Worst Case Traffic from Regulated Sources

Cormac Walsh

Inria/ENS-Lyon, LIP, 46 Allée d'Italie, 69364 Lyon Cedex 07, France

Abstract

We address the problem of finding the worst possible traffic a user of a telecommunications network can send. We take "worst" to mean having the highest effective bandwidth, a concept that arises in the Large Deviation theory of queueing networks. The traffic is assumed to be stationary and to satisfy "leaky bucket" constraints, which represent the *a priori* knowledge the network operator has concerning the traffic. Firstly, we show that this optimization problem may be reduced to an optimization over periodic traffic sources. Then, using convexity methods, we show that the realizations of a worst case source must have the following properties: at each instant the transmission rate must be either zero, the peak rate, or the leaky bucket rate; it may only be the latter when the leaky bucket is empty or full; each burst of activity must either start with the leaky bucket empty or end with it full.

1991 Mathematical Subject Classification: 90B22 (60K25)

Keywords: regulated traffic, worst case, statistical multiplexing, stationary independent, large deviations, effective bandwidth, leaky bucket, convex optimization.

1 Introduction

When operating a telecommunications network, it is often important to know the worst possible traffic a user can send. This is the case, in particular, when the network intends to guarantee a certain quality of service, for example bounds on the loss or delay. To do this, the network must act conservatively. In other words, when it makes a decision concerning, for example, resource allocation or call acceptance, it must assume the worst: that the user is adversarial to the maximum possible extent. The network will wish to take into account any information it has about the offered traffic, including perhaps that contained in a contract negotiated with the user. We are thus led to consider an optimization over the space of traffic processes. We wish to find the traffic source that maximizes the congestion while conforming to the given traffic parameters.

We model the network as follows. The possible realizations of a traffic source are taken to be non-negative, locally-finite Borel measures on $[0, \infty)$. We denote by $\mathbb{D}$ the set of measures with these properties. If $a \in \mathbb{D}$, then $a[t_0, t_1]$ gives the amount of traffic arriving in the time interval $[t_0, t_1]$.

A traffic source is a probability measure on $\mathbb{D}$. We use, on $\mathbb{D}$, the σ-algebra generated by the topology of weak convergence. Realizations are shifted forward in time with the operator $\Theta_h : \mathbb{D} \to \mathbb{D}$ defined, for each $h \geq 0$, by $(\Theta_h a)[\mathcal{B}] := a[\mathcal{B} + h]$ for all Borel sets $\mathcal{B} \subset [0, \infty)$. When we talk of stationary sources we mean stationary with respect to this operator.

There are many functionals that could represent the degree of congestion in the network. In this paper, we are concerned with the *effective bandwidth* functional, defined for a source μ to be

$$\mathcal{E}_{\vartheta,T}(\mu) := \frac{1}{\vartheta T} \log \mathbb{E}_\mu e^{\vartheta a[0,T]}.$$

The two parameters $\vartheta \in \mathbb{R}$ and $T \in [0,\infty)$ are called the *space-scale* and the *time-scale* respectively. The effective bandwidth is exactly what is needed to calculate the asymptotics of the loss rate as the number of sources becomes infinite. Suppose there are J types of source, with type $j \in \{1,\ldots,J\}$ having law μ_j. Let n_j be the proportion of sources of type j; thus $\sum_{j=1}^J n_j = 1$. Denote by $Loss(S,B,N)$ the loss rate when a single server queue with service rate S and buffer size B is fed by N independent sources, where the proportion of each type is fixed by the n_j. As was shown in [5], *Loss* obeys the large deviations asymptotics

$$\lim_{N\to\infty} \frac{1}{N} \log Loss(sN,bN,N) = -\inf_{T>0} \sup_{\vartheta\in\mathbb{R}} \Big[(b+sT)\vartheta - \vartheta T \sum_{j=1}^J n_j \mathcal{E}_{\vartheta,T}(\mu_j)\Big]. \tag{1}$$

These asymptotics are also obeyed by the fraction of time the buffer is full and by the probability that the queue length exceeds a level bN in an infinite buffer [3]. The values T^* and ϑ^* where the infimum and supremum are attained define the operating point of the system. The time-scale parameter T^* will be the typical duration of the busy period before overflow. The space-scale parameter ϑ^* indicates the degree of statistical multiplexing obtained: for low values of ϑ, $\mathcal{E}_{\vartheta,T}(\mu)$ approaches the mean rate of source μ, while for high values it approaches the maximum rate that μ can sustain over an interval of length T. For an overview of the effective bandwidth functional, see the paper of Kelly [9].

The most commonly used traffic parameters in the literature are those defined by the "leaky bucket" mechanism, described first in [21]. This is the standard policing mechanism defined by the ATM Forum [1] under the name Generic Cell Rate Algorithm and is also used in the Intserv architecture for the Internet. In this paper, we use a fluid version of the leaky bucket. For $a \in \mathbb{D}$, $s > 0$, and $t \geq 0$, define

$$Q_s(a,t) := \sup_{t_1\in[0,t]} \big(a[t_1,t] - (t-t_1)s\big).$$

This is the queue length at time t in a buffer when a gives the arrivals and s the service rate. The leaky bucket constraint requires that $Q_\sigma(a,t) \leq \beta$ for all $t \geq 0$. We will also employ a peak rate constraint: $Q_\rho(a,t) = 0$ for all $t \geq 0$, where ρ is greater than σ. It is easy to see that these two constraints are equivalent to, respectively,

$$a[t_1,t_2] \leq (t_2-t_1)\sigma + \beta \tag{2}$$

$$\text{and} \qquad a[t_1,t_2] \leq (t_2-t_1)\rho, \tag{3}$$

for all $t_2 \geq t_1 \geq 0$. We denote by C the elements of $\mathbb{D}$ that satisfy both. We are interested in the set of stationary sources that meet the constraints almost surely. These are represented by the set $\mathcal{C}$ of stationary probability measures μ on $\mathbb{D}$ for which $\mu[C] = 1$.

With these preliminaries, we can now state our optimization problem:

$$\text{maximise } \mathcal{E}_{\vartheta,T}(\mu), \qquad \text{subject to } \mu \in \mathcal{C}. \tag{4}$$

We conjectured the solution in [22]. For each γ in the range $[0,T]$, let μ^*_γ be the source that transmits a periodic traffic pattern with uniformly distributed phase and whose realizations are composed of following four sections (see Figure 1):

- a peak rate burst of duration $\tau_\rho := \beta/(\rho - \sigma)$,
- an interval in which the source transmits at rate σ, of duration γ,
- a silent interval of duration $\tau_0 := \beta/\sigma$,
- another interval at rate σ, again of duration γ.

Note that τ_ρ and τ_0 are, respectively, just long enough to fill the leaky bucket from empty and to empty it from full. The conjecture in [22] is the following.

Conjecture 1 *For each ϑ and T, the stationary traffic source that maximizes the effective bandwidth $\mathcal{E}_{\vartheta,T}$ under the peak rate and leaky bucket constraints is one of the sources $\{\mu^*_\gamma\}_{0\leq\gamma\leq T}$.*

Note that this conjecture allows different sources to be worst case for different values of ϑ and T. If the conjecture is true, the problem of finding the worst case effective bandwidth for any given ϑ and T reduces to a one parameter optimization over γ.

Apart from the effective bandwidth, there are many other functionals that one is interested in maximizing. The simplest are those that arise when the source being optimized is the only one present. For many functionals of this type, the worst case is known. For example, Lee [12] shows that the worst case average queueing delay for a single source occurs when the source behaves in a periodic manner, transmitting firstly at the peak rate, then at rate σ, and finally at rate zero. Kesidis and Konstantopoulos [10] have a similar result for the source that maximizes the time the queue length spends above any given level.

More generally, one wishes to consider several constrained sources transmitting simultaneously. If the sources are not assumed to be stationary and independent, then in the worst case they conspire by synchronizing their bursts of traffic. The traffic patterns sent are similar to the single source worst case pattern [6, 20].

In many situations however, independence is a natural assumption. Since it may greatly increase the multiplexing gain achieved, it is necessary to take it into account. Examples of simple functionals incorporating independence include those arising from buffer-less systems. In general, the worst case sources with respect to these are "on-off", that is, alternate between transmitting at the peak rate and rate zero [8, 14]. With buffering, however, this is not true: counterexamples have been found which are worse than on-off sources [8, 15].

To deal with systems in which there is both independence and buffering, several authors [17, 13, 2, 11] have sought bounds on the functional of interest rather than attempting to calculate the exact worst case.

However, given the relatively simple form of the sources appearing in Conjecture 1, an exact analysis for the effective bandwidth functional should be possible. We remark that Kelly [4] has obtained results for a discrete-time packetized version of the problem. We have previously [22] proved Conjecture 1 in the case when T is less than both τ_0 and τ_ρ. In that paper, the methods used were specific to the effective bandwidth functional. Here our techniques are more general. They exploit properties, such as continuity and convexity, that are shared by many of the other functionals described above and may therefore be useful in determining the worst case traffic of these also.

The rest of the paper is organized as follows. In Section 2, we use an Ergodic Theory result to reduce the problem to an optimization over sources with periodic realizations. In Section 3, we define a linear structure on the set of periodic realizations with fixed period. Using convexity arguments, we show that the worst case source must have realizations which are extreme points of the constraint set. In Section 4, we characterize these points. In Section 5, we define an alternative linear structure. Under this, the constraint set has a different set of extreme points, which turns out to be significantly smaller. We show that the realizations of a worst case source

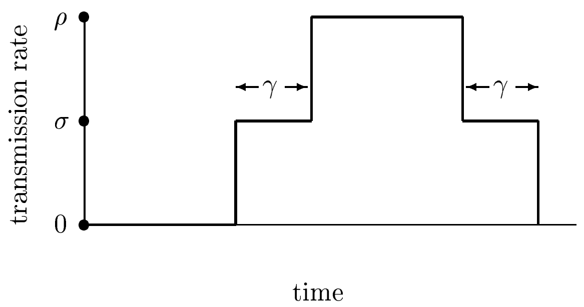

Figure 1: The “wedding cake” pattern, conjectured to have the worst case effective bandwidth.

must be extreme in this linear structure also. Finally, in Section 6, we characterize this second set of extreme points.

2 Periodicity

We say a realization $a \in \mathbb{D}$ is periodic with period $p > 0$ if $\Theta_p a = a$. Given such a realization, we may construct a stationary traffic source μ by taking a uniform distribution over the phase. More precisely,

$$\mu[\mathcal{B}] := \frac{1}{p}\,\mathrm{Leb}\{t \in [0,p) : \Theta_t a \in \mathcal{B}\}, \qquad \text{for all measurable subsets } \mathcal{B} \text{ of } \mathbb{D}.$$

We denote by $\mathcal{P}_p$ the set of sources formed in this way from elements of C that are periodic with period p. Our aim in this section is to show that the supremum of the effective bandwidth over $\cup_{p>0}\mathcal{P}_p$ is the same as that over $\mathcal{C}$.

The technique we use requires us to consider a discrete version of our optimization problem. For each $n \in \mathbb{N}$, we have the following model. A traffic realization is represented by a sequence $(a_i)_{i\in\mathbb{N}}$ with values in $\{1,\ldots,n\}$. We denote the set of such sequences by $\mathbb{D}_{(n)}$ and use on this product topology on this set. We relate the discrete model to the fluid one by interpreting $a_i\rho T/n^2$ as the amount of traffic in the interval $[iT/n,(i+1)T/n)$, for $i \in \mathbb{N}$. So, the set of discrete realizations that satisfy the constraints is

$$C^{(n)} := \Big\{a \in \mathbb{D}_{(n)} : \frac{\rho}{n}\sum_{i=i_0}^{i_1-1} a_i \le (i_1 - i_0)\sigma + \frac{n}{T}\beta \quad \text{for all } i_0, i_1 \in \mathbb{N},\ i_1 > i_0\Big\}.$$

We define the discrete time shift operator by $(\theta a)_i := a_{i+1}$ for all $i \in \mathbb{N}$. Let $\mathcal{C}^{(n)}$ be the set of stationary sources that satisfy the constraints almost surely, that is the set of probability measures ν on $\mathbb{D}_{(n)}$ such that $\nu[C^{(n)}] = 1$ and $\nu[\mathcal{B}] = \nu[\theta^{-1}\mathcal{B}]$ for all Borel subsets $\mathcal{B}$ of $\mathbb{D}_{(n)}$. A realization $a \in \mathbb{D}_{(n)}$ is periodic with period $p \ge 1$ if $\theta^p a = a$. A source μ is periodic if there is some $p \ge 1$ and $a \in C^{(n)}$ such that a is periodic with period p and μ has mass $1/p$ on $a, \theta a, \ldots, \theta^p a$.

The Ergodic Theory result we wish to apply requires the concept of *specification*. Recall that a *subshift* is a subset of $\mathbb{D}_{(n)}$ that is compact and shift invariant. A subshift Y is said to have the specification property if there exists an $N \in \mathbb{N}$ such that for any pair of elements

$a_1, a_2 \in Y$ and integer intervals $[t_1, s_1], [t_2, s_2] \subset \mathbb{N}$, with $t_1 > N$ and $t_2 > s_1 + N$, there exists a periodic $a \in Y$ of period s_2 that agrees with a_1 on $[t_1, s_1]$ and with a_2 on $[t_2, s_2]$. Although this concept is rather technical, it is easy to see that it applies to $C^{(n)}$.

Lemma 1 *$C^{(n)}$ is a subshift and has the specification property.*

Proof. Clearly $C^{(n)}$ is compact and shift invariant. We take N to be such that $N\sigma \geq n\beta/T$. Let $a_1, a_2 \in C^{(n)}$ and $[t_1, s_1], [t_2, s_2] \subset \mathbb{N}$ be as in the definition of the specification property. For each $i \in [1, s_2]$, define

$$a_i := \begin{cases} (a_1)_i, & \text{for } i \in [t_1, s_1] \\ (a_2)_i, & \text{for } i \in [t_2, s_2] \\ 0, & \text{otherwise} \end{cases}$$

and extend the definition to the rest of $\mathbb{N}$ in such a way that a is periodic with period s_2. It is not to hard to see that a is in $C^{(n)}$. □

Define the function $f_n(a) := \exp(\vartheta \sum_{i=0}^{n-1} a_i \rho T/n^2)$ for each $n \in \mathbb{N}$ and $a \in \mathbb{D}_{(n)}$.

Corollary 1 *For each $n \in \mathbb{N}$, the supremum of $\mathbb{E}_\nu f_n$ over $\nu \in \mathcal{P}^{(n)}$ is the same as its supremum over $\nu \in \mathcal{C}^{(n)}$.*

Proof. An Ergodic Theory result [7, Prop. 21.8] allows us to conclude from the previous lemma that $\mathcal{P}^{(n)}$ is dense in $\mathcal{C}^{(n)}$. This means that, for any $\nu \in \mathcal{C}^{(n)}$, there is a sequence $(\nu_i)_{i \in \mathbb{N}}$ in $\mathcal{P}^{(n)}$ converging to ν. Since the functional f_n is bounded and continuous, $\mathbb{E}_{\nu_i} f_n$ converges to $\mathbb{E}_\nu f_n$ as $i \to \infty$. The conclusion follows. □

The proof of the following theorem is based on the fact that, as the discretization parameter n goes to infinity, the discrete model approaches the continuous one.

Theorem 1 *We have that* $\sup\{\mathcal{E}_{\vartheta,T}(\mu) : \mu \in \mathcal{C}\} = \sup\{\mathcal{E}_{\vartheta,T}(\mu) : \mu \in \mathcal{P}_p, p > 0\}$.

Proof. We define the *discretizing map* $\Phi_n : \mathbb{D} \to \mathbb{D}_{(n)}$ by

$$(\Phi_n(a))_i := \left\lfloor \frac{n^2}{\rho T} a\Big[\frac{iT}{n}, \frac{(i+1)T}{n}\Big] \right\rfloor,$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x. For all $a \in \mathbb{D}$,

$$\left| a[0,T] - \frac{\rho T}{n^2} \sum_{i=0}^{n-1} (\Phi_n(a))_i \right| \leq \frac{\rho T}{n}.$$

Therefore, $|f(a) - f_n(\Phi_n(a))| \leq \vartheta e^{\vartheta \rho T} \rho T/n$ for all $a \in C$, where $f(a) := \exp(\vartheta a[0,T])$.

The map Φ_n can be extended to act on measures: $\Phi_n(\mu)$ is defined by $[\Phi_n(\mu)][\mathcal{B}] := \mu[\Phi_n^{-1}(\mathcal{B})]$ for each Borel subset $\mathcal{B}$ of $\mathbb{D}_{(n)}$. Note that if $\mu \in \mathcal{C}$, then $\Phi_n(\mu) \in \mathcal{C}^{(n)}$. It follows that

$$\sup\{\mathbb{E}_\mu f : \mu \in \mathcal{C}\} \leq \sup\{\mathbb{E}_\nu f_n : \nu \in \mathcal{C}^{(n)}\} + \vartheta e^{\vartheta \rho T} \frac{\rho T}{n}. \tag{5}$$

Now we define the *continuizing map* $\Psi_n : \mathbb{D}_{(n)} \to \mathbb{D}$ by

$$(\Psi_n(a))\Big[\frac{iT}{n}, \frac{iT}{n} + t\Big] := t a_i \frac{\rho}{n}, \qquad \text{for } t \in [0, T/n], i \in \mathbb{N}.$$

Suppose that $\nu \in \mathcal{P}^{(n)}$ has period $p \in \mathbb{N}$. Then ν has mass $1/p$ on elements $a, \theta a, \ldots, \theta^{p-1} a$ of $C^{(n)}$. So, $\mathbb{E}_\nu f_n = p^{-1} \sum_{i=0}^{p-1} f_n(\theta^i a)$. Now consider the source obtained by taking $\Psi_n(a)$ and choosing the phase uniformly in the interval $[0, pT/n]$:

$$\mu[\mathcal{B}] := \frac{n}{pT} \operatorname{Leb}\{t \in [0, pT/n) : \Theta_t \Psi_n(a) \in \mathcal{B}\}$$

for all measurable subsets $\mathcal{B}$ of $\mathbb{D}$. We have

$$\mathbb{E}_\mu f := \frac{n}{pT} \int_0^{pT/n} e^{\vartheta \Psi_n(a)[t, t+T]} \, dt.$$

At each time $t = iT/n$ with $i \in \mathbb{N}$, the integrand is equal to $f_n(\theta^i a)$. Between any two such times, it is either non-increasing or non-decreasing. Therefore

$$|\mathbb{E}_\mu f - \mathbb{E}_\nu f_n| \leq \frac{1}{2} \max_i |f_n(\theta^i a) - f_n(\theta^{i+1} a)| \leq \vartheta e^{\vartheta \rho T} \frac{\rho T}{n}. \tag{6}$$

Therefore

$$\sup\{\mathbb{E}_\nu f_n : \nu \in \mathcal{P}^{(n)}\} \leq \sup\{\mathbb{E}_\mu f : \mu \in \mathcal{P}_p, p > 0\} + \vartheta e^{\vartheta \rho T} \frac{\rho T}{n}. \tag{7}$$

Combining this with Inequality (5) and Corollary 1, and using the fact that n is arbitrary, we get that $\sup\{\mathcal{E}_{\vartheta,T}(\mu) : \mu \in \mathcal{C}\} \leq \sup\{\mathcal{E}_{\vartheta,T}(\mu) : \mu \in \mathcal{P}_p, \, p > 0\}$. The opposite inequality is trivial. $\square$

3 Optimization with Fixed Period

In this section we consider the problem of maximizing $\mathcal{E}_{\vartheta,T}(\mu)$ over $\mu \in \mathcal{P}_p$ for fixed p. Recall that $\mathcal{P}_p$ is the set of traffic sources formed by taking a realization of period p that meets the constraints, and choosing the phase uniformly. We first develop some notation that will be helpful in this setting. For $t_0, t_1 \in [0, p]$, define

$$[t_0, t_1]_p := \begin{cases} [t_0, t_1], & \text{if } t_0 \leq t_1 \\ [t_0, p] \cup [0, t_1], & \text{otherwise.} \end{cases}$$

If either t_0 or t_1 is outside the range $[0, p]$, then $[t_0, t_1]_p$ is interpreted to mean $[t_0 (\bmod\ p), t_1 (\bmod\ p)]_p$. We define $(t_0, t_1]_p := [t_0, t_1]_p - \{t_0\}$. The sets $(t_0, t_1)_p$ and $[t_0, t_1)_p$ are defined similarly. Another useful piece of notation will be $t_1 \ominus t_0 := \operatorname{Leb}[t_0, t_1]_p$.

We define $\mathbb{D}_p$ to be the set of non-negative, finite measures on $[0, p]$. The set of periodic realizations with period p may be identified with the set $\mathbb{D}_p$ modulo the equivalence relation $\sim$, where $a \sim b$ if there exists $h \in \mathbb{R}$ such that $a(t_0, t_1)_p = b(t_0 + h, t_1 + h)_p$ for all $t_0, t_1 \in [0, p]$. Motivated by this identification, we define the queue length at time $t \in [0, p]$, with arrivals $a \in \mathbb{D}_p$ and service rate $s \geq 0$, to be

$$W_s(a, t) := \begin{cases} \sup_{t_1 \in [0, p]} \Big(a[t_1, t]_p - (t \ominus t_1) s \Big), & \text{if } a[0, p] \leq sp, \\ +\infty, & \text{otherwise.} \end{cases} \tag{8}$$

Constraints (2) and (3) translate into the condition that an element of $\mathbb{D}_p$ be in the set

$$C_p := \{a \in \mathbb{D}_p : W_\sigma(a, t) \leq \beta \text{ and } W_\rho(a, t) = 0, \text{ for all } t \in [0, p]\}.$$

If μ is the periodic source identified with the equivalence class of $a \in \mathbb{D}_p$, then the effective bandwidth of μ is $(\vartheta T)^{-1} \log E(a)$, where

$$E(a) := \frac{1}{p} \int_0^p e^{\vartheta a[t, T+t]_p} \, dt. \tag{9}$$

Note that if $a \sim b$ then $E(a) = E(b)$. The problem of maximizing the effective bandwidth over $\mathcal{P}_p$ is the same as that of maximizing $E(a)$ over $a \in C_p$.

We wish to introduce the following supplementary constraint: $a[0,p] = \sigma p$. We denote by C_p^{full} those elements of C_p that satisfy it. This condition requires that the mean rate be equal to the leaky bucket rate. In [22], it was shown that if $a \in C_p$ then there exists some $b \in C_p^{\mathrm{full}}$ such that $E(b) \geq E(a)$. Thus, including this constraint does not change the solution of the optimization problem.

We consider $\mathbb{D}_p$ to be embedded in the linear space $\overline{\mathbb{D}}_p$, the space of finite signed measures on $[0,p]$. Here addition and multiplication by scalars are set-wise. With this linear structure, E is obviously a convex function.

From the convexity of E and the fact that it is bounded above in the neighborhood of any point, it follows that E is continuous [18, p. 93].

We will also need some properties of the constraint set C_p^{full}. For each $s \geq \sigma$, the functional $\sup_{t \in [0,p]} W_s(\cdot, t)$ is convex on $\mathbb{D}_p$ since it can be written as the supremum of a family of affine functions. Again, from this and the fact that it is bounded above in the neighborhood of any point, we conclude that it is continuous. Therefore C_p^{full} is closed and convex. The peak rate constraint implies that this set is bounded, and thus compact.

We denote by $\overline{\operatorname{conv}}\, S$ the closed convex hull of a set S. By the Krein-Milman theorem, $C_p^{\mathrm{full}} = \overline{\operatorname{conv}}\, \xi_p$, where ξ_p is the set of extreme points of C_p^{full}. Recall that a point q is said to be an extreme point of a set S if $q \in S$ and q is not an interior point of any line segment in S.

Theorem 2 $\sup\{\mathcal{E}_{\vartheta,T}(\mu) : \mu \in \mathcal{C}\} = \sup\{(\vartheta T)^{-1} \log E(a) : a \in \xi_p,\, p > 0\}$

Proof. Since E is convex on C_p^{full}, its supremum over ξ_p the same as its supremum over the convex hull of ξ_p. Since it is continuous, this second supremum is the same as the supremum over $\overline{\operatorname{conv}}\, \xi_p$. But we have seen above that $\overline{\operatorname{conv}}\, \xi_p$ is equal to C_p^{full}. The proof is complete when we apply Theorem 1. $\square$

4 The Extreme Points of the Constraint Set

Recall the definition of W_s given in Equation (8). We also define

$$U_s(a,t) := \begin{cases} \sup\limits_{t_1 \in [0,p]} \big(a[t,t_1]_p - (t_1 \ominus t)s\big), & \text{if } a[0,p] \leq sp, \\ +\infty, & \text{otherwise.} \end{cases}$$

for $t \in [0,p]$, $a \in \mathbb{D}_p$, and $s \geq 0$.

Theorem 3 *A point $a \in C_p^{full}$ is an extreme point if and only if for almost all $t \in [0,p]$ one of the following hold:*

(X1) a is differentiable at t and $Da(t) = \rho$,

(X2) a is differentiable at t and $Da(t) = 0$,

(X3) $W_\sigma(a,t) = \beta$,

(X4) $U_\sigma(a,t) = \beta$.

Remark 1 *Note that if $a \in C_p^{full}$ obeys either (X3) or (X4) at an instant t and a is differentiable there, then $Da(t) = \sigma$.*

Proof. a is extreme $\implies$ condition holds. Let $a \in C_p^{\text{full}}$. Suppose that there is a set $S \subset [0,p]$ with positive Lebesgue measure such that none of the four conditions *(X1–4)* hold at any point in S. We shall construct $\overline{a}$ and $\underline{a}$ in C_p^{full} such that $a = (\overline{a} + \underline{a})/2$.

Since $W_\sigma(a,\cdot)$ and $U_\sigma(a,\cdot)$ are continuous on $[0,p]$, the set $G := \{t \in [0,p] : W_\sigma(a,t) < \beta, U_\sigma(a,t) < \beta\}$ is open (in $[0,p]$). Therefore it can be expressed as a countable disjoint union of open intervals $G_i; i \in \mathbb{N}$. There must be some $j \in \mathbb{N}$ such that $\text{Leb}[G_j \cap S] > 0$, for otherwise $G \cap S$, which is equal to S, would have Lebesgue measure zero.

In what follows it will be convenient to work on a compact interval. It is obvious that there exists a compact interval $K \subset G_j$ such that $\text{Leb}[K \cap S] > 0$. Furthermore, we can find two disjoint compact intervals $K_1, K_2 \subset K$, such that both $\text{Leb}(K_1 \cap S)$ and $\text{Leb}(K_2 \cap S)$ are positive. We may take K_1 to be on the left of K_2 in K; that is if $t_1 \in [0,p] - K$ and $[t_1,t]_p$ contains a point of K_2, then $K_1 \subset [t_1,t]_p$. Since K is compact and $W_\sigma(a,\cdot)$ and $U_\sigma(a,\cdot)$ are continuous, these two functions attain their supremum over K. Therefore $\omega := \sup_{t\in K} W_\sigma(a,t)$ and $\upsilon := \sup_{t\in K} U_\sigma(a,t)$ are both strictly less than β.

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

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

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

Note that $\overline{r}_\delta(x) \ge x$ and $\underline{r}_\delta(x) \le x$ for all $x \in [0,\rho]$, and that $(\overline{r}_\delta + \underline{r}_\delta)/2 = \text{Id}$, where Id is the identity function. Both maps leave 0 and ρ unchanged. Let

$$k_1(\delta) := \int_{K_1} ((\overline{r}_\delta - \text{Id}) \circ Da)(x)\, dx$$

$$\text{and} \quad k_2(\delta) := \int_{K_2} ((\overline{r}_\delta - \text{Id}) \circ Da)(x)\, dx.$$

Here I_A denotes the indicator function which takes the value 1 at x if $x \in A$ and the value 0 otherwise. Both k_1 and k_2 are continuous and strictly increasing, and therefore invertible in a neighborhood of 0. Choose ϵ in this neighborhood such that $0 < \epsilon < \min(\beta - \omega, \beta - \upsilon)$. Then we can find $\delta_1, \delta_2 \in (0,1)$ such that $k_1(\delta_1) = k_2(\delta_2) = \epsilon$.

Write $K' := [0,p] - (K_1 \cup K_2)$ and consider the measures $\overline{a}$ and $\underline{a}$ defined by

$$\overline{a}[\mathcal{B}] := \int_{\mathcal{B}} \Big(Da.\mathrm{I}_{K'} + (\overline{r}_{\delta_1} \circ Da).\mathrm{I}_{K_1} + (\underline{r}_{\delta_2} \circ Da).\mathrm{I}_{K_2}\Big)(x)\, dx \tag{10}$$

$$\text{and} \quad \underline{a}[\mathcal{B}] := \int_{\mathcal{B}} \Big(Da.\mathrm{I}_{K'} + (\underline{r}_{\delta_1} \circ Da).\mathrm{I}_{K_1} + (\overline{r}_{\delta_2} \circ Da).\mathrm{I}_{K_2}\Big)(x)\, dx. \tag{11}$$

Clearly $a = (\overline{a} + \underline{a})/2$. Since $\overline{r}_\delta(x) > x$ for all $x \in (0,\rho)$, and K_1 and K_2 each contain a set of positive measure in which $0 < Da < \rho$, we have that $\overline{a}[K_1] > a[K_1]$ and that $\underline{a}[K_1] < a[K_1]$. Thus a, $\overline{a}$, and $\underline{a}$ are distinct. We will show that both $\overline{a}$ and $\underline{a}$ are in C_p^{full}.

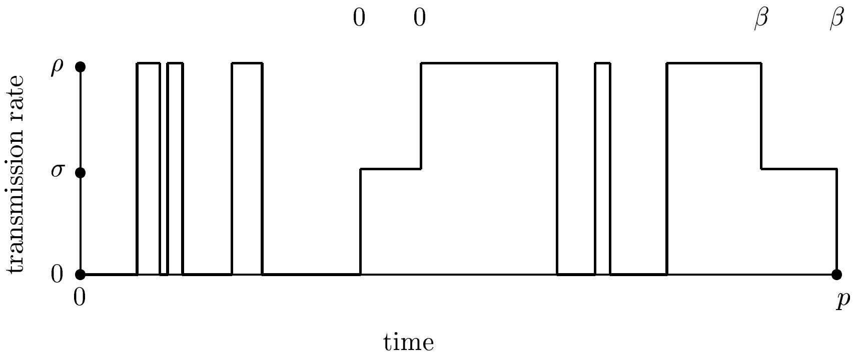

Figure 2: A typical extreme point of C_p^{full}. The values above the figure give the content of the leaky bucket at that time.

Since neither $\overline{r}_\delta$ nor $\underline{r}_\delta$ take values outside $[0,\rho]$, we have that $0 \le D\overline{a}(x) \le \rho$ for almost all $x \in [0,p]$. We conclude that $\overline{a}$ is a non-negative measure on $[0,p]$ and satisfies the peak rate constraint. The same is true for $\underline{a}$.

Let $t_1, t_2 \in [0,p]$ and write $\kappa := (t_2 \ominus t_1)\sigma$. If $t_2 \in [0,p]-K$, then $\overline{a}[t_1,t_2]_p \le a[t_1,t_2]_p \le \beta+\kappa$. If, on the other hand, $t_2 \in K$ then $a[t_1,t_2]_p \le \omega + \kappa$ and therefore $\overline{a}[t_1,t_2]_p \le a[t_1,t_2]_p + \epsilon \le \omega + \kappa + \epsilon < \beta + \kappa$. We conclude that $\overline{a}$ obeys the leaky bucket constraint.

A similar argument shows that $\underline{a}$ also obeys the leaky bucket constraint.

condition holds $\Longrightarrow$ a is extreme. Suppose that a obeys at least one of *(X1–4)* at almost every time $t \in [0,p]$ and that $a = (\overline{a}+\underline{a})/2$ for two traffic realizations $\overline{a}$ and $\underline{a}$ in C_p^{full}. Since a, $\overline{a}$, and $\underline{a}$ are non-negative measures, the derivatives of all three exist almost everywhere. Let $t \in [0,p]$ be such that a obeys one of the conditions *(X1–4)* at t, and $\overline{a}$ and $\underline{a}$ are both differentiable at t. Depending on which condition a obeys at t, we have one of the following four cases.

- $Da(t) = 0$. In this case $D\overline{a}(t) = D\underline{a}(t) = 0$, since $Da(t) = [D\overline{a}(t) + D\underline{a}(t)]/2$ and both $D\overline{a}(t)$ and $D\underline{a}(t)$ lie in the interval $[0,\rho]$.
- $Da(t) = \rho$. Similarly, we must have that $D\overline{a}(t) = D\underline{a}(t) = \rho$.
- $W_\sigma(a,t) = \beta$. In this case $W_\sigma(\overline{a},t) = W_\sigma(\underline{a},t) = \beta$ since $W_\sigma(a,t) \le (W_\sigma(\overline{a},t) + W_\sigma(\underline{a},t))/2$ and both $W_\sigma(\overline{a},t)$ and $W_\sigma(\underline{a},t)$ are no greater than β.
- $U_\sigma(a,t) = \beta$. Similarly, $U_\sigma(\overline{a},t) = U_\sigma(\underline{a},t) = \beta$.

Thus, if a obeys any one of the four conditions at t, then $\overline{a}$ and $\underline{a}$ also obey the same condition at t. We conclude that $Da = D\overline{a} = D\underline{a}$ 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^{full} and is therefore an extreme point of this set. □

At each instant, a source with realizations in ξ_p may only transmit at one of three rates: zero, the peak rate ρ, or the leaky bucket service rate σ. The source can transmit at rate zero or ρ at any time, but may transmit at rate σ only when the leaky bucket is either empty of full. Figure 2 shows a typical extreme point of C_p^{full}. The figures 0 and β above the diagram indicate, respectively, the times when the leaky bucket is empty and full.

We see that C_p^{full} has many extreme points. In fact we will show that the extreme points form a dense subset of C_p^{full}. Heuristically, any source behavior can be approximated by a source transmitting at rates zero and ρ. All that is needed is for the source to switch between these two rates sufficiently quickly and to spend the appropriate proportion of time transmitting at each. This result means that we did not need to use the convexity of the functional E to show that it approaches its supremum on the set ξ_p; we could have deduced this from its continuity. Thus, Theorem 3 does not reduce the domain of the optimization problem significantly. However, in the next section we will see how, in another linear structure on the space of traffic processes, convexity is crucial and enables us to make a much more meaningful reduction.

Proposition 1 *The constraint set C_p^{full} is the closure of the set of its extreme points ξ_p.*

Proof. Let $n \in \mathbb{N}$ be such that $\beta > 6p\rho/n$. Define the realization b^n by $b^n[\mathcal{B}] := \rho \operatorname{Leb}\big[\mathcal{B} \cap \bigcup_{i=0}^{n-1} B_i^n\big]$, where

$$B_i^n := \begin{cases} \Big[\dfrac{ip}{n}, \dfrac{ip}{n} + \dfrac{a_i^n}{\rho}\Big], & \text{if } W_\sigma(a,t) \geq \beta/2 \text{ for some } t \in \Big[\dfrac{ip}{n}, \dfrac{(i+1)p}{n}\Big], \\ \Big[\dfrac{(i+1)p}{n} - \dfrac{a_i^n}{\rho}, \dfrac{(i+1)p}{n}\Big], & \text{otherwise} \end{cases}$$

and $a_i^n := a[ip/n, (i+1)p/n]$.

Clearly, b^n satisfies the peak rate constraint and the conditions of Theorem 2. Also $b^n[0,p] = p\sigma$. It remains to show that b^n satisfies the leaky bucket constraint.

Let $t_0, t_1 \in [0,p]$. Then $|b[t_0,t_1]_p - a[t_0,t_1]_p| \leq 2p\rho/n$. We wish to show that $b[t_0,t_1]_p \leq \beta + (t_1 \ominus t_0)\sigma$. If $a[t_0,t_1]_p - (t_1 \ominus t_0)\sigma \leq \beta - 2p\rho/n$, this follows immediately. So suppose the contrary. Then $W_\sigma(a,t_1) \geq \beta/2$ and $W_\sigma(a,t_0) \leq \beta/2$. We conclude that $b[t_0,t_1]_p \leq a[t_0,t_1]_p$, which implies the required statement.

It is easy to verify that, as $n \to \infty$, the difference between a and b^n goes to zero on every set. $\square$

5 Another Linear Structure

In this section we consider another way of representing traffic realizations. Here the arrival time of each parcel of traffic is specified, rather than the amount of traffic that arrives within each interval. This representation leads to a linear structure on the set of realizations which is different from the one considered in the previous two sections. It turns out that the functional E is 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 optimization must be performed.

Let

$$\mathbb{D}_{p,q} := \{a \in \mathbb{D}_p : a[0,p] = q\}.$$

We define the map $\triangleleft : \mathbb{D}_{p,q} \to \mathbb{D}_{q,p}$ by

$$(\triangleleft a)[0,c] := \sup\{t \in [0,p] : a[0,t] \leq c\}, \qquad \text{for } c \in [0,q].$$

Note that $\triangleleft$ is a bijection and is its own inverse. If the distribution function of a, which is given by $a[0,\cdot]$, is strictly increasing and continuous, then so is the distribution function of $\triangleleft a$ and the two functions are inverse to one another.

In this setting, the constraint set of interest is $\cup_{q>0} C_q^{\triangleleft}$, where $C_q^{\triangleleft} := \{d \in \mathbb{D}_{q,q/\sigma} : \triangleleft d \in C_{q/\sigma}^{\text{full}}\}$. Note that now elements of the constraint set may have atoms. These represent time intervals during which the source is silent. We wish to maximize the functional $E^{\triangleleft}(d) := E(\triangleleft d)$ over this set. Note that if $a \in C_p^{\text{full}}$ and $d = \triangleleft a$, then

$$\sup_{t\in[0,p]} W_s(a,t) = \sup_{c_1,c_2\in[0,q]} \Big(c_2 \ominus c_1 - d[c_1,c_2]_q s \Big), \qquad \text{for all } s \geq p/q.$$

It follows that $C_q^{\triangleleft}$ is convex.

It is shown in [16, Lemma A4] that the map $\triangleleft$ is a homeomorphism between $\mathbb{D}_{p,q}$ and $\mathbb{D}_{q,p}$. An immediate corollary is that the functional $E^{\triangleleft}$ is continuous on $\mathbb{D}_{q,p}$ and $C_q^{\triangleleft}$ is closed.

Establishing convexity of the functional to be maximized was trivial in the linear structure of the previous two sections, but here it is considerably more difficult. We will first establish the convexity of $E^{\triangleleft}$ in the space of *simple* realizations and then extend the result to $\mathbb{D}_{q,p}$ using a continuity argument. By a simple realization, we mean one that can be written

$$a = \sum_{i=1}^{n} m_i \delta_{t_i},$$

where $n \in \mathbb{N}$, and $t_i \in [0,q]$ and $m_i \in [0,p]$ for each $i \in \{1,\dots,n\}$. We denote by $\mathbb{A}_{q,p}$ the set of elements of $\mathbb{D}_{q,p}$ that can be written in this form. Note that if a is in $\mathbb{A}_{p,q}$ then $\triangleleft a$ is in $\mathbb{A}_{q,p}$. Also note that $\mathbb{A}_{q,p}$ is a convex subset of $\mathbb{D}_{q,p}$.

Simple realizations are useful because they may be used to approximate any other realization, in other words $\mathbb{A}_{q,p}$ is dense in $\mathbb{D}_{q,p}$ for each $p, q > 0$.

Lemma 2 *The functional $E^{\triangleleft}$ is convex on $\mathbb{A}_{q,p}$.*

Proof. Write $g(x) := \exp(\vartheta x)$. The only property of g we will use will be its convexity. It will suffice to show that the one-dimensional function

$$f : [0,1] \to \mathbb{R} : \alpha \mapsto E^{\triangleleft}((1-\alpha)d_1 + \alpha d_2)$$

is convex for any two points d_1 and d_2 in $\mathbb{A}_{q,p}$. For $\alpha \in (0,1)$, let a_α be the realization in $\mathbb{A}_{p,q}$ such that $\triangleleft a_\alpha := (1-\alpha)d_1 + \alpha d_2$. This realization may be written as $a_\alpha = \sum_{i=1}^n m_i \delta_{t_i(\alpha)}$, where the m_i are constant and the t_i are affine in α. We write v_i for the slope of t_i.

Denote by J_α the set of ordered pairs $(c_1,c_2) \in \{1,\dots,n\}^2$ such that $t_{c_2}(\alpha) \ominus t_{c_1}(\alpha) = T$ and $v_{c_1} \neq v_{c_2}$. Also, denote by K the set of ordered pairs (c_1,c_2) such that $t_{c_2}(\alpha) \ominus t_{c_1}(\alpha) = T$ and $v_{c_1} = v_{c_2}$. The set K needs no subscript since it does not change with α.

There are only a finite number of values of α for which J_α is nonempty. For any other value of α, let $\eta \in (0,T)$ be such that $2\max_c |v_c \eta|$ is less than both

$$\min_{(c_1,c_2)\notin K} \Big[t_{c_2}(\alpha) \ominus (t_{c_1}(\alpha) + T) \Big] \qquad \text{and} \qquad \min_{c_1,c_2\in\{1,\dots,n\}^2} \Big[t_{c_2}(\alpha) \ominus t_{c_1}(\alpha) \Big].$$

Consider first the case when K is empty. Then, for each set $[x, x+T]_p$, with $x \in [0,p]$, there may be at most one $i \in \{1,\dots,n\}$ such that $t_i(\gamma)$ passes across the boundary of the set as γ ranges from α to $\alpha + \eta$. We conclude that

$$f(\alpha+\eta) - f(\alpha) = \sum_{c=1}^{n} v_c \eta \Big[g(z_c + m_c) - g(z_c) + g(y_c) - g(y_c + m_c) \Big], \tag{12}$$

where $z_i(\alpha) := a(t_i, t_i + T)_p$ and $y_c(\alpha) := a(t_i - T, t_i)_p$. Here we are suppressing the dependence on α for clarity.

Now suppose that K is non-empty and contains a pair (c_1, c_2). In this case $z_{c_1}(\alpha) = y_{c_2}(\alpha)$ for all $\alpha \in (0,1)$ and $v_{c_1} = v_{c_2}$. Without loss of generality, we may take v_{c_1} to be positive. The contribution made by the intervals $\{[x, x+T]_p : t_{c_1} \leq x \leq t_{c_1} + v_{c_1}\eta\}$ to $f(\alpha)$ is $v_{c_1}\eta\, g(z_{c_1} + m_{c_2})$, and to $f(\alpha + \eta)$ is $v_{c_1}\eta\, g(z_{c_1} + m_{c_1})$. The increase is equal to

$$v_{c_1}\eta\big[g(z_{c_1} + m_{c_1}) - g(z_{c_1})\big] - v_{c_2}\eta\big[g(y_{c_2} + m_{c_2}) - g(y_{c_2})\big].$$

Now we sum over $(c_1, c_2) \in K$ and add, as before, the contributions from those sets $x, x+T]_p$; $x \in [0,p]$ for which a single t_i crosses the boundary. We find that Equation (12) holds in case also.

We conclude from the arguments above that f is piecewise affine and that changes in its slope occur at those values of α for which J_α is non-empty. Now fix α at such a value. For each pair $(c_1, c_2) \in J_\alpha$, let $z_{c_1 c_2}(\alpha) := z_{c_1}(\alpha) = y_{c_2}(\alpha)$. Then the change in the derivative of f at α is given by

$$\begin{aligned}
\Delta f'(\alpha) &= \sum_{(c_1,c_2)\in J_\alpha} v_{c_1}\,\mathrm{sgn}(v_{c_1} - v_{c_2})\Big[g(z_{c_1c_2} + m_{c_1} + m_{c_2}) - g(z_{c_1c_2} + m_{c_1}) \\
&\qquad\qquad - g(z_{c_1c_2} + m_{c_2}) + g(z_{c_1c_2})\Big] \\
&\quad + \sum_{(c_1,c_2)\in J_\alpha} v_{c_2}\,\mathrm{sgn}(v_{c_1} - v_{c_2})\Big[g(z_{c_1c_2} + m_{c_1}) - g(z_{c_1c_2}) \\
&\qquad\qquad - g(z_{c_1c_2} + m_{c_1} + m_{c_2}) + g(z_{c_1c_2} + m_{c_2})\Big] \\
&= \sum_{(c_1,c_2)\in J_\alpha} |v_{c_1} - v_{c_2}|\Big[g(z_{c_1c_2} + m_{c_1} + m_{c_2}) - g(z_{c_1c_2} + m_{c_1}) \\
&\qquad\qquad - g(z_{c_1c_2} + m_{c_2}) + g(z_{c_1c_2})\Big],
\end{aligned}$$

where $\mathrm{sgn}(x)$ is 1, 0, or -1, depending on whether x is positive, zero, or negative. From the convexity of g we have that

$$g(z_{c_1c_2} + m_{c_1}) \leq \frac{m_{c_1}}{m_{c_1} + m_{c_2}} g(z_{c_1c_2} + m_{c_1} + m_{c_2}) + \frac{m_{c_2}}{m_{c_1} + m_{c_2}} g(z_{c_1c_2})$$

and

$$g(z_{c_1c_2} + m_{c_2}) \leq \frac{m_{c_2}}{m_{c_1} + m_{c_2}} g(z_{c_1c_2} + m_{c_1} + m_{c_2}) + \frac{m_{c_1}}{m_{c_1} + m_{c_2}} g(z_{c_1c_2}).$$

Therefore $\Delta f'(\alpha) \geq 0$, and so f is convex. □

Lemma 3 *Let S be a convex subset of a metrizable topological vector space and let $D \subset S$ be convex and dense in S. If $f : S \to \mathbb{R}$ is a continuous function and its restriction to D is convex, then f is also convex on S.*

Proof. Let $x, y \in S$ and let $\alpha \in [0,1]$. Then there exist sequences x_n and y_n in D such that $x_n \to x$ and $y_n \to y$. Consider the sequence $w_n := (1-\alpha)x_n + \alpha y_n$ which lies in D since this set is convex. The operations of addition and multiplication by a scalar are continuous, and so this sequence must approach the limit $w := (1-\alpha)x + \alpha y$. Now, the continuity of f implies that $f(x_n) \to f(x)$, that $f(y_n) \to f(y)$, and that $f(w_n) \to f(w)$. Furthermore, the convexity of f implies that $f(w_n) \leq (1-\alpha)f(x_n) + \alpha f(y_n)$ for all $n \in \mathbb{N}$. It follows that $f(w) \leq (1-\alpha)f(x) + \alpha f(y)$, and therefore f is convex on S. □

We can now extend the convexity result to the whole of $\mathbb{D}_{q,p}$.

Theorem 4 *The functional $E^{\triangleleft}$ is convex on $\mathbb{D}_{q,p}$.*

Proof. We have shown that $E^{\triangleleft}$ is continuous on $\mathbb{D}_{q,p}$ and convex on a dense convex subset $\mathbb{A}_{q,p}$ of $\mathbb{D}_{q,p}$. The conclusion follows from an application of the previous lemma. $\square$

The following theorem, combined with Theorem 6, is the main result of this paper.

Theorem 5 *The worst case effective bandwidth* $\sup\{\mathcal{E}_{\vartheta,T}(\mu) : \mu \in \mathcal{C}\}$ *is equal to*

$$\sup\left\{\frac{1}{\vartheta T}\log E^{\triangleleft}(d) : q > 0, d \in \varphi_q\right\},$$

where φ_q is the set of extreme points of $C_q^{\triangleleft}$.

Proof. The proof of this result is analogous to that of Theorem 2. $\square$

6 Extreme Points in the Alternative Representation

We will need the following definitions. Given $d \in \overline{\mathbb{D}}_{q,p}$, write

$$z(c_1, c_2)_q := c_2 \ominus c_1 - \sigma d(c_1, c_2)_q, \qquad \text{for } c_1, c_2 \in [0, q]. \tag{13}$$

For $d \in \overline{\mathbb{D}}_{q,p}$ and $c \in [0, q]$, we define

$$\begin{aligned} W^{\triangleleft}(d, c) &:= \sup_{c_1 \in [0,q]} z(c_1, c)_q \\ \text{and} \quad U^{\triangleleft}(d, c) &:= \sup_{c_1 \in [0,q]} z(c, c_1)_q. \end{aligned}$$

In contrast to the elements of C_p^{full}, an element of $C_q^{\triangleleft}$ may have a singular part. Recall [19] that every measure d can be written as the sum of a singular measure d_s and an absolutely continuous measure d_a. Also, the support of a measure is the intersection of all closed subsets of full measure.

We have the following characterization of extreme points of $C_q^{\triangleleft}$.

Theorem 6 *$d \in C_q^{\triangleleft}$ is an extreme point if and only if both the following conditions hold:*

- *(Y) for almost every $c \in [0, q]$, one of the following hold*
  - *(Y1) d is differentiable at c and $Dd(c) = 1/\rho$,*
  - *(Y2) $W^{\triangleleft}(d, c) = \beta$,*
  - *(Y3) $U^{\triangleleft}(d, c) = \beta$.*
- *(Z) for each pair of distinct points c_1 and c_2 in the support of d_s, the supremum over the set $(c_1, c_2)_q$ of either $W^{\triangleleft}(d, \cdot)$ or of $U^{\triangleleft}(d, \cdot)$ is equal to β.*

Condition (Y) says that, like the extreme points of C_p^{full}, an extreme point of F_q may only transmit at rates ρ, zero, and σ, and may only transmit at the latter rate when the leaky bucket is either empty or full. To gain an understanding of Condition (Z), it is useful to consider those realizations that become silent a finite number of times per period, in other words those for which d_s has finite support. The intervals between these points of support may be considered to be "bursts" of traffic. Since the source may only transmit at rates ρ and σ during the burst, the content of the leaky bucket cannot decrease. Therefore Condition (Z) implies that either the leaky bucket is empty at the start of the burst or full at the end. Of course, we have no control over the number of bursts in any time interval. An example of a realization that is extreme is given in Figure 3.

We will postpone the proof of Theorem 6. Instead, we will now show that the set of extreme points of $C_q^{\lhd}$ is strictly smaller than that of C_p^{full}.

Theorem 7 *If d is an extreme point of $C_q^{\lhd}$ then $\lhd d$ is an extreme point of C_p^{full}, where $p := q/\sigma$.*

Proof. Define $a := \lhd d$ and let $g : [0,p] \to \mathbb{R} : t \mapsto a[0,t]$ be its distribution function. Since

$$\begin{aligned} d[0,c] &= \sup\{t \in [0,p] : g(t) \leq c\} \\ &= \operatorname{Leb} g^{-1}[0,c], \end{aligned}$$

we have that $d[\mathcal{B}] = \operatorname{Leb}[g^{-1}\mathcal{B}]$ for all Borel subsets $\mathcal{B}$ of $[0,q]$.

We define the following subsets of $[0,p]$:

$$\begin{aligned} Z &:= \{t \in [0,p] : Da(t) \text{ exists and is zero}\}, \\ P &:= \{t \in [0,p] : Da(t) \text{ exists and is strictly positive}\}. \end{aligned}$$

Since Da exists almost everywhere, $\operatorname{Leb} Z + \operatorname{Leb} P = p$. Note that g is a bijection from P to $g(P)$. If $t \in P$ and $c := g(t)$ then $g'(t) = Da(t) = 1/(Dd(c))$. We also have that $W(a,t) = W^{\lhd}(d,c)$ and that $U(a,t) = U^{\lhd}(d,c)$. (Recall that W_σ and U_σ were defined in Sections 3 and 4.) It follows that a obeys $(X1)$, $(X3)$, or $(X4)$ at t if and only if d obeys, respectively, $(Y1)$, $(Y2)$, or $(Y3)$ at c.

It remains to show that $\operatorname{Leb}[g^{-1}(K)] = 0$, where

$$K := \{c \in g(P) : \text{none of } (Y1\text{–}3) \text{ hold for } d \text{ at } c\ \}.$$

Let d_a and d_s be the absolutely continuous and singular parts respectively of d. Since K has Lebesgue measure zero, $d_a[K] = 0$. It is well known [19] that any singular measure μ has infinite derivative μ-almost everywhere. Thus $d_s[g(P)] = 0$ since Dd is finite on this set. But $K \subset g(P)$ and so we conclude that $d[K] = 0$. Thus $\operatorname{Leb}[g^{-1}(K)]$ is indeed zero. □

It is easy to find realizations that are extreme in $C_{q/\sigma}^{\text{full}}$ but not in extreme in $C_q^{\lhd}$. For example, if $\rho = 2\sigma = 1$, $q = 5$, and $\beta > 1/4$, then such a realization is given by

$$a[\mathcal{B}] := \operatorname{Leb}\Big[\mathcal{B} \cap \bigcup_{i=0}^{9}[i, i+\frac{1}{2}]\Big], \qquad \text{for all Borel subsets } \mathcal{B} \text{ of } [0,10].$$

This realization switches on at integer times and switches off at half-integer times.

We break the proof of Theorem 6 into the following three lemmas.

Lemma 4 *If d is an extreme point of $C_q^{\lhd}$, then condition (Y) holds.*

Proof. The proof of this lemma is very similar to that of Theorem 3 and we will only describe the main differences. Firstly, $W^{\lhd}(d,\cdot)$ and $U^{\lhd}(d,\cdot)$ are not continuous, only upper-semicontinuous. This does not significantly alter the proof however. Another difference is that the functions $\overline{r}_\delta$ and $\underline{r}_\delta$ should now map the interval $[1/\rho, \infty)$ to itself. A suitable choice of maps is

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

Finally, analogues of Expressions (10) and (11) give only absolutely continuous measures. To find $\overline{d}$ and $\underline{d}$ such that $d = (\overline{d} + \underline{d})/2$, we must add the singular part of d. □

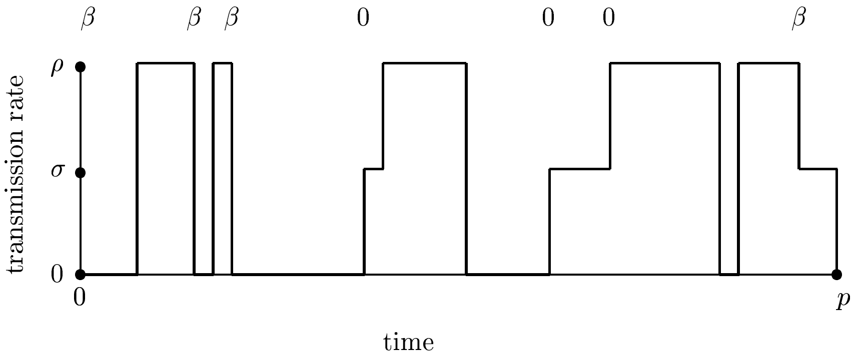

Figure 3: A typical extreme point of $C_q^{\triangleleft}$. The values above the figure give the content of the leaky bucket at that time. Note that each burst must either start with the leaky bucket empty or end with it full.

Lemma 5 *If d is an extreme point of $C_q^{\triangleleft}$, then condition (Z) holds.*

Proof. Let c_1 and c_2 be distinct points in the support of d_s. Suppose $d \in C_q^{\triangleleft}$ is such that both

$$\sup_{c \in (c_1, c_2)_q} W^{\triangleleft}(d, c) \qquad \text{and} \qquad \sup_{c \in (c_1, c_2)_q} U^{\triangleleft}(d, c)$$

are strictly less than γ, where $\gamma < \beta$. We will construct $\overline{d}, \underline{d} \in C_q^{\triangleleft}$ such that $\overline{d}$, d, and $\underline{d}$ are distinct and $(\overline{d} + \underline{d})/2 = d$.

If d_s has no atom at c_1, then both $W^{\triangleleft}(d, \cdot)$ and $U^{\triangleleft}(d, \cdot)$ are continuous there and so there is some open interval I_1 containing c_1 such that both $W^{\triangleleft}(d, c_1)$ and $U^{\triangleleft}(d, c_1)$ are less than γ for all $c \in I_1$. In this case, define a positive measure Δ_1 such that $\Delta_1[\mathcal{B}] := d_s[I_1 \cap \mathcal{B}]$ for all Borel subsets $\mathcal{B}$ of $[0, q]$.

If, on the other hand, d_s has an atom at c_1, then define Δ_1 to be the measure on $[0, q]$ with total mass 1, concentrated at c_1. In this case, $I_1 := \{c_1\}$.

Replacing c_1 with c_2 in the definition above, we similarly define an interval I_2 and a positive measure Δ_2. Note that we may choose I_1 and I_2 in such a manner that they do not overlap.

Choose $\delta_1 > 0$ and $\delta_2 > 0$ such that $\delta_1 \Delta_1[0, q] = \delta_2 \Delta_2[0, q]$ and such that this number is less than $\beta - \gamma$. We now define

$$\begin{aligned} \overline{d} &:= d - \delta_1 \Delta_1 + \delta_2 \Delta_2 \\ \text{and} \quad \underline{d} &:= d + \delta_1 \Delta_1 - \delta_2 \Delta_2. \end{aligned}$$

Let $x, y \in [0, q]$. Suppose y is in the interior of $I_1 \cup (c_1, c_2)_q \cup I_2$. Then, $z(x, y)_q \leq W^{\triangleleft}(d, y) < \gamma$ and we have that $\overline{z}(x, y)_q \leq z(x, y)_q + \beta - \gamma < \beta$. Suppose, on the other hand, that y is not in the interior of that set. Then $\overline{z}(x, y)_q \leq z(x, y)_q \leq \beta$. Either way, we conclude that $\overline{d}$ is in $C_q^{\triangleleft}$.

A similar argument shows that $\underline{d}$ is also in $C_q^{\triangleleft}$.

□

Lemma 6 *If $d \in C_q^{\triangleleft}$ satisfies Conditions (Y) and (Z), then d is an extreme point of $C_q^{\triangleleft}$.*

Proof. Assume that d satisfies the given conditions and that $d = (\overline{d} + \underline{d})/2$ with $\overline{d}, \underline{d} \in C_q^{\triangleleft}$. Let $c \in [0, q]$ be such that both $\overline{d}$ and $\underline{d}$ are differentiable at c. The argument proceeds as in the proof of Theorem 3. We show that if d satisfies one of the conditions $(Y1\text{–}3)$ at c, then both $\overline{d}$ and $\underline{d}$ satisfy the same condition. Since both $(Y2)$ and $(Y3)$ imply that the derivative at c is $1/\sigma$, we conclude that $Dd = D\overline{d} = D\underline{d}$ almost everywhere. Thus the absolutely continuous parts of d, $\overline{d}$, and $\underline{d}$ are equal.

Let $\mathcal{W}$ and $\mathcal{U}$ be defined by

$$\begin{aligned} \mathcal{W} &:= \{y \in [0,q] : W^{\triangleleft}(d,y) = \beta\} \\ \mathcal{U} &:= \{x \in [0,q] : U^{\triangleleft}(d,x) = \beta\}. \end{aligned}$$

Suppose x and y are such that $x \in \mathcal{U}$ and $y \in \mathcal{W}$. Then $z(x,y)_q = \beta$. To see this, observe that there exists $x' \in \mathcal{U}$ and $y' \in \mathcal{W}$ such that $z(x,y')_q = z(x',y)_q = \beta$. Then, since $z[y',x']_q \geq -\beta$,

$$\beta \geq z(x,y)_q = z(x,y')_q + z[y',x']_q + z(x',y)_q \geq \beta.$$

Now define $\overline{z}$ and $\underline{z}$ as in Equation (13), but replacing d with $\overline{d}$ and $\underline{d}$ respectively. Then $z = (\overline{z}+\underline{z})/2$. Since both $\overline{z}$ and $\underline{z}$ are bounded above by β, we conclude that $\overline{z}(x,y)_q = \underline{z}(x,y)_q = \beta$. This implies that $d(x,y)_q = \overline{d}(x,y)_q = \underline{d}(x,y)_q$. Since the absolutely continuous parts of these measures are equal, their singular parts satisfy $d_s(x,y)_q = \overline{d}_s(x,y)_q = \underline{d}_s(x,y)_q$.

We denote the support of d_s by S and define the following set of subsets of $[0,q]$:

$$\Pi := \{S \cap (x,y)_q : x \in \mathcal{U},\, y \in \mathcal{W}\}.$$

Let Π^* be the set of all subsets of $[0,q]$ that can be expressed as the set difference of two elements of Π. Then d_s, $\overline{d}_s$, and $\underline{d}_s$ agree on all elements of Π^*.

To show that these three measures agree on every Borel subset of S, it will be sufficient to show that they agree on every open subset of S. Each of these can be expressed as a countable disjoint union of sets of the form $S \cap (a,b)_q$, with $a, b \in [0,q]$. Hence we need only consider sets of this form. Let G be such a set. We will define an increasing sequence of sets J_i such that $J_i \cap S$ is contained in Π^* and $\bigcup_i J_i \cap S = G$.

Suppose there exists a sequence v_i in S which decreases strictly to a. Applying Condition (Z) to the points v_i and v_{i+1}, and using the upper semicontinuity of $W^{\triangleleft}$ and $U^{\triangleleft}$, we can find a point a_i that is in either $\mathcal{W}$ or $\mathcal{U}$. This will be the left boundary point of J_i. Take J_i to be closed on the left if a_i is in $\mathcal{W}$ and open if it is in $\mathcal{U}$.

Suppose, on the other hand, that there is no such sequence v_i. Then we can find $l, r \in S$ such that $a \in [l,r)_q$ and $(l,r)_q$ contains no point of S. On the interval $(l,r)_q$, both $W^{\triangleleft}(d,\cdot)$ and $U^{\triangleleft}(d,\cdot)$ are continuous. Since $Dd \geq 1/\rho$, these two functions are, respectively, non-decreasing and non-increasing on the same interval. Thus, either $r \in \mathcal{W}$ or $l \in \mathcal{U}$. If the former is the case then, for each $i \in \mathbb{N}$, we define the left boundary point of J_i to be r and take J_i to be closed on the left. If the latter, then the left boundary point of J_i is defined to be l, and J_i is taken to be open on the left.

The right boundary point of J_i is defined, for each $i \in \mathbb{N}$, in an analogous way. It may be easily verified that each member of the sequence $\{S \cap J_i : i \in \mathbb{N}\}$ is in Π^*, that this sequence is increasing, and that the union of all its members is equal to G. $\square$

References

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

[2] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Statistical service assurances for traffic scheduling algorithms. *IEEE Journal for Selected Areas in Communications*, 18(12):2651–2664, 2000.

[3] 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.

[4] C. Courcoubetis, F. Kelly, and R. Weber. Measurement-based usage charges in communication networks. *Operations Research*, 48:535–548, 2000.

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

[6] 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.

[7] Manfred Denker, Christian Grillenberger, and Karl Sigmund. *Ergodic Theory on Compact Spaces.* Springer-Verlag, Berlin, 1976. Lecture Notes in Mathematics, Vol. 527.

[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] 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.

[10] G. Kesidis and T. Konstantopoulos. Extremal shape-controlled traffic patterns in high-speed networks. *IEEE Trans. Comm.*, 48(5):813–819, May 2000.

[11] G. Kesidis and T. Konstantopoulos. Worst-case performance of a buffer with independent shaped arrival processes. *IEEE Communication Letters*, 4(1):26–28, 2000.

[12] 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.

[13] F. LoPresti, Z. Zhang, D. Towsley, and J. Kurose. Source time scale and optimal buffer/bandwidth trade-off for regulated traffic in an ATM node. In *Proceedings of IEEE Infocom '97*, pages 347–355, Kobe, Japan, April 1997.

[14] 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.

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

[16] A. Puhalskii. Large deviation analysis of the single server queue. *Queueing Systems*, 21:5–65, 1995.

[17] M. Reisslein, K. W. Ross, and S. Rajagopal. Packet multiplexers with adversarial regulated traffic. In *Proceedings of IEEE Infocom '98*, pages 347–355, San Francisco, April 1998.

[18] A. Wayne Roberts and Dale E. Varberg. *Convex functions.* Academic Press, New York-London, 1973. Pure and Applied Mathematics, Vol. 57.

[19] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, 1987.

[20] S. Shioda. Worst-case performance of ATM multiplexer. *J. Oper. Res. Soc. Japan*, 41(1):3–20, 1998.

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

[22] Cormac Walsh. Maximal effective bandwidth of constrained traffic. *Queueing Systems*, 44:161–182, 2003.