Applied Probability Trust (April 22, 1999)

DISTRIBUTION-FREE CONFIDENCE INTERVALS FOR MEASUREMENT OF EFFECTIVE BANDWIDTH

LÁSZLÓ GYÖRFI,* *Technical University of Budapest*

ANDRÁS RÁCZ,* *Technical University of Budapest*

KEN DUFFY,** *Dublin Institute for Advanced Studies*

JOHN T. LEWIS,** *Dublin Institute for Advanced Studies*

FERGAL TOOMEY,** *Dublin Institute for Advanced Studies*

Abstract

Hoeffding's inequality can be used in conjunction with the declared parameters of a traffic source, such as its peak rate, to obtain confidence intervals for measurements of the traffic's effective bandwidth. We describe a variety of interval-estimation procedures based on this idea, designed to provide differing degrees of robustness against non-stationarity. We also discuss how to compute confidence intervals for the effective bandwidth of an aggregate of traffic sources.

EFFECTIVE BANDWIDTH; CONFIDENCE INTERVALS; HOEFFDINGS INEQUALITY; LARGE DEVIATIONS

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60G35
SECONDARY 60G50;93E10

1. Introduction

The notion of effective bandwidth has become widely accepted as a measure of the resource requirements of bursty traffic in queuing networks. Intuitively, the effective bandwidth of a traffic source at a given network resource determines the quantity of

* Postal address: Department of Computer Science and Information Theory, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary.

** Postal address: Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland.

resource capacity which must be reserved for it in order to achieve a specified rate of data-loss. This quantity depends on the statistical properties of the traffic source, on the properties of other traffic which may be sharing the resource in question, and on the nature of the resource itself (for example, buffered or unbuffered). It has been realised, through the work of a number of authors, that the complex relationships between these different factors can be unravelled using a family of large deviation limit results for the data-loss probability. These results lead to the *effective bandwidth function*, as defined by F. Kelly [14].

The effective bandwidth σ of a stationary stochastic traffic source is given by

$$\sigma(s,t) := \frac{1}{st} \log \mathbb{E} e^{sX(t)},$$

where $X(t)$ is a random variable representing the amount of data generated by the source during intervals of length t. The parameter s in this definition is an *inverse space scale*, that is, $1/s$ may be measured in units of bits, bytes, or cells. To illustrate the significance of this function, imagine a buffered resource which is shared by L stationary, independent, and statistically identical traffic sources. If the resource has a capacity of Lc units of data per unit time, and a buffer to hold Lb units of data, then the steady-state rate of data-loss R_L satisfies

$$\lim_{L\to\infty} \frac{1}{L} \log R_L = \sup_{t>0} \inf_{s\geq 0} \Big(t\sigma(s,t) - tc - b \Big).$$

Here σ is the effective bandwidth function of each identical source. Results of this type, and generalisations to the case of non-identical sources, have been proved by A. Simonian and J. Guibert [18], D. D. Bottvich and N. G. Duffield [1], and C. Courcoubetis and R. Weber [4]. More details on the origin of effective bandwidth and its uses can be found in [14].

Given a stochastic model of a traffic source, the associated effective bandwidth function can be calculated more-or-less easily from the model's parameters (see [14] for a number of examples and references to the literature). For many purposes, however, it may be more practical to determine effective bandwidths directly from traffic measurements, thereby eliminating the need to fit a stochastic model to recorded data. N. Duffield *et al.* have presented arguments for this approach in [6]. Measurements

of effective bandwidth may be useful both for off-line characterisation of traffic and for real-time control of multiplexing systems. An analysis of recorded traffic traces in terms of their measured effective bandwidths has been carried out by R. Gibbens in [9], while algorithms for connection admission control in connection-oriented networks, based on measurements of effective bandwidth, have been proposed by G. de Veciana *et al.* [7] and J. T. Lewis *et al.* [16]. A measurement-based approach to admission control and resource pricing, which is motivated by effective bandwidths but does not measure them directly, is described in [10] and [3].

In this paper we address the extent of sampling error in measured values of effective bandwidth. Sampling error is not a critical issue in off-line traffic characterisation, but assumes greater importance when measurements are used for the purpose of dynamic resource allocation. For example, M. Grossglauser and D. Tse [11] have studied the impact of sampling error on a resource operating with a measurement-based admission control system. In this setting, under-estimation of resource requirement causes the admission control to accept too many new connections, leading to violation of the data-loss target, while over-estimation leads to a reduction in both data-loss and utilisation. Grossglauser and Tse observe that the negative effects on data-loss of under-estimation exceed the positive effects of over-estimation, so that on the average the effect of sampling error is to increase the data-loss rate. The use of a certainty-equivalent point-estimate of resource requirement can therefore be expected to yield an average rate of data-loss *somewhat in excess* of the desired target. [11] describes an extreme case (not based on effective bandwidths), in which a simple system using point-estimates misses its performance target by two orders of magnitude. An important consideration is the fact that the large deviation results in which the effective bandwidth function has its origins are intended to control the frequencies of extremely rare events, whose probabilities may be of the order of 10^{-6} or less. At this level of liklihood even small sampling errors can have a significant impact; the $1 - 10^{-6}$ quantile for an estimator of effective bandwidth may be considerably larger than its mean.

Interval-estimates of bandwidth requirement are therefore desirable: if the target data-loss rate is 10^{-6}, then a $1 - 10^{-6}$ upper confidence limit for bandwidth require-

ment can be used safely as a basis for resource allocation. Approximate confidence intervals can be obtained from a Gaussian approximation in the usual manner, but this approach does not seem appropriate here due to the very low liklihood levels which are of interest. Instead we turn to concentration inequalities designed to provide rigorous upper bounds on the probabilities of rare events. Hoeffding's inequality is particularly attractive for our problem: to use it we require only an upper bound on the random variables of interest, and this can obtained directly, without further measurement, if traffic sources declare a peak rate or other token bucket constraint.

Theorem 1 (W. Hoeffding [13]) Let $Z_1, \ldots, Z_n$ be independent bounded random variables such that $Z_k \in [a_k,\ b_k]$ with probability one. Then for any $t > 0$,

$$\mathbb{P}\Big(|\sum_{k=1}^{n}(Z_k - \mathbb{E}Z_k)| \geq t\Big) \leq \exp\Big(-2t^2/\sum_{k=1}^{n}(b_i - a_i)^2\Big).$$

A succint proof of theorem 1 is given in [5]. S. Floyd uses Hoeffding's inequality in [8] to obtain an upper bound on the effective bandwidth of an aggregate of traffic sources. In section 2 we use it to compute confidence intervals for the effective bandwidth of a single source, based on measurements of source activity. The confidence limits can be chosen to converge almost surely to the true value of σ as the sample size increases, assuming that the source is stationary. Alternatively they can be chosen to provide robustness against violations of the stationarity hypothesis. Confidence intervals for the effective bandwidth of aggregate traffic are obtained in section 3.

For our purposes, the use of Hoeffding's inequality has two draw-backs. It requires independent observations of source behaviour; and it becomes tight only at large sample sizes. In practice measurements of the activity of a source made at different times cannot be assumed completely independent, although they may be approximately so if the elapsed time between measurements is large. Taking widely-separated measurements will of course increase the time required to obtain a narrow confidence interval. The numerical results in sections 2 and 3 show that the sample size required to obtain a useful confidence interval may be very large, so that the approach we describe here may not be practical for small data sets.

2. Effective Bandwidth of a Single Source

Throughout this section we let $X(t)$ be a random variable representing the amount of work generated by a stationary stochastic traffic source during intervals of length t. We assume that $X(t)$ takes values between 0 and some upper limit $p(t) > 0$, with probability one. When the peak rate P of the source is known we may take $p(t) = Pt$; more generally, if the source is policed by a token bucket with bucket size b and token fill rate c, then $X(t)$ cannot exceed $p(t) = ct + b$. In the case of multiple token bucket constraints, we set

$$p(t) = \min_i (c_i t + b_i),$$

where (b_i, c_i), $i = 1, 2, \ldots$, are the token bucket parameters.

Our aim is to estimate the value of the effective bandwidth function

$$\sigma(s,t) := \frac{1}{st} \log \mathbb{E} e^{sX(t)},$$

for given s and t, from independent observations $X(t,1), \ldots, X(t,n)$ of $X(t)$. Thus each $X(t,k)$, $k = 1, \ldots, n$, is assumed to have the same distribution as $X(t)$. Given $q > 0$ we construct a $1 - e^{-q}$ confidence interval for $\sigma(s,t)$ as follows. Let ϕ be the value of the moment generating function of $X(t)$ at s, and let $\phi(n)$ be the weighted average

$$\phi(n) := \sum_{k=1}^{n} w(k,n) e^{sX(t,k)},$$

where the weights $w(k,n)$ satisfy $0 \le w(k,n) \le 1$ for each k and $w(1,n) + \ldots + w(n,n) = 1$. $\phi(n)$ is then an unbiased estimate of ϕ. Define

$$\varepsilon(n) := \left[\frac{q}{2} (e^{sp(t)} - 1)^2 \sum_{k=1}^{n} w_n^2(k) \right]^{\frac{1}{2}},$$

and let $\alpha(n)$, $\beta(n)$ be given by

$$\begin{aligned} \alpha(n) &:= \frac{1}{st} \log\Big([\phi(n) - \varepsilon(n)] \vee 1 \Big), \\ \beta(n) &:= \frac{1}{st} \log\Big(\phi(n) + \varepsilon(n) \Big) \wedge \frac{p(t)}{t}. \end{aligned}$$

Proposition 1 $\alpha(n) < \sigma(s,t) < \beta(n)$ *with probability at least* $1 - e^{-q}$.

Proof. Since $\sigma(s,t)$ takes values between 0 and $p(t)/t$ we have

$$\mathbb{P}\Big(\sigma(s,t) \le \alpha(n) \text{ or } \beta(n) \le \sigma(s,t)\Big) = \mathbb{P}\Big(|\phi(n) - \phi| \ge \varepsilon(n)\Big).$$

Set $Z_k := w(k,n)e^{sX(t,k)}$ for $k = 1, \ldots, n$, so that $\phi(n) = Z_1 + \ldots + Z_n$. By assumption the Z_k's are independent random variables satisfying

$$w(k,n) \le Z_k \le w(k,n)e^{sp(t)} \qquad k = 1, \ldots, n.$$

Hoeffding's inequality therefore yields

$$\begin{aligned}\mathbb{P}\Big(|\phi(n) - \phi| \ge \varepsilon(n)\Big) &= \mathbb{P}\Big(|\sum_{k=1}^{n} Z_k - \mathbb{E}Z_k| \ge \varepsilon(n)\Big) \\ &\le \exp\Big(-2\varepsilon^2(n)/\sum_{k=1}^{n} w^2(k,n)(e^{sp(t)} - 1)^2\Big),\end{aligned}$$

and, inserting $\varepsilon(n)$, the right-hand side is just e^{-q}.

If the error term $\varepsilon(n)$ tends to zero as n becomes large then so does the difference between the upper and lower estimates $\beta(n)$ and $\alpha(n)$. Since $\phi(n) \ge 1$ almost surely we have for $\varepsilon(n) < 1$,

$$\begin{aligned}\beta(n) - \alpha(n) &= \frac{1}{st}\log\left(1 + \frac{2\varepsilon(n)}{\phi(n) - \varepsilon(n)}\right) \\ &\le \frac{1}{st}\log\left(1 + \frac{2\varepsilon(n)}{1 - \varepsilon(n)}\right) \qquad \text{a.s.}\end{aligned}$$

Using the inequality $\log(1 + x) \le x$,

$$\beta(n) - \alpha(n) \le \frac{2\varepsilon(n)}{st(1 - \varepsilon(n))} \qquad \text{a.s.},$$

which is of order $\varepsilon(n)$ as $\varepsilon(n) \to 0$. The size of the error depends on the choice of the weights $w(k,n)$, the optimal choice being $w(k,n) = 1/n$ for each k. In this case

$$\varepsilon(n) = (e^{sp(t)} - 1)\sqrt{\frac{q}{2n}}$$

and

$$\beta(n) - \alpha(n) \le \frac{2(e^{sp(t)} - 1)\sqrt{q}}{st[\sqrt{2n} - (e^{sp(t)} - 1)\sqrt{q}]}$$

for $n > (e^{sp(t)} - 1)^2 q/2$.

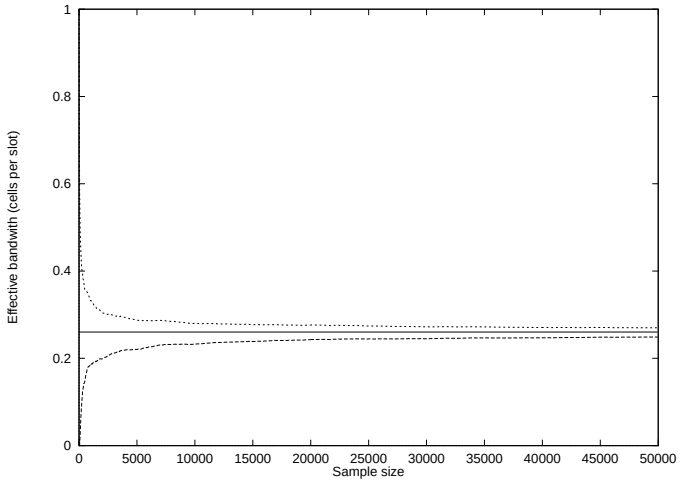

Figure 1. Estimated $1 - 10^{-6}$ confidence intervals for the effective bandwidth of traffic from a simple stochastic source. The central line is the source's true effective bandwidth.

Figure 1 illustrates the type of results which can be obtained using this scheme of equal weight for every measurement. Shown in the figure is a $1 - 10^{-6}$ confidence interval plotted against sample size, made using traffic from a two-state Markovian source. In its 'on' state this source transmits at a constant rate of 1 unit of work ('cell') per unit time ('slot'), and in its 'off' state transmits nothing. Sojourn times in both states are geometrically distributed, with a mean on-time of 5.333 slots and a mean off-time of 16 slots. The effective bandwidth estimates in the figure are for the values $s = 0.0184$ per cell and $t = 20$ slots; also shown is the true effective bandwidth of the source, for these values of s and t. The measurements of source activity used to compute the estimates were taken from back-to-back blocks of length 20 slots in a single long traffic sample. They cannot therefore be considered independent observations, and the effect of this is to introduce a small amount of bias into the estimates. Except possibly at the very largest sample sizes, this bias is more than offset by the conservative nature of the bounds.

Although the choice of equal weight for each observation of $X(t)$ minimises the value of $\varepsilon(n)$, it may not be appropriate in situations where the effective bandwidth of a source is to be monitored continously over time. This choice is not robust against

possible deviations from stationarity. We would like to be able to fix a timescale over which the estimated value of $\sigma(s,t)$ will track any changes in the true value: this can be achieved, for example, by using the weights $w(k,n)$ to implement an autoregressive filter. Each measurement $X(t,k)$ of $X(t)$ comes from a block of length t; let $\tau_k \geq t$ be the time between the end of the kth block and the end of its successor. To obtain a first-order autoregressive filter we set $w(1,1) := 1$ and, for $n \geq 2$,

$$\begin{aligned} w(1,n) &:= \gamma^{\tau_1+\cdots+\tau_n-1}, \\ w(k,n) &:= (1-\gamma^{\tau_k-1})\gamma^{\tau_k+\cdots+\tau_n-1} \qquad k \geq 2, \end{aligned}$$

where $\gamma \in (0,1)$. Just after the end of the nth block our estimate $\phi(n)$ of the moment generating function of $X(t)$ is

$$\phi(n) = \gamma^{\tau_1+\cdots+\tau_n-1} e^{sX(t,1)} + \sum_{k=2}^{n} (1-\gamma^{\tau_k-1})\gamma^{\tau_k+\cdots+\tau_n-1} e^{sX(t,k)}.$$

Thus $\phi(1) = e^{sX(t,1)}$ and for $n \geq 2$,

$$\phi(n) = (1-\gamma^{\tau_n-1})e^{sX(t,n)} + \gamma^{\tau_n-1}\phi(n-1).$$

Setting $a = 1-\gamma^{\tau_n-1}$,

$$\phi(n) = \phi(n-1) + a(e^{sX(t,n)} - \phi(n-1)),$$

which is a special case of the constant gain stochastic approximation applied in control and communication problems. Here a is typically a negative integer power of 2, so that the computation of the product on the right-hand side reduces to bit-shifting. If the sequence $X(t,1)$, $X(t,2)$, ... is i.i.d. then $\phi(1)$, $\phi(2)$, ... is a homogeneous Markov process with limit distribution concentrated around $\mathbb{E}e^{sX(t,1)}$ for 'small' a [17, 12, 15].

Assume that observations of $X(t)$ are made periodically, so that $\tau_k = \tau \geq t$ for each k. After the nth observation the error term $\varepsilon(n)$ in the upper and lower bounds of proposition 1 is given by

$$\varepsilon^2(n) = (e^{sp(t)}-1)^2 \frac{q}{2} \left(\frac{1-\gamma^\tau + 2\gamma^{(2n-1)\tau}}{1+\gamma^\tau} \right),$$

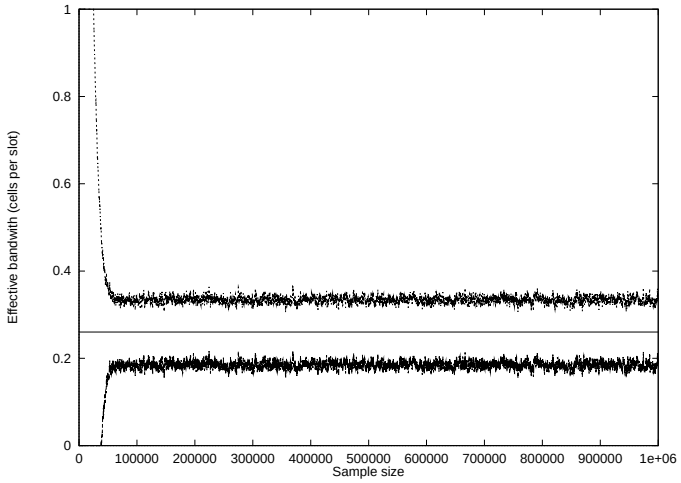

Figure 2. Estimated $1-10^{-6}$ confidence intervals for traffic from a simple stochastic source, made using a first-order autoregressive filter. The central line is the source's true effective bandwidth.

and as $n \to \infty$, $\varepsilon(n)$ converges to the non-zero value

$$\varepsilon_\infty := \lim_{n\to\infty} \varepsilon(n) = (e^{sp(t)} - 1)\sqrt{\frac{q}{2}}\left(\frac{1-\gamma^\tau}{1+\gamma^\tau}\right)^{\frac{1}{2}}.$$

Thus a large value for γ^τ (closer to one) results in a smaller confidence interval in the limit $n \to \infty$, but one which takes longer to converge, and longer to respond to any changes in the pattern of traffic from the source. The minimal value of γ^τ required to achieve a confidence interval of given size can be determined from the above equations. By varying both γ and τ the size of the limiting confidence interval and the rate of convergence can be controlled independently to some extent, but always subject to the constraint $\tau \geq t$. Note that if

$$\gamma^\tau > \frac{q(e^{sp(t)} - 1)^2 - 2}{q(e^{sp(t)} - 1)^2 + 2}$$

then $\varepsilon_\infty < 1$, and the error between the upper and lower effective bandwidth estimates $\beta(n)$ and $\alpha(n)$ satisfies

$$\lim_{n\to\infty} \beta(n) - \alpha(n) \leq \frac{2\varepsilon_\infty}{st(1-\varepsilon_\infty)} \qquad \text{a.s.}$$

Applying the auto-regressive estimation procedure to a sample of Markovian traffic produces results such as those in figure 2. This plot was made using the same traffic

source as that used to make figure 1. The period τ between measurements was set equal to $t = 20$ slots, and the value of γ was 0.999; with these parameters a reasonably narrow final confidence interval is achieved but the rate of convergence is slow compared to that in figure 1.

This estimation scheme suffers from a remaining defect, namely that the presence of periodicities in the traffic may lead to bias in the estimate $\phi(n)$ of the moment generating function. To avoid this, it is neccessary to choose the observations randomly, rather than periodically, from the available data. Let us suppose that the inter-block time τ_k is equal to $t + r_k$, where $r_1, r_2, \ldots$ are independent exponentially distributed random variables with mean $1/\lambda$ (this type of procedure was suggested by R. Gibbens in [9]). The bounds of proposition 1 continue to hold (with probability at least $1 - e^{-q}$), but the error term $\varepsilon(n)$ is now a random variable because it depends on the inter-block times through the weighting function $w(\cdot, n)$.

The sum of squared weights $w^2(1,n) + \ldots + w^2(n,n)$ satisfies

$$\begin{aligned}\sum_{k=1}^{n} w^2(k,n) &= \gamma^{2\tau_1+\ldots+2\tau_{n-1}} + \sum_{k=2}^{n}(1-\gamma^{\tau_{k-1}})^2\gamma^{2\tau_k+\ldots+2\tau_{n-1}} \\ &= \gamma^{2\tau_{n-1}}\sum_{k=1}^{n-1} w^2(k,n-1) + (1-\gamma^{\tau_{n-1}})^2.\end{aligned}$$

Therefore the sequence $\{\varepsilon^2(n) : n \geq 1\}$ evolves through the stochastic recursive equations

$$\begin{aligned}\varepsilon^2(n) &= \frac{q}{2}(e^{sp(t)}-1)^2\sum_{k=1}^{n} w^2(k,n) \\ &= a_{n-1}\varepsilon^2(n-1) + b_{n-1}, \qquad (1)\end{aligned}$$

where $\{(a_n, b_n) : n \geq 1\}$ is the i.i.d. sequence given by

$$a_n := \gamma^{2\tau_n}, \qquad b_n := \frac{q}{2}(e^{sp(t)}-1)^2(1-\gamma^{\tau_n})^2.$$

Recursive systems of this kind have been studied by W. Vervaat [19] and A. Brandt [2], who show that if

$$-\infty \leq \mathbb{E}\log|a_1| < 0 \qquad \text{and} \qquad \mathbb{E}(\log|b_1|)^+ < \infty, \qquad (2)$$

or if $\mathbb{P}(a_1 = 0) > 0$, then there is a unique stationary process $\{S(n) : n \in \mathbb{Z}\}$ satisfying equations (1), given by

$$S(n) := \sum_{k=0}^{\infty} b_{n-k-1} a_{n-k} \cdots a_{n-1} \qquad n \in \mathbb{Z}$$

(here we assume that $\{a_n\}$ and $\{b_n\}$ have been extended to become integer-indexed sequences). If $\{S'(n) : n \in \mathbb{Z}\}$ is any other solution of (1) then $|S'(n) - S(n)|$ tends to zero almost surely in n, and the distribution of $S'(n)$ converges to that of $S(0)$.

Conditions (2) are easily verified for the particular case at hand, and we conclude that $\varepsilon(n)$ converges almost surely to $\varepsilon_\infty := \sqrt{S(0)}$. If μ_a and μ_b denote, respectively, the expected values of a_1 and b_1:

$$\begin{aligned} \mu_a &:= \mathbb{E}a_1 = \frac{\lambda\gamma^{2t}}{\lambda - 2\log\gamma}, \\ \mu_b &:= \mathbb{E}b_1 = \frac{q}{2}(e^{sp(t)} - 1)^2 \left(1 - \frac{2\lambda\gamma^t}{\lambda - \log\gamma} + \frac{\lambda\gamma^{2t}}{\lambda - 2\log\gamma}\right), \end{aligned}$$

the $\mathbb{E}\varepsilon(n)$ and $\mathbb{E}\varepsilon_\infty$ satisfy

$$\mathbb{E}\varepsilon(n) \le \sqrt{\mathbb{E}\varepsilon^2(n)} = \left(\mu_b \frac{1 - \mu_a^n}{1 - \mu_a}\right)^{\frac{1}{2}}$$

and

$$\mathbb{E}\varepsilon_\infty \le \sqrt{\mathbb{E}S(0)} = \left(\frac{\mu_b}{1 - \mu_a}\right)^{\frac{1}{2}}.$$

As γ^t increases from zero to one, the right-hand side of this last inequality decreases from $(e^{sp(t)} - 1)\sqrt{q/2}$ to zero. Thus the limiting value of the error term can again be made as small as desired, at the cost of slower convergence.

3. Effective Bandwidth of Aggregate Traffic

We now assume that we are given separate data for each of several independent traffic sources which share a link, our task being to estimate the effective bandwidth of the aggregate link traffic. Let $X_l(t)$ be a random variable representing the quantity of work generated by source $l = 1, \ldots, L$ during an interval of length t, and let $p_l(t)$ be a known upper bound for $X_l(t)$, obtained in the same manner as $p(t)$ in section

2. We assume that $\{X_l(t),\ l = 1, \ldots, L\}$, are independent random variables, so that the effective bandwidth $\sigma(s,t)$ of the sum $X_1(t) + \ldots + X_L(t)$ satisfies

$$\sigma(s,t) = \sum_{l=1}^{L} \sigma_l(s,t),$$

where

$$\sigma_l(s,t) := \frac{1}{st} \log \mathbb{E} e^{sX_l(t)}$$

is the effective bandwidth of source l.

Let ϕ_l be the value of the moment generating function of $X_l(t)$ at s. Given n_l independent observations $X_l(t,1), \ldots, X_l(t,n_l)$ of $X_l(t)$, and a weighting function $w(\cdot, n_l)$, we form the estimate

$$\phi_l(n_l) := \sum_{k=1}^{n_l} w^2(k,n_l) e^{sX_l(t,k)}.$$

As before, $\phi_l(n_l)$ is an unbiased estimate of ϕ_l so long as the weights $w(\cdot, n_l)$ are chosen appropriately. Let $q > 0$ be given, and let $q_1, \ldots, q_L$ be positive numbers satisfying

$$e^{-q_1} + \ldots + e^{-q_L} = e^{-q}.$$

For each l define

$$\varepsilon_l(n_l) := \left[\frac{q_l}{2} (e^{sp_l(t)} - 1)^2 \sum_{k=1}^{n_l} w^2(k,n_l) \right]^{\frac{1}{2}}.$$

From the results of the last section we know that $\alpha_l(n_l)$

and $\beta_l(n_l)$, defined by

$$\begin{aligned} \alpha_l(n_l) &:= \frac{1}{st} \log\Big([\phi_l(n_l) - \varepsilon_l(n_l)] \vee 1 \Big), \\ \beta_l(n_l) &:= \frac{1}{st} \log\Big([\phi_l(n_l) + \varepsilon_l(n_l)] \Big) \wedge \frac{p_l(t)}{t}, \end{aligned}$$

bracket the value of $\sigma_l(s,t)$ with probability at least $1 - e^{-q_l}$.

Proposition 2 Set

$$\begin{aligned} \alpha &:= \alpha_1(n_1) + \ldots + \alpha_L(n_L), \\ \beta &:= \beta_1(n_1) + \ldots + \beta_L(n_L). \end{aligned}$$

Then $\alpha < \sigma(s,t) < \beta$ holds with probability at least $1 - e^{-q}$.

Proof. $\sigma(s,t) \leq \alpha$ or $\sigma(s,t) \geq \beta$ holds only if $\sigma_l(s,t) \leq \alpha_l(n_l)$ or $\sigma_l(s,t) \geq \beta_l(n_l)$ for some l. Therefore

$$\begin{aligned}
&\mathbb{P}(\sigma(s,t) \leq \alpha \text{ or } \sigma(s,t) \geq \beta) \\
&\quad \leq \sum_{l=1}^{L} \mathbb{P}(\sigma_l(s,t) \leq \alpha_l(n_l) \text{ or } \sigma_l(s,t) \geq \alpha_l(n_l)) \\
&\quad \leq \sum_{l=1}^{L} e^{-q_l} = e^{-q}.
\end{aligned}$$

As an illustration let us take $e^{-q_l} = e^{-q}/L$ for each l and compare proposition 2 with the results of section 2. The error terms are given by

$$\varepsilon_l(n_l) = \left[\frac{q + \log L}{2} (e^{sp_l(t)} - 1)^2 \sum_{k=1}^{n_l} w^2(k, n_l) \right]^{\frac{1}{2}}.$$

Note that the upper estimate $\beta_l(n_l)$ of the effective bandwidth of source l satisfies

$$\beta_l(n_l) \leq \frac{1}{st} \log[\phi_l(n_l) + \varepsilon_l(n_l)] \leq \frac{1}{st} \log \phi_l(n_l) + \frac{\varepsilon_l(n_l)}{st\phi_l(n_l)},$$

and hence

$$\beta = \beta_1(n_1) + \ldots + \beta_L(n_L) \leq \frac{1}{st} \sum_l \log \phi_l(n_l) + \frac{1}{st} \sum_l \frac{\varepsilon_l(n_l)}{\phi_l(n_l)}.$$

For large L the sum of the error terms on the right-hand side of this inequality grows proportionately with $L\sqrt{\log L}$. For comparison, if the aggregate traffic were treated as a single traffic stream, the results of the previous section would yield an error term proportional to $e^{sp(t)}$ with $p(t) = p_1(t) + \ldots + p_L(t)$. Thus the estimation error would grow approximately exponentially in the number of sources.

Figures 3 and 4 depict results obtained by applying this procedure to an aggregate of 100 Markovian sources of the type described in section 2. The weighting function $w(k,n) = 1/n$ was used to make the estimates in figure 3, while those of figure 4 were made using an auto-regressive filter. The results are seen to be qualitively very similar to the corresponding single source results, indicating little or no loss of precision due to aggregation.

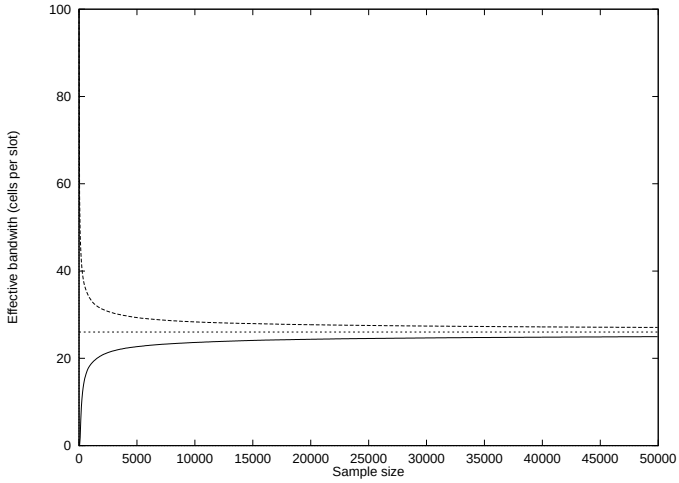

Figure 3. Estimated $1-10^{-6}$ confidence intervals for the effective bandwidth of an aggregate of 100 sources.

Other choices for the coefficients q_l are also possible; for example, we may wish to choose smaller q_l where n_l is small, or where $p_l(t)$ is large. Using the weighting function $w(k, n_l) = 1/n_l$, $k = 1, \ldots, n_l$, the error $\varepsilon_l(n_l)$ is given by

$$\varepsilon_l(n_l) = (e^{sp_l(t)} - 1)\sqrt{\frac{q_l}{2n_l}}.$$

This can be made independent of l by choosing choosing

$$q_l = \frac{cn_l}{(e^{sp_l(t)} - 1)^2}$$

where c satisfies

$$\sum_{l=1}^{L} \exp(-cn_l/(e^{sp_l(t)} - 1)) = e^{-q}. \tag{3}$$

Then

$$\varepsilon_l(n_l) = \sqrt{c/2}$$

for each $l = 1, \ldots, L$. Thus the error terms are equalised by allowing more latitude to sources with higher peak rates, or from which fewer observations have been obtained. The transcendental equation (3) must however be solved numerically.

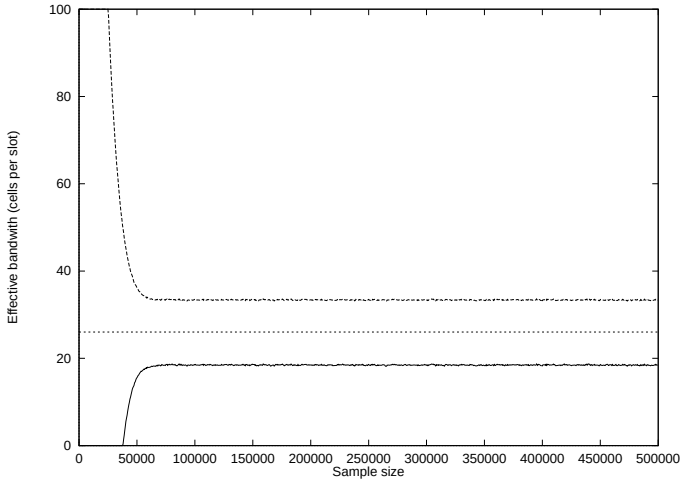

Figure 4. Estimated $1-10^{-6}$ confidence intervals for the effective bandwidth of an aggregate of 100 sources, made using a first-order autoregressive filter.

References

[1] Bottvich, D. D. and Duffield, N. G. (1996) Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. Queueing Systems, **20**.

[2] Brandt, A. (1986) The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. Adv. Appl. Prob., **18**.

[3] Courcoubetis, C., Kelly, F. P., and Weber, R. (1997) Measurement-based charging in communications networks. Preprint.

[4] Courcoubetis, C. and Weber, R. (1996) Buffer overflow asymptotics for a switch handling many traffic sources. J. Appl. Prob., **33**.

[5] Devroye, L., Györfi, L., and Lugosi. G. (1996) *A Probabilistic Theory of Pattern Recognition.* Springer.

[6] Duffield, N. G., Lewis, J. T., O'Connell, N., Russell, R., and Toomey, F. (1995) Entropy of ATM traffic streams: a tool for estimating

QoS parameters. *IEEE Journal on Selected Areas in Communications*, **13(6)**.

[7] DE VECIANA, G., KESIDIS, G., AND WALRAND, J. (1995) Resource management in wide-area ATM networks using effective bandwidths. *IEEE Journal on Selected Areas in Communications*, **13(6)**.

[8] FLOYD, S. Comments on measurement-based admissions control for controlled-load services. Technical report, ICSI, 1996. http://www.aciri.org/floyd/.

[9] GIBBENS, R. J. (1996) Traffic characterisation and effective bandwidths for broadband network traces. In F. P. Kelly, S. Zachary, and I. B. Ziedins, editors, *Stochastic Networks: Theory and Applications*. Oxford University Press.

[10] GIBBENS, R. J. AND KELLY, F. P. (1997) Measurement-based connection admission control. In V. Ramaswami and P. E. Wirth, editors, *Teletraffic Contributions for the Information Age: Proceedings of the 15th International Teletraffic Conference, Washington, D. C.* Elsevier, Amsterdam.

[11] GROSSGLAUSER, M. AND TSE, D. (1997) A framework for robust measurement-based admission control. In *Proceedings of ACM SIGCOMM '97*.

[12] GYÖRFI, L. AND WALK, H. (1996) On the averaged stochastic approximation for linear regression. *SIAM Journal of Control and Optimization*, **34**.

[13] HOEFFDING, W. (1963) Probability inequalities for sums of bounded random variables. J. Amer. Statist. Soc., **58**.

[14] KELLY, F. P. (1996) Notes on effective bandwidths. In F. P. Kelly, S. Zachary, and I. B. Ziedins, editors, *Stochastic Networks: Theory and Applications*. Oxford University Press.

[15] KUSHNER, H. J., AND SHWARTZ, A. (1984) Weak convergence and asymptotic properties of adaptive filters with constant gains. *IETIT*, **IT-30**.

[16] LEWIS, J. T., RUSSELL, R., TOOMEY, F., MCGURK, B., CROSBY, S., AND LESLIE, I. (1998) Practical connection admission control for ATM networks based on on-line measurements. *Computer Communications*, **21**.

[17] PFLUG, G. CH. (1986) Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal of Control and Optimization*, **24**.

[18] SIMONIAN, A. AND GUIBERT, J. (1995) Large deviations approximation for fluid queues fed by a large number of on/off sources. *IEEE Journal on Selected Areas in Communications*, **13(6)**.

[19] VERVAAT, W. (1979) On a stochastic difference equation and a representation of non-negative infinitely-divisble random variables. Adv. Appl. Prob., **11**.