Simulation Methods for Estimation of Blocking Probabilities in Cellular Telecommunication Networks

Felisa J. Vázquez-Abad *

Department of Computer Science and Operations Research

University of Montreal, Montreal, Canada H3C 3J7

Email: vazquez@iro.umontreal.ca

also *Principal Fellow*, Department of Electrical and Electronic Engineering

The University of Melbourne

Lachlan L. H. Andrew †

ARC Special Research Centre for Ultra-Broadband Information Networks

Department of Electrical and Electronic Engineering

The University of Melbourne, Victoria 3010, Australia

Email: l.andrew@ee.mu.oz.au

January 5, 2001

CUBIN Technical Report 2001-01-01

*Supported in part by NSERC-Canada grant # WFA0184198

†Supported by the Australian Research Council (ARC).

Contents

Abstract

Blocking probabilities in FDMA/TDMA cellular mobile communication networks using dynamic channel assignment are hard to compute for realistic sized systems. This computational difficulty is due to the structure of the state space, which imposes strong coupling constraints amongst components of the occupancy vector. Tractable models for dynamically reconfigurable networks have been proposed, and for those, the stationary distribution of the occupancy vector is a product of truncated Poisson random variables. Nonetheless, even in such cases realistic network sizes prevent computation of the closed form within reasonable time and the only viable way to estimate blocking seems to be through simulation.

Alas! Simulation as a means for estimating blocking probability suffers from the fact that the relative error of the estimates can grow without bound for the same number of samples, as the blocking probability diminishes. Small blocking probabilities thus typically require an enormous amount of CPU time for the estimates to be meaningful. Advanced simulation approaches use importance sampling (IS) to overcome this problem. This is known in the simulation literature as "rare event simulation".

Two simulation approaches can be identified. The first one attempts to use simulation as a means for generating the stationary distribution *directly*. We review the Acceptance/Rejection (A/R) method and a fast simulation approach applied to it. While it does give a remarkable variance reduction, this method requires solving a complex optimisation problem before IS can be applied. Next we describe a Markov Chain Monte Carlo method that we have call the Filtered Gibbs Sampler (FGS), which dramatically outperforms A/R and does not need any set-up to perform the simulations.

The second simulation method is to simulate the actual occupancy process and estimate blocking from the measurements. This method can in principle be more robust than estimation of the product form probabilities, because realistic channel assignments can be dealt with, and not just models that satisfy the product form. In this paper we study two regimes under which blocking is a rare event: low utilisation and high capacity. Our simulations use the Standard Clock (SC) method that generates directly the birth and death process and we propose a change of measure that we call *static ISSC* which has bounded relative error: as the traffic intensity decreases, the relative efficiency of this method becomes infinitely better than the FGS method. For high capacity, we use a change of measure that depends on the current state of the network occupancy. This is the *dynamic ISSC* method, for which we can prove optimality in single clique models and we empirically show the advantages of this method over naïve simulation for networks of moderate size and traffic loads.

1 Introduction

Efficient design of communications networks requires the ability to determine the quality of service provided by a particular network configuration. A common quality of service measure is the *blocking probability*, which is the probability that a new call will not be admitted to the network due to insufficient network resources. This paper will consider techniques for determining the blocking probability in cellular telephony systems with frequency reuse, including first generation systems such as the Advanced Mobile Phone System, AMPS (Lee, 1995), and second generation systems such as the Global System for Mobile communication, GSM (Mouly and Pautet, 1992, Redl, Weber, and Oliphant, 1995).

In cellular networks, each mobile station communicates with a base station connected to the wireline telephone network. The region in which mobiles connect to a given station is called a *cell*. Each mobile station communicates with its base station using a specific frequency pair or frequency/timeslot pair known as a "channel". To avoid interference, this channel cannot be used in nearby cells; however, it may be reused in cells sufficiently remote that interference caused by the reused channel is below a specified threshold.

In static assignment schemes, each cell is allocated a fixed subset of the available channels, and calls arriving in a cell are connected only when there are free channels available from those channels. While simple to implement, this strategy may result in wasted resources: all the channels for one cell may be in use, but adjacent cells may have free capacity that could be used to connect incoming calls without causing interference. Network capacity can be improved by *dynamic channel assignment* (Cox and Reudink, 1972, Cox and Reudink, 1973), in which channels not

currently in use in the nearby cells may be used. It is these systems which are the focus of this paper.

Many techniques have been developed for determining the performance of such networks. For Markov models (Poisson arrivals and exponential holding times), when the system is reversible (Kelly, 1979), the stationary state distribution has a simple product form expression on a state space $\mathcal{S}$ which is a small subset of a hypercube H. When there is no mobility of users, this is the case under maximum packing (Everitt and Macfadyen, 1983 and see below), in which calls in progress can be rearranged. There are also models of mobility which preserve this property (see Pallant and Taylor (1995), Boucherie and Mandjes (1998)). Moreover, the result remains valid even when call holding times have non-exponential distributions (Kelly, 1979).

The product form expression involves a normalising constant, from which the blocking probability can be determined directly, without needing to determine specific state probabilities. However, it is computationally prohibitive to calculate explicitly for "large" systems. Product form systems have been studied extensively (see for example the survey of Nelson (1993)). When the number of channels is large but the number of cells is small, it can be solved exactly by recursive methods (Dziong and Roberts, 1987, Pinsky and Conway, 1992), mean value analysis (Reiser and Lavenberg, 1980), generating function inversion

methods (Choudhury, Leung, and Whitt, 1995) or uniform asymptotic approximation (Mitra and Morrison, 1994). However, these techniques all have exponential complexity in the number of cells.

Systems with a large number of cells can be analysed by Monte Carlo techniques, either to estimate the normalising constant (Ross and Wang, 1992 or to avoid the need to calculate it. The simplest approach of the second type is acceptance/rejection (A/R) method, in which states are generated in the full hypercube H: those lying outside the state space $\mathcal{S}$ are rejected, while for those on the boundary of the feasible region, the proportion of blocked cells is recorded (see for example Everitt and Macfadyen, 1983). As the number of cells grows, generation of a sample point inside the state space $\mathcal{S} \subset H$ may become a rare event, and so importance sampling has been applied to these methods (see Ross, Tsang, and Wang, 1994, Lassila and Virtamo, 2000, Mandjes, 1997). An alternative approach is to use Markov chain Monte Carlo (MCMC) techniques such as the Gibbs sampler used by Vázquez-Abad and Andrew (2000), Lassila and Virtamo (1998a) and Lassila and Virtamo (1998b). These generate a Markov chain whose state probabilities satisfy the target product form, and they may be simulated more

efficiently.

Most dynamic channel assignment implementations do not have such a product form solution. It is common in such cases is to use closed form approximations, such as the ubiquitous reduced load approximation Kelly, 1991, developed for circuit switched networks. This approximation works well if there is minimal correlation between blocking due to conflicts with different reuse constraints, but poorly if there is significant correlation. Due to the spatial nature of the reuse constraints in the cellular case, it can be expected that there will be significant correlation. Other approximations are described in Harvey and Hills (1979) and Zahorjan, Eager, and Sweillam (1988).

A very flexible, straightforward and hence common approach is simply to simulate the direct arrival and departure process of calls. This allows any parameters of the system to be measured, and allows arbitrary channel allocation schemes to be compared directly. For these reasons, this is the approach most commonly taken by engineers investigating different dynamic channel assignment systems. However, this approach can be very slow, especially when blocking rates are low. Then the method is not suitable for the dimensioning of large real world networks. In this paper, we present two importance sampling schemes for the efficient direct simulation of systems with low blocking probabilities.

2 Blocking Probabilities

A cellular network is a surface in a two-dimensional space that is covered by a partition of K subsets (called cells). Each cell has a base station that supports C channels, each of which allows one user to communicate with the base station. The principle behind cellular networks is that the limited number of channels available can be reused across the network, provided that the so-called "reuse constraints" are satisfied. These constraints ensure that the performance in any given cell is not excessively degraded by the interference caused by other cells using the same channel elsewhere. The reuse constraints hence depend on the precise

layout of the cells. For the examples in this paper, we shall assume that the cells form a hexagonal grid and 3-cell reuse is employed. That is, the reuse constraints are that no channel may be used more than once in any group of three mutually adjacent cells. In general, a set of cells in which a channel may only be used once is called a "clique". Figure 1 shows a simple seven-cell network, with one clique highlighted.

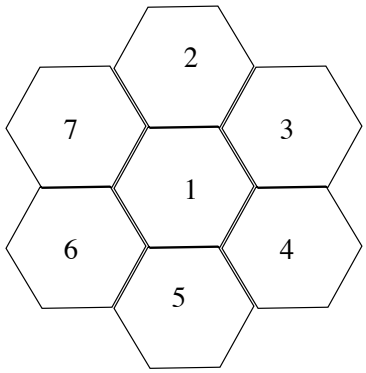


Figure 1: Simple cellular network model

Calls can arrive at the cell in one of two ways. They may be new calls or they may be existing calls being handed off from neighbouring cells due to user mobility. Using dynamic channel assignment, calls arriving to a cell are assigned one of the available channels. If no channel can be allocated without violating a reuse constraint, then the call is *blocked*. Otherwise it is accepted, and uses the selected channel. In practice, the call will generally use the same channel until it leaves the cell. Thus in general the state of the system depends not only on the number of calls in each cell, but also on which particular channels they use.

Let K be the number of cells in the network, M be the number of cliques, and C be the number of channels. Let c_j be the jth clique, $j = 1, \ldots, M$. For the seven cell network of Figure 1, $\{c_j : j = 1, \ldots, M\} = \{(1,2,3),(1,3,4),(1,4,5),(1,5,6),(1,6,7),(1,2,7)\}$.

When the channel assignment cannot be changed while a call is in progress, very little can be said about the occupancy of cell i in the states when blocking occurs, except that it must of course be less than C. In particular, it is possible for calls to be blocked when there are no calls at all in cell i, if all the channels are used elsewhere in the cliques to which i belongs. However, since some channels may be used in one clique and some in another, there is not even a useful lower bound on the number of calls in any of the individual cliques containing i. Define the "cluster" associated with cell i to be the union of all cliques containing i:

$$\mathcal{C}_i = \bigcup_{c_j \ni i} c_j.$$

It is then possible to say that the occupancy of the *cluster* $\mathcal{C}_i$ must be at least C when calls arriving to cell i are blocked, since each channel must be blocked by at least one of the cliques containing cell i. This is the fundamental property of blocking states on which the methods presented here rely.

Most of the techniques described in the introduction rely on having a known closed form for the blocking probability. There is such a closed form for the maximum packing bound proposed by Everitt and Macfadyen (1983), in which channels may be reassigned on the arrival of a new call. However, the operation of reassigning calls is not feasible in practice, and so this closed form does not apply to real channel assignment algorithms. The techniques presented later on in Sections 6 and 7 are applicable to real channel assignment algorithms and are thus of more general applicability than most of the techniques described in the introduction. However, it will simplify the description to describe the algorithm initially for the maximum packing case. This removes the dependence on which particular channels are occupied; the behaviour of the system is determined entirely by the number of calls in each cell. Moreover, we will make the clique packing approximation of Everitt and Macfadyen (1983), Raymond (1991), which only considers constraints local to each clique.

Under clique packing, a state is feasible if *each* of the cliques contains no more calls than there are channels:

$$n^{(c_j)} \leq C \quad \forall j = 1, \ldots, M \tag{1}$$

where $n^{(A)}$ is the number of calls in a set of cells, A, in a given network state, n.

In general, the state of the process is $\tilde{n}(t) = (\tilde{n}_{1,1}(t), \ldots, \tilde{n}_{K,C}(t))$, where $\tilde{n}_(i,j)(t) = 1$ if channel j is used in cell i at time t, and zero otherwise. Under maximum packing, this can be simplified to $n(t) = (n_1(t), \ldots, n_K(t))$, where $n_i(t)$ represents the number of channels in use at cell i at time t, At each cell i calls arrive following independent Poisson process with

corresponding intensities $\lambda_i, i = 1, \ldots, K$. Upon arrival of a call at cell i at time t, it is accepted if there is still at least one channel available. Under maximum packing, the requirement is that

$$\max_{c_j \ni i}(n^{(c_j)}) \leq C - 1.$$

An accepted call on channel j causes $n_{i,j}(t) = n_{i,j}(t^-)+1$ ($n_i(t) = n_i(t^-)+1$ under maximum packing), all other components of the state remaining unchanged. We say that at this time the call is

connected. If an incoming call to cell i finds no channels available (under maximum packing, the current state satisfies (1) with equality for some $c_j \ni i$) then all channels are used and the call is blocked, with no change to the state.

Calls stay connected for a random length of time called the "holding" time, assumed to be exponentially distributed and independent of the rest of the process history. All holding times are identically distributed with mean $1/\mu$. When a call using a channel on cell i terminates, the corresponding occupancy component is decreased by one unit. Although the holding times are assumed to be exponential in this paper, the network performance is in fact independent of the holding time distribution for many channel assignment schemes, including maximum packing (Kelly, 1979). This model gives rise to a continuous-time Markov process.

Because the process consists of independent arrivals and departures, it may be expected that it is a form of birth and death process. Indeed it can be expressed as a *quasi birth and death* (QBD) process (Neuts, 1981). In QBD processes, states can be arranged in layers, such that transitions from layer n can only be to states in layers $n-1$, n or $n+1$. If the nth layer consists of states in which a cluster $\mathcal{C}_i$ contains n calls, then the system is clearly a QBD, since a call arrival within the cluster causes a transition from layer n to layer $n+1$, a departure within the cluster causes a transition from n to $n-1$, and an arrival or departure outside the cluster causes a transition entirely within level n. In this representation, all blocking states are in layers C or higher.

Let $\mathcal{S}$ be the state space consisting of all integer vectors $n = (n_1, \ldots, n_K) \in \mathbb{N}^K$ satisfying (1). The process can be seen as a truncated multidimensional birth and death process for each component: a particular case of a QBD where the rates and barriers depend on all components of the process. When the process is in state $n \in \mathcal{S}$, the birth rate of component i is $\lambda_i \mathbf{1}_{\{n \notin \mathcal{B}_i\}}$, where $\mathbf{1}_{\{A\}}$ is the indicator function of event A and the set $\mathcal{B}_i$ is the set of blocking states for cell i, that is:

$$\mathcal{B}_i = \{n \in \mathcal{S} : \exists c_j \ni i, n^{(c_j)} = C\}, \quad i = 1, \ldots, K \tag{2}$$

The death rate at cell i in state $n \in \mathcal{S}$ is $n_i\mu$.

The performance measure of interest is called the *blocking probability*, and it is defined as the long term probability that an incoming arrival is lost:

$$B = \lim_{t\to\infty} \frac{\sum_{i=1}^K Y_i(t)}{A(t)} = \sum_{i=1}^K \left(\frac{\lambda_i}{\lambda}\right) B_i = \sum_{i=1}^K \left(\frac{\lambda_i}{\lambda}\right) \pi(\mathcal{B}_i) \tag{3}$$

where $Y_i(t)$ is the total number of calls lost in cell i up to time t and $A(t)$ is the total number of arrivals up to time t. The term B_i is the long term proportion of calls arriving to cell i that are blocked, and $\pi(\mathcal{B}_i)$ is the stationary probability that the state is in the blocking set $\mathcal{B}_i$.

Evaluating blocking probabilities using (3) is a difficult numerical problem. Everitt and Macfadyen (1983) propose a methodology that identifies geometrical structures corresponding

to the cliques within the network and characterizes the constraint sets in terms of the geometry of the cells. This is used to calculate the normalization factor $G(C)$, where the sums have to respect (1). This method for calculating blocking probabilities is computationally prohibitive when the number of cells is large, which means that simulation is often superior.

The renewal theorem can be used to re-write (3) in terms of expectations within regenerative cycles. However, as the system state has many components, regeneration cycles are frequently too long to be a feasible basis for simulation. The concept of *quasi-regeneration* was introduced to calculate stationary averages for such systems, as explained in Gaivoronski and Messina (1996) and Chang, Heidelberger, and Shahabuddin (1995). Consider a random process, n, and assume it starts with the stationary distribution $\mathsf{P}[n(0) = n] = \pi(n)$. Consider a set of points, A, such that there is an a.s. finite stopping time T_1 defined as the first entry time to the set A from its complement A' and such that the distribution of the process $\{n(t + T_1); t > 0\}$ is identical to that of the process $\{n(t); t > 0\}$ with $n(0) \in A$ drawn from the stationary distribution of the process, conditioned on being on the boundary of the set A. The set A is called a quasi-regenerative set. Because the channel occupancy process $n(t)$ described above is an irreducible Markovian process on a finite state space, all subsets of the state space are quasi-regenerative sets and a unique stationary measure π exists. The times between consecutive entries to the set A are termed "A-cycles". Unlike true regenerative cycles, A-cycles may not be independent, but they are still identically distributed.

It will be useful to consider different quasi-regeneration sets, A_i, for different cells i. Let $T^{(i)}$ be the time of the first reentry to set A_i, and $X_i(T^{(i)})$ be the amount of time within an A_i-cycle that the process spends in $\mathcal{B}_i$. Then (Breiman, 1992)

$$B_i = \frac{\mathsf{E}[X_i(T^{(i)})]}{\mathsf{E}(T^{(i)})}. \tag{4}$$

The sets A_i will be chosen in such a way as to minimise the required simulation time.

3 Fast Simulation Methods

3.1 Relative Efficiency

To approximate the value of B, simulation can be used to produce a sample of random variables $Y_s, s = 1, \ldots, S$ whose sample average is consistent for B, namely,

$$\mathsf{E}[\hat{Y}(S)] = \frac{1}{S}\sum_{s=1}^{S} \mathsf{E}[Y_s] \to B$$

as $S \to \infty$. In some cases $\hat{Y}(S)$ is unbiased, that is, $\mathsf{E}[\hat{Y}(S)] = B$. When consecutive samples come from independent replicas of a simulation, the variance of $\hat{Y}(S)$ can easily be estimated. Under some conditions on the boundedness of this variance, Alexopoulos and Seila (1998) show that the Central Limit Theorem (CLT) holds:

$$\frac{\hat{Y}(S) - B}{\mathsf{Var}[\hat{Y}(S)]} \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0, 1)$$

where $\stackrel{\mathcal{L}}{\Longrightarrow}$ represents convergence in distribution and $\mathcal{N}(0,1)$ is the standard normal distribution. Confidence intervals (see Larson, 1973) can then be approximated using:

$$\lim_{S\to\infty} \mathsf{P}\left\{ B \notin \hat{Y}(S) \pm z_{(1-\alpha/2)}\sqrt{\mathsf{Var}[\hat{Y}(S)]} \right\} \leq \alpha,$$

where $z_{1-\alpha/2}$ is the upper $(1-\alpha/2)$ quantile of the normal distribution. In order to achieve a prespecified precision within a

confidence level α, S must be large enough for $\mathsf{Var}[\hat{Y}(S)]$ to be within the specified limits. Some estimators have less variance than others, yet they may require extremely long simulation times. A commonly used measure which takes this into account (see for example Glynn and Whitt, 1992) is that of the *efficiency* of an estimator. When estimating probabilities, one often requires estimating the probabilities within a given *relative error*, in terms of a percentage of the (unknown) probability. In these cases, the sample size must grow in order for

$$\frac{\sqrt{\mathsf{Var}[\hat{Y}(S)]}}{B}$$

to remain at the required level.

Definition 1 *The* relative efficiency *of a consistent estimator $\hat{Y}(S)$ is:*

$$\mathcal{E}_r(\hat{Y}(S)) = \frac{B^2}{\mathrm{CPU}[\hat{Y}(S)]\mathsf{Var}[\hat{Y}(S)]}, \tag{5}$$

where $\mathrm{CPU}[\hat{Y}(S)]$ *denotes the average CPU time of the simulation that produces the S samples. The* asymptotic relative efficiency *is defined as* $\lim_{S\to\infty}\mathcal{E}_r(\hat{Y}(S))$*, if this limit exists.*

In particular, if the simulation has produced a sample of i.i.d. random variables, then the relative efficiency is independent of S: CPU times tend to grow linearly with the sample size and $\mathsf{Var}[\hat{Y}(S)]$ decreases as $1/S$. When estimating efficiencies, we shall often use the estimated values of B and of $\mathsf{Var}[\hat{Y}(S)]$ in the definition of $\mathcal{E}_r$. If the estimators are known not to be consistent, i.e., $\lim_{S\to\infty}\mathsf{E}[\hat{Y}(S)] \neq B$, then the actual definitions of efficiency use the mean square errors instead of the variances. In this paper we deal with consistent or unbiased estimation, so the problem of estimating the MSE is not present.

In this paper we will compare several simulation methods for estimating B, in terms of their efficiencies.

3.2 Rare Events

The occupancy process $\{n(t); t \geq 0\}$, as described in Section 2 is a continuous time Markov process on $(\Omega, \mathbb{P})$ referring to a truncated birth and death process with state-dependent reflecting boundaries. The arrival rate of calls into cell i is λ_i, $i = 1, \ldots, K$, and the average holding time per call is $1/\mu$.

Fix any cell i and consider expression (4) of the blocking probability B_i using A-cycles. Then $T^{(i)}$ is the duration of an A-cycle, which is adapted to the *natural filtration* of the process $\{n(t); t > 0\}$; that is, knowledge of the process history up to time t is sufficient to determine

if the A-cycle terminates at this epoch, and note that $T^{(i)}$ must necessarily coincide with an arrival to one of the cliques where cell i belongs.

Define now τ_i to be the stopping time when blocking states are reached and $\mathcal{R}_i$ to be the event that the occupancy process enters the blocking set $\mathcal{B}_i$ before the cycle is over:

$$\tau_i = \min\{k : S_k \leq T^{(i)} \text{ and } n(S_k) \in \mathcal{B}_i\}, \quad \text{and} \quad \mathcal{R}_i = \{S_{\tau_i} < T^{(i)}\}. \tag{6}$$

Equation (4) requires estimation of the proportion of time that the process spends in blocking states. For all ω with $\mathbf{1}_{\{\mathcal{R}_i\}}(\omega) = 0$, $X_i(T^{(i)})(\omega) = 0$, thus the numerator in (4) can be estimated with:

$$\mathsf{E}[X_i(T^{(i)})] = \mathsf{E}\left(\mathsf{E}[X_i(T^{(i)})|\mathcal{R}_i]\,\mathbf{1}_{\{\mathcal{R}_i\}}\right).$$

Definition 2 *Let $\{n(t)\}$ be a process defined on a probability space $(\Omega, \{\mathfrak{F}_k, k \geq 0\}, \mathbb{P})$ and let $\epsilon > 0$ be a parameter of the distribution of the process $\{n(t)\}$. Denote by P^ϵ the corresponding distribution. The event $\mathcal{R} \subset \Omega$ is called a* rare event *if $\lim_{\epsilon \to 0} \mathsf{P}^\epsilon(\mathcal{R}) = 0$.*

EXAMPLE: In the cellular network problem, $\mathcal{R}_i$ is a rare event under several different parameterizations. In particular, if $\lambda_i(\epsilon) \to 0$ as $\epsilon \to 0$, then the blocking probability tends to zero. We will study this "light traffic" regime. As well, it is of importance to study the case when $C \to \infty$, where C is the maximum number of available channels in the cellular network. Another "high capacity" regime, $GI/G/s$ queues with buffer size C has been thoroughly studied (see for example Sadowsky, 1991), but the methods for that system are not applicable to our case.

⁂

If $p(\epsilon) = \mathsf{P}^\epsilon(\mathcal{R}_i)$ is estimated via simulation using $\mathbf{1}_{\{\mathcal{R}_i\}}$ for S consecutive A-cycles, then the variance of the estimator is at least $p(\epsilon)(1-p(\epsilon))/S$ because consecutive A-cycles introduce positive correlation on the Bernoulli sequence. The relative error in the estimation is bounded below by:

$$\frac{\sqrt{p(\epsilon)(1-p(\epsilon))/S}}{\hat{p}(\epsilon)} \approx \sqrt{\frac{(1-p(\epsilon))}{S\,p(\epsilon)}} \to \infty$$

as $\epsilon \to 0$. Larger and larger sample sizes S must be used to obtain the same accuracy in the estimation for different parameter values. The efficiency in the estimation of rare events can be improved using a change of measure approach via importance sampling (see Bratley, Fox, and Schrage, 1987, Asmussen and Nielsen, 1995, and Devetsikiotis and Townsend, 1993 among others).

3.3 Importance Sampling and BRE

Assume an underlying Markovian process $\{U_k\}$ with natural filtration $\mathfrak{F}_k = \sigma(U_1, \ldots, U_k)$, such that the occupancy process $\{n(t)\}$ can be embedded into a discrete event process

adapted to the filtration $\{\mathfrak{F}_k, k = 1, 2, \ldots\}$. Use the notation $\{n(k), k = 1, 2, \ldots\}$ for the embedded process (with the obvious abuse in notation).

EXAMPLE:

Discrete event simulation is commonly performed by generating, for each cell, the interarrival times $A_k(i), k = 1, 2, \ldots$ according to a Poisson process with rate λ_i, and the holding

times $H_k(i), k = 1, 2, \ldots$ as independent exponential variables with mean $1/\mu$. At the arrival epochs to cell i, given by the expression:

$$S_t(i) = \sum_{k=1}^{t} A_k(i),$$

the occupancy process n_i increases by one, provided $n \notin \mathcal{B}_i$. At the call termination epochs $S_k(i)+H_k(i)$ the corresponding component decreases by one. Identify $U_k(i) = (A_k(i), H_k(i))$. Then the embedded process is adapted to the filtration $\mathfrak{F}_k = \sigma(U_1(i), \ldots, U_{k(i)}(i); i = 1, \ldots, K)$, where $k(i)$ is the number of arrivals occurring at cell i amongst the first k total arrivals.

⁂

The general structure of the evolution of the Markovian process $\{U_k\}$ is assumed to satisfy: $\mathsf{P}[U_{k+1} \in A|\mathfrak{F}_k] = \mathsf{P}[U_{k+1} \in A|n(k)]$. In the example above, U_{k+1} is actually independent of the state $n(k)$; the more general formulation will prove useful for the standard clock technique in Section 5.3. Let $f_n(u)$ be the conditional density of U_{k+1} given the state $n(k) = n$ (the case where U_k has discrete components can be treated in a similar way). Then for any real valued functional of the process $\Psi(U_1, U_2, \ldots)$ with finite expectation, and any finite integer T,

$$\mathsf{E}[\Psi(U_1, \ldots, U_T)] = \int \Psi(u_1, \ldots, u_T) \left[\prod_{k=0}^{T-1} f_{n(k)}(u_{k+1})\right] du_1, \ldots du_T, \quad n_0 : \text{initial state.}$$

Let now $f_n^*(\cdot)$ be *another* conditional density. Multiplying and dividing each term in the product above by the new densities, we obtain:

$$\begin{aligned}
\mathsf{E}[\Psi(U_1, \ldots, U_T)] &= \int \Psi(u_1, \ldots, u_k) \left(\prod_{k=0}^{T-1} \frac{f_{n(k)}(u_{k+1})}{f^*_{n(k)}(u_{k+1})}\right) \left[\prod_{k=0}^{T-1} f^*_{n(k)}(u_{k+1})\right] du_1, \ldots du_k \\
&= \mathsf{E}^*[\Psi(U_1, \ldots, U_T) L(U_1, \ldots, U_T)]; \\
L(U_1, \ldots, U_T) &= \prod_{k=0}^{T-1} \frac{f_{n(k)}(U_{k+1})}{f^*_{n(k)}(U_{k+1})} \equiv L_T \qquad (7)
\end{aligned}$$

where E^* denotes the expectation w.r.t. a different probability: here the process $\{U_k\}$ evolves according to the transition densities f^*. When L_k is interpreted as a statistics depending on the observations, it is called the "likelihood ratio". The above expression is valid for general functionals only when the support of the old distributions is contained within that of the new distributions, i.e., if for every $u > 0$, $f_n(u) > 0$ implies $f_n^*(u) > 0$ for the corresponding densities. In such cases one says that the original measure P is *absolutely continuous* w.r.t. P^*, denoted by $\mathsf{P} << \mathsf{P}^*$. In the case of rare event estimation for the cell blocking probability, we will identify $T = \tau_i$ and $\Psi = \mathbf{1}_{\{\mathcal{R}_i\}}$. The change of measure approach can be specialized to the "important" region and (7) can be used for densities such that $\mathsf{P}^*(\mathcal{R}_i)$ is as close to 1 as desired, even if P is not absolutely continuous w.r.t. P^*, as long as $\mathsf{P}\rfloor_{\mathcal{R}_i}$ is, where $\mathsf{P}\rfloor_{\mathcal{R}_i}$ the restriction of P to the "important set" $\mathcal{R}_i$ (see Vázquez-Abad and LeQuoc, 2001). In this case the Radon-Nikodym derivative is no longer called a *likelihood ratio*, but it still satisfies $\mathsf{E}[\mathbf{1}_{\{\mathcal{R}_i\}}] = \mathsf{E}^*[L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}}]$.

Remark: The random process $\{L_k, k = 1, 2 \ldots\}$ is what is known as a (multiplicative) P^*-martingale adapted to the filtration, because it satisfies $\mathsf{E}^*[L_{k+1}|\mathfrak{F}_k] = L_k$. This martingale

defines the appropriate Radon-Nikodym derivative for functionals of the process $\{U_k\}$, and the change of measure (7) is valid also when T is a stopping time, which is a consequence of Wald's identity, as explained in detail in Asmussen and Nielsen (1995), Sadowsky (1991).

Definition 3 *An unbiased IS estimator for the rare event probability* $\mathsf{P}(\mathcal{R}_i) = p(\epsilon)$, $\mathsf{E}^*[L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}}]$ *has asymptotic* bounded relative error (BRE) *if there are constants* $b < \infty, \epsilon_0 > 0$ *such that:*

$$\sup_{\epsilon \leq \epsilon_0} \frac{\sqrt{\mathsf{Var}^*[L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}}]}}{\mathsf{E}^*[L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}}]} \leq b. \tag{8}$$

It is usually desirable to find a new probability P^* such that

$\mathsf{P}^*(\mathcal{R}_i) \to 1$ as $\epsilon \to 0$, with $L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}} \leq 1$ w.p.1. for all ϵ. It is in this sense that the estimation can be performed faster and with variance reduction. Indeed, the variance of the estimator without IS is $\mathsf{Var}(\mathbf{1}_{\{\mathcal{R}_i\}}) = p(\epsilon)[1 - p(\epsilon)]$, but under the new measure,

$$\mathsf{Var}^*(L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}}) = \mathsf{E}^*[L_{\tau_i}^2\mathbf{1}_{\{\mathcal{R}_i\}}] - p^2(\epsilon) = \mathsf{E}[L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}}] - p^2(\epsilon) \leq p(\epsilon) - p^2(\epsilon),$$

because (7) can be applied to $\Psi = L_{\tau_i}$ and we have assumed that $L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}} \leq 1$ w.p.1. The following lemma is a direct consequence of Definition 3.

Lemma 1 *If there are constants* d_1, d_2 *and* b *such that* $p(\epsilon) \geq d_1\epsilon^b$ *and* $L\,\mathbf{1}_{\{\mathcal{R}\}} \leq d_2\epsilon^b$ *a.s., then the IS estimator* $L\,\mathbf{1}_{\{\mathcal{R}\}}$ *is BRE for* $p(\epsilon)$.

3.4 Large Deviations and Asymptotic Optimality

By construction, under *any* change of measure with bounded Radon-Nikodym derivative it holds that $p(\epsilon) = \mathsf{P}^\epsilon(\mathcal{R}) = \mathsf{E}^*[L\,\mathbf{1}_{\{\mathcal{R}\}}]$. Often the new measure depends on ϵ, and L will in general also be a function of ϵ, although we will not make this explicit in our notation. Because variances are non negative, it must always be true that:

$$\mathsf{E}^*[L^2\,\mathbf{1}_{\{\mathcal{R}\}}] \geq p^2(\epsilon). \tag{9}$$

Estimators that satisfy (9) with equality are optimal.

Definition 4 *Suppose that there is a constant* v *and a function* $h(v, \epsilon)$ *such that* $h(v, \epsilon) \to 0$ *as* $\epsilon \to 0$*, at a rate which increases with increasing* v*, and*

$$\lim_{\epsilon \to 0} h^{-1}(p(\epsilon), \epsilon) = v,$$

where $h^{-1}(p, \cdot)$ *is the inverse of* $h(v, \cdot)$ *with respect to the first argument, so that* $h^{-1}(h(v, \epsilon), \epsilon) = v$*. An IS estimator of* $p(\epsilon)$ *is is said to be* asymptotically optimal *(a.o) if it satisfies:*

$$\lim_{\epsilon \to 0} h^{-1}\left(\sqrt{\mathsf{E}^*[L^2\,\mathbf{1}_{\{\mathcal{R}\}}]}, \epsilon\right) = v. \tag{10}$$

In particular, if the rare event probability is exponentially decreasing, $p(\epsilon) \approx e^{-v/\epsilon}$, then $h(v, \epsilon) = e^{-v/\epsilon}$ and $h^{-1}(p, \epsilon) = -\epsilon \log(p)$. If the rare event probability is geometrically decreasing, $p(\epsilon) \approx \epsilon^v$, then $h(v, \epsilon) = \epsilon^v$ and $h^{-1}(p, \epsilon) = \log(p)/\log(\epsilon)$.

It follows that a BRE estimator of rare event probability will be asymptotically optimal.

Definition 5 *For a distribution $F(\cdot)$, the* exponentially twisted *distribution is defined by:*

$$dF^{(\alpha)}(x) = \frac{e^{-\alpha x}}{M(-\alpha)} dF(x) = e^{-\alpha x - \Gamma(\alpha)} dF(x),$$

where $M(\alpha) = \mathsf{E}[e^{\alpha X}]$ is the moment generating function (MGF) of F and $\Gamma(\alpha) = \log(M(-\alpha))$ is correspondingly called the log-MGF.

In the case that exponential twisting is used for a change of measure, the likelihood ratio is given by $L(X) = e^{\alpha X + \Gamma(\alpha)}$. In many known problems, exponential twisting can yield BRE and asymptotically optimal estimators.

EXAMPLE: The model is a $GI/G/s/\infty$ queueing system: s parallel servers with infinite queueing capacity. Interarrival times $\{A_i\}$ are i.i.d. following a distribution F_A with log-MGF $\Gamma_A(\alpha)$ and all services are i.i.d. with distribution F_B with log-MGF $\Gamma_B(\alpha)$. Sadowsky (1991) shows the large deviations result for the overflow probability:

$$\lim_{n\to\infty} \frac{1}{n} \log p(n) = \Gamma_A(s\alpha^*),$$

where $p(n)$ is the probability that the queue size will reach n customers in waiting within each A-cycle, and α^* satisfies $\Gamma_A(s\alpha^*) + \Gamma_B(-\alpha^*) = 0$. Sadowsky uses the following definition: the start of an A-cycle is the arrival time of a customer that finds $s-1$ servers busy. Let the initial distribution of the s residual service times $H_0^{(i)}, i = 1, \ldots, s$ be the stationary distribution at the start of the A-cycle. For the Markovian model all residual service times are exponentially distributed. Define T as the index of the next customer that finds again $s-1$ servers busy. Next define τ as the index of the first customer that finds n customers in the queue. Necessarily, all the s servers must be busy at this time and other n customers wait in queue. This implies that on the set $\tau < T$, for each server i:

$$\sum_{k=0}^{N_i(\tau)} H_k^{(i)} < S_\tau \leq \sum_{k=0}^{N_i(\tau)+1} H_k^{(i)}, \tag{11}$$

where $H_k^{(i)}$ are the consecutive service times at server i and $N_k(i)$ are the number of customers that have departed from server i by the time of the k-th arrival. Furthermore, because at time S_τ there are still n customers waiting, then necessarily:

$$\sum_{i=1}^{s} \sum_{k=0}^{N_i(\tau)+1} H_k^{(i)} = \sum_{k=0}^{\tau-n} \tilde{H}_k$$

where $\tilde{H}_k$ is the service time of the k-th customer (regardless of which server it goes to).

Theorem 1 (Sadowsky) *The IS estimator obtained using the exponentially twisted measures $F_A^*(x) = F_A^{s\alpha^*}(x)$, $F_B^*(y) = F_B^{(-\alpha^*)}(y)$ is the only asymptotically optimal IS estimator.*

Consider the filtration of the process $\mathfrak{F}_t = \sigma((A_k, H_k); k = 0, \ldots, t)$, and notice that knowledge of the interarrival and service requirements is sufficient to determine which server is

assigned to each customer (FCFS). Using independence between A_i, H_i, under the proposed change of measure, the Radon-Nikodym derivative is the adapted process:

$$L_t = \prod_{k=0}^{t} e^{-s\alpha^* A_k + \Gamma_A(-s\alpha^*)} \, e^{\alpha^* H_k + \Gamma_B(\alpha^*)}$$

which, evaluated at time τ, can be re-written as:

$$\begin{aligned} L_\tau \mathbf{1}_{\{\tau<T\}} &= \exp\left\{ \alpha^* \sum_{i=1}^{s} \left(S_\tau - \sum_{k=0}^{N_i(\tau)+1} H_k^{(i)} \right) - \alpha^* \sum_{k=\tau-n+1}^{\tau} H_k \right\} \\ &= \exp\left\{ -\alpha^* \left(\sum_{i=1}^{s} H_0^{(s)} + \sum_{k=\tau-n+1}^{\tau} H_k \right) \right\}, \end{aligned}$$

where $H_0^{(i)}$ is the residual service time at server i at the time of the τ arrival S_τ. For the exponential model, these are simply the exponentials. For Poisson arrivals, these would have the stationary distribution at a full busy cycle. Call $R(s,\alpha^*) = \log \prod_i \mathsf{E}^*[e^{\alpha^* H_0^{(i)}}]$ its log-MGF and assume that it is bounded (as would be in the case of exponential residual times). By assumption, the service times of the customers in queue are independent of these residual times (under the original as well as the "star" measure). We will use the fact that under the new measure, overflow is asymptotically certain, that is, $\mathsf{P}^*(\mathcal{R}) \to 1$ as $n \to \infty$. The large deviations rate can be calculated using:

$$p(n) = \mathsf{E}^*[L_\tau \, \mathbf{1}_{\{\mathcal{R}\}}] = \mathsf{E}^*[L_\tau | \mathbf{1}_{\{\mathcal{R}\}}] \mathsf{P}^*(\mathcal{R}) \to e^{R(s,\alpha^*)} [\mathsf{E}^*(e^{-\alpha^* H_1})]^n = e^{R(s,\alpha^*) + n\Gamma_A(s\alpha^*)},$$

where we have used that $(H_k; k = \tau - n, \ldots, n)$ are i.i.d. independent of τ. The last identity is a consequence of the definition of the conjugate twisted exponentials:

$$\mathsf{E}^*\big[e^{-\alpha^* H_k + \Gamma_B(-\alpha^*) - \Gamma_B(-\alpha^*)}\big] = e^{-\Gamma_B(\alpha^*)} = e^{\Gamma_A(s\alpha^*)}.$$

It is now straightforward to evaluate $\lim_{n\to\infty} \frac{1}{n} \log p(n)$.

In the special case of an $M/M/s/\infty$ server, where $F_A(\cdot) \sim \exp(\lambda), F_B(\cdot) \sim \exp(\mu)$ the conjugate exponential twisting yields:

$$\begin{aligned} M_A(\alpha) = \mathsf{E}[e^{\alpha} A_1] = \frac{\lambda}{\lambda - \alpha}, \qquad F_A^{(\alpha)} \sim \exp(\lambda - \alpha), \qquad \alpha > -\lambda \\ \alpha^* = \frac{s\mu - \lambda}{s}, \qquad F_A^* \sim \exp(s\mu), \qquad F_B^* \sim \exp(\lambda/s), \end{aligned}$$

and it is known as "rate swapping": one again simulates an $M/M/s/\infty$ server, with the new rates:

$$\lambda^* = s\mu, \quad \mu^* = \lambda/s. \tag{12}$$

In Sections 6 and 7 we discuss the implementation of the IS that swaps arrival and service rates for the cellular network problem.

⁂

4 Estimation of Closed Form Probabilities

This section presents two methods for estimating blocking probabilities by means of (3), when the channel assignment model satisfies the product form solution. In particular, for the maximum packing model, the stationary cell occupancy probability is a truncated multivariate Poisson distribution

(Boucherie and Mandjes, 1998, Everitt and Macfadyen, 1983)

$$\pi(n) = \frac{1}{G(C)} \prod_{i=1}^{K} \left(\frac{\rho_i^{n_i}}{n_i!} \right), \quad n \in \mathcal{S} \tag{13}$$

where $\rho_i = \lambda_i/\mu$ is the "offered traffic" to cell i, and

$$G(C) = \sum_{n \in \mathcal{S}} \prod_{i=1}^{K} \left(\frac{\rho_i^{n_i}}{n_i!} \right)$$

is the normalizing constant. This section studies the estimation of the blocking probabilities in (3) by generating a sample of random variables $X_s \in \mathcal{S}$ with distribution $\pi(\cdot)$ to calculate:

$$B = \sum_{i=1}^{K} \left(\frac{\lambda_i}{\lambda} \right) \pi(\mathcal{B}_i).$$

4.1 Acceptance/Rejection Method

The method of acceptance/rejection explained in Ross (1997) and Bratley, Fox, and Schrage (1987) is used for generating random variables. In the case of truncated distributions, it is a natural method of generating a random variable n with the stationary probability $\pi(n)$ of (13). The method follows the recursion:

Algorithm 1: A/R

1. Repeat

   (a) Generate $M_1, \ldots, M_K$ independent variables with $M_i \sim$Poisson(ρ_i).

   Until $N = (M_1, \ldots, M_K) \in \mathcal{S}$

2. $X_s = N$

3. $s \leftarrow s + 1$, go to 1.

The resulting random variables $\{X_s\}$ have distribution π in (13).

This method is used in Everitt and Macfadyen (1983) and Yates (1997) to calculate (3) as follows. Random variables $\{X_s\}$ are generated according to Algorithm 1. Next, compute:

$$Y_s = \sum_{i=1}^{K} \left(\frac{\lambda_i}{\lambda} \right) \mathbf{1}_{\{X_s \in \mathcal{B}_i\}} \tag{14}$$

which identifies for which cells the state X_s is in a blocking state and weights the probability accordingly. By definition, $\mathcal{B}_i \cap \mathcal{B}_j \neq \emptyset$ for $i \neq j$, and this overlapping means that some states may block several adjacent cells. By construction, Y_s is an unbiased estimator of the blocking probability: $\mathsf{E}[Y_s] = B$. Clearly the samples Y_s, $s = 1, \ldots, S$ are independent, and the CLT can be used to build confidence intervals (35) for the estimation.

The acceptance probability $G(C)$ can be very low for moderately large networks. To see this, imagine the feasible region of the seven-cell network projected on (n_i, n_{i+1}, n_{i-1}) as shown in Figure 2.

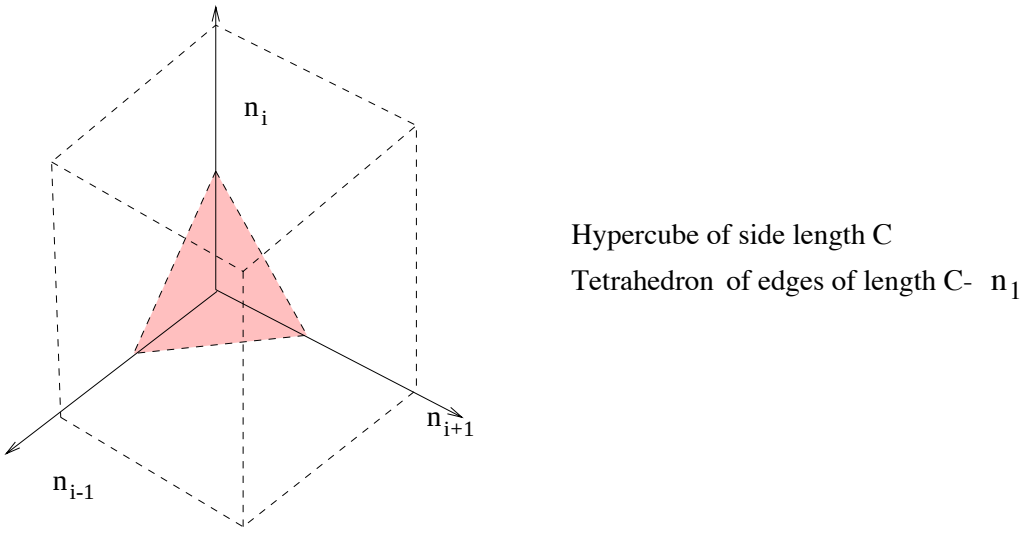


Figure 2: Projection of the feasibility region $\mathcal{S}$.

For each value of $n_1 \leq C$ a tetrahedron contained in the hypercube of side length C symbolizes the feasible region on the projected coordinates. Clearly other coordinates will be also constrained by the values of n_{i-1} and n_{i+1}. Even if a pre-calculated table is used to generate Poisson random variables truncated to the hypercube, that is, $M_i \sim$Poisson$(\rho_i)\rfloor_C, M_i \leq C$ a.s., the volume of the feasible region can be considerably smaller than the whole space. This suggests the use of IS as follows (refer to Section 6.2).

EXAMPLE: The canonical large deviations example deals with sample averages, as in Boucherie and Mandjes (1998), Bucklew (1990). Suppose that $S_n = \sum_{i=1}^n X_i$ where $X_i, i = 1, 2, \ldots$ are zero-mean i.i.d. random variables with well defined MGF $M(\alpha)$. The rare event $\mathcal{R} = \{S_n/n > a\}$, for $a > 0$ is exponentially decreasing with increasing n, and it is well known that:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathsf{P}[S_n/n > a] = -\alpha^* a + \Gamma(-\alpha^*) \equiv -v(a),$$

where α^* satisfies $\alpha^* a - \Gamma(\alpha^*) = \sup_{\alpha\in\mathbb{R}}(\alpha a - \Gamma(\alpha))$. The rate of convergence is clearly the same as the exponent of the likelihood ratio for the exponential twisting with parameter α^*. Using this change of measure for each X_i, by independence the Radon-Nikodym derivative is:

$$L(X_1, \ldots, X_n) = \prod_{i=1}^n e^{-\alpha^* X_i + \Gamma(-\alpha^*)} = e^{-\alpha^* S_n + n\Gamma(-\alpha^*)}$$

so that $\log L\mathbf{1}_{\{\mathcal{R}\}} \leq -\alpha^*(na) + n\Gamma(-\alpha^*)$, where we have used $\mathbf{1}_{\{\mathcal{R}\}} = \{S_n \geq na\}$. Non-negativity of the variance implies that

$$\frac{1}{n}\log \mathsf{E}^*[L^2\mathbf{1}_{\{\mathcal{R}\}}] \geq -2v(a), \quad \text{as } n \to \infty.$$

Identifying $\epsilon = 1/n$, it then follows that this change of measure yields an asymptotically optimal IS estimator.

Boucherie and Mandjes (1998) and Mandjes (1997) use a scaling large deviations result from Kelly to establish the limit rate of the exponential decay of the blocking probability for route i when both the capacity and the load are scaled, namely nC and $n\rho_i$, respectively. Both the state space

$\mathcal{S}^{(n)}$ and the blocking sets $\mathcal{B}_i^{(n)}$ depend on n through the capacity nC. It is shown in the above references that

$$\lim_{n\to\infty}\frac{1}{n}\log \mathsf{P}[M^{(n)} \in \mathcal{B}_i^{(n)}] = -\inf_{x\in\mathcal{B}_i^{(n)}}\sum_{r=1}^{R}\left(x_r\log\left(\frac{x_r}{\rho_r}\right) - x_r + \rho_r\right),$$

which gives rise to the heuristic argument for the change of measure: use a multivariate Poisson distribution with parameter nx^*, where x^* is the minimizer of the expression in the right hand side. Indeed under this change of measure, the likelihood ratio is:

$$\prod_{i=1}^{R}\frac{e^{-n\rho_i}(n\rho_i)^{m_i}}{m_i!}\frac{e^{nx_i^*}m_i!}{(nx_i^*)^{m_i}} = \exp\left[-n\sum_{i=1}^{R}\left(\frac{m_i}{n}\log\left(\frac{x_i^*}{\rho_i}\right) - (\rho_i - x_i^*)\right)\right].$$

Notice that it is not necessarily true that $m_i \leq nx_i^*$ on the set $\{m \in \mathcal{B}_i^{(n)}\}$, which would parallel the canonical arguments for a.o. when we dealt with a sample average. While Mandjes (1997) does not show asymptotic optimality of this IS estimator, numerous simulation results seem to exhibit BRE.

⁂

4.2 Markov Chain Monte Carlo Methods

The computational inefficiency of the acceptance/rejection method for generating X_k is mainly due to the fact that most of the generated variables lie outside the region $\mathcal{S}$. Although the IS estimator proposed by Mandjes (1997) seems to overcome this problem under the scaling regime, the potential advantages of variance reduction may be outweighed by the amount of time and programming effort to set up the simulation methods: the optimisation problem to be solved may require a lot of computational effort. An algorithm based on the Metropolis Hastings method (see Ross, 1997) is built here. It defines a Markov chain that evolves within the state space $\mathcal{S}$ and whose stationary probabilities are the probabilities π given by (13). First we present the metropolis-Hastings algorithm, next we build the so-called Filtered

Gibbs Sampler proposed in Vázquez-Abad and Andrew (2000)

For each state $n \in \mathcal{S}$ define a *neighborhood* $N(n) \subset \mathcal{S}$. The requirement of the neighborhoods is that for every $m_0, m_f \in \mathcal{S}$ there exists a finite sequence of states $m_1, m_2, \ldots, m_p =$

m_f such that $m_k \in N(m_{k-1})$. In addition, if $n \in N(m)$ then $m \in N(n)$. The name "neighborhood" therefore does not necessarily mean that the elements within are adjacent to each other. To generate the successive values of the process $\{X_n\}$ use:

Algorithm 2: Metropolis Hastings

1. Set $n = X_k$,
2. Choose $m \in N(n)$ uniformly,
3. Generate $U \sim U[0,1]$ and define the next state:

$$X_{k+1} = \begin{cases} m & \text{if } U < \alpha(n,m) \\ n & \text{otherwise} \end{cases}$$

4. $k \leftarrow k+1$, go to 1.

where:

$$\alpha(n,m) = \min\left(1, \frac{\pi(m)\|N(n)\|}{\pi(n)\|N(m)\|}\right), \quad m \in N(n)$$

and $\|N(n)\|$ is the number of elements in the neighborhood of n.

Then the Markov chain $\{X_k, k \geq 1\}$ has stationary probabilities π given by (13). To see this, first notice that the chain is irreducible by the requirement on the neighborhoods: all states communicate. Since it is a finite state chain, all the states are positive recurrent. Using the reversibility theorem of Ross (1993), the process is reversible with stationary probabilities π^* if and only if

$$\pi^*(n)P_{n,m} = \pi^*(m)P_{m,n} \tag{15}$$

where $P_{n,m} = \mathsf{P}\{X_{k+1} = m | X_k = n\}$ is the transition matrix. Proving (15) with $\pi^* = \pi$ is not hard. Let $m \in N(n)$ be any element of the neighborhood of n. Then by construction

$$P_{n,m} = \begin{cases} \dfrac{1}{\|N(n)\|}\,\alpha(n,m) & \text{if } m \neq n \\ \dfrac{1}{\|N(n)\|}\,\alpha(n,n) + \dfrac{1}{\|N(n)\|}\displaystyle\sum_{k\neq n}(1-\alpha(n,k)) & \text{if } m = n \end{cases}.$$

If $m \neq n$ in (15), assume w.l.o.g. that $\alpha(n,m) < 1$ and notice that this implies that $\alpha(m,n) = 1$. Then

$$\pi(n)P_{n,m} = \pi(n)\frac{1}{\|N(n)\|} \times \frac{\pi(m)\|N(n)\|}{\pi(n)\|N(m)\|}$$

and

$$\pi(m)P_{m,n} = \pi(m)\frac{1}{\|N(m)\|}$$

so that (15) is satisfied and the claim that the stationary probabilities are indeed $\pi(n)$ is verified.

Remark: Implementation of our Metropolis-Hastings algorithm still leaves the choice of the neighborhoods $N(n), n \in \mathcal{S}$, which significantly influences the efficiency of the resulting estimator.

A sequential version of the Metropolis-Hastings algorithm that uses algorithm can be implemented as explained in Gilks, Richardson, and Spiegelhalter (1996): when calculating X_{k+1} from a value X_k, the updates are performed one component at a time. Let

$$X_k^{-j} = (X_{k+1}(1), \ldots, X_{k+1}(j-1), X_k(j+1), \ldots, X_k(K))$$

be the current vector at the j-th stage of iteration k, when component $j = 1, \ldots, K$ is to be updated: all previous components $i < j$ have their new value while all components with $i > j$ are yet to be updated. Then generate a new jth component from the conditional target distribution,

$$q_j(m|X_k^{-j}) = \pi(m|X_k^{-j}).$$

Once the K stages of the k-th iteration are finished, the resulting value X_{k+1} is the next state of the Markov chain. This is the Gibbs sampler, and it is a variant of the Metropolis Hastings algorithm that uses a distribution $q(\cdot)$ to draw the next sample, instead of a uniform distribution. Under this interpretation, the new generated sample is always accepted, as explained in Chapter 7, Brémaud (1999).

Algorithm 3: Sequential Gibbs Sampler

1. For each $j = 1, \ldots, K$ do:
   (a) Generate $X_{k+1}(j) \sim \pi(\cdot|X_k^{-j})$
2. $k \leftarrow k + 1$, go to 1.

The ensuing process $\{X_k\}$ also possesses stationary distribution π. Irreducibility of the finite state chain ensures the existence of a unique ergodic measure for this example. Suppose now that $X_k \sim \pi(\cdot)$ has the target distribution π given in (13). Then by construction

$$\begin{aligned} \mathsf{P}[X_{k+1} = m] &= \sum_{n \in \mathcal{S}} \pi(n)\, \mathsf{P}[X_{k+1} = m | X_k = n] \\ &= \sum_{n \in \mathcal{S}} \pi(n) \prod_{j=1}^{K} \mathsf{P}[X_{k+1}(j) = m_j | X_k^{-j} = (m_1, \ldots, m_{j-1}, n_{j+1}, \ldots, n_K)] \\ &= \pi(m) \end{aligned}$$

which can be shown by proving that the sum above is proportional to $\rho_i^{m_i}/m_i!$.

As in the case of A/R, if $Y_s =$ is given by (14) then $\lim_{s\to\infty} \mathsf{E}[Y_s] = B$. Once the chain $\{X_s\}$ has been simulated for a "warm-up" stage, $X_{-k_0}, \ldots, X_{-1}$, the distribution of the chain is nearly stationary. Sample averages of $\{Y_s, s \geq 0\}$ can then be used to derive a confidence interval for B (see the Appendix). To achieve variance reduction, a variant of the method, called FGS, was studied in Vázquez-Abad and Andrew (2000) and Lassila and Virtamo (1998a). Instead of (14) it uses:

$$\hat{Y}(S) = \frac{R}{S} \sum_{k=1}^{S} \left(\frac{\lambda_{\sigma_k}}{\lambda} \right) \mathsf{P}[X_{k+1} \in \mathcal{B}_{\sigma_k} | X_k], \tag{16}$$

with $\sigma_k = k \text{mod}(K)$, where $\text{mod}(K)$ takes values $1, \ldots, K$.

Load per cell, ρ_i	7 cells	19 cells	37 cells
14	0.42	2.4	13
16	4.2	69	7000
18	17	16660	5.3e08

Table 1: Average number of candidate states N rejected per accepted state.

The FGS algorithm becomes very simple to implement. Define:

$$P_c(-1) = 0, \quad P_c(i) = \sum_{n=0}^{i} \frac{\rho_c^n}{n!}, \text{ for all } i \leq \Lambda.$$

Algorithm 4: Filtered Gibbs Sampler

1. For each $k = 1, \ldots, N$ do:
   (a) Set $j = \sigma_k$
   (b) Generate $X_{k+1}(j) \sim \pi(\cdot | X_k^{-j})$
   (c) Set $c = \Lambda - \max_{i: j \in C_i, i \neq j} \sum_{k \in C_i} n_k$
   (d) $Y_k = (P_c(c) - P_c(c-1))/P_c(c)$
2. Evaluate weighted average $\frac{1}{S} \sum_{k=1}^{S} \left(\frac{\lambda_{\sigma_k}}{\lambda} \right) Y_k$

4.3 Numerical Results

In order to evaluate the methods described above, numerical experiments were carried out to determine the relative efficiency using $S = 1000$ samples, $\mathcal{E}_r(\hat{Y}(S))$. In all cases, a batch size of $\beta = 1000$ was used, to facilitate comparison (see Appendix). These experiments were conducted on a 266 MHz Pentium II processor using the gnu C++ compiler under the Linux operating system. In the simulations that follow, the arrival rate was the same over all of the cells, so that $\rho_i \equiv \rho$ is constant. The clique limit was $C = 50$, typical of Advanced Mobile Phone System (AMPS) networks (Lee, 1995).

Acceptance/rejection: For efficiency, after each component M_i was generated, the sample was immediately rejected if any clique constraints (1) were violated, without generating M_j, $j > i$.

Table 1 shows the super-exponential increase in the average number of candidate states N rejected by the acceptance/rejection method.

Gibbs Sampler: For this problem, the conditional distribution $\pi(\cdot | X_k^{-i})$ is a one dimensional truncated Poisson random variable on $\{0, \ldots, C_k(i)\}$, with:

$$C_k(i) = C - \max_{c_j \ni i} n^{(c_j)} + n_i \tag{17}$$

A look-up table, P, of the cumulative probabilities of the Poisson distribution from 0 to C can be used to generate $M \sim \pi(\cdot | X_k^{-i})$. Let

$$P_m(j) = \sum_{n=0}^{m} \frac{\rho_j^n}{n!} \times \left(\sum_{n=0}^{C} \frac{\rho_j^n}{n!} \right)^{-1}, \quad m = 1, \ldots, C. \tag{18}$$

The distribution of a Poisson random variable M truncated at $c < C$ satisfies $\mathsf{P}[M \leq m] = P_m(j)/P_c(j), m = 0, \ldots, c$, by which M can easily be generated.

Step 1(a) of Algorithm 3 thus becomes

i) Set $c = C_k(j)$ as in (17),

ii) Generate $U \sim U[0,1]$ and set $n_j = \min\{m : P_m(j) \geq U\, P_c(j)\}$, using (18),

iii) Set $X_{k+1}(j) = n_j$.

The relative efficiency of the acceptance/rejection and Gibbs sampler methods on hexagonal grids of 7, 19 and 37 cells are shown in Figure 3 as a function of the load. As the blocking probability or size of the grid increases, the relative efficiency of the acceptance/rejection algorithm initially increases, but then degrades dramatically. The initial increase is due to the difficulty in estimating small blocking probabilities, which is the focus of Sections 5 to 7. As the load decreases, the reduction in the number of observed blocking events degrades efficiency. The degradation at high loads is because the CPU time required is approximately inversely proportional to the probability that a state is feasible. For small blocking probability, this is approximately unity even for small networks, but as the blocking probability or network size increases, the probability decreases rapidly.

When the probability of a state being infeasible is negligible, the Gibbs sampler performs similarly to the acceptance/rejection method, since the only overhead required is determining the occupancy of the (six or fewer) cliques containing each newly generated cell. Consecutive samples are approximately independent since the truncation of the Poisson distribution to preclude infeasible states is negligible. As the blocking probability increases, the correlation between successive samples increases, but this effect is dominated by the variance reduction mentioned above.

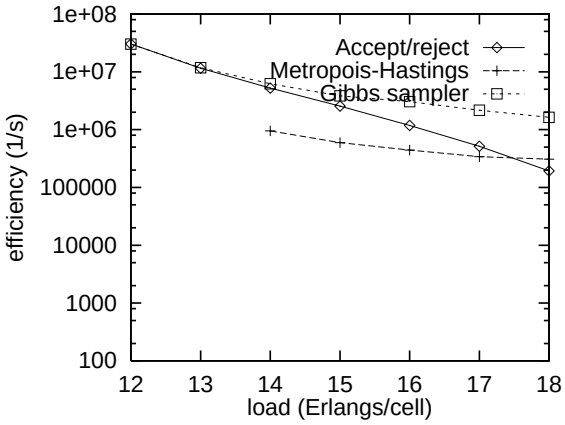
Let $\mathcal{P}$ be the probability that the candidate state N is feasible under acceptance/rejection. The dependence of $\mathcal{P}$ on load, capacity and network size is qualitatively captured by

$$P(\text{feasible}) \sim (1 - V)^K,$$

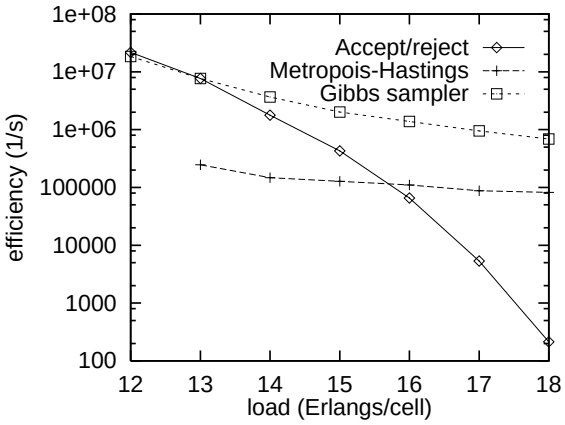
where K is the number of cells in the network and V is the probability that a given cell will be "excessively" loaded, which depends on the load and the capacity. The impact of this can be seen in Figure 4, which shows the relative efficiency of each estimator against the blocking probability, 7-, 19- and 37-cell grids. In contrast, the computational time of the Gibbs sampler grows linearly with network size.

5 Model for Fast Simulation

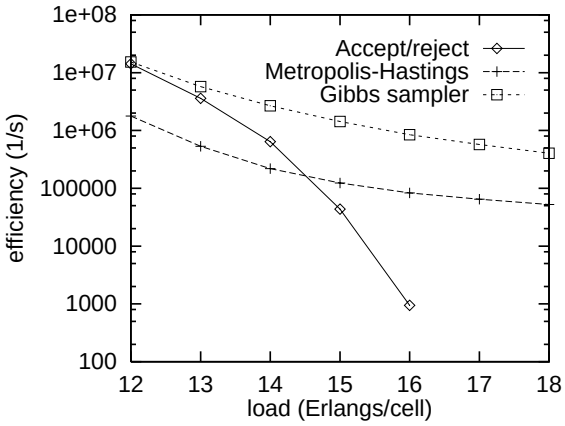
Estimation of blocking probabilities can be especially difficult in the case that blocking is studied as a rare event, where the MCMC techniques of the previous section will not be appropriate. The following sections address fast simulation of blocking probabilities for two regimes under which blocking is a rare event: low utilisation and high capacity. For both problems we use the same simulation model, described shortly.



(a) 7 cells

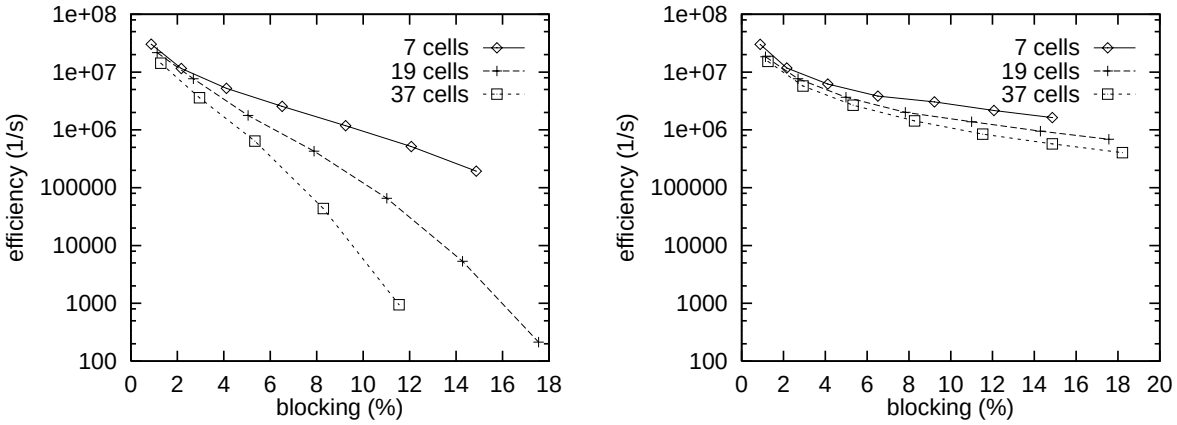


(b) 19 cells



(c) 37 cells

Figure 3: Relative efficiency vs. load for three algorithms and different grid sizes.



(a) Accept/reject

(b) Gibbs sampler

Figure 4: Relative efficiency vs. blocking probability for three grid sizes and different algorithms.

As stated in Section 2, the blocking probability can be expressed in terms of the blocking time within A-cycles, that is:

$$B_i = \frac{\mathsf{E}[X_i(T^{(i)})]}{\mathsf{E}(T^{(i)})},$$

for each cell i, where $T^{(i)}$ is the length of the A-cycle and $X_i(T^{(i)}$ is the total time spent in the blocking set $\mathcal{B}_i$ within the cycle. As in (6), $\mathcal{R}_i$ is the event that the blocking set

$\mathcal{B}_i$ is reached before the next A-cycle starts, and therefore:

$$B_i = \frac{\mathsf{E}[X_i(T^{(i)})|\mathcal{R}_i]\,\mathsf{P}(\mathcal{R}_i)}{\mathsf{E}(T^{(i)})} \tag{19}$$

For each cell i, a change of measure is performed to build the estimator for B_i. Fix the index i throughout the remaining sections. An A-cycle is defined here by the set A_i:

Definition 6 *Let the quasi-regenerative set A_i be set of all states in which the cluster occupancy, $n^{(\mathcal{C}_i)}$, satisfies $n^{(\mathcal{C}_i)} > \theta_i$ for some threshold $0 \leq \theta_i < C$.*

The optimal choice of θ_i will be considered in Section 7.2.2.

The following sections will derive changes of measure which are asymptotically optimal for estimating $\mathsf{P}(\mathcal{R}_i)$, the probability that blocking occurs within an A-cycle associated with an arbitrary cell i of the cellular network model. Fast simulation via Importance Sampling (IS) requires simulating the process with cluster acceleration *only until blocking is arrived at*. Once the state enters $\mathcal{B}_i$ one computes the time that the state remains blocked within the current A-cycle, or $\mathsf{E}[X(T^{(i)})|\mathcal{R}_i]$. To estimate the numerator of (19), the process is accelerated until blocking occurs, that is, up to time τ_i, while the standard measure will be used to resume the simulation and to estimate the time spent in blocking states within the cycle. As well, the original process is simulated without changing the measure in order to estimate the denominator. This technique is common in queueing network simulation, as has been implemented by Chang, Heidelberger, and Shahabuddin (1995), L'Ecuyer and Champoux (1996).

Let $X(t)$ be the occupancy process of the *cluster* $\mathcal{C}_i$ of all the cliques where i belongs. Use the aggregate arrival process at this cluster of surrounding cells, which is Poisson with intensity $\lambda \equiv \sum_{j \in \mathcal{C}_i} \lambda_j$. The original process is modeled now using this arrival process to generate the inter-arrival times $A_k(i) \sim \exp(\lambda)$. Under this relabeling, $S_k(i)$ is the epoch of the k-th *local* arrival at cluster $\mathcal{C}_i$. Next, at arrival epochs $S_k(i)$, the cell to which the arrival to the cluster is assigned is chosen to be cell $j \in \mathcal{C}$ with probability λ_j/λ. Holding times $H_k(i) \sim \exp(\mu)$ are i.i.d. exponential, as before. These quantities correspond to the actual holding times of the k-th call to be connected within the cluster, regardless of the particular cell $j \in \mathcal{C}_i$ where it happens to belong. If the k-th arriving call to the cluster is blocked we set $H_k(i) = 0$ by convention (the call is not connected), rather than queueing the customer for future service. The process $X(t)$ can be seen as a queueing model with parallel servers, but it behaves very differently from the $GI/G/s/n$ model that was treated in Section3.4. The latter can be viewed as a truncation of a $GI/G/s/\infty$ queue, while the former is a truncation of an *infinite* server queue.

In many queueing studies the discrete event model is used for simulation as well as for IS estimation. The discrete event simulation model for the process considers the generation of inter-arrival and holding times as needed, that is, $U_k(i) = (A_k(i), H_k(i), D_k(i))$ where $D_k(i) \in \mathcal{C}_i$ is the cell where the arrival occurs within the cluster. Let $U_k(j) = (A_k(j), H_k(j))$ be the inter-arrival and holding times for cells $j \notin \mathcal{C}_i$. The natural filtration is given by:

$$\mathfrak{F}_k^{DE} = \sigma(U_1(i), \ldots, U_{k(i)}(i); U_1(j), \ldots, U_{k(j)}(j); j \notin \mathcal{C}_i); \quad k = 1, 2, \ldots \tag{20}$$

where exactly $k(i)$ arrivals have occurred within the cluster $\mathcal{C}_i$ and $k(j)$ arrivals at cell $j \notin \mathcal{C}_i$ at the epoch of the k-th global arrival. The embedded occupancy process $\{n(k)\}$ can be evaluated from the information in $\mathfrak{F}_k^{DE}$ and is therefore adapted to the filtration (20), as required.

The usual problem for these systems is that the acceleration affects several holding times even after blocking has been detected and may slow down termination of the cycles. Sadowsky (1991) stops the change of measure for the service times from the arrival $(\tau - n)$ of the customer that will be in service when customer τ

arrives finding n in waiting. Although this is mathematically convenient, it turns out to be impossible to simulate using the above DES model because τ_i is not measurable w.r.t. $\mathfrak{F}_k^{DE}, k < \tau_i$: at the time of arrivals of the customers presently in the queue when τ finds blocking, this fact is unknown. Devetsikiotis and Townsend (1993) propose to use another change of measure from the moment that blocking occurs, in order to *empty* the occupancy quickly for fast termination of the A-cycle. L'Ecuyer and Champoux (1996) seek to find a smaller stopping time in the hope of stopping the acceleration before blocking occurs. Instead we propose to do it using a different simulation model which in addition is

much simpler to code than the Discrete Event model.

The simulation model is the Standard Clock technique of Vakili (1991), which corresponds to the dynamical description of a multidimensional birth and death process. When the total occupancy of the current state is n, an exponential random variable with intensity Λ_n is used to determine the *inter-event time*, or the time for the next event. Here

$$\Lambda_n = \sum_j \lambda_j + n\mu.$$

Next, the event *type* is determined as a discrete random variable with distribution:

$$D = \begin{cases} a_j : \text{arrival at cell } j & \text{w.p. } \dfrac{\lambda_j}{\Lambda_n} \\ s_j : \text{termination of call at cell } j & \text{w.p. } \dfrac{n_j \mu}{\Lambda_n} \end{cases}$$

where n_j is the number of calls at cell j, so that $n = \sum_j n_j$.

The appropriate filtration for this process is

$$\mathfrak{F}_k = \sigma(T_1, \ldots, T_k; D_1, \ldots, D_k), \quad k = 1, 2, \ldots \tag{21}$$

where T_k are the successive inter-event times and D_k the corresponding event types or "decisions". The embedded occupancy process is updated by setting

$$\begin{aligned} n_j(k+1) &= \begin{cases} n_j(k) + 1 & \text{if } n(k) \notin \mathcal{B}_i \text{ and } D_{k+1} = a_j, \\ n_j(k) - 1 & \text{if } D_{k+1} = s_j \end{cases} \\ n_k^{(\mathcal{C}_i)} &= \sum_{j \in \mathcal{C}_i} n_j(k), \quad \text{total cluster occupancy} \end{aligned}$$

Work has been done on applying importance sampling to other problems though uniformization (Heidelberger, Shahabuddin, and Nicola, 1994), which creates events a a maximal rate $\bar{\Lambda}$ independent of the state, and then adjusts the probabilities by adding fictitious, or "null" events. Use of the SC model for simulation may be more efficient because there is no need for simulating null events. To emphasize that we apply Importance Sampling to the Standard Clock simulation model, use the acronym ISSC.

Remark: Notice that, although the two models describe the same process (in distribution), the partial histories $\mathfrak{F}_k^{DE}$ and $\mathfrak{F}_k$ are not equivalent. Filtration (20) contains filtration (21) but clearly, H_k is not measurable w.r.t. $\sigma(T_1, \ldots, T_k; D_1, \ldots, D_k)$. If a particular change of measure (such as rate-swapping) is BRE for the DES formulation, it does not necessarily mean that it is also BRE for the SC formulation and *vice versa*, however for some models this may well be the case.

6 Static ISSC Estimation for Light Traffic

6.1 Change of Measure

In the light traffic regime, assume that $\lambda_i = k_i \epsilon, i = 1, \ldots K$. Unlike the $GI/G/s/\infty$ case, (11) need not hold for each channel on the trajectories where blocking occurs before the A-cycle ends and swapping rates as in (12) will not be optimal. Instead, consider the change of measure that swaps aggregate arrival rates per cluster and inverse holding times.

Proposition 1 *Consider the ISSC simulation model with initial state as the start of an A-cycle. Arrivals at the cluster $\mathcal{C}_i$ have rate $\lambda^* = \mu$ and holding times for the calls in the cluster have*

rate $\mu^* = \lambda$. Other inter-arrival and holding times (outside the cluster) have the original exponential distribution. Call the underlying measure P^*. Then:

$$\mathsf{P}(\mathcal{B}_i) = \mathsf{E}^* \left[e^{-(\mu-\lambda)\sum_{k=\theta+1}^{\tau_i-1}(n^{(\mathcal{C}_i)}(k)-1)T_{k+1}} \left(\frac{\lambda}{\mu}\right)^{k_1-k_2} \mathbf{1}_{\{\mathcal{R}_i\}} \right],$$

where $k_1(j)$ is the total number of arrivals to cell j up to event number τ_i (including blocked calls), $k_1 = \sum_{j\in\mathcal{C}_i} k_1(j)$ and k_2 is the corresponding number of call completions (excluding blocked calls).

Proof : Given the history of the process up to the k-th event (that is, given $\mathfrak{F}_k$) $T_{k+1} \sim \exp(\Lambda^*_{n(k)})$, for $n(k) = n_1(k) + \ldots + n_K(k)$ the total occupancy of the process after event k has occurred. The new event rate is a $\mathfrak{F}_k$-measurable random variable:

$$\Lambda^*_{n(k)} = \mu + \sum_{j\notin\mathcal{C}_i} \lambda_j + \lambda \sum_{j\in\mathcal{C}_i} n_j(k) + \mu \sum_{j\notin\mathcal{C}_i} n_j(k),$$

and the event types now follow:

$$D_{k+1} = \begin{cases} \text{arrival at cell } j \in \mathcal{C}_i & \text{w.p. } \dfrac{\lambda_i}{\lambda}\dfrac{\mu}{\Lambda^*_n} \\ \text{arrival at cell } j \notin \mathcal{C}_i & \text{w.p. } \dfrac{\lambda_j}{\Lambda^*_n} \\ \text{termination of call at cell } j \in \mathcal{C}_i & \text{w.p. } \dfrac{n_j\lambda}{\Lambda^*_n} \\ \text{termination of call at cell } j \notin \mathcal{C}_i & \text{w.p. } \dfrac{n_j\mu}{\Lambda^*_n} \end{cases}.$$

Within an A-cycle, we perform this change of measure until the τ_i-th event occurs, which is the first time that the state is in $\mathcal{B}_i$. Because the

distributions of T_{k+1} and D_{k+1} are conditional on $\mathfrak{F}_k$, the change of measure defines a multiplicative martingale which is adapted to the filtration $\{\mathfrak{F}_k, k \geq 1\}$, and the corresponding Radon-Nikodym derivative is given by:

$$\begin{aligned} L_{\tau_i} \;=\; & \prod_{k=1}^{\tau_i-1} \left(\frac{\Lambda_{n(k)}}{\Lambda^*_{n(k)}}\right) e^{(\Lambda^*_{n(k)}-\Lambda_{n(k)})T_{k+1}} \\ & \times \prod_{k=1}^{\tau_i-1} \left(\frac{\Lambda^*_{n(k)}}{\Lambda_{n(k)}}\right) \left(\sum_{j\in\mathcal{C}_i} \left(\frac{\lambda_j}{\mu(\lambda_j/\lambda)}\right) \mathbf{1}_{\{D_{k+1}=a_j\}} + \left(\frac{\mu}{\lambda}\right) \mathbf{1}_{\{D_{k+1}\in\mathcal{S}_i\}} + \mathbf{1}_{\{D_{k+1}\notin\mathcal{S}_i\cap\mathcal{A}_i\}} \right) \end{aligned} \tag{22}$$

where $\mathcal{A}_i = \{a_j : j \in \mathcal{C}_i\}$ is the set of event types which are arrivals to the cluster and similarly, $\mathcal{S}_i = \{s_j : j \in \mathcal{C}_i\}$ is the set of event types which are termination of calls within the cluster. From their definitions, it follows that:

$$\Lambda^*_n - \Lambda_n = (\mu-\lambda) + (\lambda-\mu)\sum_{j\in\mathcal{C}_i} n_j = -(\mu-\lambda)(n^{(\mathcal{C}_i)} - 1)$$

where $n^{(\mathcal{C}_i)} = \sum_{j\in\mathcal{C}_i} n_j$ is the total occupancy of the cluster.

Simplifying the expression above,

$$\begin{aligned} L_{\tau_i} &= e^{-(\mu-\lambda)\sum_{k=\theta+1}^{\tau_i-1}(n^{(\mathcal{C}_i)}(k)-1)T_{k+1}} \left[\prod_{j\in\mathcal{C}_i}\left(\frac{\lambda}{\mu}\right)^{k_1(j)}\right]\left(\frac{\mu}{\lambda}\right)^{k_2} \\ &= e^{-(\mu-\lambda)\sum_{k=\theta+1}^{\tau_i-1}(n^{(\mathcal{C}_i)}(k)-1)T_{k+1}}\left(\frac{\lambda}{\mu}\right)^{k_1-k_2}, \end{aligned} \tag{23}$$

which proves the claim. ◁

Lemma 2 *When $\lambda < \mu$,*

$$L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}} < (\lambda/\mu)^{C-\theta}, \mathsf{P}^* - \textit{w.p.1.} \tag{24}$$

In particular, $L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}} < 1, \mathsf{P}^* -$ w.p.1., guaranteeing variance reduction of the Importance Sampling estimator with the Standard Clock (ISSC).

Proof : Clearly $\lambda_j < \lambda = \sum_{j\in\mathcal{C}_i}\lambda_j$, and there are at least C on-going calls within the cluster at the blocking time τ_i. At the time the A-cycle begins, there are (by Definition 6) θ calls within the cluster. Thus acceleration begins when there are $\theta+1$ calls in the cluster. Because the stopping time within an A-cycle counts only the transitions from the start of the A-cycle up until blocking, on the set $\{\omega : \tau_i < T^{(i)}\}$ we have $n^{(\mathcal{C}_i)}(k) \geq \theta+1, k = 1,\ldots,\tau_i$, whence

$$(\mu-\lambda)\sum_{k=1}^{\tau_i-1}(n^{(\mathcal{C}_i)}(k)-1)T_{k+1} \geq 0. \tag{25}$$

Moreover, $k_1 = \sum_{j\in\mathcal{C}_i}k_1(j) \geq k_2 + C - \theta$, giving the result. ◁

Theorem 2 *The ISSC estimation that swaps the rates λ and μ is BRE for $\mathsf{P}[\mathcal{R}_i]$ as $\epsilon \to 0$, when $\lambda_i = k_i\epsilon$, for all cells i and $\theta_i = 0$.*

Proof : The proof is an application of Lemma 1. The upper bound is obtained with the result of Lemma 2:

$$L_{\tau_i} \leq \left(\frac{\lambda}{\mu}\right)^C = d_2\epsilon^C, \tag{26}$$

where $d_2 = \sum_{j\in\mathcal{C}_i}k_j$.

It remains to show $p(\epsilon) \geq d_1\epsilon^C$. In order for blocking of cell i arrivals to occur, the occupancy of one of the cliques in the cluster must necessarily attain level C. Let a "minimal path" be a trajectory in which the first C events in an A-cycle which involve the accelerated cluster are arrivals to the same clique. All minimal paths will cause blocking, and thus their probability is a lower bound for $p(\epsilon)$. The probability of such minimal paths is the probability that each of the first C events be an arrival, and so:

$$p(\epsilon) \geq \left(\frac{\bar{\lambda}_i}{\Lambda_n}\right)^C \geq \left(\frac{\bar{k}_i\epsilon}{C\mu}\right)^C \geq d_1\epsilon^C, \tag{27}$$

where $\bar{\lambda}_i$ is the smallest aggregate clique rate within cluster $\mathcal{C}_i$ and $d_1 = (\bar{k}_i/C\mu)^C$, with $\bar{k}_i$ the smallest of $\sum_{s\in c_j}k_s$ over the cliques $c_j \in \mathcal{C}_i$.

It follows by Lemma 1 that in this case, ISSC is BRE. ◁

Note that the condition $\theta_i = 0$ is not a major limitation, since the modal cluster occupancy will be zero, and so $\theta_i = 0$ gives the shortest A-cycles, which is desirable for simulations.

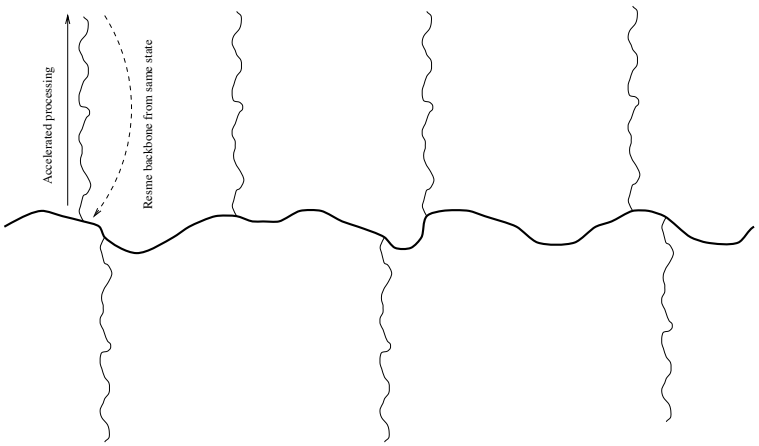


Figure 5: "Backbone and ribs" simulation framework.

6.2 Numerical Results

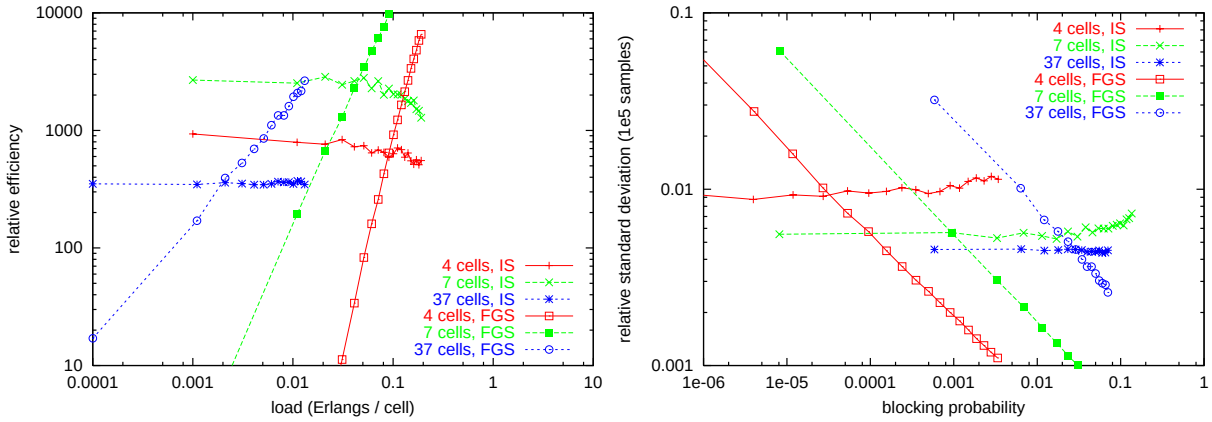
An A-cycle must start with the state distribution being the equilibrium distribution conditional on being on the boundary between A and A'. In the absence of importance sampling, this will also be the distribution at

the end of the A-cycle. That means that multiple A-cycles can be simulated simply by simulating the system for a long period, and marking off the A-cycles. However, importance sampling disturbs the distribution, and the distribution of the state at the end of an A-cycle cannot be used as the starting state for the next A-cycle. Instead of this, when the start of an A-cycle is detected, the simulation will be suspended, and a second simulation will be started from the current state n. This second simulation will be accelerated until blocking occurs (or the current A-cycle ends) and the original measure is used to determine the time spent in blocking states until the end of the current A-cycle. The original simulation is then recommenced from state n, until the start of the next A-cycle. This gives rise to the "backbone and ribs" arrangement, seen in Figure 5. A second advantage of this approach is that the length of an A cycle, as required by (4), may be estimated more accurately (with lower variance) from the unaccelerated A-cycles.

Each A-cycle only estimates the blocking probability in a single cell. Thus in order to estimate the blocking probability in each cell, it is necessary to run separate simulations for each cell. Fortunately, it is not necessary to simulate the "backbone" separately. Instead, separate "ribs" can be started each time the backbone starts an A_i-cycle for *any* cell i. Note that a single event may be the start of A_i-cycles for more than one i, and each of these must be considered.

The ribs may be further simplified by noting that the set A_i' contains no blocking states for cell i. Thus, once an accelerated A_i-cycle has left the set A_i, the entire time spent in blocking states will already have occurred. This means that the rest of the A_i-cycle need not be simulated, even though a considerable amount of time may be spent in the set A_i' before the A_i cycle is truly over.

The ISSC with $\theta_i = 0$ for $\rho \to 0$ was tested on three sizes of network: 4, 7 and 37 cells. Figure 6 compares these results with those of the filtered Gibbs sample of Section 4.2. The results clearly show that the relative error of the estimated network blocking probability is



(a) efficiency against load

(b) error against blocking

Figure 6: Relative efficiency and relative error for importance sampling (IS) and filtered Gibbs sampler (FGS) for light loads

bounded as $\rho \to 0$.

7 Dynamic ISSC Estimation for High Capacity

7.1 Change of Measure

It is unlikely that future networks will be operated at extremely low utilisation, as was assumed in the previous section. Engineers are more interested in the behaviour as the capacity increases. This is particularly true of wavelength continuous WDM trunk networks, which are mathematically analogous to the cellular networks described so far (see Andrew and Vázquez-Abad, 2001). It is possible to overcome the restriction that $\lambda < \mu$ in the previous section by allowing the arrival and service rates to be state dependent.

This section addresses the simulation of the important regimes of $C \to \infty$ with ρ constant, and $C \to \infty$ with ρ/C constant, both of which cause blocking to become a rare event. Swapping λ and μ as in the previous section will not yield BRE for high capacity regimes (even when $\lambda < \mu$) because as $C \to \infty$ the lower bound on $\mathsf{P}[\mathcal{R}_i]$ goes to zero, and blocking does not become more likely under the star measure P^*. A change of measure which is optimal in the case of a C parallel server system with no waiting will be presented, and applied to the cellular case. It is suboptimal in this latter context, but it provides a dramatic improvement over simulation using the original measure.

Consider again the Standard Clock simulation model and let the change of measure be such that the total event rate is the same as for the original measure: $\Lambda^*(n) = \Lambda(n)$, when the state is n. When an event occurs, it is a departure (or an arrival) to cell $j \notin \mathcal{C}_i$ with probability λ_j/Λ_n (equivalently, $n_j\mu/\Lambda_n$) just as for the original measure. Arrivals (departures) to the cluster will now occur with probability $\lambda^*(n)/\Lambda_n$ (or $\mu^*(n)/\Lambda_n$, respectively). The proportion of arrivals to the cluster that go to cell $j \in \mathcal{C}_i$ remains fixed at λ_j/λ. Under this change of measure, the rates are no longer constant, but depend on the state, hence the name "dynamic ISSC".

It is straightforward to calculate now the Radon-Nikodym derivative:

$$L_{\tau_i} = \prod_{k=1}^{\tau_i - 1} \left(\sum_{j \in \mathcal{C}_i} \left(\frac{\lambda}{\lambda^*(n(k))} \right) \mathbf{1}_{\{D_{k+1} = a_j\}} + \left(\frac{\mu}{\mu^*(n(k))} \right) \mathbf{1}_{\{D_{k+1} \in \mathcal{S}_i\}} + \mathbf{1}_{\{D_{k+1} \notin \mathcal{S}_i \cap \mathcal{A}_i\}} \right),$$

independent of the inter-arrival times.

Specifically, if $n = n^{(\mathcal{C}_i)}(k)$ is the *cluster* occupancy of the process at event k, let the rates be defined by the recurrence relation:

$$\mu^*(n) = \frac{\lambda\mu}{\lambda^*(n-1)} \tag{28a}$$

$$\lambda^*(n) = \lambda + n(\mu - \mu^*(n)). \tag{28b}$$

Notice that the rates under the star measure (28) only depend on the cluster occupancy, rather than the whole state of the process $n(k)$. This is related to swapping the arrival and service rates of (12) in the following sense.

Lemma 3 *Under the update rule (28), for any initial* $0 < \mu^*(1) < \mu$*,*

$$\lim_{n \to \infty} \lambda^*(n)/n = \mu \tag{29a}$$

$$\lim_{n \to \infty} \mu^*(n)n = \lambda. \tag{29b}$$

Proof : We will first show that $\mu^*(n) \to 0$ as $n \to \infty$. By induction, $0 < \mu^*(n) < \mu$ for all $n \geq 1$, and hence $\{\mu^*(n)\}$ has a convergent subsequence. To see that μ is not an accumulation point, note that this would imply $\mu^*(n) = \mu - \phi(n)$ for some $\phi(n) = o(1)$, $\phi(n) \not\equiv 0$. Then by (28a), $\lambda^*(n) - \lambda \sim \phi(n)$, but by (28b), $\lambda^*(n) - \lambda \sim n\phi(n)$, which is a contradiction. Thus there is a strictly increasing sequence $n(m) \in \mathbb{N}$ and a $\mu^* \in [0, \mu)$ such that $\mu^*(n(m)) \to \mu^*$ as $m \to \infty$. Thus there is a $\delta = \mu - \mu^* > 0$ such that

$$\lambda^*(n(m+1)) - \lambda^*(n(m)) = (n(m+1) - n(m))(\mu - \mu^*) + o(1) > \delta + o(1).$$

Thus $\lambda^*(n(m)) \to \infty$, and by (28a), $\mu^*(n(m)) \to \mu^* = 0$. Since the sequence $m(i)$ was arbitrary, 0 is the unique accumulation point, and $\mu^*(n) \to 0$ as $n \to \infty$.

By (28b) and the convergence of $\{\mu^*(n)\}$,

$$\lambda^*(n+1) - \lambda^*(n) = \mu + o(1).$$

Thus $\lambda^*(n) = \lambda^*(1) + \sum_{k=2}^{n} (\mu + o(1)) = n\mu + o(n)$, and $\lambda^*(n)/n \to \mu$ as $n \to \infty$. The result then follows from (28a). ◁

This asymptotic form is analogous to the static change of measure for an $M/M/n/\infty$ queue. An $M/M/s/\infty$ queue has a maximum of s servers active in the limit, and so the change of measure is static in that case. The change of measure of (28) is dynamic, reflecting the fact that the number of active servers increases continually as the size of the system increases.

Theorem 3 *Consider a birth and death process with rates* $\lambda(n) \equiv \lambda, \mu(n) = n\mu$ *on* $\{\theta, \ldots, C\}$*. Let* τ *be the first hitting time of* C*,* T *the first return to state* θ *and* $\mathcal{R} = \{\tau < T\}$*. Then the ISSC estimator for* $\mathsf{P}(\mathcal{R})$ *using the dynamic rates in (28) starting with* $\mu^*(\theta + 1) = 0$ *is asymptotically optimal in the limit of* $C \to \infty$*. Moreover, it is exactly optimal: the variance of the estimate is zero even for finite* C*.*

Proof : With $\mu^*(\theta+1)=0$, $\mathsf{P}^*(\mathcal{R})=1$ since the A-cycle is not allowed to end until blocking occurs. This violates the absolute continuity condition that for every $x \in \mathcal{S}$, $\mathsf{P}(x) > 0 \Rightarrow \mathsf{P}^*(x) > 0$. However, as mentioned in Section 6, $\mathsf{P}\rfloor_{\mathcal{R}_i} << \mathsf{P}^*$. This follows because on the event $\mathcal{R}_i$, blocking occurs before the A-cycle is over, thus no trajectory on $\mathcal{R}_i$ can have a transition from $\theta+1$ back to θ (this would start a new A-cycle). Thus for every $x \in \mathcal{R}_i$, $\mathsf{P}(x) > 0 \Rightarrow \mathsf{P}^*(x) > 0$, and the change of measure is valid for the estimation of $\mathsf{P}(\mathcal{R}_i)$.

On any path leading to the blocking boundary C, any transition due to a call termination from state n to $n-1$ must necessarily be followed in some future stage by a matching transition from $n-1$ to n: otherwise it is impossible to achieve full occupancy. The corresponding factors contributing to L_τ are then:

$$\left(\frac{\mu(n)}{\mu^*(n)}\right)\left(\frac{\lambda(n-1)}{\lambda^*(n-1)}\right) = \left(\frac{\lambda(n-1)}{\lambda}\right)\left(\frac{\lambda}{\lambda^*(n-1)}\right) = 1,$$

therefore all such loops cancel out their contributions. The only remaining contributions to L_τ are the factors for the "minimal" blocking path $\theta+1 \to \theta+2 \to \theta+3 \to \ldots \to C$, which yields:

$$L_\tau = \prod_{k=\theta+1}^{C-1} \frac{\lambda}{\lambda + k(\mu - \mu^*(k))}, \tag{30}$$

which is a deterministic function of C, thus optimal. ◁

Note that for a fixed $\mu^*(\theta+1)$, the rates are independent of C. The optimal adaptivity to C comes from the fact that the rates change as the actual current occupancy changes.

In the cellular network case, we identify the birth and death process as the cluster occupancy process: while it remains true that to achieve blocking at a state $n_\tau \in \mathcal{B}_i$ all backward transitions from n to $n-1$ will cancel out from forward transitions from $n-1$ to n, now the ISSC estimator is:

$$L_{\tau_i} = \prod_{k=\theta}^{n_\tau^{(\mathcal{C}_i)}-1} \frac{\lambda}{\lambda + k(\mu - \mu^*(k))},$$

and the final cluster occupancy $n_\tau^{(\mathcal{C}_i)}$ satisfies $C \le n_\tau^{(\mathcal{C}_i)} \le mC$, where m is the number of cliques in the given cluster, which depends on the interconnectivity of the network. The variance of L_{τ_i} is thus dependent on the variation of the distribution of the cluster occupancy at blocking.

7.2 Implementation considerations

7.2.1 Subsampling the ribs

The correlation between consecutive A-cycles in the backbone can be very significant. In order to reduce this, it is possible to subsample the A-cycles, and only start a rib for every kth A-cycle. This greatly increases the amount of work required to simulate the backbone. However, an A-cycle which is blocked may be very much longer than a "typical" A-cycle, especially when the load is a small fraction of the number of channels, and so the backbone is often a small proportion of the simulation time. Moreover, the backbone is shared between many cells, making subsampling very worth while. Estimating the variance of a ratio can be

performed following Bratley, Fox, and Schrage (1987), when both numerator and denominator are sample averages of Markov processes with exponentially decaying covariances, instead of iid random variables. In this case, if $B_i = E[Y]/E[T]$, then the random variable obtained during the l-th A-cycle:

$$Z_l = kY_l \mathbf{1}_{\{l \bmod k=0\}} - B_i T_l$$

has zero expectation. Here Y_i represents the estimate of the numerator obtained from the l-th rib (of which only one out of k is used) and T_l is the estimate of the cycle length (which we called $T^{(i)}$ before). Then the estimator obtained with S consecutive A-cycles satisfies:

$$\mathsf{Var}[\hat{B}_i(S)] = \mathsf{Var}\left(\frac{\frac{k}{S}\sum_{l=1}^{S} Y_l \mathbf{1}_{\{l \bmod k=0\}}}{\frac{1}{S}\sum_{l=1}^{S} T_l}\right) \approx K_1 \,\mathsf{Var}\left(\frac{1}{S}\sum_{l=1}^{S} Z_l\right),$$

where K_1 is a constant depending on $E[T]^2$, but independent of the sample size S. The optimal value of k can be obtained by noting that

$$\mathsf{Var}\left(\frac{1}{S}\sum_{l=1}^{S} Z_l\right) \approx K_Y(k/S) + K_T/S,$$

where K_Y depends on $\mathsf{Var}[Y]$ and K_T depends on $B_i^2\mathsf{Var}[T]$. The covariance term decreases as S^2. On the other hand,

$$\text{CPU}[\hat{B}_i(S)] = l_T(k/S) + l_Y S,$$

where l_T and l_Y are the mean lengths of unaccelerated and accelerated A-cycles respectively. The efficiency (5) is then maximised by setting

$$k^* \approx \sqrt{\frac{K_Y\, l_Y}{K_T\, l_T}}. \tag{31}$$

These values can be estimated coarsely from a short pilot simulation, which can also serve as the "warm-up" to achieve steady state. When acceleration

is used, k^* is of the order of 10, to within one order of magnitude. Without acceleration, k^* is actually often less than 1, indicating that there may be value in running multiple ribs from the same point in the backbone.

7.2.2 Choice of quasi-regenerative cycles

When the load, λ/μ, is not negligible, the probability that the cluster will be completely empty is small. Thus if θ is too small, like $\theta = 0$ as is used for single server queues, then the A-cycles become unmanageably long. This has several implications. The most obvious result is that the simulation time increases in proportion. The seriousness of this is to some extent alleviated by the fact that longer A-cycles produce better estimates of the proportion of time spent in blocking states within an A-cycle.

The more serious problem with long A-cycles is that the blocking states become a small proportion of the A-cycles, even given that blocking occurs. The assumption behind the IS scheme proposed here is that A-cycles in which blocking occurs are rare events, but if blocking does occur, it does so for a significant proportion of the A-cycle. Thus the system is accelerated until the point when blocking first occurs, and is then allowed to relax back to finish its A-cycle.

If λ and C are both large, then empty clusters become rarer than blocking, and most A-cycles contain a period of blocking, reducing the effectiveness of the acceleration.

The length of A-cycles is minimised by maximising the rate of crossing the boundary between sets A and A' of the embedded Markov Chain. Since the stationary rate at which the process crosses from any state with $n^{(\mathcal{C}_i)} = \theta$ to $n^{(\mathcal{C}_i)} = \theta_i + 1$ equals the stationary rate at which it crosses from $n^{(\mathcal{C}_i)} = \theta_i + 1$ to $n^{(\mathcal{C}_i)} = \theta$. This rate is state dependent:

$$\frac{\lambda\pi(\theta_i)}{\Lambda_n}, \quad n \in \mathcal{S}(\theta_i)$$

where $\mathcal{S}(\theta_i)$ is the subset of the state space $\mathcal{S}$ that has $n^{(\mathcal{C}_i)} = \theta$ channels busy in cluster $\mathcal{C}_i$. Naturally, one seeks a unique value θ_i that maximises this rate, such as a maxmin optimiser. A simple observation however simplifies this task: for many systems with several cells, for the regime where both C and λ grow, the value of Λ_n will be nearly constant, and the value that maximises the numerator is the mode of the stationary distribution $\pi(\cdot)$, or $\theta_i \approx \sum_{j\in\mathcal{C}_i} \lambda_j/\mu$ (assuming that blocking does not significantly distort the state distribution, and that the mean and mode of π are very close).

Next, in order to guarantee that $L_{\tau_i} < 1$ when blocking occurs, it was required that $\theta_i \leq C$, since L_{τ_i} can exceed 1 whenever the number of departures exceeds the number of arrivals during the acceleration period. Thus we use the value

$$\theta_i = \max\left(\sum_{j\in\mathcal{C}_i} \lambda_j/\mu, C\right). \tag{32}$$

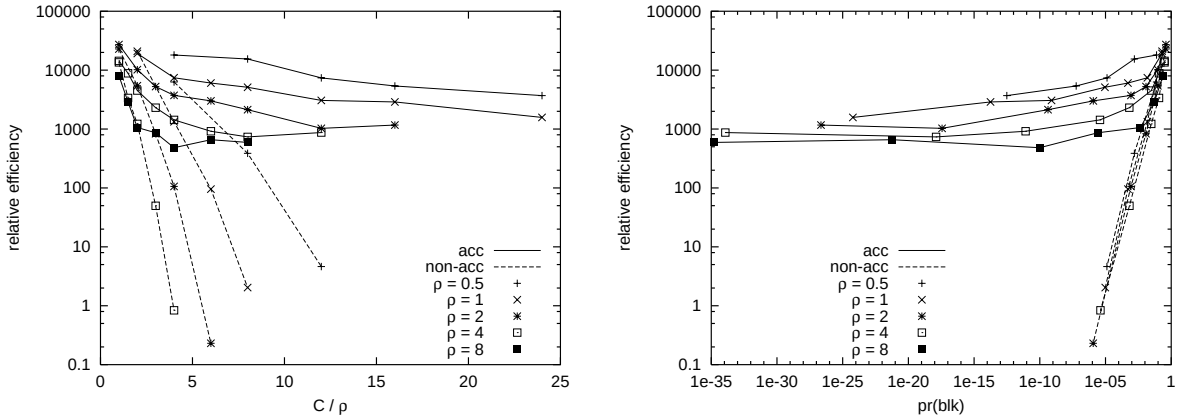
It is possible that in some instances the overall variance can be reduced by using $\theta_i > C$, since the probability of a greater number of departures than arrivals leading to blocking may be small. This is the subject of ongoing research. There may also be benefit in using values of θ_i larger than $\sum_{j\in\mathcal{C}_i} \lambda_j/\mu$

(but less than C). Although A cycles become longer, the time spent in the set A' ($n^{(\mathcal{C}_i)} < \theta_i$) need not be simulated in the "ribs", since blocking cannot occur in this case. It must still be simulated in the backbone to estimate the mean duration of an A-cycle, but this is shared between all cells.

7.3 Simulation results

The dynamic change of measure was shown to be a.o. for the probability of blocking occurring within an A-cycle, $\mathsf{P}(\mathcal{R}_i)$, in the case of a single cell. Figure 7 shows the relative efficiency for the true probability of blocking in this case. As C increases, the proportion of each time spent in a blocking state decays, even on those A-cycles which contain blocking. This accounts for the slight reduction in efficiency as the blocking rate decreases. However, this reduction is very much smaller than occurs without acceleration.

Figure 8 shows the relative efficiency of the accelerated and non-accelerated methods for a seven-cell cellular system, for a range of loads and a range of capacities (normalised by the load). The ISSC subsamples the A-cycles in the backbone by a factor of $k = 10$. The simulations without acceleration use $k = 1$, as the rib A-cycles are shorter and the variance of the corresponding estimator is higher. These results do not suggest that ISSC is a.o. for network



(a) efficiency against load

(b) efficiency against blocking

Figure 7: Relative efficiency for importance sampling (ISSC), A-cycle framework without IS, and the simple simulation, all in a single cell.

blocking as $C \to \infty$. However, IS substantially reduces the rate at which the performance degrades for large C.

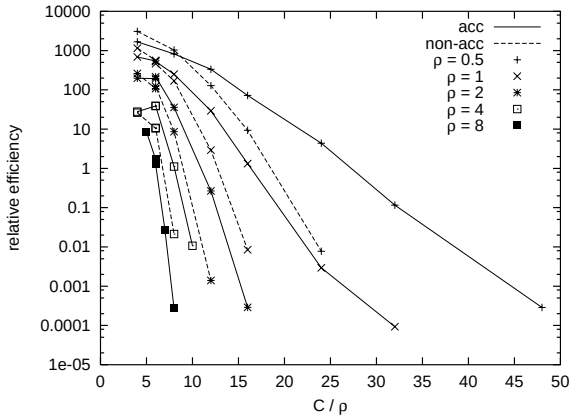
The reason for the reduced efficiency is that the acceleration is applied to all cells in a cluster. For a constant load, as C increases the (true)

expected cluster occupancy on blocking satisfies $\mathsf{E}[n^{(\mathcal{C}_i)}|\mathcal{B}_i]/C \to 1$, since arrivals at each cell are independent, and only one clique need be full. However, because the acceleration is applied to all cells in the cluster, the cells outside the clique which caused blocking are also filled up. Thus the expected cluster occupancy at blocking under the new measure is significantly larger than under the original measure. That is, outcomes with a high cluster occupancy are accelerated too much, increasing the variance.

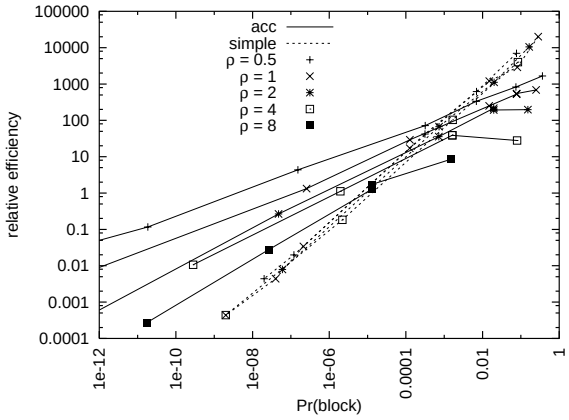
It seems from Figure 8(a) that the improvement decreases as the load increases. However, this is largely due to the fact that the rate at which the blocking decreases for increasing C is different. Figure 8(b) shows the relative efficiency against the blocking probability. This shows that the change in slope of the curves is similar over a range of loads. Note that in the range which is of most interest to engineers, with blocking between 10^{-6} and 10^{-2}, the acceleration consistently outperforms the non-accelerated simulation.

Figure 8(b) also includes results for a "simple" estimation scheme, which does not simulate A-cycles and use (4), but instead simulates only the backbone and simply counts the proportions of calls which are blocked. This shows a constant factor improvement over the non-accelerated A-cycles, which is largely due to its ability to estimate blocking over the entire network at once, rather than focusing on a single cell in each A-cycle.

The key advantage of ISSC over the filtered Gibbs sampler is that it can be used even when the stationary distribution does not have a product form, or the closed form is not know. This is the case when existing calls cannot be rearranged. Figure 9 shows the performance of ISSC for a seven cell network when existing calls are not rearranged.

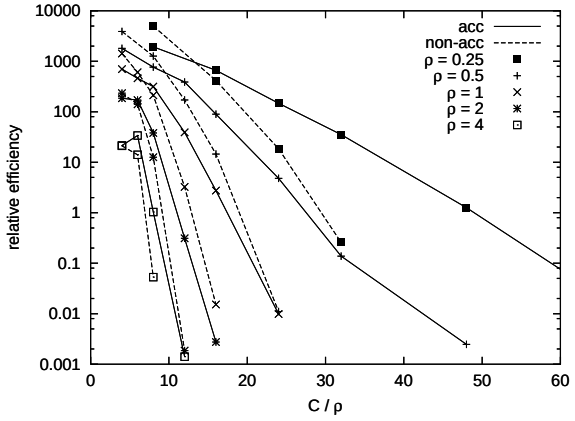


(a) efficiency against load

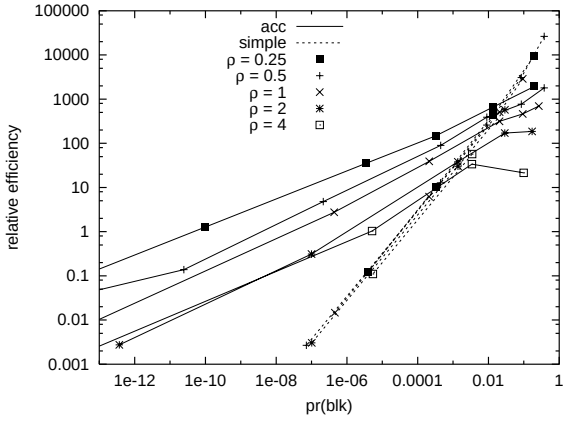


(b) efficiency against blocking

Figure 8: Relative efficiency for importance sampling (ISSC), A-cycle framework without IS, and the simple simulation, all in a 7 cell network.



(a) efficiency against load



(b) efficiency against blocking

Figure 9: Relative efficiency for importance sampling (ISSC), A-cycle framework without IS, both in a 7 cell network without call rearrangement.

8 Concluding Remarks

This paper has addressed two aspects of rare events in the calculation of network blocking probabilities. The first, the rarity of feasible states in acceptance/rejection Monte Carlo techniques, is merely an artifact of a common technique. This is removed by using a MCMC method, namely the Gibbs sampler. The Gibbs sampler clearly outperform the usual acceptance/rejection method and its relative efficiencies grow with problem size and with increasing load.

The second class of rare events is intrinsic to the problem. It is that blocking is a rare event when the load is relatively low. To overcome this problem, we implemented the main ideas of fast simulation using standard queueing-like results under the discrete event framework for simulation. Next we suggest using this change of measure for the standard clock framework for simulation, which entails significantly less computational effort.

When dealing with cellular network lay-outs, it may be that some regions of the space satisfy the conditions for which the MCMC methods are best suited, while other regions may fall under the rare event framework. We believe that hybrid methods may well prove to be the most efficient ones in practice.

References:

Alexopoulos, C., and A. Seila. 1998. Output data analysis. chapter 7. In *Handbook of Simulation*, ed. J. Banks, 225–272. John Wiley & Sons.

Andrew, L., and F. Vázquez-Abad. 2001. Fast simulation of wavelength-continuous WDM networks. (In preparation).

Asmussen, S., and H. Nielsen. 1995. Ruin probabilities via local adjustment coefficients. *J. Appl. Probab.*, 32:736–755.

Boucherie, R. J., and M. Mandjes. 1998. Estimation of performance measures for product form cellular mobile communications networks. *Telecommunication Systems*, 10:321–354.

Bratley, P., B. Fox., and L. Schrage. 1987. *A Guide to Simulation*. Second ed. New York: Springer-Verlag.

Breiman, L. 1992. *Probability*. Classics in Applied Mathematics, SIAM.

Brémaud, P. 1999. *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues*. Texts in Applied Mathematics, 31, New York: Springer.

Bucklew, J. A. 1990. *Large Deviations Techniques in Decision Simulation and Estimation*. New York: Wiley.

Chang, C.-S., P. Heidelberger., and P. Shahabuddin. 1995. Fast simulation of packet loss rates in a shared buffer communications switch. *ACM Trans. Model. Comput. Simul.*, 5(4):306–325.

Choudhury, G. L., K. K. Leung., and W. Whitt. 1995. An algorithm to compute blocking probabilities in multi-rate multi-class multi-resource loss models. *Adv. Appl. Prob.*, 27:1104–1143.

Cox, D., and D. Reudink. 1972. Dynamic channel assignment scheme in large cellular-structured mobile communication systems. *IEEE Trans. Commun.*, COM-26:432–438.

Cox, D., and D. Reudink. 1973. Increasing channel occupancy in large scale mobile radio systems: dynamic channel reassignment. *IEEE Trans. Vehic. Technol.*, VT-22:218–222.

Devetsikiotis, M., and K. Townsend. 1993. Statistical optimization of dynamic importance sampling parameters in efficient simulation of communication networks. *IEEE/ACM Trans.*

Networking, 1(3):293–305.

Dziong, Z., and J. W. Roberts. 1987. Congestion probabilities in a circuit-switched integrated services network. *Perf. Eval.*, 7:267–284.

Everitt, D., and N. Macfadyen. 1983. Analysis of multicellular mobile radiotelephone systems with loss. *Br. Telecom Technol. J.*, 1(2):37–45.

Gaivoronski, A., and E. Messina. 1996. Optimization of stationary behavior of general stochastic discrete event dynamic systems. In *Proceedings of WODES '96*. IEE Editors.

Gilks, W., S. Richardson., and D. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Glynn, P., and W. Whitt. 1992. The asymptotic efficiency of simulation estimators. *Operations Research*, 40(3):505–520.

Harvey, C., and C. Hills. 1979. Determining grades of service in a network. In *9th International Teletraffic Congress ITC9*, paper 626.

Heidelberger, P., P. Shahabuddin., and V. Nicola. 1994. Bounded relative error in estimating transient measures of highly dependable non-Markovian systems. *ACM Trans. Model. Comput. Simul.*, 4(2):137–164.

Kelly, F. P. 1979. *Reversibility and Stochastic Networks*. New York: Wiley.

Kelly, F. P. 1991. Loss networks. *The Annals of Probability*, 1:319–378.

Larson, H. J. 1973. *Introduction to the Theory of Statistics*. New York: John Wiley & Sons.

Lassila, P., and J. Virtamo. 2000. Nearly optimal importance sampling for Monte Carlo simulation of loss systems. COST report COST257TD(00), Helsinki University of Technology.

Lassila, P. E., and J. T. Virtamo. 1998a. Efficient Monte Carlo simulation of product form systems. In *Proc. Nordic Teletraffic Seminar (NTS) 14*, Copenhagen Denmark.

Lassila, P. E., and J. T. Virtamo. 1998b. Variance reduction in Monte Carlo simulation of product form systems. *Electronics Letters*, 34(12):1204–1205.

L'Ecuyer, P., and Y. Champoux. 1996. Importance sampling for large ATM-type queueing networks. In *Proceedings of the 1996 Winter Simulation Conference*, 309–316. IEEE Press.

Lee, W. 1995. *Mobile Cellular Telecommunications*. 2nd ed. McGraw Hill.

Mandjes, M. 1997. Fast simulation of blocking probabilities in loss networks. *European Journal of operations Research*, 101:393–405.

Mitra, D., and J. A. Morrison. 1994. Erlang capacity and uniform approximations for shared unbuffered resources. In *Proc. 14th Int. Teletraffic Congress*, 875–886, Amsterdam. Elsevier.

Mouly, M., and M.-B. Pautet. 1992. *The GSM System for Mobile Communications*. Telecom Publishing. ISBN 2-9507190-0-7.

Nelson, R. 1993. The mathematics of product form queuing networks. *Computing Surveys*, 25(3):339–369.

Neuts, M. F. 1981. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore, MD: Johns Hopkins University Press.

Pallant, D. L., and P. G. Taylor. 1995. Modeling handovers in cellular mobile networks with dynamic channel allocation. *Operations Research*, 43(1):33–42.

Pinsky, E., and A. E. Conway. 1992. Computational algorithms for blocking probabilities in circuit-switched networks. *Ann. Operat. Res.*, 35:31–41.

Raymond, P.-A. 1991. Performance analysis of cellular networks. *IEEE Trans. Commun.*, 39(12):1787–1793.

Redl, S., M. Weber., and M. Oliphant. 1995. *An Introduction to GSM*. Artech House.

Reiser, M., and S. Lavenberg. 1980. Mean-value analysis of closed multichain queuing networks. *J. ACM*, 27(2):313–322.

Ross, K., and J. Wang. 1992. Monte-Carlo summation applied to product-form loss networks. *Probability in the Engineering and Information Sciences*, 6:323–348.

Ross, K. W., D. H. K. Tsang., and J. Wang. 1994. Monte Carlo summation and integration applied to multiclass queuing networks. *J. ACM*, 41(6):1110–1135.

Ross, S. 1993. *Introduction to Probability Models*. Fifth ed. Boston: Academic Press.

Ross, S. 1997. *Simulation*. Second ed. Boston: Academic Press.

Sadowsky, J. S. 1991. Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Trans. Autom. Control*, 36(12):1383–1394.

Vakili, P. 1991. Using a standard clock technique for efficient simulation. *Operations Research Letters*, 10:445–452.

Vázquez-Abad, F., and L. Andrew. May, 2000. Filtered Gibbs sampler for estimating blocking probabilities in WDM optical networks. In *Proc. 14th European Simulation Multiconference*, ed. D. Landeghem, 548ff, Ghent, Belgium. Society for Comuter Simulation.

Vázquez-Abad, F., and P. LeQuoc. 2001. Sensitivity analysis for ruin probabilities. *Journal of the Operational Research Society*, 52(1):71–81.

Yates, J. 1997. Performance analysis of dynamically-reconfigurable wavelength division multiplexed networks. PhD thesis, University of Melbourne, Australia.

Zahorjan, J., D. Eager., and H. Sweillam. 1988. Accuracy, speed and convergence of approximate mean value analysis. *Perf. Eval.*, 8:255–270.

A Estimating Variance

Whenever each sample Y_s is unbiased, their mean $Y(S)$ will also be unbiased. However, the variance of $Y(S)$ may be difficult to estimate when the samples are highly correlated. We are interested in estimating blocking probabilities. The general simulation framework used is to simulate some underlying process $\{X_k, k = 1, 2, \ldots\}$, such as a Markov Chain Monte Carlo (MCMC) process or the birth and death process of Section 2. Then it is possible to set $Y_s \in \{0, 1\}$ to be an indicator of X_s being a blocking state. Some underlying processes introduce considerable correlation, so that from a point when $Y_s = 0$, it may take a long time t before an observation yields a non-zero value for Y_{s+t}. In order to calculate a variance of $Y(S)$ in the presence of correlation, the batch means method reviewed in Alexopoulos and Seila (1998) can be used. It is based on regrouping β consecutive samples to form a "batch mean" $\bar{Y}_s, s = 1, \ldots, S$ and simulating a total of S batches. The idea is that if β is large enough, then consecutive batch means will be approximately independent and the CLT is applied to these quantities. An extensive study on the choice of β, S for accurate estimation of the confidence interval is provided in Alexopoulos and Seila (1998) and suggests that choosing $\beta \approx S$ may be optimal.

To alleviate the problem of high correlations of consecutive values,

periodic sampling of the sequence Y_s could — in principle — be used to reduce the variance of the estimation. Let t denote the sampling period. Then if $\{Y_s, s = 1, 2, \ldots\}$ are unbiased,

the average of the observations $\{Y_{kt}, k = 1, 2, \ldots\}$ is also unbiased, yielding the estimator:

$$\bar{Y}_s = \frac{1}{\beta}\sum_{k=1}^{\beta} Y_{s\beta t+kt} \tag{33}$$

$$\bar{Y}(S) = \frac{1}{S}\sum_{s=1}^{S} \bar{Y}_s \tag{34}$$

$$\hat{V}(S) = \frac{1}{S-1}\sum_{s=1}^{S}(\bar{Y}_s - \bar{Y}(S))^2,$$

where $\hat{V}(S)$ is the estimated variance of a single batch mean, $\bar{Y}_s$, derived from S batches.

For $S \geq 10$ the normal approximation gives the approximate confidence interval:

$$CI : \bar{Y}(S) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{V}(S)}{S}}. \tag{35}$$

As t grows, it may suffice to consider smaller values of β, as the samples themselves become less correlated.