*Proceedings of the 2005 Winter Simulation Conference*
*M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.*

# EFFICIENT IMPORTANCE SAMPLING HEURISTICS FOR THE SIMULATION OF POPULATION OVERFLOW IN JACKSON NETWORKS

Victor F. Nicola
Tatiana S. Zaburnenko

Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente, P.O. Box 217
7500 AE Enschede, THE NETHERLANDS

## ABSTRACT

In this paper we propose state-dependent importance sampling heuristics to estimate the probability of population overflow in Jackson networks with arbitrary routing. These heuristics approximate the "optimal" state-dependent change of measure without the need for costly optimization involved in other recently proposed adaptive algorithms. Experimental results on tandem, feed-forward and feed-back networks with a moderate number of nodes yield asymptotically efficient estimates (often with bounded relative error) where no other state-independent importance sampling techniques are known to be efficient.

## 1 INTRODUCTION

Efficient simulation of queueing networks has long been the focus of much research, owing to its applicability in the modeling, analysis and dimensioning of logistic, production and communication networks. Among the most effective methodologies researched and applied so far are those based on importance sampling (see, e.g., Asmussen and Rubinstein 1995, Heidelberger 1995, Juneja and Nicola 2004, Parekh and Walrand 1989).

Until recently, only state-independent importance sampling heuristics were developed and considered for analysis. In these heuristics, the change of measure is "static" and independent of the network state (i.e., the number of customers at each node in a Jackson network). A relatively simple (and well known) heuristic change of measure for simulations of population overflow in queueing networks is that proposed in Parekh and Walrand (1989) and further investigated in Frater et al. (1991). However, even for the simplest Jackson queueing network (e.g., 2-node tandem network), the effectiveness of this heuristic is limited to only some region of the (arrival and service) parameters space (see Glasserman and Kou 1995, de Boer 2004). (We use the term "effectiveness" interchangeably with "asymptotic efficiency," see Section 2.2 for a precise definition.)

Based on Markov additive process formulation of a two-node tandem network and large deviations arguments, work in Kroese and Nicola (2002) reveals that a state-dependent change of measure is effective where no effective state-independent change of measure exists. Since then, there has been increasingly more research on methodologies to obtain efficient state-dependent importance sampling heuristics. In de Boer and Nicola (2002) an adaptive optimization technique based on the method of cross-entropy (Rubinstein 2002) is used to approximate the "optimal" state-dependent change of measure. A similar adaptive approach based on stochastic approximation is introduced in Ahamed et al. (2004). A drawback of these adaptive approaches, however, is the excessive computational and storage demands for large state-space models associated with large networks.

In Zaburnenko and Nicola (2005) and Nicola and Zaburnenko (2005), the aim is to develop a (heuristic) state-dependent change of measure which is sufficiently close to the "optimal" without the need for a costly optimization. The key observation is that the "optimal" change of measure depends on the network state only along and close to the boundaries (when one or more nodes are empty), and tends to become state-independent in the interior of the state-space. Therefore, if we can determine the change of measure along the boundaries and at the interior of the state-space, then we may be able to combine them appropriately to construct a state-dependent change of measure that approximates the "optimal" one in the entire state-space. The proposed methodology is dubbed "state-dependent heuristic" or SDH in short. Experimental results with the so obtained heuristic for tandem networks with multiple nodes yield estimates with a bounded relative error (see Zaburnenko and Nicola 2005, Nicola and Zaburnenko 2005).

In this paper we propose extensions and generalizations of the heuristics in Nicola and Zaburnenko (2005) to efficiently simulate Jackson networks with a moderate

number of nodes and arbitrary routing. Experimental results to estimate the probability of population overflow in tandem, feed-forward and feed-back networks (with up to 4 nodes) produce asymptotically efficient estimates, often with bounded relative error. The proposed heuristics are effective, yet easier to implement and could be more efficient than those based on adaptive methodologies (e.g., de Boer and Nicola 2002), particularly for large networks.

In Section 2 we give some preliminaries, introduce the basic model and define the probability of interest. The importance sampling technique is briefly reviewed. Also, the change of measure to simulate buffer overflow at an arbitrary network node is outlined, as it plays a key role in our heuristics. In Section 3 we motivate the proposed SDH and give its formal representation for tandem, feed-forward and feed-back networks, respectively. In Section 4 we present experimental results and comparisons with other known methods to estimate the probability of population overflow in some example networks. We conclude in Section 5.

## 2 PRELIMINARIES

The queueing network model and associated notation are introduced in Section 2.1. A brief review of importance sampling and some properties of simulation estimators are provided in Section 2.2. A change of measure to simulate overflow at a network node is introduced in Section 2.3.

### 2.1 Model and Notation

Consider a Jackson network consisting of $n$ nodes (queues), each having its own buffer of infinite size. Customers arrive at node $i$ ($1 \leqslant i \leqslant n$) according to a Poisson process with rate $\lambda_i$. The service time of a customer at node $i$ is exponentially distributed with rate $\mu_i$ ($1 \leqslant i \leqslant n$). Customers that leave node $i$ join node $j$ with probability $p_{ij}$ ($1 \leqslant i, j \leqslant n$) or leave the network with probability $p_{ie}$ ($1 \leqslant i \leqslant n$). We also assume that the queueing network is stable, i.e., $\gamma_i < \mu_i$ for all $1 \leqslant i \leqslant n$, where $\gamma_i$ is the total arrival rate at node $i$, as determined from the traffic equations

$$\gamma_i = \lambda_i + \sum_{\forall j} \gamma_j \; p_{ji} \; .$$

Let $X_{i,t}$ ($1 \leqslant i \leqslant n$) denote the number of customers at node $i$ at time $t \geqslant 0$ (including those in service). Then the vector $\mathbf{X}_t = (X_{1,t}, X_{2,t}, ..., X_{n,t})$ is a Markov process representing the state of the network at time $t$. Denote by $S_t$ the total number of customers in the network (network population) at time $t$, i.e., $S_t = \sum_{i=1}^{n} X_{i,t}$.

Assuming that the initial network state is $\mathbf{X}_0$ (usually, $\mathbf{X}_0 = (0, 0, ..., 0)$ corresponding to an empty network), we are interested in the probability that the network population reaches some high level $L \in \mathbb{N}$ before becoming empty. We denote this probability by $\gamma(L)$ and refer to it as the *population overflow probability*, starting from the initial state $\mathbf{X}_0$. Since the associated event is typically rare, importance sampling may be used to efficiently estimate this probability.

### 2.2 Importance Sampling

Importance sampling involves simulating the system under different underlying probability distributions so as to increase the frequency of typical sample paths leading to the rare event. Formally, let $w$ be a sample path over the interval $[0, t]$. Then, the likelihood ratio associated with $w$ is given by $W_t(w) = \frac{P(w)}{\tilde{P}(w)}$, where $P(w)$ and $\tilde{P}(w)$ are the probabilities (or likelihoods) of sample path $w$ under the original and the new measure, respectively. Obviously, $\tilde{P}(w) > 0$ whenever $P(w) > 0$. Starting from $\mathbf{X}_0$, define $\tau$ as the first time $S_t$ hits level $L$ or level 0, then

$$\gamma(L) \; = \; \mathbb{E}\, I_{\{S_\tau = L\}} \; = \; \tilde{\mathbb{E}}\, W_\tau \, I_{\{S_\tau = L\}} \, , \tag{1}$$

where $W_\tau$ is the likelihood ratio over the interval $[0, \tau]$; $\mathbb{E}$ and $\tilde{\mathbb{E}}$ are the expectations under the original and the new change of measures, respectively. The variance of the estimator $\tilde{\mathbb{E}}\, W_\tau \, I_{\{S_\tau = L\}}$ is given by

$$\tilde{\mathbb{E}}\, {W_\tau}^2 \, I_{\{S_\tau = L\}} \; - \; (\gamma(L))^2 \, . \tag{2}$$

The relative error is the ratio of the standard deviation of the estimator over its expectation, i.e.,

$$\sqrt{\frac{\tilde{\mathbb{E}}\, {W_\tau}^2 \, I_{\{S_\tau = L\}}}{\gamma(L)^2} \; - \; 1} \, . \tag{3}$$

The estimator $\tilde{\mathbb{E}}\, W_\tau \, I_{\{S_\tau = L\}}$ is said to be *asymptotically efficient* if its relative error grows at sub-exponential (e.g., polynomial) rate as $L \to \infty$ (i.e., as $\gamma(L) \to 0$). Formally, let $\lim_{L\to\infty} \frac{1}{L} \log \gamma(L) = \theta$. That is, $\theta$ is the asymptotic decay rate of the overflow probability $\gamma(L)$ as $L \to \infty$. Then, asymptotic efficiency is obtained if

$$\lim_{L\to\infty} \frac{1}{L} \log \tilde{\mathbb{E}}\, {W_\tau}^2 \, I_{\{S_\tau = L\}} \; = \; 2\theta \, . \tag{4}$$

The estimator is said to have *bounded relative error* if its relative error is bounded in $L$ as $\gamma(L) \to 0$.

It is important to note that a change of measure may, in general, depend on the state of the system, even if the original underlying distributions do not depend on the system state. For instance, the arrival and service rates in a Markovian queueing network are typically fixed and independent of the network state (i.e., the buffer content at each node). However, a change of measure to be used in importance sampling simulation may involve new arrival and service rates that depend on the state of the network. State-dependent change of measures are generally more effective

in simulations of rare events in queueing networks (see, e.g., Kroese and Nicola 2002, de Boer and Nicola 2002). Therefore, in this paper (as in Zaburnenko and Nicola 2005) we aim at developing heuristics to approximate the "optimal" state-dependent change of measure.

### 2.3 Buffer Overflow at a Network Node

The "optimal" change of measure to simulate the build-up at any specific network node plays a key role in the heuristics proposed in this paper. In this section we give a brief discussion and characterize such a change of measure. Some more notation is necessary. Consider a Jackson network as described in Section 2.1 and let all nodes in the network be indexed by the set $\mathcal{H}$. These nodes are further categorized by one "target" node indexed by $t$ and the remaining "feeder" nodes indexed by the set $\mathcal{F}$. Thus, $\mathcal{H} \equiv \{t\} \cup \mathcal{F}$. In Juneja and Nicola (2004) a state-independent change of measure is proposed to estimate the probability that the buffer content at the target node exceeds a large level during its busy period (a busy period of the target queue is initiated when an arrival to it finds it empty, and ends when the target queue subsequently re-empties). Under this change of measure, the simulated queueing network is again a Jackson network in which the original inter-arrival and service time distributions are exponentially twisted so as to achieve asymptotic efficiency. Moreover, only the target node $t$ becomes unstable while each of the other (feeder) nodes is either critical (in the set $\mathcal{C} \subseteq \mathcal{F}$) or stable (in the set $\mathcal{S} = \mathcal{F} - \mathcal{C}$). Let $\tilde{\lambda}_i$, $\tilde{\mu}_i$, and $\tilde{p}_{ij}$ $(1 \leqslant i, j \leqslant n)$ be the new external arrival rates, service rates, and routing probabilities, respectively. Also, define the constants $\tilde{c}_i \geqslant 1$ for $i \in \mathcal{H}$, and let $\mathcal{D} \subset \mathcal{H}$ denote the set $\{i : \lambda_i = 0\}$. The change of measure in Juneja and Nicola (2004) is characterized as follows:

- The new external arrival rates are given by $\tilde{\lambda}_i = \tilde{c}_i \lambda_i$, for $i \in \mathcal{H}$. (Thus, $\tilde{\lambda}_i = 0$, for $i \in \mathcal{D}$).
- The new routing probabilities are given by $\tilde{p}_{ij} = \frac{\tilde{c}_j}{\tilde{c}_i} \frac{\mu_i}{\tilde{\mu}_i} p_{ij}$ and $\tilde{p}_{ie} = \frac{1}{\tilde{c}_i} \frac{\mu_i}{\tilde{\mu}_i} p_{ie}$, for all $i, j \in \mathcal{H}$.
- The new service rates $\tilde{\mu}_i$, along with the unknown constants $\tilde{c}_i$ $(1 \leqslant i \leqslant n)$ are determined from the non-linear program (NLP) given below.
- Maximize $\tilde{c}_t$ subject to the following constraints:

$$\sum_{i \in \mathcal{H}-\mathcal{D}} \tilde{\lambda}_i + \sum_{i \in \mathcal{H}} \tilde{\mu}_i = \sum_{i \in \mathcal{H}-\mathcal{D}} \lambda_i + \sum_{i \in \mathcal{H}} \mu_i. \quad (5)$$

  For all $i \in \mathcal{H}$, the new routing probabilities and the new total arrival rate $\tilde{\gamma}_i$ must satisfy

$$\sum_{j \in \mathcal{H}} \tilde{p}_{ij} + \tilde{p}_{ie} = 1, \quad (6)$$

  and

$$\tilde{\gamma}_i = \tilde{\lambda}_i + \sum_{j \in \mathcal{F}} \tilde{p}_{ji} \tilde{\gamma}_j + \tilde{\mu}_t \tilde{p}_{ti}. \quad (7)$$

  For all $i \in \mathcal{F}$, a feeder node $i$ is either stable (i.e., $\tilde{\mu}_i = \mu_i > \tilde{\gamma}_i$, $i \in \mathcal{S}$) or critical (i.e., $\tilde{\mu}_i = \tilde{\gamma}_i$, $i \in \mathcal{C}$). The target node is unstable (i.e., $\tilde{\mu}_t < \mu_t$).

Assuming that the queue lengths at the feeder nodes are initially bounded, the change of measure characterized above is asymptotically efficient for estimating the probability of overflow in the target node $t$. In the sequel of this paper we refer to it as the $\mathbf{JN}_t$ change of measure, where $t$ is the index of the target node.

**Remark 1.** When the service rates at the feeder nodes are sufficiently large (for example, when the target node is the bottleneck), this change of measure can be determined explicitly and is identical to that proposed in Parekh and Walrand (1989) and in Frater et al. (1991) to simulate network population overflow.

Formally, denote by $P = (p_{ij} : i, j \in \mathcal{H})$ the matrix with the routing probabilities, and let $R = (r_{ij} : i, j \in \mathcal{H})$ equals $(I - P)^{-1}$. Since the network is stable, $r_{ij}$ is the expected number of visits to queue $j$ by a customer starting from queue $i$, before it leaves the system. Note that $r_{it} \leqslant r_{tt}$. If, for each $i \in \mathcal{F}$, the service rates at the feeder nodes satisfy the inequality

$$\mu_i > \gamma_i (1 + \frac{r_{it}}{r_{tt}} (\frac{\mu_t}{\gamma_t} - 1)), \quad (8)$$

then, the change of measure in Juneja and Nicola (2004) is determined explicitly as follows:

- All feeder nodes are stable (i.e., the set $\mathcal{C}$ is empty) and $\tilde{\mu}_i = \mu_i$ for $i \in \mathcal{F}$. The target node is unstable and $\tilde{\mu}_t = \frac{(r_{tt}-1)\mu_t + \gamma_t}{r_{tt}}$.
- For each $i \in \mathcal{H}$, $c_i = (1 + \frac{r_{it}}{r_{tt}} (\frac{\mu_t}{\gamma_t} - 1))$.
- For each $i \in \mathcal{H}$, $\tilde{\lambda}_i = c_i \lambda_i$ and $\tilde{\gamma}_i = c_i \gamma_i$.
- For $i, j \in \mathcal{H}$, $\tilde{p}_{ij} = \frac{c_j}{c_i} \frac{\mu_i}{\tilde{\mu}_i} p_{ij}$ and $\tilde{p}_{ie} = \frac{1}{c_i} \frac{\mu_i}{\tilde{\mu}_i} p_{ie}$.

While being asymptotically efficient to simulate overflow at the bottleneck node (as proved in Juneja and Nicola 2004), the above change of measure is not always asymptotically efficient to simulate network population overflow (as shown in Glasserman and Kou 1995 and in de Boer 2004).

## 3 STATE-DEPENDENT HEURISTICS

Theoretical and empirical results in Kroese and Nicola (2002) and de Boer and Nicola (2002) indicate that the "optimal" change of measure depends on the network state, i.e., the number of customers at the network nodes. Fur-

thermore, this dependence is strong along the boundaries of the state-space (i.e., when one or more buffers are empty) and diminishes in the interior of the state-space (i.e., when contents of all buffers are sufficiently large). Dependencies along the boundaries have shown to be very crucial for the asymptotic efficiency (or "optimality") of the change of measure.

The above observation suggests that if we know the "optimal" change of measure along the boundaries and in the interior of the state-space, then we might be able to construct a change of measure that approximates the "optimal" one over the entire state-space. In Nicola and Zaburnenko (2005), heuristics based on combining known large deviations results and time-reversal arguments are used to construct such a change of measure for the 2-node tandem network. Empirical results there has shown that it produces asymptotically efficient estimators with a bounded relative error for all feasible network parameters. In this section we propose heuristic state-dependent changes of measure for the efficient simulation of Jackson networks with more general topologies. The change of measure in Section 3.1 is a generalization of that in Nicola and Zaburnenko (2005) to tandem networks with any number of nodes. In Sections 3.2 and 3.3 we give heuristic changes of measures for the efficient simulation of feed-forward and feed-back networks, respectively.

### 3.1 SDH for the $n$-node Tandem Network

Let $\lambda$ and $\mu_i$ $(i = 1, \ldots, n)$ be the arrival rate at the first node and the service rate at the $i-th$ node, respectively. Denote by $\rho_i = \frac{\lambda}{\mu_i}$ the traffic intensity at node $i$, and assume that $\rho_1 \leqslant \rho_2 \leqslant \ldots \rho_n < 1$. We note, however, that this ordering is not a restriction, since the probability of population overflow is invariant with respect to the placement order of nodes in a Jackson tandem network (Weber 1979). Without loss of generality we assume that $\lambda + \sum_{i=1}^{n} \mu_i = 1$. Denote by $\tilde{\lambda}, \tilde{\mu}_i$ the corresponding rates under the new change of measure, and by $\mathbf{SDH}_n$ the $(n+1) \times (n+1)$ SDH transformation matrix for the $n$-node tandem network. Thus, $\mathbf{SDH}_n$ is a linear operator transforming the original rates into the new rates. Define $[a]^+ = \max(a, 0)$ and $[a]^1 = \min(a, 1)$, then for $n = 2$, the change of measure in Nicola and Zaburnenko (2005) may be expressed as:

$$\begin{bmatrix} \tilde{\lambda} \\ \tilde{\mu}_1 \\ \tilde{\mu}_2 \end{bmatrix} = \mathbf{SDH}_2 \begin{bmatrix} \lambda \\ \mu_1 \\ \mu_2 \end{bmatrix}, \tag{9}$$

$$\begin{aligned} \mathbf{SDH}_2 = & \left[\frac{b - x_2}{b}\right]^+ \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ & + \left[\frac{x_2}{b}\right]^1 \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}. \end{aligned} \tag{10}$$

The first matrix is the identity matrix with the first and the second rows interchanged; this corresponds to interchanging the arrival rate $\lambda$ with the service rate $\mu_1$. The second matrix is the identity matrix with the first and the third rows interchanged; this corresponds to interchanging the arrival rate $\lambda$ with the service rate $\mu_2$. The above heuristic can be generalized for an $n$-node tandem network. Let $\Theta$ be a vector with the original network parameters, i.e., $\Theta^T = [\lambda, \mu_1, \ldots, \mu_n]$. Similarly, $\tilde{\Theta}$ is a vector with the new network parameters. Define the transformation matrix $\mathbf{SDH}_n$ recursively as follows:

$$\mathbf{SDH}_k = \left[\frac{b - x_k}{b}\right]^+ \mathbf{SDH}_{k-1} + \left[\frac{x_k}{b}\right]^1 \mathbf{I}_k, \; k = 2, \ldots, n. \tag{11}$$

$\mathbf{I}_k$ is the identity matrix of dimension $(n+1)$ with the first and the $(k+1)$-st rows interchanged. Then the SDH for an $n$-node tandem network is given by

$$\tilde{\Theta} = \mathbf{SDH}_n \; \Theta.$$

Note that for $n = 1$ (a single queue), SDH corresponds to the well known heuristic of interchanging the arrival rate $\lambda$ and the service rate $\mu_1$ (Parekh and Walrand 1989). From Equation 11, the transformation matrix for $n = 3$ is given by:

$$\begin{aligned} \mathbf{SDH}_3 = & \left[\frac{b - x_3}{b}\right]^+ \left( \left[\frac{b - x_2}{b}\right]^+ \mathbf{I}_1 + \left[\frac{x_2}{b}\right]^1 \mathbf{I}_2 \right) \\ & + \left[\frac{x_3}{b}\right]^1 \mathbf{I}_3 \,. \end{aligned} \tag{12}$$

Here the first matrix ($\mathbf{I}_1$) corresponds to interchanging $\lambda$ and $\mu_1$ ($\lambda \leftrightarrow \mu_1$), the second matrix ($\mathbf{I}_2$) corresponds to interchanging $\lambda$ and $\mu_2$ ($\lambda \leftrightarrow \mu_2$), and the third matrix ($\mathbf{I}_3$) corresponds to interchanging $\lambda$ and $\mu_3$ ($\lambda \leftrightarrow \mu_3$). Initially, the network is empty and we start by interchanging the arrival rate $\lambda$ and $\mu_1$, i.e., overloading the first node. As soon as a customer arrives at node 2, we also start overloading the second node by gradually increasing the weight of matrix $\mathbf{I}_2$ and reducing the weight of matrix $\mathbf{I}_1$. When the number of customers at node 2 is sufficiently large ($x_2 \geqslant b$), the weight of matrix $\mathbf{I}_1$ becomes 0. In the meantime, as soon as customers start to arrive at node 3, we start overloading the third node and gradually increase the weight of matrix

$\mathbf{I}_3$ and reduce the weights of matrices $\mathbf{I}_1$ and $\mathbf{I}_2$. When the number of customers at node 3 is sufficiently large ($x_3 \geqslant b$), the weights of matrices $\mathbf{I}_1$ and $\mathbf{I}_2$ become 0.

**Remark 2.** Note that $b$ is the number of boundary levels for which the change of measure depends on the network state (we also refer to it as the dependence range). It is the only variable parameter in the above heuristic, and its proper selection is crucial for achieving asymptotic efficiency. In general, the best value of $b$ (yielding estimates with lowest variance) may depend on the set of network parameters as well as the overflow level $L$. Empirical results indicate that for some regions in the parameter space, the best $b$ is quite robust and does not change with the level $L$. For other regions, the best $b$ may vary slightly from one parameter point to another, and may also depend on $L$.

### 3.2 SDH for a Feed-Forward Network

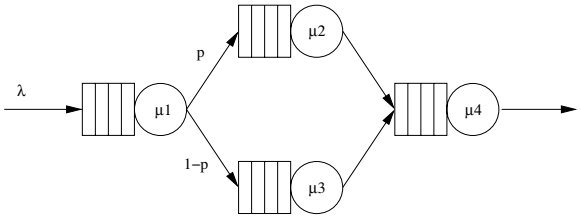


Figure 1: 4-Node Feed-Forward Network

To describe our state-dependent heuristic for feed-forward Jackson networks, we use the specific example depicted in Figure 1.

Let $\Theta^T = [\lambda, \mu_1, \mu_2, \mu_3, \mu_4, p]$ be a vector with the original network parameters. Without loss of generality we assume that $\lambda + \sum_{i=1}^{4} \mu_i = 1$. The traffic intensity at node $i$ is $\rho_i = \frac{\gamma_i}{\mu_i}$, where $\gamma_i$ is the total arrival rate at node $i$ ($i = 1, 2, 3, 4$). We also assume that $\rho_1 \leqslant \rho_2 \leqslant \rho_3 \leqslant \rho_4$. Denote by $\tilde{\Theta}_{JN_i}$ the "optimal" change of measure to simulate buffer overflow at node $i$ (as given in Section 2.3). And, let $\tilde{\Theta}^T = \left[\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3, \tilde{\mu}_4, \tilde{p}\right]$ be a vector with the corresponding network parameters under the new change of measure to simulate network population overflow. Then, for the feed-forward network in Figure 1

$$\begin{aligned}\tilde{\Theta} = &\left[\frac{x_4}{b}\right]^1 \tilde{\Theta}_{JN_4} \\ &+ \left[\frac{b-x_4}{b}\right]^+ \left(\left[\frac{x_3}{b}\right]^1 \tilde{\Theta}_{JN_3}\right. \\ &+ \left[\frac{b-x_3}{b}\right]^+ \left(\left[\frac{x_2}{b}\right]^1 \tilde{\Theta}_{JN_2}\right. \\ &\left.\left. + \left[\frac{b-x_2}{b}\right]^+ \tilde{\Theta}_{JN_1}\right)\right).\end{aligned} \tag{13}$$

In the above, the nesting of $\tilde{\Theta}_{JN_i}$s is in the same order as the traffic intensities at the corresponding nodes, with node 4 at the highest level. That is, dependence on $x_4$ supersedes dependence on $x_3$ which supersedes denpendence on $x_2$, and so on.

As for the tandem network, the above change of measure depends on the number of customers at all nodes (except the first). And, again, the choice of the variable $b$ (dependence range) is crucial for the effectiveness of the heuristic, particularly for networks with a large number of nodes.

### 3.3 SDH for a Feed-Back Network

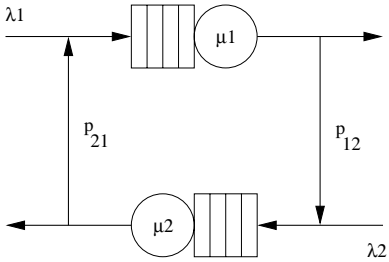


Figure 2: 2-Node Feed-Back Network

To describe our state-dependent heuristic for feed-back Jackson networks, we use the specific example depicted in Figure 2 (similar feed-back network is considered in Randhawa and Juneja 2004).

Let $\Theta$ be a vector with the original network parameters, i.e., $\Theta^T = [\lambda_1, \lambda_2, \mu_1, \mu_2, p_{12}, p_{21}]$. Without loss of generality we assume that $\sum_{i=1}^{2} \lambda_i + \mu_i = 1$. We also assume that $\rho_1 \leqslant \rho_2$, where $\rho_i = \frac{\gamma_i}{\mu_i}$, $i = 1, 2$. Let $\tilde{\Theta}^T = \left[\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{p}_{12}, \tilde{p}_{21}\right]$ be a vector with the corresponding network parameters under the new change of measure to simulate network population overflow. Then, for the feed-back network in Figure 2

$$\tilde{\Theta} = \left[\frac{x_2}{b}\right]^1 \tilde{\Theta}_{JN_2} \left[\frac{b-x_2}{b}\right]^+ \tilde{\Theta}_{JN_1}, \tag{14}$$

where $\tilde{\Theta}_{JN_i}$ is the change of measure (described in Section 2.3) to simulate buffer overflow at node $i$ ($i = 1, 2$). In the above, nodes are nested in the order of their traffic intensities, with node 2 at the highest level. Accordingly, for this particular network, the above heuristic depends only on $x_2$ (the content of the bottleneck node). Descriptively: with the network initially empty, we start by overloading node 1. As the number of customers at node 2 ($x_2$) increases, we gradually and proportionately increase overloading node 2 and decrease overloading node 1. When the number of customers at node 2 is sufficiently large ($x_2 \geqslant b$), only

node 2 is overloaded. Here too, the choice of the variable $b$ is crucial and must be set appropriately.

## 4 EXPERIMENTAL RESULTS

Importance sampling to estimate the probability of population overflow ($\gamma(L)$) involves generating, say, $N$, independent and identically distributed (i.i.d.) busy cycles (i.e., starting with an empty network). Starting a cycle at time 0, define $\tau_L$ as the instant when the network population reaches level $L$ for the first time. Similarly, define $\tau_0$ as the instant when the network population returns to 0 for the first time. The indicator function $I_i(\tau_L < \tau_0)$ takes the value 1 if the population overflow (level $L$) is reached in cycle $i$, otherwise it takes the value 0.

In each cycle, the change of measure is applied until either the population overflow event is reached or the network population returns to 0. Let $W_i$ be the likelihood ratio associated with cycle $i$, then an unbiased estimator $\tilde{\gamma}$ of $\gamma(L)$ is given by

$$\tilde{\gamma} = \frac{1}{N} \sum_{i=1}^{i=N} I_i \, W_i \,, \tag{15}$$

The second moment of $I\,W$ is estimated by

$$\tilde{\gamma^2} = \frac{1}{N-1} \sum_{i=1}^{i=N} I_i \, {W_i}^2 \,. \tag{16}$$

The variance and the relative error of the importance sampling estimator $\tilde{\gamma}$ are given by $VAR(\tilde{\gamma}) = (\tilde{\gamma^2} - (\tilde{\gamma})^2) \,/\, N$ and $RE(\tilde{\gamma}) = \sqrt{VAR(\tilde{\gamma})} \,/\, \tilde{\gamma}$, respectively. Another useful measure for comparing the efficiency of different estimators is the "relative time variance" ($RTV$) product, which is defined as the simulation time (in seconds) multiplied by the squared relative error of the estimator. As the estimate becomes more stable, its $RTV$ tends to a constant value, which is smaller for a more efficient estimator. For example, if $RTV_2$ (for Estimator 2) is larger than $RTV_1$ (for Estimator 1), then it will take Estimator 2 a longer simulation time to reach the same accuracy. For efficiency comparisons we use the variance reduction ratio, $VRR = RTV_2 \,/\, RTV_1$, which represents the efficiency gain when using Estimator 1 relative to that when using Estimator 2.

In the following sections, three sets of experiments (for a tandem, a feed-forward and a feed-back networks) are presented. Each set consists of two experiments corresponding to two points of feasible network parameters. In order to illustrate the utility of our approach, all points are chosen in the region where the well-known heuristic in Parekh and Walrand (1989) is shown to be ineffective. In all simulation experiments, the same number of replications, namely, $10^6$, is used to obtain estimates of the population overflow probability $\gamma(L)$. For each estimate in these tables, we include the relative error (in percentage). For the purpose of comparing the heuristics in this paper (termed SDH) and the adaptive methodology (termed SDA) in de Boer and Nicola (2002), we also include $VRR$ (relative to SDA). Hence, $VRR > 1$ implies efficiency gain of SDH over SDA. Estimates obtained using the well known heuristic in Parekh and Walrand (1989) (termed PW) are also presented, although these are not necessarily accurate or stable. In general, numerical results are difficult to obtain for larger and/or higher overflow levels (i.e., for larger sate-space). Whenever feasible, numerical results (for example, using the algorithm outlined in de Boer 2000) are included to verify the correctness of the simulation estimates. Otherwise, the corresponding table entry is marked with a "$*$". In these cases, agreement of different estimators may be an indication of correctness.

### 4.1 Simulation of Tandem Networks

The experiments in this section are designed to demonstrate that the state-dependent change of measure proposed in Section 3.1 always yield asymptotically efficient estimates (mostly with bounded relative error), also in those regions where no state-independent change of measure is known to be asymptotically efficient. Similar to SDH, SDA assumes state-dependence only over a (small) number of boundary layers (say, $b$) which must be properly determined to ensure the effectiveness and efficiency of these methods. Too small $b$ may not capture crucial dependencies close to the boundaries. Too large $b$ may render SDH ineffective, but it will only reduce the efficiency of SDA. In either SDH or SDA, the "optimal" $b$ which maximizes the efficiency (minimizes the $RTV$) may be determined by repeating the simulation for increasing $b$ starting with, say, $b = 0$ (i.e., no state-dependence). Experimental results with SDH and SDA presented in this section are obtained using the corresponding "optimal" $b$.

For the 2-node tandem network, it is proven or shown empirically (Glasserman and Kou 1995, de Boer 2004) that the state-independent heuristic (PW) in Parekh and Walrand (1989) yields estimates with bounded relative error only in some (non-contiguous) regions of the feasible parameter space. (The feasible parameter space is that corresponding to stable networks.) Thus, the feasible parameter space may be divided into 2 regions, depending on the asymptotic properties of the PW estimator (see de Boer 2004):

- **BRE** Region - PW is asymptotically efficient (with bounded relative error).
- **NAE** Region - PW is not asymptotically efficient.

Empirical studies seem to confirm that the above division of the feasible parameter space holds also for tandem

networks with any number of nodes (i.e., for any feasible set of network parameters, PW is either BRE or NAE). For the $n$-node tandem network, in Glasserman and Kou (1995) sufficient conditions are given for the asymptotic (and non-asymptotic) efficiency of the PW heuristic. These conditions are rather strong and do not cover the entire parameter space (i.e., not all feasible parameter points may be determined as BRE or NAE).

We experiment with two tandem networks having 3 and 4 nodes, respectively. Network parameters are chosen in the NAE Region, with $\lambda = 0.04$ and equal service rates at all nodes; $\mu = 0.32$ for the 3-node tandem network and $\mu = 0.24$ for the 4-node tandem network. (Typically, it is most difficult to efficiently estimate the probability of overflow when the service rates are equal.)

Experimental results in Tables 1 and 2 show that unlike PW, SDH (as described in Section 3.1) yields correct and asymptotically efficient estimates with bounded relative error. To converge properly, our basic (non-optimized) implementation of SDA may require many iterations, each with a large number of cycles (i.e., long simulation time). On the other hand, if and when it converges, it gives very small relative error. (For more on SDA and its implementation details see de Boer and Nicola 2002.) For the examples presented here, SDH typically requires only a few minutes to achieve relative errors less than 1%, and could be more efficient than SDA ($VRR > 1$) even though its relative error may be higher. See Zaburnenko and Nicola (2005) for more comprehensive experimentations with tandem networks.

### 4.2 Simulation of a Feed-Forward Network

In this section we present experimental results performed on the feed-forward network depicted in Figure 1. In our example, we consider two sets of feasible network parameters in the NAE Region (this is verified empirically). In the first set, $\lambda = 0.0455, \mu_1 = 0.7272, \mu_2 = 0.0455, \mu_3 = \mu_4 = 0.0909, p = 0.1$. In the second set, $\lambda = 0.064, \mu_1 = 0.564, \mu_2 = 0.039, \mu_3 = 0.192, \mu_4 = 0.141, p = 0.1$. For compatibility with the heuristic change of measure presented in Section 3.3, nodes are indexed according to their traffic intensities, such that $\rho_1 \leqslant \rho_2 \leqslant \rho_3 \leqslant \rho_4$.

Experimental results in Tables 3 and 4 show that for the considered parameter sets, SDA seems to work very well and yields stable estimates with small relative errors (although correctness could not be verified numerically for the displayed values of $L$). Estimates using SDH (as described in Section 3.2) seem to agree with those of SDA, however, with larger and less stable relative errors. Estimates using PW also seem to agree with those of SDA, but the relative errors are even larger and less stable than those of SDH.

The heuristic in Section 3.2 is particularly sensitive to the dependence range $b$ which is conveniently chosen to be the same at different nodes. This is not necessarily optimal, and more robust performance may be achieved by allowing different values of $b$ at at different nodes. It seems to us that further tuning and/or refinement of the heuristic are needed.

### 4.3 Simulation of a Feed-Back Network

In this section we present experimental results performed on the feed-back network depicted in Figure 2. For the same network Randhawa and Juneja (2004) identify some region in the parameter space in which the PW heuristic is provably not efficient (i.e., they identify only a subset of the NAE region). In our example, we consider two sets of feasible network parameters in the NAE Region. In the first set, $\lambda_1 = 0.01, \mu_1 = 0.13, \lambda_2 = 0.09, \mu_2 = 0.77, p_{12} = 0.9, p_{21} = 0.05$. In the second set, $\lambda_1 = 0.01, \mu_1 = 0.14, \lambda_2 = 0.25, \mu_2 = 0.55, p_{12} = 0.9, p_{21} = 0.05$. For compatibility with the heuristic change of measure in Section 3.3, nodes are indexed such that $\rho_1 \leqslant \rho_2$.

Experimental results in Tables 5 and 6 show that while PW gives wrong estimates, SDH (as described in Section 3.3) yields correct and asymptotically efficient estimates with relative error less than 1% (simulations with a larger number of samples suggest bounded relative error). Moreover, SDH compares well with SDA and seems to yield some efficiency gains for the second set of parameters (Table 6). Again, we must note that neither SDA or SDH implementation is optimized. Therefore, the relative efficiency gains indicated in the presented tables may not be conclusive.

The example feed-back network considered above is relatively small, yet it helps to illustrate that our approach may indeed be useful where no other existing methods are known to be effective. Development of (and experimentation with) similar heuristics for larger feed-back networks is currently underway.

## 5 CONCLUSIONS AND FURTHER WORK

In this paper we have proposed and experimented with a heuristic approach to approximate the “optimal" state-dependent change of measure for the efficient simulation of Jackson queueing networks. The developed changes of measure (which we refer to as SDH) are used to estimate (using importance sampling) the probability of population overflow in tandem, feed-forward and feed-back networks.

Experimental results indicate that the heuristics yield asymptotically efficient estimates, often with bounded relative error. The efficiency of the obtained change of measure compares well with those determined using adaptive methodologies. Moreover, our approach does not require costly pre-computation, and its effectiveness is not diminished for networks with larger state-space.

Table 1: 3-Node Tandem Network ($\lambda = 0.04$, $\mu_1 = \mu_2 = \mu_3 = 0.32$)

| **L** | **Numerical** | **PW** | **SDA** | | **SDH** | | |
|---|---|---|---|---|---|---|---|
| | $\gamma(L)$ | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | $VRR$ |
| 25 | 5.9531e-020 | 4.1433e-020 ± 6.95 | 3 | 5.9625e-020 ± 0.21 | 4 | 5.9491e-020 ± 0.05 | 19.1 |
| 50 | 6.2176e-042 | 2.2129e-042 ± 44.1 | 3 | 6.2260e-042 ± 0.09 | 4 | 6.2264e-042 ± 0.06 | 2.20 |
| 100 | * | 9.6025e-088 ± 12.9 | 3 | 1.7254e-086 ± 0.12 | 5 | 1.7268e-086 ± 0.11 | 1.38 |

Table 2: 4-Node Tandem Network ($\lambda = 0.04$, $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0.24$)

| **L** | **Numerical** | **PW** | **SDA** | | **SDH** | | |
|---|---|---|---|---|---|---|---|
| | $\gamma(L)$ | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | $VRR$ |
| 25 | 5.0207e-016 | 3.7499e-016 ± 11.2 | 3 | 5.0222e-016 ± 0.09 | 4 | 5.0084e-016 ± 0.15 | 0.38 |
| 50 | * | 2.9876e-035 ± 38.2 | 4 | 1.3111e-034 ± 0.13 | 4 | 1.3548e-034 ± 0.11 | 1.58 |
| 100 | * | 6.9845e-074 ± 59.5 | 3 | 1.3020e-072 ± 0.24 | 5 | 1.3076e-072 ± 0.11 | 0.82 |

Table 3: 4-Node Feed-Forward Network ($\lambda = 0.0455$, $\mu_1 = 0.7272$, $\mu_2 = 0.0455$, $\mu_3 = \mu_4 = 0.0909$, $p = 0.1$)

| **L** | **Numerical** | **PW** | **SDA** | | **SDH** | | |
|---|---|---|---|---|---|---|---|
| | $\gamma(L)$ | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | $VRR$ |
| 25 | * | 3.9347e-007 ± 1.90 | 3 | 4.0000e-007 ± 0.07 | 1 | 4.1613e-007 ± 5.68 | 0.03 |
| 50 | * | 1.2630e-014 ± 3.05 | 3 | 1.3283e-014 ± 0.12 | 1 | 1.2725e-014 ± 4.07 | 0.08 |
| 100 | * | 1.3025e-029 ± 10.3 | 3 | 1.2531e-029 ± 0.13 | 2 | 1.2572e-029 ± 5.56 | 0.02 |

Table 4: 4-Node Feed-Forward Network ($\lambda = 0.064$, $\mu_1 = 0.564$, $\mu_2 = 0.039$, $\mu_3 = 0.192$, $\mu_4 = 0.141$, $p = 0.1$)

| **L** | **Numerical** | **PW** | **SDA** | | **SDH** | | |
|---|---|---|---|---|---|---|---|
| | $\gamma(L)$ | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | $VRR$ |
| 25 | * | 1.8357e-008 ± 3.85 | 3 | 1.9687e-008 ± 0.03 | 1 | 1.8217e-008 ± 3.00 | 0.02 |
| 50 | * | 4.7725e-017 ± 2.98 | 3 | 5.2205e-017 ± 0.03 | 1 | 5.0114e-017 ± 3.45 | 0.01 |
| 100 | * | 3.9708e-034 ± 8.03 | 3 | 3.6806e-034 ± 0.03 | 1 | 3.4001e-034 ± 2.04 | 0.01 |

Table 5: 2-Node Feed-Back Network ($\lambda_1 = 0.01$, $\mu_1 = 0.13$, $\lambda_2 = 0.09$, $\mu_2 = 0.77$, $p_{12} = 0.9$, $p_{21} = 0.05$)

| **L** | **Numerical** | **PW** | **SDA** | | **SDH** | | |
|---|---|---|---|---|---|---|---|
| | $\gamma(L)$ | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | $VRR$ |
| 25 | 7.9508e-021 | 2.4772e-020 ± 77.4 | 5 | 7.8930e-021 ± 0.02 | 5 | 8.0126e-021 ± 0.88 | 0.02 |
| 50 | 1.3885e-042 | 8.7117e-043 ± 7.25 | 5 | 1.3883e-042 ± 0.01 | 5 | 1.3964e-042 ± 0.96 | 0.01 |
| 100 | 3.9811e-086 | 2.3596e-086 ± 2.60 | 5 | 3.9535e-086 ± 0.01 | 6 | 3.9790e-086 ± 0.71 | 0.01 |

Table 6: 2-Node Feed-Back Network ($\lambda_1 = 0.01$, $\mu_1 = 0.14$, $\lambda_2 = 0.25$, $\mu_2 = 0.55$, $p_{12} = 0.9$, $p_{21} = 0.05$)

| **L** | **Numerical** | **PW** | **SDA** | | **SDH** | | |
|---|---|---|---|---|---|---|---|
| | $\gamma(L)$ | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | **b** | $\tilde{\gamma}(L) \pm RE\%$ | $VRR$ |
| 25 | 9.9890e-006 | 7.3235e-006 ± 9.21 | 5 | 9.9834e-006 ± 0.09 | 13 | 9.9668e-006 ± 0.32 | 1.49 |
| 50 | 1.4634e-011 | 8.4150e-012 ± 10.4 | 5 | 1.4585e-011 ± 0.29 | 13 | 1.4669e-011 ± 0.78 | 1.96 |
| 100 | 2.0500e-023 | 1.6348e-023 ± 33.3 | 5 | 2.0064e-023 ± 0.57 | 13 | 2.0426e-023 ± 0.62 | 6.16 |

Needless to say, the utility of the approach needs to be tested on larger networks and more complex topologies. This requires further investigations on how to approximate and combine (interpolate) the "optimal" change of measures along the boundaries and in the interior of the state-space. Extensive testing and experimentations would also be required. Also, simple and robust guidelines for selecting the number of boundary layers (dependence range) is another challenge that needs to be addressed.

## REFERENCES


Ahamed, T.P.I., V.S. Borkar and S.K. Juneja. 2004. Adaptive importance sampling technique for Markov chains using stochastic approximation. *Operations Research*. Accepted.

Asmussen, S., and R.Y. Rubinstein. 1995. Steady state rare events simulation in queueing models and its complexity properties. In *Advances in Queueing: Theory, Methods and Open problems,* ed. J.H. Dshalalow, 429–461. CRC Press, New York.

de Boer, P.T. 2000. Analysis and efficient simulation of queueing models of telecommunication systems. PhD Thesis, University of Twente.

de Boer, P.T., and V.F. Nicola. 2002. Adaptive state-dependent importance sampling simulation of Markovian queueing networks. *European Transactions on Telecommunications* 13 (4): 303–315.

de Boer, P.T. 2004 . Analysis of sate-independent IS measures for the two-node tandem queue. *International Workshop on Rare Event Simulation (RESIM'04),* Budapest, Hungary.

Frater, M.R., T.M. Lenon, and B.D.O. Anderson. 1991. Optimally efficient estimation of the statistics of rare events in queueing networks. *IEEE Trans. Autom. Control* 36: 1395–1405.

Glasserman, P., and S-G. Kou. 1995. Analysis of an importance sampling estimator for tandem queues. *ACM Transactions of Modeling and Computer Simulation* 5 (1): 22–42.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Transactions of Modeling and Computer Simulation* 5 (1): 43–85.

Juneja, S.K., and V.F. Nicola. 2004. Efficient simulation of buffer overflow probabilities in Jackson networks with feedback. *ACM Transactions of Modeling and Computer Simulation.* Under final revision.

Kroese, D.P., and V.F. Nicola. 2002. Efficient simulation of a tandem Jackson network. *ACM Transactions of Modeling and Computer Simulation* 12 (2): 119–141.

Nicola, V.F., and T.S. Zaburnenko. 2005. Importance sampling simulation of population overflow in two-node tandem networks. In *Proceedings of the 2nd International Conference on the Quantitative Evaluation of Systems (QEST'05),* Torino, Italy.

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

Randhawa, R.S., and S.K. Juneja. 2004. Combining importance sampling and temporal difference control variates to simulate Markov chains. *ACM Transactions of Modeling and Computer Simulation* 14 (1): 1–30.

Rubinstein, R.Y. 2002. The cross-entropy method and rare events for maximal cut and bipartition problems. *ACM Transactions of Modeling and Computer Simulation* 12 (1): 27–53.

Weber, R.R. 1979. The interchangeability of $M/M/1$ queues in series. *Journal of Applied Probability* 16: 690–695.

Zaburnenko, T.S., and V.F. Nicola. 2005. Efficient heuristics for simulating population overflow in tandem networks. In *Proceedings of the 5th St. Petersburg Workshop on Simulation (SPWS'05),* ed. S.M. Ermakov, V.B. Melas, and A.N. Pepelyshev, 755–764. St. Petersburg University Publishers.


## AUTHOR BIOGRAPHIES


**VICTOR F. NICOLA** is an Associate Professor at the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands. Before that he held positions at IBM Thomas J. Watson Research Center, New York, at Duke University, North Carolina, and at Eindhoven University, The Netherlands. He was also a Visiting Professor at the Norwegian University of Science and Technology and at Simula Research Laboratory, Norway. He is on the Editorial Board of the *International Journal of Simulation Modelling,* and served as a Guest Editor for the *ACM Transactions on Modeling and Computer Simulation.* His research interests include performance and reliability modeling and analysis; (rare event) simulation and optimization methodologies; with applications to high performance networked computing systems and broadband wireless/mobile communication. His e-mail address is: <v.f.nicola@ewi.utwente.nl>.

**TATIANA S. ZABURNENKO** is a PhD candidate at the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands. Her research interests are in the area of (rare-event) simulation with applications in computer and communication networks. Her e-mail address is: <t.s.zaburnenko@ewi.utwente.nl>.