(1990), 579-588.

[32] Sadowsky, J.S. and Szpankowski, W., The probability of large queue length and waiting times in a heterogeneous multiserver queue, Part I: Tight limits, Manuscript, School of Elect. Eng. and Dept. of Comp. Sci., Purdue Univ., West Lafayette, IN 47907 USA (1992).

[33] Sadowsky, J.S. and Szpankowski, W., The probability of large queue length and waiting times in a heterogeneous multiserver queue, Part II: Positive recurrence and logarithmic limit, Manuscript, School of Elect. Eng. and Dept. of Comp. Sci., Purdue Univ., West Lafayette, IN 47907 USA (1992).

[34] Sadowsky, J., On the optimality and stability of exponential twisting in Monte Carlo simulation, *IEEE Trans. Inform. Theory* **IT-39** (1993), 119-128.

[35] Sadowsky, J., Monte Carlo estimation of large deviation probabilities, Manuscript, School of Elect. Eng. and Dept. of Comp. Sci., Purdue Univ., West Lafayette, IN 47907 USA (1994).

[36] Shwartz, A. and Weiss, A., Large deviation for performance analysis: queues, communication and computers, Manuscript, Dept. of Elect. Eng., Technion-IIT, Haifa 32000, Israel (1994).

[37] Stockmeyer, L.J., Computation complexity, *Handbooks in OR & MS* **3** (1992).

[38] Tsoucas, P., Rare events in series of queues, *J. of Appl. Prob.* (1992), 168-175.

Chapter 16

Parametric estimation of tail probabilities for the single-server queue[1]

Peter W. Glynn and Marcelo Torres

ABSTRACT In this chapter, we consider the question of how long the arrival process to the single-server queue needs to be observed in order to accurately estimate the long-run fraction of time that the workload exceeds y. We assume that the arrival process can be modeled parametrically. In such a parametric context, our results suggest that one typically needs to observe the arrival process over a time horizon that is large relative to y^2. This conclusion appears to hold regardless of whether the arrival process model exhibits complex dependencies or not.

CONTENTS

16.1 INTRODUCTION

Consider a packet-switched communications network, in which admission control of additional sources to the network is determined by constraints on the long-run fraction of time, which network switches expect to see packet buffer content that exceeds predetermined levels. This call acceptance criteria is natural when end-to-end network delay and packet loss probabilities are of principle importance in measuring network performance. In such a setting, one expects that the admissions control policy will need to either implicitly or explicitly estimate the long-run fraction of time that the packet buffer content is greater than a given level y. Unless a great deal is known about the source traffic to the network, such estimates will need to be based on real-time estimates of traffic.

In this chapter, we consider a simplified version of the above problem, in the hope that it offers some general insight into the issues likely to be encountered in developing such network estimators. We simplify the problem in three fundamental ways. Firstly, rather than dealing with a network, we consider only the traffic congestion as modeled by a single station with a constant rate server. Secondly, we approximate the behavior of the finite buffer workload process with that of the

[1]This research was supported by the Army Research Office under contract No. DAAL03-91-G-0319.

0-8493-8076-6/97/$0.00+$.50

infinite capacity analog. For instance, when considering the long-run fraction of packets lost due to buffer overflow, we replace the finite-buffer loss probability by the corresponding probability that the infinite-capacity workload process has a steady state workload greater than the finite system's buffer capacity; this probability is commonly used as a surrogate to the loss probability desired and is generally a good approximation when the buffer is (very) large relative to the utilization. Finally we assume, in this paper that the arrival process to the queue can be modeled parametrically.

In particular, Section 16.2 is concerned with estimation issues related to the tail of the steady-state workload distribution of the M/M/1 queue. The uncertainty in the source traffic to the queue is captured as uncertainty in the underlying Poisson arrival rate to the system. Thus, Section 16.2 discusses the question of how to construct estimators for the tail probabilities in such a queue, when the arrival rate is unknown and needs to be estimated from real-time traffic. In Section 16.3, we consider the workload tail estimation problem again, except that now the source traffic is modeled as a Markov-modulated fluid, with a parametrically determined generator. This extension is of some interest, in view of the fact that such Markov-modulated source models are widely used to characterize the complex dependencies exhibited by packet arrivals in modern communication networks; see for example [1, 2, 3, 14, 19]. Section 16.4 considers how our theory is affected by the traffic intensity. Finally, Section 16.5 describes open research problems.

The most important single qualitative conclusion reached in our analysis is that the amount of real-time traffic that one must observe in order to accurately estimate the steady-state probability that workload exceeds y must be large relative to y^2, when y is out in the extreme right tail. This behavior is exhibited in both our M/M/1 analysis and Markov-modulated analysis. While the amount of data that one needs to collect might appear to be large, it turns out that much more data needs to be collected if the source traffic is modeled non-parametrically. In particular, the amount of data collected must increase much faster than y^2 in such problems; see Glynn and Torres [10] for details. Thus, assuming parametric modeling of the source traffic is appropriate to the application considered, it can provide better performance predictions to the admission control entity during the real-time operation of a network.

We note that some of the theory developed in this paper also relates to the question of how "continuous" a queue is, as a function of the arrival process (see Theorem 16.3, for example). Thus, some of our results can be viewed as being in the same spirit as work previously published by, for example, Stoyan [20] and Borovkov [7].

16.2 PARAMETRIC ESTIMATION OF TAIL PROBABILITIES FOR THE M/M/1 QUEUE

In this section, we will offer an analysis of the basic estimation issues in the context of the simplest possible (interesting) single-server queueing model, namely the M/M/1 queue. To illustrate our basic theory, we choose to focus on the workload process $W = (W(t): t \geq 0)$ associated with the single-server queue.

To be more specific, let $N = (N(t): t \geq 0)$ be a Poisson process running at rate λ^* and let $V = (V_n: n \geq 1)$ be an independent sequence of i.i.d. exponential random variables having mean $1/\mu^*$. We can interpret $N(t)$ as the total number of customers to arrive to the queue in $[0,t]$, whereas V_n can be viewed as the processing time of the nth customer. Assuming the queue starts off with no work present, it follows that

$$W(t) = S(t) - \min_{0 \leq u \leq t} S(u),$$

where

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

If $\rho \equiv \lambda^*/\mu^*$, it is well known (see, for example, Asmussen [4], pg. 96) that

$$W(t) \Rightarrow W(\infty)$$

as $t \rightarrow \infty$, where

$$P(W(\infty) \geq y) = \rho \exp(-(\mu^* - \lambda^*)y) \tag{16.1}$$

for $y \geq 0$. Thus, if λ^* and μ^* are known, (16.1) offers a complete solution as to how to compute the steady-state probability that the workload in the M/M/1 queue exceeds y.

However, in many real-world settings, one may have imperfect knowledge of the statistical characteristics of the source that is generating arrivals to the system. In the M/M/1 queue context, this might translate into λ^* and/or μ^* being unknown *a priori*. To fix our ideas, let us suppose that only λ^* is unknown and that we wish to predict the long-run fraction of time that the workload will exceed y. By observing the arrival process N over $[0,t]$, we can estimate the arrival rate λ^* via the empirically observed rate

$$\lambda(t) \equiv \frac{1}{t} N(t).$$

($\lambda(t)$ is, in fact, the maximum likelihood estimator of λ^*.) Based on (16.1), it therefore appears reasonable to compute our prediction for the long-run fraction of time that the workload will exceed y via

$$\alpha(t;y) \equiv g(\lambda(t), y),$$

where

$$g(\lambda, y) = \frac{\lambda}{\mu^*} \exp(-(\mu^* - \lambda)y).$$

The asymptotic theory associated with maximum likelihood estimation suggests that our estimator $\alpha(t;y)$ for $\alpha(y) \triangleq P(W(\infty) \geq y)$ is, in some sense, asymptotically optimal; see for example, Ibragimov and Has'minskii [12].

A natural question that arises here is the extent to which uncertainty in our estimate of λ^* propagates into uncertainty in our prediction of $\alpha(y)$. But note that

$$\begin{aligned} \alpha(t;y) - \alpha(y) &= g(\lambda(t), y) - g(\lambda^*, y) \\ &\approx g'(\lambda^*, y)(\lambda(t) - \lambda^*) \\ &= (1 + \lambda^* y) \exp(-(\mu^* - \lambda^*)y)/\mu^* \cdot (\lambda(t) - \lambda^*). \end{aligned}$$

Since $\lambda(t) - \lambda^*$ is asymptotically normal, this suggests the following result.

Proposition 16.1. *If $\rho < 1$, then*

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

where

$$\sigma^2(y) = \rho(1 + \lambda^* y)^2 \exp(-2(\mu^* - \lambda^*)y)/\mu^*.$$

A rigorous proof is straightforward; see, for example, Serfling [18].

Suppose now that we are particularly interested in computing the fraction of time that the steady-state workload exceeds y, where y is large. More specifically, suppose that we wish to offer such a prediction that is valid to a high level of relative accuracy. Thus, we are concerned with the magnitude of $\alpha(t;y)/P(W(\infty) \geq y)$ when y is large.

Our next result addresses this issue.

Proposition 16.2. *Suppose* $\rho < 1$.

If $t_y/y^2 \to +\infty$, *then*

$$\frac{\sqrt{t_y}}{y}\left(\frac{\alpha(t_y;y)}{\alpha(y)} - 1\right) \Rightarrow \sqrt{\lambda^*}N(0,1) \tag{16.2}$$

as $y \to \infty$.

If $t_y/y^2 \to \gamma > 0$, *then*

$$\frac{\alpha(t_y;y)}{\alpha(y)} \Rightarrow \exp\left(\sqrt{\tfrac{\lambda^*}{\gamma}}N(0,1)\right) \tag{16.3}$$

as $y \to \infty$.

Proof. Note that

$$\alpha(t;y)/\alpha(y) = g(\lambda(t),y)/g(\lambda^*,y) \tag{16.4}$$

$$= \frac{\lambda(t)}{\lambda^*}\exp((\lambda(t) - \lambda^*)y). \tag{16.5}$$

Assuming $t_y/y^2 \to +\infty$, we obtain

$$g(\lambda(t_y),y)/\alpha(y) = \frac{\lambda(t_y)}{\lambda^*}(1 + \exp(\xi(t_y)))(\lambda(t_y) - \lambda^*)y$$

$$= (1 + O_p(t_y^{-1/2}))(1 + \exp(\xi(t_y)))(\lambda(t_y) - \lambda^*)y,$$

where $\xi(t_y) \Rightarrow 0$ as $y \to \infty$. Hence,

$$\frac{\sqrt{t_y}}{y}\left(\frac{\alpha(t_y,y)}{\alpha(y)} - 1\right) = O_p(y^{-1}) + \exp(\xi(t_y))\,(1 + O_p(t_y^{-1/2}))\sqrt{t_y}(\lambda(t_y) - \lambda^*).$$

Since $t^{1/2}(\lambda(t) - \lambda^*) \Rightarrow \sqrt{\lambda^*}N(0,1)$ as $t \to \infty$, (16.2) follows. For (16.3), observe that if $t_y/y^2 \to \gamma > 0$, then

$$(\lambda(t_y) - \lambda^*)y \Rightarrow \sqrt{\tfrac{\lambda^*}{\gamma}}N(0,1)$$

as $y \to \infty$. The result is then immediate from (16.5). □

Proposition 16.2 makes clear that in order to predict the workload tail probability when y is large, the arrival process N must be observed for a period of time t_y that is large relative to y^2. Without observing N over such a time scale, the tail probability cannot be reasonably estimated to a given degree of relative accuracy. Thus, even under the stringent M/M/1 assumptions made here, the amount of time over which the arrival process must be observed to make such workload predictions increases rapidly as a function of y (with the critical time scale being of order y^2.)

16.3 PARAMETRIC ESTIMATION OF TAIL PROBABILITIES FOR MARKOV MODULATED QUEUES

We study here the issue of the degree to which the results of Section 16.2 generalize to Markov-modulated queues. Such queues are widely viewed as offering the modeling flexibility required to cope with the complexity of traffic sources typical of modern communications networks.

To describe such a queue, let $X = (X(t): t \geq 0)$ be a continuous-time Markov chain living on a finite state space S. The infinitesimal generator of X is assumed to belong to the set $\{\overline{A}(\theta): a < \theta < b\}$; we denote the "true" value of $\theta \in (a,b)$ as θ^*. We require $\overline{A}(\theta^*)$ to be irreducible; we further assume that $\overline{A}(\cdot)$ is continuously differentiable.

For $f: S \to (0,\infty)$, let

$$\Lambda(t) = \int_0^t f(X(s))ds; \tag{16.6}$$

we interpret $\Lambda(t)$ as the total work to arrive to the queue in $[0,t]$. As (16.6) suggests, we view the work as arriving in "fluid" form. Set

$$S(t) = \int_0^t f(X(s))ds - t.$$

Then, assuming that the queue starts off with no work present, the workload process $W = (W(t): t \geq 0)$ is given by

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

Let $\overline{P}_\theta(\cdot)$ be the probability measure on the path-space of X, under which X has generator $\overline{A}(\theta)$ and initial distribution $\pi(\theta)$, where $\pi(\theta)$ is the (unique) stationary distribution associated with $\overline{A}(\theta)$. Then,

$$\overline{P}_\theta(S(t) - \min_{0 \leq u \leq t} S(u) \in \cdot) = \overline{P}_\theta\left(\max_{0 \leq u \leq t}\left(\int_u^t [f(X(s)) - 1]ds\right) \in \cdot\right) \tag{16.7}$$

$$= \overline{P}_\theta\left(\max_{0 \leq u \leq t}\left(\int_0^u [f(X(t-s)) - 1]ds\right) \in \cdot\right) \tag{16.8}$$

$$= P_\theta\left(\max_{0 \leq u \leq t}\left(\int_0^u [f(X(s)) - 1]ds\right) \in \cdot\right) \tag{16.9}$$

where $P_\theta(\cdot)$ is the probability measure on the path-space of X, under which X has initial distribution $\pi(\theta)$ and generator $A(\theta) = (A(\theta,x,y): x,y \in S)$, where $A(\theta,x,y) = \overline{A}(\theta,y,x)\pi(\theta,y)/\pi(\theta,x)$. (Thus, P_θ describes the time-reversed dynamics of X under $\overline{P}_\theta(\cdot)$).

Let $M = (M(t): t \geq 0)$ be the maximum process defined via

$$M(t) \equiv \max_{0 \leq u \leq t} S(u).$$

Relations (16.7)-(16.9) assert that

$$\overline{P}_\theta(W(t) \in \cdot) = P_\theta(M(t) \in \cdot).$$

Since M is nondecreasing, it follows that $M(t) \uparrow M(\infty)$ a.s., and consequently,

$$\bar{P}_\theta(W(t) \in \cdot\,) \Rightarrow P_\theta(M(\infty) \in \cdot\,)$$

as $t \rightarrow \infty$. It follows that if one knows θ^*, one can predict the long-run fraction of time that the workload exceeds y via the equation

$$\alpha(y) \equiv P_{\theta^*}(M(\infty) \geq y).$$

Our interest here is, however, in the case where the arrival source is imperfectly known. In particular, suppose that we have only an estimator $\theta(t)$ for θ^* available, rather than θ^* itself. (If X itself were observed under $\bar{P}_{\theta^*}$, one can apply standard maximum likelihood estimation methods to construct such an estimator $\theta(t)$; see, for example, Billingsley [6]. If f is not one-to-one and only $(f(X(s))$: $0 \leq s \leq t)$ is observed, then more complicated estimators of θ^* would need to be considered, as described in Bickel and Ritov [5] and Ryden [17]). Our prediction for the probability that the workload exceeds y in steady-state is then

$$\alpha(t; y) = g(\theta(t), y),$$

where

$$g(\theta, y) = P_\theta(M(\infty) \geq y). \tag{16.10}$$

In order that $M(\infty)$ have a proper distribution (and hence $W(\infty)$) under the true parameter θ^*, we need to assume that the long-run rate at which work enters the system is less than one. Specifically, we shall assume that $\rho(\theta^*) < 1$, where

$$\rho(\theta) \triangleq \sum_{x \in S} \pi(\theta, x) f(x).$$

Set $T(y) = \inf\{t \geq 0: S(t) \geq y\}$. From (16.10), it is evident that we can express $g(\theta, y)$ as a "level-crossing" probability, namely

$$g(\theta, y) = P_\theta(T(y) < \infty).$$

In order to proceed further, we shall now invoke a "change-of-measure" argument. Let D be the diagonal matrix defined by $D = \text{diag}(f(x): x \in S)$, and let

$$A(\theta, \gamma) = A(\theta) + \gamma(D - I)$$

for $\gamma \in \Re$. Because $A(\theta, \gamma)$ is an irreducible M-matrix, it follows that there exists a strictly positive column eigenvector $h(\theta, \gamma)$ and a real eigenvalue $r(\theta, \gamma)$ such that

$$A(\theta, \gamma) h(\theta, \gamma) = r(\theta, \gamma) h(\theta, \gamma).$$

Using a similar argument to that presented on page 126 of Bucklew [9] for discrete-time Markov chains, we can prove that $r(\theta, \cdot)$ is convex with $r(\theta, \gamma) \rightarrow +\infty$ as $\gamma \rightarrow +\infty$. Furthermore, $\frac{\partial}{\partial \gamma} r(\theta, \gamma)\,|_{\gamma = 0} < 0$ for θ in a neighborhood of θ^*. Hence, it is clear that there exists a positive root $\gamma(\theta)$ to the equation

$$r(\theta, \gamma) = 0.$$

Set

$$B(\theta, x, y) = A(\theta, \gamma(\theta), x, y) \frac{h(\theta, \gamma(\theta), y)}{h(\theta, \gamma(\theta), x)}$$

and note that $B(\theta) = (B(\theta, x, y): x, y \in S)$ is the generator of an irreducible continuous-time Markov chain on S. Furthermore,

$$A(\theta, x, y) = \frac{h(\theta, \gamma(\theta), x)}{h(\theta, \gamma(\theta), y)} K(\theta, x, y),$$

where $K(\theta) = B(\theta) - \gamma(\theta)(D - I)$. It is trivial to verify that

$$(\exp(A(\theta)t))(x, y) = \frac{h(\theta, \gamma(\theta), x)}{h(\theta, \gamma(\theta), y)} (\exp(K(\theta)t))(x, y). \tag{16.11}$$

Set $u_\theta(t, x, y) = (\exp(K(\theta)t))(x, y)$ and note that $u_\theta(t) = (u_\theta(t, x, y): x, y \in S)$ solves the linear system of ordinary differential equations defined by

$$u_\theta'(t) = (B(\theta) - \gamma(\theta)(D - I)) u_\theta(t) \tag{16.12}$$

$$u_\theta(0) = I.$$

Let $\underline{E}_\theta(\cdot)$ and $\underline{P}_\theta(\cdot)$ be the expectation operator and probability distribution on the path-space of X associated with the generator $B(\theta)$ and initial distribution $\pi(\theta)$. It is well known that the solution to (16.12) can be expressed probabilistically as

$$u_\theta(t, x, y) = \underline{E}_\theta[\exp(-\gamma(\theta) S(t)) I(X(t) = y) \mid X(0) = x]. \tag{16.13}$$

Combining (16.11) and (16.13), we conclude that

$$P_\theta(X(t) = y \mid X(0) = x) = \underline{E}_\theta[\exp(-\gamma(\theta) S(t)) I(X(t) = y)$$

$$\cdot \frac{h(\theta, \gamma(\theta), X(0))}{h(\theta, \gamma(\theta), X(t))} \mid X(0) = x]. \tag{16.14}$$

Letting $E_\theta(\cdot)$ be the expectation operator associated with $P_\theta(\cdot)$, it is straightforward to extend (16.14) to the identity

$$E_\theta \zeta = \underline{E}_\theta \zeta \exp(-\gamma(\theta) S(t)) \frac{h(\theta, \gamma(\theta), X(0))}{h(\theta, \gamma(\theta), X(t))}, \tag{16.15}$$

for any non-negative r.v. ζ that is measurable with respect to $\sigma(X(s): 0 \leq s \leq t)$. Since $T(y)$ is a stopping time, (16.15) in turn easily implies the identity

$$P_\theta(T(y) < \infty) = \underline{E}_\theta \exp(-\gamma(\theta) S(T(y))) \frac{h(\theta, \gamma(\theta), X(0))}{h(\theta, \gamma(\theta), X(T(y)))} I(T(y) < \infty)$$

$$= \exp(-\gamma(\theta) y) \underline{E}_\theta \frac{h(\theta, \gamma(\theta), X(0))}{h(\theta, \gamma(\theta), X(T(y)))} I(T(y) < \infty).$$

Let $\nu(\theta) = (\nu(\theta, x): x \in S)$ be the (unique) stationary distribution associated with $B(\theta)$. Again, by following an argument similar to that given in Bucklew [9] for discrete-time chains, one can show that

$$\frac{\partial}{\partial \gamma} r(\theta, \gamma)\,|_{\gamma = \gamma(\theta)} = \sum_{x \in S} (f(x) - x) \nu(\theta, x).$$

Since $\frac{\partial}{\partial \gamma} r(\theta, \gamma)$ must clearly be positive at the root $\gamma(\theta)$, it follows that $S(\cdot)$ must have positive drift under $\underline{P}_\theta$ for θ in a neighborhood of θ^*. Consequently, we may simplify (16.16) to the identity

$$P_\theta(T(y) < \infty) = \exp(-\gamma(\theta) y) \underline{E}_\theta \frac{h(\theta, \gamma(\theta), X(0))}{h(\theta, \gamma(\theta), X(T(y)))}.$$

In other words,

$$g(\theta, y) = \exp(-\gamma(\theta) y) \underline{E}_\theta \frac{h(\theta, \gamma(\theta), X(0))}{h(\theta, \gamma(\theta), X(T(y)))}. \tag{16.16}$$

Our analysis of Section 16.2 suggests that the derivative $g'(\theta^*, y)$ plays an important role in the estimation theory associated with predicting $\alpha(y)$. Our next results proves that $g(\cdot, y)$ is smooth. It can be viewed as a strengthening of the various “continuity” results available for the single-server queue (see, for example, page 194 of Asmussen [4]) to “differentiability”. Set $\Gamma(\theta) = \{(x,y): \overline{a}(\theta,x,y) \neq 0\}$ and $h(\theta,x) = h(\theta,\gamma(\theta),x)$. Also, let $J(t)$ denote the number of jumps of X over $[0,t]$, and let $Y = (Y_n: n \geq 0)$ denote the embedded discrete-time Markov chain associated with X.

Theorem 16.3. *Assume that $\rho(\theta^*) < 1$ and that $\overline{A}(\theta^*)$ is irreducible. If $\overline{A}(\cdot)$ is continuous differentiable in a neighborhood of θ^* and $\Gamma(\theta) = \Gamma(\theta^*)$ in a neighborhood of θ^*, then $g(\cdot,y)$ is differentiable at θ^* for each $y \geq 0$. Furthermore,*

$$g'(\theta^*,y) = \exp(-\gamma(\theta^*)y) \cdot \underline{E}_{\theta^*} \frac{h(\theta^*,X(0))}{h(\theta^*,X(T(y)))}$$

$$\cdot \Big[\frac{\pi'(\theta^*,X(0))}{\pi(\theta^*,X(0))} + \sum_{i=0}^{J(T(y))-1} \frac{A'(\theta^*,Y_i,Y_{i+1})}{A(\theta^*,Y_i,Y_{i+1})} + \int_0^{T(y)} A'(\theta^*,X(s),Xs))ds\Big]. \quad (16.17)$$

Proof. Our “absolute continuity” condition involving the set $\Gamma(\cdot)$ implies the existence of a likelihood ratio process $(L(\theta,t): t \geq 0)$ such that

$$g(\theta,y) = \exp(-\gamma(\theta)y)\underline{E}_\theta \frac{h(\theta,X(0))}{h(\theta,X(T(y)))} L(\theta,T(y));$$

see Brémaud [8]. Setting $\lambda(\theta,x) = -B(\theta,x,x)$, $L(\theta,t)$ is given by

$$L(\theta,t) = \frac{\pi(\theta,X(0))}{\pi(\theta^*,X(0))} \prod_{i=0}^{J(t)-1} \frac{B(\theta,Y_i,Y_{i+1})}{B(\theta^*,Y_i,Y_{i+1})}$$

$$\cdot \exp\Big(-\int_0^t [\lambda(\theta,X(s)) - \lambda(\theta^*,X(s))]ds\Big).$$

But

$$B(\theta,Y_i,Y_{i+1}) = A(\theta,Y_i,Y_{i+1}) \frac{h(\theta,Y_{i+1})}{h(\theta,Y_i)}$$

and

$$\int_0^t \lambda(\theta,X(s))ds = -\Big(\int_0^t A(\theta,X(s),X(s))ds + \gamma(\theta)S(t)\Big).$$

Thus,

$$g(\theta,y) = \exp(-\gamma(\theta^*)y)\underline{E}_\theta \beta(\theta),$$

where

$$\beta(\theta) \triangleq \exp((\gamma(\theta^*) - \gamma(\theta))y) \frac{h(\theta,X(0))}{h(\theta,X(T(y)))} L(\theta,T(y)) \quad (16.18)$$

$$= \frac{\pi(\theta,X(0))}{\pi(\theta^*,X(0))} \frac{h(\theta^*,X(0))}{h(\theta^*,X(T(y)))} \prod_{i=0}^{J(t)-1} \frac{A(\theta,Y_i,Y_{i+1})}{A(\theta^*,Y_i,Y_{i+1})}$$

$$\cdot \exp\Big(\int_0^{T(y)} [A(\theta,X(s),X(s)) - A(\theta^*,X(s),X(s))]ds\Big). \quad (16.19)$$

Consequently, assuming that we can interchange the derivative and expectation operators, it follows that $g'(\theta,y) = \exp(-\gamma(\theta^*)y)\underline{E}_{\theta^*}\beta'(\theta)$, where

$$\frac{\beta'(\theta)}{\beta(\theta)} = \frac{d}{d\theta} \log \beta(\theta) \quad (16.20)$$

$$= \frac{\pi'(\theta,X(0))}{\pi(\theta,X(0))} + \sum_{i=0}^{J(T(y))-1} \frac{A'(\theta,Y_i,Y_{i+1})}{A(\theta,Y_i,Y_{i+1})} + \int_0^{T(y)} A'(\theta,X(s),X(s))ds. \quad (16.21)$$

Since $\beta(\theta^*) = h(\theta^*,X(0))/h(\theta^*,X(T(y)))$, this leads immediately to (16.17). So, it remains to justify the interchange of derivative and expectation.

We need to show that the difference quotient

$$D(h) = \frac{\beta(\theta+h) - \beta(\theta)}{h}$$

can be dominated (uniformly in h small) by a $\underline{P}_{\theta^*}$ integrable r.v. Note that $r(\cdot)$ is continuous [13] and $r(\theta,\cdot)$ is convex. Hence, $\gamma(\cdot)$ is continuous in a neighborhood of θ^*, as is $h(\cdot,x)$ for $x \in S$. It is then straightforward to show that for each $\epsilon > 0$, there exist (deterministic) positive constants h_0 and c such that

$$\sup_{|h| < h_0} |D(h)| \leq \sup_{|h| < h_0} |\beta'(h)| \quad (16.22)$$

$$\leq c\exp(\epsilon T(y) + \epsilon J(T(y))). \quad (16.23)$$

To show that the right-hand side of (16.23) is integrable, we use a martingale argument. Recall that

$$A(\theta^*,\gamma(\theta^*)/2)h(\theta^*,\gamma(\theta^*)/2) = r(\theta^*,\gamma(\theta^*)/2)h(\theta^*,\gamma(\theta^*)/2).$$

In other words,

$$[B(\theta^*) - (\gamma(\theta^*)/2)(D-I)]k = \eta k,$$

where $k(x) = h(\theta^*,\gamma(\theta^*)/2,x)/h(\theta^*,\gamma(\theta^*),x)$ and $\eta = r(\theta^*,\gamma(\theta^*)/2)$. Let $C = (C(x,y)\colon x,y \in S)$ be defined by

$$C(x,y) = B(\theta^*,x,y) = [\gamma(\theta^*)(f(x)-1)/2 + \eta]\delta_{xy},$$

and note that $(C(x,y)k(y)/k(x): x,y \in S)$ is the generator of a continuous-time Markov chain in S. Hence,

$$\sum_y \exp(Ct)(x,y) \cdot \frac{k(y)}{k(x)} = 1.$$

As in the derivation leading to (16.13), this can be written probabilistically as

$$\underline{E}_{\theta^*}[\exp(-\gamma(\theta^*)S(t)/2 - \eta t)\frac{k(X(t))}{k(X(0))}] = 1.$$

It follows easily that

$$\exp(-\gamma(\theta^*)S(t)/2 - \eta t)\frac{k(X(t))}{k(X(0))}$$

is a $\underline{P}_\theta^*$-martingale. The optional sampling theorem then yields

$$\underline{E}_{\theta^*}M(t \wedge T(y)) = 1$$

for $t \geq 0$. But $S(t \wedge T(y)) \leq y$ and $\gamma(\theta^*) > 0$, so evidently,

$$M(t \wedge T(y)) \geq \exp(-\gamma(\theta^*)y/2) \inf_{x,z}\frac{k(x)}{k(z)} \cdot \exp(-\eta(t \wedge T(y))).$$

Hence,

$$\underline{E}_{\theta^*}\exp(-\eta(t\wedge T(y))) \le \exp(\gamma(\theta^*)y/2)\sup_{x,z}\frac{k(x)}{k(z)}.$$

Since $r(\theta^*,0)=r(\theta^*,\gamma(\theta^*))=0$ and $r(\theta^*,\cdot)$ is convex, we can infer that $\eta<0$. The monotone convergence theorem then establishes the finiteness of $\underline{E}_{\theta^*}\exp(-\eta T(y))$, yielding the existence of a positive ϵ for which $\underline{E}_{\theta^*}\exp(\epsilon T(y))$ is finite.

To handle $\exp(\epsilon J(T(y)))$, note that for $\nu>0$,

$$\underline{E}_{\theta^*}\exp(\nu J(T(y))) = \sum_{n=0}^{\infty}\underline{E}_{\theta^*}\exp(\nu J(T(y)))I(n\le T(y)<n+1) \tag{16.24}$$

$$\le \sum_{n=0}^{\infty}\underline{E}_{\theta^*}\exp(\nu J(N+1))I(T(y)\ge n) \tag{16.25}$$

$$\le \sum_{n=0}^{\infty}\underline{E}^{1/2}_{\theta^*}\exp(2\nu J(n+1))\underline{P}^{1/2}_{\theta^*}(T(y)\ge n). \tag{16.26}$$

But $J(\cdot)$ can be stochastically dominated by a Poisson process having rate $\tilde{\lambda}\triangleq\max(\lambda(\theta^*,x):x\in S)$, and so

$$\underline{E}^{1/2}_{\theta^*}\exp(2\nu J(n+1))\le\exp(\tilde{\lambda}(n+1)(e^{2\nu}-1)/2).$$

The finiteness of $\underline{E}_{\theta^*}\exp(\epsilon T(y))$ then guarantees the finiteness of $\underline{E}_{\theta^*}\exp(\nu J(T(y)))$ for some positive ν. The Cauchy-Schwartz inequality then yields (16.23) for some suited chosen positive ϵ. □

Theorem 16.3 proves that $g(\cdot,y)$ is differentiable for each $y\ge 0$, and offers a representation for the derivative that is surprisingly simple. In particular, note that the function f plays no explicit role in the expression for $g'(\theta^*,y)$ (although it does serve to implicitly define the expectation $\underline{E}_{\theta^*}(\cdot)$ under which the expectation is computed).

Assume now that θ^* is unknown and that it must be statistically estimated from data collected over the time interval $[0,t]$. Let $\theta(t)$ be the corresponding estimator for θ^* Most well-behaved estimators obey a central limit theorem (CLT) of the form

$$t^{1/2}(\theta(t)-\theta^*)\Rightarrow\delta N(0,1) \tag{16.27}$$

as $t\to\infty$, for some suitably defined constant $\delta>0$; see, for example, the extensive central limit theory available for maximum likelihood estimators in [12].

Corollary 16.4. *Assume* $(\theta(t):t\ge 0)$ *obeys* (27). *Then, under the conditions of Theorem* 16.3, *it follows that*

$$t^{1/2}(\alpha(t;y)-\alpha(y))\Rightarrow g'(\theta^*,y)\delta N(0,1)$$

as $t\to\infty$.

Corollary 16.4 is the Markov-modulated analog of Proposition 16.1.

We turn now to the analog of Proposition 16.2, developed earlier in the M/M/1 context. Equation (16.16) suggests that the r.v. $X(T(y))$ plays an important role in determining the behavior of $g(\theta,y)$. With this in mind, note that $S(t)\to+\infty$ $\underline{P}_\theta$ a.s. for all θ in a neighborhood of θ^* and hence $T(y)<\infty$ $\underline{P}_\theta$ a.s. for all $y\ge 0$ under such values of θ. Set $\underline{X}(y)=X(T(y))$ for $y\ge 0$. Then, $\underline{X}=(\underline{X}(y):y\ge 0)$ is a well-defined stochastic process under $\underline{P}_\theta$ for θ in a neighborhood of θ^*. Furthermore, it turns out that $\underline{X}$ is itself a time-homogeneous continuous-time Markov chain with a single closed communicating class $\underline{S}\subset S$; [2, 3, 16].

Let $\underline{\pi}(\theta)=(\underline{\pi}(\theta,x):x\in S)$ be the (unique) stationary distribution of $\underline{X}$ under $\underline{P}_\theta$. Since $\underline{X}(y)$ is asymptotically independent of $\underline{X}(0)$, this suggests the following approximation for $g(\theta,y)$, derived from (16.16):

$$g(\theta,y)\approx\exp(-\gamma(\theta)y)\cdot\sum_{x\in S}\pi(\theta,x)h(\theta,\gamma(\theta),x)\cdot\underline{\pi}(\theta,z)/h(\theta,\gamma(\theta),z)$$

$$\triangleq\exp(-\gamma(\theta)y)\cdot v(\theta).$$

From the above approximations, the conclusions of Theorem 16.5 are intuitively clear.

Theorem 16.5. *Assume the condition of Theorem* 16.3 *and suppose that* $(\theta(t):t\ge 0)$ *satisfies* (27). *If* $t_y/y^2\to+\infty$ *and* $y/\log t_y\to+\infty$, *then*

$$\frac{\sqrt{t_y}}{y}\left(\frac{\alpha(t_y;y)}{\alpha(y)}-1\right)\Rightarrow\delta\frac{v'(\theta^*)}{v(\theta^*)}N(0,1)$$

as $y\to\infty$. *If* $t_y/y^2\to\gamma>0$, *then*

$$\frac{\alpha(t_y;y)}{\alpha(y)}\Rightarrow\exp(\delta\gamma'(\theta^*)\gamma^{-1/2}N(0,1))$$

as $y\to\infty$.

Proof. Our starting point is (16.16). Let $v(\theta,y)=\underline{E}_\theta h(\theta,\gamma(\theta),X(0))/h(\theta,\gamma(\theta),X(T(y)))$. We need to show that replacing $v(\theta,y)$ by $v(\theta)$ holds uniformly in a neighborhood of θ^* as $y\to+\infty$. To this end, we write

$$v(\theta,y)=\sum_{x,z}\pi(\theta,x)h(\theta,\gamma(\theta),x)\underline{P}_\theta(\underline{X}(y)=z\mid X(0)=x)\cdot h(\theta,\gamma(\theta),z)^{-1}.$$

We will show that for ϵ sufficiently small, there exists $a>0$ such that

$$\sup_{|\theta-\theta^*|<\epsilon}|\underline{P}_\theta(\underline{X}(y)=z\mid\underline{X}(0)=x)-\underline{\pi}(\theta,z)|=O(e^{-ay}) \tag{16.28}$$

as $y\to\infty$. The finiteness of the state space then guarantees that

$$\sup_{|\theta-\theta^*|<\epsilon}|v(\theta,y)-v(\theta)|=O(e^{-by}) \tag{16.29}$$

as $y\to\infty$, for some suitably chosen $b>0$. To verify (16.28), we invoke a "uniform coupling" argument. Note that

$$\underline{P}_\theta(\underline{X}(y)=z\mid\underline{X}(0)=x)=\underline{P}_\theta(X(T(y))=z\mid X(0)=x).$$

The same argument as that applied in the proof of Theorem 16.3 establishes that the right-hand side is differentiable at θ^* for each $y\ge 0$. Now,

$$\inf_{\substack{x\in S\\ z\in\underline{S}}}\underline{P}^*_\theta(X(T(1))=z\mid X(0)=x)>0$$

follows as a consequence of the fact that $\underline{X}$ is a continuous-time Markov chain on finite state space. Because the transition probabilities are differentiable at θ^*, there therefore exists $\epsilon>0$ such that

$$\inf_{\substack{x\in S\\ z\in\underline{S}\\ |\theta-\theta^*|<\epsilon}}\underline{P}_\theta(\underline{X}(1)=z\mid\underline{X}(0)=x)\triangleq m>0.$$

A standard coupling argument (see, for example, Lindvall [15]) then proves that

$$\inf_{\substack{x \in S \\ z \in \underline{S} \\ |\theta - \theta^*| < \epsilon}} | \underline{P}_\theta(\underline{X}(y) = z \mid \underline{X}(0) = x) - \underline{\pi}(\theta, z) | \leq (1-c)^{\lfloor y \rfloor},$$

yielding (16.28). In view of (16.16) and (16.29), we can write

$$\frac{\alpha(t_y; y)}{\alpha(y)} = \exp(-(\gamma(\theta(t)) - \gamma(\theta^*))y) \cdot \left\{ \frac{v(\theta(t))}{v(\theta^*)} + O(e^{-by}) \right\}.$$

The conclusions of Theorem 16.5 then follow from precisely the same style of argument as that used for Proposition 16.2, provided only that we show $v(\cdot)$ to be differentiable at θ^*. (Note that the additional hypothesis that $y/\log t_y \to +\infty$ serves only to guarantee that $O(e^{-by})$ is of (much) smaller magnitude than $t_y^{-1/2}$, the magnitude of $\theta(t_y) - \theta^*$, thereby permitting us to discard $O(e^{-by})$ in our asymptotic analysis.)

To prove that $v(\cdot)$ is differentiable at θ^*, recall our earlier remark that the matrix of transition probabilities

$$(\underline{P}_\theta(\underline{X}(y) = z \mid \underline{X}(0) = x); x, z \in \underline{S})$$

is differentiable at θ^*. The stationary distribution $\underline{\pi}(\theta)$ can be viewed as the (unique) stationary distribution of the above transition matrix. Since the transition matrix is differentiable at θ^*, we may conclude that $\underline{\pi}(\cdot)$ is differentiable at θ^*; see Glynn [10]. It then trivially follows that $v(\cdot)$ is differentiable at θ^*. □

Theorem 16.5 reinforces the qualitative insight gained from Section 16.2. If the arrival process is modeled parametrically, then it must be observed over a time scale t_y, which is large relative to y^2, in order that (relatively) accurate predictions of the tail workload probability $P_{\theta^*}(W(\infty) > y)$ be available.

16.4 FURTHER ASYMPTOTIC ANALYSIS

In Sections 2 and 3, we considered the question of how much data needs to be collected in order to parametrically estimate tail and workload probabilities in two related single-server models, as a function of the parameter y. In this section, we briefly consider the issue of how much data needs to be collected, as a function of the traffic intensity of the queue. We restrict our attention to the M/M/1 queue, because its parametrization lends itself naturally to such analysis (whereas, in the Markov-modulated setting, there is no obvious connection between θ and the traffic intensity of the queue).

The starting point is relation (16.5) of Section 16.2. Note that $\alpha(t;y)/\alpha(y)$ is independent of μ^*, so that the relative error associated with estimating $\alpha(y)$ is independent of μ^*. In particular, the relative error is (perhaps surprisingly) independent of how close μ^* is to λ^*, and thereby, independent of the traffic intensity. Of course, the magnitude of the workloads y that are of interest depend heavily on the traffic intensity ρ.

For example, as $\rho \to 1$ in "heavy traffic", it is clear that y must be of order $(1-\rho)^{-1}$, in order that $P_{\theta^*}(W(\infty) \geq y)$ be held constant. The analysis of Section 16.16.2 then suggests that the time horizon t_y over which the arrival process should be observed ought to be large relative to $(1-\rho)^{-2}$. Since $(1-\rho)^{-2}$ is the time-scaling over which $W(\cdot)$ exhibits significant (relative) fluctuations, this indicates that the non-parametric estimator

$$\alpha_N(t;y) \triangleq \frac{1}{t} \int_0^t I(W(s) \geq y) ds$$

should also provide good estimation behavior over time horizons that are large relative to $(1-\rho)^{-2}$. In other words, parametric modeling of the arrival process does not have a dramatic effect on the ability to estimate tail workload probabilities when the value y being considered is in "the middle" of the distribution. Its primary impact is in estimation of the extreme tail.

16.5 OPEN RESEARCH PROBLEMS

A number of different research problems are suggested by the line of inquiry pursued in this chapter.

Given that this chapter focuses on parametric modeling of queues, an important question relates to the types of parametric models that lend themselves best to specific application contexts. A particularly intriguing variant on this issue is the question of how to effectively model complex dependent arrival processes via stochastic processes that are parametrized by low-dimensional vectors. Ideally, the parametrized class of processes chosen should give rise to a class of queues that is numerically tractable, and for which the parameter estimation can be implemented in a satisfactory fashion. Furthermore, one prefers, everything else equal, a parametrization in which each of the parameters entering the model has some physically intuitive meaning.

Given our emphasis in this chapter on parametric queueing models, it is natural to ask about how the results developed here would be affected by a non-parametric formulation. In particular, if one is unwilling to assume that the arrival process is well-described via some (finite-dimensional) parametric model, how long must one observe the arrival process in order to accurately estimate tail probabilities for (very) large buffer levels?

Turning now to the theory developed in this chapter, it is worth noting that computing the quantity $g(\theta(t), y)$ (our estimator for $\alpha(y)$), given $\theta(t)$, may be itself a highly non-trivial activity. For example, there is no closed-form for this quantity in the Markov-modulated context. Instead, one needs to compute $g(\theta(t), y)$ either via some conventional numerical solver, or via a Monte Carlo simulation of some kind. Given the real-time applications that motivate this class of problem, the issue of how to rapidly compute $g(\theta(t), y)$ could potentially be of importance. In addition, if confidence intervals for $\alpha(y)$ are desired, rapid computation of $g'(\theta(t), y)$ (see Proposition 16.1 and Theorem 16.3) plays an important role. Such computational issues should ideally be considered in concert with the statistical questions that were central in this chapter.

BIBLIOGRAPHY

[1] Anick, D., Mitra, D. and Sondhi, M.M., Stochastic theory of a data handling system with multiple sources, *Bell Sys. Tech. J.* **61** (1982), 1871-1894.

[2] Asmussen, S., Stationary distributions for fluid flow models with or without

Brownian noise, *Stochastic Models* **11**:1 (1995), 21-49.
[3] Asmussen, S., Busy period analysis, rare events and transient behavior in fluid flow models, *J. of Appl. Math. and Stoch. Anal.* **7**:3 (1994), 269-299.
[4] Asmussen, S., *Applied Probability and Queues*, Wiley, New York 1987.
[5] Bickel, P.J. and Ritov, Y., Inference in hidden Markov models I. Local asymptotic normality in the stationary case, *Bernoulli* (1996), to appear.
[6] Billingsley, P., *Statistical Inference for Markov Processes*, University of Chicago Press, Chicago 1961.
[7] Borovkov, A.A., *Stochastic Processes in Queueing Theory*, Springer-Verlag, New York 1976.
[8] Brémaud, P., *Point Processes and Queues: Martingale Dynamics*, Springer-Verlag, New York 1981.
[9] Bucklew, J.A., *Large Deviations Techniques in Decision, Simulation, and Estimation*, Wiley, New York 1990.
[10] Glynn, P.W., Sensitivity analysis for the stationary probabilities of a Markov chain, In: *Proc. of the 4th Army Conference on Applied Math and Comp.* (1986), 917-932.
[11] Glynn, P.W. and Torres, M., Nonparametric estimation of tail probabilities for the single-server queue, Forthcoming technical report, Dept. of Operations Res., Stanford University, Stanford, CA 1996.
[12] Ibragimov, I.A. and Has'minskii, R.Z., *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York 1981.
[13] Kato, T., *Perturbation Theory for Linear Operators*, Wiley, New York 1976.
[14] Kesidis, G., Walrand, J. and Chang, C.S., Effective bandwidths for multiclass Markov fluids and other ATM sources, *IEEE/ACM Trans. on Networking* 1:4 (1993), 424-428.
[15] Lindvall, T., *Lectures on the Coupling Method*, Wiley, New York 1982.
[16] Rogers, C.L.G., Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains, submitted.
[17] Ryden, T., Estimates for hidden Markov models, *Ann. Stat.* **22**:4 (1994), 1884-1945.
[18] Serfling, R.J., *Approximation Theorems of Mathematical Statistics*, Wiley, New York 1980.
[19] Stern, T. and Elwalid, E., Analysis of separable Markov-modulated rate models for information-handling systems, *Adv. Appl. Prob.* **23** (1991), 105-139.
[20] Stoyan, D., *Comparison Methods for Queues and Other Stochastic Models*, Wiley, New York 1983.

Index