# THE ESTIMATION OF DELAY GRADIENTS FOR PURPOSES OF ROUTING IN DATA - COMMUNICATION NETWORKS

by

Martin Glen Bello

B.S. Cornell University

1975

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February, 1977

Signature of Author . . Martin G. Bello . . . . . . . . . . . . . . . .
Department of Electrical Engineering, February , 1977

Certified by . . . . Adrian Segall . . . . . . . . . . . . . . . . .
Thesis Supervisor

Accepted by . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Chairman, Departmental Committee on Graduate Students

# THE ESTIMATION OF DELAY GRADIENTS FOR PURPOSES OF ROUTING IN DATA - COMMUNICATION NETWORKS

by

Martin Glen Bello

Submitted to the Department of E.E. and C.S.

on

January 20, 1977

in

Partial Fulfillment of the Requirements for the
Degree of Master of Science

## ABSTRACT

Appealing to current and past work on the routing problem in data - communication networks, we motivate the need for algorithms that estimate the derivative with respect to flow, of the total message delay on each of the links. We then cast the problem in a queueing theory framework and, making no statistical assumptions other than stationarity, we propose three algorithms that process the record of arrivals and departures of a single-server queue to derive an estimate of the derivative, with respect to arrival rate, of the total delay accumulated per unit time. Through simulation and analysis we show that all three algorithms are asymptotically unbiased and efficient for M/D/1 queues. By simulation of other queues we investigate the relative robustness of the three procedures. Finally, through examination of the storage and computational requirements we identify a single most promising algorithm.

Thesis Supervisor: Adrian Segall
Associate Professor

# TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

## LIST OF ILLUSTRATIONS

## LIST OF ILLUSTRATIONS (CONTINUED)

# LIST OF TABLES

# SECTION 1
# INTRODUCTION

## 1.1 The Message-Routing Problem in Data-Communication Networks

There are several major analytical problems in the design of a modern data communication network. Given a set of nodes, different topological configurations may be considered. Having specified the manner in which the nodes are connected, there remains the question of how to assign capacities to the communication links. For each problem, different constraints and optimization criteria are appropriate. Once we have resolved the issue of the structure of the network, the central problem that remains is how to route a given message from a source to a destination node. We are interested in store-and-forward computer networks where messages, or segments of messages called packets, travel from a source to a destination node, waiting in queues for retransmission at each intermediate node. One way to specify a routing policy is by providing a routing table for each node i listing what fraction of the traffic destined for node j is to be sent on each of the outgoing links from node i. Routing strategies vary in character from purely static, the routing fractions being fixed in time and determined on the basis of average arrival statistics, to the completely dynamic case where the fractions vary continuously in time according to the "state" of the network. The philosophy for implementing any strategy varies between centralized and decentralized extremes. In the centralized case a special node in the network receives information and does all the routing computations, communicating changes in the routing fractions to all other nodes. In a decentralized scheme, each node computes its own routing table on the basis of "locally" available information.

An intermediary between the strictly static and dynamic routing schemes is termed quasi-static routing. Here we overcome the static procedures insensitivity to gradual traffic changes and the failures of links and nodes by up-dating the routing strategy periodically, or when a special need arises. In the dynamic routing case, messages that have been segmented into packets may arrive out of order at the destination node, necessitating a "reassembly" operation. In the quasi-static procedure, most messages would be delivered in order since the time intervals between routing charges will be relatively long. Hence, quasi-static routing procedures suggest a sensible mid-point between the static and dynamic extremes.

## 1.2 Purpose of the Thesis

Routing procedures that have been derived to optimize system performance in the sense of minimizing the total delay $D_T$ accumulated per unit time, work with an objective function of the following form:

$$D_T = \sum_{(i,k)} D_{ik}(f_{ik}) \qquad (1.1)$$

The assumptions inherent in (1.1) are discussed by Kleinrock in [5]. $D_{ik}(f_{ik})$ denotes the average delay/unit time on link i-k and $f_{ik}$ denotes the total link flow in bits/sec. An extra term may be added to include the effect of propagation delays on link i-k. These previous approaches to routing have employed closed form expressions for the $D_{ik}(f_{ik})$'s, derived from queueing theory by making many simplifying assumptions. This approach to the routing problem has been taken by Kleinrock [5], Cantor and Gerla [2], and Schwartz and Cheung [7].

The departure of this thesis is from the search for closed-form expressions to finding efficient algorithms to estimate the quantities of interest. In particular, for static and quasi-static routing it has been shown in the previously mentioned works, as well as others like Gallager [3], Agrew [1], and Segall [6], that the routing procedure should be based on knowledge of the derivative of the total delay/unit time $D'_{ik}(f_{ik})$ of messages passing through a link i-k with respect to the total flow $f_{ik}$. Rather than differentiating closed from expressions for $D_{ik}(f_{ik})$, we propose processing the queues at the links to estimate the $D'_{ik}(f_{ik})$'s directly. In this manner we can disassociate the optimality of a given routing procedure from all the assumptions necessary to the closed form formulae for delay. In this thesis we derive three different estimation procedures for the marginal delays $D'_{ik}(f_{ik})$ by making no assumptions as to the structure of the queues. However, to study the properties of each estimator through analysis and simulation methods, we make very specific assumptions about the structure and underlying statistics of queues to which the estimation algorithms are to be applied. Hence, in the following paragraphs we motivate the importance of directly estimating the incremental delays $D'_{ik}(f_{ik})$ by reviewing the previous work in designing routing strategies, with special emphasis on those results relevant to quasi-static routing procedures.

## 1.3 Previous Work

The most common model for routing problems in data networks is that derived by Kleinrock [5]. He makes the following assumptions:

1) Poisson arrival at nodes

2) Exponential distribution of message length

3) Independence of arrival processes at different nodes

4) The "independence" assumption of service times at successive nodes. Each time a message arrives at a node a new service requirement is chosen from the same exponential distribution.

On the basis of these assumptions he derives an explicit formula for the total delay/unit time accumulated on the (i-k)-th link. If $f_{ik}$ denotes the amount of traffic passing over the (i-k)-th link in bits/sec. and $C_{ik}$ is the capacity of link (i-k) in bits/sec., the average total delay will be given by

$$D_{ik}(f_{ik}) = \frac{f_{ik}}{C_{ik} - f_{ik}} \tag{1.2}$$

To illustrate how Kleinrock's result in (1.2) is used in routing problems, we outline the static routing scheme presented by Cantor and Gerla [2]. The problem of finding an optimal set of routes is posed as a nonlinear multi-commodity flow problem, where we want to derive the flow vector f*, whose entries are $f_{ik}$'s, that minimizes the following objective function:

$$T = \frac{1}{\gamma} \sum_{(i,k)} \frac{f_{ik}}{C_{ik} - f_{ik}} \tag{1.3}$$

The summation in (1.3) is taken over all (i,k) pairs that are connected and $\gamma$ is the total external arrival rate in packets/ unit time. T is interpreted as the average packet delay. The set of flows that satisfy multi-commodity, capacity, and non-negativity constraints is shown to be a convex polyredral set and hence any f in that set may be expressed as a convex combination of extremal flows $\varphi^{(i)}$

$$f = \sum_{i=1}^{r} \lambda_i \varphi^{(i)} \quad \sum_{i=1}^{r} \lambda_i = 1 \tag{1.4}$$

Letting $\nabla T(f^*)$ denote the gradient of the objective function in (1.3) evaluated at $f^*$, Cantor and Gerla propose an algorithm that finds the optimal $f = f^*$ in the sense of (1.3) for a given basis of external flows ($\varphi^{(1)} \ldots \varphi^{(k)}$), and then generates a new basis vector $\varphi^{(k+1)}$ that minimizes $\langle \nabla T(f^*), \varphi \rangle$. $\varphi^{(k+1)}$ is the new extremal flow that will help us reduce T the fastest. The procedure continues iteratively until we are as close to the optimal flow as desired. As part of the calculation of $\varphi^{(i)}$, we generate a set of routing tables that realize the given flow.

Schwartz and Cheung [7] describe a gradient type algorithm for calculating the optimal flow vector which motivates a possible stochastic approximation algorithm for quasi-static routing. Let $f_{ij}$ denote the total bit rate on link (i-j) and $f_{ij}^{(m,n)}$ denote the bit rate of messages with source m and destination n on link i-j. Let $T'_{ij}$ denote the propagation time for link i-j and $\frac{1}{\mu}$ be the average message size in bits. Then the objective function which Schwartz and Cheung define is the average message delay

$$T = \frac{1}{\gamma} \sum_{(i,j)} f_{ij} \left( \frac{1}{C_{ij} - f_{ij}} + \mu T'_{ij} \right) \tag{1.5}$$

$\gamma$ is the expected total external message arrivals/unit time and $C_{ij}$, the capacity of link (i-j) in bits/sec. If NN denotes the number of nodes in the network and $\gamma_{mn}$ is the expected number of arrivals/unit time at node m with destination n, the multicommodity, non-negativity, and capacity constraints on the flows are stated as follows:

$$\sum_{k=1}^{NN} f_{ik}^{(m,n)} - \sum_{\ell=1}^{NN} f_{\ell i}^{(m,n)} = \begin{cases} \gamma_{mn}/\mu & i = m \\ 0 & i \neq m, \ i \neq n \\ -\gamma_{mn}/\mu & i = n \end{cases} \tag{1.6}$$

$$f_{ij}^{(m,n)} \geq 0 \tag{1.7}$$

$$f_{ij} = \sum_{(m,n)} f_{ij}^{(m,n)} < C_{ij} \tag{1.8}$$

Defining a commodity flow vector f, whose entries are the $f_{ij}^{(m,n)}$'s, the conservation of flow constraints (1.6) may be expressed as

$$A_q f = b \ . \tag{1.9}$$

b is a vector whose entries are either 0, $\gamma_{mn}/\mu$, or $-\gamma_{mn}/\mu$. $A_q$ is a matrix consisting of submatrices corresponding to each (m,n) commodity.

$$A_q = \begin{pmatrix} A^{(1,2)} & & & & \\ & A^{(1,3)} & & & \\ & & \ddots \ A^{(m,n)} & & \\ & & & \ddots & A^{(NN,NN-1)} \end{pmatrix} \tag{1.10}$$

Given a flow $f^i$ satisfying constraints (1.6) through (1.8), we can obtain a feasible direction of descent for the objective function in (1.5) by projecting $\nabla T(f^i)$ onto the constraint surface defined in (1.9). Hence, Schwartz and Cheung propose the iteration

$$f^{i+1} = f^i - hP_q \nabla T(f^i) \ , \tag{1.11}$$

where h is a step size and $P_q$ a projection operator defined by

$$P_q = I - A_q^T(A_qA_q^T)^{-1}A_q \ . \tag{1.12}$$

The capacity constraint is handled implicitly by the penalty function in the objective function (1.5). Schwartz and Cheung derive an h' such that for $0 \leq h \leq h'$, the non-negativity of flows is preserved. The actual h is determined by appealing to the convexity of the objective function.

Now we recast the procedure of Schwartz and Cheung [7] to apply to a quasi-static routing situation. Suppose we redefine the objective function in (1.5) by not using Kleinrock's formula

$$T = \frac{1}{\gamma} \sum_{(i,j)} (D_{ij}(f_{ij}) + f_{ij}\mu T'_{ij}) \tag{1.13}$$

We assume that the current vector of flows is $f^i$ and we have available estimates for $\left.\frac{\partial D}{\partial f_{ij}}\right|_{f=f^i}$ and the $T'_{ij}$ are known constants. Hence, we can calculate $\nabla T(f^i)$ and apply the iteration

$$f^{i+1} = f^i - \eta h' P_q \nabla T(f^i) \ , \tag{1.14}$$

where h' is the upper bound on the step size to insure non-negativity of flows and $\eta$ is some scale factor $0 \leq \eta \leq 1$. Equation (1.14) could be termed a stochastic approximation algorithm since our values for $\left.\frac{\partial D_{ij}}{\partial f_{ij}}\right|_{f=f^i}$ would necessarily be inexact and hence our gradient $\nabla T(f^i)$ would only be an estimate.

In the algorithm offered by Cheung and Schwartz, the routing fractions are determined by knowledge of the commodity flows $f_{ij}^{(m,n)}$. In addition, the procedure is centralized, namely we assume the routing computations are performed at a special node and then communicated with the rest of the network. For many reasons, decentralized algorithms where the computation is distributed through the netwoik are more desirable. We next discuss a quasi-static routing algorithm derived by Gallager [3] that not only is decentralized, but works directly with the routing fractions.

Gallager uses a static model with stationary inputs and proposes an algorithm that seeks to minimize the total delay $D_T$ in the network specified by Eq. (1.1). He assumes the functions $D_{ik}(f_{ik})$ are increasing and convex U functions of the flow $f_{ik}$. Let $t_i(j)$ denote the total expected traffic at node i destined for node j and $r_i(j)$ the total expected external arrivals at node i destined for node j. The routing variables are defined as $\phi_{ik}(j)$, the fraction of traffic $t_i(j)$ that is routed over link (i-k). Conservation of flow for traffic with destination j at node i is expressed with these variables as

$$t_i(j) = r_i(j) + \sum_{\ell} t_{\ell}(j)\phi_{\ell i}(j) \; . \tag{1.15}$$

Hence, the link flows $f_{ik}$ are given as

$$f_{ik} = \sum_{j} t_i(j)\phi_{ik}(j) \; . \tag{1.16}$$

Gallager next derives two quantities, $\frac{\partial D_T}{\partial r_i(j)}$ and $\frac{\partial D_T}{\partial \phi_{ik}(j)}$, that appear in his algorithm and are used to characterize the conditions for a minimum of $D_T$ with respect to the $\phi$ variables.

$$\frac{\partial D_T}{\partial r_i(j)} = \sum_{k} \phi_{ik}(j)\left[D'_{ik}(f_{ik}) + \frac{\partial D_T}{\partial r_k(j)}\right] \tag{1.17}$$

$$\frac{\partial D_T}{\partial \phi_{ik}(j)} = t_i(j)\left[D'_{ik}(f_{ik}) + \frac{\partial D_T}{\partial r_k(j)}\right] \tag{1.18}$$

He then shows that a necessary condition for $\phi$ to achieve the minimum $D_T$ is

$$\frac{\partial D_T}{\partial \phi_{ik}(j)} = \begin{cases} \lambda_{ij} & ; \quad \phi_{ik}(j) > 0 \\ \geq \lambda_{ij} & ; \quad \phi_{ik}(j) = 0 \end{cases} \tag{1.19}$$

Gallager's algorithm consists of two parts: a protocol between nodes to calculate marginal delays $\frac{\partial D_T}{\partial r_i(j)}$ and keep track of a number of sets which he terms $B_i(j)$, and a procedure for up-dating the routing variables $\phi$. The procedure for adjusting the routing variables is defined as a mapping $\phi^1 = A(\phi)$ that attempts to move closer to the optimal equalibrium condition specified by (1.19). The sets $B_i(j)$ denote nodes for which $\phi_{ik}(j) = 0$ and the algorithm is not permitted to increase $\phi_{ik}(j)$ from zero. The way the $B_i(j)$'s are defined insures the "looplessness" of routes from any given source i to destination j, i.e., we can never go from a node i to some intermediate node m, back to node i, and finally to our destination node j.

We can see that the marginal delays $D'_{ik}(f_{ik})$ are fundamental to Gallager's procedure. While we could obtain them by differentiating Kleinrock's formula [3], it would then be necessary to estimate the flow $f_{ik}$. Hence, it is important to make the algorithm independent of Kelinrock's assumptions and estimate $D'_{ik}(f_{ik})$ directly by locally processing the queue for traffic using link (i-k).

We next review Carson Agnew's discussion in [1] of the ARPA scheme for routing, since his analysis reveals the reason for the sub-optimality of that method and he suggests an ARPA type routing strategy employing marginal delays. In the ARPA procedure, each node in the network maintains a table whose (i,j)-th entry is an estimate of the minimum time to reach the j-th node through the i-th neighbor. These estimates are based on the queue sizes at intermediate nodes, and hence the time it takes to empty those queues. When a message arrives addressed to the j-th destination, we look down column j and send the message to the neighbor with the smallest estimated delay.

Agnew introduces a simple single-commodity network with input flow $\lambda$ to be split into n routes with $\lambda_i$ arrivals/sec. each and stationary M/M/1 queues (exponential service requirements and exponential inter-arrival times all mutually independent). The flow, capacity, and non-negativity constraints together with the objective function corresponding to average message delay are represented as follows:

$$\lambda = \sum_{i=1}^{n} \lambda_i \tag{1.20}$$

$$0 \leq \lambda_i \leq \mu C_i \tag{1.21}$$

$$\overline{T} = \frac{1}{\lambda} \sum_{i=1}^{n} \lambda_i (\overline{T}_i + T_i') \tag{1.22}$$

The average message length is specified by $\frac{1}{\mu}$. $\overline{T}_i$ denotes the average total delay/message for the i'th queue and $T_i'$ denotes some remaining constant delay to get to the destination, such as a propagation time. Agnew's analysis is not relevant to more general networks since the $T_i'$ should be functions of $\lambda_i$.

For this simple one destination model, to implement the ARPA technique we would have a table with i'th entry $T_i' + (1+L_i)\mu C_i$, where $L_i$ denotes the number of messages in the queue and being serviced. After a long time, the distribution of traffic would be determined by the equalibrium conditions

$$\begin{aligned} \overline{T}_i + T_i' &= \overline{T}_j + T_j' \qquad \text{for} \quad \lambda_i, \lambda_j > 0 \\ \overline{T}_i + T_i' &> \overline{T}_j + T_j' \qquad \text{for} \quad \lambda_i = 0, \lambda_j > 0 \end{aligned} \tag{1.23}$$

However, these do not correspond to the conditions for obtaining a minimum $\overline{T}$ in (1.22), which Agnew shows to be

$$\begin{aligned} \frac{\partial}{\partial\lambda_i}(\lambda_i\overline{T}_i) + T_i' &= \frac{\partial}{\partial\lambda_j}(\lambda_j\overline{T}_j) + T_j' \qquad \lambda_i, \lambda_j > 0 \\ \frac{\partial}{\partial\lambda_i}(\lambda_i\overline{T}_i) + T_i' &> \frac{\partial}{\partial\lambda_j}(\lambda_j\overline{T}_j) + T_j' \qquad \lambda_i = 0, \lambda_j > 0 \end{aligned} \tag{1.24}$$

These differing equalibrium conditions (1.23) and (1.24) reflect the difference between system and user optimization.

For his single-commodity model, Agnew suggests a way to obtain an ARPA-like scheme that will approach the conditions for system optimality defined in (1.24). What we need is a quantity that has an expectation equal to the marginal delay $\frac{\delta}{\delta\lambda_i}(\lambda_i\overline{T}_i)$. For an M/M/1 queue he shows that $S_i = \frac{1}{\mu C_i}(1 + L_i)(1 + L_i/2)$ satisfies the desired property. Hence, he proposes that we do ARPA-type routing with revised table entries $S_i + T_i'$.

In [6] Segall proposes for a general network an ARPA-like routing strategy that uses marginal delays. Suppose that the objective function we wish to minimize is the toral delay accumulated per unit time over the network defined in (1.1). If we denote by $r_i(j)$ the average bit rate of external arrivals at node i with destination j, and given a small change $\delta r_i(j)$ in $r_i(j)$, where should we direct the extra traffic? Assuming we direct $\delta r_i(j)$ on a path P from i to j, the associated change in the total delay $D_T$ is given up to first-order terms by

$$\delta D_T = \sum_{(\ell,m)\epsilon P} \frac{\delta D_T}{\delta f_{\ell m}} \delta r_i(j) = \sum_{(\ell,m)\epsilon P} D'_{\ell m}(f_{\ell m})\delta r_i(j) \ . \tag{1.25}$$

Hence, this suggests choosing P so as to minimize $\frac{\delta D_T}{\delta r_i(j)}$. The routing procedure motivated by (1.25) is analogous to the ARPA-scheme, but the (i,j)-th location in the routing table would now list the estimate of the minimum $\frac{\delta D_T}{\delta r_i(j)}$ for directing extra traffic to destinaton j through neighbor i.

## 1.4 Formulation of Thesis as Queueing Theory Problem

In all of the quasi-static type routing algorithms presented, the incremental delays $D'_{ik}(f_{ik})$ are essential quantities. Rather than differentiate queueing theoretic formulae, with all their implied statistical assumptions, we propose estimating $D'_{ik}(f_{ik})$ by operating on the record of the queue associated with the outgoing link from node i to k. We are interested in finding recursive estimation procedures that process the queueing record and converge to $D'_{ik}(f_{ik})$ as the observation interval becomes sufficiently large. Hence, our problem can be formulated in the context of single server queueing theory. In our case the customers are identified with messages. The service time becomes the transmission time for the message due to the finite capacity of the communications link.

Hence, in this thesis we derive three algorithms that process the record of arrivals and departures of a queue to generate estimates of the derivative with respect to arrival rate of the average total delay per unit time. Since we make no assumption as to the exact form of the queue, we are interested in robust techniques that are as insensitive to statistical assumptions of the queueing process as possible. However, the performance of the algorithms can be analyzed only for relatively simple queues. Consequently, although the proposed algorithms can be applied in practice in very general situations, their explicit analysis is only done for queues like M/M/1, M/D/1, etc.

## 1.5 Summary of Thesis

The plan of the thesis is to examine the behavior of three estimation procedures, which we term the customer-addition, customer-removal, and time-contraction algorithms, for a variety of queues. The algorithms are described in detail in Section 2.1 but we may say now that each procedure corresponds to a different technique for imagining a hypothetical alteration of the queueing record to reflect a differential change in arrival rate $\delta\lambda$. In the customer-addition algorithm we conceptually add a customer at a random time in the observation period to simulate an increase in arrival rate. In the customer-removal algorithm we randomize the conceptual removal of customers from the queue, achieving the effect of a decrement in arrival rate. In the time-contraction procedure we redefine the arrival times of customers to simulate a compression in time scale and hence a differential increase in arrival rate.

Section 2 contains the main results of the thesis. First we give a detailed description of the way the notions for altering the queueing record indicated above are refined into the actual estimation procedures. This motivation leads then to the derivation of the algorithms and their realization in flow chart form is indicated. The analysis of the algorithms is performed in detail for special queues in Sections 2.3, 2.6, and 2.8. For the customer-addition and time-contraction procedures we are able to prove asymptotic unbiasedness for an M/D/1 queue. For the time-contraction and customer-removal procedures we define the calculation of the asymptotic bias as a power series in $\rho$, the utilization factor $\lambda\overline{x}$, in the case of M/G/1 queues. Employing this power series representation, we show that for an M/D/1 queue the bias for the customer-removal algorithm may only contain terms of third-order or higher in $\rho$. We also show that for an M/M/1 queue, both the bias for the time-contraction and customer-removal algorithm contain terms with powers of $\rho$ of all orders. Since the calculation of the variance associated with each of the estimators is too cumbersome, we derive Cramer-Rao bounds for each algorithm in the case of an M/D/1 queue. Since for purposes of practical implementation, routing calculations are secondary to the actual transmission of data, it is important to analyze and compare the storage and computation requirements of the three algorithms, which we accomplish in the final section of Section 2.

In Section 3 we present the results of simulating all three algorithms for an M/D/1 queue and the customer-removal and time contraction procedures for M/M/1, D/M/1, and U/M/1 queues. The performance measures we use to compare the algorithms are the relative bias and fractional rms error. Since the queueing record is segmented into busy and idle periods during which the server is occupied and unoccupied respectively, the variable we use to quantify the observation interval is the number of busy periods N included in the period. Hence, to investigate the convergence of the algorithms, for each queue of interest we present curves of the fractional rms error for N = 10, 100, and 1,000 busy periods and similarly present tables of the fractional bias. Employing our Cramer-Rao bounds for the M/D/1 case, we find that all three algorithms are both consistent and asymptotically efficient. We examine the robustness of the customer-removal and time-contraction algorithms by comparing their performance for M/D/1, M/M/1, D/M/1, and U/M/1 queues. The only significant difference in the two procedures performance occurs in the case of a D/M/1 queue, where the customer-removal procedure does worse.

# SECTION 2

# THREE ESTIMATION ALGORITHMS

## 2.1 Introduction

The main goal of this thesis is to propose and evaluate algorithms which process the record of a single server queueing system to estimate the derivative of the total delay/unit time with respect to arrival rate $\lambda$. The available record consists of exact knowledge of arrivals of customers to the queue and their departures after service is completed. Time is segmented into alternate intervals, busy periods, during which the server is occupied, and idle periods, when the server is free. The observation interval which is used to form our estimate consists of a number of busy periods and the intervening idle periods.

A simple thought-experiment motivates all three estimation algorithms. Consider our single-server queueing system with its average arrival rate of $\lambda$ customers per unit time. For a given observation period $T_E$, if we can compute the total system time S, i.e., the sum of all customer's service and waiting times, then the average delay/unit time is given by $D = S/T_E$. Suppose we could actually alter the input flow by some $\delta\lambda$. Then, on the basis of an earlier D, by computing D* for the next observation period, we can estimate the derivative of the total delay/unit time by calculating

$$\hat{D}' = \frac{D^* - D}{\delta\lambda} \; . \qquad (2.1)$$

However, in any actual queueing system it would be undesirable to change flows just for measurement purposes. Even if we could implement (2.1), the independent statistical fluctuations in D and D* would probably make it a very poor estimator. Hence, what we need is some mathematical formalism for an imaginary increment in flow $\delta\lambda$, which will allow us to compute the corresponding change in delay without actually perturbing the arrival rate.

According to intuition, an increase in arrival rate should result in additional customers entering the system. An extra customer arriving in a time interval $T_E$ with probability $\varepsilon$ will increase the effective rate by $\delta\lambda = \varepsilon/T_E$. If extra arrivals are mutually independent events, the probability of two or more customers will be of second-order in $\varepsilon$ and hence of second-order 'n $\delta\lambda$. Therefore, only the effect of a single extra arrival has to be considered explicitly. We also assume that the arrival time of the extra customer is uniformly distributed over the observation period $T_E$. In addition, in order to explicitly compute the change in total system time over the observation interval due to an extra arrival, we must assume that the additional customer has some known service requirement. These assumptions allow us to compute an expected increase in system time conditioned on the arrival of a new customer, and the resulting estimation procedure will be called the <u>customer-addition algorithm</u>.

In a second algorithm, an incremental decrease in the effective rate $\lambda$ is simulated. This is done by assuming that each customer arriving to the system is allowed to indeed enter the queue with probability $1 - \varepsilon$, and is eradicated with probability $\varepsilon$, independently from customer to customer. In this way we simulate an arrival process with rate $\lambda(1 - \epsilon)$. Hence $\epsilon$ is determined as follows:

$$\lambda(1 - \epsilon) = \lambda + \delta\lambda \tag{2.2}$$

$$\epsilon = - \frac{\delta\lambda}{\lambda} \tag{2.3}$$

We then estimate $\lambda$ by appealing to the law of large numbers. If $M'$ is the total number of customers in a period $T_E$, then $\lambda T_E \sim M'$. Hence, we use

$$\varepsilon = \frac{-T_E}{M'} \delta\lambda \tag{2.4}$$

Again, the probability of removal of two or more customers from the same period $T_E$ is second-order in $\delta\lambda$ and hence the reduction of total system time that has to be considered explicitly is due to removal of only one customer. This reduction consists of its own system time and the effect on other customers. The estimation procedure motivated by this second technique for making a "virtual" change in flow $\delta\lambda$ is termed the customer-removal algorithm.

A second characteristic that we associate with an increase in arrival rate, besides the fact that more customers appear in a given time period, is that there is less time between successive arrivals and therefore the customers are more "compressed" together. To make this argument quantitative, we note that an average arrival rate of $\lambda$ customers per second means an average inter-arrival time of $\frac{1}{\lambda}$. The change in the average inter-arrival time due to an increment in $\lambda$ is given by

$$\delta(\frac{1}{\lambda}) = - \frac{1}{\lambda} (\frac{\delta\lambda}{\lambda}) \quad . \tag{2.5}$$

If $\tau_n$ denotes the arrival time of the n-th customer, we define a new set of arrival times $\tau'_n = \tau_n(1 - \frac{\delta\lambda}{\lambda})$. If $E(\tau_{n+1} - \tau_n) = \frac{1}{\lambda}$, then $E(\tau'_{n+1} - \tau'_n) = \frac{1}{\lambda} (1 - \frac{\delta\lambda}{\lambda})$. This result is consistent with the change in inter-arrival time predicted by (2.5) due to an increase in flow $\delta\lambda$. We compute the resulting increment in system time by considering first the fact that customers arrive a little earlier and second, the fact that, given our fixed observation period $T_E$, the redefinition of arrival times results in the trailing edge of the interval being contracted and leaving a gap $\frac{\delta\lambda}{\lambda} T_E$ during which extra customers could have arrived. The estimation method suggested here is termed the _time-contraction algorithm_.

The present section contains the derivation and realization in flow chart form of the three algorithms. In addition, an extensive analysis of the algorithms is performed. We give a proof of the asymptotic unbiasedness of the _customer-addition_ and _time-contraction_ algorithms for a queue with Poisson arrivals and deterministic service requirements (M/D/1). The asymptotic bias behavior of the _customer-removal_ and _time-contraction_ algorithms for M/G/1 systems are examined as a power series in the utilization factor $\rho = \lambda\bar{x}$. For an M/D/1 queue we show explicitly that the _customer-removal_ algorithm is asymptotically unbiased up to the third power of $\rho$ and give a construction to prove asymptotic unbiasedness up to an arbitrary power. Also for an M/D/1 queue, Cramer-Rao bounds for any unbiased estimator of the delay gradient are derived. They will be used in Section 3 to determine the asymptotic efficiency of the algorithms applied to M/D/1 queues. We complete the present section by analyzing and comparing the storage and computation requirements of each method.

## 2.2 Derivation and Realization in Flow Diagram Form of Customer-Addition Algorithm

In the customer-addition algorithm we simulate an increase $\delta\lambda$ in the arrival rate. The following assumptions will be made:

1) The probability of an extra arrival in the interval $T_E$ is $\delta\lambda T_E$.
2) Each extra arrival is independent of all other arrivals.
3) The extra arrival is uniformly distributed over the interval $T_E$.
4) The service requirement of the extra customer is known; we denote it by x.

Let $T_k$ and $I_k$ denote the duration of the k-th busy and idle periods, respectively. Let $\delta S$ denote the increase in system time over N busy periods associated with the arrival of an extra customer. We let $\delta\hat{S}$ denote the expected increase in system time associated with an increase in arrival rate $\delta\lambda$ and conditioned on the record of arrivals and departures. By conditioning on the random arrival time t being in each $T_k$ and $I_k$ we can compute $\delta\hat{S}$ as

$$\delta\hat{S} = E(\delta S|\text{Queueing Record})T_E\delta\lambda$$

$$\delta\hat{S} = \left\{\sum_{k=1}^{N} E(\delta S|t\epsilon T_k, \text{ Queueing Record})\frac{T_k}{T_E} + \sum_{k=1}^{N-1} E(\delta S|t\epsilon I_k, \text{ Queueing Record})\frac{I_k}{T_E}\right\}T_E\delta\lambda \ . \qquad (2.6)$$

The $T_E\delta\lambda$ outside the brackets is the probability of an extra arrival. The increment in system time due to an increase in flow $\delta\lambda$ is zero if no additional arrival occurs. The factors $T_k/T_E$ and $I_k/T_E$ represent the probabilities of t being in the k-th busy and idle periods, respectively. This is a consequence of the assumption that t is uniformly distributed over $T_E$. Since we are interested in the derivative of the total delay/unit time with respect to the flow rate, our estimator is given by

$$\hat{D}' = \frac{1}{T_E} \frac{\delta\hat{S}}{\delta\lambda} . \tag{2.7}$$

We focus next on the calculation of $E(\delta S \mid t \epsilon T_k$, Queueing Record) and $E(\delta S \mid t \epsilon I_k$, Queueing Record). These expected increments in system time are composed of the average effect on existing customers plus the average system time of the additional customer. In considering additional arrivals in a busy period, we can distinguish between effects on the customers in that busy period and interactions with succeeding busy periods. First we examine the part of $E(\delta S \mid t \epsilon T_k$, Queueing Record), call it $\Delta S_k$, that comes from considering the k-th busy period in isolation. To facilitate discussion, the following notation is defined.

$\tau_i \triangleq$ Arrival time of i-th customer in the busy period (relative to the start of the busy interval)

$x_i \triangleq$ Service requirement of the i-th customer

$S_i \triangleq$ System time of the i-th customer

$S_i^* \triangleq$ System time the i-th customer would have had if an additional customer arrived at time t

$M \triangleq$ Number of customers in the busy period

$\Delta \triangleq$ System time of the additional customer

$t \triangleq$ Arrival time of the additional customer

$T \triangleq$ Duration of the busy period. (2.8)

We break up the calculation of the expected increase in system time into expectations conditioned on an arrival in the interval $[\tau_i, \tau_{i+1}]$ for i = 1 ... M. $\tau_1$ is zero and $\tau_{M+1}$ is defined as the duration of the busy period T. Then

$$\Delta S = E\{\Sigma S_n^* + \Delta - \Sigma S_n | t \epsilon [0,T]\}$$

$$= \sum_{i=1}^{M} E\{S_n^* + \Delta - \Sigma S_n | t \epsilon [\tau_i, \tau_{i+1}]\} \Pr\{t \epsilon [\tau_i, \tau_{i+1}] | t \epsilon [0,T]\} \tag{2.9}$$

By the assumption of uniformly distributed arrival time

$$\Pr\ \{t\epsilon[\tau_i,\tau_{i+1}]\,|t\epsilon[0,T]\} = \frac{\tau_{i+1} - \tau_i}{T}\ . \tag{2.10}$$

The system time of a given customer is equal to his service plus waiting time. The waiting time is equal to the sum of the service requirements of those who entered the busy period before him minus his arrival time. Hence, the system time of the n-th customer is given by

$$S_n = \sum_{i=1}^{n} x_i - \tau_n\ . \tag{2.11}$$

Now consider the new total system time due to an arrival at time $t\epsilon[\tau_i,\tau_{i+1}]$

$$\sum_{n=1}^{M} S_n^* + \Delta = \sum_{n=1}^{i} S_n + \sum_{n=i+1}^{M} (S_n + x) + \left(\sum_{n=1}^{i} x_n + x - t\right). \tag{2.12}$$

The first term represents the first i customers whose system times are unaffected by the new arrival. The second term shows that each customer ahead of the new arrival will suffer an additional delay x. The final term represents the system time Δ of the new customer. Since conditioned on being in $[\tau_i,\tau_{i+1}]$, the random variable t is uniformly distributed over that interval, taking appropriate conditional expectations in (2.12) yields

$$E\,\{\Sigma\, S_n^* + \Delta - \Sigma\, S_n \,|\, t\epsilon[\tau_i,\tau_{i+1}]\} = x + \sum_{n=1}^{i} x_n - \left(\frac{\tau_{i+1} + \tau_i}{2}\right) + (M-i)x \;. \qquad (2.13)$$

Substituting (2.13) and (2.10) into (2.9) results in the following expression for $\Delta S$:

$$\Delta S = x + \sum_{i=1}^{M} \left\{ \sum_{n=1}^{i} x_n - \frac{\tau_i + \tau_{i+1}}{2} \right\} \left\{ \frac{\tau_{i+1} - \tau_i}{T} \right\} + \sum_{i=1}^{M} (M-i)x \left\{ \frac{\tau_{i+1} - \tau_i}{T} \right\}. \qquad (2.14)$$

The first and second terms represent the service requirement of the new customer and his expected waiting time, respectively. The third term is the expected delay suffered by the existing customers due to the new arrival.

For the special case where all the service requirements are the same, namely $x_i = x$, Eq. (2.14) simplifies. This can be seen by direct substitution of $x_i = x$ into Eq. (2.14), but it will be illuminating to re-examine the computation that led to (2.14). Suppose our extra customer arrives at time t after the k-th and before the (k+1)-st customer. The M-k customers ahead of the new arrival suffer an additional delay x. According to the rule described by (2.11), the system time of the new customer is given by (k+1)x-t. Hence, for an arrival at time t, there is an additional system time Mx-t+x. This amount of time is equivalent to allowing the extra customer to wait till the end of the busy period and then be served. Since t is uniformly distributed over an interval Mx, we have $\bar{t} = Mx/2$. Therefore, $\Delta S$ is given by

$$\Delta S = \frac{1}{2} Mx+x \ . \tag{2.15}$$

We complete the calculation of $E(\delta S \mid t \epsilon T_k$, Queueing Record) by looking at the additional system time that may result from one busy period overlapping onto another. The following will hold for arbitrary service times $x_i$. No matter where an additional customer arrives in the k-th busy period, that period will be extended by the extra service time x. The value of x relative to the following idle period durations will determine the number of succeeding busy periods that will be affected by an arrival in $T_k$. However, the number is always finite. If $x \leq I_k$, no following busy periods suffer additional delay. If $I_k \leq x \leq I_k + I_{k+1}$, only the (k+1)-st busy period is affected. The exact effect on a given busy period j in the future, depends on how much an arrival in $T_k$ causes the (j-1)-st busy period to overlap onto the j-th busy period. For example, if $x > I_k$ then each customer in $T_{k+1}$ will suffer an additional delay $(x-I_k)$. Letting $M_k$ denote the number of customers served in the k-th busy period, the preceding reasoning leads to the following rule for computing $E(\delta S \mid t\epsilon T_k$, Queueing Record):

$$E(\delta S \mid t \epsilon T_k, \text{Queueing Record}) = \begin{cases} \Delta S_k & x \leq I_k \\ \Delta S_k + M_{k+1}(x - I_k) & I_k \leq x \leq I_k + I_{k+1} \\ \cdot & \\ \cdot & \\ \cdot & \\ \Delta S_k + \sum\limits_{j=1}^{\ell+1} M_{k+j}\left(x - \sum\limits_{m=0}^{j-1} I_{k+m}\right) & \sum\limits_{j=0}^{\ell} I_{k+j} \leq x \leq \sum\limits_{j=0}^{\ell+1} I_{k+j} \end{cases} \tag{2.16}$$

The last relation in (2.16) refers to the case when an arrival in $T_k$ affects ($\ell$+1) busy periods into the future.

To complete our description of the customer-addition algorithm, we must now evaluate the average increase in system time $E(\delta S \mid t\epsilon I_k$, Queueing Record) associated with arrivals in idle periods. The effect of an arrival in $I_k$ on the j-th busy period again depends on how much the (j-1)-st busy period slides onto the j-th busy interval. As the effect on customers in $T_j$ is not independent of the exact arrival time, we must average over all times t in the k-th idle period. Let $\alpha_{k+1}$ denote the time instant corresponding to the beginning of the (k+1)-st busy period. If

$x \leq I_{k+1}$, an arrival in $I_k$ affects only the (k+1)-st busy period. Hence $E(\delta S | t \epsilon I_k$, Queueing Record) is computed as

$$E(\delta\hat{S} | t\epsilon I_k, \text{Queueing Record}) = x + \int_{\max\{\alpha_{k+1} - I_k,\ \alpha_{k+1} - x\}}^{\alpha_{k+1}} M_{k+1}(x + t - \alpha_{k+1}) \frac{dt}{I_k}$$

$$\text{for } x \leq I_{k+1} . \qquad (2.17)$$

Here, $x + t - \alpha_{k+1}$ is the amount that the additional customers service time overlaps onto the (k+1)-st busy period. The "max" is necessary in the lower limit of integration since our arrival t must be in the interval $I_k$. If $x < I_k$, then $\alpha_{k+1} - x$ represents the earliest time at which an arrival can occur and influence the (k+1)-st busy period. By a simple change of variable, $t' = t - \alpha_{k+1}$, the dependence on $\alpha_{k+1}$ disappears. The relation (2.17) may be generalized to

$$E(\delta S | t\epsilon I_k, \text{Queueing Record}) = x + \sum_{j=1}^{\ell+1} \int_{\max\{-I_k, S_j - x\}}^{0} M_{k+j}(x + t' - S_j) \frac{dt'}{I_k}$$

$$\text{for } \sum_{j=1}^{\ell} I_{k+j} \leq x \leq \sum_{j=1}^{\ell+1} I_{k+j} . \qquad (2.18)$$

$$S_j = \begin{cases} 0 & j=1 \\ \sum_{m=1}^{j-1} I_{k+m} & j \neq 1 \end{cases} \tag{2.19}$$

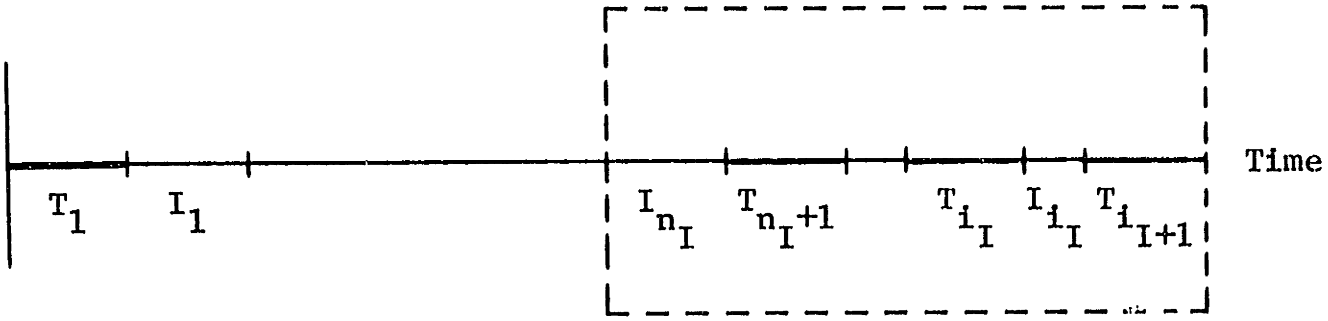
Equation (2.18) corresponds to the case when an arrival in $I_k$ influences $\ell+1$ succeeding busy periods. The value of the integrals in the summation are given by

$$\int_{\max\{-I_k, S_j-x\}}^{0} M_{k+j}(x+t'-S_j) \frac{dt'}{I_k} = \begin{cases} M_{k+j}(x - \frac{1}{2} I_k - S_j) & -I_k > S_j - x \\ \frac{M_{k+j}}{2I_k} (x-S_j)^2 & S_j - x > -I_k \end{cases} \tag{2.20}$$

Employing (2.18) and (2.16) in (2.6) and (2.7) we can conceive of a processor which up-dates an estimate for the delay gradient at the end of each busy period. Let $\ell_k$ denote the time from the start of the observation period to the end of the k-th busy period. Let $\hat{D}'_{(k+1)}$ denote the estimate for the delay gradient based on k+1 busy periods. Let $\Delta_k$ be the incremental expected delay suffered by the (k+1)-st busy period due to an additional arrival in the current queueing record. Hence, the up-dating at the end of the (k+1)-st busy period assumes the form

$$\hat{D}'_{(k+1)} = \frac{\ell_k}{\ell_{k+1}} \hat{D}'_{(k)} + \Delta_k . \qquad (2.21)$$

Since x is finite, we look back in time a finite number of busy and idle periods to compute $\Delta_k$. Suppose that the busy and idle periods are numbered consecutively from the beginning of the observation time and let $i_I$ denote the index of the most recent idle period. Let $n_I$ denote the index of the first idle period at which an arrival with service requirement x can influence the current $(i_I+1)$-st, busy period. Hence, we need only store idle period and busy period information for $(I_{n_I} \ldots I_{i_I+1})$ and $(I_{n_I+1} \ldots T_{i_I+1})$. At the end of every idle period, $n_I$ must be up-dated to reflect how far back we must look to compute effects on the newest busy period. Additional simplification is possible due to the fact that both $E(\delta S|t\epsilon I_k$, Queueing Record) and $E(\delta S|t\epsilon I_k$, Queueing Record) include an x term. Since the incremental expected system time due to an arrival in either a busy or idle period is weighted in the estimator by the probability for arrival in that time slot, we can add x at the end. These ideas are realized in the flow diagrams of Figures 2.2 and 2.3. Figure 2.1 pictures the relationship between $n_I$ and $i_I$, and defines the variables that appear in the flow diagrams of Figures 2.2 and 2.3.



$i_I+1$ = Number of the current busy period.

$n_I$ = Number of the first idle period at which an arrival with service requirement x can cause customers in busy period $i_I+1$ to suffer additional delay.

$\ell, \ell'$ = Variables denoting the elapsed time from the beginning of the observation period to the end of busy periods $i_I$, $i_{I+1}$, respectively.

$\Delta$, S, C = Auxilliary variables.

q = Index of current busy period $i_I+1$.

$\tau_n^i$ = Arrival time of n-th customer in i-th busy period relative to the beginning of that busy period.

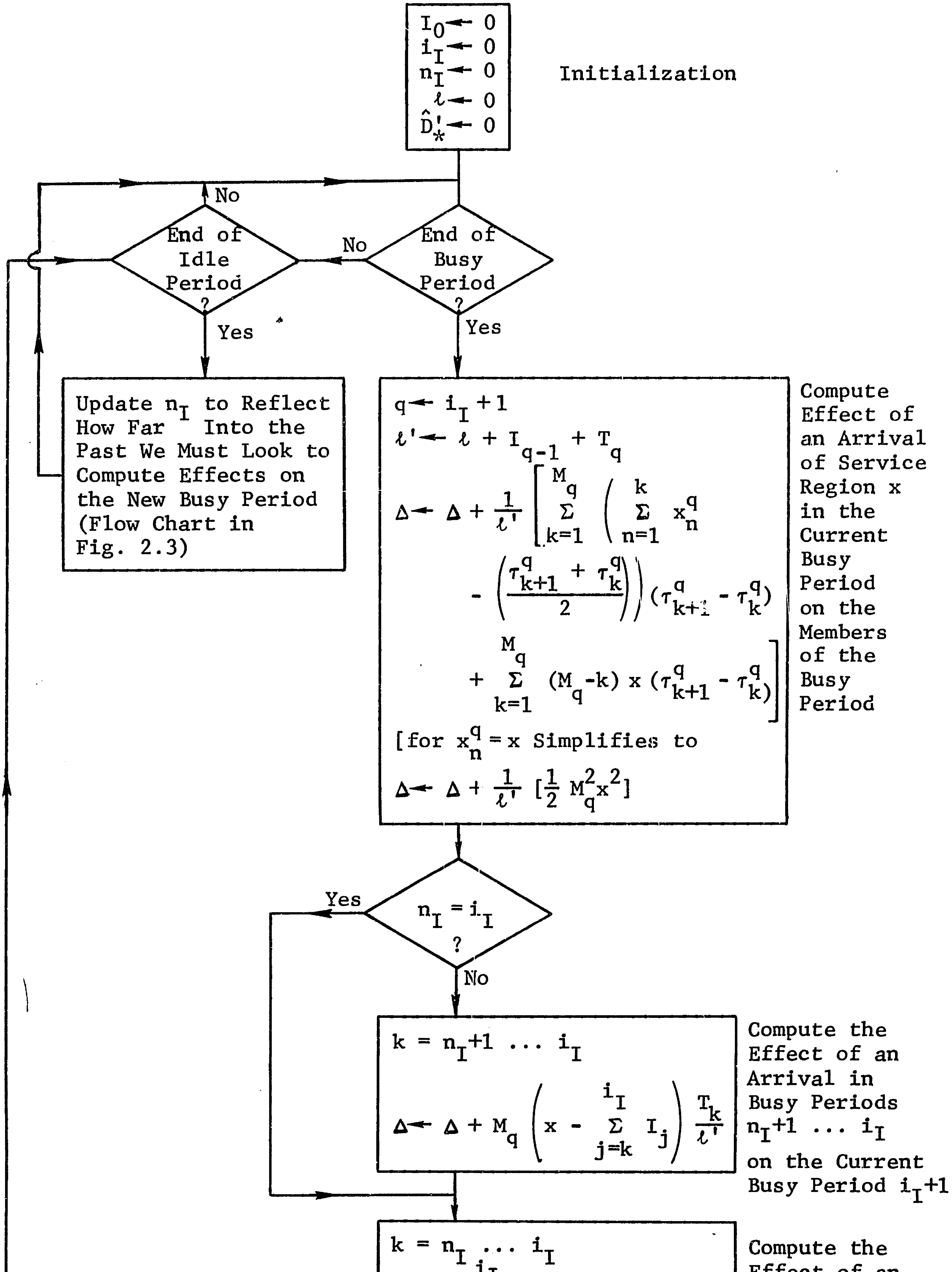
$x_n^i$ = Service requirement of n-th customer in the i-th busy period.

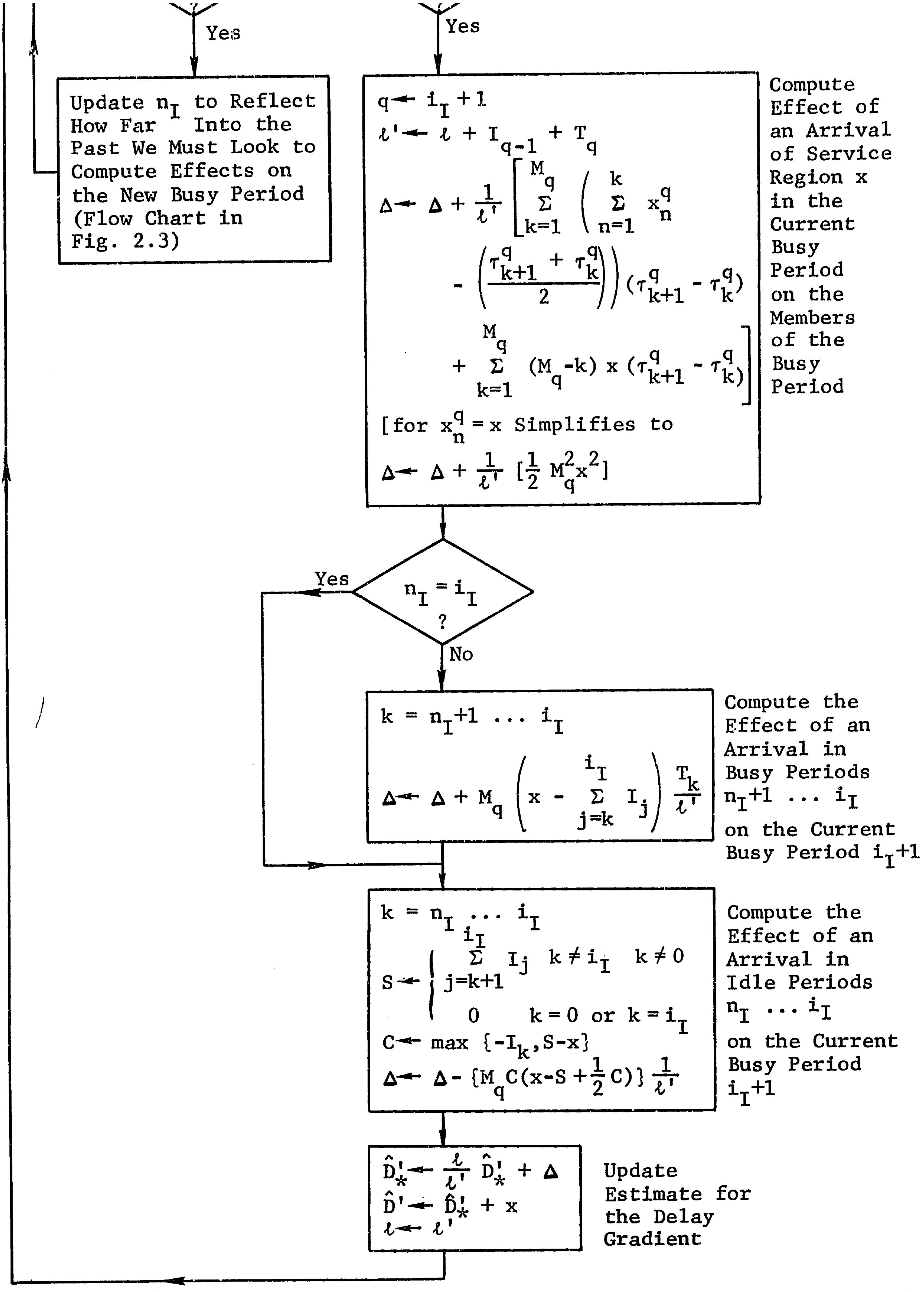
$M_i$ = Number of customers in the i-th busy period.

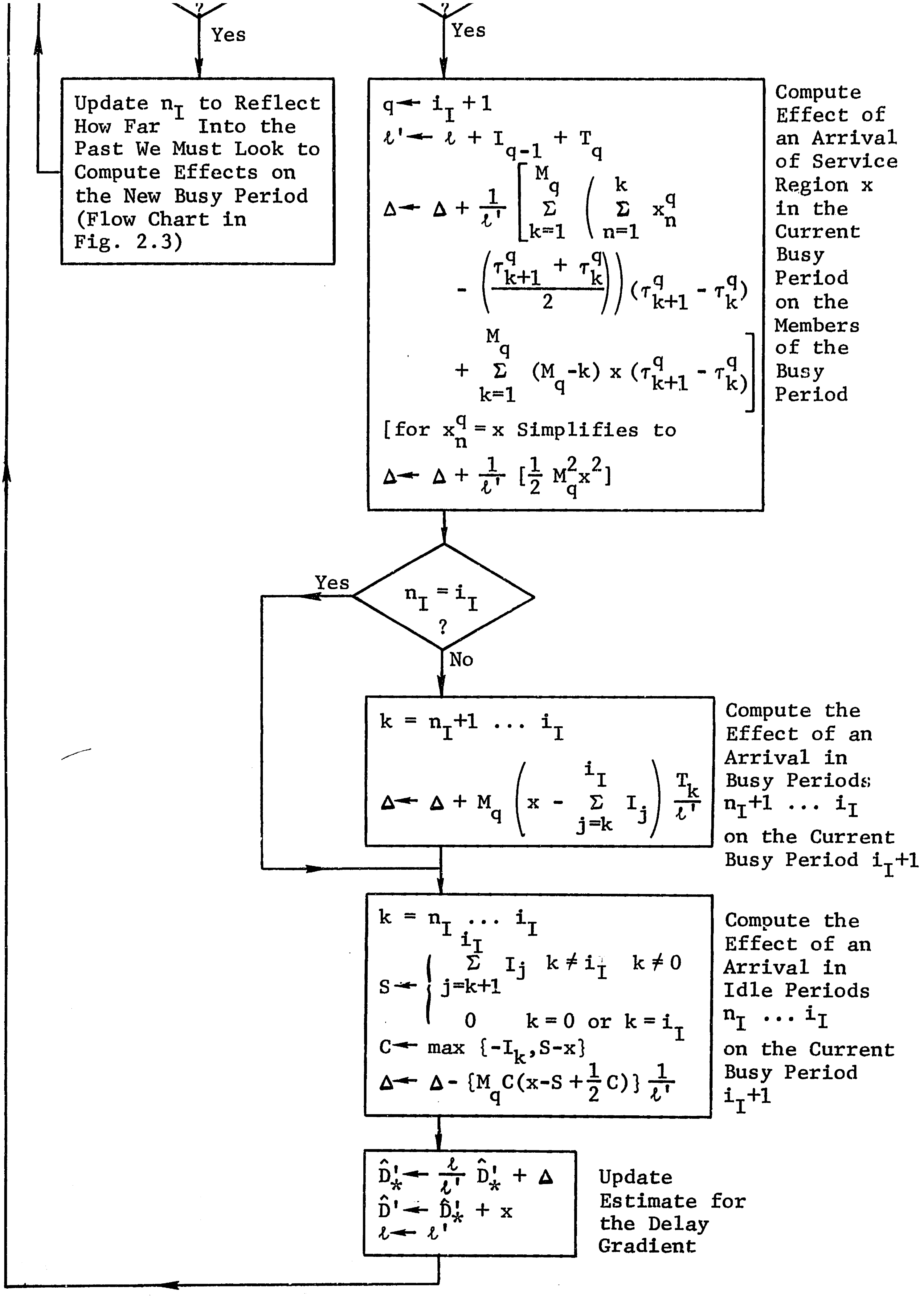
$\hat{D}'$ = Delay gradient estimator.

$\hat{D}'_*$ = Delay gradient estimator minus service requirement of additional customer x.

Figure 2.1 Queueing Record Structure and Definition of Variables Relevant to Flow Chart Realization of Customer-Addition Algorithm


$I_0 \leftarrow 0$
$i_I \leftarrow 0$
$n_I \leftarrow 0$
$\ell \leftarrow 0$
$\hat{D}'_* \leftarrow 0$
Initialization
No
End of Idle Period ?
No
End of Busy Period ?
Yes
Yes
Update $n_I$ to Reflect How Far Into the Past We Must Look to Compute Effects on the New Busy Period (Flow Chart in Fig. 2.3)
$q \leftarrow i_I + 1$
$\ell' \leftarrow \ell + I_{q-1} + T_q$
$\Delta \leftarrow \Delta + \frac{1}{\ell'} \left[ \sum_{k=1}^{M_q} \left( \sum_{n=1}^{k} x_n^q - \left( \frac{\tau_{k+1}^q + \tau_k^q}{2} \right) \right) (\tau_{k+1}^q - \tau_k^q) + \sum_{k=1}^{M_q} (M_q - k) x (\tau_{k+1}^q - \tau_k^q) \right]$
[for $x_n^q = x$ Simplifies to
$\Delta \leftarrow \Delta + \frac{1}{\ell'} [\frac{1}{2} M_q^2 x^2]$
Compute Effect of an Arrival of Service Region x in the Current Busy Period on the Members of the Busy Period
$n_I = i_I$ ?
Yes
No
$k = n_I+1 \ldots i_I$
$\Delta \leftarrow \Delta + M_q \left( x - \sum_{j=k}^{i_I} I_j \right) \frac{T_k}{\ell'}$
Compute the Effect of an Arrival in Busy Periods $n_I+1 \ldots i_I$ on the Current Busy Period $i_I+1$
$k = n_I \ldots i_I$
Compute the Effect of an


Yes
Yes
Update $n_I$ to Reflect How Far Into the Past We Must Look to Compute Effects on the New Busy Period (Flow Chart in Fig. 2.3)
$q \leftarrow i_I + 1$
$\ell' \leftarrow \ell + I_{q-1} + T_q$
$\Delta \leftarrow \Delta + \frac{1}{\ell'} \left[ \sum_{k=1}^{M_q} \left( \sum_{n=1}^{k} x_n^q - \left( \frac{\tau_{k+1}^q + \tau_k^q}{2} \right) \right) (\tau_{k+1}^q - \tau_k^q) + \sum_{k=1}^{M_q} (M_q - k) \; x \; (\tau_{k+1}^q - \tau_k^q) \right]$
[for $x_n^q = x$ Simplifies to
$\Delta \leftarrow \Delta + \frac{1}{\ell'} [\frac{1}{2} M_q^2 x^2]$
Compute Effect of an Arrival of Service Region x in the Current Busy Period on the Members of the Busy Period
$n_I = i_I$ ?
Yes
No
$k = n_I + 1 \; \ldots \; i_I$
$\Delta \leftarrow \Delta + M_q \left( x - \sum_{j=k}^{i_I} I_j \right) \frac{T_k}{\ell'}$
Compute the Effect of an Arrival in Busy Periods $n_I + 1 \; \ldots \; i_I$ on the Current Busy Period $i_I + 1$
$k = n_I \; \ldots \; i_I$
$S \leftarrow \begin{cases} \sum_{j=k+1}^{i_I} I_j & k \neq i_I \quad k \neq 0 \\ 0 & k = 0 \text{ or } k = i_I \end{cases}$
$C \leftarrow \max \{ -I_k, S - x \}$
$\Delta \leftarrow \Delta - \{ M_q C (x - S + \frac{1}{2} C) \} \frac{1}{\ell'}$
Compute the Effect of an Arrival in Idle Periods $n_I \; \ldots \; i_I$ on the Current Busy Period $i_I + 1$
$\hat{D}_*' \leftarrow \frac{\ell}{\ell'} \hat{D}_*' + \Delta$
$\hat{D}' \leftarrow \hat{D}_*' + x$
$\ell \leftarrow \ell'$
Update Estimate for the Delay Gradient


?
Yes
Update $n_I$ to Reflect How Far Into the Past We Must Look to Compute Effects on the New Busy Period (Flow Chart in Fig. 2.3)
?
Yes
$q \leftarrow i_I + 1$
$\ell' \leftarrow \ell + I_{q-1} + T_q$
$\Delta \leftarrow \Delta + \frac{1}{\ell'}\left[\sum_{k=1}^{M_q}\left(\sum_{n=1}^{k} x_n^q - \left(\frac{\tau_{k+1}^q + \tau_k^q}{2}\right)\right)(\tau_{k+1}^q - \tau_k^q) + \sum_{k=1}^{M_q}(M_q - k)\, x\, (\tau_{k+1}^q - \tau_k^q)\right]$
[for $x_n^q = x$ Simplifies to
$\Delta \leftarrow \Delta + \frac{1}{\ell'}\left[\frac{1}{2} M_q^2 x^2\right]$
Compute Effect of an Arrival of Service Region x in the Current Busy Period on the Members of the Busy Period
$n_I = i_I$ ?
Yes
No
$k = n_I + 1 \ldots i_I$
$\Delta \leftarrow \Delta + M_q \left(x - \sum_{j=k}^{i_I} I_j\right)\frac{T_k}{\ell'}$
Compute the Effect of an Arrival in Busy Periods $n_I+1 \ldots i_I$ on the Current Busy Period $i_I+1$
$k = n_I \ldots i_I$
$S \leftarrow \begin{cases} \sum_{j=k+1}^{i_I} I_j & k \neq i_I \quad k \neq 0 \\ 0 & k = 0 \text{ or } k = i_I \end{cases}$
$C \leftarrow \max\{-I_k, S - x\}$
$\Delta \leftarrow \Delta - \{M_q C(x - S + \frac{1}{2} C)\}\frac{1}{\ell'}$
Compute the Effect of an Arrival in Idle Periods $n_I \ldots i_I$ on the Current Busy Period $i_I+1$
$\hat{D}_*' \leftarrow \frac{\ell}{\ell'}\hat{D}_*' + \Delta$
$\hat{D}' \leftarrow \hat{D}_*' + x$
$\ell \leftarrow \ell'$
Update Estimate for the Delay Gradient

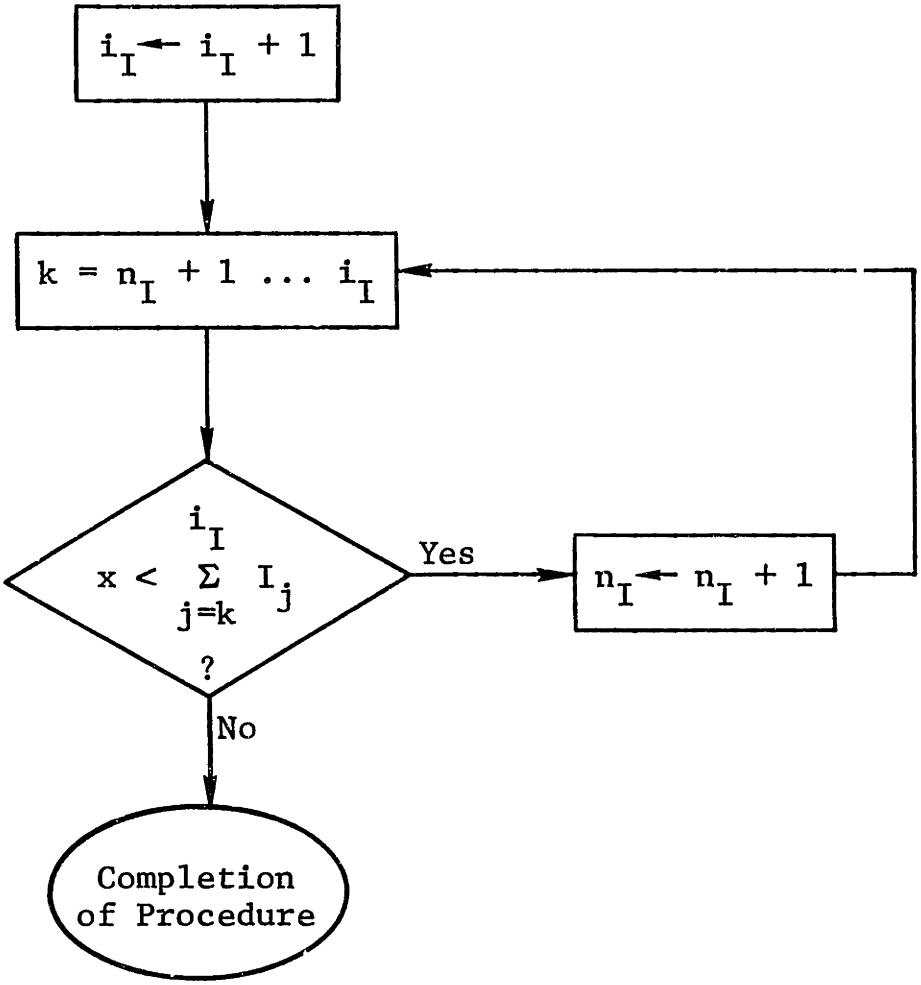


Figure 2.3 Up-Dating Procedure for $n_I$

We now pose the question of whether the customer addition algorithm can be generalized to be applicable to a wider class of queues than those where all customers have the same service requirement x. As formulated, the algorithm is limited by an assumption of a fixed service requirement x for the additional customer. Hence, we can conceive of extending the algorithm by doing a final averaging over x,

$$\hat{D}' = \frac{1}{T_E} \int_{x=0}^{\infty} B(x) \frac{\delta \hat{S}}{\delta \lambda} (x) \; dx \tag{2.22}$$

B(x) denotes the service time density and $\frac{\delta \hat{S}}{\delta \lambda}$ (x) refers to the unnormalized incremental delay as a function of the assumed extra customer service requirement x. While possible in principle, the scheme implied by (2.22) is unacceptable for practical reasons. The evaluation of (2.22) necessitates saving the entire queueing record and doing all our processing at the end.

Only when the service time density consists of a discrete set of values would it be reasonable to implement (2.22). In this situation B(x) is given as a train of impulses.

$$B(x) = \sum_{k=1}^{L} \rho_k \delta(x-x_k) \tag{2.23}$$

Then (2.22) would become

$$\hat{D}' = \frac{1}{T_E} \sum_{k=1}^{L} \rho_k \frac{\delta\hat{S}}{\delta\lambda} (x_k) \ . \tag{2.24}$$

Hence, for each $x_k$, k = 1 ... L we would process the queueing record in parallel, employing the algorithm given in the flow diagram of Figure 2.2.

## 2.3 Proof of Asymptotic Unbiasedness of Customer-Addition Algorithm for an M/D/1 Queue

We now examine the bias of the customer-addition algorithm as the number of busy periods in the observation period, N, become unbounded. For the algorithm to be asymptotically unbiased we must prove that

$$\lim_{N\to\infty} \left\{ E \left\{ \frac{1}{T_E} \ \frac{\delta\hat{S}}{\delta\lambda} \right\} \right\} = \frac{\partial D}{\partial\lambda} \ , \tag{2.25}$$

where D is the average total delay/unit time.

To check (2.25) we must first define the quantity $\frac{\partial D}{\partial\lambda}$ . The average total delay/unit time D is equal to λ times the average total delay/customer $D_c$. Hence, $\frac{\partial D}{\partial\lambda}$ may be expressed in terms of $\frac{\partial D_c}{\partial\lambda}$ as

$$\frac{\partial D}{\partial\lambda} = D_c + \lambda \frac{\partial D_c}{\partial\lambda} \ . \tag{2.26}$$

The average total delay/customer is expressible in terms of the average service time $\overline{x}$ and the average waiting time $\overline{w}$ as

$$D_c = \overline{x} + \overline{w} . \tag{2.27}$$

Hence, $\frac{\partial D}{\partial \lambda}$ can be reformulated in terms of the average waiting time and service requirement.

$$\frac{\partial D}{\partial \lambda} = \overline{x} + \overline{w} + \lambda \frac{\partial \overline{w}}{\partial \lambda} \tag{2.28}$$

We can evaluate the above expression for all queues for which an explicit form of the waiting time distribution is available.

Reviewing the assumptions inherent in the customer-addition algorithm, we can expect that the procedure will be asymptotically unbiased in the case of an M/D/1 queue. The descriptor "M/D/1" means the arrival process is Poisson, and the service requirements deterministic. Since all customers in an M/D/1 queue have the same service requirement, the assumption that the additional customer has a fixed service time is harmless. The two other assumptions, uniform arrival time distribution for the extra customer and the probability density for the arrival of extra customers, are both consistent with a Poisson arrival process. For a Poisson process the probability density for the time of occurrence of the i-th event given $k \geq i$ events did occur in $[0,T]$

is uniform on the interval. Let $p(k,\lambda)$ denote the probability of k arrivals in an interval T given that the arrival rate is $\lambda$. If we let $(\delta\lambda T)^k$ be the probability that k additional customers arrive in an interval T due to an increase in rate $\delta\lambda$, $p(k,\lambda+\delta\lambda)$ must satisfy

$$p(k,\lambda+\delta\lambda) = \sum_{i=1}^{k} p(k-i,\lambda)(\delta\lambda T)^i + p(k,\lambda)\left(1 - \frac{\delta\lambda T}{1-\delta\lambda T}\right). \tag{2.29}$$

After some manipulation, dividing both sides by $\delta\lambda$ and taking the limit as $\delta\lambda$ approaches zero, we obtain

$$\frac{\partial p(k,1)}{\partial\lambda} = T\{p(k-1,\lambda) - p(k,\lambda)\} . \tag{2.30}$$

By direct substitution we can verify that the Poisson process formula for the probability of occurrence of k events in a time T given below satisfies (2.30).

$$p(k,\lambda) = \frac{(\lambda T)^k i^{-\lambda T}}{k!} \tag{2.31}$$

Motivated by the preceding arguments, we proceed to prove that the customer-addition algorithm is asymptotically unbiased for an M/D/1 queue. Since for an M/D/1 queue, the average waiting time is given in [4] as

$$\overline{w} = \frac{\rho x}{2(1-\rho)} , \tag{2.32}$$

where $\rho = \lambda x$ is the utilization factor, formula (2.28) dictates that we must prove the expectation of our estimate (2.7) converges as $N \to \infty$ to

$$\frac{\delta D}{\delta \lambda} = x + \frac{\rho x}{2(1-\rho)} + \frac{\rho x}{2(1-\rho)^2} . \tag{2.33}$$

Since by the law of large numbers we have $\lim_{N\to\infty} T_E = \lim_{N\to\infty} \frac{N\overline{M}}{\lambda}$, interchanging the limit and expectation operations in (2.25) we must show that

$$\lambda \lim_{N\to\infty} \left( \frac{1}{N\overline{M}} E \frac{\delta \hat{S}}{\delta \lambda} \right) = \frac{\delta D}{\delta \lambda} . \tag{2.34}$$

Employing (2.6), we can break the problem of computing $E \frac{\delta \hat{S}}{\delta \lambda}$ into evaluating the expectation of two types of terms as below.

$$\begin{aligned} E \frac{\delta \hat{S}}{\delta \lambda} = & \sum_{k=1}^{N} E\{E(\delta S \mid t \epsilon T_k, \text{ Queueing Record}) T_k\} \\ & + \sum_{k=1}^{N-1} E\{E(\delta S \mid t \epsilon I_k, \text{ Queueing Record}) I_k\} \end{aligned} \tag{2.35}$$

The terms $E(\delta S \mid t \epsilon T_k$, Queueing Record) and $E(\delta S \mid t \epsilon I_k$, Queueing Record) are given by Eqs. (2.16) and (2.18), respectively. Since the queueing record is given by the number of customers served in each busy period $(M_1 \ldots M_N)$ and the idle period durations $(I_1 \ldots I_{N-1})$, we further break the calculations by first conditioning on $(M_1 \ldots M_N)$ and averaging over $(I_1 \ldots I_{N-1})$ and then averaging over $(M_1 \ldots M_N)$ as below.

$$E\{E(\delta S \,|\, t\epsilon T_k, \text{ Queueing Record})T_k\}$$

$$= E\{E[E(\delta S \,|\, t\epsilon T_k, M_1 \ldots M_N, I_1 \ldots I_{N-1})T_k / M_1 \ldots M_N]\} \quad (2.36)$$

$$E\{E(\delta S \,|\, t\epsilon I_k, \text{ Queueing Record})I_k\}$$

$$= E\{E[E(\delta S \,|\, t\epsilon I_k, M_1 \ldots M_N, I_1 \ldots I_{N-1})I_k / M_1 \ldots M_N]\} \quad (2.37)$$

We organize our calculations by first computing $\delta S_{T_1}$ and $\delta S_{I_1}$ defined below as

$$\delta S_{T_1} \triangleq E[E(\delta S \,|\, t\epsilon T_1, M_1 \ldots M_N, I_1 \ldots I_{N-1})T_1 \,|M_1 \ldots M_N] \quad (2.38)$$

and

$$\delta S_{I_1} \triangleq E[E(\delta S \,|\, t\epsilon I_1, M_1 \ldots M_N, I_1 \ldots I_{N-1})I_1 (M_1 \ldots M_N] , \quad (2.39)$$

and then generalizing our result to $\delta S_{T_k}$ and $\delta S_{I_k}$ specified similarly. Finally, we sum $\delta S_{T_k}$ and $\delta S_{I_k}$ over all k's and average over $(M_1 \ldots M_N)$.

<u>Calculation of $\delta S_{T_1}$</u>

We calculate $\delta S_{T_1}$ by first illustrating the thinking involved when the number of busy periods included in the observation period, N, is three and then generalizing the procedure. Figure 2.4 depicts the queueing record for N=3 and a partitioning of $(I_1,I_2)$ space into three regions: $R_1$, $R_2$, and $R_3$. According to (2.16), $T_1E(\delta S|t\epsilon T_1$, Queueing Record) is given by

$$T_1E(\delta S|t\epsilon T_1, \begin{matrix}\text{Queueing}\\ \text{Record}\end{matrix}) = \begin{cases} \Delta S_1T_1 & x \leq I_1 \text{ or} \\ & (I_1,I_2)\epsilon R_1 \\ \Delta S_1T_1 + (x-I_1)M_2T_1 & I_1 \leq x \leq I_2 \text{ or} \\ & (I_1,I_2)\epsilon R_2 \\ \Delta S_1T_1 + (x-I_1)M_2T_1 & x > I_1 + I_2 \text{ or} \\ \qquad + (x-I_1-I_2)M_3T_1 & (I_1,I_2)\epsilon R_3 \end{cases} \tag{2.40}$$

The key to computing the desired expectation is in noting which region of the $(I_1,I_2)$ space corresponds to each term. $\Delta S_1T_1$ is

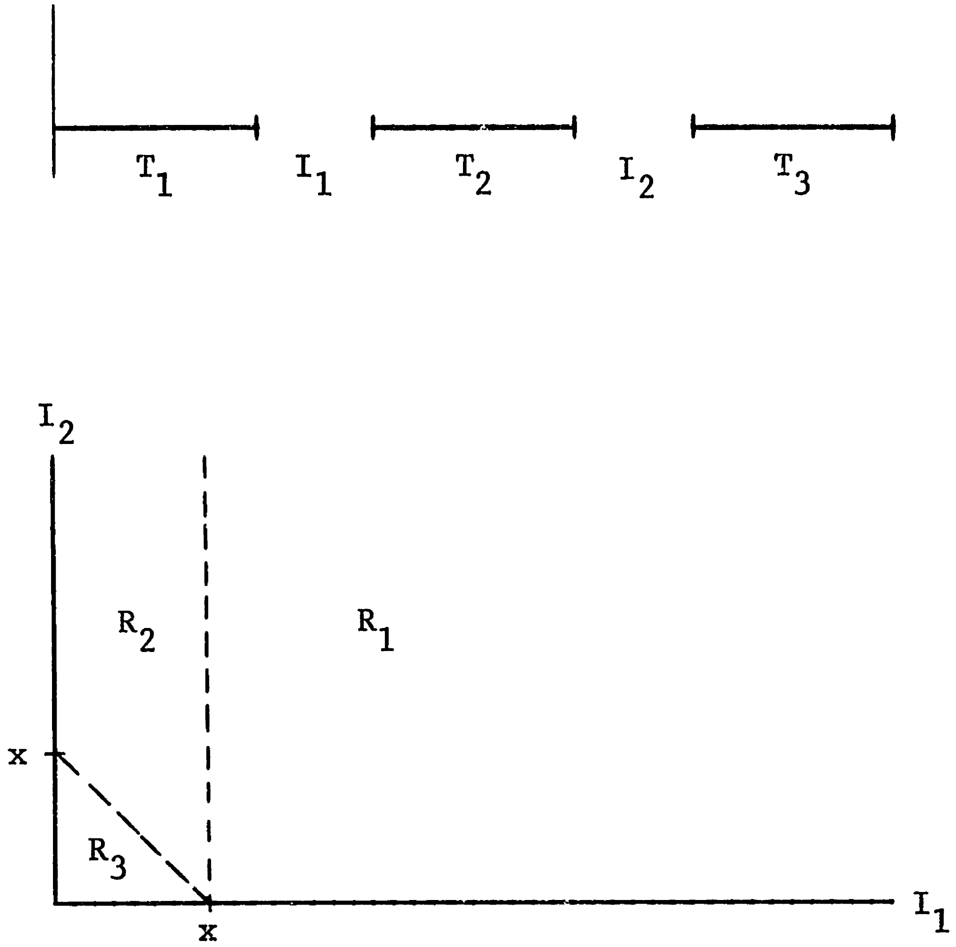


Figure 2.4 Queueing Record for N=3 and Division of $(I_1,I_2)$ Space

averaged over the whole space. $(x-I_1)M_2T_1$ is averaged over $I_1 \leq x$ and $(x-I_1-I_2)M_3T_1$ is averaged over $I_1+I_2 \leq x$. Substituting $T_1 = M_1x$ and employing (2.15) for $\Delta S_1$, we break up the expectation by averaging each term over the appropriate region.

$$\delta S_{T_1} = (M_1x^2 + \frac{1}{2} M_1^2x^2) + M_1M_2xE(x-I_1|x \geq I_1) \Pr (x \geq I_1)$$

$$+ M_1M_3xE(x-I_1-I_2|x \geq I_1 + I_2) \Pr (x \geq I_1 + I_2) \qquad (2.41)$$

We can now extend the arguments that led to (2.36) to the case of an arbitrary number of busy periods N. The necessary condition that there be a contribution to the incremental delay due to the effect of an arrival in $T_1$ on the (k+1)-st busy period is

$$\sum_{j=1}^{k} I_j \leq x \ . \qquad (2.42)$$

Hence, (2.41) generalizes in the case of N busy periods to

$$\delta S_{T_1} = (M_1x^2 + \frac{1}{2} M_1^2x^2) + \sum_{k=1}^{N-1} M_1M_{k+1}xE\left(x - \sum_{j=1}^{k} I_j \,\middle|\, \sum_{j=1}^{k} I_j \leq x\right)$$

$$\Pr \left\{ \sum_{j=1}^{k} I_j \leq x \right\}. \qquad (2.43)$$

We now proceed to define the statistics of quantities which we need to calculate (2.43). For any M/G/1 queue, the idle period lengths $I_j$ are independent, identically distributed exponential random variables with parameter $\lambda$ [4]. Hence, their sum has a gamma distribution. The density and distribution function associated with the sum Y of k idle period durations are defined below.

$$Y = \sum_{j=1}^{k} I_j \tag{2.44}$$

$$f_Y(y) = \frac{y^{k-1}\lambda^k e^{-\lambda y}}{(k-1)!} \tag{2.45}$$

$$F_Y(y) = 1 - e^{-\lambda y} \sum_{j=0}^{k-1} \frac{(\lambda y)^j}{j!} \tag{2.46}$$

To calculate (2.43), we need to evaluate $E(Y|y \leq x)$.

$$E(Y|y \leq x) = \frac{\int_0^x y f_Y(y)\, dy}{\Pr\{y \leq x\}} = \frac{\frac{k}{\lambda}\left(1 - e^{-\lambda x} \sum_{j=0}^{k} \frac{(\lambda x)^j}{j!}\right)}{\left(1 - e^{-\lambda x} \sum_{j=0}^{k-1} \frac{(\lambda x)^j}{j!}\right)} \tag{2.47}$$

Using (2.47) and (2.46), Eq. (2.43) becomes

$$\delta S_{T_1} = (M_1 x^2 + \frac{1}{2} M_1^2 x^2) + \sum_{k=1}^{N-1} x M_1 M_{k+1} \left[ (x - \frac{k}{\lambda})\left(1 - e^{-\lambda x} \sum_{j=0}^{k-1} \frac{(\lambda x)^j}{j!}\right) + e^{-\lambda x} \frac{\lambda^{k-1} x^k}{(k-1)!} \right] \tag{2.48}$$

The first term represents the effect of the additional arrival on the first busy period and the terms in the summation show effects on the remaining N-1 busy periods.

Calculation of $\delta S_{I_1}$

The preceding procedure, of examining each type of term separately and imposing conditions on the space or $(I_1 \ldots I_{N-1})$ such that the term appears in $E(\delta S | t \epsilon T_1, \text{Queueing Record})T_1$, may be applied to computing $\delta S_{I_1}$. A more compact statement of Eqs. (2.18) through (2.20) for $E(\delta S | t \epsilon I_k$, Queueing Record) will make the identification of the relevant terms clearer.

$$E(\delta S | t \epsilon I_k, \begin{matrix}\text{Queueing}\\ \text{Record}\end{matrix}) = \left\{ \begin{aligned} & x + \sum_{j=1}^{\ell+1} \left\{ \gamma_j M_{k+j} (x - \frac{1}{2} I_k - S_j) \right. \\ & \qquad \left. + (1-\gamma_j) \frac{M_{k+j}}{2I_k} (x-S_j)^2 \right\} \\ & \\ & \text{for} \quad \sum_{j=1}^{\ell} I_{k+j} < x < \sum_{j=1}^{\ell+1} I_{k+j} \end{aligned} \right. \qquad (2.49)$$

$$S_j = \begin{cases} 0 & j=1 \\ \sum_{m=1}^{j-1} I_{k+m} & j\neq 1 \end{cases} \qquad (2.50)$$

$$\gamma_j = \begin{cases} 1 & -I_k \geq S_j - x \\ 0 & S_j - x > -I_k \end{cases} \qquad (2.51)$$

From (2.44), there are three terms in $E(\delta S | t \epsilon I_1$, Queueing Record$)I_1$ to consider.

$$xI_1 \qquad (2.52)$$

$$M_{1+j}(x - \frac{1}{2} I_1 - S_j)I_1 \qquad (2.53)$$

$$\frac{M_{1+j}}{2} (x-S_j)^2 \qquad (2.54)$$

The first term (2.52) always appears and hence is averaged over the whole space of $(I_1 \ldots I_{N-1})$. We next consider the j=1 terms specified by Eqs. (2.53) and (2.54). There is no condition on $(I_1 \ldots I_{N-1})$ needed to guarantee contributions to the incremental delay due to the effect of an arrival in $I_1$ on the second busy period. Relation (2.51) implies that expression (2.53) will appear if $I_1 < x$ and expression (2.54) if $I_1 \geq x$. For $j > 1$, $\sum_{\ell=2}^{j} I_\ell < x$ is the necessary condition that arrivals in $I_1$ affect the (j+1)-st busy period. Taken together with (2.50) and (2.51) this implies the following rule for computing the expectation of terms (2.53) and (2.54) for $j > 1$.

$$\text{Average } M_{1+j}(x - \tfrac{1}{2} I_1 - \sum_{\ell=1}^{j} I_\ell) I_1 \text{ Over } \left\{ \sum_{\ell=1}^{j} I_\ell \leq x \right\} \qquad (2.55)$$

$$\text{Average } \frac{M_{1+j}}{2} (x - \sum_{\ell=2}^{j} I_\ell)^2 \text{ Over } \left\{ \begin{matrix} \sum_{\ell=1}^{j} I_\ell > x \\ \\ \sum_{\ell=2}^{j} I_\ell < x \end{matrix} \right\} \qquad (2.56)$$

This discussion of the conditions for the appearance of all the terms in $E(\delta S | t \epsilon I_1, \text{Queueing Record}) I_1$ is summarized in the formulation of its expectation.

$$\delta S_{I_1} = xEI_1 + \sum_{j=1}^{N-1} M_{1+j} E\{(x - \tfrac{1}{2} I_1 - S_j) I_1 | S_j + I_1 < x\} \Pr\{S_j + I_1 < x\}$$

$$+ \sum_{j=1}^{N-1} \frac{M_{1+j}}{2} E\{(x - S_j)^2 | S_j + I_1 > x, S_j < x\} \Pr\{S_j + I_1 > x, S_j < x\}$$

(2.57)

$$S_j = \begin{cases} 0 & j=1 \\ \sum\limits_{\ell=2}^{j} I_\ell & j\neq 1 \end{cases} \tag{2.58}$$

$EI_1$ is the unconditional mean of an exponential random variable with parameter $\lambda$ and hence is $\frac{1}{\lambda}$. The computation for the j=1 terms in the summations are lumped together and the result listed below.

$$\frac{1}{\lambda} M_2(x + \frac{1}{\lambda} (e^{-\lambda x} - 1)) \tag{2.59}$$

This term is the part of the incremental system time due to the effect on the busy period following $I_1$. The terms for $j > 1$ represent contributions due to effects on busy periods more than one removed from $I_1$. To compute the terms in (2.57) for $j > 1$ we can rewrite the expectations implied in (2.55) and (2.56) as

$$\text{Average } M_{1+j}\left(xI_1 - \left(\sum_{\ell=1}^{j} I_\ell\right)I_1 + \frac{1}{2} I_1^2\right) \text{ Over } \left\{\sum_{\ell=1}^{j} I_\ell \leq x\right\} \tag{2.60}$$

$$\text{Average } M_{1+j}\left(x - \sum_{\ell=1}^{j} I_\ell + I_1\right)^2 \text{ Over } \left\{ \begin{array}{c} \sum_{\ell=1}^{j} I_\ell \geq x \\ \\ \sum_{\ell=1}^{j} I_\ell - I_1 \leq x \end{array} \right\} \qquad (2.61)$$

By defining two random variables $Y_1$, $Y_2$ we can reformulate the evaluation of the expectations implied by (2.60) and (2.61).

$$Y_1 = I_1 \qquad (2.62)$$

$$Y_2 = I_1 + \sum_{\ell=2}^{j} I_\ell = \sum_{\ell=1}^{j} I_\ell \qquad (2.63)$$

The joint density for $Y_1$, $Y_2$ is computed from the density for the sum of j-1 independent exponential variates.

$$f_{Y_1,Y_2}(y_1,y_2) = f_{Y_1}(y_1) f_{Y_2|Y_1}(y_2|y_1) = (\lambda e^{-\lambda y_1})\left(\frac{\lambda^{j-1}(y_2-y_1)^{j-2}e^{\lambda(y_2-y_1)}}{(j-2)!}\right)$$

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{\lambda^j}{(j-2)!}(y_2-y_1)^{j-2}e^{-\lambda y_2}$$

$$\text{for } y_2 \geq y_1 \geq 0 \qquad (2.64)$$

We now define two regions in $(Y_1,Y_2)$ space.

$$R_1 \triangleq \{(Y_1,Y_2): \; Y_2 \leq x\} \qquad (2.65)$$

$$R_2 \triangleq \{(Y_1,Y_2): \quad Y_2 \geq x, \qquad Y_2-Y_1 \leq x\} \tag{2.66}$$

By making identifications between $Y_1$, $Y_2$ and the variables in (2.60) and (2.61), the desired expectations may be expressed as follows:

$$M_{1+j}(x(\overline{Y}_1|R_1) - (\overline{Y_1Y_2}|R_1) - \frac{1}{2}(\overline{Y_1^2}|R_1))Pr\{R_1\} \tag{2.67}$$

$$\frac{M_{1+j}}{2}(x^2 + (\overline{Y_1^2}|R_2) + (\overline{Y_2^2}|R_2) - 2x(\overline{Y}_2-\overline{Y}_1|R_2) - 2(\overline{Y_1Y_2}|R_2))Pr\{R_2\} \tag{2.68}$$

The notation $(\overline{g(Y_1,Y_2)}|R_1)$ denotes $E(g(Y_1,Y_2)|(Y_1,Y_2)\epsilon R_1)$ and $Pr\{R_2\}$ denotes the probability that $(Y_1,Y_2)$ lies in $R_2$. Results of the calculations in (2.67) and (2.68) are listed below.

$$Pr\{R_1\} = \int_0^x \int_{y_1}^x f_{Y_1,Y_2}(y_1,y_2)\, dy_2\, dy_1$$

$$= 1 - e^{-\lambda x} \sum_{\ell=0}^{j-1} \frac{(\lambda x)^{\ell}}{\ell!} \tag{2.69}$$

$$(\overline{Y}_1|R_1) = \int_0^x \int_{y_1}^x y_1 \frac{f_{Y_1,Y_2}(y_1,y_2)}{\Pr\{R_1\}} dy_2 dy_1$$

$$= \frac{\frac{1}{\lambda}(1 - e^{-\lambda x}) - xe^{-\lambda x} - \lambda e^{-\lambda x} \sum_{\ell=0}^{j-2} \frac{\lambda^{\ell} x^{\ell+2}}{(\ell+2)!}}{\Pr\{R_1\}} \tag{2.70}$$

$$(\overline{Y}_1^2|R_1) = \int_0^x \int_{y_1}^x y_1^2 \frac{f_{Y_1,Y_2}(y_1,y_2)}{\Pr\{R_1\}} dy_2 dy_1$$

$$= \frac{\frac{2}{\lambda^2}(1-e^{-\lambda x}) - \frac{2}{\lambda} xe^{-\lambda x} - x^2 e^{-\lambda x} - \lambda e^{-\lambda x} \sum_{\ell=0}^{j-2} \frac{2\lambda^{\ell} x^{\ell+3}}{(\ell+3)!}}{\Pr\{R_1\}} \tag{2.71}$$

$$(\overline{Y_1 Y_2}|R_1) = \int_0^x \int_{y_1}^x y_1 y_2 \frac{f_{Y_1,Y_2}(y_1,y_2)}{\Pr\{R_1\}} dy_2 dy_1$$

$$= \frac{\left\{ \frac{2}{\lambda^2}(1 - e^{-\lambda x}) - \frac{2}{\lambda} xe^{-\lambda x} - x^2 e^{-\lambda x} + \frac{j-1}{\lambda}\left[\frac{1}{\lambda}(1 - e^{-\lambda x}) - xe^{-\lambda x}\right] - \frac{e^{-\lambda x}}{\lambda^2}\left[(j+1) \sum_{\ell=0}^{j-2} \frac{(\lambda x)^{\ell+3}}{(\ell+3)!} + (j-1) \frac{(\lambda x)^2}{2}\right]\right\}}{\Pr\{R_1\}} \tag{2.72}$$

$$\Pr\{R_2\} = \int_{y_2=x}^{\infty} \int_{y_1=y_2-x}^{y_2} f_{Y_1,Y_2}(y_1,y_2)\, dy_1 dy_2$$

$$= \frac{(\lambda x)^{j-1}}{(j-1)!} e^{-\lambda x} \tag{2.73}$$

$$(\overline{Y}_1 | R_2) = \int_{y_2=x}^{\infty} \int_{y_1=y_2-x}^{y_2} y_1 \frac{f_{Y_1,Y_2}(y_1,y_2)}{\Pr\{R_2\}}\, dy_1 dy_2$$

$$= \frac{e^{-\lambda x} \dfrac{(\lambda x)^{j-1}}{j!} \left(x + \dfrac{j}{\lambda}\right)}{\Pr\{R_2\}} \tag{2.74}$$

$$(\overline{Y_1^2} | R_2) = \int_{y_2=x}^{\infty} \int_{y_1=y_2-x}^{y_2} y_1^2 \frac{f_{Y_1,Y_2}(y_1,y_2)}{\Pr\{R_2\}}\, dy_1 dy_2$$

$$= \frac{\dfrac{(\lambda x)^{j-1}}{(j+1)!} \left[2x^2 + \dfrac{2}{\lambda}(j+1)x + \dfrac{2}{\lambda^2}(j^2+j)\right] e^{-\lambda x}}{\Pr\{R_2\}} \tag{2.75}$$

$$(\overline{Y}_2|R_2) = \int_{y_2=x}^{\infty} \int_{y_1=y_2-x}^{y_2} y_2 \frac{f_{Y_1,Y_2}(y_1,y_2)}{\Pr\{R_2\}} dy_1 dy_2$$

$$= \frac{\frac{(\lambda x)^{j-1}}{(j-1)!} (x + \frac{1}{\lambda}) e^{-\lambda x}}{\Pr\{R_2\}} \tag{2.76}$$

$$(\overline{Y_2^2}|R_2) = \int_{y_2=x}^{\infty} \int_{y_1=y_2-x}^{y_2} y_2^2 \frac{f_{Y_1,Y_2}(y_1,y_2)}{\Pr\{R_2\}} dy_1 dy_2$$

$$= \frac{\frac{(\lambda x)^{j-1}}{(j-1)!} (x^2 + \frac{2}{\lambda} x + \frac{2}{\lambda^2}) e^{-\lambda x}}{\Pr\{R_2\}} \tag{2.77}$$

$$(\overline{Y_1 Y_2}|R_2) = \int_{y_2=x}^{\infty} \int_{y_1=y_2-x}^{y_2} y_1 y_2 \frac{f_{Y_1,Y_2}(y_1,y_2)}{\Pr\{R_2\}} dy_1 dy_2$$

$$= \frac{\frac{(\lambda x)^{j-1}}{j!} (x^2 + \frac{(j+1)}{\lambda} x + j \frac{2}{\lambda^2}) e^{-\lambda x}}{\Pr\{R_2\}} \tag{2.78}$$

Substituting (2.69) - (2.78) into (2.67) and (2.68) and using (2.59), we can finally evaluate the expectation of $E(\delta S|t \epsilon I_1$, Queueing Record$)I_1$ over $(I_1 \ldots I_{N-1})$ outlined in (2.57).

$$\delta S_{I_1} = \frac{x}{\lambda} + \frac{1}{\lambda} M_2 \left(x + \frac{1}{\lambda} \left(e^{-\lambda x} - 1\right)\right)$$

$$+ \sum_{j=2}^{N-1} M_{1+j} \left\{ \begin{array}{l} \left(\frac{x}{\lambda} - \frac{j}{\lambda^2}\right) + \frac{j}{\lambda^2} e^{-\lambda x} + \frac{(j-1)}{\lambda} x e^{-\lambda x} + \frac{j-2}{2} x^2 e^{-\lambda x} \\ \\ - \frac{1}{\lambda} x e^{-\lambda x} \sum_{\ell=0}^{j-2} \frac{(\lambda x)^{\ell+2}}{(\ell+2)!} + \frac{j}{\lambda^2} e^{-\lambda x} \sum_{\ell=0}^{j-2} \frac{(\lambda x)^{\ell+3}}{(\ell+3)!} \end{array} \right\}$$

$$+ \sum_{j=2}^{N-1} M_{1+j} \left\{ \frac{(\lambda x)^j}{(j+1)!} e^{-\lambda x} \frac{x}{\lambda} \right\} \tag{2.79}$$

Calculation of $\delta S_{T_k}$, $\delta S_{I_k}$ and $E \left\{ \frac{\delta \hat{S}}{\delta \lambda} | M_1 \ldots M_N \right\}$

To use our results (2.48) and (2.79) for $\delta S_{T_1}$ and $\delta S_{I_1}$ respectively, in order to derive the mean of the unnormalized estimator formulated in (2.34), we note that in computing $\delta S_{T_k}$ and $\delta S_{I_k}$, only the N-k idle periods following $T_k$ enter into the averaging. Hence, we use our answers for $\delta S_{T_1}$ and $\delta S_{I_1}$ adjusted to correspond to a N-k+1 busy period case. By the preceding argument, the expectation of the unnormalized estimator over $(I_1 \ldots I_{N-1})$ is given by

$$E\left\{\frac{\delta\hat{S}}{\delta\lambda}\middle| M_1 \ \ldots \ M_N\right\} = \sum_{k=1}^{N} (M_k x^2 + \frac{1}{2} M_k^2 x^2) + \sum_{j=1}^{N-1} \sum_{k=1}^{N-j} M_j M_{j+k} a_k$$

$$+ \sum_{k=1}^{N-1} (\frac{x}{\lambda} + \frac{1}{\lambda} M_{k+1} (x + \frac{1}{\lambda} (e^{-\lambda x} - 1))$$

$$+ \sum_{k=1}^{N-2} \sum_{j=2}^{N-k} M_{k+j} b_j$$

$$+ \sum_{k=1}^{N-2} \sum_{j=2}^{N-k} M_{k+j} C_j \ . \qquad (2.80)$$

$$a_k = x\left\{(x - \frac{k}{\lambda})\left(1 - e^{-\lambda x} \sum_{\ell=0}^{K-1} \frac{(\lambda x)^\ell}{\ell!}\right) - e^{-\lambda x} \frac{\lambda^{k-1} x^k}{(k-1)!}\right\} \qquad (2.81)$$

$$b_j = \left\{\begin{array}{l} (\frac{x}{\lambda} - \frac{j}{\lambda^2}) + \frac{j}{\lambda^2} + \frac{(j-1)}{\lambda} x e^{-\lambda x} + (\frac{j-2}{2}) x^2 e^{-\lambda x} \\ - \frac{1}{\lambda} x e^{-\lambda x} \sum_{\ell=0}^{j-2} \frac{(\lambda x)^{\ell+2}}{(\ell+2)!} + \frac{j}{\lambda^2} e^{-\lambda x} \sum_{\ell=0}^{j-2} \frac{(\lambda x)^{\ell+3}}{(\ell+3)!} \end{array}\right\}$$

$$(2.82)$$

$$C_j = \frac{(\lambda x)^j}{(j+1)!} e^{-\lambda x} \frac{x}{\lambda} \qquad (2.83)$$

The behavior of $a_k$ and $b_j$ are made clearer by replacing each summation in their definition by $e^{\lambda x}$ minus some quantity.

$$a_k = x(x - \frac{k}{\lambda})e^{-\lambda x} \sum_{\ell=k}^{\infty} \frac{(\lambda x)^{\ell}}{\ell!} + x^2 e^{-\lambda x} \frac{(\lambda x)^{k-1}}{(k-1)!}$$

$$b_j = e^{-\lambda x} \left\{ \frac{x}{\lambda} \sum_{\ell=j-1}^{\infty} \frac{(\lambda x)^{\ell+2}}{(\ell+2)!} - \frac{j}{\lambda^2} \sum_{\ell=j-1}^{\infty} \frac{(\lambda x)^{\ell+3}}{(\ell+3)!} \right\} \qquad (2.84)$$

<u>Evaluation of Limit in (2.34)</u>

Having almost evaluated the expectation of the unnormalized estimator, we are nearly ready to examine the limit in (2.34). We complete the expectation of $\frac{\delta \hat{S}}{\delta \lambda}$ by averaging over $(M_1 \ldots M_N)$. This amounts to replacing $M_i$ by $\bar{M}$, $M_i^2$ by $\overline{M^2}$, and noting that due to the independence of the M's, $EM_jM_{j+k} = \bar{M}^2$. Carrying out the expectation over the M's, changing the order of summation in the double sums, and dividing by $N\bar{M}$ we are left with

$$\frac{1}{N\overline{M}} E \frac{\delta \hat{S}}{\delta \lambda} = x^2 + \frac{1}{2} \frac{\overline{M^2}}{\overline{M}} x^2 + \overline{M} \frac{1}{N} \sum_{k=1}^{N-1} (N-k)a_k$$

$$+ \frac{N-1}{N} \left\{ \frac{x}{\lambda \overline{M}} + \frac{1}{\lambda} \left(x + \frac{1}{\lambda} (e^{-\lambda x} - 1)\right)\right\}$$

$$+ \frac{1}{N} \sum_{j=2}^{N-1} (N-j)b_j + \frac{1}{N} \sum_{j=2}^{N-1} (N-j)C_j \ . \qquad (2.85)$$

To examine the behavior of (2.85) as N → ∞ we must evaluate the following limits:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N-1} ka_k \qquad (2.86)$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=2}^{N-1} jb_j \qquad (2.87)$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=2}^{N-1} jC_j \qquad (2.88)$$

$$\lim_{N\to\infty} \sum_{k=1}^{N-1} a_k \qquad (2.89)$$

$$\lim_{N\to\infty} \sum_{j=2}^{N-1} b_j \tag{2.90}$$

$$\lim_{N\to\infty} \sum_{j=2}^{N-1} C_j \tag{2.91}$$

To prove that the three limits (2.86) - (2.88) are zero, it is sufficient to show that the unnormalized infinite sums are finite. The infinite sums implied by (2.86) - (2.91) are evaluated by switching the order of summation and looking for terms that correspond to the exponential power series. The results are expressed in terms of $\rho = \lambda x$.

$$\sum_{k=1}^{\infty} ka_k = x^2[\frac{1}{2}\rho + \frac{1}{6}\rho^2] \tag{2.92}$$

$$\sum_{j=2}^{\infty} jb_j = \frac{x^2e^{-\rho}}{\rho^2}\{\frac{2}{3}\rho^3e^{\rho} - \frac{1}{2}\rho^4 + \rho^2 + 2e^{\rho} - 2\rho e^{\rho} - 2\} \tag{2.93}$$

$$\sum_{j=2}^{\infty} jC_j = \frac{x^2e^{-\rho}}{\rho^2}\{\rho e^{\rho} - e^{\rho} + 1 - \frac{\rho^2}{2}\} \tag{2.94}$$

$$\sum_{k=1}^{\infty} a_k = \frac{1}{2}\rho x^2 \tag{2.95}$$

$$\sum_{j=2}^{\infty} b_j = \frac{1}{2} x^2 + \frac{1}{2} x^2 e^{-\rho} + \frac{x}{\lambda} (e^{-\rho} - 1) \tag{2.96}$$

$$\sum_{j=2}^{\infty} C_j = \frac{1}{\lambda^2} - \frac{1}{\lambda^2} e^{-\rho} - \frac{x}{\lambda} e^{-\rho} - \frac{1}{2} x^2 e^{-\rho} \tag{2.97}$$

Examination of the power series in $\rho$ that corresponds to (2.92) - (2.94) shows that each is a bounded function of $\rho$ on [0,1]. We are interested in $\rho$ on [0,1] since the statistics of the queueing system are stationary for this range. The boundedness of (2.92) - (2.94) implies that the limits (2.86) - (2.88) must be zero.

To complete the description of (2.85), we list from [4] expressions for the first and second moment of the number served in a busy period for an M/D/1 system.

$$\overline{M} = \frac{1}{1-\rho} \tag{2.98}$$

$$\overline{M^2} = \frac{2\rho - \rho^2}{(1-\rho)^3} + \frac{1}{1-\rho} \tag{2.99}$$

Employing our knowledge of the limits (2.86) - (2.91) provided by (2.92) - (2.97) and using (2.98) and (2.99), we can calculate the limit suggested by (2.85).

$$\lim_{N\to\infty}\left\{\frac{1}{\overline{NM}}\ E\ \frac{\delta\hat{S}}{\delta\lambda}\right\} = \frac{3}{2}\ x^2 + \frac{\frac{1}{2}\ \rho x^2}{(1-\rho)} + \frac{x}{\lambda}\ (1-\rho)\ + \frac{1}{2}\ x^2\left(\frac{1}{1-\rho}\right)^2 \qquad (2.100)$$

Multiplying Eq. (2.85) by λ and with some minor rearranging we obtain

$$\lambda\ \lim_{N\to\infty}\left\{\frac{1}{\overline{NM}}\ E\ \frac{\delta\hat{S}}{\delta\lambda}\right\} = x + \frac{\rho x}{2(1-\rho)} + \frac{\rho x}{2(1-\rho)^2}\ . \qquad (2.101)$$

This is the desired delay gradient for an M/D/1 queue derived in (2.33). Hence, on the basis of the thinking leading to (2.34), we have proven the asymptotic unbiasedness of the customer-addition algorithm.

## 2.4 Cramer-Rao Bound for Customer-Addition Algorithm in Case of M/D/1 Queue

Since the calculation of the exact variance associated with the customer-addition algorithm is too cumbersome, we derive a Cramer-Rao bound. If we have an observation vector R, a parameter A we want to estimate, and a conditional density $P_{ra}(R|A)$, the Cramer-Rao bound for the variance of any unbiased estimator $\hat{a}(R)$ of A is stated as follows:

$$\mathrm{Var}(\hat{a}(R)\ -\ A) \geq \frac{1}{-\ E\left\{\dfrac{\partial^2 \ln P_{r|a}(R|A)}{\partial A^2}\right\}} \qquad (2.102)$$

In the case of an M/D/1 queue, the observation vector which the customer-addition algorithm works with is the concatenation of two sets of variables. If Y denotes the total observation vector, then $Y \triangleq (Y_1 \mid Y_2)$ where $Y_1$ consists of $(M_1 \ldots M_N)$ and $Y_2$ of $(I_1 \ldots I_{N-1})$. We know that $M_i$ is independent of $M_j$ and $I_k$ independent of $I_\ell$ for all $i \neq j$, $k \neq \ell$. The length of idle period $I_k$ is determined by an "end" effect in the k-th busy period. Hence, the only conceivable place we could find statistical dependence is between $M_k$ and $I_k$. We resolve this question by considering the density for the length of idle period $I_1$ conditioned on $M_1 = m_1$. The dynamics of successive waiting times in a queue are described by the recursion

$$\omega_{n+1} = \max\{0, \omega_n + x_n - \theta_n\} \text{ with initial condition } \omega_1 = 0. \tag{2.103}$$

$x_n$ is the n-th service requirement and $\theta_n$ the inter-arrival time between the n-th and (n+1)-st customer. If there are $m_1$ customers in the first busy period, the following relations must hold.

$$\omega_i + x_i - \theta_i > 0 \qquad i = 1 \ldots m_1 - 1 \tag{2.104}$$

$$\omega_{m_1} + x_{m_1} - \theta_{m_1} < 0 \tag{2.105}$$

$$I_1 = \theta_{m_1} - (x_{m_1} + \omega_{m_1}) \tag{2.106}$$

For any M/G/1 queue, the unconditional density for $\theta_1$ is exponential with parameter $\lambda$. Hence, the density for $I_1$ conditioned on $m_1$ customers in the first busy period is related by (2.105) and (2.106) to the density for $\theta_{m_1}$, conditioned on $\theta_{m_1}$ being greater than the sum of the $m_1$-th service and waiting time.

$$P_{\theta_{m_1} | \theta_{m_1} \geq x_{m_1} + w_{m_1}}(\tau) = \frac{e^{-\lambda\tau}}{e^{-\lambda(x_{m_1} + \omega_{m_1})}} = \lambda e^{-\lambda(\tau - (x_{m_1} + \omega_{m_1}))}$$

$$\text{for } \tau \geq x_{m_1} + \omega_{m_1} \tag{2.107}$$

From (2.106) and (2.107) we calculate the conditional density of $I_1$ as

$$P_{I_1 | M_1}(I_1) = e^{-\lambda I_1} . \tag{2.108}$$

Hence, since the density for $I_1$ conditioned on $m_1$ customers being served in the first busy period is identical to the unconditional density, $M_1$ and $I_1$ are statistically independent random variables. This independence is a property of the "memoryless" inter-arrival density. Based on the preceding arguments, all the variables in the observation vector $Y \triangleq (M_1 \ldots M_N \; I_1 \ldots I_{N-1})$ are mutually independent.

Hence, the joint density of Y may be expressed as

$$P(Y) = \prod_{i=1}^{N} \Pr\{M_i = m_i\} \prod_{j=1}^{N-1} \lambda e^{-\lambda I_j} . \qquad (2.109)$$

For M/G/1 systems, queueing theory has calculated the probability of k customers being served in a busy period [4]. The result for an M/D/1 queue is given in [4] as

$$\Pr\{M_i = m_i\} = \frac{(m_i \rho)^{m_i - 1}}{m_i!} e^{-m_i \rho} . \qquad (2.110)$$

For the moment, we pretend that the parameter of interest is $\rho$ and using (2.110) we rewrite (2.109) in its terms.

$$P(Y|\rho) = \left( \prod_{i=1}^{N} \frac{m_i^{m_i - 1}}{m_i!} \right) \rho^{\sum_{i=1}^{N} m_i - N} e^{-\left(\sum_{i=1}^{N} m_i\right)\rho} \left(\frac{\rho}{x}\right)^{N-1} e^{-\frac{\rho}{x} \sum_{j=1}^{N-1} I_j} \qquad (2.111)$$

We let $\gamma$ denote the delay gradient $\frac{\partial D}{\partial \lambda}$. For $\rho$ on [0,1] Eq. (2.33) specifies a 1-1 correspondence between $\gamma$ and $\rho$. Relation (2.33) may be inverted to find $\rho$ as a function of $\gamma$.

$$\rho = 1 - \sqrt{1 - \frac{2(\gamma - x)}{(2\gamma - x)}} \tag{2.112}$$

We can evaluate the second partial derivative of the logarithm of the joint density required in (2.97) by applying the chain-rule of differentiation

$$\frac{\partial^2 \ln P(Y|\gamma)}{\partial \gamma^2} = \frac{\partial^2 \ln P(Y|\rho)}{\partial \rho^2} \left(\frac{\partial \rho}{\partial \gamma}\right)^2 + \frac{\partial \ln P(Y|\rho)}{\partial \rho} \frac{\partial^2 \rho}{\partial \gamma^2} \tag{2.113}$$

Performing the above manipulations and employing the two following expectations:

$$E \sum_{i=1}^{N} m_i = \frac{N}{1-\rho} \tag{2.114}$$

$$E \sum_{j=1}^{N-1} I_j = \frac{N-1}{\lambda} \tag{2.115}$$

We can evaluate (2.92) for the customer-addition algorithm.

$$\mathrm{Var}\ (\gamma - \hat{\gamma}) \geq \frac{x^2\rho^2}{(N-1+\rho)(1-\rho)^5} \tag{2.116}$$

The result (2.116) behaves as expected for $\rho$ near zero and one. As the utilization factor $\rho$ nears 1, the queue becomes nonstationary. The means and variances of variables such as the number in the system, the waiting time, and the number served in a busy period become infinite. Hence, any estimation algorithm which is a function of these queueing variables might be expected to diverge as $\rho$ goes to 1. If we conceive of $\rho$ approaching zero by fixing x and letting $\lambda$ go to zero, the average idle period duration becomes unbounded. In addition, Var $\{M_i\}$ goes to zero as $\rho \to 0$. Hence, since the queueing variables become "known" as $\rho \to 0$, it is reasonable to expect that the variance of the estimator goes to zero for $\rho = 0$.

## 2.5 Derivation and Realization in Flow Diagram Form of Customer-Removal Algorithm

Since the function $D(\lambda)$ for the average delay/unit time accumulated by the queue is continuously differentiable on $0 \leq \lambda < \frac{1}{x}$ $(0 \leq \rho < 1)$, the limit defining $\frac{\partial D}{\partial \lambda}(\lambda)$ at a given $\lambda^*$ is independent of the direction from which $\lambda$ approaches $\lambda^*$.

$$\left.\frac{\partial D}{\partial \lambda}\right|_{\lambda=\lambda^*} = \lim_{\lambda \to \lambda^*} \frac{D(\lambda) - D(\lambda^*)}{\lambda - \lambda^*} \tag{2.117}$$

In the customer-addition algorithm we let $\lambda = \lambda^* + \delta\lambda$ and allow $\delta\lambda$ to approach zero through positive values. For the customer-removal algorithm we equivalently let $\delta\lambda$ go to zero through negative values. We simulate a decrement in arrival rate $\delta\lambda$ by removing customers from the queue with probability $\varepsilon$ and computing the resulting decrement in total system time.

The value for $\varepsilon$ is motivated by the fact that the expected change $\delta\lambda T_E$ in the number of customers arriving in an interval $T_E$ caused by a decrement $\delta\lambda$ in incoming flow equals the negative of the expected number of customers removed by Bernoulli trials, $M'\varepsilon$. Here $M'$ is the total number of customers arriving in a period $T_E$. Hence, $\varepsilon = -\frac{\delta\lambda T_E}{M'}$. Letting $\delta S_j^{(i)}$ denote the change in system time of the i-th busy period due to the removal of the j-th customer, the expected change in system time $\delta S$ in a time $T_E$ due to a decrement in flow $\delta\lambda$ and conditioned on the queueing record is formulated as

$$E(\delta S \,|\text{Queueing Record}) = \sum_{i=1}^{N} \sum_{j=1}^{M_i} -\delta S_j^{(i)} \frac{T_E \delta\lambda}{M'} , \qquad (2.118)$$

where N is the number of busy periods in the $T_E$ long observation period. Hence, the desired delay gradient estimator is given by

$$\hat{D}' = \frac{1}{T_E} \frac{E(\delta S \,|\text{Queueing Record})}{\delta\lambda} = \frac{1}{M'} \sum_{i=1}^{N} \sum_{j=1}^{M_i} -\delta S_j^{(i)}. \tag{2.119}$$

We compute $\delta S_j^{(i)}$ by working with more microscopic quantities. Let $C_{m,i}^{n}$ denote the amount of system time saved for the M-th customer in the i-th busy period by the removal of the n-th customer in that busy period. Since the removal of the n-th customer can have no effect on customers that preceded him, $C_{m,i}^{n} = 0$ for m = 1 ... n-1. Hence, $\delta S_j^{(i)}$ can be computed as

$$\delta S_j^{(i)} = - \sum_{m=j}^{M_i} C_{m,i}^{j}. \tag{2.120}$$

We now develop a systematic procedure for calculating the $C_{k,i}^{n}$'s. To simplify the notation, we drop the i denoting the index of the busy period. Let $\omega_n$, $S_n$, and $x_n$ denote the waiting time, system time and service requirement, respectively of the n-th customer in the busy period. Let $d_n$ and $a_n$ denote the corresponding departure and arrival time of the n-th customer. Since the system time the n-th customer saves by the removal of the n-th customer is $S_n$, we have $C_n^n = S_n$. In considering the effect of removing the n-th customer on the (n+1)-st customer, either a new busy period begins with the (n+1)-st customer or the (n + 1)-st customer remains part of the busy period formed by customers 1 to n-1. The condition for customer n + 1 beginning a new busy period is that the arrival time $a_{n+1}$ of the (n + 1)-st customer is greater than the departure time $d_{n-1}$ of the (n-1)-st customer. In this case, customer n + 1 will save its waiting time $\omega_{n+1}$. If $d_{n-1} > a_{n+1}$, customer n + 1 does not start a new busy period, and saves an amount of time $x_n$ since it need no longer wait for customer n to be served. This rule for $C^n_{n+1}$ is summarized by the following:

$$C^n_{n+1} = \begin{cases} d_n - a_{n+1} = \omega_{n+1} & \text{for} \quad a_{n+1} \geq d_{n-1} \\ d_n - d_{n-1} = x_n & \text{for} \quad d_{n-1} > a_{n+1} \end{cases} \tag{2.121}$$

Relation (2.121) is more succinctly stated as

$$C^n_{n+1} = d_n - \max \{a_{n+1}, d_{n-1}\} \tag{2.122}$$

Noting that $\max \{a_{n+1}, d_{n-1}\} = -\min \{-a_{n+1}, -d_{n-1}\}$, $C^n_{n+1}$ may be restated in a final form as

$$C^n_{n+1} = \min \{\omega_{n+1}, x_n\}. \tag{2.123}$$

Similar reasoning to that employed in calculating $C^n_{n+1}$ applies to the computation of $C^n_m$. The removal of customer n either causes customers m and m-1 to be in the same busy period, or customer m may begin a new busy period. The removal of customer n causes customer m-1 to save system time $C^n_{m-1}$. Hence, customer m-1 departs at an earlier time $d_{m-1} - C^n_{m-1}$. If this new departure time for customer m-1 is greater than the arrival time $a_m$ of customer m, customers m and m-1 remain in the same busy interval and customer m is saved a system time $C^n_{m-1}$. However, if $a_m > d_{m-1} - C^n_{m-1}$, customer m begins a new busy period and saves its waiting time $\omega_m$. These relationships are summarized in the following rule for computing $C^n_m$:

$$C^n_m = \begin{cases} C^n_{m-1} & \text{for} \quad d_{m-1} - C^n_{m-1} > a_m \\ \\ d_{m-1} - a_m = \omega_m & \text{for} \quad a_m > d_{m-1} - C^n_{m-1} \end{cases} \tag{2.124}$$

This rule may be expressed more compactly as

$$C^n_m = \min \{C^n_{m-1}, \omega_m\}. \tag{2.125}$$

Hence, the algorithm for computing the $C^n_m$'s may be summarized as

$$C_n^n = S_n$$

$$C_{n+1}^n = \min\,\{x_n,\ \omega_{n+1}\} \tag{2.126}$$

$$C_m^n = \min\,\{C_{m-1}^n,\ \omega_m\}\quad m = n+2\ \ldots\ M$$

M is the number of customers served in the given busy period. The customer-removal algorithm is completely specified by (2.126) and the following form for the delay-gradient estimator derived by substituting (2.120) into (2.119).

$$\hat{D}' = \frac{1}{M'} \sum_{i=1}^{N} \sum_{n=1}^{M} \sum_{m=n}^{M} C_{m,i}^n = \frac{1}{M'} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{n=1}^{m} C_{m,i}^n \tag{2.127}$$

The second form suggests calculating and summing $C_m^n$ for $n = 1 \ldots m$ when the m-th customer arrives. Hence, to calculate the inner two summations in (2.127), we need only M variables $\xi_i$ to store $C_j^i$ as j is varied. This idea is realized in a flow diagram for the customer-removal algorithm in Fig. 2.6. The variables in the flow chart are defined in Fig. 2.5.

$\hat{D}'$ = Current estimate for delay gradient

$M'$ = Current total number of customers in observation period

TS = Running sum of service times in most recent busy period

$x$ = Service requirement of most recent customer

$\tau$ = Arrival time of most recent customer relative to beginning of busy period

$\omega$ = Waiting time of most recent customer

$j$ = Index of most recent customer in current busy period

$M$ = Total number of customers in most recent busy period

$\xi_i$ = Storage location for $C_j^i$ as $j$ is varied

$S = \sum_{n=1}^{M} \sum_{m=n}^{M} C_m^n$ for most recent busy period. It is the cumulative system time saved by the removal of each customer in the current busy period.

Figure 2.5 Definition of Variables for Customer-Removal Algorithm Flow Diagram

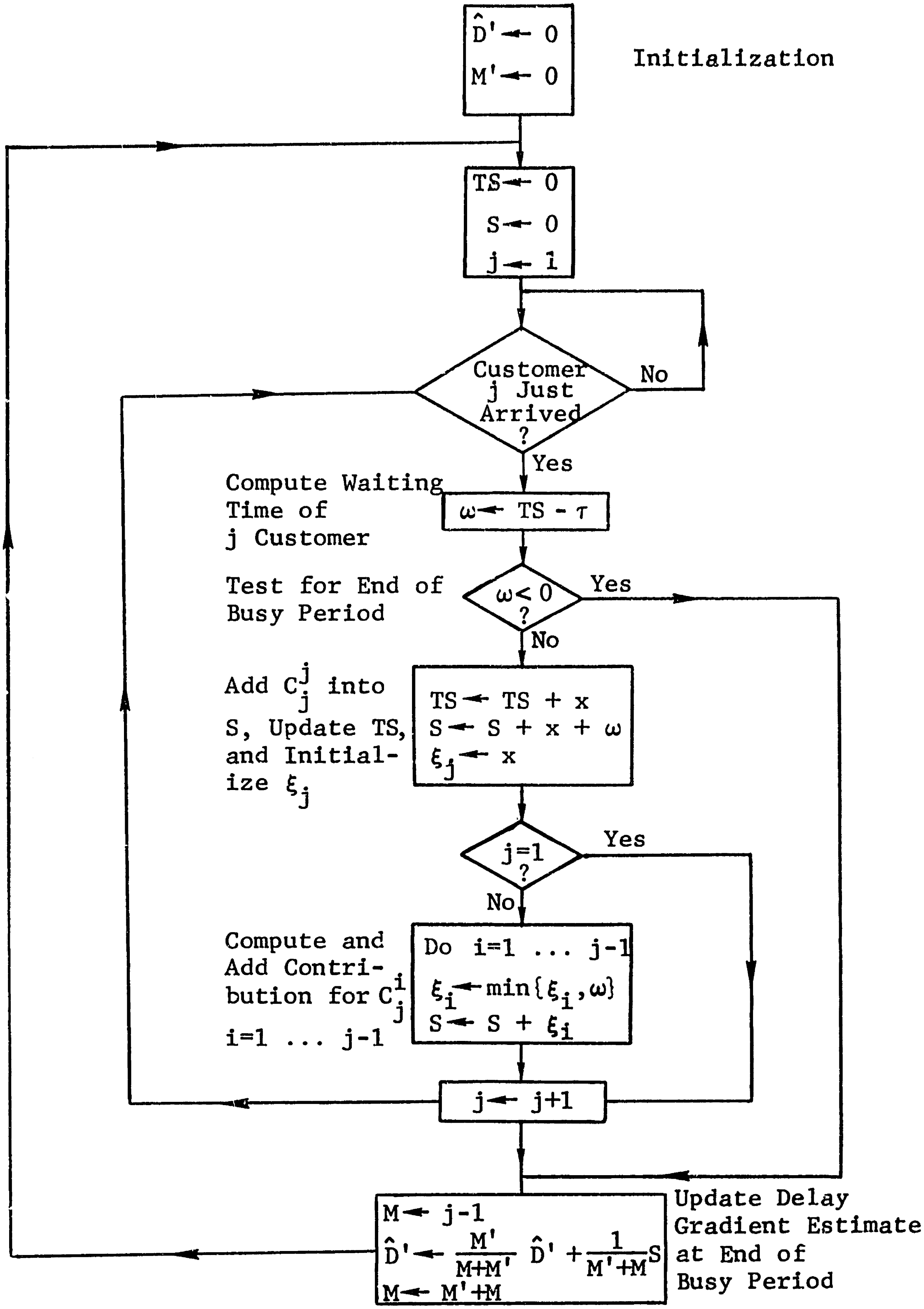

$\hat{D}' \leftarrow 0$
$M' \leftarrow 0$
Initialization
$TS \leftarrow 0$
$S \leftarrow 0$
$j \leftarrow 1$
Customer j Just Arrived ?
No
Yes
Compute Waiting Time of j Customer
$\omega \leftarrow TS - \tau$
Test for End of Busy Period
$\omega < 0$ ?
Yes
No
Add $C_j^j$ into S, Update TS, and Initialize $\xi_j$
$TS \leftarrow TS + x$
$S \leftarrow S + x + \omega$
$\xi_j \leftarrow x$
$j=1$ ?
Yes
No
Compute and Add Contribution for $C_j^i$ i=1 ... j-1
Do i=1 ... j-1
$\xi_i \leftarrow \min\{\xi_i, \omega\}$
$S \leftarrow S + \xi_i$
$j \leftarrow j+1$
$M \leftarrow j-1$
$\hat{D}' \leftarrow \frac{M'}{M+M'} \hat{D}' + \frac{1}{M'+M} S$
$M \leftarrow M'+M$
Update Delay Gradient Estimate at End of Busy Period

Figure 2.6 Flow Diagram for Customer Removal Algorithm

## 2.6 Calculation of Asymptotic Bias for Customer-Removal Algorithm for M/G/1 Queues

We now investigate the asymptotic properties of the customer-removal algorithm by first interpreting the terms in the estimator. For a given busy period, the inner two summations in (2.127) may be grouped into two terms representing the sum of all the service times of the customers in that busy period and the cumulative service time saved by all other customers due to the removal of each customer separately. If $S_j^{(i)}$ denotes the system time of the j-th customer in the i-th busy period, the customer-removal delay gradient estimator may be expressed as follows:

$$\hat{D}' = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} S_j^{(i)}}{\sum_{i=1}^{N} M_i} + \frac{\sum_{i=1}^{N} P_i}{\sum_{i=1}^{N} M_i} , \qquad (2.128)$$

where $P_i$ is defined by

$$P_i = \begin{cases} 0 \quad \text{if} \quad M_i = 1 \\ \sum_{n=1}^{M_i - 1} \sum_{m=n+1}^{M_i} C_{m,i}^{n} \quad \text{if} \quad M_i \neq 1. \end{cases} \qquad (2.129)$$

We examine the asymptotic behavior of the mean of the estimator specified in (2.128) by interchanging the expectation and limit operation. By appealing to the law of large numbers, the limiting form of the estimator as N becomes unbounded is

$$\lim_{N \to \infty} E\hat{D}' = E \lim_{N \to \infty} \hat{D}' = D_c + \frac{\overline{P}}{\overline{M}} \qquad (2.130)$$

where $D_c$ is the average system time per customer and $\overline{P}$ is the expectation of the quantity defined in (2.124). For an M/G/1 queue, the mean of $P_i$ is independent of i since the $C_m^n$'s depend on waiting times and service times which are statistically independent from one busy period to another. $\overline{M}$ denotes the average number of customers served per busy period.

Using the general relation derived in (2.26) that the delay gradient is equal to $D_c$ plus $\lambda \frac{\partial D_c}{\partial \lambda}$, we can formulate the asymptotic bias of the customer-removal algorithm as

$$b = \lim_{N \to \infty} E\hat{D}' - \frac{\partial D}{\partial \lambda} = \frac{\overline{P}}{\overline{M}} - \lambda \frac{\partial D_c}{\partial \lambda} . \qquad (2.131)$$

We can break up the calculation of $\overline{P}$ by conditioning on $M = i$ for $i = 2 \ldots \infty$ and find

$$\overline{P} = \sum_{i=2}^{\infty} Q(i) f_i \ . \tag{2.132}$$

where

$f_i \triangleq$ Probability that i customers are served in a busy period (2.123)

$$Q(i) \triangleq \sum_{n=1}^{i-1} \sum_{m=n+1}^{i} E(C_m^n | M=i) \tag{2.134}$$

For an M/G/1 queue, $D_c$ is given in [4] as

$$D_c = \overline{x} \left( 1 + \frac{\rho(1 + C_b^2)}{2(1 - \rho)} \right) , \tag{2.135}$$

where

$$C_b^2 = \frac{\sigma_b^2}{\overline{x}^2} \ . \tag{2.136}$$

$\sigma_b^2$ denotes the variance and $\overline{x}$ is the mean of the service time distribution. Hence, $\lambda \frac{\partial D_c}{\partial \lambda}$ may be expressed as

$$\lambda \frac{\partial D_c}{\partial \lambda} = \overline{x} \ \frac{(1 + C_b^2)}{2} \ \frac{\rho}{2(1 - \rho)^2}$$

$$= (1 - \rho) \ \frac{\overline{x}}{2} \ (1 + C_b^2) \sum_{k=0}^{\infty} \frac{(k+2)(k+1)}{2} \rho^{k+1} . \tag{2.137}$$

The average number of customers served per busy period is given in [4] by

$$\overline{M} = \frac{1}{1-\rho} \ . \tag{2.138}$$

The z-transform for the probability density of the number of customers served per busy period is described in [2.4] by the following functional equation:

$$F(z) = zB^*[\lambda - \lambda F(z)] \ . \tag{2.139}$$

B* is the one-sided Laplace transform of the service time density and F(z) is defined by

$$F(z) = \sum_{n=1}^{\infty} f_n z^n . \tag{2.140}$$

Having described the quantities that compose the asymptotic bias b, one may express it as

$$b = (1-\rho)\left[\sum_{n=2}^{\infty} Q(n)f_n - \sum_{k=0}^{\infty} \overline{x}(1+C_b^2)\,\frac{(k+2)(k+1)}{4}\,\rho^{k+1}\right]. \tag{2.141}$$

Now suppose $\overline{P}$ in (2.132) may be expressed as a power series in $\rho$.

$$\overline{P} = \sum_{n=2}^{\infty} Q(n) f_n = \sum_{j=0}^{\infty} \alpha_j \rho^{j+1} \qquad (2.142)$$

then based on (2.141), to prove the customer-removal algorithm is asymptotically unbiased, we must show that

$$\alpha_n = \overline{x}(1 + C_b^2) \frac{(n+2)(n+1)}{4} \quad \text{for} \quad n = 0,1,2, \ldots \qquad (2.143)$$

In addition, using (2.142), the power series for b is given by

$$b = [\alpha_0 - \frac{1}{2} \overline{x}(1 + C_b^2)]\rho + \sum_{j=1}^{\infty} [\alpha_j - \alpha_{j-1} - \frac{1}{2} \overline{x}(1 + C_b^2)(j+1)]\rho^{j+1}. \qquad (2.144)$$

For an M/D/1 queue, we were able to prove (2.143) for $n = 0, 1, 2$, which shows that for this case, the estimation algorithm is asymptotically unbiased at least up to third order in $\rho$. Although we could not prove (2.143) for arbitrary n, based on our intuition we believe that for M/D/1 the algorithm is asymptotically unbiased. We have also examined (2.143) for M/M/1 queues. For this case it turns out that (2.143) does not hold even for $n = 0$, which shows that the asymptotic bias contains terms of order $\rho$.

In the remainder of this section we give the detailed proofs indicated above for M/D/1 and M/M/1 queues. The main part is the calculation of Q(i) defined in (2.134), so that we divide the proofs into several steps:

a) Since by (2.121), $C^n_{n+k}$ is given as a function of waiting times and service requirements

$$C^n_{n+k} = \min \{x_n, \omega_{n+1}, \omega_{n+2} \cdots \omega_{n+k}\} \tag{2.145}$$

we calculate the joint density of $(x_n, \omega_{n+1} \cdots \omega_{n+k})$ conditioned on M.

b) Evaluate $E(C^n_{n+k} | M)$.

c) Proof of (2.143) for M/D/1 for n = 0, 1, 2.

d) Calculation of first and second order terms in $\rho$ in the asymptotic bias for M/M/1 queues.

Calculation of Joint Density of $(x_n,\omega_{n+1} \cdots \omega_{n+k})$ Conditioned on M

We approach the problem of deriving $p(x_n,\omega_{n+1} \cdots \omega_{n+k}|M)$ by deriving the joint density of the service and inter-arrival times conditioned on M customers being served in the busy period. Let $\theta_j$ denote the inter-arrival time between the j-th and (j+1)-st customer. By the rules for conditional probabilities, we can break up the joint density of $(x_1, \ldots, x_M, \theta_1, \cdots \theta_M)$ conditioned on M as follows:

$$p(x_1 \ldots x_M, \theta_1 \ldots \theta_M | M) = \frac{p(M|x_1 \ldots x_M, \theta_1 \ldots \theta_M) p(x_1 \ldots x_M, \theta_1 \ldots \theta_M)}{\Pr \{M\}} \tag{2.146}$$

$p(M|x_1 \ldots x_M, \theta_1 \ldots \theta_M)$ is either one or zero depending on whether the variables $(x_1 \ldots x_M, \theta_1 \ldots \theta_M)$ satisfy the constraints such that exactly M customers are served in the busy period. With no conditioning, the service times and inter-arrival times are mutually independent random variables. Since our queue is M/G/1, the inter-arrival times are exponential random variables with parameter $\lambda$. Hence, $p(x_1 \ldots x_M, \theta_1 \ldots \theta_M)$ is given by

$$p(x_1 \ldots x_M, \theta_1 \ldots \theta_M) = \prod_{i=1}^{M} B(x_i) \prod_{i=1}^{M} \lambda e^{-\lambda \theta_i} \tag{2.147}$$

The conditions on the $x_j$'s and $\theta_j$'s that guarantee M customers are served in the busy period come from requiring that the waiting times of customers 2 ... M are greater than zero and satisfying a terminal condition that customer M+1 falls outside the busy period. Earlier we defined the waiting time of the k-th customer as the sum of the service requirements of preceding customers minus his arrival time relative to the start of the busy period. This relative arrival time may be expressed as the sum of the first k-1 inter-arrival times $\theta_j$. Hence, the condition for the waiting times of customers 2 ... M being greater than zero may be expressed as

$$\omega_k = \sum_{j=1}^{k-1} x_j - \sum_{j=1}^{k-1} \theta_j \geq 0 \quad \text{for} \quad k = 2 \ldots M . \qquad (2.148)$$

For the (M+1)-st customer to fall outside the busy period, the arrival time of the (M+1)-st customer relative to the start of the busy period must be greater than the departure time of the M-th customer. Noting that the M-th customer departs when all M service requirements have been satisfied, we define a dummy variable $\omega_{M+1}$ to state the terminal condition.

$$\omega_{M+1} = \sum_{j=1}^{M} x_j - \sum_{j=1}^{M} \theta_j < 0 \qquad (2.149)$$

Hence, the joint density of M service times and inter-arrival times conditioned on M customers being served in the busy period is as follows:

$$p(x_1 \ldots x_M, \theta_1 \ldots \theta_M | M) = \frac{\prod_{i=1}^{M} B(x_i) \prod_{i=1}^{M} \lambda e^{-\lambda \theta_i}}{f_M}$$

$$K=1 \ldots M-1 \quad 0 \leq \theta_K \leq \sum_{j=1}^{k} x_j - \sum_{j=1}^{k-1} \theta_j$$

$$\theta_M > \sum_{j=1}^{M} x_j - \sum_{j=1}^{M-1} \theta_j . \qquad (2.150)$$

As defined by (2.140), $f_M$ denotes the probability that a busy period has M customers.

We can now calculate $p(x_1 \ldots x_M, \omega_2 \ldots \omega_M | M)$ by working with the linear relationship between $(\theta_1 \ldots \theta_M)$ and $(\omega_2 \ldots \omega_{M+1})$ implied by Eqs. (2.148) and (2.149). The inverse relations for the $\theta_j$'s as a function of the $\omega_j$'s are

$$\theta_1 = x_1 - \omega_2$$

$$\theta_k = x_k + \omega_k - \omega_{k+1} \qquad k = 2 \ldots M \qquad (2.151)$$

The non-negativity of the $\vartheta_j$'s for j = 1 ... M and the non-negativity of $\omega_j$ for j = 1 ... M, together with the relation (2.149) for $\omega_{M+1}$, result in the following set of constraints on the waiting times:

$$0 \leq \omega_2 \leq x_1$$

$$0 \leq \omega_{k+1} \leq x_k + \omega_k \qquad k = 2 \ldots M-1 \qquad (2.152)$$

$$\omega_{M+1} < 0$$

The Jacobian J of the inverse transformation between waiting times and inter-arrival times described by (2.151) is the following matrix:

$$J = \begin{pmatrix} -1 & 0 & \cdots & & \\ 1 & -1 & & & \\ 0 & 1 & -1 & & \\ \vdots & & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix} \qquad (2.153)$$

Since J is a lower triangular matrix, the diagonal elements are eigenvalues and hence $|\det J| = 1$. Thus, we calculate $p(x_1 \ldots x_M, \omega_2 \ldots \omega_{M+1} | M)$ by substituting the relations for $\theta_j$ defined by (2.151) into (2.150) and combining this with the constraints described by (2.152).

$$p(x_1 \ldots x_M, \omega_2 \ldots \omega_{M+1} | M) = \frac{\prod_{i=1}^{M} B(x_i)\lambda \; e^{-\lambda\left(\sum_{i=1}^{M} x_i - \omega_{M+1}\right)}}{f_M}$$

$$0 \leq \omega_2 \leq x_1$$

$$0 \leq \omega_{k+1} \leq x_k + \omega_k \qquad k = 2 \ldots M-1$$

$$\omega_{M+1} < 0 \qquad (2.154)$$

The final step is to integrate out the dummy variable $\omega_{M+1}$ and service requirement $x_M$.

$$p(x_1 \ldots x_{M-1}, \omega_2 \ldots \omega_M | M) = \frac{\lambda^{M-1} B^*(\lambda)}{f_M} \prod_{i=1}^{M-1} B(x_i) e^{-\lambda x_i}$$

$$0 \leq \omega_2 \leq x_1$$

$$0 \leq \omega_{k+1} \leq x_k + \omega_k \qquad k = 2 \ldots M-1 \qquad (2.155)$$

$B^*(\lambda)$ denotes the single-sided Laplace transform of the density $B(x)$ evaluated at $\lambda$.

We observe parenthetically that (2.155) takes a particularly simple form from M/D/1 queue. Since the service requirements are deterministic, $B(x_j) = \delta(x_j - x)$ and $B^*(S) = e^{-Sx}$. We can integrate over all the impulses to obtain the joint density of waiting times conditioned on M. Employing the explicit formula for $f_M$, the probability of M customers being served in a busy period, found by solving the functional equation of (2.139) and listed earlier as Eq. (2.110), we find that $p(\omega_2 \ldots \omega_M | M)$ is given by

$$p(\omega_2 \ldots \omega_M | M) = M! \left(\frac{1}{Mx}\right)^{M-1} .$$

$$0 \leq \omega_2 \leq x$$

$$0 \leq \omega_{k+1} \leq x + \omega_k \qquad k = 2 \ldots M-1 \qquad (2.156)$$

## Evaluation of $E(C^n_{n+k}|M)$

We now return to the more general case of an M/G/1 queue and approach the calculation of the mean of $C^n_{n+k}$ conditioned on M by deriving the joint density of $(x_n, \omega_{n+1} \ldots \omega_{n+k})$ conditioned on M. We use this joint density to compute the distribution function of $C^n_{n+k}$. Let $F_{C^n_{n+k}|M}(\tau)$ denote the probability that $C^n_{n+k} \leq \tau$ and $P_{C^n_{n+k}|M}(\tau)$ be the probability density of $C^n_{n+k}$ conditioned on M. Since $C^n_{n+k}$ is defined as the minimum of the variables $(x_n, \omega_{n+1} \ldots \omega_{n+1})$, we compute the probability that $C^n_{n+k} \leq \tau$ as one minus the probability of the complementary event that each variable is greater than $\tau$.

$$F_{C^n_{n+k}|M}(\tau) \triangleq \Pr\{C^n_{n+k} \leq \tau|M\} = 1 - \Pr\{x_n > \tau : \omega_{n+1} > \tau : \ldots \omega_{n+k} > \tau|M\} \tag{2.157}$$

Now we calculate the density of $C^n_{n+k}$ by differentiating with respect to $\tau$.

$$P_{C^n_{n+k}|M}(\tau) = \frac{d}{d\tau} F_{C^n_{n+k}|M}(\tau) \tag{2.158}$$

By integrating the density defined in (2.158) we can calculate the desired conditional mean of $C^n_{n+k}$.

$$E(C^n_{n+k|M}) \triangleq \overline{C}^n_{n+k|M} = \int_{\tau=0}^{\infty} \tau\, p_{C^n_{n+k|M}}(\tau)\, d\tau \qquad (2.159)$$

We obtain the joint density for $(x_n, \omega_{n+1} \cdots \omega_{n+k})$ conditioned on M by integrating first over the variables $(\omega_2 \cdots \omega_n)$, $(\omega_{n+k+1} \cdots \omega_M)$ and then $(x_1 \cdots x_{n-1}, x_{n+1} \cdots x_M)$ in $p(x_1 \cdots x_{M-1}, \omega_2 \cdots \omega_M|M)$ defined by (2.155). For $n \neq 1$, the suggested integrations are performed in two steps and the results listed below.

$$p(x_1 \cdots x_{M-1}, \omega_{n+1} \cdots \omega_{n+k}|M) = \int_{\omega_2=\psi_2}^{x_1} \cdots \int_{\omega_n=\psi_n}^{x_{n-1}+\omega_{n-1}} \int_{\omega_{n+k+1}=0}^{x_{n+k}+\omega_{n+k}} \cdots \int_{\omega_M=0}^{x_{M-1}+\omega_{M-1}}$$

$$p(x_1 \cdots x_{M-1}, \omega_2 \cdots \omega_M|M)\, d\omega_M \cdots d\omega_{n+k+1} \cdots d\omega_2$$

$$\text{for } 0 \leq \omega_{n+1} \leq \sum_{j=1}^{n-1} x_j + x_n$$

$$0 \leq \omega_{n+2} \leq x_{n+1} + \omega_{n+1} \qquad \psi_k \triangleq \max\left\{\omega_{n+1} - \sum_{j=k}^{n} x_j, 0\right\}$$

$$\vdots$$

$$0 \leq \omega_{n+k} \leq x_{n+k-1} + \omega_{n+k-1} \qquad (2.160)$$

$$p(x_n, \omega_{n+1} \cdots \omega_{n+k}|M) = \underbrace{\int \cdots \int}_{\begin{cases} \sum_{i=1}^{n-1} x_i \geq \max\{\omega_{n+1} - x_n, 0\} \\ x_j \geq 0 \qquad j = 1 \ldots n-1 \end{cases}} \int_{x_{n+1} = \varkappa_{n+1}}^{\infty} \cdots \int_{x_{n+k-1} = \varkappa_{n+k-1}}^{\infty}$$

$$\int_{x_{n+k} = 0}^{\infty} \cdots \int_{x_{M-1} = 0}^{\infty}$$

$$p(x_1 \ldots x_{M-1} \omega_{n+1} \cdots \omega_{n+k}|M)$$

$$dx_1 \ldots dx_{n-1} dx_{n+1} \cdots dx_{M-1} \tag{2.161}$$

$$\varkappa_\ell \triangleq \max\{\omega_{\ell+1} - \omega_\ell, 0\}$$

Note that the resulting density in (2.161) has no constraints remaining on any of the variables except non-negativity. The case n=1 is worth distinguishing since in the final density one constraining relation remains between $\omega_2$ and $x_i$.

$$p(x_1 \ldots x_{M-1}, \omega_2 \cdots \omega_{1+k}|M) = \int_{\omega_{2+k}=0}^{x_{1+k} + \omega_{1+k}} \cdots \int_{\omega_M = 0}^{x_{M-1} + \omega_{M-1}}$$

$$p(x_1 \ldots x_{M-1}, \omega_2 \cdots \omega_M|M) d\omega_M \ldots d\omega_{2+k} \tag{2.162}$$

for $0 \leq \omega_2 \leq x_1$

$$0 \leq \omega_{1+\ell} \leq x_\ell + \omega_\ell \qquad \ell = 2 \ldots k$$

$$p(x_1,\omega_2 \cdots \omega_{1+k}|M) = \int_{x_2=\varkappa_2}^{\infty} \cdots \int_{x_k=\varkappa_k}^{\infty} \int_{x_{1+k}=0}^{\infty} \cdots \int_{x_{M-1}=0}^{\infty} p(x_1 \cdots x_{M-1},\omega_2 \cdots \omega_{1+k}|M)\, dx_{M-1} \cdots dx_{1+k} \cdots dx_2 \tag{2.163}$$

for $0 \leq \omega_2 \leq x_1$

$$\varkappa_\ell \triangleq \max\,\{\omega_{\ell+1} - \omega_\ell, 0\}$$

Hence, the distribution function $F_{C^n_{n+k|M}}(\tau)$ is obtained by applying (2.157) to the results in (2.161) and (2.163).

$$F_{C^n_{n+k|M}}(\tau) = \begin{cases} n\neq 1 & 1 - \displaystyle\int_{x_n=\tau}^{\infty} \int_{\omega_{n+1}=\tau}^{\infty} \cdots \int_{\omega_{n+k}=\tau}^{\infty} p(x_n,\omega_{n+1} \cdots \omega_{n+k}|M)\, d\omega_{n+k} \cdots dx_n \\ \\ n=1 & 1 - \displaystyle\int_{x_1=\tau}^{\infty} \int_{\omega_2=\tau}^{x_1} \int_{\omega_3=\tau}^{\infty} \cdots \int_{\omega_{1+k}=\tau}^{\infty} p(x_1,\omega_2 \cdots \omega_{1+k}|M)\, d\omega_{1+k} \cdots d\omega_2\, dx_1 \end{cases} \tag{2.164}$$

$P_{C^n_{n+k|M}}(\tau)$ could now be obtained by differentiating the above integrals with respect to the parameter $\tau$ using Leibnitz's rule.

Simpler results for $\overline{C}^n_{n+k|M}$ are possible when we specialize the preceding calculations to the case of an M/D/1 queue. Since the service requirements are all the same, $x_i=x$ and $C^n_{n+k}$ can be expressed as

$$C^n_{n+k} = \min\ \{x,z\}\ , \tag{2.165}$$

where

$$z = \min\ \{\omega_{n+1},\ \dots\ \omega_{n+k}\}\ . \tag{2.166}$$

By inspection we write the density of $C^n_{n+k}$ conditioned on M in terms of the distribution function of z, $F_{z|M}(\tau)$, and the density of z, $P_{z|M}(\tau)$.

$$P_{C^n_{n+k|M}}(\tau) = (1 - F_{z|M}(x))\delta(\tau-x) + U_{-1}(x-\tau)P_z(\tau) \tag{2.167}$$

There is a finite probability that $C^n_{n+k} = x$, hence the impulse in the density (2.167). For $z < x$, $C^n_{n+k} = z$ and hence the density of $C^n_{n+k}$, $P_{C^n_{n+k|M}}(\tau)$, is the same as the density of z for

$0 \leq \tau < x$. From (2.167), we can express the conditional mean of $C^n_{n+k}$ as

$$\overline{C}^n_{n+k|M} = (1 - F_{z|M}(x))x + \int_0^x \tau P_{z|M}(\tau)\, d\tau . \qquad (2.168)$$

Hence, for an M/D/1 queue, the problem of computing $\overline{C}^n_{n+k}|M$ reduces to calculating the density and distribution function of z defined in (2.166), conditioned on M. We need the joint distribution of $(\omega_{n+1} \cdots \omega_{n+k})$ conditioned on M to characterize the distribution function of z. In a manner similar to the derivation of (2.160) from (2.155), we integrate out the unneeded $\omega_i$'s from the density $p(\omega_2 \cdots \omega_M|M)$ given in Eq. (2.156) to obtain $p(\omega_{n+1} \cdots \omega_{n+k}|M)$.

$$p(\omega_{n+1} \cdots \omega_{n+k}|M) = M!\left(\frac{1}{Mx}\right)^{M-1} \int_{\omega_2=\psi_2}^{x} \int_{\omega_3=\psi_3}^{x+\omega_2} \cdots \int_{\omega_n=\psi_n}^{x+\omega_{n-1}} \int_{\omega_{n+k+1}=0}^{x+\omega_{n+k}}$$

$$\cdots \int_{\omega_M=0}^{x+\omega_{M-1}} d\omega_M \cdots d\omega_{n+k+1} \cdots d\omega_2$$

$$\text{for} \quad 0 \leq \omega_{n+1} \leq nx$$

$$0 \leq \omega_{n+\ell} \leq x + \omega_{n+\ell-1} \qquad \ell = 2 \ldots k$$

$$\psi_\ell \triangleq \max\{\omega_{n+1} - (n-\ell+1)x, 0\}$$

(2.169)

Hence, by the same reasoning that led to (2.166), $F_{z|M}(\tau)$ is given by

$$F_{z|M}(\tau) = 1 - \int_{\omega_{n+1}=\tau}^{nx} \int_{\omega_{n+2}=\tau}^{x+\omega_{n+1}} \cdots \int_{\omega_{n+k}=\tau}^{x+\omega_{n+k-1}} p(\omega_{n+1} \cdots \omega_{n+k}|M) \, d\omega_{n+k} \cdots d\omega_{n+1} \; . \tag{2.170}$$

Applying Leibnitz's rule successively, we can derive $P_{z|M}(\tau)$.

$$\begin{aligned} P_{z|M}(\tau) = & \int_{\omega_{n+2}=\tau}^{x+\tau} \int_{\omega_{n+3}=\tau}^{x+\omega_{n+2}} \cdots \int_{\omega_{n+k}=\tau}^{x+\omega_{n+k-1}} p(\tau,\omega_{n+2}, \cdots \omega_{n+k}|M) \, d\omega_{n+k} \cdots d\omega_{n+2} \\ & + \int_{\omega_{n+1}=\tau}^{nx} \int_{\omega_{n+3}=\tau}^{x+\tau} \cdots \int_{\omega_{n+k=\tau}}^{x+\omega_{n+k-1}} p(\omega_{n+1},\tau,\omega_{n+3} \cdots \omega_{n+k}|M) \, d\omega_{n+k} \cdots d\omega_{n+2} \\ & \vdots \\ & + \int_{\omega_{n+1}=\tau}^{nx} \int_{\omega_{n+2}=\tau}^{x+\omega_{n+1}} \cdots \int_{\omega_{n+k}=\tau}^{x+\tau} p(\omega_{n+1} \cdots \omega_{n+k-2},\tau,\omega_{n+k}|M) \, d\omega_{n+k} \cdots d\omega_{n+1} \\ & + \int_{\omega_{n+1}=\tau}^{nx} \int_{\omega_{n+2}=\tau}^{x+\omega_{n+1}} \cdots \int_{\omega_{n+k-1}=\tau}^{x+\omega_{n+k-2}} p(\omega_{n+1} \cdots \omega_{n+k-1},\tau|M) \, d\omega_{n+k-1} \cdots d\omega_{n+1} \end{aligned} \tag{2.171}$$

In the next few paragraphs we show how to compute explicitly $\overline{C}^{n}_{n+1}|M$. We were not able, however, to obtain explicit terms for a general $\overline{C}^{n}_{m}|M$ and therefore have to rely on (2.173) when other $\overline{C}^{n}_{m}|M$ will be needed. For $\overline{C}^{n}_{n+1}|M$, z becomes the single variable $\omega_{n+1}$. Note that to compute the conditional mean in (2.168) we only need $P_{z|M}(\tau)$ for $0 \leq \tau \leq x$. Hence, the approach we take is to derive the explicit form for $P_{\omega_{n+1}|M}(\tau)$ on $0 \leq \tau \leq x$. From (2.169), we write the density for $P_{\omega_{n+1}|M}(\tau)$ on $0 \leq \tau \leq x$ as

$$P_{\omega_{n+1}|M}(\tau) = \frac{M!}{(Mx)^{M-1}} \underbrace{\int_{\omega_2=0}^{x} \int_{\omega_3=0}^{x+\omega_2} \cdots \int_{\omega_n=0}^{x+\omega_{n-1}}}_{\substack{n-1\\ \text{Integrations}}} \underbrace{\int_{\omega_{n+2}=0}^{x+\tau} \cdots \int_{\omega_M=0}^{x+\omega_{M-1}}}_{\substack{M-n-1\\ \text{Integrations}}} d\omega_M \cdots d\omega_{n+2} d\omega_n \cdots d\omega_2 \qquad (2.172)$$

Note that the above integral consists of two sets of integrations defined by the brackets. We define the following set of iterated integrals

$$\tilde{I}_{n+1}(\omega) = \int_{\omega=0}^{x+\omega} \tilde{I}_n(\omega')d\omega' \text{ with } \tilde{I}_1(\omega) = 1 \qquad (2.173)$$

Hence, we can express $P_{\omega_{n+1}|M}(\tau)$ for $0 \leq \tau \leq x$ in terms of these $\tilde{I}_j$'s.

$$P_{\omega_{n+1|M}}(\tau) = M!\left(\frac{1}{Mx}\right)^{M-1}\tilde{I}_n(0)\tilde{I}_{M-n}(\tau) \qquad (2.174)$$

$$0 \le \tau \le x$$

Examining a few of the $\tilde{I}_j$'s, we assume that $\tilde{I}_n(\omega)$ can be represented as a power series in $(x+\omega)$ with powers up to (n-1)-st order.

$$\tilde{I}_n(\omega) = \sum_{i=0}^{n-1} \Omega_i^{(n)}(x+\omega)^i \qquad (2.175)$$

To be consistent with $\tilde{I}_1(\omega) = 1$, $\Omega_0^{(1)} = 1$. Employing (2.173) we can derive a set of difference equations relating $\Omega_j^{(n+1)}$ to the $\Omega_i^{(n)}$'s.

$$\begin{cases} \Omega_q^{(n+1)} = \displaystyle\sum_{i=q-1}^{n-1} \frac{1}{q}\binom{i}{q-1} x^{i-q+1}\Omega_i^{(n)} \\ \qquad q=1 \ldots n \\ \\ \Omega_0^{(n+1)} = 0 \end{cases} \qquad (2.176)$$

It is verified in Appendix A that the solution for the $\Omega_i^{(n)}$'s are given by

$$\text{for } n > 1 \;\; \Omega_i^{(n)} = \begin{cases} \dfrac{x^{n-1-i}(n-1)^{n-2-i}}{(n-1-i)!(i-1)!} & i = 1 \ldots n-1 \\ 0 & i = 0 \end{cases}$$

$$\text{for } n = 1 \;\; \Omega_0^{(1)} = 1 \tag{2.177}$$

Employing (2.168), the expression for the density $P_{\omega_{n+1}|M}(\tau)$ given in (2.174), and the definition of the $\tilde{I}_j$'s in (2.173), $\overline{C}^n_{n+1}|M$ may be expressed as

$$\overline{C}^n_{n+1}|M = (1 - M!(\frac{1}{Mx})^{M-1}\tilde{I}_n(0)\tilde{I}_{M-n+1}(0))x$$

$$+ M!(\frac{1}{Mx})^{M-1}\tilde{I}_n(0) \int_0^{\tau} \tau I_{M-n}(\tau)d\tau \; . \tag{2.178}$$

By looking at $P_{\omega_{n+1}|M}(\tau)$ for $n=1$ we can obtain an expression for $\tilde{I}_M(0)$ since the density for $\omega_2$ must integrate to 1 over the interval $[0,x]$. Hence, $\tilde{I}_M(0)$ must be given by

$$\tilde{I}_M(0) = \frac{(Mx)^{M-1}}{M!} \; . \tag{2.179}$$

To specify the only remaining unknown in (2.178), we define

$$m_n \triangleq \int_{\tau=0}^{x} \tau\tilde{I}_n(\tau)\, d\tau \; . \tag{2.180}$$

We evaluate this integral by substituting our explicit form for $\tilde{I}_n(\tau)$ defined by (2.175) and (2.177)

$$m_n = \begin{cases} \frac{1}{2}x^2 & n = 1 \\ \\ \sum_{i=1}^{n-1} \frac{x^{n+1}(n-1)^{n-2-i}}{(n-1-i)!(i-1)!} \sum_{\ell=0}^{i} \frac{1}{\ell+2} \binom{i}{\ell} & n \neq 1 \end{cases} \tag{2.181}$$

Putting together (2.179), (2.181), and (2.178), $\overline{C}^n_{n+1}|M$ may be expressed as

$$\overline{C}^n_{n+1}|M = (1 - M!(\frac{1}{Mx})^{M-1} \frac{(nx)^{n-1}}{n!} \frac{((M-n+1)x)^{M-n}}{(M-n+1)!})x$$

$$+ M!(\frac{1}{Mx})^{M-1} \frac{(nx)^{n-1}}{n!} m_{M-n} \, . \tag{2.182}$$

Bias Calculation for an M/D/1 Queue

Having defined the calculation of $\overline{C}^n_{n+k}|M=i$ and hence Q(i), we can investigate the behavior of the asymptotic bias b expressed as a power series in $\rho$ by Eq. (2.144). For simplicity, we start with an M/D/1 queue. $C_b^2 = 0$ in (2.136), since the variance of the service time density is zero. Since each $C^n_{n+k}$ is a function of $(x,\omega_{n+1} \cdots \omega_{n+k})$, the summation of $C^n_{n+k}$'s for a given busy period with M customers can be expressed as different functions of $(x,\omega_2 \cdots \omega_M)$ over regions in $(\omega_2 \cdots \omega_M)$ space. Hence Q(M) can be expressed as the sum of a set of integrals of functions of $(x, \omega_2 \ldots \omega_M)$ over regions in $(\omega_2 \ldots \omega_M)$ space weighted by the density $p(\omega_2 \ldots \omega_M | M) = M!(\frac{1}{Mx})^{M-1}$ specified in Eq. (2.156). Since the density $p(\omega_2 \ldots \omega_M)$ is a constant, independent of $\rho$, Q(M) will also be a constant with no dependence on $\rho$. $f_n$, the probability of n customers being served in a busy period, is $\frac{n^{n-1}}{n!} \rho^{n-1} e^{-n\rho}$. Hence, expanding $f_n$ as a power series in $\rho$, substituting the result into the expression for $\overline{p}$ in (2.142), and changing the order of summation, we can derive an explicit formula for $\alpha_n$, the coefficient of $\rho^{n+1}$ in a power series expansion of $\overline{p}$ in powers of $\rho$. Therefore, on the basis of (2.143), to show that the estimation procedure is unbiased up to the $\ell$-th power in $\rho$, we have to show that

$$\begin{cases} \alpha_n = \sum_{i=2}^{n+2} \frac{(-1)^{n-i} i^{n+1}}{i!(n+2-i)!} Q(i) = x \frac{(n+2)(n+1)}{4} \\ \\ n = 0 \ldots \ell-1 \end{cases} \qquad (2.183)$$

By a change of variable n = n'-2 and some manipulation, verifying conditions (2.183) may be reformulated as checking the following recursion for the Q(i)'s:

$$\begin{cases} Q(n') = \dfrac{n!}{(n')^{n'-1}} \left(\dfrac{n'(n'-1)}{4} x\right) \\ \\ \qquad - \dfrac{n!}{(n')^{n'-1}} \displaystyle\sum_{i=2}^{n'-1} \dfrac{(-1)^{n'-i} i^{n'-1}}{i!(n'-i)!} Q(i) \\ \\ n' = 3 \ldots \ell+1 \end{cases}$$

$$\text{with } Q(2) = \frac{1}{2} x \tag{2.184}$$

Employing the preceding results for an M/D/1 queue, we proceed to show that the asymptotic bias b only contains powers of $\rho$ greater than third order. For $\ell=3$, relation (2.184) dictates that Q(2), Q(3), and Q(4) assume the following values:

$$\begin{aligned} Q(2) &= \frac{1}{2} x \\ Q(3) &= \frac{5}{3} x \\ Q(4) &= \frac{57}{16} x \end{aligned} \tag{2.185}$$

By definition of the Q(i)'s given in (2.134), Q(2), Q(3), and Q(4) are given by

$$\begin{aligned} Q(2) &= \bar{C}_2^1|2 \\ Q(3) &= \bar{C}_2^1|3 + \bar{C}_3^1|3 + \bar{C}_3^2|3 \\ Q(4) &= \bar{C}_2^1|4 + \bar{C}_2^1|4 + \bar{C}_4^1|4 + \bar{C}_3^2|4 + \bar{C}_4^2|4 + \bar{C}_4^3|4 \end{aligned} \tag{2.186}$$

We can compute all $\overline{C}_m^n|M$'s of the form $\overline{C}_{n+1}^n|M$ by employing the general formula derived as Eq. (2.182). To calculate other $\overline{C}_m^n|M$'s we appeal to the procedure outlined in Eqs. (2.165) through (2.171). The results of the calculations implied in (2.186) are listed below.

$$\overline{C}_2^1|2 = \frac{1}{2}x$$

$$\overline{C}_2^1|3 = \frac{5}{9}x \qquad \overline{C}_2^1|3 = \frac{4}{9}x$$

$$\overline{C}_3^2|3 = \frac{2}{3}x$$

$$\overline{C}_2^1|4 = \frac{37}{64}x \qquad \overline{C}_3^1|4 = \frac{1}{2}x \qquad \overline{C}_4^1|4 = \frac{27}{64}x$$

$$\overline{C}_3^2|4 = \frac{3}{4}x \qquad \overline{C}_4^2|4 = \frac{19}{32}x$$

$$\overline{C}_4^3|4 = \frac{23}{32}x \tag{2.187}$$

By using the results of (2.187) in (2.186) it is easy to see that the numbers for Q(2), Q(3), and Q(4) are consistent with the values listed in (2.185).

For the sake of clarity, we go through the derivation of $\overline{C}_4^2|4$ as a sample calculation. The remaining $\overline{C}_m^n|M$'s, those not covered by relation (2.182), are calculated in Appendix B. $C_4^2$ is defined as follows:

$$C_4^2 = \min\{x,z\} \ , \qquad (2.188)$$

with

$$z = \min\{\omega_3,\omega_4\} \ . \qquad (2.189)$$

the joint distribution of $(\omega_2,\omega_3,\omega_4)$ conditioned on M=4 is given by Eq. (2.156) as

$$p(\omega_2,\omega_3,\omega_4|M=4) = \frac{3}{8x^3} \ . \qquad (2.190)$$

$$0 \leq \omega_2 \leq x$$

$$0 \leq \omega_3 \leq x + \omega_2$$

$$0 \leq \omega_4 \leq x + \omega_3$$

Hence, the joint distribution of $(\omega_2,\omega_4)$ conditioned on M=4 is given by Eq. (2.169) as

$$p(\omega_3,\omega_4|M=4) = \frac{3}{8x^3} \int_{\omega_2 = \max\{0,\omega_3-x\}}^{x} d\omega_2 \ . \qquad (2.191)$$

$$0 \leq \omega_3 \leq 2x$$

$$0 \leq {}_4 \leq x + \omega_3$$

More explicitly, (2.191) means that

$$p(\omega_3,\omega_4|M{=}4) = \frac{3}{8x^2} \quad \text{and} \quad p(\omega_3,\omega_4|M{=}4) = \frac{3}{8x^3}(2x - \omega_3) .$$

$$0 \leq \omega_3 \leq x \qquad\qquad x \leq \omega_3 \leq 2x$$

$$0 \leq \omega_4 \leq x + \omega_3 \qquad\qquad 0 \leq \omega_4 \leq x + \omega_3 \tag{2.192}$$

Hence, by (2.170) the distribution function of z conditioned on M=4 and for z in the interval [0,x] is given by

$$F_{z|M=4}(\tau) = 1 - \int_{\omega_3=\tau}^{x} \int_{\omega_4=\tau}^{x+\omega_3} \frac{3}{8x^2}\, d\omega_4 d\omega_3$$

$$0 \leq \tau \leq x$$

$$- \int_{\omega_3=x}^{2x} \int_{\omega_4=\tau}^{x+\omega_3} \frac{3}{8x^3}(2x - \omega_3)\, d\omega_4 d\omega_3 . \tag{2.193}$$

Differentiating (2.193) with respect to $\tau$ and doing the integrations yields

$$P_{z|M=4}(\tau) = \frac{15}{16x} - \frac{3}{8x^2}\tau . \tag{2.194}$$

$$0 \leq \tau \leq x$$

From Eq. (2.168), we compute $\overline{C}_4^2|4$ as follows:

$$\overline{C}_4^2|4 = x(1 - \int_0^x P_{z|M=4}(\tau)\, d\tau) + \int_0^x \tau P_{z|M=4}(\tau)\, d\tau = \frac{19}{32}x . \tag{2.195}$$

## Bias Calculation for an M/M/1 Queue

We now examine the behavior of the asymptotic bias b for the case of an M/M/1 queue. The service requirements are now exponential random variables with parameter $\mu$. Hence, B(x) and B*(s) are given by

$$\begin{aligned} B(x) &= \mu e^{-\mu x} \\ B^*(s) &= \frac{\mu}{s+\mu} \end{aligned} \qquad (2.196)$$

Since the mean service time is $\frac{1}{\mu}$ and the variance of the service requirement is $\frac{1}{\mu^2}$, the $C_b^2$ defined in Eq. (2.136) is one. The functional equation of (2.139) may be solved to obtain F(z) and hence $f_i$, the probability of i customers being served in the busy period

$$f_i = \frac{1}{i} \binom{2i-2}{i-1} \frac{\rho^{i-1}}{(1+\rho)^{2i-1}} \qquad (2.197)$$

To characterize the asymptotic bias b we need $\alpha_j$, the coefficient of $\rho^{j+1}$ in a power series expansion of $\overline{p}$, which is defined in (2.132) as a summation of the products $Q(i)f_i$ for $i = 2 \ldots \infty$. We contend that $Q(i)f_i$ can be expressed as a power series in $\rho$ whose terms are at least (i-1)-st order in $\rho$. Hence, to find $\alpha_j$ we need to collect coefficients of $\rho^{j+1}$ in $Q(2)f_2 \ldots Q(j+2)f_{j+2}$. The reason for $Q(i)f_i$ being expressible as a power series in $\rho$ with terms of at least (i-1)-st orders follows from Eq. (2.155) for $p(x_1 \ldots x_{M-1}, \omega_2 \ldots \omega_M | M)$. Q(M) represents the mean of a random variable which is expressed as a summation of $C_m^n$'s, each of which is a function of $(x_1 \ldots x_{M-1}, \omega_2 \ldots \omega_M)$. Hence, Q(M) could be expressed as an integral of various functions of $(x_1 \ldots x_{M-1}, \omega_2 \ldots \omega_M)$ weighted by $p(x_1 \ldots x_{M-1}, \omega_2 \ldots \omega_M | M)$, over $(x_1 \ldots x_{M-1}, \omega_2 \ldots \omega_M)$ space. From the density given in (2.155), this integral for Q(M) will have a leading factor of $\lambda^{M-1}/f_M$. Hence, $Q(M)f_M$ can be expressed as the product of $\rho^{M-1}$ and an appropriate integral. This shows that $Q(M)f_M$ can be represented as a sum of terms which are (M-1)-st order or higher in $\rho$.

We now proceed to compute Q(2) and Q(3) so we can calculate $\alpha_0$ and $\alpha_1$ and then the coefficients of the first and second power of $\rho$ in the asymptotic bias b defined in Eq. (2.144). Q(2) is equal to $\overline{C_2^1}|2$ and $C_2^1$ is given by

$$C_2^1 = \min \{x_1, \omega_2\} \ . \tag{2.198}$$

From (2.155), the joint density of $(x_1, \omega_2)$ conditioned on M=2 is as follows:

$$p(x_1, \omega_2 | M=2) = \frac{\lambda B^*(\lambda)}{f_2} B(x_1) e^{-\lambda x_1} \ . \tag{2.199}$$

$$0 \leq \omega_2 \leq x_1$$

Hence, by (2.164) the distribution function of $C_2^1$ conditioned on M=2 is

$$F_{C_2^1|M=2}(\tau) = 1 - \frac{\lambda B^*(\lambda)}{f_2} \int_{x_1=\tau}^{\infty} \int_{\omega_2=\tau}^{x_1} B(x_1)e^{-\lambda x_1} d\omega_2 dx_1 \ . \qquad (2.200)$$

Differentiating with respect to $\tau$ by applying Leibnitz's rule, we obtain

$$P_{C_2^1|M=2}(\tau) = \frac{\lambda B^*(\lambda)}{f_2} \int_{x_1=\tau}^{\infty} B(x_1)e^{-\lambda x_1} dx_1 \qquad (2.201)$$

Employing (2.201), we can express the desired conditional mean as a double integral. Changing the order of integration, we obtain the following result for an M/G/1 queue:

$$Q(2) = \overline{C}_2^1|2 = \frac{1}{2} \ \frac{\lambda B^*(\lambda)}{f_2} \int_{x_1=0}^{\infty} x_1^2 e^{-\lambda x_1} B(x_1)dx_1 \qquad (2.202)$$

For an M/M/1 queue, (2.202) specializes to

$$Q(2) = \overline{C}_2^1|2 = \frac{1}{2} \ \frac{1}{\mu} \ \frac{\rho}{(1+\rho)^4} \frac{1}{f_2} \qquad (2.203)$$

Next, we calculate Q(3) as the sum of $\overline{C}_2^1|3$, $\overline{C}_3^1|3$, and $\overline{C}_3^2|3$. The joint density of $(x_1, \omega_2)$ conditioned on M=3 is specified by Eqs. (2.155), (2.162), and (2.163) as

$$p(x_1,\omega_2|M{=}3) = \frac{\lambda^2 B^*(\lambda)}{f_3} \int_{x_2=0}^{\infty} \int_{\omega_3=0}^{x_2+\omega_2} B(x_1)B(x_2)e^{-\lambda(x_1+x_2)} d\omega_3 dx_2 \text{ .}$$

$$0 \leq \omega_2 \leq x_1 \qquad (2.204)$$

Performing the inner integration over $\omega_3$ and changing the order of the integration over the variables $\omega_2$ and $x_1$ specified in (2.164), the distribution function of $C_2^1$ conditioned on M=3 is given by

$$F_{C_2^1|M=3}(\tau) = 1 - \frac{\lambda^2 B^*(\lambda)}{f_3} \int_{\omega_2=\tau}^{\infty} \int_{x_1=\omega_2}^{\infty} \int_{x_2=0}^{\infty} (x_2+\omega_2)B(x_1)B(x_2)e^{-\lambda(x_1+x_2)} dx_2dx_1d\omega_2 \text{ .} \qquad (2.205)$$

Differentiating (2.205) with Liebnitz's rule we obtain $P_{C_2'|M=3}(\tau)$. Based on this $P_{C_2'|M=3}(\tau)$ we can write the following integral expression for the conditional mean:

$$\bar{C}_2^1|3 = \frac{\lambda^2 B^*(\lambda)}{f_3} \int_{\tau=0}^{\infty} \tau \int_{x_1=\tau}^{\infty} \int_{x_2=0}^{\infty} (x_2+\tau)B(x_1)B(x_2)e^{-\lambda(x_1+x_2)} dx_2dx_1d\tau \text{ .} \qquad (2.206)$$

For an M/M/1 queue we employ the B(x) specified by Eq. (2.196) and obtain the result

$$\bar{C}_2^1|3 = 3\,\frac{1}{\mu}\,\frac{\rho^2}{(1+\rho)^6}\,\frac{1}{f_3} \text{ .} \qquad (2.207)$$

We next consider the calculation of $\overline{C_3^2}|3$. $C_3^2$ is defined as

$$C_3^2 = \min\{x_2,\omega_3\} \tag{2.208}$$

By Eqs. (2.155), (2.160), and (2.161), the joint density of $(x_2,\omega_3)$ conditioned on M=3 is

$$p(x_2,\omega_3|M=3) = \frac{\lambda^2 B^*(\lambda)}{f_3} \int_{x_1=\max\{\omega_3-x_2,0\}}^{\infty} \int_{\omega_2=\max\{\omega_3-x_2,0\}}^{x_1} B(x_1)B(x_2)e^{-\lambda(x_1+x_2)} d\omega_2 dx_1 \ . \tag{2.209}$$

Doing the inner integration over $\omega_2$, and noting the form for $F_{C_3^2|M=3}(\tau)$ dictated by (2.164), differentiation of $F_{C_3^2|M=3}(\tau)$ with respect to $\tau$ by Leibnitz's rule will yield the following integral

$$P_{C_3^2|M=3}(\tau) = \int_{x_2=\tau}^{\infty} P(x_2,\tau|M=3)dx_2 + \int_{\omega_3=\tau}^{\infty} P(\tau,\omega_3|M=3)d\omega_3 \tag{2.210}$$

Employing (2.209) and (2.210), the desired conditional mean may be expressed as the following integral:

$$\overline{C}_3^2|3 = \frac{\lambda^2 B^*(\lambda)}{f_3}\left\{\int_{\tau=0}^{\infty} \tau \int_{x_2=\tau}^{\infty} B(x_2)e^{-\lambda x_2}\int_{x_1=0}^{\infty} x_1 B(x_1)e^{-\lambda x_1}\, dx_1 dx_2 d\tau\right.$$

$$\left. + \int_{\tau=0}^{\infty} \tau B(\tau)e^{-\lambda\tau}\int_{\omega_3=\tau}^{\infty}\int_{x_1=\omega_3-\tau}^{\infty} B(x_1)e^{-\lambda x_1}[x_1 - (\omega_3-\tau)]\, dx_1 d\omega_3 d\tau\right\}$$

$$(2.211)$$

Using the B(x) corresponding to an M/M/1 queue, we obtain

$$\overline{C}_3^2|3 = 2\,\frac{1}{\mu}\,\frac{\rho^2}{(1+\rho)^6}\,\frac{1}{f_3} \qquad (2.212)$$

Our final calculation is of $\overline{C}_3^1|3$. The quantity $C_3^1$ is defined by

$$C_3^1 = \min\,\{x_1, \omega_2, \omega_3\}\ . \qquad (2.213)$$

The joint density of $(x_1, \omega_2, \omega_3)$ conditioned on M=3 is determined by Eqs. (2.155), (2.162), and (2.163).

$$P(x_1,\omega_2,\omega_3|M=3) = \frac{\lambda^2 B^*(\lambda)}{f_3}\int_{x_2=\max\,\{\omega_3-\omega_2,0\}}^{\infty} B(x_1)B(x_2)e^{-\lambda(x_1+x_2)}\,dx_2$$

$$0 \le \omega_2 \le x_1$$

$$(2.214)$$

Employing (2.164) to derive the distribution function $F_{C_2^1|M=3}(\tau)$ and breaking up the integration over $\omega_3$ to get rid of the "max" in the limits of the integration over $x_2$, we obtain

$$F_{C_3^1|M=3}(\tau) = 1 - \frac{\lambda^2 B^*(\lambda)}{f_3} \int_{x_1=\tau}^{\infty} \int_{\omega_2=\tau}^{x_1} \int_{\omega_3=\tau}^{\omega_2} \int_{x_2=0}^{\infty} B(x_1)B(x_2)e^{-\lambda(x_1+x_2)} \, dx_2 d\omega_3 d\omega_2 dx_1$$

$$- \frac{\lambda^2 B^*(\lambda)}{f_3} \int_{x_1=\tau}^{\infty} \int_{\omega_2=\tau}^{x_1} \int_{\omega_3=\omega_2}^{\infty} \int_{x_2=\omega_3-\omega_2}^{\infty} B(x_1)B(x_2)e^{-\lambda(x_1+x_2)} \, dx_2 d\omega_3 d\omega_2 dx_1 \, . \qquad (2.215)$$

Differentiating (2.215) by applying Liebnitz's rule we can derive $P_{C_3^1|M=3}(\tau)$ and from that a single integral expression for $\overline{C}_3^1|3$.

$$\overline{C}_3^1|3 = \frac{\lambda^2 B^*(\lambda)}{f_3} \int_{\tau=0}^{\infty} \tau \int_{x_1=\tau}^{\infty} \int_{\omega_2=\tau}^{x_1} \int_{x_2=0}^{\infty} B(x_1)B(x_2)e^{-\lambda(x_1+x_2)} dx_2 d\omega_2 dx_1 d\tau$$

$$+ \frac{\lambda^2 B^*(\lambda)}{f_3} \int_{\tau=0}^{\infty} \tau \int_{x_1=\tau}^{\infty} \int_{\omega_3=\tau}^{\infty} \int_{x_2=\omega_3-\tau}^{\infty} B(x_1)B(x_2)e^{-\lambda(x_1+x_2)} \, dx_2 d\omega_3 dx_1 d\tau \qquad (2.216)$$

Using the B(x) for an M/M/1 queue given by (2.196), we find

$$\overline{C}_3^1|3 = 5 \, \frac{1}{\mu} \, \frac{\rho^2}{(1+\rho)^6} \, \frac{1}{f_3} \, . \qquad (2.217)$$

Now we can calculate the coefficients in the asymptotic bias b for the first and second order terms in $\rho$. Employing (2.203), (2.207), (2.212), and (2.217), we compute Q(2) and Q(3) as

$$Q(2) = \frac{1}{2}\,\frac{1}{\mu}\,\frac{\rho}{(1+\rho)^4}\,\frac{1}{f_2}$$

$$Q(3) = 10\,\frac{1}{\mu}\,\frac{\rho^2}{(1+\rho)^6}\,\frac{1}{f_3} \tag{2.218}$$

$\alpha_0$ is defined by the coefficient of $\rho$ in $Q(2)f_2$. $\alpha_1$ is determined by collecting powers of $\rho^2$ in $Q(2)f_2$ and $Q(3)f_3$. Hence, we obtain

$$\alpha_0 = \frac{1}{2}\,\frac{1}{\mu}$$

$$\alpha_1 = \frac{8}{\mu} \tag{2.219}$$

Therefore, from (2.144) we can represent the asymptotic bias b as

$$b = -\frac{1}{2}\,\frac{1}{\mu}\,\rho + \frac{11}{2}\,\frac{1}{\mu}\,\rho^2 + O(\rho^3)\ . \tag{2.220}$$

$O(\rho^3)$ denotes terms that are third order or higher in $\rho$. This shows that the algorithm is biased for M/M/1 and contains terms of all orders in $\rho$, in contradistinction with the M/D/1 case where we showed that the algorithm is asymptotically unbiased up to third order in $\rho$. In fact, we are confident that for an M/D/1 queue the algorithm is completely unbiased, although we were not able to show this.

## 2.7 Cramer-Rao Bound for Customer-Removal Algorithm in Case of M/D/1 Queue

Now we derive a Cramer-Rao bound for the variance of any unbiased estimator of the delay gradient that works with the same observations as the customer-removal algorithm in the case of an M/D/1 queue. Assuming the service requirement is a known variable x, the observations which the customer removal algorithm processes to derive the $C_m^n$'s and form the estimator are the $M_i$'s, the number of customers served in the i-th busy period, and $(\omega_2^{(i)} \ldots \omega_{M_i}^{(i)})$, the waiting times of customers in the i-th busy period. Hence, we define our observation vector Y' as

$$Y' \triangleq (M_1,\omega_2^{(1)} \ldots \omega_{M_1}^{(1)} \,\vdots\, M_2,\omega_2^{(2)} \ldots \omega_{M_2}^{(2)} \,\vdots \ldots \vdots\, M_N,\omega_2^{(N)} \ldots \omega_{M_N}^{(N)}) . \tag{2.221}$$

Since waiting times and the number of customers served per busy period are statistically independent from one busy period to another, the joint density of Y' may be expressed as

$$P(Y') = \prod_{i=1}^{N} P(M_i,\omega_2^{(i)} \ldots \omega_{M_i}^{(i)}) . \tag{2.222}$$

We compute the joint density of $(M_i, \omega_2^{(i)} \ldots \omega_{M_i}^{(i)})$ by breaking it into the product of $P(\omega_2^{(i)} \ldots \omega_{M_i}^{(i)} | M_i)$ and $\Pr(M_i)$. Hence, employing Eq. (2.156) for the joint density of waiting times conditioned on the number of customers served in a busy period, expressing the densities in terms of the parameter $\rho = \lambda x$, and applying (2.222) we obtain

$$P(Y' | \rho) = \frac{\rho^{\left(\sum\limits_{i=1}^{N} M_i - N\right)}}{x^{\left(\sum\limits_{i=1}^{N} M_i - N\right)}} e^{-\left(\sum\limits_{i=1}^{N} M_i\right)\rho} . \tag{2.223}$$

$$0 \leq \omega_2^{(i)} \leq x \qquad i = 1 \ldots N$$

$$0 \leq \omega_{k+1}^{(i)} \leq x + \omega_k^{(i)} \qquad k = 2 \ldots M_i - 1$$

Letting $\gamma$ denote the delay gradient, $\gamma$ and $\rho$ may be related by (2.112). Employing (2.113), we can derive $\frac{\partial^2 \ell n P(Y' | \gamma)}{\partial \gamma^2}$. Using the expectation in (2.114) and the general formulation given in (2.102) for a Cramer-Rao bound on the variance of any unbiased estimator of a parameter A based on an observation vector R, we find

$$\mathrm{Var}\ (\gamma - \hat{\gamma}) \geq x^2 \frac{\rho}{N(1 - \rho)^5} . \tag{2.224}$$

We can shed light on the reason for the $\rho$ in the numerator of the bound for the customer-removal algorithm as compared to the $\rho^2$ in the numerator of the bound for the customer-addition algorithm by examining the maximum likelihood estimation procedures that follow from the observation vectors Y and Y' employed by each technique. The reason for this is that maximum likelihood procedures that are asymptotically unbiased will asymptotically achieve the Cramer-Rao bound. Maximum likelihood estimators for $\rho$ are defined by the following equations:

$$\frac{\partial}{\partial\rho} \ln P(Y|\rho) \Big|_{\rho=\hat{\rho}} = 0 \qquad (2.225)$$

$$\frac{\partial}{\partial\rho} \ln P(Y'|\rho) \Big|_{\rho=\hat{\rho}} = 0 \qquad (2.226)$$

Since the delay gradient D' is expressed by (2.33) as a function of $\rho$, the maximum likelihood estimator for D' is given

$$\hat{D}' = x + \frac{\hat{\rho}x}{2(1-\hat{\rho})} + \frac{\hat{\rho}x}{2(1-\hat{\rho})^2} \, . \qquad (2.227)$$

The maximum likelihood estimator specified by (2.225) is

$$\hat{\rho} = \frac{\sum_{i=1}^{N} M_i - 1}{\sum_{i=1}^{N} M_i + \sum_{j=1}^{N-1} I_j \backslash x} . \qquad (2.228)$$

The estimator that follows from the observation vector Y' employed by the customer-removal algorithm is given by (2.226) as

$$\hat{\rho} = 1 - \frac{N}{\sum_{i=1}^{N} M_i} . \qquad (2.229)$$

Each estimator for $\rho$ can be shown to be asymptotically unbiased and together with (2.227) must determine an asymptotically unbiased estimator of the delay gradient.

Since the delay gradient estimator specified by using (2.228) in (2.227) uses idle period information in addition to the number of customers served in each busy period, we might expect that asymptotically, the variance of this estimator will be smaller than that for the estimator that follows from using (2.229) in (2.228). Hence, it is not surprising that the Cramer-Rao bound for the customer-addition algorithm is always smaller than the corresponding bound for the customer-removal algorithm.

## 2.8 Derivation and Realization in Flow Diagram Form of Time-Contraction Algorithm

In our third algorithm, we simulate an increase in rate $\delta\lambda$ by a linear contraction in time scale. Assume that $\tau_j^{(i)}$ denotes the time of arrival of the j-th customer in the i-th busy period relative to the beginning of the observation interval. We define a new set of shifted arrival times by

$$\tau_j^{*(i)} = (1 - \frac{\delta\lambda}{\lambda})\tau_j^{(i)} \; . \tag{2.230}$$

Since $\delta\lambda$ represents an infinitesimal change in rate, we can choose it sufficiently small so none of the busy periods are shifted onto other busy periods by the time contraction. A simple sufficient condition on $\delta\lambda$ so that the redefinition of arrival times in (2.230) produces no "interactions" between busy periods is given by

$$\frac{\delta\lambda}{\lambda} T_E < \min_j \{I_j\} \tag{2.231}$$

$T_E$ refers to the observation period length and $I_j$ denotes the duration of the j-th idle period.

We now consider the increment in system time that comes from each customer arriving a little earlier and hence, waiting a little longer. The waiting time of the j-th customer in the i-th busy period is defined by

$$\omega_j^{(i)} = \sum_{\ell=1}^{j-1} x_\ell^{(i)} - (\tau_j^{(i)} - \tau_1^{(i)}) \ . \qquad (2.232)$$

If we substitute the shifted arrival times given by (2.230) into (2.232), we can relate the new waiting times $\omega_j^{*(i)}$ to $\omega_j^{(i)}$ by

$$\omega_j^{*(i)} = \omega_j^{(i)} + \frac{\delta\lambda}{\lambda} [\tau_j^{(i)} - \tau_1^{(i)}] \ . \qquad (2.233)$$

We define a new variable $\tau_j^{R(i)}$ as the arrival time of the j-th customer in the i-th busy period relative to the start of that busy period. Hence, the additional system time over N busy periods that follows from Eq. (2.232) can be expressed as

$$\frac{\delta\lambda}{\lambda} \sum_{i=1}^{N} Z_i \ , \qquad (2.234)$$

where

$$Z_i \triangleq \begin{cases} 0 & M_i = 1 \\ \sum_{n=2}^{M_i} \tau_n^{R(i)} & M_i \neq 1 \end{cases} \ . \qquad (2.235)$$

A second contribution to the increment in system time follows from the fact that our time contraction procedure shifts the right edge of the observation interval, leaving a gap of duration $\frac{\delta\lambda}{\lambda} T_E$ in which additional customers could arrive. Since an average of $\lambda D_c$ total delay is accumulated per unit time, where $D_c$ is the average total delay/customer, the average increment in system time is given by

$$\left(\frac{\delta\lambda}{\lambda} T_E\right)(\lambda D_c) = \delta\lambda T_E D_c \,. \tag{2.236}$$

Employing (2.234) and (2.236), we obtain the following expected increment in system time conditioned on the queueing record.

$$E(\delta S \mid \text{Queueing Record}) = \delta\lambda T_E D_c + \frac{\delta\lambda}{\lambda} \sum_{i=1}^{N} Z_i \tag{2.237}$$

Since both $\lambda$ and $D_c$ are unknowns, we use the fact that asymptotically, by the law of large numbers

$$\lambda \sim \frac{\sum_{i=1}^{N} M_i}{T_E} \tag{2.238}$$

$$D_c \sim \sum_{i=1}^{N} \sum_{j=1}^{M_i} S_j^{(i)} / \sum_{i=1}^{N} M_i \tag{2.239}$$

$S_j^{(i)}$ denotes the system time of the j-th customer in the i-th busy period. Substituting relations (2.238) and (2.239) into

(2.237) and normalizing by $\delta\lambda T_E$ we obtain the following delay gradient estimator:

$$\hat{D}' = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} S_j^{(i)}}{\sum_{i=1}^{N} M_i} + \frac{\sum_{i=1}^{N} Z_i}{\sum_{i=1}^{N} M_i} \qquad (2.240)$$

Comparing Eq. (2.128) and Eq. (2.240), we see that the customer-removal and time-contraction estimators have an identical first term corresponding to an estimate of $D_c$. They differ only in the second term where the customer-removal algorithm employs $p_i$ defined by (2.129) and the time contraction algorithm uses $Z_i$ specified by (2.235). Hence, flow diagrams for both algorithms are very similar. The variables that appear in the flow chart are the same as those for the customer-removal algorithm defined in Fig. 2.5 except that we no longer need the $\xi_i$'s and S is redefined as the sum of all the system times and relative arrival times of customers in the current busy period. Dropping the indices for the busy period, S is given by

$$S = \sum_{j=1}^{M} (x_j + \omega_j) + \sum_{j=2}^{M} \tau_j^R . \qquad (2.241)$$

Noting that $\omega_1$=0 and expressing $\omega_j$ as the sum of the j-1 preceding service times minus the arrival time $\tau_j^R$ of the j-th customer relative to the start of the busy period, we obtain the following expression for S since the contribution due to the relative arrival times is canelled:

$$S = \sum_{i=1}^{M} x_i + \sum_{i=2}^{M} \sum_{\ell=1}^{i-1} x_\ell \ . \tag{2.242}$$

We use (2.242) in the flow diagram for the time-contraction algorithm presented in Fig. 2.7.

The expression for S given in (2.242) suggests an alternative form for the estimator in (2.240). Changing the order of summation in the double sum in (2.242) and realizing that S is the contribution to the numerator of (2.240) corresponding to a given busy period, we obtain

$$\hat{D}' = \frac{\sum_{j=1}^{N} B_j}{\sum_{j=1}^{N} M_j} , \tag{2.243}$$

where

$$B_j \triangleq \sum_{\ell=1}^{M_j} (M_j - \ell + 1) x_\ell^{(j)} \tag{2.244}$$

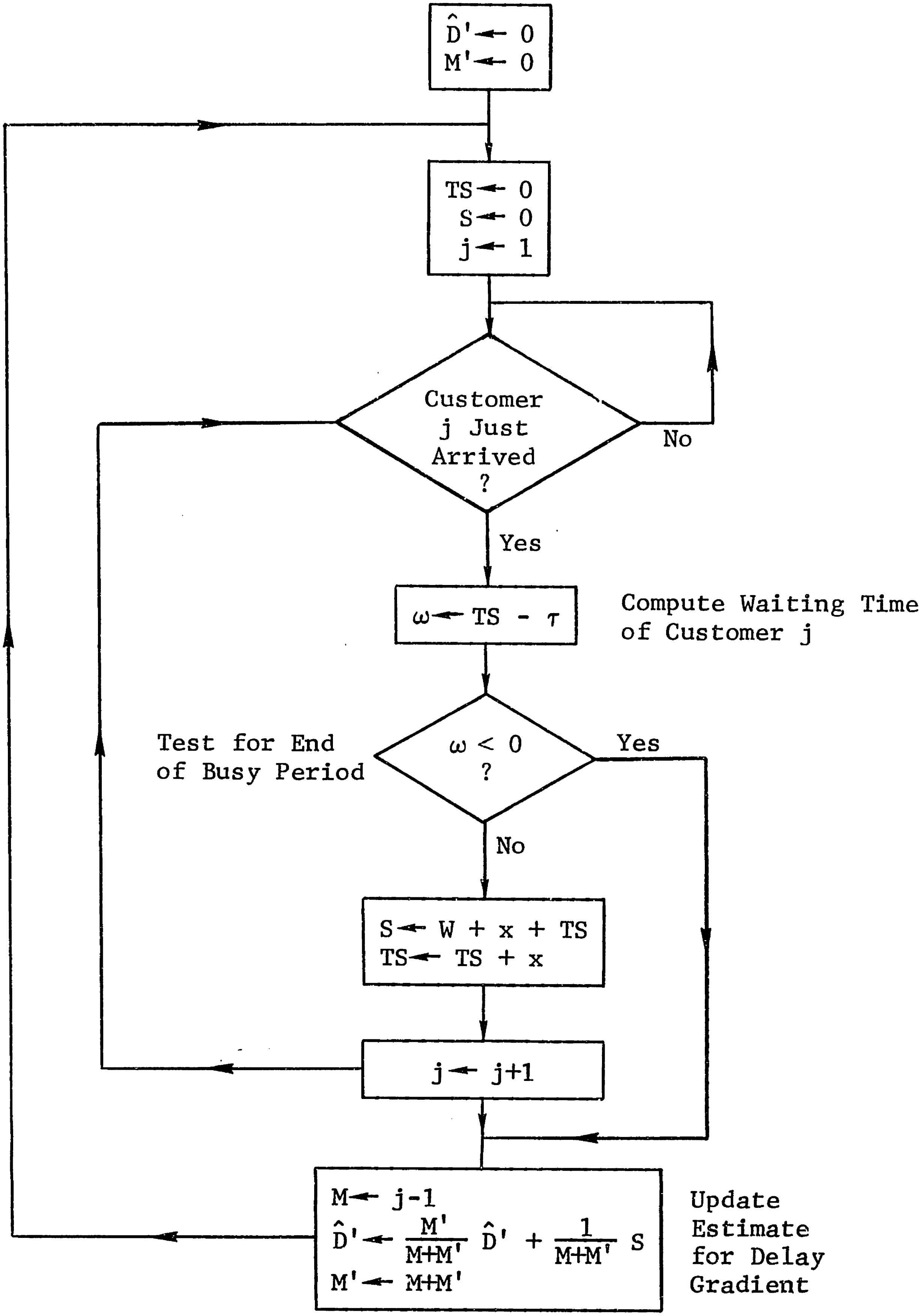

$\hat{D}' \leftarrow 0$
$M' \leftarrow 0$
$TS \leftarrow 0$
$S \leftarrow 0$
$j \leftarrow 1$
Customer j Just Arrived ?
No
Yes
$\omega \leftarrow TS - \tau$
Compute Waiting Time of Customer j
Test for End of Busy Period
$\omega < 0$ ?
Yes
No
$S \leftarrow W + x + TS$
$TS \leftarrow TS + x$
$j \leftarrow j+1$
$M \leftarrow j-1$
$\hat{D}' \leftarrow \frac{M'}{M+M'} \hat{D}' + \frac{1}{M+M'} S$
$M' \leftarrow M+M'$
Update Estimate for Delay Gradient

Figure 2.7 Flow Diagram for Time-Contraction Algorithm

## 2.9 Asymptotic Bias Calculation for Time-Contraction Algorithm in the Case of M/G/1 Queues

To investigate the asymptotic properties of the time-contraction algorithm we can employ the form for the estimator given in Eq. (2.240) or (2.243). We first examine Eq. (2.243), since the asymptotic unbiasedness of the algorithm for an M/D/1 queue follows readily. Exchanging the limit and expectation operations and appealing to the law of large numbers, we obtain

$$\lim_{N \to \infty} E\hat{D}' = E \lim_{N \to \infty} \hat{D}' = \frac{\overline{B}}{\overline{M}} . \qquad (2.245)$$

For an M/D/1 queue $x_i^{(j)} = x$ and we evaluate $\overline{B}$ by conditioning on M=j and then taking the expectation over M. Hence, $\overline{B}$ is given by

$$\overline{B} = E_M E(B \mid M{=}j) = \frac{1}{2} x \quad \overline{M^2} + \overline{M} . \qquad (2.246)$$

By Eq. (2.245), the mean of the estimator asymptotically approaches

$$\frac{1}{2} x \left( \frac{\overline{M^2}}{\overline{M}} + 1 \right) . \qquad (2.247)$$

Employing M/D/1 formulas for $\overline{M^2}$ and $\overline{M}$ given in Eqs. (2.98) and (2.99), we obtain the final result that

$$\lim_{N \to \infty} E\hat{D}' = x + \frac{\rho x}{2(1-\rho)} + \frac{\rho x}{2(1-\rho)^2} . \qquad (2.248)$$

The expression in (2.248) is identical to that in (2.33) for the delay gradient of an M/D/1 queue.

We can extend the argument of the preceding paragraph to an M/G/1 queue for calculation of the asymptotic form of the estimator as a power series in $\rho$. To evaluate the expectation of B conditioned on M we need $p(x_1 \ldots x_M|M)$ since

$$E(B|M) = E(Mx_1 + (M-1)x_2 + \ldots x_M|M) \qquad (2.249)$$

By Bayes' rule

$$p(x_1 \ldots x_M|M) = \frac{p(M|x_1 \ldots x_M)p(x_1 \ldots x_M)}{f_M} \qquad (2.250)$$

We calculate $p(M|x_1 \ldots x_M)$ as the probability of the region in $(\theta_1 \ldots \theta_M)$ space that corresponds to M customers with service requirements $(x_1 \ldots x_M)$ being in the busy period. By the constraints on the $\theta_j$'s specified in (2.148) and (2.149) for M customers to be served in a busy period and since the unconditional joint distribution of the $\theta_j$'s is a product of M exponential densities with parameter $\lambda$, $p(M|x_1 \ldots x_M)$ is given by

$$p(M|x_1 \ldots x_M) = \int_{\theta_1=0}^{x_1} \int_{\theta_2=0}^{x_1+x_2-\theta_1} \cdots \int_{\theta_k=0}^{\sum_{i=1}^{k} x_i - \sum_{i=1}^{k-1} \theta_i} \cdots \int_{\theta_M = \sum_{i=1}^{M} x_i - \sum_{i=1}^{M-1} \theta_i}^{\infty} \lambda^M e^{-\lambda \sum_{i=1}^{M} \theta_i} \, d\theta_M \ldots d\theta_1 . \tag{2.251}$$

Doing the final integration over $\theta_M$ and combining with (2.250) we obtain

$$p(x_1 \ldots x_M|M) = \frac{\lambda^{M-1}}{f_M} g(x_1 \ldots x_{M-1}) \prod_{i=1}^{M} B(x_i) e^{-\lambda x_i}, \tag{2.252}$$

where

$$g(x_1 \ldots x_{M-1}) = \begin{cases} 1 & M=1 \\ \displaystyle\int_{\theta_1=0}^{x_1} \int_{\theta_2=0}^{x_1+x_2-\theta_1} \cdots \int_{\theta_{M-1}=0}^{\sum_{i=1}^{M-1} x_i - \sum_{i=1}^{M-2} \theta_i} d\theta_{M-1} \ldots d\theta_1 & M \neq 1 \end{cases} \tag{2.253}$$

Hence, we can calculate $\overline{B}$ as

$$\overline{B} = \sum_{M=1}^{\infty} (\overline{B}|M) f_M \tag{2.254}$$

From the $\lambda^{M-1}/f_M$ factor in (2.252), we can see that $(\overline{B}|M)f_M$ will be representable as a power series in $\rho$ with terms of at least (M-1)-st order. Forming $\overline{B}|\overline{M}$ by multiplying $\overline{B}$ with $1-\rho$, we can collect the powers of $\rho$ in the asymptotic form of the estimator.

We have demonstrated that contrary to what happens for an M/D/1 queue, for the more general case of an M/G/1 queue, analysis of the expression given by (2.243) for the time-contraction estimator does not yield an explicit closed form expression for the asymptotic mean of the estimator. Hence, rather than pursuing the calculation outlined by Eqs. (2.245) through (2.254), we examine the form of the estimator in (2.240). Comparing this with (2.128), we note that $Z_i$ plays the same role as $p_i$ in the customer-removal algorithm. Therefore, the asymptotic bias of the time-contraction algorithm for an M/G/1 queue is specified by Eq. (2.134) where $\alpha_j$ is defined as the coefficient of $\rho^{j+1}$ in a power series ezpansion of $\overline{Z}$. $\overline{Z}$ is expressed using (2.235) as

$$\overline{Z} = \sum_{M=2}^{\infty} E\left(\sum_{i=2}^{M} \tau_i^R|M\right) f_M \ . \tag{2.255}$$

In the next few paragraphs we develop results useful in calculating $E\left(\sum_{i=2}^{M} \tau_i^R | M\right)$. Since $\tau_i^R$ is equal to the sum of the first (i-1) inter-arrival times, the problem of computing $E\left(\sum_{i=2}^{M} \tau_i^R | M\right)$ can be restated as

$$E\left(\sum_{i=2}^{M} \tau_i^R | M\right) = E\left(\sum_{j=1}^{M-1} \theta_j (M-j) | M\right) . \tag{2.256}$$

To evaluate the second expectation we need $p(\theta_1 \ldots \theta_{M-1} | M)$. Integrating out $\theta_M$ and $x_M$ from $p(x_1 \ldots x_M, \theta_1 \ldots \theta_M | M)$ defined in Eq. (2.150), we obtain

$$p(x_1 \ldots x_{M-1}, \theta_1 \ldots \theta_{M-1} | M) = \frac{\lambda^{M-1} B^*(\lambda)}{f_M} \prod_{i=1}^{M} [B(x_i) e^{-\lambda x_i}] .$$

$$k = 1 \ldots M-1 \qquad 0 \le \theta_k \le \sum_{j=1}^{k} x_j - \sum_{j=1}^{k-1} \theta_j \tag{2.257}$$

Employing the inequalities that the $\theta_j$'s and $x_j$'s must satisfy, we integrate out $(x_1 \ldots x_{M-1})$ and find

$$p(\theta_1 \ldots \theta_{M-1} | M) = \frac{\lambda^{M-1} B^*(\lambda)}{f_M} \int_{x_1=\theta_1}^{\infty} \int_{x_2=\varkappa_2}^{\infty} \cdots \int_{x_{M-1}=\varkappa_{M-1}}^{\infty} \prod_{i=1}^{M-1} [B(x_i) e^{-\lambda x_i}] \, dx_{M-1} \ldots dx_1 , \tag{2.258}$$

where

$$\aleph_k \triangleq \max \left\{ \sum_{i=1}^{k} \theta_i - \sum_{i=1}^{k-1} x_i , 0 \right\} .$$

By considering a series of linear transformations of $(\theta_1 \ldots \theta_{M-1})$ we can derive the statistics of the sum of relative arrival times conditioned on M and hence express the conditional mean as a single integral. We introduce the following two transformations:

$$\begin{pmatrix} \tau_2^R \\ \vdots \\ \tau_M^R \end{pmatrix} = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & & \ddots & \\ & & & 1 \\ 1 & & & & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{M-1} \end{pmatrix} \tag{2.259}$$

$$\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_{M-1} \end{pmatrix} = \begin{pmatrix} 1 & & \\ \vdots & & \\ 1 & & 1 \end{pmatrix} \begin{pmatrix} \tau_2^R \\ \vdots \\ \tau_M^R \end{pmatrix} \tag{2.260}$$

$\eta_{M-1}$ is the variable that corresponds to the sum of the relative arrival times. Employing (2.259) and (2.260), we obtain the following composite transformations from $(\theta_1 \ldots \theta_{M-1})$ to $(\eta_1 \ldots \eta_{M-1})$ and the corresponding inverse transformation.

$$\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_{M-1} \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ 2 & 1 & & & \\ 3 & 2 & 1 & & \\ & & & \ddots & \\ M-1 & M-2 & & & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{M-1} \end{pmatrix} \tag{2.261}$$

$$\begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{M-1} \end{pmatrix} = \begin{pmatrix} 1 & & & & & \\ -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_{M-1} \end{pmatrix} \tag{2.262}$$

The absolute value of the determinant of the Jacobian of the inverse transformation specified by (2.262) is one. Hence, we compute $p(\eta_1 \ldots \eta_{M-1}|M)$ by making the substitutions in $p(\theta_1 \ldots \theta_{M-1}|M)$ given by (2.258) for the $\theta_j$'s in terms of the $\eta_j$'s.

$$\begin{cases} \theta_1 = \eta_1 \\ \theta_2 = \eta_2 - 2\eta_1 \\ \theta_j = \eta_j - 2\eta_{j-1} + \eta_{j-2} \qquad j = 3 \ldots M-1 \end{cases} \tag{2.263}$$

The constraints on the $\eta_j$'s are obtained by guaranteeing $\theta_j \geq 0$.

$$\eta_1 \geq 0$$

$$\eta_2 \geq 2\eta_1 \qquad (2.264)$$

$$\eta_j \geq 2\eta_{j-1} - \eta_{j-2} \qquad j = 3 \ldots M-1$$

Hence, the desired conditional mean may be expressed as

$$E\left\{ \sum_{i=2}^{M} \tau_i^R | M \right\} = \int_{\eta_1=0}^{\infty} \int_{\eta_2=2\eta_1}^{\infty} \cdots \int_{\eta_j=2\eta_{j-1}-\eta_{j-2}}^{\infty} \cdots \int_{\eta_{M-1}=2\eta_{M-2}-\eta_{M-3}}^{\infty}$$

$$\eta_{M-1} p(\eta_1 \cdots \eta_{M-1} | M) d\eta_{M-1} \cdots d\eta_1 \qquad (2.265)$$

M/M/1 Bias Calculation

Now we calculate the first and second order terms in $\rho$ for the asymptotic bias $b_{T.C.}$ of the time-contraction algorithm in the case of an M/M/1 queue. Similar to our analysis of the customer-removal algorithm, we need to calculate $\alpha_1$ and $\alpha_0$, the coefficients of $\rho$ and $\rho^2$ in an expansion of $\overline{z}$ defined by (2.255). From the $\frac{\lambda^{M-1}}{f_M}$ term in $p(\theta_1 \cdots \theta_{M-1} | M)$, we see that $E\left\{ \sum_{j=2}^{M} (\tau_j^R | M) f_M \right\}$ will be representable as a power series in $\rho$ with terms of at least (M-1)'st order. Hence, to compute $\alpha_j$ we need to calculate $E \sum_{i=2}^{M} (\tau_i^R | M) f_M$ for $M = 2 \ldots j+2$, and collect coefficients of $\rho^{j+1}$.

We start by calculating $\overline{\tau}_2^R|2$. From (2.258), $p(\theta_1|M=2)$ is given by

$$p(\theta_1|M=2) = \frac{\lambda B^*(\lambda)}{f_2} \int_{x_1=\theta_1}^{\infty} B(x_1)e^{-\lambda x_1}dx_1 \ . \qquad (2.266)$$

Since $\tau_2^R = \theta_1$, we can express the desired conditional mean $\overline{\tau}_2^R|2$ as a double integral using (2.266). Changing the order of integration yields

$$\overline{\tau}_2^R|2 = \frac{1}{2} \ \frac{\lambda B^*(\lambda)}{f_2} \int_{x_1=0}^{\infty} x_1^2 B(x_1)e^{-\lambda x_1}dx_1 \ . \qquad (2.267)$$

Using the B(x) and B*(S) for an M/M/1 queue specified by (2.196), we find

$$\overline{\tau}_2^R|2 = \frac{1}{2} \ \frac{1}{\mu} \ \frac{\rho}{(1+\rho)^4} \ \frac{1}{f_2} \ . \qquad (2.268)$$

Next we evaluate $E\left(\sum_{i=1}^{M} \tau_i^R|M\right)$ for M=3. $p(\theta_1,\theta_2|M=3)$ is given by (2.258) as

$$p(\theta_1,\theta_2|M=3) = \frac{\lambda^2 B^*(\lambda)}{f_3} \int_{x_1=\theta_1}^{\infty} \int_{x_2=\max\{\theta_1+\theta_2-x_1,0\}}^{\infty} (B(x_1)e^{-\lambda x_1})(B(x_2)e^{-\lambda x_2})\, dx_2 dx_1 \ . \qquad (2.269)$$

Breaking up the integration over $x_1$ to get rid of the "max" in the lower limit of the $x_2$ integration and making the substitutions defined in (2.263) for $\theta_1$ and $\theta_2$, we derive $p(\eta_1,\eta_2|M=3)$ as

$$p(\eta_1,\eta_2|M=3) = \frac{\lambda^2 B^*(\lambda)}{f_3}\left\{ \int_{x_1=\eta_1}^{\eta_2-\eta_1} \int_{x_2=\eta_2-\eta_1-x_1}^{\infty} (B(x_1)e^{-\lambda x_1})(B(x_2)e^{-\lambda x_2}) \, dx_2dx_1 \right.$$

$$\left. + \int_{x_1=\eta_2-\eta_1}^{\infty} \int_{x_2=0}^{\infty} (B(x_1)e^{-\lambda x_1})(B(x_2)e^{-\lambda x_2})dx_2dx_1 \right\} \tag{2.270}$$

By (2.265) the desired conditional mean is formulated as $\overline{\eta}_2|M=3$.

$$E\left\{\sum_{i=2}^{3} \tau_i^R |M=3\right\} = \frac{\lambda^2 B^*(\lambda)}{f_3}\left( \int_{\eta_1=0}^{\infty} \int_{\eta_2=2\eta_1}^{\infty} \eta_2 \int_{x_1=\eta_1}^{\eta_2-\eta_1} \int_{x_2=\eta_2-\eta_1-x}^{\infty} (B(x_1)e^{-\lambda x_1})(B(x_2)e^{-\lambda x_2}) \; dx_2dx_1d\eta_2d\eta_1 \right.$$

$$\left. + \int_{\eta_1=0}^{\infty} \int_{\eta_2=2\eta_1}^{\infty} \eta_2 \int_{x_1=\eta_2-\eta_1}^{\infty} \int_{x_2=0}^{\infty} (B(x_1)e^{-\lambda x_1})\left(B(x_2)e^{-\lambda x_2}\right)dx_2dx_1d\eta_1d\eta_2 \right\} \tag{2.271}$$

Employing B(x) and B*(S) for an M/M/1 queue, we obtain

$$E\left(\sum_{i=2}^{3} \tau_i^R | M=3\right) = 7 \frac{1}{\mu} \frac{\rho^2}{(1+\rho)^6} \frac{1}{f_3} . \tag{2.272}$$

Using Eqs. (2.268) and (2.272), we can calculate the coefficients of the $\rho$ and $\rho^2$ terms in the asymptotic bias $b_{T.C.}$. $\alpha_0$ is derived by taking the coefficient of $\rho$ in an expansion of $(\overline{\tau}_2^R|2) f_2$. $\alpha_1$ is obtained by collecting powers of $\rho^2$ in $(\overline{\tau}_2^R|2) f_2$ and $(\overline{\tau}_2^R + \overline{\tau}_3^R|3) f_3$. Hence, we find

$$\begin{aligned} \alpha_0 &= \frac{1}{2} \frac{1}{\mu} \\ \alpha_1 &= 5 \frac{1}{\mu} \end{aligned} \tag{2.273}$$

Employing Eq. (2.137), we specify $b_{T.C.}$ as

$$b_{T.C.} = - \frac{1}{2} \frac{1}{\mu} \rho + \frac{5}{2} \frac{1}{\mu} \rho^2 + O(\rho^3) . \tag{2.274}$$

Hence, the asymptotic bias of the time-contraction algorithm for an M/M/1 queue contains all powers of $\rho$. This contrasts with the case of an M/D/1 queue, where we were able to prove asymptotic unbiasedness of the algorithm.

## 2.10 Cramer-Rao Bound for Time-Contraction Algorithm in the Case of M/D/1 Queues

We now conclude our analysis of the time-contraction algorithm by deriving a Cramer-Rao bound on the variance of the estimator for an M/D/1 queue. The form of the time contraction estimator defined in (2.243) and (2.244) tells us that the observations which the algorithm uses, in the case of an M/D/1 queue, consist of the number of customers served in each busy period. Since the number of customers served is independent from one busy period to another and for an M/D/1 queue the probability of M customers in a busy period is given by Eq. (2.110), $p(M_1 \dots M_N|\rho)$ is as follows:

$$p(M_1 ,\dots M_N|\rho) = \prod_{i=1}^{N} \frac{(M_i\rho)^{M_i-1}}{M_i!} e^{-M_i\rho} . \qquad (2.275)$$

Letting $\gamma$ denote the delay gradient, there is a 1-1 correspondence between $\gamma$ and $\rho$ specified by Eq. (2.112). Employing (2.113) and (2.112), as we did before for the customer-addition and customer-removal algorithms, we can derive a Cramer-Rao bound for the variance of any unbiased estimator of $\gamma$ that employs observations $(M_1 \dots M_N)$.

$$\text{Var } (\gamma - \hat{\gamma}) \geq x^2 \frac{\rho}{N(1-\rho)^5} \qquad (2.276)$$

The bound (2.276) is identical to that for the customer-removal algorithm. This is to be expected since the maximum-likelihood estimate for $\rho$ that follows from (2.275) is the same as that which follows from $P(Y'|\rho)$ defined in (2.223), where Y' denotes the set of observations which the customer-removal algorithm uses in the case of an M/D/1 queue.

## 2.11 Computational Complexity and Storage Requirement Analysis for the Three Algorithms

Now we proceed to analyze, compare, and contrast the computational complexity and storage requirements of the three estimation algorithms presented. The procedure for up-dating the delay-gradient estimate in all three algorithms assumes the following form:

$$\hat{D}'_{(k+1)} = \zeta \hat{D}'_{(k)} + \Delta \ . \qquad (2.277)$$

$\hat{D}'_{(k)}$ is the delay gradient estimate based on an observation interval containing k busy periods, $\zeta$ is a factor that renormalizes $\hat{D}'_{(k)}$ to correspond to a component of the (k+1) busy period estimate. $\Delta$ denotes the contribution to the (k+1) busy period estimate that comes from the (k+1)-st busy period. Hence, to compare the three algorithms, we examine the storage and computational requirements of forming $\Delta$.

In the customer-addition algorithm, $\Delta$ is the expected additional delay suffered by the customers of the (k+1)-st busy period due to an arrival in the entire (k+1) busy period record. According to the notation defined in Fig. 2.1, we need $(I_{n_I} \ldots I_{i_I})$ and $(T_{n_I+1} \ldots T_{i_I+1})$ to compute $\Delta$. $I_{n_I}$ denotes the first idle period at which an extra arrival with a service requirement x can influence the $(i_I+1)$-st or current busy period. Hence, x satisfies

$$x > \sum_{k=n_I+1}^{i_I} I_k \, . \qquad (2.278)$$

We need two buffers with $i_I - n_I+1$ storage locations for the idle period and busy period durations, respectively. We can generalize this argument to say that the probability of x being greater than the sum of $N_b$ idle period durations is equal to the probabil ity that we need two buffers with $N_b+1$ locations. Hence, for the case of an M/G/1 queueing system, the probability of x being greater than the sum of $N_b$ independent exponential variates with parameter $\lambda$ is equal to the probability of a buffer overflow give we have only $N_b$ storage locations per buffer. In designing our system, we want to choose $N_b$ sufficiently large to make the probability of an overflow less than some acceptable probability $\varepsilon$. In Eqs. (2.44) through (2.46), we derive the statistics of the sum of k exponential random variables with parameter $\lambda$.

Employing (2.46) and letting $\tilde{\rho} \triangleq \lambda x$ we obtain the following expression for the probability of an overflow given our two buffers have $N_b$ locations:

$$\Pr\ \{\text{Overflow}\,|\,N_b \text{ Storage Location}\,|\,\text{Buffer}\} = 1 - e^{-\tilde{\rho}} \sum_{j=0}^{N_b} \frac{\tilde{\rho}^j}{j!}\ . \tag{2.279}$$

By applying Taylor's remainder theorem to an expansion of $e^{\tilde{\rho}}$ we obtain

$$e^{\tilde{\rho}} \le \sum_{j=0}^{N_b-1} \frac{\tilde{\rho}^j}{j!} + \frac{e^{\tilde{\rho}}\tilde{\rho}^{N_b}}{N_b!}\ . \tag{2.280}$$

By employing (2.280) we can upper bound the probability in (2.279).

$$\Pr\ \{\text{Overflow}\,|\,N_b \text{ Location}\,|\,\text{Buffer}\} \le \frac{\tilde{\rho}^{N_b}}{N_b!} \tag{2.281}$$

Since $\tilde{\rho} = \lambda x$ and the range of interest of $\lambda$ is $0 \le \lambda \le \frac{1}{x}$, we obtain the final bound

$$\Pr\ \{\text{Overflow}\,|\,N_b \text{ Location}\,|\,\text{Buffer}\} \le \frac{(x|\bar{x})^{N_b}}{N_b!}\ . \tag{2.282}$$

We want to make the right side of the inequality in (2.282) less than some tolerable probability of overflow $\varepsilon$. As an example, for $\bar{x} = x$ and $\varepsilon = 10^{-9}$, $N_b \ge 13$ satisfies (2.282). Having arrived at a suitable value of $N_b$, the storage requirement of the customer-addition algorithm is approximately $2N_b + 12$.

We now consider the computational requirements of forming $\Delta$. For the customer-addition algorithm, $\Delta$ is composed of a contribution which represents the additional system time resulting from an arrival in the (k+1)-st busy period plus the effect on the (k+1)-st busy period of an arrival at any earlier time in the (k+1) busy period record. We consider first the contribution due to the effect of an extra arrival in the (k+1)-st busy period. This term assumes two forms depending on whether all the service requirements are identical. We employ the notation defined in Fig. 2.1 for expressing the contributions to D. If $x_n^{(q)} = x$, the term is given by

$$\frac{1}{\ell'} \left(\frac{1}{2} M^2 x^2\right) , \tag{2.283}$$

which requires four multiplications and one division to evaluate. For the more general case qhen $x_n^{(q)} \neq x$, the desired contribution is given by Eq. (2.14) normalized by $\ell'$. With some manipulation, the component of $\Delta$ due to an extra arrival in the (k+1)-st busy period is given by

$$\frac{1}{\ell'} \left\{ \sum_{k=1}^{M} \left[ \sum_{n=1}^{k} x_n - \frac{(\tau_{k+1} + \tau_k)}{2} \right] (\tau_{k+1} - \tau_k) + x \sum_{i=2}^{M} \tau_i^R \right\} . \tag{2.284}$$

The above takes M+1 multiplications, M+1 divisions, and 6M-4 additions to calculate. The $\tau_i^R$ denote arrival times relative to the start of the busy period. The expression in (2.284) is conceptualized as being evaluated as the busy period progresses. Hence, we needn't store all the service requirements and arrival times.

To enumerate the remaining calculations in forming $\Delta$ we count the operations involved in computing the effect of an extra arrival in the preceding $i_I - n_I+1$ idle periods and $i_I - n_I$ busy periods on the customers in the $(i_I+1)$-st busy period. If we let $i_I - n_I+1 \triangleq n_b$, then $n_b$ corresponds to the current number of idle period and busy period durations that we store. In terms of $n_b$, the remaining operations to compute $\Delta$ are counted as

$$\begin{array}{l} 7n_b - 6 \text{ additions} \\ 6n_b - 2 \text{ multiplications} \\ 2n_b - 1 \text{ divisions} \\ n_b \text{ comparisons} \end{array} \qquad (2.285)$$

We can upper bound the number of operations listed in (2.285) by letting $n_b = N_b$, where $N_b$ is our buffer size chosen to guarantee the probability of overflow being less than some threshold $\varepsilon$. Using our preceding example, we can let $n_b = 13$ for an $\varepsilon = 10^{-9}$.

For the general case when the service requirements are not all identical, the number of operations to evaluate (2.284) varies with M. Hence it is of interest to bound the probability of M exceeding some value t. For an M/G/1 queue, the simplest Chebyshev bound yields

$$\Pr\{M \geq t\} \leq \frac{\overline{M}}{t} = \frac{1}{(1-\rho)t} \ . \tag{2.286}$$

Hence, as $\rho \to 1$ we need to make t very large for the bound to give us any information. Therefore, the number of operations required in (2.284) becomes unbounded as $\rho \to 1$.

For the customer-removal algorithm we identify S defined in Fig. 2.5 as the corresponding $\Delta$. The number of storage locations required to calculate $\Delta$ is approximately M + 8. The M comes from the array of $\xi_i$'s in which we store $C_j^i$ as j is varied. The number of operations required is determined with reference to the flow chart in Fig. 2.6.

$$\begin{aligned} &\frac{M(M-1)}{2} \text{ Minimizations or Comparisons} \\ &\frac{M(M-1)}{2} + 4M-3 \text{ additions} \end{aligned} \tag{2.287}$$

In the time-contraction algorithm S corresponds to $\Delta$ as before. The storage requirement to compute $\Delta$ is 8. The number of operations obtained by examination of Fig. 2.7 is 3M-3 additions.

From the discussion above we see that of all three algorithms, the time-contraction algorithm has the smallest storage requirement. Unlike the customer-addition and customer-removal algorithms, the time-contraction procedure requires no buffer with randomly varying size. In addition, the computational load for the time-contraction algorithm is the least of the three. Hence, the time-contraction procedure is the least costly to implement.

# SECTION 3

## SIMULATION RESULTS

In Section 2 we derived the asymptotic bias of the three estimation algorithms. Since we are unable to calculate the variance of the estimators as a function of N, the question of consistency, whether the estimates converge asymptotically to the delay gradient, remains unanswered. In addition, for an M/D/1 queue we would like to know if our algorithms are asymptotically efficient, whether they achieve the Cramer-Rao bounds derived in (2.116), (2.224), and (2.276). We would also like to investigate the robustness of the customer-removal and time-contraction estimation schemes by seeing how they perform for a variety of queueing systems. We attempt to provide answers to these questions in the present section by presenting the results of simulating all three algorithms for an M/D/1 queue and simulating the time-contraction and customer-removal algorithm for M/M/1, D/M/1, and U/M/1 queues.

We simulate a single-server queue by the following recursion for successive waiting times:

$$\omega_{n+1} = \max\{\omega_n + x_n - \theta_n, 0\} \quad \text{with} \quad \omega_1 = 0 \, . \qquad (3.1)$$

$x_n$ and $\theta_n$ are random variables corresponding to the m-th service requirement and the inter-arrival time between the n-th and (n+1)-st customer, respectively. When $\omega_\ell$ goes to zero, this signals the start of a new busy period.

We now describe the calculations necessary to evaluate the statistics of a given estimator. To derive the mean and variance of a k-busy period estimator $\hat{D}'_{(k)}$ we generate a certain sample size $N_S$ of k-busy period records, processing each to form an estimate $\hat{D}'_{(k),i}$. We compute estimates of the bias $\hat{b}_{(k)}$ and variance $\hat{\sigma}^2_{(k)}$ associated with $\hat{D}'_{(k)}$ as follows:

$$\hat{b}_{(k)} = \frac{1}{N_S} \sum_{i=1}^{N_S} \hat{D}'_{(k),i} - \frac{\partial D}{\partial \lambda} \tag{3.2}$$

$$\hat{\sigma}^2_{(k)} = \frac{N_S}{N_S - 1} \left( \frac{1}{N_S} \sum_{i=1}^{N_S} D'^2_{(k),i} - \left( \frac{1}{N_S} \sum_{i=1}^{N_S} D'_{(k),i} \right)^2 \right) \tag{3.3}$$

The measure of performance that we use most often is the fractional rms error, which we approximate by employing our estimates for the bias and variance in (3.2) and (3.3).

$$\frac{\sqrt{E\left(D'_{(k)} - \frac{\partial D}{\partial \lambda}\right)^2}}{\frac{\partial D}{\partial \lambda}} \approx \frac{\sqrt{\hat{b}^2_{(k)} + \hat{\sigma}^2_{(k)}}}{\frac{\partial D}{\partial \lambda}} \qquad (3.4)$$

For all the queueing systems under examination, we present curves of the fractional rms error associated with $\hat{D}'_{(k)}$ for k = 10, 100, 1000 busy periods with $N_S$ = 400 Monte Carlo rms for each estimate. We experimented with $N_S$, trying $N_S$ = 10, 50, 100, 200, 400, and found that $N_S$ = 400 insured from one to two significant figures in our value for the fractional error defined in (3.4).

We do not present separate curves for our estimate of the bias $\hat{b}(k)$ defined in (3.2), since in many cases our value for $\hat{b}(k)$ was of comparable size to the statistical fluctuations in the quantity. To make this notion clearer, (3.2) can be rewritten as

$$\hat{b}_{(k)} = \left(E\hat{D}'_{(k)} - \frac{\partial D}{\partial \lambda}\right) + \delta . \qquad (3.5)$$

$\delta$ is a zero mean random variable with

$$E\delta^2 \approx \frac{\hat{\sigma}^2_{(k)}}{N_S} \qquad (3.6)$$

Hence, when our value for $\hat{b}(k)$ is of the same size as $\sqrt{E\delta^2}$, we can not possibly measure the actual bias $b_{(k)}$ with any accuracy. Careful examination of the results of the simulations support the conclusion that in the cases where we could not measure b(k) accurately, it contributed negligibly to the fractional rms error in (3.4) and when b(k) could be measured accurately, its contribution in (3.4) was non-negligible.

Having defined the measures we will use to compare the algorithms, we first discuss the simulation results for an M/D/1 queue. From earlier analytical results, we know the customer-addition and time-contraction algorithms are asymptotically unbiased for M/D/1 and that the customer-removal algorithm is unbiased at least up to the third power of $\rho$. Curves of the fractional rms error for N = 10, 100, 1000 are presented for all three algorithms in graphs 3.1, 3.2, and 3.3. The consistency of the three estimation procedures is suggested by the improved performance with increasing N. In graphs 3.4 and 3.5, for N = 100 and N = 1000, respectively, we present the lower bounds on fractional rms error that follow from our Cramer-Rao bounds for the variance of the estimators, together with the simulation results. The closeness of the simulation curves to their respective lower bounds suggest that all three algorithms are asymptotically efficient for an M/D/1 queue. Graphs 3.4 and 3.5 show that as a function of $\rho$ the fractional rms error for the customer-addition algorithm remains below that of the time-contraction and customer-removal algorithms, and hence we conclude that it performs the best of the three procedures for an M/D/1 queue. The customer-

Figure 3.1 Customer-Addition Algorithm - Fractional RMS Error for M/D/1

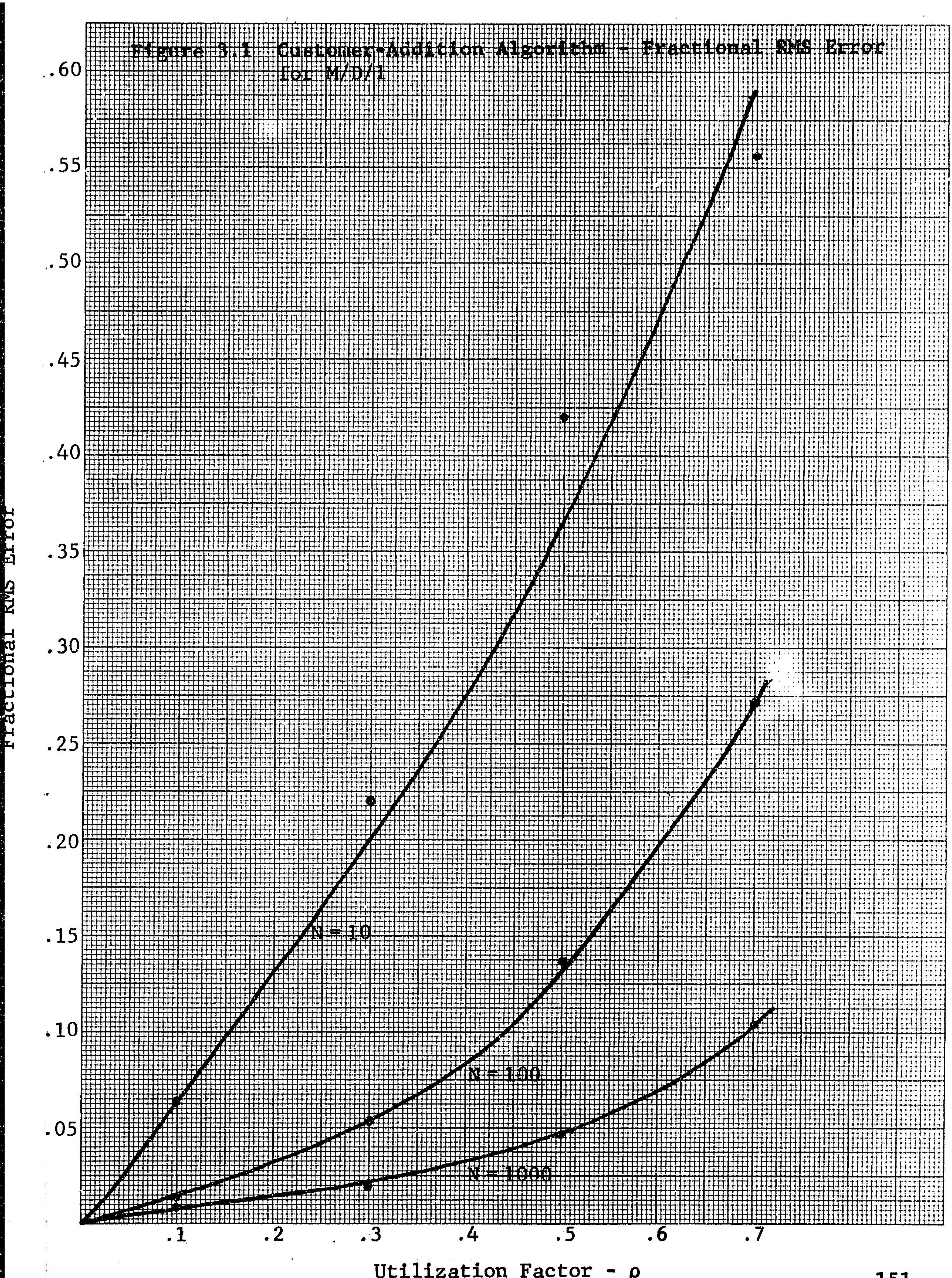


Figure 3.2 Time Contraction Algorithm - Fractional RMS Error for M/D/1

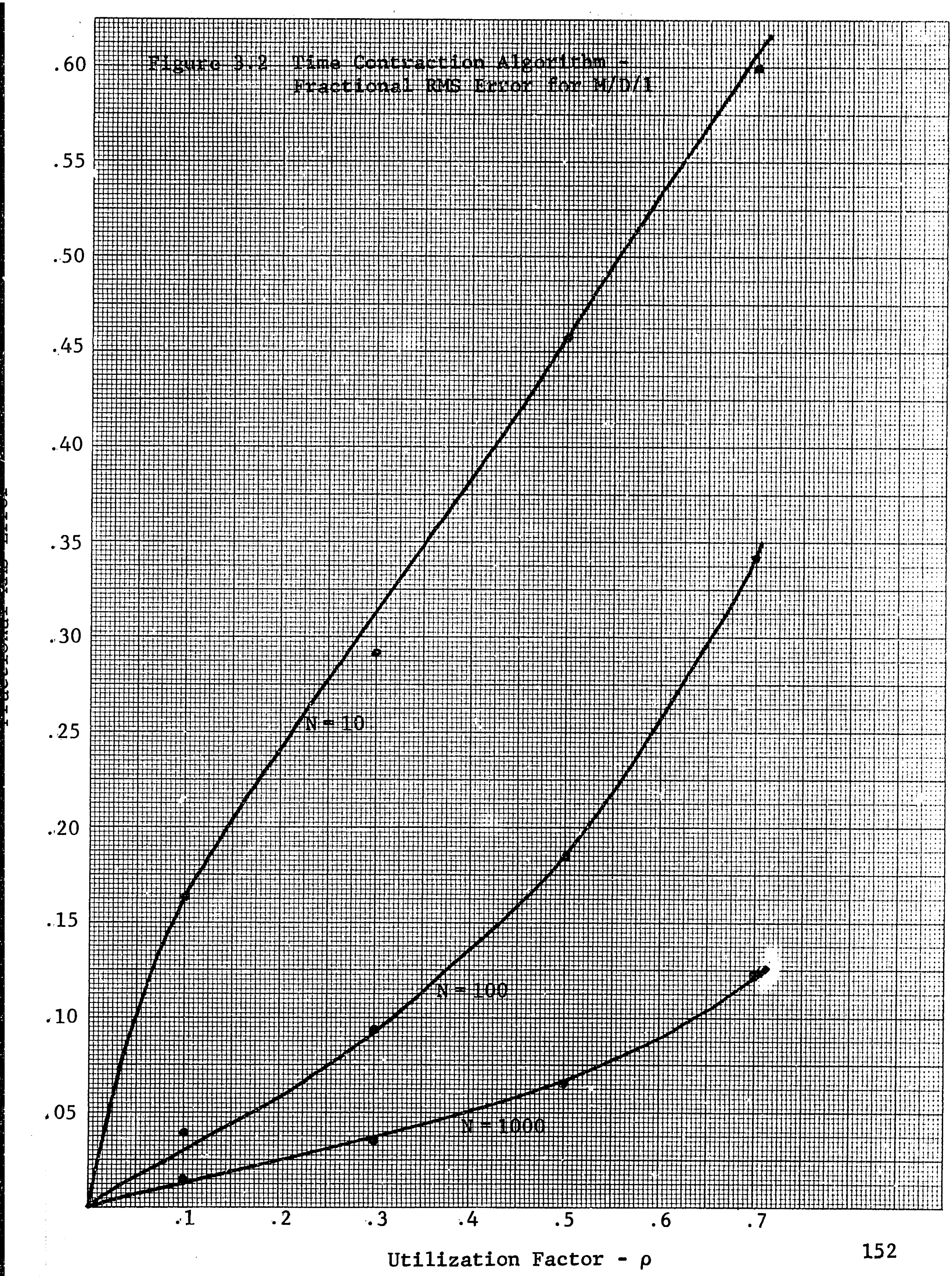


Figure 5.5 Customer-Removal Algorithm - Fractional RMS Error for M/D/1

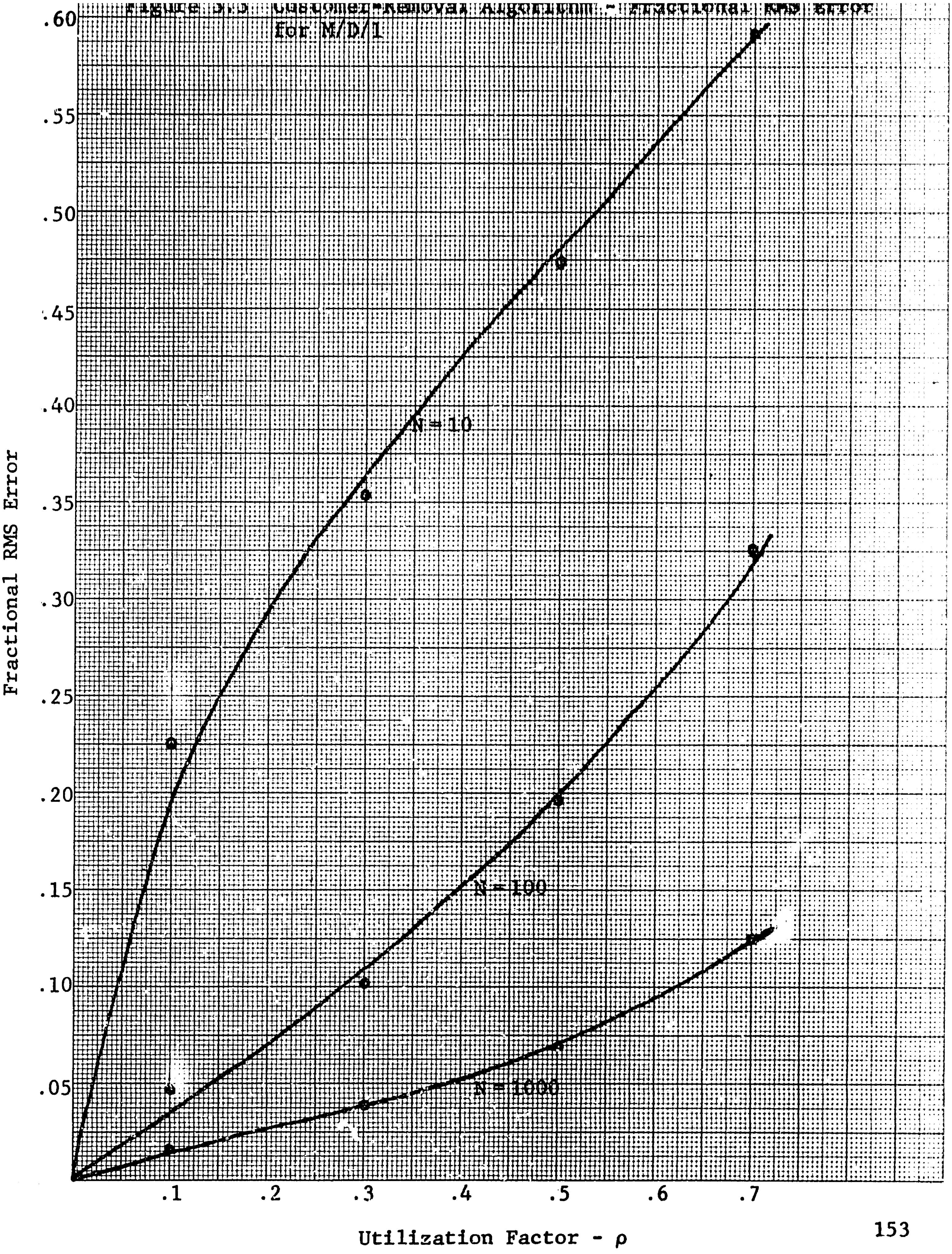
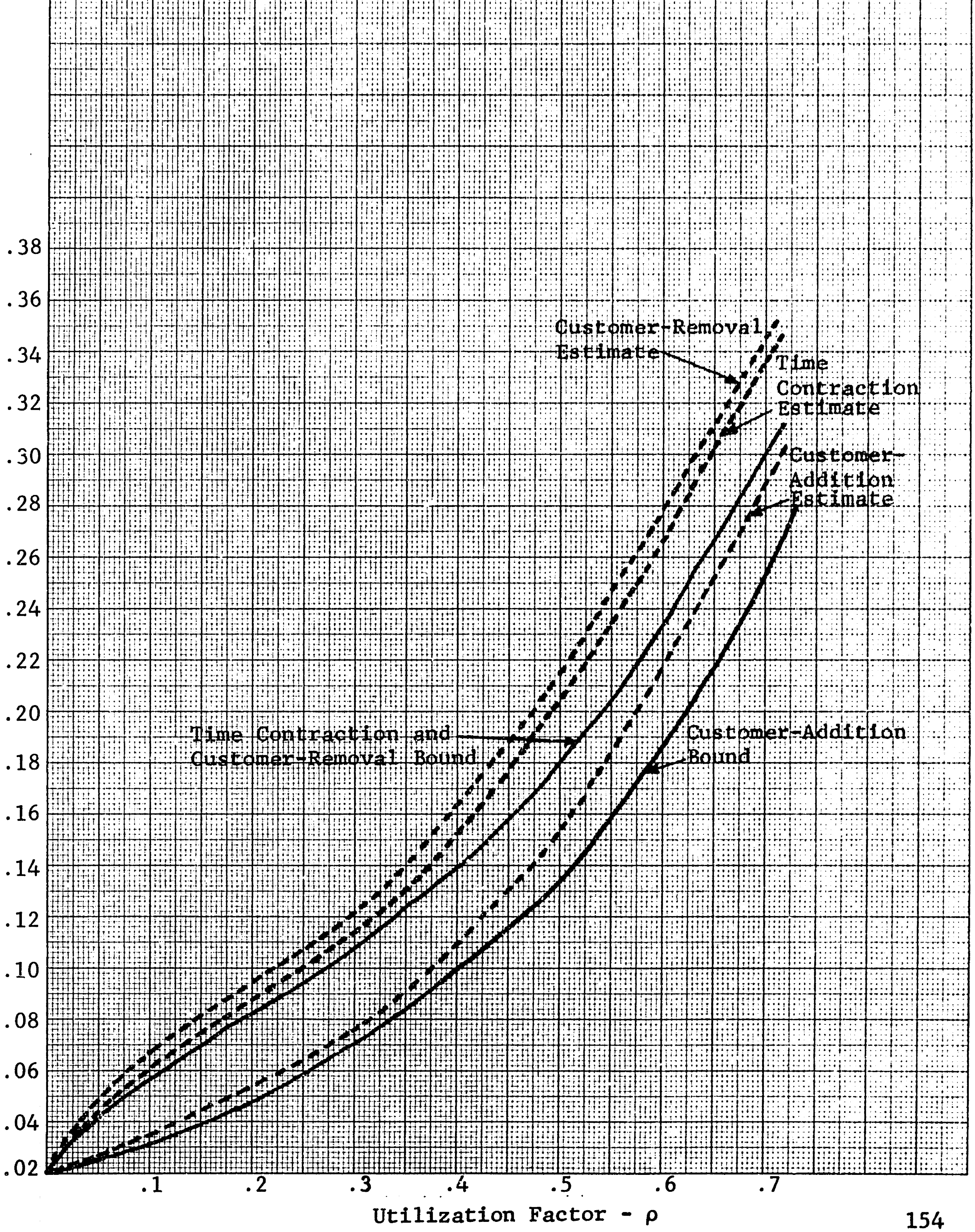


Figure 3.4 Lower Bounds on Fractional RMS Error for M/D/1 (N = 100)



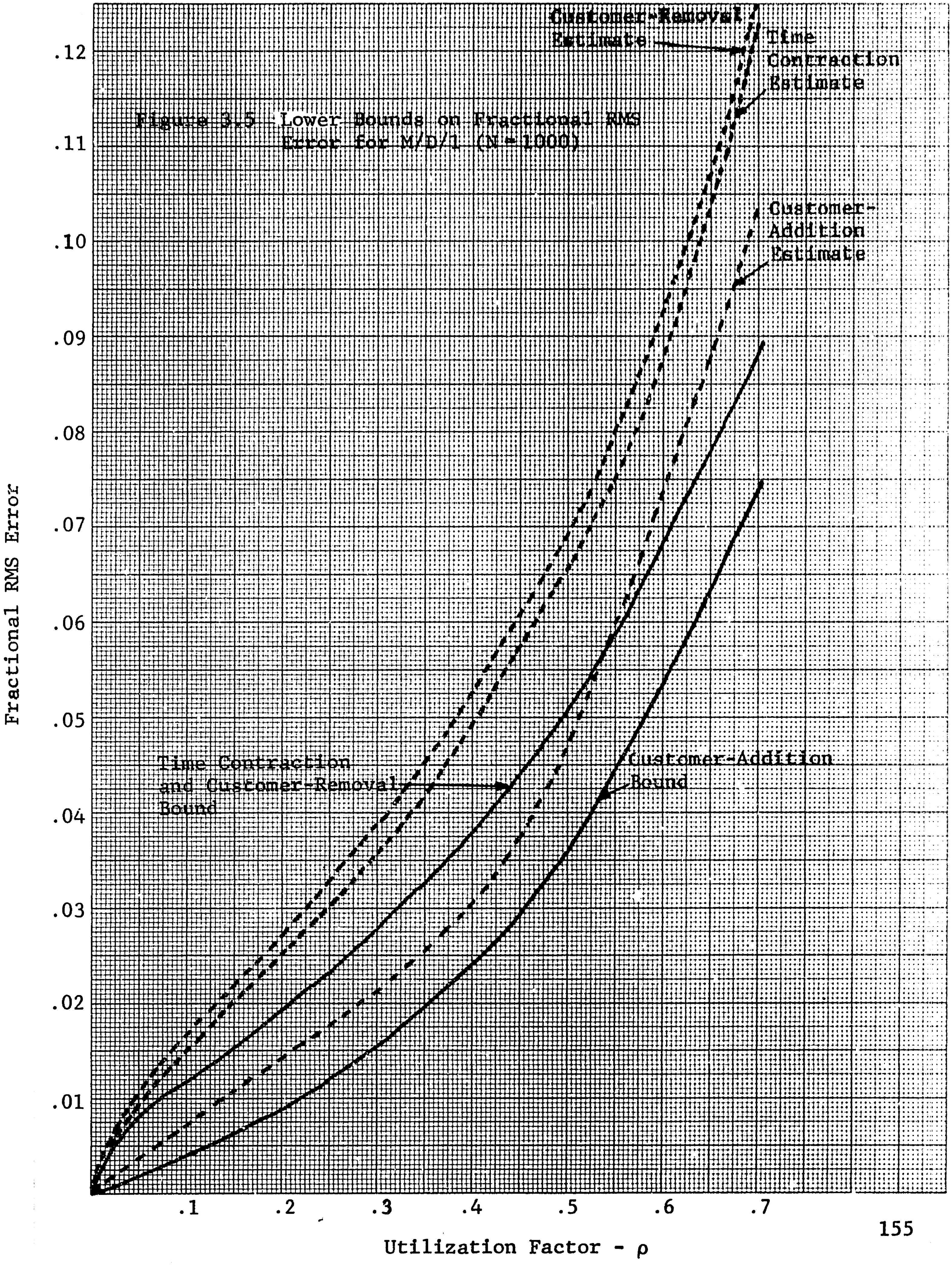


Figure 3.5 Lower Bounds on Fractional RMS Error for M/D/1 (N = 1000)

removal and time-contraction algorithm's fractional rms error curves are quite close, with the customer-removal algorithm's performance being slightly worse. The bias results for the three algorithms are displayed in Tables 3-1 and 3-2. We list our estimate for the relative bias $\hat{b}(k)/\frac{\delta D}{\delta\lambda}$ together with the size of statistical fluctuations in the relative bias, $\frac{\hat{\sigma}(k)/\sqrt{N_S}}{\frac{\delta D}{\delta\lambda}}$. Seemingly anomalous behavior of the relative bias is often explainable by looking at the size of the statistical fluctuations in the quantity.

We now proceed to demonstrate that the behavior of the algorithms for $\rho$ near zero indicated by graphs 3.1, 3.2, and 3.3 is reasonable. We examine all three algorithms for $\rho \to 0$ by keeping $\bar{x}$ constant and considering $\lambda \to 0$. As $\rho$ is made arbitrarily close to zero, the density for $M_i$, the number of customers served in the i-th busy period, becomes an impulse at 1. Hence, as $\rho \to 0$, the queueing record approaches a set of single customer busy periods. In addition, since $\bar{I}_k$, the average size of the k'th idle period is $\frac{1}{\lambda}$, the idle periods become unbounded as $\rho \to 0$. Hence, with high probability the service requirement of the additional customer x will be $\ll I_k$. Applying (2.14), (2.15), and (2.17) through (2.19), we find that with high probability $E(\delta S \mid t \epsilon T_k$, Queueing Record) and $E(\delta S \mid t \epsilon I_k$, Queueing Record) will be given as

$$E(\delta S \mid t \epsilon T_k, \text{ Queueing Record}) = x + \frac{1}{2} x \tag{3.7}$$

TABLE 3-1

RELATIVE BIAS FOR CUSTOMER-ADDITION ALGORITHM (M/D/1 QUEUE)

| $\rho$ | N | $\dfrac{\hat{b}}{\frac{\partial D}{\partial \lambda}}$ | $\dfrac{\hat{\sigma}_{\hat{D}'(N)}/\sqrt{N_S}}{\frac{\partial D}{\partial \lambda}}$ |
|---|---|---|---|
| .1 | 10 | $3.4 \times 10^{-2}$ | $2.7 \times 10^{-3}$ |
| | 100 | $1.6 \times 10^{-3}$ | $7.0 \times 10^{-4}$ |
| | 1000 | $6.5 \times 10^{-4}$ | $4.2 \times 10^{-4}$ |
| .3 | 10 | $7.2 \times 10^{-2}$ | $1.0 \times 10^{-2}$ |
| | 100 | $-1.4 \times 10^{-3}$ | $2.7 \times 10^{-3}$ |
| | 1000 | $-3.6 \times 10^{-4}$ | $9.9 \times 10^{-4}$ |
| .5 | 10 | $6.3 \times 10^{-2}$ | $2.1 \times 10^{-2}$ |
| | 100 | $-7.6 \times 10^{-3}$ | $6.8 \times 10^{-3}$ |
| | 1000 | $-1.8 \times 10^{-3}$ | $2.4 \times 10^{-3}$ |
| .7 | 10 | $-1.0 \times 10^{-1}$ | $2.7 \times 10^{-2}$ |
| | 100 | $-4.4 \times 10^{-2}$ | $1.3 \times 10^{-2}$ |
| | 1000 | $8.6 \times 10^{-4}$ | $5.2 \times 10^{-3}$ |

TABLE 3-2

RELATIVE BIAS FOR CUSTOMER-REMOVAL AND
TIME-CONTRACTION ALGORITHM (M/D/1 QUEUE)

| $\rho$ | N | $\dfrac{\hat{b}_{C.R.}}{\frac{\partial D}{\partial \lambda}}$ | $\dfrac{\hat{\sigma}_{\hat{D}'_{(N),C.R.}}/\sqrt{N_S}}{\frac{\partial D}{\partial \lambda}}$ | $\dfrac{\hat{b}_{T.C.}}{\frac{\partial D}{\partial \lambda}}$ | $\dfrac{\hat{\sigma}_{\hat{D}'_{(N),T.C.}}/\sqrt{N_S}}{\frac{\partial D}{\partial \lambda}}$ |
|---|---|---|---|---|---|
| .1 | 10 | $1.6 \times 10^{-1}$ | $8.7 \times 10^{-3}$ | $8.8 \times 10^{-2}$ | $6.8 \times 10^{-3}$ |
|  | 100 | $1.6 \times 10^{-2}$ | $2.2 \times 10^{-3}$ | $7.8 \times 10^{-3}$ | $1.9 \times 10^{-3}$ |
|  | 1000 | $4.7 \times 10^{-3}$ | $7.9 \times 10^{-4}$ | $2.8 \times 10^{-3}$ | $7.3 \times 10^{-4}$ |
| .3 | 10 | $1.2 \times 10^{-1}$ | $1.7 \times 10^{-2}$ | $5.4 \times 10^{-2}$ | $1.4 \times 10^{-2}$ |
|  | 100 | $6.3 \times 10^{-3}$ | $5.1 \times 10^{-3}$ | $-3.2 \times 10^{-3}$ | $4.7 \times 10^{-3}$ |
|  | 1000 | $5.0 \times 10^{-3}$ | $1.9 \times 10^{-3}$ | $2.8 \times 10^{-3}$ | $1.8 \times 10^{-3}$ |
| .5 | 10 | $3.1 \times 10^{-2}$ | $2.4 \times 10^{-2}$ | $-3.0 \times 10^{-3}$ | $2.3 \times 10^{-2}$ |
|  | 100 | $-1.4 \times 10^{-2}$ | $9.8 \times 10^{-3}$ | $-1.7 \times 10^{-2}$ | $9.2 \times 10^{-3}$ |
|  | 1000 | $-9.0 \times 10^{-4}$ | $3.5 \times 10^{-3}$ | $-1.4 \times 10^{-3}$ | $3.3 \times 10^{-3}$ |
| .7 | 10 | $-1.7 \times 10^{-1}$ | $2.8 \times 10^{-2}$ | $-1.8 \times 10^{-1}$ | $2.8 \times 10^{-2}$ |
|  | 100 | $-5.3 \times 10^{-2}$ | $1.6 \times 10^{-2}$ | $-5.7 \times 10^{-2}$ | $1.6 \times 10^{-2}$ |
|  | 1000 | $3.2 \times 10^{-4}$ | $6.3 \times 10^{-3}$ | $-4.6 \times 10^{-4}$ | $6.2 \times 10^{-3}$ |

$$E(\delta S \mid t \epsilon I_k, \text{ Queueing Record}) = x + \frac{1}{2} \frac{x^2}{I_k} \qquad (3.8)$$

Using the fact that $T_E \sim Nx + \sum_{k=1}^{N-1} I_k$ for $\rho \rightarrow 0$ and substituting (3.7) and (3.8) into (2.5) and (2.6) to form the delay gradient estimator, we obtain the following expression for $\hat{D}'_{(N)}$:

$$\hat{D}'_{(N)} \sim x + \frac{(2 - \frac{1}{N}) \frac{1}{2} x^2}{x + \frac{1}{N} \sum_{k=1}^{N-1} I_k} \qquad (3.9)$$

If we let $\rho$ approach zero, we can make the probability that $\frac{1}{N} \sum_{k=1}^{N-1} I_k$ is greater than any given number converging to 1 (since $\bar{I}_k = \frac{1}{\lambda}$). Hence, in the limit as $\rho \rightarrow 0$ the second term in (3.9) yields a zero contribution and $\hat{D}'_{(N)} \sim x$. Therefore, not only is the customer-addition algorithm unbiased as $\rho \rightarrow 0$, but Var $\hat{D}'_{(N)} \rightarrow 0$ for $\rho \rightarrow 0$.

Since as seen in Eqs. (2.123) and (2.242), the customer-removal and time-contraction algorithms have a similar structure, we can examine the behavior of both concurrently for $\rho \rightarrow 0$. As $\rho \rightarrow 0$, $M_i \sim 1$ and each customer has a zero waiting time. $Z_i$ and $P_i$ in Eqs. (2.123) and (2.242) become zero, and both estimators assume the following form:

$$\hat{D}'_{(N)} \sim \frac{\sum_{j=1}^{N} x_1^{(j)}}{N} \tag{3.10}$$

$x_1^{(j)}$ denotes the service requirement of the first customer in the j-th busy period. Hence, for $\rho \to 0$ $E\hat{D}'_{(N)} \sim \bar{x}$ and Var $\hat{D}'_{(N)} \sim \frac{\sigma_b^2}{N}$, where $\sigma_b^2$ denotes the variance of the service time density. Since from Eq. (2.24) it follows that $\frac{\partial D}{\partial \lambda}\Big|_{\rho=0} = \bar{x}$ for a general single-server queue, the customer-addition and time-contraction algorithms will be unbiased near $\rho = 0$, but the estimators will have a variance $\sigma_b^2/N$. Hence, since $\sigma_b = 0$ for an M/D/1 queue, we expect graphs 3.2 and 3.3 of the fractional rms error to pass through the origin.

Before examining the robustness of the time-contraction and customer-removal algorithms, we review the theoretical relations for the delay gradient in the case of M/G/1 and G/M/1 queues. Employing (2.130) for $D_c$ in Eq. (2.22), we obtain the following relation for an M/G/1 queue:

$$\frac{\partial D}{\partial \lambda} = \bar{x} \left\{ 1 + \frac{(1 + C_b^2)}{2} \; \frac{(2\rho - \rho^2)}{(1 - \rho)^2} \right\}, \tag{3.11}$$

where

$$C_b^2 \triangleq \frac{\sigma_b^2}{\bar{x}^2} \; . \tag{3.12}$$

For an M/D/1 queue $C_b^2 = 0$, while for an M/M/1 queue $C_b^2 = 1$.

In a G/M/1 queue we have a general inter-arrival time density A(x) with one-sided Laplace transform A*(s), and an exponentially distributed service requirement with parameter $\mu$. The waiting time $\overline{W}$ is specified by

$$\overline{W} = \frac{\sigma}{\mu(1-\sigma)} , \tag{3.13}$$

where $\sigma$ solves the nonlinear equation

$$\sigma = \underset{0 < \sigma < 1}{A^*} \quad (\mu - \mu\sigma) . \tag{3.14}$$

Applying Eq. (2.24), we find the delay gradient is given by

$$\frac{\partial D}{\partial \lambda} = \frac{1}{\mu}\left[1 + \frac{\sigma}{1-\sigma} + \frac{\rho}{(1-\sigma)^2} \frac{\partial \sigma}{\partial \rho}\right] . \tag{3.15}$$

$\frac{\partial \sigma}{\partial \rho}$ is determined by expressing (3.14) as a relation between $\rho$ and $\sigma$, differentiating both sides with respect to $\rho$, and solving for $\frac{\partial \sigma}{\partial \rho}$ .

The two examples of G/M/1 systems that we use are D/M/1 and U/M/1. For a D/M/1 queue the arrivals are periodic. A(x) and A*(s) are given by

$$A(x) = \delta(x - \frac{1}{\lambda}) \tag{3.16}$$

$$A^*(s) = e^{-s/\lambda} \tag{3.17}$$

Relation (3.14) is expressed as

$$\sigma = e^{-\frac{1}{\rho}} e^{\frac{\sigma}{\rho}} . \tag{3.18}$$

Differentiating (3.18) with respect to $\rho$ we find

$$\frac{\partial\sigma}{\partial\rho} = \frac{\sigma(\ell n\sigma)^2}{\sigma\ell n\sigma + (1-\sigma)} . \tag{3.19}$$

For a U/M/1 queue, the inter-arrival times are uniformly distributed. A(x) and A*(s) are given by

$$A(x) = \frac{\lambda}{2} [U_{-1}(x) - U_{-1}(x - \frac{2}{\lambda})] \tag{3.20}$$

$$A^*(s) = \frac{\lambda}{2} \frac{1}{s} \left(1 - e^{-\frac{2}{\lambda} s}\right) \tag{3.21}$$

$U_{-1}(x)$ denotes the unit step function. Relation (3.14) may be expressed as

$$\sigma = \frac{\rho\left(1 - e^{-\frac{2}{\rho}} e^{\frac{2\sigma}{\rho}}\right)}{2(1-\sigma)} . \tag{3.22}$$

Differentiating (3.22) with respect to $\rho$ we obtain $\frac{\partial\sigma}{\partial\rho}$ as

$$\frac{\partial\sigma}{\partial\rho} = \frac{1 - e^{-\frac{2}{\rho}(1-\sigma)}\left(1 + \frac{2}{\rho}(1-\sigma)\right)}{2\left(1 - 2\sigma + e^{-\frac{2}{\rho}(1-\sigma)}\right)} . \tag{3.23}$$

To construct $\frac{\partial D}{\partial \lambda}$ as a function of $\rho$ for a G/M/1 queue, we vary $\rho$, solving for the appropriate $\sigma$ by employing the fixed-point iteration method on Eq. (3.14)

$$\sigma_{n+1} = A^*(\mu - \mu\sigma_n) \tag{3.24}$$

We select some starting $0 < \sigma_0 < 1$ and apply the above iteration until $\sigma_{n+1} = \sigma_n$ to the desired accuracy.

Now we examine the robustness properties of the customer-removal and time-contraction algorithms by comparing their performance for M/M/1, U/M/1, and D/M/1 queues. The simulation results for the fractional rms error of the two algorithms for M/M/1, U/M/1, and D/M/1 queues are presented in graphs 3.6 through 3.11, respectively. The results for the relative bias are displayed in Tables 3-3 through 3-5. The time-contraction algorithm performs slightly better than the customer-removal algorithm for M/M/1 and both perform nearly the same for a U/M/1 queue. The only dramatic differences occur for the D/M/1 queue. Both the relative bias and fractional rms error show the time-contraction algorithm performing better than the customer-removal procedure. This result is reasonable since for a D/M/1 queue the time-contraction algorithm simulates a change in arrival rate in the exact way dictated by the structure of the queue.

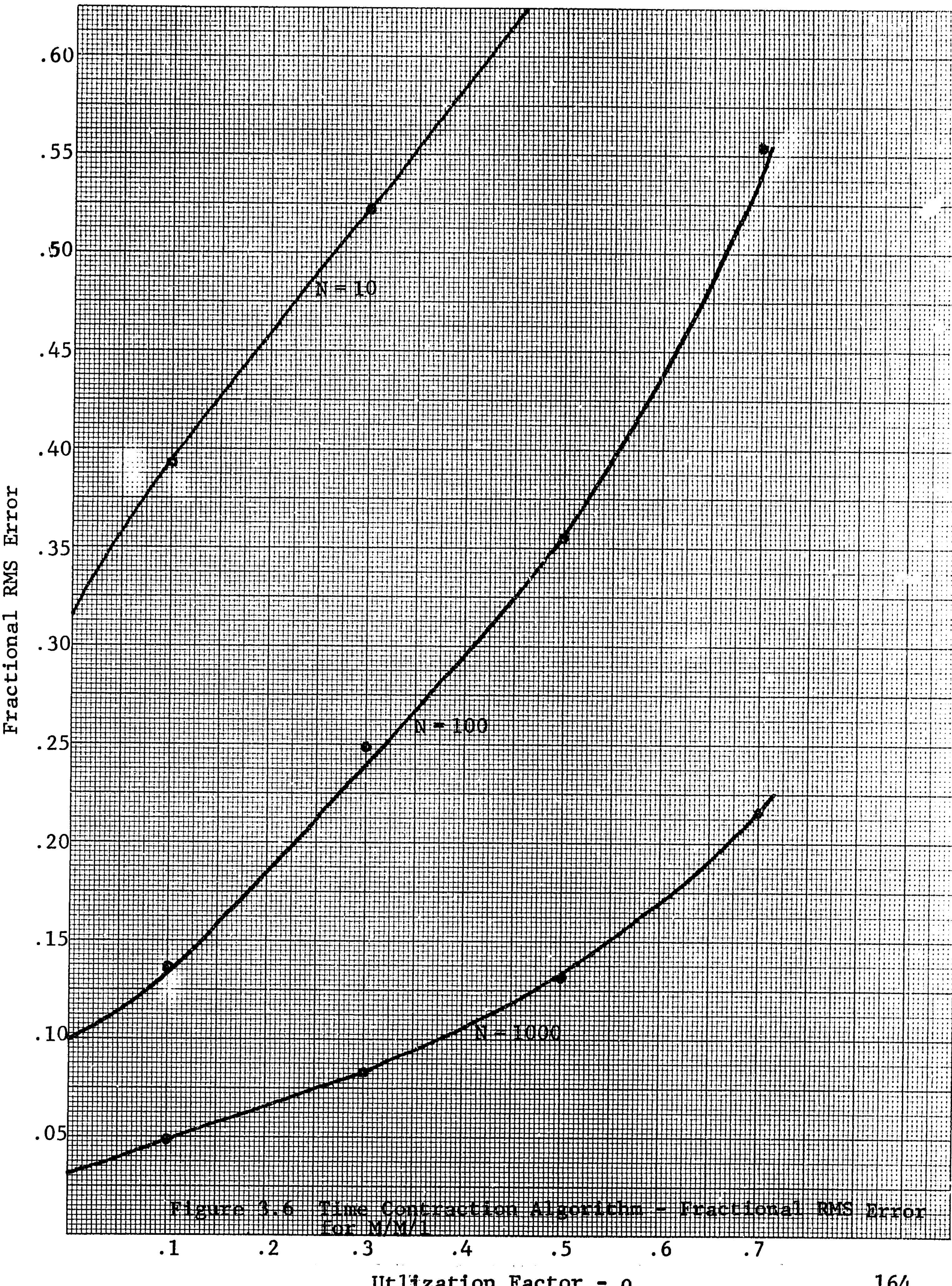


Figure 3.6 Time Contraction Algorithm - Fractional RMS Error for M/M/1

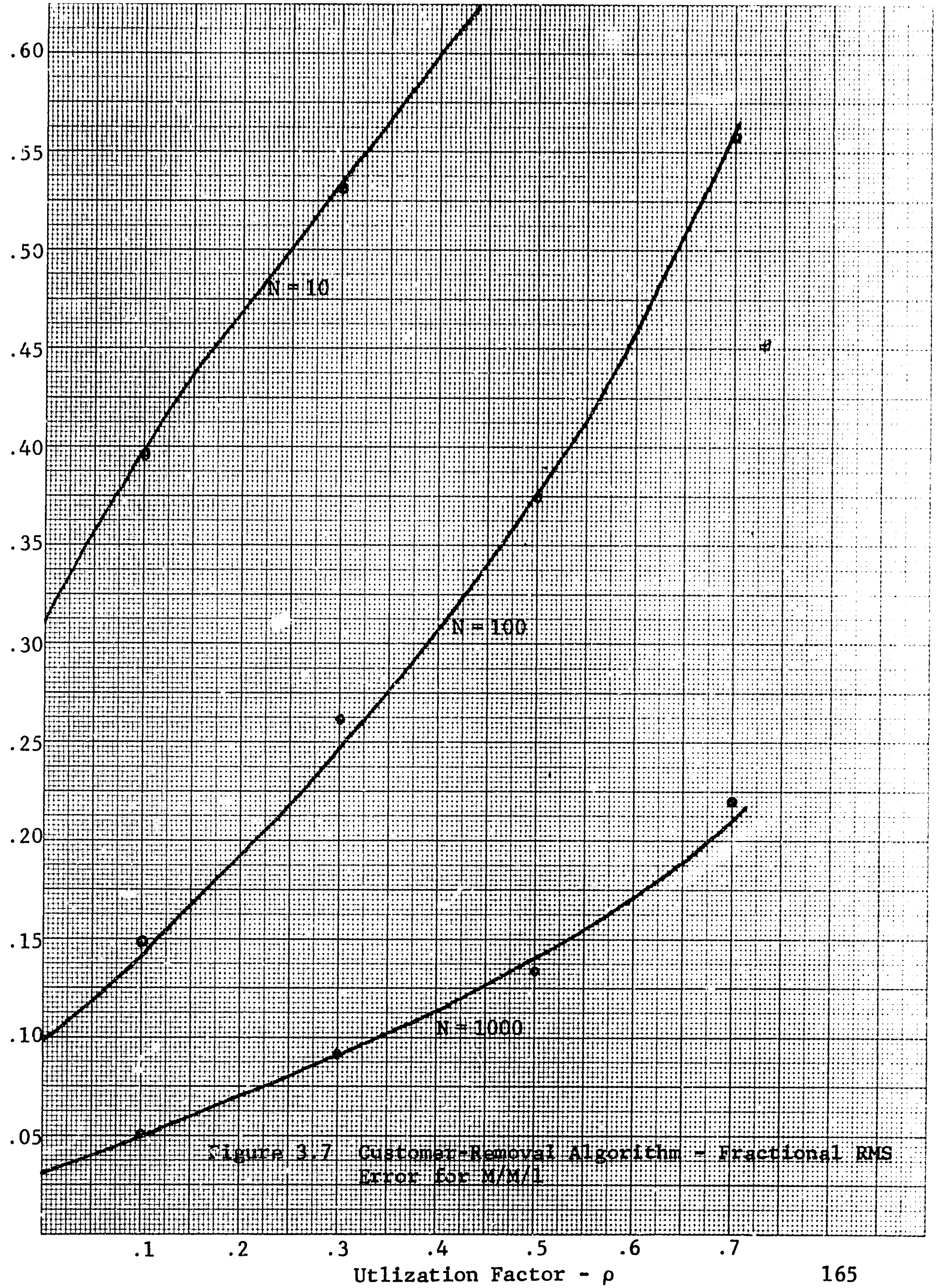


Figure 3.7 Customer-Removal Algorithm - Fractional RMS Error for M/M/1

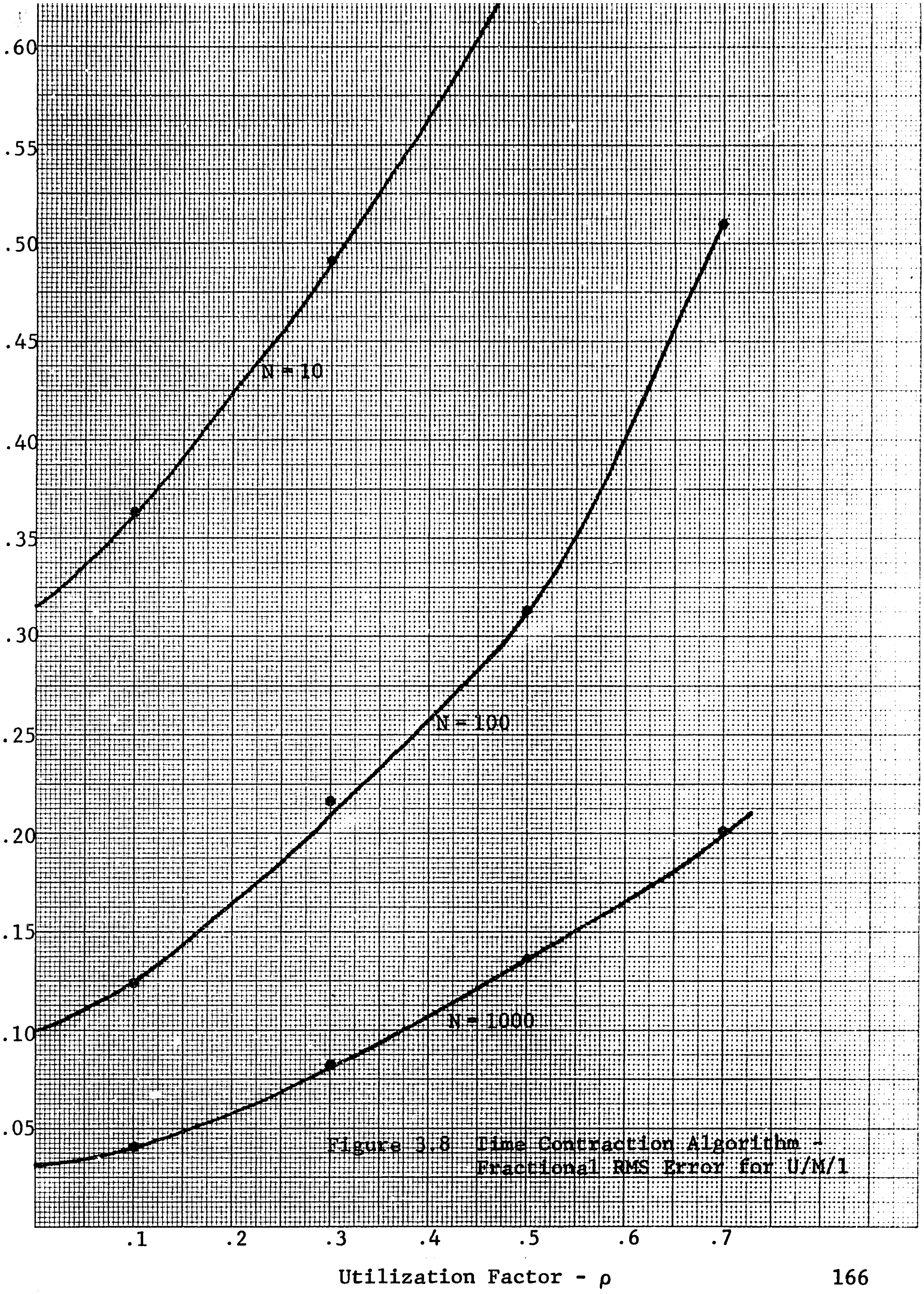


Figure 3.8 Time Contraction Algorithm - Fractional RMS Error for U/M/1

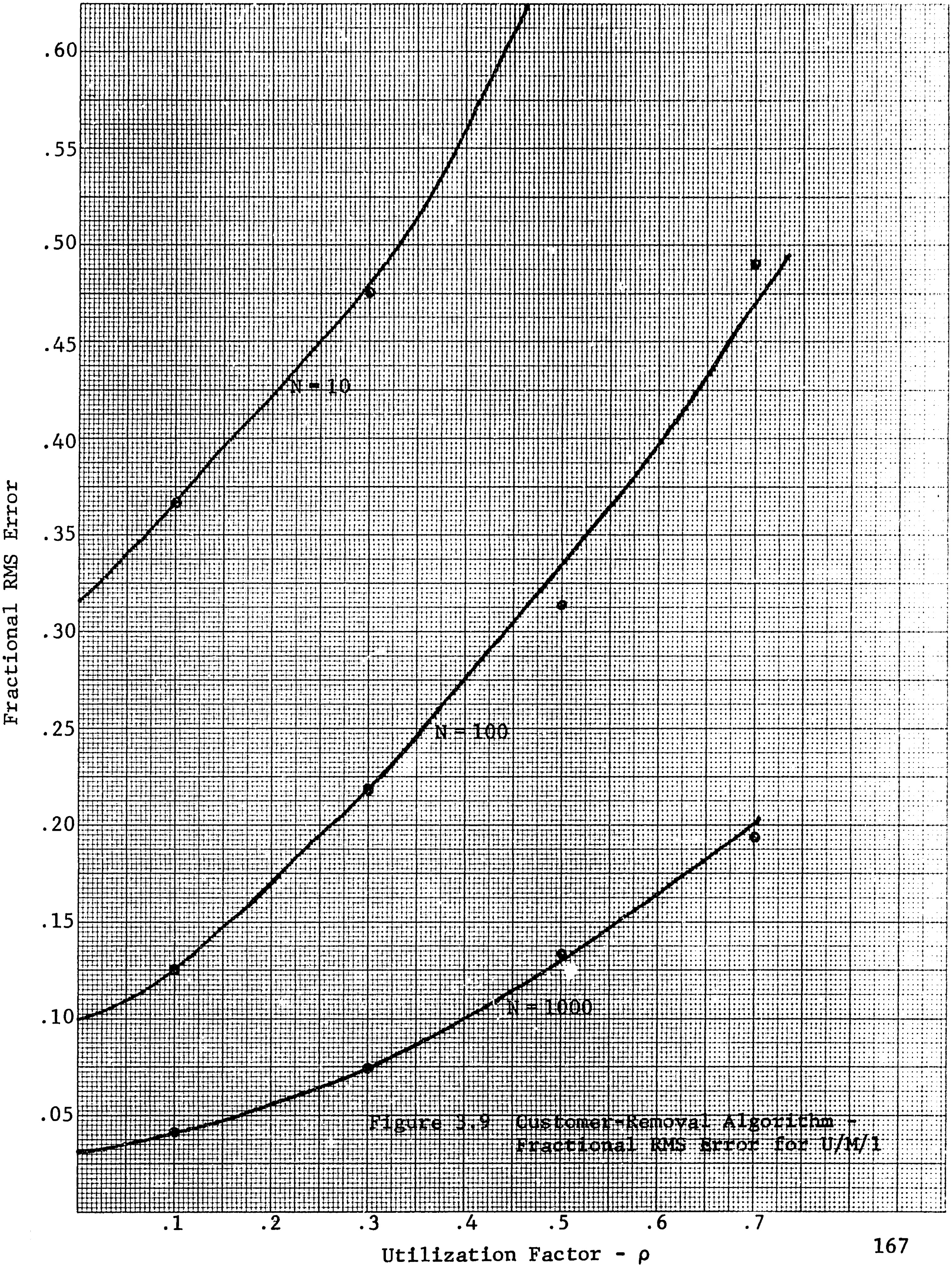


Figure 3.9 Customer-Removal Algorithm - Fractional RMS Error for U/M/1

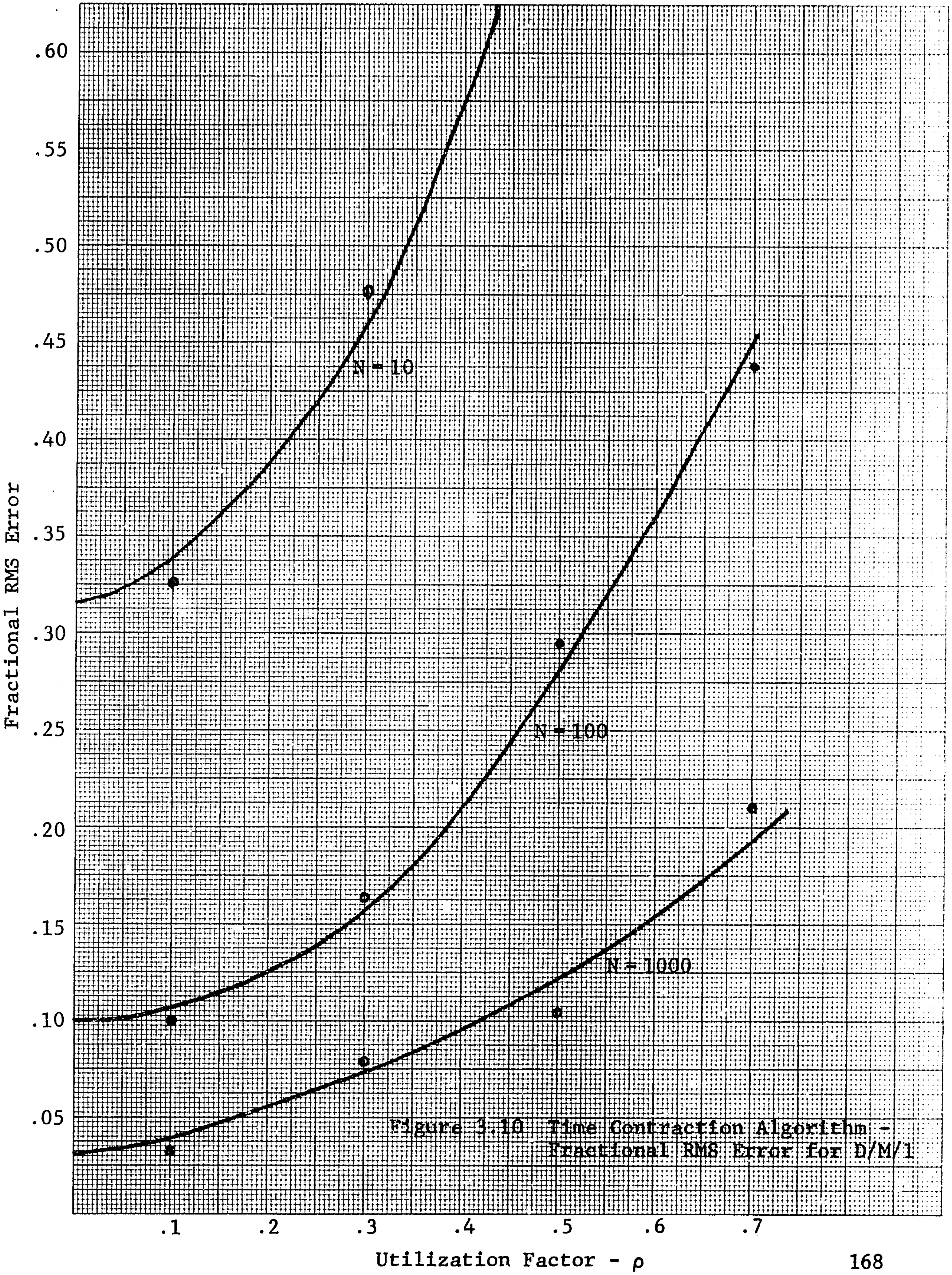


Figure 3.10 Time Contraction Algorithm - Fractional RMS Error for D/M/1

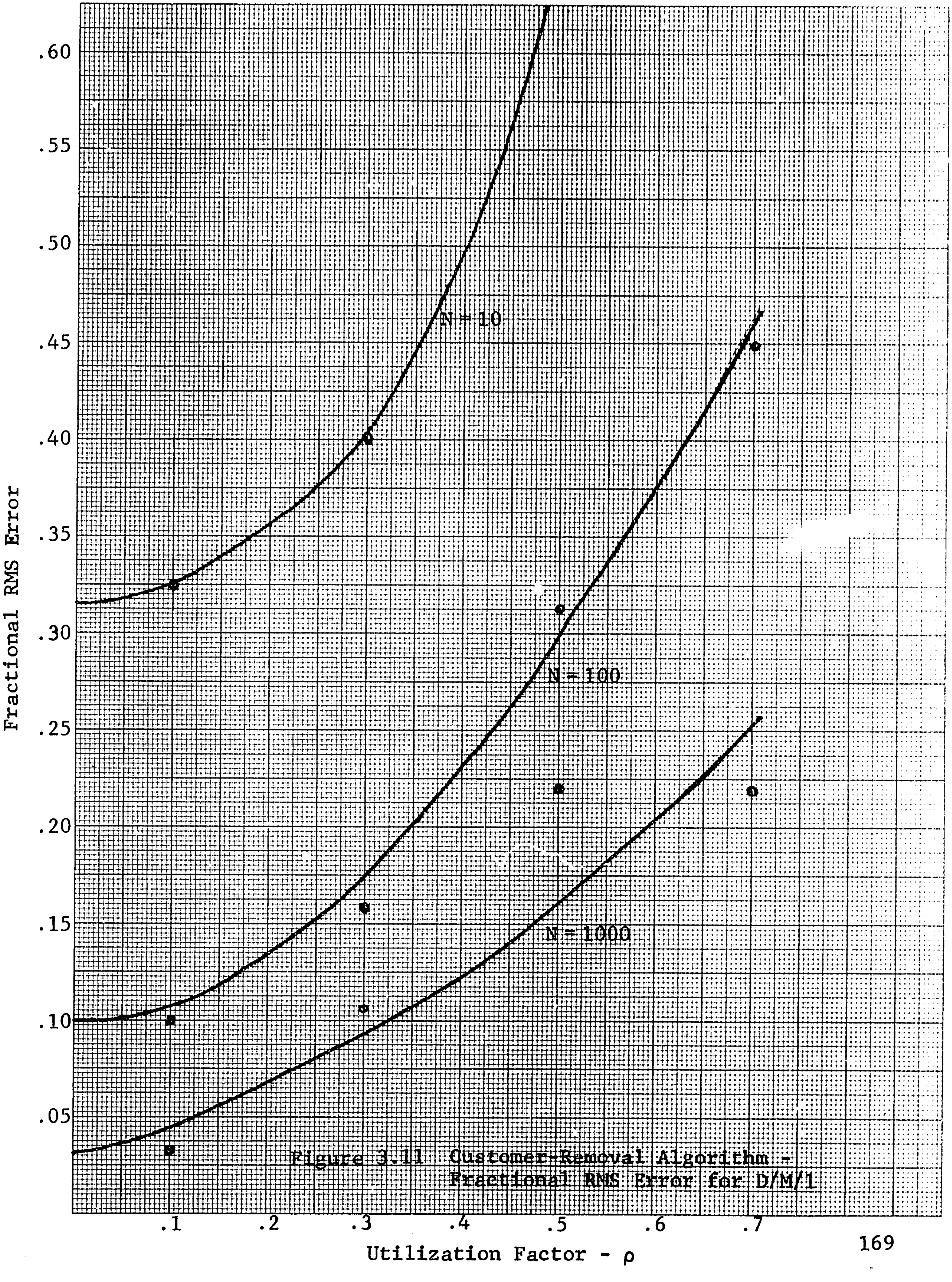


Figure 3.11 Customer-Removal Algorithm - Fractional RMS Error for D/M/1

## TABLE 3-3

## RELATIVE BIAS FOR CUSTOMER-REMOVAL AND TIME-CONTRACTION ALGORITHM (M/M/1 QUEUE)

| $\rho$ | N | $\dfrac{\hat{b}_{C.R.}}{\frac{\partial D}{\partial \lambda}}$ | $\dfrac{\hat{\sigma}_{\hat{D}(N),C.R.}/\sqrt{N_S}}{\frac{\partial D}{\partial \lambda}}$ | $\dfrac{\hat{b}_{T.C.}}{\frac{\partial D}{\partial \lambda}}$ | $\dfrac{\hat{\sigma}_{\hat{D}(N),T.C.}/\sqrt{N_S}}{\frac{\partial D}{\partial \lambda}}$ |
|---|---|---|---|---|---|
| .1 | 10 | $-1.5 \times 10^{-1}$ | $1.8 \times 10^{-2}$ | $-1.4 \times 10^{-1}$ | $1.8 \times 10^{-2}$ |
|  | 100 | $1.3 \times 10^{-3}$ | $7.4 \times 10^{-3}$ | $2.3 \times 10^{-3}$ | $6.7 \times 10^{-3}$ |
|  | 1000 | $7.2 \times 10^{-3}$ | $2.5 \times 10^{-3}$ | $8.5 \times 10^{-3}$ | $2.4 \times 10^{-3}$ |
| .3 | 10 | $-2.3 \times 10^{-1}$ | $2.4 \times 10^{-2}$ | $-2.2 \times 10^{-1}$ | $2.4 \times 10^{-2}$ |
|  | 100 | $-1.5 \times 10^{-3}$ | $1.3 \times 10^{-2}$ | $2.6 \times 10^{-3}$ | $1.2 \times 10^{-2}$ |
|  | 1000 | $2.1 \times 10^{-2}$ | $4.4 \times 10^{-3}$ | $2.4 \times 10^{-2}$ | $4.1 \times 10^{-3}$ |
| .5 | 10 | $-3.0 \times 10^{-1}$ | $3.2 \times 10^{-2}$ | $-3.0 \times 10^{-1}$ | $3.2 \times 10^{-2}$ |
|  | 100 | $-2.3 \times 10^{-2}$ | $1.9 \times 10^{-2}$ | $-2.8 \times 10^{-2}$ | $1.7 \times 10^{-2}$ |
|  | 1000 | $3.2 \times 10^{-2}$ | $6.4 \times 10^{-3}$ | $3.6 \times 10^{-2}$ | $6.3 \times 10^{-3}$ |
| .7 | 10 | $-4.4 \times 10^{-1}$ | $3.3 \times 10^{-2}$ | $-4.5 \times 10^{-1}$ | $3.1 \times 10^{-2}$ |
|  | 100 | $-3.0 \times 10^{-2}$ | $2.8 \times 10^{-2}$ | $-2.5 \times 10^{-2}$ | $2.7 \times 10^{-2}$ |
|  | 1000 | $6.4 \times 10^{-3}$ | $1.1 \times 10^{-2}$ | $8.8 \times 10^{-2}$ | $1.1 \times 10^{-2}$ |

TABLE 3-4

RELATIVE BIAS FOR CUSTOMER-REMOVAL AND TIME-CONTRACTION ALGORITHM (D/M/1 QUEUE)

| $\rho$ | N | $\frac{\hat{b}_{C.R.}}{\frac{\partial D}{\partial \lambda}}$ | $\frac{\hat{\sigma}_{\hat{D}(N),C.R.}/\sqrt{N_S}}{\frac{\partial D}{\partial \lambda}}$ | $\frac{\hat{b}_{T.C.}}{\frac{\partial D}{\partial \lambda}}$ | $\frac{\hat{\sigma}_{\hat{D}(N),T.C.}/\sqrt{N_S}}{\frac{\partial D}{\partial \lambda}}$ |
|---|---|---|---|---|---|
| .1 | 10 | $-1.0 \times 10^{-1}$ | $1.5 \times 10^{-2}$ | $-1.0 \times 10^{-1}$ | $1.5 \times 10^{-2}$ |
| | 100 | $-5.9 \times 10^{-4}$ | $5.0 \times 10^{-3}$ | $-5.9 \times 10^{-4}$ | $5.0 \times 10^{-3}$ |
| | 1000 | $-2.4 \times 10^{-3}$ | $1.6 \times 10^{-3}$ | $-1.8 \times 10^{-3}$ | $1.6 \times 10^{-3}$ |
| .3 | 10 | $-2.1 \times 10^{-1}$ | $1.7 \times 10^{-2}$ | $-1.3 \times 10^{-1}$ | $2.3 \times 10^{-2}$ |
| | 100 | $-1.0 \times 10^{-1}$ | $6.0 \times 10^{-3}$ | $-8.1 \times 10^{-3}$ | $8.2 \times 10^{-3}$ |
| | 1000 | $-9.8 \times 10^{-2}$ | $2.1 \times 10^{-3}$ | $-2.4 \times 10^{-3}$ | $2.9 \times 10^{-3}$ |
| .5 | 10 | $-3.6 \times 10^{-1}$ | $2.6 \times 10^{-2}$ | $-2.1 \times 10^{-1}$ | $3.6 \times 10^{-2}$ |
| | 100 | $-2.2 \times 10^{-1}$ | $1.1 \times 10^{-2}$ | $-1.4 \times 10^{-2}$ | $1.5 \times 10^{-2}$ |
| | 1000 | $-2.1 \times 10^{-1}$ | $4.2 \times 10^{-3}$ | $-5.0 \times 10^{-3}$ | $5.2 \times 10^{-3}$ |
| .7 | 10 | $-4.8 \times 10^{-1}$ | $3.8 \times 10^{-2}$ | $-3.4 \times 10^{-1}$ | $4.4 \times 10^{-2}$ |
| | 100 | $-2.3 \times 10^{-1}$ | $1.9 \times 10^{-2}$ | $-2.7 \times 10^{-2}$ | $2.2 \times 10^{-2}$ |
| | 1000 | $-1.2 \times 10^{-1}$ | $9.1 \times 10^{-3}$ | $7.8 \times 10^{-2}$ | $9.2 \times 10^{-3}$ |

TABLE 3-5

RELATIVE BIAS FOR CUSTOMER-REMOVAL AND
TIME-CONTRACTION ALGORITHM (U/M/1 QUEUE)

| $\rho$ | N | $\frac{\hat{b}_{C.R.}}{\frac{\partial D}{\partial \lambda}}$ | $\frac{\hat{\sigma}_{\hat{D}(N),C.R.}/\sqrt{N_S}}{\frac{\partial D}{\partial \lambda}}$ | $\frac{\hat{b}_{T.C.}}{\frac{\partial D}{\partial \lambda}}$ | $\frac{\hat{\sigma}_{\hat{D}(N),T.C.}/\sqrt{N_S}}{\frac{\partial D}{\partial \lambda}}$ |
|---|---|---|---|---|---|
| .1 | 10 | $-1.4 \times 10^{-1}$ | $1.7 \times 10^{-2}$ | $-1.3 \times 10^{-1}$ | $1.7 \times 10^{-2}$ |
|  | 100 | $-1.2 \times 10^{-3}$ | $6.4 \times 10^{-3}$ | $1.5 \times 10^{-3}$ | $6.1 \times 10^{-3}$ |
|  | 1000 | $5.2 \times 10^{-3}$ | $2.1 \times 10^{-3}$ | $7.1 \times 10^{-3}$ | $2.0 \times 10^{-3}$ |
| .3 | 10 | $-2.1 \times 10^{-1}$ | $2.1 \times 10^{-2}$ | $-1.9 \times 10^{-1}$ | $2.3 \times 10^{-2}$ |
|  | 100 | $-3.5 \times 10^{-2}$ | $1.1 \times 10^{-2}$ | $-2.7 \times 10^{-3}$ | $1.1 \times 10^{-2}$ |
|  | 1000 | $-1.3 \times 10^{-2}$ | $3.6 \times 10^{-3}$ | $3.1 \times 10^{-2}$ | $2.8 \times 10^{-3}$ |
| .5 | 10 | $-3.5 \times 10^{-1}$ | $2.7 \times 10^{-2}$ | $-2.9 \times 10^{-1}$ | $3.1 \times 10^{-2}$ |
|  | 100 | $-9.7 \times 10^{-2}$ | $1.5 \times 10^{-2}$ | $-1.2 \times 10^{-2}$ | $1.6 \times 10^{-2}$ |
|  | 1000 | $-4.3 \times 10^{-2}$ | $6.3 \times 10^{-3}$ | $4.5 \times 10^{-2}$ | $6.4 \times 10^{-3}$ |
| .7 | 10 | $-4.7 \times 10^{-1}$ | $3.2 \times 10^{-2}$ | $-4.1 \times 10^{-1}$ | $3.5 \times 10^{-2}$ |
|  | 100 | $-1.4 \times 10^{-1}$ | $2.4 \times 10^{-2}$ | $-3.7 \times 10^{-2}$ | $2.5 \times 10^{-2}$ |
|  | 1000 | $-4.2 \times 10^{-2}$ | $9.4 \times 10^{-3}$ | $5.7 \times 10^{-2}$ | $9.8 \times 10^{-3}$ |

In examining the fractional rms error curves of the customer-removal and time-contraction algorithms for M/M/1, U/M/1, and D/M/1 queues, we first note that the behavior of each algorithm is nearly the same for all three queues when $\rho$ is near 0. This follows from our demonstration that the estimator corresponding to both algorithms approaches Eq. (3.10) as $\rho \rightarrow 0$. Since for all our simulations we take $\bar{x} = \frac{1}{\mu} = 1$, the variance of the service time density is $\sigma_b^2 = 1$. Hence, our N busy period curve for the fractional rms error for each algorithm should cross the axis at approximately $\frac{1}{\sqrt{N}}$. The simulation results show this most clearly in the curves for the M/M/1 and U/M/1 queues. The points derived from the simulation of the D/M/1 queue are more scattered than those for the other two queues, making the interpretation of behavior near $\rho = 0$ uncertain.

For each queue the performance of the estimation algorithms degrades steadily with increasing $\rho$. Since the queue becomes non-stationary at $\rho = 1$, it is reasonable that the fractional rms error should become unbounded at $\rho = 1$. We can summarize the performance of the time-contraction and customer-removal procedures for each queue by listing least upper bounds on the fractional rms error for $\rho$ in the interval [0,.7]. For the N = 100 and N = 1000 busy period estimators, these least upper bounds are listed in Table 3-6.

TABLE 3-6

LEAST UPPER BOUNDS ON FRACTIONAL RMS ERROR FOR $\rho \epsilon$ [0, .7]

| | N = 1000 | | | N = 100 | | |
|---|---|---|---|---|---|---|
| | M/M/1 | D/M/1 | U/M/1 | M/M/1 | D/M/1 | U/M/1 |
| Customer-Removal Algorithm | .21 | .25 | .20 | .56 | .46 | .47 |
| Time-Contraction Algorithm | .22 | .19 | .20 | .54 | .45 | .51 |

We now pose the question of how long the observation period $T_E$ must be to contain an average number of N busy periods and hence achieve the fractional rms errors reported. A record of length $T_E$ contains an average number of customers $\lambda T_E$. From queueing theory we can derive an expression for $\overline{M}$, the average number of customers served/busy period. Hence, $\lambda T_E/\overline{M}$ will be the average number of busy periods/$T_E$ sec. observation time. Appealing to M/G/1 queueing theory, $\overline{M}$ is given by Eq. (2.93) and we can obtain the following condition on $T_E$ so the average number of busy periods contained in the observation period is greater than or equal to N.

$$T_E \geq \frac{\overline{x}\,N}{\rho(1-\rho)} \tag{3.25}$$

We now interpret (3.25) for the case of queues that occur in computer networks. In this situation, the customers are messages with a certain number of bits $\ell$. The service requirement is the time needed to transmit the message $\ell/c$, where c is the capacity of the communications link in bits/sec. Hence $\overline{x}$ is given by

$$\overline{x} = \frac{\overline{\ell}}{c} . \tag{3.26}$$

For representative purposes, values of $\overline{\ell}$, c, and $\overline{x}$ for the Arpanetwork are listed below

$$\begin{aligned} \bar{\ell} &= 1000 \text{ bits} \\ c &= 50 \times 10^3 \text{ bits/sec} \\ \bar{x} &= \frac{1}{50} \text{ sec.} \end{aligned} \tag{3.27}$$

If $\rho$ is unrestricted, (3.25) will require an unbounded observation time as $\rho$ is made arbitrarily close to 0 or 1. Hence, we hypothesize that $\rho$ is restricted to the interval [.1, .9]. Employing the value in (3.27) for $\bar{x}$ and letting $\rho \epsilon$ [.1, .9], for the observation interval $T_E$ to contain an average of N busy periods we must have

$$T_E \geq .222 \text{ N.} \tag{3.28}$$

For N = 1000 Eq. (3.28) yields $T_E \geq 222$ sec. or $T_E \geq 3.7$ min. and for N = 100, $T_E \geq 22.2$ sec. or $T_E \geq .37$ minutes.

# SECTION 4
# CONCLUSION

Hence, we have three estimation algorithms that appear from simulation and analytical results to be asymptotically unbiased, consistent, and asymptotically efficient in the case of an M/D/1 queue. We proved asymptotic unbiasedness for the time-contraction and customer-addition algorithms in the M/D/1 case. We showed that the asymptotic bias for the customer-removal algorithm, expressed as a power series in $\rho$, only contains terms of third-order or higher (for an M/D/1 queue). However, the closeness of the customer-removal simulation results to those of the time-contraction algorithm suggest that the customer-removal procedure is asymptotically unbiased. The consistency and asymptotic efficiency of the three algorithms follows from comparing the fractional rms errors derived from simulating the procedures on an M/D/1 queue with lower bounds on fractional rms errors derived from Cramer-Rao bounds on the variance of any unbiased estimator.

In evaluating the most promising algorithm as far as robustness, computational complexity, and storage requirements, we can only choose between the customer-removal and time-contraction algorithms, since the customer-addition algorithm as formulated is only applicable to queues where all customers have the same service requirement. The customer-removal algorithm requires on the order of $M^2$ minimization, $M^2$ additions, and M storage locations to form the quantity $\Delta$ needed to up-date the estimate at the end of the current busy period, where M is the number of customers served in the most recent busy period. The time-contraction algorithm requires eight storage locations and on the order of M additions. In examining the performance of the two algorithms for M/D/1, M/M/1, D/M/1, and U/M/1 queues, the only dramatic difference occurs in the case of a D/M/1 queue, where the customer-removal algorithm behaves worse as reflected in the size of the relative bias and the scatter of simulation points for the fractional rms error. Analytical results show that both algorithms are not asymptotically unbiased for an M/M/1 queue. Hence, our results suggest the time-contraction algorithm as the best candidate since it appears to be at least as robust as the customer-removal procedure, while having considerably less computational and storage requirements.

Whatever the estimation procedures we employ, there is a trade-off between observation time and the accuracy of our estimates that needs to be investigated further. If we up-date our routing variables every $T_r$ seconds, it would be desirable that our estimates should converge in approximately $\frac{1}{2} T_r$ seconds. On the basis of this criterion, and using the result for the Arpa net derived in Eq. (3.28), if we up-date our routing variables every

$T_r$ = 7.4 minutes, we will obtain estimates with performance at least as good as the N = 1000 busy period estimator. Equation (3.28) assumes that $\rho$ is confined to [.1, .9]. When the link is very free or very loaded the estimates will not be as accurate. However, this is ameliorated by the fact that in the loaded case we will try to take traffic away and in the lightly loaded case we will try to add additional traffic. In the Arpa-network, delay estimates are exchanged between nodes and new routes may be determined every $\frac{1}{2}$ second. Hence, we may not be able to up-date the routing as often as we like if we expect to do the estimation of the delay gradients with reasonable accuracy. However, since in a quasi-static procedure the changes in the routing variables are likely to be small, highly accurate estimates may be unnecessary.

One possibility for future work is in investigating the convergence of the quasi-static routing algorithms presented in Section 1 given that they use estimates, instead of exact values, for the marginal delays. The most ambitious approach would be to simulate the entire communication net with queues at each link and apply the estimation algorithms and the different "control" strategies for adjusting the routing variables. A second approach might be to approximate the estimates as some random variable we generate in the computer with a mean and variance derived from simulation results and see whether the quasi-static routing algorithms still converge to the optimal flow pattern.

# APPENDIX A

## VERIFICATION THAT (2.117) IS THE SOLUTION OF DIFFERENCE EQUATIONS (2.176)

The proof that (2.177) is the solution of the difference equations specified by (2.176) follows from substituting the solution into (2.176) and checking for equality. The following two identities are useful at certain steps in the proof:

$$\sum_{q'=0}^{k} \binom{k}{q'}(-1)^{q'} = 0 \qquad (A.1)$$

$$\sum_{q'=1}^{k} \binom{k}{q'}(-1)^{q'} q' = 0 \quad \text{(for } k > 1\text{)} \qquad (A.2)$$

Formula (A.1) follows from using the binomial representation of $(1-x)^k$ and substituting $x = 1$. Formula (A.2) comes from similarly expanding $\frac{d}{dx}(1-x)^k$ and evaluating it at $x = 1$ for $k > 1$.

Substituting our expressions for $\Omega_i^{(n)}$ and $\Omega_q^{(n+1)}$ given by (2.177) into Eq. (2.176), cancelling a common factor of $x^{n-q}/(q-1)!$ that appears on both sides, and making the change of variables $j = n-2-i$, we are left with the following relation which we must prove.

$$\frac{n^{n-q-1}}{(n-q)!} = \sum_{j=0}^{n-q-1} \frac{(n-2-j)}{q(n-1-q-j)!(j+1)!}(n-1)^j + \frac{1}{q(n-q)!} \qquad (A.3)$$

Letting q = n-k, we can recognize the binomial coefficient $\binom{k}{j+1}$, and multiplying both sides by k!(n-k) we are left with

$$n^k - kn^{k-1} = \sum_{j=0}^{k-1} \binom{k}{j+1}(n-2-j)(n-1)^j + 1. \qquad (A.4)$$

Relation (A.4) must be true for k = 0 ... n-1 with $n \geq 1$. By inspection, the relation is true for k=0 and k=1. We prove (A.4) for k > 1 by matching the coefficients of powers of n on both sides of the equation. We substitute for $(n-1)^j$ the representation given by the binomial formula and then change the order of summation, obtaining the following expression for the right side of (A.4):

$$\sum_{\ell=1}^{k} \left[ \sum_{j=\ell-1}^{k-1} \binom{k}{j+1}\binom{j}{j-\ell+1}(-1)^{j-\ell+1} \right] n^\ell$$

$$+ \sum_{\ell=0}^{k-1} \left[ \sum_{j=\ell}^{k-1} \binom{k}{j+1}\binom{j}{j-\ell}(-1)^{j-\ell+1}(2+j) \right] n^\ell + 1. \qquad (A.5)$$

The coefficient for $n^k$ in (A.5) is given by

$$\sum_{j=k-1}^{k-1} \binom{k}{j+1}\binom{j}{j-k+1}(-1)^{j-k+1} = 1 . \qquad (A.6)$$

The coefficient of $n^{k-1}$ is given by

$$\sum_{j=k-2}^{k-1} \binom{k}{j+1}\binom{j}{j-k+2}(-1)^{j-k+2} + \sum_{j=k-1}^{k-1} \binom{k}{j+1}\binom{j}{j-k+1}(-1)^{j-k+2}(2+j) = -k . \qquad (A.7)$$

The coefficient of $n^0$ is obtained by considering the second and third terms in (A.5).

$$\sum_{j=0}^{k-1} \binom{k}{j+1}\binom{j}{j}(-1)^{j+1}(2+j) + 1 \qquad \text{(A.8)}$$

Making the change of variable $j = q'-1$ and then employing (A.1) and (A.2), since we consider $k > 1$, we obtain the following expression from (A.8):

$$\sum_{q'=1}^{k} \binom{k}{q'}(-1)^{q'}(q'+1) + 1 = \sum_{q'=0}^{k} \binom{k}{q'}(-1)^{q'} + \sum_{q'=1}^{k} \binom{k}{q'}(-1)^{q'}q' = 0. \qquad \text{(A.9)}$$

The final step in the proof is to show that the coefficients of $n^{\ell}$ for $\ell = 1 \ldots k-2$ are zero. From (A.5) the desired coefficient of $n^{\ell}$ is expressed as

$$\sum_{j=\ell-1}^{k-1} \binom{k}{j+1}\binom{j}{j-\ell+1}(-1)^{j-\ell+1} + \sum_{j=\ell}^{k-1} \binom{k}{j+1}\binom{j}{j-\ell}(-1)^{j-\ell+1}(2+j) . \qquad \text{(A.10)}$$

Making the change of variables $j = q' + \ell-1$, breaking the second term into two summations, and regrouping the binomial coefficient terms in one of the new summations, we obtain

$$\sum_{q'=0}^{k-\ell} \binom{k}{\ell+q'}\binom{q'+\ell-1}{q'}(-1)^{q'} + \sum_{q'=1}^{k-\ell} \binom{k}{\ell+q'}\binom{q'+\ell-1}{q'-1}(-1)^{q'}$$

$$+ \binom{k}{\ell} \sum_{q'=1}^{k-\ell} \binom{k-\ell}{q'} q'(-1)^{q'} \ . \tag{A.11}$$

From (A.2), for $k-\ell > 1$ or $1 \leq \ell \leq k-2$, we recognize the last term in (A.11) as zero. Grouping the first two summations in (A.11) together, we are left with

$$\binom{k}{\ell} + \sum_{q'=1}^{k-\ell} \binom{k}{\ell+q'}(-1)^{q'}\{\binom{q'+\ell-1}{q'} + \binom{q'+\ell-1}{q'-1}\} \ . \tag{A.12}$$

Now we employ the following identity for binomial coefficients:

$$\binom{m}{p} + \binom{m}{p+1} = \binom{m+1}{p+1} \ . \tag{A.13}$$

Hence, (A.12) becomes

$$\binom{k}{\ell} + \sum_{q'=1}^{k-\ell} \binom{k}{\ell+q'}\binom{q'+\ell}{q'}(-1)^{q'} \ . \tag{A.14}$$

The product of the two binomial coefficients in the summation in (A.14) may be represented alternatively as

$$\binom{k}{\ell+q'}\binom{q'+\ell}{q'} = \binom{k}{\ell}\binom{k-\ell}{q'} \ . \tag{A.15}$$

Substituting (A.15) into (A.14), adding and subtracting 1 in the summation, we are left with

$$\binom{k}{\ell} + \binom{k}{\ell} \left\{ \sum_{q'=0}^{k-\ell} \binom{k-\ell}{q'} (-1)^{q'} - 1 \right\} . \tag{A.16}$$

But Eq. (A.16) is zero, since by (A.1) the summation inside the brackets is zero.

# APPENDIX B

## CALCULATION OF $\bar{C}_3^1|3$, $\bar{C}_3^1|4$, $\bar{C}_4^1|4$ FOR AN M/D/1 QUEUE

We outline here the calculation of $\bar{C}_3^1|3$, $\bar{C}_3^1|4$, and $\bar{C}_4^1|4$. $C_3^1$ is given by

$$C_3^1 = \min\{x,z\} , \tag{B.1}$$

where

$$z = \min\{\omega_2,\omega_3\} . \tag{B.2}$$

The joint density of $\omega_2$ and $\omega_3$ conditioned on M=3 is as follows:

$$p(\omega_2,\omega_3|M=3) = \frac{2}{3x^2}$$

$$0 \leq \omega_2 \leq x$$

$$0 \leq \omega_3 \leq x + \omega_2 \tag{B.3}$$

Hence, the distribution function of z is

$$F_{z|M=3}(\tau) = 1 - \frac{2}{3x^2} \int_{\omega_2=\tau}^{x} \int_{\omega_3=\tau}^{x+\omega_2} d\omega_3 d\omega_2 . \tag{B.4}$$
$$0 \leq \tau \leq x$$

Differentiating (B.4) with respect to $\tau$ we obtain

$$P_{z|M=3}(\tau) = \frac{2}{3x^2} (2x - \tau) . \tag{B.5}$$
$$0 \leq \tau \leq x$$

Since $F_{z|M=3}(x) = 1$, by (2.168) $\bar{C}_3^1|3$ is computed as

$$\bar{C}_3^1|3 = \int_0^x \tau P_{z|M=3}(\tau)\, d\tau = \frac{4}{9}x \tag{B.6}$$

We next consider the calculation of $\bar{C}_3^1|4$. $C_3^1$ and the appropriate z are defined in (B.1) and (B.2). The joint density of $(\omega_2, \omega_3)$ conditioned on M=4 is specified by (2.169) as

$$\begin{aligned} &p(\omega_2, \omega_3|M=4) = \frac{3}{8x^3} \int_{\omega_4=0}^{x+\omega_3} d\omega_4 . \\ &\quad 0 \leq \omega_2 \leq x \\ &0 \leq \omega_3 \leq x+\omega_2 \end{aligned} \tag{B.7}$$

The distribution function of z is found as

$$\begin{aligned} &F_{z|M=4}(\tau) = 1 - \frac{3}{8x^3} \int_{\omega_2=\tau}^{x} \int_{\omega_3=\tau}^{x+\omega_2} (x+\omega_3)\, d\omega_3 d\omega_2 . \\ &0 \leq \tau \leq x \end{aligned} \tag{B.8}$$

Differentiating (B.8) with respect to $\tau$, $P_{z|M=4}(\tau)$ is given by

$$\begin{aligned} &P_{z|M=4}(\tau) = \frac{3}{8x^3} [\frac{5}{2}x^2 + x\tau - \tau^2] . \\ &0 \leq \tau \leq x \end{aligned} \tag{B.9}$$

Noting again that $F_{z|M=4}(x) = 1$, by (2.168) $\bar{C}_3^1|4$ is calculated as

$$\bar{C}_3^1|4 = \int_{\tau=0}^{x} \tau P_{z|M=4}(\tau)\, d\tau = \frac{1}{2}x . \tag{B.10}$$

Our final calculation is of $\overline{C}_4^1|4$. $C_4^1$ is defined by

$$C_4^1 = \min\{x,z\}, \qquad (B.11)$$

where

$$z = \min\{\omega_2,\omega_3,\omega_4\}. \qquad (B.12)$$

From the joint density of $(\omega_2,\omega_3,\omega_4)$ conditioned on M=4 given in (2.190), we express the distribution function of z as

$$\underset{0 \leq \tau \leq x}{F_{z|M=4}(\tau)} = 1 - \frac{3}{8x^3} \int_{\omega_2=\tau}^{x} \int_{\omega_3=\tau}^{x+\omega_2} \int_{\omega_4=\tau}^{x+\omega_3} d\omega_4 d\omega_3 d\omega_2 . \qquad (B.13)$$

Differentiating (B.13) with respect to $\tau$, we obtain

$$\underset{0 \leq \tau \leq x}{P_{z|M=4}(\tau)} = \frac{3}{8x^3}[4x^2 - 3x\tau + \frac{1}{2}\tau^2]. \qquad (B.14)$$

Since $F_{z|M=4}(x) = 1$, by (2.168) $\overline{C}_4^1|4$ is computed as

$$\overline{C}_4^1|4 = \int_0^x \tau P_{z|M=4}(\tau)\, d\tau = \frac{27}{64}x. \qquad (B.15)$$

# REFERENCES

1] C. Agnew, "On the Optimality of Adaptive Routing Algorithsm," Conference Record of the National Telecommunications Conference, 1974, pp. 1021 - 1025.

2] D. G. Cantor and M. Gerla, "Optimal Routing in a Packet-Switched Computer Network," IEEE Trans. on Computers, Vol. C-23, No. 10, pp. 1062 - 1069.

3] R. Gallager, "A Minimum Delay Routing Algorithm Using Distributed Computation," IEEE Trans. on Communications, Vol. COM-25, No. 1, pp. 73 - 85.

4] L. Kleinrock, Queueing Theory, Wiley-Interscience, 1975.

5] L. Kleinrock, Communication Nets: Stochastic Message Flow and Delay, McGraw-Hill, 1964.

6] A. Segall, "The Modeling of Adaptive Routing in Data-Communication Networks," IEEE Trans. on Communications, Vol. COM-25, No. 1, pp. 85 - 95.

7] M. Schwartz and C. K. Cheung, "The Gradient Algorithm for Multiple Routing in Message-Switched Networks," Proceedings of the Fourth Data Communication Symposium, Quebec City, October 1975.