7

Nonparametric Estimation of Tail Probabilities for the Single-Server Queue

Peter W. Glynn and Marcelo Torres

ABSTRACT We consider the estimation of tail probabilities in queues via the nonparametric estimator constructed by simply computing the observed fraction of time that the queue is out in the tail. We show that for reflected Brownian motion, the M/M/1 queue-length process, and the GI/G/1 waiting time sequence that the amount of time over which one must observe the queue grows exponentially in the tail parameter when such a nonparametric estimator is used.

7.1 Introduction

This paper is concerned with the question of how long one must observe a queue in order to accurately estimate a tail probability corresponding to a performance measure like queue-length or waiting time. The motivation for the problem stems from the current interest in developing robust admission control schemes for high-speed packet-based telecommunications networks. Assuming that a principle objective of such a scheme would likely be to minimize packet loss at the buffers associated with the switches in such a network, the value of estimating the loss probabilities from observed traffic becomes clear. As indicated above, we choose here to focus on tail probabilities for queues having an infinite buffer, in the belief that such a theory is likely to also describe the asymptotic behavior of estimators for loss networks of the type arising in the above telecommunications context. We note, in passing, that replacing a finite buffer system by an infinite buffer analog is a commonly used approximation in the queueing community.

There are many different types of estimation methodologies that can be used in this context. Our focus, in this paper, will be on a nonparametric formulation. Assuming that essentially nothing is known about the traffic source to the queue, one may be tempted to try to estimate the tail probability by the observed fraction of time that the queue is out in the tail. (Any other estimator will typically take advantage of additional structure that we may be unwilling to assume.) This observed fraction of time

is precisely the nonparametric tail estimator considered in this paper. A companion paper (Torres and Glynn [12]) describes a competing estimator, based on parametric statistical modeling of the input source, that assumes more about the structure of the queue.

Because so little is assumed about the queue, one might expect such nonparametric estimators to be very inefficient. In particular, one expects that the amount of time over which the system needs to be observed in order to accurately estimate a tail probability increases as a function of the tail parameter. In this paper, we establish this result rigorously. In fact, we compute the critical rate at which the observed time horizon needs to grow as a function of the tail parameter; the growth is typically exponential. This suggests that for the finite buffer queues arising in packet networks that the amount of traffic observed needs to grow exponentially in the size of the buffers. From a pragmatic standpoint, this likely renders the nonparametric methodology considered here impractical. As a consequence, the analysis in this paper makes a strong case for the need to impose more model structure on the input sources in attempting to estimate such loss probabilities.

In addition to the above conclusion, we view the following as the main results in the paper:

1. We develop general theory (Proposition 7.2.1), relevant to non-queues as well as queues, on how much time a stochastic process needs to be observed in order to accurately estimate the probability of a rare event. In addition to the statistical implications, this result is also relevant to simulation, in which the nonparametric estimators considered in this paper are especially widely used.

2. We develop central limit theorems, with explicitly computed time-average variance constants, describing the rate of convergence of non-parametric tail estimators for both reflected Brownian motion and the M/M/1 queue. These results illustrate the use of stochastic calculus, martingale methods, regeneration, and Poisson's equation in obtaining central limit theorems for stochastic processes.

3. In the setting of the GI/G/1 waiting time sequence, we obtain a quite precise solution to the question of how long one must observe the queue in order to accurately estimate a tail probability. The critical rate is exponential and depends on the Cramér-Lundberg parameter; see Theorems 7.5.1 and 7.5.2.

This paper is organized as follows. General theory is developed in Section 7.2, whereas Sections 7.3 and 7.4 are devoted to the analysis of reflected Brownian motion and the M/M/1 queue, respectively. Section 7.5 undertakes the analysis of the GI/G/1 queue, while Section 7.6 offers conclusions, including some discussion of the competing parametric methodology.

7.2 A Bit of General Theory

Our concern, in this paper, is with the nonparametric estimation of (extreme) tail probabilities associated with the single-server queue. This can be cast in more general terms, as follows.

Let $X = (X_n : n \geq 0)$ be a (strictly) stationary stochastic process taking values in a state space S. For a given sequence of subsets $(A_m : m \geq 1)$ contained in S, consider the problem of estimating the quantity $p_m = P(X_0 \in A_m)$ for m large. We assume throughout that $p_m \downarrow 0$ as $m \to \infty$. In the single-server queue context that interests us, X_n might, for example, be the (stationary) waiting time of the n'th customer to arrive to the queue, with $A_m = [m, \infty)$, in which case p_m is an (extreme) tail probability for the steady-state waiting time distribution.

The obvious non-parametric estimator for p_m is clearly

$$\alpha(n;m) \stackrel{\Delta}{=} \frac{1}{n}\sum_{j=0}^{n-1} I(X_j \in A_m).$$

The issue, then, is how large one must choose n, as a function of m, in order that $\alpha(n;m)$ be an accurate estimator of p_m. This can be re-formulated, more precisely, in terms of the relative error given by $|\alpha(n;m)/p_m - 1|$. In particular, how fast must $n_m \to \infty$ in order that

$$\frac{1}{p_m}\alpha(n_m;n) \Rightarrow 1 \tag{7.1}$$

as $m \to \infty$? The answer is easy if the X_n's are i.i.d.

In this case, note that (7.1) holds, provided that

$$p_m^{-2}\text{var } \alpha(n_m;m) \to 0 \tag{7.2}$$

as $m \to \infty$. But var $\alpha(n;m) = n^{-1}p_m(1-p_m)$. Hence, (7.2) follows if we require that $p_m n_m \to \infty$ as $m \to \infty$. On the other hand, if $p_m n_m \to c > 0$ as $m \to \infty$, then

$$n_m\alpha(n_m;m) \stackrel{\mathcal{D}}{=} \text{Binomial}(n_m, p_m) \tag{7.3}$$

$$\Rightarrow \text{Poisson}(c) \tag{7.4}$$

as $m \to \infty$, from which it is evident that

$$\frac{1}{p_m}\alpha(n_m;m) \Rightarrow \frac{1}{c}\text{Poisson}(c)$$

as $m \to \infty$, violating (7.1). Consequently, $n_m \gg 1/p_m$ is a necessary and sufficient condition for the (relatively) accurate determination of p_m.

Our goal is now to see how far this analysis can be extended when X is a dependent sequence, as is typical of queueing applications. First, note that simple algebra verifies that

$$n \cdot \text{var } \alpha(n;m) = \text{var } I_0(m) + 2\sum_{j=1}^{n-1}(1-j/n)\text{cov } (I_0(m), I_j(m)) \quad (7.5)$$

where $I_j(m) \stackrel{\Delta}{=} I(X_j \in A_m)$. Hence, if the stationary sequence $(I_j(m) : j \geq 0)$ has non-negative autocorrelations (i.e. cov $(I_0(m), I_j(m)) \geq 0$ for $j \geq 1$), it follows that

$$\begin{aligned} \text{var } \alpha(n;m) &\geq \frac{1}{n}\text{var } I_0(m) \\ &= \frac{1}{n}p_m(1-p_m). \end{aligned}$$

Thus, assuming $\limsup_{m\to\infty} n_m p_m < +\infty$, it is evident that

$$\liminf_{m\to\infty} \text{var } [p_m^{-1}\alpha(n_m;m)] > 0.$$

So, in the presence of non-negative autocorrelations, one requires that $n_m \gg 1/p_m$ in order that the variance of $p_m^{-1}\alpha(n_m;m)$ converge to zero.

This has important implications for estimation of tail probabilities in the single-server queue. Consider, for example, the stationary waiting time sequence $W = (W_n : n \geq 0)$ associated with the GI/G/1 queue with traffic intensity lass than unity. Suppose that our interest is in estimating the tail probability $p_m = P(W_0 \geq m)$. It is well known that W is a stochastically monotone Markov chain, so that the stationary waiting times $(W_n : n \geq 0)$ form an associated sequence; see, for example, Stoyan [11]. Note that for each $m \geq 1$, $I_j(m) = I(W_j \geq m)$ is a non-decreasing function of W_j, and consequently $(I_j(m) : j \geq 0)$ is itself an associated sequence. Thus, it is evident that cov $(I_0(m), I_j(m)) \geq 0$ for $j \geq 1$, establishing the fact that our non-negative autocorrelation condition holds in this setting. Hence, in order that the variance of $p_m^{-1}\alpha(n_m;m)$ go to zero as $n \to \infty$, it is necessary that $n_m \gg p_m^{-1}$. Of course, the Cramér-Lundberg approximation states that, under general conditions, $P(W_0 \geq x) \sim a\exp(-\theta^* x)$ as $x \to \infty$, for suitable positive constants a and θ^*; see, for example, Asmussen [1]. As a result, we may conclude that we need to at least observe W over a time horizon of order $\exp(\theta^* m)$ in order to estimate p_m to a reasonable degree of accuracy. However, given the complex dependency structure of W, it is conceivable that one might need to observe W over a time scale much longer than p_m^{-1}. We address this issue next.

A standard concept used in the stochastic process setting to describe dependence is that of mixing. Without loss of generality, we may extend $(X_j : j \geq 0)$ to a "two-sided" (strictly) stationary sequence $(X_j : -\infty < j < \infty)$. Let $\mathcal{F}_j = \sigma(X_k : k \leq j)$ and $\mathcal{F}^j = \sigma(X_k : k \geq j)$. The j'th $(j \geq 1)$ uniform mixing coefficient of X is defined by

$$\varphi_\infty(j) \stackrel{\Delta}{=} \sup_{A\in\mathcal{F}^j} \sup_{\substack{B\in\mathcal{F}_0 \\ P(B)>0}} |P(A|B) - P(A)|.$$

Assume that X is uniformly mixing in the sense that

$$\sum_{j=1}^{\infty} \varphi_\infty(j) < \infty. \quad (7.6)$$

Then, a standard inequality (see, for example, Ethier and Kurtz [5]) states that

$$|\text{cov } (I_0(m), I_j(m))| \leq 2\varphi_\infty(j)EI_0(m). \quad (7.7)$$

It follows from (7.5) - (7.7) that

$$\begin{aligned} n \cdot \text{var } \alpha(n;m) &\leq \text{var } I_0(m) + 4\sum_{j=1}^{\infty}\varphi(j)EI_0(m) \quad (7.8) \\ &\leq p_m(1+4\sum_{j=1}^{\infty}\varphi(j)) \quad (7.9) \end{aligned}$$

and hence

$$\text{var } [p_m^{-1}\alpha(n_m;m)] \leq \frac{1}{n_m p_m}(1+4\sum_{j=1}^{\infty}\varphi(j)) \to 0$$

provided that $n_m p_m \to +\infty$. We summarize the above discussion with the following result.

Proposition 7.2.1 Suppose that X satisfies the uniform mixing condition (7.6). Then, if $p_m n_m \to +\infty$,

$$p_m^{-1}\alpha(n_m;m) \Rightarrow 1$$

as $m \to \infty$. Conversely, if $(I_j(m) : j \geq 0)$ has non-negative autocorrelations for each $m \geq 1$, then it is necessary that $p_m n_m \to \infty$ in order that

$$\text{var } [p_m^{-1}\alpha(n_m;m)] \to 0$$

as $m \to \infty$.

Thus Proposition 7.2.1 gives fairly general conditions under which observing X for a time period $n_m \gg 1/p_m$ is necessary and sufficient for (relatively) accurate estimation of p_m. It effectively shows that, under the conditions stated, the dependence is such that the relative magnitude of n_m is the same as in the i.i.d. case.

It turns out that the uniform mixing condition is fundamental to the sufficiency half of Proposition 7.2.1. To see this, consider the stationary Markov chain X on the non-negative integers $\{0, 1, ...\}$ with transition probabilities $P(0, y) = a_y > 0$ for $y \geq 1$, $P(y, 0) = b_y > 0$ for $y \geq 1$, $P(y, y) = 1 - b_y$ for $y \geq 1$, and $P(x, y) = 0$ for all other pairs $(x, y) \in S \times S$. Assume $\sum_{k=1}^{\infty} a_k/b_k < \infty$. Then, X is irreducible and positive recurrent with $\pi_0 = (1 + \sum_{j=1}^{\infty} a_j/b_j)^{-1}$ and

$$\pi_j = \pi_0 a_j / b_j$$

for $j \geq 1$. Let $A_m = \{m\}$, so that $p_m = \pi_m$. Then,

$$\begin{aligned} & P(\frac{1}{n_m} \sum_{j=0}^{n_m - 1} I(X_j = m) = 0) \\ \geq & \ \pi_0 P(\frac{1}{n_m} \sum_{j=0}^{n_m - 1} I(X_j = m) = 0 | X_0 = 0) \\ = & \ \pi_0 P(T_m \geq n_m | X_0 = 0) \end{aligned}$$

where $T_m = \inf\{j \geq 0 : X_j = m\}$. Observe that

$$T_m \geq \sum_{j=0}^{T_m - 1} I(X_j = 0) \stackrel{\Delta}{=} \beta_m,$$

where β_m is geometric with parameter a_m. So,

$$P(\frac{1}{n_m} \sum_{j=0}^{n_m - 1} I(X_j = m) = 0) \geq \pi_0 (1 - a_m)^{n_m}.$$

Choose $a_j = 2^{-j}$, $b_j = 1/j$, and $n_j = 2^j$. Observe that $p_m n_m \to \infty$ as $m \to \infty$. On the other hand,

$$\liminf_{m \to \infty} P(\frac{1}{n_m} \sum_{j=0}^{n_m - 1} I(X_j = m) = 0) \geq \pi_0 / e > 0,$$

violating (7.1). The difficulty here is that this chain, although strong mixing (see Ethier and Kurtz (1986), p. 345), is not uniformly mixing.

Unfortunately, it turns out that the standard processes that arise in conjunction with the single-server queue are not uniformly mixing. Take, for example, the waiting time sequence W. Because W is Markov, it is evident that

$$\varphi_\infty(j) = \sup_A \sup_x |P(W_j \in A | W_0 = x) - P(W_j \in A)|.$$

But for each $j \geq 1$ and $b > 0$, it is clear that

$$P(W_j \in [0, b] | W_0 = x) \to 0$$

as $x \to \infty$, and hence $\varphi_\infty(j) = 1$ for $j \geq 1$, violating (7.6). Consequently, we can not invoke the sufficiency half of Proposition 7.2.1, and must consider other methods of attack for studying this problem in the context of queues.

7.3 Analysis of Reflected Brownian Motion

In this section, we consider the problem of estimating tail probabilities for one-dimensional reflected Brownian motion (RBM). As is well known, this stochastic process is the limit process that arises when studying the single-server queue in heavy traffic; see, for example, Glynn [8]. Thus, we would minimally expect the theory developed for RBM to be representative of the qualitative behavior of queues in heavy traffic. In addition, RBM is a particularly tractable stochastic process from a mathematical viewpoint, making the required analysis especially straightforward.

Let $X = (X(t) : t \geq 0)$ be a stationary version of a one-dimensional RBM living on $[0, \infty]$, having drift $-\mu < 0$ and infinitesimal variance $\sigma^2 > 0$. Then, in order that X be stationary, it follows that for each $t \geq 0$, $X(t)$ has an exponential distribution with parameter $\theta^* \stackrel{\Delta}{=} 2\mu/\sigma^2$; see Asmussen [1]. Our goal here is to estimate the tail probability

$$\alpha(y) \stackrel{\Delta}{=} P(X(0) \geq y) = \exp(-\theta^* y).$$

Our focus will be on estimating $\alpha(y)$ via the nonparametric time-average

$$\alpha(t; y) \stackrel{\Delta}{=} \frac{1}{t} \int_0^t I(X(s) \geq y) ds.$$

Clearly, in practice, if we know a priori that the observed process X is such an RBM, we would choose to estimate $\alpha(y)$ via a parametric estimator of the form $\exp(-2\hat{\mu} y/\sigma)$, where $\hat{\mu}$ is (for example) the maximum likelihood estimator of μ. (Note that σ^2 can be estimated without error by computing the quadratic variation of X.) However, our interest is in considering the estimation of $\alpha(y)$ in settings in which we are unwilling to make any a priori parametric assumptions about the observed process. In such a context, $\alpha(t; y)$ is perhaps the most natural estimator for $\alpha(y)$.

To proceed with our analysis, let $f(x; y) = I(x \geq y)$ and $f_c(x; y) = f(x; y) - \alpha(y)$. Then

$$\alpha(t; y) - \alpha(y) = \frac{1}{t} \int_0^t f_c(X(s); y) ds.$$

A key idea in studying $\int_0^t f_c(X(s);y)ds$ is to represent this additive functional of X as a semimartingale. We can construct such a representation via judicious use of the Itô calculus.

Let $u(\cdot;y)$ satisfy the conditions of Itô's lemma so that, in particular, it has a continuous first derivative and second derivative existing a.e. Then, recall that

$$dX(t) = -\mu dt + \sigma dB(t) + dL(t),$$

where $B = (B(t) : t \geq 0)$ is a one-dimensional standard Brownian motion and $L = (L(t) : t \geq 0)$ is a non-decreasing process that increases only when X is zero; see Harrison [14] for details. Itô's formula states that for $u(\cdot;y)$ chosen as above,

$$du(X(t);y) = u'(X(t);y) + \frac{1}{2}u''(X(t);y)\sigma^2 dt,$$

where the derivatives are computed with respect to the first component of u. Hence,

$$\begin{aligned} u(X(t);y) - u(X(0);y) &= \int_0^t \{-\mu u'(X(s);y) + \frac{\sigma^2}{2}u''(X(s);y)\}ds \\ &\quad + \int_0^t u'(X(s);y)\sigma dB(s) \\ &\quad + \int_0^t u'(X(s))dL(s). \end{aligned} \tag{7.10}$$

Suppose that we choose $u(\cdot;y)$ to satisfy "Poisson's equation" for RBM:

$$-\mu u'(\cdot;y) + \tfrac{\sigma^2}{2}u''(\cdot;y) = -f_c(\cdot;y) \tag{7.11}$$

such that

$$u'(0;y) = 0.$$

Because $L(\cdot)$ increases only when $X(\cdot)$ is zero, evidently, the last term on the right-hand side of (7.10) vanishes. By our choice of $u(\cdot;y)$, it then follows that

$$\int_0^t f_c(X(s);y)ds + u(X(t);y) - u(X(0);y) = \int_0^t u'(X(s);y)\sigma dB(s). \tag{7.12}$$

But the stochastic integral $\int_0^t u'(X(s);y)\sigma dB(s)$ is a martingale, provided that $u'(\cdot;y)$ is appropriately integrable. Thus, (7.12) provides us with a martingale representation for our additive functional $\int_0^t f_c(X(s);y)ds$.

Note that (7.11) only determines $u(\cdot;y)$ up to an additive constant. Arbitrarily setting $u(0;y) = 0$, it is straightforward to verify that the unique solution to (7.11) is then given by

$$u(x;y) = \begin{cases} \frac{e^{-\theta^* y}}{\mu\theta^*}(e^{\theta^* x} - 1) - \frac{\alpha(y)}{\mu}x & ; 0 \leq x \leq y, \\ \frac{(1-e^{-\theta^* y} - y\theta^* e^{-\theta^* y})}{\mu\theta^*} + (\frac{1-\alpha(y)}{\mu})(x-y) & ; x \geq y. \end{cases}$$

This solution has a continuous first derivative and satisfies the necessary regularity hypothesis required for the application of Itô's lemma; see Chung and Williams [4]. In addition, since $u'(\cdot;y)$ is bounded, it follows that the stochastic integral $\int_0^t u'(X(s);y)\sigma dB(s)$ is in fact a martingale.

We are interested in knowing over how long a time interval t_y one must observe X in order to obtain a (relatively) accurate estimate of $\alpha(y)$. This amounts to asking how fast t_y must grow in order that

$$\alpha(t_y;y)/\alpha(y) \Rightarrow 1 \tag{7.13}$$

as $y \to \infty$. Relation (7.13) is equivalent to showing that

$$\frac{\alpha(y)^{-1}}{t_y}\int_0^{t_y} f_c(X(s);y)ds \Rightarrow 0$$

as $y \to \infty$. Hence by (7.12), we need to show that

$$\frac{\alpha(y)^{-1}}{t_y}\int_0^{t_y} u'(X(s);y)\sigma dB(s) + \frac{\alpha(y)^{-1}}{t_y}u(X(0);y) - \frac{\alpha(y)^{-1}}{t_y}u(X(t);y) \Rightarrow 0 \tag{7.14}$$

as $y \to \infty$. Now,

$$\begin{aligned} \alpha(y)^{-1}E|u(X(t_y);y)| &= \alpha(y)^{-1}E|u(X(0);y)| && (7.15) \\ &= \alpha(y)^{-1}|\int_0^\infty |u(x;y)|\theta^* e^{-\theta^* x}dx && (7.16) \\ &= O(y) && (7.17) \end{aligned}$$

and consequently we require that $t_y/y \to +\infty$ in order that the latter two terms in (7.14) converge to zero in expectation, implying convergence in probability to zero.

As for the first term on the left-hand side of (7.14), the L^2-isometry property of stochastic integrals implies that

$$\begin{aligned} &E[\frac{\alpha(y)^{-2}}{t_y^2}(\int_0^{t_y} u'(X(s);y)\sigma dB(s))^2] \\ &= \frac{\alpha(y)^{-2}}{t_y^2}E\int_0^{t_y} u'(X(s);y)^2\sigma^2 ds \\ &= \frac{\alpha(y)^{-2}}{t_y}\sigma^2 Eu'(X(0);y)^2 \\ &= \frac{\alpha(y)^{-2}}{t_y}\sigma^2\int_0^\infty u'(x;y)^2\theta^* e^{-\theta^* x}dx \end{aligned}$$

But

$$\alpha(y)^{-2}\int_0^\infty u'(x;y)^2\theta^* e^{-\theta^* x}dx = O(e^{\theta^* y})$$

as $y \to \infty$, so it follows that if $e^{\theta^* y}/t_y \to 0$, then the L^2-norm of the first term in (7.14) converges to zero, proving convergence in probability. We have therefore proved the following result.

Theorem 7.3.1 If $e^{\theta^* y}/t_y \to 0$ as $y \to \infty$, then

$$\alpha(t_y; y)/\alpha(y) \Rightarrow 1$$

as $y \to \infty$.

Consequently, $t_y \gg \alpha(y)^{-1}$ is sufficient for (relatively) accurate determination of $\alpha(y)$ via the nonparametric estimator $\alpha(t_y; y)$. Thus, it is enough to observe X over an amount of time that increases exponentially (at rate θ^*) in the level y.

In developing estimators for $\alpha(y)$, it is often of some interest to obtain a rate of convergence, as expressed via a central limit theorem (CLT). Our first CLT characterizes the rate of convergence, for fixed y, as $t \to \infty$.

Theorem 7.3.2 Fix $y > 0$. Then,

$$t^{1/2}(\alpha(t; y) - \alpha(y)) \Rightarrow \sigma(y)N(0, 1)$$

as $t \to \infty$, where $\sigma^2(y) = 2(\frac{\sigma}{\mu})^2 \alpha(y)^2\{\alpha(y)^{-1} - y - 1\}$.

Proof. Because of (7.12), Theorem 7.3.2 follows from an invocation of the martingale central limit theorem (see p. 339 of Ethier and Kurtz (1986)). To prove Theorem 7.3.2, the martingale CLT requires that we show

$$\frac{1}{t}[M_y](t) \Rightarrow \sigma^2(y) \tag{7.18}$$

as $t \to \infty$, where $[M_y](t)$ is the quadratic variation over $[0, t]$ of the martingale $M_y(t) \stackrel{\Delta}{=} \int_0^t u'(X(s); y)\sigma dB(s)$. But

$$[M_y](t) = \sigma^2 \int_0^t u'(X(s); y)^2 ds.$$

By the ergodic theorem, it follows that

$$\frac{1}{t}[M_y](t) \to \sigma^2 E u'(X(0); y)^2$$

a.s. as $t \to \infty$. Noting that $u'(x; y) = \mu^{-1}\alpha(y)(\exp(\theta^*(x \wedge y)) - 1)$, it is easily verified that $\sigma^2 E u'(X(0); y)^2$ equals the given expression for $\sigma^2(y)$; this establishes (7.18) and proves that

$$t^{-1/2} M_y(t) \Rightarrow \sigma(y)N(0, 1)$$

as $t \to \infty$. The theorem is then an immediate consequence of the representation (7.12) and the fact that $u(X(t); y)/t^{1/2}$ and $u(X(0); y)/t^{1/2}$ both have the same distributions and converge in probability to zero. □

Of course, given our interest in the behavior of $\alpha(t; y)$ for large y, the more interesting question is perhaps the CLT behavior of $\alpha(t_y; y)$ as $y \to \infty$. Theorem 7.3.2 suggests the approximation

$$\alpha(t_y; y) \stackrel{\mathcal{D}}{\approx} \alpha(y) + \sqrt{\frac{\sigma^2(y)}{t_y}} N(0, 1), \tag{7.19}$$

where $\stackrel{\mathcal{D}}{\approx}$ denotes "approximately equal in distribution to." Noting that $\sigma^2(y) \sim 2\sigma^2\alpha(y)/\mu^2$ as $y \to \infty$, (7.19) can be re-cast in the form

$$\frac{\alpha(t_y; y)}{\alpha(y)} \stackrel{\mathcal{D}}{\approx} 1 + \frac{\sigma}{\mu}\sqrt{\frac{2}{\alpha(y)t_y}} N(0, 1). \tag{7.20}$$

As a consequence, the relative error is of the order of $(\alpha(y)t_y)^{-1/2}$, showing clearly the benefits of observing X over a time horizon $[0, t_y]$, with $t_y \gg \alpha(y)^{-1}$. Our final result of this section is a CLT that makes (7.20) rigorous.

Theorem 7.3.3 Suppose $t_y\alpha(y) \to +\infty$ as $y \to \infty$. Then,

$$\sqrt{\alpha(y)t_y}\left(\frac{\alpha(t_y; y)}{\alpha(y)} - 1\right) \Rightarrow \sqrt{2}(\sigma/\mu)N(0, 1)$$

as $y \to \infty$.

Proof. As in the proof of Theorem 7.3.2, the key step is showing that

$$\frac{1}{t_y\alpha(y)}[M_y](t_y) \Rightarrow \frac{2\sigma^2}{\mu^2}$$

as $y \to \infty$. Equivalently, we need to prove that

$$\frac{1}{t_y\alpha(y)} \int_0^{t_y} u'(X(s); y)^2 ds \Rightarrow \frac{2\sigma^2}{\mu^2} \tag{7.21}$$

as $y \to \infty$. But $\alpha(y)^{-1}u'(x; y)^2 = \mu^{-2}\alpha(y)\sigma^2 \exp(2\theta^*(x \wedge y)) + O(1)$, where $O(1)$ is uniform in x and y. Since $t_y \to \infty$, (7.21) holds, provided that

$$\frac{\alpha(y)}{t_y} \int_0^{t_y} h(X(s); y) ds \Rightarrow 2 \tag{7.22}$$

as $y \to \infty$, where $h(x; y) \stackrel{\Delta}{=} \exp(2\theta^*(x \wedge y))$. Let $h_c(x; y) = h(x; y) - 2\alpha(y)^{-1}$ and set

$$w(x; y) = \begin{cases} \frac{(2+4\alpha(y)^{-1})}{\sigma^2\theta^*} e^{\theta^* x} - \frac{4\alpha(y)}{\sigma^2\theta^*} x - \frac{1}{\sigma^2(\theta^*)^2} e^{2\theta^* x} & ; 0 \le x \le y \\ w(y; y) + 2(\alpha(y)^{-2} - \alpha(y)^{-1})(x - y)/\mu & ; x > y. \end{cases}$$

Then, it is straightforward to verify that $w(\cdot;y)$ has a continuous first derivative and satisfies at $x \neq y$ the differential equation

$$\frac{\sigma^2}{2}w''(\cdot;y) - \mu w'(\cdot;y) = -h_c(\cdot;y)$$

subject to $w'(0;y) = 0$. Proceeding as in the development of (7.12), we can then show that

$$\int_0^{t_y} h_c(X(s);y)ds + w(X(t_y);y) - w(X(0);y) = \int_0^{t_y} w'(X(s);y)\sigma dB(s). \tag{7.23}$$

Now, $|w(x;y)| \leq a_1 + a_2\alpha(y)^{-2}x$ uniformly in x and y (for suitably chosen a_1, a_2) so that $\alpha(y)|w(X(t_y);y)|/t_y \leq |X(t_y)|/t_y\alpha(y)$. Since $t_y\alpha(y) \to \infty$ and $X(t_y)$ has a distribution independent of t_y, it follows that

$$w(X(t_y);y)\alpha(y)/t_y \Rightarrow 0$$

as $y \to \infty$. A similar, but simpler, argument shows that

$$w(X(0);y)\alpha(y)/t_y \Rightarrow 0$$

as $y \to \infty$. As for the stochastic integral,

$$E[(\int_0^{t_y} w'(X(s);y)dB(s))^2\alpha(y)^2/t_y^2] = Ew'(X(0);y)^2\alpha(y)^2/t_y \to 0$$

as $y \to \infty$, proving that $\int_0^{t_y} w'(X(s);y)dB(s)\alpha(y)/t_y \to 0$ in mean square and hence in probability. The representation (7.23) therefore proves that $\int_0^{t_y} h_c(X(s);y)ds \cdot \alpha(y)/t_y \Rightarrow 0$ as $y \to \infty$, yielding (7.22).

The martingale CLT then implies that

$$(t_y\alpha(y))^{-1/2}M_y(t_y) \Rightarrow \frac{\sigma}{\mu}\sqrt{2}N(0,1) \tag{7.24}$$

as $y \to \infty$. But since $u(x;y) \leq b_1 + b_2x$ uniformly in x and y,

$$(t_y\alpha(y))^{-1/2}u(X(t);y) \Rightarrow 0$$

and

$$(t_y\alpha(y))^{-1/2}u(X(0);y) \Rightarrow 0$$

as $y \to \infty$. Relations (7.24) and (7.12) then finish the proof of the theorem. □

7.4 Analysis of the M/M/1 Queue

In this section, we turn to the analysis of the M/M/1 queue. As perhaps the simplest model that exhibits true queueing behavior (as opposed to the "approximate" queueing behavior of RBM), we would hope to see the same qualitative behavior as that obtained in Section 7.3 for RBM.

We start with an analysis of the "number in system" process for the M/M/1 queue. To be precise, assume that $\lambda < \mu$ and let $Q = (Q(t) : t \geq 0)$ be a stationary version of the birth-death process on the non-negative integers $\{0, 1, 2, ...\}$ with birth rates $\lambda_n = \lambda(n \geq 0)$ and death rates $\mu_n = \mu(n \geq 1)$. Then, λ can, of course, be interpreted as the arrival rate to the queue, μ can be viewed as the service intensity, and $Q(t)$ represents the number of customers in the system as time t.

Suppose that we are interested in estimating the tail probability $\alpha(m) = P(Q(0) \geq m)$ via the nonparametric estimator

$$\alpha(t;m) \stackrel{\Delta}{=} \frac{1}{t}\int_0^t I(Q(s) \geq m)ds.$$

The issue here is how fast t must grow as a function of m, in order that $\alpha(t;m)$ be a (relatively) accurate estimator of $\alpha(m)$. To answer this question, we adopt the same basic approach as used in our analysis of RBM in Section 7.3.

Proceeding as in Section 7.3, let $f(x;m) = I(x \geq m)$ and set $f_c(x;m) = f(x;m) - P(Q(0) \geq m) = f(x;m) - \rho^m$, where $\rho \stackrel{\Delta}{=} \lambda/\mu$. We wish to express $\int_0^t f_c(Q(s);m)ds$ in terms of an appropriately defined martingale. To do so, we solve "Poisson's equation" for $u(\cdot;m)$:

$$\sum_{j=0}^{\infty} A_{ij}u(j;m) = -f_c(i;m), (i \geq 0)$$

where $A = (A_{ij} : i, j \geq 0)$ is the infinitesimal generator of $(Q(t) : t \geq 0)$. In other words, we must compute the solution $u(\cdot;m)$ to

$$\lambda u(i+1;m) - (\lambda+\mu)u(i;m) + \mu u(i-1;m) = -f_c(i;m),$$

for values $i \geq 1$, and

$$\lambda u(1;m) - \lambda u(0;m) = -f_c.$$

A solution to the above linear system is easily verified to be

$$u(i;m) = \begin{cases} \frac{\rho^m}{\mu(1-\rho)^2}(\rho^i - 1) - \frac{\rho^m i}{\mu(1-\rho)} & ; 0 \leq i < m \\ \frac{1-\rho^m}{\mu(1-\rho)^2} - \frac{\rho^m m}{\mu(1-\rho)} + \frac{(1-\rho^m)}{\mu(1-\rho)}(i-m) & ; i \geq m. \end{cases}$$

(Note the similarity to the solution $u(\cdot;y)$ obtained in the RBM context.) Noting that $u(Q(t);m)$ and $f_c(Q(t);m)$ are both integrable r.v.'s for each

$t \geq 0$, it is a well known fact (see, for example, p. 298 of Karlin and Taylor [10]) that for each $m \geq 1$,

$$M_m(t) \stackrel{\Delta}{=} u(Q(t);m) - u(Q(0);m) + \int_0^t f_c(Q(s);m)ds$$

is a martingale. To study the behavior of the relative error

$$\alpha(t;m)/\alpha(m)) - 1,$$

observe that

$$\frac{\alpha(t;m)}{\alpha(m)} - 1 = \frac{\alpha(m)^{-1}}{t}[M_m(t) - u(Q(t);m) + u(Q(0);m)]. \tag{7.25}$$

Now, $|u(Q(t);m))\alpha(m)^{-1}/t| \leq (c_1 + c_2|Q(t)|)/t$ independent of m (for suitable constants c_1, c_2). Since $E|Q(t)| = \rho(1-\rho)^{-1} < \infty$, it is evident that $\alpha(m)^{-1}u(Q(t_m);m)/t_m \Rightarrow 0$ as $m \to \infty$, provided that $t_m \to \infty$ (as does $\alpha(m)^{-1}u(Q(0);m)/t_m$). As for the first term on the right-hand side of (7.25), note that

$$[M_m](t) = \sum_{i=1}^{J(t)}[u(Y_j;m) - u(Y_{j-1};m)]^2,$$

where $Y = (Y_j : j \geq 0)$ is the embedded discrete-time Markov chain associated with $(Q(t) : t \geq 0)$, and $J(t)$ is the number of jumps of Q over $[0,t]$.

This quadratic variation process is a bit easier to handle if we uniformize. (Note, for example, that $Y = (Y_j : j \geq 0)$ is non-stationary, despite the fact that we have chosen Q to be stationary.) Let $\Gamma = (\Gamma(t) : t \geq 0)$ be a Poisson process having rate $\lambda+\mu$ and let $Z = (Z_n : n \geq 0)$ be a stationary version of the discrete-time Markov chain living on the non-negative integers having transition matrix $P = (P_{ij} : i,j \geq 0)$ with $P_{i,i+1} = p \stackrel{\Delta}{=} \lambda/(\lambda+\mu)(i \geq 0)$ and $P_{i,i-1} = P_{0,0} = 1-p$ for $i \geq 1$. Then, provided that Γ and Z are independent of one another,

$$[M_m](t) \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\Gamma(t)}[u(Z_i;m) - u(Z_{i-1};m)]^2,$$

where $\stackrel{\mathcal{D}}{=}$ denotes "equality in distribution". Then,

$$\begin{aligned} EM_m^2(t) &= E[M_m](t) \\ &= E\Gamma(t) \cdot E[u(Z_1;m) - u(Z_0;m)]^2 \\ &= 2(\lambda+\mu)t \cdot E[u(Z_0;m)^2 - u(Z_0;m)u(Z_1;m)] \\ &= -2(\lambda+\mu)t \cdot E[u(Z_0;m) \cdot [\sum_{j=0}^{\infty} P_{Z_0,j}u(j;m) - u(Z_0;m)]] \end{aligned}$$

By virtue of the fact that we constructed our solution $u(\cdot;m)$ so that $u(0;m) = 0$, it is straightforward to verify that

$$(\lambda+\mu)(\sum_{j=0}^{\infty} P_{ij}u(j;m) - u(i;m)) = \sum_{j=0}^{\infty} A_{ij}u(j;m) = -f_c(i;m)$$

and thus

$$EM_m^2(t) = 2tEu(Z_0;m)f_c(Z_0;m). \tag{7.26}$$

Since $Z_0 \stackrel{\mathcal{D}}{=} Q(0)$ (i.e. geometric with parameter $1-\rho$),

$$Eu(Z_0;m)f_c(Z_0;m) = \frac{(1+\rho)(1-\rho^m)\rho^m}{\mu(1-\rho)^2} - \frac{2\rho^{2m}m}{\mu(1-\rho)}. \tag{7.27}$$

Consequently, in view of (7.26), it is evident that

$$\frac{\alpha(m)^{-2}}{t_m^2}EM_m^2(t_m) \to 0,$$

provided that $t_m\rho^m \to \infty$. Thus, $\alpha(m)^{-1}M_m(t_m)/t_m \to 0$ in mean square, and therefore in probability, under the condition $t_m \gg \rho^{-m}$. Based on (7.25), we have proved the following result, which is the M/M/1 queue-length analog to Theorem 7.3.1.

Theorem 7.4.1 If $t_m\rho^m \to \infty$ as $m \to \infty$, then

$$\alpha(t_m;m)/\alpha(m) \Rightarrow 1$$

as $m \to \infty$.

We have also done essentially all the necessary work required to establish a CLT for $\alpha(t;m)$ for fixed m.

Theorem 7.4.2 Fix $m > 0$. Then,

$$t^{1/2}(\alpha(t;m) - \alpha(m)) \Rightarrow \sigma_1(y)N(0,1)$$

as $m \to \infty$, where

$$\sigma_1^2(y) = \frac{2(1+\rho)(1-\rho^m)\rho^m}{\mu(1-\rho)^2} - \frac{4\rho^{2m}m}{\mu(1-\rho)}.$$

Proof. The strong law for the chain $(Z_n : n \geq 0)$ guarantees that

$$\frac{1}{t}[M_m](t) \to 2Eu(Z_0;m)f(Z_0;m) \ \ a.s.$$

as $t \to \infty$. Since $M_m(\cdot)$ is a martingale with discontinuous sample paths, the martingale CLT requires that we verify the additional condition

$$t^{-1/2}E[\sup_{0 \leq s \leq t}|M_m(s) - M_m(s-)|] \to 0$$

as $t \to \infty$. But

$$\sup_{0 \le s \le t} |M_m(s) - M_m(s-)| = \max_{1 \le i \le \Gamma(t)} |u(Z_i; m) - u(Z_{i-1}; m)| \quad (7.28)$$

$$\le d_1 + d_2 \max_{1 \le i \le \Gamma(t)} |Z_i| \quad (7.29)$$

for suitable constructs d_1, d_2. Since $n^{-1} \sum_{i=1}^n Z_i^2 \to EZ^2 < \infty$ a.s. as $n \to \infty$, it is evident that

$$\max_{1 \le i \le n} Z_i^2/n \to 0 \text{ a.s.},$$

and thus

$$t^{-1/2} \max_{1 \le i \le \Gamma(t)} |Z_i| \to 0 \text{ a.s.}$$

as $t \to \infty$. It remains to verify that this latter limit also holds in expectation.

Now,

$$t^{-1/2} \max_{1 \le i \le \Gamma(t)} |Z_i| \le 1 + t^{-1} \max_{1 \le i \le \Gamma(t)} Z_i^2 \quad (7.30)$$

$$\le 1 + t^{-1} \sum_{i=1}^{\Gamma(t)} Z_i^2. \quad (7.31)$$

Since $t^{-1} \sum_{i=1}^{\Gamma(t)} Z_i^2 \to (\lambda + \mu) EZ_1^2 < \infty$ a.s. and

$$E \sum_{i=1}^{\Gamma(t)} Z_i^2 = EE[\sum_{i=1}^{\Gamma(t)} Z_i^2 | \Gamma(t)] = E\Gamma(t) \cdot EZ_1^2 = (\lambda + \mu) \cdot EZ_1^2 \cdot t,$$

it follows that $(t^{-1} \sum_{i=1}^{\Gamma(t)} Z_i^2 : t > 0)$ is a uniformly integrable family of r.v.'s, and consequently the same must be true of $(t^{-1/2} \max_{1 \le i \le \Gamma(t)} |Z_i| : t > 0)$.□

The above CLT suggests the approximation

$$(\rho^m t_m)^{-1/2} (\frac{\alpha(t_m; m)}{\alpha(m)} - 1) \overset{\mathcal{D}}{\approx} (\frac{2(1+\rho)}{\mu(1-\rho)^2})^{1/2} N(0,1)$$

for m large, provided that we choose the time horizon t_m so that $t_m \gg \rho^{-m}$. (A rigorous proof of this result would follow an argument similar to that used in establishing Theorem 7.3.3.)

We conclude this section with a brief discussion of the corresponding computation for the workload process $(W(t) : t \ge 0)$ associated with the M/M/1 queue. Because $(W(t) : t \ge 0)$ is neither a continuous-time Markov chain nor a diffusion, developing a martingale representation like (7.12) (or

that involving $M_m(\cdot)$) for the tail probability estimator is not straightforward. Instead, we illustrate here a different approach, involving the use of regeneration.

To precisely define $W(\cdot)$, let $N = (N(t) : t \ge 0)$ be a Poisson process running at rate $\lambda > 0$ and let $V = (V_n : n \ge 1)$ be an independent sequence of i.i.d. exponential(μ) r.v.'s, with $\mu > \lambda$. Set

$$S(t) = W(0) + \sum_{i=1}^{N(t)} V_i - t.$$

Then, assuming $W(0) = 0$, $W(\cdot)$ may be represented in terms of $S(\cdot)$ as

$$W(t) = S(t) - \min_{0 \le u \le t} S(u).$$

It is well known that $W(t) \Rightarrow W(\infty)$ as $t \to \infty$, where $P(W(\infty) \ge y) = \rho \exp(-(\mu - \lambda)y) \triangleq \alpha(y)$ for $y > 0$. The obvious nonparametric estimator for the tail probability $\alpha(y)$ is

$$\alpha(t; y) = \frac{1}{t} \int_0^t I(W(s) \ge y) ds.$$

The key computation in developing an understanding of the asymptotics of $\alpha(t; y)$ is the calculation of the time-average variance constant $\sigma_2^2(y)$ appearing in the CLT

$$t^{1/2}(\alpha(t; y) - \alpha(y)) \Rightarrow \sigma_2(y) N(0,1)$$

as $t \to \infty$. Noting that the origin is a regeneration state for $(W(t) : t \ge 0)$, the regenerative CLT expresses $\sigma_2^2(y)$ as

$$\sigma_2^2(y) = E_0 R^2 / E_0 T_0$$

where $E_x(\cdot) = E(\cdot | W(0) = x), T_x = \inf\{t \ge 0 : W(t) = x, W(t-) \ne x\}$, and $R = \int_0^{T_0} [I(W(t) \ge y) - \alpha(y)] dt$. Given that $(1 - \rho) = P(W(\infty) = 0) = \lambda^{-1}/E_0 T_0$, it is evident that $\sigma_2^2(y)$ can be computed via successive differentiation of the moment generating function of R. Set $a = -\alpha(y)$, $b = 1 - \alpha(y)$, and observe that conditioning on the time of the first jump yields the identity

$$E_0 \exp(\theta R) = \frac{\lambda}{\lambda - \theta a} \int_0^\infty E_x \exp(\theta R) \mu e^{-\mu x} dx.$$

The key to computing $\sigma_2^2(y)$ is therefore the calculation of $E_x \exp(\theta R)$ for $x > 0$. Note, first of all, that for all $x \ge y$, the absence of downward jumps in $W(\cdot)$ guarantees that y will be hit prior to the origin, and consequently, the strong Markov property guarantees that

$$E_x \exp(\theta R) = E_x \exp(\theta b T_y) E_y \exp(\theta R) \quad (7.32)$$

$$= E_{x-y} \exp(\theta b T_0) E_y \exp(\theta R). \quad (7.33)$$

Now, since $S(\cdot)$ is a compound Poisson process with exponential jumps, it is easily verified that

$$M(t) = \exp(\kappa(S(t) - S(0)) - \kappa t(\lambda - \mu + \kappa)/(\mu - \kappa))$$

is a P_x-martingale for $\kappa < \mu$. Set $\theta = -\kappa(\lambda - \mu + \kappa)/(\mu - \kappa)$ and note that $\theta \leq \mu + \lambda - 2\sqrt{\lambda\mu} \triangleq \Lambda$. So, for $\theta \leq \Lambda$,

$$\kappa_1(\theta) = [(\theta + \mu - \lambda) - ((\lambda - \mu - \theta)^2 - 4\theta\mu)^{1/2}]/2$$

and

$$\kappa_2(\theta) = [(\theta + \mu - \lambda) + ((\lambda - \mu - \theta)^2 - 4\theta\mu)^{1/2}]/2$$

both satisfy $\theta = -\kappa(\lambda - \mu + \kappa)/(\mu - \kappa)$ and hence each of the processes

$$M_i(t) = \exp(\kappa_i(\theta)(S(t) - S(0)) + \theta t)$$

$(i = 1, 2)$ are P_x-martingales for $\theta \leq \Lambda$. For $0 < x < z$, let $\tilde{T}_z = T_z \wedge T_0$ and note that $W(t) = S(t)$ for $t \leq \tilde{T}_z$. The optional sampling theorem implies that

$$E_x M_i(\tilde{T}_z \wedge t) = 1$$

for $i = 1, 2$ and $t \geq 0$. But the memoryless structure of the exponential "overshoot" above level z allows us to re-write the above identity as

$$\begin{aligned} 1 \;=\; & \exp(-\kappa_i(\theta)x)E_x[\exp(\theta T_0); T_0 < T_z \wedge t] \\ & + \exp(\kappa_i(\theta)(z - x))(\mu/(\mu - x_i(\theta))E_x[\exp(\theta T_z); T_z < T_0 \wedge t] \\ & + E_x[\exp(\kappa_i(\theta)(S(t) - S(0)) + \theta t); T_0 \wedge T_z > t] \end{aligned} \tag{7.34}$$

For $\theta \leq 0$, we can use the dominated convergence theorem and the fact that $|S(t) - S(0)| \leq z$ on $\{\tilde{T}_z > t\}$ to obtain the identity

$$\begin{aligned} \exp(\kappa_i(\theta)x) \;=\; & E_x[\exp(\theta T_0); T_0 < T_z] \\ & + \mu E_x[\exp(\theta T_z); T_z < T_0] \exp(\kappa_i(\theta)z)/(\mu - \kappa_i(\theta)) \end{aligned} \tag{7.35}$$

for $i = 1, 2$. For $\theta > 0$, Stein's lemma (see Feller [5], p. 601) guarantees that $E_x \exp(\theta\tilde{T}_z) < \infty$ for θ in a neighborhood of the origin. For such θ's, the dominated convergence theorem ensures that the third term on the right-hand side of (7.34) converges to zero as $t \to \infty$. The monotone convergence then applies to the first two terms, thereby verifying (7.35) for positive θ's in a neighborhood of the origin.

Since (7.35) holds for both $i = 1$ and $i = 2$, we have two linear equations in the two unknowns $E_x[\exp(\theta T_0); T_0 < T_z]$ and $E_x[\exp(\theta T_z); T_z < T_0]$, from which the two unknowns may be computed. Now, for $0 < x \leq y$,

$$\begin{aligned} E_x \exp(\theta R) \;=\; & E_x[\exp(\theta a T_0); T_0 < T_y] \\ & + E_x[\exp(\theta a T_y); T_y < T_0] \cdot \mu \int_0^\infty e^{-\mu r} E_{y+r} \exp(\theta R) dr. \end{aligned} \tag{7.36}$$

But $E_{y+r} \exp(\theta R) = E_r \exp(\theta b T_0) E_y \exp(\theta R)$ for $r \geq 0$. The quantity $E_r \exp(\theta b T_0)$ is, by monotone convergence, the limit of $E_r[\exp(\theta b T_0); T_0 < T_z]$ as $z \to \infty$, and is therefore calculable. Setting $x = y$ in (7.36) then leaves (7.36) as an equation in one unknown, namely $E_y \exp(\theta R)$, from which $E_x \exp(\theta R)$ can be obtained for all $x \geq 0$.

Thus, the regenerative approach outlined here provides, in principle, a closed form for the Laplace transform of R, and the potential to analytically compute its distribution.

7.5 General Theory for the GI/G/1 Queue

In Sections 7.3 and 7.4, we computed very explicit asymptotics for RBM and the M/M/1 queue. These asymptotics provided insight into the asymptotic efficiency of our proposed nonparametric estimator for the (extreme) tail probabilities associated with the single server queue. In this section, we develop some general theory for the GI/G/1 queue. As one might expect, however, our resulting asymptotic theory is somewhat less explicit than that presented in Sections 7.3 and 7.4.

We shall focus here on the waiting time sequence $W = (W_n : n \geq 0)$ of the GI/G/1 queue. It is to be expected that similar theory may be developed for the workload process and the queue-length process of the GI/G/1 queue, using methods similar to those we will exploit in the waiting time context. Assuming that $W_0 = 0$, W may be constructed from a random walk $(S_n : n \geq 0)$ via the identity

$$W_n = S_n - \min_{0 \leq k \leq n} S_k,$$

for $n \geq 0$. Letting $X_n = S_n - S_{n-1}$ be the n'th increment of the random walk, we assume that $(X_n : n \geq 1)$ is i.i.d. with $EX_1 < 0$; this "negative drift" condition on $(S_n : n \geq 0)$ is, of course, merely an assertion that the traffic intensity of the queue is strictly less than one.

Under the above "negative drift" condition, it is well known that

$$W_n \Rightarrow W_\infty$$

as $n \to \infty$. Furthermore, under certain conditions on the X_i's, it is possible to give a fairly precise characterization of the tail behavior of W_∞. Specifically, suppose that the X_i's are non-lattice r.v.'s and let $\varphi(\theta) = E \exp(\theta X_i)$ be the moment generating function of X_i. Assume that there exists a positive root θ^* to the equation $\varphi(\theta) = 1$, and assume further that $\varphi(\cdot)$ converges in a neighborhood of θ^*. Then, the Cramér-Lundberg approximation (see, for example, p. 269 of Asmussen [1]) states that

$$P(W_\infty \geq x) \sim a \exp(-\theta^* x) \tag{7.37}$$

as $x \to \infty$, for some positive constant a. The natural nonparametric estimator for $\alpha(y) \triangleq P(W(\infty) \geq y)$ is

$$\alpha(n;y) = \frac{1}{n}\sum_{j=0}^{n-1} I(W_j \geq y).$$

In view of (7.37), the theory developed thus far in this paper suggests that in order to obtain (relatively) accurate estimates of $\alpha(y)$, one must take the time horizon n large enough so that $n \gg \exp(\theta^* y)$. The remainder of this section is devoted to verifying this result rigorously.

Our argument takes advantage of the regenerative structure of W. In particular, the origin acts as a regenerative state for W. Let $T(0) = 0$, $T(n) = \inf\{m > T(n-1) : W_m = 0\}$ for $n \geq 1$, and set $\tau_i = T(i) - T(i-1)$. Put $\ell(n) = \max\{k \geq 0 : T(k) \leq n\}$, so that $\ell(n)$ counts the number of completed regenerative cycles to occur in $[0, n]$. If we let

$$\chi_j(y) = \sum_{k=T(j-1)}^{T(j)-1} (I(W_k \geq y) - \alpha(y)),$$

then $(\chi_j(y) : j \geq 1)$ is a sequence of i.i.d. random variables; furthermore, it is a standard fact of regenerative process theory that $E\chi_j(y) = 0$. (Unless otherwise stated, all expectations and probabilities in this section are computed conditionally on $W_0 = S_0 = 0$.) In addition, we may write the relative error for our nonparametric estimator in the form

$$\alpha(n;y)/\alpha(y) - 1 = n^{-1}\sum_{j=1}^{\ell(n)} \chi_j(y)/\alpha(y) + n^{-1}R(n;y) \tag{7.38}$$

where $R(n;y)$ is the "remainder term" given by

$$R(n;y) = \sum_{j=T(\ell(n))}^{n-1} [I(W_j \geq y) - \alpha(y)]/\alpha(y).$$

In view of the representation (7.38), it seems clear that the variance of $\chi_j(y)$ as $y \to \infty$ will play an important role in our subsequent analysis.

The key to this asymptotic analysis is a "change-of-measure" argument similar to that used to obtain the Cramér-Lundberg approximation itself. We start by observing that W is identical to the random walk up to time $\tau \triangleq \tau_1$, so that

$$\chi_1(y) = \sum_{j=0}^{\tau-1} (I(S_j \geq y) - \alpha(y))$$

where $\tau = \inf\{n \geq 1 : S_n \leq 0\}$. We now define a change-of-measure on the paths of the random walk. To be precise, for $y > 0$, let $T_y = \inf\{n \geq 0 : S_n \geq y\}$, and define the measure

$$\begin{aligned} &P^*(S_1 \in dx_1, \ldots, S_n \in dx_n; T_y = j) \\ &\quad \triangleq \quad e^{\theta^* x_j} P(S_1 \in dx_1, \ldots S_n \in dx_n; T_y = j), \end{aligned}$$

with j, n arbitrary integers satisfying $j \leq n$. If $E^*(\cdot)$ denotes the expectation corresponding to P^*, the above identity implies that

$$E^*[\xi; T_y < \infty] = E[\xi \exp(\theta^* S_{T_y}); T_y < \infty]$$

for all non-negative r.v.'s ξ. It immediately follows that

$$E[\xi; T_y < \infty] = E^*[\xi \exp(-\theta^* S_{T_y}); T_y < \infty].$$

Consequently,

$$\begin{aligned} E[(\sum_{j=0}^{\tau-1} I(S_j \geq y))^2] &= E^*[(\sum_{j=0}^{\tau-1} I(S_j \geq y))^2 \exp(-\theta^* S_{T_y}); T_y < \infty] \\ &= E^*[(\sum_{j=T_y}^{\tau-1} I(S_j \geq y))^2 \exp(-\theta^* S_{T_y}); T_y < \tau < \infty] \\ &= E^*[g(S_{T_y} - y; y) \exp(-\theta^* S_{T_y}); T_y < \tau < \infty], \end{aligned}$$

where $g(x;y) = E[(\sum_{j=0}^{T_{-y}} I(S_j \geq 0))^2 | S_0 = x]$ and $T_{-y} = \inf\{n \geq 0 : S_n \leq -y\}$ for $-y \leq 0$. It is a standard fact about such changes-of-measure that, conditional on $\{T_y \geq j\}$, P^* makes $(X_i : 1 \leq i \leq j)$ independent, with common distribution $\exp(\theta^* x)P(X_1 \in dx)$, whereas conditional on $\{T_y = j\}$, the X_i's are independent for $i > j$, with common distribution $P(X_1 \in dx)$. The "twisted" distribution $\exp(\theta^* x)P(X_1 \in dx)$ gives X_1 positive mean (equal to $\varphi'(\theta^*)$) so that the random walk $(S_n : n \geq 0)$ initially has positive drift under P^*. Consequently, $T_y < \infty$ P^* a.s. Subsequent to T_y, the random walk evolves according to its original dynamics (having negative drift) so that $\tau < \infty$ P^* a.s. Hence,

$$E[(\sum_{j=0}^{\tau-1} I(S_j \geq y))^2] = \exp(-\theta^* y) E^*[g(S_{T_y} - y; y) \exp(-\theta^*(S_{T_y} - y)); T_y < \tau]. \tag{7.39}$$

Now the monotone convergence theorem guarantees that

$$g(x;y) \nearrow g(x) \triangleq E[(\sum_{j=0}^{\infty} I(S_j \geq 0))^2 | S_0 = x]$$

as $y \to \infty$. In addition, $I(T_y < \tau) \downarrow I(\tau = +\infty)$ and, under our non-lattice hypothesis on the X_i's, the "overshoot" $\Psi(y) \triangleq S_{T_y} - y$ converges weakly to

a limit r.v. $\Psi(\infty)$ as $y \to \infty$ (see p.168 of Asmussen [1]). These observations suggest the approximation

$$E[(\sum_{j=0}^{\tau-1} I(S_j \geq y))^2] \tag{7.40}$$
$$\approx \exp(-\theta^* y) E^*[g(\Psi(\infty)) \exp(-\theta^* \Psi(\infty))] P^*[\tau = +\infty].$$

Because it seems reasonable to expect

$$E\chi_1(y)^2 (= \text{var } \chi_1(y)) \approx E[(\sum_{j=0}^{\tau-1} I(S_j \geq y))^2]$$

for y large, this analysis suggests the following theorem.

Theorem 7.5.1 Suppose that $(X_i : i \geq 1)$ is a sequence of bounded i.i.d. non-lattice r.v.'s. If $P(X_1 > 0) > 0$ and $EX_1 < 0$, then

$$E\chi_1^2(y) \sim \exp(-\theta^* y) E^*[g(\Psi(\infty)) \exp(-\theta^* \Psi(\infty))] \cdot P^*[\tau = +\infty] \tag{7.41}$$

as $y \to \infty$.

Proof. We start by making rigorous our approximation (7.40). Since $P(X_1 > 0) > 0$, it follows that $E \exp(\theta X_1) = \varphi(\theta) \to +\infty$ as $y \to \infty$. Since $\varphi'(0) = EX_1 < 0$, the convexity of $\varphi(\cdot)$ guarantees existence of a unique $\theta^* > 0$ such that $\varphi(\theta^*) = \varphi(0) = 1$. Furthermore, the boundedness of the X_i's implies that $\varphi(\cdot)$ is everywhere finite-valued. Hence, (7.39) is valid and $\Psi(y) \Rightarrow \Psi(\infty)$ as $y \to \infty$.

We turn now to the analysis of $g(x; y)$. We first wish to prove that $g(\Psi(y); y)$ is a uniformly bounded sequence of r.v.'s under P^* so that, in particular, $g(\Psi(\infty))$ is bounded; a sketch of the proof follows. Observe that

$$\sum_{j=0}^{\infty} I(S_j \geq 0) \leq T_{-1} + \sum_{i=1}^{\beta} T_{-1,i}, \tag{7.42}$$

where β is the number of times that the random walk proceeds from below level -1 to above the origin, and $T_{-1,i}$ is the time required, on the i'th excursion above the origin, to go below level -1 again. Each time the random walk goes below level -1, there is a probability at most equal to $P_0(T_1 < \infty) < 1$ of an additional excursion. Consequently, $P(\beta > k) \leq P_0(T_1 < \infty)^k$ so β has moments of all orders. In addition, because the X_i's are bounded and living in $[-K, K]$, say,

$$\begin{aligned} P(T_{-1,i} > t) &\leq P(T_{-1} \geq t | S_0 = K) \\ &= P(\tau \geq t | S_0 = K + 1). \end{aligned}$$

Set $\psi(\theta) = \log \varphi(\theta)$ and observe that $\exp(\theta S_n - \psi(\theta) n)$ is a martingale. Hence, the optional sampling theorem implies that

$$E[\exp(\theta S_{\tau \wedge n} - \psi(\theta)(\tau \wedge n)) | S_0 = x] = \exp(\theta x).$$

Consequently,

$$E[\exp(\theta S_\tau - \psi(\theta)\tau) I(\tau \leq n) | S_0 = x] \leq \exp(\theta x),$$

and the monotone convergence theorem then implies that for $\theta > 0$,

$$\exp(-\theta K) E[\exp(-\psi(\theta)\tau) | S_0 = x] \leq E[\exp(\theta S_\tau - \psi(\theta)\tau) | S_0 = x] \leq \exp(\theta x).$$

Thus, if we choose $\theta > 0$ so that $\psi(\theta) < 0$, we may conclude that the $T_{-1,i}$'s have uniformly bounded exponential moments, as does T_1 (provided that S_0 lies in a compact set). We conclude from (7.42) that $g(\cdot)$ is bounded on compact sets. But $g(\Psi(y); y) \leq g(\Psi(y)) \leq \sup\{g(x) : 0 \leq x \leq K\}$, and hence the $g(\Psi(y); y)$'s are a family of uniformly bounded r.v.'s.

Our next task is to prove that we can replace $g(\Psi(y); y)$ in (7.41) by $g(\Psi(y))$. This can be done provided that we show that

$$g(\Psi(y); y) - g(\Psi(y)) \to 0$$

P^* a.s. as $y \to \infty$. Because of the boundedness of $\Psi(y)$ (by K), it suffices to prove that

$$\sup_{0 \leq x \leq K} |g(x; y) - g(x)| \to 0 \tag{7.43}$$

as $y \to \infty$. Recall that

$$\begin{aligned} g(x) &= E[(\sum_{k=0}^{\infty} I(S_k \geq 0))^2 | S_0 = x] \\ &= g(x; y) + 2E[(\sum_{j=0}^{T_{-y}} I(S_j \geq 0)) \cdot (\sum_{j=T_{-y}+1}^{\infty} I(S_j \geq 0)) | S_0 = x] \\ &\quad + E[(\sum_{j=T_{-y}+1}^{\infty} I(S_j \geq 0))^2 | S_0 = x]. \end{aligned}$$

Relation (7.43) will therefore follow from the Cauchy-Schwarz inequality if we show that

$$\sup_{0 \leq x \leq K} E[(\sum_{j=T_{-y}+1}^{\infty} I(S_j \geq 0))^2 | S_0 = x] \to 0$$

as $y \to \infty$. Note that the above r.v. vanishes unless the random walk exceeds level 0 subsequent to time T_{-y}. Applying the strong Markov property at

time T_{-y} and taking advantage of the stochastic monotonicity and spatial homogeneity of random walk, we can bound the above by

$$E[(\sum_{j=0}^{\infty} I(S_j \geq y))^2; T_y < \infty].$$

Applying the strong Markov property, at time T_y, to the above expectation, we obtain

$$E[g(\Psi(y)); T_y < \infty].$$

But $g(\Psi(y))$ is bounded in y and $P(T_y < \infty) \to 0$ as $y \to \infty$, proving (7.43).

The proof of (7.40) is therefore complete if we can show that

$$\begin{aligned} & E^*[g(\Psi(y)) \exp(-\theta^* \Psi(y)) I(T_y < \tau)] \\ \to \quad & E^*[g(\Psi(\infty)) \exp(-\theta^* \Psi(\infty))] P^*[\tau = +\infty]. \end{aligned} \tag{7.44}$$

To prove the above limit relation, we first show that

$$\exp(-\theta^* \Psi(y)) g(\Psi(y)) \Rightarrow \exp(-\Psi(\infty)) g(\Psi(\infty))$$

as $y \to \infty$; this will follow if we establish that $g(\cdot)$ is continuous a.s. at $\Psi(\infty)$. Since X_1 is non-lattice, it is evident that $\Psi(\infty)$ is a continuous r.v. (see p. 168 of Asmussen [1]), and therefore it suffices to prove that $g(\cdot)$ has at most countably many discontinuities. But the stochastic monotonicity of random walk implies that $g(\cdot)$ is non-decreasing. Such a function can have at most countably many discontinuities, and thus it follows that $\exp(-\theta^* \Psi(y)) g(\Psi(y)) \Rightarrow \exp(-\Psi(\infty)) g(\Psi(\infty))$ as $y \to \infty$.

Now the left-hand side of (7.44) can be written as

$$E^*[g(\Psi(y)) \exp(-\theta^* \Psi(y))] - E^*[g(\Psi(y)) \exp(-\theta^* \Psi(y)) I(\tau < T_y)]. \tag{7.45}$$

Denote the first expectation in (7.45) as $\tilde{g}(y)$. Since $(g(\Psi(y)) \exp(-\theta^* \Psi(y)) : y > 0)$ is bounded in y and converges weakly, it is evident that $\tilde{g}(y)$ converges to $E^* g(\Psi(\infty)) \exp(-\theta^* \Psi(\infty))$ as $y \to \infty$. As for the second expectation in (7.45), applying the strong Markov property at time τ allows us to re-write it as

$$E^*[\tilde{g}(y - S_\tau) I(\tau < T_y)].$$

Since $\tilde{g}(y - S_\tau) \to E^* g(\Psi(\infty)) \exp(-\theta^* \Psi(\infty))$ as $y \to \infty$ and $I(\tau < T_y) \to I(\tau < \infty)$ P^* a.s. as $y \to \infty$, (7.44) is therefore proved. Our proof thus far yields the conclusion that

$$\begin{aligned} & E[(\sum_{j=0}^{\tau-1} I(S_j \geq y))^2] \\ \sim \quad & \exp(-\theta^* y) E^*[g(\Psi(\infty)) \exp(-\theta^* \Psi(\infty))] \cdot P^*[\tau = +\infty] \end{aligned} \tag{7.46}$$

as $y \to \infty$. But

$$\begin{aligned} E\chi_1^2(y) \quad &= \quad E[(\sum_{j=0}^{\tau-1} I(S_j \geq y))^2] \\ & \quad -2\alpha(y) E[\tau \sum_{j=0}^{\tau-1} I(S_j \geq y)] + \alpha^2(y) E\tau^2. \end{aligned} \tag{7.47}$$

We earlier showed that τ has moments of all orders. Since

$$\tau \sum_{j=0}^{\tau-1} I(S_j \geq y) \leq \tau^2,$$

it is evident, via an application of the dominated convergence theorem, that

$$E[\tau \cdot \sum_{j=0}^{\tau-1} I(S_j \geq y)] \to 0$$

as $y \to \infty$, as does $\alpha(y) E\tau^2$. Relations (7.46) and (7.47), together with the Cramér-Lundberg approximation for $\alpha(y)$, then prove the theorem. □

Theorem 7.5.1 provides a precise asymptotic for the behavior of $E\chi_1^2(y)$. This, in turn, sheds light on the asymptotic behavior of the relative error $\alpha(n; y)/\alpha(y) - 1$. The CLT for regenerative processes guarantees that for y fixed,

$$n^{1/2}(\frac{\alpha(n; y)}{\alpha(y)} - 1) \Rightarrow (\frac{E\chi_1^2(y)}{\alpha^2(y) E\tau})^{1/2} N(0, 1)$$

as $n \to \infty$. Relation (7.48) suggests that

$$\frac{\alpha(n; y)}{\alpha(y)} - 1 \overset{\mathcal{D}}{\approx} (\frac{E\chi_1^2(y)}{n\alpha^2(y) E\tau})^{1/2} N(0, 1) \tag{7.48}$$

for n large. But, according to (7.37) and Theorem 7.5.1,

$$\frac{E\chi_1^2(y)}{\alpha^2(y) E\tau} \sim \exp(\theta^* y) r \tag{7.49}$$

as $y \to \infty$, where

$$r \triangleq \frac{E^*[g(\Psi(\infty)) \exp(-\theta^* \Psi(\infty))] P^*[\tau = +\infty]}{(E^*[\exp(-\theta^* \Psi(\infty))])^2 E\tau}.$$

(See Asmussen [1], p. 269, for the exact form of the constant a appearing in (7.37).) Thus (7.48) and (7.49) together suggest the approximation

$$\frac{\alpha(n_y; y)}{\alpha(y)} - 1 \overset{\mathcal{D}}{\approx} (\frac{\exp(\theta^* y)}{n_y} r)^{1/2} N(0, 1),$$

provided $n_y \gg \exp(\theta^* y)$. Our final theorem of this section makes this approximation rigorous.

Theorem 7.5.2 Assume the conditions of Theorem 7.5.1 and suppose that $n_y \exp(-\theta^* y) \to +\infty$ as $y \to \infty$. Then,

$$(n_y \exp(-\theta^* y))^{1/2}(\frac{\alpha(n_y; y)}{\alpha(y)} - 1) \Rightarrow r^{1/2}N(0,1)$$

as $y \to \infty$.

Proof. We return to the representation (7.38) for the relative error. Our first task is to deal with the remainder term. We wish to show that

$$(n_y \exp(-\theta^* y))^{1/2} R(n_y; y)/n_y \Rightarrow 0 \tag{7.50}$$

as $y \to \infty$. Let $V_j(y) = \sum_{k=T(j-1)}^{T(j)-1} I(W_k \geq y)$ and observe that

$$|R(n_y; y)/n_y| \leq \max_{1 \leq j \leq n_y} V_j(y)/n_y\alpha(y) + \max_{1 \leq j \leq n_y} \tau_j/n_y.$$

Because $E\tau_j^2 < \infty$ (see the proof of Theorem 7.5.1), it follows that

$$\max_{1 \leq j \leq n_y} \tau_j/n_y^{1/2} \to 0 \text{ a.s.}$$

as $y \to \infty$. To finish the proof of (7.50), we therefore need to show that

$$\max_{1 \leq j \leq n_y} V_j(y)/(n_y\alpha(y))^{1/2} \Rightarrow 0.$$

We accomplish this task by showing that the sequence converges to zero in mean square error. Note that

$$\frac{1}{n_y\alpha(y)} E \max_{1 \leq j \leq n_y} V_j(y)^2 = \frac{1}{n_y\alpha(y)} \int_0^\infty P(\max_{1 \leq j \leq n_y} V_j(y)^2 > t)dt. \tag{7.51}$$

For each fixed t, $P(\max_{1 \leq j \leq n_y} V_j(y)^2 > t)/n_y\alpha(y) \to 0$ as $y \to \infty$. On the other hand,

$$\begin{aligned} P(\max_{1 \leq j \leq n_y} V_j(y)^2 > t)/n_y\alpha(y) &\leq P(V_1(y)^2 > t)/\alpha(y) \\ &\leq EV_1(y)^4/t^2\alpha(y). \end{aligned}$$

The proof of Theorem 7.5.1 establishes that $EV_1(y)^2 = O(\alpha(y))$ as $y \to \infty$. A similar change-of-measure argument proves that $EV_1(y)^p = O(\alpha(y))$ as $y \to \infty$, for $p > 0$. Consequently, we can apply the dominated convergence theorem to (7.51), thereby obtaining the asymptotic negligibility of the remainder term.

The proof is therefore complete if we can show that

$$(n_y\alpha(y))^{-1/2} \sum_{j=1}^{\ell(n_y)} \chi_j(y) \Rightarrow (ar)^{1/2}N(0,1)$$

as $y \to \infty$, where a is the constant appearing in (7.37). This, in turn, follows if we can prove that

$$(n_y\alpha(y))^{-1/2}(\sum_{j=1}^{\ell(n_y)} \chi_j(y) - \sum_{j=1}^{\lfloor \ell n_y \rfloor} \chi_j(y)) \Rightarrow 0 \tag{7.52}$$

and

$$(n_y\alpha(y))^{-1/2} \sum_{j=1}^{\lfloor \ell n_y \rfloor} \chi_j(y) \Rightarrow (ar)^{1/2}N(0,1) \tag{7.53}$$

as $y \to \infty$, where $\ell \stackrel{\Delta}{=} 1/E\tau$. We note that $\ell(n)/n \to \ell$ a.s. as $n \to \infty$, so that $I(|\ell(n)/n - \ell| < \epsilon) \to 1$ a.s. as $n \to \infty$ for any $\epsilon > 0$. To obtain (7.42), fix $\epsilon > 0$ and observe that on $\{|\ell(n)/n - \ell| < \epsilon\}$,

$$|\sum_{j=1}^{\ell(n_y)} \chi_j(y) - \sum_{j=1}^{\lfloor \ell n_y \rfloor} \chi_j(y)| \leq \max_{|k| \leq \lceil n_y\epsilon \rceil} |\sum_{j=\lceil \ell n_y \rceil}^{\ell n_y + k} \chi_j(y)|.$$

Kolmogorov's maximal inequality implies, however, that

$$P(\max_{k \leq \lceil n_y\epsilon \rceil} |\sum_{j=\lfloor \ell n_y \rfloor}^{\lfloor \ell n_y \rfloor + k} \chi_j(y)| > x(n_y\alpha(y))^{1/2}) \leq 2\lceil n_y\epsilon \rceil E\chi_1^2(y)/(x^2 n_y\alpha(y)).$$

Theorem 7.5.1 and the fact that ϵ may be made arbitrarily small then yield (7.52).

To prove (7.53), we use the Lindeberg-Feller theorem for triangular arrays of i.i.d. r.v.'s; see, for example, p. 205 of Chung [3]. The triangular array that arises here is trivially holospoudic. As for the Lindeberg condition, this requires showing that

$$E[\chi_1^2(y)/\alpha(y); \chi_1^2(y) > x\alpha(y)n_y] \to 0 \tag{7.54}$$

as $y \to \infty$. But the above expectation is bounded by $E\chi_1^4(y)/(x\alpha^2(y)n_y)$. It is easily seen that

$$E\chi_1^4(y) = EV_1^4(y) + o(\alpha(y)) = O(\alpha(y)),$$

from which (7.54) follows. Application of the Lindeberg-Feller CLT gives (7.53), proving the theorem. □

Theorem 7.5.2 provides a precise rate of convergence for the relative error of our nonparametric tail probability estimator. Thus, we have established, in significant generality, the relative magnitude of the time horizon over which a queue needs to be observed, in order that one obtain (relatively) accurate estimates of the tail probability.

7.6 Some Concluding Remarks

In the preceding sections of this paper, we analyzed a number of different queueing models, with the aim of producing as strong an argument as possible for the following claim:

> If a tail probability for a queue is to be estimated via the observed proportion of time that the corresponding performance measure is out in the tail, then, in order that the probability be (relatively) accurately estimated, it is necessary and sufficient that the time horizon over which the queue is observed be large relative to the reciprocal of the tail probability.

Given that queues typically exhibit non-negative autocorrelations, the necessity was covered by Proposition 7.2.1. As for sufficiency, an asymptotic theory in which the various limiting constants were explicitly identified was developed for both the M/M/1 queue and RBM. In addition, we verified the sufficiency of such a time horizon for uniformly mixing processes and for the waiting time sequence of the single-server queue. Because tail probabilities for queues typically decay exponentially fast, the above claim establishes that the time interval over which a queue needs to be observed in order to (relatively) accurately estimate a tail probability grows exponentially rapidly in the tail parameter. This implies that, if such a nonparametric estimation approach is adapted, then the amount of data which one must collect in order to accurately estimate the loss probability for a buffered queue is potentially enormous.

This suggests that perhaps one should explore alternative means, for estimating such tail probabilities. In particular, suppose that the queue can be well described by a (finite-dimensional) parametric model. This would typically occur when, for example, enough is known about the behavior of the input sources to the queue that a parameterized stochastic model describing the input can be developed. In such a setting, the parametric estimator of the tail probability associated with level y typically takes the form $\gamma(\hat{\theta}; y)$, where $\hat{\theta}$ is an estimator of the parametric vector θ describing (for example) the input model, and $\gamma(\theta; y)$ is the tail probability associated with level y under parameter θ. Torres and Glynn [12] show that, in contrast to the estimators described in the current paper, the tail probability associated with tail parameter y can be (relatively) accurately estimated provided that the time interval over which the queue is observed is large relative to y^2. Thus, although tail estimation continues to get harder in the parametric context, the rate at which it gets harder is (much) better behaved than in the nonparametric setting. Of course, the disadvantage of the parametric approach is the need to fit an appropriate parametric family to the (for example) input sources to the system.

When this parametric methodology is applied, it is clearly necessary to evaluate $\gamma(\theta; y)$ at $\theta = \hat{\theta}$. Furthermore, if confidence intervals for the tail probability are required, the derivative $\gamma'(\hat{\theta}; y)$ (with respect to θ) needs to be computed; while such computations are trivial for simple parametric models like the M/M/1 queue, they are decidedly less so for more complex models, such as queues in which the arrival process is described by a Markov modulated Poisson process. For more complex parametric models, simulation may be needed to compute $\gamma(\hat{\theta}; y)$ and/or $\gamma'(\hat{\theta}; y)$.

The results developed in this paper have important implications for such simulations. The conventional simulation-based estimator for a tail probability is precisely the non-parametric tail estimator considered in this paper, namely the proportion of time that the simulated queue lies out in the tail. The theory developed in this paper makes clear the enormous computation time demanded by such estimators. This suggests that one consider alternative means of computing such tail probabilities. Fortunately, this is a problem to which the simulation-based efficiency improvement technique known as importance sampling is ideally suited. Chang et al. [2] and Glynn [7] show that when this method is applied to the tail probability estimation problem, the computational difficulty increases only linearly in the tail parameter y. (The variance per run is roughly constant in y; however, the computation time per simulated run increases linearly in y because this growth describes the time required for the queue, under the change-of-measure, to hit level y.) This is to be compared with the exponential growth in difficulty associated with the conventional simulation-based estimator. Thus, enormous improvements in efficiency are to be afforded by using importance sampling in this context.

Of course, if parametric tail probability estimators are to be computed in "real time", one would ideally pre-compute $\gamma(\cdot; y)$ and $\gamma'(\cdot; y)$ off-line and store the corresponding function values in memory. This would avoid the need to (possibly) simulate the queue in real time in order to evaluate $\gamma(\hat{\theta}; y)$ and $\gamma'(\hat{\theta}; y)$.

Acknowledgments: This research was supported by the Army Research Office under contract no. DAAL03-91-G-0319.

7.7 References

[1] Asmussen, S., *Applied Probability and Queues,* John Wiley & Sons, New York, 1987.

[2] Chang, C.S., Heidelberger, P., Juneja, S., and Shahabuddin, P., Effective bandwidth and fast simulation of ATM intree networks, *Performance Evaluation,* **20**, 45–65, 1994.

[3] Chung, K.L., *A Course in Probability Theory,* Academic Press, New York, 1974.

[4] CHUNG, K.L. AND WILLIAMS R.G., *Stochastic Integration,* Birkhäuser, Boston, 1983.

[5] ETHIER, S.N. AND KURTZ, T.G., *Markov Processes: Characterization and Convergence,* John Wiley & Sons, New York, 1986.

[6] FELLER, W., *An Introduction to Probability Theory and its Applications, Volume 2,* John Wiley & Sons, New York, 1971.

[7] GLYNN, P.W., Efficiency Improvement Techniques, *Annals of Operations Research,* **53**, 175–197, 1994.

[8] GLYNN, P.W., Diffusion Approximations in *Stochastic Models, Handbook of OR and MS Volume 2,* (D. Heyman and M. Sobel, eds.,) Elsevier Science Publishers, 1990.

[9] HARRISON, J.M., *Brownian Motion and Stochastic Flow Systems,* John Wiley & Sons, New York, 1985.

[10] KARLIN, S. AND TAYLOR, H.M., *A Second Course in Stochastic Processes,* Academic Press, New York, 1981.

[11] STOYAN, D., *Comparison Methods for Queues and Other Stochastic Models,* John Wiley & Sons, New York, 1983.

[12] TORRES, M. AND GLYNN, P.W., Parametric estimation of tail probabilties for the single-server queue, Technical Report, Department of Operations Research, Stanford University, Stanford, CA, 1996.

8

Rational Interpolation for Rare Event Probabilities

Wei-Bo Gong and Soracha Nananukul

ABSTRACT We propose to use rational interpolants to tackle some computationally complex performance analysis problems such as rare-event probabilities in stochastic networks. Our main example is the computation of the cell loss probabilities in ATM multiplexers. The basic idea is to use the values of the performance function when the system size is small, together with the asymptotic behaviour when the size is very large, to obtain a rational interpolant which can be used for medium or large systems. This approach involves the asymptotic analysis of the rare-event probability as a function of the system size, the convergence analysis of rational interpolants on the positive real line, and the quasi-Monte Carlo analysis of discrete event simulation.

8.1 Motivation: Padé Approximation for the GI/GI/1 Queue

The introduction of rational interpolation for evaluating rare event probabilities in [22] was motivated by the earlier work on the application of Padé approximants to single-server queues with renewal arrival processes [7]. To obtain the Padé approximants of a performance function we first need to obtain its MacLaurin series. This has been done for several systems [6, 7, 10, 23]. The basic idea is to expand the innermost part of the performance function (for example the k-fold convolution of the interarrival-time density functions) into a MacLaurin series. Then interchange the operations (integrations, expectations, and summations) so that the summation operation is in the outermost position. The interchanges of operations are usually justifiable using the dominated convergence theorem.

We first derive the MacLaurin series of the k-fold convolution of a function. Let $f(t)$ and $g(t)$ be defined on the real line and are 0 when $t < 0$. The convolution of $f(t)$ and $g(t)$ is denoted by $(f * g)(t)$, and the k-fold convolution of $f(t)$ is denoted by $f_{*k}(t)$, where $f_{*1}(t) \triangleq f(t)$. Assume that $\sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} t^i$ and $\sum_{i=0}^{\infty} \frac{g^{(i)}(0)}{j!} t^i$ converge at any finite $t \geq 0$. The convolu-