Moderate Deviations in Queueing Theory

Damon Wischik
Statistical Laboratory
Cambridge

1 February 2001

Abstract

Moderate deviations theory concerns a collection of scales between large deviations theory and the central limit theorem. When applied to queueing problems, moderate deviations theory combines the simplicity of large deviations techniques with the parsimony of heavy traffic approximations. This leads to some very simple heuristics for traffic engineering—for example, that a traffic stream passing through several queues is not significantly smoothed except at the most congested queue.

1 Introduction

It is now widely known that the behaviour of a heavily loaded queue can be approximated rather well using reflected Gaussian processes. This is appealing because it leads to parsimonious models: all we need to know about an arrival process is its mean and covariance structure. The theory behind this is known as *heavy traffic theory* [11, 16, 24, 26], and it is based on the scaling used in the central limit theorem.

Another type of approximation is based on *large deviations theory* and the closely related idea of effective bandwidth [15, 20]. A key idea in this theory is that (in the large deviations scaling) a rare event occurs only in the most likely way. For example, to estimate the probability that a large buffer overflows, we need only consider the probability of the most likely way for the buffer to fill up. This clearly simplifies the calculation.

It is tempting to combine the two approaches: by the central limit theorem, we can model the arrival process as a Gaussian process; by large deviations theory, it is sufficient to study most likely paths. We see papers with titles like "Most probable paths and performance formulae for buffers with Gaussian input traffic" [1]. Tempting, but dangerous[1]. The central limit theorem involves one sort of scaling, and large deviations theory involves another. To what extent is it legitimate to mix the two?

In this paper I attempt to answer this question, by studying the *moderate deviations scalings*, which I shall index by a burstiness parameter $\beta \in (0, 1)$. These scalings lies between the central limit scaling ($\beta = 0$) and the large deviations scaling ($\beta = 1$). A typical traffic flow has bursts at all scales $0 <$

[1] I hasten to note that all the performance formulae in [1] represent a legitimate mixture of heavy traffic and large deviations scales, and are justified elsewhere [2] by simulation.

$\beta < 1$, with larger bursts (large β) less frequent than smaller bursts (small β). The parameter β thus plays a dual role: it measures both burst *sizes* and burst *frequencies.*

As one would expect, the resulting limit theorems mirror aspects of both heavy traffic theory and large deviations theory; these theorems lead to very simple heuristics. But some aspects are *not* mirrored. Specifically, I will describe two objects that merit great care: the effective bandwidth of a Gaussian process, and the output of a queue fed by a Gaussian process.

The most striking moderate deviations result concerns the manner in which a traffic flow is smoothed as it passes through a queue. Overflow at a queue has a certain frequency, depending on its utilization, and so we can assign to the queue its own characteristic burstiness scale β. We will see that the queue smooths out all bursts in the traffic at scales larger than β, but leaves unchanged all bursts at smaller scales. In effect, the queue acts as a low-pass filter.

The aim of this paper is not to prove new theorems but to illuminate the links between heavy traffic and large deviations. I therefore hope that readers familiar with heavy traffic theory will find some parts obvious, and that readers familiar with large deviations theory will find different parts obvious. It will however be necessary to prove some new results to fill in gaps in the existing literature: on moderate deviations for multiclass networks and for networks with many flows, and on mixed limits.

The rest of this paper is organized as follows. Section 2 briefly defines the notation. Section 3 describes the standard traffic limits (heavy traffic limits and large deviations limits) and how moderate deviations fits between them. This is the most important section. Section 4 defines moderate deviations theory; and Section 5 explains how it applies to traffic processes. Section 6 uses standard tools of large deviations theory to deduce moderate deviations results for queue size and related quantities; Section 7 uses those tools in a different way to look at systems whose various parts are scaled differently (and proves the low-pass-filter result). Section 8 considers heuristics and approximations, and ties together moderate deviations with heavy traffic and large deviations. Section 9 is the conclusion.

2 Notation

This paper refers to at least eight different scalings, so it is important straight away to be clear about the notation.

The first thing to define is the traffic process. To save some analytical complexity, we will deal only with queues operating in discrete time. Define a traffic process to be a sequence of real numbers indexed by the negative integers $\mathbf{x} = (\dots, x_{-2}, x_{-1})$. Interpret x_{-t} as the amount of work arriving at a queue at time $-t$. We apologise for the bother of negative indices, but hope that it will prevent confusion later. We will deal with a variety of scaled versions of $\mathbf{x}$, indicated by $\hat{\mathbf{x}}$. The hats indicate that we need to be careful in interpreting $\hat{\mathbf{x}}$. Since this paper is all about scaling phenomena, there will be very many hats.

Denote by $\mathbf{x}[-s, -t)$ the truncation of the process: $\mathbf{x}[-s, -t) = (x_u)_{-s \le u < -t}$, for $-s < -t$. Denote by $x[-s, -t)$ the cumulative sum process: $x[-s, -t) = x_{-s} + \dots + x_{-t-1}$, with $x[-s, -s) = 0$. When the process is random, we will write $\mathbf{X}$ and X. Let $\mathbf{1} = (\dots, 1, 1)$.

To keep things simple, we will deal only with queues with constant service rate. Consider a queue with service rate C and buffer size B fed by an input process $\mathbf{x}$. Define the queue size at time $-s$ to be the limit

$$Q_{-s}(\mathbf{x}) = \lim_{t\to\infty} Q_{-s}(\mathbf{x}[-t,0))$$

where $Q_{-s}(\mathbf{x}[-t,0))$ is given by the recursion

$$Q_{-u} = [Q_{-u-1} + x_{-u} - C]_0^B, \quad Q_{-u} = 0 \text{ for } -u \leq -t.$$

When $B = \infty$, this reduces to the familiar form

$$Q_{-s}(\mathbf{x}) = \sup_{t\geq 0} x[-s-t,-s) - Ct.$$

Later on, we shall define queue size in multiclass queues, and departure processes.

Two quantities which recur, and which merit their own symbols, are the *fast-time* and the *many-flows* versions of a random traffic process $\mathbf{X}$. Define the fast-time version $\mathbf{X}^{\otimes L}$ to be $\mathbf{X}$ speeded up by L: $X^{\otimes L}[-t,0) = X[-Lt,0)$. And define the many-flows version $\mathbf{X}^{\oplus L}$ to be the aggregate (i.e. sum) of L independent copies of $\mathbf{X}$.

3 Background

The moderate deviations scales lie between the heavy traffic scale and the large deviations scale, so we shall first review those two scales. In fact, there are four scales to review: both types of theory have two flavours, the fast-time flavour and the many-flows flavour.

In this section, we will deal with an arrival process $\mathbf{X}$, which we will assume to be stationary. Let $\mu = \mathbb{E}X_{-1}$.

3.1 Central limit theorem, fast time

This is the best known of all the scales. It is usually called the *heavy traffic limit* or the *diffusion approximation*. Let $\sigma^2 = \operatorname{Var} X_{-1}$. For now, assume that the X_{-t} are independent. By the central limit theorem, under mild conditions on the X_{-t}, $t^{-1/2}(X[-t,0) - \mu t)$ converges as $t \to \infty$ to a normal random variable with mean 0 and variance σ^2; and

$$\hat{\mathbf{X}}^L = L^{1/2}(L^{-1}\mathbf{X}^{\otimes L} - \mu\mathbf{1})$$

converges to a (discrete sample of a) Brownian motion with zero drift and variance parameter σ^2. This is often true even when the X_{-t} are not independent. (It is not appropriate here to be precise about the nature of the convergence.) Note that the limit process depends only on the variance of $\mathbf{X}$.

Consider a queue $\hat{Q}$ with service rate $\hat{C}$ and buffer size $\hat{B}$ fed by $\hat{\mathbf{X}}^L$:

$$\begin{aligned} \hat{Q}_0 &= \big[\hat{Q}_{-1} + \hat{X}^L_{-1} - \hat{C}\big]_0^{\hat{B}} \\ \implies \sqrt{L}\hat{Q}_0 &= \big[\sqrt{L}\hat{Q}_{-1} + X[-L,0) - (L\mu + \sqrt{L}\hat{C})\big]_0^{\sqrt{L}\hat{B}} \\ \implies Q^L_0 &= \big[Q^L_{-1} + X[-L,0) - LC^L\big]_0^{B^L} \end{aligned}$$

where $Q^L = \sqrt{L}\hat{Q}$ is the queue size in a queue fed by $\mathbf{X}$ and served at rate $C^L = \mu + L^{-1/2}\hat{C}$, with buffer size $B^L = \sqrt{L}\hat{B}$ (and in which time has been speeded up[2] by L). In heavy traffic theory it is more common to parameterize the sequence by the traffic intensity ρ. In these terms,

$$\rho^L = \mu / C^L \sim 1 - L^{-1/2}\hat{C}/\mu \tag{1}$$

and so $\sqrt{L}(1-\rho^L) \to \hat{C}/\mu$ and $\rho^L \to 1$.

Since $\hat{\mathbf{X}}^L$ converges to a Brownian motion, $\hat{Q}_t(\hat{\mathbf{X}}^L)$ (that is, $L^{-1/2}Q^L_t$) converges to a reflected Brownian motion. A great deal is known about this limit—see for example [12, 16, 26]. Later in this paper we will be interested in two special properties: the snapshot principle [24] and state space collapse [13, 23].

3.2 Central limit theorem, many flows

Recall that $\mathbf{X}^{\oplus L}$ is the aggregate of L independent copies of $\mathbf{X}$. Define

$$\hat{\mathbf{X}}^L = \sqrt{L}(L^{-1}\mathbf{X}^{\oplus L} - \mu\mathbf{1}).$$

By the central limit theorem we would expect $\hat{\mathbf{X}}^L$ to converge to a Gaussian process (though not necessarily to a Brownian motion, if we allow $\mathbf{X}$ to have an arbitrary covariance structure) [3].

As before, let $\hat{Q}$ be the queue size function for a queue served at rate $\hat{C}$ and with buffer size $\hat{B}$. It is easy to check that

$$\hat{Q}(\hat{\mathbf{X}}^L) = L^{-1/2}Q^L_0(\mathbf{X}^L),$$

where $Q^L_0(\mathbf{X}^L)$ is the queue size in a queue fed by an aggregate of L copies of $\mathbf{X}$, served at rate $C^L = L\mu + \sqrt{L}\hat{C}$, with buffer size $B^L = \sqrt{L}\hat{B}$. The traffic intensity is the same as before, (1).

One can conclude that $L^{-1/2}Q^L_0(\mathbf{X}^L)$ converges to a Gaussian process. The precise nature of the convergence has been studied by Addie et al. [2], who assert that "virtually all conceivable performance measures, must approach the performance of a communication system carrying Gaussian traffic with the same second order statistics". They justify this assertion for systems with a central limit scaling. It is well-known that the assertion does not extend to systems with a large deviations scaling [20, 29]; however, moderate deviations theory will show us that it does extend well beyond the central limit scaling.

3.3 Large deviations, fast time

We can now move on to a different sort of theory. The large deviations fast-time limit (also known as the large-buffer limit) is the best known of the three

[2] In this discrete time model, the effect of speeding up is to make the sampling interval more coarse. In a continuous time model, speeding up would not have this effect. It is not a priori clear which is better, though fortunately the two models give the same qualitative answers in most cases.

[3] Incidentally, this limit is *not* the same as the many-servers limit, described by Halfin and Whitt [10]. They study systems in which a single flow of customers is served by several servers, in the limit where the traffic intensity increases as the number of servers increases. (It is difficult to even describe this limit with our notation.) It shares some similarities with the many-flows limit: in particular, the limiting queue size process is Gaussian, and in general more complicated than a simple reflected Brownian motion. There are also differences: most noteably, the snapshot principle does not apply.

different large deviations scales we shall meet. Let $\hat{\mathbf{X}}^L = L^{-1}\mathbf{X}^{\otimes L}$. Under mild conditions on $\mathbf{X}$, $\hat{\mathbf{X}}^L$ satisfies a large deviations principle of the form

$$L^{-1}\log\mathbb{P}(\hat{\mathbf{X}}^L \in \hat{S}) \approx -\inf_{\mathbf{x}\in\hat{S}} I(\hat{\mathbf{x}}) \tag{2}$$

where the rate function I has the form

$$I(\hat{\mathbf{x}}) = \sum_{-t<0} \Lambda^*(\hat{x}_{-t})$$

for some convex function Λ^* (which depends on the entire distribution of $\mathbf{X}$). We shall go into much more detail about what this means later. For now, simply note the form of the right hand side of (2): it is an infimum, which means that the event $\hat{\mathbf{X}}^L \in \hat{S}$ happens only in the most likely way, $\arg\inf_{\hat{\mathbf{x}}\in\hat{S}} I(\hat{\mathbf{x}})$. This is an example of the *principle of the largest term*.

Now consider a queue with service rate $\hat{C}$ and buffer size $\hat{B}$ fed by $\hat{\mathbf{X}}^L$:

$$\begin{aligned} \hat{Q}_0 &= \big[\hat{Q}_{-1} + \hat{X}^L_{-1} - \hat{C}\big]_0^{\hat{B}} \\ \implies Q^L_0 &= \big[Q^L_{-1} + X[-L,0) - LC^L\big]_0^{B^L} \end{aligned}$$

where Q^L is the queue size in a queue fed by $\mathbf{X}$ and served at rate $C^L = \hat{C}$, with buffer size $B^L = L\hat{B}$ (in which time has been speeded up by a factor of L). The traffic intensity is just

$$\rho^L = \mu/C^L = \mu/\hat{C} < 1. \tag{3}$$

By the contraction principle, $\hat{Q}_0(\hat{\mathbf{X}}^L)$ (that is, $L^{-1}Q^L_0$) satisfies a large deviations principle of the form

$$L^{-1}\log\mathbb{P}(L^{-1}Q^L_0 \geq b) \approx -J(b).$$

In other words, the queue length has an exponentially decaying tail. There has been a great deal of work on this limit; see [20] for an account which fits in well with this paper.

3.4 Large deviations, many flows

Let $\hat{\mathbf{X}}^L = L^{-1}\mathbf{X}^{\oplus L}$. It can be shown that $\hat{\mathbf{X}}^L$ satisfies a large deviations principle of the form

$$L^{-1}\log\mathbb{P}(\hat{\mathbf{X}}^L \in \hat{S}) \approx -\inf_{\hat{\mathbf{x}}\in\hat{S}} I(\hat{\mathbf{x}})$$

for some rate function I. As before, let $\hat{Q}$ be the queue size in a queue served at rate $\hat{C}$ with buffer $\hat{B}$. It is simple to check that

$$\hat{Q}_0(\hat{\mathbf{X}}^L) = L^{-1}Q^L_0(\mathbf{X}^{\oplus L}),$$

where $Q^L_0(\mathbf{X}^{\oplus L})$ is the queue size in a queue fed by an aggregate of L copies of $\mathbf{X}$, served at rate $C^L = L\hat{C}$, with buffer size $B^L = L\hat{B}$. (The traffic intensity is the same as (3), so $\rho^L < 1$.) By the contraction principle, $L^{-1}Q_0^L$ satisfies a large deviations principle of the form

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

and hence the overflow probability decays exponentially in the degree of multiplexing.

This scale was first introduced by Weiss [25]. For a full account see Wischik [29], who shows that the fast-time result is essentially a special case of the many-flows result.

3.5 Moderate deviations, fast time

So far, we have learnt about $\hat{Q}\big(\sqrt{L}(L^{-1}\mathbf{X}^{\otimes L} - \mu\mathbf{1})\big)$ and $\hat{Q}(L^{-1}\hat{\mathbf{X}}^{\otimes L})$. The former is approximately Gaussian, governed by the variance of $\mathbf{X}$; the latter has exponential tails, and is governed by the principle of the largest term.

There is a collection of intermediate limits, the moderate deviations limits. Consider the scaled arrival process $\hat{\mathbf{X}}^L$ defined by

$$\hat{\mathbf{X}}^L = L^{(1-\beta)/2}(L^{-1}\mathbf{X}^{\otimes L} - \mu\mathbf{1})$$

where $\beta \in (0,1)$. When $\beta \approx 0$, this is close to the heavy traffic scale; when $\beta \approx 1$, it is close to the large deviations scale. We will see later that (for all $\beta \in (0,1)$, under mild conditions on $\mathbf{X}$) $\mathbf{X}^L$ satisfies a moderate deviations principle of the form

$$L^{-\beta}\log\mathbb{P}(\hat{\mathbf{X}}^L \in \hat{S}) \approx -\inf_{\hat{\mathbf{x}}\in\hat{S}} I(\hat{\mathbf{x}}) \tag{4}$$

where $I(\cdot)$ depends only on the long-term variance of $\mathbf{X}$. Note that $\hat{\mathbf{X}}^L$ is governed both by the variance of $\mathbf{X}$, and by the principle of the largest term.

What of the queue size? As usual, consider a queue with service rate $\hat{C}$ and buffer size $\hat{B}$ fed by $\hat{\mathbf{X}}^L$:

$$\begin{aligned} \hat{Q}_0 &= \big[\hat{Q}_{-1} + X^L_{-1} - \hat{C}\big]_0^{\hat{B}} \\ \implies Q^L_0 &= \big[Q^L_{-1} + X[-L,0) - LC^L\big]_0^{B^L} \end{aligned}$$

where $Q^L = L^{(1+\beta)/2}\hat{Q}$ is the queue size in a queue fed by $\mathbf{X}$ and served at rate $C^L = \mu + L^{-(1-\beta)/2}\hat{C}$, with buffer size $B^L = L^{(1+\beta)/2}\hat{B}$ (in which time has been speeded up by L). The traffic intensity is

$$\rho^L = \mu/C^L \sim 1 - L^{-(1-\beta)/2}\hat{C}/\mu, \tag{5}$$

so that $\rho^L \to 1$ but not as quickly as in the regular heavy traffic case. Perhaps inevitably, we call this *moderately heavy traffic*.

Since $\hat{\mathbf{X}}^L$ satisfies a moderate deviations principle, so does $\hat{Q}_0(\hat{\mathbf{X}}^L)$ (that is, $L^{-(1+\beta)/2}Q_0^L$), of the form

$$L^{-\beta}\log\mathbb{P}(L^{-(1+\beta)/2}Q_0^L \geq b) \approx -J(b), \tag{6}$$

or equivalently

$$L^{-2(1-\frac{1}{1+\beta})}\log\mathbb{P}(L^{-1}\hat{Q}_0^L \geq b) \approx -J(b).$$

There has been little work so far on approximations of this form. They are briefly mentioned by Wischik [28]. They are developed more fully by Chang et al. [3], Puhalskii [22] and Majewski [18]. Chang et al. study moderate deviations in a single queue, and present their results as an extension of the large deviations results of Duffield and O'Connell [8] and others. Puhalskii studies networks of queues with feedback, with an homogeneous customer population, and stressed the link with heavy-traffic theory. Majewski also studies networks of queues with an homogeneous customer population, and gives a sophisticated presentation of his results as an interchange between the heavy traffic limit and the large deviations limit.

3.6 Moderate deviations, many flows

As one might by now expect, moderate deviations can be applied to the many-flows regime. Define the scaled process $\hat{\mathbf{X}}^L$ by

$$\hat{\mathbf{X}}^L = L^{(1-\beta)/2}(L^{-1}\mathbf{X}^{\oplus L} - \mu\mathbf{1})$$

for $\beta \in (0,1)$. This is nearly identical to the fast-time moderate-deviations scale in Section 3.5, and indeed, exactly the same moderate deviations principle (4) holds (though the rate function I is more complicated in the many-flows case, since it depends on the full covariance structure of the process $\mathbf{X}$).

However, the interpretation in terms of queue size is rather different. It is simple to check that $\hat{Q}_0(\hat{\mathbf{X}}^L) = L^{-(1+\beta)/2}Q_0^L$, where Q_0^L is the queue size in a queue fed by the aggregate of L copies of $\mathbf{X}$ and served at rate $C^L = L\mu + L^{(1+\beta)/2}\hat{C}$, with buffer size $B^L = L^{(1+\beta)/2}\hat{B}$. As in (5), the traffic intensity is

$$\rho^L = \mu/C^L \sim 1 - L^{-(1-\beta)/2}\hat{C}/\mu. \tag{7}$$

One can show that Q_0^L satisfies a moderate deviations principle of the form

$$L^{-\beta}\log\mathbb{P}(L^{-(1+\beta)/2}Q_0^L \geq b) \approx -J(b)$$

where J depends on the full covariance structure of $\mathbf{X}$.

There seem to be no results on this scale in the literature. This is unfortunate, because it does seem to offer the best compromise between accuracy and simplicity of all scales we study here.

Because of this accuracy and simplicity, there are many papers which *assume* such results. Such papers typically assert that "the aggregate arrival process can be effectively characterized by a stationary Gaussian process" [4] and then go on to study asymptotics of Gaussian processes based on the principle of the largest term [1, 4, 5, 19] (finding either large deviations type asymptotics, or refined asymptotics). This mixture of limits is typically justified by simulation; see Choe and Shroff [4] for a particularly good account.

The purpose of this paper is to show that such an approach is valid, so long as the queue is in moderately heavy traffic (7). When the queue is lightly loaded (3), or very heavily loaded (1), the approach is not valid.

3.7 Summary of scales

An important point to appreciate is that by choosing a particular sort of limit theorem, we implicitly restrict attention to a particular way of scaling the resources at a queue.

Heavy traffic theory deals with a single scale $\beta = 0$, and large deviations theory deals with a different single scale $\beta = 1$, whereas moderate deviations theory covers a range of scales $\beta \in (0, 1)$. This makes it very useful in understanding the different sorts of scaling phenomena that can occur, particularly in systems which different parts are scaled differently.

Moderate deviations models are *parsimonious*, like heavy traffic models, because the limit theorems depend only on the mean and variance of the input process; large deviations results, by contrast, involve its entire statistical characteristics. Moderate deviations techniques are *simple*, like large deviations results, because they are based on the principle of the largest term; heavy traffic results, by contrast, require one to consider many possible paths to overflow.

The contortions of parameterization in stating a moderate deviations principle undoubtedly obscure the simplicity of the idea. Really, moderate deviations theory can greatly *simplify* the application of both heavy traffic theory and large deviations theory. We will see later just how this works, after we have been more precise about the technical aspects of moderate deviations.

We have chosen a particularly concrete representation of the moderate deviations limit, indexed by a parameter β. One could study moderate deviations more abstractly, by proving that the large deviations limit and the heavy traffic limit can be interchanged—indeed, Majewski [18] has taken this route in studying fast-time moderate deviations in networks with a homogeneous customer population. In these terms, the parameter β represents a particular mixture of the two limits. The advantage of our concrete representation is that it makes it easier to talk about the range of scaling phenomena that moderate deviations theory describes.

But before continuing, we should briefly note that the burstiness scale β is not related to the Hurst parameter of a long-range dependent process. Both parameters are intended to describe burstiness, but in different ways; in Section 8.6 we will look at the links between the two.

4 Moderate deviations principle

Before we can proceed, we need a proper definition of a moderate deviations principle. This section also includes some examples of arrival processes that satisfy the principle, and some useful results.

A moderate deviations principle is nothing other than a special case of a large deviations principle (LDP).

Definition 1 (Large deviations principle) *A sequence of random variables X^L taking values in a Hausdorff space $\mathcal{X}$ with σ-algebra $\mathcal{B}$ is said to satisfy a* large deviations principle *with good rate function I if for any $S \in \mathcal{B}$*

$$\begin{aligned} -\inf_{x\in S^\circ} I(x) &\le \liminf_{L\to\infty} \frac{1}{L}\log \mathbb{P}(X^L \in S) \\ &\le \limsup_{L\to\infty} \frac{1}{L}\log \mathbb{P}(X^L \in S) \le -\inf_{x\in\bar{S}} I(x), \end{aligned} \tag{8}$$

S° is the interior of S, $\bar{S}$ is the closure of S, and the function $I : \mathcal{X} \to \mathbb{R}^+ \cup \{\infty\}$ has compact level sets.

We will assume throughout that $\mathcal{B}$ contains the σ-algebra of interest, which in most cases is the Borel σ-algebra.

The two sides of (8) are known as the large deviations lower and upper bounds.

When we write an equation like

$$\frac{1}{L}\log \mathbb{P}(X^L \in S) \approx -\inf_{x\in S} I(x), \tag{9}$$

we mean that the approximation holds in the large deviations sense (8).

We say that X^L satisfies a *moderate deviations principle* (MDP) with scaling parameter $\beta \in (0,1)$ and mean μ if

$$\frac{1}{L^\beta}\log \mathbb{P}\big(L^{(1-\beta)/2}(X^L - \mu) \in S\big) \approx -\inf_{x\in S} I(x).$$

In all the examples we study in this section, the rate function I will be quadratic and depend only on the limiting covariance structure of X^L. (We will explain what this means when we decide on the space $\mathcal{X}$.) With some abuse of language, we have already use the term *moderate deviations principle* to describe large deviations principles which nearly fit this form, such as (6).

A very important tool is the contraction principle, which lets us take one LDP and derive another. This principle applies to MDPs as well as LDPs, since the former are simply special cases of the latter.

Theorem 1 (Contraction principle) *Suppose X^L satisfies an LDP of the form* (9) *in some space $\mathcal{X}$, and that $f : \mathcal{X} \to \mathcal{Y}$ is a continuous function. Then we obtain an LDP in $\mathcal{Y}$ of the form*

$$L^{-1}\log \mathbb{P}(f(X^L) \in S) \approx -\inf_{y\in S} J(y)$$

where

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

5 Moderate deviations for traffic processes

Let $\mathbf{X}^L$ be a sequence of traffic processes.

Definition 2 (Sample path MDP) *We will say that $\mathbf{X}^L$ satisfies the* sample path MDP *with mean μ and covariance structure $(\gamma_t)_{t>0}$ if the following conditions hold:*

For each $\beta \in (0,1)$, $\mathbf{X}^L$ satisfies a moderate deviations principle of the form

$$\frac{1}{L^\beta} \log \mathbb{P}\big(L^{(1-\beta)/2}(\mathbf{X}^L - \mu\mathbf{1}) \in \hat{S}\big) \approx - \inf_{\hat{\mathbf{x}} \in \hat{S}} I(\hat{\mathbf{x}}) \tag{10}$$

with good rate function $I(\cdot)$, in the space

$$\mathcal{X}_\delta = \Big\{\mathbf{x} : \frac{x[-t,0)}{t} \leq \delta \ \textit{ eventually}\Big\}$$

equipped with the uniform norm

$$\|\mathbf{x}\| = \sup_{t>0} \Big| \frac{x[-t,0)}{t} \Big|$$

for any $\delta > 0$.

The rate function $I(\cdot)$ has the form

$$I(\hat{\mathbf{x}}) = \sup_{t>0} \sup_{\theta \in \mathbb{R}^t} \boldsymbol{\theta}^\mathsf{T} \hat{\mathbf{x}}[-t,0) - \tfrac{1}{2}\boldsymbol{\theta}^\mathsf{T} \Sigma_t \boldsymbol{\theta}$$

where Σ_t is the $t \times t$ matrix

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

Furthermore, setting $V_t = \mathbf{1}^\mathsf{T}\Sigma_t\mathbf{1}$, $V_t = o(t^2/\log t)$.

5.1 Proving the MDP

In this section we find conditions under which $\mathbf{X}^L$ satisfies the sample path MDP. It may safely be skipped: most reasonable processes satisfy it.

Since a moderate deviations principle is just a special case of a large deviations principle, it is not surprising that the MDP can be proved using the same techniques as for the LDP. In particular, we appeal to the general large deviations result from Wischik [29] for most of the proofs of the following theorems.

Condition 3 (Finite-time regularity) *Define the scaled cumulant moment generating function $\mathbf{\Lambda}_t^L(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathbb{R}^t$ by*

$$\mathbf{\Lambda}_t^L(\boldsymbol{\theta}) = \frac{1}{L^\beta} \log \mathbb{E}\exp\big(\boldsymbol{\theta}^\mathsf{T} L^\beta L^{(1-\beta)/2}(L^{-1}\mathbf{X}^L[-t,0) - \mu^L\mathbf{1})\big)$$

where $\beta \in (0,1)$. Assume that for each t the limiting moment generating function exists and is given by

$$\lim_{L\to\infty} \mathbf{\Lambda}_t^L(\boldsymbol{\theta}) = \tfrac{1}{2}\boldsymbol{\theta}^\mathsf{T}\Sigma_t\boldsymbol{\theta} \tag{11}$$

where Σ_t is a $t \times t$ matrix.

Condition 4 (Covariance structure) *Assume that $(\Sigma_t)_{ij} = \gamma_{|i-j|}$ for some process γ_t. Let $V_t = \mathbf{1}^\mathsf{T}\Sigma_t\mathbf{1}$, and assume that $V_t = o(t^2/logt)$.*

Remark
It is worth noting that for stationary flows, we can recover the full covariance structure γ_t from the marginal variances V_t (given by $V_t = \mathbf{1}^\mathsf{T}\Sigma_t\mathbf{1}$). To see this, note that

$$\begin{aligned} V_{t+1} &= \mathbf{1}_{t+1}^\mathsf{T}\Sigma_{t+1}\mathbf{1}_{t+1} \\ &= \mathbf{1}_t^\mathsf{T}\Sigma_t\mathbf{1}_t + \gamma_0 + 2\mathbf{1}^\mathsf{T}(\gamma_t,\dots,\gamma_1), \end{aligned}$$

and so $\gamma_0 = V_1$, $\gamma_1 = \frac{1}{2}(V_2 - V_1)$, and for $t > 1$, $\gamma_t = \frac{1}{2}(V_{t+1} - 2V_t + V_{t-1})$.

In other words, as far as moderate deviations is concerned, the marginal distributions of the cumulative arrival process $(X[-t,0))_{t>0}$ *fully characterize the process.*

The covariance structure (γ_t) is what makes moderate deviations interesting. In the many-flows moderate deviations scale, the rate function depends on the covariance structure of the source over all timescales. This makes the technique well-suited to any sort of traffic modelling where one wants to capture the fine-grained statistical characteristics of a flow. We will continue the discussion of timescales in the next section.

Condition 5 (Long-term regularity) *Define for $\theta \in \mathbb{R}$*

$$\Lambda_t^L(\theta) = \frac{1}{t}\frac{V_t}{t}\mathbf{\Lambda}_t^L\Big(\mathbf{1}\theta\frac{t}{V_t}\Big).$$

(Note that $\mathbf{\Lambda}$ depends on β.) Assume that for θ in some neighbourhood of the origin, and for all t, the limit

$$\lim_{L\to\infty}\frac{1}{\theta^2}\big(\Lambda_t^L(\theta) - \frac{1}{2}\theta^2\big) = 0 \qquad (12)$$

is uniform.

This final condition relates the speed at which $\mathbf{X}^L[-t,0)$ converges to a Gaussian process, to the speed at which $\mathbf{X}^L[-t,0)$ becomes well-behaved as $t \to \infty$.

The following theorem is taken directly from Wischik [29].

Theorem 2 *Assume that condition 3 holds for each $\beta \in (0,1)$. Then for each t, $\mathbf{X}^L[-t,0)$ satisfies a moderate deviations principle in $\mathbb{R}^t$ with good rate function*

$$I_t(\mathbf{x}) = \sup_{\boldsymbol{\theta}\in\mathbb{R}^t}\boldsymbol{\theta}^\mathsf{T}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\theta}^\mathsf{T}\Sigma_t\boldsymbol{\theta}.$$

The following theorem is very similar to Theorem 3 in Wischik [29]. The only difference is that our long-term regularity condition, condition 5, has been phrased slightly differently, in order to make it more useful for moderate deviations.

Theorem 3 *Assume that conditions 3 and 5 hold for each $\beta \in (0,1)$, and that condition 4 holds too. Then $\mathbf{X}^L$ satisfies the sample path moderate deviations principle, Definition 2.*

Proof. The only additional claim that needs justification is this: that there exists a t_0 such that

$$\lim_{\alpha\to\infty} \limsup_{L\to\infty} \frac{1}{L} \log \sum_{t\geq t_0} \exp\bigl[-Lw_t \sup_\theta\bigl(\theta\alpha d_t - \Lambda_t^L(\theta)\bigr)\bigr] \tag{13}$$

is equal to $-\infty$, where $d_t = \surd(V_t \log t/t^2)$ and $w_t = t^2/V_t$.

Pick $\theta = d_t$. By (12), given $\varepsilon > 0$, for L and t sufficiently large,

$$\frac{1}{\theta^2}\bigl|\Lambda_t^L(\theta) - \tfrac{1}{2}\theta^2\bigr| < \varepsilon.$$

Thus

$$\sup_\theta\bigl(-Lw_t(\theta\alpha d_t - \Lambda_t^L(\theta)\bigr) \leq -Lw_t d_t^2(\alpha - \tfrac{1}{2} - \varepsilon)$$

and so (13) is less than or equal to

$$\lim_{\alpha\to\infty}(\alpha - \tfrac{1}{2} - \varepsilon) \limsup_{L\to\infty} \frac{1}{L} \log \sum_{t\geq t_0} t^{-L}.$$

It is simple to check that this is equal to $-\infty$. □

We have assumed that the conditions hold for all $\beta \in (0,1)$. One would expect there to be processes $\mathbf{X}^L$ for which the conditions hold only for β in a subset of $(0,1)$. We will not investigate tighter conditions in this paper.

5.2 Example traffic processes

For much of the paper, we have in mind the many-flows scale, which is important enough to merit this lemma.

Lemma 4 (Many flows over finite timescales) *Let* $\mathbf{X}$ *be a random stationary flow. For* $\boldsymbol{\theta} \in \mathbb{R}^t$*, let*

$$M_t(\boldsymbol{\theta}) = \log \mathbb{E} \exp \boldsymbol{\theta}^\mathsf{T}\mathbf{X}[-t,0).$$

Assume that M_t *is finite in a neighbourhood of the origin. Then* $\mathbf{X}$ *has finite mean* μ*, and* $L^{-1}\mathbf{X}^{\oplus L}$ *satisfies Condition 3 for all* $\beta \in (0,1)$*. The matrix* Σ_t *is given by* $(\Sigma_t)_{ij} = \operatorname{Cov}(X_{-i}, X_{-j})$*.*

Proof. M_t is a log moment generating function. Since it is finite in a neighbourhood of the origin, it is infinitely differentiable in that neighbourhood, and so the mean exists. Now,

$$\boldsymbol{\Lambda}_t^L(\boldsymbol{\theta}) = \frac{M_t(\boldsymbol{\theta}\delta) - \boldsymbol{\theta}^\mathsf{T}\mathbf{1}\mu\delta}{\delta^2},$$

where $\delta = L^{-(1-\beta)/2}$. Thus $\boldsymbol{\Lambda}_t^L(\boldsymbol{\theta}) = \operatorname{Var} \boldsymbol{\theta}^\mathsf{T}\mathbf{X}[-t,0)$, which is $\boldsymbol{\theta}^\mathsf{T}\Sigma_t\boldsymbol{\theta}$. □

We still need to check that condition 5 holds. There are two special cases in which it is trivial.

Example 1 (Aggregated Gaussian flows)
Suppose that bX is a stationary Gaussian process, with variance $\operatorname{Var} X[-t,0) = o(t^2/\log t)$. Since $\mathbf{X}$ has finite moment generating function, $L^{-1}\mathbf{X}^{\oplus L}$ satisfies Condition 3 by Lemma 4. It satisfies Condition 4 by stationarity, and it trivially satisfies Condition 5 since $\Lambda_t^L(\theta) = \frac{1}{2}\theta^2$. $\diamond$

Example 2 (Aggregate flows with independent increments)
Suppose that $\mathbf{X}$ has independent increments, say $X_{-t} \sim Y$, and that the moment generating function of Y is finite in a neighbourhood of the origin. Then $L^{-1}\mathbf{X}^{\oplus L}$ satisfies Condition 3 by Lemma 4. It satisfies Condition 4, with $\gamma_0 = \operatorname{Var} Y$ and $\gamma_t = 0$ otherwise, and $V_t = Vt$. It satisfies Condition 5 since $\Lambda_t^L(\theta) = \Lambda_1^L(\theta)$. $\diamond$

This characteristic of the covariance structure is important to merit a definition.

Definition 6 (Asymptotically independent increments) *We say that* $\mathbf{X}^L$ *has* asymptotically independent increments *if the covariance structure* γ *satisfies* $\gamma_0 = V$ *and all other* $\gamma_t = 0$*. Note that then* $V_t = Vt$*.*

This property means that most likely paths to overflow are linear. It often arises in the fast-time scaling.

Example 3 (Fast time, independent increments)
Let $\mathbf{X}$ be a process with independent increments, say $X_{-t} \sim Y$. Suppose that the moment generating function of Y is finite in a neighbourhood of the origin. Define $\mathbf{X}^{\otimes L}$ by $X^{\otimes L}[-t,0) = X[-Lt,0)$. Then $L^{-1}\mathbf{X}^{\otimes L}$ satisfies Condition 3 by a similar argument to Lemma 4. It has asymptotically independent increments, so satisfies Condition 4. It satisfies Condition 5, since $\Lambda_t^L(\theta) = \Lambda_1^L(\theta)$. $\diamond$

Our final example is of fractional Brownian motion in the fast-time limit. This process has attracted a great deal of attention, because of its unusual scaling behaviour, which will be discussed further in Section 8.6.

Example 4 (Fast time, fractional Brownian motion)
Define $\mathbf{X}$ by $X[-t,0) = \mu t + \sigma Z_t$, where Z_t is a standard fractional Brownian motion with Hurst parameter H, so that $\mathbf{X}$ is Gaussian and $\operatorname{Var} X[-t,0) = \sigma^2 t^{2H}$. Consider $\mathbf{X}^{\otimes N}$, where $N = L^{1/(2-2H)}$. It is easy to check that $N^{-1}\mathbf{X}^{\otimes N}$ satisfies Condition 3, since it is Gaussian. The covariance structure is

$$\gamma_t = \tfrac{1}{2}\sigma^2\big(|t-1|^{2H} - 2|t|^{2H} + |t+1|^{2H}\big)$$

and $V_t = t^{2H}$. Condition 5 is trivially satisfied, since $\Lambda_t^L(\theta) = \frac{1}{2}\theta^2$. $\diamond$

6 Queue size and related quantities

From here on, we will simply assume a process which satisfies the sample path moderate deviations principle, Definition 2. Our main tool for deriving new MDPs from this will be the contraction principle, Theorem 1. We will find MDPs for three sorts of quantity: in Section 6.1 for the total queue size, in Section 6.2 for paths to overflow, in Section 6.3 for the queue size due to each flow in a shared buffer, and in Section 6.4 for the departure process.

6.1 Total queue size

The theory for this section is largely identical to the large deviations theory, described by Wischik [29], so we will stress the results and skip the proofs.

Consider a sequence of arrival processes $\mathbf{X}^L$, assumed to satisfy the sample path MDP—that is, a moderate deviations principle of the form

$$\frac{1}{L^\beta}\log\mathbb{P}\big(L^{(1-\beta)/2}(L^{-1}\mathbf{X}^L-\mu\mathbf{1})\in\hat{S}\big)\approx-\inf_{\hat{\mathbf{x}}\in\hat{S}}I(\hat{\mathbf{x}})$$

where

$$I(\hat{\mathbf{x}})=\lim_{t\to\infty}\sup_{\boldsymbol{\theta}\in\mathbb{R}^t}\boldsymbol{\theta}^{\mathsf{T}}\hat{\mathbf{x}}[-t,0)-\tfrac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\Sigma_t\boldsymbol{\theta}.$$

The processes $\mathbf{X}^L$ could arise from the fast-time limit ($\mathbf{X}^L=\mathbf{X}^{\oplus L}$, Section 3.5) or the many-flows limit ($\mathbf{X}^L=\mathbf{X}^{\otimes L}$, Section 3.6), or indeed from any other sort of limit.

Consider a sequence of queues: let Q^L be the queue size in a queue fed by $\mathbf{X}^L$ and served at rate $C^L=L\mu+L^{(1+\beta)/2}\hat{C}$, with buffer size $B^L=L^{(1+\beta)/2}\hat{B}$ (where $\hat{B}$ may be infinite). We have chosen this scale so that

$$Q^L=L^{(1+\beta)/2}\hat{Q}\big(L^{(1-\beta)/2}(\mathbf{X}^L-\mu\mathbf{1})\big)$$

where $\hat{Q}(\cdot)$ is the queue size function for a queue with service rate $\hat{C}$ and buffer size $\hat{B}$.

By Lemma 3.7 in [28], the queue size function is continuous on $\mathcal{X}_\delta$ for $0<\delta<\hat{C}$. Pick any $0<\delta<\hat{C}$ and write $\mathcal{X}$ for $\mathcal{X}_\delta$. By applying the contraction principle, we immediately obtain an MDP for Q^L of the form

$$\frac{1}{L^\beta}\log\mathbb{P}(L^{-(1+\beta)/2}Q^L\in\hat{S})\approx-\inf_{\hat{b}\in\hat{S}}J(\hat{b})$$

where

$$J(\hat{b})=\inf_{\hat{\mathbf{x}}\in\mathcal{X}:\hat{Q}(\hat{\mathbf{x}})=\hat{b}}I(\hat{\mathbf{x}}).$$

There is a whole slew of results like this, exactly analogous to those in [29]. Here is a restatement of some of those results.

Lemma 5 *Let $J_{\hat{B}}$ be the rate function for the queue size in a queue with finite buffer $\hat{B}$, and let J be the rate function when $\hat{B}=\infty$. If $\hat{b}\le\hat{B}$ then $J_{\hat{B}}(\hat{b})=J(\hat{b})$; otherwise, $J_{\hat{B}}(\hat{b})=\infty$. Also, $J(\hat{b})$ is increasing and is given by*

$$J(\hat{b})=\inf_{t>0}\frac{(\hat{b}+\hat{C}t)^2}{2V_t}. \tag{14}$$

If $J(\hat{b})<\infty$ then the optimal t^* is attained, and it is the time period over which the queue is most likely to fill up.

Proof. The large deviations theorem gives as rate function

$$J(b)=\inf_{t>0}\sup_{\theta\in\mathbb{R}}\theta(b+Ct)-\Lambda_t(\theta\mathbf{1}),$$

where $\Lambda_t(\boldsymbol{\theta})=\frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\Sigma_t\boldsymbol{\theta}$. This reduces immediately to (14). □

Lemma 6 *If $\hat{B} > 0$, the event $\{\hat{Q} > 0\}$ has moderate deviations lower bound $-I(0^+)$ and upper bound $-I^+(0)$. If $\hat{B} < \infty$, the event that $\hat{Q}$ overflows has moderate deviations lower bound $-I(\hat{B}^+)$ and upper bound $-I(\hat{B})$ (or $-I^+(0)$ if $\hat{B} = 0$). Here, $I(\hat{b}^+) = \lim_{\hat{a}\downarrow\hat{b}} I(\hat{a})$ and $I^+(0) = \hat{C}^2/2V_1$.*

6.2 Paths to overflow

It is not very difficult to describe the most likely path to overflow. Restating the appropriate large deviations theorem,

Lemma 7 *If $J(\hat{b})$ is finite then the most likely path $\hat{\mathbf{x}}^*$ to lead to overflow, and the most likely timescale t^*, are both attained. The most likely path is given by*

$$\hat{\mathbf{x}}^*[-t,0) = \Sigma_t\Big(0\mathbf{1} + \frac{\hat{b} + \hat{C}t^*}{2V_{t^*}}\mathbf{1}_{t^*}\Big).$$

Remember: this means that the most likely *unscaled* path is $\mathbf{X}^L = L\mu\mathbf{1} + L^{(1+\beta)/2}\hat{\mathbf{x}}^*$, and it causes Q^L to fill to level $L^{(1+\beta)/2}\hat{b}$.

6.3 Shared queues

Now consider a single queue fed by several different traffic streams $\underline{\mathbf{X}} = (\mathbf{X}(i))$, where $1 \leq i \leq M$. In the next section, we will seek a moderate deviations principle for the collection of departure processes $\underline{\mathbf{D}}(\underline{\mathbf{X}})$. That is rather hard. So first we will answer a problem which is simpler, but still interesting in its own right: how much work of type i is there in the queue?

In what follows, we will write $\underline{\mathbf{X}}$ for the vector $(\mathbf{X}(i))_{i=1\ldots M}$, and $\mathbf{X}$ for the aggregate $\mathbf{X} = \sum_i \mathbf{X}(i)$.

In the last section, we looked at aggregate queue size Q^L_{-t} in a queue fed with aggregate input $\mathbf{X}^L$ and served at rate $L\mu + L^{(1+\beta)/2}\hat{C}$. (For convenience, we will take the buffer to be infinite. The modifications for finite buffers are trivial.) Then

$$Q^L_{-t} = L^{(1+\beta)/2}\hat{Q}_{-t}\big(L^{(1-\beta)/2}(L^{-1}\mathbf{X}^L - \mu\mathbf{1})\big),$$

where $\hat{Q}_{-t}(\hat{\mathbf{X}})$ is the queue size function for a queue fed by process $\hat{\mathbf{X}}$; this function is continuous, thus we were able to use the contraction principle to find an MDP for Q^L_{-t}.

Now, let the amount of work of type i in the queue be

$$Q^L_{-t}(i).$$

To make sure this is well-defined, we need to specify a service policy. Assume that work from each source $\underline{X}_{-t}$ arrives uniformly spread throughout the interval $[-t, -t-1)$, and that work is served in the order it arrives. Let $\hat{Q}^L_{-t}(i) = L^{-(1+\beta)/2}Q^L_{-t}(i)$. We will seek an MDP for $\underline{\hat{Q}}^L$.

This is harder than finding an MDP for Q^L, because $\underline{Q}^L$ is not a simple continuous function of $L^{(1-\beta)/2}(L^{-1}\underline{\mathbf{X}}^L - \underline{\mu}\mathbf{1})$. We shall see, however, that it is *very nearly* a simple continuous function, and instead of the contraction principle, we will find we can use the approximate contraction principle.

Recall that in the moderate deviations limit, the actual (unscaled) amount of work of type i that arrives at time $-t$ is $X^L_{-t}(i) = L\mu(i) + L^{(1+\beta)/2}\hat{X}^L_{-t}(i)$, where $\hat{X}^L_{-t}(i)$ is a fluctuation at the moderate deviations scale[4]. This means that the vast majority of work in the queue comes from the $L\mu$ term. So one would expect

$$Q^L_{-t}(i) \approx \frac{\mu(i)}{\mu} Q^L_{-t}. \tag{15}$$

This can be made precise using the idea of *exponential equivalence* [7]. Broadly speaking, if two random variables Y^L and Z^L are exponentially equivalent, then any differences between them are too small to be picked up by large deviations techniques. We will show that the (scaled) left and right hand sides of (15) are (moderately) exponentially equivalent.

Theorem 8 *Let*

$$Y^L = \hat{Q}^L_{-t}(i) \quad \textit{and} \quad Z^L = \hat{Q}^L_{-t}\frac{\mu(i)}{\mu}.$$

Then Y^L and Z^L are moderately exponentially equivalent, in that

$$\limsup_{L\to\infty} \frac{1}{L^\beta} \log \mathbb{P}\big(|Y^L - Z^L| > \delta\big) = -\infty \;\; \textit{for all } \delta > 0.$$

(We will as usual assume that these sets are all measurable.)

Proof. Consider how $Q_{-t}(i)$ comes about. At time $-t-1$ there was a certain amount of work Q_{-t-1} in the queue, with work from the different flows distributed somehow. Then $\underline{X}^L_{-t}$ arrives, and work from the different flows is distributed evenly. Of this total work, $L\mu + L^{(1+\beta)/2}\hat{C}$ is served, the original work Q^L_{-t-1} coming first.

So either Q^L_{-t-1} is served completely, in which case

$$Y^L = \hat{Q}^L_{-t}\frac{X^L_{-t}(i)}{X^L_{-t}},$$

or it is not, which implies that

$$Q^L_{-t-1} > L\mu + L^{(1+\beta)/2}\hat{C}.$$

Thus

$$\begin{aligned}&\mathbb{P}(|Y^L - Z^L| > \delta)\\ &\qquad \le \mathbb{P}\bigg(\hat{Q}^L_{-t-1} > L^{(1-\beta)/2}\mu + \hat{C}\bigg) + \mathbb{P}\bigg(\Big|\frac{X^L_{-t}(i)}{X^L_{-t}} - \frac{\mu(i)}{\mu}\Big|\hat{Q}^L_{-t} > \delta\bigg).\end{aligned} \tag{16}$$

By the principle of the largest term, it is sufficient to show that for each of these parts, $\limsup_L L^{-\beta}\log\mathbb{P}(\cdot) = -\infty$.

[4] When we say that $\hat{X}^L$ is a fluctuation at the moderate deviations scale, we mean that

$$L^{-\beta}\log\mathbb{P}(\hat{X}^L \in \hat{S}) \approx -\inf_{\hat{x}\in\hat{S}} I(\hat{x}).$$

We deal with the first term first. We know that $\hat{Q}^L_{-t-1}$ satisfies a moderate deviations principle, say with rate function $J(\hat{q})$. Thus

$$\limsup_{L\to\infty} \frac{1}{L^\beta} \log \mathbb{P}\big(\hat{Q}^L_{-t-1} > L^{(1-\beta)/2}\mu + \hat{C}\big) \le -J(\hat{q})$$

for every $\hat{q} > 0$. (This relies on the assumption that $\mu > 0$.) But $J(\hat{q})$ is unbounded as $\hat{q} \to \infty$. (This relies on the assumption that $V_t = o(t^2/\log t)$.) So the lim sup is equal to $-\infty$.

Now for the second term in (16). Let $\delta_1 = |L^{-1}X^L_{-t}(i) - \mu(i)|$ and $\delta_2 = |L^{-1}X^L_{-t} - \mu|$. If $\delta_2 < \mu$ then

$$\Big|\frac{X^L_{-t}(i)}{X^L_{-t}} - \frac{\mu(i)}{\mu}\Big| \le \frac{\mu\delta_1 + \mu(i)\delta_2}{\mu(\mu - \delta_2)}.$$

Thus we can break up the second term in (16):

$$\begin{aligned} &\mathbb{P}\Big(\Big|\frac{X^L_{-t}(i)}{X^L_{-t}} - \frac{\mu(i)}{\mu}\Big|\hat{Q}^L_{-t} > \delta\Big) \\ &\qquad\qquad \le \mathbb{P}(\delta_1\hat{Q}^L_{-t} > \delta\mu^L) + \mathbb{P}(\delta_2\hat{Q}^L_{-t} > \delta\mu^L) + \mathbb{P}(\delta_2 > \mu/2). \end{aligned}$$

The three terms can be dealt with similarly. We will deal with the second term by way of example. Rewriting it in full and adding in some scaling terms we get

$$\begin{aligned} &\mathbb{P}\big(L^{(1-\beta)/2}|L^{-1}X^L_{-t} - \mu|\hat{Q}^L_{-t} > L^{(1-\beta)/2}\delta\mu\big) \\ &\qquad \le \mathbb{P}\big(L^{(1-\beta)/2}|L^{-1}X^L_{-t} - \mu| > L^{(1-\beta)/4}\mu\big) + \mathbb{P}\big(\hat{Q}^L_{-t} > L^{(1-\beta)/4}\delta\big). \end{aligned}$$

Again, $\limsup L^{-\beta}\log\mathbb{P}(\cdot) = -\infty$ for each of these terms, as $L^{(1-\beta)/2}(L^{-1}X^L_{-t} - \mu)$ and $\hat{Q}^L_{-t}$ both satisfy moderate deviations principles with rate functions that tend to ∞. □

Corollary 9 *$\underline{Q}^L$ is exponentially equivalent to $Q^L\underline{\mu}/\mu$.*

Proof. This is a straightforward consequence of the fact that if Y^L_1 and Z^L_1 are exponentially equivalent, and Y^L_2 and Z^L_2 are also, then so are (Y^L_1, Y^L_2) and (Z^L_1, Z^L_2). □

This leaves us with the following picture of the evolution of the queue. At the level of the fluctuations we are interested in, a total amount of work $\hat{x}_{-t}$ arrives at time $-t$, made up as a vector $\underline{\hat{x}}_{-t}$ of work from the different flows. The queue size fluctuates according to the standard recursion: $\hat{Q}_{-t} = (\hat{Q}_{-t-1} + \hat{x}_{-t} - \hat{C})^+$. All of $\hat{Q}_{-t-1}$ is served at time $-t$, and the amount of work of type i left in the queue is $\hat{Q}_t(i) = \hat{Q}_t\mu(i)/\mu$.

6.4 Departures

While one can use the contraction principle to obtain results for departures from a queue, the results are almost invariably messy. We present the results here, and give some examples of how things can go wrong. We advise the reader not to be too disheartened: by using the various moderate deviations scales more cleverly, one can find much nicer results. This we address in the next section.

6.4.1 Aggregate departures

As usual, we are interested in a queue with service rate $L\mu + L^{(1+\beta)/2}\hat{C}$ fed by an aggregate input $\mathbf{X}^L$. Let the queue size at time $-t$ be Q^L_{-t}. Recall that

$$Q^L_{-t} = L^{(1+\beta)/2}\hat{Q}_{-t}\big(L^{(1-\beta)/2}(L^{-1}\mathbf{X}^L - \mu\mathbf{1})\big).$$

Now let $\tilde{\mathbf{X}}^L$ be the departure process, defined by

$$\tilde{X}^L_{-t} = X^L_{-t} + Q^L_{-t} - Q^L_{-t+1}.$$

Does $\tilde{\mathbf{X}}^L$ satisfy a moderate deviations principle? We must look at $L^{(1-\beta)/2}(L^{-1}\tilde{\mathbf{X}}^L - \mu\mathbf{1})$:

$$\begin{aligned} L^{(1-\beta)/2}(L^{-1}\tilde{X}^L_{-t} - \mu) = {} & L^{(1-\beta)/2}(L^{-1}X^L_{-t} - \mu) + \\ & \hat{Q}_{-t}\big(L^{(1-\beta)/2}(L^{-1}\mathbf{X}^L - \mu\mathbf{1})\big) - \hat{Q}_{-t+1}\big(L^{(1-\beta)/2}(L^{-1}\mathbf{X}^L - \mu\mathbf{1})\big). \end{aligned}$$

Consider the map $\hat{\mathbf{D}} : \mathcal{X} \to \mathcal{X}$ defined by

$$\hat{D}(\hat{\mathbf{x}})_{-t} = \hat{x}_{-t} + \hat{Q}_{-t-1}(\hat{\mathbf{x}}) - \hat{Q}_{-t}(\hat{\mathbf{x}}).$$

The scaled departure process is given by

$$L^{(1-\beta)/2}(\tilde{\mathbf{X}}^L - \mu\mathbf{1}) = \hat{\mathbf{D}}\big(L^{(1-\beta)/2}(\mathbf{X}^L - \mu\mathbf{1})\big).$$

The departure map $\hat{\mathbf{D}}(\cdot)$ is continuous; this has been shown for example by O'Connell [21]. This means that we can use the contraction principle to deduce a moderate deviations principle for the output process.

Theorem 10 *The output process $\tilde{\mathbf{X}}^L$ satisfies a moderate deviations principle of the form*

$$\frac{1}{L^\beta}\log\mathbb{P}\big(L^{(1-\beta)/2}(\tilde{\mathbf{X}}^L - \mu\mathbf{1}) \in \hat{S}\big) \approx -\inf_{\hat{\mathbf{y}}\in\hat{S}} J(\hat{\mathbf{y}})$$

where

$$J(\hat{\mathbf{y}}) = \inf_{\hat{\mathbf{x}}:\hat{\mathbf{D}}(\hat{\mathbf{x}})=\hat{\mathbf{y}}} I(\hat{\mathbf{x}}).$$

This derived rate function $J(\hat{\mathbf{y}})$ is very difficult to deal with. In one case, though, we can simplify it.

Theorem 11 *Suppose $\mathbf{X}^L$ has asymptotically independent increments. Then*

$$J(\hat{\mathbf{y}}) = \begin{cases} I(\hat{\mathbf{y}}) & \text{if } \hat{y}_{-t} \le \hat{C} \text{ for all } t \\ \infty & \text{otherwise.} \end{cases}$$

Proof. If $y_{-t} > \hat{C}$ for some t then clearly $J(\mathbf{y}) = \infty$. So restrict attention to cases where $\hat{y}_{-t} \le \hat{C}$ for all t.

Thus $\hat{\mathbf{D}}(\hat{\mathbf{y}}) = \hat{\mathbf{y}}$. If $I(\hat{\mathbf{y}}) < \infty$ then $J(\hat{\mathbf{y}}) < \infty$. So if $J(\hat{\mathbf{y}}) = \infty$ then $I(\hat{\mathbf{y}}) = \infty$. So we can restrict attention to cases where $J(\hat{\mathbf{y}}) < \infty$.

In that case, since J is good, the optimum path in the infimum is attained, say at $\hat{\mathbf{x}}$. Suppose that $\hat{\mathbf{x}}$ causes the queue size to exceed 0 in some interval of time. Then there is a time $-t_0$ such that $\hat{x}_{-t_0} > \hat{C}$ and $\hat{x}_{-t_0+1} < \hat{C}$ (unless the interval of time ends at 0, in which case the modification to the argument is simple.) By removing a small amount of work at $-t_0$ and adding it at $-t_0+1$, we obtain a new path which leads to exactly the same departure process $\hat{\mathbf{D}}(\hat{\mathbf{x}})$. But the independent increments property implies that $I(\hat{\mathbf{x}}) = (2V)^{-1}\sum_t \hat{x}_{-t}^2$, and so the new path has a strictly lower rate function, contradicting optimality. □

So, when the input flow has asymptotically independent increments, the rate function for the departure flow is the same as the rate function for the input flow, at least for all feasible departure flows. This is not surprising: the same thing happens in large deviations, and the departure map is just the same for moderate deviations as for large deviations. Note that we have assumed a constant service rate; when the service is random, this result may not hold.

6.4.2 Individual departures

Unfortunately, this nice result only holds for the aggregate departure process from a queue fed by an input flow with asymptotically independent increments, served at constant service rate. When we look at the individual departure flows from a queue fed by two separate input flows, things are more complicated. But before we go on to give counterexamples, we ought to establish an MDP for the departure flows.

Theorem 12 *Define the scaled input process* $\underline{\hat{\mathbf{X}}}^L$ *by*

$$\underline{\hat{\mathbf{X}}}^L = L^{(1-\beta)/2}(\underline{\mathbf{X}}^L - \underline{\mu}\mathbf{1}),$$

define the actual departure process $\underline{\tilde{\mathbf{X}}}^L$ *by*

$$\underline{\tilde{X}}^L_{-t} = \underline{X}^L_{-t} + Q^L_{-t-1}(i) - Q^L_{-t}(i),$$

and define the departure map $\underline{\hat{\mathbf{D}}}$ *by*

$$\underline{\hat{D}}(\underline{\hat{\mathbf{x}}})_{-t} = \underline{\hat{x}}_{-t} + \hat{Q}_{-t}(\underline{\hat{\mathbf{x}}})\underline{\mu}/\mu - \hat{Q}_{-t+1}(\underline{\hat{\mathbf{x}}})\underline{\mu}/\mu.$$

Over finite intervals, the scaled departure process is exponentially equivalent to the departure map applied to the scaled input process, in that

$$\limsup_{L\to\infty} \frac{1}{L^\beta} \log \mathbb{P}\Big(\big|L^{(1-\beta)/2}(\underline{\tilde{\mathbf{X}}}^L - \underline{\mu}\mathbf{1})[-t,0) - \underline{\hat{\mathbf{D}}}(\underline{\hat{\mathbf{X}}}^L)[-t,0)\big| > \delta\Big) = -\infty.$$

Proof. We have already proved exponential equivalence for the amount of work of each type in the queue. But the departure map, over a finite interval, is a continuous function of finitely many of these terms. Hence the result. □

If we wanted to, we might attempt to prove exponential equivalence over infinite timescales. But it hardly seems worth it, given the following counterexample.

Example 5
Consider a queue serving two independent flows, $\mathbf{X}$ and $\mathbf{Y}$, both of which have asymptotically independent increments[5]. Let $\mathbf{X}$ have mean rate μ and variance V, and let $\mathbf{Y}$ have mean rate ν and variance W. Let the service rate be $\hat{C}$. Consider the most likely path to lead to $D(\hat{\mathbf{x}})[-t,0) = \alpha$. When $t = 2$, we can perform the calculation explicitly. Let $\mu = 0.9$, $\nu = 0.1$, $V = 1$, $W = 10$, $\hat{C} = 1$, and $\alpha = -3$. Then a most likely path is $(\hat{x}_{-2}, \hat{x}_{-1}) = (-0.34, 0.63)$ and $(\hat{y}_{-2}, \hat{y}_{-1}) = (0, 4.02)$, leading to $(D_{-2}(\hat{\mathbf{x}}), D_{-1}(\hat{\mathbf{x}})) = (-0.34, -2.66)$. The most likely path with $D_{-2}(\hat{\mathbf{x}}) = D_{-1}(\hat{\mathbf{x}})$ is $(\hat{x}_{-2}, \hat{x}_{-1}) = (-1.5, -1.5)$ and $(\hat{y}_{-2}, \hat{y}_{-1}) = (0, 0)$. Thus, the most likely path to lead to $D(\hat{\mathbf{x}})[-t,0) = \alpha$ is nonlinear. $\diamond$

This example shows that even if the input flows have asymptotically independent increments, and the service rate is constant, the output flows may not.

The phenomenon observed in this counterexample is called coupling. We say that two flows are *coupled*, if the most likely path $(\hat{\mathbf{x}}^*, \hat{\mathbf{y}}^*)$ to lead to $D(\hat{\mathbf{x}})[-t,0) = \alpha$ has $I(\hat{\mathbf{y}}^*) > 0$. Otherwise, we say they are *decoupled*. Coupling makes it difficult to describe traffic flow in networks.

6.4.3 Ramified networks

The troublesome phenomenon of coupling prompts us to seek alternative limiting regimes in which there is decoupling. In the many-flows large-deviations scale, such a limit has been described by Wischik [27].

Consider a queue serving very many independent flows, and suppose that each of these flows is routed to a different destination. We call such a queue, a *switch*, to emphasize that its purpose is to route input flows to different destinations.

In the large deviations scale, the fundamental result is this: that the essential characteristics of a *single flow* of traffic are not altered by passing through a switch, in the limit where the number of flows increase and the capacity of the switch increases in proportion.

It seems likely that a similar result holds in the moderate deviations scales. We will not pursue this line of study here, since the flexibility of the moderate deviations scales allow us to prove a novel result, which we do in the following section.

7 Mixed Limits

In a moderate deviations principle of the form

$$\frac{1}{L^\beta} \log \mathbb{P}\big(L^{-(1-\beta)/2}(L^{-1}\mathbf{X}^L - \mu\mathbf{1}) \in \hat{S}\big) \approx -\inf_{\hat{\mathbf{x}} \in \hat{S}} I(\hat{\mathbf{x}}),$$

[5] This is shorthand for the following. Consider a sequence of queues, indexed by L, in which the Lth queue serves two independent flows, $\mathbf{X}^L$ and $\mathbf{Y}^L$, which both satisfy the sample path MDP with asymptotically independent increments, and in which the service rate is $L(\mu+\nu) + L^{(1+\beta)/2}\hat{C}$, where μ and ν are the mean rates of the two flows..

which we will write suggestively as

$$\frac{1}{L^\beta}\log\mathbb{P}\big(\mathbf{X}^L\in L\mu\mathbf{1}+L^{(1+\beta)/2}\hat{S}\big)\approx-\inf_{\hat{\mathbf{x}}\in\hat{S}} I(\hat{\mathbf{x}}),$$

the most important feature is the not the form of the rate function I, but the scaling behaviour—the relationship between the frequency of the rare event $L^{-\beta}$ and its magnitude $\mathbf{X}^L\in L\mu\mathbf{1}+L^{(1+\beta)/2}\hat{S}$.

In this section we will consider systems whose various parts are scaled by different β. Such a study will produce a theory of scaling phenomena in networks which, one might expect, is cruder—and more useful—than the careful estimates of the last section.

7.1 Smoothing

Suppose a traffic flow is fed into a queue whose service rate is scaled by one parameter β, and that the output of this queue is fed into another queue whose service rate is indexed by β'. What is the overflow probability at the downstream queue? We will answer this, by answering a more general question: What are the statistical characteristics of the output flow at scale β'?

The answer turns out to be astonishingly simple. The queue smooths out all bursts of scale β' for $\beta'>\beta$, and leaves unchanged all bursts of scale β' for $\beta'<\beta$. (The bursts of scale β are smoothed in a complicated way, as described in Section 6.4.) In other words, the queue acts as a low-pass filter. We will discuss the implications of this result in the conclusion, Section 9, and leave the remainder of this section for the proof.

Consider as usual a queue fed by input process $\mathbf{X}^L$ and served at rate $L\mu+L^{(1+\beta)/2}\hat{C}$, with buffer $L^{(1+\beta)/2}\hat{B}$ (possibly $\hat{B}=\infty$). Let the queue size at time $-t$ be Q^L_{-t}, and let the output process be $\tilde{\mathbf{X}}^L$, defined as in Section 6.4 by

$$\tilde{X}^L_{-t}=X^L_{-t}+Q^L_{-t}-Q^L_{-t+1}.$$

Assume as usual that $\mathbf{X}^L$ satisfies the sample path MDP.

Theorem 13 *If $\beta'<\beta$ then $\mathbf{X}^L$ and $\tilde{\mathbf{X}}^L$ are moderately exponentially equivalent at the β' scale; that is, for any $\delta>0$,*

$$\limsup_{L\to\infty}\frac{1}{L^{\beta'}}\log\mathbb{P}\big(\|L^{(1-\beta')/2}(L^{-1}\mathbf{X}^L-\mu\mathbf{1})-L^{(1-\beta')/2}(L^{-1}\tilde{\mathbf{X}}^L-\mu\mathbf{1})\|>\delta\big)=-\infty. \tag{17}$$

This means that $\tilde{\mathbf{X}}^L$ satisfies exactly the same moderate deviations principle as does $\mathbf{X}^L$, at any scale $\beta'<\beta$.

Proof. Rewriting (17), we need to show that

$$\limsup_{L\to\infty}\frac{1}{L^{\beta'}}\log\mathbb{P}\Big(\sup_t\Big|\frac{X^L[-t,0)}{t}-\frac{\tilde{X}^L[-t,0)}{t}\Big|>\delta L^{(1+\beta')/2}\Big)=-\infty.$$

Since $\tilde{X}^L[-t,0)=X^L[-t,0)+Q^L_{-t}-Q^L_0$, we need to show that

$$\limsup_{L\to\infty}\frac{1}{L^{\beta'}}\log\mathbb{P}\Big(\sup_t\Big|\frac{Q^L_0}{t}-\frac{Q^L_{-t}}{t}\Big|>\delta L^{(1+\beta')/2}\Big)=-\infty.$$

Now,

$$\sup_t \Big| \frac{Q_0^L}{t} - \frac{Q_{-t}^L}{t} \Big| \le Q_0^L + \sup_t \frac{Q_{-t}^L}{t}.$$

Since Q_0^L satisfies a moderate deviations principle at scale β,

$$\limsup_{L\to\infty} \frac{1}{L^\beta} \log \mathbb{P}\big(Q_0^L \ge \delta L^{(1+\beta)/2}\big) \le -J(\delta),$$

where $J(\delta)$ is the rate function (5), and in particular $J(\delta) > 0$ for any $\delta > 0$. As $\beta' < \beta$,

$$\limsup_{L\to\infty} \frac{1}{L^{\beta'}} \log \mathbb{P}\big(Q_0^L > \tfrac{1}{2}\delta L^{(1+\beta')/2}\big) = -\infty.$$

Also, by the following lemma,

$$\limsup_{L\to\infty} \frac{1}{L^{\beta'}} \log \mathbb{P}\big(\sup_t \frac{Q_{-t}^L}{t} > \tfrac{1}{2}\delta L^{(1+\beta')/2}\big) = -\infty.$$

Putting these two together, we obtain the result. □

Lemma 14 *For $\delta > 0$,*

$$\limsup_{L\to\infty} \frac{1}{L^{\beta'}} \log \mathbb{P}\big(\sup_s \frac{Q_{-s}^L}{s} > \delta L^{(1+\beta')/2}\big) = -\infty.$$

Proof. We may assume without loss of generality that $B = \infty$, since the queue size in a finite-buffer queue is always less than or equal to the queue size in an infinite-buffer queue. Now,

$$\begin{aligned}
&\mathbb{P}\big(\sup_s Q_{-s}^L/s > \delta L^{(1+\beta')/2}\big) \\
=&\mathbb{P}\big(\sup_{s,t} \frac{1}{s}\big(X^L[-t-s,-s) - (L\mu + L^{(1+\beta)/2}\hat{C})t\big) > \delta L^{(1+\beta')/2}\big) \\
=&\mathbb{P}\big(\sup_{s,t} \frac{1}{s}\big(L^{(1-\beta')/2}(L^{-1}X^L[-s-t,-s) - \mu t) - L^{(\beta-\beta')/2}\hat{C}t\big) > \delta\big) \\
=&\mathbb{P}\big(\sup_s \frac{1}{s}\hat{R}_{-s}(\hat{\mathbf{X}}^L, L^{(\beta-\beta')}\hat{C}) > \delta\big)
\end{aligned}$$

where $\hat{\mathbf{X}}^L$ is the β'-scaled version of $\mathbf{X}^L$,

$$\hat{\mathbf{X}}^L = L^{(1-\beta')/2}(\mathbf{X}^L - \mu\mathbf{1}),$$

and $\hat{R}_{-s}(\hat{\mathbf{x}}, \hat{C})$ is the queue size function,

$$\hat{R}_{-s}(\hat{\mathbf{x}}, \hat{C}) = \sup_t \hat{x}[-s-t,-s) - \hat{C}t.$$

Thus

$$\limsup_{L\to\infty} \frac{1}{L^{\beta'}} \log \mathbb{P}\big(\sup_s Q_{-s}^L/s > \delta L^{(1+\beta')/2}\big) \le \limsup_{L\to\infty} \frac{1}{L^{\beta'}} \log \mathbb{P}\big(\hat{\mathbf{X}}^L \in M(\hat{C}')\big)$$

for any $\hat{C}'$, where

$$M(\hat{C}) = \{\hat{\mathbf{x}} : \sup_s \hat{R}_{-s}(\hat{\mathbf{x}}, \hat{C})/s \geq \delta\}.$$

We know that $\mathbf{X}^L$ satisfies a moderate deviations principle at scale β', and so

$$\limsup_{L\to\infty} \frac{1}{L^{\beta'}} \log \mathbb{P}\big(\sup_s L^{(1-\beta')/2}(\mathbf{X}^L - \mu\mathbf{1}) \in M(\hat{C}')\big) \leq - \inf_{\hat{\mathbf{x}}\in\bar{M}(\hat{C}')} I(\hat{\mathbf{x}}).$$

The following lemma shows that this infimum tends to ∞ as $\hat{C}' \to \infty$; hence the result. □

Lemma 15 *Let*

$$M(\hat{C}) = \{\mathbf{x} : \sup_s \hat{R}_{-s}(\hat{\mathbf{x}}, \hat{C})/s \geq \delta\}$$

and let $K(\hat{C}) = \inf_{\hat{\mathbf{x}}\in\bar{M}(\hat{C})} I(\hat{\mathbf{x}})$*. Then* $K(\hat{C}) \to \infty$ *as* $\hat{C} \to \infty$*.*

Proof. As we remarked in Section 6.4, the departure map is continuous with respect to the uniform norm $\|\cdot\|$. Thus the map $\hat{\mathbf{x}} \mapsto \sup_s \hat{R}_{-s}(\hat{\mathbf{x}}, \hat{C})$ is continuous. So $\bar{M}(\hat{C}) = M(\hat{C})$.

Suppose, without loss of generality, that $K(\hat{C}) < \infty$. Since I is a good rate function and $M(\hat{C})$ is closed, the infimum in $K(\hat{C})$ is attained, say at $\hat{\mathbf{x}}^*$. Since $\hat{\mathbf{x}}^* \in M(\hat{C})$, there exists some $s > 0$ such that $\hat{R}_{-s}(\hat{\mathbf{x}}^*, \hat{C}) > \delta s/2$, and thus there exists some t such that $x^*[-s-t, -s) > \hat{C}t + \delta s/4$. In particular, there exist s and t such that $\hat{x}^*[-s, -t, -s) > \hat{C}t$. By stationarity, and by Lemma 6, $I(\hat{\mathbf{x}}^*) \geq \hat{C}^2/2V_1$. This tends to ∞ as $\hat{C} \to \infty$. □

The inverse result is trivial because of the way we have set up our queueing model, with fixed service rates. Recall that $\tilde{\mathbf{X}}^L$ has passed through a queue of scale β, that is, with service rate $L\mu + L^{(1+\beta)/2}\hat{C}$. Suppose $\tilde{\mathbf{X}}^L$ is fed into a queue scaled by β' where $\beta' > \beta$, that is, a queue with service rate $L\mu + L^{(1+\beta')/2}\hat{C}'$. Whatever the values of $\hat{C}$ and $\hat{C}'$, for large enough L, the arrival rate at the downstream queue is strictly less than its service rate, so queues never build up at all. Thus the downstream queue size is (moderately) exponentially equivalent to 0.

Note that the departure process $\tilde{\mathbf{X}}^L$ is *not* moderately exponentially equivalent to 0. It is still possible for $\mathbf{X}^L$ to make a moderately large excursion below its mean rate, leading to a correspondingly large excursion of $\tilde{\mathbf{X}}^L$ below its mean rate. The upstream queue only smooths out large *positive* bursts of traffic.

Finally, note that the results in this section apply just as well to queues serving a mixture of several traffic flows. The characteristics of each flow sharing the queue, at scales $\beta' < \beta$, are not changed by passing through the queue. Since the flows do not influence each other (except trivially, through their mean rates), we say they are *decoupled*.

This result is tantalizingly similar to a result of de Veciana et al. [6], who study decoupling in the fast-time large deviations limit. They show that, if the service rate is high enough, and if one considers a sufficiently small scale of burstiness, the flows decouple in their marginal (short-timescale) distributions. They measure scales of burstiness by the large deviations tilt parameter (rather than by the moderate deviations scale β that we are using). Their technique is not crude enough to establish decoupling in our full sample-path sense [9].

7.2 A priority queue

The next example of mixed limits is of a priority queue in which the high priority flows are scaled by $\beta = 1$ and the low priority queues are scaled by $\beta < 1$.

Consider a priority queue fed by two flows. The high priority flow is $\mathbf{X}^L$, and the low priority flow is $\mathbf{Y}^L$. Let $\mathbf{X}^L$ have mean rate $L\mu$, let $\mathbf{Y}^L$ have mean rate $L\nu$, and let the service rate be $C^L = L\mu + L\nu + L^{(1+\beta)/2}\hat{C}$. Think of the many-flows limit, in which each of the two flows $\mathbf{X}^L$ and $\mathbf{Y}^L$ is the aggregate of L independent copies of base flows $\mathbf{X}$ and $\mathbf{Y}$ (though the argument also works in the fast-time limit.)

The limiting traffic intensity of the high priority flow is $\mu/(\mu+\nu)$, which is less than 1. We may think of the high priority flow as a high quality multimedia flow, which requires low traffic intensity in order to achive good enough quality of service.

The limiting traffic intensity of the aggregate is 1. In other words, the low priority flow pushes the traffic intensity up towards 1. We may think of the low priority flow as data traffic, which seeks to take up all available capacity, and is not very sensitive to loss or jitter.

Consider the high priority traffic on its own: it sees a total service rate of $C^L = L\mu + L\nu + L^{(1+\beta)/2}C$, so $L^{-1}C^L \to \mu + \nu$. Let Q^L be the amount of high priority work in the queue. Since the traffic intensity is less than 1, we can apply large deviations results:

$$\frac{1}{L}\log\mathbb{P}\big(Q^L > L\hat{b}\big) \approx -I_Q = -\inf_{t\geq 0}\sup_{\theta}\theta(\hat{b} + (\mu+\nu)t) - \log\mathbb{E}\exp(\theta\mathbf{X}[-t,0)).$$

Now consider the aggregate traffic $\mathbf{X}^L + \mathbf{Y}^L$, and let the total amount of work in the queue be $Q^L + R^L$. We have chosen the scaling so as to give a moderate deviations principle for aggregate queue size:

$$\frac{1}{L^\beta}\log\mathbb{P}\big(Q^L + R^L) > L^{(1+\beta)/2}\hat{b}\big) \approx -I_R = -\inf_{t\geq 0}\frac{(\hat{b} + \hat{C}t)^2}{2(V_t + W_t)}$$

where $V_t = \operatorname{Var} X[-t,0)$ and $W_t = \operatorname{Var} Y[-t,0)$.

By a small extension of the argument in Section 6.3, it can be shown that (at the moderate deviations scale) the aggregate queue size $Q^L + R^L$ is entirely made up of low priority work—that is, $Q^L = 0$. This seems at first sight to conflict with the previous estimate for Q^L. What it really means is that the most likely way for $Q^L + R^L$ to fill up to level $L^{(1+\beta)/2}\hat{b}$ is for R^L to do all the work in reaching that level; this happens with probability roughly $\exp(-L^\beta I_R)$. But Q^L can still produce work: it reaches level $L\hat{b}$ with probability roughly $\exp(-LI_Q)$. At the moderate deviations scale, this sort of event is completely swamped.

It is confusing to have to worry about all the different scalings: of arrivals, of service rates, of queue sizes, and of buffer sizes. To make things a little simpler, suppose that the buffer size for each type of traffic is very small: $\hat{b} \approx 0$. Then,

$$\mathbb{P}(Q \text{ overflows}) \approx \exp(-LI_Q),$$
$$\text{and } \mathbb{P}(R \text{ overflows}) \approx \exp(-L^\beta I_R).$$

One interpretation is that, by scaling the system in the right way, one can give significantly better service to the high priority flow. A large deviations analysis would not have revealed this: it would have found overflow probabilities which decay exponentially in L for both the high and low priority flows.

A different interpretation is that by adding a tiny amount of extra service $L\varepsilon$, both flows will have overflow probability which decays exponentially in L—in other words, both will have very good service, and so there is no point in using priorities.

8 Interpretation

We have finished with the formal part of the study. The purpose of this section is to highlight some of the properties of the moderate deviations scale.

In Sections 8.1–8.3 we will look at heuristics and approximations based on moderate deviations theory. We will see in which cases it is appropriate to use large deviations, moderate deviations, and heavy traffic theory.

In Section 8.4 we will look at the connection between large deviations and moderate deviations. After the last section, we hardly need to comment that the two theories are closely linked—the proofs are largely the same, and the main difference is the form of the rate function. We will dwell on effective bandwidth theory, and explain why it is not useful at the moderate deviations scale.

In Section 8.5 we will look at the connection between heavy traffic and moderate deviations, dwelling especially on the snapshot principle and on state space collapse.

In Section 8.6 we make some comments about the relationship between the burstiness of long-range-dependent traffic, described by the Hurst parameter, and the burstiness described by our parameter β.

8.1 Moderate deviations approximations

Forget for a moment that moderate deviations results are limiting results. What sort of estimates do they lead to for finite systems? Here is the heuristic for the many-flows scale.

Fix L large, and let Q^L be the queue size in a queue fed by $\mathbf{X}^L = \mathbf{X}^{\oplus L}$ (the aggregate of L independent copies of $\mathbf{X}$), served at rate $C^L = L\mu + L^{(1+\beta)/2}\hat{C}$, where $\mu = \mathbb{E}X_1$. Let $b^L = L^{(1+\beta)/2}\hat{b}$ and consider the event that $Q^L \geq b^L$. From Section 6.1,

$$\frac{1}{L^\beta}\log\mathbb{P}(L^{-(1+\beta)/2}Q^L \geq \hat{b}) \approx -\inf_{t\geq 0}\frac{(\hat{b}+\hat{C}t)^2}{2V_t} \tag{18}$$

where $V_t = \operatorname{Var} X[-t,0)$.

Now rewrite this estimate in terms of $\mathbf{X}^L$, C^L and b^L, and let $\mu^L = \mathbb{E}X_1^L$ and $V_t^L = \operatorname{Var} X^L[-t,0)$:

$$\begin{aligned}\log\mathbb{P}(Q^L \geq b^L) &\approx -L^\beta \inf_{t\geq 0}\frac{(L^{-(1+\beta)/2}b^L + L^{-(1+\beta)/2}(C^L-\mu^L)t)^2}{2L^{-1}V_t^L}\\ &= \inf_{t\geq 0}\frac{(b^L+(C^L-\mu^L)t)^2}{2V_t^L}.\end{aligned}$$

Conveniently, β has disappeared. This estimate should be good when L is large and $\mathbf{X}^L$, b^L and C^L are scaled as in the previous paragraph.

Now let us drop the superscript L: let Q be the queue size in a queue fed by $\mathbf{X}$ and served at rate C. The event that $Q > b$ may be estimated by

$$\log \mathbb{P}(Q \geq b) \approx -I \tag{19}$$

where

$$I = \inf_{t \geq 0} \frac{(b + (C - \mu)t)^2}{2V_t}. \tag{20}$$

This is precisely (18) with L set to 1. It is to be understood that this approximation is justified when $\mathbf{X}$, C and b stand in a certain relation (though of course one can compute the approximation for any $\mathbf{X}$, C and b).

I have belaboured this explanation, as the argument "we make this approximation for $L \to \infty$, thus for L large, thus for $L = 1$" might appear strange at first sight.

There is also a moderate deviations estimate for the fast-time limit. It turns out to be

$$\log \mathbb{P}(Q \geq b) \approx - \inf_{t \geq 0} \frac{(b + (C - \mu)t)^2}{2Vt} = 2\frac{b(C-\mu)}{V} \tag{21}$$

where

$$V = \lim_{t \to \infty} t^{-1} \operatorname{Var} X[-t, 0). \tag{22}$$

Here we have taken the infimum over $t \in \mathbb{R}^+$, for simplicity of calculation. (There are modifications for when the limit does not exist.) This estimate is generally worse than the many-flows estimate, even when there is only a single flow, because it does not take into account the short-timescale variance structure V_t.

These are estimates for isolated queues. For networks of queues, the low-pass filter result suggests the following approximation. The loss probability at a queue of scale β will be of order $L^{-\beta}$. Thus the loss probability for a flow through a network will be dominated by the loss probability at the queue along its path with the smallest scale $\beta_{\min}$. We will call this the *bottleneck link* of the flow. (In case of priority discipline, the result about state-space collapse shows that all flows except that with lowest priority will effectively see scale $\beta = 1$.) We can empirically identify the bottleneck link very simply: it is the queue with the highest frequency of overflow.

By the low-pass filter result, traffic is essentially unchanged (at scales less than or equal to $\beta_{\min}$) until it reaches the bottleneck link. Therefore we can use the estimate (20) for loss probability at the bottleneck link, without having to take account of any smoothing. Thus we can approximate the loss probability of a flow through a network, by the loss probability at its bottleneck link.

In fact, we do not even have to compare loss probabilities. We can simply apply the formula (20) to each queue along a path, and add up these estimates. The queue with the smallest scale will automatically dominate (as $L \to \infty$).

8.2 Other approximations

In the next section we will compare these moderate deviations estimates to those from large deviations and heavy traffic theory. First, a short summary.

The many-flows large deviations estimate, described in Wischik [29], is

$$\log \mathbb{P}(Q \geq b) \approx -I \tag{23}$$

where

$$I = \inf_{t \geq 0} \sup_{\theta} \theta(b + Ct) - \Lambda_t(\theta) \tag{24}$$

and

$$\Lambda_t(\theta) = \log \mathbb{E} \exp(\theta X[-t, 0)).$$

Refinements to this approximation have been studied by Likhanov and Mazumdar [17][6]. More on this approximation and its interpretation in terms of effective bandwidth in Section 8.4.

There is also a fast-time large deviations estimate, but it is less accurate.

For many-flows heavy traffic, one would approximate

$$\mathbb{P}(Q(\mathbf{X}) \geq b) \approx \mathbb{P}(Q(\mathbf{Y}) \geq b)$$

where $\mathbf{Y}$ is a Gaussian process with the same mean and covariance structure as $\mathbf{X}$. The right-hand side can be hard to calculate in the many-flows limit, so often we are restricted to the fast-time heavy traffic estimate, in which $\mathbf{Y}$ is a Brownian motion with drift, with parameters chosen so that $\mathbb{E}Y[-t, 0) = \mathbb{E}X[-t, 0)$ and $\operatorname{Var} Y[-t, 0) = Vt$, with V as in (22). By standard results for Brownian motion with drift,

$$\mathbb{P}(Q \geq b) \approx \exp\Big(-\frac{b(C - \mu)}{2V^2}\Big). \tag{25}$$

8.3 Numerics

It is illuminating to see how well these three estimates (many-flows moderate deviations, many-flows large deviations, and fast-time heavy traffic) do in practice.

[6] A better approximation is

$$\mathbb{P}(Q \geq b) \approx \frac{1}{\theta^* \sqrt{2\pi \Lambda''_{t^*}(\theta^*)}} e^{-I}$$

where I is as in (24), and t^* and θ^* are the optimizing values in I. This suggests a refinement of the moderate deviations approximation:

$$\mathbb{P}(Q \geq b) \approx \frac{1}{\theta^* \sqrt{2\pi V_{t^*}}} e^{-I}$$

where I is as in (20), t^* is the optimizing value in I, and $\theta^* = (b + (C - \mu)t^*)/2V_{t^*}$. This paper is concerned with moderate deviations, not with refined asymptotics, so we will not attempt to prove this conjecture.

Example 6 (Independent increments)
Let $\mathbf{X}$ be a random process in which all the X_{-t} are independent and identically distributed, $X_{-t} \sim \mathrm{Bin}(2,p)$. This gives

$$\log \mathbb{E} \exp(\theta X[-t,0)) = 2(pe^{\theta} + 1 - p),$$

$\mu = 2p$, $V_t = 2tp(1-p)$, and $V = 2p(1-p)$. Suppose $\mathbf{X}$ is fed into a queue with service rate 1. One can calculate exactly the queue size distribution:

$$\mathbb{P}(Q \geq b) = \Big(\frac{p}{1-p}\Big)^{2b}.$$

This may be compared to the large deviations estimate (23), the moderate deviations estimate (19), and the heavy traffic estimate (25). Figure 1 plots these three estimates as a function of traffic intensity ρ (where $p = \frac{1}{2}\rho$). (Actually, we plot the refined large and moderate deviations estimates.)

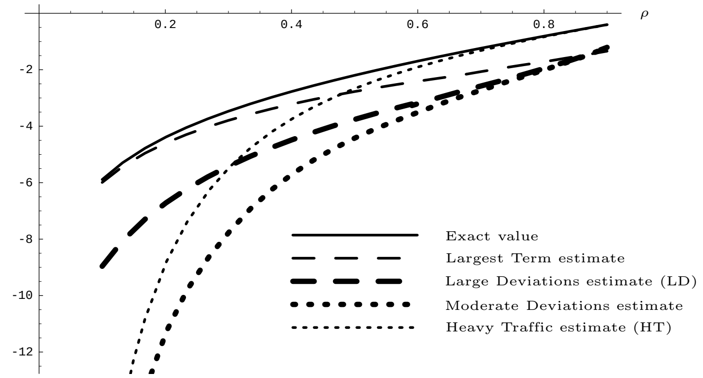

Figure 1: Estimates of $\log \mathbb{P}(Q \geq 1)$. The queue is fed by a source with independent increments and traffic intensity ρ, as described in Example 6—When traffic intensity is low, overflow depends on higher moments than just mean and variance (so HT is poor), and the principle of the largest term is a good approximation (so LD is reasonably good, though the exponential approximation part of LD is not perfect). When traffic intensity is high, overflow depends on just mean and variance (so HT is good), but the principle of the largest term is not a good approximation (so LD is poor). The moderate deviations estimate suffers from the flaws of both HT and LD.

When traffic intensity is greater than 30%, the heavy traffic estimate does well. This means that events $Q \geq b$ are well explained by just the mean and variance of $\mathbf{X}$. In the same range, the large deviations estimate does poorly: partly because the exponential approximation is inaccurate; mainly because the principle of the largest term is inaccurate.

When traffic intensity is lower than 30%, the large deviations estimate is better, because $Q \geq b$ comes about through peaks in the input, which are governed by higher moments. The heavy traffic estimate does not attempt to describe these peaks, so it does poorly.

The moderate deviations estimate is worst over the entire range of traffic intensities. When traffic intensity is higher than 30% it fails for the same reason as the large deviations estimate; when traffic intensity is lower, it fails for the same reason as the heavy traffic estimate. ◇

This looks like bad news for moderate deviations. We would expect the moderate deviations estimate to be worse than large deviations in all cases, since it is based on the same theory but uses a simplified traffic model, namely that the traffic is approximately Gaussian. When that is true, the heavy traffic estimate should be better.

Example 7 (Correlations)
Let $\mathbf{X}$ be a Markov chain on the state space $\{0,2\}$, with transition probabilities $\mathbb{P}(X_{t+1}=2 \mid X_t=0)=p$ and $\mathbb{P}(X_{t+1}=0 \mid X_t=2)=q$. In words, $\mathbf{X}$ is an on/off source which jumps from on to off with probability p, and from off to on with probability q, and when on produces 2 units of work each timestep. Consider feeding $\mathbf{X}$ into a queue with service rate 1. One can can calculate exactly the queue size distribution: for $b>0$,

$$\mathbb{P}(Q \geq b) = \frac{\alpha}{\lambda(1-\lambda)}\lambda^b$$

where $\lambda=(1-p)/(1-q)$ and

$$\alpha = \frac{1-\lambda+\lambda(p+q)}{(p+q)(\frac{1-q}{q}+\frac{1}{1-\lambda})}.$$

Again, compare this to the large deviations estimate (23), the moderate deviations estimate (19), and the heavy traffic estimate (25). Figure 2 plots these three estimates as a function of traffic intensity ρ, where the $\mathbf{X}$ is parameterized by $p=0.1$ and $q=p\rho/(2-\rho)$.

When traffic intensity is less than 80%, the large deviations estimate does well, because it takes into account the correlation structure of $\hat{\mathbf{X}}$. When traffic intensity is between 50% and 80%, the moderate deviations does nearly as well, meaning that the event $Q \geq b$ is governed largely by the mean and variance of $\mathbf{X}$. When traffic intensity is less than 50%, the event is governed by higher moments, so the moderate deviations estimate is poor. Note that the moderate deviations estimate is significantly easier to calculate.

The heavy traffic estimate is poor over most of the range of traffic intensities. This is because we are using the fast-time heavy traffic estimate, which ignores the short-timescale correlation structure of $\mathbf{X}$—which, in this example, governs the behaviour of the queue. The many-flows heavy traffic estimate would be much better, but it does not in general have a closed form solution. ◇

These examples show us when moderate deviations should be useful in practice. When traffic intensity is low, the behaviour of the queue is governed by higher moments than just the mean and variance; large deviations is best, because it takes into account the full distribution of the input process. When traffic intensity is high, the behaviour of the queue is governed largely by the first two moments; that is, the input process can be well-approximated by a Gaussian process, matched to have the same mean and covariance structure. If, over the timescale over which the queue overflows, the correlation structure is insignificant, use the fast-time heavy traffic approximation. If not, use the many-flows moderate deviations approximation.

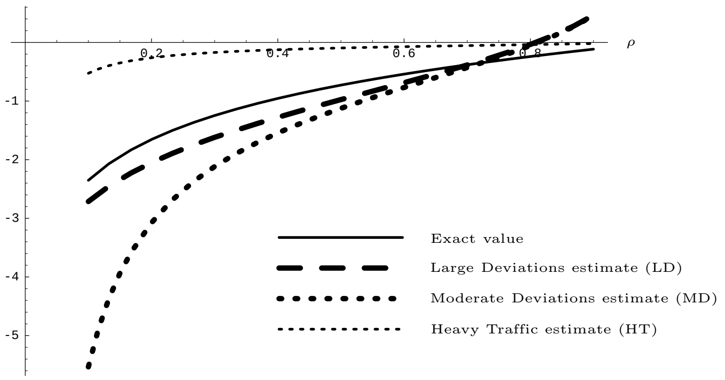

Figure 2: Estimates of $\log \mathbb{P}(Q > 1)$. The queue is fed by a source with strong correlations and traffic intensity ρ, as described in Example 7—At all traffic intensities, HT is bad because it ignores the correlation structure of the source. When traffic intensity is low, the principle of the largest term is a good approximation (so LD is reasonably good) though overflow depends on higher moments than just mean and variance (so MD is poor). When traffic intensity is moderate, the principle of the largest term is still a good approximation (so LD is reasonably good), and overflow depends only on mean and variance (so MD is good too). When traffic intensity is high, the principle of the largest term is not a good approximation (so LD and MD are poor).

Lest the reader be disheartened by the seemingly poor showing of moderate deviations estimates, it is appropriate here to reiterate some of their benefits. Moderate deviations estimates are parsimonious—they involve only the mean and covariance structure—so they are easier to work with than large deviations estimates. Moderate deviations estimates are simple—they are based on the principle of the largest term—so they are easier to work with than heavy traffic estimates. (Indeed, when the correlation structure is significant, the heavy traffic estimates are almost intractable.)

Furthermore, moderate deviations estimates are much easier to apply to networks. In both heavy traffic theory and large deviations theory, traffic is smoothed before it reaches the bottleneck link, and so one is simply not justified in applying the estimates we have described to downstream queues. (There are modified estimates, which involve knowledge of the entire structure of the network and all traffic flows through it; as one might expect, the calculation is significantly more complicated.) By contrast, the straightforward moderate deviations estimate can justifiably be applied to the bottleneck link, thanks to our result about queues as low-pass filters.

8.4 Large and moderate deviations

Let us return to the large deviations estimate

$$\log \mathbb{P}(Q \geq b) \approx -\inf_{t \geq 0} \sup_{\theta} \theta(b + Ct) - \Lambda_t(\theta) \tag{26}$$

where

$$\begin{aligned}\Lambda_t &= \log \mathbb{E} \exp(\theta X[-t,0)) \\ &= \theta\mu t + \tfrac{1}{2}\theta^2 V_t + O(\theta^3).\end{aligned}$$

The optimal θ^* and t^* are known as the *operating point*.

To understand this estimate better, it is helpful to think of it as based on two ideas: the *principle of the largest term*, and *exponential approximations*. For the first, the event $\{Q \geq b\}$ can be written $\{\exists t : X[-t,0) \geq b + Ct\}$, and the principle of the largest term says

$$\mathbb{P}(Q \geq b) \approx \sup_t \mathbb{P}(X[-t,0) \geq b + Ct).$$

For the second, the exponential approximation is that

$$\log \mathbb{P}(X[-t,0) \geq b + Ct) \approx -\sup_{\theta>0} \theta(b+Ct) - \log \mathbb{E}\exp(\theta X[-t,0)).$$

These two elementary ideas are a necessary part of any large deviations result; putting them together gives precisely the large deviations many-flows estimate (26).

If one replaces the full log moment generating function Λ_t by its second-order approximation one obtains the moderate deviations estimate (19).

More deeply, the large deviations heuristic (26) can guide us to the full moderate deviations approximation (18) as follows. Replace the parameters according to the moderate deviations scale: replace b by $L^{(1+\beta)/2}\hat{b}$, C by $L\mathbb{E}X_{-1} + L^{(1+\beta)/2}\hat{C}$ and $\mathbf{X}$ by the aggregate of L copies of $\mathbf{X}$. If Λ_t is sufficiently nice, then in the limit as $L \to \infty$,

$$\frac{1}{L^\beta} \sup_\theta \theta(b+Ct) - \Lambda_t(\theta) \to \sup_\phi \phi(\hat{b} + \hat{C}t) - \tfrac{1}{2}\phi^2 V_t.$$

(The optimal θ^L is roughly $L^{-(1-\beta)/2}\phi$.) At a purely symbolic level, ignoring all the complications of probability theory, this is exactly the moderate deviations limit theorem (18).

The estimate (26) can help us interpret the idea of effective bandwidth. The effective bandwidth of a source with log moment generating function Λ_t is $\alpha(\theta,t) = (\theta t)^{-1}\Lambda_t(\theta)$. This is a convenient parameterization: $\alpha(\theta,t)$ lies between the mean and the peak of $X[-t,0)$; and it has the interpretation that (in a queue with many input processes, with operating point θ^* and t^*) replacing some flows by constant-rate flows of rate $\alpha(\theta^*,t^*)$ does not alter the overflow probability. For more details see Wischik [29] and Kelly [15].

In the moderate deviations scale, indexed by L, we have just seen that $\theta^L \sim L^{-(1-\beta)/2}\phi$. In other words, in the moderate deviations scale, we are looking at behaviour that is ever closer to the mean. The relevant effective bandwidth is thus ever closer to the mean: $\alpha(\theta^L,t) = \mu + \frac{1}{2}L^{-(1-\beta)/2}\phi V_t/t$.

Thus, at the moderate deviations scale, the aggregate process is approximated well by a Gaussian process with the same mean and covariance structure. If one takes a large deviations estimate, and simply replaces the log moment generating function Λ_t by its second-order approximation $\theta V_t/t$, one obtains exactly the moderate deviations estimate.

Now, 'real' traffic processes are never Gaussian—there is no such thing as negative work arriving at a queue. When we study large deviations of Gaussian processes, we are therefore implicitly dealing with some moderate-deviation-like scaling.

We must therefore exercise caution in interpreting large deviations results for Gaussian processes, especially in interpreting the effective bandwidth of a Gaussian process.

8.5 Moderately heavy traffic

Two very important ideas from heavy traffic theory—state space collapse, and the snapshot principle—apply also to moderate deviations theory.

In Section 6.3 we considered a single queue fed by several independent flows $\mathbf{X}(i)$, $i = 1..N$, with mean arrival rates $\mu(i)$ and aggregate mean arrival rate $\mu = \sum_i \mu(i)$. We assumed that work arriving from a flow i at time $-t$, $X_{-t}(i)$, was distributed uniformly throughout the timeslot $[-t, -t+1)$, and that work was served in the order it arrived. Let the total queue size be Q, and let the amount of work in the queue due to flow i be $Q(i)$. The conclusion was that $Q(i)$ is approximately $Q\mu(i)/\mu$. (The precise sense of the approximation is that the two quantities are moderately exponentially equivalent, which means that they satisfy the same moderate deviations principles.)

Consider instead a priority queue discipline: suppose that all work $X_{-t}(i)$ arrives at time $-t$ and is served in order of priority, $X_{-t}(1)$ first then $X_{-t}(2)$ and so on. With a small modification to the proof in Section 6.3, one can show that $Q(i) \approx 0$ for $i < N$, and so all the work in the queue belongs to flow N. (Again, this is in the moderately exponentially equivalent sense.)

In each case, the single variable Q dictates the values of the other quantities $Q(i)$. This phenomenon is known in heavy traffic theory as *state space collapse*—see Harrison and Van Mieghem [13] and Reiman [23]. We have just seen two simple examples of state space collapse in systems with fixed service policies; in fact, the concept is most powerful when applied to systems in which the service policy can be changed dynamically, based on the state of the system. We conjecture that there is a similar control theory for queueing systems at the moderate deviations scale.

In large deviations, there is no state space collapse. For example, [29] shows that in a priority queue with two input flows, the distribution of work in the queue will depend on the queue size.

Section 6.3 has another insight to give. In the course of the proof, we showed that all the work arriving at time $-t$ is served either at $-t$ or at $-t+1$. (Again, this is a statement about moderate exponential equivalence.)

Our decision to work in discrete time obscures the key idea here. Perhaps a more helpful way to describe it is as follows. Suppose the queue size has reached $Q = L^{(1+\beta)/2}\hat{b}$, in the usual moderate deviations scaling. This is most likely to happen over a critical timescale t^*—the optimizing value of t in (14), Section 6.1. The service rate is $C = L\mu + L^{(1+\beta)/2}\hat{C}$. If the service were continuous, an arriving piece of work would depart the queue in roughly $L^{-(1-\beta)/2}\hat{b}/\mu$.

Thus the timescale over which work passes through a queue is of order $L^{(1-\beta)/2}$ quicker than the timescale over which the queue size fluctuates. In a feedforward network of queues, the state of the network (that is, the amount of work at each of the queues) will barely change in the time it takes a piece of work to pass completely through the network. It is as if each piece of work, as it passes through the network, observes a snapshot of the current state. This is called the *snapshot principle*.

The snapshot principle in heavy traffic theory has been described by Reiman [24]. As might be expected, in heavy traffic ($\beta = 0$) the timescale of queueing delay is of order $L^{1/2}$ quicker than the timescale of queue size fluctuations. In large deviations theory ($\beta = 1$), the two timescales are roughly similar, and the snapshot principle does not apply. One must therefore take great care in applying large deviations theory about departure processes to moderate deviations.

8.6 Long-range dependence

A common model for long-range dependence in Internet traffic is *fractional Brownian motion*. In discrete time, this is characterized as follows. A standard fractional Brownian motion with Hurst parameter H is a Gaussian process $\mathbf{Z}$ which has mean 0 and variance $\operatorname{Var} Z_t = \sigma^2 t^{2H}$. Consider the arrival process $\mathbf{X} = \mu\mathbf{1} + \sigma\mathbf{Z}$.

Because $\mathbf{X}$ does not have independent increments it is difficult calculate, for example, the distribution of $Q(\mathbf{X})$. This has motivates the study of large deviations for long-range dependent processes.

In the many-flows limit, things are simple. Wischik [29] shows that $\mathbf{X}^{\oplus L}$ satisfies exactly the same sort of large deviations principle as any other Gaussian process; and Example 1 shows the same for moderate deviations.

In the fast-time limit, $\mathbf{X}$ is more interesting. Processes which satisfy the normal fast-time large deviations principle of Section 3.3 must have variance growing linearly; this is not the case for fractional Brownian motion. Nonetheless it is possible to recover a large deviations principle, by choosing a different scaling. Let $\hat{\mathbf{X}}^L = N^{-1}\mathbf{X}^{\otimes N}$, where $N = L^{1/2(1-H)}$. Then one can obtain a large deviations principle of the form

$$L^{-1} \log \mathbb{P}(\hat{\mathbf{X}}^N \in \hat{S}) \approx -\inf_{\hat{x} \in \hat{S}} I(\hat{\mathbf{x}}).$$

Rewriting in terms of the natural scale of queue size,

$$\frac{1}{L^{2\left(1+\frac{H}{1-H}\right)^{-1}}} \log \mathbb{P}(Q_0 \geq L\hat{b}) \approx -J(\hat{b}).$$

This means that for $H > \frac{1}{2}$, the tail of the queue size is subexponential. This fact has attracted much attention and caused some alarm. See Duffield and O'Connell [8] for a good mathematical account. One interpretation is that one needs very large buffers to give low overflow probability. A more robust interpretation is that large buffers are not a good way to reduce loss probability—that loss probability is best reduced by aggregating more flows.

The reciprocal of the prefactor is known as the *speed* of the large deviations principle. In this case, the speed is $L^{2(1-H)}$, whereas for regular large deviations the speed is L. For this reason, Chang et al. [3] refer to this as a moderate deviations limit. Since the speed can be arbitrarily changed, by reparameterizing the limit, I prefer to restrict the term 'moderate deviations' to cases where the speed is not essentially determined by the input process $\hat{\mathbf{X}}$.

Moderate deviations is a special case of large deviations, so it should not be surprising to see the equivalent moderate deviations result, proved in Example 4:

$$\frac{1}{L^\beta} \log \mathbb{P}\big(L^{(1-\beta)/2}(N^{-1}\mathbf{X}^{\otimes N} - \mu\mathbf{1}) \in \hat{S}\big) \approx - \inf_{\hat{\mathbf{x}} \in \hat{S}} I(\hat{\mathbf{x}}).$$

This yields estimates for queue size of the form

$$\frac{1}{L^{2\left(1+\frac{H}{1-H}\frac{1}{\beta}\right)^{-1}}} \log \mathbb{P}(Q_0^L \geq L\hat{b}) \approx -J(\hat{b}), \tag{27}$$

where Q_0^L is a queue with service rate

$$\mu + L^{-(1-(H+\beta-H\beta)^{-1})}\hat{C},$$

and thus traffic intensity

$$\rho \sim 1 - L^{-(1-(H+\beta-H\beta)^{-1})}\hat{C}/\mu.$$

This indicates a tradeoff between utilization and long-range dependence.

While H and β play similar roles in (27), it is important to be clear on their differences. The Hurst parameter H is a single quantity which describes the degree of self-similarity of a traffic flow. Scaling arguments suggest that when a flow with Hurst parameter H passes through a queue, the output flow has exactly the same H. On the other hand, a typical traffic flow has bursts over many scales $\beta \in (0, 1)$, and these scales coexist. When a flow passes through a queue with a certain burst scale β', the output has bursts over scales $\beta < \beta'$ but none over scales $\beta > \beta'$.

9 Conclusion

The technical contribution of this paper has been the presentation of a family of moderate deviations probability estimates, intermediate between heavy traffic and large deviations theory. Moderate deviations estimates share the benefits of both extremes: they are parsimonious, like heavy traffic models; and they are easy to work with, like large deviations models. They reflect a regime in which utilization is high, as it is in heavy traffic models; yet in which overflow is rare, as it is in large deviations models.

The real goal of the paper is more ambitious. I hope to have drawn attention to the importance of scaling phenomena in queueing networks. It is typical for mathematicians to rescale, centre, and otherwise massage problems into cases with interesting mathematical structure. We might assume for example that a system is scaled in an appropriate way so as to yield a limiting Brownian motion, and then study in great detail the characteristics of systems fed by Brownian motion. However, it is often the case that the scaling is far more important than the detailed study of the limit.

Moderate deviations theory describes a *family* of scales, indexed by a parameter $\beta \in (0, 1)$, whereas both heavy traffic theory and large deviations theory restrict attention to a single scale. We have therefore been able to study a range of scaling phenomena, using a single set of tools.

In particular, we have seen that traffic has bursts at many scales, with large bursts less frequent than small bursts. A queue, which overflows with a certain frequency, has its own characteristic scale; and overflow is governed by the burstiness of the traffic at that scale. We have seen that a queue acts as a low-pass filter on traffic bursts. This allows us to identify the bottleneck link on a path—the link with the highest frequency of overflow—and to justify the common assumption that traffic is essentially unsmoothed until it reaches that link. This effect is *not seen* in either large deviations theory or heavy traffic theory.

References

[1] Ron Addie, Petteri Mannersalo, and Ilkka Norros. Performance formulae for queues with Gaussian input. In *Proceedings of ITC 16, Edinburgh*, pages 1169–1178, 1999. URL `http://www.vtt.fi/tte/tte23/traffic/papers/Gaussperf.ps`.

[2] Ronald G. Addie, Moshe Zukerman, and Timothy D. Neame. Application of the central limit theorem to communication networks. Technical Report SC-MC-9819, University of Southern Queensland, 1998.

[3] Cheng-Shang Chang, David D. Yao, and Tim Zajic. Large deviations, moderate deviations, and queues with long-range dependent input. *Advances in Applied Probability*, 31:254–278, 1999. URL `http://news.ee.nthu.edu.tw/cschang/md.ps`.

[4] Jinwoo Choe and Ness B. Shroff. A central limit theorem based approach for analyzing queue behaviour in ATM networks. *IEEE/ACM Transactions on Networking*, 6(5):659–671, 1998. URL `http://yara.ecn.purdue.edu/~shroff/NEWS/abstracts/j4.html`.

[5] Jinwoo Choe and Ness B. Shroff. Supremum distribution of Gaussian processes and queueing analysis including long-range dependence and self-similarity. Submitted to Stochastic Models, 1999.

[6] G. de Veciana, C. Courcoubetis, and J. Walrand. Decoupling bandwidths for networks: a decomposition approach to resource management for networks. In *Proceedings of IEEE Infocom*, volume 2, pages 466–474, 1994. URL `http://walrandpc.eecs.berkeley.edu/Papers/dec.pdf`.

[7] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Springer, 2 edition, 1998.

[8] N. G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Mathematical Proceedings of the Cambridge Philosophical Society*, 118:363–374, 1995.

[9] A. J. Ganesh and Neil O'Connell. The linear geodesic property is not generally preserved by a FIFO queue. *Annals of Applied Probability*, 8(1): 98–111, 1998. URL `ftp://hplose.hpl.hp.com/pub/noc/papers/9606.ps`.

[10] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–588, 1981.

[11] J. Michael Harrison. *Brownian motion and stochastic flow systems*. John Wiley and Sons, New York, 1985.

[12] J. Michael Harrison. Brownian models of queueing networks with heterogeneous customer populations. In W. Fleming and P.-L. Lions, editors, *Stochastic differential systems, stochastic control theory and applications*, volume 10 of *IMA Volumes in Mathematics and its Applications*, pages 147–186. Springer, 1988.

[13] J. Michael Harrison and Jan A. Van Mieghem. Dynamic control of brownian networks: state space collapse and equivalent workload formulations. *Annals of Applied Probability*, 7(3), 1997.

[14] F. P. Kelly, S. Zachary, and I. Ziedins, editors. *Stochastic Networks: Theory and Applications*. Royal Statistical Society Lecture Note Series. Oxford, 1996.

[15] Frank Kelly. Notes on effective bandwidths. In Kelly et al. [14], chapter 8, pages 141—168. URL `http://www.statslab.cam.ac.uk/~frank/eb.html`.

[16] J. F. C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society B*, 24:383–392, 1962.

[17] N. Likhanov and R. R. Mazumdar. Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *Journal of Applied Probability*, 36:86–96, 1999.

[18] Kurt Majewski. Path-wise heavy traffic convergence of single class queueing networks and consequences. Preprint, 2000.

[19] Ilkka Norros. Most probable paths in Gaussian priority queues. COST257TD(99)16, 1999.

[20] Neil O'Connell. Queue lengths and departures at single-server resources. In Kelly et al. [14], chapter 5. URL `ftp://hplose.hpl/hp.com/pub/noc/papers/9604.ps`.

[21] Neil O'Connell. A large deviation principle with queueing applications. Technical Report HPL-BRIMS-97-05, BRIMS, Hewlett Packard Labs, Bristol, March 1997. URL `ftp://hplose.hpl/hp.com/pub/noc/papers/9705.ps`.

[22] Anatolii A. Puhalskii. Moderate deviations for queues in critical loading. *Queueing Systems*, 31:359–392, 1999.

[23] M. I. Reiman. Some diffusion approximations with state space collapse. In F. Baccelli and G. Fayolle, editors, *Modelling and performance evaluation methodology*, number 60 in Lecture notes in control and information sciences. INRIA, Springer-Verlag, 1984.

[24] Martin I. Reiman. Open queueing networks in heavy traffic. *Mathematics of Operations Research*, 9(3):441–458, 1984.

[25] A. Weiss. A new technique for analyzing large traffic systems. *Advances in Applied Probability*, 18:506–532, 1986.

[26] R. J. Williams. On the approximation of queueing networks in heavy traffic. In Kelly et al. [14], chapter 3, pages 35–56.

[27] Damon Wischik. The output of a switch, or, effective bandwidths for networks. *Queueing Systems*, 32:383–396, 1999. URL `http://www.statslab.cam.ac.uk/~djw1005/Stats/Research/output.html`.

[28] Damon Wischik. *Large deviations and Internet congestion.* PhD thesis, University of Cambridge, 2000. URL `http://www.statslab.cam.ac.uk/~djw1005/Stats/Research/phd.html`.

[29] Damon Wischik. Sample path large deviations for queues with many inputs. URL `http://www.statslab.cam.ac.uk/~djw1005/Stats/Research/sampleldp.html`. To appear in Annals of Applied Probability, 2001.