Bias Correction in Effective Bandwidth Estimation

A. J. Ganesh*

Department of Computer Science

University of Edinburgh

JCMB, Kings Buildings, Edinburgh EH9 3JZ

Email : *A.J.Ganesh@bristol.ac.uk*

May 30, 1996

Abstract

Call admission in ATM networks involves a trade-off between ensuring an adequate quality of service to users and exploiting the scale efficiencies of statistical multiplexing. Achieving a good trade-off requires some knowledge of the source traffic. Its effective bandwidth has been proposed as a measure that captures characteristics which are relevant to quality of service provisioning. The effective bandwidth of a source is not known *a priori*, but needs to be estimated from an observation of its output. We show that direct estimators that have been proposed for this purpose are biased when the source traffic is autocorrelated. By explicitly computing the bias for auto-regressive and Markov sources, we devise a bias correction scheme that does not require knowledge of the model parameters. This is achieved by exploiting a scaling property of the bias that is insensitive to model parameters, and that has the same form for both auto-regressive and Markov sources. This leads us to conjecture that the scaling property may be valid in greater generality and can be used to obtain unbiased effective bandwidth estimates for real traffic. Use of our bias correction technique enables us to obtain accurate estimates of effective bandwidths using relatively short block lengths. The latter is important both because the variance of the estimator increases with the block length, and because real traffic may well be non-stationary, requiring that estimates be obtained from short data records.

*Research supported by a BP/Royal Society of Edinburgh postdoctoral fellowship.

1 Introduction

The traffic in an ATM (asynchronous transfer mode) network is packaged into cells of fixed size (53 bytes) and carried over links between switches in the network. Traffic sources are bursty and so for periods of time cells may arrive at a switch faster than they can be switched to output links. Switches are buffered to cope with overflow traffic but cells will be lost when buffers are full. Cells arriving when there is a large backlog in the buffer will suffer large delays. The network limits call acceptance in order to ensure an adequate quality of service (QoS), specified as a bound on the probability of cell loss, or of cell delays exceeding a threshold. A bound of 10^{-8} on the cell loss probability is a typical requirement. In order to achieve efficient network utilization while maintaining QoS requirements, knowledge of traffic characteristics is essential.

Traditional approaches to traffic characterisation have relied on modelling. A statistical model of the source is proposed, whose parameters are estimated from observations of its output. The estimated parameters are used to compute cell loss probabilities which form the basis of call admission decisions. Such an approach suffers from several shortcomings. Automating model selection is difficult. The number of model parameters needed is usually large, which increases the computational cost and reduces the statistical accuracy of the estimates. The effect of errors in parameter estimates on computed cell loss probability is not easy to see or to incorporate in the estimation procedure.

A number of recent papers [1, 5, 8] show that, in networks with large buffer capacities, cell loss and delay probabilities depend only on the large deviations rate function of the traffic. The detailed description of the traffic obtained by estimating a model of it contains a great deal of information which is superfluos to QoS provisioning. In view of this, Courcoubetis et al. [2] and Duffield et al. [7] suggest characterising the large deviations behaviour of the traffic, which can be done without reference to a specific source model and does not suffer from the drawbacks mentioned above. The approach in [2, 7] is to directly estimate either the large deviations rate function of the traffic stream or a related quantity, its effective bandwidth. Variants of this quantity were introduced by Guerin et al. [9], Hui [10], Kelly [11] and de Veciana and Walrand [5] as a measure of the resource requirements of a source.

In the next section we briefly review the effective bandwidth concept and its relevance to call admission control, and discuss some of the problems in estimating it from observed traffic. In Section 3, we show that the effective bandwidth estimator proposed in [7] is biased for auto-regressive (AR) sources. By obtaining an expression for the bias, we suggest a procedure to correct for it. These results are extended to a Markovian source model in Section 4. It needs to be emphasized that our bias correction method does not require estimation of the model parameters; it exploits scaling properties that are common to all models in a certain class. Our results imply that both AR and Markov sources fall into this class; what other sources do remains an open question. We describe some simulation results in Section 5 and conclude in Section 6.

2 Estimating effective bandwidths

2.1 Effective bandwidths

An output buffer in an ATM switch can be modeled as a single server queue in discrete time, with stationary ergodic arrivals $\{X_n\}$ and constant service rate C. Suppose for now that the buffer capacity is infinite. For stability we require that $EX_1 < C$, i.e., the service rate exceeds the mean arrival rate. The limiting cumulant generating function (cgf) of the input traffic stream is defined as

$$\Lambda(\theta) = \lim_{n\to\infty} \Lambda_n(\theta), \quad \text{where} \quad \Lambda_n(\theta) = \frac{1}{n}\log E\left[\exp\, \theta(X_1 + \ldots + X_n)\right]. \tag{1}$$

We assume that this limit exists as an extended real number for all $\theta \in I\!R$, and satisfies the assumptions of the Gärtner-Ellis theorem, [4]. This assumption is not very restrictive and is satisfied if, for instance, the traffic is an ARMA process, or a Poisson or fluid process modulated by a finite state Markov chain. Then, it was shown in [5] that the tail of the queue length distribution satisfies the following condition.

$$\lim_{B\to\infty} \frac{1}{B}\log \mathbf{P}(Q \geq B) < -\theta \quad \Longleftrightarrow \quad \alpha(\theta) \triangleq \frac{\Lambda(\theta)}{\theta} < c. \tag{2}$$

Here, Q denotes a random variable with the stationary queue length distribution, and $\alpha(\cdot)$ is called the effective bandwidth function. If the actual buffer capacity, B, is large compared to the number of cells arriving in one time slot, then the cell loss rate is well approximated by $\mathbf{P}(Q > B)$, the probability that the queue size in the corresponding infinite buffer queue exceeds B.

For a given buffer size B, and a desired bound p on the cell loss probability, let $\theta = -(\log p)/B$. If we accept an additional call only when the effective bandwidth, $\alpha(\theta)$, of the resulting traffic stream is less than the service capacity C, then (2) suggests that if B is large, the cell loss rate won't exceed p. In order to implement this call admission policy, we need to estimate $\alpha(\theta)$, or equivalently, $\Lambda(\theta)$, from observations of the arrival process $\{X_n\}$.

2.2 Estimation

Suppose we have a record $\{X_1, \ldots, X_N\}$ of the arrival process over N time slots, for some large N. Fix $n, m \in I\!N$, and define $K = \lfloor N/m - 1\rfloor$, where $\lfloor x\rfloor$ denotes the largest integer less than or equal to x. Let

$$S_k \;=\; \sum_{j=km+1}^{km+n} X_j, \quad 0 \leq k \leq K, \tag{3}$$

denote the total traffic arriving in the k^{th} block, where the blocks are of length n and the total number of blocks is K. Adjacent blocks overlap in $n-m$ time slots if $m < n$ and are non-overlapping if $m \geq n$. The S_k are identically distributed by stationarity, but not necessarily independent. Define

$$\hat{M}_{n,m}(\theta) \;=\; \frac{1}{K}\sum_{k=1}^{K} e^{\theta S_k}, \qquad \hat{\Lambda}_{n,m}(\theta) \;=\; \frac{1}{n}\log \hat{M}_{n,m}(\theta). \tag{4}$$

$\hat{M}_{n,m}(\theta)$ is an unbiased estimator of the moment generating function of the total arrivals in a block, to which it converges, as $N \to \infty$, by the strong law of large numbers. Note that, for fixed n, the bias of $\hat{\Lambda}_{n,m}(\theta)$ as an estimator of $\Lambda(\theta)$ may not go to zero as $N \to \infty$. However, $\hat{\Lambda}_{n,m}(\theta)$ is an asymptotically unbiased estimator of $\Lambda_n(\theta)$, defined in (1). So, by (1), the bias in the estimate of $\Lambda(\theta)$ can be made arbitrarily small by choosing n sufficiently large. With this justification, $\hat{\Lambda}_{n,m}(\theta)$, for sufficiently large n, has been proposed as an estimator for $\Lambda(\theta)$ by Crosby et al. [3] and Duffield et al. [7].

In practice, however, it is hard to know when n is large enough. Furthermore, it can be shown that the variance of the estimator increases rapidly with n, so the use of large values of n is undesirable. Such is also the case if the traffic statistics are not stationary but vary slowly over time, as may well be true of real traffic. Duffield et al. [7] suggest choosing n to achieve a trade-off between bias and variance. Apart from the difficulty of obtaining estimates of the bias and variance, we feel that this does not fully exploit the information available in the traffic data. Instead, we derive an explicit relationship between the bias and the block length n, whose form does not depend on the model parameters. Using this, we show how estimates $\hat{\Lambda}_{n,m}(\theta)$, obtained for a number of different values of n, can be combined to yield an unbiased estimate. This is done in the context of an AR traffic model in the next section, and of a Markov model in Section 4. We also obtain estimates of the variance as a function of m and n.

3 Autoregressive sources

3.1 Bias of the estimator

Let the traffic be modeled as an AR(M) process, i.e.,

$$X_n = \sum_{i=1}^{M} a_i X_{n-i} \,+\, U_n, \qquad U_n \;\text{ i.i.d. } \;\sim N(\mu, \sigma^2), \tag{5}$$

for given constants, a_i, and a given innovations process $\{U_n\}$, which is white Gaussian with mean μ and variance σ^2. We assume that the AR process is stable, i.e., that all roots of the characteristic equation $1 - \sum_{k=1}^{M} a_k z^{-k} \;=\; 0$ lie within the unit circle in the complex plane. Fix a block length n and a shift m, and let S_k be as in (3), for all $k \in \mathbb{Z}$. Define

$$H_1(\omega) = \sum_{k=1}^{M} a_k e^{i\omega k}, \qquad H_2(\omega) = \sum_{k=0}^{n-1} e^{i\omega k}. \tag{6}$$

Then,

$$H_2(\omega) = \exp\left[\frac{i\omega(n-1)}{2}\right] \frac{\sin(n\omega/2)}{\sin(\omega/2)}. \tag{7}$$

Let $U(\omega) = \sum_{k=-\infty}^{\infty} U_k e^{i\omega k}$ denote the Fourier transform of the innovations process U_k. Similarly, define $X(\omega)$ and $S(\omega)$ to be the Fourier transforms of the sequences X_k and S_k respectively. $H_1(\cdot)$, $H_2(\cdot)$ are deterministic whereas $U(\cdot)$, $X(\cdot)$ and $S(\cdot)$ are random functions; the infinite sums defining them converge almost surely. We have

$$S(\omega) = H_2(\omega)X(\omega) = H_1(\omega)H_2(\omega)U(\omega). \tag{8}$$

Define $H(\omega) = H_1(\omega)H_2(\omega)$. Observe that S_k and X_k, being linear combinations of Gaussian random variables U_k, are Gaussian. Hence,

$$\Lambda_n(\theta) \triangleq \frac{1}{n}\log E\left[\exp(\theta S_0)\right] = \frac{1}{n}\Big[\theta E(S_0) + \frac{\theta^2}{2}\mathrm{Var}(S_0)\Big]. \tag{9}$$

But $S_0 = (1/2\pi)\int_{-\pi}^{\pi} S(\omega)d(\omega)$ and so, by (8),

$$E[S_0] = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(\omega)E[U(\omega)]d\omega.$$

Since $U(\omega)$ is the Fourier transform of U_k, which are i.i.d. $\sim N(\mu, \sigma^2)$, we have $E[U(\omega)] = 2\pi\mu\delta(\omega)$, where $\delta(\cdot)$ denotes the Dirac delta function. Hence,

$$E[S_0] = \mu H(0) = \mu n H_1(0), \tag{10}$$

where the second equality is because $H_2(0) = n$. Let $\mathcal{S}_U(\omega)$, $\mathcal{S}_S(\omega)$ denote the power spectral densities of $\{U_k\}$ and $\{S_k\}$ respectively. Then, $\mathcal{S}_U(\omega) \equiv \sigma^2$ for all $\omega \in [-\pi, \pi]$ and $\mathcal{S}_S(\omega) = H(\omega)H(-\omega)\mathcal{S}_U(\omega)$. Hence, by Parseval's theorem,

$$\begin{aligned}
\mathrm{Var}(S_0) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{S}_S(\omega)d\omega \\
&= \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi} H(\omega)H(-\omega)d\omega \\
&= \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi} H_1(\omega)H_1(-\omega)\frac{\sin^2(n\omega/2)}{\sin^2(\omega/2)}d\omega \\
&= n\sigma^2 H_1^2(0) + \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi} \frac{H_1(\omega)H_1(-\omega) - H_1^2(0)}{\sin^2(\omega/2)}\Big[1 - \cos(nw)\Big]d\omega,
\end{aligned}$$

where the third equality is obtained by substituting $H(\omega) = H_1(\omega)H_2(\omega)$, and the last equality is because $\int_{-\pi}^{\pi}[\sin^2(n\omega/2)/\sin^2(\omega/2)]d\omega = 2\pi n$. Observe from (6) that $H_1(\omega)$ is a C^∞ function of ω on $[-\pi,\pi]$, i.e., it has derivatives of all orders. Hence, so is $[H_1(\omega)H_1(-\omega) - H_1^2(0)]/\sin^2(\omega/2)$. This is obvious when $\omega \neq 0$; when $\omega = 0$, it can be seen by expanding $H_1(\omega)$ in a Taylor series around 0. Therefore, the integral of this function over $[-\pi,\pi]$ is some constant, c, that does not depend on the block length n. In addition, over any one period of $\cos(n\omega)$, it varies by no more than $2\pi L/n$, where

$$L = \max_{-\pi<\omega<\pi}\left|\frac{d}{d\omega}\left\{\frac{H_1(\omega)H_1(-\omega) - H_1^2(0)}{\sin^2(\omega/2)}\right\}\right|$$

is finite. Therefore, we have

$$\left|\int_{-\pi}^{\pi}\frac{H_1(\omega)H_1(-\omega) - H_1^2(0)}{\sin^2(\omega/2)}\cos(n\omega)d\omega\right| \leq \frac{\pi^2 L}{n}.$$

Combining the above results, we get

$$\mathrm{Var}(S_0) = n\sigma^2 H_1^2(0) + c + O\left(\frac{1}{n}\right), \tag{11}$$

for a constant c that does not depend on n. Therefore, by (9), (10) and (11),

$$\Lambda_n(\theta) = \theta\mu H_1(0) + \frac{\theta^2\sigma^2}{2}H_1^2(0) + \frac{c}{n} + O(\frac{1}{n^2}). \tag{12}$$

Letting $n \to \infty$ above, we have by (1) that

$$\Lambda(\theta) = \theta\mu H_1(0) + \frac{\theta^2\sigma^2}{2}H_1^2(0). \tag{13}$$

Since $\hat{\Lambda}_{n,m}(\theta)$ is an asymptotically unbiased estimator of $\Lambda_n(\theta)$, we have from (12), 13) that its asymptotic bias as an estimator of $\Lambda(\theta)$ is $c/n + O(1/n^2)$. Ignoring the last term for moderately large n, the asymptotic bias is a constant times $1/n$; this form does not depend on the model parameters though the value of the constant does. Based on this result, we suggest the following improved estimator. Obtain $\hat{\Lambda}_{n,m}(\theta)$ for a range of values of n, plot them against $1/n$ and find a linear least squares fit. The intercept of this line at $1/n = 0$ gives an estimate $\hat{\Lambda}(\theta)$, which the above analysis suggests will be unbiased. An alternative approach is possible: observe from (12) that if we fit a straight line to a plot of $n\hat{\Lambda}_{n,m}(\theta)$ versus n, it should have intercept approximately equal to c and slope approximately equal to $\Lambda(\theta)$. It remains to be seen which of these estimators performs better; simulation studies can throw light on the question.

The estimator suggested above has a number of advantages. Because bias is corrected for, we can use moderately small values of n, for which the bias is large.

It can be shown that the variance of the estimator increases exponentially in n; therefore, it is desirable not to use large n. In addition, real traffic may well be non-stationary, in which case estimates must be obtained from data records that are sufficiently short that the source can be thought of as being stationary over that time. To do this, it is essential that we be able to work with small values of n.

3.2 Variance of the estimator

Observe from (4) that

$$E\left[\hat{M}^2_{n,m}(\theta)\right] = \frac{1}{K^2}\sum_{i,j=1}^{K} E\left[e^{\theta(S_i+S_j)}\right],$$

for S_k defined in (3). Using the stationarity of the source, we can rewrite the above as

$$E\left[\hat{M}^2_{n,m}(\theta)\right] = \frac{1}{K^2}\sum_{k=-K}^{K}(K-|k|)E\left[e^{\theta(S_0+S_k)}\right].$$

Let $Y_k(n) = S_n + S_{k+n}$ and note that the $Y_k(n)$ are Gaussian since the S_n are. Therefore

$$E\left[\hat{M}^2_{n,m}(\theta)\right] = \frac{1}{K^2}\sum_{k=-K}^{K}(K-|k|)\exp\left[\theta EY_k(0) + \frac{\theta^2}{2}\mathrm{Var}(Y_k(0))\right]. \tag{14}$$

Now, $E[Y_k(0)] = 2E[S_0]$ for all k, by stationarity. Also,

$$E[\hat{M}_{n,m}(\theta)] = E[\exp(\theta S_0)] = \exp\left(\theta E[S_0] + \frac{\theta^2}{2}\mathrm{Var}(S_0)\right).$$

Noting that $(1/K^2)\sum_{k=-K}^{k}\left(K-|k|\right) = 1$, we see from (14) that

$$\mathrm{Var}(\hat{M}_{n,m}(\theta)) = e^{2\theta ES_0}\frac{1}{K^2}\sum_{k=-K}^{K}(K-|k|)\left[e^{\theta^2\mathrm{Var}(Y_k(0))/2} - e^{\theta^2\mathrm{Var}(S_0)}\right],$$

and so,

$$\mathrm{cv}^2(\hat{M}_{n,m}(\theta)) = \frac{1}{K^2}\sum_{k=-K}^{K}(K-|k|)\left[e^{\theta^2(\mathrm{Var}(Y_k(0))-2\mathrm{Var}(S_0))/2} - 1\right], \tag{15}$$

where, for a random variable X, $\mathrm{cv}^2(X) = \mathrm{Var}(X)/(EX)^2$ denotes its squared coefficent of variation. We shall now evaluate the right hand side of (15). Fix k and define $\mathcal{S}_Y(\omega)$ to be the power spectral density of the sequence $Y_k(n)$. Then

$\mathcal{S}_Y(\omega) = H_3(\omega)H_3(-\omega)\mathcal{S}_S(\omega)$, where $H_3(\omega) = 1+\exp(-i\omega km)$ and m is the shift between adjacent blocks in the sum. Hence, by Parseval's theorem,

$$\begin{aligned}
&\mathrm{Var}(Y_k(0)) - 2\mathrm{Var}(S_0) \\
= \; & \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\mathcal{S}_Y(\omega) - 2\mathcal{S}_S(\omega)\right] d\omega \\
= \; & \frac{\sigma^2}{\pi}\int_{-\pi}^{\pi} H_1(\omega)H_1(-\omega)\frac{\sin^2(n\omega/2)}{\sin^2(\omega/2)}\cos(mk\omega)d\omega \\
= \; & \frac{\sigma^2}{\pi}\int_{-\pi}^{\pi}\left[H_1^2(0) + O(\omega^2)\right]\frac{\sin^2(n\omega/2)}{\sin^2(\omega/2)}\cos(mk\omega)d\omega.
\end{aligned}$$

Neglecting the $O(\omega^2)$ term and using the fact that $\sin^2(n\omega/2)/\sin^2(\omega/2) = (1 + \ldots + \exp(i\omega n))(1 + \ldots + \exp(-i\omega n))$, we can rewrite the above as:

$$\mathrm{Var}(Y_k(0)) - 2\mathrm{Var}(S_0) \;=\; 2\sigma^2 H_1^2(0)(n - |k|m)^+, \tag{16}$$

where x^+ denotes $\max\{x, 0\}$. For notational convenience, we define the quantity $\gamma = \theta^2\sigma^2 H_1^2(0)$. Recall that K was defined to be $\lfloor N/m - 1\rfloor$, so $K \approx N/m$. Using this, we get on substituting (16) in (15) and simplifying that

$$\begin{aligned}
\mathrm{cv}^2(\hat{M}_{n,m}(\theta)) \;=\; & \frac{m(1+e^{-\gamma m})}{1-e^{-\gamma m}}\frac{e^{\gamma n}-1}{N} + \frac{m}{1-e^{-\gamma m}}\frac{2n}{N} \\
& - \frac{2m^2 e^{-\gamma m}}{(1-e^{-\gamma m})^2}\frac{e^{\gamma n}-1}{N^2} - \frac{2n}{N} + \frac{n^2}{N^2}.
\end{aligned} \tag{17}$$

The last two terms above don't depend on m. It can be verified by differentiation that the second term is increasing and the third term is decreasing for all positive m, whereas the first term reaches its minimum at the solution of $\gamma m = 1 - e^{-\gamma m}$. Let m^* denote the solution. Then $m^* = c/\gamma$ for some constant $c \in [0, 1]$, and so it does not depend on n or N. It is clear that the minimum over m of $\mathrm{cv}^2(\hat{M}_{n,m}(\theta))$ is achieved at some $m < m^* < 1/\gamma$. It is not possible to find the exact value of m that minimizes the variance of the estimator without knowledge of the model parameters. However, the above calculations imply that this value is small, and that it does not grow with n or N. We feel that the choice $m = 1$ is reasonable in practice. The more important point to note is that the squared coefficient of variation of the estimator depends on n as $e^{\gamma n} - 1$, for fixed m and N. It can be shown that the variance of $\hat{\Lambda}_{n,m}$, defined in (4), is approximately the same as $\mathrm{cv}^2(\hat{M}_{n,m})$. Therefore, this variance grows rapidly with the block length n, making the use of large values of n undesirable.

4 Markov-modulated sources

Let $\{\xi_0, \xi_1, \ldots\}$ be a Markov chain on a finite state space $\{1, \ldots, M\}$. The traffic process $\{X_n\}$ is specified by the vector $(\mu_1, \ldots, \mu_M)$; if $\xi_n = j$, then $X_n = \mu_j$.

Let P be the transition probability matrix of the Markov chain, assumed to be irreducible and aperiodic, and let π denote its unique stationary distribution. Given $\theta > 0$, define $Q = P\,\mathrm{diag}\,(\exp\theta\mu)$, where $\mathrm{diag}\,(\exp\theta\mu)$ denotes the matrix whose j^{th} diagonal entry is $\exp(\theta\mu_j)$, and whose off-diagonal entries are zero. Define $f_i(n) = E[\exp\theta(X_1 + \ldots + X_n)|\xi_0 = i]$. Then, by the Markov property,

$$f_i(n) = \sum_{j=1}^{M} p_{ij} e^{\theta j} f_j(n-1), \quad \text{i.e.,} \quad f(n) = Qf(n-1). \tag{18}$$

Combined with the fact that $f(0) = \mathbf{1}$, the vector of ones, the above implies that in stationarity,

$$E\left[e^{\theta(X_1+\ldots+X_n)}\right] = \pi Q^n \mathbf{1}. \tag{19}$$

Define ρ to be the spectral radius of Q. Since Q is primitive, i.e., $Q \geq 0$ and $Q^k > 0$ for some k, we have by the Perron-Frobenius theorem (see [12] for example) that ρ is an eigenvalue of Q and that all other eigenvalues are strictly smaller than ρ in absolute value. Writing $Q = S^{-1}JS$, where J is in Jordan form, it is clear from (19) that

$$E\left[e^{\theta(X_1+\ldots+X_n)}\right] = c\rho^n + \sum_{j=2}^{m} \sum_{k=0}^{r_j-1} c_{jk} \lambda_j^{n-k}, \tag{20}$$

for some constants $c \neq 0$, c_{jk}. Let λ_2 be the second largest, in absolute value, of the eigenvalues of Q. Then $|\lambda_2| < \rho$, and it follows from (20) that

$$\frac{1}{n} \log E\left[e^{\theta(X_1+\ldots+X_n)}\right] = \log\rho + \frac{\log c}{n} + o\left(\frac{\alpha^n}{n}\right), \tag{21}$$

for any $\alpha > |\lambda_2|/\rho$. Ignoring the last term above for moderately large n, we have from (1) that

$$\Lambda(\theta) = \log\rho, \quad \Lambda_n(\theta) = \log\rho + \frac{c}{n}, \tag{22}$$

for some finite constant c. Since $\hat{\Lambda}_{n,m}(\theta)$, defined in (4), is an asymptotically unbiased estimator of $\Lambda_n(\theta)$, it follows from the above that its bias as an estimator of $\Lambda(\theta)$ is approximately a constant times $1/n$. Note that the this form does not depend on the model parameters, even though the exact value of the constant c does. Also, the form is identical to what we obtained in (12), (13) for autoregressive sources. Therefore, the same techniques for bias correction that were developed in the autoregressive setting are also applicable to Markov modulated sources.

5 Simulation results

5.1 Autoregressive source

We simulated traffic from the AR(3) model

$$X_n = 1.4X_{n-1} - 0.73X_{n-2} + 0.2X_{n-3} + U_n,$$

where U_n is a white noise process with mean 0.52 and variance 0.1566. These values were chosen so that the resulting AR process has mean 4 and variance 1. The AR process has a real pole at 0.8 and a pair of complex poles at $0.3 \pm 0.4i$. The results reported below are based on a record of 10,000 observations of the above process taken after it had reached stationarity.

Figure 1 shows estimates of the cumulant generating function, $\Lambda_n(\theta)$, corresponding to a range of blocklengths n between 10 and 50, for $\theta = 0.10$, 0.15, 0.20 and 0.25. These estimates were obtained using the procedure outlined in Section 2.2. The dotted horizontal line in each plot depicts the true value of $\Lambda(\theta)$, calculated using (13). Observe that the estimated values of the cumulant generating function (cgf), and hence of the effective bandwidth, underestimate the true value. The error grows larger as the blocklength increases, and also as the parameter θ increases. If we take the value corresponding to the largest blocklength in the experiment, then the underestimate of the cgf ranges from 2.2% for $\theta = 0.1$ to 9.6% for $\theta = 0.25$. A more conservative approach would be to use the largest value of the estimated cgf, over all choices of blocklengths. This procedure yields estimates that fall short of the true value by 2% when $\theta = 0.1$ and by 5.8% when $\theta = 0.25$.

It may appear that these underestimates are small. Nevertheless, they are significant for the following reason. The aim is to implement a call admission policy which can ensure a guaranteed quality of service in the ATM network. This guarantee usually takes the form of a bound on the cell loss probability, a typical value of the bound being around 10^{-8}. Around such low values, the actual cell loss probabilities are very sensitive to the service rate. Therefore, even a small underestimate of the effective bandwidth can result in a large degradation in the quality of service. On the other hand, small overestimates are harmless, resulting only in a small loss of efficiency in network utilization.

Observe from Figure 1 that the cgf estimates decrease progressively with increasing blocklength. The reason for this is that the cgf estimates are sensitive to the tail behaviour of the corresponding random variables. Given a finite data sample, we are less likely to observe large deviations from the mean in long blocks than in short ones. Since it is such large deviations that provide the main contribution to the cgf estimate, it is no surprise that the estimates decrease with blocklength. This observation points up a deficiency of the traditional approach to estimation, which is based on choosing a fairly large blocklength, in keeping with the asymptotic nature of the formula, (1).

We now turn to the estimates obtained using our proposed approach to bias correction, see Figure 2. Here, we have plotted the cgf estimates obtained earlier against $1/n$, where n denotes the blocklength. Keeping only those values of n that are less than the value at which the maximum estimate obtains, we fit a straight line to the corresponding estimates and extrapolate it to $1/n = 0$. The intercept is our corrected estimate of the cgf. The true value of the cgf is again shown by a dotted horizontal line. It is clear from the figures that the bias correction yields a substantial reduction in the estimation error. For the corrected cgf values, the underestimates range from 0.7% for $\theta = 0.1$ to 2.2% for $\theta = 0.25$. Our findings are summarised in Table 5.1 below. We show for each value of θ the true cgf, the estimate corresponding to the largest blocklength, the conservative estimate which is the largest over all blocklengths, and finally our corrected estimate, denoted $\Lambda^*(\theta)$. For each estimate, the percentage by which it underestimates the true value is shown in brackets.

θ	$\Lambda(\theta)$	$\hat{\Lambda}_{48}(\theta)$	$\max_n \hat{\Lambda}_n(\theta)$	$\Lambda^*(\theta)$
0.10	0.446	0.436 (2.2)	0.437 (2.0)	0.443 (0.7)
0.15	0.704	0.673 (4.4)	0.681 (3.3)	0.697 (1.0)
0.20	0.985	0.917 (6.9)	0.940 (4.6)	0.971 (1.4)
0.25	1.290	1.166 (9.6)	1.215 (5.8)	1.262 (2.2)

Table 1. *Comparison of different estimators of the cgf.*

5.2 Markov-modulated source

Traffic was simulated by multiplexing the output of ten independent Markovian On-Off sources. Each source produces one unit of output when it is On and none when it is Off. The transition probability from Off to On is 0.1, and that from On to Off is 0.2 in each time slot, independent of the past, and sources remain in their current state with the residual probability. Consequently, in stationarity, each source is On with probability 1/3 and Off with probability 2/3. The aggregate traffic can be modelled using a Markov chain with eleven states, the states corresponding to the number of sources that are On. It is however simpler to compute the effective bandwidth of the aggregate traffic by making use of the additivity of the effective bandwidth for independent sources. Using this fact and the results of Section 4, we have calculated $\Lambda(\theta)$ for $\theta = 0.10$, 0.15, 0.20 and 0.25. The calculated values are shown in Table 5.2, as are estimates based on a record of 10,000 observations of the simulated aggregate traffic.

Figure 3 shows estimates of the cumulant generating function, $\Lambda_n(\theta)$, corresponding to a range of blocklengths n between 10 and 50, for $\theta = 0.10$, 0.15, 0.20 and 0.25. These estimates were obtained using the procedure outlined in Section 2.2. The dotted horizontal line in each plot depicts the true value of $\Lambda(\theta)$, calculated using (22). Observe that the estimated values of the cumulant generating function (cgf), and hence of the effective bandwidth, underestimate the true value. The estimates obtained using our proposed method for bias correction are shown in Figure 2. Here, we have plotted the cgf estimates obtained earlier against $1/n$, where n denotes the blocklength. Using only the ten smallest blocklengths, we fit a straight line to the corresponding estimates and extrapolate it to $1/n = 0$. The intercept is our corrected estimate of the cgf. The true value of the cgf is again shown by a dotted horizontal line.

Our findings are summarised in Table 5.2 below. We show for each value of θ the true cgf, the estimate corresponding to the largest blocklength, the conservative estimate which is the largest over all blocklengths, and finally our corrected estimate, denoted $\Lambda^*(\theta)$. For each estimate, the percentage by which it underestimates the true value is shown in brackets. It is clear from these figures that bias correction achieves considerable improvement over just using some large block length. However, in this example at least, the method of using the most conservative estimate over all block lengths works almost equally well.

θ	$\Lambda(\theta)$	$\hat{\Lambda}_{48}(\theta)$	$\max_n \hat{\Lambda}_n(\theta)$	$\Lambda^*(\theta)$
0.10	0.4014	0.4017(-0.07)	0.4017(-0.07)	0.4008 (0.15)
0.15	0.6568	0.6396 (2.6)	0.6447 (1.8)	0.6510 (0.9)
0.20	0.9489	0.8841 (6.8)	0.9143 (3.7)	0.9188 (3.2)
0.25	1.275	1.132 (11.2)	1.196 (6.2)	1.190 (6.7)

Table 2. *Comparison of different estimators of the cgf.*

6 Conclusions

We studied the estimation of effective bandwidths for autoregressive and Markov models of the source traffic. These are popular models for variable bit rate sources in ATM networks. We showed that a direct effective bandwidth estimator, which has been used by a number of researchers, is biased and suggested an alternative, based on a scaling property of the bias. Our bias correction procedure does not require estimation of the model parameters. Simulation studies reported here suggest that this procedure can yield significant improvements in the accuracy of effective bandwidth estimates. Our results are preliminary, and a more thorough simulation study is needed to confirm the findings.

Acknowledgements I would like to think Profs. Peter Green and Bernard Silverman for their warm hospitality at the Dept. of Mathematics, University of Bristol, where this work was done.

References

[1] C.S. Chang, Stability, queue length and delay of deterministic and stochastic queueing networks, *IEEE Trans. Autom. Contr.* 39(5): 913–931, 1994.

[2] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand and R. R. Weber, Admission control and routing in ATM networks using inferences from measured buffer occupancy, *IEEE Trans. on Comm.* 43: 1778-1784, 1995.

[3] S. Crosby, I. Leslie, J. T. Lewis, N. O'Connell, R. Russell and F. Toomey, Bypassing modelling: an investigation of entropy as a traffic descriptor in the Fairisle ATM network, *Proc. 12th U.K. Teletraffic Symposium*, London 1995.

[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett, 1993.

[5] G. de Veciana and J. Walrand, Effective bandwidths: call admission, traffic policing and filtering for ATM networks, *Queueing Systems* 20: 37-59, 1995.

[6] N. G. Duffield, Exponential bounds for queues with Markovian arrivals, *Queueing Systems* 17: 413-430, 1994.

[7] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russell and F. Toomey, Entropy of ATM traffic streams: A tool for estimating QoS parameters, *IEEE J. Sel. Areas in Comm.* 13(6): 981-990, 1995.

[8] P. Glynn and W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue, *J. Appl. Probab.* 31A: 131-156, 1994.

[9] R. Guerin, H. Ahmadi and M. Nagshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE J. Sel. Areas in Comm.* 9: 968-981, 1991.

[10] J. Y. Hui, Resource allocation for broadband networks, *IEEE J. Sel. Areas in Comm.* 6: 1598-1608, 1988.

[11] F. P. Kelly, Effective bandwidths at multi-class queues, *Queueing Systems* 9: 5-15, 1991.

[12] E. Seneta, *Non-negative Matrices*, George Allen and Unwin, 1973.

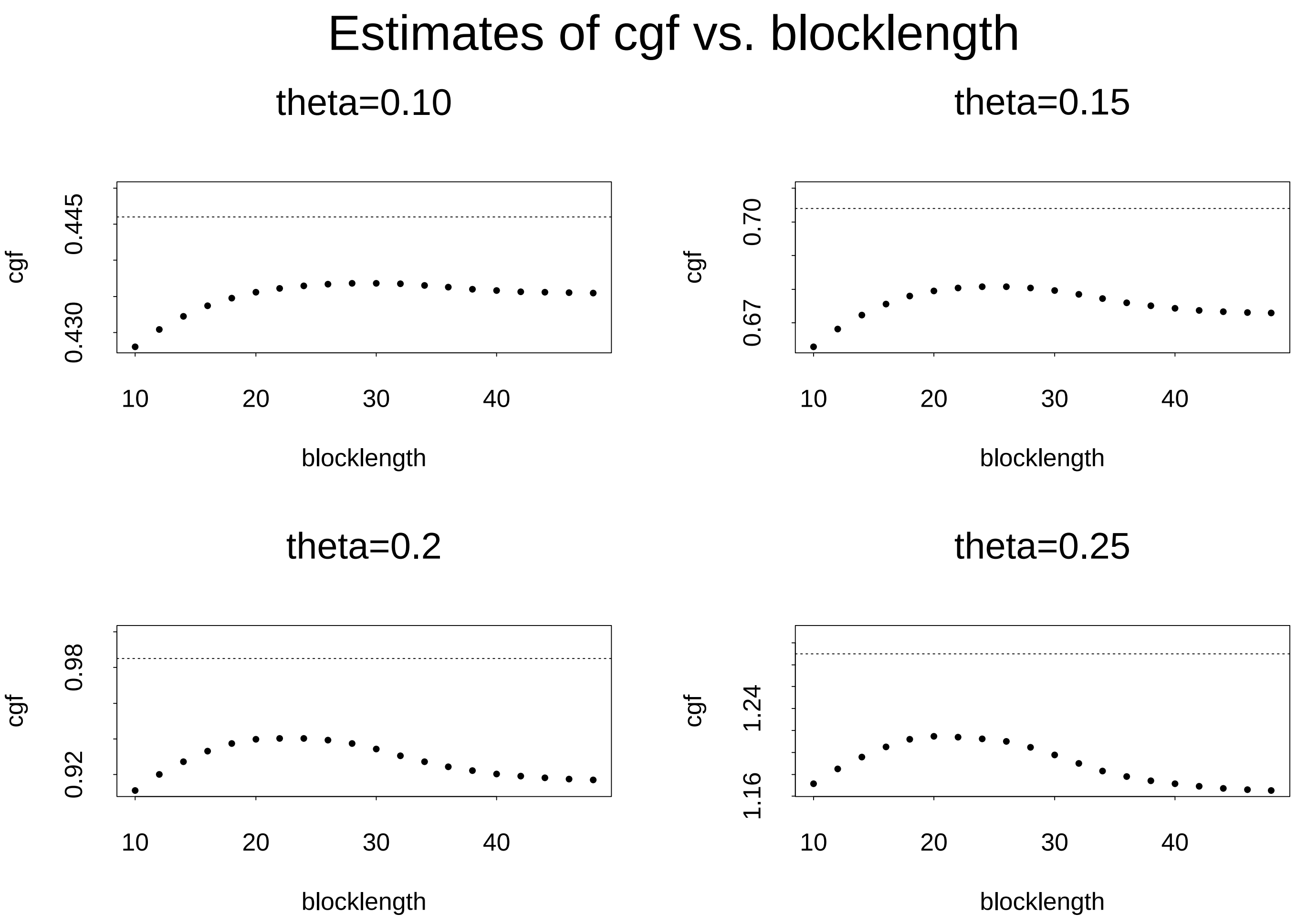

Figure 1: Cgf estimates for the autoregressive source

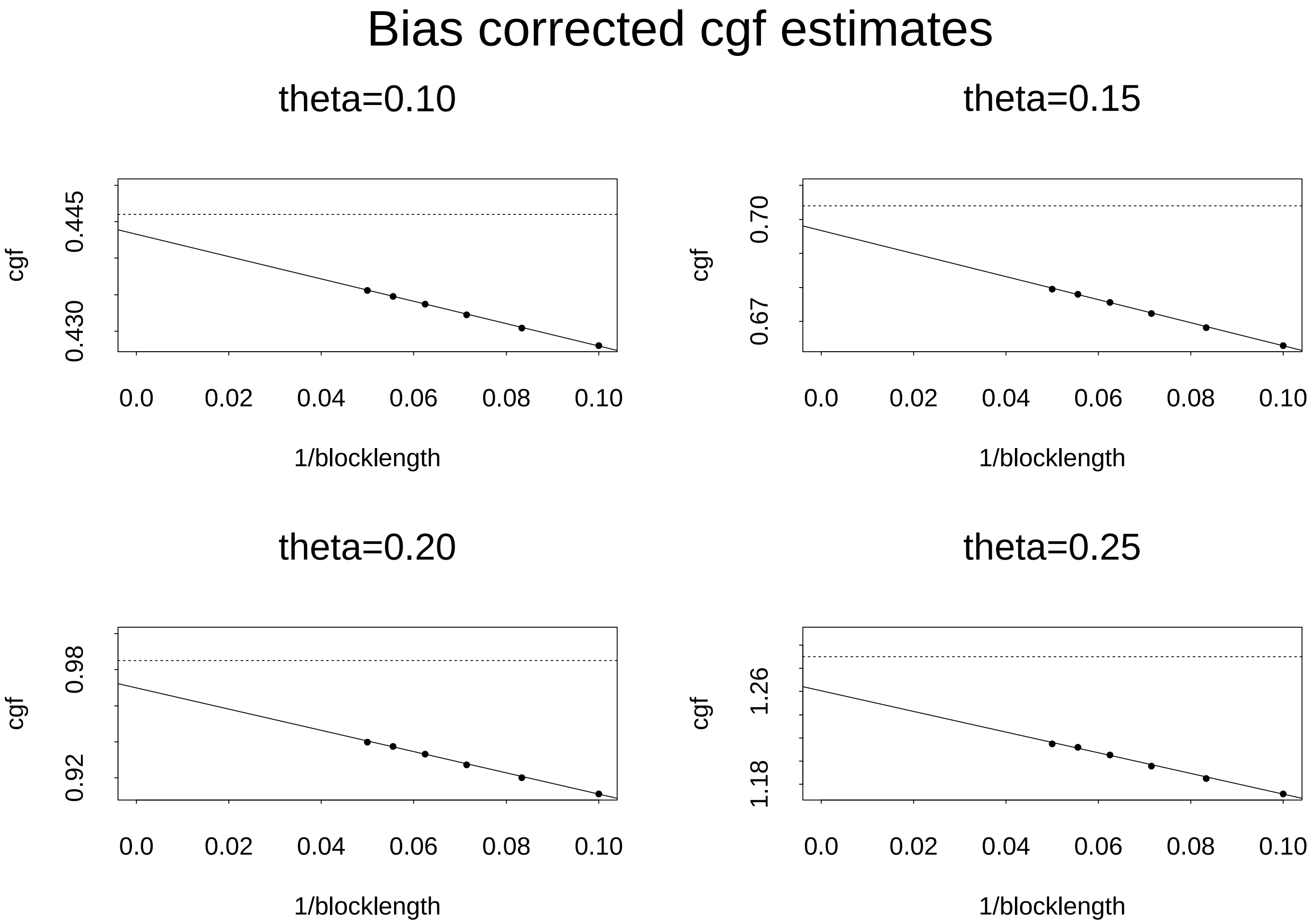

Figure 2: Bias correction for the autoregressive source

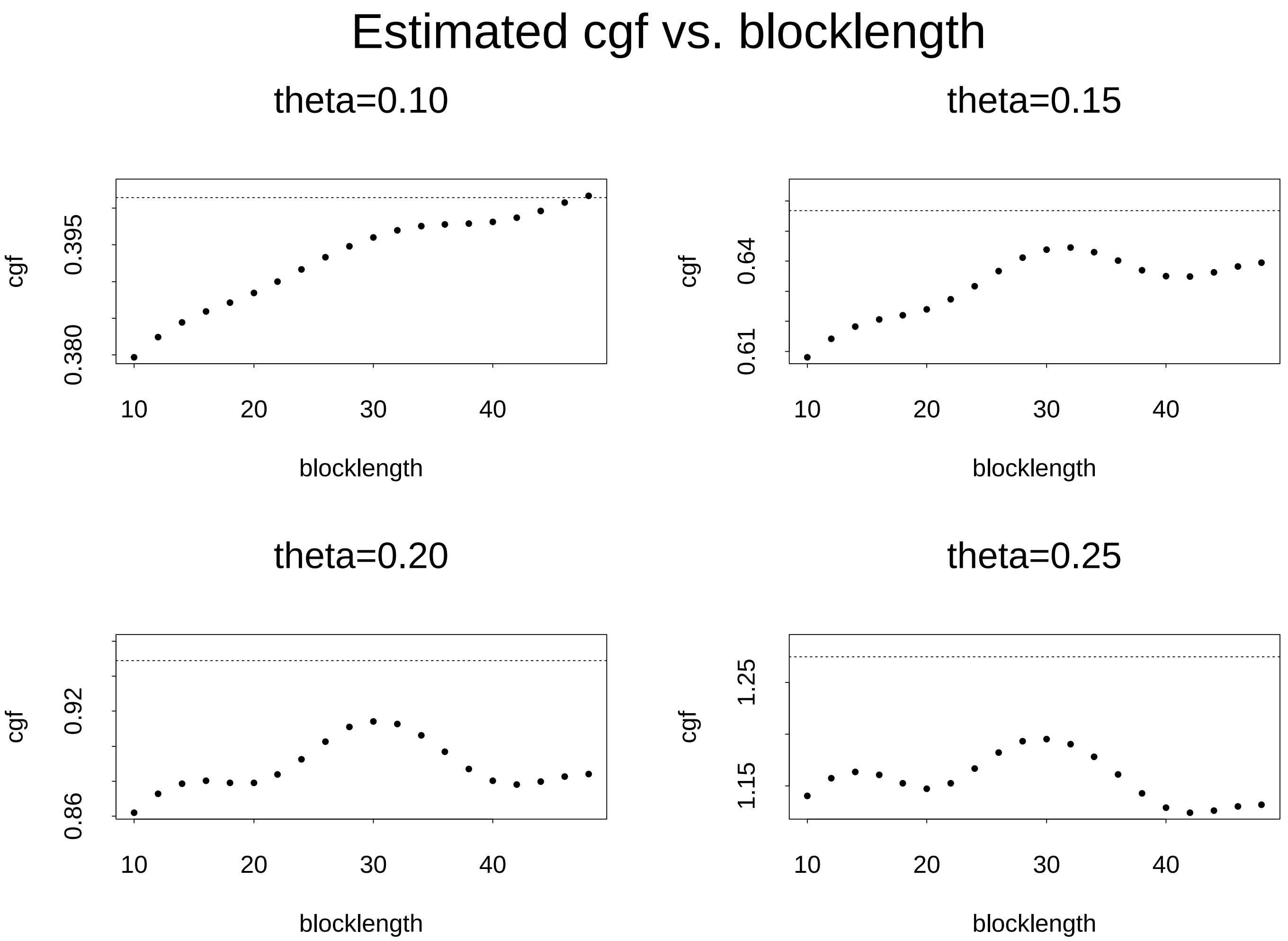

Figure 3: Cgf estimates for the Markov source

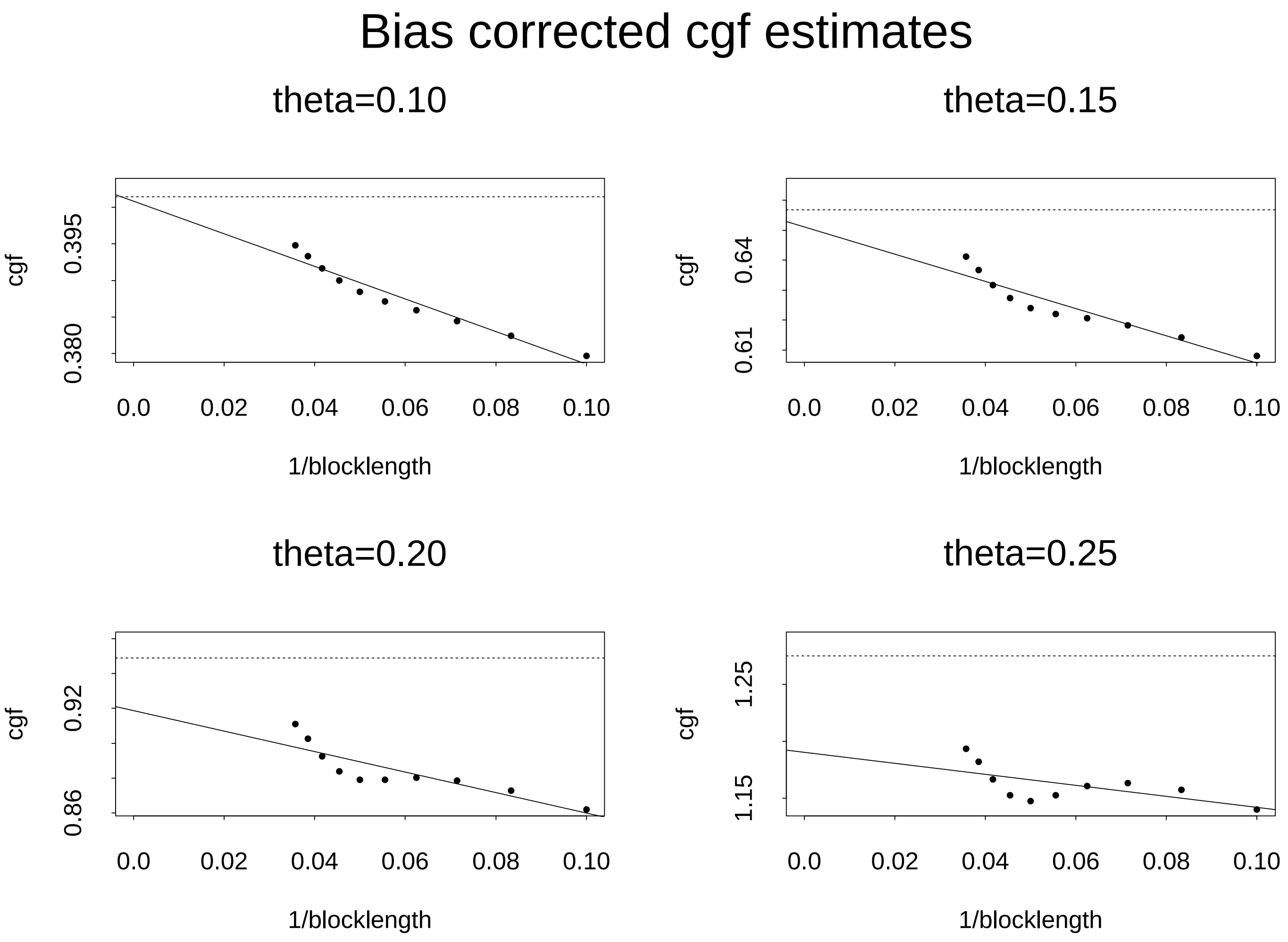

Figure 4: Bias correction for the Markov source