UNIVERSITY OF OSLO
Department of Informatics

Queuing systems: Modeling, analysis and simulation

Frode B. Nilsen

Research Report 259

ISBN 82-7368-185-8
ISSN 0806-3036

April 1998

Queuing systems: Modeling, analysis and simulation

Frode B. Nilsen

Research Report 259
ISBN 82-7368-185-8, ISSN 0806-3036
Department of Informatics, University of Oslo.

April 1998

Abstract

This document gives an overview the standard methods used to model and analyze the performance of queuing systems. This includes both analytical techniques and simulation methodology. Simulation must be used when the complexity of the system makes mathematical analysis intractable. This normally happens when the underlying model fails to have Markov behavior.

Attention is restricted to single-queue systems and the $M/M/1$ queue is used as an illustrative example regarding analytical work. A number of specific results concerning performance figures can then be derived. Simulation is discussed in the setting of the $G/G/1$ queue and various sampling-strategies are central to the discussion. The need to perform statistical output analysis is also emphasized.

Contents

Preface

This report was written as a part of my work with a doctoral dissertation at the Department of Informatics, University of Oslo. The work is supported by grant no. 100722/410 from the Norwegian Research Council. The contact address of the author is

Frode B. Nilsen
<froden@ifi.uio.no>
<http://www.ifi.uio.no/~froden>
Department of Informatics, University of Oslo
P.O. Box 1080, Blindern
N-0316 Oslo, Norway

Chapter 1

Introduction

This document gives an overview the standard methods used to model and analyze the performance of queuing systems. Queuing is always due to unpredictable demands for a limited set of resources. It is customary to use an abstraction where "customers" are being "served". Typical performance figures are the mean number of customers in the queue and the expected waiting time before access to the service facility is granted.

Unpredictability is best described in statistical terms and stochastic modeling is central to the discussion. Depending on the complexity of the model, qualitative or quantitative results can be derived by mathematical analysis. If the performance figures of interest cannot be determined by analytical means, simulation is inevitable.

Both analytical techniques and simulation methodology are discussed in this document. For simplicity attention is restricted to *single-queue* systems. The $M/M/1$ model is used as illustrative example for analytical work. Due to mathematical tractability fundamental principles and notions can be introduced in a clear-cut way. These concepts carry over to more general situations where results are otherwise often obscured by mathematical complexity. Simulation is discussed in the setting of a more general $G/G/1$ queuing model.

It is important to be aware that the tractability of the $M/M/1$ system is due to simplifying assumptions. We point at what makes the model tractable and what will typically break the tractability in more complex models. Based on the analysis of the $M/M/1$ queue we also outline some principal extensions applicable to more complex systems.

Note that simulation issues are restricted to experimental design, sampling strategies and output analysis. Discussion of appropriate modeling and simulation tools are outside the scope of this report. Note also that we discuss performance evaluation under the assumption that there is no real systems available for measurements.

The reader is assumed to be familiar with basic mathematical probability and statistics. This includes the concept of stochastic processes and statistical inference. For unfamiliar readers [3, 21, 33] is a suite of introductory references. Any prior exposure to stochastic discrete-event simulation methodology is also useful. Pointers to general texts on this subject are [1, 2, 8, 23, 28]. In addition to reading this document the reader is urged to taking a look at textbooks on queuing theory. Two useful references in this respect are [6, 19]

1.1 Single-queue systems

Consider a system where customers arrive at some service facility. The reader may think of an post-office. After being served the customers depart from the system. By assumption the customer inter-arrival intervals and also customer service periods are subject to statistical fluctuations. Hence, a queue will occasionally form in front of the service facility. If there are no customers waiting to be served the system is said to be idle. Otherwise the system is busy. Under normal circumstances the system will cyclically alternate between idle and busy periods. A customer arriving at an empty system terminates an idle period and a busy period begins. Correspondingly, a customer leaving an otherwise empty system terminates a busy period and initiates an idle period.

A single-queue system is characterized by having a single waiting line organized according to some queuing discipline. We restrict attention to FIFO queues with no priorities among customers. It is also assumed that customers will not balk from the queue once they have arrived. Equipped with these assumptions, a single-queue system is classified according to the notation $A/B/s/N$. Here A characterize the input process and refers to the probability distribution of inter-arrival times of customers. Likewise, B characterize the service process and refers to the probability distribution of service time for customers. The s component is the number of parallel stations at the service facility. Finally, N specifies the capacity of the waiting line in front of the service facility. For infinite capacity systems the N component is normally omitted.

1.2 Organization

The rest of this document is organized as follows. Chapter 2 establishes the idea of a birth-death Markov process which is at the core of analytical queuing theory. A number of key concepts like transition probability, memoryless property and transient vs. steady-state behavior are also introduced.

This is followed by an analysis of the $M/M/1$ system in chapter 3. Assuming that both inter-arrival and service times are independent and exponentially distributed gives the simplest example of a birth-death process. The most important parameter for this system is the normalized load. The performance figures subject to analysis are the number of customers in the queue and the expected waiting time. The former is based on a state-based view whereas the latter builds on a transaction-based formulation. Little's law which relates the two views is also discussed.

Chapter 4 looks beyond the $M/M/1$ model and discusses how analytical tractability depends on the memoryless property of Markov models. For intractable systems stochastic simulation must be used. This is addressed in chapter 5. Estimators, point estimates and confidence intervals are central to the discussion.

In chapter 6 steady-state simulation of a general $G/G/1$ queue is considered. This provides an opportunity to discuss various sampling strategies like replicate-runs, equally-spaced observations, asynchronous observations and regenerative cycles. The former is time-consuming but always work. The other are more efficient single-run strategies, with asynchronous sampling being most efficient. Regenerative sampling is convenient as it has no problems associated with the transient warm-up period. The chapter is closed with a discussion of indirect estimation by way of Little's law.

Issues related to run-length control and estimation efficiency are discussed in chapter 7. The report is concluded in chapter 8.

Chapter 2

Birth-death Markov processes

Stochastic birth-death Markov processes turns out to be a highly suitable modeling tool for many queuing processes. Several examples will be considered throughout this document. In this section we shortly preview the general features of birth-death Markov processes. More information on the subject can be found in [4, 6, 19, 33].

Let $N(t)$ be an integer-valued continuous-time stochastic process. The discrete state space of the process comprises non-negative integer values $0, 1, \ldots, \infty$. At this point we discuss the $N(t)$ process without any particular physical meaning attached. However, as a conceptual aid the reader may think of $N(t)$ as being the random number of members in some population as a function of time.

By assumption the classical Markov property is imposed as a restriction on the process $N(t)$. I.e. given the value of $N(s)$ the values for $N(s+t)$ for $t > 0$ are *not* influenced by the values of $N(u)$ for $u < s$. In words, the way in which the entire past history affects the future of the process is completely summarized in the current state of the process. Expressed analytically the Markov property may be written as

$$P[N(t_{m+1}) = n_{m+1} \mid N(t_m) = n_m, \ldots, N(t_1) = n_1] = P[N(t_{m+1}) = n_{m+1} \mid N(t_m) = n_m] \quad (2.1)$$

and it should be valid for all $t_1 < t_2 < \cdots < t_m < t_{m+1}$ and any m.

2.1 Transition probabilities

In equation (2.1) set $t_m = s$, $n_m = i$, $t_{m+1} = s+t$ and $n_{m+1} = j$. Then the right-hand side of the equation expresses the probability that the the process makes a transition from state i at time s to state j in time t relative to s. Such a probability, denoted $p_{i,j}(s,t)$, is referred to as a state transition probability for the Markov process. In this document we only consider transition probabilities being independent of absolute time s. I.e. for all $s > 0$ we have

$$p_{i,j}(s,t) = p_{i,j}(t) = P[N(t) = j \mid N(0) = i] = P[N(s+t) = j \mid N(s) = i]$$

This is called time-homogeneous or stationary transition probabilities. Henceforth time-homogeneity is tacitly assumed. It is generally assumed that the transition probabilities $p_{i,j}(t)$ are well behaved in the sense that they are all continuous and the derivative exists.

For a Markov process with time-homogeneous transition probabilities the so-called Chapman-Kologomorov equation applies

$$p_{i,j}(t+s) = \sum_{k=0}^{\infty} p_{i,k}(t) p_{k,j}(s) \tag{2.2}$$

This equation states that in order to move from state i to j in time $(t+s)$, the queue size process $N(t)$ moves to some intermediate state k in time t and then from k to j in the remaining time s. It also says how to compute the long-interval transition probability from a sum of short-interval transition probability components.

An infinitesimal transition probability, denoted $p_{i,j}(dt)$, specifies the *immediate* probabilistic behavior of a Markov process in that $dt \to 0$. By help of equation (2.2) it turns out that any transition probability $p_{i,j}(t)$ can in principle be determined if the infinitesimal transition probabilities are known. Hence, the overall probabilistic behavior of a Markov process is ultimately given by the infinitesimal transition probabilities. Together they define the transition kernel of the process.

A birth-death Markov process is characterized by the fact that the discrete state variable changes by at most one, if it changes at all, during an infinitely small time interval. Reflecting this fact, the following postulations specify the transition kernel of a general birth-death Markov process

$$\begin{array}{lcll} p_{i,i+1}(dt) & = & \lambda_i\, dt + \mathrm{o}(dt) & i = 0, 1, \ldots, \infty \\ p_{i,i-1}(dt) & = & \mu_i\, dt + \mathrm{o}(dt) & i = 0, 1, \ldots, \infty \\ p_{i,i}(dt) & = & [1 - (\lambda_i + \mu_i)]dt + \mathrm{o}(dt) & i = 0, 1, \ldots, \infty \\ p_{i,j}(dt) & = & \mathrm{o}(dt) & |i-j| = 2, 3, \ldots, \infty \end{array} \tag{2.3}$$

Here $\mathrm{o}(dt)$ is a quantity such that $\lim_{dt\to 0} \mathrm{o}(dt)/dt = 0$. The first equation handles the case when the state variable increases by one. This is referred to as a single birth. Here λ_i is a proportionality constant such that the product $\lambda_i\, dt$ should reflect the probability for a single birth to happen during the infinitesimal time interval. We may treat λ_i as a parameter without any particular meaning attached to it. However, it is customary to interpret λ_i as the instantaneous birth rate. Likewise, the second equation is for the case when the state variable is reduced by one. This is referred to as single death. The product $\mu_i\, dt$ signifies the probability that a single death takes place. Then μ_i denote the instantaneous death rate. The third equation handles the case when the state variable does not change. I.e. $[1 - (\lambda_i + \mu_i)]dt$ reflects the probability that neither a single birth nor a single death occur during the infinitely small time interval. Multiple births, multiple deaths and simultaneous births and deaths are taken care of by the $\mathrm{o}(dt)$ terms in the equations. This should be interpreted such that the probability for these events to happen is negligible as $dt \to 0$. We say that multiple events are prohibited.

Note that the infinitesimal transition probabilities from (2.3) are in general state dependent. This is so since the instantaneous birth rate λ_i and also the death rate μ_i may depend on the departing state i. A small comment also applies to the second and third equations. Since no deaths can occur if the state variable is already zero, i.e. if $i = 0$, we always define $\mu_0 = 0$.

2.2 Model equation

By combining equation (2.2) with the infinitesimal transition probabilities from (2.3), we may write

$$p_{i,j}(t+dt) = p_{i,j-1}(t)\lambda_{j-1}\,dt + p_{i,j}(t)[1-(\lambda_j+\mu_j)]dt + p_{i,j+1}(t)\mu_{j+1}\,dt + \mathrm{o}(dt)$$

where all $\mathrm{o}(dt)$ terms from (2.3) are now summarized in a single term. By rearranging, division by dt and taking the limit as $dt \to 0$ we arrive at the following where the time-derivative of the transition probability now appears on the left-hand side.

$$p'_{i,j}(t) = \lambda_{j-1}p_{i,j-1}(t) - (\lambda_j+\mu_j)p_{i,j}(t) + \mu_{j+1}p_{i,j+1}(t) \tag{2.4}$$

This is the general model equation for a birth-death Markov process and it essentially captures the probabilistic dynamics of the process. The equation is a differential equation in the continuous time variable t and a difference equation[1] in the discrete state variable j.

Depending on the particular values of λ_i and μ_i in equation (2.4) it may be possible to solve the model equation so as to get a closed-form expression for $p_{i,j}(t)$. This is referred to as the transient solution of the stochastic process model. The transient solution completely characterizes the time-dependent probabilistic behavior of a birth-death Markov process. In the next section we consider a different kind of solution of the model equation. As opposed to the transient solution this is called a steady-state solution.

Note that the model equation (2.4) is valid for $t > 0$ and $i = 0, 1, \ldots, \infty$. For $t = 0$ we have a boundary condition and it is customary to define

$$p_{i,j}(0) = \delta_{i,j} \tag{2.5}$$

where $\delta_{i,j}$ is the Kronecker delta defined as 1 if $i = j$ and 0 otherwise. Hence, in zero time the process will certainly not move.

2.3 Steady-state solution

Consider an arbitrary point s in time at which the process $N(s) = i$. From this point on the time-dependent probabilistic behavior of the process is given by the transient solution $p_{i,j}(t)$ where t is taken relative to s. In this context the transition probability $p_{i,j}(t)$ represents the probability that the process will be in state j after an incremental time t. Henceforth we refer to $p_{i,j}(t)$ as a *state probability* and it is tacitly understood that it is conditioned on the fact that the observation of the process started in state i at time s.

By now considering the limit of the transient solution as $t \to \infty$ it is interesting to see if the state probabilities eventually settle down. I.e. for a given departing state i we are interested in a family of limits

$$\lim_{t\to\infty} p_{i,j}(t) = p_{i,j} \tag{2.6}$$

[1] Also often called a recurrence equation.

for $j = 0, 1, \ldots, \infty$. If such a family exists in the sense that $p_{i,j} > 0$ for all j and if $\sum_{j=0}^{\infty} p_{i,j} = 1$, then $p_{i,j}$ represents the limiting probability distribution of the state variable $N(t)$, given that we started in state i. Alternatively we may say that as $t \to \infty$ the stochastic process $N(t)$ converge in distribution [3, 24] to a random variable N_i having $p_{i,j}$ as its probability distribution over j. This is written

$$N(t) \Rightarrow N_i$$

In some cases it may be that a family of limits does exist but that every member approaches zero. Then $p_{i,j}$ is called a degenerate limiting distribution. Henceforth is it tacitly assumed that a limiting distribution refers to the non-degenerate case.

If $N(t) \Rightarrow N_i$ for some random variable N_i we say that a statistical equilibrium or steady-state is associated with the process. The corresponding limiting distribution $p_{i,j}$ is referred to as a steady-state solution of the stochastic process model.

It should be emphasized that based on the above discussion we can *not* conclude that $p_{i_1,j} = p_{i_2,j}$ for $i_1 \neq i_2$ and all j. The possibility exists that that the limiting distribution is not unique but depends on the initial state i. Fortunately, for a birth-death Markov process model it can be shown that if a limiting distribution *do* exist then it is unique. I.e.

$$p_{i,j} = p_j \tag{2.7}$$

for all i and the process converges towards the same limiting distribution regardless of initial state i. In other words the effect of the initial state is not apparent under steady-state.

The limiting distribution is always asymptotically stationary or invariant in the sense that

$$p_j = \sum_{k=0}^{\infty} p_k p_{k,j}(t)$$

for all t when steady-state prevails. This equation follows easily from equation (2.2) by taking the limit as $s \to \infty$ and then employing the definitions from equations (2.6) and (2.7). This tells us that when the state probabilities first equals the stationary distribution, then at any additional time t into the future the state probabilities will remain unchanged. The reader is warned at this point. A stationary distribution does *not* mean that the process has lost its probabilistic behavior. Even if the state probabilities become time-independent constant values, they are still probabilities.

Note carefully that the concept of statistical equilibrium relates not only to the properties of the process itself, but also to the observer's knowledge of the process. In the above discussion we have assumed that an observer finds the process in state i if he were to look at any time s. If he were to look again at a later time $(s + t)$, where t is a finite incremental time, the probability that he will find the process in state j is given by the transient solution. As opposed to this, if equilibrium had prevailed at time s and the observer had *not* looked, then the corresponding probability would be given by the associated steady-state solution.

Assuming the existence of a limiting distribution p_j we may consider the corresponding transient solution $p_{i,j}(t)$ and take the limit as $t \to \infty$ in order to arrive at an expression for the limiting distribution. In many cases, however, it is impossible to solve the model equation (2.4) for the transient solution. Then we must use the following approach to find the limiting distribution. Recall that the derivative $p'_{i,j}(t)$ appears

on the left-hand side in equation (2.4). Under steady-state conditions this derivative must be zero. Consequently, we have

$$\begin{aligned} \lim_{t\to\infty} p'_{i,j}(t) &= \lim_{t\to\infty} [\lambda_{j-1}p_{i,j-1}(t) - (\lambda_j + \mu_j)p_{i,j}(t) + \mu_{j+1}p_{i,j+1}(t)] \\ 0 &= \lambda_{j-1}p_{j-1} - (\lambda_j + \mu_j)p_j + \mu_{j+1}p_{j+1} \end{aligned} \tag{2.8}$$

so that for steady-state the original differential-difference model equation reduces to a time-independent difference equation which is in general much easier to solve.

2.4 Memoryless property

Since the future probabilistic behavior of a Markov process depends only on the current state as expressed by equation (2.1), it is customary to say that a Markov process is memoryless. This fact is clearly illustrated if we consider the holding times or sojourn times [33] of a Markov process. At entrance to a specific state the corresponding sojourn time is defined as the time spent in that state before the process makes a transition to a different state.

For a transition to an arbitrary state, let S be a random variable denoting the corresponding sojourn time in that state. By help of the Markov property alone it can be shown [19, 33] that any sojourn time S must be distributed according to an exponential function

$$P[S \leq s] = 1 - e^{-\gamma_i s} \tag{2.9}$$

where γ_i is generally left as an unspecified parameter which may depend upon the sojourning state i. In the case of birth-death Markov processes it can be shown that this parameter relates to the infinitesimal transition probabilities by $\gamma_i = (\lambda_i + \mu_i)$. Figure 2.1 shows a plot of the exponential probability distribution for $\gamma_i = 1$ along

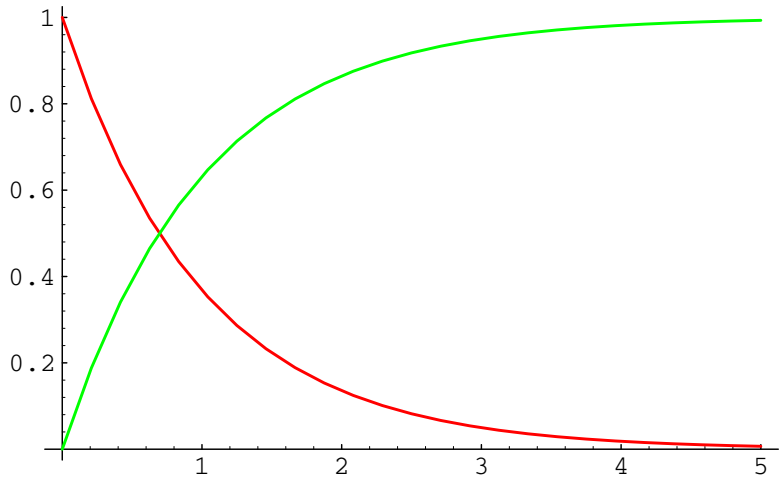

Figure 2.1: *The rising curve is a plot of the (cumulative) exponential probability distribution function* $1 - e^{-\gamma s}$ *for* $\gamma = 1$. *The falling curve is a plot of the corresponding exponential probability density function* $e^{-\gamma s}$.

with the corresponding probability density function.

The exponential distribution of sojourn time is amazing in the sense that it is truly memoryless. To see this, consider the following argument. Say that a transition to state i has just occured for the Markov process $N(t)$. The associated sojourn time S is then exponentially distributed as expressed by equation (2.9). Now let some time pass, say s_0, during which no transition away from state i occurs. At this point let S' be a random variable signifying the *remaining* sojourn time so that $S = s_0 + S'$. With this at hand the following identity can be established [19]

$$P[S' \leq s \mid S > s_0] = P[S \leq s] \tag{2.10}$$

showing that the distribution of remaining sojourn time, given that a time of s_0 has already elapsed, is identically equal to the unconditional exponential distribution of the total sojourn time. The impact of this statement is that our probabilistic feeling regarding the time until the termination of the sojourn period is independent of how long it has been since the sojourn period was actually initiated. This clearly illustrates the memoryless property of the underlying Markov process.

Chapter 3

Analyzing the $M/M/1$ queue

The $M/M/1$ queue is characterized by the features of its arrival and service processes. These processes are discussed in the next two subsections, respectively, and we will see that both processes are modeled as memoryless Markov processes. The M designation in $M/M/1$ actually refers to this memoryless/Markov feature of the arrival and service processes. Then in section 3.3 we consider an analysis of how the number of customers in the queue behave probabilistically. If we are interested in other features of the $M/M/1$ queue we must change our stochastic process view of the system. At the end we briefly consider some important cases.

3.1 Arrivals

The input process of the $M/M/1$ queue is modeled as a pure Markov birth process with state independent birth rates. An arrival plays the role of a birth and $N_a(t)$ denotes the number of arrivals in time t. With respect to (2.3) we now define $\mu_i = 0$ and $\lambda_i = \lambda$ for all i. In this case the model equation (2.4) can be solved for the transient solution giving [4, 19, 33]

$$p_{i,j}(t) = \frac{(\lambda t)^{j-i+1}}{(j-i+1)!} e^{-\lambda t}, \qquad j \geq i \geq 0$$

This is the celebrated Poisson distribution. Hence, the arrival process is a Poisson process. For a fixed departing state i and a specific time interval t, the above equation gives the (discrete) distribution of the number of arrivals $(j-i)$ in that time interval. Note that this distribution is independent of the departing state i and depends only on the difference $(j-i)$. In figure 3.1 we have plotted the Poisson distribution of $(j-i)$ for $\lambda = 0.5$ and two different time intervals. The applicability of Poisson processes in practical arrival situations is well proven [21, 33] thereby justifying the model.

With the Poisson distribution at hand it can easily be shown [19, 33] that the inter-arrival times of customers are represented by mutually i.i.d. random variables[1]. If A denote the time between any two customer arrivals we have that A is exponentially distributed

$$P[A \leq t] = 1 - e^{-\lambda t} \tag{3.1}$$

[1] Here i.i.d. denotes "independent and identically distributed".

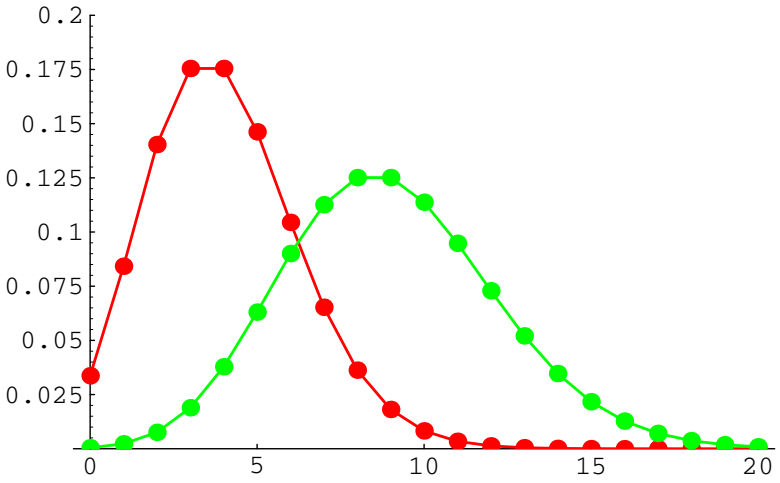

Figure 3.1: *Plot of the Poisson probability distribution of number of arrivals* $(j-i)$ *for* $\lambda = 0.5$ *and two different time intervals* $t = 10$ *and* $t = 20$*. The darkest curve corresponds to the shortest time interval.*

where λ, denoting the instantaneous arrival rate, now appears as a parameter to the exponential distribution. The mean inter-arrival time is $E[A] = 1/\lambda$. The significance of independent and exponentially distributed inter-arrival times will become apparent later.

Up to this point we have considered λ as an instantaneous arrival rate. This may seem fictitious to the reader. The following result for a Poisson process [4, 19, 33] explains the rationale of this interpretation

$$E[N_a(t)] = \lambda t$$

We see that λ reflects the expected number of arrivals in an interval of unit length, or in other words, λ is the arrival rate.

Since the number of arrivals $N_a(t)$ clearly grows without bounds as $t \to \infty$ the arrival process never reaches statistical equilibrium and no limiting distribution is associated with the process.

3.2 Departures

The service process of the $M/M/1$ queue is modeled much the same way as the arrival process. Specifically, the service process is modeled as a pure Markov death process with state independent death rates. A service completion plays the role of a death. With respect to (2.3) we now define $\lambda_i = 0$ for all i and $\mu_i = \mu$ for $i = 1, 2, \ldots, \infty$. Note that the instantaneous service rate μ is only defined for a busy system.

There are two minor complication with the service process. The first arise from the fact that the service process is typically intervened by idle periods in which the system is empty and no departures take place. This problem is resolved simply by freezing the running time variable, denoted t_e, during idle periods. The service process is then essentially built by merging the busy periods. The second complication has to do with the fact that the state variable is monotonically decreasing in a pure death process.

This problem is resolved by a redefinition of the discrete state space. Hence, let $N_s(t_e)$ denote the *negative* (integer-valued) number of service completions as a function of *effective* busy time t_e.

With this in mind, the features of the service process is completely analogous to that of the arrival process. Specifically it is a Poisson process and the service times of customers are represented by mutually i.i.d. random variables. If B denote the service time of any customer we have that B is exponentially distributed

$$P[B \leq t_e] = 1 - e^{-\mu t_e} \tag{3.2}$$

where μ, denoting the service rate, appears as a parameter to the exponential distribution. The mean service is $E[B] = 1/\mu$.

3.3 Queue size

In this section we pay attention to the statistical fluctuations of the *size* of the queue in the $M/M/1$ model. Let the continuous-time stochastic process $N(t)$ denote the (integral) number of customers in the system. Note that the number of customers in the system is defined as the number of customers queued *plus* the one in service, if any. The process $N(t)$ is modeled as a birth-death Markov process now incorporating both customer arrivals and service completions. By assumption the arrival process $N_a(t)$ and the departure process $N_s(t)$ are mutually stochastically independent. Then the process $N(t)$ essentially becomes a superposition of the $N_a(t)$ and $N_s(t)$ processes. It should be emphasized that the fact that we can model $N(t)$ by the proposed Markov process is a direct consequence of the memoryless property possessed by both the arrival and service processes.

Equipped with these definitions the general model equation (2.4) now becomes

$$p'_{i,j}(t) = \lambda p_{i,j-1}(t) - (\lambda + \mu)p_{i,j}(t) + \mu p_{i,j+1}(t), \qquad j = 1, 2, \ldots \tag{3.3}$$

Note that this equation is not defined for $j = 0$. This particular case, corresponding to the fact that customers will not depart from an empty system, leads to a boundary condition

$$p'_{i,0}(t) = -\lambda p_{i,0}(t) + \mu p_{i,1}(t) \tag{3.4}$$

3.3.1 Traffic intensity

We will soon see that the ratio between the arrival rate λ and the service rate μ plays an important role in the analysis of the queue size process. Therefore, we define the new parameter

$$\rho = \frac{\lambda}{\mu} \tag{3.5}$$

which can be interpreted as the load on the system. The load ρ is also referred to as offered load or traffic intensity and provides a *relative* measure of the demand placed on the system. Recall that μ is actually not defined for an empty system. Consequently, the traffic intensity parameter ρ should be interpreted as the load conditioned on the fact that the system is already loaded. During idle periods in which the system is unloaded an arrival can always be served immediately.

3.3.2 Transient solution

To completely characterize the time-dependent probabilistic behavior of the queue size process $N(t)$ we should find the transient solution $p_{i,j}(t)$ from the appropriate model equation. Proceeding from this point on typically involves transforming equation (3.3) with associated boundary conditions (equation (2.5) and equation (3.4)) both by a Laplace-transform step and a generating function transform step. The transformed equation is then algebraically manipulated before it is inversely transformed twice. The details of this procedure are beyond the scope of this document, though. We merely state the result [19]:

$$p_{i,j}(t) = e^{-(\lambda+\mu)t}\Big[\rho^{(j-i)/2}I_{j-i}(at) + \rho^{(j-i-1)/2}I_{j+i+1}(at) + (1-\rho)\rho^{j}\sum_{k=j+i+2}^{\infty}\rho^{-k/2}I_k(at)\Big] \quad (3.6)$$

where

$$\begin{aligned} a &= 2\mu\sqrt{\rho} \\ I_k(x) &= \sum_{m=0}^{\infty}\frac{(x/2)^{k+2m}}{(k+m)!m!} \end{aligned}$$

Here $I_k(x)$ is the modified Bessel[2] function of the first kind of order k. The traffic intensity ρ is previously defined in equation (3.5). At this point the following quote from [19] concerning equation (3.6) is appropriate:

> This last expression is most disheartening. What it has to say is that an appropriate model for the simplest interesting queuing system leads to an ugly expression for the time-dependent behavior of its state probabilities. As a consequence, we can only hope for a greater complexity and obscurity in attempting to find time-dependent behavior of more general queuing systems.

Consider the $M/M/1$ system at start-up where the queue is assumed to be empty at time $s = 0$. Then $p_{0,j}(t)$ denote the probability that there are j customers in the system at time t from start-up. In this context we refer to $p_{0,j}(t)$ as a state probability instead of a transition probability. Now it is very instructive to plot the time-dependent behavior of the state probabilities $p_{0,j}(t)$. This is shown[3] in figure 3.2 for the case $\rho = \lambda/\mu = 0.5/1.0 = 0.5$ and for $j = 0, 1, 2$. The topmost curve corresponds to $p_{0,0}(t)$ signifying the probability that there are no customers in the system at time t. Initially we certainly[4] have no customers in the system. Then this probability gradually decreases and seemingly approaches a constant level as t grows. The curve in the middle corresponds to $p_{0,1}(t)$ and the bottommost curve corresponds to $p_{0,2}(t)$. Initially these probabilities are both zero, of course. Then they grow gradually before they both seem to flatten. The fact that the state probabilities seem to converge towards distinct

[2] Bessel functions often appear in the solution of differential equations. Consult any text book on advanced calculus

[3] Note that $p_{i,j}(t)$ from equation (3.6) contains infinite sums. In plotting $p_{i,j}(t)$ such sums must be truncated. The plots in this document have been generated by the `Mathematica` program performing such numerical truncations automatically.

[4] I.e. the probability is 1.

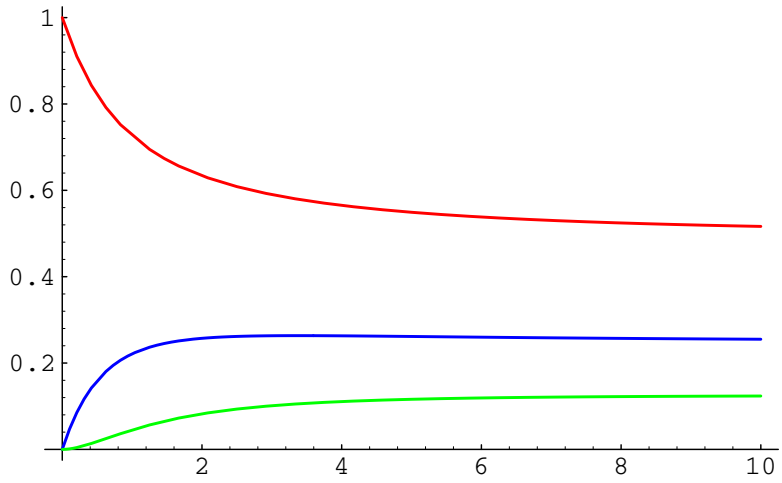

Figure 3.2: *Plot of state probabilities $p_{0,j}(t)$ for $j = 0, 1, 2$ and $\rho = \lambda/\mu = 0.5/1.0 = 0.5$. The topmost curve corresponds to $p_{0,0}(t)$ and the bottommost to $p_{0,2}(t)$. The onset of statistical equilibrium is easily identified.*

constant levels indicate that statistical equilibrium or steady-state is reached. We will return to the issue of steady-state solution in the next section.

An interesting point about figure 3.2 is the time it takes before each state probability settles down. We clearly see that $p_{0,1}(t)$ converge faster than both $p_{0,0}(t)$ and $p_{0,2}(t)$. Hence, the rate of convergence varies among the state probabilities. To get an aggregated view of the time it takes before steady-state prevails we therefore consider

$$E[N(t)] = \sum_{j=0}^{\infty} j p_{0,j}$$

being the mean number of customers in the system as a function of time t. Note that $E[N(t)]$ takes all transient state probabilities into account. Under the same conditions as in figure 3.2 the middle curve in figure 3.3 is a plot of $E[N(t)]$. Taking the different time scales of the two figures into account we conclude that $E[N(t)]$ converges slower than the individual state probabilities. The steady-state level suggests that for this case there is on the average one customer in the system when statistical equilibrium prevails.

The bottommost curve in figure 3.3 is also a plot of $E[N(t)]$ but this time for the case $\rho = \lambda/\mu = 0.25/0.5 = 0.5$. Note that the traffic intensity ρ is unchanged from the preceding case but that the absolute value of λ and μ has now changed. The figure suggests an unchanged steady-state level but a slower rate of convergence for the latter case. From this we conclude that convergence is slower with decreasing arrival intensity and service intensity. Assuming the same traffic intensity, it is intuitively reasonable that a slowly operating system reaches steady-state more slowly than a quickly operating system.

The topmost curve in figure 3.3 corresponds to a plot of $E[N(t)]$ for the case $\rho = \lambda/\mu = 0.7/1.0 = 0.7$. Compared to the other two cases the traffic intensity is now higher. The figure illustrates two points. First, the steady-state level for the average number of customers in the system increases with increasing traffic intensity. We return to this fact in the next section. Next, the rate of convergence is slower with increased

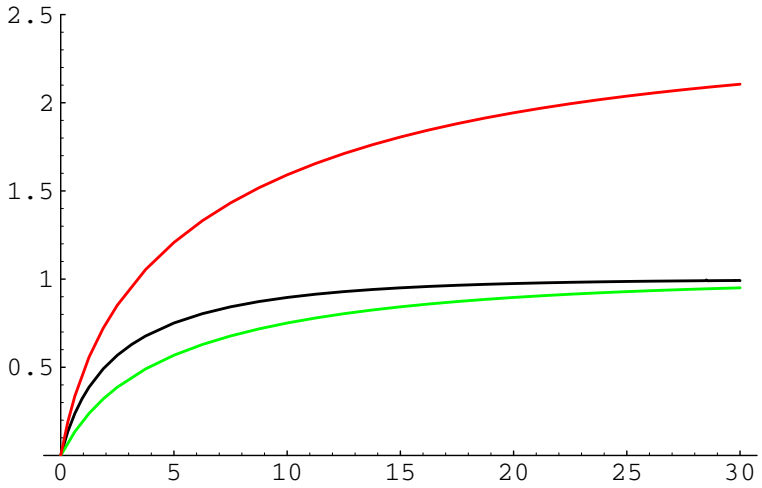

Figure 3.3: *Plot of expected number of customer $E[N(t)]$ as a function of time t. The uppermost curve corresponds to the case $\rho = \lambda/\mu = 0.7/1.0 = 0.7$. The middle curve is for the case $\rho = \lambda/\mu = 0.5/1.0 = 0.5$. The bottommost curve corresponds to the case $\rho = \lambda/\mu = 0.25/0.5 = 0.5$. Note the rate of convergence for each case.*

traffic intensity. It is intuitively reasonable that it takes more time for a highly loaded system to settle down (probabilistically) than it takes for a less loaded system.

Before leaving the transient behavior, consider figure 3.4 showing $p_{0,j}(t)$ for $j = 0, 5, 10, 15$. In this case $\rho = \lambda/\mu = 1.0/0.8 = 1.25$. Compared to figure 3.2 note that the time scale has now changed and that different state probabilities are plotted. Anyway, we observe that the characteristics of the curves are now quite different in that they cross each other and do not seem to converge. This indicates that a steady-state does not exist for the latter case.

3.3.3 Steady-state solution

As suggested by the plots from section 3.3.2 the $N(t)$ process seems to settle down probabilistically under certain circumstances. In this section we focus on the issue of statistical equilibrium and steady-state behavior.

Assuming the existence of a limiting distribution p_j we may use the general method from section 2.3 to arrive at an expression for p_j. This time, however, we must take the model equations (3.3)-(3.4) as our starting point. The resulting set of time-independent difference equations is easily solvable by several methods [19, 33] and the result turns out to be

$$p_j = (1 - \rho)\rho^j \tag{3.7}$$

Henceforth the limiting p_j distribution is alternatively referred to as the p-distribution. Note that the limiting distribution component $p_j = (1 - \rho)\rho^j$ is recognized as a term in the transient solution from equation (3.6). This is not accidental. In taking the limit of equation (3.6) as $t \to \infty$ we should end up with p_j, of course.

If figure 3.5 we have plotted the limiting p_j distribution for $j = 0, \ldots, 10$ and for two different traffic intensities. The steepest curve corresponds to the lowest

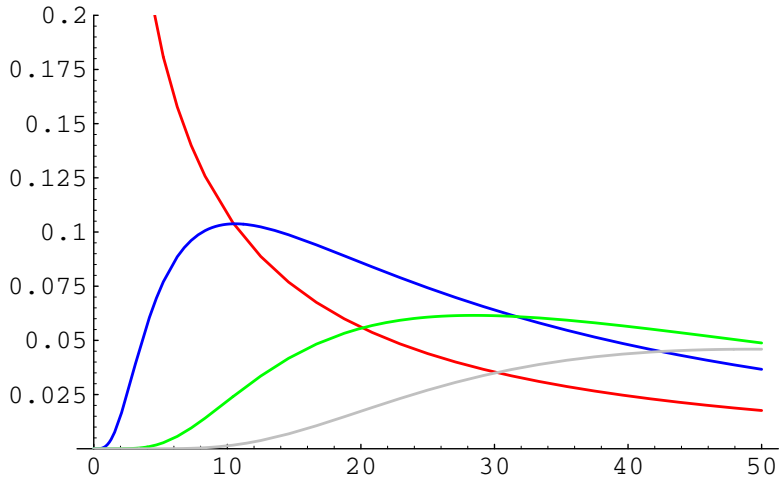

Figure 3.4: *Plot of state probabilities* $p_{0,j}(t)$ *for* $j = 0, 5, 10, 15$ *and* $\rho = \lambda/\mu = 1.0/0.8 = 1.25$. *The crossing curves suggests that statistical equilibrium does not exist in this case.*

traffic intensity. As intuitively expected we see that the probability p_0 that there are no customers in the system under steady-state is greater for the less loaded case. At the other end of the range we see that the probability p_{10} that there are ten customers is greater for the higher loaded case.

From the limiting distribution p_j the steady-state mean number of customers in the system $E[N] = \sum_{j=0}^{\infty} jp_j$ can be found. This yields [19]

$$E[N] = \frac{\rho}{1-\rho} \tag{3.8}$$

In figure 3.6 we have plotted $E[N]$ for traffic intensities ρ in the range $0 - 1$. We see that the steady-state mean number of customers in the system is comfortable for moderate traffic intensities. As the traffic intensity approaches 1 the mean number of customers in the system increases dramatically. The knee-like curve profile shown in the figure is characteristic for many queuing systems.

For the sake of the discussion we have up to this point *assumed* the existence of steady state for the queue size process. Now it is time to consider the condition under which a statistical equilibrium actually exists. Recall that $\rho = \lambda/\mu$ denote the instantaneous traffic intensity. Clearly, if $\rho > 1$ sustained, the queue will grow without bounds. Then arrivals sustain-ably occur more rapidly than departures. In that case it is reasonable to expect that steady-state will not exists. It can be shown [19, 33] that this is actually so. Likewise it can be shown that the condition for existence of a non-degenerate steady-state is

$$\rho = \frac{\lambda}{\mu} < 1 \tag{3.9}$$

for the $M/M/1$ queue. The boundary case $\rho = \lambda/\mu = 1$ corresponds to a degenerate kind of steady-state.

Note that existence of steady-state, and also the corresponding limiting distribution from equation (3.7), depends only on ρ or the *ratio* of λ and μ. As opposed to this

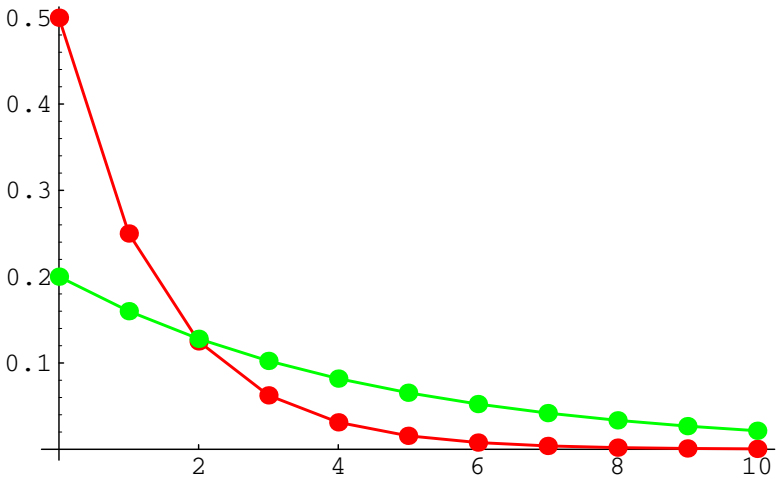

Figure 3.5: *Plot of limiting distribution probabilities p_j for $j = 0, \ldots, 10$. The steepest curve corresponds to $\rho = 0.5$. The other curve corresponds to $\rho = 0.8$.*

the transient solution from equation (3.6) also depends on the *absolute* values of λ and μ. The latter was commented on when we discussed the rate of convergence to steady-state.

3.4 Queue size: a refined view

Consider the following question: assuming that steady-state prevails, what is the probability that an arriving customer finds j customers already in the system? Intuitively the answer is p_j as defined by the p-distribution from equation (3.7). This is initially wrong, however, since we are now asking for the state probability at a *restricted* discrete set of time points, namely at the arrival instants. The p-distribution originating from the stochastic process $N(t)$ reflects the number of customers in the system at a *totally arbitrary* instant in continuous time t.

To properly analyze the situation we must therefore consider a entirely new stochastic process $M_a(k)$ denoting the number of customers in the system immediately *before* the k'th arrival. Note that $M_a(k)$ is a discrete-time stochastic process as opposed to the continuous-time process $N(t)$. Without going into details it can be shown [6, 19] that as $k \to \infty$ a unique steady-state exists for the $M_a(k)$ process under the same circumstances as for the $N(t)$ process. Then $M_a(k) \Rightarrow M_a$ where M_a is the limiting random variable. The distribution of M_a is denoted by π_j and is referred to as the π-distribution. Thus, the probability that an *arriving* customer finds j customers already in the system is given by π_j. In the case of the $M/M/1$ system it fortunately turns out [6, 19] that $\pi_j = p_j$ for all j, but this is in general *not* true for an arbitrary queuing system.

The instants of service completions is another restricted set of points in time at which the queue size is often of special interest. Hence, let the discrete-time process $M_s(k)$ denote the number of customers in the system immediately *after* the departure of the k'th customer. Concerning the existence of steady-state the same applies for this process as for the $M_a(k)$ process. Consequently, as $k \to \infty$ we have that $M_s(k) \Rightarrow M_s$

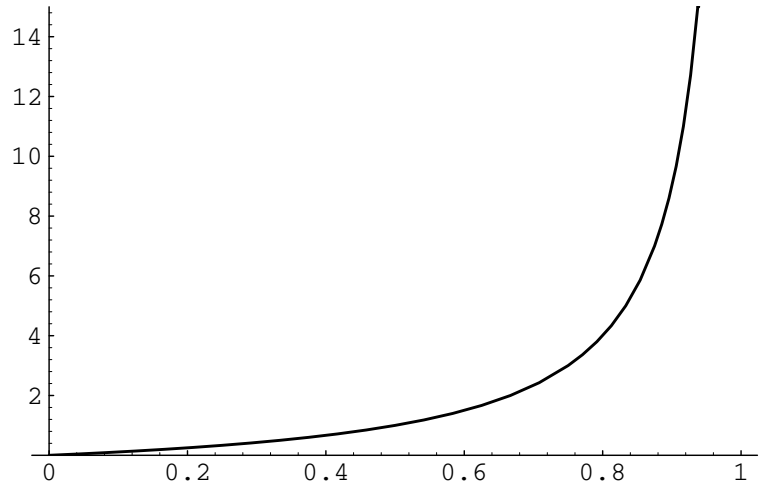

Figure 3.6: *Plot of steady-state mean number of customers in the system $E[N]$ as a function of $0 \leq \rho < 1$.*

where M_s is the limiting stochastic variable. The distribution of M_s is referred to as the q-distribution. I.e. in steady-state q_j represent the probability that a departing customer leaves behind j customer in the system. Again we have that $q_j = p_j$ for all j.

To conclude this section we have in case of the $M/M/1$ queue that the identity

$$\pi_j = p_j = q_j$$

holds for all j. For a more general single-queue system it can be shown [6] that $\pi_j = q_j$ still holds for all j. The p_j distribution may be significantly different, however.

3.5 Waiting time

The waiting time of arriving customers is a very important feature of a queuing system. After all, if we arrive at a queue we are essentially more interested in a probabilistic feeling of how long we have to wait to be served than we are interested in the number of customer in front of us. However, the latter is clearly an indication of the former. By convention the waiting time of a customer often refers to the time spent in the queue *plus* the service time of the customer. This convention is tacitly assumed throughout.

For the $M/M/1$ queue at least three different approaches can be taken in order to describe the probabilistic behavior of waiting time.

- We can model the time spent in the system for each individual customer by a stochastic process $W(k)$. This process will be indexed by a discrete parameter $k = 1, 2, \ldots, \infty$ corresponding to the subsequently arriving customers. Since $W(k)$ signifies waiting time its range will be the continuum of non-negative real numbers. If we are interested in waiting time under steady-state we assume that $W(k) \Rightarrow W$ and then consider the (continuous) distribution of the limiting random variable W.

- If we are only interested in steady-state conditions we can consider an arbitrary customer arrival assuming that statistical equilibrium already prevails. By help

of the π-distribution and the service time distribution we can then derive an expression for the distribution of W.

- If we are only interested in steady-state conditions and if we are satisfied in knowing the mean waiting time $E[W]$ without knowledge of the distribution of W, we can employ the so-called Little's law. We will return to this issue in a moment.

3.6 State-orientation vs. transaction-orientation

Compared to the queue size process $N(t)$ discussed in section 3.3, note that the $W(k)$ process from section 3.5 radically changes the stochastic process view of the system. This is so since $W(k)$ directly accounts for the *individual* customers while $N(t)$ summarize the history of arrivals and departures in a *state variable*. Hence, $W(k)$ is referred to as a customer-oriented stochastic process as opposed to $N(t)$ being a state-oriented stochastic process. In general simulation literature [2] the terms transaction-oriented or process-oriented are used to distinguish it from a state-oriented view.

3.6.1 Little's law

As already pointed out, the number of customers queued in front of an arriving customer clearly gives an indication of the time the arriving customer has to wait in order to be served. Little's law in its classical form [6, 19] make use of this fact, and applied on the $M/M/1$ queue the following relationship concerning steady-state variables can be established

$$E[N] = \lambda\, E[W] \tag{3.10}$$

By now substituting for $E[N]$ from equation (3.8) we arrive at the following expression for mean waiting time in the $M/M/1$ queue under steady-state conditions.

$$E[W] = \frac{1/\mu}{1-\rho} \tag{3.11}$$

As a function of ρ the mean waiting time shows a knee-like profile similar to that in figure 3.6 for the mean number of customers in the system. Note however that $E[W]$ depends on the absolute value of μ in addition to the traffic intensity ρ. The mean number of customers in the system depends only on the latter.

The utility of Little's law is due to the fact that the state-oriented stochastic process $N(t)$ is more tractable than the customer-oriented stochastic process $W(k)$. By focusing on the more tractable process we can by Little's law indirectly say something about the features of the less tractable process. The amount of information about the less tractable process is limited by this indirect approach, however.

Note that the applicability of Little's law go beyond the $M/M/1$ queue. For arbitrary queuing systems there exists generalized results that entail a deep relationship [11, 12, 29, 31] between the state-oriented and transaction-oriented views.

Chapter 4

Beyond the $M/M/1$ model

The analytical tractability of the $M/M/1$ queuing model is primarily due to the following (assumed) features.

- The independence between inter-arrival intervals.
- The memoryless property of the arrival process.
- The independence between service periods.
- The memoryless property of the service process.
- The mutual independence between the arrival and service processes.

For other queuing models the typical case is that one or more of these features are not longer automatically applicable thereby complicating the analysis. Nevertheless, the $M/M/1$ model make an excellent point of departure for two reasons.

First, the analysis of the $M/M/1$ model establishes various *performance measures* for a queuing system. E.g. the steady-state mean waiting time, the steady-state π-distribution and the transient $p_{i,j}(t)$ probabilities are all useful performance measures. There is no reason why the definition and significance of such measures do not carry over to more general cases.

Second, the stochastic process models of the $M/M/1$ queue is a natural starting point from which generalizations and extensions can be made. In the following we consider two such extensions. The first extension models a restricted form of dependency. The second extension deals with lack of memoryless property.

4.1 Dependency

The independency features of the $M/M/1$ model are unrealistic in many queuing situations. E.g. most people are discouraged by long queues and it is reasonable to think that customer inter-arrival intervals generally increase with growing queue size. This simply means that the customer inter-arrival intervals are not independent; a sequence of short intervals tends to generate longer intervals.

One readily available approach to model this kind of dependence arise from the definition of a birth-death Markov process. Recall that the instantaneous arrival rate λ_i is in general state dependent as expressed by (2.3). We may therefore define $\lambda_i =$

$\lambda/(i+1)$ to model discouraged arrivals. In this way we model dependency amongst arrivals via the state variable. Note that the analytical tractability is principally retained by this approach.

As another example consider a "Russian queuing discipline" where customers are encouraged by queue length and tend to arrive more rapidly as the queue grows. For this case we may define $\lambda_i = (i+1)\lambda$ to model the dependency inherent in the arrival process.

In a similar way we may define the instantaneous service rate $\mu_i = (i+1)\mu$ to model a situation where service completions generally occur more rapidly as the queue grow. This may be a reasonable scenario for a stressed clerk at an post-office.

Note that a combination of state-dependent arrivals rates and state-dependent service rates is also feasible within the framework of a birth-death Markov process. Such a combination effectively also models a mutual dependency between the arrival and departure processes.

In the general taxonomy of single-queue systems the notation M^i/M^i/1N refers to a system where dependency are modeled via the state variable as explained here.

4.2 Supplementary variables

Recall that the exponentially distributed inter-arrival intervals and service periods of the $M/M/1$ queue made it possible to model the queue size process $N(t)$ by a Markov process as discussed in section 3.3. This is due to the memoryless property of the exponential probability distribution. If the arrival process and/or the service process fails to be memoryless we get into trouble since $N(t)$ can no longer be modeled as a Markov process. Then mathematically difficulties arise immediately. A conceptually simple method to escape from this situation is to reestablish the Markov property by augmenting the state description with one or more supplementary variables [6, 19].

To illustrate the supplementary variable technique say that the service periods are no longer exponentially distributed but instead distributed according to a general probability density function. Then the $N(t)$ process becomes intractable due to the missing Markov property. At this point we introduce a new random variable $Y(t)$ denoting the *remaining service time* for the customer in service at time t. Then $(N(t), Y(t))$ denotes a vector-valued stochastic process. Note that $N(t)$ is still discrete but $Y(t)$ is a non-negative continuous-valued stochastic variable. The point is that by augmenting the state description by the supplementary variable $Y(t)$ it can be shown that the compound two-dimensional stochastic process $(N(t), Y(t))$ becomes a Markov process. By considering this augmented process the memoryless property is reestablished and this new process is more tractable as opposed to the now non-Markovian process $N(t)$. Based on an analysis of the compound $(N(t), Y(t))$ process certain features of the component process $N(t)$ can then be derived indirectly.

Chapter 5

Stochastic simulation

Queuing models beyond the $M/M/1$ system often turn out to be analytically intractable. Then stochastic discrete-event simulation [1, 2, 8, 23, 28] is a useful tool for gaining insight. A stochastic simulation is characterized by the fact that the same (correct) program produces different (but correct) output data from each run. The random nature of the output data can not be ignored and procedures for making statistical inferences from the output data are of absolute necessity [18].

Output analysis from a stochastic queuing simulation is most often concerned about estimating various quantities of the underlying stochastic process machinery. This inferential problem is almost always casted in terms of a point estimate along with an associated confidence interval [3]. The next subsections outline the prototypical steps taken and also discuss associated problems. It should be emphasized that statistical inference procedures are strongly problem dependent. The effectiveness of any particular inferential method depends on the level of a priori knowledge of the system behavior. Therefore, the establishment of an underlying (at least approximate) stochastic process model is often crucial to any inference methodology.

5.1 Point estimate

Let $(X_1, X_2, \ldots, X_n)$ denote (random) observations gained from a stochastic simulation. At this point we discuss inference methodology generically without any particular meaning attached to the observations. Later we shall see several examples of what may comprise an observation in a queuing simulation. Initially we assume nothing special about the observations. In the most general case they are correlated and have different distributions. In the most trivial case they are i.i.d. For simplicity we consider the X_i observations to be univariates in this section. Generally the observations may be multivariates, however.

Now, let θ denote some quantity of interest subject to estimation. Based on n sample data the objective is to estimate θ by some statistics $\hat{\theta}(n) = h(X_1, X_2, \ldots, X_n)$ referred to as the estimator. Note that the estimator $\hat{\theta}(n)$ being some function of the random observation variables is itself a random variable. As a prototypical example, consider the case when the observations are i.i.d. with mean μ_X and variance σ_X^2. Then

the ordinary sample mean

$$\bar{X}(n) = 1/n \sum_{i=1}^{n} X_i \tag{5.1}$$

is a statistics serving as an estimator of μ_X. Throughout this is referred to as the classical case.

There are three important figures of merit for the goodness or quality of an estimator.

Bias defined by $\text{Bias}[\hat{\theta}(n)] = E[\hat{\theta}(n) - \theta]$ measures the systematic deviation of the estimator from the true value of the estimated quantity. Ideally the estimator should be unbiased so that $E[\hat{\theta}(n)] = \theta$ for all n. E.g. for the classical case, $\bar{X}(n)$ is an example of an unbiased estimator.

Variance of the estimator itself $\text{Var}[\hat{\theta}(n)] = E[(\hat{\theta}(n) - E[\hat{\theta}(n)])^2]$ measures the mean (squared) deviation of the estimator from its expected value. The smaller variance the better, of course. For the classical case we have that $\text{Var}[\bar{X}(n)] = \sigma_X^2/n$. Note that in this case the variance of the estimator is directly related to the variance of the individual observations.

MSE (Mean Square Error) is defined by $\text{MSE}[\hat{\theta}(n)] = E[(\hat{\theta}(n) - \theta)^2] = \text{Bias}[\hat{\theta}(n)]^2 + \text{Var}[\hat{\theta}(n)]^2$ and is an aggregate measure incorporating both bias and variance. A small mean square error is desirable, of course.

The asymptotic features of an estimator are of special interest. With respect to the above figure of merits the quality of an estimator, should improve as n grows. Various laws of large numbers [16] are central in this respect. Particularly, an estimator $\hat{\theta}(n)$ is said to be (weakly) consistent if it converges in probability [3, 24] to the estimated quantity θ as $n \to \infty$. A *strongly* consistent estimator converges almost surely[1] [3, 24] to the estimated quantity. E.g. $\bar{X}(n)$ is a strongly consistent estimator of μ_X for the classical case discussed above.

For a particular finite sequence of observations, i.e. for a particular realization of the random variables $(X_1, X_2, \ldots, X_n)$, the corresponding realization of the statistics $\hat{\theta}(n)$ is called a point estimate of θ. Depending on the quality of the estimator and also the number of observations n we expect the point estimate to be "close" to the true value of the estimated quantity θ. To determine "how close", however, it is essential to assess the precision of the point estimate. This is the purpose of the confidence interval.

5.2 Confidence interval

The natural way to assess the precision of a point estimate is to consider the (random) difference $(\hat{\theta}(n) - \theta)$ reflecting the estimation error. Assuming that the estimator $\hat{\theta}(n)$ is consistent and behaves according to some law of large numbers, we expect this error to become smaller as n grows.

Computing confidence intervals requires knowledge of how the random error $(\hat{\theta}(n) - \theta)$ itself is distributed. Hence, we are seeking second order results about some law of large numbers which by assumption is at play. Such results are generally referred to as central limit theorems [15]. The point is that working with the exact distribution

[1] Also called convergence with probability one.

of $(\hat{\theta}(n) - \theta)$ is in general complicated, if at all possible. Thus, approximations must be employed. Specifically, the condition of asymptotic normality [17] is usually imposed[2]. Then it is either proved or conjectured that

$$\sqrt{n}\,(\hat{\theta}(n) - \theta) \Rightarrow \sigma\, N(0,1) \tag{5.2}$$

holds asymptotically where $N(0,1)$ refers to the standard normal distribution. The σ parameter appearing at the right-hand side is called an asymptotic variance parameter. Note that for the asymptotic normality assumption to be useful the above equation should become approximately valid for fairly large n. The exact definition of σ is generally strongly problem dependent. However, an asymptotic statement of the following form [7] can usually be established

$$\lim_{n\to\infty} n\,\mathrm{Var}[\hat{\theta}(n)] = \sigma^2 \tag{5.3}$$

relating the asymptotic variance parameter σ to the asymptotic variance of the primary estimator $\hat{\theta}(n)$. For the classical case discussed in the previous section equation 5.3 is in fact true for all n and $\sigma = \sigma_X$ reduces to the common variance of the i.i.d. observations. In this case equation 5.2 also reduces to the ordinary central limit theorem [3,21].

Even if the variance parameter σ from equation 5.2 is left unspecified at this point, note that it neatly reflects the asymptotic efficiency of the estimator. E.g. say that $\hat{\theta}_1(n)$ and $\hat{\theta}_2(n)$ are alternative estimators for θ. Now if σ_1 and σ_2 signify the corresponding asymptotic variance parameters and if $\sigma_1 < \sigma_2$, then the former estimator is more efficient than the latter since it leads to a more compressed distribution in equation 5.2 for the same (asymptotic) n.

With equation 5.2 at hand an asymptotic confidence interval for $\hat{\theta}(n)$ is easily given by

$$\left(\hat{\theta}(n) - \delta/2, \hat{\theta}(n) + \delta/2\right) \tag{5.4}$$

where

$$\delta = 2\, z_{1-\alpha/2}\, \frac{\sigma}{\sqrt{n}} \tag{5.5}$$

refers to the width of the confidence interval. Here $0 < \alpha < 1$ and $(1-\alpha)$ specifies the level of confidence. The quantity $z_{1-\alpha/2}$ refers to the $100(1-\alpha/2)$ percentile of the normal distribution. I.e. if $\Phi(z)$ is the (cumulative) distribution function of the standard normal, then $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$.

For fairly large n we expect the confidence interval given by equation 5.4 to be an approximate confidence interval for the estimator $\hat{\theta}(n)$. The interpretation of the confidence interval is as follows. If the width 2δ of the confidence interval is found for a specified confidence level of $(1-\alpha)$ and the simulation experiment were repeated[3] a number of times, the confidence interval would contain the unknown quantity θ in (approximately) $100(1-\alpha)\%$ of the cases and would not in $100\alpha\%$ of the cases.

An small but important point escaped so far is that the general process dependent variance parameter σ is almost always an unknown quantity. To proceed then, we must use an estimator $\hat{\sigma}(n)$ in its place. E.g. for the classical case previously discussed the

[2]Note that asymptotic normality and central limit theorems only applies when the primary estimator is given as some sum of the observations. This is almost always the case, however.

[3]Do not confuse the number of repetitions with n. At each repetition n observations are collected.

unknown variance parameter is consistently estimated by the ordinary sample variance of i.i.d. observations

$$\hat{\sigma}^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}(n))^2 \tag{5.6}$$

In any case, using a consistent estimator $\hat{\sigma}(n)$ in place of σ, the conclusions from equation 5.2 and 5.5 remains.

5.3 Inferential errors

In making statistical inferences as outlined in section 5.1 there are many sources of errors. Here we emphasize the most prominent ones.

- As already pointed out the effectiveness of any particular inferential method depends on the level of a priori knowledge of the system behavior. Hence, at the outset it is important to have a good model of the underlying stochastic process.
- It is important to use a high-quality primary estimator $\hat{\theta}(n)$. Ideally, an estimator should be unbiased, consistent and having a small variance for each n. The unbiasness and small variance requirements translate into a small mean square error. Especially, the asymptotic variance parameter σ should be small so as to give an efficient estimator. Any deviation from these requirements leads to point estimates of lower precision.
- In imposing an asymptotic normality assumption on the distribution of $(\hat{\theta}(n)-\theta)$ recall than an approximation is really made for finite n.
- If the problem dependent asymptotic variance parameter σ is unknown and must itself be estimated, care must taken. First, it is again important to use a high-quality estimator $\hat{\sigma}(n)$. Next, recall that the resulting confidence interval specification is really a twicely approximated confidence interval.

To conclude it is important to be aware that a confidence interval specification potentially suffers from several errors and should be considered only as an *approximate* statement of the precision of the actual inference procedure.

Chapter 6

Steady-state simulation

To illustrate some of the questions associated with a simulation approach in queuing theory, we use the $G/G/1$ queuing model as an framework. The G designation refers to general inter-arrival time and service time distributions, respectively. For a $G/G/1$ queue it is implicitly assumed that there is no dependency between the arrival and the service processes. Likewise the inter-arrival times and service times are mutually independent. The $M/M/1$ queue is a special case of the $G/G/1$ queue, of course.

In the following we use the same notational conventions as we did in discussing the $M/M/1$ queue. Specifically, $N(t)$ denotes the queue size process. Assuming the existence of statistical equilibrium we focus on inference methodology associated with the steady-state behavior as $N(t) \Rightarrow N$. For an excellent general overview of this subject the reader is referred to [27]. Specifically we discuss four different inferential methods referred to as

- Replicated runs
- Equally spaced observations
- Asynchronous observations
- Regenerative cycles

As opposed to the first method, the latter three methods are based on making inferences from a single simulation run only.

Note that making inferences about the transient behavior from simulations are methodologically simple since we can always conduct a series of n replicated finite-simulations [2, 10]. Classical estimation procedures can then always be employed due to the assumed independency of the replicated runs.

The reader may argue that inference about steady-state features based on simulations will always fail since any simulation must be stopped in finite time while steady-state is asymptotically defined. Strictly speaking this is true, of course. However, the notion that a simulation eventually reaches steady-state after an initial transient warm-up period can be regarded as a convenient fiction that is at least approximately true.

Finally, note that we will say nothing about implementation issues associated with the various inferential methods in this document. This not to neglect the importance of the subject but rather as a result of limited scope. The interested reader is refered to [27] and references therein.

6.1 Replicated runs: Ensemble averages

Consider the following definition of the expectation of a functional f of the limiting random variable N.

$$e_f = E[f(N)] = \sum_{j=0}^{\infty} p_j f(j) \tag{6.1}$$

Two examples are illustrative here. If $f = I$ where I denote the identity function, equation 6.1 reduces to the ordinary mean number of customers in the system under steady-state. If $f = I_j$ where I_j signifies a (discrete) indicator function, equation 6.1 reduces to p_j being the probability that there are j customers in the system at an arbitrary point in time when steady-state prevails. Note that equation 6.1 in any case reflects some property of N that can be interpreted as an average measure over the complete sample space comprising the ensemble of all possible trajectories the process may take. Hence, e_f is referred to as an ensemble average.

In a steady-state simulation we are essentially interested in estimating various ensemble averages e_f. The obvious way to proceed with estimation is to perform n replicated independent simulation runs. For each run i one observation

$$X_i = f(N(t_i)) \tag{6.2}$$

of the quantity of interest is sampled at time t_i when steady-state is assumed to prevail. By assumption then, the observations X_i are independent all having the same distribution, namely that of $f(N)$.

With respect to the inference procedure outlined in section 5.1, the classical case now applies due to the independency. I.e. if $\hat{e}_f(n)$ denotes an estimator for the ensemble average we are seeking, we employ the ordinary sample mean from equation 5.1

$$\hat{e}_f(n) = \bar{X}(n) \tag{6.3}$$

The corresponding asymptotic variance parameter, now denoted σ_a, is simply defined by the common variance of the individual observations

$$\sigma_a^2 = \text{Var}[X_i] \tag{6.4}$$

As suggested by the discussion in section 5.2, it is usually difficult to find an explicit expression for σ_a. Hence, an estimator $\hat{\sigma_a}(n)$ must be used in its place. Due to the independent observations, $\hat{\sigma_a}(n)$ is naturally given by the the ordinary sample variance from equation 5.6.

6.2 Warm-up

There is a problem associated with the inference procedure described in section 6.1. This is due to the warm-up phase or initial transient period. Ideally this period should be discarded for each replicated run in the sense that the simulator should scan past it before taking the observation X_i at time t_i when steady-state supposedly prevails. However, if the rate of convergence to steady-state is slow, it may take prohibitively long time to achieve sufficiently many replications. Few replications usually leads to a large variance estimate and a correspondingly wide confidence interval. It is reasonable

then to try to collect observations prematurely, i.e. before the "real" onset of statistical equilibrium, so as to increase the number of observations and reduce the variance estimate within the same time budget. Unfortunately, the sample mean estimator then becomes biased due to influence from the initial condition. The mean square error of the estimator includes both bias and variance terms, so in either case the replicated run approach suffers from having an estimator with a significant mean square error.

It is interesting to note that a steady-state hypothesis simplified the analytical solution of the $M/M/1$ queue considered in section 3.3.3. In this section we have seen that a similar steady-state hypothesis complicates the analysis of simulation results due to the inevitable influence of the warm-up phase.

6.3 Single run: Time averages

Performing replicated runs as explained in section 6.1 is not the only way to make inferences. Alternative methods less sensitive to warm-up effects exist. In this section we discuss one such method. The new method involves long-run time averages of the process $N(t)$, generally defined by

$$r_f = \lim_{t\to\infty} \frac{1}{t} \int_0^t f(N(s))\, ds \tag{6.5}$$

where r_f should be interpreted as an random variable at this point. Two examples are readily available by considering the same two functions $f = I$ and $f = I_j$ as in section 6.1. If $f = I$ equation 6.5 reduces to the long-run time-averaged number of customers in the system. For $f = I_j$ equation 6.5 corresponds to the long-run fraction of time there is j customer in the system.

Due to the assumed existence of a steady-state for the $N(t)$ process, we now have the following important result

$$r_f \to e_f = E[f(N)] \qquad \text{a.s.} \tag{6.6}$$

Equation 6.6 states that various steady-state ensemble averages e_f of the process can be replaced by corresponding long-run time averages r_f. When this relation holds the process is said to be ergodic [8, 14]. Ergodicity is closely related to (asymptotically) stationary processes and essentially assures that a law of large numbers applies [16,32]. The utility of equation 6.6 is obvious. By estimating the long-run time average r_f we essentially estimate the corresponding ensemble average e_f being the real objective of the simulation.

6.4 Equally spaced observations

The natural way to estimate a long-run time average r_f is to sample the $N(t)$ process regularly. Hence, let

$$X_i = f(N(i\Delta)) \tag{6.7}$$

denote subsequent observations taken from a single simulation run of $N(t)$. Here Δ signifies the fixed spacing between successive observations. Equally spaced observations like this is also called a time series.

As before let $\hat{e}_f(n)$ denote an estimator for the ensemble average e_f we ultimately are seeking. By way of equation 6.6 we now set $\hat{e}_f(n) = \hat{r}_f(n)$ where the right-hand refers to an estimator for the corresponding long-run time average r_f. In turn we set $\hat{r}_f(n) = \bar{X}(n)$ being the usual sample mean. In sum we have that

$$\hat{e}_f(n) = \bar{X}(n) \tag{6.8}$$

gives a strongly consistent estimator for e_f. Unfortunately it is also a biased estimator due to the influence from the initial transient period. However, the biasness becomes less pronounced with increasing n, and asymptotically the estimator is unbiased. The obvious way to reduce the bias effect is to discard the initial observations from the warm-up phase. But this leads to fewer observations and possibly a larger estimator variance which in turn gives a wider confidence interval. Qualitatively we are in the same situation as discussed in section 6.2 for the replicated run approach. This time, however, we only have to deal with a single initial transient period and the problems are significantly reduced.

For the sake of the remaining discussion we make a stationarity assumption. I.e. we assume that the initial transient period is discarded so that the remaining observations X_i can be considered to be taken from a strictly stationary stochastic process [8]. By assumption then, the observations X_i all have the same distribution namely that of $f(N)$. As already stated the normal sample mean is a suitable estimator for the ensemble average we are seeking. Assuming stationarity the estimator is also unbiased. Note, however, that the observations are now in general correlated or dependent since they are taken from the same simulation run. Due to this dependency the classical inference procedure used for the replicated run approach fails.

Nevertheless, under certain conditions the inference procedure outlined in section 5.1 still applies. Hence, if the asymptotic variance parameter is now denoted by σ_b, we have [2, 8]

$$\sigma_b^2 = \text{Var}[X_i] + 2\sum_{k=1}^{\infty} \text{Cov}[X_i, X_{i+k}] \tag{6.9}$$

Note here that this equation holds for all i due to the stationarity assumption. This is trivial for the $\text{Var}[X_i]$ term. Regarding the infinite sum of covariances, keep in mind that a strictly stationary process is also wide-sense or covariance stationary [8].

Compared to the corresponding equation 6.4 for the replicated-run approach it should come as no surprise that the definition of σ_a is simpler than σ_b. This is due to the fact that the observations are correlated in the latter case while being independent in the former case. Anyway, as previously pointed out it is still generally hard to arrive at a closed-form expression for σ_b, hence a corresponding estimator $\hat{\sigma}_b(n)$ must be employed. Several standard approaches exists, and two commonly used techniques are batched-means and spectral methods [2, 7, 8, 27]. The details are beyond the scope of this document, though.

Note that the inference method outlined in this section essentially is an application of inferential procedures associated with time series from wide-sense stationary stochastic processes [8]. As already indicated specific conditions must be satisfied by such processes for the inferential procedures to hold. As an intuitive rule of thumb they do hold if the correlation between two observations X_{i_1} and X_{i_2} diminishes with the distance $|i_1 - i_2|$ between them. In queuing simulations this is typically the case.

In discussing inferential procedures based on long-run time averages we have implicitly assumed that the observations X_i comprise a time series sampled at regularly spaced intervals Δ. More can be said about this. E.g. observations defined by

$X_i = 1/\Delta \int_{i\Delta}^{(i+1)\Delta} f(N(t))\,dt$ may equally well be used. The length of the time interval Δ is also of significance. Larger spacing generally reduce serial correlation but also reduce the number of observations, and by that the estimated variance parameter, within the same time budget.

6.5 Embedded event process

Classically, equally spaced observations of $N(t)$ collected from a single simulation run has been used in output analysis of queuing simulations. However, another way of collecting observations from a single simulation run with associated inference procedures do exist. This new approach arise by changing the stochastic process view as discussed next.

Figure 6.1 shows a typical sample path of the queue size process $N(t)$ of a $G/G/1$ queue. The piecewise continuous step-like trajectory is characteristic for state-oriented

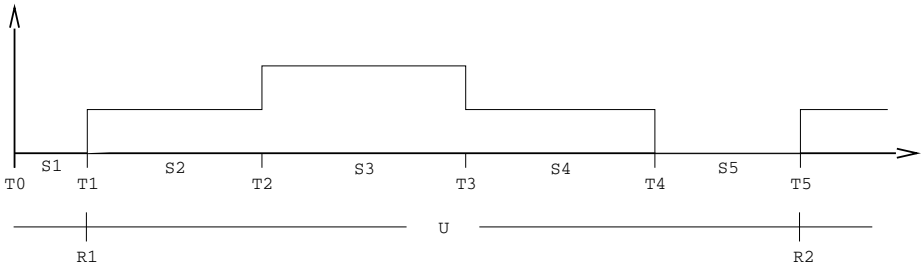

Figure 6.1: A typical sample path of the queue size process $N(t)$ in a $G/G/1$ queuing simulation. Sojourn times are denoted by S_k and event times by T_k. Regeneration points are denoted by R_j and associated regeneration cycles by U_j.

queuing simulations. Now, let S_k denote the subsequent random sojourn times for the process as illustrated. Further, define

$$T_k = \sum_{j=1}^{k} S_j$$

simply being the random time at which the k'th state change takes place. Note that each state change corresponds to the occurrence of an event, either an arrival or a service completion. Consequently, the random time points T_k are refered to as event times.

The sequence of event times effectively define an embedded discrete-time stochastic process. Hence, let $M_k = N(T_k)$ take the value of $N(t)$ at these selected time points. The precise relation between the two processes is

$$N(t) = \sum_{k=0}^{\infty} M_k I_{[T_k, T_{k+1}]}(t)$$

where the indicator function $I_A(t)$ is 1 or 0 depending on whether or not $t \in A$.

Note that the discrete-time compound process formed by (M_k, S_k) is only a reformulation of the same phenomenon described by the queue size process $N(t)$. In this

sense the two process descriptions are really equivalent. Especially, under the same conditions as the $N(t)$ process has a steady-state, a steady-state $(M_k, S_k) \Rightarrow (M, S)$ will exist for the compound process as $k \to \infty$. Another evidence of the fact that the two processes are inherently equivalent is clearly displayed by a relation between long-run time averages for the two processes.

The concept of a long-run time average r_f for the process $N(t)$ was defined in section 6.3. For the two-component discrete-time process (M_k, S_k) this definition do not carry over directly. Instead we consider the following two kinds of (random) long-run time averages [2, 9, 10]

$$q_f = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(M_{k-1}) S_k \tag{6.10}$$

$$s = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} S_k \tag{6.11}$$

now for discrete-time but for the same function f. Here q_f is the analog to r_f but the average is now computed relative to the the number of state transitions instead of total elapsed time. The long-run average sojourn time is given by s. Note that by considering the ratio q_f/s we conclude that this ratio and r_f essentially reflects the same thing in the long-run.

Analogous to equation 6.6 the assumed existence of a steady-state for the compound process ensures that this process is also ergodic, hence

$$q_f \to E[f(M)S] \quad \text{a.s.} \tag{6.12}$$

$$s \to E[S] \quad \text{a.s.} \tag{6.13}$$

holds. I.e. the long-run averages converge almost surely to the corresponding steady-state ensemble averages. Note here that $E[S]$ is the mean sojourn time in steady-state.

Equipped with these definition the following result [9, 10]

$$r_f \to \frac{E[f(M)S]}{E[S]} \quad \text{a.s.} \tag{6.14}$$

shows that the $N(t)$ and (M_k, S_k) processes are really long-run equivalent since r_f and q_f/s both converge almost surely to the same ratio.

By now combining this statement with the result from equation 6.6 we arrive at the important conclusion that

$$e_f = \frac{E[f(M)S]}{E[S]} \tag{6.15}$$

This result states that various steady-state ensemble averages e_f of the process $N(t)$ can be replaced by the ratio of two associated steady-state ensemble averages averages on the equivalent process (M_k, S_k). The ensemble averages for the latter process can in turn be computed from corresponding long-run averages by way of equation 6.12 and 6.13. The utility of this result is obvious, then. By estimating long-run averages q_f and s we effectively arrive at an estimate of the corresponding ensemble average e_f really being the objective of the simulation.

6.6 Asynchronous observations

To estimate the long-run averages q_f and s we must take observations from the compound (M_k, S_k) process. Since this is a discrete-time process the observations are naturally defined by the subsequent readings of the process. However, rather than using the readings directly, we define the following transformed pairs

$$(X_i, S_i) = (f(M_i)S_i, S_i) \tag{6.16}$$

and refer to them as the observations of the process. The reason for performing this transformation is due to equation 6.10 since the X_i's now match the summands of the long-run measure q_f we are interested in.

Relative to the $N(t)$ process note that the observations are now taken asynchronously in that the observations are randomly spaced by the subsequent sojourn times. Therefore this is referred to as asynchronous observations [2, 9].

As previously let $\hat{e}_f(n)$ signify an estimator for the ensemble average e_f we ultimately are seeking. Due to equation 6.15 combined with equations 6.12 and 6.13 we now define

$$\hat{e}_f(n) = \frac{\bar{X}(n)}{\bar{S}(n)} \tag{6.17}$$

being a strongly consistent estimator for e_f. Keeping in mind the definition of an observation pair from equation 6.16, $\bar{X}(n)$ and $\bar{S}(n)$ denote the usual sample mean estimators corresponding to the long-run averages q_f and s, respectively. They are both strongly consistent estimators. Note, however, that in the same way as discussed in section 6.4 these estimators are also biased due to initial warm-up effects. Likewise, the same tradeoffs applies regarding deletion of initial observations in order to reduce biasness.

For the sake of the discussion we again assume that the initial observations are discarded so that we can impose a stationarity condition on the remaining observations. By assumption then, the X_i's are identically distributed. The same applies for the S_i observations. In addition, the observations are in general correlated since they are taken from the same simulation run.

Assuming stationarity the estimators $\bar{X}(n)$ and $\bar{S}(n)$ now become unbiased. Despite this the primary estimator from equation 6.17 is still biased. This is so since the expectation of a ratio is in general not equal to the ratio of the expectations. Nevertheless, the estimator is consistent and we continue to use it.

Taking the correlated observations into account [9] shows that an inferential procedure similar to that described in section 5.1 applies. This time, however, the width of the confidence interval is given by

$$\delta_c = 2z_{1-\alpha/2}\frac{\sigma_c}{E[S]\sqrt{n}} \tag{6.18}$$

Compared to equation 5.5 note that the steady-state mean sojourn time $E[S]$ now appears in the interval specification. The asymptotic variance parameter, denoted σ_c in this case, may be expressed as [9]

$$\sigma_c^2 = c_1 - e_f(c_2 + c_3)e_f^2 c_4 \tag{6.19}$$

where

$$
\begin{aligned}
c_1 &= \text{Var}[X_i] + 2\sum_{k=1}^{\infty} \text{Cov}[X_i, X_{i+k}] \\
c_2 &= \text{Cov}[X_i, S_i] + 2\sum_{k=1}^{\infty} \text{Cov}[X_i, S_{i+k}] \\
c_3 &= \text{Cov}[S_i, X_i] + 2\sum_{k=1}^{\infty} \text{Cov}[S_i, X_{i+k}] \\
c_4 &= \text{Var}[S_i] + 2\sum_{k=1}^{\infty} \text{Cov}[S_i, S_{i+k}]
\end{aligned}
$$

Again, due to the stationarity assumption, these equations holds for all i. A corresponding estimator is given by

$$\hat{\sigma}_c(n) = \hat{c}_1(n) - \hat{e}_f(n)(\hat{c}_2(n) + \hat{c}_3(n))\left(\hat{e}_f(n)\right)^2 \hat{c}_4(n) \tag{6.20}$$

where $\hat{e}_f(n)$ is given by equation 6.17 and each of the terms $\hat{c}_1(n)$, $\hat{c}_2(n)$, $\hat{c}_3(n)$, $\hat{c}_4(n)$ can be consistently estimated by standard techniques like bathed-means and spectral methods [2, 7, 8, 27].

6.7 Regenerative method

A key part of the inference procedure outlined in section 6.6 was estimation of the long-run averages q_f and s by the estimators $\bar{X}(n)$ and $\bar{S}(n)$, respectively. In this section we consider yet another inferential method in which these estimators are replaced by a new pair. This is called the regenerative approach [2,27,30] and relies on identification of an embedded renewal process [5, 33] in the compound (M_k, S_k) process.

The *G*/*G*/1 queue-size process is regenerative and the regeneration points are associated with arrivals at an otherwise empty system. This is illustrated in figure 6.1 by the random time points R_j. At these instants in time the process becomes memoryless and restarts probabilistically. The random length of an regeneration cycle is denoted by $U_j = R_{j+1} - R_j$

Note that a regeneration point R_j is always associated with an event time T_k. In the following let $k(R_j)$ denote the index k of the event time T_k corresponding to the regeneration time R_j. E.e. with respect to figure 6.1 we have $k(R_1) = 1$ and $k(R_2) = 5$. With this at hand we may express the length of the j'th regeneration cycle by

$$U_j = \sum_{k=k(R_j)}^{k(R_{j+1})} S_k$$

Correspondingly we define

$$Y_j = \sum_{k=k(R_j)}^{k(R_{j+1})} X_k$$

as the sum of the X_k observations over the same regeneration cycle. For each regeneration cycle note that these variables are easily computed from asynchronous observations (X_i, S_i) collected during the actual cycle.

Due to the regenerative property we have that the U_j's are i.i.d. random variables. The same applies for the Y_j variables. In addition, if we define

$$\tau_j = k(R_{j+1}) - k(R_j$$

as the number of state transitions within a regenerative cycle, we have that the τ_j variables are also i.i.d. Note however, that for any j we have that U_j, Y_j and τ_j are dependent variables.

The regenerative property permits us to re-express the long-run limits for q_f and s from equations 6.12 and 6.13, respectively. Specifically, it can be shown [10, 30] that the following important equalities holds

$$E[f(M)S] = \frac{E[Y]}{E[\tau]} \tag{6.21}$$

$$E[S] = \frac{E[U]}{E[\tau]} \tag{6.22}$$

$$\tag{6.23}$$

where $E[Y]$ refers to the common mean of the Y_i variables, $E[U]$ signify the common mean of the U_i variables and $E[\tau]$ denote the common mean of the τ_j variables. A substitution into equation 6.15 in turn yields

$$e_f = \frac{E[Y]}{E[U]} \tag{6.24}$$

which essentially says that the ensemble average e_f we ultimately seeks is determined by the behavior of the process within a single regeneration cycle. This equation also lays the foundation for the regenerative inference methodology. A strongly consistent estimator for e_f is now naturally given by

$$\hat{e}_f(n) = \frac{\bar{Y}(n)}{\bar{U}(n)} \tag{6.25}$$

where $\bar{Y}(n)$ and $\bar{U}(n)$ denote the obvious sample mean estimators. Note that these are both strongly consistent and unbiased due to the independency amongst regeneration cycles.

Utilizing the regenerative structure of the process we now have arrived at equation 6.25 as an estimator for e_f instead of the original equation 6.17. Effectively, we have replaced the original estimators $\bar{X}(n)$ and $\bar{S}(n)$ with a new pair $\bar{Y}(n)$ and $\bar{U}(n)$. As commented on when discussing the former pair, they are both biased due to initialization effects. As opposed to this the latter pair is unbiased. Hence they do not suffer from initialization problems and deletion of initial observations is not an issue. This is the advantageous feature of regenerative simulation methodology. Note, however, that the primary estimator $\hat{e}_f(n)$ itself is still biased for the same reasons as pointed out in section 6.6.

For the regenerative estimator from equation 6.25 the width of an associated asymptotic confidence interval is given by [2, 9, 30]

$$\delta_d = 2z_{1-\alpha/2}\frac{\sigma_d}{E[U]\sqrt{n}} \tag{6.26}$$

Compared to the interval specification in equation 6.18 note that the mean regeneration cycle length $E[U]$ replaces $E[S]$. For the regenerative method the asymptotic variance parameter, now denoted σ_d^2, is defined by [2, 9, 30]

$$\sigma_d^2 = \text{Var}[Y_j] - 2e_f\text{Cov}[Y_j, U_j] + e_f^2\text{Var}[U_j] \tag{6.27}$$

Trivially, this equation holds for all j. Compared to equation 6.19 note that the definition of σ_d is considerably simpler than σ_c. This is due to the identification of regeneration cycles which in turn gives independency. A corresponding estimator $\hat{\sigma}_d(n)$ is straightforward. An estimator for e_f is given by equation 6.24. The ordinary sample variances are used as estimators for $\text{Var}[Y_j]$ and $\text{Var}[U_j]$ and the ordinary sample covariance [21] is used as an estimator for $\text{Cov}[Y_j, U_j]$. In [30] elaborate numerical techniques are given for computing these estimates.

6.8 Waiting time

Up to this point we have only considered steady-state inference procedures associated with the queue size process as $N(t) \Rightarrow N$. However, as discussed in section 3.5 for the $M/M/1$ queue the behavior of steady-state waiting time as $W(k) \Rightarrow W$ is also an important feature of the $G/G/1$ queue. Recall that $W(k)$ is a discrete-indexed processes denoting the waiting time for the k'th customer.

The natural way for making inferences about steady-state waiting time, assuming its existence, is to perform a discrete-event simulation of the process $W(k)$. As discussed in section 3.5 note here that $W(k)$ is customer-oriented as opposed to $N(t)$ being state-oriented. Hence, the inner workings of a simulation program corresponding to $W(k)$ is rather different from a program corresponding to $N(t)$. Parallel to equation 6.1 we define the following steady-state ensemble average as the objective of the simulation

$$w_f = E[f(W)] = \int_0^\infty P[s \leq W \leq s + ds]f(s)\,ds \tag{6.28}$$

Keep in mind that W is a continuous-valued random variable as opposed to N being discrete-valued.

Estimation of w_f may now proceed in one of two ways. First, a replicated-run approach completely analogous to that described in section 6.1 can be used if we define an observation by

$$X_i = f(W(k_i))$$

Here k_i is assumed to be sufficiently large for steady-state to prevails. Compared to the corresponding definition 6.2 recall that k_i is discrete as opposed to t_i. Alternatively, we may use a single-run approach parallel to that described in sections 6.3 and 6.4. In this case an observation is simply defined by

$$X_i = f(W(i))$$

Clearly, both approaches rely on a stationarity assumption and suffer from warm-up effects in the same ways as previously discussed.

6.9 Indirect estimation

Little's law was introduced in section 3.6.1 in the context of an $M/M/1$ queue. As mentioned the law is very widely applicable and in particular its holds for the $G/G/1$ queue. Properly interpreted the law entails a close relation among the processes $W(n)$ and $N(t)$. This is discussed in [11] and in essence it leads to an alternative approach for making inferences about steady-state mean waiting time. An early reference on this subject is [22]. More recent references are [11–13].

Note first that by equation 6.1 and 6.28, Little's law from equation 3.10 can be written

$$w_I = \lambda e_I$$

where $f = I$ is taken to be the identity function. Now the idea is to estimate w_I by way of e_I. I.e. an estimator $\hat{w}_I(n)$ is constructed by letting

$$\hat{w}_I(n) = \lambda \hat{e}_I(n)$$

In words we can make inferences about steady-state mean waiting time from a simulation of the state-oriented queue size process $N(t)$. Concerning $\hat{e}_I(n)$ any of the consistent estimators given by equations 6.3, 6.8, 6.17 or 6.25 may be used.

For the $G/G/1$ queue λ is a parameter of the model. Hence it is an a priori known quantity. However, to emphasize the fact that λ need not be known for the outlined inference procedure to work, we will somewhat artificially threat it as an unknown quantity. Consequently, we will need an estimator $\hat{\lambda}(n)$ in its place. As a stand-alone issue several approaches exist for finding such an estimator [11]. However, depending on the way in which the estimator $\hat{e}_I(n)$ is constructed, a corresponding natural estimator $\hat{\lambda}(n)$ can often be identified. E.g. if we for the sake of the discussion assume that equation 6.8 and equally spaced observations are used for estimating $\hat{e}_I(n)$, then a particularly suitable estimator for λ is due to the following result where $N_a(t)$ signifies the (random) number of arrivals up to time t.

$$\frac{N_a(t)}{t} \to \lambda \qquad \text{a.s.}$$

Hence, the long-run average arrival rate converges almost surely to λ reflecting the instantaneous arrival rate. This is an intuitively reasonable result. By now defining

$$A_i = N_a(i\Delta) - N_a((i-1)\Delta)$$

as a second set of observations in addition to X_i defined in equation 6.7, we arrive at the following strongly consistent estimator

$$\hat{\lambda}(n) = \frac{\bar{A}(n)}{\Delta}$$

where $\bar{A}(n)$ refers to the ordinary sample mean of the A_i observations. Once again we assume that the warm-up period is discarded so that we can impose a stationarity assumption.

To sum up at this point we have that by setting

$$\hat{w}_I(n) = \hat{\lambda}(n)\hat{e}_I(n) \tag{6.29}$$

the steady-state mean waiting time can be consistently estimated by a pair of suitable estimators $\hat{\lambda}(n)$ and $\hat{e}_I(n)$. It remains, however, to assess the precision of the resulting point estimate. Fortunately it turns out that an inference procedure similar to that described in section 5.1 applies. Specifically, we have that the width of the confidence is given by

$$\delta_e = 2z_{1-\alpha/2}\frac{\sigma_e}{\lambda\sqrt{n}} \tag{6.30}$$

The asymptotic variance parameter, now denoted σ_e, can be expressed as [11, 13]

$$\sigma_e^2 = (c1 - w_I(c_2 + c_3)w_I^2 c_4) \tag{6.31}$$

where

$$\begin{aligned}
c_1 &= \text{Var}[X_i] + 2\sum_{k=1}^{\infty}\text{Cov}[X_i, X_{i+k}] \\
c_2 &= \text{Cov}[X_i, A_i] + 2\sum_{k=1}^{\infty}\text{Cov}[X_i, A_{i+k}] \\
c_2 &= \text{Cov}[A_i, X_i] + 2\sum_{k=1}^{\infty}\text{Cov}[A_i, X_{i+k}] \\
c_4 &= \text{Var}[A_i] + 2\sum_{k=1}^{\infty}\text{Cov}[A_i, A_{i+k}]
\end{aligned}$$

Note that the form of equation 6.30 and 6.31 is identical to the pair of equations 6.18 and 6.19. This becomes apparent by considering the underlying mathematics for the two cases [9, 13]. Accordingly, parallel to equation 6.20 we have that σ_e is estimated by

$$\hat{\sigma}_e(n) = \hat{c}_1(n) - \hat{w}_I(n)(\hat{c}_2(n) + \hat{c}_3(n))\left(\hat{w}_I(n)\right)^2\hat{c}_4(n)$$

where $\hat{w}_I(n)$ is given by equation 6.29 and each of the terms $\hat{c}_1(n)$, $\hat{c}_2(n)$, $\hat{c}_3(n)$, $\hat{c}_4(n)$ again can be consistently estimated by standard techniques.

Regarding estimation by way of Little's law some final comments applies. First, it can also be used in the opposite way. I.e. inferences about mean queue size e_I can be made from a simulation of the customer-oriented waiting time process $W(k)$. Second, if λ is an a priori known quantity it turns out that it is asymptotically most efficient to make inferences from a simulation of the $W(k)$ process. However, if λ is unknown and must itself be estimated both inferential approaches have the same asymptotic efficiency. For more information on the issue of Little's law and relative asymptotic efficiency, the reader is referred to [13].

Chapter 7

Run-length and efficiency

7.1 Sequential procedures

With the exception of regenerative simulation methodology, all inference procedures previously discussed ideally assume the sequence of observations to be stationary. The correctness, i.e. the bias, of these methods is sensitive to observations collected during the non-stationary warm-up period. However, in an attempt to reduce biasness the problem to discard or not to discard initial observations is a perennial dilemma of stochastic simulation practice. This is so since deletion also leads to fewer observations within the same simulation-time budget and by that possibly a larger inferential error as discussed in section 6.2. Hence, there is a tradeoff but usually the question of when to start sampling is an important issue.

Normally, it is impossible to determine a priori how many observations should be discarded. The only way is to monitor the running process and start sampling when a set of specified conditions is first met. This is refered to as a sequential procedure. Depending on the nature of the particular inference method employed various solutions to this problem exist. The interested reader is referred to [2, 27] and the references therein.

At the other end it is also difficult to decide a priori how many observations n to take before terminating a simulation. Again sequential procedures which automatically control the length of a simulation experiment must be used, and there are two important issues. First, with respect to the discussion in section 5.2 we must ensure that n is large enough for the normality approximation to hold. There exists standard techniques for testing normality [2]. Next, it is desirable to continue the simulation until a pre-specified accuracy of the point estimators is obtained. Naturally this is formulated as a requirement on the relative width $\epsilon(n)$ of the resulting confidence interval

$$\epsilon(n) = \frac{\delta}{\hat{\theta}(n)}$$

The definition of relative width $\epsilon(n)$ should not be confused with the confidence level $(1-\alpha)$. The latter says something about the precision of a resulting point estimate for *arbitrary*[1] n. As n grows, however, the relative precision $\epsilon(n)$ of the confidence interval improves but the level of confidence remains. Hence, the simulation experiment

[1] Assuming the validity of the normality approximation.

is stoped at the first checkpoint for which $\epsilon(n) \leq \epsilon_{\text{max}}$ where $0 < \epsilon_{\text{max}} < 1$ is the required relative precision limit of the results at the specified confidence level.

Usually the question of relative interval precision is the dominating one in sequential termination procedures. Hence normality is often simply assumed and an explicit test is omitted. In addition such procedures are often also governed by specifying a maximal simulation-time budget. If the required relative precision ϵ_{max} can not be met within the budget, the simulation is stopped unconditionally. For more information on sequential procedures for simulation termination the reader is referred to [2, 27].

7.2 Efficiency

We have described four different methods for making statistical inferences from a queuing simulation. Along we have also said a few words about the pros and cons of the methods. The natural question arises however, which method is generally most efficient? Note that efficiency is naturally defined as the reciprocal product of final relative precision, denoted ϵ, and simulation-time t

$$\text{Efficiency} = \frac{1}{t\,\epsilon} \tag{7.1}$$

Thus, a high efficiency value is desirable. E.g. for a given simulation-time budget the most efficient method is the one achieving the best precision[2] within the limits of the budget. For a fixed precision requirement the most efficient method is the one reaching this level of precision most quickly.

It seems to be no definite answer to the efficiency question and no method can be rated as universally best. Depending on the situation one method may be more efficient than the other. It is therefore very useful, if not to say necessary, to develop a notion of what makes influence of efficiency and how the various methods differ in this respect. The following subsections point at two major factors refered to as asymptotic efficiency and observation hardness.

7.2.1 Asymptotic efficiency

The concept of asymptotic efficiency was defined in section 5.2 and refers to the magnitude of the asymptotic variance parameter σ. In essence we have that a smaller σ gives a better relative precision $\epsilon(n)$ for the same (large) n. By way of equation 7.1 this in turn leads to a higher overall efficiency.

However, σ is usually an unknown quantity and it is difficult to compare the relative performance of inference procedures in this respect. To illustrate that the methods *do* vary, consider the mean queue length of the $M/M/1$ system as an example. For this exceptional case an explicit expressions for σ can actually be found. Particularly, for the replicated-run approach we have [19] that

$$\sigma^2_{a_*} = \rho/(1-\rho)^2 \tag{7.2}$$

corresponding to equation 6.4. In the case of asynchronous observations taken from a single simulation run we have [13]

$$\sigma^2_{b_*} = 2\rho^3(1+4\rho-4\rho^2+\rho^3)/(1-\rho)^4 \tag{7.3}$$

[2] I.e. smallest ϵ

corresponding to equation 6.9. For large ρ it is easily seen that $\sigma^2_{a_*} < \sigma^2_{b_*}$. Hence, in this range the replicated run approach is asymptotically more efficient than the single-run approach. The reader is warned at making a rushed conclusion at this point. Keep in mind that other factors also must be taken into account in order to make a statement of the total relative efficiency of the two methods.

Despite the fact that the asymptotic variance parameter σ is typically unknown making it difficult to assess the asymptotic efficiency of an inference method, some reasoning can still be done by shifting focus to the corresponding estimator $\hat{\sigma}(n)$ being employed.

Restricting attention to inferential procedures based on a single simulation run, [2, 9] argue that methods based on asynchronous observations, including the regenerative method,is preferable since such methods operates on the inherent natural time-scale of the process. It is reasonable to think that the correlation structure of the process is more suitably or efficiently estimated on the natural time scale $(T_1, T_2, \dots)$ than some arbitrary sequence $(\Delta, 2\Delta, \dots)$ of equally spaced instants. E.g. if the time between events tends to be large then one would prefer a large Δ to avoid highly correlated observations. However, by using the T_i's instead one automatically compensates for this correlation effect, without any need to deal with choice of the parameter Δ.

In finding an estimator $\hat{\sigma}(n)$ for the asymptotic variance parameter there is often plenty of room for ingenuity. Variance reduction techniques generally refers to methods aiming at reducing this variance estimate so as to improve the efficiency. We will say nothing special about variance reduction techniques here except to emphasize its impact. Note, however, that such techniques are often closely associated with the particular inferential approach being used. In addition there are techniques being more generally applicable. For an overview of the subject, the reader is refered to [2, 23].

7.2.2 Observation hardness

As previously pointed out the replicated-run approach suffers from the fact that it must deal with a new warm-up period for each subsequent observation taken. Clearly, single-run methods perform better in this respect and we generally expect the time between observations to be shorter. It must be emphasized that time here actually refers to the number of simulated events between the takings of two observations. In the following we use the term observation hardness to describe this. Hence, the observation hardness of the replicated-run approach is more prominent. Obviously, observation hardness is undesirable since it leads to an increased real simulation time to achieve the same number of observations. With reference to equation 7.1 this in turn gives a reduced overall efficiency.

Another manifestation of the observation hardness problem can be seen by considering the regenerative method. The point is that the length of regeneration cycles increases as the traffic intensity grows. This is intuitively reasonable. For traffic intensities close to one the occurrence of a regeneration point, i.e. an empty queue, is really a rare event. Consequently, the observation hardness becomes high and the efficiency of the regenerative method drops. Keep in mind, however, that the regenerative method do not suffer from warm-up effects and under normal circumstances the observation hardness of the *initial* observation is smaller than for the other methods.

Finally, note that the method based on asynchronous observations is naturally adaptable with respect to observation hardness. This is so since the number of simulated events between any two takings is always one. In addition, operating on the intrinsic natural time scale of the system there is no need to introduce an artificial sample-

spacing parameter Δ. This gives an computational advantage with respect to data collection.

Chapter 8

Concluding remarks

The most important conclusion is that analytical methods have limited applicability. It takes significant efforts to perform a mathematical analysis even of the most trivial $M/M/1$ queuing system. The key point is that analytical tractability depends on stochastic independence and Markov behavior. If such assumptions cannot be justified, simulation is the preferred tool for performance evaluation.

It must be recognized that a mathematical model seldom do an exact job in representing the system subject to analysis. This is especially due to the simplifying assumption that are often being made. Hence, an analytical method provides an *exact solution* of an *approximate model*. In contrast, a simulation provides an *approximate solution*, in terms of an estimate, of a more *exact model*.

It is important to be aware that *both* approaches end up with approximate results [20]. In general, it is hard to say which approach is most appropriate. After all, the final test is when the predictions are compared to actual measurements of a real system. One argument in favor of simulation is that the method is, in principle, applicable to systems of arbitrary complexity. The primary advantage of an analytical method is that a closed-form expression covers a large parameter space in a bold stroke.

Another point is that we have discussed analytical work under the provision that it should yield an explicit closed-formed performance expression. There are also generalized techniques that prepares a model which can be solved numerically by an algorithmic approach [25, 26]. The resulting solution is exact but otherwise this represents an intermediate case between analytical methods and simulation. It is interesting to note that [10] argues that simulation is often more computationally efficient. The reason is that a numerical approach will suffer when the underlying state-space grows. Complex models are usually characterized by a combinatorial exploding state-space.

Bibliography

[1] BANKS, J., AND CARSON, J. *Discrete-Event System Simulation*. Prentice-Hall, 1984.

[2] BRATLEY, P., FOX, B., AND SCHRAGE, L. *A Guide to Simulation*. Springer-Verlag, 1987.

[3] CASELLA, G., AND BERGER, R. *Statistical Inference*. Wadsworth, Brooks & Cole Publishing Company, 1990.

[4] CINLAR, E. *Introduction to Stochastic Processes*. Prentice-Hall, Inc., 1975.

[5] CINLAR, E. Regenerative processes. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. Johnson, Eds. John Wiley & Sons, 1982, pp. 673–676.

[6] COOPER, R. *Introduction to Queuing Theory*. Edward Arnold Ltd., 1981.

[7] DAMERDJI, H. Strong consistency of the variance estimator in steady-state simulation output analysis. *Mathematics of Operations Research 19*, 2 (1994), 494–512.

[8] FISHMAN, G. *Concepts and Methods in Discrete Digital Simulation*. John Wiley and Sons, 1973.

[9] FOX, B., AND GLYNN, P. Estimating time averages via randomly-spaced observations. *SIAM Journal on Applied Mathematics 47*, 1 (1987), 186–200.

[10] GLYNN, P., AND IGLEHART, D. Simulation methods for queues: An overview. *Queuing Systems: Theory and Applications 3* (1988), 221–256.

[11] GLYNN, P., AND WHITT, W. A central-limit theorem version of $L = \lambda W$. *Queuing Systems 2* (1986), 191–215.

[12] GLYNN, P., AND WHITT, W. Ordinary CLT and WLLN versions of $L = \lambda W$. *Mathematics of Operations Research 13*, 4 (1988), 674–692.

[13] GLYNN, P., AND WHITT, W. Indirect estimation via $L = \lambda W$. *Operations Research 37*, 1 (1989), 82–103.

[14] GRENANDER, U., AND ROSENBLATT, M. *Statistical Analysis of Stationary Time Series*. John Wiley and Sons, Inc., 1957.

[15] HEYDE, C. Central limit theorems. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. Johnson, Eds. John Wiley & Sons, 1982, pp. 651–655.

[16] HEYDE, C. Laws of large numbers. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. Johnson, Eds. John Wiley & Sons, 1982, pp. 566–568.

[17] HOEFFDING, W. Asymptotic normality. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. Johnson, Eds. John Wiley & Sons, 1982, pp. 673–676.

[18] KELTON, W. Statistical issues in simulation. In *Proc. of the 1996 Winter Simulation Conference* (1996), pp. 47–54.

[19] KLEINROCK, L. *Queuing Systems: Vol. 1 Theory*. Wiley, 1975.

[20] KLEINROCK, L. On the modeling and analysis of computer networks. *Proc. of the IEEE 81*, 8 (1993), 1179–1191.

[21] LARSEN, R., AND MARX, M. *An Introduction to Mathematical Statistics and Its Applications*, second ed. Prentice-Hall, 1986.

[22] LAW, A. Efficient estimators for simulated queuing systems. *Management Science 22*, 1 (1975), 30–41.

[23] LAW, A., AND KELTON, W. *Simulation Modeling and Analysis*, second ed. McGraw-Hill, 1991.

[24] LUKACS, E. Convergence of sequences of random variables. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. Johnson, Eds. John Wiley & Sons, 1982, pp. 183–187.

[25] NEUTS, M. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins Univ. Press, 1981.

[26] NEUTS, M. *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*. Marcel Dekker, Inc., 1989.

[27] PAWLIKOWSKI, K. Steady-state simulation of queuing processes: A survey of problems and solutions. *ACM Computing Surveys 22*, 2 (1990), 123–170.

[28] RIPLEY, B. *Stochastic Simulation*. John Wiley and Sons, 1987.

[29] ROLSKI, T., AND STIDHAM, S. Continuous versions of the queuing formulas $L = \lambda W$ and $H = \lambda G$. *Operations Research Letters 2*, 5 (1983), 211–215.

[30] SHEDLER, G. *Regeneration and Networks of Queues*. Springer-Verlag, 1987.

[31] STIDHAM, S. A last word on $L = \lambda W$. *Operations Research 22*, 2 (1974), 417–421.

[32] SYSKI, R. Stochastic processes. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. Johnson, Eds. John Wiley & Sons, 1982, pp. 836–851.

[33] TAYLOR, H., AND KARLIN, S. *An Introduction to Stochastic Modeling*. Academic Press, Inc., 1984.