ENTROPY OF ATM TRAFFIC STREAMS: A TOOL FOR ESTIMATING QoS PARAMETERS

N.G.Duffield [1,2], J.T.Lewis [2], Neil O'Connell [2],
Raymond Russell [2] and Fergal Toomey [2]
Dublin Applied Probability Group

[1] School of Mathematical Sciences,
Dublin City University, Dublin 9, Ireland.

[2] Dublin Institute for Advanced Studies
10 Burlington Road, Dublin 4

September, 1994

1 Introduction

How will ATM carriers allocate the band-width required to guarantee the quality-of-service promised in their customer contracts? How can customers exploit to their advantage the tariff structures offered by the carriers? Both carrier and customer will need to measure QoS (quality-of-service) parameters. Existing proposals involve modelling: fitting a statistical model to the input traffic and calculating QoS parameters from the model. Doubts have been expressed about this procedure because data traffic is bursty and cannot be described by a model with a small number of parameters. Our approach is more radical: we estimate directly the thermodynamic entropy of the data-stream at an input-port; from this, the QoS

parameters can be calculated rapidly. The algorithms are simple enough to be programmed onto the port-controllers of an ATM switch.

2 The Problem

ATM switches are buffered. Cells may be lost if a buffer overflows; cells may be delayed by being stacked in a buffer. In this paper, we are concerned with the components of cell-loss and cell-delay which are attributable to a single buffer of finite size. The QoS parameters we are concerned with are:

- cell-loss ratio;
- cell-delay variation;
- mean cell-delay;
- 'jitter' (the variance of the cell-delay).

The problem we address is the estimation of these parameters for ATM traffic which is stacked in a buffer emptied at a fixed service-rate. This is a queueing problem: the QoS parameters can be estimated easily provided we know the tail of the queue-length distribution. Estimating the tail is a problem in large deviation theory; as we shall see, the large deviation rate-function of the arrivals process (the number of cells entering the buffer in each clock-cycle) yields an estimate of the tail of the queue-length distribution. The current practice is to model the arrivals process:

- choose a statistical model;
- fit the parameters of the model to the traffic (using moments, for example);
- compute QoS parameters, using the model.

There are objections to the implementation of this programme:

- it is difficult to automate the selection of a model;
- bursty traffic cannot be modelled using a small number of parameters;

- the computational requirements make it difficult to perform the estimation in real time;
- it wastes resources – a good model contains more information about the arrivals process than is required for the estimation of QoS parameters.

This has triggered the search for alternatives to modelling (see Courcoubetis et al [5] for another proposal). Since all that is required for the estimation of QoS parameters is a knowledge of the large deviation rate-function of the arrivals process, why not estimate the rate-function itself? There are good reasons for believing this to be possible: since the work of Ruelle [13] and Lanford [11], it has been well known (but not widely) that the rate-function of large deviation theory is the same kind of mathematical object as the entropy-function of equilibrium thermodynamics. (The connection between large deviation theory and equilibrium thermodynamics is explained briefly in Appendix 2.) The rate-function and the entropy-function have this in common: they encapsulate concisely the relevant information about the system. For an ideal gas, the entropy-function can be calculated from first principles; for a real gas, one could choose a statistical model, fit the parameters of the model to measured properties of the gas (virial coefficients, for example), compute the entropy-function from the model and use the entropy-function to compute the bulk properties of the gas. This is not the practice of chemical engineers: they measure the entropy-function or use the tables of measured values available in the literature.

Our claim is this:

> *for the purposes of estimating QoS parameters, it is enough to know the rate-function of the ATM traffic stream; the modelling procedure can be by-passed if we can estimate the rate-function directly.*

3 Basic theory

In the previous section we stated somewhat vaguely that QoS parameters for a traffic stream passing through a buffer can be estimated using the rate function of the arrivals process. We will now make that statement more precise with an overview of the underlying theory.

3.1 Estimating QoS parameters

Suppose we have a single server queue with stationary arrivals (X_k) and constant service rate s. For stability we require that $s > EX_1$; in other words, the service rate exceeds the mean input rate. The rate function of the arrivals process is defined, for $x > 0$, by

$$I(x) := \lim_{n\to\infty} \frac{1}{n} \log P\left(\sum_{k=1}^{n} X_k > nx\right), \tag{1}$$

whenever this limit exists. The theory of large deviations tells us that, provided I satisfies some technical conditions [8, 9, 10], the tails of the queue-length distribution are asymptotically log-linear:

$$\lim_{q\to\infty} \frac{1}{q} \log P(Q > q) = -\delta; \tag{2}$$

moreover,

$$\delta = \inf_{w>0} I(w + s)/w. \tag{3}$$

A variation of this result is that for a large finite buffer, the log-frequency of cell-loss is approximately linear in the buffer-size, with slope $-\delta$. To obtain an upper bound on the log-frequency of cell-loss, we suggest using the straight line $-\mu - \delta q$, where q denotes the buffer-size and $-\mu$ is the log-frequency with which the queue is non-empty[1]. Using this bound, one can estimate the cell-loss ratio, cell-delay variation, mean cell-delay, variance of the cell-delay ('jitter'). The cell-loss ratio is given approximately by the frequency of cell-loss divided by the mean input rate; the cell-delay variation is just the distribution of the queue-length divided by the service rate (up to 'round off' error), and the mean and variance of the cell-delay are just the mean and variance of this distribution.

Now that we have established the role of the rate function in the problem of estimating QoS parameters, we turn to the question of how to estimate it using traffic observations. It turns out that it is easier to estimate a transform of the rate function, namely the *scaled cumulant generating function* (cgf), rather than the rate function itself. This is defined by

$$\lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E \exp\left(\theta \sum_{k=1}^{n} X_k\right), \tag{4}$$

[1]This is not a rigorous upper bound in general, but we suspect that a sufficient condition for it to be so is that the arrivals are non-negatively correlated, which is typically the case in practice.

whenever this limit exists; then the scaled cgf λ is related to the rate-function $I(x)$ by

$$\lambda(\theta) = \sup_x \{x\theta - I(x)\}; \tag{5}$$

moreover, δ can be calculated directly from the scaled cgf using the formula

$$\delta = \sup\{\theta : \ \lambda(\theta) \leq s\theta\}. \tag{6}$$

The above theory is valid whenever the arrivals process is both stationary and mixing (that is, there is no long-range dependence).

3.2 Estimating the scaled cgf

The mixing condition has a second consequence which we exploit in constructing an estimator for λ : there exists a block size b for which the block sums

$$\tilde{X}_1 := \sum_{k=1}^{b} X_k, \quad \tilde{X}_2 := \sum_{k=b+1}^{2b} X_k, \quad \ldots \tag{7}$$

are approximately independent and identically distributed. Furthermore,

$$\lambda(\theta) \approx \lambda_b(\theta) := \frac{1}{b} \log E e^{\theta \tilde{X}_1}, \tag{8}$$

and so the problem of estimating λ is approximately equivalent to the problem of estimating the distribution of $\tilde{X}_1$. This suggests using the (normalised) cgf of the empirical distribution of the block sums as an estimator for λ and the corresponding solution to (6) as an estimator for δ:

$$\hat{\lambda}_b^n(\theta) := \frac{1}{b} \log \frac{b}{n} \sum_{i=1}^{[n/b]} e^{\theta \tilde{X}_i}, \tag{9}$$

$$\hat{\delta}_b^n := \sup\{\theta : \ \hat{\lambda}_b^n(\theta) \leq s\theta\}. \tag{10}$$

To apply this method in practice, one is faced with the following questions:

- how much data do we need to get a good estimate?
- what is a suitable block size?
- can we assume stationarity?

The analytic and simulation results that follow are intended to provide some insight and rough heuristics for the first two. The question of stationarity is not specific to this problem: it is a fundamental requirement for prediction. (That is not to say that data with trends, cycles and 'external forces' cannot be dealt with; it is often the case that non-stationarity can be 'removed' from the data once it is 'explained'.) We anticipate that this method will be most useful for short term prediction, where stationarity is only required over relatively short time periods. For example, it provides a basis for characterising traffic 'on the fly' so that resources can be allocated dynamically.

Acknowledgement. The idea of using (9) as an estimator for λ was suggested to us by Amir Dembo (private communication).

3.3 Sampling properties of $\hat{\delta}_b^n$

3.3.1 Analytic results

For a given model, the sampling distribution of $\hat{\delta}_b^n$ can be determined using the formula

$$P(\hat{\delta}_b^n < d) = P(\hat{\lambda}_b^n(d) > sd), \tag{11}$$

or approximated using the fact (see Appendix 1) that for large n, $\sqrt{n}(\hat{\delta}_b^n - \delta_b)$ is approximately normal with zero mean and variance given by

$$\sigma_b^2 := \lim_{n\to\infty} \frac{b}{\phi_b'(\delta_b)^2} \left[\phi_b(2\delta_b) + \sum_{k=1}^{n/b-1} 2\left(1 - \frac{kb}{n}\right) E e^{\delta_b(\tilde{X}_1 + \tilde{X}_{1+k})} - \frac{n}{b} \right], \tag{12}$$

where

$$\delta_b = \sup\{\theta : \ \lambda_b(\theta) \le s\theta\} \tag{13}$$

and

$$\phi_b(\theta) = e^{b\lambda_b(\theta) - sb\theta}. \tag{14}$$

As $n \to \infty$, $\hat{\delta}_b^n$ converges in probability to δ_b, and $\delta_b = \delta$ if the block sums are independent; in general we have $\delta_b \approx \delta$ for large b. The sampling distribution tells us how good the estimator is; it can also be used to obtain confidence intervals for δ_b and hence approximate confidence intervals for δ. From the approximation (12) we can immediately see the effect of increasing

the sample size: as $n \to \infty$ the variance of our estimator decays like σ_b^2/n. However, this is not strictly true, because the variance of $\hat{\delta}_b^n$ is generally infinite! This is a technical hitch due to the fact that in general there is a small but positive probability that most or all of the block sums do not exceed bs, leading to huge or even infinite values of $\hat{\delta}_b^n$; this probability goes to zero as $n \to \infty$ provided $P(\tilde{X}_1 > bs) > 0$, and the effect disappears in the normal approximation. Incidentally, it can be shown (see Appendix 1) that if the block sums are independent and we condition on at least *three* block sums exceeding bs, the variance of $\hat{\delta}_b^n$ becomes finite; in general we have finite k^{th} moments when we condition on there being $k+1$ block sums exceeding bs.

We will now apply the normal approximation to an example to see the effect of varying the block size and service rate on the precision of our estimator. A good model to work with for this purpose is the (stationary) first order auto-regressive process: let ϵ_k be a sequence of independent normal random variables with zero mean and unit variance, $0 \leq \alpha < 1$, and define the X's recursively by

$$X_k = \alpha X_{k-1} + \epsilon_k, \tag{15}$$

with $X_1 = \epsilon_1/\sqrt{1-\alpha^2}$. Note that the arrivals may be positive or negative, and the mean input rate is zero: this is only a mathematical convenience; what we call the 'service rate' should really be thought of as the difference between the actual service rate and the actual mean input rate. The parameter α reflects the 'burstiness' of the traffic: the larger the value of α the more strongly correlated the arrivals are and this has the effect of encouraging trends, or 'bursts'; if $\alpha = 0$ we have independent, Gaussian arrivals. For this model,

$$\phi_b(\theta) = \exp\left[\frac{1}{2}\left(b - 2\alpha\left(\frac{1-\alpha^b}{1-\alpha^2}\right)\right)\left(\frac{\theta}{1-\alpha}\right)^2 - bs\theta\right], \tag{16}$$

$$\delta_b = 2(1-\alpha)^2 s\left[1 - \frac{2\alpha(1-\alpha^b)}{b(1-\alpha^2)}\right]^{-1} \tag{17}$$

and

$$\delta = 2(1-\alpha)^2 s. \tag{18}$$

The asymptotic variance is given by

$$\sigma_b^2 = \frac{b}{\phi_b'(\delta_b)^2}\left[\phi_b(2\delta_b) - 1 + 2\sum_{k=0}^{\infty}\left(e^{C\alpha^{bk}} - 1\right)\right], \tag{19}$$

where

$$C = \frac{1}{2}\frac{\alpha}{1-\alpha^2}(1-\alpha^b)^2\left(\frac{\delta_b}{1-\alpha}\right)^2(1+\alpha^{b+1}). \tag{20}$$

(For details of these calculations, see Appendix 1.)

In Figure 1 we have plotted the (approximate) inter-quartile range (IQR) of $\hat{\delta}_b^n$ against block size b, for fixed service rate $s = 1$, $\alpha = 0.99$ and sample size $n = 10^7$. This illustrates the trade-off between

- increasing the block size to reduce bias, and
- decreasing the block size to reduce variance.

One way of optimising this trade-off is to minimise the *mean squared error* of the estimator. Again, strictly speaking this does not exist, but for the normal approximation it is given by

$$\sigma_b^2/n + (\delta_b - \delta)^2. \tag{21}$$

This is plotted in Figure 2 (against block size). From this we can read off the optimal block size, which in this case is about 350/400. Note that the optimal block size depends on the sample size (n), service rate (s) and burstiness (α); in Figures 3–5 we have plotted the optimal block size against each of these parameters.

The next thing we consider is the effect of increasing the service rate on the precision of the estimator. Figure 6 is a plot of the approximate IQR of $\hat{\delta}_b^n$ against s, for fixed block size $b = 5$, sample size $n = 800$ and $\alpha = 0.5$. Clearly the estimator looses precision as the service rate increases. Intuitively, this is because a higher service rate gives rise to shorter queues and hence less information about the tails of the queue-length distribution.

3.3.2 Simulation results

The results of this section are intended to compliment the analytic approximations of the previous section; here we investigate the sampling properties of $\hat{\delta}_b^n$ with 'exact' calculations based on simulations of two kinds of traffic:

Bernoulli. The arrivals X_k are independent and identically distributed with

$$P(X_1 = 1) = 1 - P(X_1 = 0) = 1/4.$$

Two-state Markov. Here the arrivals are modelled by a two-state Markov chain with transition probabilities

$$P(X_2 = 1|\ X_1 = 0) = 1/16; \qquad P(X_2 = 0|\ X_1 = 1) = 3/16.$$

Note that both traffic streams have the same mean activity.

First we consider the sample size: Figures 7 and 8 show how the empirical interquartile range varies with sample size for each model. In both cases the service rate is fixed ($s = 0.3$) and for the two-state Markov traffic we have used the block size $b = 500$ (by independence there is no need to aggregate the Bernoulli traffic). In the Bernoulli case we have superimposed the analytic approximation of §3.3.1: clearly it is a good approximation; the analytic approximation is difficult to compute for the two-state Markov model.

Next we consider the service rate: Figure 9 is a plot, for fixed sample size $n = 1024$, of the empirical interquartile spread against service rate for the Bernoulli traffic, with the analytic approximation. Again we see that for large service rates, the precision of our estimator is reduced quite dramatically.

Finally we consider the effect of varying the block size for correlated traffic, with a view to finding an optimal block size. In Figure 10 we have plotted, for fixed sample size $n = 4.2(10)^6$ and service rate $s = 0.3$, the empirical interquartile range of the estimator against block size for the two-state Markov traffic: just as we expect, increasing the block size reduces bias on one hand, but increases spread on the other. Since we are now dealing with a finite sample where the variance does not exist, we consider minimising the *median* squared error of the estimator as a criterion for choosing the optimal block size. This is asymptotically equivalent to the mean squared error of §3.3.1, and is a more robust quantity than the sample mean squared error. The empirical median squared error, for fixed sample size $n = 4.2(10)^6$ and service rate $s = 0.3$, is plotted in Figure 11: this is minimised at about $b = 250$.

3.4 Watermarking

Recall that the parameter δ we are trying to estimate is given by

$$\delta = -\lim_{q\to\infty} \frac{1}{q} \log P(Q > q), \tag{22}$$

where $P(Q > q)$ is the frequency with which the queue-length exceeds the level q. In other words, if we plot $\log P(Q > q)$ against q, the asymptotic slope is $-\delta$. To observe this empirically, we feed simulated data through a virtual buffer and plot the log-frequencies with which each level is exceeded; we call this a *watermark plot*. Given sufficient data, a watermark plot will typically have a 'straight part' with slope close to $-\delta$ before becoming 'wobbly' at levels which are rarely exceeded. It is important to keep in mind that a watermark plot is a random object.

In this section we compare the variation in our δ-estimates with the variation in the corresponding watermark plots using simulated two-state Markov arrival streams with transition probabilities 1/16 (0 to 1) and 3/16 (1 to 0). The estimates are based on optimal block sizes which, in each case, were found using the procedure described in §3.3.2. Figure 12 shows the results of 30 simulations for different sample sizes and service rates; for each simulation, the watermark is plotted along with our estimated value of δ. The variation in the δ-estimates is comparable with the variation in the slopes of the (straight part of the) watermarks in all cases.

3.5 The relation of cell-loss ratio to watermarking

In §3.1 we proposed the linear upper bound $-\mu - \delta q$ on the log-frequency of cell-loss from a finite buffer of size q; this, in turn, provides an upper bound on the log of the cell-loss ratio. In this section, we compare the watermark plot (the log-tail-frequencies of the queue-length distribution in an *infinite buffer*), the cell-loss ratios at each finite buffer-size, and the queue-length distribution in a finite buffer, using simulated Bernoulli and two-state Markov traffic.

In Figure 13, we demonstrate the relation between watermark and cell-loss ratio in a simulation of 10^7 cycles of Bernoulli traffic with activity 0.38 and service rate 0.4 up to buffer-size 40, and the queue-length distribution in a buffer of size 40.

In Figure 14, we do the same for 10^6 cycles of two-state Markov traffic up to buffer-size 100 (in this case the log-tail-frequencies of the queue-length distribution in a buffer of size 100 are plotted). As before, the parameters of the source are 1/16, 3/16, and the service rate is 0.26.

4 Conclusions

This investigation has shown that the proposal to use an empirical entropy function to estimate QoS parameters is theoretically sound. Preliminary experiments, made on the Fairisle network at the University of Cambridge by Simon Crosby, have established that it is feasible to collect the required data in real time.

Our experience so far with real traffic has been promising. For our method to be applicable, the traffic must be stationary and mixing over periods long enough to ensure adequate precision. The requirement of stationarity is not specific to this method: inconsistent behaviour cannot be predicted. We anticipate that our method will be most useful for short-term prediction where stationarity is only required only for relatively short time-periods. For example, it provides a basis for characterising traffic on the fly, so that resources can be allocated dynamically. The question of the applicability of our analysis to real switches is currently being investigated by running simulated data through the Fairisle switch.

Appendix 1

The central limit theorem

An alternative expression for $\hat{\delta}_b^n$ is

$$\hat{\delta}_b^n = \sup\{\theta : \ \hat{\phi}_b^n(\theta) \leq 1\}, \tag{23}$$

where

$$\hat{\phi}_b^n(\theta) = \frac{b}{n} \sum_{i=1}^{[n/b]} e^{\theta(\tilde{X}_i - bs)}. \tag{24}$$

Thus,

$$P(\hat{\delta}_b^n < d) = P(\hat{\phi}_b^n(d) > 1). \tag{25}$$

Set $\phi_b(\theta) := E\hat{\phi}_b^n(\theta)$. By the central limit theorem for weakly dependent stationary processes we have that, for each θ, $\sqrt{n}[\hat{\phi}_b^n(\delta_b) - 1]$ converges weakly as $n \to \infty$ to a normal distribution with zero mean and variance given by

$$\tau_b^2 \ := \ \lim_{n\to\infty} n\text{var } \hat{\phi}_b^n(\delta_b)$$

$$= \lim_{n\to\infty} b\left[\phi_b(2\delta_b) + \sum_{k=1}^{n/b-1} 2\left(1-\frac{kb}{n}\right) E e^{\delta_b(\tilde{X}_1+\tilde{X}_{1+k})} - \frac{n}{b}\right],$$

provided this limit exists. Assuming ϕ_b is smooth on the interior of its effective domain (the set where it is finite) and finite in a neighbourhood of δ_b, the weak law of large numbers tells us that $\hat{\phi}_b^{n\prime}(\delta_b)$ converges in probability to $\phi_b'(\delta_b)$ and $\hat{\phi}_b^{n\prime\prime}(\theta)$ converges in probability to $\phi_b''(\theta)$ for $\theta \le \delta_b$. Combining these facts with (25) we have that as $n \to \infty$,

$$\begin{aligned}
P(\sqrt{n}(\hat{\delta}_b^n - \delta_b) < d) &= P(\hat{\delta}_b^n < d/\sqrt{n} + \delta_b) \\
&= P(\hat{\phi}_b^n(\delta_b + d/\sqrt{n}) > 1) \\
&= P\left(\hat{\phi}_b^n(\delta_b) + \frac{d}{\sqrt{n}}\hat{\phi}_b^{n\prime}(\delta_b) + \frac{1}{n}O_p(1) > 1\right) \\
&= P\left(\sqrt{n}[\hat{\phi}_b^n(\delta_b) - 1] > -d\hat{\phi}_b^{n\prime}(\delta_b) - \frac{1}{n}O_p(1)\right) \\
&\to P(Z > -\phi_b'(\delta_b)/\tau_b),
\end{aligned}$$

where $O_p(1)$ means 'convergent in probability' and Z is a standard normal random variable. The last step follows from Slutzky's theorem (see, for example, [3]). It follows that $\sqrt{n}(\hat{\delta}_b^n - \delta_b)$ converges weakly to a normal random variable with zero mean and variance $\tau_b^2/\phi_b'(\delta_b)^2$.

The conditional moments of $\hat{\delta}_b^n$

For notational convenience set $m = n/b$, $Z_i = \tilde{X}_i - bs$, and drop the subscripts and superscripts on $\hat{\delta}_b^n$ and $\hat{\phi}_b^n$. Let N denote the number of block sums exceeding bs, that is

$$N = \#\{i : \ Z_i > 0\}. \tag{26}$$

Note that $\hat{\delta} = \infty$ on $\{N = 0\}$. Now, assuming the block sums are independent and identically distributed,

$$\begin{aligned}
P(\hat{\delta} \ge \theta |\ N \ge k) &= P(\hat{\phi}(\theta) \le 1 |\ N \ge k) \\
&= P(\hat{\phi}(\theta) \le 1 |\ Z_1, \ldots, Z_k > 0) \\
&= P(\sum_{i=2}^{m} e^{\theta Z_i} \le m - e^{\theta Z_1},\ e^{\theta Z_1} \le m |\ Z_1, \ldots, Z_k > 0) \\
&\le P(\sum_{i=2}^{m} e^{\theta Z_i} \le m,\ e^{\theta Z_1} \le m |\ Z_1, \ldots, Z_k > 0)
\end{aligned}$$

$$
\begin{aligned}
&= P(\sum_{i=2}^{m} e^{\theta Z_i} \leq m |\ Z_2, \ldots, Z_k > 0) P(e^{\theta Z_1} \leq m |\ Z_1 > 0) \\
&\vdots \\
&\leq P(\sum_{i=k+1}^{m} e^{\theta Z_i} \leq m) P(e^{\theta Z_1} \leq m |\ Z_1 > 0)^k
\end{aligned}
$$

Thus, to ensure that

$$E(\hat{\delta}^r |\ N \geq k) < \infty, \tag{27}$$

it suffices to have, for some $\epsilon > 0$,

$$\lim_{\theta \to \infty} \theta^{r+\epsilon} P(0 < Z_1 \leq \log m/\theta) = 0; \tag{28}$$

if we assume that there exists $\gamma > 0$ for which Z_1 has a density on $(0, \gamma)$ then (28) is guaranteed once $k > r$.

Calculations for the auto-regressive process

The stationary AR(1) (first order autoregressive) process $\{X_i\}_{i \geq 1}$ can be defined recursively by

$$
\begin{aligned}
X_1 &= \alpha \frac{\xi_0}{\sqrt{1-\alpha^2}} + \xi_1 \\
X_{i+1} &= \alpha X_i + \xi_{i+1} \qquad\qquad \forall\ i \geq 1
\end{aligned}
$$

where $\{\xi_j\}_{j \geq 1}$ is a sequence of independent Gaussian random variables with zero mean and unit variance: $\xi_j \sim N(0,1),\ \forall j \geq 1$. Thus

$$X_i = \sum_{j=1}^{i} \alpha^{i-j} \xi_j + \frac{\alpha^i}{\sqrt{1-\alpha^2}} \xi_0.$$

The stationary distribution of this process is Gaussian with mean 0 and variance $(1-\alpha^2)^{-1}$, so we can start it off in the stationary regime by choosing ξ_0 to be also Gaussian with zero mean and unit variance. We choose a block size b, and define the aggregated process $\{\tilde{X}_k\}_{k \geq 1}$ by

$$
\begin{aligned}
\tilde{X}_{k+1} &:= \sum_{i=kb+1}^{kb+b} X_i \\
&= \sum_{j=kb+1}^{kb+b} \frac{1-\alpha^{kb+b+1-j}}{1-\alpha} \xi_j + \frac{1-\alpha^b}{1-\alpha} \sum_{j=1}^{kb} \alpha^{kb+1-j} \xi_j + \frac{1-\alpha^b}{1-\alpha} \frac{\alpha^{kb+1}}{\sqrt{1-\alpha^2}} \xi_0
\end{aligned}
$$

We wish to calculate $E[e^{\theta \tilde{X}_k}]$ and $E[e^{\theta(\tilde{X}_k+\tilde{X}_l)}]$. The $\tilde{X}_i$'s are stationary, since the X_i's are, and so

$$\begin{aligned} E[e^{\theta \tilde{X}_k}] &= E[e^{\theta \tilde{X}_1}] \\ E[e^{\theta(\tilde{X}_k+\tilde{X}_l)}] &= E[e^{\theta(\tilde{X}_{|k-l|+1}+\tilde{X}_1)}] \end{aligned}$$

If X and Y are two independent Gaussian random variables with zero mean and unit variance, then

$$\log E[e^{\theta(aX+bY)}] = \frac{1}{2}(a^2+b^2)\theta^2.$$

We use this fact repeatedly on the ξ_j's which make up the $\tilde{X}_k$'s to get

$$\begin{aligned} \log E[e^{\theta \tilde{X}_1}] &= \frac{1}{2}\left(b - \frac{2\alpha}{1-\alpha^2}(1-\alpha^b)\right)\left(\frac{\theta}{1-\alpha}\right)^2 \\ \log E[e^{\theta(\tilde{X}_{k+1}+\tilde{X}_1)}] &= \frac{1}{2}\frac{\alpha}{1-\alpha}(1-\alpha)^2(1+\alpha^{b+1})\alpha^{(k-1)b}\left(\frac{\theta}{1-\alpha}\right)^2 + 2\log E[e^{\theta \tilde{X}_1}] \end{aligned}$$

for $k \neq 1$. Thus

$$\begin{aligned} \phi_b(\theta) &:= E[\frac{b}{n}\sum_{k=1}^{n/b} e^{\theta(\tilde{X}_k - s)}] \\ &= \exp\left\{\frac{1}{2}\left(b - \frac{2\alpha}{1-\alpha^2}(1-\alpha^b)\right)\left(\frac{\theta}{1-\alpha}\right)^2 - \theta s\right\}, \end{aligned}$$

and

$$\begin{aligned} \delta_b &:= \sup\{\theta : \phi_b(\theta) \leq 1\} \\ &= \frac{2(1-\alpha)^2 s}{b - \frac{2\alpha}{1-\alpha^2}(1-\alpha^b)}. \end{aligned}$$

Now $E[e^{\theta(\tilde{X}_{k+1}+\tilde{X}_1)}] = e^{C\beta^{k-1}}$, where $\beta = \alpha^b$ and

$$C = \frac{1}{2}\frac{\alpha}{1-\alpha^2}(1-\beta)^2\left(\frac{\delta_b}{1-\alpha}\right)^2(1+\alpha\beta).$$

Finally,

$$\begin{aligned} \sigma_b^2 &= \lim_{n\to\infty} \frac{b}{\phi_b'(\delta_b)}\left[\phi_b(2\delta_b) - 1 + 2\sum_{k=1}^{n/b-1}\left(1-\frac{kb}{n}\right)\left(E[e^{\delta_b(\tilde{X}_{k+1}+\tilde{X}_1)}] - 1\right)\right] \\ &= \frac{b}{\phi_b'(\delta_b)}\left[\phi_b(2\delta_b) - 1 + 2\sum_{k=0}^{\infty}\left(e^{C\beta^k} - 1\right)\right], \end{aligned}$$

since

$$\lim_{n\to\infty}\sum_{k=1}^{n/b-1}\frac{kb}{n}\left(e^{C\beta^k} - 1\right) = 0,$$

for $\beta < 1$.

Appendix 2

Here we recall the rudiments of thermodynamics and explain briefly the connection with the theory of large deviations. For a readable account of this approach to thermodynamics (which goes back to Gibbs in 1873), see Callen [4].

For simplicity, we consider the thermodynamics of a gas consisting of molecules of a single chemical species. The entropy per unit volume of the gas is a function $s(u, \rho)$ of u, the internal energy per unit volume, and ρ, the mass per unit volume. The function $s(u, \rho)$ is continuously differentiable (expressing the continuity of the intensive variables, such as pressure) and concave (expressing thermodynamic stability). All bulk properties of the gas can be derived from the entropy function, using standard formulae (see [4]). For example, the entropy function of an ideal gas is given (up to an additive constant) by

$$s(u, \rho) = k\rho \left\{ \frac{f}{2} \ln u - \left(1 + \frac{f}{2}\right) \ln \rho \right\}, \tag{29}$$

where k is Boltzmann's constant and f is a constant depending on the chemical species. From this function we may derive, for example, the two equations of state:

$$p = kT\rho, \quad u = \frac{f}{2} kT. \tag{30}$$

Boltzmann's remarkable formula

$$S = k \ln W, \tag{31}$$

relating the thermodynamic entropy S of a macroscopic equilibrium state to the number W of microscopic states which correspond to the macroscopic state, gives rise to problems of interpretation. To have any hope of giving the formula a precise meaning, we should pass to the limit in which the volume V of the system become infinite. In this limit we might expect a formula:

$$s(u, \rho) = \lim_{V \to \infty} \frac{k}{V} \ln W \left[H_V = uV, N_V = \rho V \right] \tag{32}$$

where $W \left[H_V = uV, N_V = \rho V \right]$ is the phase-space volume of those macroscopic states for which the Hamiltonian H_V takes the value uV and the number of particles N_V takes the value ρV. However, the existence of the limit on the right-hand side of (32) poses obvious difficulties. These were resolved by Ruelle [13] in 1965.

Ruelle's idea can be illustrated simply: let B_a be a disc of radius a centered on the point (u, ρ); one can prove that the limit

$$s[B_a] = \lim_{V\to\infty} \frac{k}{V} \ln W[(H_V/V, N_V/V) \in B_a] \tag{33}$$

exists. Now let B_a shrink to a point, defining

$$s(u, \rho) := \inf_{a>0} s[B_a]. \tag{34}$$

The argument used to establish the existence of the limit (33) proves also that $s(u, \rho)$ is a concave function. In the case of an ideal gas, it is easy to verify (using Stirling's Formula) that Ruelle's procedure yields (29).

This simple idea was developed by Ruelle and Lanford to provide a rigorous treatment of statistical thermodynamics, described in detail in Lanfords's 1971 Battelle Lectures [11]. Ruelle's idea turned out to have a surprising ramification in probability theory: Lanford used it to give a completely new proof of Cramèr's Theorem; this was the first step in an important development in the theory of large deviations.

The modern theory of large deviations began with Cramèr's refinement [6] of the weak law of large numbers.

Theorem 1 *Let $X_1, X_2, \ldots$ be a sequence of bounded, identically distributed independent random variables. There exists a concave function s such that, for every open interval J,*

$$\lim_{n\to\infty} \frac{1}{n} \ln P\left[n^{-1}(X_1 + X_2 + \cdots + X_n) \in J\right] = \sup_{x\in J} s(x). \tag{35}$$

Lanford first proved that, for each open interval $B_a := (x - a, x + a)$, the limit

$$\lim_{n\to\infty} \frac{1}{n} \ln P\left[n^{-1}(X_1 + X_2 + \cdots + X_n) \in B_a\right] \tag{36}$$

exists (the value $-\infty$ is allowed). Lanford then defined $s(x)$ by

$$s(x) := \inf_{a>0} s[B_a], \tag{37}$$

and proved that (34) holds for each open interval J.

This approach to the theory of large deviations was taken up by Bahadur and Zabell [2]; they developed it to prove a powerful generalisation of Cramèr's Theorem. Azencott [1]

and, later, Deuschel and Stroock [7], systematised these developments. A detailed review of thermodynamic aspects of large deviation theory, including an account of the part played by the grand canonical pressure (the scaled cumulant generating function used in §3), is given in [12].

We have illustrated how, from the mathematical point of view, the rate-function of large deviation theory is the same kind of object as the entropy function of thermodynamics. They have this in common: they encapsulate concisely the relevant information about the system; for this reason it makes sense to measure them.

References

[1] R. Azencott (1980). Grandes déviations et applications, in *Lecture Notes in Mathematics* **774** 2-176 Springer, Berlin.

[2] R.R. Bahadur and S.L. Zabell (1979). Large deviations of the sample mean in general vector spaces. *Ann. Prob.* **7** 587-621

[3] Patrick Billingsley (1968). *Convergence of Probability Measures.* Wiley, New York.

[4] H.B. Callen (1985). *Thermodynamics and an introduction to thermostatics.* Wiley, New York.

[5] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand and R. Weber (1993). Admission control and routing in ATM networks using inferences from measured buffer occupancy. To appear in *IEEE Trans. Comm.*

[6] H. Cramèr (1938). Sur un théorème-limite de la théorie des probabilités. *Actualités scientifiques et industrielles* **736** 5-23.

[7] J.-D. Deuschel and D.W. Stroock (1989). *Large deviations.* Academic Press, New York.

[8] G. de Veciana, C. Courcoubetis and J. Walrand (1993). Decoupling bandwidths for networks: a decomposition approach to resource management. Memorandum No. UCB/ERL M93/50, University of California.

[9] N.G. Duffield and Neil O'Connell (1993). Large deviations and overflow probabilities for the general single server queue, with applications. To appear in *Proc. Camb. Phil. Soc.*

[10] Peter W. Glynn and Ward Whitt (1993). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. To appear in *J. Appl. Prob.*.

[11] O.E. Lanford (1973). Entropy and equilibrium states in classical statistical mechanics, in *Lecture Notes in Physics* **20** 1-113 Springer, Berlin.

[12] J.T. Lewis and C.-E. Pfister (1993). Thermodynamic probability theory: some aspects of large deviations. To appear in *Theor. Prob. Appl.*

[13] D. Ruelle (1965). Correlation functionals. *J. Math. Phys.* **6** 201-220.

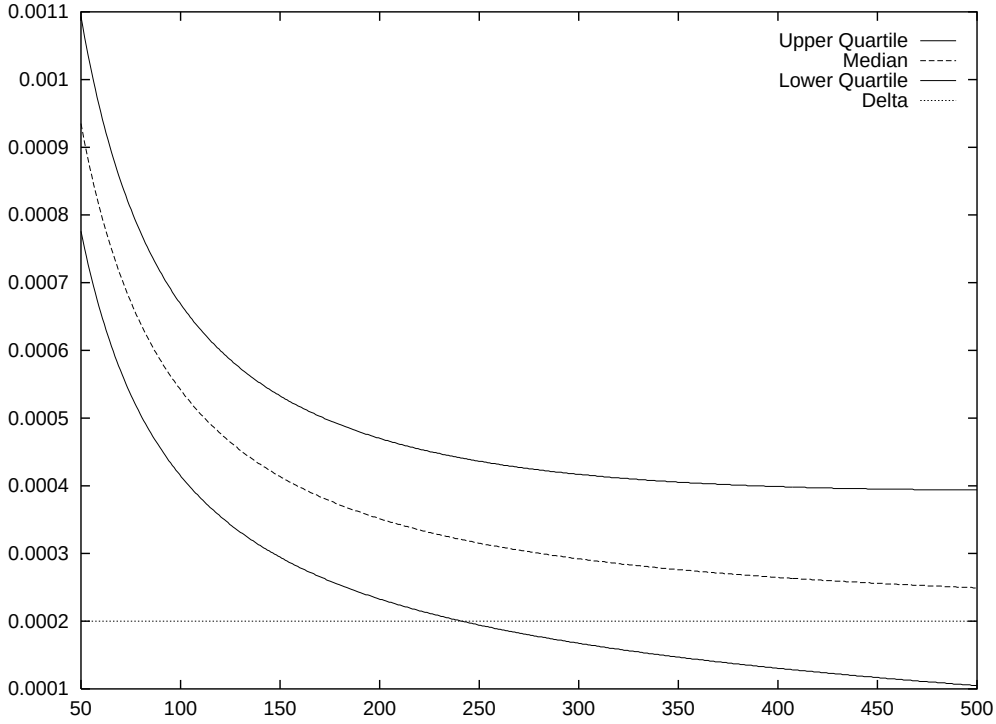

Figure 1: Gaussian first-order autoregressive process: IQR v. block size

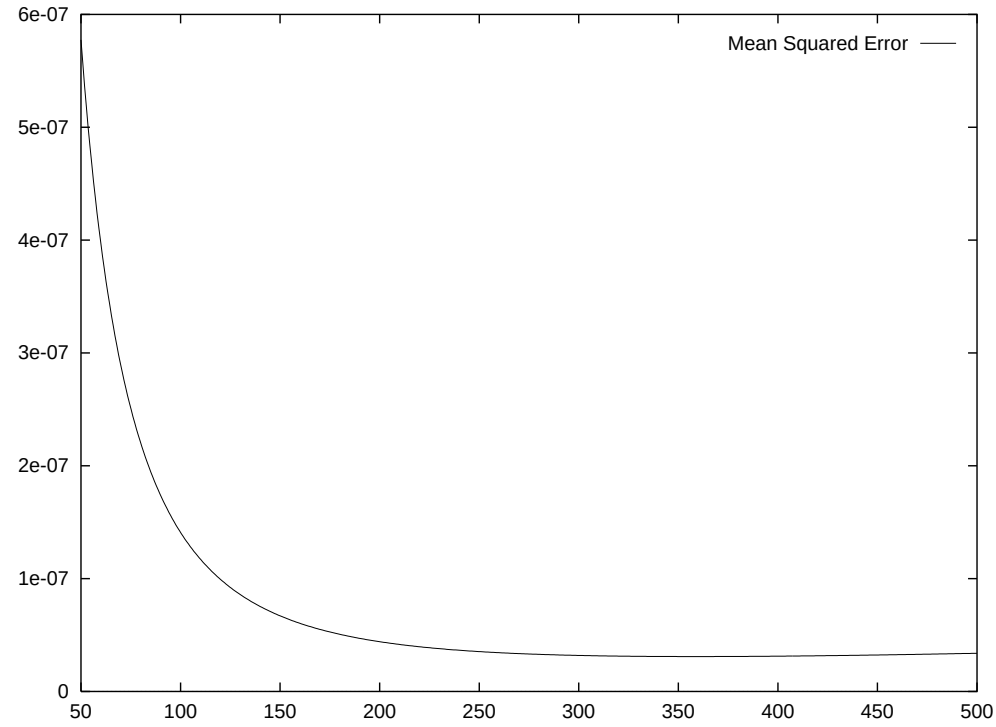

Figure 2: Gaussian first-order autoregressive process: mean squared error v. block size

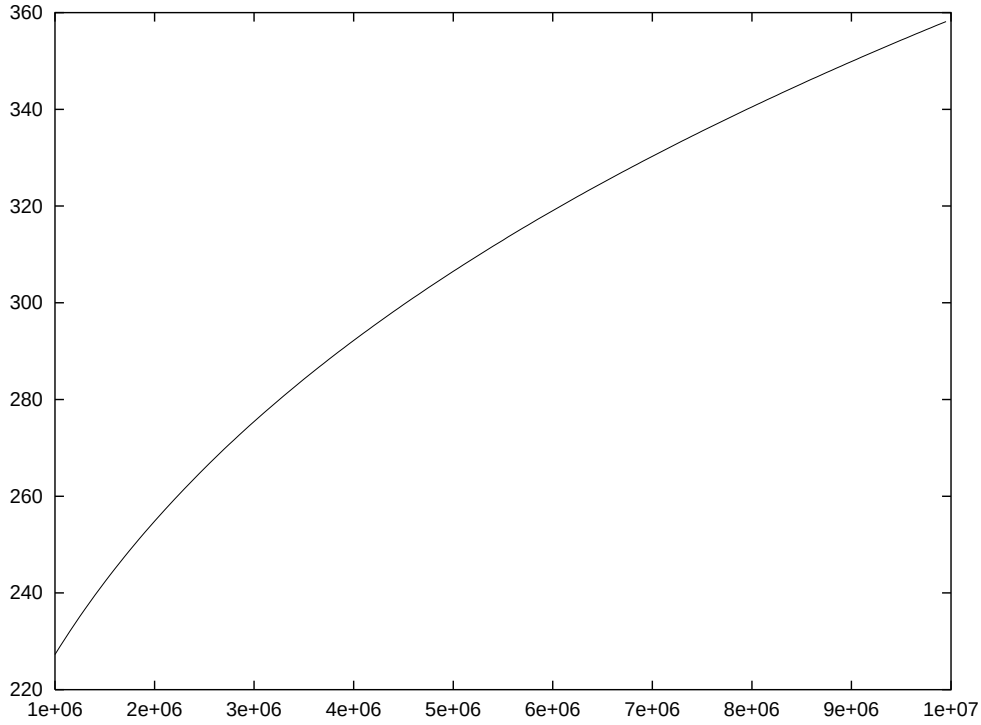

Figure 3: Gaussian first-order autoregressive process: optimal block size v. sample size

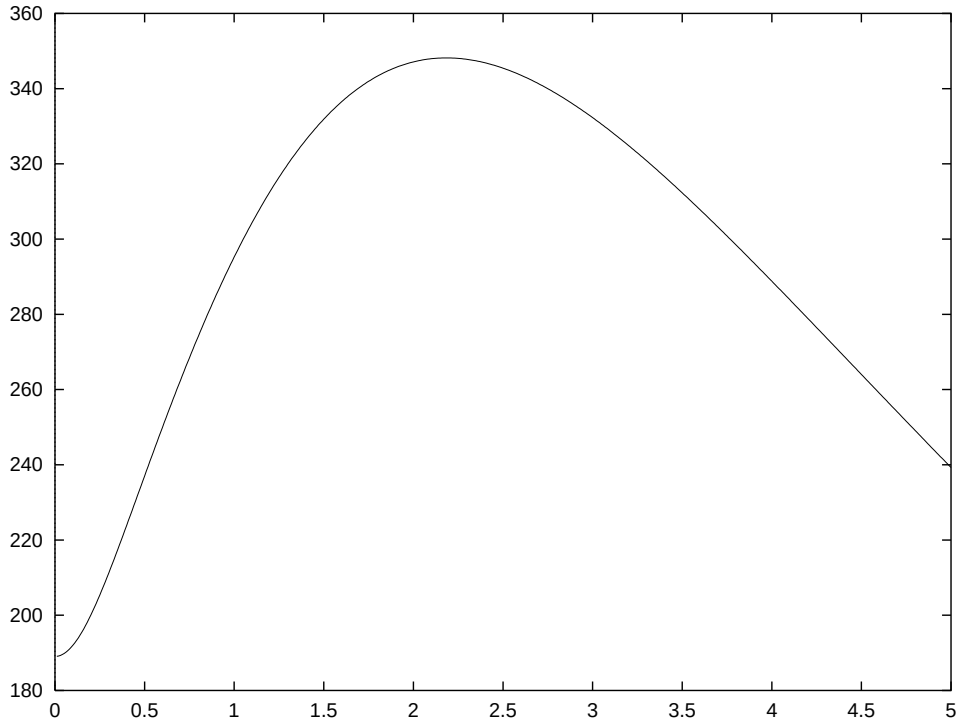

Figure 4: Gaussian first-order autoregressive process: optimal block size v. service rate

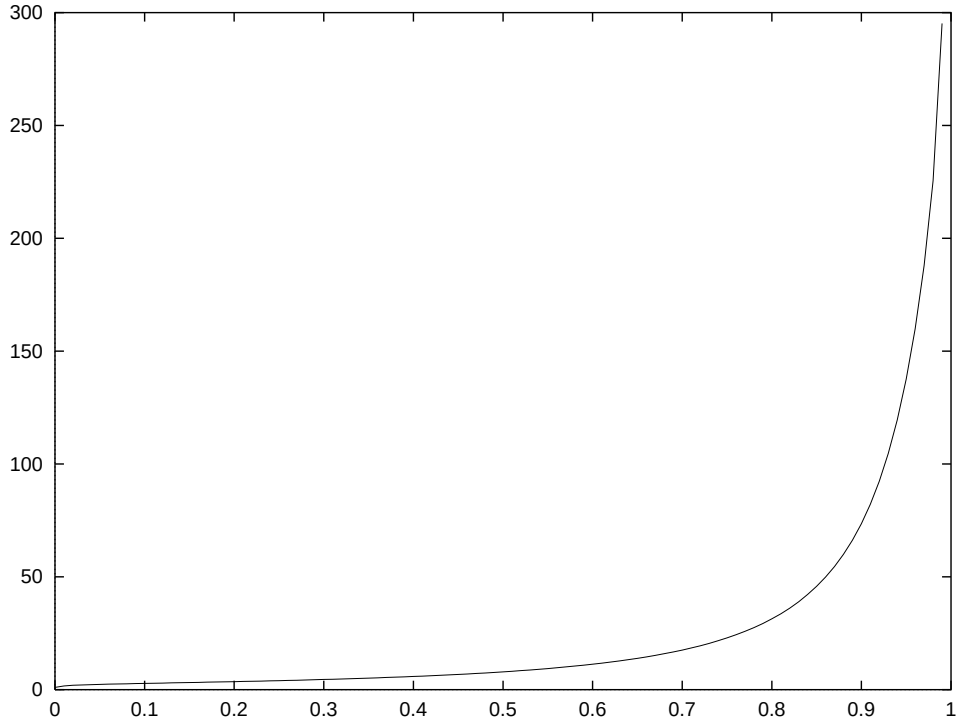

Figure 5: Gaussian first-order autoregressive process: optimal block size v. α

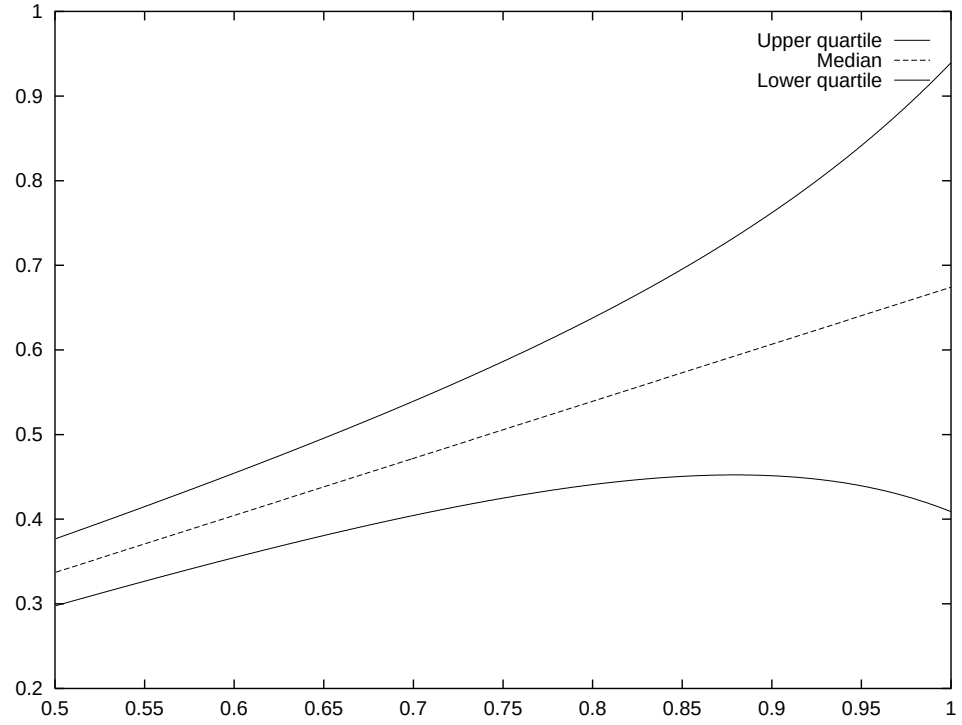

Figure 6: Gaussian first order autoregressive process: IQR v. service rate

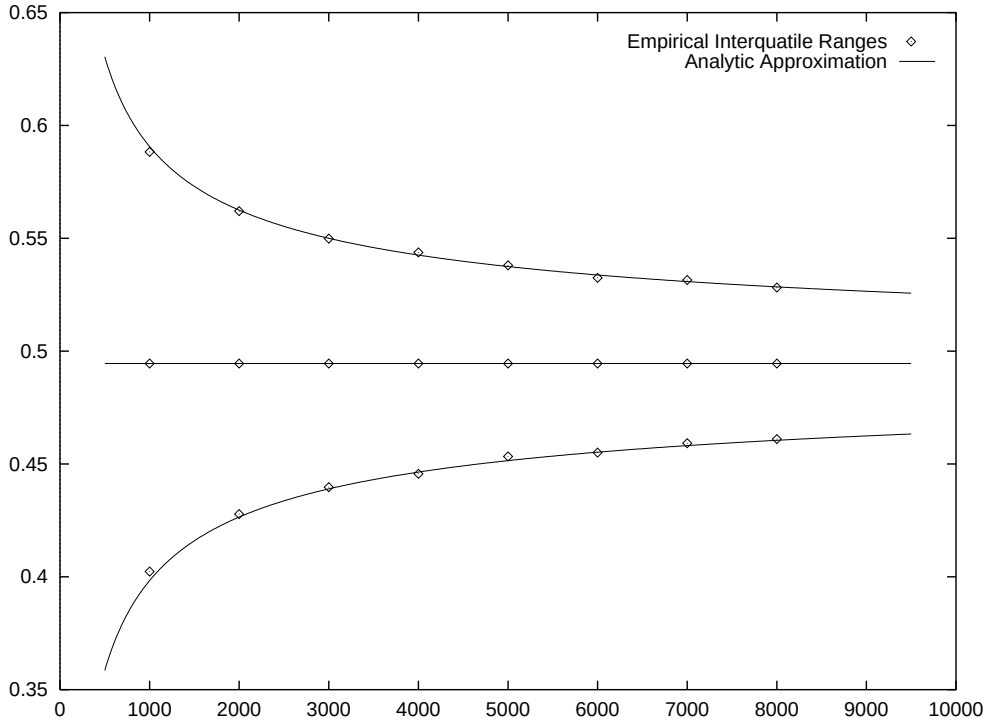

Figure 7: Interquartile range v. sample size for Bernoulli traffic

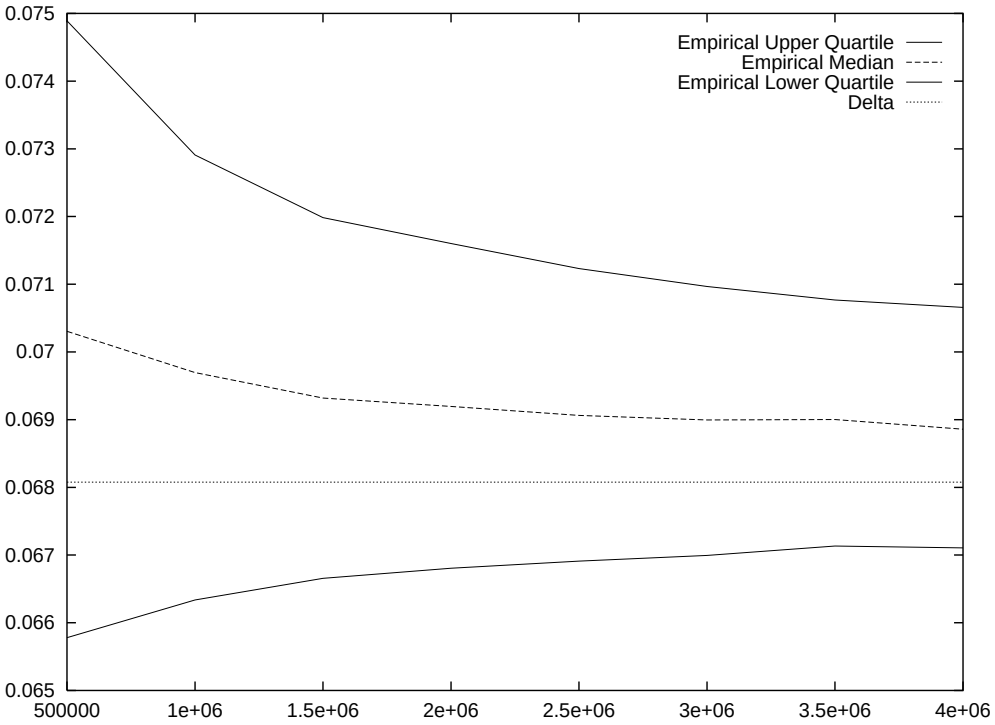

Figure 8: Interquartile range v. sample size for two-state Markov traffic

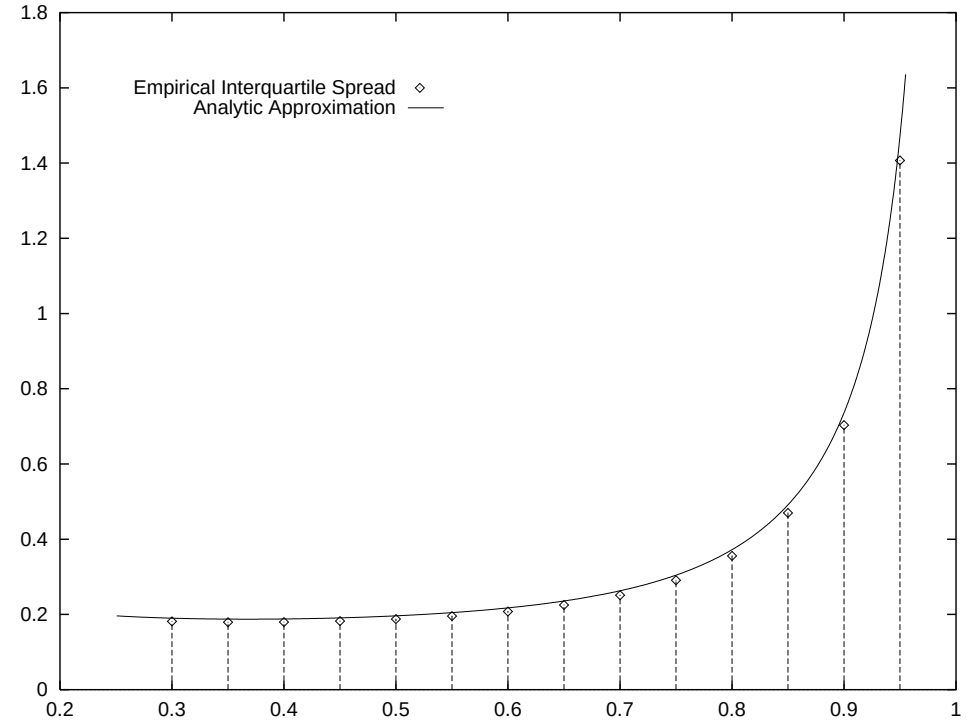

Figure 9: Interquartile spread v. service rate for Bernoulli traffic

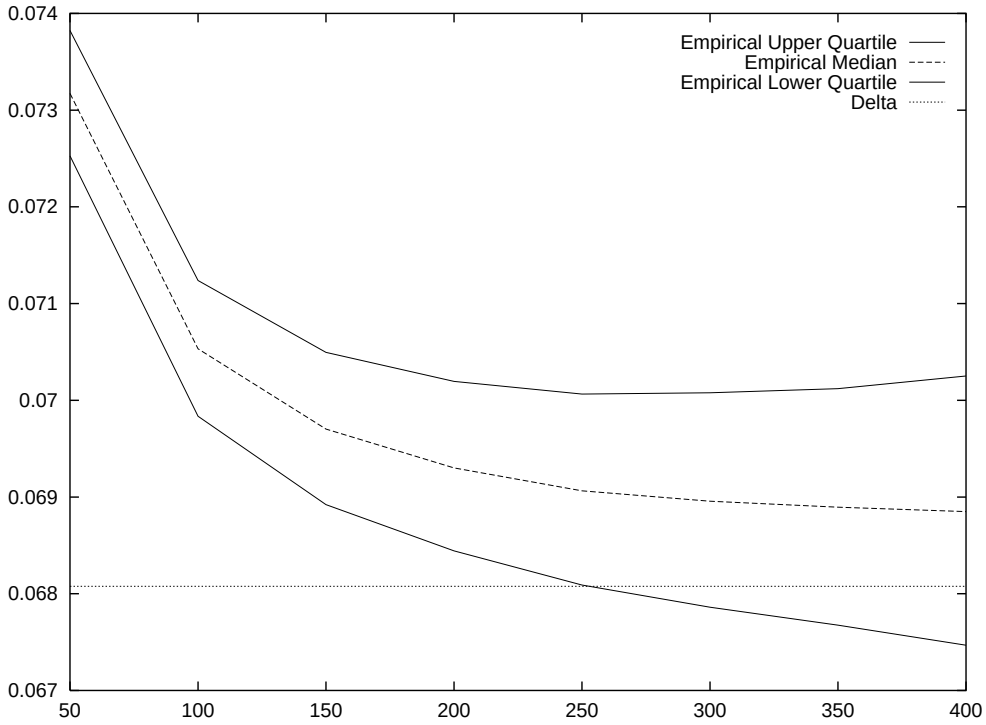

Figure 10: Interquartile v. block size for two-state Markov traffic

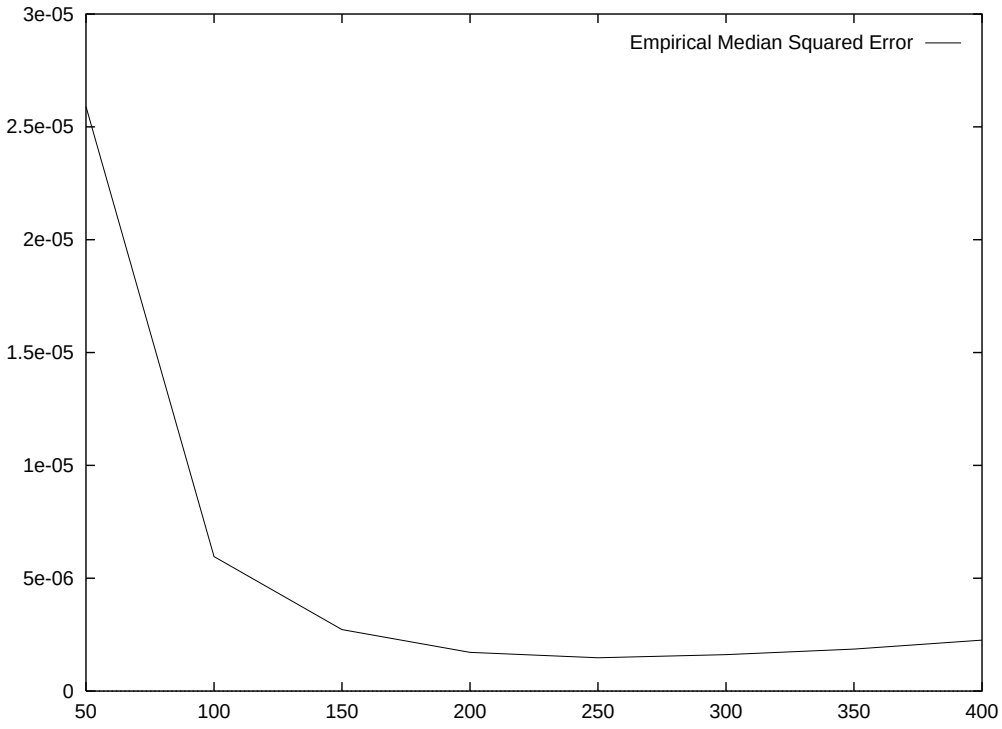

Figure 11: Median squared error v. block size for two-state Markov traffic

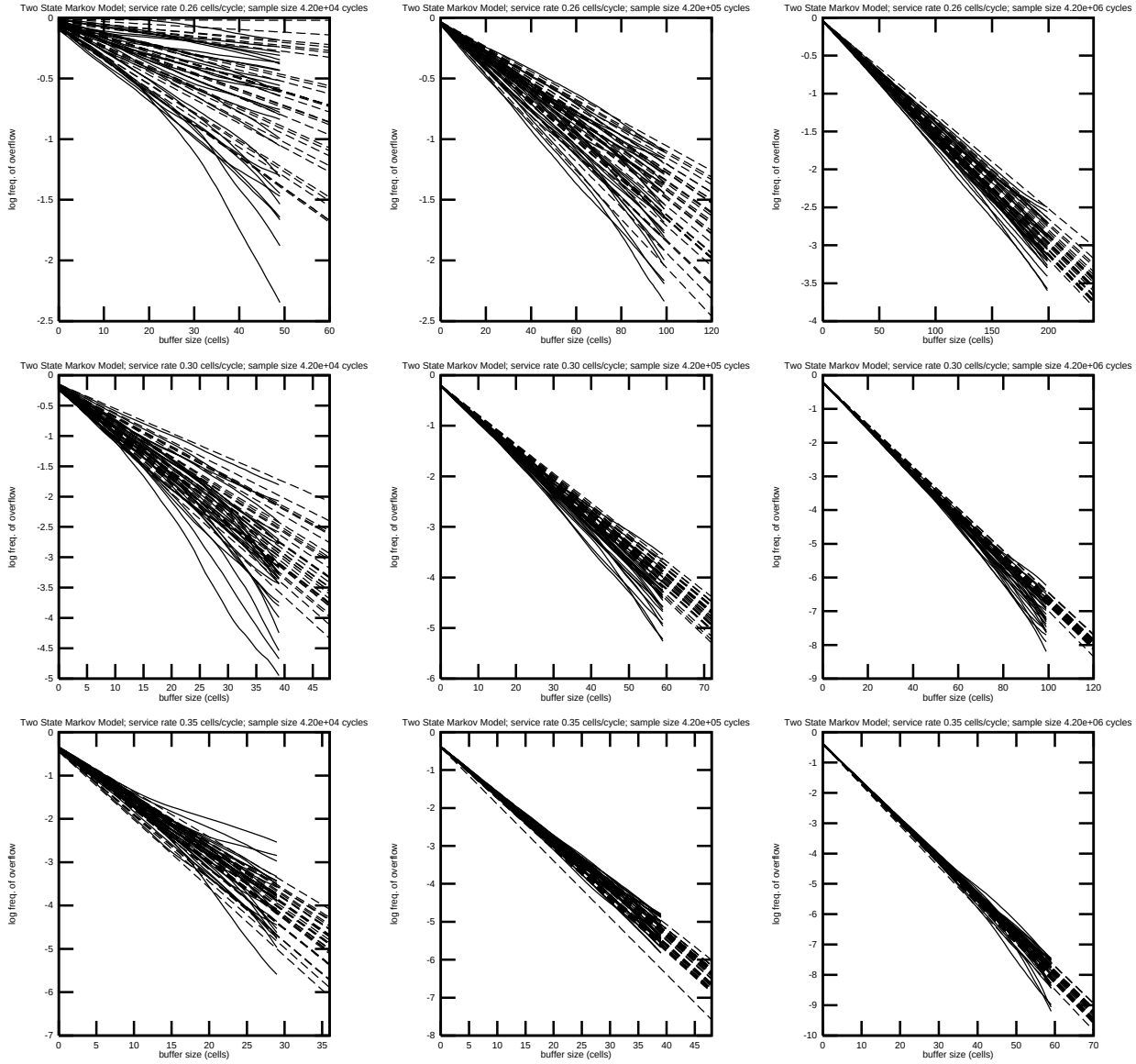

Figure 12: Watermarks and δ-estimates for a variety of sample sizes and service rates

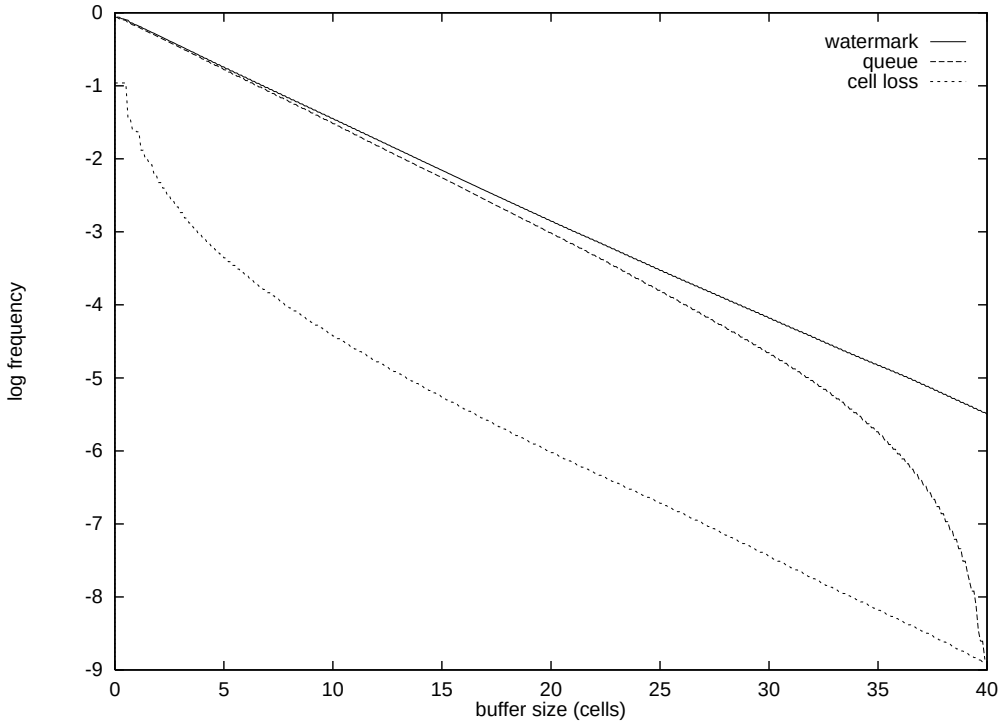

Figure 13: Watermark, finite buffer queue-length and cell-loss ratio for Bernoulli arrivals.

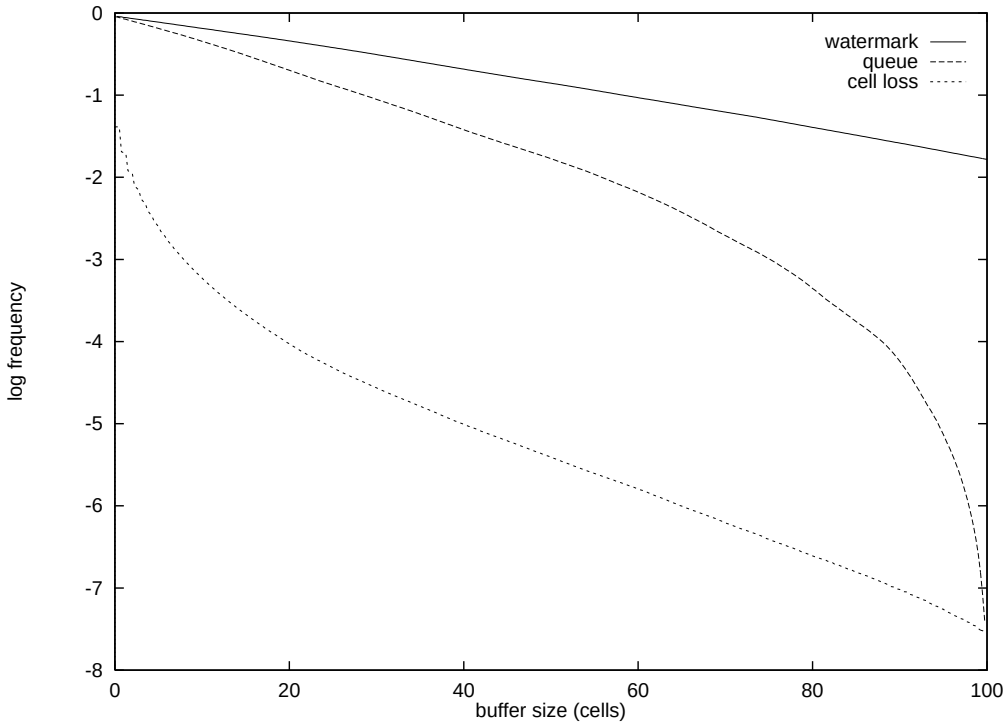

Figure 14: Watermark, finite buffer queue-length and cell-loss ratio for two-state Markov arrivals.