states have finite and equal, expectation. For example, with $B = 1$, see figure 3-3 for a graph of $K^w_{t,t}(\cdot)$.

On the scale $V(L) = L$, by the contraction principle,

$$J_{t,t}(y) = \inf\{J^u_{t,t}(y_1) + J^w_{t,t}(y_2) : y_1 + y_2 = y + c\}.$$

Hence, as $J^w_{t,t}(x) = 0$ for $x \in [0, B]$,

$$J_{t,t}(y) = \begin{cases} +\infty & \text{if } y < -c \\ J^u_{t,t}(y + c) & \text{if } y \in [-c, Ap - c] \\ 0 & \text{if } y \in [Ap - c, B + Ap - c] \\ J^u_{t,t}(y + c - B) & \text{if } y \in [B + Ap - c, B + A - c] \\ +\infty & \text{if } y > B + A - c. \end{cases}$$

As $J_{t,t}(\cdot)$ is convex and the sources are time independent, $J_{r,s}(\cdot)$ is defined by

$$J_{r,s}(y) := (s - r + 1) J_{t,t}\left(\frac{y}{s - r + 1}\right). \tag{3.4}$$

Note that $J_{r,s}(\cdot)$ only depends on r and s through $s-r$ as the sources are i.i.d. in time. Note also that, as $J_{t,t}=0$ if $y = Ap+B-c$,

$$J_{r,s}(y) = 0 \qquad \text{if} \qquad y = (s-r+1)(Ap+B-c). \tag{3.5}$$

In order to see many-scale behavior it is necessary to assume that $(Ap+B-c)>0$. Then the highly correlated part of the sources can cause a deviation on their own scale, despite any downward pull from the uncorrelated part less service rate (premium). By equation (3.5), assuming $(Ap+B-c)>0$, $J_{1,T}(y)$ is zero for $y=T(Ap+B-c)$, and hence takes the value zero for a larger y than any other $J_{r,s}(y)$ does. Moreover, as $J_{t,t}(\cdot)$ is convex and $(Ap+B-c)>0$,

$$(s-r+1)J_{t,t}\left(\frac{y}{s-r+1}\right) \leq J_{t,t}(y),$$

for all $y \geq 0$. Hence

$$J_{1,T}(y) \leq J_{r,s}(y) \qquad \text{for all } y \geq 0.$$

Thus, assuming $(Ap+B-c)>0$, on the scale $V(L)=L$,

$$\inf_{y>b} I_Q(y) = \inf_{y>b} \min_{\{1\leq r<s\leq T\}} J_{r,s}(y) = \inf_{y>b} J_{1,T}(y) = J_{1,T}(b),$$

for $b>0$, and

$$\inf_{y>b} I_C(y) = \inf_{y>b} \min_{\{1\leq s\leq T\}} J_{1,s}(y) = \inf_{y>b} J_{1,T}(y) = J_{1,T}(b),$$

for $b>0$.

For example if $A=10$, $p=1/2$, $B=1$, $c=5\frac{3}{4}$ and $T=10$, see figure 3-4 for a graph of $\inf_{y>b} I_Q(y)$ against b on the scale $V(L)=L$. Note that, by equation (3.5), the rate-function is zero for b below $T(Ap+B-c)=2\frac{1}{2}$. On this scale exhaustion of the resource is a concentration set if the resource space scale is below $2\frac{1}{2}$.

By the contraction principle, on the scale $V(L)=\sqrt{L}$,

$$J_{t,t}(y) = \inf\{K^u_{t,t}(y_1) + K^w_{t,t}(y_2) : y_1+y_2 = y+c\}.$$

$K^u_{t,t}(y)$ is infinite except at $M^u_{t,t} := Ap$ where it is zero, therefore

$$J_{t,t}(y) = K^w_{t,t}(y+c-Ap).$$

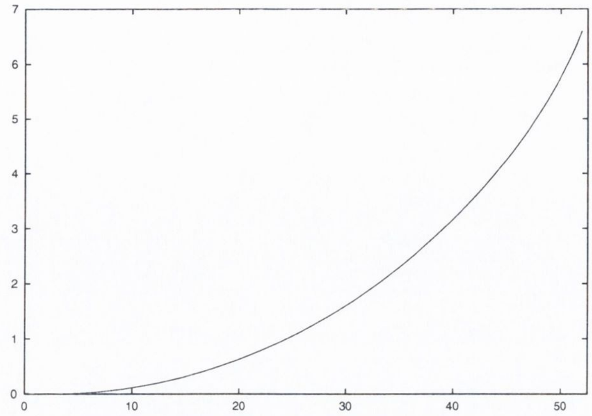

Figure 3-4: $\inf_{y>b} I_Q(y)$ vs. b on the scale L.

Note that $J_{t,t}(y)$ is infinite for $y > (Ap + B - c)$, as $K^w_{t,t}(x)$ is infinite for $x > B$.

By the comments at the end of the last section,

$$J_{r,s}(y) = \inf \left\{ \sum_{i=r}^{s} J_{i,i}(y_i) : \sum_{i=r}^{s} y_i = y \right\}.$$

Due to the concave nature of $J_{t,t}(\cdot)$, and using the fact that $J_{t,t}(B+Ap-c) = J_{t,t}(Ap-c) = 1$, we have that $J_{r,s}(y + (r-s+1)(Ap + \frac{B}{2} - c))$ equals

$$\begin{cases} -\lceil \frac{2y}{B} \rceil + J_{t,t}\left(y + \frac{B}{2} + Ap - c - \lceil \frac{2y}{B} \rceil \frac{B}{2}\right) & \text{if } (s-r+1)(-\frac{B}{2}) \leq y \leq 0 \\ +\lfloor \frac{2y}{B} \rfloor + J_{t,t}\left(y + \frac{B}{2} + Ap - c - \lfloor \frac{2y}{B} \rfloor \frac{B}{2}\right) & \text{if } 0 \leq y \leq (s-r+1)(\frac{B}{2}) \\ \infty & \text{otherwise.} \end{cases}$$

If $y > (s-r+1)(B + Ap - c)$, then $J_{r,s}(y) = +\infty$, and it is not possible for the sources to cause a large deviation on this scale. As $K^w_{t,t}(\cdot)$ is concave, we have the following structure: the minimum amount of time is used to cause the deviation. If it is possible for one time instance to cause all of the deviation, then due to the concavity this has a lower rate than sharing the deviation over time. This is in stark contrast to the convex case where the deviation is shared equally over time.

On the scale $V(L) = \sqrt{L}$, assuming $(Ap + B - c) > 0$,

$$\inf_{x \geq b} I_Q(x) = \inf_{x \geq b} \min_{\{1 \leq r \leq s \leq T\}} J_{r,s}(x) \qquad \text{and} \qquad \inf_{x \geq b} I_C(x) = \inf_{x \geq b} \min_{\{1 \leq s \leq T\}} J_{1,s}(x)$$

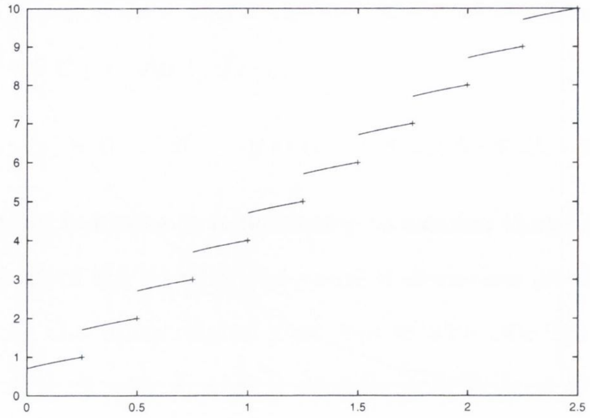

Figure 3-5: $\inf_{y \geq b} I_Q(y)$ vs. b on the scale $\sqrt{L}$.

both equal

$$\begin{cases} J_{t,t}(b) & \text{if } b \in (0, B + Ap - c] \\ 1 + J_{t,t}(b - (B + Ap - c)) & \text{if } b \in (B + Ap - c, 2(B + Ap - c)] \\ 2 + J_{t,t}(b - 2(B + Ap - c)) & \text{if } b \in (2(B + Ap - c), 3(B + Ap - c)] \\ \vdots & \\ (T - 1) + J_{t,t}(b - (T - 1)(B + Ap - c)) & \text{if } b \in ((T - 1)(B + Ap - c), T(B + Ap - c)] \\ \infty & \text{if } b > T(B + Ap - c), \end{cases}$$

for $b > 0$.

For example, if $A = 10$, $p = 1/2$, $B = 1$, $c = 5\frac{3}{4}$ and $T = 10$, see figure 3-5 for a graph of $\inf_{y \geq b} I_Q(y)$ vs. b on the scale $V(L) = \sqrt{L}$. Note that the rate-function is infinite for b above $T(B + Ap - c) = 2\frac{1}{2}$. On this scale exhaustion of the resource will not happen (in probability) if the resource space-scale is above $2\frac{1}{2}$.

Hence if $(B + Ap - c) > 0$ we observe many-scale behavior. The scale on which large deviations are observed depends upon the size of the deviation in question.

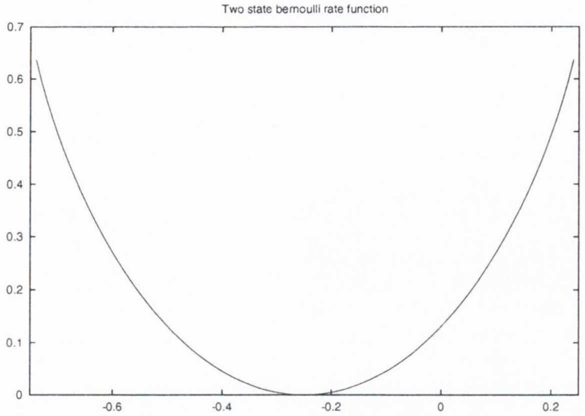

Figure 3-6: $J_{r,r}(y)$ vs. y.

3.7.3 Example 3: Non-Stationarity

The purpose of this example is to highlight the effect of non-stationarity. Many-scale behavior also appears. The rate of decay of probability of overflow and rate of decay of probability of bankruptcy differ on the slower scale.

We consider fixed service rate (or premium), c per source per unit time, and model the arrivals (or claims) process $\{Y_{\cdot}(\cdot)\}$, that is $X_t(L) := Y_t(L) - c$. At all times bar one the sources are uncorrelated. At one instant they are highly correlated.

$Y_t(L)$ will take one of two values, $\{0, A\}$, for all t and all L. Fix $t \in [1, T]$. At all times $r \neq t$ we model $Y_r(L)$ by independent Bernoulli random variables taking the values $\{0, A\}$ with $\mathbb{P}[Y_r(L) = A] = p$ and $\mathbb{P}[Y_r(L) = 0] = 1 - p$. $M_{r,r} = Ap - c$. The partial sums of $X_r(L)$ satisfy a large deviation principle, with non-trivial rate-function $J_{r,r}(\cdot)$, on the scale $V(L) = L$, where $J_{r,r}(y) = J^u_{t,t}(y + c)$, and $J^u_{t,t}(y)$ is defined in the previous example. With $c = 3/4$ and $A = 1$, see figure 3-6 for a graph of $J_{r,r}(y)$ vs. y. On the scale $V(L) = \sqrt{L}$ the rate-function is trivial, it is zero at $M_{r,r} = Ap - c$ and infinite elsewhere.

At time t we model the source by the heavy tailed sojourn source described in the previous example, setting $B := A$. On the scale $V(L) = L$ the rate-function for $X_t(\cdot)$ is trivial, it

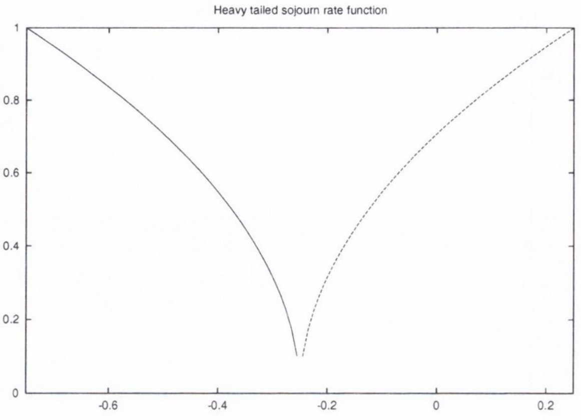

Figure 3-7: $J_{t,t}(y)$ vs. y.

is zero in $[-c, A-c]$ and infinite elsewhere. On the scale $V(L) = \sqrt{L}$ the rate-function is non-trivial, $J_{t,t}(y) = K^{w}_{t,t}(y+c)$, where $K^{w}_{t,t}(y)$ is defined in the previous example. With $c = 3/4$ and $B := A = 1$, see figure 3-7 for a graph of $J_{t,t}(y)$ vs. y.

If $t \notin [r,s]$, then on the scale $V(L) = L$,

$$J_{r,s}(y) = (s-r+1)J_{r,r}\left(\frac{y}{s-r+1}\right),$$

as $J_{r,r}(\cdot)$ is convex and the sources are time independent.

If $t \in [r,s]$, then on the scale $V(L) = L$,

$$J_{r,s}(y) = \begin{cases} (s-r)J_{r,r}\left(\frac{y}{s-r}\right) & \text{if } y \leq (Ap-c)(s-r)-c \\ 0 & \text{if } y \in [(Ap-c)(s-r)-c, (Ap-c)(s-r)+A-c] \\ (s-r)J_{r,r}\left(\frac{y-A+c}{s-r}\right) & \text{if } y \geq (Ap-c)(s-r)+A-c, \end{cases}$$

as at t a deviation in $[-c, A-c]$ has rate zero. Note that, whether t is in $[r,s]$ or not, $J_{r,s}(y)$ is infinite if $y > (s-r+1)(A-c)$, as it is not possible for the sources to create a deviation that large. Also note that $J_{t,t}(y) = 0$ for $y \in (0, A-c]$. If $M_{i,i} < 0$ for all $i \in [1,T]$, the minima

$$\min_{\{1\leq r\leq s\leq T\}} J_{r,s}(y) \qquad \text{and} \qquad \min_{\{1\leq s\leq T\}} J_{1,s}(y)$$

may occur at different values of $[r,s]$ for different values of the model parameters (A, p and c), but can be easily evaluated once the model parameters are known.

On the scale $V(L) = \sqrt{L}$, if $t \notin [r,s]$,

$$J_{r,s}(y) = \begin{cases} 0 & \text{if } y = (s-r+1)(Ap-c) \\ +\infty & \text{otherwise.} \end{cases}$$

If $t \in [r,s]$, then

$$J_{r,s}(y) = J_{t,t}\left(y - (s-r)(Ap-c)\right).$$

Assume that $M_{r,r} := Ap - c < 0$. As $J_{t,t}(y) = +\infty$ for $y > A - c$, if $(s-r) > \frac{(A-c)}{(Ap-c)}$, then $J_{r,s}(y) = +\infty$ for all $y > 0$ and all r, s.

Thus, on the scale $V(L) = \sqrt{L}$,

$$\inf_{y>b} I_Q(y) = \inf_{y>b} \min_{\{1\le r\le s\le T\}} J_{r,s}(y) = \inf_{y>b} J_{t,t}(y) = J_{t,t}(b),$$

as $J_{t,t}(y) \le J_{r,s}(y)$ for $y > 0$ and for all r, s. Note that $J_{t,t}(x)$ is infinite if $x > A - c$, thus this scale is only appropriate for deviations in the range $b \in (0, A-c]$.

$I_C(\cdot)$ satisfies

$$\inf_{y>b} I_C(y) = \inf_{y>b} \min_{\{1\le s\le T\}} J_{1,s}(y) = \inf_{y>b} J_{1,t}(y) = J_{1,t}(b),$$

as $J_{1,t}(y) \le J_{1,s}(y)$ for $y > 0$ and for all s. If $t-1$ is greater than $(A-c)/(Ap-c)$, then the heavy tailed sojourn effect is not enough to cause a large deviation on this scale as it cannot compensate for the downward pull of the earlier Bernoulli effects. In this case, no deviation is seen on the slower scale.

Chapter 4

On the Large Deviations of a Class of Stationary on/off Sources which Exhibit Long Range Dependence

"Perspective is lost in the spirit of the chase."

Dave Mustaine

4.1 Introduction

Long range dependence has been of interest to phenomenologists since Hurst published his 1951 paper [36] on a time-series of water levels of the Nile. His findings showed that statistical tests could imply a complicated, long range, correlation structure in this time-series. In the late 'sixties Mandelbrot and Wallis [49, 50, 51, 52], and Mandelbrot and Van Ness [48], proposed fractional Brownian motion, which is stationary and exhibits long range dependence, as a model for Hurst's time-series. In 1971 O'Connell [58] proposed an ARIMA model as an explanation of Hurst's phenomenon. In 1974 Klemes [40] objected strenuously to long range dependence as an explanation of Hurst's findings and demonstrated that non-stationarities, which seem physically more plausible than infinite memory, could lead to the observed phenomena.

Long range dependence has been of interest to teletraffic engineers since its proposal as an explanation of phenomena, similar to Hurst's, found in data-sets; for example in Leland *et al.* [42], Crovella and Bestavros [9], and Beran *et al.* [2]. Klemes' remarks were reiterated in the teletraffic setting by Duffield *et al.* [16, 14], where again it can be demonstrated that non-stationarities can lead to the observed phenomena. Fractional Brownian motion, which is long range dependent, has been proposed by several authors (see, for example, Norros [56] and Leyland *et al.* [42]) as a model of multiplexed Internet data.

From a phenomenology point of view, without addressing Klemes' remarks, fractional Brownian motion has two drawbacks: it is unbounded and it takes negative values. Boundedness arrives naturally in networks from bandwidth restrictions and negative arrivals have a dubious physical interpretation. We present a class of stationary two state sources whose sojourn times are distributed so that the source exhibits long range dependence. By using the term long range dependent source we mean that, given any $C > 0$, correlations within the source decay more slowly in time than $\exp(-Ct)$, for all sufficiently large t. We use techniques developed by Russell in [62] to relate the large deviations of the sojourn times to the large deviations of the sources themselves.

In section 4.2 we set up our basic notation and introduce Russell's [62] results. In section 4.3 we construct our class of two state sources which possess long range dependence and prove two simple lemmas to increase the ease of application of *Russell's random time-change.* In section 4.4 we present an example where the sojourn times spent in the 'on' and 'off' states are determined by an i.i.d. sequence with semi-exponential distribution. An explicit form is found for the source's rate-function, on a nonlinear scale.

4.2 Notation and Background

We follow a prescription set down by Russell in [62]. Let a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ be given. Let $\{X_t : t \in T\}$ be a stochastic process where T is $\mathbb{R}_+$ or $\mathbb{Z}_+$. For each $t \in T$, define the random function (the sample path) $S_t(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ by

$$S_t(x) := \int_0^{tx} X.\, d\lambda,$$

where λ is Lebesgue measure if $T = \mathbb{R}_+$, and $\lambda = \sum_{k=1}^{\infty} \delta_k$, where δ_k is Dirac measure at k, if $T = \mathbb{Z}_+$. We also define the partial sums process $\{S_t : t \in T\}$ by

$$S_t := S_t(1) = \int_0^t X.\, d\lambda.$$

Let $\{T_n : n \in \mathbb{Z}^+\}$ be a sequence of random times and $\{N_t : t \in T\}$ be its adjoint counting process, that is $N_t := \sup\{n : T_n \leq t\}$. For each $n \in \mathbb{Z}_+$ we define the sample path of T_n to be the function, $T_n(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$, defined by

$$T_n(x) := T_{\lfloor nx \rfloor}.$$

Similarly for each $t \in T$ we define the sample path of N_t to be the function, $N_t(\cdot) : \mathbb{R}_+ \to \mathbb{Z}_+$, defined by

$$N_t(x) := N_{tx}.$$

Large deviation results relating $\{T_n\}$ and $\{N_t\}$ have been proved by Duffield and Whitt in [18], by Glynn and Whitt in [27], and by Russell in [61].

We consider large deviation principles both for sample paths (SP-LDP) and for partial sums (1D-LDP; one dimensional LDP). See Lewis and Pfister [43] for a review of large deviation theory, and Dembo and Zeitouni [11] for a general reference to large deviation techniques. We follow Russell in considering our SP-LDPs in the topology of pointwise convergence (see Kelly [38] section 3).

We define a scale, $v : \mathbb{R}^+ \to [0, \infty]$, to be a non-decreasing function with $\lim_{t\to\infty} v(t) = \infty$ and define $\mathcal{M}(\mathbb{R}^+, \mathbb{R})$ be the space of right-continuous functions from $\mathbb{R}^+$ to $\mathbb{R}$ endowed with the topology of pointwise convergence.

Definition 4.1 *$\{S_t(\cdot)\}$ satisfies a SP-LDP on the scale $v(t)$ with rate-function $I : \mathcal{M}(\mathbb{R}_+, \mathbb{R}) \to [0, \infty]$ if $I(\zeta)$ is lower semi-continuous, has compact level sets,*

$$\lim_{t\to\infty} \frac{1}{v(t)} \log P\left[\frac{S_t(\cdot)}{t} \in F\right] \leq -\inf_{\zeta\in F} I(\zeta)$$

for all F closed in $\mathcal{M}(\mathbb{R}_+, \mathbb{R})$, and

$$\lim_{t\to\infty} \frac{1}{v(t)} \log P\left[\frac{S_t(\cdot)}{t} \in G\right] \geq -\inf_{\zeta\in G} I(\zeta)$$

for all G open in $\mathcal{M}(\mathbb{R}_+, \mathbb{R})$.

Definition 4.2 *$\{S_t\}$ satisfies an 1D-LDP on the scale $v(t)$ with rate-function $I^{(1)} : \mathbb{R} \to [0, \infty]$ if $I^{(1)}(x)$ is lower semi-continuous, has compact level sets,*

$$\lim_{t\to\infty} \frac{1}{v(t)} \log \mathbb{P}\left[\frac{S_t}{t} \in F\right] \leq -\inf_{x\in F} I^{(1)}(x)$$

for all F closed in $\mathbb{R}$, and

$$\lim_{t\to\infty} \frac{1}{v(t)} \log \mathbb{P}\left[\frac{S_t}{t} \in G\right] \geq -\inf_{x\in G} I^{(1)}(x)$$

for all G open in $\mathbb{R}$.

In conjunction with Lemma's 5.2 and 5.3, Theorem 5.1 in [62] proves that if $(S_{T_n}(\cdot), T_n(\cdot))$ satisfies a joint SP-LDP on the scale $v(t) := t$ (with rate-function $U(x, y)$ which satisfies

$U(x, 0) = \infty$ so that, on the scale of large deviations, the probability that the sampling process re-samples the same point indefinitely is zero, with rate ∞), then $(S_t(\cdot), N_t(\cdot))$ satisfies a SP-LDP on the scale $v(t) := t$. Moreover, in Theorem 5.10 he gives a simple relationship between the one dimension rate-functions $U^{(1)}(\cdot, \cdot)$ and $W^{(1)}(\cdot, \cdot)$, for (S_{T_n}, T_n) and (S_t, N_t),

$$W^{(1)}(x, y) = yU^{(1)}\left(\frac{x}{y}, \frac{1}{y}\right).$$

In his work he considers only scaling functions which are linear, that is $v(t) := t$. If we wish to describe a source that exhibits long range dependence we will require scaling functions which are nonlinear so that correlations within the source decay slower than exponentially. As it turns out the condition that we shall require is that $v(t)$ be regularly varying (see Bingham *et al.* [3] for a general reference).

$v(t)$ being regularly varying implies that $\lim_{t\to\infty} v(ct)/v(t)$ exists as an extended real number for all $c > 0$. As $v(t)$ is non-decreasing and diverging to $+\infty$ this implies that there exists $G \geq 0$ such that

$$\lim_{t\to\infty} \frac{v(ct)}{v(t)} = c^G,$$

for all $c > 0$.

With this hypothesis in mind the only alteration necessary to Russell's work is a trivial one in Lemma 5.3. The rest of his proofs remain unchanged with the final relationship between the one dimensional contractions changing slightly to be

$$W^{(1)}(x, y) = y^G U^{(1)}\left(\frac{x}{y}, \frac{1}{y}\right). \tag{4.1}$$

We call this transformation *Russell's random time-change.* We have the following diagram describing the relationship between the sample path large deviation principles and the one dimensional large deviation principles:

$$\begin{tikzcd}
\text{Sample path:} & (S_{T_n}, T_n) \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & (S_t, N_t) \arrow[d, Rightarrow] \\
\text{One dimensional:} & (S_{T_n}, T_n) & (S_t, N_t)
\end{tikzcd}$$

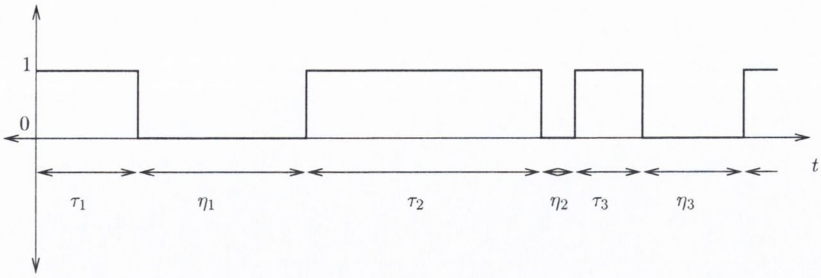

Figure 4-1: Construction of X_t.

If we start with a joint large deviations of the sample paths then we can deduce a relationship between the one dimensional large deviations of (S_{T_n}, T_n) and (S_t, N_t).

4.3 A Class of Stationary On/Off Sources

Let $\{\tau_i\}$ and $\{\eta_i\}$ be two stationary sequences of random variables taking values in T. $\{\tau_i\}$ are the sojourn times spent by a source in the 'on' state and $\{\eta_i\}$ are the sojourn times spent by a source in the 'off' state. We define the source's activity X_t at time $t \in T$ to be

$$X_t := \begin{cases} 0 & \text{if} \quad \tau_1 + \eta_1 + \cdots + \tau_k \leq t < \tau_1 + \eta_1 + \cdots + \tau_k + \eta_k \\ 1 & \text{if} \quad \tau_1 + \eta_1 + \cdots + \eta_k \leq t < \tau_1 + \eta_1 + \cdots + \eta_k + \tau_k, \end{cases}$$

see figure 4-1 for an illustration.

Define the processes $S_n^\tau := \sum_{i=1}^n \tau_i$, $S_n^\eta := \sum_{i=1}^n \eta_i$, and $T_n := S_n^\tau + S_n^\eta$. S_n^τ is the total time spent in the 'on' state after n 'on' periods, S_n^η is the total time spent in the 'off' state after n 'off' periods and T_n is the time after n 'on' and 'off' periods. We note that $S_{T_n} = S_n^\tau$.

Assumption 4.1 *$(S_{T_n}(\cdot), T_n(\cdot))$ satisfies a joint SP-LDP on the regularly varying scale $v(n)$.*

We now prove two simple lemmas which allow us to relate the large deviations of (S_{T_n}, T_n) and (S_n^τ, S_n^η).

Lemma 4.1 *$(S_{T_n}(\cdot), T_n(\cdot))$ satisfies a joint SP-LDP if and only if $(S_n^\tau(\cdot), S_n^\eta(\cdot))$ satisfies a joint SP-LDP.*

PROOF As $(x,y) \to (x, x+y)$ and $(x,y) \to (x, x-y)$ are continuous functions in the topology of pointwise convergence it follows directly from the contraction principle (see theorem 6.4 of [43]) that $(S_{T_n}(\cdot), T_n(\cdot)) = (S_n^\tau(\cdot), S_n^\tau(\cdot) + S_n^\eta(\cdot))$ satisfies a joint SP-LDP if and only if $(S_n^\tau(\cdot), S_n^\eta(\cdot))$ satisfies a joint SP-LDP.

Lemma 4.2 *(S_{T_n}, T_n) satisfies a joint 1D-LDP with rate-function $U^{(1)}(\cdot,\cdot)$ if and only if (S_n^τ, S_n^η) satisfies a joint 1D-LDP with rate-function, $I^{(1)}(\cdot,\cdot)$, given by*

$$I^{(1)}(x,y) = U^{(1)}(x, x+y).$$

PROOF Define $f : (x,y) \to (x, x-y)$. As f is continuous and (S_{T_n}, T_n) satisfies a joint 1D-LDP it follows directly from the contraction principle that (S_n^τ, S_n^η) satisfies a joint 1D-LDP with rate-function, $I^{(1)}(\cdot,\cdot)$, given by

$$\begin{aligned} I^{(1)}(x,y) &= \inf\{U^{(1)}(a,b) : f(a,b) = (x,y)\} \\ &= \inf\{U^{(1)}(a,b) : (a, a-b) = (x,y)\} \\ &= U^{(1)}(x, x+y). \end{aligned}$$

Similarly as $g : (x,y) \to (x, x+y)$ is continuous we have the converse.

Under assumption 4.1 we can apply *Russell's random time-change* and see that (S_t, N_t) satisfies a 1D-LDP on the scale $v(t)$ with rate-function

$$W^{(1)}(x,y) = y^G U^{(1)}\left(\frac{x}{y}, \frac{1}{y}\right) = y^G I^{(1)}\left(\frac{x}{y}, \frac{1}{y} - \frac{x}{y}\right).$$

In order to get the large deviations for $\{S_t\}$ all that is left to do is to contract out the effect of $\{N_t\}$. By the contraction principle $\{S_t\}$ satisfies a large deviation principle with rate-function $K^{(1)}(\cdot)$ given by,

$$K^{(1)}(x) = \inf_{y \in \mathbb{R}_+} W^{(1)}(x, y).$$

Note that if $\{\tau_i\}$ and $\{\eta_i\}$ are independent of each other, and if $\{S_n^\tau\}$ and $\{S_n^\eta\}$ satisfy 1D-LDP's on the scale $v(n)$ with rate-functions $I^\tau(\cdot)$ and $I^\eta(\cdot)$ respectively, then (S_n^τ, S_n^η) satisfies a joint large deviation principle with rate-function $I^{(1)}(x, y) = I^\tau(x) + I^\eta(y)$. Hence, by Lemma 4.2, $U^{(1)}(\cdot, \cdot)$ is given by $U^{(1)}(x, y) = I^\tau(x) + I^\eta(y - x)$. Applying *Russell's random time-change formula* (4.1) we get

$$W^{(1)}(x, y) = y^G I^\tau\left(\frac{x}{y}\right) + y^G I^\eta\left(\frac{1}{y} - \frac{x}{y}\right).$$

We now have the following diagram describing the relationship between the SP-LDPs and 1D-LDPs:

$$\begin{tikzcd}
\text{SP-LDP:} & (S_n^\tau, S_n^\eta) \arrow[r, Leftrightarrow] & (S_{T_n}, T_n) \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & (S_t, N_t) \arrow[d, Rightarrow] & \\
\text{1D-LDP:} & (S_n^\tau, S_n^\eta) \arrow[r, Leftrightarrow] & (S_{T_n}, T_n) & (S_t, N_t) \arrow[r, Rightarrow] & (S_t)
\end{tikzcd}$$

If we start with a joint large deviations of the sample paths (S_n^τ, S_n^η), then we can deduce a relationship between the joint one dimensional large deviations of (S_n^τ, S_n^η) and the one dimensional large deviations of $\{S_t\}$.

In order to use *Russell's random time-change* we must prove that $(S_n^\tau(\cdot), S_n^\eta(\cdot))$ satisfies a joint SP-LDP. In section two of [62], Russell proves a SP-LDP on the scale $v(t) := t$ under the assumption of a mixing condition adapted from Lewis *et al.* [44]. This condition does not move to the case of more general scalings as it relies on the use of the sub-additivity lemma.

4.4 Example: Semi-Exponential Tails

Let $\{\tau_i\}$ and $\{\eta_i\}$ be independent sequences of i.i.d. random variables with

$$\mathbb{P}[\tau_1 \geq x] := \mathbb{P}[\eta_1 \geq x] := a(x)\exp(-b(x)x^r),$$

where $r \in (0,1)$ and, $a(\cdot)$ and $b(\cdot)$ are slowly varying (see Bingham *et al.* [3]), that is

$$\lim_{x\to\infty} \frac{a(cx)}{a(x)} = \lim_{x\to\infty} \frac{b(cx)}{b(x)} = 1,$$

for all $c > 0$. Note that all the moments of τ_1 are finite but the cumulant generating function does not exist in a neighborhood of the origin. Define $M := \mathbb{E}[\tau_1]$.

In [23] Gantert proves a SP-LDP, in the projective limit topology, for $\{S_n^\tau(\cdot)\}$ on the scale $v(n) := b(n)n^r$ with a rate-function that has compact level sets. By the characterisation theorem, Theorem 1.4.1 of Bingham *et al.* [3], $v(x) := b(x)x^r$ is a regularly varying function. Hence, by Lemma 4.1, $(S_{T_n}(\cdot), T_n(\cdot))$ satisfies a joint SP-LDP on the scale $v(n)$. Thus the LDP and scaling hypotheses are satisfied and we may use *Russell's random time change.*

In [22] Gantert proves that $\{S_n^\tau\}$ satisfies a 1D-LDP on the scale $v(n)$ with rate-function $I^\tau(x)$, given by

$$I^\tau(x) = I^\eta(x) = \begin{cases} (x-M)^r & \text{if } x \geq M \\ \infty & \text{if } x < M. \end{cases}$$

$I(x)$ has compact level sets. It is possible to get more precise expansions for the tail probabilities of sums of these random variables, see Nagaev [55], when one is not just interested in logarithmic asymptotics. The exponential rate suffices for our needs. This is quite an unusual 1D-LDP, the rate-function, $I(x)$, is not convex. This is as large deviations are caused by the tail of individual random variables rather than aggregate behavior of sums.

By Lemma 4.2, (S_{T_n}, T_n) satisfies a joint 1D-LDP on the scale $v(n)$ with rate-function, $U^{(1)}(x,y)$, given by

$$U^{(1)}(x,y) = I^\tau(x) + I^\eta(y-x),$$

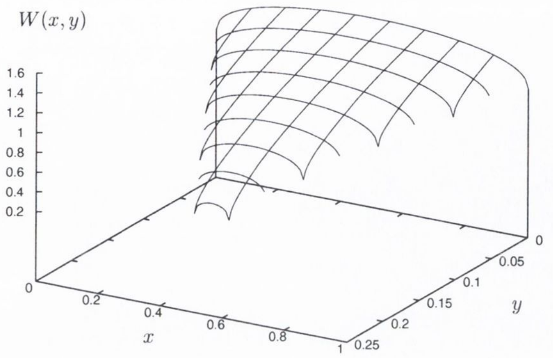

Figure 4-2: Rate function for (S_t, N_t) on the scale $v(t) := \sqrt{t}$.

that is

$$U^{(1)}(x,y) = \begin{cases} \infty & \text{if } x < M \text{ or } y - x < M \\ (x-M)^r + (y-x-M)^r & \text{otherwise.} \end{cases}$$

By *Russell's random time-change* (4.1), (S_t, N_t) satisfies a 1D-LDP on the scale $v(t)$ with rate-function

$$W^{(1)}(x,y) = y^r U^{(1)}\left(\frac{x}{y}, \frac{1}{y}\right),$$

that is

$$W^{(1)}(x,y) = \begin{cases} \infty & \text{if } \frac{x}{y} < M \text{ or } \frac{1}{y} - \frac{x}{y} < M \\ (x-My)^r + (1-x-My)^r & \text{otherwise.} \end{cases}$$

For a graph of $W^{(1)}(x,y)$, with $a(x) = b(x) = 1$ and $r = 1/2$, see figure 4-2.

We now wish to contract down to remove the y dependence in order to evaluate $K^{(1)}(x)$, the rate-function for $\{S_t\}$. As $U^{(1)}(x,y)$ is concave in both its arguments, is increasing at its left boundary, and is decreasing at its right boundary, its minimum is attained at one or other of its boundaries. Hence

$$K^{(1)}(x) = \begin{cases} (1-2x)^r & \text{if } 0 \le x \le 1/2 \\ (2x-1)^r & \text{if } 1/2 \le x \le 1 \\ \infty & \text{otherwise.} \end{cases}$$

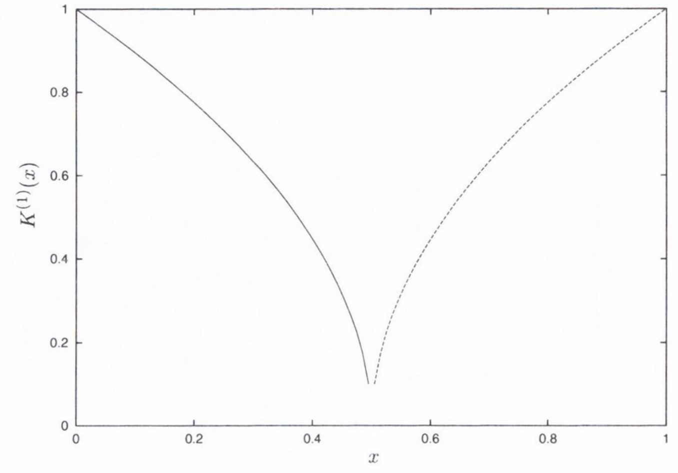

Figure 4-3: Rate function for $\{S_t\}$ on the scale $v(t) := \sqrt{t}$.

Note that $K^{(1)}(x)$ is zero at $x = 1/2$. This is as the mean sojourn times spent in both the on, and off, states are finite and equal. $K^{(1)}(x)$ is graphed, with $a(x) = b(x) = 1$ and $r = 1/2$, in figure 4-3. We note that $K^{(1)}(x)$ has non-convex structure, and that this is quite unusual for a large deviation rate-function.

Chapter 5

Distribution-free Confidence Intervals for Measurement of Effective Bandwidth

"Times fun when you're having flies."

Kermit the Frog

5.1 Introduction

The notion of effective bandwidth has become widely accepted as a measure of the resource requirements of bursty traffic in queueing networks. Intuitively, the effective bandwidth of a traffic source at a given network resource determines the quantity of resource capacity which must be reserved for it in order to achieve a specified rate of data-loss. This quantity depends on the statistical properties of the traffic source, on the properties of other traffic which may be sharing the resource in question, and on the nature of the resource itself (for example, buffered or unbuffered). It has been realised, through the work of a number of authors, that the complex relationships between these different factors can be unravelled using a family of large deviation limit results for the data-loss probability. These results lead to the *effective bandwidth function*, as defined by F. Kelly [37].

The effective bandwidth σ of a stationary stochastic traffic source is given by

$$\sigma(s,t) := \frac{1}{st} \log \mathbb{E}e^{sX(t)},$$

where $X(t)$ is a random variable representing the amount of data generated by the source during intervals of length t. The parameter s in this definition is an *inverse space scale*, that is, $1/s$ may be measured in units of bits, bytes, or cells. To illustrate the significance of this function, imagine a buffered resource which is shared by L stationary, independent, and statistically identical traffic sources. If the resource has a capacity of Lc units of data per unit time, and a buffer to hold Lb units of data, then the steady-state rate of data-loss R_L satisfies

$$\lim_{L\to\infty} \frac{1}{L} \log R_L = \sup_{t>0} \inf_{s\geq 0} \Big(t\sigma(s,t) - tc - b \Big).$$

Here σ is the effective bandwidth function of each identical source. Results of this type, and generalisations to the case of non-identical sources, have been proved by A. Simonian and J. Guibert [63], D. D. Botvich and N. G. Duffield [4], and C. Courcoubetis and R. Weber [8]. More details on the origin of effective bandwidth and its uses can be found in [37].

Given a stochastic model of a traffic source, the associated effective bandwidth function can be calculated more-or-less easily from the model's parameters (see [37] for a number of examples and references to the literature). For many purposes, however, it may be more practical to determine effective bandwidths directly from traffic measurements, thereby eliminating the need to fit a stochastic model to recorded data. N. Duffield *et al.* have presented arguments for this approach in [17]. Measurements of effective bandwidth may be useful both for off-line characterisation of traffic and for real-time control of multiplexing systems. An analysis of recorded traffic traces in terms of their measured effective bandwidths has been carried out by R. Gibbens in [25], while algorithms for connection admission control in connection-oriented networks, based on measurements of effective bandwidth, have been proposed by G. de Veciana *et al.* [10] and J. T. Lewis *et al.* [45]. A measurement-based approach to admission control and resource pricing, which is motivated by effective bandwidths but does not measure them directly, is described in [26] and [7].

In this paper we address the extent of sampling error in measured values of effective bandwidth. Sampling error is not a critical issue in off-line traffic characterisation, but assumes greater importance when measurements are used for the purpose of dynamic resource allocation. For example, M. Grossglauser and D. Tse [30] have studied the impact of sampling error on a resource operating with a measurement-based admission control system. In this setting, under-estimation of resource requirement causes the admission control to accept too many new connections, leading to violation of the data-loss target, while over-estimation leads to a reduction in both data-loss and utilisation. Grossglauser and Tse observe that the negative effects on data-loss of under-estimation exceed the positive effects of over-estimation, so that on the average the effect of sampling error is to increase the data-loss rate. The use of a certainty-equivalent point-estimate of resource requirement can therefore be expected to yield an average rate of data-loss *somewhat in excess* of the desired target. [30] describes an extreme case (not based on effective bandwidths), in which a simple system using point-estimates misses its performance target by two orders of magnitude. An important consideration is the fact that the large deviation results in which the effective bandwidth function has its origins are intended to control the frequencies of extremely rare events, whose probabilities may be of the order of 10^{-6} or less. At this level of likelihood even small sampling errors can have a significant impact; the $1 - 10^{-6}$ quantile for an estimator of effective bandwidth may be considerably larger than its mean.

Interval-estimates of bandwidth requirement are therefore desirable: if the target data-loss rate is 10^{-6}, then a $1-10^{-6}$ upper confidence limit for bandwidth requirement can be used safely as a basis for resource allocation. Approximate confidence intervals can be obtained from a Gaussian approximation in the usual manner, but this approach does not seem appropriate here due to the very low likelihood levels which are of interest. Instead we turn to concentration inequalities designed to provide rigorous upper bounds on the probabilities of rare events. Hoeffding's inequality is particularly attractive for our problem: to use it we require only an upper bound on the random variables of interest, and this can obtained directly, without further measurement, if traffic sources declare a peak rate or other token bucket constraint.

Theorem 5.1 (W. Hoeffding [35]) *Let $Z_1, \dots, Z_n$ be independent bounded random variables such that $Z_k \in [a_k,\ b_k]$ with probability one. Then for any $t > 0$,*

$$\mathbb{P}\Big(|\sum_{k=1}^{n}(Z_k - \mathbb{E}Z_k)| \geq t\Big) \leq 2\exp\Big(-2t^2/\sum_{k=1}^{n}(b_i - a_i)^2\Big).$$

A succinct proof of theorem 5.1 is given in [12]. S. Floyd uses Hoeffding's inequality in [21] to obtain an upper bound on the effective bandwidth of an aggregate of traffic sources. In section 5.2 we use it to compute confidence intervals for the effective bandwidth of a single source, based on measurements of source activity. The confidence limits can be chosen to converge almost surely to the true value of σ as the sample size increases, assuming that the source is stationary. Alternatively they can be chosen to provide robustness against violations of the stationarity hypothesis. Confidence intervals for the effective bandwidth of aggregate traffic are obtained in section 5.3.

For our purposes, the use of Hoeffding's inequality has two draw-backs. It requires independent observations of source behavior; and it becomes tight only at large sample sizes. In practice measurements of the activity of a source made at different times cannot be assumed completely independent, although they may be approximately so if the elapsed time between measurements is large. Taking widely-separated measurements will of course increase the time required to obtain a narrow confidence interval. The numerical results in sections 5.2 and 5.3 show that the sample size required to obtain a useful confidence interval may be very large, so that the approach we describe here may not be practical for small data sets.

5.2 Effective Bandwidth of a Single Source

Throughout this section we let $X(t)$ be a random variable representing the amount of work generated by a stationary stochastic traffic source during intervals of length t. We assume that $X(t)$ takes values between 0 and some upper limit $p(t) > 0$, with probability one. When the peak rate P of the source is known we may take $p(t) = Pt$; more generally, if the source is policed by a token bucket with bucket size b and token fill rate c, then $X(t)$ cannot exceed $p(t) = ct + b$. In the case of multiple token bucket constraints, we set

$$p(t) = \min_i (c_i t + b_i),$$

where (b_i, c_i), $i = 1, 2, \ldots$, are the token bucket parameters.

Our aim is to estimate the value of the effective bandwidth function

$$\sigma(s,t) := \frac{1}{st} \log \mathbb{E}e^{sX(t)},$$

for given s and t, from independent observations $X(t,1), \ldots, X(t,n)$ of $X(t)$. Thus each $X(t,k)$, $k = 1, \ldots, n$, is assumed to have the same distribution as $X(t)$. Given $r > 0$, and setting $q := r - \log 2$, we construct a $1 - e^{-q}$ confidence interval for $\sigma(s,t)$ as follows. Let ϕ be the value of the moment generating function of $X(t)$ at s, and let $\phi(n)$ be the weighted average

$$\phi(n) := \sum_{k=1}^{n} w(k,n) e^{sX(t,k)},$$

where the weights $w(k,n)$ satisfy $0 \leq w(k,n) \leq 1$ for each k and $w(1,n) + \ldots + w(n,n) = 1$. $\phi(n)$ is then an unbiased estimate of ϕ. Define

$$\varepsilon(n) := \left[\frac{q}{2} (e^{sp(t)} - 1)^2 \sum_{k=1}^{n} w_n^2(k) \right]^{\frac{1}{2}},$$

and let $\alpha(n)$, $\beta(n)$ be given by

$$\begin{aligned}\alpha(n) &:= \frac{1}{st}\log\Big([\phi(n)-\varepsilon(n)]\vee 1\Big),\\ \beta(n) &:= \frac{1}{st}\log\Big(\phi(n)+\varepsilon(n)\Big)\wedge\frac{p(t)}{t}.\end{aligned}$$

Proposition 5.1 *$\alpha(n) < \sigma(s,t) < \beta(n)$ with probability at least $1-e^{-q}$.*

PROOF Since $\sigma(s,t)$ takes values between 0 and $p(t)/t$ we have

$$\mathbb{P}\Big(\sigma(s,t)\le\alpha(n) \text{ or } \beta(n)\le\sigma(s,t)\Big) = \mathbb{P}\Big(|\phi(n)-\phi|\ge\varepsilon(n)\Big).$$

Set $Z_k := w(k,n)e^{sX(t,k)}$ for $k=1,\ \ldots,\ n$, so that $\phi(n) = Z_1+\ldots+Z_n$. By assumption the Z_k's are independent random variables satisfying

$$w(k,n)\le Z_k\le w(k,n)e^{sp(t)} \qquad k=1,\ \ldots,\ n.$$

Hoeffding's inequality therefore yields

$$\begin{aligned}\mathbb{P}\Big(|\phi(n)-\phi|\ge\varepsilon(n)\Big) &= \mathbb{P}\Big(|\sum_{k=1}^{n} Z_k-\mathbb{E}Z_k|\ge\varepsilon(n)\Big)\\ &\le 2\exp\Big(-2\varepsilon^2(n)/\sum_{k=1}^{n} w^2(k,n)(e^{sp(t)}-1)^2\Big),\end{aligned}$$

and, inserting $\varepsilon(n)$, the right-hand side is just e^{-q}.

If the error term $\varepsilon(n)$ tends to zero as n becomes large then so does the difference between the upper and lower estimates $\beta(n)$ and $\alpha(n)$. Since $\phi(n)\ge 1$ almost surely we have for $\varepsilon(n)<1$,

$$\begin{aligned}\beta(n)-\alpha(n) &= \frac{1}{st}\log\left(1+\frac{2\varepsilon(n)}{\phi(n)-\varepsilon(n)}\right)\\ &\le \frac{1}{st}\log\left(1+\frac{2-\varepsilon(n)}{1-\varepsilon(n)}\right) \qquad \text{a.s.}\end{aligned}$$

Using the inequality $\log(1+x)\le x$,

$$\beta(n)-\alpha(n)\le\frac{2\varepsilon}{st(1-\varepsilon(n))} \qquad \text{a.s.,}$$

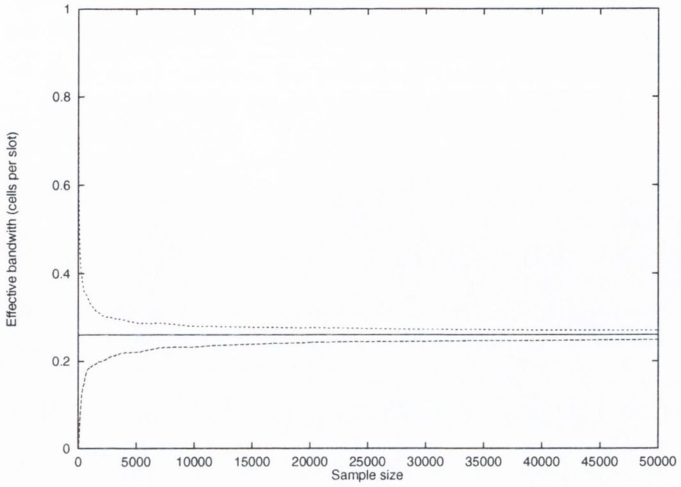

Figure 5-1: Estimated $1 - 10^{-6}$ confidence intervals for the effective bandwidth of traffic from a simple stochastic source. The central line is the source's true effective bandwidth.

which is of order $\varepsilon(n)$ as $\varepsilon(n) \to 0$. The size of the error depends on the choice of the weights $w(k, n)$, the optimal choice being $w(k, n) = 1/n$ for each k. In this case

$$\varepsilon(n) = (e^{sp(t)} - 1)\sqrt{\frac{q}{2n}}$$

and

$$\beta(n) - \alpha(n) \leq \frac{2(e^{sp(t)} - 1)\sqrt{q}}{st[\sqrt{2n} - (e^{sp(t)} - 1)\sqrt{q}]}$$

for $n > (e^{sp(t)} - 1)^2 q/2$.

Figure 5-1 illustrates the type of results which can be obtained using this scheme of equal weight for every measurement. Shown in the figure is a $1 - 10^{-6}$ confidence interval plotted against sample size, made using traffic from a two-state Markovian source. In its 'on' state this source transmits at a constant rate of 1 unit of work ('cell') per unit time ('slot'), and in its 'off' state transmits nothing. Sojourn times in both states are geometrically distributed, with a mean on-time of 5.333 slots and a mean off-time of 16 slots. The effective bandwidth estimates in the figure are for the values $s = 0.0184$ per cell and $t = 20$ slots; also shown is the true effective bandwidth of the source, for these values of s and t. The measurements of source activity used to compute the estimates were taken from back-to-back blocks of length 20 slots in a single long traffic sample. They cannot therefore be considered independent observations, and the effect of this is to introduce a small amount of bias into the estimates. Except possibly at the very largest sample sizes, this bias is more than offset by the conservative nature of the bounds.

Although the choice of equal weight for each observation of $X(t)$ minimises the value of $\varepsilon(n)$, it may not be appropriate in situations where the effective bandwidth of a source is to be monitored continuously over time. This choice is not robust against possible deviations from stationarity. We would like to be able to fix a timescale over which the estimated value of $\sigma(s,t)$ will track any changes in the true value: this can be achieved, for example, by using the weights $w(k,n)$ to implement an auto-regressive filter. Each measurement $X(t,k)$ of $X(t)$ comes from a block of length t; let $\tau_k \geq t$ be the time between the end of the kth block and the end of its successor. To obtain a first-order autoregressive filter we set $w(1,1) := 1$ and, for $n \geq 2$,

$$\begin{aligned} w(1,n) &:= \gamma^{\tau_1+\ldots+\tau_{n-1}}, \\ w(k,n) &:= (1-\gamma^{\tau_{k-1}})\gamma^{\tau_k+\ldots+\tau_{n-1}} \qquad k \geq 2, \end{aligned}$$

where $\gamma \in (0,1)$. Just after the end of the nth block our estimate $\phi(n)$ of the moment generating function of $X(t)$ is

$$\phi(n) = \gamma^{\tau_1+\ldots+\tau_{n-1}} e^{sX(t,1)} + \sum_{k=2}^{n} (1-\gamma^{\tau_{k-1}})\gamma^{\tau_k+\ldots+\tau_{n-1}} e^{sX(t,k)}.$$

Thus $\phi(1) = e^{sX(t,1)}$ and for $n \geq 2$,

$$\phi(n) = (1-\gamma^{\tau_{n-1}})e^{sX(t,n)} + \gamma^{\tau_{n-1}}\phi(n-1).$$

Setting $a = 1-\gamma^{\tau_{n-1}}$,

$$\phi(n) = \phi(n-1) + a(e^{sX(t,n)} - \phi(n-1)),$$

which is a special case of the constant gain stochastic approximation applied in control and communication problems. Here a is typically a negative integer power of 2, so that the computation of the product on the right-hand side reduces to bit-shifting. If the sequence $X(t,1)$, $X(t,2)$, ... is i.i.d. then $\phi(1)$, $\phi(2)$, ... is a homogeneous Markov process with limit distribution concentrated around $\mathbb{E}e^{sX(t,1)}$ for 'small' a [59, 32, 41].

Assume that observations of $X(t)$ are made periodically, so that $\tau_k = \tau \geq t$ for each k. After the nth observation the error term $\varepsilon(n)$ in the upper and lower bounds of proposition 5.1 is given by

$$\varepsilon^2(n) = (e^{sp(t)} - 1)^2 \frac{q}{2} \left(\frac{1 - \gamma^\tau + 2\gamma^{(2n-1)\tau}}{1 + \gamma^\tau} \right),$$

and as $n \to \infty$, $\varepsilon(n)$ converges to the non-zero value

$$\varepsilon_\infty := \lim_{n\to\infty} \varepsilon(n) = (e^{sp(t)} - 1) \sqrt{\frac{q}{2}} \left(\frac{1 - \gamma^\tau}{1 + \gamma^\tau} \right)^{\frac{1}{2}}.$$

Thus a large value for γ^τ (closer to one) results in a smaller confidence interval in the limit $n \to \infty$, but one which takes longer to converge, and longer to respond to any changes in the pattern of traffic from the source. The minimal value of γ^τ required to achieve a confidence interval of given size can be determined from the above equations. By varying both γ and τ the size of the limiting confidence interval and the rate of convergence can be controlled independently to some extent, but always subject to the constraint $\tau \geq t$. Note that if

$$\gamma^\tau > \frac{q(e^{sp(t)} - 1)^2 - 2}{q(e^{sp(t)} - 1)^2 + 2}$$

then $\varepsilon_\infty < 1$, and the error between the upper and lower effective bandwidth estimates $\beta(n)$ and $\alpha(n)$ satisfies

$$\lim_{n\to\infty} \beta(n) - \alpha(n) \leq \frac{2\varepsilon_\infty}{st(1 - \varepsilon_\infty)} \qquad \text{a.s.}$$

Applying the auto-regressive estimation procedure to a sample of Markovian traffic produces results such as those in figure 5-2. This plot was made using the same traffic source as that used to make figure 5-1. The period τ between measurements was set equal to $t = 20$ slots, and the value of γ was 0.999; with these parameters a reasonably narrow final confidence interval is achieved but the rate of convergence is slow compared to that in figure 5-1.

This estimation scheme suffers from a remaining defect, namely that the presence of periodicities in the traffic may lead to bias in the estimate $\phi(n)$ of the moment generating function. To avoid this, it is necessary to choose the observations randomly, rather than periodically, from the available data. Let us suppose that the inter-block time τ_k is equal to $t + r_k$, where $r_1, r_2, \ldots$ are independent exponentially distributed random variables with

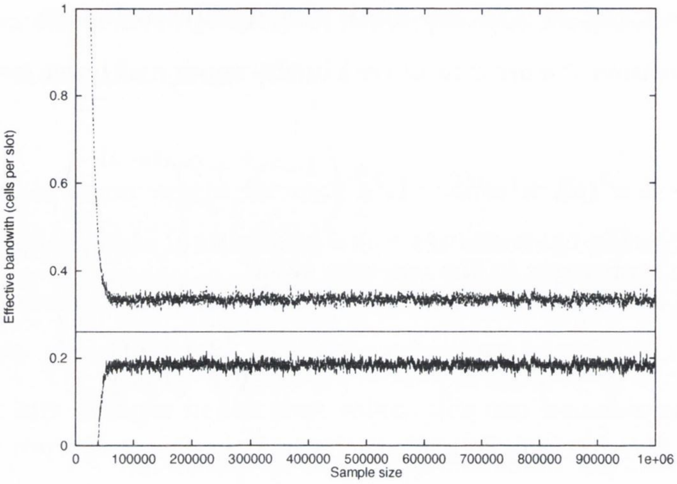

Figure 5-2: Estimated $1 - 10^{-6}$ confidence intervals for traffic from a simple stochastic source, made using a first-order autoregressive filter. The central line is the source's true effective bandwidth.

mean $1/\lambda$ (this type of procedure was suggested by R. Gibbens in [25]). The bounds of proposition 5.1 continue to hold (with probability at least $1 - e^{-q}$), but the error term $\varepsilon(n)$ is now a random variable because it depends on the inter-block times through the weighting function $w(\cdot, n)$.

The sum of squared weights $w^2(1,n) + \ldots + w^2(n,n)$ satisfies

$$\begin{aligned}\sum_{k=1}^{n} w^2(k,n) &= \gamma^{2\tau_1+\ldots+2\tau_{n-1}} + \sum_{k=2}^{n}(1-\gamma^{\tau_{k-1}})^2\gamma^{2\tau_k+\ldots+2\tau_{n-1}} \\ &= \gamma^{2\tau_{n-1}}\sum_{k=1}^{n-1} w^2(k,n-1) + (1-\gamma^{\tau_{n-1}})^2.\end{aligned}$$

Therefore the sequence $\{\varepsilon^2(n) : n \geq 1\}$ evolves through the stochastic recursive equations

$$\begin{aligned}\varepsilon^2(n) &= \frac{q}{2}(e^{sp(t)}-1)^2\sum_{k=1}^{n} w^2(k,n) \\ &= a_{n-1}\varepsilon^2(n-1) + b_{n-1}, \qquad (5.1)\end{aligned}$$

where $\{(a_n, b_n) : n \geq 1\}$ is the i.i.d. sequence given by

$$a_n := \gamma^{2\tau_n}, \qquad b_n := \frac{q}{2}(e^{sp(t)}-1)^2(1-\gamma^{\tau_n})^2.$$

Recursive systems of this kind have been studied by W. Vervaat [64] and A. Brandt [5], who show that if

$$-\infty \leq \mathbb{E}\log|a_1| < 0 \qquad \text{and} \qquad \mathbb{E}(\log|b_1|)^+ < \infty, \qquad (5.2)$$

or if $\mathbb{P}(a_1 = 0) > 0$, then there is a unique stationary process $\{S(n) : n \in \mathbb{E}\}$ satisfying equations (5.1), given by

$$S(n) := \sum_{k=0}^{\infty} b_{n-k-1} a_{n-k} \cdots a_{n-1} \qquad n \in \mathbb{E}$$

(here we assume that $\{a_n\}$ and $\{b_n\}$ have been extended to become integer-indexed sequences). If $\{S'(n) : n \in \mathbb{E}\}$ is any other solution of (5.1) then $|S'(n) - S(n)|$ tends to zero almost surely in n, and the distribution of $S'(n)$ converges to that of $S(0)$.

Conditions (5.2) are easily verified for the particular case at hand, and we conclude that $\varepsilon(n)$ converges almost surely to $\varepsilon_\infty := \sqrt{S(0)}$. If μ_a and μ_b denote, respectively, the expected values of a_1 and b_1:

$$\begin{aligned} \mu_a &:= \mathbb{E}a_1 = \frac{\lambda\gamma^{2t}}{\lambda - 2\log\gamma}, \\ \mu_b &:= \mathbb{E}b_1 = \frac{q}{2}(e^{sp(t)} - 1)^2 \left(1 - \frac{2\lambda\gamma^t}{\lambda - \log\gamma} + \frac{\lambda\gamma^{2t}}{\lambda - 2\log\gamma}\right), \end{aligned}$$

the $\mathbb{E}\varepsilon(n)$ and $\mathbb{E}\varepsilon_\infty$ satisfy

$$\mathbb{E}\varepsilon(n) \leq \sqrt{\mathbb{E}\varepsilon^2(n)} = \left(\mu_b \frac{1 - \mu_a^n}{1 - \mu_a}\right)^{\frac{1}{2}}$$

and

$$\mathbb{E}\varepsilon_\infty \leq \sqrt{\mathbb{E}S(0)} = \left(\frac{\mu_b}{1 - \mu_a}\right)^{\frac{1}{2}}.$$

As γ^t increases from zero to one, the right-hand side of this last inequality decreases from $(e^{sp(t)} - 1)\sqrt{q/2}$ to zero. Thus the limiting value of the error term can again be made as small as desired, at the cost of slower convergence.

5.3 Effective Bandwidth of Aggregate Traffic

We now assume that we are given separate data for each of several independent traffic sources which share a link, our task being to estimate the effective bandwidth of the aggregate link traffic. Let $X_l(t)$ be a random variable representing the quantity of work generated by source $l = 1, \ldots, L$ during an interval of length t, and let $p_l(t)$ be a known upper bound for $X_l(t)$, obtained in the same manner as $p(t)$ in section 5.2. We assume that $\{X_l(t),\ l = 1, \ldots, L\}$, are independent random variables, so that the effective bandwidth $\sigma(s,t)$ of the sum $X_1(t) + \ldots + X_L(t)$ satisfies

$$\sigma(s,t) = \sum_{l=1}^{L} \sigma_l(s,t),$$

where

$$\sigma_l(s,t) := \frac{1}{st} \log \mathbb{E} e^{sX_l(t)}$$

is the effective bandwidth of source l.

Let ϕ_l be the value of the moment generating function of $X_l(t)$ at s. Given n_l independent observations $X_l(t,1), \ldots, X_l(t,n_l)$ of $X_l(t)$, and a weighting function $w(\cdot, n_l)$, we form the estimate

$$\phi_l(n_l) := \sum_{k=1}^{n_l} w^2(k, n_l) e^{sX_l(t,k)}.$$

As before, $\phi_l(n_l)$ is an unbiased estimate of ϕ_l so long as the weights $w(\cdot, n_l)$ are chosen appropriately. Let $q > 0$ be given, and let $q_1, \ldots, q_L$ be positive numbers satisfying

$$e^{-q_1} + \ldots + e^{-q_L} = e^{-q}.$$

For each l define

$$\varepsilon_l(n_l) := \left[\frac{q_l}{2} (e^{sp_l(t)} - 1)^2 \sum_{k=1}^{n_l} w^2(k, n_l) \right]^{\frac{1}{2}}.$$

From the results of the last section we know that $\alpha_l(n_l)$

and $\beta_l(n_l)$, defined by

$$\begin{aligned} \alpha_l(n_l) &:= \frac{1}{st} \log\Big([\phi_l(n_l) - \varepsilon_l(n_l)] \vee 1 \Big), \\ \beta_l(n_l) &:= \frac{1}{st} \log\Big([\phi_l(n_l) + \varepsilon_l(n_l)] \Big) \wedge \frac{p_l(t)}{t}, \end{aligned}$$

bracket the value of $\sigma_l(s,t)$ with probability at least $1 - e^{-q_l}$.

Proposition 5.2 *Set*

$$\begin{aligned}\alpha &:= \alpha_1(n_1)+\ldots+\alpha_L(n_L),\\ \beta &:= \beta_1(n_1)+\ldots+\beta_L(n_L).\end{aligned}$$

Then $\alpha < \sigma(s,t) < \beta$ holds with probability at least $1-e^{-q}$.

PROOF $\sigma(s,t) \le \alpha$ or $\sigma(s,t) \ge \beta$ holds only if $\sigma_l(s,t) \le \alpha_l(n_l)$ or $\sigma_l(s,t) \ge \beta_l(n_l)$ for some l. Therefore

$$\begin{aligned}&\mathbb{P}(\sigma(s,t) \le \alpha \text{ or } \sigma(s,t) \ge \beta)\\ &\le \sum_{l=1}^{L} \mathbb{P}(\sigma_l(s,t) \le \alpha_l(n_l) \text{ or } \sigma_l(s,t) \ge \alpha_l(n_l))\\ &\le \sum_{l=1}^{L} e^{-q_l} = e^{-q}.\end{aligned}$$

As an illustration let us take $e^{-q_l} = e^{-q}/L$ for each l and compare proposition 5.2 with the results of section 5.2. The error terms are given by

$$\varepsilon_l(n_l) = \left[\frac{q+\log L}{2}(e^{sp_l(t)}-1)^2 \sum_{k=1}^{n_l} w^2(k,n_l)\right]^{\frac{1}{2}}.$$

Note that the upper estimate $\beta_l(n_l)$ of the effective bandwidth of source l satisfies

$$\beta_l(n_l) \le \frac{1}{st}\log[\phi_l(n_l)+\varepsilon_l(n_l)] \le \frac{1}{st}\log\phi_l(n_l) + \frac{\varepsilon_l(n_l)}{st\phi_l(n_l)},$$

and hence

$$\beta = \beta_1(n_1)+\ldots+\beta_L(n_L) \le \frac{1}{st}\sum_l \log\phi_l(n_l) + \frac{1}{st}\sum_l \frac{\varepsilon_l(n_l)}{\phi_l(n_l)}.$$

For large L the sum of the error terms on the right-hand side of this inequality grows proportionately with $L\sqrt{\log L}$. For comparison, if the aggregate traffic were treated as a single traffic stream, the results of the previous section would yield an error term proportional to $e^{sp(t)}$ with $p(t) = p_1(t)+\ldots+p_L(t)$. Thus the estimation error would grow approximately exponentially in the number of sources.

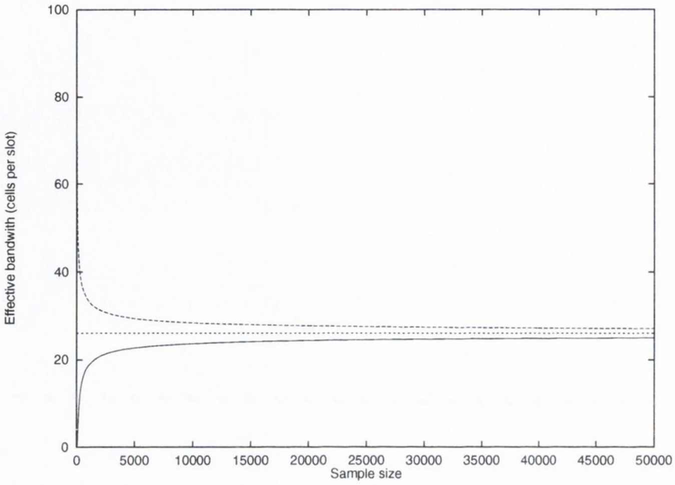

Figure 5-3: Estimated $1 - 10^{-6}$ confidence intervals for the effective bandwidth of an aggregate of 100 sources.

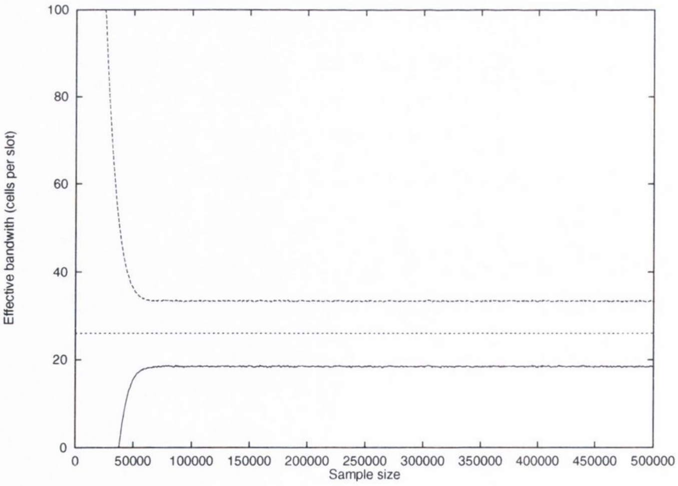

Figure 5-4: Estimated $1 - 10^{-6}$ confidence intervals for the effective bandwidth of an aggregate of 100 sources, made using a first-order autoregressive filter.

Figures 5-3 and 5-4 depict results obtained by applying this procedure to an aggregate of 100 Markovian sources of the type described in section 5.2. The weighting function $w(k, n) = 1/n$ was used to make the estimates in figure 5-3, while those of figure 5-4 were made using an auto-regressive filter. The results are seen to be qualitively very similar to the corresponding single source results, indicating little or no loss of precision due to aggregation.

Other choices for the coefficients q_l are also possible; for example, we may wish to choose smaller q_l where n_l is small, or where $p_l(t)$ is large. Using the weighting function $w(k, n_l) = 1/n_l$, $k = 1, \ldots, n_l$, the error $\varepsilon_l(n_l)$ is given by

$$\varepsilon_l(n_l) = (e^{sp_l(t)} - 1)\sqrt{\frac{q_l}{2n_l}}.$$

This can be made independent of l by choosing choosing

$$q_l = \frac{cn_l}{(e^{sp_l(t)} - 1)^2}$$

where c satisfies

$$\sum_{l=1}^{L} \exp(-cn_l/(e^{sp_l(t)} - 1)) = e^{-q}. \tag{5.3}$$

Then

$$\varepsilon_l(n_l) = \sqrt{c/2}$$

for each $l = 1, \ldots, L$. Thus the error terms are equalised by allowing more latitude to sources with higher peak rates, or from which fewer observations have been obtained. The transcendental equation (5.3) must however be solved numerically.

Appendix A

Appendix: The Loynes Theorems

Let a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ be given. Let $S : \Omega \to \Omega$ be an invariant left shift operator which is stationary (that is, $\mathbb{P}[S^{-1}F] = \mathbb{P}[F]$ for all $F \in \mathcal{F}$) and ergodic (that is, the shift invariant events, $\mathcal{B} \subseteq \mathcal{F}$, defined by $B \in \mathcal{B}$ if $S^{-1}[B] = B$, have probability 1 or 0). Let $\{\sigma_n : n \in \mathbb{Z}\}, \{\theta_n : n \in \mathbb{Z}\}$ be two sequences of non-negative, almost surely finite, random variables. Define the sequence $\{Z_n : n \in \mathbb{Z}\}$ by $Z_n := \sigma_n - \theta_n$ and the sequence $\{W_n : n \in \mathbb{Z}_+\}$ by $W_0 := 0$ and $W_n := \sum_{i=1}^{n} Z_{-i}$, for $n \geq 1$.

Assumption A.1 *The sequence $\{Z_n : n \in \mathbb{Z}\}$ is stationary with respect to S.*

Theorem A.1 *Define the queueing recursion by $w(n+1) := (w(n) + Z_{n+1}) \vee 0$, where we define $a \vee b$ to be the maximum of a and b. Then there exists a stationary sequence of random variables $\{\Upsilon_n : -\infty < n < \infty\}$, with distribution equal to $\Theta = \sup_{n \geq 0} W_n$, satisfying $\Upsilon_{n+1} := (\Upsilon_n + Z_{n+1}) \vee 0$ such that the distribution of $w(n)$ tends monotonically to that of Θ as $n \to \infty$.*

PROOF Define $w(-N,-N) := 0$ and define $w(-N,n)$, for $n > -N$, by

$$w(-N,n+1) := (w(-N,n) + Z_{n+1}) \vee 0.$$

We shall show that $w(-N,n)$ is monotone increasing in N and that for each n

$$\lim_{N\to\infty} w(-N,n) = \Upsilon_n,$$

almost surely. Note that for $t = -N+1$

$$\begin{aligned} w(-N,-N+1) &= (w(-N,-N) + Z_{-N+1}) \vee 0 \\ &= Z_{-N+1} \vee 0. \end{aligned}$$

Hence for $n = -N+2$,

$$\begin{aligned} w(-N,-N+2) &= (w(-N,-N+1) + Z_{-N+2}) \vee 0 \\ &= 0 \vee Z_{-N+2} \vee (Z_{-N+2} + Z_{-N+1}). \end{aligned}$$

By induction,

$$\begin{aligned} w(-N,n+1) &= (w(-N,n) + Z_{n+1}) \vee 0 \\ &= 0 \vee Z_{n+1} \vee (Z_{n+1} + Z_n) \vee \ldots \vee (Z_{n+1} + \cdots + Z_{-N+1}). \end{aligned}$$

Now, letting $N \to \infty$, we get that, for each n, $w(-N,n)$ tends monotonically to a limit Υ_n which satisfies $\Upsilon_{n+1} := (\Upsilon_n + Z_{n+1}) \vee 0$ and

$$\Upsilon_n := \lim_{N\to\infty} w(-N,n) = 0 \vee \sup_{r\geq 0} \sum_{i=0}^{r} Z_{n-i}.$$

To show the sequence $\{\Upsilon_n\}$ is stationary with respect to the left shift operator S we must show that $\Upsilon_{n+1} = S^{n+1}\Upsilon_0$. We do so by induction for $n \geq 0$. From 0 to 1 we have,

$$\begin{aligned} S\Upsilon_0 &= S \lim_{N\to\infty}\{0 \vee Z_0 \vee (Z_0 + Z_{-1}) \vee \ldots \vee (Z_0 + Z_{-1} + \cdots + Z_{-N})\} \\ &= \lim_{N\to\infty}\{0 \vee Z_1 \vee Z_1 + Z_0 \vee \ldots \vee (Z_1 + Z_0 + Z_{-1} + \cdots + Z_{-N+1})\} \\ &= (\Upsilon_0 + Z_1) \vee 0 \\ &= \Upsilon_1. \end{aligned}$$

Now, from n to $n+1$, we have,

$$\begin{aligned} S\Upsilon_n &= S \lim_{N\to\infty}\{0 \vee Z_n \vee (Z_n + Z_{n-1}) \vee \ldots \vee (Z_n + Z_{n-1} + \cdots + Z_{n-N})\} \\ &= \lim_{N\to\infty}\{0 \vee Z_{n+1} \vee (Z_{n+1} + Z_n) \vee \ldots \vee (Z_{n+1} + Z_n + \cdots + Z_{n-N+1})\} \\ &= (\Upsilon_n + Z_{n+1}) \vee 0 \\ &= \Upsilon_{n+1}. \end{aligned}$$

So we have shown that $\Upsilon_{n+1} = S^{n+1}\Upsilon_0$, for each $n \geq 0$, and hence the sequence $\{\Upsilon_n\}$ is stationary with respect to the left shift operator S.

We note that, because the sequence $\{\Upsilon_n\}$ is stationary, each Υ_n has the same distribution. For convenience, we define Θ by

$$\Theta := \Upsilon_0 := \lim_{N\to\infty} w(-N,0) = \sup_{n\geq 0} W_n.$$

Assumption A.2 *The sequence $\{Z_n : n \in \mathbb{Z}\}$ satisfies*

$$\limsup_{N\to\infty} \sum_{i=0}^{N} Z_i = -\infty,$$

almost surely.

Note that if $\{Z_n\}$ is stationary, ergodic and $\mathbb{E}[Z_n] := \mathbb{E}[\sigma_n] - \mathbb{E}[\theta_n] < 0$, then the Birkoff strong law of large numbers says that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} Z_i = \mathbb{E}[Z_0] < 0,$$

almost surely, and thus

$$\limsup_{N\to\infty} \sum_{i=0}^{N} Z_i = -\infty,$$

almost surely. Hence assumption A.2 is satisfied if $\{Z_n\}$ is stationary, ergodic and $\mathbb{E}[Z_n] < 0$.

Definition A.1 *We say the sequence of random variables $\{\Upsilon'_n : n \in \mathbb{Z}\}$ is coupled with the sequence $\{\Upsilon_n : n \in \mathbb{Z}\}$ if $\{\Upsilon_n : n \in \mathbb{Z}\}$ is stationary and there exists an almost surely finite random variable, τ, such that $\Upsilon'_n = \Upsilon_n$ for all $n > \tau$.*

Theorem A.2 *If assumptions A.1 and A.2 are satisfied, then any solution $\{\Upsilon'_n\}$ of $w(n+1) = (w(n) + Z_n) \vee 0$, for $n \geq 0$, couples to $\{\Upsilon_n\}$, where $\{\Upsilon_n\}$ is defined in theorem A.1.*

PROOF By the recursion, we have

$$\Upsilon'_n = 0 \vee Z_n \vee (Z_n + Z_{n-1}) \vee \ldots \vee (Z_n + Z_{n-1} + \cdots + Z_0 + \Upsilon'_0)$$

and

$$\Upsilon_n = 0 \vee Z_n \vee (Z_n + Z_{n-1}) \vee \ldots \vee (Z_n + Z_{n-1} + \cdots + Z_0 + \Upsilon_0).$$

By assumption A.2,

$$\limsup_{N \to \infty} \sum_{i=0}^{N} Z_i = -\infty,$$

almost surely. Therefore there exists an almost surely finite random variable, τ, such that

$$(Z_\tau + Z_{\tau-1} + \cdots + Z_1 + \Upsilon'_0) \vee (Z_\tau + Z_{\tau-1} + \cdots + Z_1 + \Upsilon_0) < 0.$$

Thus we have that, for all $n \geq \tau$,

$$\Upsilon_n = \Upsilon'_n = 0 \vee Z_n \vee (Z_n + Z_{n-1}) \vee \ldots \vee (Z_n + Z_{n-1} + \cdots + Z_1),$$

and $\{\Upsilon'_n\}$ is coupled to $\{\Upsilon_n\}$.

Bibliography

[1] S. Asmussen and J. F. Collamore. Exact asymptotics for a large deviations problem for the GI/G/1 queue. *Markov Processes and Related Fields*, 5(4):451–476, 1999.

[2] J. Beran, R. Sherman, W. Willinger, and M. Taqqu. Long range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, 43:1566–1579, 1995.

[3] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation.* Cambridge University Press, 1987.

[4] 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, 1996.

[5] A. Brandt. The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability*, 18:69–86, 1986.

[6] W. Bryc and A. Dembo. Large deviations and strong mixing. *Annales de l'I.H.P. Probabilitiés and Statistiques*, 32:549–569, 1996.

[7] C. Courcoubetis, F. P. Kelly, and R. Weber. Measurement-based charging in communications networks. Preprint, 1997.

[8] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a switch handling many traffic sources. *Journal of Applied Probability*, 33:886–903, 1996.

[9] M. Crovella and A. Bestavros. Self-simmilarity in world wide web traffic: evidence and possible causes. *Performance Evaluation Review*, 24:160–169, 1996.

[10] G. de Veciana, G. Kesidis, and J. Walrand. Resource management in wide-area ATM networks using effective bandwidths. *IEEE Journal on Selected Areas in Communications*, 13(6):1081–1090, 1995.

[11] A. Dembo and O. Zeitouni. *Large Deviation Techniques and Applications.* Springer, 1998.

[12] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

[13] N. G. Duffield. Economies of scale in queues with sources having power-law large deviation scalings. *Journal of Applied Probability*, 33:840–857, 1996.

[14] N. G. Duffield, J.T. Lewis, N. O'Connell, R. Russell, and F. Toomey. Statistical issues raised by the bellcore data. In *Proceedings of the 11th IEE UK Teletraffic Symposium*, 1994.

[15] 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 (special issue on Advances in the Fundamentals of Networking)*, 13:981–990, 1995.

[16] N. G. Duffield, J.T. Lewis, N. O'Connell, R. Russell, and F. Toomey. Predicting quality of service for traffic with long-range fluctuations. In *Proceedings of the IEEE International Conference on Communications*, 1995.

[17] 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(2):363–374, 1995.

[18] N. G. Duffield and W. Whitt. Large deviations of inversely related processes with non-linear scalings. *Annals of Applied Probability*, 8:995–1026, 1998.

[19] K. Duffy. On the large deviations of a class of stationary on/off sources which exhibit long range dependence. Technical report, DIAS, 2000.

[20] R. Ellis. Large deviations for a general class of random vectors. *Annals of Probability*, 12:1–12, 1984.

[21] S. Floyd. Comments on measurement-based admissions control for controlled-load services. Technical report, ICSI, 1996. http://www.aciri.org/floyd/.

[22] N. Gantert. Large deviations for a heavy tailed mixing sequence. Technical report, Technische Universitat Berlin, 1996.

[23] N. Gantert. Functional Erdös-Renyi laws for semi-exponential random variables. *Annals of Probability*, 26(3):1356–1269, 1998.

[24] J. Gartner. On large deviations from the invariant measure. *Theor Prob. Appl.*, 22:24–39, 1977.

[25] R. J. Gibbens. Traffic characterisation and effective bandwidths for broadband network traces. In F. P. Kelly, S. Zachary, and I. B. Ziedins, editors, *Stochastic Networks: Theory and Applications*, pages 169–179. Oxford University Press, 1996.

[26] R. J. Gibbens and F. P. Kelly. Measurement-based connection admission control. In V. Ramaswami and P. E. Wirth, editors, *Teletraffic Contributions for the Information Age: Proceedings of the 15th International Teletraffic Conference, Washington, D. C.* Elsevier, Amsterdam, 1997.

[27] P. Glynn and W. Whitt. Large deviations behaviour of counting processes and their inverses. *Queueing Systems*, 17:107–128, 1994.

[28] P. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability*, 31A:413–430, 1994.

[29] J. Grandell. *Aspects of Risk Theory*. Springer, Berlin, 1991.

[30] M. Grossglauser and D. Tse. A framework for robust measurement-based admission control. In *Proceedings of ACM SIGCOMM '97*, 1997.

[31] L. Györfi, A. Rácz, K. Duffy, J. T. Lewis, and F. Toomey. Distribution-free confidence intervals for measurement of effective bandwidths. To appear in the Journal of Applied Probability, 2000.

[32] L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal of Control and Optimization*, 34:31–61, 1996.

[33] P. R. Halmos. *Ergodic Theory.* Chelesa Publishing Company, 1956.

[34] J. M. Harrison. *Brownian Motion and Stochastic Flow Systems.* Wiley, 1985.

[35] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Society*, 58:13–30, 1963.

[36] H. E. Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770–808, 1951.

[37] F. P. Kelly. Notes on effective bandwidths. In F. P. Kelly, S. Zachary, and I. B. Ziedins, editors, *Stochastic Networks: Theory and Applications*, pages 141–168. Oxford University Press, 1996.

[38] J. L. Kelly. *General Topology.* D. Van Nostrand Company, Inc., 1955.

[39] J. F. C. Kingman. On continuous time models in the theory of dams. *Journal of the Australian Mathematical Society*, 3:480–487, 1963.

[40] V. Klemes. The Hurst phenomenon: a puzzle? *Water Resources Research*, 10:675–688, 1974.

[41] H. J. Kushner and A. Shwartz. Weak convergence and asymptotic properties of adaptive filters with constant gains. *IEEE Transactions on Information Theory*, IT-30:177–182, 1984.

[42] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Willson. On the self-simmilar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2:1–15, 1994.

[43] J. T. Lewis and C. E. Pfister. Thermodynamic probability theory: some aspects of large deviations. *Russian Mathematical Surveys*, 50(2):279–317, 1995.

[44] J. T. Lewis, C. E. Pfister, and W. G. Sullivan. Entropy, concentration of probability and conditional limit theorems. *Markov Processes and Related Fields*, 1:319–386, 1995.

[45] J. T. Lewis, R. Russell, F. Toomey, B. McGurk, S. Crosby, and I. Leslie. Practical connection admission control for ATM networks based on on-line measurements. *Computer Communications*, 21:1585–1596, 1998.

[46] D. V. Lindley. The theory of queues with a single server. *Proceedings of the Cambridge Philosphical Society*, 48:277–289, 1952.

[47] R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Proceedings of the Cambridge Philosphical Society*, 58:497–520, 1962.

[48] B. B. Mandelbrot and J. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437, 1968.

[49] B. B. Mandelbrot and J. Wallis. Noah, joseph and operational biology. *Water Resources Research*, 4:909–918, 1968.

[50] B. B. Mandelbrot and J. Wallis. Computer experiments with fractional Gaussian noises, 1, Averages and variances; 2, Rescaled ranges and spectra; 3, Mathematical apendix. *Water Resources Research*, 5(1):228–267, 1969.

[51] B. B. Mandelbrot and J. Wallis. Robustness of the rescaled R/S in the measurment of noncylic long run statistical dependence. *Water Resources Research*, 5:967–988, 1969.

[52] B. B. Mandelbrot and J. Wallis. Some long-run properties of geophysical records. *Water Resources Research*, 5:321–340, 1969.

[53] A. Martin-Löf. Entropy, a useful concept in risk theory. *Scandinavian Actuarial Journal*, 3-4:223–235, 1986.

[54] T. Mikosch and A. V. Nagaev. Large deviations of heavy-tailed sums with applications in insurance. *Extremes*, 1(1):81–110, 1998.

[55] S. V. Nagaev. Large deviations of sums of independent random variables. *Annals of Probability*, 7:745–789, 1979.

[56] I. Norros. A storage model with self-simmilar input. *Queueing systems*, 16:387–396, 1994.

[57] H. Nyrhinen. Large deviations for the time of ruin. *Journal of Applied Probability*, 36:733–746, 1999.

[58] P. E. O'Connell. A simple stochastic modeling of Hurst's law. In *Proceedings of the International Symposium on Mathematical Models in Hydrology*, 1971.

[59] G. Ch. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal of Control and Optimization*, 24 (4):655–660, 1986.

[60] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[61] R. Russell. The large deviations of inversely related processes. Technical report, DIAS, 1995.

[62] R. Russell. *The Large Deviations of Random Time Changes.* PhD thesis, Trinity College, Dublin, 1997.

[63] A. Simonian and J. Guibert. Large deviations approximation for fluid queues fed by a large number of on/off sources. *IEEE Journal on Selected Areas in Communications*, 13(6):1017–1027, 1995.

[64] W. Vervaat. On a stochastic difference equation and a representation of non-negative infinitely-divisble random variables. *Advances in Applied Probability*, 11:750–783, 1979.