Convexity Properties of Loss and Overflow Functions

K. Kumaran M. Mandjes[1] A.L. Stolyar
Bell Labs, Lucent Technologies,
Murray Hill, NJ 07974, United States
{kumaran,stolyar}@research.bell-labs.com, michel@cwi.nl

December 11, 2001

Abstract

We show that the fluid loss ratio in a fluid queue with finite buffer b and constant link capacity c is always a jointly convex function of b and c. This generalizes prior work [6] which shows convexity of the (b, c) trade-off for large number of i.i.d. multiplexed sources, using the large deviations rate function as approximation for fluid loss. Our approach also leads to a simpler proof of the prior result, and provides a stronger basis for optimal measurement-based control of resource allocation in shared resource systems.

1 Introduction

A queueing system can be described as a set of resources, typically service capacities (or *bandwidth* in communication networks), and buffers (where excess traffic can be stored temporarily). The traffic offered to the queues is determined by the routing and scheduling disciplines, which in turn determines system performance. In general, the resources at the queues have to be designed such that a performance criterion is satisfied, for instance the buffer overflow probability has to be below some small fraction ϵ. For a fixed performance level, the resources involved often *trade off*. As each of these resources has its specific price, it is important to have more detailed knowledge of this trade-off. Hence, efficient (i.e., economical) network design is enabled through insight into the shape of the resulting 'iso-performance curve'.

IIn this note we focus on one of the simplest queueing networks: a single queue, with finite buffer size b and constant service rate c, fed by a stationary fluid (arbitrarily divisible) source. We show that the fluid *loss ratio* function $\Phi(c, b)$ is jointly convex on c and b. (This in particular implies that for a given $\epsilon > 0$, the trade-off curve between b and c, defined by $\Phi(c, b) = \epsilon$ is convex.) A crucial observation in our analysis is the following general principle:

[1] M. Mandjes is currently with both Faculty of Mathematical Sciences at University of Twente (P.O. Box 217, 7500 AE Enschede, the Netherlands, email: m.r.h.mandjes@math.utwente.nl) and CWI (P.O. Box 94079, 1090 GB Amsterdam, the Netherlands, email: michel@cwi.nl).

Sharing resources is always better than partitioning. More precisely: compare (1) a 'partitioned' system in which random input processes f_i feeding queues with resources (c_i, b_i) (with $i = 1, 2$), with (2) a 'shared' system in which $f_1 + f_2$ feeds a queue with buffer $b_1 + b_2$ and link rate $c_1 + c_2$. The amount of traffic lost in the shared system is dominated by the (total) amount of traffic lost in the partitioned system.

Our analysis shows that this general principle easily implies the joint convexity of $\Phi(c, b)$. In the *large deviations regime*, which involves scaling up the resources as $b \to nb$ and $c \to nc$, and replacing the input process by n i.i.d. copies, previous work established that the *overflow probability* decays exponentially in n [1, 2, 9], with decay *rate function* $I(c, b)$. It is proved in [6] that the trade-off curve defined by $I(c, b) = \delta$ is convex. Based on our observations about $\Phi(c, b)$, we also give a simpler proof of that result.

However, the main advance of this work is the generalization of [6] by proving the joint convexity of the loss ratio $\Phi(c, b)$, even away from the asymptotic large n regime. This generalization can offer several practical advantages. Firstly, it permits more optimal resource partitioning. As in [6], this can be done optimally using our results for inhomogeneous traffic and differential QoS requirements, and unlike [6], even when each traffic class has only a few connections. Secondly, joint convexity permits use of more complex QoS metrics without losing the computational tractability of the multi-class resource optimization problem. Finally, when measurement-based control is feasible, the loss ratio $\Phi(c, b)$ is a more appropriate, and more directly measurable, parameter in comparison to the rate function $I(c, b)$. These considerations make it desirable to directly prove the universal joint convexity of $\Phi(c, b)$ without using $I(c, b)$ as approximation, which is what we accomplish here using arguments that are simpler than those of [6].

The rest of this note is organized as follows. Section 2 gives a number of key sample path relations, and Section 3 establishes the convexity of $\Phi(b, c)$. In Section 4 the large deviations regime is examined, and the convexity of the trade-off curve for buffer overflow probability is proved before conclusion in Section 5.

2 Key sample path relations

Consider a single server with (constant) processing capacity c, and a buffer size b. The server is fed by a fluid input flow. We denote by $f(t)$ the cumulative amount of fluid arrived by time $t \in \mathbb{R}$. In this section we assume that $f(\cdot)$ is a fixed function (interpreted later as a sample path of a random process), which is non-decreasing continuous. Also, we consider the evolution of such (fluid) system in the interval $[0, \infty)$, so $f = (f(t),\ t \geq 0)$ and, by convention, $f(0) = 0$.

Suppose the buffer size is finite, $b < \infty$, and the initial queue length (buffer content) is $q(0)$, where $0 \leq q(0) \leq b$. Then, the queue length at every time $t \geq 0$ is determined as follows:

$$q(t) \quad = \quad q(0) + f(t) - ct + h[q(0), f, c, b](t) - g[q(0), f, c, b](t), \tag{1}$$

where the lower and upper *regulations* h and g are unique non-decreasing continuous functions such that

$\mathrm{A}_b \qquad 0 \leq q(t) \leq b, \;\; t \geq 0,$

$\mathrm{B}_b \qquad h(0) = g(0) = 0,$

$$\text{C}_b \qquad \int_0^t I\{q(s) > 0\}\mathrm{d}h(s) = 0, \;\; t \geq 0, \;\; \text{and}$$

$$\text{D}_b \qquad \int_0^t I\{q(s) < b\}\mathrm{d}g(s) = 0, \;\; t \geq 0,$$

see Harrison [5], Proposition 2.4.6.
The functions h and g we will also refer to as the *(cumulative) idleness* and *(cumulative) loss* function, respectively.
In case of an infinite buffer, $b = \infty$, the queue length is defined similarly:

$$q(t) = q(0) + f(t) - ct + h[q(0), f, c](t), \tag{2}$$

with

$$\text{A}_\infty \qquad 0 \leq q(t), \;\; t \geq 0,$$

$$\text{B}_\infty \qquad h(0) = 0,$$

$$\text{C}_\infty \qquad \int_0^t I\{q(s) > 0\}\mathrm{d}h(s) = 0, \;\; t \geq 0.$$

In addition, for infinite buffer the expression for the idleness h is explicit (see [5], Proposition 2.2.3):

$$h(t) = -[\inf_{s\in[0,t]}(q(0) + f(s) - cs) \wedge 0] \;,$$

where $\wedge$ denotes the minimum of two numbers. Note that, for any $t \geq 0$, $q(t)$ is monotone non-decreasing on the initial condition $q(0)$.
The following two simple lemmas are needed to prove the key Lemma 2.3 below.

Lemma 2.1 *Let q_1, q_2, q, be the queue length functions in three infinite buffer systems, defined by the triples $(q_1(0), f_1, c_1)$, $(q_2(0), f_2, c_2)$, and $(q(0), f, c)$, respectively, where $q(0) = q_1(0) + q_2(0)$, $f = f_1 + f_2$, $c = c_1 + c_2$. Then,*

$$q(t) \leq q_1(t) + q_2(t), \;\; \forall t \geq 0 \;. \tag{3}$$

Proof. If

$$\inf_{s\in[0,t]}(q(0) + f(s) - cs) \leq 0 \tag{4}$$

then

$$\begin{aligned} h[q(0), f, c](t) &= -\inf_{s\in[0,t]}(q(0) + f(s) - cs) \\ &\leq -\inf_{s\in[0,t]}(q_1(0) + f_1(s) - c_1 s) - \inf_{s\in[0,t]}(q_2(0) + f_2(s) - c_2 s) \\ &\leq h[q_1(0), f_1, c_1](t) + h[q_2(0), f_2, c_2](t) \;, \end{aligned}$$

which implies (3). The case when condition (4) does not hold is trivial, because then $h[f, c](t) = 0$. ∎

Lemma 2.2 *Consider a 4-tuple* $(q(0), f, c, b)$, $0 \le q(0) \le b$, *and suppose non-decreasing contuinuous functions (of* t*)* $\hat{h}$ *and* $\hat{g}$ *are such that*

$$\hat{q}(t) = f(t) - ct + \hat{h}[f, c, b](t) - \hat{g}[f, c, b](t),$$

with

$$\begin{array}{ll} \mathrm{A}'_b & 0 \le \hat{q}(t) \le b, \ \ t \ge 0, \\ \mathrm{B}'_b & \hat{h}(0) = \hat{g}(0) = 0, \\ \mathrm{C}'_b & \displaystyle\int_0^t I\{\hat{q}(s) > 0\} d\hat{h}(s) = 0, \ \ t \ge 0. \end{array}$$

(The condition is absent that $\hat{g}$ *can only increase when* $\hat{q}(t)$ *is equal to* b*.) Then,*

$$\hat{g}(t) \ge g[q(0), f, c, b](t), \ \ \forall t \ge 0. \tag{5}$$

Proof. This result can be easily obtained for example by slightly extending the proof of Proposition 2.4.6 in [5]. Namely, the pair $(h[q(0), f, c, b], g[q(0), f, c, b])$ is the *unique* minimal fixed point of a *monotone* operator mapping a pair $(h^{(1)}, g^{(1)})$ into a pair $(h^{(2)}, g^{(2)})$. It is easy to see that $(\hat{h}, \hat{g})$ is a fixed point of that operator, and therefore the majorization (5) does hold. We omit details. ∎

The following lemma is the key in establishing the convexity of $\Phi(c, b)$, where $\Phi(c, b)$ is the fraction of lost fluid in a finite buffer system with buffer b and server rate c.

Lemma 2.3 *Consider three finite buffer systems defined by 4-tuples* $(q_1(0), f_1, c_1, b_1)$, $(q_2(0), f_2, c_2, b_2)$, *and* $(q(0), f, c, b)$, *respectively, where* $q(0) = q_1(0) + q_2(0)$, $f = f_1 + f_2$, $c = c_1 + c_2$, *and* $b = b_1 + b_2$. *Then,*

$$g[q(0), f, c, b](t) \le g[q_1(0), f_1, c_1, b_1](t) + g[q_2(0), f_2, c_2, b_2](t) \ \ \forall t \ge 0 \ .$$

Proof. Using Lemma 2.1, it is easy to verify directly that $q(0), f, \hat{g} = g[q_1(0), f_1, c_1, b_1] + g[q_2(0), f_2, c_2, b_2]$, and

$$\hat{h}(t) = -\inf_{s \in [0,t]} (q(0) + f(s) - \hat{g}(s) - cs) \ ,$$

satisfy the conditions of Lemma 2.2. Therefore, $g(t) \le \hat{g}(t)$ for all t. ∎

3 Convexity of the fluid loss ratio

In this section we prove the convexity result for the fluid loss ratio. The importance and universality of this result is in the fact that the convexity holds for a system with any fixed parameters – not only in an asymptotic regime.

Let $f = (f(t), \ t \ge 0)$ now be a *random process* with continuous non-decreasing paths and stationary increments. We can use normalization $f(0) = 0$ without loss of generality. The mean (fluid) arrival rate

$$m := \mathbb{E} f(1)$$

is well defined. The steady state fraction of lost fluid in a finite buffer system defined by (f, c, b) is defined as follows:

$$\Phi(c, b) := \frac{1}{m} \lim_{t \to \infty} \frac{\mathbb{E}\{g[q(0), f, c, b](t)\}}{t},$$

assuming that the right-hand side is well defined and does not depend on $q(0)$. (It is easy to see that this assumption holds in virtually all cases of interest, and so the above definition matches most of the common definitions of steady state 'loss probability' or 'fluid loss ratio' in systems with finite buffers.)

Theorem 3.1 *Suppose the random input flow process f with stationary increments is fixed. Then the function $\Phi(c, b)$ is convex jointly on c and b.*

Proof. Consider a fixed sample path f of the input process, and a fixed initial buffer content $q(0)$. Let us fix positive constants α_1 and α_2, such that $\alpha_1 + \alpha_2 = 1$, and positive constants c, b, c_1, b_1, c_2, b_2, such that $c = \alpha_1 c_1 + \alpha_2 c_2$ and $b = \alpha_1 b_1 + \alpha_2 b_2$. We have for any $t \geq 0$:

$$\begin{aligned} g[q(0), f, c, b](t) &\leq g[\alpha_1 q(0), \alpha_1 f, \alpha_1 c_1, \alpha_1 b_1](t) + g[\alpha_2 q(0), \alpha_2 f, \alpha_2 c_2, \alpha_2 b_2](t) \\ &= \alpha_1 g[q(0), f, c_1, b_1](t) + \alpha_2 g[q(0), f, c_2, b_2](t), \end{aligned}$$

where the inequality follows directly from Lemma 2.3 and the equality follows from the fact that an (auxillary) system defined by $(\alpha_i q(0), \alpha_i f, \alpha_i c_i, \alpha_i b_i)$, $i = 1, 2$, is the system defined by $(q(0), f, c_i, b_i)$ 'scaled down' by a factor α_i. ∎

4 Convexity of the trade-off curve for buffer overflow probability in the large deviations regime

In this section we consider a system with *infinite* buffer space (i.e., no traffic loss). We consider traffic from n independent, statistically identical, sources feeding into a buffered resource. This resource is modeled as a queue with constant depletion rate nc. As in the previous section, a single traffic source is described by a *random process* $f = (f(t), t \geq 0)$ continuous non-decreasing paths and stationary increments. (It will be clear from our development that the continuity assumption is not important.) We remind that $f(t) - f(0)$ represents the amount of traffic arrived by time t, so we assume that $f(0) = 0$.

Since we consider an infinite buffer system we assume that the mean arrival rate (for a single source) therefore, the mean (fluid) arrival rate

$$m = \mathbb{E}f(1) < c\,,$$

so that the system is stable.

We are interested in the steady-state *overflow probability* $p_n(c, b)$, which is the probability of the queue length exceeding level nb. (Here we use notation b for a 'threshold value' rather than the buffer size.) Under non-restrictive conditions, this probability decays *exponentially* in the scaling parameter n. We define the corresponding exponential *decay rate*:

$$I(c,b) := -\lim_{n\to\infty} \frac{1}{n} \log p_n(c,b).$$

The key result on $I(b,c)$, based on large deviations arguments, is given below in Theorem 4.3. A major contribution to the development of these results was given by Botvich and Duffield [1], whereas related results were derived in [2, 9]. Recently, a significant improvement was made by Likhanov and Mazumdar [7].

Assumption 4.1 *See* [7] *– Assume $I_t(c,b)$ is larger than $\alpha \log t$, for t large enough, and a positive α, where*

$$I_t(c,b) := \sup_{\theta>0} \left(\theta(b+ct) - \log \mathbb{E}e^{\theta f(t)}\right). \tag{6}$$

An additional assumption has to be made on the regularity of the traffic. For this purpose we here use Hypothesis 1(iv) from [1], which essentially implies that the decay rate for discrete time carries over to continuous time. This is proven analogously to the proof of Theorem 1 of [1, p. 302].
Hypothesis 1(iv) is stated as follows. Define

$$f^n_{t,r} := \sup_{0<r'<r} \sum_{i=1}^{n} f_i(t) - f_i(t-r'),$$

where f_i's are i.i.d. copies of a single source process. (To be precise, the above notation also assumes that the process f is extended to be defined for all real t.) Then it is required that

$$\limsup_{r\downarrow 0} \limsup_{n\to\infty} \frac{1}{n} \sup_{t\geq 0} \log \mathbb{E}\exp\left(\theta f^n_{t,r}\right) \leq 0.$$

It is easily verified that, due to the stationarity and i.i.d. assumptions, this requirement reduces to the following.

Assumption 4.2 *See* [1] *– Assume that for all $\theta \in \mathbb{R}$,*

$$\limsup_{r\downarrow 0} \log \mathbb{E}\exp\left(\theta \sup_{0<r'<r} f(r')\right) \leq 0.$$

Theorem 4.3 *Under Assumptions 4.1 and 4.2,*

$$I(c,b) = \inf_{t>0} I_t(c,b). \tag{7}$$

Applicability. The striking feature of the above theorem is that it is applicable for almost all types of fluid traffic sources of practical interest. Short-range dependent sources satisfy the conditions (Markov fluid sources, on-off sources with light-tailed on-times), see [1], but also long-range dependent processes (fractional Brownian motion, on-off sources with heavy tailed on-times), see [4, 7, 8].
Now, let us recall the meaning of $I_t(c,b)$. By the classical Cramér theorem (cf. Theorem 2.2.3 in [3]), $I_t(c,b)$ is nothing else but the value of the large deviations *rate function* for the sequence of distributions of

$$\frac{1}{n}\sum_{k=1}^{n} f_k(t)$$

at point $b+ct$. This in particular means that, for a given t, $I_t(c,b)$ is essentially a function of a single variable $b+ct$. Using standard properties of rate functions, we can rewrite (7) as

$$I(c,b) = \inf\{I_t(\xi,\beta) \mid t>0,\ \xi \geq 0, \beta \geq 0,\ \beta+\xi t \geq b+ct\}. \tag{8}$$

Theorem 4.4 *Defined implicitly by $I(c,b)=\delta$ with $\delta>0$, buffer b as a function of service capacity c, is* convex.

Proof. Let us fix two (positive) pairs (c_1,b_1) and (c_2,b_2) such that $I(c_1,b_1)=I(c_2,b_2)=\delta$. Let us fix positive α_1 and α_2, such that $\alpha_1+\alpha_2=1$. Denote $c=\alpha_1 c_1+\alpha_2 c_2$ and $b=\alpha_1 b_1+\alpha_2 b_2$. We need to show that $I(c,b) \geq \delta$.
Let $t^\star \in \mathbb{R}$ be the value of t that optimizes (7) for the pair (c,b), and therefore $(t^\star, c, b)$ optimizes the right-hand side of (8). We obviously have the following property:

$$\text{Either } \ b+ct^\star \geq b_1+c_1 t^\star \ \text{ or } \ b+ct^\star \geq b_2+c_2 t^\star\ . \tag{9}$$

Without loss of generality, let us assume that the former inequality holds.
Property (9) is of course a matter of simple arithmetic. However, it is key here, and can be interpreted the same way as Lemma 2.1 (and in fact can be obtained as a simple corollary from it):
If the amount of fluid $b+ct^\star$ arrives in the interval $[0,t^\star]$ into the system with parameters (c,b), then $q(t^\star) \geq b$. Therefore, if we partition the system into two, with parameters $(\alpha_i c_i, \alpha_i b_i)$, $i=1,2$, and feed the amounts of fluid $\alpha_i(b+ct^\star)$ in $[0,t^\star]$ into them, then the queue length must be at or above threshold $\alpha_i b_i$ in at least one of them.
By (8) and (9), we have

$$\delta = I(c_1,b_1) \leq I_{t^\star}(c,b) = I(c,b)\ .$$

∎

5 Conclusion

In this work, we have proved that the loss ratio is always a jointly convex function of buffer and bandwidth sizes in buffer-bandwidth queuing systems, which also implies convexity of buffer-bandwidth trade-off curves. This result generalizes previous work [6] that proves convexity of the loss function, approximated by the large deviations rate function, for large numbers of multiplexed flows. Our results hence establish that the same practical advantages mentioned in [6] regarding optimal design of partitioned buffer-bandwidth systems continue to apply if loss ratio were the criterion instead of the rate function, and without the need for large numbers of multiplexed flows. Apart from theoretical generality, this may offer some practical advantages in terms of more efficient resource allocation, better use of loss measurements, and the ability to admit more complex QoS requirements.

Acknowledgements. The authors would like to thank Iraj Saniee for insightful comments.

References

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

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

[3] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* 2nd ed. Springer, 1998.

[4] N. Duffield. Queueing at large resources driven by long-tailed M/G/∞-modulated processes. *Queueing Systems*, 28: 245 – 266, 1998.

[5] J.M. Harrison. *Brownian Motion and Stochastic Flow Systems.* Wiley, 1985.

[6] K. Kumaran and M. Mandjes. The buffer-bandwidth trade-off curve is convex. *Queueing Systems*, 38: 471 – 483, 2001

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

[8] M. Mandjes and S. Borst. Overflow behavior in queues with many long-tailed inputs. *Advances in Applied Probability*, 32: 1150 – 1167, 2000.

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