Worst Case Traffic from Sources Constrained by Leaky Buckets

by

C. P. Walsh
B.A. (Mod.)

A thesis submitted to
Dublin City University
for the degree of
Doctor of Philosophy

Supervisor: Emmanuel Buffet
School of Mathematical Sciences
Dublin City University

April 1999

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Cormac Walsh

ID NO.: 94970904

Date: September 30, 2000

Abstract

This thesis investigates the worst possible behaviour of a source if the traffic emanating from it is constrained to pass through a selection of leaky buckets. Various criteria for judging the worst case are considered, including the average queue length when the traffic is passed through an infinite buffer served at a constant rate and the rate of loss when the traffic is passed through a finite buffer, again served at a constant rate. Both of these criteria may be used when the source traffic is buffered alone or in combination with other traffic. Another functional considered is the effective bandwidth function which governs the asymptotic loss rate when the number of sources becomes infinite. This functional turns out to be the most tractable and we concentrate our attention on it. In all cases considered it is found that the functional to be maximised is convex in the space of traffic processes. This leads to the use of convex optimisation methods to characterise the worst case traffic process. In addition, Optimal Control Theory is used to show that the worst case traffic exhibits periodic behaviour.

Acknowledgments

First and foremost I would like to thank John Lewis for all the guidance he has given me. His support and understanding have been invaluable.

Fergal Toomey suggested my thesis problem to me, for which I will be forever grateful. I doubt he could have suspected at the time how much mileage I would get from it.

I would like to thank Mark Dukes, my heroic proof reader, and apologise for not being able to respond in kind.

Thanks also to the other members of the APG who include: Raymond Russell, Meriel Huggard, Brian McGurk (who showed me how to juggle), Ken Duffy, Marc Corluy, and Sean Coffey. Together they have made DIAS a lively and interesting place to work.

A special thanks go to the librarian of DIAS, Anne Goldsmith, and its secretary, Margaret Matthews. I would also like to thank Nick Duffield and Neil O'Connell for the many interesting discussions I have had with each. Thank you Ian Dowse for keeping my machine running, and Emmanuel Buffet for doing the unrewarding job (from his point of view) of liaising with DCU.

Lastly, my family deserves special mention for putting up with me through stressful times.

Contents

Chapter 1

Introduction

We discuss here in general terms the problem addressed in this thesis.

1.1 Motivation

Our problem takes its motivation from the emerging technology of Asynchronous Transfer Mode. This is the proposed standard for the future Broadband Integrated Services Digital Network. In this technology all traffic (data, video, voice) is sent in small, fixed length packets called cells. These cells are routed on the basis of information contained in their headers. This information is not a global address, it is merely a label to identify the cell. Explicit addressing is not practical because the small cell size would make the overhead too great. When a cell arrives at a switch the switch uses a lookup table to decide which output port to send it to. It also changes the label to something understood by the next switch along the cell's route. This new label will be used by the next switch to send the cell to the appropriate output port and to change the label again. All routing and creation of lookup tables is done at connection setup. This arrangement allows routing and Connection Acceptance Control (CAC) to be done once per connection, allowing individual cells to be switched with the minimum overhead.

Quality of Service (QoS) is guaranteed using CAC: a connection is accepted only if there is enough spare bandwidth at each node along the connection's route to guarantee the appropriate cell delay and loss probabilities. To carry out CAC the network must have a reasonable description of the characteristics of the offered traffic. This is given

at connection setup when a traffic contract is agreed. The user contracts that his traffic will conform to certain parameters, perhaps defining a policing algorithm that bounds properties of the traffic. In return the network contracts to carry this traffic with a particular QoS which may be specified in terms of delay and loss. The traffic contract gives the network information that will bound the network resources that will be required to carry the call.

Any charging scheme will typically be based on both the parameters of the contract and on quantities measured once the user has started transmitting: an incentive is needed for the user to accurately describe its anticipated usage and also to minimise this usage once transmission is underway.

The network may have no additional information about user traffic and must therefore infer its characteristics from the descriptors specified in the contract. In any decisions the network makes concerning resource allocation or call acceptance, it would be prudent of the network to assume the worst case, that is that the user is adversarial to the maximum extent permitted by the policer.

1.2 The Problem

The considerations above motivate an optimisation problem over the space of traffic processes. We must choose a functional on this space to represent the performance of the network, and a constraint to represent the action of the policer. We will discuss some possible choices of functionals and constraints presently.

A major decision is whether to insist that the traffic sources be stationary and independent. Some authors do not, for example [1, 2]. Without this restriction the character of the optimisation problem becomes deterministic. These authors can guarantee tight bounds on delay and queue length. Often, however, it is reasonable to assume that the sources know nothing about each other. In this case the network may extract significant multiplexing gains by allowing small loss probabilities. We shall therefore focus on this case.

Functionals

There are many functionals that could plausibly represent the degree of congestion in a network. We shall concentrate on those associated with the simplest queueing system: the single server queue. Other possibilities include functionals relating to systems such as the fair queueing system introduced by Parekh and Gallager [2] and those relating to various priority schemes. In Chapter 2 we shall define our functionals rigorously; for the moment we will be content to describe them informally. Amongst these is the class of functionals relating to the queue length in an infinite buffer being served at constant rate. For example, a simple functional is the average queue length when the source passes alone through the buffer. The worst case traffic for this functional is somewhat trivial as will be shown in Chapter 6. More generally we are interested in the situation where there is some fixed number of identical stationary independent sources feeding the buffer.

Another situation is where the source is combined with a number of stationary sources with fixed statistics. By this we mean that their behaviour is fixed and is not to be optimised over. Since the sum of two independent stationary processes is a stationary process, the case of many fixed sources reduces to that of one. The following simple argument shows that the problem of maximising the average queue length of a source multiplexed with another fixed source is the same as that of maximising the average queue length of a single source when the service is a certain stationary stochastic process. Since the average value of the total queue length is our only concern, we may think of the fixed source as receiving priority. The buffer space occupied by this source will then be independent of the statistics of the source over which we are optimising. The only thing which changes is the average queue length of the optimised source. This will be equal to the average queue length when the source is passed alone through a buffer being served at a rate equal to the service that is not being used by the fixed source. This unused service rate will be a stationary stochastic process.

Rather than being concerned with the delay in an infinite buffer, we may be concerned with the loss from a finite buffer. Again, we may consider a single source, a finite number of identical independent sources, or a multiplex of the source with a fixed source. By reasoning in a manner similar to above, we see that maximising this latter functional is equivalent to maximising the loss when the buffer size and service capacity are particular

stationary processes with fixed statistics.

Another functional, the maximisation of which turns out to be particularly tractable, is motivated by the study of the asymptotic loss rate as the number of sources becomes large. We can study the asymptotic loss in this regime using the techniques of large deviations and in Section 2.7 we show that finding the worst case traffic in this case reduces to minimising the value of a functional called the *effective bandwidth* of the source.

Constraints

In [3] Turner describes the "leaky bucket" mechanism for traffic regulation. This consists of a counter for each connection that is incremented whenever a cell from that connection arrives and is decremented periodically. If the counter exceeds a fixed threshold, the cell responsible is declared a non-conforming cell. These cells are discarded, delayed, or tagged as a low priority cells. The user specifies both the rate σ at which the counter is incremented and the threshold β. We say that a traffic stream conforms to the leaky bucket constraint (σ, β) if none of its cells are non-conforming.

So far we have been considering discrete cells. However in this thesis we will mainly work with fluid models. There is an obvious continuous version of the leaky bucket regulator. Here continuous fluid is added at a constant rate to a buffer, and is taken from the buffer at a rate equal to that of the arriving traffic stream. Again, a stream is conforming if the fluid never exceeds a level β.

The discrete cell leaky bucket is the standard policing mechanism defined by the ATM Forum [4] under the name Generic Cell Rate Algorithm. There are various traffic classes defined by the ATM Forum, two of which are the Constant Bit Rate (CBR) and Variable Bit Rate (VBR) classes. CBR traffic is only required to meet a single constraint of the form $(\rho, 0)$. Such a constraint is just a bound on the peak rate: the instantaneous transmission rate at any time must be less than ρ. In practice the leaky bucket threshold is not zero but some small amount called the *cell delay variation tolerance*. This allows a traffic stream to meet the constraint even if it encounters some jitter. The worst case source under a single peak rate constraint is trivial: it is the source that transmits at a constant rate ρ. Our main concern will therefore be the next simplest case, that of two constraints. Again, ignoring the CDV tolerance, there is a peak rate constraint $(\rho, 0)$. However, there is now

also a mean rate constraint (σ, β). In the definition of the VBR traffic class the parameters σ and β are called the *sustainable cell rate* and the *intrinsic burst tolerance.*

1.3 Previous Results

Deterministic Bounds

Parekh and Gallager [2] consider the deterministic worst case. They model the multiplexor as a finite number of queues, one for each input, being served by a single server. Each input is constrained by its own pair of leaky bucket parameters (σ_i, β_i) and $(\rho_i, 0)$. In addition each input has a weight ϕ_i associated with it that governs the amount of service it will receive. The service is divided up amongst the non-empty queues in proportion to their weights so that input i will receive rate $\phi_i / \sum_{j \in I} \phi_j$, where I is the set of queues that are not empty. There are two functionals of interest: the maximum delay experienced by any one of the sources, and the maximum queue length it experiences. No assumptions are made about the stationarity or independence of the sources. They show that the worst case for both of these functionals occurs when the sources are greedy, that is when each source i transmits at rate ρ_i for time $\beta_i/(\rho_i - \sigma_i)$ and then transmits rate σ_i continuously. Note that the sources are behaving collusively by synchronising the start of their bursts.

The deterministic worst case is similar to the statistical worst case when the functional to be maximised represents a system in which there is only one source. This was investigated by Lee [5] for average queueing delay.

Stationary, Independent Sources

Mitra and Morrison [6] investigate the worst case loss probability in a bufferless server with constant service rate. They use a large deviations estimate for the asymptotic loss as the number of sources becomes infinite, and show that this estimate depends only on the distribution of the transmission rate. The leaky bucket policers place constraints on this distribution. Firstly, the average transmission rate must not exceed the smallest leaky bucket rate m. Secondly, the essential supremum of the distribution must not be exceed the peak rate constraint ρ. The distribution satisfying these two constraints that maximises the large deviation estimate of the loss is shown to have mass m/ρ at 0 and

mass $1 - m/\rho$ at ρ. Furthermore they show that there is a periodic traffic source with random phase whose transmission rate has this distribution.

Oechslin [7] considers the special case where the server has a service rate equal to the sum of mean rates of the input traffic and the buffer size is close (within ϵ) to the total burst size of the sources. He finds an upper bound on the expected loss rate experienced by any fixed number of independent on-off sources and a lower bound on that experienced by symmetric sources. By making ϵ small enough, he can make the lower bound exceed the upper bound and thus prove that there are cases where on-off traffic is not the worst case.

Doshi [8] tackles essentially the same problem we do: he tries to maximise the loss ratio for a finite number of independent sources. In addition to considering the situation where there are N identical sources (the homogeneous case) he also tackles the inhomogeneous case, that is where $N - 1$ of the sources remain fixed and the maximum is taken with respect to the behaviour of the remaining source. He shows that when the buffer size is zero, the worst case source is always on-off. We give details in Section 6.1. He also demonstrates with a counter-example that the loss rate for two identical and independent sources is not maximised when they are on-off.

1.4 The Conjecture

Consider the loss rate when a single source passes through a single server queue. We will show in Section 6.2 that this functional is maximised by a source transmitting a stream of traffic composed of bursts at the peak rate followed by silences (Figure 1.2a). The bursts are just long enough to empty the token buffer and the silences are just long enough to allow it to fill again. This means that the mean rate of the source is the same as the leaky bucket rate. For the source to be stationary, the phase must be uniformly distributed. Heuristically, the loss rate is maximised when the traffic is bursty to the maximum extent permitted by the leaky bucket regulators.

When we multiplex sources the situation is somewhat different. We find that the loss rate can be increased by preceding or following the peak rate burst with an interval in which the source transmits at the leaky bucket rate (which again is the mean rate of the

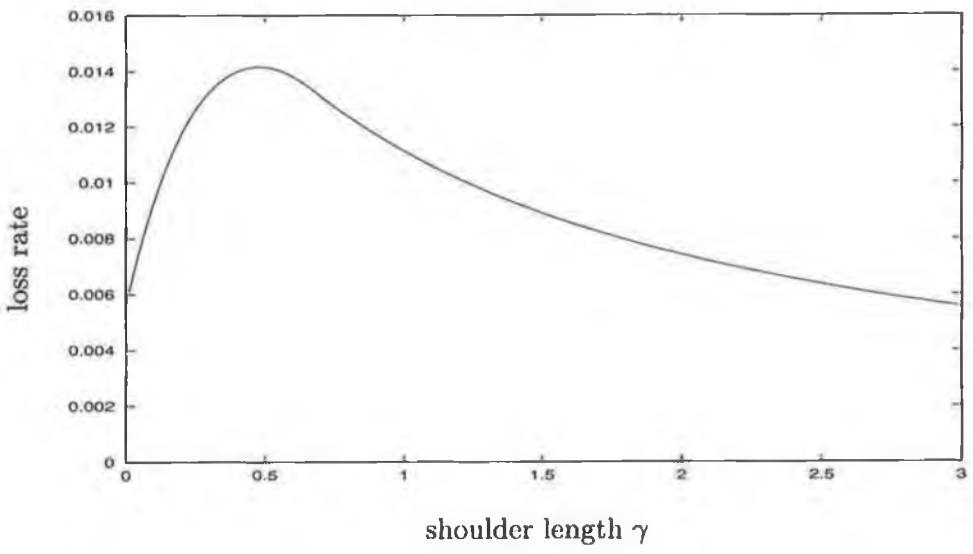

Figure 1.1: The loss rate vs. shoulder length for two independent sources.

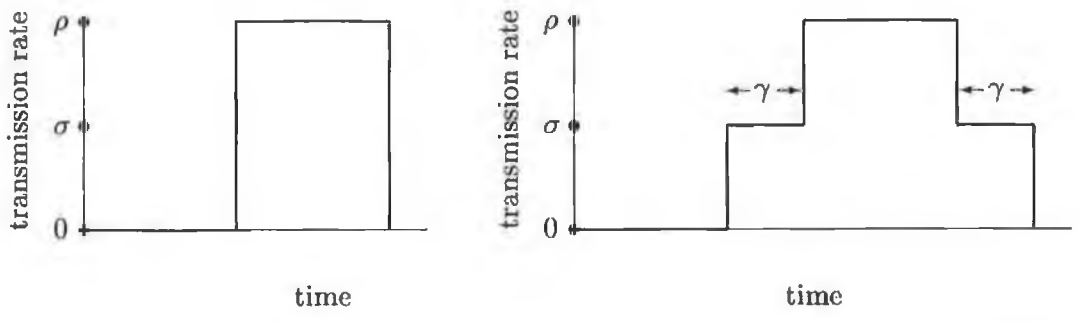

Figure 1.2: a) the on-off pattern, and b) the wedding cake pattern, conjectured to be the worst traffic under several performance measures.

source). For instance, Figure 1.1 shows the cell loss rate when two independent sources of the type depicted in Figure 1.2b are multiplexed through a buffer, as the shoulder length γ is varied. These results were obtained numerically. The maximum loss rate is clearly not attained at $\gamma = 0$.

Conjecture 1 *For each of the functionals we have discussed, under leaky bucket constraints (σ, β) and $(\rho, 0)$, the worst case traffic consists of a periodic pattern with random phase and its sample paths are composed of following four phases:*

- *a peak rate burst of length $\beta/(\rho - \sigma)$,*
- *an interval in which the source transmits at rate σ, the length of which depends on the particular functional under consideration,*
- *a silent interval of length β/σ,*
- *another interval at rate σ, the length of which again depends on the particular functional.*

Note that the time intervals in which the source is transmitting at the peak rate and is silent are, respectively, just long enough to fill and empty the leaky bucket. We have not managed to prove the truth of this conjecture, however we have made some progress towards this end by restricting the class of sources over which the optimisation must be performed. Firstly we show in Chapter 3 that the optimisation of the effective bandwidth may be restricted to an optimisation over the periodic processes, that is processes whose sample paths are periodic and have uniformly distributed phase. This has a number of interesting consequences, for example it shows that the worst case traffic is ergodic. It also allows us to consider the source behaviour over a compact interval which introduces considerable technical simplification. The effective bandwidth functional is then shown to be convex on the space of periodic processes. Using this fact and choosing an appropriate topology on the space, we show how the optimisation may be reduced to one over the extreme points of the set of processes that obey the constraints. We give a characterisation of these extreme points in Section 4.5. In Chapter 5 we apply the same techniques again, but using a different linear structure on the set of periodic processes. In this new linear structure the set of processes that obey the constraints has a strictly smaller set of extreme points

which we again characterise. In Chapter 6 we give examples of some simple functionals and their worst case sources. We investigate more closely the effective bandwidth functional and show that in this case the optimum leading and trailing shoulder lengths are equal, and that their common value is always less than one of the parameters taken by this functional (the timescale parameter). We calculate numerically the optimum shoulder lengths for several complicated functionals, including the effective bandwidth. Finally, we describe an application of our work to Connection Admission Control.

We do not deal with questions of existence or uniqueness of the optimising source. We merely conjecture that for each of the functionals considered in Section 1.2 the supremum over the set of sources that meet the constraints is attained by one of those sources. We shall see in Chapter 6 that the functionals that represent a bufferless server do not have a unique maximum. The defining characteristic of these functionals is that they depend only on the distribution of the rate of transmission of a source and not on the structure of the transmitted traffic. This means that any source which has the same distribution of transmission rate as a worst case source will also be worst case. Also there is more than one behaviour of a source that maximises its loss rate in a finite buffer. This will be discussed in Section 6.2. However, the remaining functionals under consideration appear to have a unique worst case.

Chapter 2

Finite Buffers

In this chapter we give a more precise definition of the space of traffic processes we will be working in, and of the performance functionals on this space.

2.1 The Space of Realisations

We work in a continuous time setting. The techniques required in continuous time are more sophisticated than those in discrete time but the results should be simpler because there are no discrete effects.

Let $\mathbb{D}^+$ be the space of functions $\mathbb{R}^+ \to \mathbb{R}^+$ that are are non-decreasing and right continuous. If $a \in \mathbb{D}^+$ then a is a possible realisation of the traffic from a source: $a(t)$ represents the amount of traffic that has arrived in the interval $[0, t]$. Since the elements of $\mathbb{D}^+$ are right continuous and non-decreasing, they have left limits. We shall denote the left limit of a function f at a point t by $f(t^-) := \lim_{x \uparrow t} f(x)$.

There is an interesting interpretation of the elements of $\mathbb{D}^+$ which comes from the one-to-one correspondence between these functions and the set of σ-finite measures on $\mathbb{R}^+$. For given any σ-finite measure μ on $\mathbb{R}^+$, define $a(t) := \mu[0, t]$. It can be shown that a is right continuous and non-decreasing. Conversely, for any $a \in \mathbb{D}^+$ there is a unique measure μ on the measurable space $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ such that $\mu[0, t] = a(t)$. We may interpret the measure μ as a set function telling us the amount of traffic transmitted by the source during any Borel set of instants.

It will be prove useful to embed $\mathbb{D}^+$ in a linear space. To do this we take its linear

completion $\mathbb{D} := \mathbb{D}^+ - \mathbb{D}^+$, that is

$$\mathbb{D} := \{a - b : a, b \in \mathbb{D}^+\}.$$

We find that the elements of this set are right continuous and have bounded variation. Indeed any function $\mathbb{R}^+ \to \mathbb{R}$ that has these properties is in $\mathbb{D}$. For any $w \in \mathbb{D}$, the Jordan decomposition [9] defines the least a and b in $\mathbb{D}^+$ such that $w = a - b$. We denote the positive variation of $w \in \mathbb{D}$ by w^+, and the negative variation by w^-. A σ-finite signed measure is a countably additive set function on the measurable sets such that the empty set is mapped to 0 and for which there exists a countable collection of sets of finite measure the union of which is the whole space. Again, for each σ-finite signed measure ν there is a corresponding element $w \in \mathbb{D}$ defined by $w(t) := \nu[0,t]$. Also, for each $w \in \mathbb{D}$ such that either w^+ or w^- is bounded, there is a σ-finite signed measure. To see this, take the Jordan decomposition of w, form the positive measure associated with each part and then subtract one from the other to get a σ-finite signed measure. This measure will obey $\nu[0,t] = w(t)$ and can be the only measure to do so since $\{[0,t]\}_{t\geq 0}$ is a π-system.

2.2 Continuous Time Queues

In discrete time, the queue length in an infinite buffer evolves from its initial value q_0 according to the recursion

$$q(t+1) = [q(t) + w(t)]^+, \qquad t \in \mathbb{N}.$$

Here, the workload $w(t)$ is the excess of arrivals over service capacity at the time instant $t \in \mathbb{N}$, and x^+ denotes $\max(0,x)$. We will also write $x \vee y := \max(x,y)$ and $w[t_1,t_2] := w(t_1) + \cdots + w(t_2)$. The recursion above may be easily solved by iteration:

$$\begin{aligned} q(t) &= [q(t-1) + w(t-1)] \vee 0 \\ &= [q(t-2) + w(t-2) + w(t-1)] \vee w(t-1) \vee 0 \\ &\vdots \\ &= (q_0 + w[0,t]) \vee \max_{0\leq t_1\leq t} w[t_1,t]. \end{aligned}$$

Defining the queue length in continuous time is more difficult however. A number of approaches have been taken in the literature, each appropriate to the characteristics of

the sample paths of the particular model under investigation. One common method, used for example in [10], is to define the queue length to be the continuous time analogue of the expression above:

$$q(t) := (q_0 + w[0,t]) \vee \sup_{0 \le t_1 \le t} w[t_1, t]. \tag{2.1}$$

However, the corresponding formula for the queue length in an infinite buffer is considerably more complex and it does not seem natural to define it in this way. Also, we are interested in the amount of traffic lost and for this there is no known formula.

Another early approach was the use of limiting methods. These were used by Moran [11] and later by Gani and Pyke [12]. Here the units of time and of traffic are scaled by $1/n$. For each n, the queue length is defined recursively as above. In the limit as $n \to \infty$, a continuous time fluid model is obtained. This approach is rather cumbersome and we find that more direct methods are preferable.

In [13], Reich defines the virtual waiting time in the M/G/1 queueing system. In this model batch arrivals with some general distribution arrive in a Poisson manner at the queue and are served at constant rate. The workload is thus given by $w(t) = a(t) - st$, where a is non-decreasing and right continuous. Reich shows that the equation

$$q(t) = q_0 + a(t) - st + s \int_0^t \mathrm{I}_{\{x \ge 0 : q(x) \le 0\}}\, dx,$$

where I is the indicator function, has a unique solution and it is this he defines to be the queue length. The integral term acts as a compensator to account for the service that goes unused. Reich shows that this definition leads to the expression for the queue length in (2.1). This approach works not only for the M/G/1 queue but also for any input that is singular with respect to the Lebesgue measure. The reason that Reich's approach only works for singular sample paths is that the compensator must be either on or off. For example, if the input is $a(t) = st/2$ then there is no solution to the above equation. Kingman [14] supplied a modification that allows this approach to work in the more general situation where the workload has locally bounded total variation and no downward jumps. His queue length is the unique solution to the equation

$$q(t) = q_0 + w(t) + \int_0^t \mathrm{I}_{\{x \ge 0 : q(x) \le 0\}}\, dw^-(x),$$

where w^- is the negative variation of w. This definition can also be shown to yield the expression for queue length in (2.1). It is difficult, however, to generalise this method

further to the case where the workload is not of locally bounded variation or there are atoms of service.

For continuous workload, Harrison [15] defines the queue length $q(t)$ and the unused service $u(t)$ to be the unique pair of functions that satisfy

- u is continuous and $u(0) = 0$,
- $q(t) := q_0 + w(t) + u(t) \geq 0$ for all $t \geq 0$,
- u increases only at those t for which $q(t) = 0$.

He also applies the same method to the definition of the queue length in a finite buffer of size b. In this case the loss $l(t)$ must also be considered. Harrison defines q, u, and l to be the unique solutions to

- u and l are continuous and $u(0) = l(0) = 0$,
- $q(t) := q_0 + w(t) + u(t) \in [0, b]$ for all $t \geq 0$,
- u increases only at those t for which $q(t) = 0$, and l increases only at those t for which $q(t) = b$.

Pacheco [16] defines the unused service in an infinite buffer as the least, non-decreasing, right-continuous function u such that $q_0 + w(t) + u(t) \geq 0$ for all $t \geq 0$. The queue length will then be $q(t) := q_0 + w(t) + u(t)$. He shows that the problem of finding this unused service function is equivalent to finding a non-decreasing right continuous function u such that $u(t) > u(t_1) \Rightarrow \exists t_2 \in [t_1, t] : (w + u)(t_2) = 0$, and that both of these problems have the solution

$$u(t) = -\inf_{0 \leq t_1 \leq t} (q_0 + w(t_1)) \vee 0.$$

The expression for the queue length resulting from this solution is the same as in (2.1).

2.3 Finite Buffers

In a finite buffer, the discrete time recursion analogous to the infinite buffer recursion above is

$$q(t+1) = \Big[\min\left[q(t) + w(t), b\right]\Big]^+, \qquad t \in \mathbb{N}.$$

The solution was found by Toomey [17] to be

$$q(t) = \max_{t_1 \le t} \min_{t_1 \le t_2 \le t} \Big(w[t_1, t], b + w[t_2, t] \Big) \bigvee \min_{t_1 \le t} \Big(q_0 + w[0, t], w[t_1, t] \Big).$$

We shall attempt to define the queue length in a finite buffer in continuous time along the lines of [16].

Our definition of the loss from a finite buffer is based on the following heuristics. Consider the traffic that is lost. If this traffic had been removed from the traffic stream before it entered the buffer then the buffer would not overflow. While a finite buffer does not overflow, it acts exactly like an infinite buffer. So if we subtract some function (corresponding to the loss) from the arrivals so that the resulting pattern does not cause the queue length in an infinite buffer to exceed the level b, then this function is a candidate for the loss function. If the server is efficient, it will only lose traffic when it must: it will lose as few cells as possible and it will drop them as late as possible.

We are thus led to consider the partial ordering of $\mathbb{D}^+$ given by

$$a_1 \le a_2 \qquad \text{iff} \qquad a_1(t) \le a_2(t) \text{ for all } t \in \mathbb{R}^+.$$

Note that $\mathbb{D}^+$ is a *conditionally complete* lattice under this partial order. In such a lattice each non-empty subset that is bounded below has an infimum (greatest lower bound), and each non-empty subset that is bounded above has a supremum (least upper bound). For example, the lattice of real numbers $\mathbb{R}$ is conditionally complete. If $S \subset \mathbb{D}^+$ is non-empty and bounded below then its infimum is given by the pointwise infimum

$$(\inf S)(t) = \inf\{p(t) : p \in S\}.$$

The infimum is finite because $\{p(t) : p \in S\}$ is non-empty and bounded below for each $t \in \mathbb{R}^+$. The pointwise supremum $\sup\{p(t) : p \in S\}$ is not necessarily in $\mathbb{D}^+$ since it may not be right continuous. However $\sup S$ must exist: it can be shown that any partially ordered set in which each bounded non-empty subset has an infimum also has the property that each bounded non-empty subset has a supremum. We find that the supremum is the right continuous regularisation of the pointwise supremum:

$$(\sup S)(t) = \lim_{t_1 \searrow t} \sup\{p(t_1) : p \in S\}.$$

Let $q_0 \geq 0$. For each $w \in \mathbb{D}$ define

$$\begin{aligned} & \mathbb{U}^{(q_0)}w \;:=\; \inf\{u \in \mathbb{D}^+ : w + u + q_0 \geq 0\} \\ \text{and} \qquad & \mathbb{Q}^{(q_0)}w \;:=\; w + \mathbb{U}w + q_0. \end{aligned}$$

This is the way Pacheco defined the unused service and the queue length in an infinite buffer. Where no ambiguities arise we will drop the superscript specifying the initial queue length.

For any $w \in \mathbb{D}$, define the set L_w to be

$$L_w := \{l \in \mathbb{D}^+ : l \leq w^+, \mathbb{Q}(w-l)(t) \leq b \text{ for all } t \geq 0\}$$

This is the set of realisations which, when subtracted from the incoming traffic, yield a traffic stream that does not cause the buffer to overflow. We intend to define the loss as the least element of this set under the partial ordering defined above. We must show that such a least element exists.

Theorem 1 *For any $w \in \mathbb{D}$, the set L_w is a complete lattice.*

Proof. Clearly L_w is non-empty since $\mathbb{Q}(w^+ - w^+)(t) = \mathbb{Q}0(t) = 0 \leq b$ and so $w^+ \in L_w$. Let $S \subset L_w$ be non-empty. Since $\mathbb{D}^+$ is a conditionally complete lattice and S is bounded below by 0, we have that $m(t) := \inf_{l \in S} l(t)$ is in $\mathbb{D}^+$. We will show that $m \in L_w$. For all $l \in S$ and $t_2 \geq 0$ we have that

$$\sup_{t_1 \leq t_2} [(w-l)(t_2) - (w-l)(t_1)] \vee [q_0 + (w-l)(t_2)] \leq b.$$

So $(w-l)(t_2) \leq b + (w-l)(t_1)$ and $q_0 + (w-l)(t_2) \leq b$, for all $l \in S$, $t_2 \geq 0$, and $t_1 \leq t_2$. Taking the supremum over $l \in S$, we find that $w(t_2) - \inf_{l \in S} l(t_2) \leq b + w(t_1) - \inf_{l \in S} l(t_1)$ and $q_0 + (w - \inf_{l \in S} l)(t_2) \leq b$. Thus $w(t_2) - m(t_2) \leq b + w(t_1) - m(t_1)$ and $q_0 + (w-m)(t_2) \leq b$, for all $t_2 \geq 0$ and $t_1 \leq t_2$. Equivalently,

$$\sup_{t_1 \leq t_2} [(w-m)(t_2) - (w-m)(t_1)] \vee [q_0 + (w-m)(t_2)] \leq b, \qquad \text{for all } t_2 \geq 0,$$

and we conclude that $m \in L_w$.

We have shown that every non-empty subset of L_w has an infimum in L_w. In addition L_w has a top element w^+. We conclude that L_w is a complete lattice. □

In particular, $\inf L_w$ must itself exist and be an element of L_w. We take this as our definition of the loss function, that is $l^* := \inf\{l \leq w^+ : \mathbb{Q}_{(w-l)} \leq b\}$. The queue length and the unused service are then given by

$$\begin{aligned} q &:= \mathbb{Q}(w - l^*), \\ u^* &:= \mathbb{U}(w - l^*) = q - w + l^*. \end{aligned}$$

2.4 Some Properties of the Queue Length

We will now show that the operators we have defined have some simple but useful properties. These will be found to agree with our intuition as to how a queue should behave.

The Memoryless Property

Pacheco [16] proves the memoryless property of the infinite buffer queue, that is if for some $t_1 \geq 0$ we let $\widehat{w}(t) := w(t + t_1) - w(t_1^-)$ and $\widehat{q_0} := q(t_1^-)$, then $\mathbb{Q}\widehat{w}(t) = \mathbb{Q}w(t + t_1)$ and $\mathbb{U}\widehat{w}(t) = \mathbb{U}w(t + t_1) - \mathbb{U}w(t_1)$, for all $t \geq 0$. Intuitively this means that the length of the queue and amount of service capacity that is unused at any time after a fixed time t_1 depends only on the queue length at time t_1 and the arrivals afterwards. We use this result to prove the corresponding property for finite buffer queues.

Theorem 2 *For any $t_1 \geq 0$, let $\widehat{w}(t) := w(t + t_1) - w(t_1^-)$ and $\widehat{q_0} := q(t_1^-)$. Then the memoryless property holds:*

$$\widehat{q}(t) = q(t + t_1), \qquad \widehat{l}(t) = l(t + t_1) - l(t^-), \qquad \widehat{u}(t) = u(t + t_1) - u(t^-).$$

Proof. Let $l' := l(t + t_1) - l(t^-)$. Using the memorylessness of $\mathbb{Q}$ we have that

$$\mathbb{Q}^{(q_0)}(w - l) = \mathbb{Q}^{(q(t_1^-))}(\widehat{w} - l').$$

Since $\mathbb{Q}^{(q_0)}(w - l) \leq b$, we have that $l' \in \widehat{L} := \{z \leq \widehat{w} : \mathbb{Q}^{(q(t_1^-))}(\widehat{w} - z) \leq b\}$. Thus $\widehat{l} := \inf \widehat{L} \leq l'$. Now let

$$\bar{l}(t) := \begin{cases} l(t), & t < t_1 \\ \widehat{l}(t - t_1) + l(t_1^-), & t \geq t_1 \end{cases}.$$

Then $\mathbb{Q}(w - \bar{l})(t) = \mathbb{Q}(w - l)(t) \leq b$ for all $t \leq t_1$. Using the infinite buffer memoryless property again, we find that $\mathbb{Q}(w - \bar{l}) = \mathbb{Q}^{(q(t_1^-))}(w - \widehat{l}) \leq b$ for $t \geq t_1$. Thus $\bar{l}$ is in L_w and so

$\widehat{l} \geq l$. It follows that $\widehat{l}(t-t_1)+l(t_1^-) \geq l(t)$ for $t \geq t_1$, or equivalently $\widehat{l}(t) \geq l(t+t_1)-l(t_1^-)$ for $t \geq 0$. □

Isotonicity

Intuitively, the more traffic that arrives at the buffer the longer the queue will be, and there will be less service unused and more traffic lost. However, the partial order we have been considering is too coarse to capture this effect: even if traffic arrives earlier, this does not necessarily mean that less will be admitted to the queue. We define a finer partial order $\preceq$:

$$a_1 \preceq a_2 \text{ iff } a_1(t_2)-a_1(t_1) \leq a_2(t_2)-a_2(t_1), \text{ for all } t_1, t_2 \in \mathbb{R}^+ \text{ such that } t_1 \leq t_2.$$

Under this definition, $a_1 \preceq a_2$ means that a_1 has less traffic in every time interval than a_2. This partial order is strictly finer than the previous one: if $a_1 \preceq a_2$ then $a_1 \leq a_2$, but the converse does not hold. Pacheco proves that if $a_1 \preceq a_2$ then $\mathbb{Q}a_1 \leq \mathbb{Q}a_2$ and $\mathbb{U}a_1 \succeq \mathbb{U}a_2$. We use these results to prove the following theorem.

Theorem 3 *If $x_1 \preceq x_2$ then $q_{x_1} \leq q_{x_2}$, $u_{x_1} \succeq u_{x_2}$, and $l_{x_1} \preceq l_{x_2}$.*

Proof. It will be straightforward to show that $q_{x_1} \leq q_{x_2}$ from the formula for the queue length given in Section 2.6. The following argument shows that $l_{x_1} \leq l_{x_2}$:

$$\begin{aligned} x_1 \preceq x_2 &\Rightarrow x_1 - l_{x_2} \preceq x_2 - l_{x_2} \\ &\Rightarrow \mathbb{Q}(x_1 - l_{x_2})(t) \leq \mathbb{Q}(x_2 - l_{x_2})(t) \leq b \qquad \text{for all } t \geq 0 \\ &\Rightarrow l_{x_2} \in L_{x_1} \\ &\Rightarrow l_{x_2} \geq \inf L_{x_1} = l_{x_1}. \end{aligned}$$

We extend this result to the partial order $\preceq$ using the memoryless property. First note that the infinite buffer queue length $\mathbb{Q}^{(q_0)}w$ at any time t is isotonic in the initial queue length q_0. This means that the loss in a finite buffer $l_w^{(q_0)}$ is also isotonic in q_0. Let $q_1 := q_{x_1}(t)$ and $q_2 := q_{x_2}(t)$. For some $t_1 \geq 0$, let $\widehat{w}_1(t) := x_1(t+t_1) - x_1(t_1^-)$ and $\widehat{w}_2(t) := x_2(t+t_1) - x_2(t_1^-)$. So,

$$l_{x_1}(t_1+t) - l_{x_1}(t_1^-) = \widehat{l}_{x_1}^{(q_1)}(t) \leq \widehat{l}_{x_1}^{(q_2)}(t) \leq \widehat{l}_{x_2}^{(q_2)}(t) = l_{x_2}(t_1+t) - l_{x_2}(t_1^-).$$

□

2.5 An Equivalent Definition of the Queue Length

We have defined the loss, unused service and queue length functions to be

$$l^* := \inf\{l \in \mathbb{D}^+ : \mathbb{Q}(w-l)(t) \leq b\}, \qquad q := \mathbb{Q}(w-l^*), \qquad u^* := q - w + l^*.$$

However, the barriers at 0 and at b are symmetrical and we could just as well have defined these functions to be

$$u^* := \inf\{u \in \mathbb{D}^+ : \widehat{\mathbb{Q}}(w+u)(t) \geq 0\}, \qquad q := \widehat{\mathbb{Q}}(w+u^*), \qquad l^* := w + u^* - q,$$

where $\widehat{\mathbb{Q}}w := w + \inf\{l \geq 0 : w(t) + l(t) \leq b \text{ for } t \geq 0\}$. We will show that these two definitions are equivalent and furthermore that both are equivalent to the following set of definitions:

$$(l^*, u^*) := \inf\{(l,u) : 0 \leq w - l + u \leq b\}, \qquad q := w - l^* + u^*.$$

The order on $\mathbb{D}^+ \times \mathbb{D}^+$ we are using here is the usual product order, that is

$$(l_{x_1}, u_1) \leq (l_{x_2}, u_2) \quad \text{iff} \quad l_{x_1} \leq l_{x_2} \text{ and } u_1 \leq u_2, \qquad \text{for all } l_{x_1}, l_{x_2}, u_1, u_2 \in \mathbb{D}^+.$$

Let l^* and u^* be defined according to the first of the three definitions above and let $S := \{(l,u) : 0 \leq w - l + u \leq b\}$. We proved above that $(l^*, u^*) \in S$ and so clearly $\inf S \leq (l^*, u^*)$. Let $(l,u) \in S$. Define $u_l := \mathbb{U}(w-l)$. Then $u_l \leq u$. Also $w - l + u \leq b \Rightarrow \mathbb{Q}(w-l) = w - l + u_l \leq b$. Therefore $l^* \leq l$. Since $w - l^* \geq w - l$ we have that $u^* \leq u_l$ by the previous section. Thus we conclude that $(l^*, u^*) \leq (l, u_l) \leq (l,u)$ and hence that $(l^*, u^*) \leq \inf S$. This proves that the first definition above is equivalent to the third. The equivalence of the second and third definitions may be proved in a similar manner.

2.6 An Expression for the Queue Length

In this section we obtain a pair of expressions for the queue length defined in the manner above. These are the continuous time version of the formulas given by Toomey in [17]. We do not know of any expression for the amount of traffic lost or the amount of service that goes unused. The fact that there are two expressions again reflects the symmetry of

the barriers at 0 and at b. For $w \in \mathbb{D}$, define $w[t_1, t_2] := w(t_2) - w(t_1^-)$. The expressions are

$$Q(t) := \inf_{0 \le t_1 \le t} \Big(\sup_{t_1 \le t_2 \le t} w[t_2, t] \vee (b + \inf_{t_1 \le z \le t} w[z, t]) \Big) \wedge \Big(\sup_{0 \le t_1 \le t} (q_0 + w[0, t]) \vee w[t_1, t] \Big),$$

$$\widehat{Q}(t) := \sup_{0 \le t_1 \le t} \Big(\sup_{t_1 \le t_2 \le t} w[t_2, t] \wedge (b + \inf_{t_1 \le z \le t} w[z, t]) \Big) \vee \Big(\inf_{0 \le t_1 \le t} (q_0 + w[0, t]) \wedge (b + w[t_1, t]) \Big).$$

We will need two lemmas to prove that these expressions give the queue length. The first states that these two expressions are indeed equivalent.

Lemma 1 *$Q(t) = \widehat{Q}(t)$ for all $t \ge 0$.*

Proof. Let $t > 0$. To simplify formulas we will write $s(t_1) := \sup_{0 \le t_1 \le t} w[t_1, t]$ and $i(t_1) := \inf_{t_1 \le t_2 \le t} w[t_2, t]$ for all $t_1 \in [0, t]$.

Suppose that $s(0) \le b + i(0)$. Then $s(t_1) \le b + i(t_1)$ for all $t_1 \in [0, t]$. This means that $Q(t) = (b + i(0)) \wedge [(q_0 + w[0, t]) \vee s(0)]$ and $\widehat{Q}(t) = s(0) \vee [(q_0 + w[0, t]) \wedge (b + i(0))]$. Since $s(0) \le b + i(0)$, these two expressions are equal.

Now suppose that $s(0) \ge b + i(0)$. Then $s(0) \vee (q_0 + w[0, t]) \ge s(0) = s(0) \vee (b + i(0)) \ge \inf_{0 \le t_1 \le t}[s(t_1) \vee (b + i(t_1))]$. Thus $Q(t) = \inf_{0 \le t_1 \le t}[s(t_1) \vee (b + i(t_1))]$. Similarly $\widehat{Q}(t) = \sup_{0 \le t_1 \le t}[s(t_1) \wedge (b + i(t_1))]$. Note that for every $0 \le t_1 \le t$ either $s(t_1)$ or $b + i(t_1)$ is continuous as a function of t_1, and that $s(t_1)$ is non-increasing and $i(t_1)$ is non-decreasing. Let $c := \sup\{t_1 : s(t_1) \ge i(t_1)\}$ be the point at which s and i cross. Then $\inf_{0 \le t_1 \le t}[s(t_1) \vee (b + i(t_1))] = s(c^-) \wedge (b + i(c))$ and $\sup_{0 \le t_1 \le t}[s(t_1) \wedge (b + i(t_1))] = s(c) \vee (b + i(c^-))$. Either s or i is continuous at c. If s is that is continuous, then $b + i(c) \ge s(c) = s(c^-)$ and $b + i(c^-) \le s(c^-) = s(c)$, and therefore $s(c^-) \wedge (b + i(c)) = s(c) \vee (b + i(c^-)) = s(c) = s(c^-)$. Similarly, if i is continuous at c, then $s(c^-) \wedge (b + i(c)) = s(c) \vee (b + i(c^-)) = i(c) = i(c^-)$. □

Lemma 2 *If for some $t_1 \ge 0$ we have that $\sup_{t_1 \le t_2 \le t} w[t_2, t] < \mathbb{Q}w(t)$, then $\mathbb{Q}w(t_2) + w[t_2, t] = \mathbb{Q}w(t)$ for all $t_2 \in [t_1, t]$.*

Proof. The memoryless property for the infinite buffer queue implies that

$$\mathbb{Q}w(t) = \sup_{t_1 \le t_2 \le t} w[t_2, t] \vee \Big(\mathbb{Q}w(t_1) + w[t_1, t] \Big).$$

So, if $\mathbb{Q}w(t) < \sup_{t_1 \le t_2 \le t} w[t_2, t]$, then we must have that $\mathbb{Q}w(t) = \mathbb{Q}w(t_1) + w[t_1, t]$. □

Theorem 4 *The queue length in a finite buffer is given by $Q(t)$.*

Proof. Suppose $Q(t) < \mathbb{Q}(w - l^*)(t)$ for some $t \geq 0$. Then, since $\mathbb{Q}(w - l^*)(t) \leq \mathbb{Q}w(t) = \sup_{0 \leq t_1 \leq t}(q_0 + w[0,t], w[t_1,t])$, we must have that

$$Q(t) = \inf_{0 \leq t_1 \leq t} \Big(\sup_{t_1 \leq t_2 \leq t} w[t_2,t] \vee (b + \inf_{t_1 \leq z \leq t} w[z,t]) \Big).$$

Let δ be such that $0 < \delta < \mathbb{Q}(w - l^*)(t) - Q(t)$. Then there exists $t_1 \in [0,t]$ such that

$$\sup_{t_1 \leq t_2 \leq t} w[t_2,t] \vee (b + \inf_{t_1 \leq z \leq t} w[z,t]) < Q(t) + \delta < \mathbb{Q}(w - l^*)(t).$$

So $\sup_{t_1 \leq t_2 \leq t} w[t_2,t] < \mathbb{Q}(w - l^*)(t)$, and therefore $\sup_{t_1 \leq t_2 \leq t}(w - l^*)[t_2,t] < \mathbb{Q}(w - l^*)(t)$. Applying the previous lemma, we find that

$$\mathbb{Q}(w - l^*)(t_2) + (w - l^*)[t_2,t] = \mathbb{Q}(w - l^*)(t), \qquad \text{for all } t_2 \in [t_1,t].$$

Now,

$$\begin{aligned} \mathbb{Q}(w - l^*)(t) > b + \inf_{t_1 \leq z \leq t} w[z,t] \quad &\Rightarrow \quad \mathbb{Q}(w - l^*)(t) > b + \inf_{t_1 \leq z \leq t} (w - l^*)[z,t] \\ &\Rightarrow \quad \sup_{t_1 \leq z \leq t} \Big(\mathbb{Q}(w - l^*)(t) - (w - l^*)[z,t] \Big) > b \\ &\Rightarrow \quad \sup_{t_1 \leq z \leq t} \mathbb{Q}(w - l^*)(z) > b, \end{aligned}$$

which is a contradiction since $\mathbb{Q}(w - l^*)(z) \leq b$ for all $z \geq 0$. We conclude that $Q \geq \mathbb{Q}(w - l^*)$. By a symmetrical argument it may be shown that $\widehat{Q} \leq \widehat{\mathbb{Q}}(w + u^*)$. Since we have proved in Section 2.5 that $\widehat{\mathbb{Q}}(w + u^*) = \mathbb{Q}(w - l^*)$ and above that $Q(t) = \widehat{Q}(t)$, the equivalence of Q and $\mathbb{Q}(w - l^*)$ is established. □

2.7 Large Deviations and the Effective Bandwidth

Asymptotics in the number of sources

Let $a(t)$ be the the amount of work that a stationary source z transmits in the interval $[0,t]$. We define the *effective bandwidth* of the source to be

$$\alpha(\theta,t) := \frac{1}{\theta t} \log \mathbb{E} e^{\theta a(t)}.$$

The effective bandwidth is exactly what we need to calculate the behaviour of the loss rate when the number of sources is large. The loss rate obeys the large deviations asymptotics

$$\lim_{N\to\infty} \log \mathbb{L}(sN, bN, N) = -NI(s,b),$$

where the function I is given by

$$I(s,b) := \inf_{t>0} (\theta t \alpha(\cdot, t))^*(b + st).$$

Here f^* denotes the *Legendre-Fenchel* transform of a function f, which is defined to be

$$f^*(y) := \sup_{x\in\mathbb{R}} [xy - f(x)].$$

The asymptotics above are also obeyed by the probability that the queue length exceeds a level Nb in an infinite buffer and by the proportion of time buffer is full.

Anick, Mitra, and Sondhi [18] introduced a Markov modulated fluid model and proved large deviation results as the number of sources becomes infinite. In [19] Courcoubetis and Weber consider the case of several independent sources in discrete time. They prove that the fraction of time a finite buffer stays full has the above asymptotics. Simonian and Guibert [20] prove the result for on-off Markov fluid sources. Weakening the independence assumption, Botvich and Duffield [10] give more general Gartner-Ellis type conditions under which the result holds. They work in both discrete and continuous time and prove the result for the stationary probability that the queue length in an infinite buffer is greater than the level Nb.

For an overview of the many applications of the effective bandwidth function see the paper of Kelly [21].

Minimisation of $I(s,b)$

The asymptotics above motivate the optimisation program

$$\text{minimise } I^{(\mathrm{z})}(s,b), \qquad \text{subject to } \mathrm{z} \in \mathcal{S},$$

where $\mathcal{S}$ represents the set of stationary traffic sources that obey some choice of constraints. This is related to another optimisation program

$$\text{maximise } \alpha^{(\mathrm{z})}(\theta, t), \qquad \text{subject to } \mathrm{z} \in \mathcal{S}. \tag{2.2}$$

Indeed if

$$\inf_{z\in\mathcal{S}} I^{(z)}(s,b) = \inf_{t>0}(t\theta \sup_{z\in\mathcal{S}} \alpha^{(z)}(\cdot,t))^*(b+st)$$

then these two optimisation problems are the essentially the same.

Let us review some convex analysis. For a reference see [22]. The convex hull of a (not necessarily convex) function g is defined to be

$$(\mathrm{conv}g)(x) := \inf\{\alpha_1 g(x_1) + \alpha_2 g(x_2) : \alpha_1+\alpha_2 = 1, \alpha_1 x_1 + \alpha_2 x_2 = x, \alpha_1, \alpha_2 \in [0,1]\}.$$

The convex hull of a family of convex functions is defined to be the convex hull of their pointwise infimum:

$$(\mathrm{conv}\{f_i\}_{i\in I}) := \mathrm{conv}g$$

where $g(x) := \inf_{i\in I} f_i(x)$. Under the pointwise partial order, the collection of convex functions on $\mathbb{R}$ form a complete lattice. The infimum of a family of convex functions is their convex hull $\inf\{f_i\}_{i\in I} = \mathrm{conv}\{f_i\}_{i\in I}$, and the supremum is the pointwise supremum $\sup\{f_i\}_{i\in I}(x) = \sup_{i\in I} f_i(x)$. The following lemmas will prove useful.

Lemma 3 *If $\{f_i\}_{i\in I}$ is a family of real-valued proper convex functions on $\mathbb{R}$, then the operations of taking the supremum and of taking the convex hull are conjugate, that is*

$$\begin{aligned} (\mathrm{conv}\{f_i\}_{i\in I})^* &= \sup\{f_i^*\}_{i\in I}, \\ \textit{and} \qquad (\sup\{f_i\}_{i\in I})^* &= \mathrm{conv}\{f_i^*\}_{i\in I}. \end{aligned}$$

Lemma 4 *Suppose $f(\theta) = \sup_i f_i(\theta)$ and the supremum is attained for each θ, say at i_θ. Then $f'_-(\theta) \le {f_{i_\theta}}'_-(\theta)$ and $f'_+(\theta) \ge {f_{i_\theta}}'_+(\theta)$.*

Lemma 5 *If f is convex and there exists some continuous function g such that $f'_- \le g \le f'_+$, then f is differentiable.*

Lemma 6 *Let $f(x,y)$ be a function of two real variables. If f is continuous in each variable and monotone in one then it is jointly continuous.*

We wish to show that $\inf_{z\in\mathcal{S}} I^{(z)}(s,b) = \inf_{t>0}(t\theta \sup_{z\in\mathcal{S}} \alpha(\cdot,t))^*(b+st)$. This will be accomplished using the following theorem, which makes some assumptions about the family of functions.

Theorem 5 *Let $\{f_i\}_{i\in I}$ be a family of real-valued proper convex functions on $\mathbb{R}$. Assume that for all x the supremum $f(x) := \sup f_i(x)$ is attained by one of the functions and that f is differentiable. Then $f^*(y) = \inf_{i\in I}[f_i^*(y)]$.*

Proof. Take any $y \in \mathbb{R}$. Clearly $f^*(y) = \text{conv}\{f_i^*\}_{i\in I}(y) \leq \inf_{i\in I}[f_i^*(y)]$. Define $z := \text{argsup}_{x\in\mathbb{R}}[xy - f(x)]$ and $j := \text{argsup}_{i\in I} f_i(z)$. From Lemma 4, f_j is differentiable at z with the same derivative as f. Also $f_j(z) = f(z)$. Thus $f_j^*(y) = f^*(y)$ and so $f^*(y) \geq \inf_{i\in I}[f_i^*(y)]$. This completes the proof. □

If we could show that Optimisation Problem (2.2) has a solution for each θ and t and that $\max_{z\in\mathcal{S}} \alpha^{(z)}(\theta, t)$ is differentiable in θ, then we could conclude that

$$\begin{aligned}
\inf_{z\in\mathcal{S}} I^{(z)}(s,b) &= \inf_{z\in\mathcal{S}} \inf_{t>0} (\theta t \alpha^{(z)}(\cdot,t))^*(b+st) \\
&= \inf_{t>0} \inf_{z\in\mathcal{S}} (\theta t \alpha^{(z)}(\cdot,t))^*(b+st) \\
&= \inf_{t>0} (\theta t \sup_{z\in\mathcal{S}} \alpha^{(z)}(\cdot,t))^*(b+st).
\end{aligned}$$

In Chapter 6 we will see that the effective bandwidth optimisation problem is one dimensional. In this case we can use the following theorem to show differentiability.

Theorem 6 *Let f_i be a family of convex functions parameterised by $i \in \mathbb{R}$ and let $f(\theta) = \sup_{i\in I} f_i(\theta)$ which we assume is attained for each θ. If f_i' exists and is continuous in i for fixed θ and $i_\theta := \text{argsup}_i f_i(\theta)$ is continuous as a function of θ, then f is differentiable.*

Proof. Since $f_i(\theta)$ is convex and differentiable in θ, we have that $f_i'(\theta)$ is nondecreasing and continuous in θ. But we have assumed that $f_i'(\theta)$ is continuous in i and so, by Lemma 6, $f_i'(\theta)$ is jointly continuous in i and θ. We conclude that $f_{i_\theta}'(\theta)$ is continuous in θ. But f is convex and $f_-'(\theta) \leq f_{i_\theta}'(\theta) \leq f_+'(\theta)$, and so f is differentiable by Lemma 5. □

To summarise, we need to show that $\alpha(\theta, t)$ attains its maximum for every θ and t, that $\alpha(\theta, t)$ is continuously differentiable with respect to θ for all t and z, and that the value of γ for which $\alpha(\theta, t)$ is maximised is continuous in θ for all t. We will consider these matters again in Chapter 6.

2.8 The Optimisation Problem

We are now ready to give a precise definition of the optimisation problem that we shall be tackling.

The space of realisations of a source will be $\mathbb{D}^+$, the set of right-continuous, non-negative, non-decreasing functions from $[0,\infty) \to \mathbb{R}^+$. The σ-algebra we use is the standard one employed when dealing with continuous time processes, that is the σ-algebra generated by the coordinate functions $\{X_t\}_{t\in[0,\infty)}$. This σ-algebra, which we denote by Σ, is the smallest in which the coordinate functions are measurable. It is useful to use the smallest since the smaller the σ-algebra, the more measures on it that exist.

It can be shown [23] that the difference of two random variables and the infimum of a countable set of random variables are also random variables. Since the coordinate functions are random variables, it follows that so to is the mapping

$$M_\sigma : \mathbb{D}^+ \to \mathbb{R}^+ : a \to \sup\{a(t_2) - a(t_1) - \sigma t : t_1, t_2 \in \mathbb{Q}^+, t_1 \leq t_2\}.$$

But the elements of $\mathbb{D}^+$ are right continuous, and so we have that $M_\sigma(a) = \sup\{a(t_2) - a(t_1) - \sigma t : t_1, t_2 \in [0,\infty), t_1 \leq t_2\}$, where this time the supremum is taken over all real pairs of times. This is the expression for the maximum queue length in an infinite buffer being served at rate σ when the initial queue length is zero. Thus the set of realisations $C := \{a \in \mathbb{D}^+ : M_\sigma(a) \leq \beta, M_\rho(a) = 0\}$ which meet the leaky bucket constraint (σ,β) and the peak rate constraint ρ is a measurable set of Σ since it is the intersection of a level set of M_σ and a level set of M_ρ.

Recall that a probability measure μ on the measurable space $(\mathbb{D}^+, \Sigma)$ is a σ-additive set function from $\Sigma \to [0,1]$ such that $\mu[\mathbb{D}^+] = 1$. We denote the set of such measures by $\mathcal{M}$. We define the time shift operator $\theta : \mathbb{D}^+ \to \mathbb{D}^+$ to be

$$(\theta_h a)(t) = a(t+h) - a(h^-),$$

where $a(h^-)$ is the left limit of a at h. Note that this is a measurable map, that is $\theta^{-1}S$ is Σ-measurable for every Σ-measurable set S. A probability measure μ on $(\mathbb{D}^+, \Sigma)$ is said to be stationary if $\mu[\theta^{-1}S] = \mu[S]$ for each measurable set S; we denote the set of such measures by $\mathcal{S}$. We are interested in the set of stationary probability measures whose realisations almost surely obey the leaky bucket constraints, that is the set $\mathcal{C} := \{\mu \in \mathcal{S} : \mu[C] = 1\}$.

We wish to define the average queue length and the long term loss rate. If the queueing process was stationary it would suffice to define the average queue length to be the expectation of the queue length at any instant and the average loss rate to be the expectation of the loss in the interval $[0,1]$. Unfortunately the queueing process is not stationary because we assume that buffer is initially empty. To start with the correct initial queue length we shall construct the unique two sided process of which the one sided process is a projection. Given any measure stationary probability measure on $(\mathbb{D}^+, \Sigma)$ we can construct a two-sided process as follows. For any finite sequence $a_1, \ldots, a_k$ of non-negative times, denote by $\mu_{a_1,\ldots,a_k}$ the distribution in $\mathbb{R}^k$ of the k-dimensional random vector $(X_{a_1}, \ldots, X_{a_k})$. We also use the more flexible notation $D_\mu(Y_1, \ldots, Y_k)$ to denote the distribution of the k-tuple of random variables $(Y_1, \ldots, Y_k)$. This latter notation includes the former since $\mu_{a_1,\ldots,a_k}$ is the same as $D_\mu(X_1, \ldots, X_k)$. We call the measures $\mu_{a_1,\ldots,a_k}$ the finite-dimensional distributions of μ.

Lemma 7 *A measure μ on $(\mathbb{D}^+, \Sigma)$ is stationary with respect to θ if and only if*

$$\mu_{a_1,\ldots,a_k} = D_\mu(X_{a_1+h} - X_h^-, \ldots, X_{a_k+h} - X_h^-)$$

for any finite sequence of non-negative times $\{a_i\}_{i=1,\ldots,k}$ and for any $h > 0$.

Proof. Suppose μ is stationary. Let $Z \in \mathcal{B}(\mathbb{R}^k)$. We denote by π the map $\pi : \mathbb{D}^+ \to \mathbb{R}^k$: $a \to (X_{a_1} \ldots, X_{a_k})$. Note that $X_t \circ \theta_h = X_{t+h} - X_h^-$. So

$$\begin{aligned} D_\mu(X_{a_1+h} - X_h^-, \ldots, X_{a_k+h} - X_h^-)Z &= D_\mu(X_{a_1} \circ \theta_h, \ldots, X_{a_k} \circ \theta_h)Z \\ &= \mu(\theta_h^{-1}\pi^{-1}Z) \\ &= \mu(\pi^{-1}Z) \\ &= \mu_{a_1,\ldots,a_k}Z. \end{aligned}$$

Conversely, suppose that the finite dimensional distributions are the same. Let S be a measurable set in Σ. Then S can be expressed in terms of a countable union of sets of the form $X_t^{-1}B$ where $t \in \mathbb{R}^+$ and B is a Borel set in $\mathbb{R}$. But for any $h \geq 0$,

$$\mu[\theta_h^{-1}X_t^{-1}B] = D_\mu(X_t \circ \theta_h)B = D_\mu(X_{t+h} - X_h^-)B = D_\mu(X_t)B = \mu[X_t^{-1}B]$$

and, since μ is σ-additive, we have that $\mu[\theta_h^{-1}S] = \mu[S]$. □

For any finite sequence of possibly negative times $\{a_i\}_{i=1,\ldots,k}$ we define the distributions $\lambda_{a_1,\ldots,a_k}$ to be

$$\lambda_{a_1,\ldots,a_k} := D_\mu(X_{a_1-b} - X_b^-, \ldots, X_{a_k-b} - X_b^-),$$

where $b := \min_{1\leq i\leq k} a_i$. These finite-dimensional distributions form a consistent family since

$$\mu_{a_1,\ldots,a_k}(S_1 \times \cdots \times S_k) = \mu_{a_{\pi(1)},\ldots,a_{\pi(k)}}(S_{\pi(1)} \times \cdots \times S_{\pi(k)})$$

for any permutation π of $(1,\ldots,k)$, and

$$\mu_{a_1,\ldots,a_{k-1}}(S_1 \times \cdots \times S_{k-1}) = \mu_{a_1,\ldots,a_k}(S_1 \times \cdots \times S_{k-1} \times \mathbb{R}).$$

By Kolmogorov's existence theorem there exists a probability measure λ on $\mathbb{R}^{\mathbb{R}}$ having the $\lambda_{a_1,\ldots,a_k}$ as its finite dimensional distributions. A similar result to Lemma 7 holds for the space $\mathbb{R}^{\mathbb{R}}$ and so λ will be stationary since $\lambda_{a_1,\ldots,a_k}$ obeys the $\mathbb{R}^{\mathbb{R}}$ version of the stationarity criterion. In addition the projection on the positive half line of λ will have the same marginals as the original process.

However, although the measure λ has the right marginals, it may not have the sample path behaviour we require. Fortunately, the stronger result holds that there is a *separable* process on the same space with the same marginals [24]. If D is a dense countable subset of $\mathbb{R}$, a function $a : \mathbb{R} \to \mathbb{R}$ is said to be separable with respect to D if for each $t \in \mathbb{R}$ there exists a sequence of times $t_1, t_2, \ldots$ in D that converges to t such that $a(t_1), a(t_2), \ldots$ converges to $a(t)$. Clearly, if $a \in \mathbb{R} \to \mathbb{R}$ is separable then $\sup\{a(t) : t \in G\} = \sup\{a(t) : t \in D \cap G\}$ for any open set $G \subset \mathbb{R}$. A process is called separable if for some countable dense subset D of $\mathbb{R}$, there is a measurable set of probability zero outside of which the realisations are separable with respect to D.

When constructing independent processes, a useful theorem is that of Fubini. Given two measurable spaces (Ω_1, Σ_1) and (Ω_2, Σ_2), we form the product space $(\Omega_1 \times \Omega_2, \Sigma_1 \times \Sigma_2)$, where $\Omega_1 \times \Omega_2$ is the usual Cartesian product and $\Sigma_1 \times \Sigma_2 := \sigma(\rho_1, \rho_2)$ is the σ-algebra generated by the coordinate functions ρ_1 and ρ_2. Suppose we have a probability measure μ_1 on (Ω_1, Σ_1) and a probability measure μ_2 on (Ω_2, Σ_2). Fubini's theorem states that there is a unique probability measure μ on $(\Omega_1 \times \Omega_2, \Sigma_1 \times \Sigma_2)$ such that $\mu[S_1 \times S_2] = \mu_1[S_1]\mu_2[S_2]$ for all $S_1 \in \Sigma_1$ and $S_2 \in \Sigma_2$.

We define the following functionals from $\mathcal{S} \to \mathbb{R}^+$, each of which in some way represents the performance of a queueing network.

- *The average queue length with a stationary stochastic service process.* Assume that X a stationary probability measure on the measurable space $(\mathbb{D}^+, \Sigma)$ is given. X will represent the service process. Then for each $\mu \in \mathcal{S}$ we form the product measure $X \times \mu$ on the space $(\mathbb{D}^+ \times \mathbb{D}^+, \Sigma \times \Sigma)$. The map $\mathbb{D}^+ \times \mathbb{D}^+ \to \mathbb{D} : (a, s) \to a - s$ is measurable and so induces a measure W on $\mathbb{D}$. We extend W to the two-sided separable process λ in the manner described above. The mean queue length of traffic process μ with independent service X is then defined to be $\sup_{t \leq 0}(w(0) - w(t))$. Then we define $\mathcal{Z}_X(\mu) := \mathbb{E}[\sup_{t \leq 0}(X_0 - X_t)]$, where X_t are the coordinate functions on $\mathbb{R}^{\mathbb{R}}$.

- *The average queue length with N identical independent sources.* For any stationary measure μ on $(\mathbb{D}^+, \Sigma)$ we form the N-fold product measure μ^N. The map $(\mathbb{D}^+)^N \to \mathbb{D} : (a_1, \ldots, a_N) \to a_1 + \cdots + a_N - sN$ is measurable and induces a measure W on $\mathbb{D}$. Again we extend this measure to a two-sided separable process λ on $\mathbb{R}^{\mathbb{R}}$. We define $\mathcal{Q}_N^{\text{ave}}(\mu) := \mathbb{E}[\sup_{t \leq 0}(X_0 - X_t)]$ as before.

- *The average loss rate with a stationary stochastic service process.* Just as for the infinite buffer queue, we take the product $X \times \mu$, form the measure induced by the map $(a, s) \to a - s$ and extend it to a separable process λ on $\mathbb{R}^{\mathbb{R}}$. But now we define the random variable

$$q_0(\omega) := \inf_{t \leq 0} \left(\sup_{z \leq 0}(\omega(0) - \omega(z)) \vee (b + \inf_{t \leq z \leq 0}(\omega(0) - \omega(z))) \right),$$

  which may be interpreted as the queue length at time zero in a buffer of size b. The mean loss rate is defined to be $\mathcal{Y}_X(\mu) := \mathbb{E}_\lambda[\mathbb{L}^{(q_0)}\omega(1)]$, the expected loss in the interval $[0, 1]$.

- *The average loss rate with N identical independent sources.* This functional is defined using a combination of techniques already employed. The induced measure ${}_w(\mu^N)$ is extended to a separable process λ on $\mathbb{R}^{\mathbb{R}}$. The random variable q_0 is defined as before and the mean loss rate is defined to be $\mathcal{L}_N(\mu) := \mathbb{E}_\lambda[\mathbb{L}^{(q_0)}\omega(1)]$

- *The effective bandwidth.* This is the easiest of the functionals to define and requires none of the apparatus developed in this section. We merely define $\mathcal{E}_{\vartheta,T}(\mu) := \mathbb{E}_\mu e^{\vartheta X_T}$. The effective bandwidth of the source μ is then $(\vartheta T)^{-1} \log \mathcal{E}_{\vartheta,T}(\mu)$.

Chapter 3

Periodicity and Markov Decision Procedures

In this chapter we show how to reduce the optimisation over all stationary processes to an optimisation over periodic processes. As well as being interesting in itself, this also greatly simplifies the theory in the next two chapters. We shall use the ergodic theorem to reformulate our problem as a Markov Decision Procedure; Sections 3.1 and 3.3 review the necessary background material. To illustrate the essential ideas in a technically simple setting we will initially work in discrete time and assume that the source emits traffic in discrete cells. It will also be necessary to place restrictions on the leaky bucket constraints. Later we will show how to extend these ideas, first to general source behaviour and then to the continuous time framework we are interested in.

Our strategy is to first use the Ergodic theorem to show that the problem of optimising the effective bandwidth functional is equivalent to that of optimising another invariant functional. Then we bound the value of this functional on the constraint set using ideas from Dynamic Programming. We show that this bound can be approached arbitrarily closely by an (ergodic) stationary periodic source.

3.1 The Pointwise Ergodic Theorem

As stated our optimisation problem involves maximising the expectation of a given functional over the space of stationary processes. Using ergodic theory we will show that there

exists an invariant functional with has the same expectation as the original functional for every stationary process. Thus the problem of optimising this functional is equivalent to the original problem.

We will work for the moment in discrete time. Let $(\Omega, \mathcal{M})$ be a measurable space on which there is a time shift operator $\theta : \Omega \to \Omega$. We write $\theta^{-1}M$ to mean $\{\omega \in \Omega : \theta\omega \in M\}$. We must insist that θ is measurable, that is $\theta^{-1}M \in \mathcal{M}$ for all $M \in \mathcal{M}$. Recall that a measure ν is said to be stationary if $\nu[\theta^{-1}M] = \nu[M]$ for all $M \in \mathcal{M}$. The ergodic theorem concerns integrable and invariant functions.

Definition 1 *A measurable function $f : \Omega \to \mathbb{R}$ is said to be integrable (also known as L_1) if $\mathbb{E}|f|$ is finite. It is said to be invariant if $f(\theta\omega) = f(\omega)$ for all $\omega \in \Omega$.*

Theorem 7 (Birkhoff 1931) *If f is integrable and ν is a stationary probability measure, then $n^{-1}\sum_{i=0}^{n-1} f(\theta^i\omega)$ converges ν-almost surely as $n \to \infty$ to an integrable and invariant function $\bar{f}(\omega)$, and $\mathbb{E}_\nu \bar{f} = \mathbb{E}_\nu f$.*

For a reference see [25].

Ergodicity of the Worst Case Traffic

An invariant subset of a measurable space is a measurable set M such that $\theta^{-1}M = M$. We say that a measure is ergodic if it is stationary and all invariant sets have measure either 0 or 1. The space of finite signed measures on $(\Omega, \mathcal{M})$ is a linear space under the operations of set-wise addition and scalar multiplication. In this space the set of probability measures is a convex set and it turns out that its extreme points are exactly those probability measures that are ergodic. All the functionals we have been considering are linear in this space since they are all expectations of some quantity. If we were to choose an appropriate topology on the space of finite signed measures and show that any stationary measure can be expressed as a convex combination of ergodic measures, then we could exploit this linearity to show that the worst case traffic is ergodic. The theorem of Choquet-Bishop-de Leeuw [26] may be useful in this regard. We shall not pursue this matter here however—by showing the worst case traffic is periodic we will also be able to prove in passing that it is ergodic.

3.2 The Discrete Time Version of the Optimisation Problem

Here time is indexed by the natural numbers $\mathbb{N}$. We make the further simplifying assumption that the output from the source is discrete, in other words that at each time instant a whole number of cells are emitted. The emissions will thus be in $A := \{0, \ldots, \rho\}$, where $\rho \in \mathbb{N}$ is the peak rate, and the space of outcomes is therefore $\Omega := A^{\mathbb{N}}$. The set A has a natural σ-algebra, its power set (set of all subsets) $\Sigma := \mathcal{P}\{0, \ldots, \rho\}$. We choose as our σ-algebra for Ω the infinite product $\prod_{n=0}^{\infty} \Sigma_n$ of copies of Σ, which is the smallest σ-algebra such that the coordinate maps are measurable. Our shift operator is $(\theta\omega)_n := \omega_{n+1}$. For all $\omega \in \Omega$, define the random variable

$$f(\omega) := e^{\vartheta \sum_{i=1}^{T} \omega_i}.$$

where $T \in \mathbb{N}$ and $\vartheta \in \mathbb{R}$ are fixed. The expected value of $(T\vartheta)^{-1} \log f$ is the effective bandwidth of the source. Define the invariant function

$$\overline{f}(\omega) := \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} f(\theta^i \omega)$$

on the set of outcomes $\omega \in \Omega$ where the limit exists. By the pointwise ergodic theorem, if ν is a probability measure which is stationary with respect to θ, then $\overline{f}$ is defined ν-almost everywhere and $\mathbb{E}_\nu \overline{f} = \mathbb{E}_\nu f$. The set of outcomes that obey the leaky bucket constraint (σ, β) is

$$C := \left\{ \omega \in \Omega : \sup_{n,m \in \mathbb{N}} \Big(\sum_{i=m+1}^{n} \omega_i - (n-m)\sigma \Big) \leq \beta \right\}.$$

Let $\mathcal{C}$ be the set of stationary probability measures on Ω that are supported by C. Assume that the service rate of the leaky bucket is rational. Then the leaky bucket queue length may only take values in a finite set Q. The discrete version of the optimisation problem is

$$\text{maximise } \mathbb{E}_\nu f, \qquad \text{subject to } \nu \in \mathcal{C}.$$

3.3 Markov Decision Procedures

A finite deterministic Markov Decision Procedure (MDP) consists of the following data:

- a finite set of states S,

- a reward function $r : S \times S \to \mathbb{R} \cup \{-\infty\}$ on the set of transitions.

It is customary to call a sequence of states a path. For any $n \geq 1$, define the n step average reward function on paths of length $n+1$ to be

$$R_n(x) := \frac{1}{n} \sum_{i=1}^{n} r(x_i, x_{i+1})$$

Also define the long term average reward $R(x) := \lim_{n\to\infty} R_n(x)$ if the limit exists. We denote the set of paths for which the limit does exist by X.

A cycle of length n is a finite path $x = x_1, \ldots, x_n, x_1$ where the last state is identical to the first. The maximal cycle mean of a Markov Decision Procedure is the maximum reward per cycle length $R_n(x)$ over all cycles. Since there is a finite number of states, this maximum is attained. A *cyclic* path x is a path such that $\theta^p x = x$ for some integer $p > 1$. If p is the smallest such number then it is called the period of the path. Obviously a cyclic path with period p is composed of a cycle of length p which is endlessly repeated. The following lemma shows that the supremum of the long term average reward over all paths where it is defined is attained by a cyclic path.

Lemma 8 *There is a cyclic path $x^* := x_1, x_2, \ldots, x_p, x_1, \ldots$ such that $R(x) \leq R(x^*)$ for all paths x where $R(x)$ exists.*

Proof. Denote the maximal cycle mean by c. By concatenating infinitely many copies of the cycle with maximal reward per unit length, we obtain a cyclic path x^* such that $R(x^*) = c$. Let x be any path such that $R(x)$ exists. Since there is a finite number of states, there exists a state s such that x returns to s infinitely often. Let t_0 be the first time that x is in state s. Consider the path $y_n := x_{n+t_0-1}$ for $n \in \mathbb{N}$. Then $R(y)$ exists and is equal to $R(x)$. Let t_n be the nth time that y is in state s. Since $R_n(y) \to R(y)$, any subsequence must also converge to the same limit, in particular the subsequence $R_{t_n}(y)$. But for each t_n, the finite path $\{y_i\}_{1 \leq i \leq t_n}$ is a cycle and so $R_{t_n}(y) \leq c$. We conclude that $R(x) = R(y) \leq c$. □

See [27] for more discussion about finite state MDPs and their relation to $(\max, +)$ algebra. The idea of representing the optimisation of the effective bandwidth as an MDP can be found in [28].

3.4 Periodicity of the Worst Case Traffic

We will now apply the theory in the previous section to cast our optimisation problem in the form of a Markov decision procedure.

To calculate the block sum in the formula for the effective bandwidth it is necessary to know the number of cells emitted in T consecutive time slots. Therefore a state of our MDP must contain this information. Also, to ensure that only paths which meet the leaky bucket constraint have reward greater than $-\infty$, the content of the leaky bucket must be included in the state. We may include either the content of the leaky bucket at the start of the block of T time slots, or the content at the end—one may be calculated from the other. We choose the former.

Our states are thus of the form $(q, a_1, \ldots, a_{T-1}, a_T)$ where $q \in Q$ and $a_i \in A$ for $1 \leq i \leq T$. However not everything of this form can be a state—a state for which $q + a_1 + \cdots + a_n - n\sigma > \beta$ for some $n \leq T$ would be unreachable, creating technical difficulties. We therefore exclude these states from consideration. Our Markov Decision Procedure is then

- *State Space*: $S := \{(q_0; a_1, \ldots, a_{T-1}, a_T) : q_0 \in Q, a_n \in A, q_n \leq \beta \text{ for all } 1 \leq n \leq T\}$, where q_n in the condition is defined recursively by $q_n := [q_{n-1} + a_n - \sigma]^+$ for $1 \leq n \leq T$.

- *Reward Function*: The transitions

$$(q, a_1, \ldots, a_{T-1}, a_T) \to ([q + a_1 - \sigma]^+, a_2, \ldots, a_T, a_{T+1})$$

  have reward $\exp(\vartheta \sum_{i=1}^{T} a_i)$ provided that both these elements of $Q \times A^T$ are in S. To all other transitions we give weight $-\infty$; they are thus forbidden.

Define the mapping $\Xi_n : Q \times A^{n+T} \to S^{n+1}$ by

$$\Xi_n(q_0; a_1, \ldots, a_{n+T}) := \Big((q_0; a_1, \ldots, a_T), (q_1, a_2, \ldots, a_{T+1}), \ldots, (q_n, a_{n+1}, \ldots, a_{n+T})\Big),$$

where q_m is defined as above. If $q_m \leq \beta$ for $1 \leq m \leq n + T$, in other words if the (finite length) realisation $a = a_1, \ldots, a_{n+T}$ obeys the leaky bucket constraint when the initial queue length is q_0, then the finite path $\Xi_n(a)$ will have finite average reward. Also, for any path x of length $n + 1$ with finite reward starting in some state $(q_0, a_1, \ldots, a_n)$, there is a

sequence $a_1, \ldots, a_{n+T}$ such that $\Xi_n(q_0; a_1, \ldots, a_{n+T}) = x$. Thus this mapping is a one-to-one correspondence between the paths through S that have no forbidden transitions and the realisations of the stochastic process that satisfy the constraints. Furthermore, the average reward per time unit of a path equals the value of $\overline{f}$ of the corresponding outcome if either exist. Thus we have that $\sup_{\omega \in C} \overline{f}(\omega) = \sup_{x \in X} R(x)$. It is this correspondence that enables us to apply results about MDPs to our optimisation problem.

Definition 2 *A measure μ on Ω is said to be periodic with period p if $\theta^p \omega = \omega$ for almost all $\omega \in \Omega$ and p is the smallest positive integer for which this holds.*

Note that given any cyclic path x there is a natural way to construct a periodic measure: let $(q, \omega) := \Xi_\infty^{-1}(x)$ and choose $\omega, \theta\omega, \theta^2\omega, \ldots$, or $\theta^{p-1}\omega$ with probability $1/p$. More formally, we take a probability measure μ on Ω given by

$$\mu := \frac{1}{p} \sum_{n=0}^{p-1} \delta_{\theta^n \omega},$$

where δ_ω represents the probability measure concentrated on $\omega \in \Omega$. It is obvious that any probability measure constructed from cyclic path in this way is stationary and ergodic and that $\mathbb{E}_\mu \overline{f} = R(x)$. We use $\mathcal{P}$ to denote the set of ergodic periodic measures on Ω.

The following theorem is the main result of this section.

Theorem 8 *The supremum $\sup_{\nu \in C} \mathbb{E}_\nu f$ is attained by a measure contained in $\mathcal{P}$.*

Proof. The correspondence between realisations in C and paths with no forbidden transitions means that $\sup_{\omega \in C} \overline{f}(\omega) = \sup_{x \in X} R(x)$. We know from Lemma 8 that there is a cyclic path x^* with some cycle length p such that $R(x) \leq R(x^*)$ for all other paths x. Using this path to construct a periodic measure μ from x^* in the manner described above, we find that $\sup_{\omega \in C} \overline{f}(\omega) = R(x^*) = \mathbb{E}_\mu \overline{f}$.

But now let $\nu \in \mathcal{C}$. Since ν is stationary, Birkhoff's theorem applies and thus $\overline{f}$ is defined ν-almost surely and is integrable. Also, $\mathbb{E}_\nu \overline{f} = \mathbb{E}_\nu f$. Since we have assumed that $\nu[C] = 1$, we have that $\mathbb{E}_\nu \overline{f} \leq \sup_{\omega \in C} \overline{f}(\omega)$. Therefore $\mathbb{E}_\nu f \leq \mathbb{E}_\mu \overline{f} = \mathbb{E}_\mu f$ and the conclusion follows since $\mu \in \mathcal{P}$. □

We have thus restricted our search to the set of processes that are ergodic and periodic.

A More Sophisticated Approach

Consider the finite horizon optimisation problem. Let $f_n : S \to \mathbb{R}^+$ be the function that assigns to each state the maximum total reward achievable starting from that state over a horizon of length n:

$$f_n(u_0) := \max_{u_1,\ldots,u_n \in S} \sum_{i=1}^{n} r(u_{i-1}, u_i), \qquad n \geq 1.$$

The optimum reward over the horizon n is then $\max_{u \in S} f_n(u)$. Clearly f_n obeys the recursion

$$f_{n+1}(u) = \max_{v \in S} \Big[r(u,v) + f_n(v) \Big], \qquad n \geq 1.$$

This is Bellman's famous Optimality Principle. It has been pointed out that this equation looks like a linear recursion in an exotic algebra on $\mathbb{R}$. Write $a \oplus b := \max(a,b)$ and $a \otimes b := a + b$ for any $a, b \in \mathbb{R} \cup \{-\infty\}$. Under these two operations $\mathbb{R} \cup \{-\infty\}$ is an idempotent semiring. This means that both operations are associative and have identity elements, that $a \oplus a = a$ for all $a \in \mathbb{R} \cup \{-\infty\}$, and that $\otimes$ distributes over $\oplus$. The identity element of $\oplus$ is $-\infty$ and of $\otimes$ is 0.

Multiplication of a vector by a matrix or of two matrices is defined in the usual way, but using $(\max, +)$ arithmetic. Thus

$$(MN)_{uv} := \bigoplus_{w \in S} (M_{uw} \otimes N_{wv}) \qquad \text{and} \qquad (Mv)_u := \bigoplus_{w \in S} (M_{uw} \otimes v_w).$$

Defined in this way, matrix multiplication is a semi-group.

The recursion for f_n above can now be written

$$f_{n+1} = R f_n, \qquad n \geq 1,$$

where R is a $|S| \times |S|$ matrix with entries $R_{uv} = r(u,v)$.

In the usual algebra on $\mathbb{R}$, the asymptotic behaviour of powers of a matrix is governed by its largest eigenvalue. In the $(\max, +)$ algebra, the asymptotic behaviour is governed by the $(\max, +)$ eigenvalue λ, which is the solution of the equation

$$Rf = \lambda \otimes f.$$

It is this eigenvalue that is the maximum long term average reward that we are trying to find. There are analogues of the Perron-Frobenius theorem that guarantee the existence

of a $(\max, +)$ eigenvalue for irreducible matrices; however, we shall proceed immediately to the infinite dimensional case.

3.5 Infinite State MDPs

Our first embellishment on the theory above is to drop the requirements of discrete arrivals and rational leaky bucket service rate. We will now allow ρ, σ and β to assume any positive values such that $\sigma < \rho$, and at each integral time instant we will allow the source to emit any real valued amount of traffic in the range $A := [0, \rho]$. Now $Q := [0, \beta]$.

The definition of the MDP is similar to the finite state space case.

- *State Space:* $S = \{(q, a_1, \ldots, a_{T-1}, a_T) \in Q \times A^T : q_n \leq \beta \text{ for all } 1 \leq n \leq T\}$, where q_n is defined as before.
- *Reward Function:* The transitions

$$(q, a_1, \ldots, a_{T-1}, a_T) \to ([q + a_1 - \sigma]^+, a_2, \ldots, a_T, a_{T+1})$$

  have reward $\exp(\vartheta \sum_{i=1}^{T} a_i)$ provided both these elements of $Q \times A^T$ are in S. Again, all other transitions have reward $-\infty$. We denote this reward function by $r(\cdot, \cdot)$.

The state space S is no longer finite—instead we will choose a topology on S which makes it compact and the reward function continuous. The natural choice is the restriction of the product topology on $Q \times A^T$. Since Q and A are compact, $Q \times A^T$ is compact by Tychonoff's theorem. This topology is metrisable since there are only a finite number of factors. A compatible metric is $d(u, v) := \sum_{i=1}^{T} |a_v^i - a_u^i| + |q_v - q_u|$, where $u := (q_u, a_u)$ and $v := (q_v, a_v)$ are any two states. We will investigate questions of continuity later.

Bellman Operators

We use the usual norm $||f|| := \max_{s \in S} |f(s)|$ on the space $C(S)$ of continuous real valued functions on S. A reward function b may be regarded as the kernel of a nonlinear operator on $C(S)$:

$$Bf(u) := \max_{v \in S} \{b(u, v) + f(v)\}. \tag{3.1}$$

It is easy to verify that this operator is linear in the $(\max, +)$ sense:

$$B(f \vee g) = Bf \vee Bg, \qquad B(\lambda + f) = \lambda + B(f), \qquad \forall f, g \in C(S), \lambda \in \mathbb{R}.$$

Indeed it may be shown [29] that if an operator is $(\max, +)$ linear and obeys

$$B(\sup_\alpha f_\alpha) = \sup_\alpha Bf_\alpha,$$

for any uniformly bounded family of functions $\{f_\alpha\}$, then there exists a kernel $b(u, v)$ such that B may be represented in the form (3.1). Operators that can be represented in this form are called Bellman operators. We will now discuss some of their properties. References for this section include [29] and [30].

Lemma 9 *Every Bellman operator is isotone.*

Proof. This follows from the linearity of Bellman operators. Let $f, g \in C(S)$. If $f \geq g$ then $f = f \vee g$, and so $Bf = Bf \vee Bg$ for any Bellman operator B. Thus $Bf \geq Bg$. □

A corollary to this lemma is that Bellman operators are non-expansive. If $||f - g|| \leq c$ for two functions $f, g \in C(S)$ and $c > 0$, then $f \leq g + c$ and $g \leq f + c$. Therefore, $Bf \leq B(g + c) = Bg + c$ and *vice versa*. Thus $||Bf - Bg|| \leq c$. This is a special case of a theorem proved in [31] that states that an homogeneous operator is isotone if and only if it is non-expansive. The non-expansiveness of Bellman operators means that they are continuous on $C(S)$.

An important class of Bellman operators are those that are compact.

Definition 3 *A Bellman operator on a compact metric space is said to be compact if its kernel is jointly continuous in its two entries and nowhere takes the value $-\infty$.*

This condition may be considered a replacement of the finiteness condition of the previous section. The justification for calling these operators compact is that they are compact in the usual sense of the term: they take bounded sets of $C(S)$ to precompact ones. The proof uses the following lemma.

Lemma 10 *The image of the set of continuous functions under a compact Bellman operator is uniformly equicontinuous.*

Proof. Since the kernel is continuous and $S \times S$ is compact, the kernel is uniformly continuous in its first argument equicontinuously in its second. Thus for any $\epsilon > 0$, there exists $\delta > 0$ such that $|b(u_1, v) - b(u_2, v)| < \epsilon$, for all $u_1, u_2, v \in S$ such that $d(u_1, u_2) < \delta$. Therefore if $d(u_1, u_2) < \delta$, then

$$Bf(u_1) = \max_{v \in S}\{b(u_1, v) + f(v)\} \leq \max_{v \in S}\{b(u_2, v) + \epsilon + f(v)\} = Bf(u_2) + \epsilon.$$

Similarly, $Bf(u_2) \leq Bf(u_1) + \epsilon$ and so $|Bf(u_1) - Bf(u_2)| \leq \epsilon$ for all $f \in C(S)$ whenever $d(u_1, u_2) < \delta$. □

Theorem 9 *A compact Bellman operator is continuous on the Banach space $C(S)$ and takes bounded sets of continuous functions to precompact ones.*

Proof. The continuity has been demonstrated above. The image of any bounded set $U \in C(S)$ of continuous functions is uniformly bounded by

$$\max_{u,v \in S} |b(u, v)| + \sup_{f \in U} ||f||.$$

Furthermore its image $B(U)$ is equicontinuous by the previous lemma. The precompactness of $B(U)$ follows from an application of the Ascoli-Arzela theorem. □

As we have seen in the discussion of the finite state MDP, the maximum long term reward is given by the $(\max, +)$-eigenvalue of the Bellman operator that has the reward function as its kernel. The following theorem, which we state without proof, is crucial when dealing with operators of this sort.

Theorem 10 *Every compact Bellman operator possesses a unique eigenvalue.*

We have said that the eigenvalue of an operator governs the long term behaviour of its iterates. This is expressed formally in the next theorem.

Theorem 11 *If B is a compact Bellman operator with eigenvalue λ, and $f \in C(S)$, then the iterates $B^n f$ grow arithmetically:*

$$||B^n f - f - n\lambda|| = O(1) \qquad \text{as } n \to \infty.$$

We define the *trace* of an operator to be

$$\mathrm{tr}B := \sup_{s \in S} b(s,s).$$

This may be interpreted as the maximum reward over a cycle of length one. Similarly $\mathrm{tr}B^n$ is the maximum reward over a cycle of length n. The corollary to the following theorem shows that the maximum cycle mean of an MDP is nothing other than the eigenvalue of its reward operator.

Theorem 12 *If B is a compact Bellman operator with eigenvalue 0, then*

$$\mathrm{tr}\left(\bigoplus_{n=1}^{\infty} B^n\right) = 0.$$

Corollary 1 *The eigenvalue λ of a compact Bellman operator B is equal to the maximal cycle mean of the associated MDP:*

$$\lambda = \sup_{n \geq 1} n^{-1} \mathrm{tr}B^n.$$

Quasi-compact Operators

The operator R associated with our infinite state MDP is not compact since some transitions are forbidden. Fortunately, we are interested in the reward over the long term and it suffices that R be *quasicompact.*

Definition 4 *We call an operator quasicompact if some power of it is compact.*

We shall prove that R is quasicompact presently.

Let $u = (q_u, a_1, \ldots, a_T)$ and $v = (q_v, b_1, \ldots, b_T)$ be two states in S such that $u \to v$ is a forbidden transition, in other words $r(u,v) = -\infty$. Then either $q_v \neq [q_u + a_1 - \sigma]^+$ or $b_{k-1} \neq a_k$ for some k in the range $2 \leq k \leq T$. It is clear therefore that the set of forbidden transitions is open. But, on the set of allowed transitions, r is continuous and so we conclude that that r is upper semicontinuous. A useful lemma about operators having kernels with this property is the following.

Lemma 11 *On a compact state space, the product of two Bellman operators, each of whose kernels are upper semicontinuous, also has an upper semicontinuous kernel.*

Proof. Let B_1 and B_2 be Bellman operators with upper semicontinuous kernels b_1 and b_2 respectively. The kernel of B_1B_2 is $b_1 \otimes b_2(u,v) := \sup_{z\in S}(b_1(u,z)+b_2(z,v))$. Let $l > -\infty$. Suppose that there are two sequences of states u_n and v_n that converge respectively to u and v in S such that $b_1 \otimes b_2(u_n, v_n) \geq l$ for all $n \geq 1$. Then, given any $\epsilon > 0$, there exists a sequence of states z_n such that $b_1(u_n, z_n) + b_2(z_n, v_n) \geq l - \epsilon$ for all $n \geq 1$. Since S is compact, some subsequence z_{n_k} converges to a state $z \in S$. Now

$$\begin{aligned} b_1 \otimes b_2(u,v) &\geq b_1(u,z) + b_2(z,v) \\ &\geq \limsup_{k\to\infty} \left[b_1(u_{n_k}, z_{n_k}) + b_2(z_{n_k}, v_{n_k})\right] \\ &\geq l - \epsilon, \end{aligned}$$

by the upper semicontinuity of b_1 and b_2. Since ϵ is arbitrary, $b_1 \otimes b_2(u,v) \geq l$. Thus the upper level set of $b_1 \otimes b_2$ is closed for every level l, and so $b_1 \otimes b_2$ is upper semicontinuous. □

Theorem 13 *The operator R^p is compact for $p = T + \lceil \beta / \min(\sigma, \rho - \sigma) \rceil + 1$.*

Proof. We must show that $r^p(u,v) > -\infty$ for all $u, v \in S$, and that r^p is continuous. Let $u := (q_u; u_1, \ldots, u_T)$ and $v := (q_v; v_1, \ldots, v_T)$ be states in S. For any state $s := (q_s, s_1, \ldots, s_T)$, define $\widehat{q}_s := \max_{2\leq n\leq T} \sum_{i=n}^{T}(s_i - \sigma) \vee (q_s + \sum_{i=1}^{T}(s_i - \sigma))$, which we interpret as the queue length at the end of the block of T emissions defined by the state. Let $z := (q_v - \widehat{q}_u)/(p - T) + \sigma$. Then $0 \leq z \leq \rho$ and the path

$$\Xi_p(q_u; u_1, \ldots, u_T, z, \ldots, z, v_1, \ldots, v_T)$$

has a finite total reward.

The upper semicontinuity of r^p follows from the previous lemma by induction.

To establish lower semicontinuity we again exploit the correspondence between valid paths and traffic realisations that obey the leaky bucket constraint. Let $u, v \in S$. We know from above that $r^p(u,v) > -\infty$. As a sum of upper semicontinuous functions, R_p is is upper semicontinuous on S^p, and so reward $R_p(x) = r^p(u,v)$ is attained by some path $x = u, x_1, \ldots, x_{p-1}, v$ of length p between u and v. Denote by $a_n := (\Xi_p^{-1}(x))_n$ the realisation corresponding to x. A simple argument using the isotonicity of our functional shows that $q_v - q_u = \sum_{i=T+1}^{p}(a_n - \sigma)$, that is to say no service goes unused in the middle

section. For, if an amount s of service is unused at time m, then we could add s to a_m and construct a path x' of length p between u and v for which $R_p(x') > R_p(x)$. We will construct a continuous function that assigns to any pair of states (u', v') near (u, v) a path near x in the product topology on S^{p+1}.

The coordinate map $S \to \mathbb{R}^+ : (q_s; s_1, \ldots, s_n) \to q_s$ and the function $S \to \mathbb{R}^+ : s \to \widehat{q}_s$ defined above are both continuous. Therefore the map $\Delta : S^2 \to \mathbb{R}^+ : (u', v') \to q_{v'} - q_v - \widehat{q}_{u'} + \widehat{q}_u$ is continuous and $\Delta(u, v) = 0$.

Consider the set of maps from $[0, \rho] \to [0, \rho]$ defined by

$$z_\delta(x) := \begin{cases} (1+\delta)x, & \delta < 0 \\ (1-\delta)x + \delta\rho, & \delta \geq 0, \end{cases}$$

for δ in the range $[-1, 1]$. Define $Z : [-1, 1] \to A^{p+T}$ by

$$(Z(\delta))_n := z_\delta(a_n).$$

Each component of $Z(\delta)$ lies in $A := [0, \rho]$, and Z is a continuous function of δ using the product topology on A^{p+T}. The function $\mathcal{Z} : [-1, 1] \to \mathbb{R} : \delta \to \sum_{n=1}^{p+T} (Z(\delta))_n$ is clearly continuous. Since there must be some m in the range $T + 1 \leq m \leq p$ for which $a_m > 0$ and some n in the same range for which $a_n < \rho$, we conclude that $\mathcal{Z}$ is strictly increasing. Therefore $\mathcal{Z}$ is invertible and its inverse $\mathcal{Z}^{-1}$ is continuous.

Define the operator $\Psi : A^{p+T} \to A^{p+T}$:

$$(\Psi c)_n := \begin{cases} \psi_n - \psi_{n-1} + \sigma, & T < n \leq p \\ c_n, & n \leq T \text{ or } n > p. \end{cases}$$

where

$$\psi_n := \Big(\beta \wedge (q_u + \sum_{i=1}^{n} (c_i - \sigma)) \vee 0\Big).$$

Clearly, Ψ is continuous in c and $\Psi(a) = a$. Furthermore, every point in the image of Ψ will obey the leaky bucket constraints.

We now combine the functions we have defined. The function $\Phi : \Psi \circ Z_a \circ \mathcal{Z}^{-1} \circ \Delta : S \times S \to A^{p+T}$ is a composition of continuous functions and is therefore continuous. Also $\Phi(u, v) = a$. The image obeys the leaky bucket constraints. Let $W : S \times S \to \mathbb{R}$ be defined by $W := R \circ \Xi^{-1} \circ \Phi$ which is the reward of the corresponding path as a function of the end points. This function is continuous. Furthermore, $W(u, v) = r^p(u, v)$ and

$W(u',v') \leq r^p(u',v')$ for all $u',v' \in S$. It follows that r^p is lower semicontinuous at (u,v). □

Theorem 14 *If B is a compact Bellman operator with kernel b, then for any path x such that $B(x) := \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} b(x_i,x_{i+1})$ exists, we have that $B(x) \leq \sup_{y\in P} B(y)$, where P is the set of cyclic paths.*

Proof. Since B is compact it has a unique eigenvalue λ. The maximum reward over n transitions is $\sup_{u\in S}(B^n 0)(u)$, where 0 is the function that maps each state to 0. Therefore

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} b(x_i,x_{i+1}) - \lambda &\leq \frac{1}{n}\sup_{u\in S}(B^n 0)(u) - \lambda \\ &\leq \frac{1}{n}||B^n 0 - n\lambda||.\end{aligned}$$

From Theorem 11 we have that $n^{-1}||B^n 0 - n\lambda|| \to 0$ as $n \to \infty$. Thus if $B(x)$ exists, then $B(x) \leq \lambda$. But the eigenvalue of B equals the maximal cycle mean of the associated MDP and so $\sup_{y\in P} B(y) = \lambda$. The conclusion follows. □

Corollary 2 *Even if B is merely quasicompact the same conclusion holds.*

Proof. Denote by X the set of paths x for which $B(x)$ exists. Since B is quasicompact, there exists $p > 0$ for which B^p is compact. Consider the MDP with state space S and reward function

$$\overline{b}(u,v) := \max\{b(u,x_1) + b(x_1,x_2) + \cdots + b(x_{p-1},v) : x_1,\ldots,x_{p-1} \in S\}.$$

Then the Bellman operator $\overline{B}$ corresponding to this MDP is equal to B^p. We apply the theorem above and deduce that $\sup_{x\in X} \overline{B}(x) = \sup_{y\in P} \overline{B}(y)$.

For any path $x = \{x_i\}_{i\in\mathbb{N}}$, the path $\overline{x} = x_0, x_p, x_{2p}, \ldots$ satisfies $\overline{B}(\overline{x}) \geq pB(x)$ and so $\sup_{x\in X} \overline{B}(x) \geq p \sup_{x\in X} B(x)$. Conversely, by choosing $p-1$ states $x_1^{u,v},\ldots,x_{p-1}^{u,v}$ for each $u,v \in S$, such that $b(u,x_1^{u,v}) + b(x_1^{u,v},x_2^{u,v}) + \cdots + b(x_{p-1}^{u,v},v) = \overline{b}(u,v)$, we can take any cyclic path $\overline{y}$ and construct another cyclic path

$$y = \overline{y}_0, x_1^{\overline{y}_0,\overline{y}_1},\ldots,x_{p-1}^{\overline{y}_0,\overline{y}_1},\overline{y}_1,x_1^{\overline{y}_1,\overline{y}_2},\ldots,x_{p-1}^{\overline{y}_1,\overline{y}_2},\overline{y}_2,\ldots$$

such that $B(y) = p^{-1}\overline{B}(\overline{y})$. Thus $\sup_{x\in P}\overline{B}(x) \leq p\sup_{x\in P}B(x)$. Putting all of this together we find

$$\sup_{x\in X}B(x) \leq p^{-1}\sup_{x\in X}\overline{B}(x) = p^{-1}\sup_{x\in P}\overline{B}(x) \leq \sup_{x\in P}B(x).$$

□

We may use these theorems and reasoning similar to that of the previous section to reduce the optimisation of the effective bandwidth over all stationary processes to an optimisation over periodic processes. Again the correspondence between realisations in C and paths in that have finite average reward means that $\sup_{\omega\in C}\bar{q}(\omega) = \sup_{x\in X}R(x)$. Applying Corollary 2 to the operator R, the latter supremum is equal to $\sup_{x\in P}R(x)$. For each cyclic path x, we may construct a periodic process ν in the manner described in Section 3.3 such that $R(x) = \mathbb{E}_\nu\bar{q}$. Thus

$$\mathbb{E}_\mu q = \mathbb{E}_\mu\bar{q} \leq \sup_{\omega\in C}\bar{q}(\omega) = \sup_{\nu\in\mathcal{P}}\mathbb{E}_\nu\bar{q} = \sup_{\nu\in\mathcal{P}}\mathbb{E}_\nu q$$

for any stationary measure μ.

3.6 Continuous Time Optimisation

Finally we are ready to tackle the continuous time optimisation problem that is our main interest. First we must consider ergodic theory in this setting.

Ergodic Theory in Continuous Time

In this setting we have a one-parameter semigroup $\{\theta_t\}_{t\geq 0}$ of measure preserving mappings. This means that $\theta_{t+s} = \theta_t\theta_s$. We assume that θ is a measurable function of both its arguments, that is for every measurable subset M of $\Omega\times\mathbb{R}^+$, the set $\{(\omega,t) : \theta_t\omega \in M\}$ is also measurable in the product σ-algebra on $\Omega\times\mathbb{R}^+$.

The following is the continuous ergodic theorem.

Theorem 15 *If f is an integrable function and μ is a stationary probability measure, then $h^{-1}\int_0^h f(\theta_t\omega)\,dt$ converges μ-almost surely as $h\to\infty$ to an integrable and invariant function $\widehat{f}(\omega)$, and $\mathbb{E}_\mu\widehat{f} = \mathbb{E}_\mu f$.*

MDPs in Continuous Time

Continuous time variational problems generally take place in a state space that is a differentiable manifold. The paths take the form of absolutely continuous trajectories through this space. The role of transition function is played by a Lagrangian function $L(u, v)$, so that the reward of a path x over an interval $[0, h]$ is

$$\int_0^h L[x(t), \dot{x}(t)]\, dt.$$

In our variational problem the Lagrangian is purely a function of the state: it does not depend on its second parameter. This will allow us to dispense with an investigation of the differentiable structure of our space. Also we do not require that paths be absolutely continuous.

The MDP we wish to consider is

- *State Space.* Let A be that set of right continuous, nondecreasing functions from $[0, T) \to \mathbb{R}^+$ that obey the peak rate constraint ρ. Then the state space S is

$$S := \{(q, a) \in Q \times A : \mathbb{Q}^{(q)} a(t) \leq \beta \text{ for all } t \in [0, T)\}$$

- *Reward Function.* The Lagrangian is $L : S \to \mathbb{R}^+ : (a, q) \to \exp(a(T^-))$. We define the reward of any valid path $x(t)$ of length h to be

$$\mathcal{R}(x) := \int_0^h L[x(t)]\, dt.$$

  The valid paths are those for which the arrivals and queue lengths are consistent. In other words, denoting by (q_t, a_t) the state that x is in at time t, then the valid paths are those for which $a_t(s) - a_t^-(u) = a_{t+z}(s - z) - a_{t+z}^-(u - z)$ and $q_{t+z} = \mathbb{Q}^{(q_t)} a_t(z)$ for any $t \in [0, h]$, $s \in [0, T)$, $u \in [0, s)$, $0 \leq z \leq \min(s, u, T - t)$. We denote the set of valid paths by V.

A Bellman operator $R^h : C(S) \to C(S)$ is defined for each $h \in \mathbb{R}^+$ as follows:

$$(R^h f)(u) := \sup_{x \in V} \left\{ \int_0^h L[x(t)]\, dt + f(x(h)) : x(0) = u \right\}$$

We also define R^0 to be the identity operator on $C(S)$. This operator has the kernel

$$r^0(u, v) = \begin{cases} 0, & u = v \\ -\infty, & u \neq v \end{cases}.$$

For $h > 0$, the operator R^h has the kernel

$$r^h(u,v) = \sup_{x \in V} \left\{ \int_0^h L[x(t)]\, dt : x(0) = u, x(h) = v \right\}.$$

The operators R^h, $h > 0$ together form a semigroup of operators, that is for any $s, t \geq 0$ we have $R^{s+t} = R^s R^t = R^t R^s$. The identity element of this semigroup is R^0.

Definition 5 *A function $f \in C(S)$ is said to be a $(\max, +)$ eigenfunction of a semigroup of Bellman operators B^h with eigenvalue λ, if $B^h f = f + h\lambda$ for all $h \geq 0$.*

If B^h is a compact Bellman operator for each $h \geq 0$, then each operator can have at most one eigenvalue and we have the following proposition.

Proposition 1 *If a semigroup of compact Bellman operators has an eigenvalue then it is unique.*

There is a continuous time version of Theorem 10 which guarantees the existence of an eigenvalue when the semigroup is continuous.

Definition 6 *A semigroup B^h of Bellman operators is called a continuous semigroup if each operator is compact and*

$$\lim_{h \to 0^+} ||B^h f - f|| = 0 \qquad \text{for all } f \in C(S).$$

Theorem 16 *Every continuous semigroup of Bellman operators has a continuous eigenfunction.*

The semigroup R^h is not continuous: for small values of h, R^h is not compact. However, just as in discrete time all we require is that R^h be compact for some $h > 0$. The development parallels that of the previous section.

Theorem 17 *The operator R^p is compact for $p = T + \beta / \min(\sigma, \rho - \sigma) + 1$.*

Proof. Positivity. Let $u = (q_u, a_u)$ and $v = (q_v, a_v)$ be in S. For any state $x := (q_x, a)$, define $\overline{q}_x := \mathbb{Q}^{(q_x)} a(T^-)$. Define the realisation $a : [0, p+T) \to \mathbb{R}^+$ by

$$a(t) := \begin{cases} a_u(t), & 0 \leq t < T \\ a(T^-) + (t-T)\Big((q_v - \overline{q}_u)/(p-T) + \sigma\Big), & T \leq t < p \\ a(p^-) + a_v(t-p), & p \leq t < p+T \end{cases}.$$

Define $y_t : [0,T) \to \mathbb{R}^+ : s \mapsto a(s-t) - a(t^-)$. Then the path $x(t) := (\mathbb{Q}a(t), y_t)$ is a path of length p from u to v such that $\mathcal{R}(x) > -\infty$, and so $r^p(u,v) > -\infty$.

Upper semicontinuity. Let $l > -\infty$. Suppose there exists a sequence of pairs of states (u_i, v_i) converging to some pair of states (u,v) such that $R^p(u_i,v_i) \geq l$ for all $i \geq 1$. Given any $\epsilon > 0$, there exists a sequence of paths $x_h : [0,p] \to S$ such that $x_h(0) = u_h$, $x_h(p) = v_h$, and $\mathcal{R}(x_h) \geq l - \epsilon$ for all $h \geq 1$. We have seen above that the set of paths over a finite time interval is compact and so some subsequence x_{h_k} converges to a path x. Now the set of valid paths is closed and so x is valid. Furthermore $\mathcal{R}$ is continuous on the set of valid paths. Thus

$$R(u,v) \geq \mathcal{R}(x) = \lim_{k\to\infty} \mathcal{R}(x_{h_k}) \geq l - \epsilon.$$

Since ϵ is arbitrary, $R(u,v) \geq l$. Therefore, the upper level set is closed for each level l and so R is upper semicontinuous.

Lower semicontinuity. Since R is upper semicontinuous, there is a path $x : [0,p] \to S$ from u to v which attains the supremum of the reward over all paths from u to v. Clearly, $\mathbb{U}x(T+p) = \mathbb{U}x(T)$ since otherwise, by the isotonicity of the effective bandwidth functional, we could find a path which still obeys the leaky bucket parameters but has a higher reward. In other words, no service goes unused in the time interval $[T, T+p]$. This path will correspond to a realisation $a \in \mathbb{D}^+_{p+T}$ of emissions from the source. We will construct a continuous function that assigns to any pair of states (u',v') near (u,v) a realisation near a in the uniform topology on $\mathbb{D}^+_{p+T}$.

The coordinate map $S \to \mathbb{R}^+ : (q_x, x) \to q_x$ and the function $S \to \mathbb{R}^+ : x \to \overline{q}_x$ are both continuous. Therefore the map $\Delta : S^2 \to \mathbb{R}^+ : (u',v') \to q_{v'} - q_v - \overline{q}_{u'} + \overline{q}_u$ is continuous and $\Delta(u,v) = 0$.

As a nondecreasing function a is differentiable almost everywhere. However a' can not be ρ at almost all points of $[0,p]$, for if it is then $q_v = q_u + p(\rho - \sigma) \geq \beta$ and a' does not obey the leaky bucket constraint. Similarly, a' can not be 0 at almost all points of $[0,p]$. For δ in the range $[-1,1]$, we define the set of maps $z_\delta : [0,\rho] \to [0,\rho]$ by

$$z_\delta(x) := \begin{cases} (1+\delta)x, & \delta < 0 \\ (1-\delta)x + \delta\rho, & \delta \geq 0 \end{cases}$$

Consider the mapping $Z : [-1,1] \to \mathbb{D}^+_{p+T}$ defined by

$$\Big(Z(\delta)\Big)(t) := \int_0^t z_\delta[a'(x)]\,dx.$$

Clearly, $Z(\delta)$ obeys the peak rate constraint and is non-decreasing for each $\delta \in [-1,1]$. Also, Z is a continuous function of δ, using the uniform topology on $\mathbb{D}^+_{p+T}$. The function $\mathcal{Z} : [-1,1] \to \mathbb{R}^+ : \delta \to (Z(\delta))(p^-)$ is continuous and strictly increasing in δ. Therefore this function is invertible and its inverse $\mathcal{Z}^{-1}$ is continuous.

Define the operator $\psi : \mathbb{D}^+_{p+T} \to \mathbb{D}^+_{p+T}$ by

$$(\psi c)(t) := \Big(\beta \wedge (q_u + c(t) - \sigma t) \vee 0\Big) + \sigma t - q_u.$$

Clearly, ψc is continuous in c and $\psi a = a$. Furthermore, every point in the image of ψ will obey the leaky bucket constraints.

We now combine the functions we have defined. The function $\psi \circ Z \circ \mathcal{Z}^{-1} \circ \Delta : S \times S \to \mathbb{D}^+_{p+T}$ is a composition of continuous functions and is therefore continuous. Also $(\psi \circ Z \circ \mathcal{Z}^{-1} \circ \Delta)(u,v) = a$ and each element of the image obeys the leaky bucket constraints. □

The further development is almost exactly the same as that for discrete time, and so we merely provide a summary. Consider the discrete time MDP with state space S and operator R^p. Since R^p is compact, it has a unique eigenvalue λ, and there are periodic paths with long term average reward arbitrarily close to λ. For each of these, we may interpolate between consecutive states to form a periodic path x in continuous time such that $\mathcal{R}(x)$ is arbitrarily close to long term average reward of the discrete time path. Also the eigenvalue of R^p is an upper bound on $\mathcal{R}(z)$ for any other path z. Thus the supremum of $\mathcal{R}$ over all paths is the same as its supremum over periodic paths.

The continuous time ergodic theorem implies that

$$\overline{f}(a) := \lim_{t\to\infty} \frac{1}{t} \int_0^t e^{\vartheta(a(x+T)-a(x^-))}\,dx$$

has the same expectation with respect to any stationary measure on $\mathbb{D}^+$ as $f(a) := \exp[\vartheta(a(T))]$. Again, the correspondence between valid paths of states and realisations that obey the leaky bucket constraints means that, for any $\epsilon > 0$, there exists some periodic realisation a^* in C such that $\overline{f}(a) \le \overline{f}(a^*) + \epsilon$ for all $a \in C$. Constructing a measure

μ_{a^*} on $\mathbb{D}^+$ by choosing the phase of a^* uniformly, wee see that $\mathbb{E}_\nu \overline{f} \leq \mathbb{E}_{\mu_{a^*}} \overline{f} + \epsilon$ for any measure ν supported by C. We conclude that the supremum of $\mathcal{E}_{\vartheta,T}(\nu)$ over all probability measures supported on C is equal to its supremum over the ergodic and periodic probability measures supported on C.

3.7 Other Functionals

The methods of this chapter may be applied to other functionals apart from the effective bandwidth. As an example we deal with the functional that represents the mean queue length in an infinite buffer with stochastic service rate. This functional is more difficult to deal with than the previous one because even if the sample path output of the source is given there will still be some randomness in the form of the service process. This randomness is different from the sort usually encountered in stochastic optimal control theory. There a decision is made based on the current state of the system. However the source in our problem receives no feedback from the queuing system—it must decide in advance on the traffic which is the worst on average and send it regardless of what the service process ends up doing. Thus whenever we apply dynamic programming ideas they will always be of the deterministic variety. We will need a method of dealing with the fact that the behaviour of the server is stochastic however.

It is technically simpler to assume the service process is two sided. This means that the time shift operator θ is a bijection. We have seen in Section 2.8 that a separable stationary two sided process may be constructed from any one sided process. We will also work in discrete time—the extension to continuous time is obvious. Assume that the service capacity is represented by π, a stationary measure on the space $\Omega := (\mathbb{R}^+)^{\mathbb{Z}}$. If outcome $\omega \in \Omega$ occurs then each clock tick i an amount $\omega(i)$ of work can be done. The possible states of the leaky bucket are $L := [0, \beta]$. Denote by V the set of non-negative random variables on the space Ω. Here we identify random variables that agree almost surely. The coordinate random variables are denoted $X_i(\omega) := \omega(i)$. We define the following Markov Decision Procedure:

- *State Space*: $X := L \times V$,
- *Reward Function*: For any $a \in [0, \rho]$, the transition $(l, Q) \to ([l + a - \sigma]^+, [Q \circ \theta^{-1} +$

$a - X_0 \circ \theta^{-1}]^+$) has reward $\mathbb{E}_{\pi\circ\theta^{-1}}[Q]$, provided that $[l + a - \sigma]^+ \in L$. Otherwise the reward is $-\infty$.

Heuristically, a state has knowledge of the instantaneous content of the leaky bucket and the queue length for each outcome of the service process.

If π was not stationary we would get an *inhomogeneous* MDP, that is the reward function would change with time. Under such circumstances we would not expect to see periodic behaviour in the path that maximises the long term reward—the optimum path would be likely to react to changes in the reward function. However, if μ is stationary then the MDP described above is homogeneous. Thus the methods we have employed previously to show that the worst case traffic is periodic are applicable here also. However it is likely that showing that the reward operator for this MDP is quasicompact will prove to be difficult.

For the loss rate functional the state space is the same and the reward function is replaced by

- *Reward Function (loss rate)*: For any $a \in \mathbb{R}$ the transition $(l, Q) \to ([l + a - \sigma]^+, b \wedge [Q \circ \theta^{-1} + a - X_0 \circ \theta^{-1}]^+)$ has reward $r_i := \mathbb{E}_{\pi\circ\theta^{-1}}[Q \circ \theta^{-1} + a - X_0 \circ \theta^{-1} - b]^+$

A stationary stochastic buffer size, experienced for example by low priority traffic, may be dealt with in a similar manner. As yet we are unable to represent as an MDP the maximisation of any of the functionals describing the behaviour of the network fed by a finite number of independent identical sources.

Chapter 4

Convexity and Extreme Points

In the previous chapter we have shown how the optimisation problem may be reduced from examining all stationary traffic processes to just examining those that are periodic. We can break the optimisation into an optimisation over all periodic processes with fixed period p, followed by an optimisation over p. It is the former that we will consider here. The fact that all realisations are periodic means that we may concentrate our attention on the compact interval $[0, p]$. This introduces considerable technical simplification.

The main technique we use is to exploit the convexity of the functionals. If a convex functional on a convex set attains its supremum, then it does so at an extreme point of the set. If we could show that a functional f is convex and continuous and that the constraint set C is the closure of the convex hull of the set ξ of its extreme points, then we could conclude that $\sup_{\xi} f = \sup_{\text{conv}\xi} f$ since f is convex, and that $\sup_{\text{conv}\xi} f = \sup_{\overline{\text{conv}}\xi} f = \sup_C f$, since f is continuous. We will then have reduced the problem from an optimisation over the entire constraint set to an optimisation over its extreme points.

4.1 Convexity and Topological Vector Spaces

We collect here some background material we will need. As our setting is that of a topological vector space, it will be useful to discuss their general features.

A *topological vector space* is a linear space V endowed with a topology under which the operations

$$V \times V \to V \quad : \quad (a, b) \to a + b$$

$$\text{and} \qquad \mathbb{R} \times V \to V \quad : \quad (\lambda, a) \to \lambda a$$

are continuous. A set $S \subset V$ is said to be convex if it is closed under the operation of taking convex combinations of pairs of points, that is if

$$(1-\alpha)a + \alpha b \in S$$

for all $a, b \in S$ and $\alpha \in [0,1]$. A topological vector space is said to be a *locally convex* space if it is Hausdorff and the origin possesses a base of convex neighbourhoods. Normed spaces are locally convex since the balls centered at the origin form a neighbourhood base and each ball is convex by the triangle inequality. If the origin possesses a base of convex neighbourhoods then so will every other point since neighbourhoods may be translated: if N is a neighbourhood of a then $b + N := \{b + p : p \in N\}$ is a neighbourhood of $a + b$.

A function $f : V \to \overline{\mathbb{R}}$ is said to be *convex* if

$$f((1-\alpha)a + \alpha b) \leq (1-\alpha)f(a) + \alpha f(b)$$

for all $a, b \in V$ and $\alpha \in [0,1]$. The space of convex functions on V is closed under the operations of addition and scalar multiplication. Furthermore, if $\{f_\gamma\}$ is a family of convex functions then their pointwise supremum $\sup_\gamma f_\gamma$ is also convex. An interesting feature of convex functions is that their one-sided derivatives always exist. For example, let $f : I \to \mathbb{R}$ be convex function on an interval $I \subset \mathbb{R}$. Then $(f(b) - f(a))/(b - a)$ is nondecreasing in b and thus has a limit $f'_+(a)$ as $b \searrow a$. Similarly the left hand derivative $f'_-(a) := \lim_{b \nearrow a} (f(a) - f(b))/(a - b)$ also exists. Both of these derivatives can be shown to be non-decreasing functions.

The effective domain of a convex function is the set on which it is finite. In the finite dimensional space $\mathbb{R}^d$ equipped with the usual topology, it can be shown that every convex function is continuous in the interior of its effective domain. However, in an infinite dimensional space this in not in general true. A useful theorem is the following which gives a condition for a convex function to be continuous.

Theorem 18 *In a topological vector space V, a convex function $f : V \to \overline{\mathbb{R}}$ is continuous if it is bounded above in a neighbourhood of some point.*

The *convex hull* of a subset S of a vector space is defined to be the set of all finite convex combinations of its elements. Thus

$$\operatorname{conv} S := \left\{ \sum_{i=1}^{I} \alpha_i p_i : \sum_{i=1}^{I} \alpha_i = 1,\ \alpha_i \geq 0,\ p_i \in S,\ I \in \mathbb{N} \right\}$$

The convex hull of S can be shown to be the intersection of all convex sets containing S, and is therefore the smallest such set. Note that the supremum of a convex function over a set is the same as the supremum over the convex hull of the set. For if $p \in \operatorname{conv} S$ then p is a convex combination $\sum_i \alpha_i p_i$ of points in S. Therefore $f(p) \leq \sum_i \alpha_i f(p_i)$ and so $f(p) \leq \max_i f(p_i) \leq \sup_{a \in S} f(a)$.

When considering the supremum of a convex function over a convex set S, the extreme points play an important role.

Definition 7 *We say that a point p is an extreme point of a convex set $S \subset V$ if $p \in S$ and p is not an interior point of any line segment contained in S.*

It is amongst the extreme points that we look for a maximum. However, some sets are deficient in extreme points; for example, any open set has none. To ensure that there are enough extreme points to fully represent the constraint set we will need to show that the closure of the convex hull of its extreme points is the constraint set itself. One theorem that is useful in this regard is that of Krein and Milman.

Theorem 19 (Krein-Milman) *Every compact convex subset of a locally convex t.v.s. is the closed convex hull of the set of its extreme points.*

Thus it will prove advantageous to choose a topology such that the constraint set is compact.

A subset of a metric space is said to be precompact (totally bounded) if for any $\epsilon > 0$ it can be covered by a finite number of open balls of radius ϵ, each of which is centered on a point of the set. There is also an equivalent notion in a topological vector space. Here the linear structure provides the necessary uniformity to define precompactness. Let V be a locally convex space. Instead of using open balls to cover sets, we use neighbourhoods of the origin. For any neighbourhood U of the origin, we say that a set $S \in V$ is small of order U if $p - q \in U$ for all $p, q \in S$. A set $S \in V$ is then said to be precompact if for

any neighbourhood U of the origin it can be covered by a finite number of sets which are small of order U.

In a locally convex space which is also a metric space the two definitions are equivalent. To see this let assume that $S \in V$ is precompact in the metric sense. For any neighbourhood U of the origin we can find some radius ϵ such that any open ball with this radius is small of order U, and S can be covered by a finite number of these. Conversely, for any $\epsilon > 0$, the ball B_ϵ of radius ϵ about the origin is a neighbourhood of the origin. Therefore, if S is compact in the t.v.s. sense, there is a finite collection $\{S_i\}_{i \in 1,\ldots i}$ of set in V that cover S each of which is small of order B_ϵ. Choose $q_i \in S_i$ for each $i \in 1, \ldots i$. Then $S_i \in B_\epsilon(q_i)$ for each $i \in 1, \ldots i$ and therefore these open balls cover S.

A metric space is called complete if every Cauchy sequence converges. Again there is a similar notion in topological vector spaces. We need to use nets over arbitrary directed sets rather than sequences to define completeness because the t.v.s. may not be metrisable. A net $\{x_n\}_{n \in D}$ in a t.v.s. is called a Cauchy net if, for every neighbourhood U of the origin, there is an $N \in D$ such that $x_n - x_m \in U$ for all $n, m \geq N$. If every Cauchy net in a subset of the t.v.s. converges to a point in the subset, the subset is said to be complete. It can be shown [32] that in a metrisable t.v.s. a set is complete in this sense if and only if it is complete in the metric sense.

We have the following useful facts about precompactness and completeness. A subset of a locally convex space is compact if and only if it is precompact and complete. Also a closed subset of a complete t.v.s. is complete. Putting these together, we have that a closed precompact subset of a complete t.v.s. is compact.

4.2 Optimisation for Fixed p

Consider a stationary ergodic process with periodic realisations, each of period p. It is obvious that to specify this process it is enough to specify the behaviour of one of its realisations in the interval $[0, p)$. The behaviour of this sample path over any other interval may be reconstructed since it is periodic, and every other realisation is just this realisation shifted in time. The space of possible behaviours of the source in $[0, p)$ is the space $\mathbb{D}_p^+$ of right-continuous nondecreasing functions from $[0, p) \to \mathbb{R}^+$. To make this into a linear

space, we take the linear completion $\mathbb{D}_p := \mathbb{D}_p^+ - \mathbb{D}_p^+ = \{a - b : a, b \in \mathbb{D}^+\}$. The elements of $\mathbb{D}_p$ are the right-continuous functions from $[0,p) \to \mathbb{R}$ of bounded variation. We will consider possible choices of topology on this space in a later section. In what follows, if $a \in \mathbb{D}_p$ and $t \notin [0,p)$, then $a(t)$ is understood to mean $\lfloor t/p \rfloor a(p^-) + a(t - \lfloor t/p \rfloor p)$.

The space $\mathbb{D}_p$ is a linear space under the usual operations of addition and multiplication by scalars. Let $q_0 := \mathbb{Q}_s^0 a(p^-)$ and $q_0^b := \mathbb{Q}_s^0 a(p^-)$. We define the following functionals on $\mathbb{D}_p$:

- average queue length $\mathcal{Q}_s^{\text{ave}}(a) := \dfrac{1}{p}\displaystyle\int_0^p \mathbb{Q}_s^{(q_0)} a(t)\, dt$,
- maximum queue length $\mathcal{Q}_s^{\max}(a) := \displaystyle\sup_{t \in [0,p)} \mathbb{Q}_s^{(q_0)} a(t)$,
- loss $\mathcal{L}_{b,s}(a) := \dfrac{1}{p}\mathbb{L}_s^{(q_0^b)} a(p^-)$,
- effective bandwidth $\mathcal{E}_{\vartheta,T}(a) = \dfrac{1}{p}\displaystyle\int_0^p e^{\vartheta[a(T+t)-a(t)]}\, dt$.

It may be easily verified that each of these functionals agrees with its version defined on the set of periodic ergodic processes. For example, consider the average queue length functional. If we write $\overline{a}_x(t) = \theta^x a(t)$, where $\theta^x a(t) := a(t+x) - a(x^-)$, then the process obtained by choosing x uniformly in $[0,p)$ is seen to be the double sided process corresponding to a. The average queue length as defined in Section 2.8 is $\int_0^p Q\overline{a}_x(0)\, dx$. But this is equal to $\int_0^p \mathbb{Q}_s^{(q_0)} a(t)$ by the memoryless property. Similarly for the other functionals.

The set $C := \{a \in \mathbb{D}_p^+ : \mathcal{Q}_\sigma^{\max} a \le \beta, \mathcal{Q}_\rho^{\max} a = 0\}$ of elements of $\mathbb{D}_p$ that obey the leaky bucket constraint (σ, β) and the peak rate constraint ρ is intersection of two lower level sets of $\mathcal{Q}^{\max}$.

The optimisation for fixed p is now a deterministic optimisation problem:

$$\text{Maximise } \mathcal{E}_{\vartheta,T}(a), \quad \text{subject to } a \in C.$$

4.3 Convexity

We prove here some convexity results that are fundamental to our programme.

Queue Length

Recall for $a \in \mathbb{D}$,

$$\left(\mathbb{Q}_s^{(q_0)} a\right)(t) = \sup_{r \in [0,t]} [a(t) - a(r^-) - (t-r)s] \vee (q_0 + a(t))$$

In addition to the isotonicity result of Chapter 2 we have the following properties of $\mathbb{Q}$:

- homogeneity: $\mathbb{Q}_{\alpha s}^{(\alpha q_0)} \alpha a = \alpha \mathbb{Q}_s^{(q_0)} a$,
- joint convexity in a and s:

$$\mathbb{Q}_{\alpha s_2 + (1-\alpha)s_1}(\alpha a_2 + (1-\alpha)a_1) \leq \alpha \mathbb{Q}_{s_2} a_2 + (1-\alpha)\mathbb{Q}_{s_1} a_1.$$

The latter property follows since, for fixed t, the queue length is the supremum of a family of affine functions of s and $a(t)$.

It is a immediate consequence that Q^{ave} is a convex functional on $\mathbb{D}_p$ since an equivalent expression for Q^{ave} to that above is

$$Q_s^{\text{ave}}(a) = \frac{1}{p} \int_p^{2p} (\mathbb{Q}_s^{(0)} a)(t)\, dt.$$

In this form it is clear that Q^{ave} is convex. Similarly,

$$Q_s^{\max}(a) = \sup_{t \in [p,2p]} (\mathbb{Q}_s^{(0)} a)(t),$$

and so $Q^{\max}$ is convex.

Loss Rate

In Chapter 2 we defined the loss path l associated with an arrival path a. We shall show that the function $\mathbb{D}_p \to \mathbb{D}_p^+ : a \to l$ is jointly convex in a, s, and b.

Theorem 20 *Let $b_1, s_1, b_2, s_2 \geq 0$ and let $a_1, a_2 \in \mathbb{D}^+$. Then the loss l^* associated with buffer size $\alpha b_2 + (1-\alpha)b_1$, service rate $\alpha s_2 + (1-\alpha)s_1$ and arrivals $\alpha a_2 + (1-\alpha)a_1$ obeys*

$$(l^*)(t) \leq (\alpha l_2 + (1-\alpha)l_1)(t) \qquad \text{for all } t \geq 0.$$

Proof. Let $l_1 := \{l \in \mathbb{D}^+ : l \leq a_1, Q_{s_1}(a_1) \leq b_1\}$ and $l_2 := \{l \in \mathbb{D}^+ : l \leq a_2, Q_{s_2}(a_2) \leq b_2\}$. For any $\alpha \in [0,1]$, define $a_\alpha := \alpha a_2 + (1-\alpha)a_1$, $b_\alpha := \alpha b_2 + (1-\alpha)b_1$, $s_\alpha := \alpha s_2 + (1-\alpha)s_1$, and $l_\alpha := \alpha l_2 + (1-\alpha)l_1$. Then

$$\begin{aligned} Q_{s_\alpha}(a_\alpha - l_\alpha) &\leq Q_{\alpha s_2}(\alpha a_2 - \alpha l_2) + Q_{(1-\alpha)s_1}((1-\alpha)a_1 - (1-\alpha)l_1) \\ &= \alpha Q_{s_2}(a_2 - l_2) + (1-\alpha)Q_{s_1}(a_1 - l_1) \\ &\leq \alpha b_2 + (1-\alpha)b_1. \end{aligned}$$

Therefore

$$l_\alpha \in L_\alpha := \{l \in \mathbb{D}^+ : l \leq a_\alpha, Q_{s_\alpha}(a_\alpha) \leq b_\alpha\}.$$

Thus $l^* := \inf L_\alpha \leq l_\alpha$. □

Effective Bandwidth

As the mean of a family of convex functionals, the functional $\mathcal{E}_{\vartheta,T}$ is clearly convex for all $\vartheta, T \in \mathbb{R}^+$.

4.4 Choosing a Topology

We need to choose a topology on the space $\mathbb{D}_p$. There are some natural criteria for an appropriate topology:

- The topology must make $\mathbb{D}_p$ into a locally convex topological vector space.
- The various functionals under consideration should be continuous in the topology.
- The constraint set should be compact in the topology. This will enable us to use the Krein-Milman theorem to show that the constraint set is the closure of the convex hull of its extreme points.

The Uniform Topology

Since we have insisted that the source obeys a peak rate constraint, all elements of the constraint set C are continuous. We may therefore consider the subspace $\mathbb{D}_p^c$ consisting of

the continuous elements of $\mathbb{D}_p$. A natural topology to take on this space is that induced by the norm

$$||w|| := \sup_{x \in [0,p)} w(x).$$

Since it is induced by a norm, this topology automatically makes $\mathbb{D}^c$ into a locally convex space. It is well known [32] that this space is complete and separable.

Let $w \in \mathbb{D}_p^c$ and let $\{w_i\}_{i \in \mathbb{N}}$ be a sequence in $\mathbb{D}_p^c$ such that $||w_i - w|| \to 0$ as $i \to \infty$. Then $f_i(t) := \exp[\vartheta(w_i(T+t) - w_i(t))]$ converges uniformly to $f(t) := \exp[\vartheta(w(T+t) - w(t))]$. Thus $\mathcal{E}_{\vartheta,T}(w_i) = \int_0^p f_i(t)\,dt$ converges to $\mathcal{E}_{\vartheta,T}(w) = \int_0^p f(t)\,dt$ and so $\mathcal{E}_{\vartheta,T}$ is continuous. In a similar manner we may also show that $\mathcal{Q}^{\max}$ is continuous. Since C is a level set of $\mathcal{Q}^{\max}$, it follows that C is closed in the uniform topology. Since $\mathbb{D}_p^c$ is complete and C is closed, C is complete. A useful theorem for establishing compactness in this space is the following.

Theorem 21 (Arzela-Ascoli) *If X is a compact metric space, then a subset of $C(X, \mathbb{R})$, the set of continuous real valued functions on X, is precompact if and only if it is uniformly bounded and equicontinuous.*

Because of the peak rate constraint C is equicontinuous; indeed, it is Lipschitz with constant ρ. It is also uniformly bounded: $w(t) \leq \rho p$ for all $w \in \mathbb{D}_p^c$ and $t \in [0, p)$. Applying the Arzela-Ascoli theorem, C is therefore precompact. Since it is also closed, it is compact and we can immediately use the Krein-Milman theorem to conclude that $C = \overline{\text{conv}}\xi$, where ξ is the set of extreme points of C. Moreover, we have that $\sup_{w \in \xi} f(w) = \sup_{w \in C} f(w)$ for all continuous convex functionals f on $\mathbb{D}_p^c$.

The Topology of Weak Convergence

To make the constraint set compact in the above topology, we needed to restrict our attention to the case where there is a peak rate constraint. However, such a constraint is not fundamental to the optimisation problem, which is just as meaningful without a peak rate constraint. We will define another topology on $\mathbb{D}_p$ that will allow us to drop this restriction. This topology is motivated by the interpretation of $\mathbb{D}^+$ discussed earlier. Recall that $\mathbb{D}_p$ can be thought of as the space of distribution functions of finite signed

measures on $[0,p)$. The most commonly used topology on this space is the topology of weak convergence.

Consider the set $C[0,p)$ of continuous functions on $[0,p)$. Endow this space with the uniform topology. Consider the dual of this space, that is the set of continuous real-valued linear functionals on the this space. The Riesz representation theorem states that this dual space is the same as the space of finite signed measures on $[0,p)$.

Theorem 22 (Riesz) *If K is compact then every continuous linear functional on the linear space $C[K]$ can be represented as a signed measure on the measurable space $(K, \mathcal{B}(K))$.*

The interval $[0,p)$ may be regarded as compact if we identify the points 0 and p.

Definition 8 *The* weak* *topology on the dual space of $C[0,p)$ is the coarsest topology under which all the elements of $C[0,p)$, regarded as linear functionals, are continuous.*

The weak* topology is metrisable since the space of continuous functions on $[0,p)$ is separable [32].

An important theorem concerning weak convergence is the following characterisation of compact sets.

Theorem 23 (Helly) *A set of measures is precompact in the topology of weak convergence if and only if the total variation of the measures is bounded.*

The following characterisation of convergent sequences is also useful.

Theorem 24 *The sequence $\{w_i\}_{i\in\mathbb{N}}$ in $\mathbb{D}_p^+$ converges to $w \in \mathbb{D}_p^+$ if and only if $w_i(t) \to w(t)$ for each $t \in [0,p)$ at which w is continuous.*

We shall now prove the continuity of the effective bandwidth functional on $\mathbb{D}_p^+$ in this topology.

Theorem 25 *The functional $\mathcal{E}_{\vartheta,T}$ is continuous in the weak topology.*

Proof. Note that the topology of weak convergence is a metric topology. Let $\{w_i\}_{i\in\mathbb{N}}$ be a sequence of paths in $\mathbb{D}_p^+$ that converges in the weak topology to some $w \in \mathbb{D}_p^+$. Then $w_i(x) \to w(x)$ converges at every point x at which w is continuous. Consider the sequence of functions

$$f_i : [0,p) \to \mathbb{R} : x \to e^{\vartheta[w_i(x+T)-w_i(x)]}.$$

Then $f_i(x)$ converges to $f(x) = \exp(\vartheta[w(x+T) - w_i(x)])$ at every point x such that w is continuous at both x and $x+T$. Since w may be discontinuous at no more than a countable number of points, it follows that $f_i \to f$ almost everywhere. Each f_i is bounded above by $\exp[\vartheta w_i(p^-)]$ which converges to $\exp[\vartheta w(p^-)]$. This means that for i large enough the functions f_i are uniformly bounded and we can apply the Bounded Convergence Theorem to conclude that

$$\int_0^p f_i(x)\,dx \longrightarrow \int_0^p f(x)\,dx.$$

Thus $\mathcal{E}_{\vartheta,T}(w_i) \to \mathcal{E}_{\vartheta,T}(w)$, and so this functional is continuous. □

An interesting fact, although one we do not use, is that the weak topology is compatible with the σ-algebra we have been using, in the sense that the Borel σ-algebra $\mathcal{B}(\mathbb{D}^+)$ that it generates is also the σ-algebra generated by the coordinate functions.

Theorem 26 *$\mathcal{B}(\mathbb{D}^+) = \sigma(\{X_t\}_{0\le t\le p})$, where $X_t(w) := w(t)$ for $w \in \mathbb{D}_p$.*

Proof. In the weak topology the coordinate functions are upper-semicontinuous. They are thus measurable. Therefore $\sigma(\{X_t\}_{t\ge 0}) \subset \mathcal{B}(\mathbb{D}^+)$.

Conversely, consider the projection mapping $\phi : \mathbb{D}^+ \to \mathbb{R}^{\mathbb{Q}}$ where $(\phi w)(q) := w(q)$ for each $q \in \mathbb{Q}$. This mapping is a bijection between $\mathbb{D}^+$ and $\phi(\mathbb{D}^+)$. It seems natural to consider the topology $\tau := \{\phi(G) : G \text{ is weakly open}\}$ on $\phi(\mathbb{D}^+)$. On $\mathbb{D}^+$ the weak topology is coarser than that relativised from the product topology. Therefore τ is coarser than $\{\phi(G) : G \text{ is open in the product topology}\}$ which is the topology on $\phi(\mathbb{D}^+)$ relativised from the product topology on $\mathbb{R}^{\mathbb{Q}}$. Let $G \subset \mathbb{D}^+$ be open in the weak topology. Then G is open in the product topology. So $\phi(G)$ is open in the product topology on $\mathbb{R}^{\mathbb{Q}}$ and hence is a measurable set of the σ-algebra $\mathcal{B}(\mathbb{R}^{\mathbb{Q}})$. But $\mathbb{R}^{\mathbb{Q}}$ is a countable product of separable terms and therefore $\mathcal{B}(\mathbb{R}^{\mathbb{Q}}) = \mathcal{B}(\mathbb{R})^{\mathbb{Q}}$. Thus G is measurable $\mathcal{B}(\mathbb{R})^{\mathbb{Q}}$ which is the σ-algebra generated by the rational coordinate functions. It follows that G is measurable in the σ-algebra generated by all the coordinate functions. □

4.5 The Extreme Points of C_p

Recall that a point q is said to be an extreme point of a set H if $q \in H$ and q is not an interior point of any line segment in H. For $a \in \mathbb{D}^+$, $t_1, t_2 \in \mathbb{R}$, define $w(t_1, t_2) :=$

$a(t_2) - a(t_1) - (t_2 - t_1)\sigma$. Then $a \in C_p$ if and only if $w(t_1, t_2) \leq \beta$ for all $t_1, t_2 \in \mathbb{R}$. For $t \in \mathbb{R}$, we define $Q_a(t) := \sup_{t_1 \leq t} w(t_1, t)$ and $R_a(t) := \sup_{t_1 \geq t} w(t, t_1)$. We will now characterize the set of extreme points of the set C_p.

Theorem 27 *A point $a \in \mathbb{D}_p^+$ is an extreme point of C_p iff for almost all points $t \in [0, p)$ one of the following hold*

(X1) a is differentiable at t and $a'(t) = \rho$,

(X2) a is differentiable at t and $a'(t) = 0$,

(X3) $Q_a(t) = \beta$,

(X4) $R_a(t) = \beta$.

Proof.

a is extreme $\Rightarrow$ condition holds. Assume that there is a set $S \subset [0, p)$ such that none of the four conditions (X1)–(X4) hold at any point in S and that $\text{Leb}(S) > 0$. We shall construct two paths $\overline{a}$ and $\underline{a}$ that satisfy the peak rate and leaky bucket constraints such that $a = (\overline{a} + \underline{a})/2$.

Since Q_a and R_a are continuous, the set $G := \{t \in (0, p) : Q_a(t) < \beta, R_a(t) < \beta\}$ is open. It is a property of the real numbers that every open set can be expressed as a countable disjoint union of open intervals. Let $G = \bigcup_{n \in \mathbb{N}} G_n$ where each G_n is an open interval. Since $\bigcup_{n \in \mathbb{N}} G_n \cap S \supset S$, there exists $N \in \mathbb{N}$ such that $\text{Leb}(G_N \cap S) > 0$. In what follows it will be convenient to work on a compact interval. It is obvious that there exists a compact interval $K \subset G_N$ such that $\text{Leb}(K \cap S) > 0$. Furthermore, since the Lebesgue measure is non-atomic, we can find compact intervals $K_1, K_2 \subset K$, such that $\text{Leb}(K_1 \cap S) > 0$ and $\text{Leb}(K_2 \cap S) > 0$. Since K is compact and Q_a and R_a are continuous, these two functions attain their supremum over K. Therefore $T := \sup_{t \in K} Q_a(t) < \beta$ and $B := \sup_{t \in K} R_a(t) < \beta$.

We shall perturb a in the following way. For each $\delta \in (0, 1)$ define the following two maps from $[0, \rho] \to [0, \rho]$:

$$\overline{z}_\delta(x) := \begin{cases} (1 + \delta)x, & x \leq \rho/2 \\ (1 - \delta)x + \delta\rho, & x > \rho/2 \end{cases},$$

$$\underline{z}_\delta(x) := \begin{cases} (1-\delta)x, & x \le \rho/2 \\ (1+\delta)x - \delta\rho, & x > \rho/2 \end{cases}.$$

Note $\overline{z}_\delta(x) \ge x$ and $\underline{z}_\delta(x) \le x$ for all $x \in [0,\rho]$ and that $(\overline{z} + \underline{z})(x) = x$. Both maps leave 0 and ρ unchanged. Let

$$k_1(\delta) := \int_{K_1} ((\overline{z}_\delta - \mathrm{Id}) \circ a')(x)\, dx = \int_{K_1} ((\mathrm{Id} - \underline{z}_\delta) \circ a')(x)\, dx$$

and

$$k_2(\delta) := \int_{K_2} ((\overline{z}_\delta - \mathrm{Id}) \circ a')(x)\, dx = \int_{K_2} ((\mathrm{Id} - \underline{z}_\delta) \circ a')(x)\, dx.$$

Clearly both of these functions are continuous and strictly increasing, and therefore they are both invertible. Choose $\epsilon > 0$ such that $0 < \epsilon < \min(\beta - T, \beta - B, k_1(1), k_2(1))$. Then we can find $\delta_1, \delta_2 \in (0,1)$ such that $k_1(\delta_1) = k_2(\delta_2) = \epsilon$.

Now consider

$$\overline{a}_\epsilon(t) := \int_0^t \left(a' \, \mathrm{I}_{K_1' \cap K_2'} + (\overline{z}_{\delta_1} \circ a') \, \mathrm{I}_{K_1} + (\underline{z}_{\delta_2} \circ a') \, \mathrm{I}_{K_2} \right)(x)\, dx$$

and

$$\underline{a}_\epsilon(t) := \int_0^t \left(a' \, \mathrm{I}_{K_1' \cap K_2'} + (\underline{z}_{\delta_1} \circ a') \, \mathrm{I}_{K_1} + (\overline{z}_{\delta_2} \circ a') \, \mathrm{I}_{K_2} \right)(x)\, dx,$$

where δ_1 and δ_2 are such that $k_1(\delta_1) = k_2(\delta_2) = \epsilon$. Clearly $a = (\overline{a}_\epsilon + \underline{a}_\epsilon)/2$. Since $\overline{z}_\delta(x) > x$ for all $0 < x < \rho$ and K_1 and K_2 each contain a set of positive measure in which $0 < a' < \rho$, we have that $\overline{a}_\epsilon > a$ for all $\epsilon > 0$. Similarly $\underline{a} < a$ and thus a, $\overline{a}_\epsilon$, and $\underline{a}_\epsilon$ are all distinct for each ϵ. Therefore, a is in the interior of the line segment $[\overline{a}_\epsilon, \underline{a}_\epsilon]$. We will show that both $\overline{a}_\epsilon$ and $\underline{a}_\epsilon$ are in C_p if ϵ is small enough.

Since neither $\overline{z}$ nor $\underline{z}$ take values outside $[0,\rho]$, we have that $0 \le \overline{z}(a'(x)) \le \rho$ for almost all $x \in \mathbb{R}^+$. We conclude that both $\overline{a}$ and $\underline{a}$ are non-decreasing and satisfy the peak rate constraint: $\overline{a}(z) \le \overline{a}(y) + \rho(z-y)$ and $\underline{a}(z) \le \underline{a}(y) + \rho(z-y)$ for $y \le z$.

Let $t_1, t_2 \in \mathbb{R}$. If $(t_2 \bmod p) \notin K$, then $w_{\overline{a}}(t_1,t_2) \le w(t_1,t_2) \le \beta$. Otherwise $(t_2 \bmod p) \in K$ and so $w(t_1,t_2) \le T$. Therefore $w_{\overline{a}}(t_1,t_2) \le w(t_1,t_2) + \epsilon \le T + \epsilon < \beta$. Thus $\overline{a}$ obeys the leaky bucket constraint.

Similarly, if $(t_1 \bmod p) \notin K$ then $w_{\underline{a}}(t_1,t_2) \le w(t_1,t_2) \le \beta$, while if $t_1 \in K$, then $w(t_1,t_2) \le B$. Thus $w_{\underline{a}}(t_1,t_2) \le w(t_1,t_2) + \epsilon \le B + \epsilon < \beta$ for $t_1 \in K$. Therefore $\underline{a}$ also obeys the leaky bucket constraint.

condition holds $\Rightarrow$ *a is extreme.* Suppose that a obeys at least one of (X1)–(X4) at almost every time t and that $a = (\overline{a} + \underline{a})/2$ for two traffic paths $\overline{a}$ and $\underline{a}$, each of which obeys the leaky bucket constraints. Since a, $\overline{a}$, and $\underline{a}$ are nondecreasing, the derivatives of all three exist almost everywhere. Let $t \in [0, p)$ be such that a obeys one of the conditions (X1)–(X4) at t, and $\overline{a}$ and $\underline{a}$ are both differentiable at t.

- If $a'(t) = 0$ then $\overline{a}'(t) = \underline{a}'(t) = 0$, since $a'(t) = [\overline{a}'(t) + \underline{a}'(t)]/2$ and $0 \leq \overline{a}'(t) \leq \rho$ and $0 \leq \underline{a}'(t) \leq \rho$.
- Similarly, if $a'(t) = \rho$ then $\overline{a}'(t) = \underline{a}'(t) = \rho$.
- If $Q_a(t) = \beta$ then $Q_{\overline{a}}(t) = Q_{\underline{a}}(t) = \beta$ since $Q_a(t) \leq (Q_{\overline{a}}(t) + Q_{\underline{a}}(t))/2$ and both $Q_{\overline{a}}(t)$ and $Q_{\underline{a}}(t)$ are less than β.
- Similarly, if $R_a(t) = \beta$ then $R_{\overline{a}}(t) = R_{\underline{a}}(t) = \beta$

We have shown that if a obeys any one of the four conditions at t, then $\overline{a}$ and $\underline{a}$ also obey the same condition at t. Since either of the last two conditions imply that the derivative is the leaky bucket service rate σ, we conclude that $a' = \overline{a}' = \underline{a}'$ almost everywhere. Applying the Fundamental Theorem of Calculus and using the absolute continuity of a, $\overline{a}$, and $\underline{a}$, we find that $a = \overline{a} = \underline{a}$. Thus, a is not in the interior of any line segment contained in C_p and is therefore an extreme point of C_p. □

Note that if $a \in C_p$ obeys either (X3) or (X4) at an instant t and a is differentiable there, then $a'(t) = \sigma$. Thus, at any instant t the source may only transmit at one of three rates: 0, the peak rate ρ, or the leaky bucket service rate σ. The source can transmit at rate 0 or ρ at any time, but may transmit at rate σ only when the leaky bucket buffer is full, or is empty and will fill before any service goes unused. Figure 4.1 shows a typical extreme point of C_p. The figures 0 and β above the diagram indicate times when the leaky bucket is full and when it is empty.

We see that C_p has many extreme points. In fact we will show that the extreme points form a dense subset of C_p in the uniform topology. Since this is a finer topology than the weak topology, the result holds in the weak topology also. Heuristically, any source behaviour can be approximated by the source transmitting at rates 0 and ρ. All that is

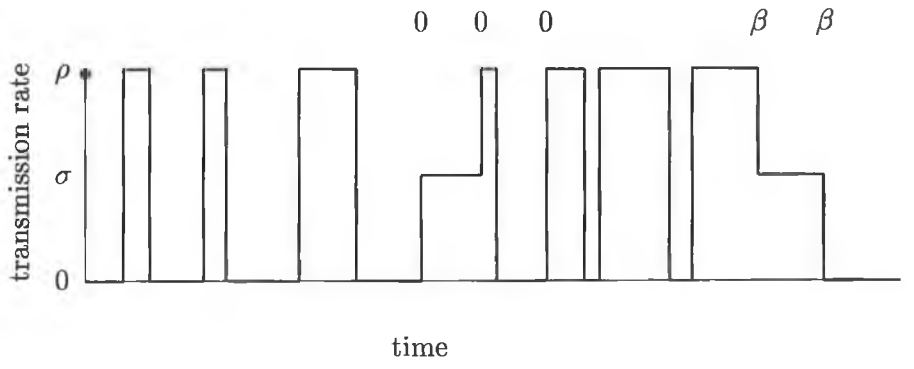

Figure 4.1: A typical extreme point of C. The values above the figure give the content of the leaky bucket at that time.

needed is for the source to switch between these two rates sufficiently quickly and to spend an appropriate proportion of time in each state. This result means that we do not need to use the convexity of the effective bandwidth functional to show that this functional approaches its supremum on ξ_p, all we need is its continuity. Unfortunately, we have not yet reduced the domain of our optimisation significantly. In the next chapter we will see how, in another linear structure on the space of traffic processes, convexity is crucial and enables us to make a more meaningful reduction in the domain of our optimisation problem.

Theorem 28 *The constraint set C_p is the closure of the set of its extreme points ξ_p in the uniform topology.*

Proof. Let $a \in C_p$ and let $\epsilon > 0$ be given. Choose $\eta < \epsilon/(2a(p))$ and define

$$b := (1-\eta)a.$$

Since a satisfies the leaky bucket constraint (σ, β), we have that b satisfies the leaky bucket constraint $((1-\eta)\sigma, (1-\eta)\beta)$. The maximum queue length functional is continuous in the uniform topology and so we can find $\zeta > 0$ such that $|\mathcal{Q}_\sigma^{\max} a - \mathcal{Q}_\sigma^{\max} b| < \eta\beta$ whenever $\|a - b\| < \zeta$. Choose $\kappa < \min(\epsilon/2, \zeta)$. Define the sequences of times $\{x_n\}$ and $\{y_n\}$ recursively to be

$$x_0 := 0$$

$$\begin{aligned} y_n &:= \inf\{t > x_n : b(t) \le b(x_n) + (t - x_n)\rho - \kappa\}, \qquad n \ge 0 \\ x_n &:= \inf\{t > y_{n-1} : b(t) \ge b(y_{n-1}) + \kappa\}, \qquad n \ge 1. \end{aligned}$$

Since a is non-decreasing and satisfies the peak rate constraint, we have that $y_n - x_n \ge \kappa/\rho$ and $x_{n+1} - y_n \ge \kappa/\rho$ for all $n \in \mathbb{N}$. Thus for any $t \in [0, \rho)$, there will be an x_n such that $x_n > t$. Define the path $c \in \mathbb{D}^+$ to be

$$c(t) := \int_0^t \rho \, \mathrm{I}_{\bigcup_n [x_n, y_n]}(x)\, dx, \qquad \text{for all } t \in [0, p).$$

Since b is continuous we have that $b(x_n) = c(y_{n-1}) + \kappa$ and that $b(y_n) = c(x_n) + (t - x)\rho - \kappa$. We have that $|b(t) - c(t)| \le \kappa$ for all $t \in [0, p)$. It follows that

$$\begin{aligned} \mathcal{Q}_\sigma^{\max} c &\le \mathcal{Q}_\sigma^{\max} b + \eta\beta \\ &\le \mathcal{Q}_{\sigma(1-\eta)}^{\max} b + \eta\beta \\ &\le (1 - \eta)\beta + \eta\beta \\ &= \beta. \end{aligned}$$

Thus c obeys the leaky bucket constraint (σ, β). Since c also clearly obeys the peak rate constraint and the conditions of Theorem 27, we conclude that c is in ξ_p. Now, $\|a - b\| = \eta a(p)$ and $\|b - c\| \le \kappa$, and so

$$\|a - c\| \le \eta a(p) + \kappa < \epsilon.$$

We have shown that any path in C_p may be approximated arbitrarily closely in the uniform topology by a path in ξ_p. □

Chapter 5

An Alternative Linear Structure

In this chapter we consider another way of representing traffic streams. Here the time taken for a certain amount of traffic to arrive is specified rather than the amount of traffic that arrives within an interval. This representation leads to a linear structure on the set of arrival paths which is different to that considered in the previous chapter. It turns out that all of the performance measures we have been considering are convex in this linear structure also. Moreover, the set of extreme points of the constraint set in this linear structure is strictly smaller. This enables us to derive a stronger result, restricting even further the set over which the optimisation must be performed.

5.1 The Space of Adjoints

Let a be any right continuous, non-decreasing function $\mathbb{R}^+ \to \mathbb{R}^+$. We define the *adjoint* of a to be

$$(\triangleleft a)(c) := \inf\{t \in \mathbb{R}^+ : a(t) \geq c\}.$$

Then $\triangleleft a$ will belong to the class of non-decreasing, left continuous, functions that have left limit 0 at $c = 0$. We denote the set of such functions by $\mathbb{B}^+$. Note that $\triangleleft$ is a bijection between $\mathbb{D}^+$ and $\mathbb{B}^+$. The inverse mapping is

$$(\triangleright d)(t) := \sup\{c \in \mathbb{R}^+ : d(c) \leq t\}$$

If either of a or $\triangleleft a$ is strictly increasing and continuous then the other is also and $\triangleleft a = a^{-1}$.

In the previous chapter we exploited the periodicity of the realisations. We may do the same here. If a is periodic with period p then $\triangleleft a$ is periodic also with period $q := a(p^-)$. For each $q > 0$, let $\mathbb{B}_q^+$ be the subspace of periodic paths with period q.

We define a linear structure on $\mathbb{B}_q := \mathbb{B}_q^+ - \mathbb{B}_q^+$ in the natural way:

$$\begin{aligned} (d_1 + d_2)(c) &:= d_1(c) + d_2(c) \\ \text{and} \qquad (\lambda d)(c) &:= \lambda d(c), \end{aligned}$$

for any $d_1, d_2 \in \mathbb{B}_q^+$ and $\lambda \in \mathbb{R}$.

We use the bijection $\triangleright$ to define the performance functionals on $\mathbb{B}_q^+$. For example $\mathcal{E}_{\vartheta,T}(d) := \mathcal{E}_{\vartheta,T}(\triangleright d)$ for $d \in \mathbb{B}_q^+$. Although we use the same symbol to represent functions on different spaces, there should be no confusion. We define the constraint set in $\mathbb{B}_q^+$ to be $F_q := \{d \in \mathbb{B}_q^+ : \triangleright d \in C_{d(q^-)}\}$.

5.2 Convexity

We conjecture that all of the functionals defined in Section 2.8 are convex in this linear structure also. However, while establishing convexity was trivial in the linear structure of the previous chapter, here it is considerably more difficult. We will remain content to prove convexity of the average queue length of a single source which is still relatively trivial, and of the effective bandwidth which is our main concern.

Single Source, Average Queue Length

Little's law [33] states that the average queue length over time Q and the average delay per customer D are related by $Q = Dm$, where m is the average arrival rate. (Here customers are infinitesimal amounts of fluid.) Since the service rate s is constant, the delay experienced by the cth piece of fluid is $1/s$ times the length of the queue when the piece of fluid arrives. Thus

$$D_d(c) := \sup_{c' \le c}[(c - c')/s - d(c) - d(c')].$$

This expression is clearly convex in d for each $c \in \mathbb{R}$.

We restrict our attention to the set $\{d \in F : d(q) = q/\sigma\}$. On this set $\mathcal{Q}^{\text{ave}}$ is clearly convex.

As the convexity proof for the effective bandwidth functional relies on a topological argument, we will discuss the choice of topology on $\mathbb{B}$ first.

5.3 Topologies

We will not discuss the topology of uniform convergence in this linear structure. Since discontinuities of the elements of $\mathbb{B}^+$ correspond to times when the source is silent, the paths may be discontinuous even in the presence of a peak rate constraint. For example, the sample paths of the process that we have conjectured to be the worst case are discontinuous. Thus we can not expect the constraint set to be compact in this topology. We therefore concentrate our attention on the topology of weak convergence. Again we interpret the elements of ${\mathbb{B}^+}_q$ as finite signed measures on $(0, q]$.

Consider $D_{q,p} := \{d \in \mathbb{B}_q^+ : d(q) = p\}$, for some $p \geq 0$. The elements of $D_{q,p}$ may also be considered to be members of $A_{p,q} := \{a \in \mathbb{D}_p^+ : a(p^-) = q\}$ through the embedding $d \leftrightarrow \triangleright d$. This mapping is a bijection between $D_{q,p}$ and $A_{p,q}$. In fact, if we use the weak topology on both sets then this bijection is a homeomorphism.

Theorem 29 *The bijection $a \leftrightarrow \triangleleft a$ is a homeomorphism between $A_{p,q}$ and $D_{q,p}$.*

Proof. Let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of points in $A_{p,q}$ which converges to $a \in A_q$. Let $d := \triangleleft a$. Then $a_n(t) \to a(t)$ for all $t \in [0, p)$ at which a is continuous. Let $c^* \in [0, q)$ be such that d is continuous at c^* and let $\epsilon > 0$ be given. Define $t^* := d(c^*)$. Since d is continuous at c^*, we have that a is strictly increasing at $d(c^*)$. Also $a(t) = \sup\{c \in \mathbb{R} : d(c) \leq t\}$ and a in non-decreasing, and so we conclude that $a(t) > c^*$ whenever $t > t^*$. Since a may only be discontinuous at a countable set of points, we may choose δ_1 in the range $0 < \delta_1 < \epsilon$ such that a is continuous at $t^* + \delta_1$. For n large enough,

$$|a_n(t^* + \delta_1) - a(t^* + \delta_1)| < a(t^* + \delta_1) - c^*.$$

Therefore $a_n(t^* + \delta_1) > c^*$ and so $d_n(c^*) := \inf\{c \in \mathbb{R} : a_n(c) \geq c^*\} \leq t^* + \delta_1 = d(c^*) + \delta_1 < d(c^*) + \epsilon$, for n large enough.

In a similar manner we may choose δ_2 such that $0 < \delta_2 < \epsilon$ and a is continuous at $t^* - \delta_2$. Again, for n large enough, $a_n(t^* - \delta_2) < c^*$ and so $d_n(c^*) \geq t^* - \delta_2 = d(c^*) - \delta_2 > d(c^*) - \epsilon$. We have thus shown that, for n large enough, $|d_n(c^*) - d(c^*)| < \epsilon$ for arbitrary $\epsilon > 0$; in

other words that $d_n(c^*) \to d(c^*)$ as $n \to \infty$. Since c^* was any point of continuity of d, the criterion for convergence is satisfied and $d_n \to d$.

That $d_n \to d$ implies $a_n \to a$ may be shown in a similar manner. The two spaces are thus homeomorphic. □

We have an immediate corollary.

Corollary 3 *The functionals $\mathcal{Q}^{ave}$, $\mathcal{L}$, and $\mathcal{E}_{\vartheta,T}$ defined above are continuous in the weak topology on $D_{q,p}$.*

The constraint set F_q is not compact in the topology of weak convergence. To see this, one need only consider the sequence $d_n(c) := nc\sigma$ which is in F_q and is unbounded. However, the isotonicity of each of our functionals enables us to restrict our attention to the subset $D_{q,q/\sigma} := \{d \in F_q : d(q) = q/\sigma\}$. This subset is closed since F_q is closed and the function $d \to d(q)$ is continuous. It is also precompact by the Helly-Bray theorem and therefore compact since $\mathbb{B}_q$ is complete. We may now apply the Krein-Milman theorem to conclude that $D_{q,q/\sigma}$ is equal to the closure of the convex hull of its extreme points. We have already seen that $\mathcal{E}_{\vartheta,T}$ is continuous on $\mathbb{B}_q$. To reduce to optimisation problem to an optimisation over the extreme points of $D_{q,q/\sigma}$, we need only show that $\mathcal{E}_{\vartheta,T}$ is convex on $D_{q,q/\sigma}$.

5.4 Convexity of the Effective Bandwidth

In this section we prove the convexity of the effective bandwidth functional in the alternative linear structure. We restrict our attention to the set $D_{q,q/\sigma}$. Recall that

$$\mathcal{E}_{\vartheta,T}(d) := \frac{1}{p}\int_0^p e^{\vartheta[a(t+T)-a(t)]}\,dt,$$

where $p := q/\sigma$ and $a = \triangleright d$. We will first prove the convexity of $\mathcal{E}_{\vartheta,T}$ in the space of *simple* realisations and then extend the result to $D_{q,q/\sigma}$ using a continuity argument.

Simple Paths

By a simple path, we mean one that is piecewise constant, in other words one which has a finite number of discontinuities but is otherwise constant. We denote by $A_{q,p}$ the set of

elements of $D_{q,p}$ which have this property. Since the elements of $D_{q,p}$ are non-decreasing, the discontinuities of any element of $A_{q,p}$ must be simple.

Note that the adjoint of a simple path is also simple; the intervals on which the path is constant become the discontinuities of the adjoint and *vice versa*.

For any $\alpha \in (0,1)$ and $d_1, d_2 \in A_{q,p}$, the convex combination $z_\alpha := (1-\alpha)d_1 + \alpha d_2$ is also in $A_{q,p}$. If the set of discontinuity points of d_1 is $\{a_n\}_{1\leq n\leq N}$ and the set of discontinuity points of d_2 is $\{b_n\}_{1\leq n\leq M}$, then the set of discontinuity points of z_α is the union of these two sets. The size of the jump in z_α at c is $\alpha m_{d_2} + (1-\alpha)m_{d_1}$, where m_{d_1} and m_{d_2} are the size of the jumps, if any, in d_1 and d_2 at c. Thus the $A_{q,p}$ is a convex subset of $D_{q,p}$.

Simple paths are useful because in the topology of weak convergence any path may be approximated by simple paths. In other words, for every $d \in \mathbb{B}^+$ there exists a sequence of simple paths $\{a_n\}_{n\in\mathbb{N}}$ such that $a_n \to d$.

Lemma 12 *The set $A_{q,p}$ is dense in $D_{q,p}$ for all $p, q > 0$.*

Proof. Let $d \in D_{q,p}$ Define the sequence of simple paths

$$d_n(c) := p\lfloor nd(c)/p \rfloor / n$$

for $n \in \mathbb{N}$. Clearly, $d_n \in D_{q,p}$. Now $||d - d_n|| \leq p/n$ and so $d_n \to d$ in the uniform topology on $D_{q,p}$. Since the uniform topology is finer than the weak topology, we have that $d_n \to d$ in the weak topology also. □

Theorem 30 *The functional $\mathcal{E}_{\vartheta,T}$ is convex on $A_{q,p}$.*

Proof. Write $g(x) := \exp(\vartheta x)$. The only property of g we will use is its convexity. To show that $\mathcal{E}_{\vartheta,T}$ is convex on $A_{q,p}$, it will suffice to show that the one-dimensional function $f(\alpha) : [0,1] \to \mathbb{R} := \mathcal{E}_{\vartheta,T}((1-\alpha)d_1 + \alpha d_2)$ is convex for any two points d_1 and d_2 in $A_{q,p}$.

Let $d_1, d_2 \in A_{q,p}$. Then $\alpha d_2 + (1-\alpha)d_1$ has discontinuities at $\{x_n\}_{1\leq n\leq N}$, a set which is independent of α. Define $m_n := x_{n+1} - x_n$ and $t_n(\alpha) := \alpha d_2(x_n^+) + (1-\alpha)d_1(x_n^+)$. Let $v_n := d_2(x_n^+) - d_1(x_n^+)$.

Denote by J_α the set of ordered pairs (i,j) such that $t_j(\alpha) - t_i(\alpha) = T$ and $v_i \neq v_j$. There are only a finite number of values of α for which J_α is nonempty. For any other

value of α, let δ be such that

$$2\max_c |v_c\delta| < \min_{i,j} |t_j(\alpha) - t_i(\alpha) + T| \wedge \min_{i,j} |t_j(\alpha) - t_i(\alpha)|.$$

Let $z_c(\alpha) := a_\alpha(t_c+T^-) - a_\alpha(t_c)$ and $b_c(\alpha) := a_\alpha(t_c^-) - a_\alpha(t_c-T)$, where $a_\alpha := \triangleright(\alpha d_2 - (1-\alpha)d_1)$. Then

$$f(\alpha+\delta) - f(\alpha) = \sum_c v_c\delta\Big[g(z_c+m_c) - g(z_c) + g(b_c) - g(b_c+m_c)\Big],$$

suppressing the dependence on α for clarity. So f is piecewise affine and the changes in its slope occur when $(1-\alpha)d_1 + \alpha d_2$ has two flat pieces with a difference in height of T, moving at different rates. For each pair $(i,j) \in J_\alpha$, let $z_{ij}(\alpha) := z_i(\alpha) = b_j(\alpha)$. Then the magnitude of the change in the derivative of f at α is given by

$$\begin{aligned}\Delta f'(\alpha) &= \sum_{(i,j)\in J} v_i \operatorname{sgn}(v_i-v_j)\Big[g(z_{ij}+m_i+m_j) - g(z_{ij}+m_i) - g(z_{ij}+m_j) + g(z_{ij})\Big] \\ &\quad + \sum_{(i,j)\in J} v_j \operatorname{sgn}(v_i-v_j)\Big[g(z_{ij}+m_i) - g(z_{ij}) - g(z_{ij}+m_i+m_j) + g(z_{ij}+m_j)\Big] \\ &= \sum_{(i,j)\in J} |v_i-v_j|\Big[g(z_{ij}+m_i+m_j) - g(z_{ij}+m_i) - g(z_{ij}+m_j) + g(z_{ij})\Big],\end{aligned}$$

where $\operatorname{sgn}(x) := x/|x|$ is the signum function. From the convexity of g we have that

$$g(z_{ij}+m_i) \le \frac{m_i}{m_i+m_j} g(z_{ij}+m_i+m_j) + \frac{m_j}{m_i+m_j} g(z_{ij})$$

and

$$g(z_{ij}+m_j) \le \frac{m_j}{m_i+m_j} g(z_{ij}+m_i+m_j) + \frac{m_i}{m_i+m_j} g(z_{ij}).$$

Therefore $\Delta f'(\alpha) \ge 0$, and the convexity of f follows. □

Lemma 13 *Let S be a convex subset of a metrisable topological vector space and let $T \subset S$ be convex and dense in S. If $f : S \to \mathbb{R}$ is a continuous function and its restriction to T is convex, then f is also convex on S.*

Proof. Let $p, q \in S$ and let $\alpha \in [0,1]$. Then there exist sequences p_n and q_n in T such that $p_n \to p$ and $q_n \to q$. Consider the sequence $r_n := (1-\alpha)p_n + \alpha q_n$ which lies in T since this set is convex. The operations of addition and multiplication by a scalar are continuous, and so this sequence must approach the limit $r := (1-\alpha)p + \alpha q$. Now the continuity of f

implies that $f(p_n) \to f(p)$, that $f(q_n) \to f(q)$ and that $f(r_n) \to f(r)$. Furthermore, the convexity of f implies that $f(r_n) \leq (1-\alpha)f(p_n) + \alpha f(q_n)$ for all $n \in \mathbb{N}$. It follows that $f(r) \leq (1-\alpha)f(p) + \alpha f(q)$ and therefore f is convex on S. □

We can now extend the convexity result to the whole of $D_{q,p}$.

Theorem 31 *The functional $\mathcal{E}_{\vartheta,T}$ is convex on $D_{q,p}$.*

Proof. We have shown that $\mathcal{E}_{\vartheta,T}$ is continuous on $D_{q,p}$ and convex on a dense convex subset $A_{q,p}$ of $D_{q,p}$. The conditions of the previous lemma therefore hold and the conclusion follows. □

5.5 The Extreme Points of F

We will now characterise the set of extreme points of the set F in the linear structure described in this chapter. In contrast to the elements of C, the elements of F need not be Lipschitz or even continuous. However, they are nondecreasing and therefore differentiable almost everywhere. It can be shown [24] that every distribution function can be decomposed into a sum of a singular and an absolutely continuous function. We denote the decomposition of $d \in \mathbb{B}^+$ by $d = d_s + d_a$, where d_s is singular and d_a is absolutely continuous. In Section 2.1 we discussed the correspondence between distribution functions and measures. We denote by μ_d the measure corresponding to a distribution function d. We define $z(c_1, c_2) := c_2 - c_1 - (d(c_2) - d(c_1))\sigma$. Clearly, $d \in F$ if and only if $z(c_1, c_2) \leq \beta$ for all $c_1, c_2 \in \mathbb{R}$. For $d \in \mathbb{B}^+$ and $c \in [0, q)$, we write

$$\begin{aligned} Q_d(c) &:= \sup_{c_1 \in (-\infty, c)} z(c_1, c) \\ \text{and} \quad R_d(c) &:= \sup_{c_1 \in (c, \infty)} z(c, c_1). \end{aligned}$$

The support of a measure is the intersection of all closed subsets of full measure.

Lemma 14 *If d is an extreme point of F_q then one of the conditions*

(DX1) $d'(c) = 1/\rho$,

(DX2) $Q_d(c) = \beta$,

(DX3) $R_d(c) = \beta$.

hold for almost every $c \in [0, q)$.

Proof. Suppose that there is a subset S of $[0, q)$ of positive Lebesgue measure in which none of the conditions (DX1),(DX1), nor (DX3) hold at any point. We shall construct two paths $\overline{d}$ and $\underline{d}$ that satisfy the leaky bucket constraints such that $d = (\overline{d} + \underline{d})/2$.

Since Q_d and R_d are upper semicontinuous, the set $\{c \in [0, q) : Q_d(c) < \beta, R_d(c) < \beta\}$ is open and therefore is the countable disjoint union of open intervals. The intersection of a least one of these with S must have positive measure, for otherwise the countable union, which contains S, would have measure zero. Let G be this interval. In what follows it will be convenient to work on a compact interval. It is obvious that there exists a compact interval with the above property. Furthermore, since the Lebesgue measure is non-atomic, we can decompose the compact set into two intervals K_1 and K_2, each with positive measure. Let $U := \sup_{c \in K} Q_d(c)$ and $B := \sup_{c \in K} R_d(c)$. Clearly, $U < \beta$ and $B < \beta$ since K is compact and Q_d and R_d are upper semicontinuous.

Define the following two maps from $[1/\rho, \infty) \to [1/\rho, \infty)$:

$$\begin{aligned} \overline{z}_\delta(x) &:= (1+\delta)x - \delta/\rho \\ \text{and} \qquad \underline{z}_\delta(x) &:= (1-\delta)x + \delta/\rho. \end{aligned}$$

Note that $\overline{z} \geq \mathrm{Id}$ and $\underline{z} \leq \mathrm{Id}$, and that $\overline{z} + \underline{z} = \mathrm{Id}$. Both maps leave $1/\rho$ unchanged. Let

$$k_1(\delta) := \int_{K_1} ((\overline{z}_\delta - \mathrm{Id}) \circ d')(x)\, dx = \int_{K_1} ((\mathrm{Id} - \underline{z}_\delta) \circ d')(x)\, dx$$

$$k_2(\delta) := \int_{K_2} ((\overline{z}_\delta - \mathrm{Id}) \circ d')(x)\, dx = \int_{K_2} ((\mathrm{Id} - \underline{z}_\delta) \circ d')(x)\, dx.$$

Clearly, both of these functions are continuous and strictly increasing, and therefore they are both invertible. Choose $\epsilon > 0$ such that $\epsilon < \min(\beta - U, \beta - B, k_1(1), k_2(1))$. Then we can find $\delta_1, \delta_2 \in (0, 1)$ such that $k_1(\delta_1) = k_2(\delta_2) = \epsilon$.

Now consider

$$\begin{aligned} \overline{d}_\epsilon(t) &:= d_s(t) + \int_0^t \left(d' \, \mathrm{I}_{K_1' \cap K_2'} + (\underline{z}_{\delta_1} \circ d') \, \mathrm{I}_{K_1} + (\overline{z}_{\delta_2} \circ d') \, \mathrm{I}_{K_2} \right)(x)\, dx. \\ \text{and} \qquad \underline{d}_\epsilon(t) &:= d_s(t) + \int_0^t \left(d' \, \mathrm{I}_{K_1' \cap K_2'} + (\underline{z}_{\delta_2} \circ d') \, \mathrm{I}_{K_1} + (\overline{z}_{\delta_1} \circ d') \, \mathrm{I}_{K_2} \right)(x)\, dx. \end{aligned}$$

Clearly $d = (\overline{d}_\epsilon + \underline{d}_\epsilon)/2$. Since $\overline{z}(x) > x$ for all $x > 1/\rho$ and both K_1 and K_2 contain a set in which $d' > \rho$, we have that $\overline{d}_\epsilon > d$. Similarly $\underline{d} < d$ and thus d, $\overline{d}$, and $\underline{d}$ are distinct. Therefore, d is in the interior of the line segment $[\overline{d}, \underline{d}]$. We will show that both $\overline{d}$ and $\underline{d}$ are in F and therefore so also is the line segment.

Since neither $\overline{z}$ nor $\underline{z}$ take values below $1/\rho$, we have that $\overline{z} \circ d'(x) \geq 1/\rho$ for almost all $x \in [0, q)$. Furthermore, d_s is non-decreasing and we conclude that $\overline{d}$ and $\underline{d}$ are non-decreasing and satisfy the peak rate constraints $\overline{d}(z) \geq \overline{d}(y) + (z - y)/\rho$ and $\underline{d}(z) \geq \underline{d}(y) + (z - y)/\rho$ for $y \leq z$.

Let $c_1, c_2 \in \mathbb{R}$. If $(c_2 \bmod q) \notin K$ then $z_{\overline{d}}(c_1, c_2) \leq z(c_1, c_2) \leq \beta$. Otherwise $(c_2 \bmod q) \in K$ and so $z(c_1, c_2) \leq U$. Therefore $z_{\overline{d}}(c_1, c_2) \leq z(c_1, c_2) + \epsilon \leq U + \epsilon < \beta$. Thus $\overline{d}$ obeys the leaky bucket constraint.

Similarly, if $(c_1 \bmod q) \notin K$ then $z_{\overline{d}}(c_1, c_2) \leq z(c_1, c_2) \leq \beta$, while if $c_1 \in K$, then $z(c_1, c_2) \leq \beta - B$. Thus $z_{\overline{d}}(c_1, c_2) \leq z(c_1, c_2) + \epsilon \leq \beta - B + \epsilon < \beta$ for $c_1 \in K$. Therefore, in this case also, $\overline{d}$ obeys the leaky bucket constraint and the theorem is proved. □

Note that if d obeys (DX2) or (DX3) at c and d is differentiable there, then $d'(c) = 1/\sigma$.

Lemma 15 *If d is an extreme point of F_q then $d(q^-)/q = 1/\sigma$, in other words the mean arrival rate is the leaky bucket token rate.*

Proof. Define the function

$$u(c) := -\inf_{x \in [0,c]} [z(0, x) + Q_d(0)].$$

which is the unused service in the interval $[0, c)$. This function clearly has the memoryless property:

$$u(c_2) - u(c_1) = -\inf_{x \in [c_1, c_2]} [z(c_1, x) + Q_d(c_1)].$$

We will need the following two facts. Firstly, since d' equals either $1/\sigma$ or $1/\rho$ almost everywhere, $d_a(c_2) - d_a(c_1) \leq (c_2 - c_1)/\sigma$ for all $c_1, c_2 \in \mathbb{R}$. Therefore

$$\begin{aligned} u(c) &= -\inf_{c_1 \in [0,c]} [c - (d_a(c_1) + d_s(c_1))\sigma + Q_d(0)] \\ &\leq [d_a(c) + d_s(c)]\sigma \\ &\leq d_s(c)\sigma. \end{aligned}$$

We may use the memoryless property to conclude that $u(c_2) - u(c_1) \leq [d_s(c_2) - d_s(c_1)]\sigma$.

Secondly,

$$z(0,c) + u(c) = \sup_{c_1 \in [0,c]} [z(c_1,c) + Q_d(0)] = Q_d(c) \leq \beta.$$

We can again use the memoryless property to deduce that

$$z(c_1,c_2) \leq \beta - [u(c_2) - u(c_1)].$$

Suppose that $d(q^-) < q/\sigma$. Let $i := \lceil \beta/(q - \sigma d(q^-)) \rceil$. Then $Q_d(0) \geq z(-iq, 0) = (q - \sigma d(q))i \geq \beta$, and d is not in F.

Suppose now that $d(q^-) > q/\sigma$. Since $u[0,q) = q - \sigma d(q^-)$, the support of μ_u must be non-empty. In other words there is some $x \in [0,q)$ such that $\mu_u[N] > 0$ for every neighbourhood N of x.

If u is discontinuous at x then so also is d, and the size of the jump in d is greater than that in u, that is to say $\mu_d(\{x\}) \geq \mu_u(\{x\})$. We use the following construction. Let $\delta < \mu_u(\{x\})$. Define

$$\overline{d} := \begin{cases} d(c), & 0 \leq c < x \\ d(c) + \delta/\sigma, & x \leq c < q \end{cases}$$

$$\text{and} \qquad \underline{d} := \begin{cases} d(c), & 0 \leq c < x \\ d(c) - \delta/\sigma, & x \leq c < q. \end{cases}$$

Then $\overline{d}$ and $\underline{d}$ are in $\mathbb{B}_q^+$ and $\overline{d} + \underline{d} = d$. Clearly $z_{\overline{d}}(c_1,c_2) \leq z_d(c_1,c_2) \leq \beta$ for all $c_1, c_2 \in \mathbb{R}$ and so $\overline{d}$ obeys the leaky bucket constraint. Also $z_{\underline{d}}(c_1,c_2) = z_d(c_1,c_2)$ for $c_1, c_2 \in \mathbb{R}$ such that $x \notin (c_1,c_2)$. For $x \in (c_1,c_2)$, $z_{\underline{d}}(c_1,c_2) = z(c_1,c_2) + \delta \leq \beta - u(x) + \delta \leq \beta$, and so $\underline{d}$ also obeys the leaky bucket constraint.

Suppose now that d is continuous at x. Then we can find three intervals I_1, I_2, and I_3 each of positive measure with respect to μ_u, such that $\sup I_3 - \inf I_1 < \beta$. We use the following construction. Define the measure Δ on $[0,d)$ by $\Delta(A) := \mu_{d_s}(A \cap \mathrm{I}_{I_2})$ for any measurable set A. Choose $\delta < \min(\mu_u[I_1], \mu_u[I_3])/\mu_{d_s}[I_2]$. Let $\overline{d}(c) := d(c) + \delta\Delta[0,c)$ and let $\underline{d}(c) := d(c) - \delta\Delta[0,c)$. Clearly, $\overline{d}$, $\underline{d}$, and d are distinct since $\Delta \neq 0$. Again, $\overline{d}$ clearly obeys the leaky bucket constraint.

Let $c_1, c_2 \in \mathbb{R}$ be such that $0 \leq c_2 - c_1 \leq d$. There are three cases to consider. Firstly, if $c_1, c_2 \in (\inf I_1, \sup I_3)$, then $z_{\underline{d}}(c_1,c_2) \leq \sup I_3 - \inf I_1 < \beta$. Secondly, if $I_2 \cap (c_1,c_2) = \emptyset$

then $z_{\underline{d}}(c_1, c_2) = z(c_1, c_2) \leq \beta$. Finally, if $I_2 \cap (c_1, c_2) \neq \emptyset$ and $I_2' \cap (c_1, c_2) \neq \emptyset$, then either I_1 or I_3 is a subset of (c_1, c_2). Either way,

$$z_{\underline{d}}(c_1, c_2) \leq z(c_1, c_2) + \delta\mu_{d_s}[I_2] \leq \beta - \min(\mu_u[I_1], \mu_u[I_3]) + \delta\mu_{d_s}[I_2] \leq \beta.$$

Thus $\underline{d}$ obeys the leaky bucket constraint. □

Lemma 16 *If d is an extreme point of F_q, then for each pair of points c_1 and c_2 in the support of μ_{d_s}, either $\sup_{c\in(c_1,c_2)} Q_d(c) = \beta$ or $\sup_{c\in(c_1,c_2)} R_d(c) = \beta$.*

Proof. Suppose $d \in F$ is such that $0 < Q_d(c) < \beta$ for $c \in (c_1, c_2)$ for some $c_1, c_2 \in \mathbb{R}^+$ with $c_1 < c_2 < c_1 + q$. Let $U := \sup\{Q_d(c) : c \in (c_1, c_2)\}$ and $B := \sup\{R_d(c) : c \in (c_1, c_2)\}$. In the following manner we construct $\overline{d}, \underline{d} \in F$ such that $\overline{d}$, d, and $\underline{d}$ are distinct and $\overline{d} + \underline{d} = d$.

If d is continuous at c_1 then Q_d is also continuous at c_1, and $Q_d(c) \leq (U + \beta)/2$ for all c in some open interval I_1 about c_1. In this case define a measure Δ_1 on $[0, q)$ by $\Delta_1[A] := \mu_{d_s}[I_1 \cap A]$. If, on the other hand, d has a discontinuity at c_1 then it must be simple. Its magnitude is $\mu_{d_s}[\{c_1\}]$. In this case we define $\Delta_1[A] := \mu_{d_s}[A]\, \mathrm{I}_{\{c_1\}}$.

Similarly, if d is continuous at c_2 then we define $\Delta_2[A] := \mu_{d_s}[I_2 \cap A]$, where I_2 is an open set about c_2 for which $Q_d(c) \geq B/2$ when $c \in I_2$. Again, if d is discontinuous at c_2 then we define $\Delta_2[A] := \mu_{d_s}[A]\, \mathrm{I}_{\{c_2\}}$.

Let $k_1 := \Delta_1[0, q)$ and $k_2 := \Delta_2[0, q)$. Choose $\delta_1 > 0$ and $\delta_2 > 0$ such that $\delta_1 k_1 = \delta_2 k_2 < \min((\beta + U)/2, B/2)$. We now define

$$\begin{aligned} \overline{d}(c) &:= d(c) + \delta_1\Delta_1[0, c) - \delta_2\Delta_2[0, c) \\ \text{and} \qquad \underline{d}(c) &:= d(c) - \delta_1\Delta_1[0, c) + \delta_2\Delta_2[0, c). \end{aligned}$$

Let x, y be in $\mathbb{R}$. If y is in the region between c_1 and c_2, then $z_{\underline{d}}(x, y) \leq z(x, y) + \delta_1 k_1 \leq (U + \beta)/2 + \delta_1 k_1 < \beta$. If y is outside the region between c_1 and c_2, then $z_{\underline{d}}(x, y) \leq z(x, y) \leq \beta$. Therefore, $\underline{d}$ obeys the leaky bucket constraint.

Recall that $z(x', y') \leq \beta - Q_d(y')$. for any $x' \in \mathbb{R}^+$ and $y' \geq x'$. Therefore, if x is in the region between c_1 and c_2, then $z(x, y) \leq \beta - B/2$. Therefore, $z_{\overline{d}}(x, y) \leq z(x, y) + \delta_2 k_2 \leq \beta - B/2 + \delta_2 k_2 < \beta$. If y is outside the region between c_1 and c_2, then $z_{\underline{d}}(x, y) \leq z(x, y) \leq \beta$. Therefore, $\overline{d}$ obeys the leaky bucket constraint. □

The preceding three lemmas establish that each of three conditions hold if d is extreme. We shall now prove the converse.

Theorem 32 *$d \in F_q$ is an extreme point if and only if each of the following three conditions hold:*

(DX) for almost every $c \in [0, q)$, one of the following hold

(DX1) $d'(c) = 1/\rho$,

(DX2) $Q_d(c) = \beta$,

(DX3) $R_d(c) = \beta$.

(DY) $d(q^-)/q = 1/\sigma$,

(DZ) for each pair of points c_1 and c_2 in the support of μ_{d_s}, either $\sup_{c \in (c_1, c_2)} Q_d(c) = \beta$ or $\sup_{c \in (c_1, c_2)} R_d(c) = \sigma$.

Proof. Assume that d obeys the given conditions and that $d = (\overline{d} + \underline{d})/2$ for $\overline{d}, \underline{d} \in F_d$. Let $c \in \mathbb{R}^+$ be such that d obeys one of (DX1), (DX2), or (DX3) at c and $\overline{d}$ and $\underline{d}$ are both differentiable at c.

- If $d'(c) = 1/\rho$ then $\overline{d}'(c) = \underline{d}'(c) = 1/\rho$ since $d'(c) = (\overline{d}'(c) + \underline{d}'(c))/2$ and $\overline{d}'(c) \geq 1/\rho$ and $\underline{d}'(c) \geq 1/\rho$.
- If $Q_d(c) = \beta$ then, since $Q_d \leq (Q_{\overline{d}} + Q_{\underline{d}})/2$, we have $Q_{\overline{d}}(c) = Q_{\underline{d}}(c) = \beta$.
- Similarly, if $R_d(c) = \beta$ then, since $R_d \leq (Q_{\overline{d}} + Q_{\underline{d}})/2$, we have $R_{\underline{d}}(c) = R_{\overline{d}}(c) = \beta$.

We have proved that if d obeys one of the three sub-conditions of (DX) at c, then both $\overline{d}$ and $\underline{d}$ obey the same condition. Since either of the last two conditions imply that the derivative is the leaky bucket service rate σ, we conclude that $d' = \overline{d}' = \underline{d}'$ almost everywhere. Thus the absolutely continuous parts $\overline{d}_a$, $\underline{d}_a$, and d_a of d, $\overline{d}$, and $\underline{d}$, respectively, are equal.

To show that the singular parts are equal, note that $\overline{d}_s(c) = \underline{d}_s(c) = d_s(c)$ for each point c for which either $Q_d(c) = \beta$ or $R_d(c) = \beta$. But we have assumed that between every two points of the support of d_s there is such a point.

We have shown that d is not in the interior of any line segment contained in F and is therefore an extreme point of F. □

We will now show that the set of extreme points in this linear structure is strictly smaller than that of the previous chapter.

Theorem 33 *If $d \in \varphi_q$ then $\triangleright d \in \xi_p$ where $p := d(q^-)$.*

Proof. Let $a := \triangleright d$. Then $a \in C_p$ and so $a(t)$ is differentiable at almost all $t \in [0,p)$. Let $X := \{t \in [0,p) : a \text{ is differentiable at } t, \text{ and } a'(t) = 0\}$. Let Z be the set of $t \in [0,p)$ such that a is differentiable at t, and $a'(t) > 0$. Then $\text{Leb}X + \text{Leb}Z = p$. Let K be the set of $c \in [0,q)$ for which neither of (DX1), (DX1), nor (DX3) hold for d at c. Then K has measure zero. Let Y be $\{t \in Z : a(t) \notin K\}$. Since a is strictly increasing at every point in Z, for each c in $[0,q)$ there is at most one $t \in Z$ for which $a(t) = c$. Therefore $\text{Leb}Y = \text{Leb}Z$, and so $\text{Leb}X + \text{Leb}Y = p$. If $t \in X$ then $(X1)$ holds for a at t. If $t \in Y$ and $d'(a(t)) = 1/\rho$ then $a'(t) = \rho$. We define $Q_a(t) := \sup_{t_1 \le t}[a(t) - a(t_1) - (t - t_1)\sigma]$ and $R_a(t) := \sup_{t_1 \ge t}[a(t_1) - a(t) - (t_1 - t)\sigma]$ as in the previous chapter.

If $t \in Y$ and $Q_d(c) = \beta$ then $Q_a(t) = \beta$. If $t \in Y$ and $R_d(c) = \beta$ then $R_a(t) = \beta$. We have proved that $X \cup Y$ has measure p and that for each point t in this set, a obeys one of (X1),(X2), (X3), or (X4) at t. Thus a is an extreme point of C_p. □

For any $p > 0$, let $t_\rho := \min(\beta/2(\rho - \sigma), \sigma p/\rho)$. Then

$$z(t) := \begin{cases} \rho t, & 0 \le t < t_\rho \\ \rho t_\rho, & t_\rho \le t < p, \end{cases}$$

is an extreme point of C_p, but $\triangleleft z$ is not an extreme point of $F_{\rho t_\rho}$.

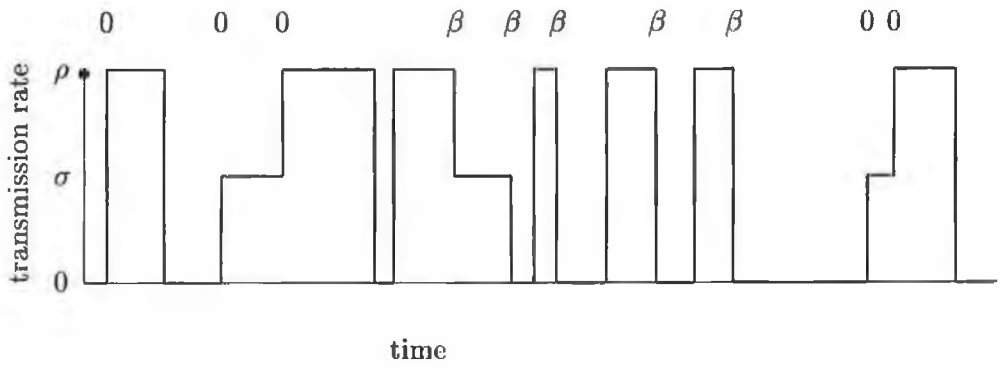

Figure 5.1: A typical extreme point of F in this linear structure. The values above the figure give the content of the leaky bucket at that time. Note that each burst must either start when the leaky bucket is empty or end when it is full.

Chapter 6

Examples and Applications

We deal in this chapter with applications of our results and examples of functionals where the maximum is easily computed.

6.1 Bufferless Resources

A set of functionals with a particularly simple worst case behaviour is the set of functionals which represent the performance of bufferless systems. The defining characteristic of these systems is that the functional depends only on the distribution of the arrival rate and not on the details of the sample paths. In other words, a real valued functional F on the set of stationary probability measures on $\mathbb{D}^+$ will represent a bufferless resource if there is some function $G : M \to \mathbb{R}^+$ such that $F(\mu) = G({}_\pi\mu)$ for all stationary probability measures μ on $\mathbb{D}^+$. Here M is the set of probability measures on $\mathbb{R}^+$ and ${}_\pi\mu$ is the image law of μ under the mapping $\pi : \mathbb{D}^+ \to \mathbb{R}^+ : a \to a'(1)$, which evaluates the derivative at $t = 0$.

A partial order on the set of random variables is called a stochastic order. A useful example is the following: $R_1 \preceq R_2$ if and only if $\mathbb{E}f(R_1) \leq \mathbb{E}f(R_2)$ for every non decreasing convex function $f : \mathbb{R} \to \mathbb{R}$. We find that the bufferless functionals considered in this thesis are all isotonic with respect to this stochastic order, that is to say $G(R_1) \leq G(R_2)$ if $R_1 \preceq R_2$. For example

- the loss rate in a buffer of size zero, served at constant rate s. In this case $G(R) := \mathbb{E}(R - s)^+$. In general, if $G(R) = \mathbb{E}f(R)$ for some nondecreasing convex function f then G will be isotonic with respect to $\preceq$.

- the loss rate in a buffer of size zero, served at constant rate s when multiplexed with a fixed, independent, stationary source. Here $G(r) := \mathbb{E}(R + X - s)^+$ where X is a real valued random variable which is independent of R.

- the loss rate in a buffer of size zero, served at constant rate s when N identical and independent stationary sources are multiplexed. This functional was considered by Doshi in [8]. Here $G(R) := \mathbb{E}(R_1 + \cdots + R_N - s)^+$, where $R_1, \ldots, R_N$ are N independent copies of R.

- the function $G(R) := \mathbb{E}e^{\vartheta R}$ for some constant $\vartheta \in \mathbb{R}$, which was investigated by Mitra and Morrison [6]. This functional can be related to the finite time moment generating function discussed in Section 2.7. Duffield and Botvich [10] show that the loss rate in a bufferless server has the asymptotics

$$\lim_{N\to\infty} \frac{1}{N} \log L(N, Ns, 0) = -I(0),$$

where

$$I(0) = \left(\lim_{t\to 0} \log \mathbb{E}e^{\vartheta a(t)/t}\right)^* = \left(\mathbb{E}e^{\vartheta a'(0)}\right)^*.$$

Assume we have a finite collection $\{(0, \rho), (b_1, \sigma_1), \ldots, (b_n, \sigma_n)\}$ of leaky bucket constraints which includes a bound ρ on the peak rate. If μ is stationary measure on $\mathbb{D}^+$ which satisfies the leaky bucket constraints almost surely, and R is a random variable whose distribution is the image law of μ under π, then clearly we must have that $\mathbb{P}(R > \rho) = 0$ and that $\mathbb{E}R \leq m := \min_i \sigma_i$. In fact the converse also holds: given any real valued random variable R that satisfies these two conditions we may find a stationary process a that satisfies the leaky bucket constraints such that R is equal in distribution to $a'(1)$. Thus the problem of finding the worst case behaviour of a bufferless system reduces to finding the distribution of R that maximizes $G(R)$ under the constraints $\mathbb{E}R \leq m$ and $\mathbb{P}(R > \rho) = 0$.

Theorem 34 *Let R^* be a random variable taking value ρ with probability m/ρ and value zero with probability $1 - m/\rho$. If $\mathbb{E}R \leq m$ and $\mathbb{P}(R > \rho) = 0$ then $R \preceq R^*$.*

Proof. Let f be any convex nondecreasing $\mathbb{R}^+$-valued function on $\mathbb{R}^+$. We have that

$$\mathbb{E}f(R) \quad \leq \quad \mathbb{E}\left[\frac{R}{\rho} f(\rho) + \left(1 - \frac{R}{\rho}\right) f(0)\right]$$

$$
\begin{aligned}
&= \frac{\mathbb{E}R}{\rho} f(\rho) + \left(1 - \frac{\mathbb{E}R}{\rho}\right) f(0) \\
&\leq \frac{m}{\rho} f(\rho) + \left(1 - \frac{m}{\rho}\right) f(0) \\
&= \mathbb{E}f(R^*).
\end{aligned}
$$

□

Consider the periodic source that transmits at rate ρ for time $t_\rho := \min_i\{\beta_i/(\rho - \sigma_i)\}$ and is silent for time $t_0 := \rho t_\rho/m$. Clearly none of the leaky buckets overflow during the time the source is transmitting and they are all empty at time $t_\rho + t_0$. The latter statement follows since the contents of leaky bucket i are less than that of leaky bucket $i_m := \text{argmin}_i\sigma_i$ at any time, and leaky bucket i_m is empty at time $t_\rho + t_0$. For this process $a'(1)$ has the same distribution as R^*. We immediately conclude that such a process is the worst case traffic for all the performance functionals discussed above.

6.2 A Single Source

When the functional to be maximised represents the queueing behaviour of a single source the optimisation problem may be tackled using elementary methods.

Loss Rate

It follows from Theorem 20 that the loss rate of any source is jointly convex in the service rate and buffer size. Assume that a single leaky bucket constraint (β, σ) is given, along with a peak rate constraint ρ. Consider the loss rate $l(b, s)$ of a stationary source that satisfies these constraints, as a function of the buffer size b and the service rate s. We know that

- at (β, σ) the loss rate is 0,
- at $(0, \rho)$ the loss rate is 0.
- at $(x, 0)$ the loss rate is less than σ, for all $x \geq 0$.

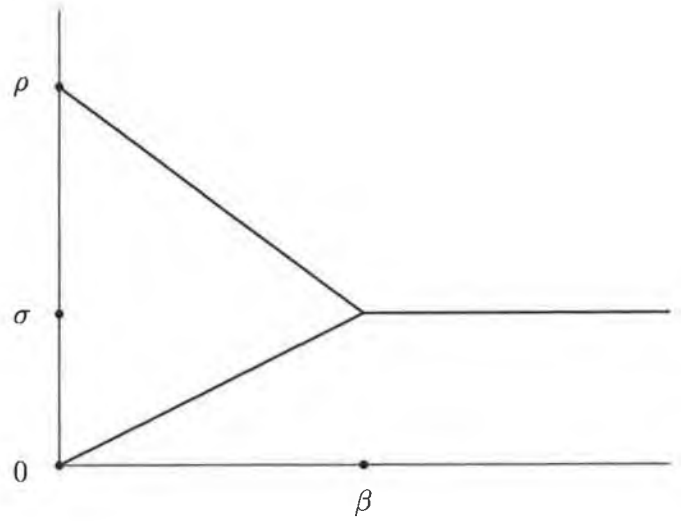

Figure 6.1: The three regions of the (b, s) plane in which $l(b, s)$ is affine.

We may use this knowledge to bound l since the it is jointly convex in b and s. Thus

$$l(b,s) \le \widehat{l}(b,s) := \begin{cases} \sigma - s, & s \le \sigma, \frac{s}{b} \le \frac{\sigma}{\beta} \\ \sigma - \left(1 - \frac{\sigma}{\rho}\right)\frac{\sigma}{\beta}b - \frac{\sigma}{\rho}s, & \frac{s}{b} \ge \frac{\sigma}{\beta}, \frac{\rho - s}{b} \ge \frac{\rho - \sigma}{\beta} \\ 0, & \frac{\rho - s}{b} \le \frac{\rho - \sigma}{\beta}, s \ge \sigma. \end{cases}$$

Note that the bound is affine in each of the three regions that are depicted in Figure 6.1. We will show that there is a stationary traffic source whose loss function is exactly $\widehat{l}(b, s)$. This traffic source is therefore the worst case traffic for this functional.

Theorem 35 *An on-off source with on periods of length $\beta/(\rho - \sigma)$ and off periods of length β/σ has a loss rate equal to the convex hull mentioned above.*

Proof. In the region of the (b, s) plane where $b/s \ge \beta/\sigma$ and $s \le \sigma$, the buffer will fill up but will never empty. It follows that the rate of loss over a period will be $\sigma - s$. In the region where $b/s \le \beta/\sigma$ and $b/(\rho - s) \le b/(\rho - \sigma)$, the buffer will fill up and then empty, and so service will be wasted. The loss rate will be $(\rho - s)\sigma/\rho - b(\rho - \sigma)\sigma/\beta\rho$. If $s \ge \sigma$ and $b/(\rho - s) \ge b/(\rho - \sigma)$, then the buffer will never overflow. The loss rate is thus found to be an affine function of b and s in each of the three regions. □

We have found that when we have only one leaky bucket parameter in addition to the peak rate, there is a single traffic source that is the worst case for any buffer size and

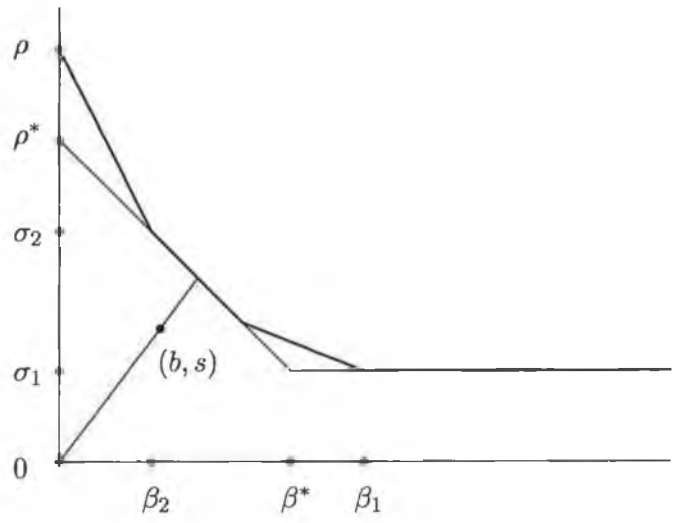

Figure 6.2: An illustration of the construction of the worst case traffic for more than one leaky bucket constraint.

service rate. When extra leaky bucket parameters are given the situation is more complex. We find in general that the convex hull bound is attained, but that the traffic source that attains the bound differs for different values of b and s. A construction of a source that attains the convex hull bound is now given.

We allow a possibly infinite number of constraints of the form $l(b_i, s_i) = 0$. Without loss of generality we may assume that $\mathcal{C} := \{(\beta_i, \sigma_i)\}$ is convex. Assume that the mean rate is not constrained to be zero, in other words that $\sigma := \inf\{\beta_i : (\beta_i, \sigma_i) \in \mathcal{C}\} > 0$. Extend a line from origin through (b, s) to intersect the boundary of $\mathcal{C}$ at (x, y). This set is convex, therefore it has a line of support line at the (x, y). Let ρ^* be the ordinate of the intersection of this line with the y-axis. and let (β^*, σ) be the point of intersection with the line $y = \sigma$. Consider the on-off source S^* that has peak rate ρ^*, on-time $\beta^*/(\rho^* - \sigma)$, and off-time β^*/σ. Let $l^*(b, s)$ be the loss rate of this source as a function of b and s. S^* clearly obeys the constraints $\mathcal{C}$. Also, if $l(b, s)$ is the loss rate of any other source that obeys the constraints, then $l(x, y) = 0$ and $l(0, 0) \leq \sigma$. Therefore $l(b, s) \leq \sigma(1 - b/x) = \sigma(1 - s/y)$. But $l^*(b, s) = \sigma(1 - b/x) = \sigma(1 - s/y)$ and so S^* is the worst case traffic source. This construction is illustrated in Figure 6.2.

Since the peak rate of S^* is less than ρ, this source is not of the form conjectured in Section 1.4 to be the worst case. However, for the single source loss rate functional the worst case not unique and we will now show how to find a source of this form that

obeys the constraints and has a loss rate in a buffer of size b served at rate s equal to that of S^*. For simplicity we assume that there are a finite number of constraints (β_i, σ_i) arranged in order of increasing β, and that there is a peak rate constraint $(0, \sigma_0)$. Let $I := \max\{i : \beta_i \leq x\}$, where x is defined above. Define $t_0 := 0$, and $t_i := \beta_i/(\sigma_{i-1} - \sigma_i)$ for $i \in \{1, \ldots, I\}$. Consider the source that transmits at rate σ_i in the interval of $[t_i, t_{i+1}]$ for $i \in \{0, \ldots, I-1\}$ followed by a silent interval of length $\sum_{i=1}^{I}(\sigma_{i-1} - \sigma_i)t_i$. This source obeys the constraints, has the same loss rate as S^* in a buffer of size b served at rate s, and is of the form discussed in Section 1.4.

Average Queue Length

Here the functional to be maximised is the average queue length when a single source passes through a single server queue with constant service rate s and infinite storage capacity. Again we assume there is a peak rate constraint ρ and another leaky bucket constraint (β, σ). This problem was considered by Lee [5]. There are a number of differences between his approach and ours:

- Lee models the traffic as a point process and his traffic policer creates discrete tokens at fixed intervals. Our results correspond to the fluid limit, that is where the token bucket size, the leaky bucket rates, and the service rate are large compared with the size of the token.

- Lee considers only a single leaky bucket constraint; there is no peak rate constraint.

- Lee considers the case of more than one source, however he does not assume that they are independent. This is equivalent to the one source case where there is an arbitrary periodic pattern of token generation.

- Lee maximises the average delay per cell rather than the average queue length over time. The two problems are equivalent since we know from the isotonicity of both functionals that the mean rate of the worst case traffic of either is equal to the service rate of the leaky bucket policer. Applying Little's law, the two functionals are related by a constant factor on the set of sources that have this mean rate.

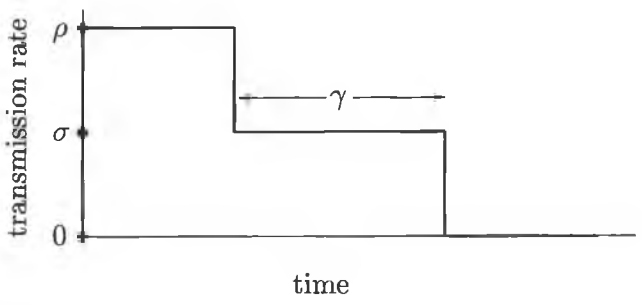

Figure 6.3: The tri-state source

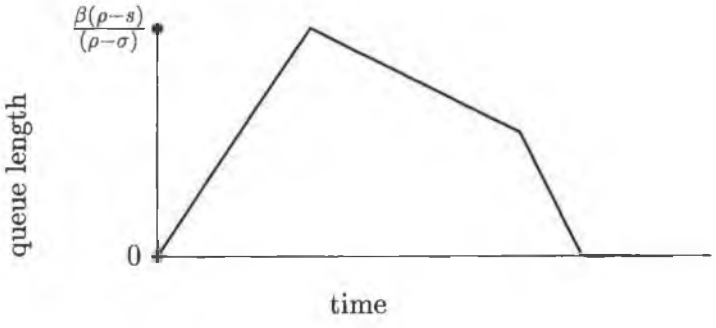

Figure 6.4: Queue length vs. time for the tri-state source

Lee reasons that the average delay over all cells is bounded by the maximum delay averaged within each busy period. One can therefore reduce the problem of finding the worst case traffic to that of maximising the average over a single cycle. This will be accomplished when each cell in the busy period, apart from the first, arrives as early as possible. Therefore the worst case source will be the one that transmits the tri-state pattern shown in Figure 6.3, for some value of the shoulder length γ. We shall calculate the average queue length for this type of source as a function of the shoulder length, and then maximise over the shoulder length.

Let $t_\rho := \beta/(\rho - \sigma)$ be the length of time the source transmits at the peak rate and let $t_0 := \beta/\sigma$ be the length of time it is silent. The average queue length over time is

$$\mathcal{Q}^{\text{ave}} = \frac{\rho(\rho - s)t_\rho^2 + 2\sigma(\rho - s)t_\rho\gamma - \sigma(s - \sigma)\gamma^2}{2s(\gamma + t_\rho + t_0)}.$$

The derivative of this expression with respect to γ is

$$\frac{d\mathcal{Q}^{\text{ave}}}{d\gamma} = \frac{-\sigma(s - \sigma)\gamma^2 - 2\sigma(s - \sigma)(t_\rho + t_0)\gamma + \rho(\rho - s)t_\rho^2}{2s\left(\gamma + t_\rho + t_0\right)^2}.$$

The numerator is quadratic in γ and its root may be found using the quadratic formula:

$$\gamma_{\text{opt}} = \frac{\beta}{\sigma}\left[-\frac{\rho}{\rho-\sigma} + \sqrt{\frac{s\rho}{(s-\sigma)(\rho-\sigma)}}\right].$$

Substituting this expression back into the formula for the mean queue length we find the maximum to be

$$Q_{\text{opt}}^{\text{ave}} = \beta\left(1 - \sqrt{\frac{\rho(s-\sigma)}{s(\rho-\sigma)}}\right).$$

6.3 Calculations for the Worst Case Effective Bandwidth

In this section we assume that the conjecture in Section 1.4 is true, and based on this we attempt to calculate the optimal shoulder lengths for the effective bandwidth functional.

Define

$$h^{(x,y)}(t) := \begin{cases} 0, & 0 \le t < t_0, \\ \sigma, & t_0 \le t < t_0 + x, \\ \rho, & t_0 + x \le t < t_0 + x + t_\rho, \\ \sigma, & t_0 + x + t_\rho \le t < t_0 + x + t_\rho + y, \end{cases}$$

which is the rate of transmission at time t of a source with shoulder lengths x and y. The corresponding realisation is given by

$$a^{(x,y)}(t) := \int_0^t h^{(x,y)}(z)\,dz.$$

The period of the source is $p := t_0 + x + t_\rho + y$. Again we use the convention that $a(t) := \lfloor t/p \rfloor a(p^-) + a(t - \lfloor t/p \rfloor p)$ for $t \notin [0,p)$. The effective bandwidth of the source is $(\vartheta T)^{-1} \log E_{\vartheta,T}^{(x,y)}$ where

$$E_{\vartheta,T}^{(x,y)} := \frac{1}{p}\int_0^p e^{\vartheta[a^{(x,y)}(z+T) - a^{(x,y)}(z)]}\,dz.$$

We want to find the x and y that maximise $E_{\vartheta,T}^{(x,y)}$.

The effective bandwidth function is symmetric in the shoulder lengths, in other words $E_{\vartheta,T}^{(x,y)} = E_{\vartheta,T}^{(y,x)}$. Alone this is not enough to imply that the two shoulder lengths of the worst case traffic are equal. However we can also make use of the following theorem.

Theorem 36 *For fixed y, ϑ, and T, the function $E_{\vartheta,T}^{(x,y-x)}$ is concave in x in the range $[0,y]$.*

Proof. Write $g_x(t) := a^{(x,y-x)}(t+T) - a^{(x,y-x)}(t)$ and $f_\gamma(t) := g^\gamma(t-\gamma)$. Differentiating, we find that

$$\begin{aligned}\frac{d}{dx}E_{\vartheta,T}^{(x,y-x)} &= \frac{1}{\tau}\int_x^{x+t_\rho}\vartheta(\rho-\sigma)e^{\vartheta g_x(z)}\,dz - \frac{1}{\tau}\int_{x-T}^{x+t_\rho-T}\vartheta(\rho-\sigma)e^{\vartheta g_x(z)}\,dz \\ &= \frac{1}{\tau}\int_0^{t_\rho}\vartheta(\rho-\sigma)e^{\vartheta f_x(z)}\,dz - \frac{1}{\tau}\int_{-T}^{t_\rho-T}\vartheta(\rho-\sigma)e^{\vartheta f_x(z)}\,dz.\end{aligned}$$

Let $0 \le \gamma_1 \le \gamma_2 \le y$. Then $f_{\gamma_2}(t) \le f_{\gamma_1}(t)$ for $t \in [0, t_\rho]$ and $f_{\gamma_2}(t) \ge f_{\gamma_1}(t)$ for $t \in [-T, t_\rho - T]$. Thus

$$\int_0^{t_\rho} e^{\vartheta f_{\gamma_1}(z)}\,dz \le \int_0^{t_\rho} e^{\vartheta f_{\gamma_2}(z)}\,dz, \quad \text{and} \quad \int_{-T}^{t_\rho-T} e^{\vartheta f_{\gamma_1}(z)}\,dz \ge \int_{-T}^{t_\rho-T} e^{\vartheta f_{\gamma_2}(z)}\,dz.$$

Therefore

$$\frac{d}{dx}E_{\vartheta,T}^{(x,y-x)}\bigg|_{x=\gamma_2} \le \frac{d}{dx}E_{\vartheta,T}^{(x,y-x)}\bigg|_{x=\gamma_1}.$$

and the result follows. □

We have that $E_{\vartheta,T}^{(x,y-x)} = E_{\vartheta,T}^{(y-x,x)}$, and so $E_{\vartheta,T}^{(x,y-x)}$ is symmetrical in x about $x = y/2$. It follows that $E_{\vartheta,T}^{(x,y-x)}$ attains its maximum at $x = y/2$.

We will now show that this optimum shoulder length is shorter than the time-scale T.

Theorem 37 *For $x > T$ we have that $E_{\vartheta,T}^{(x,x)} \le E_{\vartheta,T}^{(T,T)}$.*

Proof. Write $p_x := t_0 + t_\rho + 2x$ and $p_T := t_0 + t_\rho + 2T$. Comparing $g(t) := a^{(x,x)}(t+T) - a^{(x,x)}(t)$ with $f(t) := a^{(T,T)}(t+T) - a^{(T,T)}(t)$, we find that

$$g(t) = \begin{cases} f(t), & 0 \le t < t_0 \\ T\sigma, & t_0 \le t < t_0 + x - T \\ f(t-x+T), & t_0 + x - T \le t < t_0 + x + t_\rho \\ T\sigma, & t_0 + x + t_\rho \le t < t_0 + t_\rho + 2x - T \\ f(t-2x+2T), & t_0 + t_\rho + 2x - T \le t < t_0 + t_\rho + 2x. \end{cases}$$

Therefore

$$\int_0^{p_x} e^{\vartheta g(t)}\,dt = \int_0^{p_T} e^{\vartheta f(t)}\,dt + 2(x-T)e^{\vartheta T\sigma}.$$

Since $\exp(\vartheta x)$ is convex in x, we have that

$$\frac{1}{p_T}\int_0^{p_T} e^{\vartheta f(t)}\,dt \ge \exp\left(\frac{\vartheta}{p_T}\int_0^{p_T} f(t)\,dt\right) = e^{\vartheta T\sigma},$$

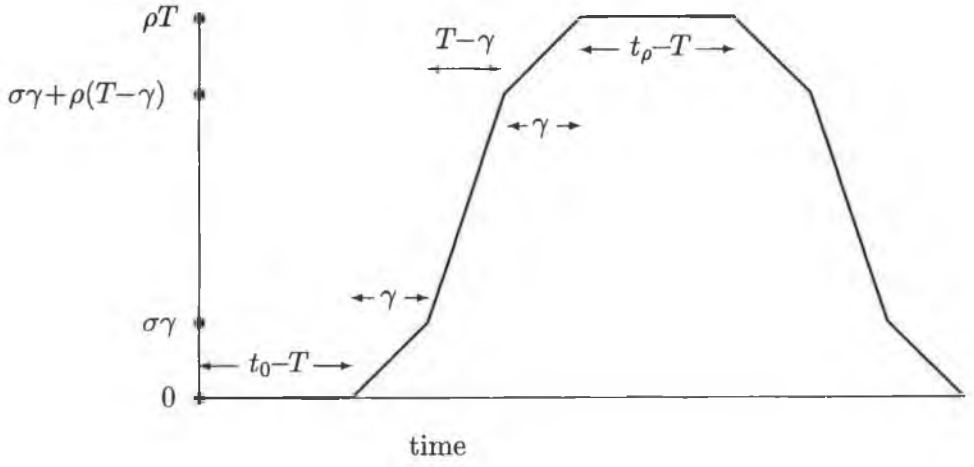

Figure 6.5: Graph of $a(t+T)-a(t)$ in region $[0,\ t_0+t_\rho+2\gamma]$

and so

$$\begin{aligned}\int_0^{p_x} e^{\vartheta g(t)}\,dt &\leq \int_0^{p_T} e^{\vartheta f(t)}\,dt + 2\frac{x-T}{p_T}\int_0^{p_T} e^{\vartheta f(t)}\,dt \\ &= \left(1+\frac{2(x-T)}{t_0+t_\rho+2T}\right)\int_0^{p_T} e^{\vartheta f(t)}\,dt \\ &= \frac{t_0+t_\rho+2x}{t_0+t_\rho+2T}\int_0^{p_T} e^{\vartheta f(t)}\,dt.\end{aligned}$$

Thus $E_{\vartheta,T}^{(x,x)} \leq E_{\vartheta,T}^{(T,T)}$. □

Calculation of the effective bandwidth can be tricky as there are a number of cases depending on the relative sizes of the time-scale T, the shoulder lengths, and the burst and silence lengths. We will concentrate on the case when the time-scale is less than the burst length and silence lengths. This is the simplest case since, from the theorem above, we know that the optimum shoulder lengths are shorter than the time-scale and therefore smaller than t_0 and t_ρ.

Take $T \leq \min(t_0, t_\rho)$. The graph of $a(t+T)-a(t)$ is shown in Figure 6.5. Using the identity

$$\int_a^b e^{\vartheta(c+mx)} = \frac{1}{m\vartheta}\left[e^{\vartheta(c+mb)} - e^{\vartheta(c+ma)}\right],$$

it is easy to verify that

$$E_{\vartheta,T}^{(\gamma,\gamma)} = \sum_{i=0}^{3} \alpha_i e^{\vartheta A_i},$$

where

$$\begin{aligned}\alpha_0 &= (t_0-T) - \frac{2}{\vartheta m}, & \alpha_1 &= \frac{2}{\vartheta m} - \frac{2}{\vartheta\rho}, \\ \alpha_2 &= \frac{2}{\vartheta\rho} - \frac{2}{\vartheta(\rho-m)}, & \alpha_3 &= \frac{2}{\vartheta(\rho-m)} + (t_\rho - T),\end{aligned}$$

and

$$\begin{aligned} A_0 &= 0, & A_1 &= \gamma m, \\ A_2 &= \gamma m + \rho(T-\gamma), & A_3 &= \rho T. \end{aligned}$$

To find the maximum, this expression may be differentiated and set equal to zero. However this gives rise to a transcendental equation which is best solved numerically.

Since $E_{\vartheta,T}^{(\gamma,\gamma)}$ is clearly continuous in γ, its supremum over $\gamma \in [0,T]$ is attained. Thus, assuming that the conjecture in the Introduction is true, the effective bandwidth functional provides an example where the optimisation problem has a solution. We also have that $E_{\vartheta,T}^{(\gamma,\gamma)}$ is differentiable in ϑ with continuous derivative. The only remaining property required in Section 2.7 is that the γ for which the supremum is attained is continuous in ϑ for each $T \in \mathbb{R}^+$. As yet however we are unable to prove this.

6.4 Numerical Results

We report here on some numerical investigations that we have carried out. Assuming that the conjecture in Section 1.4 is correct, we numerically optimise the shoulder lengths to maximise various functionals. The functionals we consider are: the average queue length for a small number of identical and independent sources, the loss rate for identical independent sources, and the effective bandwidth.

First we investigate the worst case average queue length in an infinite buffer queue served at constant rate for N independent sources. We use a program that calculates the mean queue length averaged over all phases of the sources. For $N = 1$, we find that the leading shoulder length γ_1 that maximises the average queue length is zero, which is in accord with the results of Section 6. The optimum trailing shoulder length γ_2 is plotted against the service rate in Figure 6.6. The parameters used in all experiments are $\beta = 0.5$, $\sigma = 0.5$, and $\rho = 1.0$. The service rate ranges from the mean rate 0.5 to the peak rate 1. When there is more than one source the optimum leading shoulder length is no longer zero. Figures 6.7 and 6.8 show the optimum leading and trailing shoulder lengths for the case of $N = 2$ and $N = 3$ respectively. Further values of N are impractical because the computational effort required grows exponentially in N. In all cases γ_2 becomes infinite as the service rate decreases towards the mean rate. This is what we would expect because, for a service rate that is only slightly larger than the mean rate, the buffer empties very

slowly and it pays to have a long trailing shoulder. Also, it may be seen that γ_2 goes to zero as the service rate approaches the peak rate. Again, this is understandable because at service rates just below the peak rate, the maximum queue length is small and so the buffer empties quickly. The optimum trailing shoulder length will be less than the time it takes for the buffer to empty, and so it will be short in this case. For both $N = 2$ and $N = 3$, the optimum length of the leading shoulder goes to the limit 1 time unit as the service rates approaches the mean rate. This is possibly related to the fact that our choice of leaky bucket parameters mean that the burst and silent periods have length 1 time unit. Interestingly, above a critical service rate the optimum value of γ_1 is zero. This critical service rate is higher for $N = 3$ than for $N = 2$. The actual worst case value of the average queue length is shown in Figure 6.9, for $N = 1$, $N = 2$, and $N = 3$. The value has been scaled by the number of sources to facilitate comparison. Unsurprisingly, for each value of N the optimised average queue length varies between 0.5 traffic units/source when the service rate is near the mean rate and zero traffic units when it is near the peak rate. Note that these curves are convex as predicted.

We turn now to investigate the worst case of the loss rate functional. When there is just a single source the results are trivial: the worst case is the on-off source discussed in Section 6.2. For more than one source, we find in general that the optimum values of the leading and trailing shoulder lengths are still equal. The optimum shoulder length (both leading and trailing) for $N = 2$ and $N = 3$ are shown in Figures 6.10 and 6.11 respectively. In both of these graphs, the optimum shoulder length is plotted against service rate for several values of the buffer size. It is seen to change slowly as the service rate is increased but then to drop suddenly. The optimum shoulder length is always less than 1 time unit, which is the length of the on period and of the silent period. The actual value of the maximum loss rate (Figures 6.12 and 6.13) is seen to be decreasing in the service rate and the buffer size and is jointly convex in these parameters, as expected.

For the effective bandwidth functional the optimum values of the leading and trailing shoulder lengths are again equal for all parameter values. In Figure 6.14 we see the optimum shoulder length plotted against the space scale θ for several values of the time scale T. We see that it is decreasing in θ and increasing in T. As T gets larger, the optimum value of shoulder length increases without bound. Note that optimum shoulder

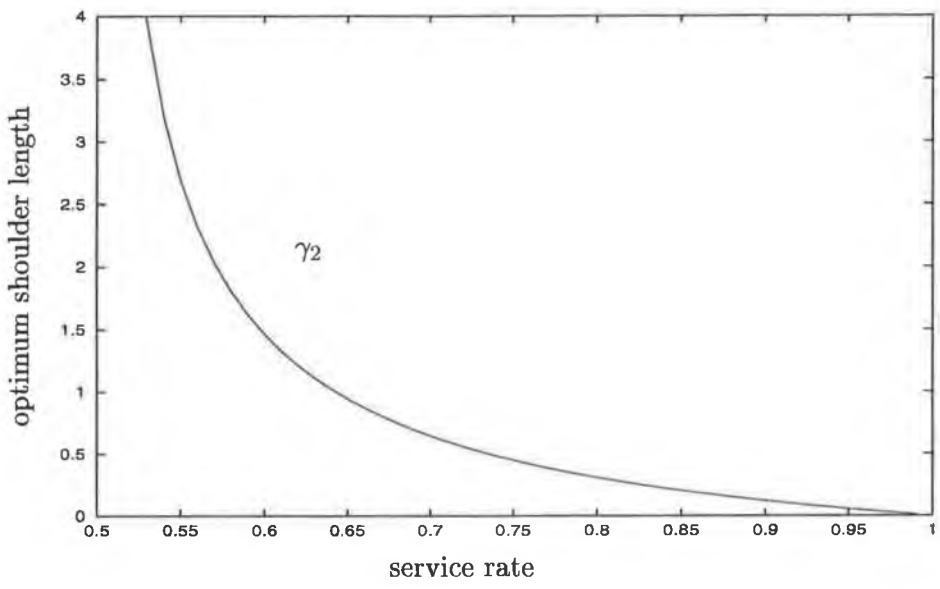

Figure 6.6: Trailing shoulder length that maximises the average queue length for a single source. The optimum leading shoulder length in zero.

length is always less than T as predicted. The maximum value of the effective bandwidth can be seen in a 3-dimensional plot against T and θ in Figure 6.15. The effective bandwidth of a periodic on-off source with on time $\beta/(\rho - \sigma) = 1$, off time $\beta/\sigma = 1$, and uniformly distributed phase is shown in Figure 6.16 for comparison. We see that the two graphs agree for small values of T. This is because, for small T, the optimum shoulder length is small and so the worst case traffic will be similar in nature to the on-off source. The main difference between the two plots is that the worst case effective bandwidth plot does not have the valleys seen in the on-off traffic plot. The valleys occur when T is close to a multiple of the period of the on-off traffic pattern because here the effective bandwidth functional only sees the average behaviour of the source over its period. In the worst case plot, the period of the source increases along with T and so this effect does not occur.

6.5 Connection Admission Control in ATM networks

We now describe an application of the results on the thesis to the problem of Connection Admission Control (CAC) in ATM networks. Recall that the network must make a decision whether to admit a connection based on the parameters supplied by the customer as

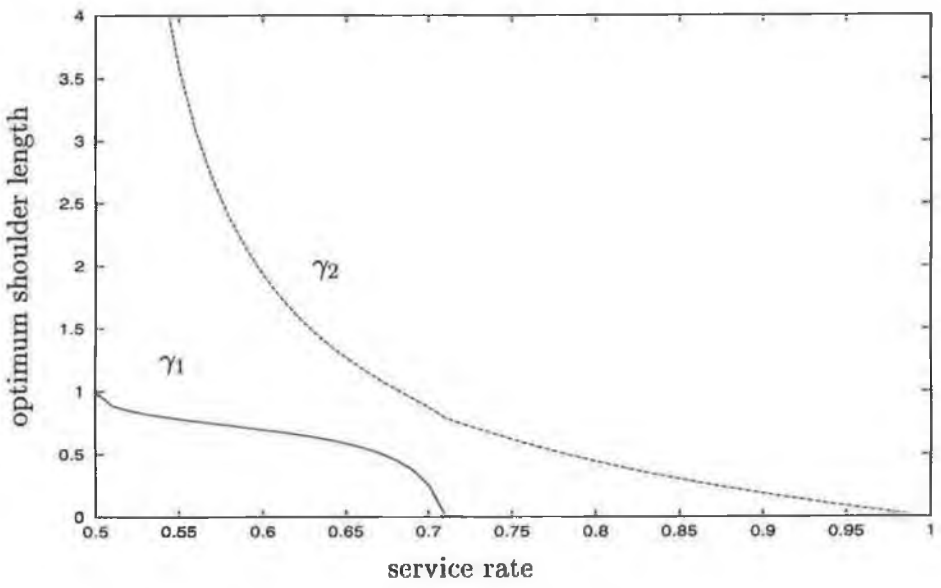

Figure 6.7: Shoulder lengths that maximise the average queue length for two independent sources.

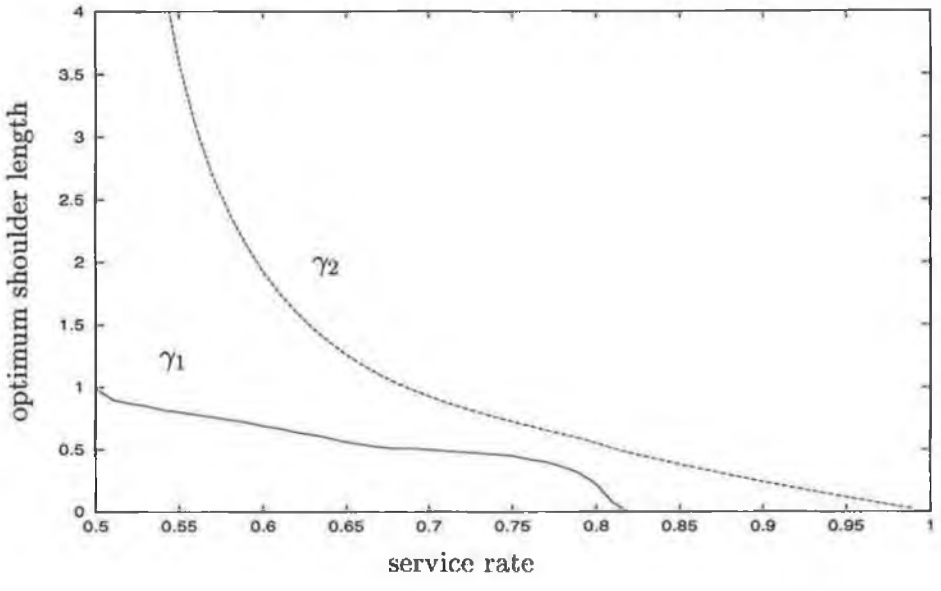

Figure 6.8: Shoulder lengths that maximise the average queue length for three independent sources.

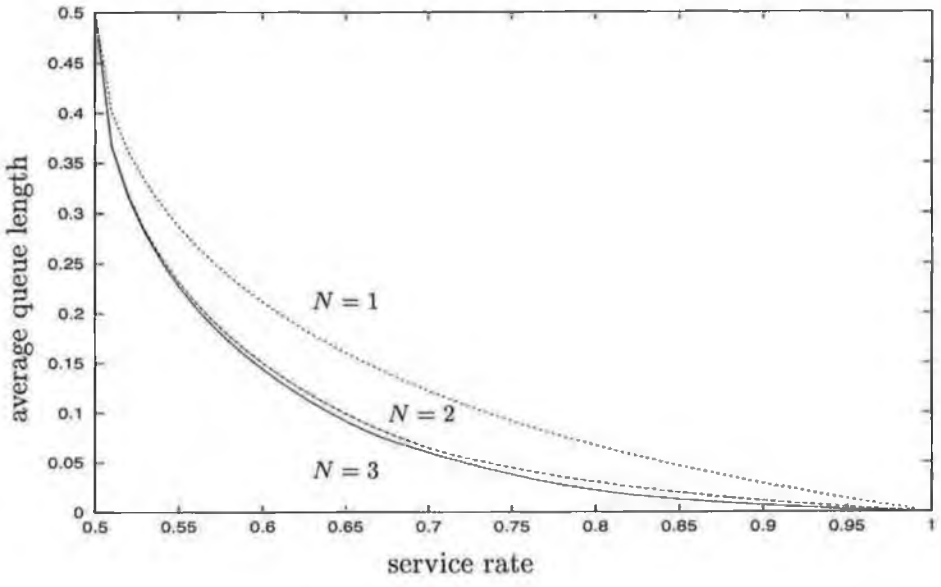

Figure 6.9: Maximum value of the average queue length for one, two, and three independent sources.

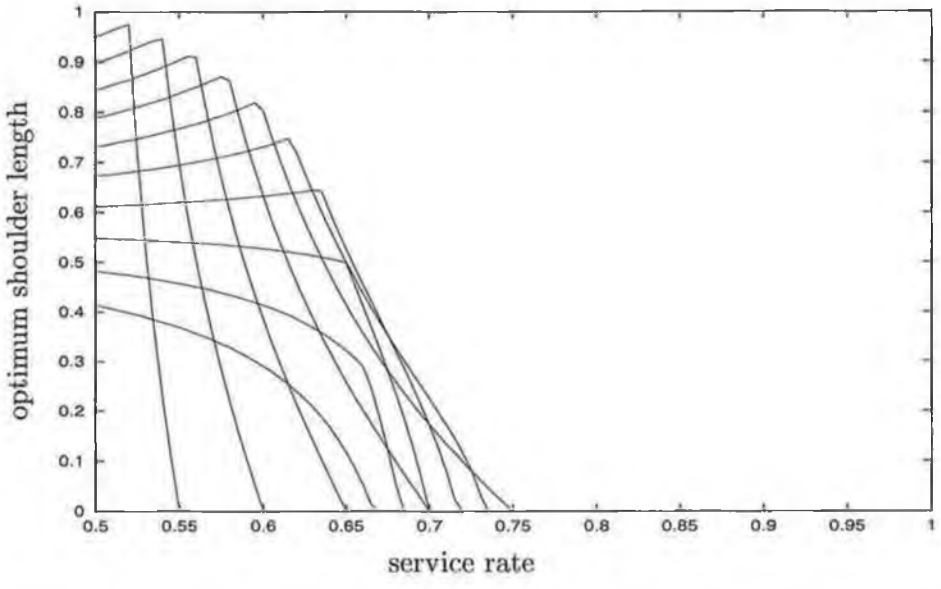

Figure 6.10: The shoulder length that maximises the loss rate of two independent sources, plotted against service rate for several values of the buffer size.

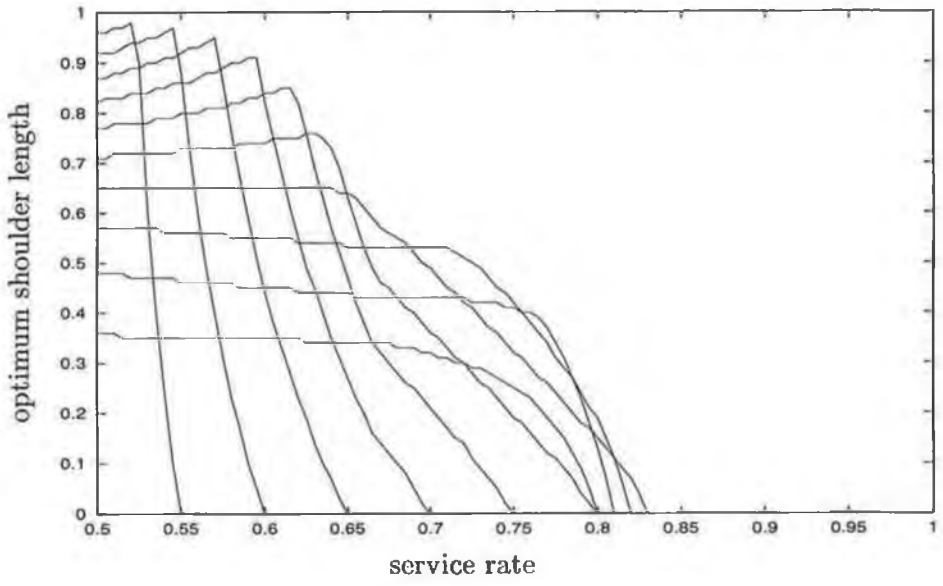

Figure 6.11: The shoulder length that maximises the loss rate of three independent sources, plotted against service rate for several values of the buffer size.

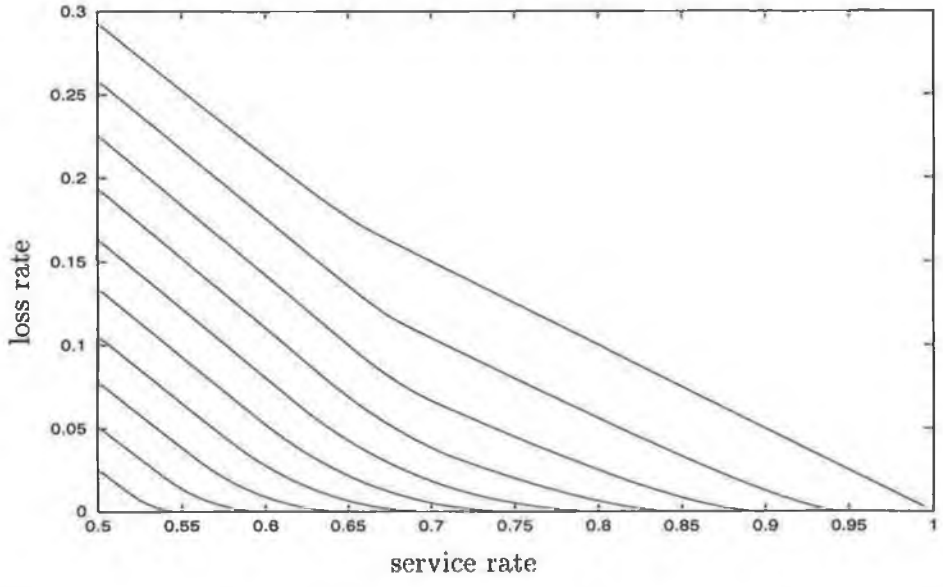

Figure 6.12: Maximum loss rate as a function of service rate for several values of the buffer size for two independent sources.

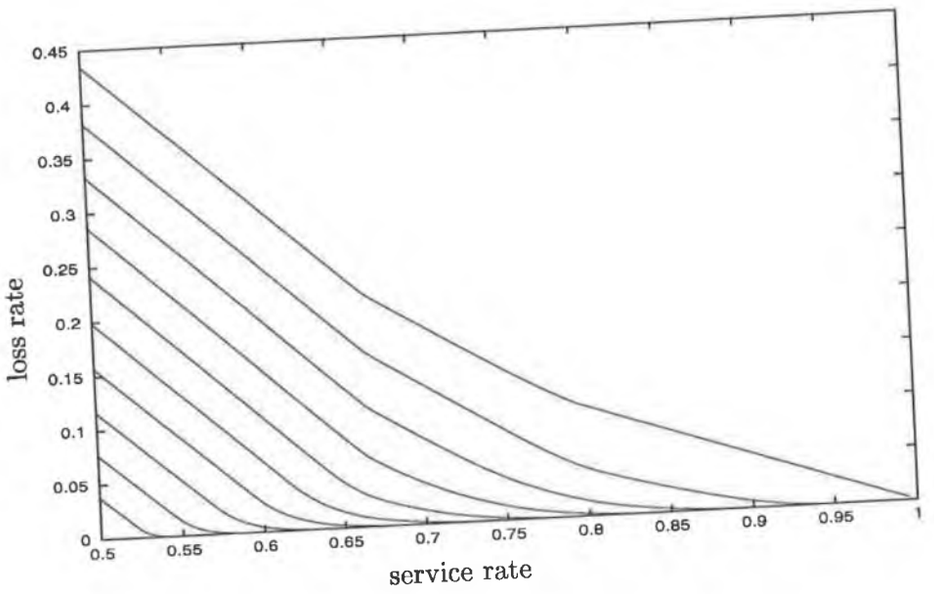

Figure 6.13: Maximum loss rate as a function of service rate for several values of the buffer size for three independent sources.

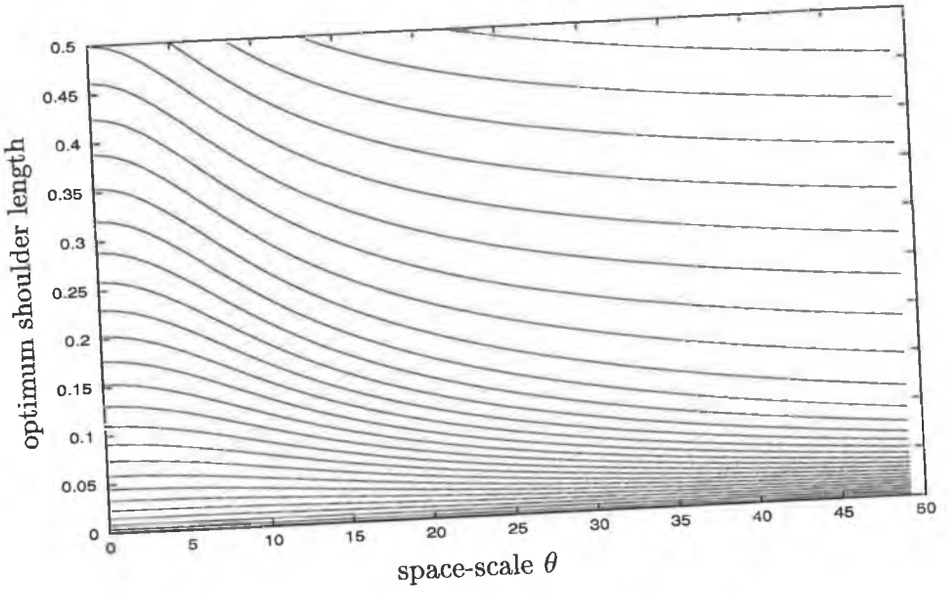

Figure 6.14: The optimum shoulder length for the effective bandwidth functional for several values of the time-scale T.

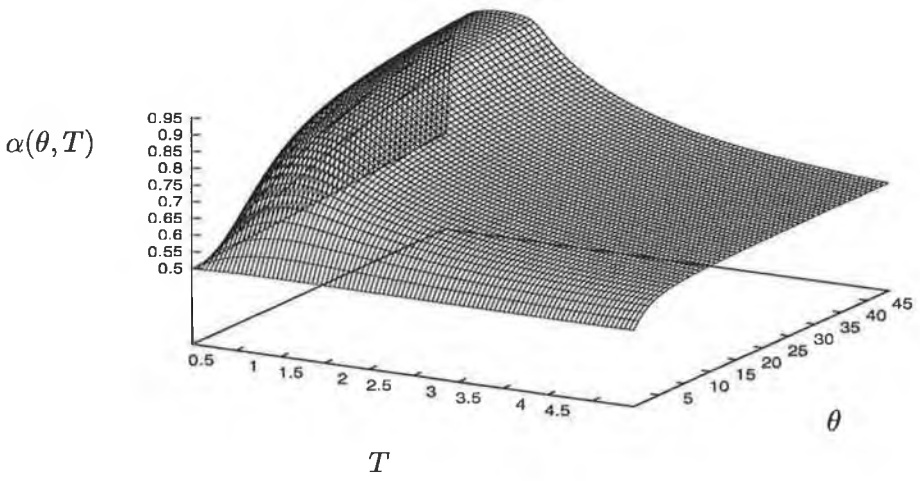

Figure 6.15: The worst case effective bandwidth against the time-scale T and the space-scale θ.

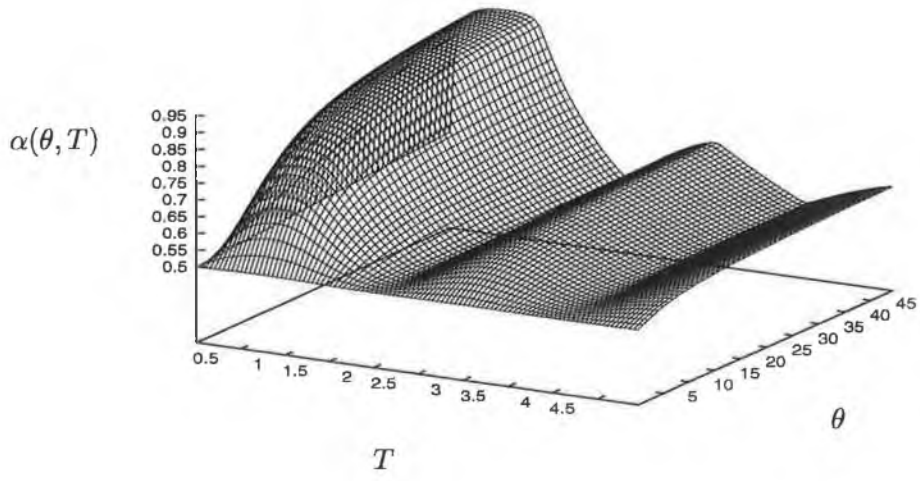

Figure 6.16: The effective bandwidth of the on-off periodic source.

part of the traffic contract. Many algorithms for doing this have been proposed in the literature [34, 35, 36]. The algorithm described here is based on that in [37]. This algorithm makes use of the declared peak rate of a connection request along with measurements of the current multiplex of connections. The ATM switch is modelled as a single server queue with a constant service rate S and a fixed buffer size B. Let $L(X, Y)$ denote the average loss rate in a buffer of size X and a service rate of Y fed with the current multiplex of connections. Since the peak rate ρ declared by the new connection is an upper bound on its rate of transmission, $L(B, S-\rho)$ will be a conservative bound on the total loss rate should the connection request be granted. The problem of CAC is thus reduced to that of estimating this quantity. We have seen in Section 2.7 that the loss rate obeys the following asymptotics in the number of sources N:

$$\lim_{N\to\infty} \frac{1}{N} \log L(Nb, Ns) = -I(b, s),$$

where I is given by

$$I(b, s) := \inf_{t\geq 0} (\theta t \alpha(\cdot, t))^*(b, s)$$

and the effective bandwidth is defined to be

$$\alpha(\theta, t) := \lim_{N\to\infty} \frac{1}{N\theta t} \log \mathbb{E} e^{\sum_{n=1}^{N} \theta A_t^n}. \tag{6.1}$$

We shall assume that the traffic sources are independent but not necessarily identical. In this case the effective bandwidth reduces to $\alpha(\theta, t) := N^{-1} \sum_{n=1}^{N} \alpha^{(n)}(\theta, t)$, where $\alpha^{(n)}$ is the effective bandwidth of the nth source defined analogously to (6.1) as $\alpha^{(n)}(\theta, t) := (\theta t)^{-1} \log \mathbb{E} e^{\theta A_t^{(n)}}$. We can use this as the basis for an approximation scheme: for N finite but large we use the approximation

$$\begin{aligned} \log L(B, S) &\approx -NI(B/N, S/N) \\ &= -N \inf_{t\geq 0} (\theta t \alpha(\cdot, t))^*(B/N + St/N) \\ &= -\inf_{t\geq 0} (\theta t \sum_n \alpha^{(n)}(\cdot, t))^*(B + St). \end{aligned}$$

In [37] a method of estimating $\alpha^{(n)}$ is proposed. This leads to an estimate

$$\log \widehat{L}(B, S) \approx -\inf_{t\geq 0} (\theta t \sum_n \widehat{\alpha}^{(n)}(\cdots, t))^*(B + St).$$

The decision procedure is that a connection is accepted if this estimate of the loss rate is less than the Quality of Service (QoS) bound r on the loss rate guaranteed to the connections.

However, it is likely that the customer making the request will be obliged to declare other parameters in addition to the peak rate, for example the leaky bucket parameters with which we have been dealing. Adapting the CAC algorithm proposed in [37] to deal with parameters such as these in addition to the peak rate would allow greater link usage while still maintaining QoS guarantees.

Finding a tight bound on the effective bandwidth is made even more important by the fact that the CAC algorithm must continue to use the declared parameters of a connection for some time after it has been accepted while measurements are being made on the statistics of the new connection. If connections are short and there are many of them, there may be a significant number of calls in the system for which no measurements are available.

The work of this thesis suggests a method of using declared leaky bucket parameters to predict the effect of new connections. The effective bandwidth of the new call is bounded by that of the worst case traffic and we have conjectured what this worst case is. When a new connection makes its request, the ATM switch can calculate the worst case bound $\overline{\alpha}$ on its effective bandwidth based on its declared leaky bucket parameters using the methods of Section 6.3. It can then accept the connection if

$$\inf_{t\geq 0} (\theta t\widehat{\alpha}(\cdot,t) + \theta t\overline{\alpha}(\cdot,t))^*(B+St) \geq -\log r$$

In practice a complete optimisation over θ and t will not be necessary since we would not expect the critical space-scale and time-scale to change appreciably with the addition of a single connection. An occasional updating of θ and t should be sufficient.

Bibliography

[1] Rene L. Cruz. A calculus for network delay, part I: Network elements in isolation. *IEEE Transactions on Information Theory*, 37(1):114–131, January 1991.

[2] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks—the single-node case. *IEEE/ACM Trans. Networking*, 1(3):344–357, June 1993.

[3] Jonathan S. Turner. New directions in communications (or which way to the information age?). *IEEE Communications Magazine*, 24(10):8–15, October 1986.

[4] ATM Forum. Traffic management specification version 4.0. Technical Report 95-0013R10, ATM Forum, February 1996.

[5] Daniel Chonghwan Lee. Effects of leaky bucket parameters on the average queueing delay: Worst case analysis. In *Proceedings of IEEE Infocom'94*, pages 482–489, Toronto, Canada, June 1994.

[6] Debasis Mitra and John A. Morrison. Multiple time scale regulation and worst case processes for ATM network control. In *Proceedings of the 34th Conference on Decision & Control*, pages 353–358, New Orleans, LA, December 1995.

[7] Philippe Oechslin. In search of the worst case arrivals of independent leaky bucket constrained sources. In *Proceedings UKPEW*, pages 100–106, Bradford, July 1997.

[8] B. T. Doshi. Deterministic rule based traffic descriptors for broadband ISDN: Worst case behavior and connection acceptance control. In *Proceedings of the 14th ITC*, pages 591–600, Antibes Juan-les-Pins, June 1994.

[9] Karl R. Stromberg. *An introduction to classical Real Analysis.* Wadsworth, 1981.

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

[11] P. A. P. Moran. A probability theory of a dam with a continuous release. *Quart. J. Math. (Oxford, 2)*, 7:130–137, 1956.

[12] J. Gani and R. Pyke. The content of a dam as the supremum of an infinitely divisible process. *J. Math. and Mech.*, 9:639–652, 1960.

[13] E. Reich. On the integro-differential equation of Takács, I. *Annals of Mathematical Statistics*, 29:563–570, 1958.

[14] J. F. C. Kingman. On continuous time models in the theory of dams. *Journal of the Australian Mathematical Society*, 3:480–487, 1963.

[15] J. Michael Harrison. *Brownian Motion and Stochastic Flow Systems.* Wiley, 1985.

[16] António Pacheco. A storage equation on continuous time. Technical report, Technical University of Lisbon, June 1994.

[17] Fergal Toomey. Bursty traffic and finite capacity queues. *Annals of Operations Research*, 79:45–62, 1998.

[18] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61:1872–1894, 1982.

[19] Costas Courcoubetis and Richard Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33:886–903, 1996.

[20] A. Simonian and J. Guibert. Large deviations approximations for fluid queues fed by a large number of on/off sources. *IEEE Journal of Selected Areas in Communications*, 13:1017–1027, 1995.

[21] Frank P. Kelly. Notes on effective bandwidths. In Kelly, Zachary, and Ziedins, editors, *Stochastic Networks: Theory and Applications*, pages 141–168. Oxford University Press, 1996.

[22] R. Tyrrell Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[23] David Williams. *Probability with Martingales.* Cambridge University Press, 1991.

[24] Patrick Billingsley. *Probability and Measure.* Wiley, 1995.

[25] Paul R. Halmos. *Ergodic Theory.* Chelsea Publishing Company, 1956.

[26] Robert R. Phelps. *Lectures on Choquet's Theorem.* Van Nostrand, 1966.

[27] Stéphane Gaubert and Max Plus. Methods and applications of (max,+) linear algebra. Research Note 3088, INRIA Rocquencourt, January 1997.

[28] Costas Courcoubetis, Frank Kelly, and Richard Weber. Measurement-based charging in communications networks. Research Report 1997-19, Statistical Laboratory, Cambridge, 1997.

[29] V. P. Maslov and S. N. Samborskiĭ, editors. *Idempotent Analysis*, volume 13 of *Advances in Soviet Mathematics.* American Mathematical Society, Rhode Island, 1992.

[30] I. V. Romanovskiĭ. Optimization of stationary control of discrete deterministic processes in dynamic programming. *Kibernetika*, 2:66–78, 1967. English translation in Cybernetics 3 (1967).

[31] M. G. Crandall and L. Tartar. Some relations between nonexpansive and order preserving maps. *Proceedings of the AMS*, 78(3):385–390, 1980.

[32] A. P. Robertson and W. J. Robertson. *Topological Vector Spaces.* Cambridge University Press, 1964.

[33] Donald Gross and Carl M. Harris. *Fundamentals of Queueing Theory.* Wiley, 1985.

[34] R. J. Gibbens and F. P. Kelly. Measurement-based connection admission control. In *Proceedings of the 15th International Teletraffic Congress*, pages 879–888, Washington, DC, June 1997.

[35] S. Floyd. Comments on measurement-based admissions control for controlled-load services. Preprint, Berkeley, July 1996. To be submitted to Computer Communications Review.

[36] Sugih Jamin, Peter B. Danzig, Scott J. Shenker, and Lixia Zhang. A measurement based admission control algorithm for integrated services packet networks. *IEEE/ACM Transactions on Networking*, 5(1):56–70, February 1997.

[37] Brian McGurk and Cormac Walsh. Investigations of a measurement-based connection admission control algorithm. In *Proceedings of the Fifth IFIP Workshop*, Bradford, July 1997.