

A DYNAMIC IMPORTANCE SAMPLING METHOD FOR QUICK SIMULATION OF RARE EVENTS

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND
ELECTRONICS ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCES
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By

Alper Erdoğan

August, 1993

QA
276.6
.E73
1993

A DYNAMIC IMPORTANCE SAMPLING METHOD
FOR QUICK SIMULATION OF RARE EVENTS

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND
ELECTRONICS ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCES
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By

Alper Erdoğan

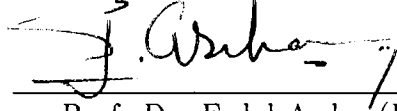
August, 1993

Alper Erdoğan
tarafından onaylanmıştır.

B.14222

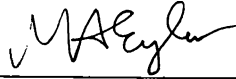
QH
276.6
.E73
1993

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.




Assoc. Prof. Dr. Erdal Arıkan(Principal Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Prof. Dr. M. Akif Eýler

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Asst. Prof. Dr. Billur Barshan

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Approved for the Institute of Engineering and Sciences:



Prof. Dr. Mehmet Baray
Director of Institute of Engineering and Sciences

ABSTRACT

A DYNAMIC IMPORTANCE SAMPLING METHOD FOR QUICK SIMULATION OF RARE EVENTS

Alper Erdoğan

M.S. in Electrical and Electronics Engineering

Supervisor: Assoc. Prof. Dr. Erdal Arıkan

August, 1993

Simulation of low-probability events may take extremely long times since they occur very rarely. There are various variance reduction methods used to speed up simulations in such cases. In this thesis, a new variance reduction technique is proposed, which is based on expressing the desired probability as the product of a number of greater probabilities and estimating each term in the product in a recursive manner. It turns out that the resulting estimator, when feasible, uses an importance sampling distribution at each step to constrain the samples into a sequence of larger sets which shrink towards the rare set gradually. Moreover, the important samples used in each step are obtained automatically from the outcomes of the experiments in the previous steps. The method is applied to the estimation of overflow probability in a network of queues and remarkable speed-ups with respect to standard simulation are obtained.

Keywords : quick simulation, rare event, importance sampling, large deviations, variance reduction, queueing network

ÖZET

ENDER OLAYLARIN HIZLI BENZETİMİ İÇİN BİR DİNAMİK ÖNEMSEL ÖRNEKLEME METODU

Alper Erdoğan

Elektrik ve Elektronik Mühendisliği Bölümü Yüksek Lisans

Tez Yöneticisi: Assoc. Prof. Dr. Erdal Arıkan

Ağustos, 1993

Çok ender gerçekleştikleri için, düşük olasılıklı olayların benzetimi aşırı uzun süreler alabilir. Benzetimi hızlandırmak için kullanılan muhtelif varyans azaltma metodları vardır. Bu tezde, sözkonusu olasılığı daha büyük bir takım olasılıkların çarpımı şeklinde ifade edip, her terimi ayrıca tahmin etme esasına dayalı bir varyans azaltma tekniği önerilmektedir. Sonuçta ortaya çıkan tahmin metodu, örnekleri ender olaya doğru daralmakta olan bir dizi daha büyük kümeyle sınırlamak için her aşamada bir önemsel örnekleme dağılımı kullanmaktadır. Ayrıca, bir aşamada kullanılan önemsel örnekler, önceki aşamalardaki deneylerin sonuçlarına göre otomatik olarak elde edilmektedir. Metod, bir kuyruklama şebekesinde taşma olasılığını tahmin etmekte kullanılmış ve standard benzetime göre kayda değer hızlanmalar kaydedilmiştir.

Anahtar Kelimeler : hızlı benzetim, ender olay, önemsel örnekleme, büyük sapmalar, varyans azaltma, kuyruklama şebekesi

ACKNOWLEDGEMENT

I would like to express my gratitude to Assoc. Prof. Dr. Erdal Arıkan for his supervision throughout the development of this thesis, his patience in our long-run discussions which helped me in many ways during my graduate studies, and above all, his understanding which has always been encouraging and invaluable for me.

I would also like to mention here Ata Seçkin Sezer for his cooperation in the Tasmus project which unintentionally formed the basis of this study.

I am also thankful to Nail Akar for his guidance during the writing of this thesis, and to Prof. Dr. M. Akif Eyler and Asst. Prof. Dr. Billur Barshan for their consideration of my work.

Contents

1	Introduction	1
1.1	General Background	1
1.2	Quick Simulation Methods	2
1.3	Objectives and Outline of the Thesis	3
2	Importance Sampling	6
2.1	Change of Measure	6
2.2	Large Deviations and Slow Markov Walk	8
2.3	Applications	11
2.4	A Heuristic Approach	14
3	Dynamic Importance Sampling	16
3.1	Theory	17
3.2	Application: Tandem Queues	21
3.3	Variance Analysis	24
4	Simulation Results	28
5	Conclusion	34
	Appendix	36

List of Figures

2.1	A typical walk and dominant exit point	10
2.2	M/M/1 Queue	11
2.3	State transition diagram for Example 2	13
3.1	$A = A_1 \cap A_2 \cap \cdots \cap A_n$	17
4.1	Empirical convergence curves for $\lambda = 0.20$, $\mu_1 = 0.30$, $\mu_2 = 0.50$, $n = 30$	32
4.2	Empirical convergence curves for $\lambda = 0.20$, $\mu_1 = 0.38$, $\mu_2 = 0.42$, $n = 25$	32

List of Tables

4.1	Simulation results for the (0.20, 0.30, 0.50)-network	31
4.2	Simulation results for the (0.20, 0.38, 0.42)-network	31
4.3	Empirical speed-up factors for the (0.20, 0.30, 0.50)-network . . .	33
4.4	Empirical speed-up factors for the (0.20, 0.38, 0.42)-network . . .	33

Chapter 1

Introduction

1.1 General Background

Study of stochastic systems often require the use of simulation as a performance evaluation tool. This is generally the case when all other tools fail. Among analysis methods, the theoretical approach involves obtaining an analytical or numerical solution of the problem posed by the system. However, many a physical systems and phenomena are so complex in nature that an analytical solution is too difficult or impossible and the numerical solution requires extensive computational resources, which might not be available. Moreover, even if the system is simple enough, some of its parameters may be unknown. Another tool for understanding the behavior of a system is experimentation, which is often impractical as the system might not yet exist or direct experimentation on the system might be dangerous or too costly. As a consequence, one often has to resort to simulation as the only available evaluation methodology, which Nobel laureate Ken Wilson characterizes as the third paradigm of science.

Stochastic simulation has a wide variety of application areas [1] such as placement of VLSI circuit components, performance evaluation of computer and communication networks, flexible manufacturing systems, global optimization and random search, molecular dynamics methods in chemical physics and Monte Carlo solutions to matrix problems. The simulation of a system may have a number of objectives including

- understanding the qualitative behavior of the system
- obtaining estimates of average performance measures
- evaluation of a set of design parameters
- model fitting to measurements of the system

Depending on the objective, the simulation model should have certain desired properties, but without exception a critical factor is the efficiency of the simulation method, i.e. its ability to generate reliable estimates within a given CPU time. It is generally the case that, simulations of complex stochastic systems are exceedingly slow, because a sufficiently high number of typical system evolutions must be generated to obtain a prescribed accuracy, and a typical evolution may take considerable computer time. Moreover, if the simulation is to be run for a long time, the period of the pseudo-random number generator may be exceeded, putting doubt on the accuracy of the resulting estimates.

1.2 Quick Simulation Methods

There are a number of techniques used to speed up simulations, which fall into two broad categories. In the first category are the variance reduction techniques (VRT) which aim to improve the statistical efficiency as measured by the variances of the output random variables (i.e. estimates). If the variance of an output random variable can be reduced without disturbing its mean, then a specified precision can be achieved with less simulation. All of these techniques, in essence, involve putting into work our knowledge about the system in one way or the other, and with regard to the system complexity, they are difficult to set up and application specific. A brief survey of existing VRTs can be found in [2]. The techniques in the second category are based on distributing the simulation to a multiprocessor system instead of using a single processor. Since most systems encountered in practice possess an inherent parallelism, it would be also natural to carry this property to simulations. In addition, when it is the case that a number of independent simulation runs are to be performed, these can be efficiently done in parallel. See [3] for an overview of distributed simulation.

Importance sampling is a VRT that has attracted a lot of research in the

last two decades. In this technique, the inputs to the system are biased in such a way that, a specific system response occurs with greater frequency, which gives chance to study that kind of response with less simulation. Since the simulated system was initially biased, the output variables have to be scaled appropriately to go back to the original system, where the scaling is determined by the original and biased input statistics. Though it presents some difficulties in practice, there has been a great deal of interest in applying importance sampling to estimating

- probabilities of rare events in Markovian systems [4], [5], [6],
- error rates in communication systems and detection [7], [8], [9], [10],
- probabilities of excessive backlogs in queueing networks [11], [12].

1.3 Objectives and Outline of the Thesis

In this thesis, we will be concerned with a subset of quick simulation problems, namely estimation of probabilities of rare events. Such events usually represent failures or overflows in stochastic systems. Although they have very low probability of occurrence, they can seriously impair the system functioning whenever they occur. As a consequence, one might wish to know the probability of such events quite accurately.

In the following, we give a precise definition of the problem:

Let (Ω, P) be a probability space, $A \subset \Omega$ an event and $I_A(\cdot)$ the indicator of A . The probability of A is given by

$$p_A = \int_{\Omega} I_A(\omega) dP(\omega) \quad (1.1)$$

We shall be concerned with the case where A is a rare event, i.e. p_A is very small. The standard Monte Carlo estimator for p_A is

$$\hat{p}_A = \frac{1}{L} \sum_{j=1}^L I(\omega^j) \quad (1.2)$$

where ω^j are i.i.d. outcomes of the experiments on (Ω, P) . An estimator is called *unbiased* if its expected value is equal to the true value, and *consistent*

if, as $L \rightarrow \infty$, it converges in probability to the true value. Note that, \hat{p}_A is unbiased and consistent, with variance

$$\text{Var}[\hat{p}_A] = \frac{1}{L}(p_A - p_A^2) \quad (1.3)$$

A commonly used measure of the accuracy of an estimator is its confidence interval. For instance, an (ϵ, β) -confidence estimator guarantees that the estimate is within $\pm\epsilon\%$ of the true value with probability β . An equally good measure is the relative precision defined in terms of the squared coefficient of variation

$$C_{\hat{p}_A}^2 = \frac{\text{Var}[\hat{p}_A]}{E[\hat{p}_A]^2} \quad (1.4)$$

Remembering that, \hat{p}_A is approximately normal for sufficiently large L , the accuracy measured by the coefficient of variation can be expressed in terms of confidence intervals. As an example, an estimator with $C_{\hat{p}_A}^2 = 10^{-2}$ is equivalent to a $(20, 0.95)$ -confidence estimator.

From (1.3) and (1.4), the coefficient of variation of the standard estimator can be computed as

$$C_{\hat{p}_A}^2 = \frac{1}{L} \left(\frac{1}{p_A} - 1 \right) \quad (1.5)$$

It is clear from (1.5) that, if p_A is very small, then a large number of i.i.d. outcomes ω^j must be generated in order to meet a specified precision and when Ω is a complex sample space imposed by a complex system, the simulation may take extremely long times and some kind of quick simulation technique is called for.

Importance sampling has been successfully used for rare event simulation in a number of applications [5], [12], [13], [14]. In this work, we will present a new variance reduction technique which is closely related to importance sampling and is based on dividing the simulation into smaller parts. This division should not be confused with distributed or parallel simulation however, because the purpose here is not to distribute the simulation but to obtain a statistical efficiency in terms of variance. Although importance sampling appears to be one of the key ideas in this method, it is not used in the sense put forward by the theory, rather it is implicit in the procedure.

The outline of the thesis is as follows:

In Chapter 2, we give the mathematical preliminaries of importance sampling and consider the class of exponentially twisted family of distributions applied to the study of overflow probabilities in networks of queues. Only Section 3.1 is crucial to the understanding of the subsequent chapters, the rest of Chapter 2 being an overview of some known results in theory and application.

In Chapter 3, we introduce our method, apply it to the example of the previous chapter and examine some conditions on the efficiency of the resulting estimator.

In Chapter 4, we present the results of our simulation experiments on the network of tandem queues and compare the three simulation methodologies discussed so far.

Finally, Chapter 5 gives conclusions and suggestions for further research.

Chapter 2

Importance Sampling

2.1 Change of Measure

Recall the standard estimator (1.2):

$$\hat{p}_A = \frac{1}{L} \sum_{j=1}^L I(\omega^j)$$

where ω^j are i.i.d. outcomes of the experiments on (Ω, P) . The basic idea of importance sampling is to change P in such a way that the estimator variance is reduced. Denoting this new measure by P^* , the importance sampling estimator is given by

$$\bar{p}_A = \frac{1}{L^*} \sum_{j=1}^{L^*} I(\omega^j) \frac{dP}{dP^*}(\omega^j) \quad (2.1)$$

where ω^j are i.i.d. outcomes of the experiments on (Ω, P^*) . It is assumed that P^* is absolutely continuous with respect to P so that the likelihood ratio dP/dP^* is finite. Note that \bar{p}_A is unbiased

$$E[\bar{p}_A] = \int_{\Omega} I_A(\omega) \frac{dP}{dP^*}(\omega) dP^*(\omega) = \int_{\Omega} I_A(\omega) dP(\omega) = p_A$$

and has variance

$$\begin{aligned} \text{Var}[\bar{p}_A] &= \frac{1}{L^*} \int_{\Omega} I_A^2(\omega) \left[\frac{dP}{dP^*}(\omega) \right]^2 dP^*(\omega) - p_A^2 \\ &= \frac{1}{L^*} \int_{\Omega} I_A(\omega) \frac{dP}{dP^*}(\omega) dP(\omega) - p_A^2 \end{aligned} \quad (2.2)$$

For the estimator (2.1) to be more efficient than the standard estimator, we should have

$$\text{Var}[\bar{p}_A] \leq \text{Var}[\hat{p}_A]$$

or equivalently

$$\int_{\Omega} I_A(\omega) \frac{dP}{dP^*}(\omega) dP(\omega) < \int_{\Omega} I_A(\omega) dP(\omega)$$

If $dP/dP^*(\omega) < 1$ whenever $\omega \in A$, this condition is satisfied, which means that a good choice of P^* should put its mass mainly on the set A . That is, the rare event should occur more frequently under the new measure.

THEOREM 1. The choice of

$$dP^*(\omega) = \frac{dP(\omega) \cdot I_A(\omega)}{p_A} \quad (2.3)$$

achieves the minimum variance for the importance sampling estimator \bar{p}_A .

See [15] for a proof.

Note that the event A occurs with probability 1 under this measure. Substituting (2.3) into (2.2), we get an interesting result

$$\text{Var}[\bar{p}_A] = 0 \quad (2.4)$$

Unfortunately, the optimum distribution is impractical for a couple of reasons. First, $P(\cdot)$ is not usually specified in closed form. Second, and more important is, in order to evaluate (2.3), we need to know p_A , which is the parameter we are trying to estimate. Hence, the result (2.4) is unachievable in practice.

Although the optimum distribution (2.3) is impractical, one can still devise sub-optimum solutions which approximate it. Typically, these solutions are chosen from a parametric class of distributions, so as to minimize the estimator variance. What kind of a constraint class should be used depends on the character of the rare event under consideration. In the following sections, we will consider the family of exponentially twisted distributions, which has been successfully used in large deviation examples. Although the underlying idea is the same, the derivation of the results for different applications exhibits big differences, so we will be concerned with the slow Markov walk as a specific example.

2.2 Large Deviations and Slow Markov Walk

Consider the Markov chain $\{X_k^\epsilon\} \in \mathfrak{R}^m$ defined by

$$\begin{aligned} X_0^\epsilon &= x_0 \in \mathfrak{R}^m \\ X_{k+1}^\epsilon &= X_k^\epsilon + \epsilon V(X_k^\epsilon, \xi_k) \end{aligned} \quad (2.5)$$

where ϵ is the parameter defining the Markov chain, $V(\cdot, \cdot)$ is a function from $\mathfrak{R}^m \times \mathfrak{R}$ to \mathfrak{R}^m and ξ_k are i.i.d. random variables on \mathfrak{R} . The results will be asymptotic when $\epsilon \rightarrow 0$.

Let F_x be the distribution of $V(x, \xi_k)$, $b(x)$ the mean of F_x and $M_x(s) = E[\exp\langle s, y \rangle]$ the Laplace transform of F_x . Let P^ϵ denote the probability measure associated with the process $\{X_k^\epsilon\}$. Assume that

1. $M_x(s) < +\infty$ for each x , i.e. F_x has a finite Laplace transform.
2. $d(F_y, F_x) \leq C\|y - x\|$ where $c > 0$ and d is the Prohorov distance [16], i.e. F_x is Lipschitz smooth in x .

THEOREM 2. Let $x^0(t)$ be the solution of the ODE

$$\frac{dx^0(t)}{dt} = b(x^0(t)) \quad \text{and} \quad x^0(0) = x_0$$

If $X_0^\epsilon = x_0$, then

$$\forall \eta > 0, \forall T < +\infty, \quad P\left(\max_{0 \leq k\epsilon \leq T} |X_k^\epsilon - x^0(k\epsilon)| > \eta\right) \rightarrow 0$$

when $\epsilon \rightarrow 0$.

Define a continuous process $X^\epsilon(t)$ from X_k^ϵ as $X^\epsilon(t) = X_{\lfloor t/\epsilon \rfloor}^\epsilon$ so that $X^\epsilon(t)$ is the interpolated version of the discrete process. With this, the result of Theorem 2 means that the process $X^\epsilon(t)$ converges uniformly in each interval $[0, T]$ to the deterministic trajectory $x^0(t)$, which is tangent to the mean-field of F_x . See Cottrell [5] for a more detailed discussion and references.

Let $l_x(s)$ and $h_x(u) = \sup_s (\langle s, u \rangle - l_x(s))$ denote the logarithm and Cramer transform of $M_x(s)$, respectively. Let \mathcal{C}_T be the set of continuously piecewise differentiable functions $\varphi : [0, T] \rightarrow \mathfrak{R}^m$ such that $\varphi(0) = x_0$ is fixed.

For $\varphi \in \mathcal{C}_T$, the action integral along φ is defined by

$$I(\varphi) = \int_0^T h_{\varphi(t)}(\dot{\varphi}(t)) dt$$

where $\dot{\varphi}(t)$ is the derivative of φ at point t .

With this definition, consider the following theorem

THEOREM 3 Let $\delta > 0$, φ be a path of \mathcal{C}_T . Let $T_\delta^\epsilon(\varphi)$ be a tube around φ with diameter δ ; that is, the set of the trajectories of $X^\epsilon(t)$, issued from x_0 , such that

$$\forall t \in [0, T], \quad |X^\epsilon(t) - \varphi(t)| < \delta$$

Then, there exists δ_0 such that for $0 < \delta < \delta_0$, we have

$$\lim_{\epsilon \rightarrow 0} (-\epsilon \log P^\epsilon(T_\delta^\epsilon(\varphi))) = I(\varphi) + \alpha(\delta) \quad \text{with} \quad \lim_{\delta \rightarrow 0} \alpha(\delta) = 0$$

Theorem 3 says that, the probability that the process $X^\epsilon(t)$ will stay inside the tube $T_\delta^\epsilon(\varphi)$ is approximately equal to $\exp(-I(\varphi)/\epsilon)$, from which we see that, $I(\varphi)/\epsilon$ is a measure of the resistance of the process to follow the path φ . See [5] for comments and [17] for a proof.

Let A be a subset of \mathcal{C}_T . By making δ smaller, one can discriminate the tubes around each $\varphi \in A$. Then, the probability of set A is approximately the sum of the probability of different tubes

$$\sum_i \exp(-I(\varphi_i)/\epsilon)$$

As $\epsilon \rightarrow 0$, the term with the smallest coefficient $I(\varphi_i)$ will become dominant due to the exponential. This means that, whenever event A occurs, it will most likely occur along an optimal path φ_{opt} , for which the action integral is minimum. With some more technical assumptions from [18], this observation is formalized by the following corollary.

COROLLARY 1. If $A \in \Sigma$ satisfies $\inf\{I(\varphi) : \varphi \in \text{int}(A)\} = \inf\{I(\varphi) : \varphi \in \text{cl}(A)\}$, then

$$\lim_{\epsilon \rightarrow 0} (-\epsilon \log P^\epsilon(A)) = \inf_{\varphi \in A} I(\varphi)$$

PROBABILITY CHANGE. The above results are generally applied to the case where A is a collection of trajectories starting from x_0 , traversing a stable domain and reaching an exterior region \mathcal{T} . Since the exit will follow approximately φ_{opt} , the change of probability measure should be made in such a way

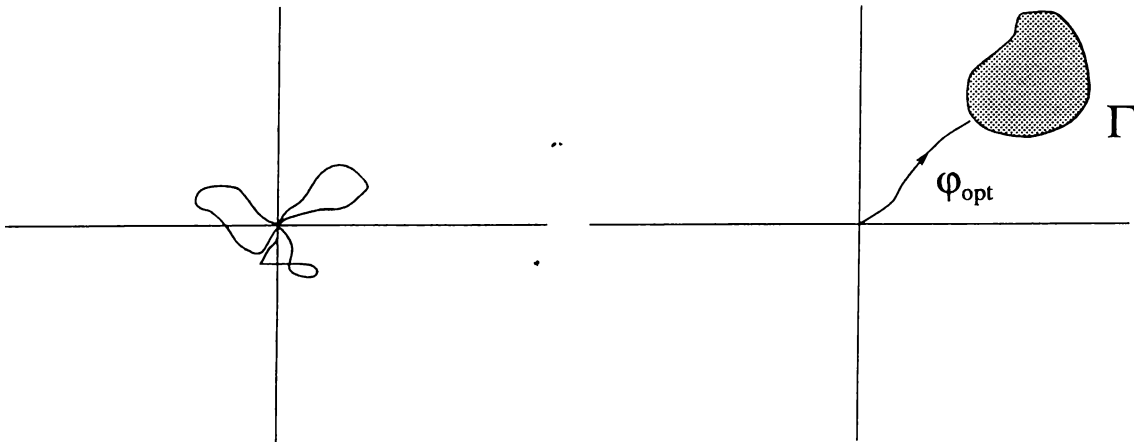


Figure 2.1. A typical walk and dominant exit point

that φ_{opt} becomes the most probable path. By recalling the result of Theorem 2, the problem is then to transform the measure F_x into F_x^* so that the transformed mean-field will be tangent to φ_{opt} .

More precisely, F_x^* is defined by

$$dF_x^*(y) = \frac{\exp(\theta_x \cdot y) dF_x(y)}{M_x(\theta_x)} \quad (2.6)$$

where θ_x is to be chosen so that, if $x = \varphi_{opt}(t)$, $E[F_x^*] = \dot{\varphi}_{opt}(t)$. Let P^{ϵ^*} be the corresponding probability induced on $\{X_k^\epsilon\}$.

The above problem has been solved for a particular situation in [5]. Suppose that $x_0 = 0$ and $b(x)$ has the sign of $-x$ for each x so that 0 is an attraction point of the process. Define A to be the set of all trajectories, starting at 0, crossing a positive boundary a before coming back to 0. If Assumptions 1 and 2 hold, we have the following theorem.

THEOREM 4. If θ_x in (2.6) is chosen as the solution of

$$l_x(\theta_x) = 0 \quad \text{and} \quad \theta_x \neq 0 \quad \text{for } x \in]0, a[\quad (2.7)$$

then among all exponential changes of probability, the transformation $P^\epsilon \rightarrow P^{\epsilon^*}$ defined by (2.6) is asymptotically optimal in the sense of the variance, i.e.

$$\lim_{\epsilon \rightarrow 0} \int_A \left[\frac{dP^\epsilon}{dP^{\epsilon^*}}(\omega) \right]^2 dP^{\epsilon^*}(\omega)$$

is minimum.

The proof can be found in [5].

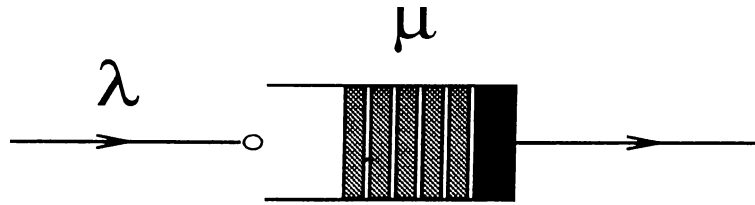


Figure 2.2. M/M/1 Queue

2.3 Applications

EXAMPLE 1. Consider a single M/M/1 queue with arrival rate λ and service rate μ . Assume $\lambda + \mu = 1$ without loss of generality and $\lambda < \mu$ so that the queue is stable. Let $\{Z_k : k = 0, 1, \dots\}$ denote the embedded Markov chain representing the number of customers in the system. The transition probabilities are given by

$$\begin{aligned} p_{i,i+1} &= \lambda \text{ for } i \geq 1 \text{ and } p_{0,1} = 1 \\ p_{i,i-1} &= \mu \text{ for } i \geq 1 \end{aligned}$$

We are interested in the probability p_A that, starting with an empty system, the number of customers reaches n before returning to 0 again. Knowledge of such a probability is useful in finding the mean buffer overflow time in queueing networks [12].

To be able to apply the results of the previous section, we need to represent the Markov chain in the form of equation (2.5). For this define $Z_k^n = Z_k/n$. Then,

$$Z_{k+1}^n = Z_k^n + \frac{1}{n}V(Z_k, \xi_k) \quad (2.8)$$

Note p_A is also equal to the probability that, starting with a *single customer*, the system reaches n before returning to 0. In this case, $V(0, \xi_k)$ will never have to be evaluated since state 0 can only be reached at the termination (final step). So, the chain can be assumed to have an homogeneous jump distribution

$$P(V(\xi_k) = 1) = \lambda \quad \text{and} \quad P(V(\xi_k) = -1) = \mu$$

for all states, satisfying the continuity requirement in Assumption 1. Equation (2.8) now becomes

$$Z_{k+1}^n = Z_k^n + \frac{1}{n}V(\xi_k)$$

The Laplace transform of $V(\xi_k)$ is

$$M_x(s) = \lambda \exp(s) + \mu \exp(-s)$$

Solving for $M_x(\theta_x) = 0$ gives

$$\theta_x = \log\left(\frac{\mu}{\lambda}\right)$$

Substituting θ_x into (2.6), we get

$$\begin{aligned} P(V^*(\xi_k) = 1) &= \lambda \exp\left(\log\left(\frac{\mu}{\lambda}\right)\right) = \mu \quad \text{and} \\ P(V^*(\xi_k) = -1) &= \mu \exp\left(-\log\left(\frac{\mu}{\lambda}\right)\right) = \lambda \end{aligned}$$

which is interesting in the sense it dictates the interchange of λ and μ in the original system.

The direct estimator of p_A is

$$\hat{p}_A = \frac{1}{L} \sum_{j=1}^L I(\omega^j)$$

where ω^j are i.i.d. evolutions of the original system. Actually, ω^j is a sequence of states starting from 0, ending either in n or 0, and is also called a *cycle*. $I(\omega^j) = 1$ if ω^j reaches n , 0 otherwise.

The importance sampling estimator on the other hand is

$$\bar{p}_A = \frac{1}{L^*} \sum_{j=1}^{L^*} I(\omega^j) \frac{dP}{dP^*}(\omega^j)$$

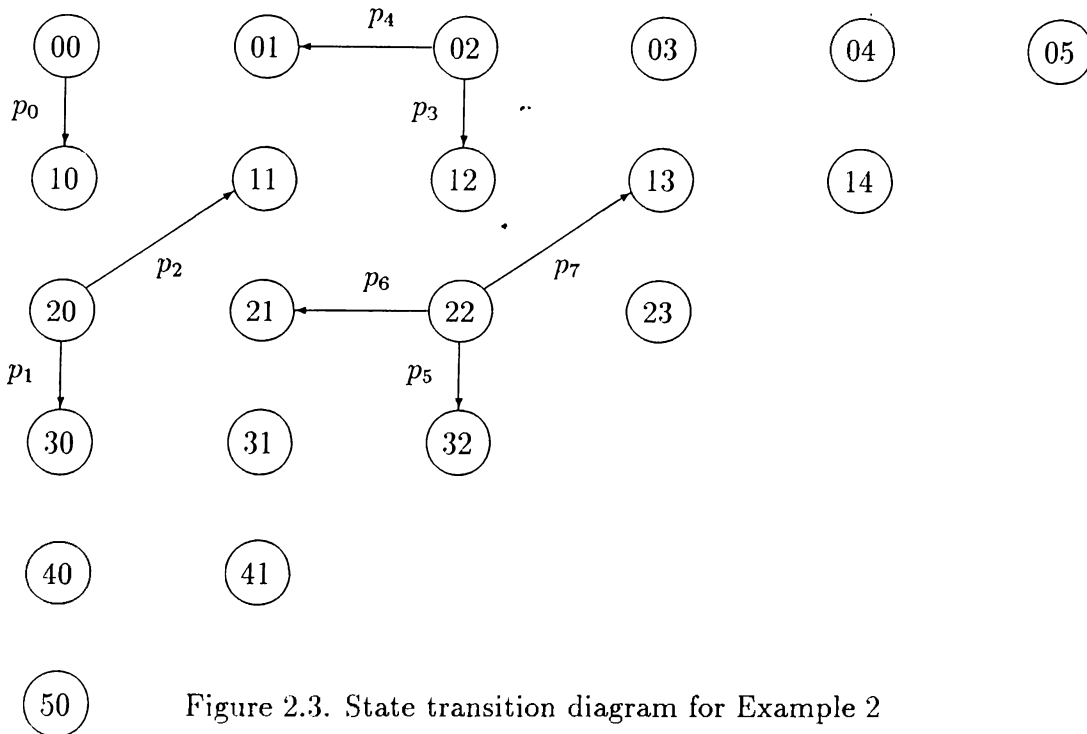
where ω^j are i.i.d. evolutions of the system in which λ and μ are interchanged and $\frac{dP}{dP^*}(\omega^j)$ is ratio of probability of occurrence of ω^j under P to that under P^* .

Let M denote the expected number of Markovian jumps in a cycle of the original system and M^* denote that of the transformed system. Then, the speed-up factor for this change of measure is given by

$$S = \frac{L M}{L^* M^*}$$

where L and L^* are to be chosen such that $\text{var}[\bar{p}_A] = \text{var}[\hat{p}_A]$. Since the system is simple enough, analytical expressions for $\text{var}[\bar{p}_A]$, M , and M^* can be obtained and the speed-up can be computed as

$$S \approx \left[n \cdot \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \right]^{-1} \quad (2.9)$$



50 Figure 2.3. State transition diagram for Example 2

See [12] for the derivation of 2.9.

As an example, let $\lambda = 0.3$, $\mu = 0.7$ and $n = 20$. Then, S is approximately 2×10^6 , showing the power of the above probability change.

EXAMPLE 2. Now consider two M/M/1 queues in tandem. Let λ be the arrival rate to the first queue and μ_1, μ_2 be the respective service rates. Call such a network a (λ, μ_1, μ_2) -network. Assume $\lambda < \mu_1$ and $\lambda < \mu_2$ for stability. Let $\{Z_k : k = 0, 1, \dots\}$ be the embedded two dimensional Markov chain taking values over the state space $\mathcal{S} = \{(n_1, n_2) : n_1 \geq 0, n_2 \geq 0\}$, where n_i represents the number of customers in queue i . Also assume without loss of generality that $\lambda + \mu_1 + \mu_2 = 1$. We are again interested in the probability that, the number of customers $n_1 + n_2$ in the system reaches n before reaching 0 again. The state diagram of the chain is shown in Figure 2.3 where the transition probabilities are

$$\begin{aligned}
 p_0 &= 1 \\
 p_1 &= \frac{\lambda}{\lambda + \mu_1} & p_2 &= \frac{\mu_1}{\lambda + \mu_1} \\
 p_3 &= \frac{\lambda}{\lambda + \mu_2} & p_4 &= \frac{\mu_2}{\lambda + \mu_2} \\
 p_5 &= \lambda & p_6 &= \mu_2 & p_7 &= \mu_1
 \end{aligned}$$

Note that the jump distributions at the boundary states are different from those at the interior states and the change is abrupt at the boundaries, violating the continuity assumption. So, the results of Section 2.2 are not directly applicable here.

It has been shown in [12] that, neglecting the discontinuities at the boundaries, the large deviations theory suggests the interchange of λ and μ_2 . However, experiments on the $(\lambda = 0.20, \mu_1 = 0.30, \mu_2 = 0.50)$ -network have shown that the above is not an optimum change of measure. It is reported in [12] that, for $n = 20$, where the true value of p_A is 3.759×10^{-4} , simulating the $(\lambda = 0.50, \mu_1 = 0.30, \mu_2 = 0.20)$ -network for 1000 cycles gave $\bar{p}_A = 8.388 \times 10^{-5}$, while simulating the $(\lambda = 0.30, \mu_1 = 0.20, \mu_2 = 0.50)$ -network for the same number of cycles gave a better result, $\bar{p}_A = 3.595 \times 10^{-4}$.

To take care of the above theoretical difficulty, Parekh and Walrand [12] proposed a heuristic method which we consider briefly in the next section.

2.4 A Heuristic Approach

The fact that the large deviations results are not applicable to discontinuous kernels does not mean that there is no optimal exit path φ_{opt} on those kernels. Thus, any other method which finds φ_{opt} and centers the probability around it would be equally useful. Such a method has been proposed in [12] based on Borovkov heuristics [19].

Consider a G/G/1 queue with arrival rate λ and service rate μ . Let $h_\lambda(u)$ and $h_\mu(u)$ be the Cramer transforms of the interarrival time distribution and service time distribution respectively. We want to estimate the probability of the backlog exceeding n in a cycle.

To find the most probable path of overflow, one reasons as follows: For the queue length to exceed n in a cycle, there must exist a time T such that $n = T(\lambda' - \mu')$ where λ' and μ' are the empirical (observed) arrival rate and departure rate until T . Using a large deviation theorem by Chernoff [20], the probability of such a behavior can be approximately evaluated as

$$\exp\{-n(\lambda' - \mu')^{-1}[\lambda' h_\lambda(1/\lambda') + \mu' h_\mu(1/\mu')]\} \quad (2.10)$$

Maximization of (2.10) with respect to λ' and μ' reveals the most likely trajectory that the original system would follow to reach an overflow. One can then replace λ by λ^* and μ by μ^* to make this behavior more probable (λ^* and μ^* are the values that maximize (2.10)), which is what we seek. Actually, $\lambda^* = \mu$ and $\mu^* = \lambda$, the solution in agreement with the large deviation results for an $M/M/1$ queue. See [20] for more details.

A similar approach is possible $M/M/1$ queues in tandem, yielding a change of measure where λ is interchanged with the smaller of μ_1 and μ_2 [12], which is quite reasonable, because if the system is to be filled up, it will most likely fill up due to the queue which has lower service rate. This result also explains the superiority of the $(\lambda = 0.30, \mu_1 = 0.20, \mu_2 = 0.50)$ -network over the $(\lambda = 0.50, \mu_1 = 0.30, \mu_2 = 0.20)$ -network, as exemplified at the end of the previous section.

The above heuristic is generalized to more complex Jackson networks in [12] and the analytical solution of the resulting maximization problem is obtained in [21].

Chapter 3

Dynamic Importance Sampling

In this chapter, we are going to introduce a variance reduction technique which essentially utilizes the idea of importance sampling, but in a somewhat different way.

As before, we are concerned with estimating probabilities of rare events, or equivalently, expectations of associated indicator functions. The technique is based on expressing the desired expectation as the product of a set of expectations. The estimation is then performed recursively on each term. This decomposition may lead to substantially low coefficients of variation on samples of the components, and hence, obtaining good estimates of the *parts* may be much easier than obtaining a comparable estimate of the *whole*.

The technique is called *dynamic*, because it has an evolutionary nature, where the statistics obtained in each stage are used as inputs in the following stages. From this point of view, the resulting algorithm can be considered as forcing samples in a stepwise manner into the rare set of interest.

3.1 Theory

Let (Ω, P) be a probability space and $A \subset \Omega$ an event whose probability $P(A)$ is to be estimated.

Introduce a number of random variables X_1, \dots, X_n on (Ω, P) , i.e. functions $X_i : \Omega \rightarrow \mathcal{Z}$, such that A can be written in the form

$$A = A_1 \cap A_2 \cap \dots \cap A_n \quad (3.1)$$

where

- i) $A_1 \supseteq A_2 \supseteq \dots \supseteq A_n$
- ii) A_i belongs to the σ -algebra of events generated by X_1, \dots, X_i , for each $i = 1, \dots, n$.

Condition ii) is equivalent to assuming that the occurrence of A_i is determined by the knowledge of the values of X_1, \dots, X_i , i.e. there exists a set $B_i \subset \mathcal{Z}^i$ such that

$$\omega \in A_i \iff (X_1(\omega), \dots, X_i(\omega)) \in B_i \quad (3.2)$$

Although we have chosen X_i 's to be discrete, the following derivations can be easily extended to the continuous case. From (3.1) and (i), we have by Bayes' rule

$$P(A) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_{n-1}) \quad (3.3)$$

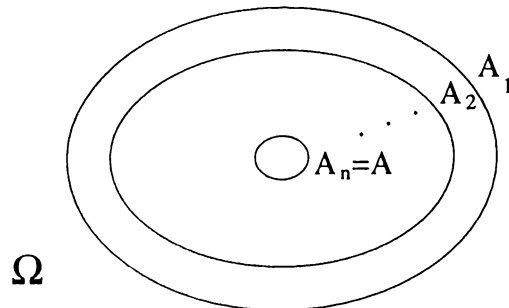


Figure 3.1. $A = A_1 \cap A_2 \cap \dots \cap A_n$

The idea is to estimate $P(A)$ by estimating each term $P(A_i|A_{i-1})$. Since A_i are not arbitrary sets, one can expect $P(A_i|A_{i-1})$ to have some nice form. Noting that $A_i \subseteq A_{i-1}$, we have

$$\begin{aligned} P(A_i|A_{i-1}) &= \frac{P(A_i)}{P(A_{i-1})} \\ &= \frac{P((X_1, \dots, X_i) \in B_i)}{P((X_1, \dots, X_{i-1}) \in B_{i-1})} \\ &= \frac{\sum \cdots \sum_{B_i} p(x_1, \dots, x_i)}{P((X_1, \dots, X_{i-1}) \in B_{i-1})} \\ &= \frac{\sum \cdots \sum_{B_i} p(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})}{P((X_1, \dots, X_{i-1}) \in B_{i-1})} \end{aligned}$$

Let $I_{B_{i-1}}$ denote the indicator of set B_{i-1} . Since $I_{B_{i-1}}(x_1, \dots, x_{i-1}) = 1$ whenever $(x_1, \dots, x_{i-1}) \in B_{i-1}$, it can be inserted into the summation of the numerator without changing the value of the sum, giving

$$P(A_i|A_{i-1}) = \sum \cdots \sum_{B_i} p^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1}) \quad (3.4)$$

where

$$p^*(x_1, \dots, x_{i-1}) = \frac{I_{B_{i-1}}(x_1, \dots, x_{i-1})p(x_1, \dots, x_{i-1})}{P((X_1, \dots, X_{i-1}) \in B_{i-1})}, \quad i = 1, \dots, n \quad (3.5)$$

It is interesting to note that $p^*(x_1, \dots, x_{i-1})$ is the optimum importance sampling distribution of (X_1, \dots, X_{i-1}) on the set B_{i-1} (see (2.3)). Equation (3.3) can now be written in a more compact form as

$$p_A = P(A) = \prod_{i=1}^n p_i \quad (3.6)$$

where

$$p_i = P(A_i|A_{i-1}) = E_{p^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})}[I_{B_i}(X_1, \dots, X_i)]$$

We consider the estimator

$$\tilde{p}_A = \prod_{i=1}^n \hat{p}_i \quad (3.7)$$

with

$$\hat{p}_i = \frac{1}{L_i} \sum_{j=1}^{L_i} I_{B_i}((X_1, \dots, X_i)^j) \quad (3.8)$$

where $\{(X_1, \dots, X_i)^j : i=1, \dots, n, j=1, \dots, L_i\}$ are independent random vectors, $(X_1, \dots, X_i)^j$ chosen from the distribution $p^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})$. The independence assumption guarantees that \tilde{p}_A is an unbiased estimator and it is easy to check that it is also consistent. We call such an estimator as a *product-form estimator*.

The feasibility of the estimator (3.8) hinges on the ability to

- i) generate samples from $p^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})$ and
- ii) recognize whether $(x_1, \dots, x_i) \in B_i$ for arbitrary (x_1, \dots, x_i) .

We discuss the first item in some more detail. In the following, we write $P(B_i)$ to denote $P((X_1, \dots, X_i) \in B_i)$. By conditioning, we have from (3.5)

$$\begin{aligned} p^*(x_1, \dots, x_i) &= \frac{p(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})I_{B_i}(x_1, \dots, x_i)}{P(B_{i-1})P(B_i|B_{i-1})} \\ &= \frac{p(x_1, \dots, x_{i-1})I_{B_{i-1}}(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})I_{B_i}(x_1, \dots, x_i)}{P(B_{i-1})P(B_i|B_{i-1})} \\ &= \frac{p^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})I_{B_i}(x_1, \dots, x_i)}{P(B_i|B_{i-1})} \end{aligned} \quad (3.9)$$

Note from (3.9) that $p^*(x_1, \dots, x_i)$ is proportional to

$$p^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1}) \quad (3.10)$$

and is concentrated on B_i . This means that if we have true samples drawn from $p^*(x_1, \dots, x_{i-1})$ and if we can sample X_i from $p(x_i|x_1, \dots, x_{i-1})$ given X_1, \dots, X_{i-1} , we can generate samples from $p^*(x_1, \dots, x_i)$. So it is generally our ability to sample from the conditional which determines the applicability of the estimator, and when we are able to do so, p^* can be generated recursively. Actually, this generation would be automatic in the procedure, because (3.10) can be recognized as the sampling distribution in the i^{th} step, so the set of samples which fall into B_i in the i^{th} step can be used as samples of $p^*(x_1, \dots, x_i)$ in the $(i+1)^{\text{th}}$ step. However, there is a technical difficulty that should be noted here: If the number of samples of $p^*(x_1, \dots, x_i)$ obtained in the above manner is less than L_{i+1} , the number necessary in the $(i+1)^{\text{th}}$ step, then there will be shortage of true samples of $p^*(x_1, \dots, x_i)$. An immediate remedy to this problem would be to draw at random L_{i+1} samples from the available set with replacement. This would be mathematically equivalent to constructing an empirical distribution $\hat{p}^*(x_1, \dots, x_i)$ from the available samples and then

generating (X_1, \dots, X_i) from $\hat{p}^*(x_1, \dots, x_i)$. With these considerations, the estimation procedure can be stated as follows:

PROCEDURE. By performing experiments on (Ω, P)

STEP 1. Generate L_1 independent samples from $p(x_1)$ to obtain X_1^j for $j = 1, \dots, L_1$. Estimate p_1 using (3.8). Record those X_1^j 's which fall into B_1 .

STEP 2. Set $i = 2$. Choose (X_1, \dots, X_{i-1}) at random among the values recorded in the previous step. Generate a sample from $p(x_i | x_1, \dots, x_{i-1})$ with (X_1, \dots, X_{i-1}) as a condition.

STEP 3. Repeat Step 2 for L_i times to obtain $(X_1, \dots, X_i)^j$ for $j = 1, \dots, L_i$. Estimate p_i using (3.8). Record those $(X_1, \dots, X_i)^j$'s which fall into B_i .

STEP 4. Repeat Step 3 and Step 4 for $i = 3, \dots, n$.

Finally, form the product $\tilde{p}_A = \prod_{i=1}^n \hat{p}_i$.

Notice that, the above procedure deviates from the theoretical estimator given in (3.8), in that, $(X_1, \dots, X_{i-1})^j$ in the i^{th} step are not sampled from the true distribution $p^*(x_1, \dots, x_{i-1})$, but from an estimate $\hat{p}^*(x_1, \dots, x_{i-1})$ of the true distribution. We demonstrate in the Appendix that the resulting estimator \tilde{p}_A in this case is biased, however the bias becomes insignificant for sufficiently large values of L_i 's. Actually, the accuracy of the estimates $\hat{p}^*(x_1, \dots, x_i)$ for $i = 1, \dots, n$ turns out to be proportional to the accuracy of \hat{p}_i 's and when L_i 's are so chosen to yield accurate estimates of p_i 's, which is generally the case, \hat{p}^* 's come out to be quite accurate. So in our following derivations, we assume that L_i 's are sufficiently large so that the bias in \tilde{p}_A is negligible.

As a final point, note from (3.6) that

$$p_i = \frac{p_A}{\prod_{j \neq i} p_j} \geq p_A, \quad i = 1, \dots, n$$

The condition $p_i \geq p_A$ assures that, the number of samples required to estimate *each* p_i for a given precision is less than or equal to the number required to estimate p_A for the same precision. However, there are n of these p_i 's now. Whether there would be net gain in terms of simulation time depends

on how p_i 's are distributed and average sampling times in each step, which in turn depend on the nature of the problem being studied. Therefore, it is not possible to draw general conclusions on the efficiency of the product-form estimator. However, as will be shown in the next section, the more uniform the p_i 's are, the more advantageous is the above estimation scheme.

COMMENTS AND REMARKS.

1. The product-form estimator utilizes optimum change of measure in each step, which is known to be unachievable as it requires the knowledge of the parameter to be estimated (see Section 2.1). However, one need not know the optimum distribution in this case, because the likelihood ratio p/p^* does not appear in the equations. Recall that in standard importance sampling the likelihood ratio appears as a weighting factor and must be evaluated for each sample. What is needed in this technique is only a set of samples drawn from the optimum importance sampling distribution $p^*(x_1, \dots, x_i)$, which can be obtained in the way described before.

2. The probability of set A is estimated by using a sequence of sets A_i shrinking towards A . At each step, the sampling domain is reduced with respect to the previous step, in other words, the samples are forced gradually into the set A . It is this forcing behavior that is represented by the importance sampling distributions appearing in the equations. Moreover, the resulting sampling algorithm has a dynamic character, in the sense that, the important samples are learned at each step from the system itself, so as to be used as input to the next step.

3. Ability to sample from the conditional $p(x_i|x_1, \dots, x_{i-1})$ amounts in stochastic systems to the ability to impose certain conditions on the system. Although this may not always be possible with real systems, by appropriate modeling, it may be possible to control the parameters and inputs of the simulated system arbitrarily to create a desired condition.

3.2 Application: Tandem Queues

We now apply the results to the simulation of tandem queues studied in Section 2.3. Let $\{Z_k : k = 0, 1, \dots\}$ be the embedded Markov chain taking values over the state space $\mathcal{S} = \{(n_1, n_2) : n_1 \geq 0, n_2 \geq 0\}$ where n_1 and n_2 are the

number of customers in each queue. Assume that the system is initially empty, i.e. $Z_0 = (0, 0)$ with probability 1. Let (Ω, P) be the underlying probability space and A the event that the number of customers reaches n before returning to zero again.

We define the following subsets of \mathcal{S} :

$$S_i = \{(n_1, n_2) : n_1 + n_2 = i\}, \quad i = 1, \dots, n \quad (3.11)$$

We say that $\{Z_k(\omega)\}$ hits S_i at (n_1, n_2) if $Z_k(\omega) = (n_1, n_2)$ for some $k \geq 1$ and $Z_l(\omega) \notin S_i$ for $l = 1, \dots, k - 1$. To estimate $P(A)$ we define the random variables X_1, \dots, X_n so that for each $\omega \in \Omega$

$$X_i(\omega) = \begin{cases} (n_1, n_2) & \text{if } \{Z_k(\omega)\} \text{ hits } S_i \text{ before hitting } S_0 \\ (0, 0) & \text{otherwise} \end{cases} \quad (3.12)$$

Note that if we regard S_0 and S_i as imaginary boundaries for the random walk $\{Z_k(\omega)\}$, X_i equals the point $\{Z_k(\omega)\}$ first hits the boundary. Let

$$A_i = \{\omega \in \Omega : X_i(\omega) \neq (0, 0)\}$$

Clearly,

$$\begin{aligned} A &= A_1 \cap A_2 \cap \dots \cap A_n \\ A_1 &\supseteq A_2 \supseteq \dots \supseteq A_n \end{aligned}$$

and for each $i = 1, \dots, n$,

$$\omega \in A_i \iff X_i(\omega) \in B_i \quad (3.13)$$

Applying the results of the previous section with $B_i = S_i$ and noting that A_i is measurable w.r.t. X_i only, we have a simpler form than given in (3.4):

$$\begin{aligned} p_i &= P(A_i | A_{i-1}) = P(X_i \in S_i | X_{i-1} \in S_{i-1}) \\ &= \sum_{S_{i-1}} \sum_{S_i} p^*(x_{i-1}) p(x_i | x_{i-1}), \quad i = 2, \dots, n \end{aligned}$$

The estimator for each p_i can now be written as

$$\hat{p}_i = \frac{1}{L_i} \sum_{j=1}^{L_i} I_{S_i}(X_i^j) \quad (3.14)$$

where X_i^j are i.i.d. copies from the distribution $p^*(x_{i-1})p(x_i | x_{i-1})$. For this case, the sampling process takes the following form:

- i) Sampling from $p(x_i|x_{i-1})$ is equivalent to starting $\{Z_k\}$ in state x_{i-1} and recording its final state as X_i when $\{Z_k\}$ reaches S_i or S_0 .
- ii) Samples of $p^*(x_{i-1})$ are those points X_{i-1} on the boundary S_{i-1} that were hit by $\{Z_k\}$ in the $(i-1)^{th}$ step.

For $\omega \in \Omega$, the part of $\{Z_k(\omega)\}$ which lies beyond the first visit of $\{Z_k(\omega)\}$ to S_{i-1} , if there exists any such visit, is called the i^{th} cycle of $\{Z_k(\omega)\}$. The i^{th} cycle of $\{Z_k(\omega)\}$ is said to be successful if $\{Z_k(\omega)\}$ hits S_i before hitting S_0 . With this definition, the simulation algorithm can be stated in simpler terms as follows:

STEP 1. Start with an empty system. Generate L_1 cycles of type 1 to obtain X_1^j . Estimate p_1 . Record the final states of successful cycles.

STEP 2. Set $i = 2$. Start the system in S_{i-1} . Choose at random among those states recorded in the $(i-1)^{th}$ step, to be a starting state for this step. Generate a cycle of type i .

STEP 3. Repeat step 2 for L_i times to obtain X_i^j 's. Estimate p_i . Record the final states of successful cycles.

STEP 4. Repeat step 2 and 3 for $i = 3, \dots, n$.

STEP 5. Form the product $\tilde{p}_A = \prod_{i=1}^n \hat{p}_i$.

SIMULATION RESULT. The $(\lambda = 0.10, \mu_1 = 0.40, \mu_2 = 0.50)$ -network has been simulated. The true overflow probability is 2.104×10^{-7} for $n = 13$. Direct simulation for 2.55 seconds gave $\hat{p}_A = 0$ while simulation using our algorithm gave in the same duration $\tilde{p}_A = 1.815 \times 10^{-7}$.

The estimates of p_i 's also are listed below:

$$[\hat{p}_1 \cdots \hat{p}_{13}] = [1.000 \ .334 \ .326 \ .303 \ .287 \ .277 \ .270 \ .264 \ .263 \ .258 \ .258 \ .256 \ .252]$$

Detailed simulation results will be presented in Chapter 4 for a more complete comparison of the present method with other methods.

3.3 Variance Analysis

In this section, we compare the performances of the standard estimator and the product-form estimator, which we repeat below for convenience.

The direct estimator of p_A is given by

$$\hat{p}_A = \frac{1}{L} \sum_{j=1}^L I_A((X_1, \dots, X_n)^j) \quad (3.15)$$

where $(X_1, \dots, X_n)^j$ are i.i.d. samples from $p(x_1, \dots, x_n)$, and the product-form estimator is

$$\tilde{p}_A = \prod_{i=1}^n \hat{p}_i \quad (3.16)$$

with

$$\hat{p}_i = \frac{1}{L_i} \sum_{j=1}^{L_i} I_{B_i}((X_1, \dots, X_i)^j) \quad (3.17)$$

where $(X_1, \dots, X_i)^j$ are independent copies drawn from the distribution $p^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})$.

As noted before, the relative magnitudes of p_i 's depend on the specific parameters of the problem under consideration. Hence, we assume throughout the section that p_A and p_i for $i = 1, \dots, n$ are known.

In the following, the squared coefficient of variation is used as a figure of merit for the estimators. The question is which estimator would achieve a lower coefficient of variation given that $\sum_{i=1}^n L_i = L$. We also assume that average sampling times in both estimators are the same so that the above is a valid measure of efficiency.

To start with, we need a fact from elementary probability theory

FACT. Let X and Y be two independent random variables with squared coefficients of variation C_X^2 and C_Y^2 respectively. Then

$$C_{XY}^2 = C_X^2 + C_Y^2 + C_X^2 C_Y^2 \quad (3.18)$$

The proof is direct from the definition of the coefficient of variation (1.4). Next, we make a simplifying approximation

APPROXIMATION. $C_{\tilde{p}_i \tilde{p}_j}^2 \cong C_{\tilde{p}_i}^2 + C_{\tilde{p}_j}^2$ for $i \neq j$.

Note that the cross-term in (3.18) is dropped. So, the approximation is valid whenever $C_{\tilde{p}_i}^2 \ll 1$ and $C_{\tilde{p}_j}^2 \ll 1$; which is generally case for estimators of acceptable precision.

With this approximation and (1.5), the coefficients of variation of both estimators can now be written

$$C_{\tilde{p}_A}^2 = \frac{1}{L} \left(\frac{1}{p_A} - 1 \right) \quad (3.19)$$

and

$$C_{\tilde{p}_A}^2 = \sum_{i=1}^n \frac{1}{L_i} C_i^2 \quad (3.20)$$

where

$$C_i^2 = \left(\frac{1}{p_i} - 1 \right) \quad (3.21)$$

Given $\sum_{i=1}^n L_i = L$, the next question is how to allocate L_i so as to get maximum performance on \tilde{p}_A .

THEOREM 1. (Optimum Allocation)

$$C_{\tilde{p}_A}^2 = \sum_{i=1}^n \frac{1}{L_i} C_i^2 \quad \text{subject to} \quad \sum_{i=1}^n L_i = L \quad \text{and} \quad L_i \geq 0, \quad i = 1, \dots, n$$

is minimized for

$$L_i = \frac{C_i}{\sum_{i=1}^n C_i} L \quad (3.22)$$

Proof. Treating L_i as continuous, we can write the Lagrangian as

$$\mathcal{L} = \sum_{i=1}^n \frac{1}{L_i} C_i^2 + \lambda \left(\sum_{i=1}^n L_i - L \right)$$

Differentiating with respect to L_i gives

$$\frac{\partial \mathcal{L}}{\partial L_i} = -\frac{C_i^2}{L_i^2} + \lambda = 0 \quad \Rightarrow \quad L_i = \frac{C_i}{\lambda^{1/2}}, \quad i = 1, \dots, n \quad (3.23)$$

Substituting L_i into the constraint, we get

$$\frac{1}{\lambda^{1/2}} \sum_{i=1}^n C_i = L \quad \Rightarrow \quad \lambda^{1/2} = \frac{\sum_{i=1}^n C_i}{L}$$

which upon back substitution into (3.23) gives

$$L_i = \frac{C_i}{\sum_{j=1}^n C_j} L$$

With the result of Theorem 1, equation (3.20) becomes

$$C_{\bar{p}_A}^2 = \sum_{i=1}^n \left(\frac{C_i}{\sum_{j=1}^n C_j} L \right)^{-1} C_i^2 = \frac{1}{L} \left(\sum_{i=1}^n C_i \right)^2 \quad (3.24)$$

Before stating the next theorem, we express (3.6) in terms of C_i 's using (3.21)

$$p_A = \prod_{i=1}^n \frac{1}{C_i^2 + 1} \quad \text{and} \quad C_i \geq 0, \quad i = 1, \dots, n \quad (3.25)$$

THEOREM 2. The minimum value of $C_{\bar{p}_A}^2 = \frac{1}{L} (\sum_{i=1}^n C_i)^2$ subject to (3.25) is achieved at $C_1 = \dots = C_n$.

Proof. The problem is to

$$\text{minimize } \frac{1}{L} \left(\sum_{i=1}^n C_i \right)^2 \quad \text{subject to} \quad \prod_{i=1}^n C_i^2 + 1 = \frac{1}{p_A} \quad \text{and} \quad C_i \geq 0, \quad i = 1, \dots, n$$

which is equivalent to

$$\text{minimize } \sum_{i=1}^n C_i \quad \text{subject to} \quad \sum_{i=1}^n \log(C_i^2 + 1) = \log\left(\frac{1}{p_A}\right) \quad \text{and} \quad C_i \geq 0, \quad i = 1, \dots, n$$

since $\sum_{i=1}^n C_i \geq 0$. The Lagrangian in this case is

$$\mathcal{L} = \sum_{i=1}^n C_i + \lambda \left(\sum_{i=1}^n \log(C_i^2 + 1) - \log\left(\frac{1}{p_A}\right) \right)$$

Differentiating with respect to C_i , we get

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial C_i} = 1 + \lambda \frac{2C_i}{C_i^2 + 1} = 0 &\Rightarrow C_i^2 + 2\lambda C_i + 1 = 0 \\ &C_i = -\lambda \pm \sqrt{\lambda^2 - 1} \end{aligned} \quad (3.26)$$

Note that, the right hand side of (3.26) does not depend on i , which proves that the minimum is achieved at $C_1 = \dots = C_n$, and since $C_{\bar{p}_A}^2$ is convex everywhere, it is the global minimum.

Theorem 2 with (3.21) also shows that the product-form estimator is optimum when p_i are uniformly distributed. We call this optimum value $C_{\bar{p}_A}^2(\min)$.

DEFINITION. The *feasible region* of $C_{\bar{p}_A}^2$ is the set of (p_1, \dots, p_n) for which $C_{\bar{p}_A}^2 \leq C_{\bar{p}_A}^2(\min)$.

Obviously, for the feasible set to be non-empty, one should have

$$C_{\bar{p}_A}^2(\min) \leq C_{\bar{p}_A}^2$$

or, from (3.24), (3.21) and Theorem 2

$$n^2 \left(\frac{1}{p_A^n} - 1 \right) = \frac{1}{p_A} - 1$$

which is satisfied for

$$p_A \leq p_n^c \quad (3.27)$$

where p_n^c is the solution of

$$n^2 \left(\frac{1}{p_A^{1/n}} - 1 \right) = \frac{1}{p_A} - 1$$

within the interval $[0, 1]$. We call p_n^c , the *critical probability*. Values of p_n^c for several n are listed below:

$$p_2^c \approx 0.1111 \quad p_3^c \approx 0.0749 \quad p_4^c \approx 0.0558 \quad p_{20}^c \approx 0.0094$$

It has been shown that, when p_A is less than or equal to a critical value, there exists a feasible region, which is defined by

$$\left[\sum_{i=1}^n \left(\frac{1}{p_i} - 1 \right)^{1/2} \right]^2 \leq \frac{1}{p_A} - 1 \quad \text{and} \quad \prod_{i=1}^n p_i = p_A, \quad p_i \geq p_A, \quad i = 1, \dots, n \quad (3.28)$$

using (3.24), (3.19) and (3.21).

The existence of an infeasible region is not a serious drawback, because for $p_A \ll 1$, which is the case with rare events, the extent of the feasible region is very large. For example, for $n = 2$ and $p_A = 10^{-3}$, the feasible set is given by

$$\{(p_A, 1), (1, p_A)\} \cup \{(p_1, p_2) : 0.001004 \leq p_1 \leq 0.995996, p_2 = p_A/p_1\}$$

Note that the bounds of p_1 are very close to unconstrained bounds, p_A and 1, hence there is a large allowable region for (p_1, p_2) such that the product-form estimator performs better than the direct estimator.

Chapter 4

Simulation Results

In this chapter, we present the results of our simulation experiments on the network of tandem queues studied in the previous chapters. Specifically we compare the results of three simulation methodologies discussed so far, namely

- i) Direct simulation based on standard Monte Carlo estimation, which we briefly call Direct Simulation,
- ii) Quick simulation based on exponential change of measure (λ interchanged with $\min(\mu_1, \mu_2)$), which we briefly call Quick Simulation, and
- iii) Quick simulation based on product-form estimation, which we briefly call Dynamic Simulation.

The above are abbreviated from now on by SS, QS and DS, respectively.

The simulations have been run to estimate the overflow probability p_A . First, we fixed the run-times to be able to compare the convergence rates of estimates. The results are shown in Table 4.1 and Table 4.2 for the $(0.20, 0.30, 0.50)$ -network and the $(0.20, 0.30, 0.50)$ -network respectively for various values of n .

It is apparent that both QS and DS perform much better than SS. Moreover, in the $(0.20, 0.30, 0.50)$ -network, QS seems to be superior to DS, while in the $(0.20, 0.38, 0.42)$ -network, they yield comparably good estimates. The empirical convergence rates of SS, QS and DS estimates are shown in Figure 4.1

and Figure 4.2 for some set of parameters in each network. It can also be seen from these figures that the performance of QS is remarkably degraded in the (0.20, 0.38, 0.42)-network. These results are not coincidences resulting from the randomness of the estimates, as we will see in a while when we consider the empirical speed-up factors.

We define the speed up factor between two simulation methods as the ratio of expected number of Markovian jumps that must be generated in each to obtain the same variance for the output estimates. To evaluate speed-up factors, we recall the variance expressions for the three types of estimators:

Standard Estimator:

$$\text{Var}[\hat{p}_A] = \frac{1}{L}(p_A - p_A^2) \quad \text{and} \quad C_{\hat{p}_A}^2 = \frac{1}{L}\left(\frac{1}{p_A} - 1\right) \quad (4.1)$$

Importance Sampling Estimator:

$$\text{Var}[\bar{p}_A] = \frac{1}{L^*}(\eta - p_A^2) \quad \text{and} \quad C_{\bar{p}_A}^2 = \frac{1}{L^*}\left(\frac{\eta}{p_A^2} - 1\right) \quad \text{with} \quad \eta = E_{P^*}\left[\frac{dP}{dP^*}\right] \quad (4.2)$$

Product-form Estimator:

$$C_{\bar{p}_A}^2 = \sum_{i=1}^n \frac{1}{L_i} C_i^2 \quad \text{with} \quad C_i^2 = \left(\frac{1}{p_i} - 1\right) \quad (4.3)$$

Let M and M^* be the mean cycle lengths in SS and QS respectively and let M_i for $i = 1, \dots, n$ be the mean length of the i^{th} cycle in DS (see Section 2.3 and 3.2 for the definition of cycles). Then the speed-up between QS and SS is

$$S_{QS-SS} = \frac{LM}{L^*M^*}$$

where L and L^* are to be chosen such that $\text{Var}[\hat{p}_A] = \text{Var}[\bar{p}_A]$. On the other hand, the speed up between DS and SS is

$$S_{DS-SS} = \frac{LM}{\sum_{i=1}^n L_i M_i} \quad (4.4)$$

where L and L_i , $i = 1, \dots, n$ are to be chosen such that $C_{\hat{p}_A}^2 = C_{\bar{p}_A}^2$.

From (4.1) and (4.2), S_{QS-SS} can be easily computed as

$$S_{QS-SS} = \frac{(p_A - p_A^2)M}{(\eta - p_A^2)M^*} \quad (4.5)$$

Recall that, an optimum allocation expression for L_i 's were given in (3.22) assuming that average sampling times (cycle length, in this case) were equal at each step. Dropping this assumption, a similar derivation can be carried out to find the optimal allocation for the case of variable M_i 's, which after some algebra, leads to the following speed-up expression for DS

$$S_{DS-SS} = \frac{\cdot \left(\frac{1}{p_A} - 1\right)M}{\left[\sum_{i=1}^n C_i M_i^{1/2}\right]^2} \quad (4.6)$$

Simulations have been run for extensively long times to get accurate estimates of M , M^* , η and M_i , $i = 1, \dots, n$. The resulting empirical values have been inserted into (4.5) and (4.6) to obtain the empirical speed-ups. The results are listed in Table 4.3 and Table 4.4.

REMARK 1. Note from Table 4.3 and Table 4.4 that S_{DS-SS} increases very fast with n . To understand the reason for this behavior, it is sufficient to consider how p_i 's change with n . Actually, p_i 's do not change with n , since p_i is the exit probability from boundary S_{i-1} to S_i , and enlarging the final boundary S_n does not affect the transition probabilities from the sub-boundaries. It is only the number of p_i 's that changes, which increases simulation time roughly linearly. Moreover, each added p_i should be approximately equal to the previous p_i 's so as to result in an exponential decrease in their product p_A , as suggested by the large deviation theory and observed in the simulation results.

REMARK 2. We see that the performance of QS in the (0.20, 0.38, 0.42)-network is very bad for small n , but recovers as n gets larger. Recall that the exponential change of measure concentrates the probability on the most dominant exit tube. In this case, however, the most dominant tube cannot be isolated since μ_1 is close to μ_2 , i.e. there exist sub-dominant exit tubes which contribute considerably to the exit probability. Therefore, the asymptotic results of large deviation theory are not valid in this network when n is on the order of 20. Actually, it has been observed by long simulation experiments that, the convergence rate of QS is very slow for small n . DS, on the other hand, is insensitive to the existence of a dominant exit point, hence its performance remains almost unchanged in the (0.20, 0.38, 0.42)-network.

$\lambda = 0.20, \mu_1 = 0.30, \mu_2 = 0.50$		Direct Simulation	Quick Simulation	Dynamic Simulation
$n=20$ $p_A \approx 3.76 \times 10^{-4}$	# of jumps	314,213	287,838	288,390
	CPU time	3.40 sec.	3.47 sec.	3.46 sec.
	Estimate	5.71×10^{-4}	3.73×10^{-4}	3.19×10^{-4}
$n=25$ $p_A \approx 4.96 \times 10^{-5}$	# of jumps	1,187,642	1,048,736	1,085,630
	CPU time	12.99 sec.	12.72 sec.	12.94 sec.
	Estimate	1.25×10^{-5}	4.94×10^{-5}	4.98×10^{-5}
$n=30$ $p_A \approx 6.52 \times 10^{-6}$	# of jumps	2,091,330	1,864,668	1,922,855
	CPU time	23.11 sec.	23.11 sec.	23.01 sec.
	Estimate	0	6.46×10^{-6}	6.48×10^{-6}

Table 4.1. Simulation results for the (0.20, 0.30, 0.50)-network

$\lambda = 0.20, \mu_1 = 0.38, \mu_2 = 0.42$		Direct Simulation	Quick Simulation	Dynamic Simulation
$n=15$ $p_A \approx 4.68 \times 10^{-4}$	# of jumps	241,793	223,329	226,973
	CPU time	2.81 sec.	2.72 sec.	2.73 sec.
	Estimate	3.00×10^{-4}	4.16×10^{-4}	4.31×10^{-4}
$n=20$ $p_A \approx 2.15 \times 10^{-5}$	# of jumps	966,414	885,648	849,988
	CPU time	10.62 sec.	10.97 sec.	10.05 sec.
	Estimate	0	1.77×10^{-5}	2.14×10^{-5}
$n=25$ $p_A \approx 9.02 \times 10^{-7}$	# of jumps	1,453,114	1,333,135	1,373,159
	CPU time	16.17 sec.	16.22 sec.	16.14 sec.
	Estimate	0	7.21×10^{-7}	8.98×10^{-7}
$n=30$ $p_A \approx 3.81 \times 10^{-8}$	# of jumps	2,038,117	1,846,449	1,997,591
	CPU time	23.63 sec.	23.45 sec.	23.97 sec.
	Estimate	0	2.85×10^{-8}	4.17×10^{-8}

Table 4.2. Simulation results for the (0.20, 0.38, 0.42)-network

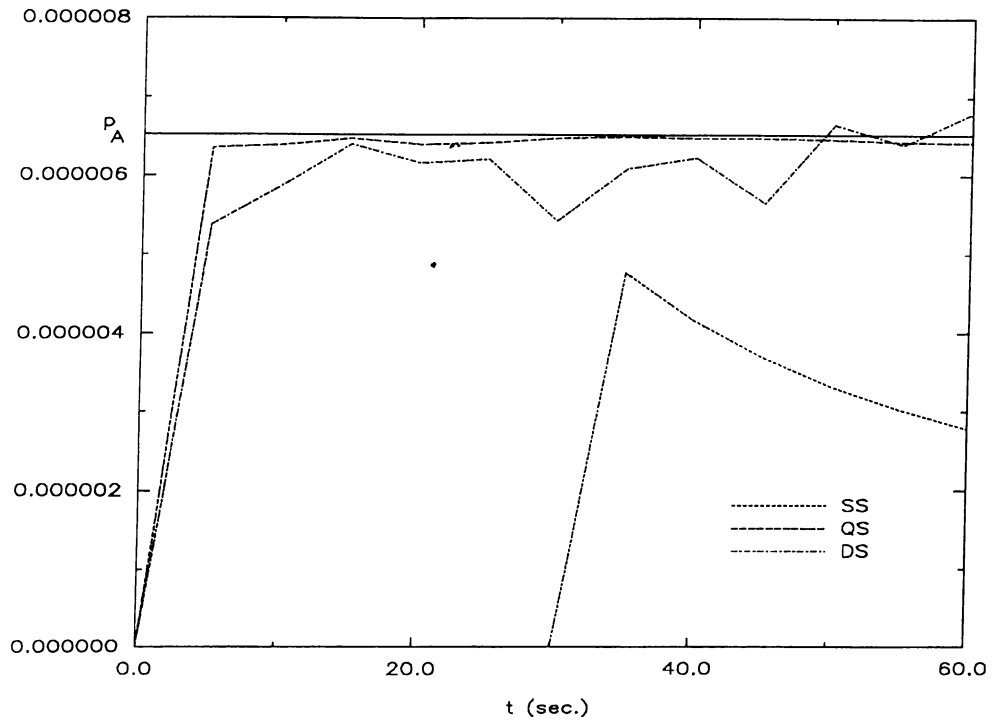


Figure 4.1. Empirical convergence curves for $\lambda = 0.20$, $\mu_1 = 0.30$, $\mu_2 = 0.50$, $n = 30$

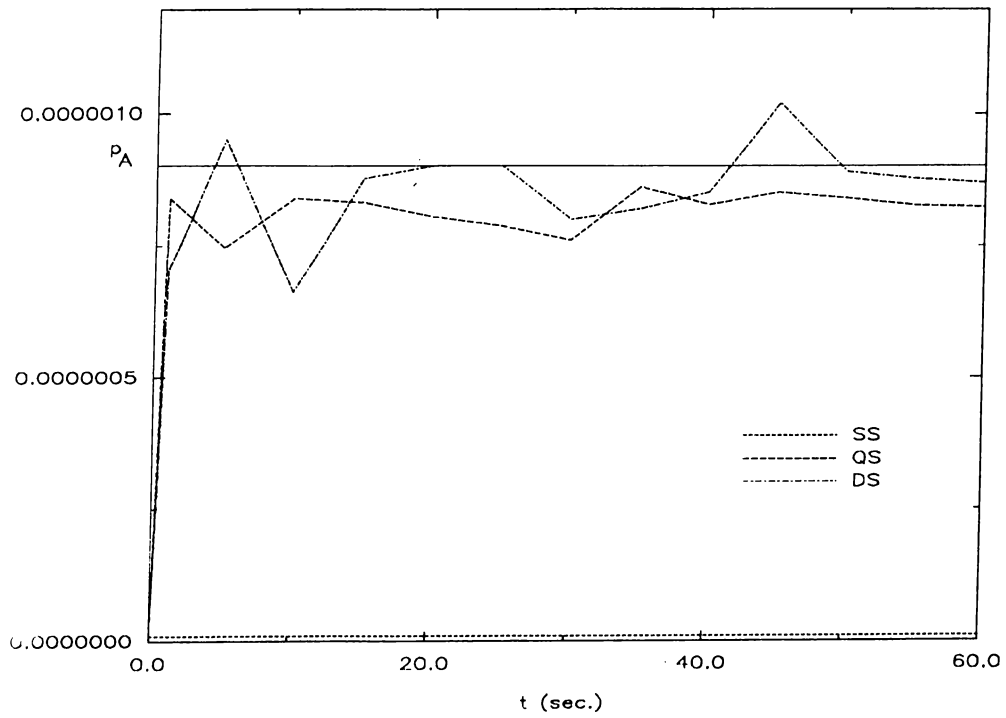


Figure 4.2. Empirical convergence curves for $\lambda = 0.20$, $\mu_1 = 0.38$, $\mu_2 = 0.42$, $n = 25$

$\lambda = 0.20, \mu_1 = 0.30, \mu_2 = 0.50$							
n	\hat{M}	\hat{M}^*	$\hat{\eta}$	$\hat{M}_i, i = 2, \dots, n$	$\hat{p}_i, i = 2, \dots, n$	\hat{S}_{QS-ss}	\hat{S}_{DS-ss}
20	15.00	57.50	7.51×10^{-7}	1.59, ..., 43.04	0.59, ..., 0.69	160	11
25	14.00	75.00	1.31×10^{-8}	1.61, ..., 64.13	0.58, ..., 0.65	870	37
30	14.95	94.39	2.19×10^{-10}	1.59, ..., 72.59	0.57, ..., 0.68	5851	170

Table 4.3. Empirical speed-up factors for the (0.20, 0.30, 0.50)-network

$\lambda = 0.20, \mu_1 = 0.38, \mu_2 = 0.42$							
n	\hat{M}	\hat{M}^*	$\hat{\eta}$	$\hat{M}_i, i = 2, \dots, n$	$\hat{p}_i, i = 2, \dots, n$	\hat{S}_{QS-ss}	\hat{S}_{DS-ss}
15	12.09	36.79	4.17×10^{-4}	1.66, ..., 32.76	0.56, ..., 0.55	0.37	12
20	12.04	52.16	9.34×10^{-6}	1.65, ..., 45.12	0.56, ..., 0.50	0.53	94
25	12.04	67.13	5.94×10^{-8}	1.67, ..., 58.09	0.54, ..., 0.51	2.72	1100
30	12.04	81.65	8.57×10^{-11}	1.65, ..., 67.16	0.56, ..., 0.55	65	15,058
35	12.04	97.41	5.60×10^{-15}	1.66, ..., 87.22	0.55, ..., 0.50	32,444	234,050

Table 4.4. Empirical speed-up factors for the (0.20, 0.38, 0.42)-network

Chapter 5

Conclusion

In this thesis, we proposed a variance reduction technique for the estimation of rare event probabilities. We obtained simulation speed-ups that are well comparable to the those of the existing techniques.

An important feature of our method is its relation to importance sampling. Actually, importance sampling is theoretically the most powerful VRT, however it presents some practical difficulties. Our sampling algorithm aims to avoid these difficulties by squeezing the samples into a sequence of sets shrinking towards the rare set, in a way, getting at each step the sampling information from the system itself. That is why, we have chosen to formulate our technique in the way we did in Chapter 3, emphasizing its dynamic character as well as its relation to importance sampling.

The feasibility of our estimator relies upon the assumptions made in Section 3.1, concerning the measurability of the introduced sets with respect to a partial set of observations and ability of sampling from the conditionals. We do not yet know how restrictive these requirements would be in practical situations, however we see clearly that the conditional sampling requirement is met whenever X_i 's are independent, yet the probability of the rare set may be difficult to estimate due to its complex nature.

The exponentially twisted estimators are known to be efficient when the rare event under consideration is governed by a large deviation principle, but even so, the asymptotic results cannot be achieved whenever the minimum rate point is not dominant enough. The performance of our estimator, on

the other hand, is not heavily dependent on this characteristic of the rare event, however, we believe that when a large deviation principle is in effect, it helps p_i 's to be distributed almost uniformly, improving the efficiency of DS. A further desirable property of DS is that it does not require the use of a change of measure, mathematics of which can be quite involved.

Although we developed the product-form decomposition to estimate probabilities, i.e expectations of indicator functions, one may naturally suspect whether it would work for arbitrary functionals of random variables other than the indicator functions. To study this problem, we think, the starting point should be expressing the functional as a product of functionals with lower variance. We leave this problem as a further research topic.

A still open question with the product form decomposition is how the distribution of p_i 's depends on the characteristics of the rare set and the underlying probability distribution. We think certain conditions on those characteristics may be developed in order to be helpful to decide whether the product-form estimator would achieve a variance reduction.

Finally, we would like to emphasize that our method is yet a new one and its feasibility in a number of more practical situations should be investigated.

Appendix

In the actual product-form estimator described in Section 3.1, the samples $(X_1, \dots, X_i)^j$ in the $(i+1)^{th}$ step are not generated from the true distribution $p^*(x_1, \dots, x_i)$ but drawn from a set of samples obtained in the i^{th} step, which is mathematically equivalent to constructing an empirical distribution $\hat{p}^*(x_1, \dots, x_i)$ from the available samples and using it as an estimate of $p^*(x_1, \dots, x_i)$. In the following, we derive a recursive relation between the estimates \hat{p}^* used in each step.

In the $(i+1)^{th}$ step, $\hat{p}^*(x_1, \dots, x_i)$ is constructed from the samples in the i^{th} step which fall into B_i . So, we have

$$\hat{p}^*(x_1, \dots, x_i) = I_{B_i}(x_1, \dots, x_i) \left[\frac{\sum_{j=1}^{L_i} I_{(x_1, \dots, x_i)}((X_1, \dots, X_i)^j)}{\sum_{j=1}^{L_i} I_{B_i}((X_1, \dots, X_i)^j)} \right] \quad (\text{A.1})$$

where $(X_1, \dots, X_i)^j$ for $j = 1, \dots, L_i$ are sampled from the distribution $\hat{p}^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})$.

For simplicity of notation, we abbreviate the second term as follows:

$$\frac{\sum_{j=1}^{L_i} I_{(x_1, \dots, x_i)}((X_1, \dots, X_i)^j)}{\sum_{j=1}^{L_i} I_{B_i}((X_1, \dots, X_i)^j)} = \frac{N_x}{N_{B_i}} \quad (\text{A.2})$$

First, we evaluate the expectation of (A.2).

$$\begin{aligned} E\left[\frac{N_x}{N_{B_i}}\right] &= \sum_m P\{N_{B_i} = m\} \frac{1}{m} E[N_x | N_{B_i} = m] \\ &= \sum_m P\{N_{B_i} = m\} \frac{1}{m} E[m I_{(x_1, \dots, x_i)}(X_1, \dots, X_i) | (X_1, \dots, X_i) \in B_i] \\ &= E[I_{(x_1, \dots, x_i)}(X_1, \dots, X_i) | (X_1, \dots, X_i) \in B_i] \\ &= \frac{\hat{p}^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})}{\sum \cdots \sum_{B_i} \hat{p}^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})} \end{aligned} \quad (\text{A.3})$$

Substituting (A.3) into (A.1) and taking expectation once more, we get

$$E[\hat{p}^*(x_1, \dots, x_i)] = I_{B_i}(x_1, \dots, x_i) E\left[\frac{\hat{p}^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})}{\sum \dots \sum_{B_i} \hat{p}^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1})}\right]$$

Assuming that $\hat{p}^*(x_1, \dots, x_{i-1})$ is unbiased, the denominator inside the expectation will be very close to $P(B_i|B_{i-1})$ for sufficiently large L_i . Replacing this term, we get an approximation

$$\begin{aligned} E[\hat{p}^*(x_1, \dots, x_i)] &\approx I_{B_i}(x_1, \dots, x_i) \frac{E[\hat{p}^*(x_1, \dots, x_{i-1})]p(x_i|x_1, \dots, x_{i-1})}{P(B_i|B_{i-1})} \\ &= p^*(x_1, \dots, x_i) \end{aligned} \quad (\text{A.4})$$

where the last inequality follows from (3.9). The validity of (A.4) is directly related to the validity of the approximation

$$\sum \dots \sum_{B_i} \hat{p}^*(x_1, \dots, x_{i-1})p(x_i|x_1, \dots, x_{i-1}) \approx P(B_i|B_{i-1})$$

or equivalently that of

$$E[\hat{p}_i] \approx p_i$$

which means that, as long as \hat{p}_i is a good estimate of p_i , which should naturally be the case, $\hat{p}^*(x_1, \dots, x_{i-1})$ is a good estimate of $p^*(x_1, \dots, x_{i-1})$. So if, disregarding the bias mentioned above, L_i 's are chosen so as to get accurate estimates of p_i 's, the resulting bias will be insignificant.

References

- [1] V. J. Rego and V. S. Sunderam, "Experiments in concurrent stochastic simulation: The Eclipse Paradigm", *Journal of Parallel and Distributed Computing*, vol. 14, pp. 66-84, 1992.
- [2] A. M. Law and W. D. Kelton, *Simulation modeling and analysis*. McGraw-Hill, 1982.
- [3] R. Righter and J. C. Walrand, "Distributed simulation of discrete event systems", *Proc. IEEE*, vol. 77, pp. 99-113, 1989.
- [4] J. A. Bucklew, *Large deviation techniques in decision, simulation, and estimation*. Wiley, 1990.
- [5] M. Cottrell, J. C. Fort and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms", *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 907-920, 1983.
- [6] J. S. Sadowsky and J. A. Bucklew, "On large deviations theory and asymptotically efficient Monte Carlo estimation", *IEEE Trans. Inform. Theory*, vol. 36, pp. 579-588, 1990.
- [7] M. C. Jeruchim, "Techniques for estimating the bit error rate in the simulation of digital communication systems", *IEEE Selected Areas Commun.*, vol. SAC-2, no. 1, pp. 153-170, 1984.
- [8] D. Lu and K. Yao, "Improved importance sampling technique for efficient simulation of digital communication systems", *IEEE Selected Areas Commun.*, 1988.
- [9] G. Orsak and B. Aazhang, "On the theory of importance sampling applied to the analysis of detection systems", *IEEE Trans. Commun.*, vol. COM-30, no. 4, pp. 332-339, 1989.

- [10] K. S. Shanmugam and P. Balaban, "A modified Monte Carlo simulation technique for the evaluation of error rate in digital communication systems", *IEEE Trans. Commun.*, vol. COM-28, no. 11, pp. 1916-1924, 1980.
- [11] J. S. Sadowsky, "Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/m queue", *IEEE Trans. Automat. Contr.*, vol. 36, no. 12, pp. 1383-1394, 1991.
- [12] S. Parekh and J. C. Walrand, "A quick simulation method for excessive backlogs in networks of queues", *IEEE Trans. Automat. Contr.*, vol. 34, no. 1, pp. 54-66, 1989.
- [13] E. E. Lewis and F. Böhm, "Monte Carlo simulation of Markov unreliability models", *Nuclear Eng. and Design*, vol. 77, pp. 49-62, 1984.
- [14] D. Siegmund, "Importance sampling in the Monte Carlo study of sequential tests", *Ann. Statistics*, vol. 4, pp. 673-684, 1976.
- [15] G. Orsak and B. Aazhang, "Constrained solutions in importance sampling via robust statistics", *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 307-316, 1991.
- [16] P. Billingsley, *Convergence of probability measures*. New York: Wiley, 1968.
- [17] A. D. Ventsel, "Rough limit theorems on large deviations for Markov stochastic processes II", *Theory Prob. Appl. (USSR)*, vol. 21, pp. 499-512, 1976.
- [18] M. Cottrell, J. C. Fort and G. Malgouyres, "Evénements rares pour l'étude de certains algorithmes stochastiques", l'Univ. Paris-Sud, Orsay, France, Prepubl. 80 T 35.
- [19] G. Ruget, "Quelques occurrences des grands écarts dans la littérature électronique", *Asterisque*, vol. 68, pp. 187-199, 1979.
- [20] J. C. Walrand, *An introduction to queueing networks*. Prentice-Hall, 1988.
- [21] M. R. Frater, T. M. Lennon and B. D. A. Anderson, "Optimally efficient estimation of statistics of rare events in queueing networks", *IEEE Trans. Automat. Contr.*, vol. 36, no. 12, pp. 1395-1405, 1991.