Fast Simulation of Markov Modulated Fluid Models

Ad Ridder & Michel Mandjes
Dept. of Econometrics, Free University of Amsterdam
De Boelelaan 1105, NL – 1081 HV Amsterdam

Abstract

We consider fluid models consisting of finite buffers, constant capacities and input processes modulated by continuous time Markov chains. These models are continuous flow approximations for describing the actually discrete traffic and buffer behaviour in ATM networks. A main issue is to find good estimates of buffer overflow and loss probabilities when they are (very) small, i.e. of the order 10^{-6} or less. We present simulation results by applying importance sampling as a variance reduction technique. Using the concept of effective bandwidth we change the transition rates of the modulating Markov chains by an exponentially twist. When we do that in such a way that peak rates get longer durations, overflows will occur more frequently when executing the simulations. By choosing the right twist, the new simulation procedure attains asymptotic optimality.

1 Introduction

The authors of the famous AMS paper [1] have modeled sources that request bandwidth of a communication link, as on-off Markov fluids. That is, a source is characterised by a rate matrix-input pair (Q, r). Here, Q is the transition rate matrix of a finite state continuous time Markov chain, and r_i is the rate of fluid input during the sojourn time of the chain in state i. In [1] each chain has two states, labeled 0 (off) and 1 (on), $r_0 = 0, r_1 = 1$ and all Q's are identical. The link transmits the fluid at a constant rate of c. The analysis of the buffer contents (there is a buffer in front of the link) is based on solving a linear eigensystem of differential equations and yields maneageable expressions e.g. for tail probabilities. Since then, several extensions have been made towards more general Markov fluid sources. To our knowledge exact results have been found for models with independent reversible Markov sources, see [11].

A typical design issue concerns the buffer size in order to keep the loss probabilitiy below an acceptable level which may be 10^{-6} or less. The loss probability behaves according to $\eta(B)\exp(\theta B)$ for some amplitude $\eta(B)$ and decay rate θ (B is the buffer size, $\log \eta(B) = o(B)$). This exponentially decaying of tail/loss/overflow probabilities has been found in much more general queueing models, cf. [12]. The determination of the decay rate θ is a tractable numerical procedure on which we shall dwell in this paper. The amplitude factor $\eta(B)$ (for tail probabilities) can be determined in the cases mentioned when we assume an infinite buffer. There are no explicit analytical expressions available for loss probabilities of finite buffers.

Analyses get even worse in network models. To explain what we mean, let us concentrate on a tandem model. The Markov sources request bandwidth of two consecutive links. Buffers are present in front of each link so that the system can cope with the situation when the fluid input is faster than the link capacities. A system of differential equations describing the contents of the second buffer becomes intractable for solving in a similar way as the single buffer/link case. However, the asymptotics of tail/loss/overflow probabilities remain as given above, cf. [2, 9].

An approach of obtaining accurate estimates of these probabilities is by executing a simulation model of the system. However, as mentioned above, we wish to estimate small probabilities which means that long simulation runs are required. These runs consume long computer times and make (too) many calls to the random generator causing unreliable estimates. Furthermore, when we do scenario analysis, we need to execute these simulations a number of times by varying the parameter values (such as Q, r, c, B).

Therefore, we propose the application of importance sampling for speeding up the simulations. Then we draw samples according to a new statistical law and obtain the correct estimates by compensating the new realisations by the likelihood. The intuition behind this procedure lies in the following reasoning. The communication system is described by a process that evolves in time. The state at any epoch comprises the modulating Markov chains and the buffer(s) contents. We can identify a sequence of time epochs where the process statistically repeats itself, so-called regeneration times. Using these epochs we execute regenerative simulation. That is, we collect the approriate data in each regeneration cycle, and obtain estimates by averaging. However, since reaching high buffer levels is a rare event (i.e. has small probability) only relatively few cycles will give us data. Now recall that the process is statistically driven only by the modulating Markov chains. When –in a cycle– the chains behave statistically according to their equilibrium, there will be almost no buffer overflow (or reaching high levels). We have tacitly assumed that the load of a buffer is less than unity. On the other hand, given that an overflow occurred in a cycle, the behaviours of the chains were statistically atypical on the time path until that happened. To be more precise, realisations of the process upto an overflow in a cycle seem to be drawn as if the transitions of the chains were generated by rate matrices Q^*'s. So, when we actually do generate transitions according to the Q^*

matrices, each cycle will show an overflow with high probability: quick simulation.

The difficulty of this procedure lies in finding the new matrices. The intuition of 'typical atypical' behaviour of a stochastic process may be formalised in large deviations expressions and these lead to variational problems whose solutions generate the matrices. This approach is extensively studied in [8, 10] and turns out to be numerically attractive. Theoretically interesting is the question whether the proposed new statistical law is optimal in the sense of most reducing the variance of the applied estimators. It turns out that the densities of the Markov chains under the new law are exponentially twisted versions of the original ones. In fact, within the class of statistical laws such that the densities are exponentially twisted, the proposed law is optimal (again see [8, 9, 10]).

2 Effective bandwidths and new laws

Consider for the moment a single Markov source (Q, r) and a single buffer in front of a link with transmission capacity c. The total amount of fluid generated by the source during the period $[0, t]$ is denoted by $A(t)$. The asymptotic cumulant generating function is defined by

$$\psi(\theta) = \lim_{t\to\infty} \frac{1}{t} \log\, E \exp(\theta A(t)), \quad \theta \in R.$$

The average input rate equals $m = \psi'(0)$. Notice that if we assume that Q admits the invariant density π by $\pi Q = 0$, we have equivalently $m = \langle \pi, r \rangle$, where $= \langle \cdot, \cdot \rangle$ denotes the scalar product of vectors. For stability we require $m < c$.

A κ-exponentially twisted version of the source is a Markov source (Q_κ, r) (on the same state space) such that for the corresponding asymptotic cumulant generating function

$$\psi_\kappa(\theta) = \lim_{t\to\infty} \frac{1}{t} \log\, E \exp(\theta A_\kappa(t)) = \psi(\kappa + \theta) - \psi(\kappa) \tag{1}$$

holds for all θ. The average input becomes $m_\kappa = \psi'_\kappa(0) = \psi'(\kappa)$. Notice that $m_\kappa > m$ for positive κ because $\psi(\cdot)$ is convex. Although an implicit expression of $\psi(\theta)$ is available as been given in [5], viz. as the largest eigenvalue of the matrix $Q+\theta R$ (R =diag$(r_1, \ldots, r_N)$), solving (1) for Q_κ seems to be intractable. We have studied in [8] an alternative method for obtaining the Q_κ matrix which is simply based on finding an eigenvector with positive entries associated with the eigenvalue $\psi(\kappa)$ of the matrix $Q + \theta R$.

The fact that the source is exponentially twisted when executing the program just mentioned, may be explained as follows. Originally the modulating Markov chain stays an exponentially distributed amount of time in state i with mean $1/q_i$. In the twisted version the mean is $1/q_i(\kappa)$ which is is expressable in the effective bandwidth. The effective bandwidth of the source may be defined through

$$\alpha(\theta) = \psi(\theta)/\theta, \quad \theta > 0,$$

according to [5]. We have shown in [8, 10]

$$q_i(\kappa) = q_i - \kappa(r_i - \alpha(\kappa)).$$

In other words, when we indicate the sojourn time in state i by S_i,

$$P_\kappa(S_i > s) = e^{-q_i s} e^{\kappa(r_i - \alpha(\kappa))s}.$$

The original probability is multiplied with a factor having κ in the exponent. A more appealing illustration is the following. Suppose that the chain consecutevily stays in states $i_1, \ldots, i_n$ and that the durations take $t_1, \ldots, t_n$. Set $T = \sum_{\ell=1}^n t_\ell$ and assume that $\sum_{\ell=1}^n r(i_\ell) t_\ell \geq cT + x$, i.e. the buffer contents has grown on $[0, T]$ with an amount of at least x. Then the likelihood of this to happen originally with respect to the twisted version is

$$L(x) = \left(\prod_{\ell=2}^n \frac{q_{i_{\ell-1} i_\ell}}{q_{i_{\ell-1} i_\ell}(\kappa)} \right) \exp\left(- \sum_{\ell=1}^n \left(q_{i_\ell} - q_{i_\ell}(\kappa) \right) t_\ell \right). \tag{2}$$

The exponent can be overestimated by

$$-\kappa x - \kappa c T (1 - \alpha(\kappa)/c).$$

Now recall that we actually like to run simulations until high buffer levels are attained. The 'quick' simulations generate transitions of the modulating chain according to the new matrix Q_κ. The mean input rate m_κ is increasing (in κ). The data collected with the quick simulations are compensated by factors similarly to L given above. So, on the one hand we like to make m_κ as large as possible, on the other hand we like to make L as small as possible. Trading these conflicting options, the best choice becomes apparently to apply $\kappa > 0$ such that

$$\alpha(\kappa) = c. \tag{3}$$

Then $m_\kappa = c + \kappa \alpha'(\kappa)$ and $L(x) \leq G \exp(-\kappa x)$ with G the maximum of all prefactors in (2), which is a finite constant (see [10]).

The new transition matrix Q_κ that goes along with the choice of (3) has been identified differently in [8]. As has been stipulated in section 1, when there is an overflow in a cycle, the state transitions of the modulating chain until that happened, are most likely drawn from some rate matrix Q^* rather than Q. The theory of large deviations has developed asymptotic expressions to quantify probabilistically such events, e.g. in [4]. Popular writing it says that

$$\frac{1}{t} \log P(\text{the Markov chain follows approximately } Q^* \text{ during } [0, t]) \approx -I(Q^*|Q)$$

for large t, where $I(Q^*|Q)$ is the relative entropy function

$$I(Q^*|Q) = \sum_i \pi_i^* \sum_{j \neq i} q_{ij}^* \log \frac{q_{ij}^*}{q_{ij}} \; + \; \sum_i \pi_i^* (q_{ii}^* - q_{ii}),$$

according to [6]. Here π^* is the invariant density of Q^* and hence the mean input rate, while the state transitions are drawn from Q^*, equals $\langle \pi^*, r \rangle$. Again we may argue that there are two conflicting options, viz. on the one hand we like the mean input rate to be

as large as possible, on the other hand we like to minimize the entropy function. The best solution is $Q^* = Q_\kappa$ with κ given by (3), see [8].

The case of multiple independent Markov sources now readily follows. Suppose that the j-th source is characterised by the pair $(Q^{(j)}, r^{(j)})$. Then we define similarly as before the asymptotic cumulant generating function and the effective bandwidth of the source,

$$\psi^{(j)}(\theta) = \lim_{t\to\infty} \frac{1}{t} \log\, E \exp(\theta A^{(j)}(t)), \quad \alpha^{(j)}(\theta) = \psi^{(j)}(\theta)/\theta.$$

The total effective bandwidth is the sum of all the j-th effective bandwidths ([5]); hence we solve similarly to (3)

$$\sum_j \alpha^{(j)}(\kappa) = c \tag{4}$$

and use κ for twisting the sources. For instance

$$q_i^{(j)}(\kappa) = q_i(j) - \kappa(r_i(j) - \alpha^{(j)}(\kappa)).$$

Now we consider the tandem model. Suppose that a single Markov source (Q, r) generates fluid that is transmitted first by a link with capacity c_1 and next by a link with capacity c_2. There is a buffer in front of each link. We are interested in the tail/loss/overflow probabilities of the second buffer. Again we like to twist the Markov chain and apply quick simulations. Can we find the new matrix Q_κ?

The arrival process at the second buffer, $A^{(2)}(t)$, being the total amount of fluid arriving during $[0, t]$, equals the departure process of the first link, $D^{(1)}(t)$. The arrival process at the first buffer is $A^{(1)}(t)$ with asymptotic cumulant generating function $\psi^{(1)}(\theta)$ and effective bandwidth $\alpha^{(1)}(\theta)$ as before. Let us find κ_1 that satisfies $\psi^{(1)'}(\kappa_1) = c_1$. If we would twist the source with κ_1, the mean arrival rate at the first buffer becomes $m_{\kappa_1} = c_1$, i.e. the first buffer has unit load.

The effective bandwidth $\alpha^{(2)}(\cdot)$ of the arrival process at the second buffer equals the effective bandwidth of the departure process $D^{(1)}(t)$ from the first buffer and has been determined in [3] to be

$$\alpha^{(2)}(\theta) = \begin{cases} \alpha^{(1)}(\theta) & \text{if } \theta \le \kappa_1, \\ c_1 - \frac{\kappa_1}{\theta}(c_1 - \alpha^{(1)}(\kappa_1)) & \text{if } \theta > \kappa_1. \end{cases}$$

Now suppose that we apply (3) to the second buffer: find $\kappa_2 > 0$ such that

$$\alpha^{(2)}(\kappa_2) = c_2. \tag{5}$$

The solution could be used to be the twist factor for the source. However, suppose that $\kappa_2 > \kappa_1$ and so $m_{\kappa_2} > c_1$. Therefore, the actual outputs of the first buffer using twist factor κ_1 or κ_2 are asyptotically the same,

$$\lim_{t\to\infty} \frac{1}{t} \log\, E \exp(\theta D_{\kappa_1}^{(1)}(t)) = \lim_{t\to\infty} \frac{1}{t} \log\, E \exp(\theta D_{\kappa_2}^{(1)}(t)).$$

That means that there would be no speeding up of the simulations when we would increase the twist factor from κ_1 to κ_2. Conclusion: use twist factor $\kappa = \min(\kappa_1, \kappa_2)$, as also been suggested in [2, 9].

3 Optimality

Recall the simulation procedure described in section 1. The estimates are calculated by collecting the appropriate data in each regeneration cycle. Let X_i stands for these data in cycle i, with $X_1, X_2, \ldots$ i.i.d. as X. In other words when the probability to estimate iin stead ofs the tail/loss/overflow probabilities from level B, we estimate $E^{(B)}X$. The asymptotics

$$\lim_{B\to\infty} \frac{1}{B} \log\, E^{(B)}X = -\theta$$

holds as we mentioned in section 1. The decay factor θ turns out to be exactly the solutions κ (and κ_2) of (3),(4) and (5) in the various models of section 2. This is proved in [9, 10].

Suppose that the Markov sources are θ-exponentially twisted for some θ and that we execute quick simulations. Besides the data X_i, we now calculate also the likelihood L_i in each cycle. The quantity to estimate is $E_\theta^{(B)}LX = E^{(B)}X$. Since we estimate it by the average of n i.i.d. L_iX_i's, the variance of the estimator is $1/n$ times

$$\mathrm{Var}_\theta^{(B)}LX = E_\theta^{(B)}(LX)^2 - (E_\theta^{(B)}LX)^2.$$

The largest reduction is obtained by minimizing $E_\theta^{(B)}(LX)^2$. Notice that $E_\theta^{(B)}(LX)^2 \geq (E^{(B)}X)^2$ and therefore

$$\lim_{B\to\infty} \frac{1}{B} \log\, E_\theta^{(B)}(LX)^2 \geq -2\kappa,$$

for any θ and where κ is as mentioned above. Finally, we have proved in [9, 10] that the reversed inequality holds when taking the twist factor $\theta = \kappa$. Then

$$\lim_{B\to\infty} \frac{1}{B} \log\, E_\kappa^{(B)}(LX)^2 = -2\kappa$$

says that the quick simulation procedure is asymptotically optimal.

4 An example

Consider the model of [7]: two groups of on-off sources. Group 1 consists of 25 two-state modulating Markov chains with rate matrix $Q^{(1)}$ and input rate function $r^{(1)}$. Group 2 contains 50 two-state chains with $(Q^{(2)}, r^{(2)})$ input pair. The data are

$$Q^{(1)} = \begin{pmatrix} -0.4 & 0.4 \\ 1.5 & -1.5 \end{pmatrix},\ r^{(1)} = (0, 2), \quad Q^{(2)} = \begin{pmatrix} -0.6 & 0.6 \\ 0.75 & -0.75 \end{pmatrix},\ r^{(2)} = (0, 1).$$

First we take the single buffer/link model with capacity $c = 38$ (load of the buffer is 0.86). Solving (4) yields $\kappa = 0.269$ and effective bandwidths $\alpha^{(1)}(\kappa) = 0.532$ for all sources of group 1, $\alpha^{(2)}(\kappa) = 0.494$ for the sources of group 2. The new rate matrices are

$$Q_\kappa^{(1)} = \begin{pmatrix} -0.543 & 0.543 \\ 1.105 & -1.105 \end{pmatrix}, \quad Q_\kappa^{(2)} = \begin{pmatrix} -0.733 & 0.733 \\ 0.614 & -0.614 \end{pmatrix}.$$

The load becomes 1.15.

We have estimated the long run average amount of lost fluid (denoted $\zeta(B)$ for buffer level B) by simulations, both originally and κ-twisted. Table 1 shows some results. The estimates are obtained with 95% confidence and 10% relative efficiency.

	direct		quick	
B	$\zeta(B)$	#cycles	$\zeta(B)$	#cycles
15	$1.99 \cdot 10^{-4}$	610K	$2.36 \cdot 10^{-4}$	5.4K
20	$4.39 \cdot 10^{-5}$	2.6M	$4.64 \cdot 10^{-5}$	5.4K
25	$1.05 \cdot 10^{-5}$	5M	$1.04 \cdot 10^{-5}$	5.5K

Table 1: Loss fractions in single buffer model.

Typically one observes that the number of simulation cycles explodes originally while it remains almost constant in the optimal twisted procedure.

Next we suppose that the sources require two consecutive links (tandem model). The capacity of the first link is set at $c_1 = 43$ and of the second $c_2 = 36$. We estimate the long run average amount of lost fluid from the buffer in front of the second link. The first buffer is made large so that almost all fluid is transmitted through the first link. Notice that the load of the first buffer is 0.762.

Twisting the sources with $\kappa_1 = 0.254$ would yield unity load at the first buffer, whereas (5) leads to $\kappa_2 = 0.172$. So we use the optimal twist factor $\kappa = \kappa_2$ and get effective bandwidths $\alpha^{(1)}(\kappa) = 0.488$ for the sources of group 1 and $\alpha^{(2)}(\kappa) = 0.476$ for the sources of group 2. The new rate matrices are

$$Q_\kappa^{(1)} = \begin{pmatrix} -0.483 & 0.483 \\ 1.240 & -1.240 \end{pmatrix}, \quad Q_\kappa^{(2)} = \begin{pmatrix} -0.682 & 0.682 \\ 0.660 & -0.660 \end{pmatrix}.$$

The load of the first buffer becomes 0.837.

Table 2 shows some results. The estimates are obtained with 95% confidence and 15% relative efficiency.

5 Conclusions

In this paper we have considered the option of executing simulations of ATM systems that are modeled as continuous processes (both in time and space) with traffic generated

	direct		quick	
B	$\zeta(B)$	#cycles	$\zeta(B)$	#cycles
10	$2.57 \cdot 10^{-3}$	11K	$2.66 \cdot 10^{-3}$	0.9K
20	$4.97 \cdot 10^{-4}$	58K	$4.91 \cdot 10^{-4}$	0.9K
50	$2.91 \cdot 10^{-6}$	9.6M	$3.01 \cdot 10^{-6}$	1K

Table 2: Loss fractions in tandem buffer model.

by Markov chains. The issue is to design buffer sizes in order to meet service in terms of a quaranteed upper bound on loss. In order to be able to find accurate estimates of the small loss probabilities quickly, we propose importance sampling. We have shown how the effective bandwidth concept is a key issue in finding optimal new statistical laws for that procedure. From the simulation results we conclude that the time required to run the simulations, grows linearly in the buffer size when applying the optimal new laws, whereas originally it explodes exponentially.

References

[1] D. Anick, D. Mitra and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61: 1871 – 1894, 1982.

[2] C.S. Chang, P. Heidelberg, S. Juneja and P. Shahabuddin. Effective bandwidth and fast simulation of ATM intree networks. IBM Research Paper, 1992.

[3] G. de Veciana, C. Courcoubetis and J. Walrand. Decoupling Bandwidths for Networks: A Decomposition Approach to Resource Management. Memorandum UCB/ ERL M93/50, University of California, Berkeley, 1993.

[4] R.S. Ellis. *Entropy, Large Deviations and Statistical Mechanics*. Springer, New York, 1985.

[5] G. Kesidis, J. Walrand and C.S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, 1: 424 – 428, 1993.

[6] G. Kesidis and J. Walrand. Relative entropy between Markov transition rate matrices. *IEEE Transactions on Information Theory*, 39: 1056 – 1057, 1993.

[7] L. Kosten. Stochastic theory of data-handling systems with groups of multiple sources. In H. Rudin and W. Bux (eds.), *Performance of Computer-Communication Systems*, 321 – 331, Elsevier Amsterdam, 1984.

[8] M. Mandjes and A. Ridder. Finding the Conjugate of Markov Fluid Processes. To appear in *Probability in the Engineering and Informational Sciences*, 1995.

[9] M. Mandjes. Quick simulation of loss probabilities in queueing networks with Markov modulated sources. Research Paper, Free University of Amsterdam, 1995.

[10] A. Ridder. Fast Simulation of Markov Fluid Models. Research Memorandum 1993-21, Free University Amsterdam, 1993.

[11] T.E. Stern and A.I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Advances of Applied Probability*, 23: 105 – 139, 1991.

[12] W. Whitt. Tail Probabilities with Statistical Multiplexing and Effective Bandwidths in Multi-Class Queues. *Telecommunication Systems*, 2: 71 – 107, 1993.