TECHNICAL RESEARCH REPORT

Approximating a Variable Bit Rate Source by Markov Processes

by N. Rananand

T.R. 94-59

INSTITUTE FOR SYSTEMS RESEARCH

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

To appear, Proceedings GLOBECOM '94

APPROXIMATING A VARIABLE BIT RATE SOURCE BY MARKOV PROCESSES

Nol Rananand*
Department of Electrical Engineering and Institute for Systems Research
University of Maryland, College Park, MD 20742
nol@eng.umd.edu

Abstract

We consider the problem of approximating a variable bit rate (VBR) source by a simple process such that the corresponding buffer-related performance measures are close approximations of the true performance measures. Assuming that a VBR source can be modeled by a discrete-time batch Markovian arrival process (D-BMAP), we propose an approach for approximating it by a "matched" Markov process of finite memory obtained by information-theoretic techniques. We confirm analytically that the approximating performance measures become increasingly accurate with the memory of the matched Markov process. When the parameters of the D-BMAP are unknown, we estimate instead the parameters of a suitable Markov approximation from samples of an observed cell stream. We show that the estimated Markov process, with a fixed memory, comes closer with increasing sample size to the D-BMAP, as do the corresponding performance measures, in accordance with the law of iterated logarithm. Numerical examples are presented to illustrate the effectiveness of the approach.

1 Introduction

Since variable bit rate (VBR) sources will be one of the major contributors to traffic on an ATM network, it is imperative that we are able to model the sources and their superpositions, by accurate yet analytically tractable stochastic models. Many stochastic models have been proposed for a VBR source in the literature (see [6]). It has been shown in [3] that the discrete-time batch Markovian arrival process (D-BMAP) is a good candidate to represent a VBR source at the cell-level, in view of its analytical tractability and flexibility in reflecting traffic characteristics. The performance analysis of an ATM statistical multiplexer fed by a D-BMAP or its special cases has also been much studied (cf. [3, 4, 12]). We assume that any VBR source can be accurately modeled by an appropriate D-BMAP. (D-BMAPs have also been used in speech recognition, where they are referred to as hidden Markov sources (HMSs).)

The D-BMAP is governed by an unobservable state process modeled by a Markov chain. The versatility of the D-BMAP results from the introduction of this process. The probabilistic mechanism generating the number of arrivals in any time unit can then be allowed to depend on the state in which the state process resides. On the other hand, the state process also possesses some undesirable traits. When the D-BMAP is used to represent an arrival stream at a switching node, owing to the "hidden" nature of the state process, processing at the node which relies on detailed statistical information concerning the arrival stream can be difficult. Another practical problem arises when the parameters of the D-BMAP are not known in advance and need to be estimated from *real* samples of the cell stream. One technique for estimating the parameters of the D-BMAP consists of fitting some carefully chosen empirical moments from samples of a cell stream to corresponding statistical moments of the D-BMAP [3, 4, 7]. Unfortunately, this technique is applicable to simple D-BMAPs only, and not to general D-BMAPs. There also are existing algorithms, such as the EM algorithm [1], that enable us to estimate the parameters of the D-BMAP from samples of a cell stream. However, they are computationally prohibitive due to the hidden nature of the state process, and inappropriate for high-speed networks.

This has motivated us to explore the feasibility of approximating the D-BMAP by "simpler" processes which do not suffer from the aforementioned problems, but are "close" approximations of the D-BMAP in the *queueing context*, i.e., the resulting buffer-related performance measures are good approximations of the performance measures corresponding to the D-BMAP. In our initial efforts below, we concentrate on two first-order performance measures: The probability of cell loss due to buffer overflow and the average cell delay in the buffer. We approximate the D-BMAP by a "matched" Markov process of finite memory which is the Markov process closest to the D-BMAP in an *information-theoretic context* (which employs the notion of *Kullback-Leibler divergence*), and assess the validity of this approximation in the queueing context. We confirm analytically that the approximating Markov process becomes more accurate as the memory of the process increases, in both contexts. Another motivation for using Markov approximations, in addition to the fact that there are no longer any hidden states, is that the estimation of the parameters of the approximating Markov process from samples of a cell stream is much simpler, for instance, by using the maximum likelihood (ML) estimation method. We show that the estimated Markov process, with a fixed memory, comes closer with increasing sample size to the D-BMAP, as do the corresponding performance measures, in accordance with the *law of iterated logarithm (LIL)*.

The paper is composed as follows. Section 2 briefly reviews the description of the queueing system, and discusses the matched Markov process obtained from the D-BMAP. Our analytical results which show the accuracy of the Markov approximations are presented in section 3. Their proofs are omitted due to limited

*This research was supported by the Institute for Systems Research at the University of Maryland, College Park, MD, USA, under NSF Grant OIR-85-00108, and by the Center for Satellite and Hybrid Communication Networks at the University of Maryland under grant NAGW-2777S.

space and can be found in [9, 10]. Section 4 provides numerical examples to substantiate the validity of the approximation technique. Further discussion is contained in section 5.

2 Preliminaries

2.1 The queueing system

The queueing system used here is a slight modification of that presented in [12]. A discrete-time single-server queue with a finite first-in-first-out (FIFO) buffer represents a statistical multiplexer. The basic time unit, called a *slot*, is equal to the constant service time of a cell. The arrival process at the multiplexer $\{X_t\}_{t=1}^{\infty}$ is assumed to be a D-BMAP where X_t denotes the number of arriving cells in the t-th slot (in steady state), taking values in $\mathcal{X} = \{0, 1, \ldots, N-1\}$. The D-BMAP is governed by a time-homogeneous Markov chain $\{S_t\}_{t=0}^{\infty}$, where S_t denotes the state of the D-BMAP in the t-th slot, taking values in $\mathcal{S} = \{1, \ldots, M\}$. The Markov chain $\{S_t\}_{t=0}^{\infty}$ is generated by an *irreducible* and *aperiodic* $M \times M$-transition probability matrix (t.p.m.) $A = \{a_{uv}\}$, $a_{uv} = Pr(S_t = v|S_{t-1} = u)$, $u, v \in \mathcal{S}$, with the initial distribution being the corresponding unique invariant distribution $\pi = (\pi_1, \ldots, \pi_M)$, where $\pi_i = \lim_t Pr(S_t = i)$, $i \in \mathcal{S}$. (This choice of π being the initial distribution is necessary for subsequent information-theoretic comparisons.) We assume that $\{X_t\}_{t=1}^{\infty}$ is generated by $\{S_t\}_{t=0}^{\infty}$ according to a time-homogeneous transition mechanism specified by a $M \times N$-stochastic matrix $B = \{b_{il}\}$, with

$$\begin{aligned} b_{il} &\stackrel{\text{def}}{=} Pr(X_t = l|S_t = i, S_0^{t-1}, X_1^{t-1}) \\ &= Pr(X_t = l|S_t = i) \qquad i \in \mathcal{S}, l \in \mathcal{X}, \end{aligned}$$

for $t \geq 1$. (Throughout this paper, s_m^n (resp. S_m^n) refers to the subsequence $(s_n, \ldots, s_m)$ (resp. $(S_n, \ldots, S_m)$) of $\mathcal{S}$-valued symbols (resp. random variables) , $0 \leq m < n$.) We then write for $x_1^n \in \mathcal{X}^n$,

$$Pr(X_1^n = x_1^n) = \sum_{s_0 \in \mathcal{S}} \Big(\sum_{s_1^n \in \mathcal{S}^n} \prod_{t=1}^{n} b_{s_t x_t} a_{s_{t-1} s_t} \Big) \pi_{s_0}, \tag{1}$$

for $n \geq 1$. It is easy to see (as also shown in [8]) that $\{X_t\}_{t=1}^{\infty}$ thus defined is stationary and ergodic. Let $\mathcal{X}^{\infty}$ be the set of all infinite sequence of symbols from $\mathcal{X}$, and $\mathcal{B}^{\infty}$ denote a suitable σ-field on $\mathcal{X}^{\infty}$. Let P denote the (stationary ergodic) probability measure on $(\mathcal{X}^{\infty}, \mathcal{B}^{\infty})$ which generates $\{X_t\}_{t=1}^{\infty}$ in accordance with eq. (1). Before proceeding further, we introduce some additional notation. Let e_j, I_j and 0_{ij} denote respectively the row-vector of 1's of size j, the $j \times j$ identity matrix and the $i \times j$ zero matrix. Also let A be a matrix, A^T is then the transpose of A. (Hereafter, we shall suppress the subscripts if dimensions are apparent from the context.)

We next introduce the set of $M \times M$-matrices $C_l = \{c_{ij}(l)\}$, $0 \leq l \leq N-1$, where $c_{ij}(l) = Pr(X_{t+1} = l, S_{t+1} = j \mid S_t = i) = Pr(X_{t+1} = l|S_{t+1} = j) \cdot Pr(S_{t+1} = j \mid S_t = i) = b_{jl} a_{ij}$, $l \in \mathcal{X}$, $i, j \in \mathcal{S}$. Note that $\sum_{l=0}^{N-1} C_l = A$. The average number of arriving cells is denoted as $\rho = \sum_{i=1}^{M} \pi_i \sum_{j=1}^{M} \sum_{l=0}^{N-1} l c_{ij}(l)$.

Let the size of the FIFO buffer be K cells (including the cell in service). We assume that $N-1 < K$. The service provided to a cell (assuming there is at least one cell in the system) commences at the beginning and completes at the end of the slot, at which time the cell departs from the system. Let Y_t denote the number of cells in the system immediately after the end of the t-th slot (before the state changes). (See Fig.1.) Therefore, Y_t is governed by the following dynamic equation: For $t \geq 0$,

$$Y_{t+1} = (Y_t - 1)^+ + \min\{X_{t+1},\ K - Y_t\} \tag{2}$$

assuming $Y_0 = 0$, where $(x)^+$ denotes $\max\{0, x\}$. The process $\{(Y_t, S_t)\}_{t=0}^{\infty}$ is then a bivariate stationary Markov chain with state space $\{0, 1, \ldots, K-1\} \times \mathcal{S}$. Let D_t denote the $M \times M$ matrix $\sum_{l=t}^{N-1} C_l$, $1 \leq t \leq N-1$. The $KM \times KM$ t.p.m. T for the bivariate Markov chain $\{(Y_t, S_t)\}_{t=0}^{\infty}$ can be written in terms of the C- and D- matrices (cf. [10]).

The Markov chain $\{(Y_t, S_t)\}_{t=0}^{\infty}$ is stationary and ergodic hence a unique steady state distribution exists. Let $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{K-1})$ denote the steady state distribution vector of $\{(Y_t, S_t)\}_{t=0}^{\infty}$, where γ_i is a $1 \times M$-vector whose j-th element γ_{ij} is given by $\gamma_{ij} = \lim_t Pr(Y_t = i, S_t = j)$. We can obtain γ by solving a set of linear equations $\gamma = \gamma T$ together with $\sum_{i=0}^{K-1} \gamma_i e^T = 1$. A number of algorithms to calculate γ numerically have been presented [12, 4, 11]. Having solved for γ by using one of these algorithms, we proceed to compute the probability of cell loss due to buffer overflow and the average cell delay in the buffer. The probability of cell loss due to buffer overflow, P_{loss}, is defined in [12] as the ratio of the average number of discarded cells to the average number of arriving cells in a slot, and is given by

$$P_{loss} = \frac{\rho - (1 - \gamma_0 e^T)}{\rho}. \tag{3}$$

Next, let $\overline{Y}$ be the average number of cells in the system, given by $\overline{Y} = \sum_{i=0}^{K-1} i \gamma_i e^T$. By Little's law, the average cell delay in the buffer, D_b, is simply

$$D_b = \frac{\overline{Y}}{\rho(1 - P_{loss})} - 1. \tag{4}$$

As explained in the previous section, there are certain circumstances that D-BMAP creates some difficulties. We then propose an approach to approximate the D-BMAP by appropriate Markov processes, as explained next.

2.2 Markov approximations

In this subsection, we identify the stationary ergodic Markov measure of memory k, $k \geq 1$, which is the "best" approximation of the measure P (generating the D-BMAP) from among all stationary ergodic Markov measures, by using an information-theoretic technique. Given $k \geq 1$, let the set $\{r_{su}^{(k)}; s \in \mathcal{X}^k,\ u \in \mathcal{X}\}$ define a $N^k \times N$ t.p.m $R^{(k)}$. Let $\Psi^{(k)}$ be the set of all such t.p.m's which are irreducible and aperiodic, and for each $R^{(k)} \in \Psi^{(k)}$, let $\pi^{(k)}$ be the corresponding unique invariant distribution on $\mathcal{X}^k$. Let $Q^{(k)}$ be a measure on $(\mathcal{X}^{\infty}, \mathcal{B}^{\infty})$ generated from $R^{(k)}$ and $\pi^{(k)}$ as follows. For $n > k$,

$$Q^{(k)}(x_1^n) = \pi^{(k)}(x_1^k) \prod_{t=k+1}^{n} r_{x_{t-k}^{t-1} x_t}^{(k)}. \tag{5}$$

We hereafter refer to $Q^{(k)}$ as a stationary ergodic Markov measure of *memory* k, and let $\mathcal{M}^{(k)}$ be the set of stationary ergodic Markov measures corresponding to $\Psi^{(k)}$. For each $k \geq 1$, we define the stationary ergodic Markov measure $P_a^{(k)}$ in $\mathcal{M}^{(k)}$ which is the *best* approximation to P as

$$P_a^{(k)} \stackrel{\text{def}}{=} \arg \inf_{Q^{(k)} \in \mathcal{M}^{(k)}} D(P \| Q^{(k)}), \tag{6}$$

where the *Kullback-Leibler divergence* or *information divergence* $D(P\|Q^{(k)})$ is defined as

$$D(P\|Q^{(k)}) = \lim_n E_P \left[\log \frac{P(X_{n+1}|X_1^n)}{Q^{(k)}(X_{n+1}|X_1^n)} \right], \qquad (7)$$

where E_P denotes expectation with respect to the measure P. We show in [9, 10] that $\min_{Q^{(k)} \in \mathcal{M}^{(k)}} D(P\|Q^{(k)})$ exists and for $x_1^n \in \mathcal{X}^n$,

$$P_a^{(k)}(x_1^n) = \begin{cases} P(x_1^n), & n \leq k, \\ P(x_1^k) \prod_{t=k+1}^n P(x_t|x_{t-k}^{t-1}), & n > k, \end{cases} \qquad (8)$$

and

$$\lim_k D(P\|P_a^{(k)}) = 0.$$

The previous equality asserts that the best Markov approximation $P_a^{(k)}$ approaches P in the limit. Let the irreducible and aperiodic $N^k \times N$ t.p.m. of $P_a^{(k)}$ be expressed as $V^{(k)} = \{v_{su}^{(k)}; s \in \mathcal{X}^k,\ u \in \mathcal{X}\}$, where $v_{su}^{(k)} = P(X_t = u|X_{t-k}^{t-1} = s)$. We then adopt the Markov process generated by $P_a^{(k)}$—the "matched" Markov process—to be the Markov approximation (of memory k) of the D-BMAP generated by the measure P.

3 Performance measures via Markov approximations

In subsection 3.1, we study how close the approximating performance measures of the statistical multiplexer—computed by assuming the arrival process to be generated by $P_a^{(k)}$—are to the true performance measures using the D-BMAP generated by a known P. Next, in subsection 3.2, we consider the situation where the parameters of the D-BMAP are unknown. We estimate the parameters of the approximating Markov process from samples of an observed cell stream, and then use the performance measures of the statistical multiplexer corresponding to the *estimated* Markov process as approximations of the true performance measures.

3.1 Computed Markov approximations

When the D-BMAP is fully specified, i.e., the matrices A, B are known, the considered performance measures resulting from an arrival process modeled by a D-BMAP, are compared with the same set of performance measures computed on the basis of an arrival process modeled by its Markov approximation of memory k. For the approximating system, the probability of cell loss due to buffer overflow and the average cell delay in the buffer are denoted by $P_{loss}^{(k)}$ and $D_b^{(k)}$, respectively. The following results show that the approximations of the performance measures improve steadily with the memory of the approximations.

Let $y = (y_1, \ldots, y_n)$ be a $1 \times n$-vector with real entries. The l_1 norm of y, denoted by $\|y\|_1$, is defined as $\sum_i |y_i|$. Also, for a $m \times n$-matrix $B = \{b_{ij}\}$, the l_1 norm is $\|B\|_1 \stackrel{\text{def}}{=} \max_i(\sum_j |b_{ij}|)$.)

Theorem 1: For $k \geq 1$, we have

$$|P_{loss} - P_{loss}^{(k)}| \leq c I^{\frac{1}{2}}(S_{k+1} \wedge X_1|X_2^{k+1}), \qquad (9)$$

and

$$|D_b - D_b^{(k)}| \leq 2cK \frac{I^{\frac{1}{2}}(S_{k+1} \wedge X_1|X_2^{k+1})}{(1 - P_{loss}^{(k)})} \qquad (10)$$

where $c = \frac{\tau_1(Z_T)\ln 2}{\rho\sqrt{2}}$, $Z_T \stackrel{\text{def}}{=} (I - T + e^T \cdot \gamma)^{-1}$ is the *fundamental matrix* of T, $\tau_1(Z_T) \stackrel{\text{def}}{=} \sup_{\|u\|_1=1, u\cdot e^T=0} \|uZ_T\|_1 < \infty$ is the l_1 *coefficient of ergodicity* of Z_T [5], and $I(\cdot \wedge \cdot|\cdot)$ denotes conditional mutual information. Furthermore,

$$\lim_k P_{loss}^{(k)} = P_{loss} \quad \text{and} \quad \lim_k D_b^{(k)} = D_b. \quad \square$$

Theorem 1 and the limiting result of divergence in the previous section together imply that the matched Markov process becomes more accurate in both the information-theoretic and queueing contexts, as its memory increases.

3.2 Estimated Markov approximations

We next turn to the situation where we have partial information that the arrival process is a D-BMAP; however, the matrices A and B are not known. Instead, given to us are samples of a cell stream x_1^n, $n \geq 1$, comprising the first n elements of an $\omega \in \mathcal{X}^\infty$ generated by the (unknown) measure P. We then consider the Markov approximation of memory k of the unknown D-BMAP and use the ML estimation method to obtain the corresponding t.p.m. on the basis of x_1^n. The ML estimate of the t.p.m. on the basis of x_1^n, $n \geq 1$, denoted by $\hat{V}^{(k)}(\omega, n) = \{\hat{v}_{su}^{(k)}(\omega, n)\}$, is given in terms of empirical counts by (cf. [2]),

$$\hat{v}_{su}^{(k)}(\omega, n) = \frac{l_{su}^{(k)}(\omega, n)}{l_s^{(k)}(\omega, n)}, \qquad u \in \mathcal{X},\ s \in \mathcal{X}^k, \qquad (11)$$

where $l_{su}^{(k)}(\omega, n)$ and $l_s^{(k)}(\omega, n)$ are defined as follows:

$$l_{su}^{(k)}(\omega, n) \stackrel{\text{def}}{=} \sum_{t=1}^n 1(\omega : x_t = u, x_{t-k}^{t-1} = s)$$

and

$$l_s^{(k)}(\omega, n) \stackrel{\text{def}}{=} \sum_u l_{su}^{(k)}(\omega, n),$$

where $1(\cdot)$ denotes indicator function and we use the convention $x_{-k+1}^0 = x_{n-k+1}^n$ in eq. (3.2). For each $u \in \mathcal{X}$, $s \in \mathcal{X}^k$, $\hat{v}_{su}^{(k)}(n) \stackrel{\text{def}}{=} \hat{v}_{su}^{(k)}(\cdot, n)$ is a random variable on $(\mathcal{X}^\infty, \mathcal{B}^\infty)$.

Our choice of the ML estimate is motivated by the particularly simple form of the estimate $\hat{v}_{su}^{(k)}(n)$ in eq. (11), as also the fact that the estimate possesses desirable strong consistency properties (cf. [2]), namely, for $u \in \mathcal{X}$, $s \in \mathcal{X}^k$,

$$\lim_n \hat{v}_{su}^{(k)}(n) = v_{su}^{(k)} \qquad P\text{–a.s.}$$

Moreover, we also show in [9, 10] that the ML estimate also satisfies the Law of Iterated Logarithm (LIL), one of the sharpest known strong limit theorems, as follows. For $u \in \mathcal{X}$, $s \in \mathcal{X}^k$ such that $0 < v_{su}^{(k)} < 1$, it holds that

$$\limsup_n \frac{\left| \hat{v}_{su}^{(k)}(n) - v_{su}^{(k)} \right|}{\frac{\sqrt{2b}}{P(s)} \sqrt{\frac{\log \log n}{n}}} = 1 \qquad P\text{–a.s.},$$

where $0 < b < \infty$. (The case $v_{su}^{(k)} = 0$ or 1 are trivial.)

We claim without proof (the complete proof is in [9, 10]) that, based on the LIL result above, it is true for ω belonging to a set of P-measure one that the t.p.m. $\hat{V}^{(k)}(\omega, n)$ will become irreducible and aperiodic, for n "sufficiently" large. We subsequently compute the probability of cell loss due to buffer overflow, denoted by

$\hat{P}^{(k)}_{loss}(\omega, n)$, and the average cell delay in the buffer, denoted by $\hat{D}^{(k)}_b(\omega, n)$, by using a Markov process of memory k with a t.p.m. $\hat{V}^{(k)}(\omega, n)$ as the arrival process to the multiplexer. We then define respectively $\hat{P}^{(k)}_{loss}(n)$ and $\hat{D}^{(k)}_b(n)$ as random variables on $(\mathcal{X}^\infty, \mathcal{B}^\infty)$ according to $\hat{P}^{(k)}_{loss}(\cdot, n)$ and $\hat{D}^{(k)}_b(\cdot, n)$. To this end, let us introduce a definition of *eventually P–a.s.* as follows: Let $\{Z_t\}_{t=1}^\infty$ be a sequence of $\mathbb{R}$-valued random variables governed by a measure P, and $\{\alpha_t\}_{t=1}^\infty$ a sequence of real numbers. We say that $Z_t = O(\alpha_t)$ eventually P–a.s. if there exists a positive random variable N which is finite P–a.s. and a (finite) constant C such that $|Z_n| \le C\alpha_n$, for $n \ge N$.

Theorem 2: For $k \ge 1$, it holds that

$$|P_{loss} - \hat{P}^{(k)}_{loss}(n)| \le cI^{\frac{1}{2}}(S_{k+1} \wedge X_1 | X_2^{k+1}) + O\left(\sqrt{\frac{\log\log n}{n}}\right) \quad \text{eventually } P\text{–a.s.},$$

and

$$|D_b - \hat{D}^{(k)}_b(n)| \le 2cK \frac{I^{\frac{1}{2}}(S_{k+1} \wedge X_1 | X_2^{k+1})}{(1 - P^{(k)}_{loss})} + O\left(\sqrt{\frac{\log\log n}{n}}\right) \quad \text{eventually } P\text{–a.s.},$$

where c is as defined in Theorem 1. Furthermore,

$$\lim_k \lim_n \hat{P}^{(k)}_{loss}(n) = P_{loss} \qquad P\text{–a.s.},$$

and

$$\lim_k \lim_n \hat{D}^{(k)}_b(n) = D_b \qquad P\text{–a.s.}. \square$$

4 Numerical examples

The first example is of a D-BMAP with the matrices A and B specified arbitrarily by

$$A = \begin{bmatrix} 0.60 & 0.30 & 0.05 & 0.05 \\ 0.20 & 0.30 & 0.10 & 0.40 \\ 0.17 & 0.23 & 0.39 & 0.21 \\ 0.15 & 0.20 & 0.35 & 0.30 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.298 & 0.380 & 0.290 & 0.030 & 0.002 \\ 0.600 & 0.320 & 0.072 & 0.005 & 0.003 \\ 0.250 & 0.464 & 0.265 & 0.020 & 0.001 \\ 0.503 & 0.365 & 0.125 & 0.005 & 0.002 \end{bmatrix}.$$

The steady state average number of arriving cells per slot (ρ) is equal to 0.812. We choose the buffer size (K) to be 40. Figs. 2 and 3 compare $P^{(k)}_{loss}$ and $D^{(k)}_b, k = 0, \ldots, 3\}$, with P_{loss} and D_b, respectively. (For $k = 0$, we approximate P by $P^{(0)}_a$—the independent and identically distributed (i.i.d.) measure closest to P according to Section 2. We remark here that all of our previous analysis holds also for the case $k = 0$, but specific mention is intentionally omitted.) In this case, the approximating performance measures are very close to the true performance measures, even when k equal to one.

We next present a D-BMAP which is a modification of the one studied by Takine *et al.* [12]. We assume that the D-BMAP is governed by an underlying 2-state Markov chain with transition probabilities given by $a_{11} = a_{22} = a$ and $a_{12} = a_{21} = 1 - a$, $0 < a < 1$, and a $N \times N$-matrix B described by

$$\begin{aligned} b_{1l} &= \binom{8}{l} \left(\tfrac{(1+c)\rho}{8}\right)^l \left(1 - \tfrac{(1+c)\rho}{8}\right)^{8-l}, \\ b_{2l} &= \binom{8}{l} \left(\tfrac{(1-c)\rho}{8}\right)^l \left(1 - \tfrac{(1-c)\rho}{8}\right)^{8-l}, \quad 0 \le l \le 8, \end{aligned} \qquad (12)$$

where ρ is as previously defined and c is a parameter such that $0 \le c < \min\{1, \frac{8}{\rho} - 1\}$. In our experiment, we choose ρ so that $P_{loss} \le 10^{-8}$, and let a and c vary. The choice of a and c influences two important characteristics of the D-BMAP (which is the main reason we consider this particular D-BMAP). The first is the squared coefficient of variation of the number of arriving cells per slot [12], C_v^2, which, in this case, is given by

$$C_v^2 \overset{\text{def}}{=} \frac{VAR(X_t)}{E^2(X_t)} = \frac{1}{\rho} + \frac{7c^2 - 1}{8}.$$

Note that $C_v^2 \ge 0$, and increases with c. The other important characteristic is the first-order correlation coefficient, $C_c(1)$, which can be expressed as

$$C_c(1) \overset{\text{def}}{=} \frac{COV(X_t, X_{t+1})}{VAR(X_t)} = \frac{c^2(2a-1)}{C_v^2}.$$

Hence, for fixed c and ρ, $C_c(1)$ is dictated by a. ($C_c(1)$ is a measure of dependency between number of cell arrivals in consecutive time slots.) Let $K = 80$. Figs. 4 and 5 show the effect of c on our approximations of the true performance measures when ρ and a are fixed. Here, we choose $\rho = 0.9$ and $a = 0.3$. In Fig. 4, we plot $\left|\log \frac{P^{(k)}_{loss}(c)}{P_{loss}(c)}\right|$ as a function of c, for $k \le 2$, and similarly, we show in Fig. 5 the dependencies of the percentage difference of average waiting time, i.e., $\left|\frac{D_b(c) - D^{(k)}_b(c)}{D_b(c)}\right| * 100$, on c, where $P_{loss}(c)$ ($D_b(c)$) and $P^{(k)}_{loss}(c)$ ($D^{(k)}_b(c)$) denote the usual P_{loss} (D_b) and $P^{(k)}_{loss}$ ($D^{(k)}_b$) corresponding to a D-BMAP with parameter $c \in \{0.1, 0.2, \ldots, 0.9\}$. We observe in this case that by using the Markov approximation of memory 2 ($k = 2$), we obtain very close approximations to the true performance measures for the entire range of values of c. Figs. 4 and 5 also show that the i.i.d. approximation does not perform well when $c > 0$ (which implies that $C_v^2 > 0$), the reason being that the approximating i.i.d. process has $C_v^2 = 0$.

Using the same model, we study the effect of a on the approximations of performance measures when ρ and c are fixed. Here, we choose $\rho = 0.9$ and $c = 0.5$. Figs. 6 and 7 display the discrepancies between the true and approximating performance measures in the same fashion as in Figs. 4 and 5, except that they now depend on a. In this case, we also have to use the Markov approximation of memory 2 to get close approximations to the true performance measures; for $0.2 \le a \le 0.8$, and for $a = 0.1$ or 0.9, we have to increase k beyond 2 to get satisfactory approximations (not shown here). Another interesting point is that, for $a = 0.5$, we observe from Figs. 6 and 7 that all Markov approximations perform extremely well. This perfect agreement is due to the fact that the D-BMAP degenerates to an i.i.d. process when $a = 0.5$, as do all of its Markov approximations.

Next, we demonstrate the strength of our approach when the parameters of the D-BMAP are not known. We proceed by simulating a sample path (a trace of an arriving cell stream) of a D-BMAP and then estimating the parameters of approximating Markov processes, from the sample path, via the ML method. Subsequently, the approximating performance measures can then be computed. Here, we use the same D-BMAP as the one used in Figs. 4–7 with $\rho = 0.92$, $a = 0.3$ and $c = 0.6$. As before, the buffer size $K = 80$. Ten independent sample paths are generated using the OPNET[1] software package and the averages of the approximating performance measures, based on different sample sizes

[1]OPNET is a product of MIL 3, Inc., Washington D.C

(in terms of time slots), are plotted against the true performance measures in Figs. 8 and 9. The averages of the "empirical" performance measures obtained from these sample paths are also shown. For this particular example, the empirical probability of cell loss is equal to zero for all the sample sizes considered. These numerical results show that, with relatively small sample size (10^5 time slots), the approximating performance measures obtained from a Markov process of memory 1 are very close to the true performance measures, and out-perform the empirical performance measures. These results also suggest that the estimated parameters of Markov processes almost converge to the true parameters at a sample size of about 10^5 slots. For this sample size, the discrepancies between the true and approximating performance measures arise mainly from the difference between the D-BMAP and its approximating Markov process, and not much from the estimation method. Therefore, further increments of sample sizes will not improve by much the approximations.

5 Discussion

Our work suggests that it is often advantageous to represent a VBR source by a Markov process, instead of a more complicated D-BMAP, when the performance measures of interest are the probability of cell loss (due to buffer overflow) and the average cell delay. Analytical results in terms of upper-bounds on deviations from the true performance measures are presented, and validated numerically. Though these (information-theoretic) upper-bounds are somewhat "loose", our objective has been to obtain asymptotic results which is why no attempt was made to obtain tighter bounds.

In general, Markov processes, by virtue of their finite memories, are poor approximations of the D-BMAP which, typically has "infinite" memory. Since our numerical results have shown that only a relatively small memory of the matched Markov process is required in order to obtain satisfactory approximations, it appears that the discrepancies between the D-BMAP and its Markov approximations are nicely smoothed out in the process of computing these performance measures. It is likely for some other (steady state) performance measures that a D-BMAP will still be well-approximated by a suitable Markov process, although the requisite memory may no longer be small .

An immediate application is to use this estimation technique to reduce the simulation time needed in assessing the performance measures of a switching node fed by a VBR source. The issue of controlling network operations based on parameters of approximating Markov processes for any VBR source, using our technique, needs further investigation.

Acknowledgment

The author would like to thank Prof. P. Narayan for his guidance and valuable discussions during the course of this research.

References

[1] L. E. Baum et al. "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains". *Ann. Math. Stat.*, 41(1):164–171, 1970.

[2] P. Billingsley. *Statistical inference for Markov processes*. The University of Chicago Press, 1961.

[3] C. Blondia and O. Casals. "Statistical multiplexing of VBR sources: A matrix-analytic approach". *Performance Evaluation*, 16:5–20, 1992.

[4] J.-Y. Le Boudec. "An efficient solution method for Markov models of ATM links with loss priorities". *IEEE Trans. Commun.*, 9(3):408–417, 1991.

[5] E.Seneta. *Non-negative matrices and Markov chains*. Springer-Verlag, second edition, 1980.

[6] V. S. Frost and B. Melamed. "Traffic modeling for telecommunications networks". *IEEE Comm. Mag.*, 32(3):70–81, March 1994.

[7] D. S. Holtsinger and H. G. Perros. "Performance of the buffered leaky bucket policing mechanism". In H. G. Perros, editor, *High-Speed communication Networks*, 1992.

[8] B. G. Leroux. "Maximum-likelihood estimation for hidden Markov models". *Stochastic Processes and their Applications*, 40:127–143, 1992.

[9] N. Rananand. *"Analysis of switch effects on ATM network performance"*. PhD thesis, University of Maryland at College Park, 1994. In preparation.

[10] N. Rananand. "Approximating a variable bit rate source by Markov processes". 1994. Submitted for publication.

[11] W. J. Stewart. "Recursive procedures for the numerical solution of Markov chains". In H. G. Perros and T. Altiok, editors, *Queueing networks with blocking*, pages 229–247. Elsevier Science Publishers B.V. (North Holland), 1989.

[12] T. Takine et al. "Cell loss and output process analyses of a finite-buffer discrete-time ATM queueing system with correlated arrivals". In *Proceedings of the IEEE INFOCOM'93*, pages 1259–1269, 1993.

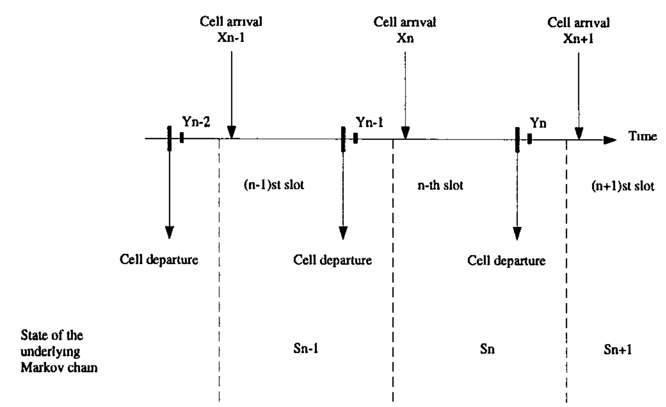

Figure 1: Time diagram

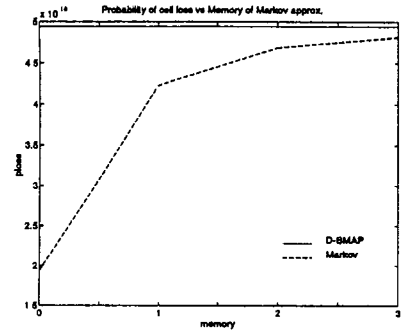

Figure 2: The probability of cell loss vs Memory of Markov approximation ($M = 4, N = 5, \rho = 0.812, K = 40$)

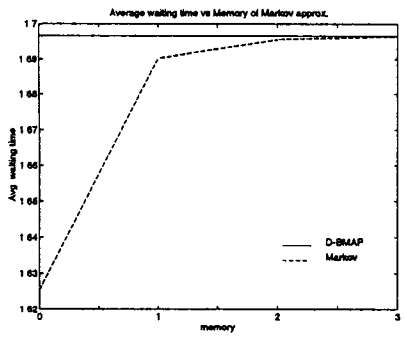

Figure 3: The average waiting time vs Memory of Markov approximation ($M = 4, N = 5, \rho = 0.812, K = 40$)

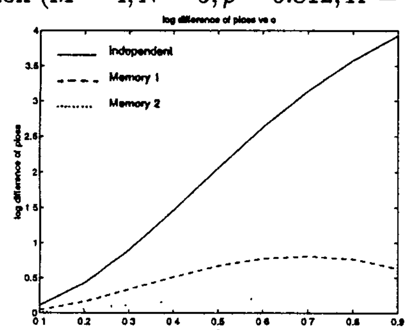

Figure 4: The absolute of log. difference of Ploss vs c ($M = 2, N = 9, \rho = 0.9, a = 0.3, K = 80$)

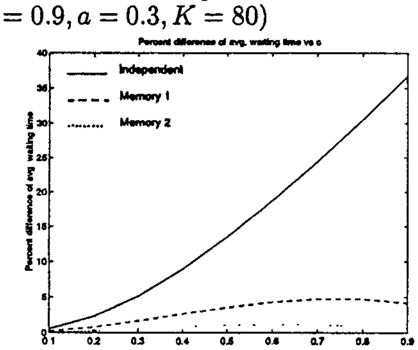

Figure 5: The percent difference of avg. waiting time vs c ($M = 2, N = 9, \rho = 0.9, a = 0.3, K = 80$)

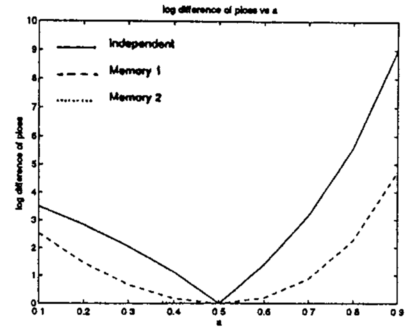

Figure 6: The absolute of log. difference of Ploss vs a ($M = 2, N = 9, \rho = 0.9, c = 0.5, K = 80$)

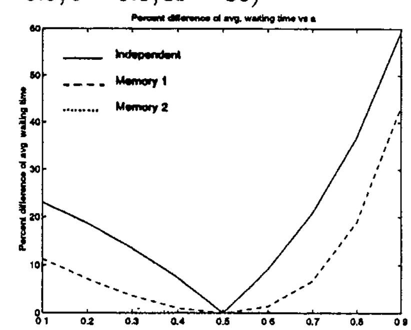

Figure 7: The percent difference of avg. waiting time vs a ($M = 2, N = 9, \rho = 0.9, c = 0.5, K = 80$)

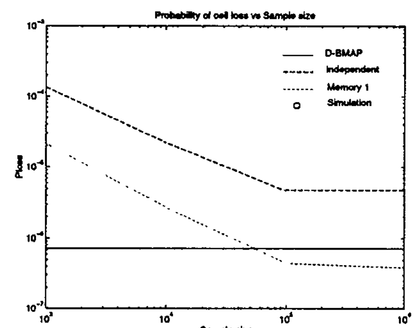

Figure 8: The probability of cell loss vs Sample size ($M = 2, N = 9, \rho = 0.92, a = 0.3, c = 0.6, K = 80$)

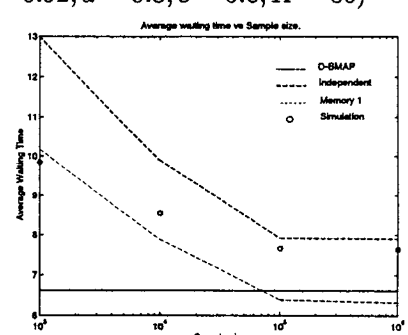

Figure 9: The average waiting time vs Sample size ($M = 2, N = 9, \rho = 0.92, a = 0.3, c = 0.6, K = 80$)