Extension of heuristics for simulating population overflow in Jackson tandem queuing networks to non-Markovian tandem queuing networks

Short paper for RESIM2008

Tatiana S. Zaburnenko, Pieter-Tjerk de Boer, Boudewijn R.H.M. Haverkort

Faculty of Electrical Engineering, Mathematics and Computer Science
Centre for Telematics and Information Technology (CTIT)
University of Twente, Enschede, The Netherlands
Email: t.s.zaburnenko@ewi.utwente.nl, ptdeboer@cs.utwente.nl, brh@cs.utwente.nl

Abstract

In this paper we extend previously proposed state-dependent importance sampling heuristics for simulation of population overflow in Markovian tandem queuing networks to *non-Markovian* tandem networks, and experimentally demonstrate the asymptotic efficiency of the resulting heuristics.

I. Introduction

In recent years, much progress has been made in the estimation of small overflow probabilities in Markovian tandem queueing networks using importance sampling simulation. With importance sampling, the distributions of the model's random variables are modified (called a change of measure) so as to make the event of interest more likely. The challenge then lies in finding a good change of measure. In the present paper, we extend a recently developed change of measure for Markovian tandem queues to non-Markovian tandem queues.

The changes of measure for Markovian tandem networks proposed in [1] and [2] and validated in [4] are state-dependent: the state-dependence involves a *linear* transition between different known state-*in*dependent changes of measure, as a function of the model's state variables (namely, the queue lengths). The changes of measure considered in the present paper inherit this linear transition, but the constituent changes of measure are replaced by the corresponding state-independent changes of measure for *non*-Markovian single queues from [3].

Two ways of calculating the new network parameters according to a linear dependence are suggested. The first one is to linearly vary the distributions' parameters (such as rates and probabilities) themselves; the second way is to linearly change the employed exponential tilting parameters (θ). Combined with the two types of state-dependence (namely partial [1] and full state-dependence [2]), this gives four possible changes of measure that can be used to estimate the overflow probability in non-Markovian tandem queuing networks. In this paper we present only the fully state-dependent changes of measure since these performed better than the ones with partial state dependence (see [4]).

In Section II the example model under consideration is described, together with some notation and background information. In Section III the general idea of our change of measure is described, followed by the exact formulas for the model considered in this paper in Section IV. At the end the experimental results (Section V) and some conclusions (Section VI) are given.

II. Preliminaries

In this section we introduce the non-Markovian tandem network model that will be used as an example; next we give the notation and describe the known results for single queues on which the change of measure for the tandem network will be based.

A. The model: H_2/Bim/1 $\rightarrow$ $\cdot$/Bim/1

The model we consider in this paper is a two-node tandem queuing network with the service times at each node being bi-modally distributed (modeling two different packets lengths, such as data and acknowledgement packets in the internet) and the inter-arrival times being hyper-exponentially distributed (resulting in bursty arrivals, as is also the case in the internet). In a real data network model, the service times at both nodes would be strongly related (because the packet's size does not change); however, for simplicity we assume them to be independent. (This is known as Kleinrock's independence assumption, and has been shown to be a reasonable approximation for networks of moderate connectivity [5].)

B. Service process

The service times in our model are *bi-modally distributed*. Consider a communication link with a transmission speed of ν (bits/sec) and packets having length l_1 with probability p and length l_2 with probability $1-p$; we call them packets of *type 1* and *2*, respectively. Then,

$$\text{service time} = \begin{cases} t_1 = l_1/\nu, & \text{with probability } p, \\ t_2 = l_2/\nu, & \text{with probability } (1-p), \end{cases}$$

and the moment generating function of the service process is

$$M_{serv}(\theta) = e^{\theta t_1} p + e^{\theta t_2}(1-p). \tag{1}$$

An Importance Sampling (IS) change of measure to be used for simulating bi-modally distributed service times changes the probability p, and neither the transmission speed, nor packet lengths can be changed, since the latter two are not random. This also follows from the definition of IS. If in the new system an event is possible such that in the original system it has a zero probability, the corresponding likelihood ratio is equal to zero. Thus, only the events (i.e., packet lengths) that are possible in the original system have a non-zero probability in the new system.

C. Arrival process

The inter-arrival times are *hyper-exponentially distributed* with parameters λ_1, λ_2, i.e., the inter-arrival time is with probability q an exponentially distributed random variable X_1 with mean $1/\lambda_1$ and with probability $(1-q)$ an exponentially distributed random variable X_2 with mean $1/\lambda_2$. Its density function is

$$f(x) = q\lambda_1 e^{-\lambda_1 x} + (1-q)\lambda_2 e^{-\lambda_2 x}, \tag{2}$$

and the moment generating function

$$M_{ar}(\theta) = \frac{q\lambda_1}{\lambda_1 - \theta} + \frac{(1-q)\lambda_2}{\lambda_2 - \theta}. \tag{3}$$

An IS change of measure changes the rates λ_1, λ_2 and the probability q.

D. Optimal change of measure for $GI/GI/1$ queue

As shown in [3] a provably asymptotically efficient change of measure for simulating a single $GI/GI/1$ queue, is *an exponential change of measure* (exponential twist), proposed in [6] and defined as

$$d\tilde{F}(x) = \frac{e^{\theta x} dF(x)}{M(\theta)}, \text{ with } \theta \in \mathbb{R}, \tag{4}$$

where $F(x)$ is the original distribution and $M(\theta) = \mathbb{E}e^{\theta x}$ is the moment generating function. The best change of measure is the one for which parameter $\theta = \theta^*$, with θ^* being a solution of the equation

$$M_{ar}(-\theta) M_{serv}(\theta) = 1, \tag{5}$$

with $M_{ar}(\theta)$ and $M_{serv}(\theta)$ being the moment generating functions, for the arrival and service process, respectively. According to (4), the new inter-arrival time distribution is given by

$$d\tilde{F}_{ar}(x) = \frac{e^{-\theta^* x} dF_{ar}(x)}{M_{ar}(-\theta^*)}, \tag{6}$$

and the new service time distribution is given by

$$d\tilde{F}_{serv}(x) = \frac{e^{\theta^* x} dF_{serv}(x)}{M_{serv}(\theta^*)}. \tag{7}$$

III. The change of measure.

A. General formulation

The proposed state-dependent change of measure is a linear combination of several state-independent changes of measure. We denote each change of measure by a vector called COM, which contains parameters of the probability distributions. The original network parameters (i.e., no change of measure) are denoted by COM_0, while COM_1 and COM_2 denote the state-independent changes of measure that "push" queues 1 and 2, respectively. Then our state-dependent change of measure can be expressed as follows:

$$\mathit{COM}_{x_1,x_2} = \left[\frac{x_2}{b_2}\right]^1 \mathit{COM}_2 + \left[\frac{b_2 - x_2}{b_2}\right]^+ \times \\ \times \left(\left[\frac{x_1}{b_1}\right]^1 \mathit{COM}_1 + \left[\frac{b_1 - x_1}{b_1}\right]^+ \mathit{COM}_0 \right), \tag{8}$$

where b_1 and b_2 are some integer numbers still to be determined, $[a]^+ = \max(a, 0)$, and $[a]^1 = \min(a, 1)$.

Equation (8) expresses the new network parameters in terms of the original ones. Note, however, that the exact representation of COM_i still needs to be defined, since parameters to be changed need to be chosen. There are two possible ways.

The first possibility is to change the arrival and the service processes *directly through their parameters*, i.e., let COM_i represent the new arrival and service parameters for the whole network. The second possibility is to change the arrival and the service processes *indirectly*, i.e., through the *twisting parameter* θ^* and let COM_i represent the new twisting parameters. In more detail these two possibilities are described in the sequel.

In order to express the COM_i, we need to introduce the following notation: θ_i^* denotes the optimal twisting parameter for simulating node i as a single node: the non-zero solution of Equation (5) with $M_{serv}(\theta) = M_{serv_i}(\theta)$ where $M_{serv_i}(\theta)$ is the moment generating function for the service distribution at node i.

B. The change of measure linear in the parameters: COM^p

The first possibility, called *the change of measure linear in the parameters* and denoted by $\mathit{COM}^p_{x_1,x_2}$ is represented in the following steps.

1) Find the twisting parameters θ_i^* for each node i.
2) Use Equations (6)–(7), for each node i to calculate the new inter-arrival and service time distributions.
3) Use the results from the previous step to define COM_i as a vector of the new arrival and service parameters *to "push" node* i as a single node.
4) Define the change of measure for the whole network as a vector of *the new arrival and service parameters* using Equation (8) and COM_i defined in the previous step.

In other words, the change of measure for the whole network is a combination *of changes of measure* for each node.

C. The change of measure linear in θ: COM^{θ}

The second possibility, called *the change of measure linear in* θ and denoted by $COM^{\theta}_{x_1,x_2}$ is defined as follows. In essence, in this approach, we change the order of the steps proposed above by moving step 2 to the last position.

1) Find the twisting parameters θ^*_i for each node i.
2) Define COM_i as a vector of *twisting parameters to "push" node* i as a single node in the network.
3) Define a vector of *twisting parameters for the whole network* using Equation (8)
4) The change of measure for the whole 2-node tandem network is defined from Equations (6)–(7) with the twisting parameter parameter found in the previous step.

In other words, the change of measure for the whole network is defined as exponential twist with parameter θ^* found as a combination *of twisting parameters* for each node.

Note that for Markovian models, the above two possibilities are equivalent, since the new (exponentially twisted) arrival and service rates depend linearly on θ^* as $\tilde{\lambda} = \lambda + \theta^*$ and $\tilde{\mu} = \mu - \theta^*$. Below we present the exact equations of the proposed heuristics for our example model.

IV. Exact calculation of COM for the model: H_2/Bim/1 → ·/Bim/1

Let $\tilde{\lambda}_{i,1}$, $\tilde{\lambda}_{i,2}$, $\tilde{q}_i$, $(1-\tilde{q}_i)$ denote the new arrival rates and probabilities for node i, and $\tilde{p}_i$ denote the new probability for the service process for node i.

A. The change of measure linear in the parameters

For the considered model $\mathrm{COM}^{p}_{x_1,x_2}$ changes the arrival rates λ_1 and λ_2, the arrival probabilities q and $(1-q)$ and the service process probability p_i (for node i). From (6)–(7) the new network parameters can be found as follows

$$\tilde{\lambda}_{i,k} = \lambda_k + \theta^*_i, \tag{9}$$

$$\tilde{q}_i = \frac{q\lambda_1(\lambda_2+\theta^*_i)}{q\lambda_1(\lambda_2+\theta^*_i)+(1-q)\lambda_2(\lambda_1+\theta^*_i)}, \tag{10}$$

$$\tilde{p}_i = \frac{e^{\theta^*_i t_{i,1}}p}{M_{serv_i}(\theta^*_i)}. \tag{11}$$

The changes of measure COM_i for each node i correspond to the new arrival and service processes and are equal to

$$COM_1 = (\tilde{\lambda}_{1,1}, \tilde{\lambda}_{1,2}, \tilde{q}_1, \tilde{p}_1, p), \tag{12}$$

$$COM_2 = (\tilde{\lambda}_{2,1}, \tilde{\lambda}_{2,2}, \tilde{q}_2, p, \tilde{p}_2), \tag{13}$$

$$COM_0 = (\lambda_1, \lambda_2, q, p, p), \tag{14}$$

where COM_0 corresponds to no change of measure. The change of measure $COM^{p}_{x_1,x_2}$ can be written by substituting (12)–(14) into (8).

B. The change of measure linear in θ

$\mathrm{COM}^{\theta}_{x_1,x_2}$ is defined as follows. COM_i denotes the new twisting parameters, i.e.,

$$COM_1 = (-\theta^*_1, \theta^*_1, 0), \tag{15}$$

$$COM_2 = (-\theta^*_2, 0, \theta^*_2), \tag{16}$$

$$COM_0 = (0, 0, 0). \tag{17}$$

Thus, COM_1 twists node 1, change of measure COM_2 twists node 2 and COM_0 corresponds to no change of measure.

The linear in θ change of measure ($COM^{\theta}_{x_1,x_2}$) is defined as the exponentially twisted distribution with parameter θ^* defined as follows

$$\begin{cases} \tilde{\theta}_{ar} = -\left[\frac{x_2}{b_2}\right]^1 \theta^*_2 - \left[\frac{b_2-x_2}{b_2}\right]^+ \left[\frac{x_1}{b_1}\right]^1 \theta^*_1, \\ \tilde{\theta}_{serv_1} = \left[\frac{b_2-x_2}{b_2}\right]^+ \left[\frac{x_1}{b_1}\right]^1 \theta^*_1, \\ \tilde{\theta}_{serv_2} = \left[\frac{x_2}{b_2}\right]^1 \theta^*_2. \end{cases} \tag{18}$$

By substituting (18) into (6)–(7) we obtain the following expressions for the arrival and service processes of the whole network:

$$\begin{cases} \tilde{\lambda}_1 = \lambda_1 + \theta^*_{ar}, \\ \tilde{\lambda}_2 = \lambda_2 + \theta^*_{ar}, \\ \tilde{q} = \frac{q\lambda_1(\lambda_2+\theta^*_{ar})}{q\lambda_1(\lambda_2+\theta^*_{ar})+(1-q)\lambda_2(\lambda_1+\theta^*_{ar})}, \\ \tilde{p}_1 = \frac{e^{\theta^*_{sr}t_{1,1}}p}{M_{serv_1}(\theta^*_{sr})}, \\ \tilde{p}_2 = \frac{e^{\theta^*_{sr}t_{2,1}}p}{M_{serv_2}(\theta^*_{sr})}. \end{cases} \tag{19}$$

V. Experiments

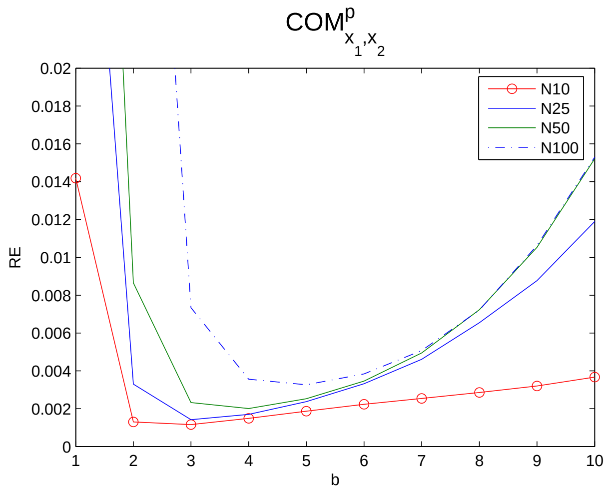
Experimental results for the model considered are given below. The inter-arrival times are hyper-exponentially distributed with mean interarrival time 1/200 s with probability 0.1 and 1/4000 s otherwise. The service times at both nodes correspond to packets having fixed lengths of 400 bits with probability 0.3, and 12000 bits otherwise (drawn independently at each queue), and being transmitted at a speed of $9.01 \cdot 10^7$ bit/s at the first server and $9 \cdot 10^7$ bit/s at the second server. These service rates were chosen almost equal, since it is known from the Markovian case that this is the most difficult parameter setting; they were not chosen precisely equal to avoid implementation complexities arising from events being scheduled at exactly the same time.

The changes of measure derived above still depend on the two parameters b_1 and b_2. We now set $b_1 = b_2 = b$, and experimentally find the value b_{opt} for b that minimizes the variance.

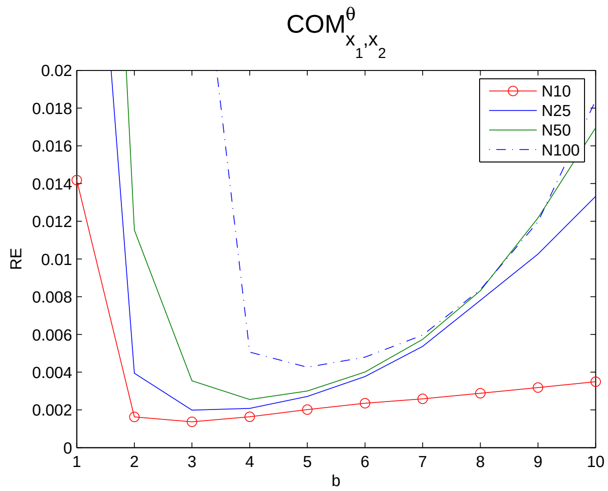
Table I shows estimated overflow probabilities obtained with $16 \cdot 10^5$ replications for the fully state-dependent versions of the heuristics, using $b = b_{opt}$. The relative errors, shown in

N	linear in parameters		linear in θ	
	b	$\tilde{\gamma}(N) \pm RE\%$	b	$\tilde{\gamma}(N) \pm RE\%$
10	3	1.3068e-05 $\pm$ 0.12	3	1.30985e-05 $\pm$ 0.14
25	3	8.1645e-16 $\pm$ 0.14	3	8.11988e-16 $\pm$ 0.20
50	4	3.8084e-33 $\pm$ 0.20	4	3.81196e-33 $\pm$ 0.25
100	5	4.4188e-68 $\pm$ 0.32	5	4.45316e-68 $\pm$ 0.43

Table I
SIMULATION RESULTS SHOWING ASYMPTOTIC EFFICIENCY



(a) linear in parameters



(b) linear in θ

Figure 1. Comparison of the sensitivity to b of the two fully state-dependent changes of measure

percent, are very small, and grow less than linearly with the overflow level N, so the heuristics show to be asymptotically efficient.

Figures 1(a–b) show the dependence of the relative error of an estimator (*RE*) on the parameter b. It is clear from the figures that the heuristics produce very stable results. The *RE* of an estimator is not very sensitive with respect to b for low level $N = 10$, and starts to be more sensitive when level N increases. Note also that *RE* is more sensitive to values of b that are lower than to those that are higher than b_{opt}.

VI. CONCLUSIONS

Although only two models have been checked so far (for the other one, see [4]), the experiments have been done for the network parameter values that are known to be the most difficult for simulation (equivalent service rates). Despite that, the resulting estimates show very good performance. This is a very promising result. More experiments are needed, however, to validate the heuristics for both considered models (i.e., for different parameter settings), and, for other non-Markovian queuing networks (other arrival and service rate distributions, and possibly other topologies).

REFERENCES

[1] T. S. Zaburnenko and V. F. Nicola. State-dependent importance sampling heuristic for simulating rare events in tandem networks. In *The 5th St. Petersburg Workshop on Simulation (SPWS '05)*, pages 755–764, St. Petersburg, Russia, 2005.

[2] V. F. Nicola and T. S. Zaburnenko. Efficient importance sampling heuristics for the simulation of population overflow in Jackson networks. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 17(2), 2007.

[3] J. S. Sadowsky. Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Transaction on Automatic Control*, 36:1383–1394, 1991.

[4] T. S. Zaburnenko. *Efficient heuristics for simulating rare events in queuing networks*. PhD thesis, University of Twente, 2008.

[5] L. Kleinrock. *Queuing Systems. Volume II: Computer Applications*. John Wiley and Sons, 1976.

[6] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control*, 34:54–66, 1989.

[7] P. T. de Boer. Rare-event simulation of non-Markovian queueing networks using a state-dependent change of measure determined using cross-entropy. *Annals of Operations Research*, 134(1):69–100, 2005.