
ELSEVIER

Performance Evaluation 48 (2002) 157–175

PERFORMANCE EVALUATION
An International Journal

www.elsevier.com/locate/peva

Effective bandwidth estimation and testing for Markov sources

Juan Pechiar, Gonzalo Perera, María Simon*

Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay

Abstract

This work addresses the resource sharing problem in broadband communication networks that can guarantee some quality of service (QoS), and develops some results about data source and traffic modelling, especially in aspects of model testing and parameter estimation. The multiplexing of variable bit rate (VBR) sources poses a mathematical and statistical problem: the estimation of the resource requirements of a source or set of sources. The estimation method shall be simple enough to be practically implemented in the connection acceptance control (CAC) function.

In this paper, the VBR video sources are taken as a typical case of variable rate, with real-time constraints. This association of requirements makes the case especially interesting. A Markov model is assumed for the VBR sources. The validity of such models is under research; they seem to be appropriate at least in certain time scales. The model is tested against real video traces. In order to estimate the resource allocation or "channel occupation" of each source, the concept of equivalent bandwidth proposed by Kelly [Notes on effective bandwidth, in: F.P. Kelly, S. Zachary, I.B. Ziedins (Eds.), Stochastic Networks: Theory and Applications, Oxford University Press, Oxford, 1996, pp. 141] is used; it is based on a consistent mathematical theory, and has proven to be robust and useful for technical applications.

A calculation of the equivalent bandwidth of a Markov source, given its parameters, can be found in the literature [IEEE ACM Trans. Networking 1 (4) (1993) 424]. But in fact, one can only estimate model and parameters. In this work, an estimation of the equivalent bandwidth is given, which can be obtained from real data. The convergence and the consistency of the estimation are studied, and practical bounds are found. Illustrative calculations are performed from real video traces that were obtained using a software MPEG coder, developed by the authors. The mathematical and statistical results are valid for whatever phenomenon that can be modelled as a Markov process.

Keywords: Markov sources; Effective bandwidths; Infinitesimal generators; Central limit theorems; Traffic on broadband networks; Real time on ATM

1. Introduction

Telecommunications are evolving towards digital integrated broadband networks (BISDN). Integration means that a common network is able to transport many kinds of information: voice, video, data, all of them in digital form. Therefore analog signals are converted into sequences of numbers and transmitted as such. The term broadband refers to the huge amount of information to be transmitted. Ideally, a common network and a common connection can fulfil the communication needs of residential or corporate users, carrying voice, video and data.

* Corresponding author.

0166-5316/02/$ – see front matter
PII: S0166-5316(02)00035-4

"Digital integrated broadband network" is the definition of an abstract service or a bearer capability; its implementation remains so far undefined. The normalisation bodies, as the ITU (International Telecommunications Union), propose the asynchronous transfer mode (ATM) as the implementation of broadband networks.

Maybe the key difference between data networks philosophies is the possibility of offering some quality of service (QoS), in front of best effort policies. Best effort designed networks, like original TCP/IP, are well suited for users that cooperate, as is the case within a same organisation. Guaranteed QoS networks, like ATM or IPv6, are adequate for public services. They can also reliably bear services with real time, maximum delay or maximum error rate constraints, as voice or video.

In such QoS networks, one major issue is to optimise the resource usage, while maintaining the quality parameters of the communications. This technological trade-off is located in the connection acceptance control (CAC) function. When accepting a new connection, its occupancy should be estimated in order to know if, carrying this new traffic, the promised QoS of the new and of the already established connections will be supported.

The general statement of the problem is that a set of resources are to be shared by a set of heterogeneous communications. When accepting a new communication, the workload of this new communication and the available resources must be estimated; once the communication is established, some parameters stated in the QoS definition, as loss probability and delay must be honoured. That means that a description of the channel capacity occupied by a source is needed. Such a description cannot be complicated, because it shall be used to take real-time actions. Effective bandwidth, a concept proposed in [1] is a good measure of channel occupancy. Its application to communication networks is also studied in [3,4], among others.

Therefore, the technical problem that motivates this work is the communications multiplexing or resource sharing problem. Resource sharing is trivial when the sources generate constant bit rate (CBR). Digitised voice, when using PCM (pulse code modulation), is an example of such sources. A channel of output rate C can accept sources s_i that produce a data rate t_i provided that $\sum t_i < C$.

When data can tolerate some delay and delay variation (for instance, computer communication), the traffic can be shaped in order to fit the available resources. The problem is not so trivial, but known protocols can be used, that adapt the data delivery of the sources to the network. Those protocols require some dialog between network and source; they can be used when the service tolerates a delay (normally a time of a few seconds).

Real time and variable bit rate (VBR) services pose the most difficult problem. As the source is variable, statistical gain is to be expected. If the maximum rate would be allotted, the bandwidth would be normally wasted. But, when real-time constraints are present, traffic shaping is possible only in a moderate form; therefore an adaptive traffic shaping is not possible.

Different signals pose specific requirements on ATM networks regarding delay or errors, and are therefore carried using different bearer capabilities. One major challenge for future ATM networks is in fact the need of carrying interactive services. These have stringent real-time requirements: maximum end to end delays must be kept sufficiently small in order not to cause annoying effects. End to end synchronisation is needed; in order to implement it, the maximum delay variation (in ATM networks, cell delay variation tolerance) must also be kept small, in order to be removed by filtering.

Digitised video is one of the more stringent services because of the amount of transmitted data, the real-time requirements and the low error tolerance. Consistent quality video signal coders produce VBR output. For these reasons, video transmission is taken as case of study in this work.

Section 2 presents the ATM paradigm in those aspects that are relevant for this paper, specially the statistical bit rate (SBR) transfer capability and the conformance algorithm. The former subject appears because SBR capability is well suited for services with VBR output and real-time constraints. The conformance algorithm is addressed because it gives a coarse description of the traffic.

The traffic produced by a video source, in this case videotelephony or videoconference, depends on the signal itself (shapes, colour, movements), on the coding algorithm that digitises video, especially on its control algorithm, and on the delivery procedure, also called traffic shaping. Section 3 gives a general description of video compression MPEG standards, and justifies the relationship between constant or consistent video quality and variable output. It also presents the coder that was used, which was developed following the standards. Its bit generation control loop allows a VBR output that fits a given SBR contract. Section 4 presents the concept of effective bandwidth, its interpretation and its application in the CAC function. It also gives its expression for a Markovian source.

Section 5 is the central part of this work; it gives estimation results for calculating the effective bandwidth from real traffic traces. The convergence and consistency of the estimation is shown, finding a close expression for the confidence interval. The details of mathematical proofs are given in Appendix A. In Section 6, real traces are analysed and compared with models. The states of a Markov source were identified; as an example, a two-state source was assumed, giving an acceptable fit. The effective bandwidth is calculated in this example. Section 7 is a general conclusion and an overview of foreseen future work.

2. Asynchronous transfer mode

This section recalls some basic concepts of ATM, which is relevant for this work. In ATM networks, the digital information is split into units called cells. Its size is fixed (53 bytes, from which 48 bytes payload and five bytes header) for technological and theoretical reasons, and small in the sense that normal information is segmented (and of course reassembled) in a large number of cells. This means that the information transmission may be regarded as a fluid process: one cell is a little quantum. This fluid approach gives normally a good approximation for traffic analysis.

The delivery address of each cell is explicitly written in the cell's header. All cells of a communication follow the same path through the network in order to ensure sequential arrivals. This means that, when establishing a connection, the control mechanisms must estimate the probability of losses along the fixed path.

The ATM network is composed by a set of switches and channels. Each channel has a propagation delay. Each switch is a device whose function is receiving cells and delivering them in the corresponding channel. A cell requires some work from the switch; in fact the switch is a server and cells form a waiting queue. A reason for taking fixed size cells comes from queuing theory, because it shows that the mean waiting time is minimum when the attention time is deterministic.

Each cell goes through its path traversing a set of queues. A total end to end delay exists, due to waiting in queues and propagation time along the channels. This delay is variable, because the occupancy of the queues is variable. The place allotted for queues is limited; therefore if the total traffic exceeds some bounds cells are lost. Lost cells cause transmission errors. Total delay and delay variation pose a problem for real-time services. The QoS states the errors, delay and delay variation accepted levels.

2.1. Transfer capabilities

In accordance with the services' requirements, different transport schemes are provided by ATM networks. Those schemes correspond with contracts between user and network, CAC algorithms and control algorithms. CBR traffic is allotted in a deterministic bit rate (DBR) contract: only the peak rate is established. As was established before, the multiplexing and CAC for DBR channels is very simple. SBR is defined in the I.371 [5] ITU recommendation. The theoretical foundations of the transmission on a channel with these constraints can be found, in its fluid version, in [6], and the analysis of the multiplexing of such channels in [7].

This connection mode is devised to support VBR services with real-time constraints. The SBR capability guarantees a minimum data rate (sustainable cell rate, SCR) and also allows the emission at higher rates during limited periods. The mean is short-term enforced. The volume of bursts is limited by the maximum burst size (MBS), and during the busts the emission rate is limited by the peak cell rate (PCR). This makes SBR especially interesting for the transmission of interactive video services. Nevertheless, some shaping on the emitted traffic is still needed in order to comply with the SBR negotiated parameters.

Two main questions arise. First, is it possible to develop a simple control algorithm which will take full advantage of SBR? Second, which will be the performance of a network carrying this type of traffic? A third issue is to devise a CAC method for allotting SBR connections within the network resources.

The first question was studied in previous work [8]; its response is a simple control generation method that can be incorporated into the video coder and that adapts the bit generation to the SBR contract. This is explained in the next section.

In order to answer the second question, a source characterisation is necessary. Software coding was performed on real sequences in order to find a parametrical characterisation of the SBR-controlled video source. Some work about multiplexing was presented in [9], taking simple Markov models of the source. The present work goes further in the analysis of the validity of the general Markov models, and in parameter estimation from real data. Effective bandwidth estimation is useful for implementing the CAC function.

2.2. The conformance algorithm

The conformance algorithm is examined here because it gives some coarse description of the traffic. In ATM networks, it is implemented in that is called the usage parameter control (UPC) function.

In an SBR connection, the user is allowed to send some mean rate and limited busts. Therefore, in the input point of the network some control function is to be implemented. Its name is UPC, and its function is to perform the "conformance test". If a cell passes the test, it is called conformant, and should be transmitted with the negotiated QoS. If it is not the case, the cell may be discarded, or may be transmitted if there is free capacity, maybe with low priority. Non-conforming traffic should be avoided in video transmission. Indeed, if the network discards a single cell, the decoder will have to discard all data until a synchronisation mark is received. This is because data transmitted is compressed, being impossible to resynchronise at any point in the bit stream.

For the fluid model, conformance at the UPC is often verified through the "leaky bucket" algorithm [10], which is a simple control algorithm that ensures a minimum rate, and a limited burst size. The "bucket" is a transmission credit reservoir or pool. For each unit of data emitted, a credit unit has to be extracted from the pool: a credit is consumed. When there is no credits, no information may be transmitted.

The source can be thought as having a credit pool which is filled at rate SCR, has a maximum of B credits (the bucket capacity) and can be emptied at a maximum rate PCR. From this maximum B comes the name of "leaky bucket": the bucket is filled but, when a certain level is reached, additional credits flow away. This mechanism avoids a user of accumulating a big amount of credits, and then sending a huge burst. It ensures a short-term mean of SCR and limited bursts at PCR maximum rate. The MBS is $\mathrm{MBS} = B/(1 - \mathrm{SCR}/\mathrm{PCR})$, where B is expressed in cells, and SCR and PCR are normally expressed in cells per second.

These constraints give a very coarse description of the traffic, and some guidance for the model design. In fact, a finite discrete model of a continuous source is not easy to define and even to discern.

The simplest model of an SBR conformant source is a two-state model. In one state, the "normal" status, the source emits at SCR, while in the "burst" status it emits at PCR. In real traces, one can identify such states but there are also rapid superimposed fluctuations that do not alter the short-term mean rate. In the "normal" status, the short-term mean equal to SCR, with rapid superimposed variations. In the "burst" status, of limited intensity and duration, the bit rate is locally constant and equal to PCR. A third, "recovery" status can be incorporated in the model, because after a burst the mean rate is enforced by a less rate period. In this work, a simple two-state source was used. Identifying states is not evident; doing it in an automatic and consistent manner is devised as future work.

3. The video signal and the generated traffic

In this section, the basics of MPEG video compression (see [11–13]) are described, in order to show how the generated data are related to the video signal.

A video sequence is a series of frames, each being represented (in the digital domain) as a matrix of pixels. The transmission of all this data as such requires a huge bandwidth. Fortunately, such data present a lot of redundancy. Subsequent frames are normally very similar to each other, and so a frame can be very well predicted from past frames, or even interpolated from the past and the future frames. This *temporal redundancy* is therefore reduced by coding only what is not predicted. This prediction is further improved by sending information on local displacements between both images (*motion compensation*). Only when big image transients occur (e.g. a scene cut), prediction fails.

In this work, as interactive services are addressed, only forward prediction is used. Interpolated frames give a higher compression, but introduce an additional delay. Within a frame, there is also what is called *spatial redundancy*: a pixel's colour is normally similar to that of its neighbours. Therefore, the image is highly compressed by using spectral transforms (i.e. the *discrete cosine transform* (DCT)). The DCT coefficients are then quantised using some quantisation parameter or step *qp*. The lower the step *qp*, the higher the coding quality, and the lower the compression factor. Data for coefficients, quantisation and motion compensation are then statistically compressed (lossless compression) by using variable length coding (VLC).

In an open loop video coder, the amount of generated data per time unit, or per frame, is not constant, but depends on the scene complexity (details, colours) and activity (movements and difference with preceding scenes). If the signal quality is expected to be consistent, video coders are VBR sources.

Most frequently implemented coders produce CBR streams, because they are designed for transmitting over CBR channels. Anyway, transmission and reception buffers are needed, in order to absorb network jitter, to smooth the delivery, and to store the information unit to be displayed in time.

How can a CBR stream be produced from video? A control loop is established, observing the transmission buffer level and acting on the quantisation step *qp*. When the buffer is lowering, because generation is lower than the negotiated CBR, *qp* is done smaller; in this way video is coded with higher quality, and a higher bit amount is produced. When the opposite situation occurs, the buffer tends to be full and a higher *qp* is selected, thus the coding is coarser and the generation decreases.

When the network allows for VBR connections, one may be tempted of implementing a completely uncontrolled coder and of transmitting the output instantaneously. However, a controller is needed, even when transmitting on a VBR connection, because uncontrolled video traffic makes it impossible to establish a contract that guarantees some QoS.

A similar question arises. How to control the generation in order to fit the SBR contract, taking a maximum profit? Should the coder know the network–user transaction status (for instance, in a leaky bucket UPC, the bucket level) to implement a good controller? In previous work [8], this group have studied the coding process for real-time video services over SBR connections. It was shown that SBR conforming traffic allows for a better image quality than CBR at a given mean rate. It was also found that the essentials of CBR buffer control strategy could be used for SBR, without need of knowing the network status. The possibility of emitting bursts, and therefore of controlling more loosely than in a CBR environment, is taken into account by means of a virtual buffer, whose size is determined knowing the admitted delay and the UPC parameters (mean and peak rates, SBR and PCR, and the MBS).

The output of such coder is a stream with mean SBR, short-term enforced, peak rate PCR and established MBS. The mean time between bursts and its size is determined by the nature of the video signal: usually, a burst corresponds to a scene cut (camera change or new program) or to a rapid movement.

Modelisation of different digital sources, in particular video sources, is a wide topic. In [14], a simple Markov model is studied for a set of ON/OFF sources (see also [15]). The video source has also been studied: in [16], different time scales are taken into account, allowing thus the usage of Markov models in the analysis of some phenomena. Some researchers found auto-similarity in video traces, and argue that no Markov model can represent the phenomenon, as [17]. In this work, a Markov model is used, and a Markovianity test is performed in Section 6.

4. Effective bandwidths

Given any multiplexing system, where many data streams share a common output port, it is of major interest to know *how many resources* will a given data stream require from this system. Think of the multiplexor as a buffer with output capacity C, fed by many different sources.

Knowing the resources needed for each connection has a direct application to, i.e. call admission control (CAC) and billing. Indeed, a new connection can be accepted if enough resources are available, and the price for the connection should be somehow proportional to these resources.

There are two obvious limits for the resources needed: reserving the peak rate for each connection leads to deterministic multiplexing, with no possibility of buffer overflow, and very poor channel utilisation. On the other hand, reserving the mean rate for each connection will lead to a 100% link utilisation, and inevitable buffer overflow. So, given an expected QoS (probability of buffer overflow), the actual resources that should be reserved lie somewhere between the mean rate and the peak rate of the connection. These resources are generally referred to as the *effective bandwidth* of the traffic source.

The approaches to actually calculating effective bandwidths are mainly inspired by results from large deviations theory. These techniques lead to asymptotic results (infinite number of sources, or limit decay rate for overflow probabilities), an aspect which has led to criticism. Anyway, these methods do provide a formal approach for directly evaluating performance issues for systems fed with heterogeneous sources.

A good interpretation and collection of results on effective bandwidths is given in Kelly's *Notes on effective bandwidths* [1], where effective bandwidth is defined as follows.

Let $X[0, t]$ be a process with stationary increments, representing the amount of work arriving from a source in the time interval $[0, t]$. Then, the effective bandwidth of the source is

$$\alpha(s, t) := \frac{1}{st}\log \mathbb{E}[e^{sX[0,t]}], \quad 0 < s, t < \infty. \tag{1}$$

Let us explain in more detail why $\alpha(s, t)$ may be taken as a measure of the resources to be assigned to the given source. First, note that $\alpha(s, t)$ lies between the mean rate (for $s \to 0$) and the peak rate (for $s \to \infty$) of the input process: assume, for the sake of simplicity that for a given time interval $[0, t]$, $X[0, t]$ takes a finite number of values $x_1 < x_2 < \cdots < x_k$ with probabilities $p_1(t), p_2(t), \ldots, p_k(t)$; then

$$\alpha(s, t) = \frac{1}{st}\log\left(\sum_{i=1}^{k} e^{sx_i} p_i(t)\right),$$

as $s \to 0$, $\alpha(s, t)$ is equivalent to

$$\frac{1}{st}\log\left(\sum_{i=1}^{k}(1 + sx_i)p_i(t)\right) = \frac{1}{st}\log\left(\sum_{i=1}^{k} p_i(t) + s\sum_{i=1}^{k} x_i p_i(t)\right) = \frac{1}{st}\log\left(1 + s\sum_{i=1}^{k} x_i p_i(t)\right),$$

which is in turn equivalent to

$$\frac{1}{st}s\sum_{i=1}^{k} x_i p_i(t) = \frac{1}{t}E(X[0, t]).$$

On the other hand, as $s \to \infty$, $\alpha(s, t)$ is equivalent to

$$\frac{1}{st}\log(e^{sx_k} p_k(t)) = \frac{1}{st}sx_k + \frac{1}{st}\log(p_k(t)),$$

which tends to x_k, maximum possible value of $X[0, t]$; finally, it is clear that for any $s > 0$ we have that

$$\alpha(s, t) \leq \frac{1}{st}\log(e^{sx_k}) = \frac{x_k}{t},$$

and that (using Jensen's inequality)

$$\alpha(s, t) \geq \frac{1}{st}E[\log e^{sX[0,t]}] = \frac{1}{t}E(X[0, t]).$$

Another important property of this coefficient $\alpha(s, t)$ is that the effective bandwidth for the superposition of independent input processes is the sum of the individual effective bandwidths: if $X^1[0, t], \ldots, X^m[0, t]$

are m independent random processes corresponding to the workload required by m independent sources, and $X[0,t]$ stands for the workload of the multiplexed system

$$X[0,t] = \sum_{i=1}^{m} X^i[0,t],$$

then calling $\alpha(s,t)$ to the effective bandwidth of X and $\alpha(s,t)^i$ to the effective bandwidth of X^i, we have that

$$\alpha(s,t) = \frac{1}{st}\log \mathbb{E}\left[e^{s\sum_{i=1}^{m} X^i[0,t]}\right] = \frac{1}{st}\log \mathbb{E}\left[\prod_{i=1}^{m} e^{sX^i[0,t]}\right] = \frac{1}{st}\log\left(\prod_{i=1}^{m} \mathbb{E}[e^{sX^i[0,t]}]\right)$$
$$= \sum_{i=1}^{m} \frac{1}{st}\log(\mathbb{E}[e^{sX^i[0,t]}]) = \sum_{i=1}^{m} \alpha(s,t)^i.$$

Another important property of effective bandwidths comes from large deviation principle (see [18,21]): if, for instance, $X[0,t]$ has independent stationary increments, then, setting $\alpha(s,\infty) = \lim_{t\to\infty}\alpha(s,t)$, for any $\delta > 0$ one has that

$$\lim_{t\to\infty}\frac{1}{t}\log P\left(\left|\frac{X[0,t]}{t} - \frac{E(X[0,t])}{t}\right| \geq \delta\right) = -(\eta\delta + \Lambda(\eta))$$

with

$$\Lambda(s) = s\alpha(s,\infty),$$

where $\alpha(s,\infty) = \lim_{t\to\infty}\alpha(s,t)$ and η is the value of s which solves

$$\frac{d\Lambda(s)}{ds} = \delta,$$

which means that $\alpha(s,t)$ controls the probability of a large deviation of the mean workload from its expected value for large values of t.

Parameters s and t are referred to as the space and time parameters, respectively. When solving for a specific performance guarantee, these parameters depend not on the source itself, but on the *context* on which this source is acting. More specifically, s and t depend on the capacity, buffer size and scheduling policy of the multiplexor, the QoS to be achieved, and the actual traffic mix (i.e. characteristics and number of the other sources).

Indeed, the amount of resources needed by a given source does not depend on the source by itself, but on the whole system. Think, i.e. sources with mean rate 10 cells per second and peak rate 30 cells per second. If the link capacity is 10 000 cells per second, then each source will take about 1/1000 of the link's resources. But, if the link capacity is 50 cells per second, each source can perfectly take half of the link's resources (i.e. no more than two sources accepted) given some QoS objective.

One interesting result is the following: consider a buffer fed with independent sources of different types $\{1,\ldots,j\}$. The overflow probability is given by

$$\lim_{N\to\infty}\frac{1}{N}\log L(C_N, b_N, n_N) = \sup_t \inf_s \left[st\sum_j n_j\alpha_j(s,t) - s(b+Ct)\right], \tag{2}$$

where N is the total number of sources, n_j the fraction of sources of type j, $L(C, b, n)$ the proportion of lost workload for a link capacity C, buffer size b and traffic mix $n = (n_1, \ldots, n_J)$.

Let s is measured in data units (cells, bits). Intuitively, the value of s which solves the previous expression is related to the degree of statistical multiplexing: larger values of s indicate lower levels of statistical multiplexing, as when the peak rates of individual sources become comparable to the link capacity.

Let t is measured in time units, and indicates the most probable busy period over which congestion is built up. Large values of t indicate a slow buffer build-up, as happens for a highly utilised link fed with non-bursty sources. Small values of t correspond to fast buffer growth, indicating that congestion most probably occurs due to burst superposition (see [18] for an intuitive overview). In general, t indicates the time scales that should be considered when studying a system.

4.1. Results for Markov-modulated streams

Let us first recall some basic facts concerning continuous-time Markov chains, since we will deal in this paper with Markov sources.

If $X = (X_t)_{t\geq 0}$ is a continuous-time homogeneous Markov chain with finite state space $S = \{1, \ldots, k\}$, and transition matrix $\mathbb{P}(t)$, with $\mathbb{P}(t)_{ij} = P(X_{s+t} = j / X_s = i)$ for any $s > 0$, $t > 0$, its infinitesimal generator is the $k \times k$ matrix Q such that $\mathbb{P}(t) = \exp(tQ) = \sum_{n=0}^{\infty}(tQ)^n/n!$ (see, for instance [19, Chapter 2] and [22]). In what follows, we will always assume Q such that X is irreducible (what implies that if $C \subset S$ and for any $i \in C$, $j \notin C$ we have $Q_{ij} = 0$, then $C = S$) and recurrent and therefore, we will deal with ergodic Markov chains. Note that this means that there exists a probability vector $\vec{\pi}$ that is invariant under $\mathbb{P}(t)$, i.e.

$$\vec{\pi}\,\mathbb{P}(t) = \vec{\pi}, \quad \text{for any } t > 0.$$

Let us present here the computation method for π that we will use here, its detailed proof is given in Lemma A.2 of Appendix A. By differentiating the last equation with respect to t at $t = 0$, we deduce that

$$\vec{\pi}\,Q = \vec{0}, \tag{3}$$

where $\vec{0} = (0, 0, \ldots, 0)$. Taking into account that

$$\langle \vec{\pi}, \vec{1} \rangle = 1 \tag{4}$$

(where $\vec{1} = (1, 1, \ldots, 1)$ and $\langle , \rangle$ stands for the scalar product $\langle x, y \rangle = \sum_{i=1}^{k} x_i y_i$), and defining $\hat{Q}_{ij} = Q_{ij}$ if $j < k$, $\hat{Q}_{ik} = 1$, we can summarise (3) and (4) in the equation

$$\vec{\pi}\,\hat{Q} = e_k, \tag{5}$$

where $e_k = (0, 0, \ldots, 0, 1)$, the element of $\mathbb{R}^k$ with all its coordinates null with the exception of the last, that equals one. It can be shown that $\hat{Q}$ is non-singular, with inverse matrix $\hat{Q}^{-1}$ (see Lemma A.1 in Appendix A), and therefore we can explicitly compute the invariant probability $\vec{\pi}$ by

$$\vec{\pi} = e_k \hat{Q}^{-1}. \tag{6}$$

Consider a source whose rate depends on the state of a finite state continuous-time Markov chain: when the chain is in state i, workload is produced at constant rate h_i. Let $H = \mathrm{diag}(h_i)$, the matrix with elements h_i in the diagonal.

Then, for such a source, the effective bandwidth is [1,2]

$$\alpha(s,t) = \frac{1}{st}\log\{\vec{\pi}\exp[(Q+Hs)t]\vec{1}\}.$$

In the next section, we find a consistent estimator for $\alpha(t,s)$ given traffic traces, and we also give the corresponding confidence intervals. Let us present here a result, due to Lebedev and Lukashuk [20], that plays an essential role in this construction, since it provides asymptotically Gaussian estimates of the infinitesimal generator Q, based on certain functionals of traffic traces.

Theorem 1 (Lebedev and Lukashuk [20]). *Let $(X_u)_{u\in\mathbb{R}^+}$ be a time continuous irreducible homogeneous Markov chain with finite state space $S = \{1,2,\ldots,k\}$ and unknown infinitesimal generator matrix $Q = (\lambda_{ij})_{1\le i,j\le k}$. Denote $D = \{(i,j) \in S\times S\colon \lambda_{ij} > 0\}$ and consider the maximum likelihood estimators of λ_{ij}, given by*

$$\hat{\lambda}^n_{ij}(x) = \frac{\nu(i,j,nx)}{\tau(i,nx)}, \tag{7}$$

where

$$\nu(i,j,h) = \textit{number of transitions of } X \textit{ from } i \textit{ to } j \textit{ in } [0,h], \tag{8}$$

$$\tau(i,h) = \textit{time spent by } X \textit{ in state } i \textit{ during the time interval } [0,h], \quad h > 0. \tag{9}$$

Then

(a) $\lim_n |\hat{\lambda}^n_{ij}(u) - \lambda_{ij}| = 0$ *a.s.*

(b) $(\sqrt{nu}(\hat{\lambda}^n_{ij}(u) - \lambda_{ij}))_{i,j\in D} \Rightarrow^w_n ((\sqrt{\lambda_{ij}/\vec{\pi}(i)})W_{ij}(u))_{i,j\in D}$ *as a stochastic process in $u \in [0,1]$ where $W = (W_{ij})_{i,j\in D}$ denotes a standard Wiener process.*

5. Estimation

Consider X as in the statement of Theorem 1. Recall that for such X, we can compute its effective bandwidth by Kesidis–Walrand–Chang formula. For $s,t > 0$,

$$\alpha(s,t) = \frac{1}{st}\log\{\vec{\pi}\exp[(Q(\Lambda)+Hs)t]\vec{1}\}, \tag{10}$$

where $\vec{1}$ stands for the element of $\mathbb{R}^k$ with all its coordinates equal one. Fix $s,t > 0$ and denote by Λ the $k(k-1)$-dimensional vector that contains the non-diagonal elements of Q, that is

$$\Lambda = (\lambda_{ij})_{1\le i\ne j\le k}. \tag{11}$$

Since

$$\sum_{j=1}^{k}\lambda_{ij} = 0, \quad i = 1,\ldots,k, \quad \lambda_{ii} < 0, \quad \lambda_{ij} \ge 0 \quad \text{for } j \ne i. \tag{12}$$

We have that $\Lambda \in (\mathbb{R}^+)^{k(k-1)}$ and that

$$Q = Q(\Lambda), \quad \text{more precisely} \quad Q_{ij} = \Lambda_{ij}, \quad \text{if } i \neq j, \qquad Q_{ii} = -\sum_{j \neq i} \Lambda_{ij}, \quad \text{if } i = 1, \ldots, k. \tag{13}$$

Let us also consider the $(k \times k)$-dimensional matrix $\hat{Q}$ defined before by

$$\hat{Q}_{ij} = \lambda_{ij}, \quad \text{if } j < k, \qquad \hat{Q}_{ik} = 1. \tag{14}$$

It is clear from (12) that Q, Λ and $\hat{Q}$ contain exactly the same information; if we are given one of them, we easily obtain the other two. Therefore, we will often think of any parameter depending on the infinitesimal generator Q as a function of Λ. In particular, taking into account that $\vec{\pi}$ is the stationary distribution associated to Q,

$$\vec{\pi} = \vec{\pi}(\Lambda). \tag{15}$$

Coming back to effective bandwidths, if we define the following functions:

$$B(\Lambda) = \exp[(Q(\Lambda) + Hs)t], \qquad g(\Lambda) = \vec{\pi}(\Lambda) B(\Lambda) \vec{1}, \qquad \psi(\Lambda) = \frac{1}{st} \log(g(\Lambda)) \tag{16}$$

with

$$B : \mathbb{R}^{k(k-1)} \to M_{k \times k}, \qquad g : \mathbb{R}^{k(k-1)} \to \mathbb{R}, \qquad \psi : \mathbb{R}^{k(k-1)} \to \mathbb{R},$$

and then (10) becomes

$$\alpha(s, t) = \psi(\Lambda). \tag{17}$$

The main result of this paper is the following.

Theorem 2. *Let $(X_u)_{u \in \mathbb{R}^+}$ be a time continuous irreducible homogeneous Markov chain with finite phase space $S = \{1, 2, \ldots, k\}$ and unknown infinitesimal generator matrix $Q = (\lambda_{ij})_{1 \leq i, j \leq k}$ and for fixed $s, t > 0$. Consider $\alpha(s, t)$ as in* (10), *ψ as in* (16) *and $\hat{\lambda}_{ij}^n(u)$ as in* (7) *and set $\Lambda^n(u) = (\hat{\lambda}_{ij}^n(u))_{1 \leq i \neq j \leq k}$ then, taking*

$$\alpha^n(s, t)(u) = \psi(\Lambda^n(u)), \tag{18}$$

we have

(a) $\sqrt{nu}(\alpha^n(s, t)(u) - \alpha(s, t)) \Rightarrow_n^w N(0, \sigma^2)$, *with* $\sigma^2 = \nabla\psi(z) \Sigma(u) \nabla\psi(z)^{\mathrm{T}}$, *where $\Sigma(u)$ is the covariance matrix of $((\sqrt{\lambda_{ij}/\vec{\pi}(i)}) W_{ij})_{(i,j) \in D}$, and $(W_{ij})_{(i,j) \in D}$ are independent Wiener processes.*

(b)
$$\sigma^2 = \sum_{(i,j) \in D} \frac{\lambda_{ij}}{\vec{\pi}(i)} \frac{\partial \psi}{\partial \lambda_{ij}}(\Lambda)^2 = \frac{1}{(st\, \vec{\pi}(\Lambda) B(\Lambda) \vec{1})^2}$$
$$\times \sum_{(i,j) \in D} \left\{ \frac{\partial \vec{\pi}(\Lambda)}{\partial \lambda_{ij}} B(\Lambda) \vec{1} + \vec{\pi}(\Lambda) \sum_{n=0}^{\infty} \sum_{r=0}^{n-1} t^{n-1} (Q(\Lambda) + Hs)^r V^{ij} (Q(\Lambda) + Hs)^{n-1-r} \right\}^2,$$

where V^{ij} is the matrix in $M_{k \times k}$ defined by $V_{lm}^{ij} = \partial \hat{Q}(\Lambda)_{lm} / \partial \lambda_{ij} = 0$ if $l \neq i$, or $m = k$; 1 if $l = i$, $m = j \neq i$; -1 if $l = m = i$.

Proof. It is essentially an application of Taylor's formula. More precisely, apply Theorem A.1 of Appendix A for ψ as in (16) and then apply Lemma A.2 (see Appendix A). Finally, use Lemma A.3 of Appendix A for the computation of $\partial\psi/\partial\lambda_{ij}$. □

Following corollary shows how to perform practical numerical computations using this result.

Corollary 1. *Let m_n be a sequence of positive integers such that* $\lim_n m_n = \infty$, *but* $m_n = o(\sqrt{n})$. *Define*

$$\sigma_n^2 = \frac{1}{S_n} \sum_{(i,j)\in D} \frac{\lambda_{ij}^n(u)}{p_n(u)(i)} \times \left(\mathrm{d}p_n^{ij}(u) B_n(u)\vec{1} + p_n(u) \sum_{l=0}^{m_n} \sum_{r=0}^{l-1} \frac{t^{l-1}(Q_n(u)+Hs)^r V^{ij}(Q_n(u)+Hs)^{l-1-r}}{l!}\vec{1} \right)^2 \tag{19}$$

with $S_n = (stp_n(u)B_n(u)\vec{1})^2$, *where* $p_n(u)$, $\mathrm{d}p_n(u)$ *are as in* Remark A.2 *and*

$$B_n(u) = \sum_{l=0}^{m_n} \frac{t^l(Q_n(u)+Hs)^l}{l!}.$$

Then

(a) $\sigma_n(u)^2$ *is a consistent estimator of* σ^2.
(b) $\sqrt{nu}(\alpha^n(s,t)(u) - \alpha(s,t))/\sigma_n(u) \Rightarrow_n^w N(0,1)$.
(c) (*Confidence intervals for effective bandwidths*) *If*

$$I_n(u) = \left[\alpha^n(s,t)(u) - z_\epsilon \frac{\sigma_n(u)}{\sqrt{nu}},\ \alpha^n(s,t)(u) + z_\epsilon \frac{\sigma_n(u)}{\sqrt{nu}} \right], \tag{20}$$

where z_ϵ *is such that*

$$P(N > z_\epsilon) = \tfrac{1}{2}\epsilon \quad \textit{for}\ \ N \sim N(0,1),$$

and then

$$\lim_n P(\alpha(s,t) \in I_n(u)) = 1 - \epsilon.$$

Proof. To prove (a), one has to compute a little bit, but the main argument is that differentiability and Theorem 1 implies that replacing Λ by its estimation induces an error that in L^2 is $O(1/\sqrt{n})$; since m_n is $o(\sqrt{n})$, this replacement induces a negligible error, while the fact that m_n goes to infinity ensures that the sum converges to the series. Here (b) and (c) are straightforward consequences of (a) and Theorem 2. □

6. Data analysis

We now illustrate how the above results can be applied to measured traffic. The traffic trace corresponds to real-time VBR videoconference communication and was obtained as follows: first, a typical video-conference clip was recorded at our studio and sampled into digital media. The sequence was then coded

with a software MPEG coder. The coder implements the policies and algorithms described in [8,9], and so as to produce a traffic trace compliant with an ATM SCR connection.

The sequence is 2200 frames long, at a rate of 25 frames per second. Taking the frame interval as the time unit, parameters u and n in expressions (18) and beyond become 1 and 2200, respectively.

Examining the trace, two states can be clearly identified. Normally, the coder produces data at a rate just below the forced mean (SCR). We call this the normal state, and in our case corresponds to 10 kbit per frame. Transients in the scene, such as movements or scene cuts, produce short bursts in the output traffic. The rate generated during the bursts is around 30 kbit per frame in our case, and depends more on the coder than on the PCR assigned to the link. Indeed, when such a burst occurs, the coder is trying to reconstruct all or part of the image, and the data rate produced is limited by what MPEG can do in this case (generally at a high *qp*: low quality). So, the trace is segmented into a two-state process. The two rates identified determine the values for matrix H.

We apply the Markov tests described in [21] and verify that our two-state trace fits into a Markovian model. The actual result of this type of test is the probability for the sample not being Markovian.

Analysis of the state transitions, the estimator (7) for $\hat{\lambda}$ results in the following generator:

$$Q_n = \begin{bmatrix} -0.024 & 0.024 \\ 0.076 & -0.076 \end{bmatrix}, \qquad \hat{Q}_n = \begin{bmatrix} -0.024 & 1 \\ 0.076 & 1 \end{bmatrix}.$$

Now that we have a Markovian model for our source.

In order to obtain 90% confidence intervals ($\epsilon = 0.1$ in expression (20)), z_ϵ should be set to 1.645. For each value of s and t, expression (18) gives an estimate for the effective bandwidth $\alpha(s, t)$ and expression (19) gives the deviation σ_n. In our case, α and the confidence interval ($\sigma/\sqrt{nu}$) have the values shown in Figs. 1 and 2 respectively. Just as an example, we take $s = 0.5\,\text{kbit}^{-1}$ and $t = 1$ frame.

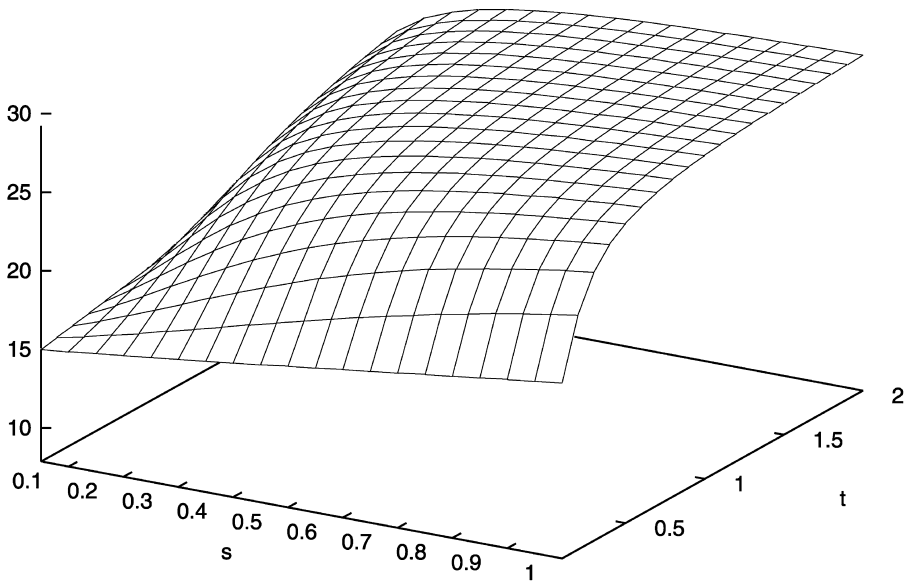

Fig. 1. Effective bandwidth, α.

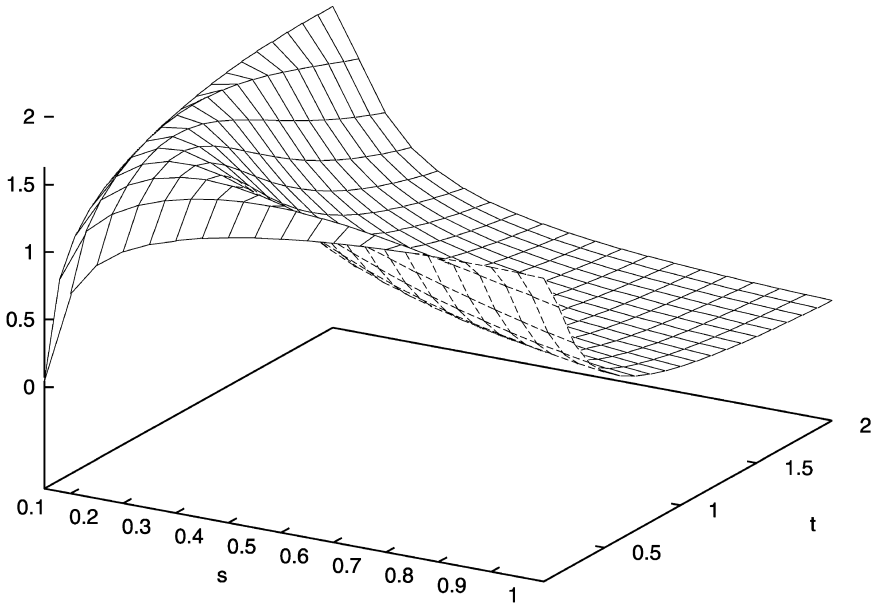

Fig. 2. Confidence interval.

The estimation for $\alpha(s,t)$ gives 27.0 kbit per frame, and σ_n^2 gives 284. This means that our 90% confidence interval is 27.0 ± 0.5. Our results allow for different time resolutions by altering parameter u. Also traces of increasing length can be taken in order to attain a desired confidence interval. In this case, sequence m_n in (19) can be taken as, i.e. $n^{1/4}$, which is what we used in the above example.

7. Conclusion

This work extends the results for the effective bandwidth for Markov-driven sources to a statistical context. A consistent estimator for the effective bandwidth and its confidence interval have been presented. These results enable the calculation of the effective bandwidth from simulated or traffic traces. Those calculations may be performed during communications.

A numerical example has been presented, where the results were applied to traffic from videoconference sessions. The traffic traces were reduced to a simple two-state Markov process from where the effective bandwidth was calculated. In a wider context, some of our intermediate results can be applied to other phenomena which can be modelled by means of Markov chains. It is expected to extend statistical effective bandwidth calculations to other stochastic phenomena, which do not need to be Markovian.

Acknowledgements

The authors express their gratitude to Prof. Carlos Falcón for his highly valuable suggestion. This work was supported by CIC, Facultad de Ingeniería, Project "Modelización matemática en telecomunicaciones".

Appendix A. Mathematical details

Some remarks on some basic elements of calculus and notation we shall use:

- For computation of limits, we will use the symbols "o" and "O" as usual; for instance o(1) denotes convergence to zero and O(1) corresponds to a bounded term.
- In $\mathbb{R}^d$ spaces, we will use Euclidean norms; in matrix spaces $M_{d_1\times d_2} = \{A : A \text{ is a } d_1 \times d_2 \text{ real matrix}\}$ we shall use operator norms, i.e. $\|A\| = \sup\|Ax\| : x \in \|x\| = 1$, which enjoys the property $\|AB\| \leq \|A\|\|B\|$ for $A \in M_{d_1\times d_2}$, $B \in M_{d_1\times d_2}$. In what follows spaces denoted by E, E_1, E_2, ... will always be Euclidean spaces or matrix spaces equipped with the operator norm.
- We will say that a function $G : E_1 \to E_2$ is differentiable at x and we will denote by $\mathrm{D}G(x)$ its differential at x if $\mathrm{D}G(x)$ is the (unique) linear transformation from E_1 to E_2 satisfying $G(y)-G(x) = \mathrm{D}G(x)(y-x) + \mathrm{o}(\|y-x\|)$ for any y in a neighbourhood of x (taking into account that a linear transformation from E_1 to E_2 corresponds to the multiplication by its associated matrix, we will indistinctly describe a linear transformation as a map or by giving its associated matrix). Chain rule applies: $\mathrm{D}(F(G))(x) = \mathrm{D}F(G(x)(\mathrm{D}G(x)))$ and most of the algebraic rules of differential calculus for real numbers are valid in this more general framework.
- Nevertheless, some significant differences appear; for instance, matrix product is not commutative, and if we consider for any $n \in \mathbb{N} G_n : M_{d\times d} \to M_{d\times d}$ defined by $G_n(A) = A^n$, then $\mathrm{D}G_n(A)(B)$ it is not $nA^{n-1}B$ as in the real case, but

$$\mathrm{D}G_n(A)(B) = \sum_{i=0}^{n-1} A^i B A^{n-1-i}, \tag{A.1}$$

as the reader can easily check from the definition. Let us also remark a consequence of this fact that will be used in the sequel: define the exponential matrix function $\exp : M_{d\times d} \to M_{d\times d}$ by $\exp(A) = \sum_{n=0}^{\infty} A^n/n!$, then

$$\mathrm{D}\exp(A)(B) = \sum_{n=0}^{\infty} \frac{\sum_{i=0}^{n-1} A^i B A^{n-1-i}}{n!} \tag{A.2}$$

(is easy to see that while $\|\mathrm{D}\exp(A)(B)\| \leq e^{\|A\|}\|B\|$, if A and B do not commute, i.e. if $AB \neq BA$, then $\mathrm{D}\exp(A)B$ may not be $\exp(A)B$ as in the real case).
- Another elementary fact we will use: if F, G are matrix-valued functions, then $\mathrm{D}(FG) = \mathrm{D}(F)G + F\,\mathrm{D}(G)$.
- Finally, when $G : \mathbb{R}^d \to \mathbb{R}$, then $\mathrm{D}G(x)(y) = \nabla G(x)y$, where ∇G denotes the gradient row vector $\nabla G(x) = ((\partial G/\partial x_1)(x), \ldots, (\partial G/\partial x_d)(x))$.

The following result is an elementary one that can be easily proved using differentiability definition.

Theorem A.1. *Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}^d$-valued random variables, $d \geq 1$, $(a_n)_{n\in\mathbb{N}}$ a sequence of positive numbers that increases to infinity and $z \in \mathbb{R}^d$ such that*

$$a_n(Z_n - z) \underset{n}{\overset{w}{\Rightarrow}} N(\vec{0}, \Sigma)$$

with $\vec{0} = (0, 0, \ldots, 0)$ *and* Σ *a covariance matrix, and consider* $G : \mathbb{R}^d \to \mathbb{R}$ *differentiable on a neighbourhood of* z; *then*

$$a_n(G(Z_n) - G(z)) \underset{n}{\overset{w}{\Rightarrow}} N(\vec{0}, \nabla G(z)\Sigma\nabla G(z)^{\mathrm{T}}).$$

We will present now the technical lemmas required for the proof and applications of Theorem 2.

Lemma A.1. *With the notation stated before,* $\hat{Q}$ *is non-singular with inverse matrix* $(\hat{Q})^{-1}$, *which is differentiable with*

$$\mathrm{D}\hat{Q}^{-1}(\Lambda)(x) = -\hat{Q}^{-1}(\Lambda)(\mathrm{D}Q(\Lambda))\hat{Q}^{-1}(\Lambda)x. \tag{A.3}$$

Proof. Assume $\hat{Q}$ singular. Then there exist $\mu_1, \ldots, \mu_k$ such that $\sum_{i=1}^{k} \mu_i^2 > 0$ and $\sum_{j=1}^{k-1} \mu_j \lambda_{ij} = -\mu_k$ for $i = 1, \ldots, k$. Since $\lambda_{ii} > 0$ for any i it follows that

$$-\mu_i + \sum_{j\neq i,\, 1\leq j\leq k} \mu_j \left(-\frac{\lambda_{ij}}{\lambda_{ii}}\right) = \frac{\mu_k}{\lambda_{ii}}, \quad i = 1, \ldots, k-1. \tag{A.4}$$

Without any loss of generality, we may assume that $\mu_k \geq 0$ what implies that the right-hand side of (A.4) is non-negative for any i. We then have

$$-\mu_i + \sum_{j\neq i,\, 1\leq j\leq k} \mu_j \left(-\frac{\lambda_{ij}}{\lambda_{ii}}\right) \geq 0 \quad \text{for any } \ i = 1, \ldots, k-1. \tag{A.5}$$

Since left-hand side of inequality (A.4) is zero for $i = k$, we have

$$-\mu_i + \sum_{j\neq i,\, 1\leq j\leq k} \mu_j \left(-\frac{\lambda_{ij}}{\lambda_{ii}}\right) \geq 0 \quad \text{for any } \ i = 1, \ldots, k. \tag{A.6}$$

Define, for $i = 1, \ldots, k$,

$$b_i = \max_{j\neq i} \mu_j, \qquad a_i = \min_{j\neq i} \mu_j. \tag{A.7}$$

Since $-\lambda_{ij}/\lambda_{ii} \geq 0$ for $j \neq i$, and $\sum_{j\neq i,\, 1\leq j\leq k}(-\lambda_{ij}/\lambda_{ii}) = 1$, we have that for any i, $c_i := \sum_{j\neq i,\, 1\leq j\leq k} \mu_j (-\lambda_{ij}/\lambda_{ii})$ belongs to the interval $[a_i, b_i]$ and, furthermore, if $c_i = a_i$ then $C_i := \{j \neq i\colon \lambda_{ij} \neq 0\} = \{j \neq i\colon \mu_j = a_i\}$ and if $c_i = b_i$ then $C_i = \{j \neq i\colon \mu_j = b_i\}$.

Let us now observe that

$$-\mu_i + b_i \geq -\mu_i + c_i \geq 0, \quad i = 1, \ldots, k, \tag{A.8}$$

what implies that $\mu_i \leq b_i$, $i = 1, \ldots, k$; hence, for any i there exists $j(i) \neq i$ such that $\mu_i \leq \mu_{j(i)}$.

Define now

$$\mu = \max_{1\leq j\leq n} \mu_j.$$

We will then have that there exist $r \geq 2$ and $i_1 < \cdots < i_r$ such that $\mu_{i_1} = \cdots = \mu_{i_r} = \mu$ what implies in turn that $b_{i_1} = \cdots = b_{i_r}$ and therefore $\mu_{i_q} = b_{i_q}$ for $q = 1, \ldots, r$. Then, for $q = 1, \ldots, r$,

$C_{i_q} = \{j \neq i_q : \mu_j = \mu\} = \{i_1, \ldots i_r\}$. Therefore, for $q = 1, \ldots r$, $\lambda_{i_q j} = 0$ if $i \notin \{i_1, \ldots, i_r\}$. This means that

$$C = \{i_1, \ldots i_r\}$$

is a closed set of states. Since our chain is irreducible, $C = \{1, \ldots, k\}$ and we then have $\mu_j = \mu$ for $j = 1, \ldots, k$. It follows that $\mu > 0$ and that

$$\sum_{j=1}^{k-1} \mu\lambda_{ij} = -\mu, \quad i = 1, \ldots, k.$$

Since $\sum_{j=1}^{k-1} \lambda_{ij} = -\lambda_{i(k-1)}$, we get $\mu\lambda_{i(k-1)} = -\mu$, $i = 1, \ldots, k$, and, taking into account that $\mu > 0$ we finally obtain

$$\lambda_{i(k-1)} = -1, \quad i = 1, \ldots, k.$$

Since $\lambda_{i(k-1)} \geq 0$ for $i = 1, \ldots, k-1$, we arrive to a contradiction, what shows that $\hat{Q}$ is non-singular.

For the second part, just note that

$$\hat{Q}(\Lambda)\hat{Q}^{(-1)}(\Lambda) = \mathbb{I}_{k\times k},$$

and differentiate to get

$$\mathrm{D}\hat{Q}(\Lambda)(\hat{Q}(\Lambda)^{(-1)}) + \hat{Q}(\Lambda)(\mathrm{D}\hat{Q}(\Lambda)^{(-1)}) = 0$$

(where "0" stands here for the null function) and the result follows. □

Remark A.1. By Lemma A.1, since $\hat{Q}$ is a continuous function of Λ and the determinant defines a continuous function on matrix spaces, by Lebedev–Lukashuk theorem, $\hat{Q}_n(u) = \hat{Q}(\Lambda_n(u))$ is non-singular for n large enough. Taking into account that inversion is also a continuous application on the space of non-singular matrices, we finally get that

$$\mathrm{D}_n^{ij}(u) = -(\hat{Q}_n(u))^{-1} V^{ij} (\hat{Q}_n(u))^{-1} \tag{A.9}$$

is a consistent estimator of $\partial\hat{Q}^{-1}(\Lambda)/\partial\lambda_{ij}$.

Lemma A.2. *With the same notation as before, $\vec{\pi}$ is a differentiable function of Λ that can be computed as $\vec{\pi}(\Lambda) = e_k^{\mathrm{T}}\hat{Q}^{-1}(\Lambda)$ and $(\partial\vec{\pi}/\partial\lambda_{ij})(\Lambda) = e_k^{\mathrm{T}}(\partial/\partial\lambda_{ij})\hat{Q}^{-1}(\Lambda)$, where e_k stands for the kth canonical (column) vector of $\mathbb{R}^k$, i.e. $e_k^{\mathrm{T}} = (0, 0, \ldots, 0, 1)$.*

Proof. First recall that, by definition of infinitesimal generator and invariant distribution we have, for any $t > 0$,

$$\vec{\pi}\exp(tQ) = \vec{\pi}. \tag{A.10}$$

Since for any $s, t > 0$, the matrices sQ and tQ commute, differentiating (A.10) and taking $t = 0$ we get

$$\vec{\pi}Q = \vec{0}, \tag{A.11}$$

where $\vec{0}$ stands for the null vector.

Taking into account that $\sum_{j=1}^{k} \lambda_{ij} = 0$ for $i = 1, \ldots, k$ and that $\sum_{i=1}^{k} \vec{\pi}(i) = 1$ we deduce from (A.11) that

$$\vec{\pi}\,\hat{Q} = e_k^{\mathrm{T}}.$$

Using now Lemma A.1, we obtain

$$\vec{\pi}(\Lambda) = e_k^{\mathrm{T}} \hat{Q}^{-1}(\Lambda), \tag{A.12}$$

and the results follow. □

Remark A.2. Using the same kind of arguments as in Remark A.1, we deduce from Lemma A.2 that

$$p_n(u) = e_k^{\mathrm{T}} (\hat{Q}_n(u))^{-1}$$

is a consistent estimator of $\vec{\pi}$, and that

$$\mathrm{d}\, p_n^{ij}(u) = e_k^{\mathrm{T}} \mathrm{D}_n^{ij}(u)$$

is a consistent estimator of $\partial\vec{\pi}/\partial\lambda_{ij}$.

Lemma A.3. *If* B, g, ψ *are as in* (16) *we have that*

$$\frac{\partial\psi(\Lambda)}{\partial\lambda_{ij}} = \frac{(1/st)(\partial g(\Lambda)/\partial\lambda_{ij})}{g(\Lambda)}, \tag{A.13}$$

$$\frac{\partial g(\Lambda)}{\lambda_{ij}} = \frac{\partial\vec{\pi}}{\lambda_{ij}} B(\Lambda)\vec{1} + \vec{\pi}(\Lambda) \sum_{n=0}^{\infty} \sum_{r=0}^{n-1} \frac{t^{n-1}(Q(\Lambda) + Hs)^r V^{ij} (Q(\Lambda) + Hs)^{n-1-r}}{n!} \vec{1}. \tag{A.14}$$

Proof. Apply chain rule and (A.2). □

References

[1] F.P. Kelly, Notes on effective bandwidth, in: F.P. Kelly, S. Zachary, I.B. Ziedins (Eds.), Stochastic Networks: Theory and Applications, Oxford University Press, Oxford, 1996, pp. 141–168.
[2] G. Kesidis, J. Walrand, C.-S. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, IEEE ACM Trans. Networking 1 (4) (1993) 424–428.
[3] G. de Veciana, J. Walrand, Effective bandwidths: call admission, traffic policing and filtering for ATM networks, Queueing Syst. 20 (1995) 37–59.
[4] R.J. Gibbens, Traffic characterisation and effective bandwidths for broadband network traces, in: F.P. Kelly, S. Zachary, I.B. Ziedins (Eds.), Stochastic Networks: Theory and Applications, Oxford University Press, Oxford, 1996, pp. 1–11.
[5] Traffic control and congestion control in B-ISDN, ITU-T Recommendation I.371, in: Proceedings of the Study Group 13 Meeting, Geneva, July 1995.
[6] R.L. Cruz, A calculus for network delay. I. Network elements in isolation, IEEE Trans. Inform. Theory 37 (1) (1991) 114–131.
[7] R.L. Cruz, A calculus for network delay. II. Network analysis, IEEE Trans. Inform. Theory 37 (1) (1991) 132–141.
[8] M. Simon, J. Pechiar, M. de Oliveira, L. Casamayou, Video coding and ATM statistical bit rate capability: ATM networks, in: D. Kouvatsos (Ed.), Performance Modelling and Analysis, Chapman & Hall, London, 1997.

[9] J. Pechiar, M. Simon, Multiplexing real time video services, in: Proceedings of the Fifth IFIP Workshop on Performance, Modelling and Evaluation of ATM Networks, Ilkley, UK, June 1997.
[10] E. Rathgeb, Modelling and performance comparison of policing mechanisms for ATM networks, IEEE J. Select. Areas Commun. 9 (3) (1991) 325–334.
[11] D. Le Gall, The MPEG Video Compression Algorithm: A Review, Image Processing Algorithms and Techniques II, Vol. 1452, SPIE, 1991.
[12] ISO/IEC JTC 1, International Standard 11172, MPEG 1.
[13] ISO/IEC 13818 Generic coding of moving pictures and associated audio (MPEG 2), November 1994, in: Proceedings of the IEEE JSAC on Mechanisms for ATM Networks, Singapore, April 1991.
[14] A. Baiocchi, N. Blefari Melazzi, M. Listani, A. Roveri, R. Winkler, Loss performance analysis of an ATM multiplexer loaded with high speed ON/OFF sources, IEEE J. Selected Areas Commun. 9 (3) (1991) 388–393.
[15] V.G. Kulkarni, Effective bandwidths for Markov regenerative sources, Queueing Syst. 24 (1996) 137–153.
[16] P. Jelenković, A. Lazar, N. Semret, The effect of multiple time scales and subexponentiallity in MPEG video streams on queuing behaviour, IEEE J. Selected Areas Commun. 15 (6) (1997) 1052–1071.
[17] V. Frost, B. Melamed, Simulating telecommunications networks with traffic modelling, IEEE Commun. Mag. (1994).
[18] V. Siris, Large deviation techniques for traffic engineering. http://www.ics.forth.gr/netgroup/msa.
[19] J.R. Norris, Markov Chains, Cambridge University Press, Cambridge, 1998.
[20] E.A. Lebedev, L.I. Lukashuk, Maximum likelihood estimation of the infinitesimal matrix of a Markov chain with continuous time, Dokl. Akad. Nauk Ukr. SSR Ser. A 1 (1986) 12–14 (in Russian with English summary).
[21] D. Dacunha Castelle, M. Dufló, Probabilités et Statistique, Vol. 2: Problèmes à Temps Mobile, Masson, Paris, 1993.
[22] I.N. Kovalenko, N.Yu. Kuznetsov, V.M. Shurenkov, Models of Random Processes: A Handbook for Mathematicians and Engineers, CRC Press, Boca Raton, FL, 1996.