# AN ALGORITHM FOR THE SOLUTION OF A CLASS OF INDUSTRIAL QUEUING PROBLEMS USING GEOMETRIC PROGRAMMING AND CONDENSATION

BY

KIMBERLY OSTER

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Mathematics).

Golden, Colorado

Date: 11/30/90

Signed: Kimberly M. Oster
Kimberly M. Oster

Approved:
Dr. R. E. D. Woolsey
Thesis Advisor

Golden, Colorado

Date: 11\30\90

Dr. Ardel J. Boes
Professor and Head
Department of Mathematics

## ABSTRACT

Many industrial queuing problems can be modelled as M/G/1 queues or M/M/s queues with parallel servers. With given parameters, the optimal service capacity can be determined through linear search, as the cost functions are unimodal.

This study presents a proof, assuming the single server is capable of providing service at the same rate as all s servers combined, that M/M/s queues are economically optimal only if $s = 1$. That is, it is economically preferable to build a single facility with a large capacity than multiple servers with the same total capacity.

The thesis then develops a convergent algorithm capable of solving economic queuing problems where the queue has unlimited queue length, unlimited calling population, first-in first-out service order and is either M/M/1 or M/G/1. The algorithm uses condensation methods to reduce the one and three degree of difficulty, constrained, signomial problems to zero degree of difficulty, posynomial problems which are then directly solvable by conventional geometric programming techniques.

The algorithm represents a more direct method for estimation of the optimum than the linear search, and in most problems, requires less total calculations. A variety of hypothetical problems were used to successfully test the algorithm and the results were verified by linear search.

## TABLE OF CONTENTS

## LIST OF FIGURES AND TABLES

## ACKNOWLEDGEMENTS

I would like to take this opportunity to thank the members of my thesis committee, Dr. Ruth Maurer and Prof. William Astle, for the assistance and advice, and to especially thank Dr. R. E. D. Woolsey for is guidance and assistance, not only in the evolution of this thesis, but throughout my post-graduate pursuits.

I would also like to thank my family for the patience and understanding. I owe much to my friends here at mines for helping me to maintain a realistic perspective, foremost Jan Caffey, Joe Katz, and Tamar Raphaeli, as well as my friends elsewhere.

Finally, I'd like to generally thank the guild; the teamwork and camaraderie is truly responsible for my achievements.

# Chapter 1

# AN INTRODUCTION TO INDUSTRIAL QUEUING MODELS

Consider a hypothetical shipping company, or division of a company. Among other methods of transport, the company uses cargo ships between several ports. This necessitates the company's ownership, or at least use, of docking and loading/unloading facilities at each port it utilizes. The related costs to the company of this docking operation include the capital costs of the docking equipment, salaries paid to the employees and other various maintenance costs. Additionally, if the efficiency of the facility is such that a ship must wait outside the dock while a previous ship is being loaded or unloaded, the maintenance costs of that waiting ship are also part of the dock's cost of operation. The company wants to know how many docks it should operate at each port, and at what service level to operate each.

This system, and systems like it, where arriving traffic accumulates in a line and demands service, are known as **queues**, or waiting lines. A queue is defined by six characteristics. The interarrival times are random variables from a given **interarrival distribution**. The service times are random variables from a given **service distribution**. There are a specified **number of servers**, which have a clearly defined **order of service** (i.e. first-in first-out (FIFO), last-in first-out (LIFO), prioritized, etc.). Finally, declaring the **size of the potential calling population** and the **maximum queue length** fully defines the queue. For this shipping example, only random arrivals (i.e. those whose interarrival times are independent, identically distributed, random variables from the memory-less exponential distribution), from a potentially infinite calling population, at a FIFO queue of unlimited length, are relevant.

The accepted notation for queues is the Kendall-Lee format (Winston 1987) which lists the six characteristics, in the order given above, separated by vertical slashes. Exponential interarrival time distributions are indicated by an M, as are exponential service time distributions. Other distributions of interest are a general distribution (G or GI), the Erlang ($E_k$), and deterministic (D) behavior. It is a common practice to list only the first three characteristics when dealing with a FIFO queue of infinite length and infinite calling population. Using this notation, M/G/1 and M/M/s queues are those relevant to this problem.

In the case of industrial queues, the operator is economically responsible for both the servers (in the above case, the docks) and the customers (the ships). The total expected cost of queue operation can be split one of two ways. Van Voorhis (1956) divided it into the cost due to customers waiting for servers (waiting time) and that due to servers waiting for customers (idle time). We assume that both these costs are linear functions of time (i.e. waiting cost is linear function of waiting time and idle cost is a linear function of idle time). Since the waiting time is inversely proportional to the service capacity (the number of servers multiplied by their service rate) while the idle time is directly proportional to the service capacity, the waiting and service costs are also inversely and directly proportional to service level, respectively. If the total cost is the sum of the two, the minimum total cost is at the intersection of these two functions where the waiting cost equals the idle cost (see fig. 1) and any change in this total cost is instigated by a change in behavior of either the servers, the customers, or both.

Hillier (1964), on the other hand, split the total cost between the same waiting cost and the cost of servicing customers. The service cost is assumed to be directly proportional to service rate. Thus, the service cost, like the idle cost, is directly proportional to the service capacity. The total cost remains the sum of the service and waiting costs. The

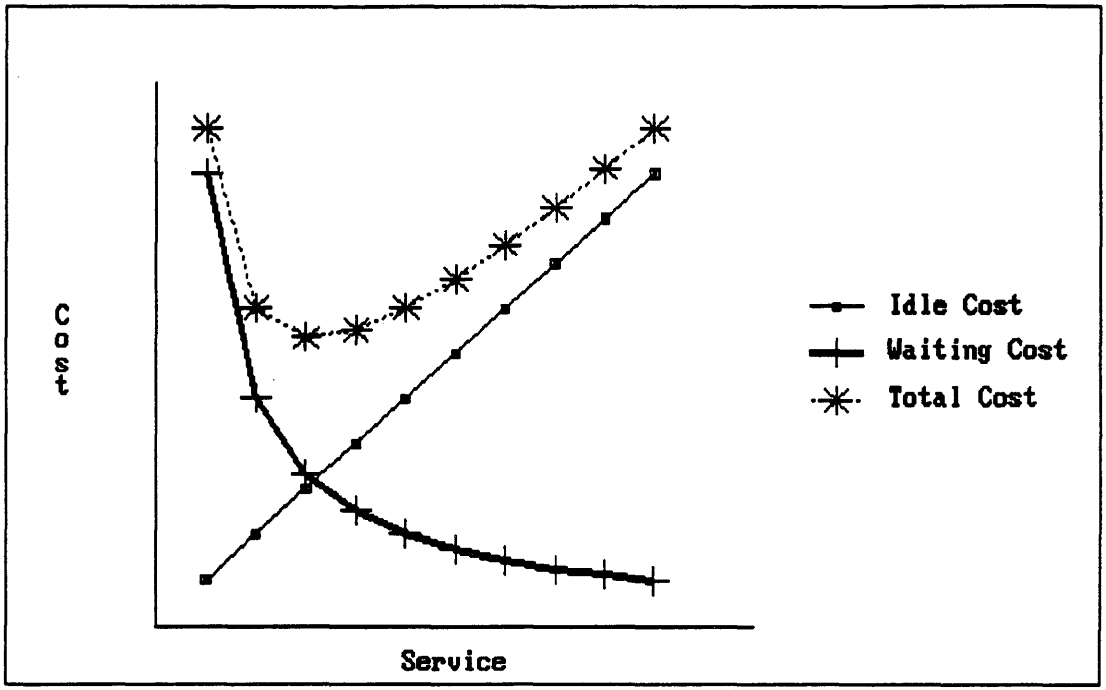

Figure 1: Operation Costs versus Service Level

*Source*: Van Voorhis, W. R. 1956. Waiting line theory as a management tool. Operations Research 4: 233.

minimum of the total cost remains at the intersection of the two functions where the waiting cost is equal to the service cost.

In his work in 1963, Hillier proposed several cost models for minimization of total expected cost (TEC); each is suited to the different decision variables common to such a model. The first is intended for a situation where the number of servers is the only unknown. The second adds the mean arrival rate to the list of unknowns. The third (and final) model is designed for the case when both the number of servers and the mean service rate are undetermined[1]. This is the appropriate model for the shipping problem introduced above, where the individual docks are the servers and the service rate is dependent on the number of workers and the equipment employed.

Hillier defined the expected service cost as

$$C_s \times s \times f(\mu),$$

where $C_s$ is the cost of service per server per customer ($/server/customer), s is the number of servers, $\mu$ is the mean service rate (customer/time), and f is the "ratio of the marginal cost of service at (service rate) $\mu$ to the marginal cost of service when the average service time is one unit of time."[2] Thus, the expected service cost has units of $/time.

The expected waiting cost (also with units of $/time) is

$$C_W \times L = C_W \times L_W + C_W \times L_S,$$

where $C_w$ is the cost of time per customer per unit time ($/customer/time), and L, the expected number of customers in the queue system at a given time is equal to the sum of $L_w$, the expected number of customers in the waiting line, and $L_s$, the expected number of

---

[1] Note the second and third models are reversed in Hillier and Lieberman (1986).

[2] This will be defined further at the end of this chapter.

customers in service.

The total expected cost is the sum of the expected service and waiting costs. The goal is to minimize the TEC (i.e. minimize the long run expected total cost per unit time). The fact that TEC is a long run expected value indicates that the system must reach a steady (or equilibrium) state. For that to occur in any queue the mean service rate ($\mu$) must be greater than the mean arrival rate ($\lambda$). If $\mu \leq \lambda$, then, in the long run, the waiting line becomes infinitely long and can never be emptied. When $\lambda < \mu$, the queue length varies and expected behavior is relatively easy to derive given an exponential interarrival time distribution.

Note that total expected cost is not the only measure of system performance. There are others such as the average wait of all the customers, the average wait of the customers who must wait, the percent of the customers who wait longer than a given acceptable time, and the total amount of idle server time. Shelton (1960) discusses these and other methods more thoroughly.

Historically, two approaches have been taken to this type of problem: simulated sampling (or simulation); and optimization through mathematical modeling. Simulation answers the question of what output results from given input. Optimization does the opposite, determining what inputs are necessary to cause given output values.

Optimization has itself been divided into two areas: control and design. Control involves determining an optimal operation policy of a given queue. That is, minimizing the costs of operation of an existing queue by turning servers off and on and varying their service rates as a response to the status of the queue; it is a matter of dynamic decision models. Designing, static decision modelling, focuses on determining an optimal queue assuming that all potential will always be used. In other words, all servers present will be turned on and functioning at the maximum possible service rates; it is assumed there is no advantage to dynamically varying the queue's serving capacity.

All three of these queue-related topics (simulation, control, and design) have been subjects of considerable research since the late 1950's.

First, it is helpful to consider a few examples of work in queue simulation. Van Voorhis (1956) provided a basic example in his simulation of a full service gas station with random service and arrivals. Setting the salary of station attendants and the average revenue per customer, while allowing that a customer arriving at a long queue may choose to leave, the author ran the model for various numbers of attendants and determined the best manning policy. This was a simple example, more easily solved by mathematical modelling and optimization but many queues are so intricate that direct mathematical modelling is extremely difficult.

The warehouse docking model of Schiller and Lavin (1956) is a more involved example. Wiebolt Stores, Inc. of Chicago was operating three warehouses but planned to consolidate to a single warehouse in the near future. There were constraints on the access and servicing ability of the single warehouse which where complicated by varying mean arrival rates at different times of the day. By running the model under different scenarios the authors determined the optimal number of new docks to add to the warehouse.

More recently, Chelst, Tilles, and Pipis (1981) used simulation in a coal transport problem. The company in question used a single unloader to service incoming coal cars. Constant use caused frequent down-time for maintenance, and coal cars waiting in cold weather often arrived with frozen coal that damaged the facility during processing. Using simulation, the authors evaluated the addition of a second unloader and the possibility of alternating operation between the two.

Control design emerged in the early 1960's when Yadin and Naor (1963) were the first in a series of mathematicians to investigate the optimal operation policy of the M/G/1 queue using mathematical modelling and optimization. The authors recognized that for a queue to attain a steady state there must be excess service capacity and therefore idle periods for the server. Their algorithm was based upon reducing the idle fraction of the server's time by shutting the server off when no customers are present. Due to set-up and shut-down costs it is not optimal to turn the server back on immediately when a new customer arrives. Yadin and Naor derived a value, R, dependent on set-up, shut-down, waiting, and idle costs as well as the six basic queue characteristics, such that when the number of waiting customers reaches or surpasses R, the server is returned to operating status.

In 1967, the same authors addressed control from the angle of varying the service rate and outlined an optimization procedure such that the service rate was a function of the number of customers waiting as well as the recent history of the system.

Heymann (1968) further investigated Yadin and Naor's M/G/1 with removable server. Discounting the costs due to the long term nature of the problem, he obtained an equation for the expected discounted cost as a function of the same value R and the interest rate. Heymann also presented a recursion relation to find the optimal dynamic policy for undiscounted problems.

Later that same year, Heymann and Marshall extended this field of dynamic control by server removal to the GI/G/1 queue. Considering undiscounted problems, they bounded the cost rate and the optimal policy.

In 1971, Bell provided the optimality proof for Heymann's discounted M/G/1 model, as well as an "improved computational algorithm" for determining R and the optimal policy.

The following year, Crabhill revived the study of queue control through variance of service rate. Assuming k possible service rates and a cost structure such that one cost is dependent upon the service rate and another on the number of customers waiting, Crabhill presented an algorithm to derive optimal policy and demonstrated it for the case k is two.

Most recently, Bell (1980) extended his research to the M/M/2 undiscounted queue and Szarkowicz (1985) analyzed both discounted and undiscounted M/M/s queues with removable servers.

Research in queue design has not been so extensive. The earliest work was done by Mangelsdorf (1955), who introduced the use of economic models to analyze waiting line problems. He used the convention of idle and waiting costs (versus service and waiting costs) and calculated optimal ratios of these costs for given, fully defined, queues. Mangelsdorf's focus was mainly on machine assignment problems which are subject to a limited calling population.

In 1963, Hillier entered the field by introducing his three queue cost models mentioned above. All of his models assume an infinite calling population, in contrast to Mangelsdorf, but the two authors agreed that one of the more difficult aspects of queue design is determining the waiting cost. Hillier drew a comparison between waiting cost and the stock out costs of inventory, hoping to simplify the problem.

Hillier (1964) describes the general approach appropriate for queuing models and summarizes the structure of the model. He also mentions a problem, similar to the shipping problem addressed here, involving a crew, performing together as a single server, whose service rate can be altered by addition of more crew members, equipment upgrade, etc.

In 1967, Hillier and Lieberman collaborated on a text which includes a summary of Hillier's 1964 paper. Additionally, they discuss two cases in the formulation of waiting cost functions. The first case covers customers who are external to the service organization. In this situation the waiting cost is due to loss of future business and is proportional to the wait experienced by each individual customer. The second case is for internal customers whose waiting cost is due to lost productivity and is, therefore, proportional to the number of customers stuck in the queue. The authors also review the effect on cost of the customers' travel time to and from the server.

Stidham (1968) (as well as his article of 1970) uses Hillier's models to attack general queues. Using stochastic minimization he proves single server optimality for the G/M/s, G/D/s, and G/$E_k$/s queues. He later extends this result to service costs that are non-linear but concave and waiting costs that are monotone increasing in the waiting time. He also addresses queue networks and non-FIFO queues.

Most of the above research (with the exception of Stidham) has made the assumption of service costs that are directly proportional to the service rate. That is, $f(\mu)=\mu$, and the service cost is $sC_s\mu$. This is not always an accurate assumption and, in the case of the shipping problem, may not be at all correct.

Andress (1954) and Yelle (1979) both provide a good introduction to the theory and uses of the learning curve. The theory behind learning is simply that the more frequently an act is performed the more efficient the performer becomes; the performer learns as the work is done. The goal of learning theory is to uncover an efficiency improvement rate regular enough to be predictable.

Andress introduces three learning curve formulas, the last of which describes learning in a queue. The formula is as follows:

$$T = kx^{n+1},$$

for T, the total man hours required to build x units, k, the cost in man hours of the first unit, and n, the learning index.[3] This effect appears in the service cost of the shipping problem where f(μ) can now be defined so that

$$\textit{Total Service Cost} = C_s \mu^{n+1}.$$

Substituting m = n + 1, f(μ) becomes the following:

$$f(\mu) = \mu^m \qquad (m \leq 1).$$

Learning effects generally occur because of learning in the literal sense and/or innovation. For this reason, learning effects are more extreme in operations made up of more worker assembly than machine

---

[3] The learning index is the ratio of the logarithm of the learning rate to the logarithm of 2 and is, therefore, always less than or equal to 0.

time and in swiftly growing or improving industries.

The docking facility under evaluation here leans toward assembly versus machine time but is not subject to frequent innovation. The potential for the learning effect is present but not guaranteed.

In the succeeding chapters of this paper we will deal with the M/G/1 and the M/M/s queues, both with potential for learning effects, with the goal of developing a cost minimizing design.

Chapter 2 will derive the mean number of customers in the queue (the previously mentioned L) for both the queues. A proof of single server optimality in the M/M/s queue will also be presented and thus simplify the optimization of that queue. Chapter 3 will describe the geometric programming method used in the actual optimization. Chapter 4 will run through sample problems and offer verification of the optimization results. Chapter 5 will conclude the paper with some discussion and recommendations for further research.

# Chapter 2

# DERIVATIONS FROM QUEUING THEORY

This chapter will specialize the Hillier cost model for more particular queue systems beginning with the single server queue.

Recalling that the model is of the form

$$TEC = sC_s f(\mu) + C_w L,$$

the only variable dependent upon the queue type is L, the mean number of customers in the system. Since the goal is to minimize with respect to μ, the mean service rate, L must be expressed in terms of μ, λ, the mean arrival rate, and the variance of the service time distribution, b(t).

## Single Server Queues

For the single sever queue with Poisson input (i.e. exponentially distributed interarrival times), it is possible to derive an expression for L in terms of only these basic parameters. Let ρ represent the ratio of λ to μ which is referred to as the traffic intensity or utilization factor for a single server queue.

The number of customers waiting at any time, n+1, is dependent upon the number of customers waiting at the preceding time n. Consider the queue system just following a customer departure. Here $X_n$ indicates the number of customers waiting at time n and A represents the number of arrivals during the past service period of length t.

$$X_{n+1} = \begin{cases} X_n - 1 + A & X_n > 0 \\ A & X_n = 0 \end{cases},$$

The expression is easier to deal with algebraically when rewritten as:

$$X_{(n+1)} = X_n - U(X_n) + A$$
$$\textit{where} \quad U(X_n) = \begin{cases} 1 & X_n > 0 \\ 0 & X_n = 0 \end{cases} . \tag{2.1}$$

In the steady state situation, the system state, at time n, is essentially independent of the initial state of the system (time zero). The number of customers at any time n, is a random variable with a single mean and variance regardless of the value of n. That is,

$$E(X_{n+1}) = E(X_n) = L^{(d)}$$
$$\sigma^2_{X_{n+1}} = \sigma^2_{X_n} . \tag{2.2}$$

where the superscript on L indicates the evaluation takes place at the moment directly after a departure.

Taking the expected value of eq 2.1 and using eq. 2.2:

$$E[U(X_n)]=E(A).$$

The expected value of A can be calculated utilizing a Riemann-Stieltjes integral:

$$E(A) = \int_0^\infty E[A\,|\,T=t]\,\partial B(t) = \int_0^\infty \lambda t\,\partial B(t) = \lambda E[t] = \lambda/\mu = \rho.$$

Squaring eq. 2.1:

$$X^2_{(n+1)} = X^2_n + [U(X_n)]^2 + A^2 + 2AX_n - 2A[U(X_n)] - 2[U(X_n)]X_n. \tag{2.3}$$

Again, invoking the equilibrium conditions,

$$\begin{gathered} \sigma^2_{X_n} = \sigma^2_{X_{n+1}} \\ E(X^2_{n+1}) + E(X_{n+1})^2 = E(X^2_n) + E(X_n)^2 \\ \textit{then} \\ E(X^2_{n+1}) = E(X^2_n). \end{gathered}$$

Also, it is intuitively obvious that

$$[U(X_n)]^2 = U(X_n),$$

and

$$X_n[U(X_n)] = X_n.$$

Therefore, taking expected values of eq. 2.3, while assuming that the number of arrivals is independent of the number of customers in the system (i.e. A is independent of $X_n$), results in the following:

$$2L^{(D)} - 2L^{(D)}\rho = E(A^2) + \rho - 2\rho^2,$$

$$L^{(D)} = \frac{\rho - 2\rho^2 + E(A^2)}{2(1-\rho)}. \tag{2.4}$$

Of course,

$$E(A^2) = var(A) + E(A)^2 = var(A) + \rho^2. \tag{2.5}$$

But, recalling that t is the length of the preceding service period, the variance of A can be expressed as follows:

$$var(A) = E[var(A\,|\,T{=}t)] + var[E(A\,|\,T{=}t)]. \quad \text{(Parzen, p.55)}$$

Where,

$$E(A\,|\,T{=}t) = \sum_{0}^{\infty} \frac{a\,e^{-\lambda t}(\lambda t)^{a}}{a!} = \lambda t \sum_{1}^{\infty} \frac{e^{-\lambda t}(\lambda t)^{a-1}}{(a-1)!} = \lambda t,$$

and

$$\begin{aligned} E[A(A-1)\,|\,T{=}t] &= \sum_{0}^{\infty} \frac{(a)(a-1)e^{-\lambda t}(\lambda t)^{a}}{a!} \\ &= \lambda^2 t^2 \sum_{2}^{\infty} \frac{e^{-\lambda t}(\lambda t)^{a-2}}{(a-2)!} = \lambda^2 t^2. \end{aligned}$$

Then,

$$\begin{aligned} var(A\,|\,T=t) &= E[A(A-1)\,|\,T=t] + E(A\,|\,T=t) - [E(A\,|\,T=t)]^2 \\ &= \lambda^2 t^2 + \lambda t - \lambda^2 t^2 = \lambda t. \end{aligned}$$

So,

$$var(A) = E(\lambda t) + var(\lambda t) = \lambda E(t) + \lambda^2 var(t) = \rho + \lambda^2 var(t). \tag{2.6}$$

The variance of t is, in this case, the variance of the service times; that is, var(t) = var(b(t)).

Finally, combining eq.s 2.4, 2.5, and 2.6, we arrive at the Pollaczek-Khinchine formula:

$$\begin{aligned} L^{(D)} &= \frac{\rho - 2\rho^2 + \rho + \rho^2 + \lambda^2 var(t)}{2(1-\rho)} \\ &= \rho + \frac{\rho^2 + \lambda^2 var(t)}{2(1-\rho)}. \end{aligned}$$

It remains now to prove that the mean number of customers in the queue the moment after a served customer departs is equivalent to the mean number of customers in the queue at any given moment.

It is true for any Poisson process, that is, any queue with input corresponding to a Poisson distribution, that the mean number of customers immediately prior to a customer arrival is equivalent to the mean number of customers at any given moment (see Cooper p.57 for a concise proof). The hypothesis that the mean just after a departure is the same as the mean just prior to an arrival is logically equivalent to the hypothesis that $d_n$, the steady state probability of n customers in the system just after a departure, is equal to $a_n$, the steady state probability of n customers in the system just prior to an arrival.

For proof of this second hypothesis let $A_n(t)$ denote the number of arrivals, during the time interval (0,t), that increase the number of customers in the system from n to n+1, and $D_n(t)$ denote the number of departures on the same interval that decrease the number of customers in the system from n+1 to n. Then, considering a long interval of length T,

$$|A_n(T)-D_n(T)| \leq 1. \tag{2.7}$$

Since a steady state solution is assumed to exist (i.e. $\lambda/\mu < 1$), the limit of the ratio of the total number of departures in time T, D(T), to the total number of arrivals in time T, A(T), must approach 1 as T goes to infinity:

$$\lim_{T\to\infty}\frac{D(T)}{A(T)} = 1. \tag{2.8}$$

Dividing eq. 2.7 by A(T) and taking the limit as T approaches infinity:

$$\lim_{T\to\infty}\left|\frac{A_n(T)}{A(T)} - \frac{D_n(T)}{A(T)}\right| \leq \lim_{T\to\infty}\frac{1}{A(T)} = 0.$$

Then, using eq 2.8,

$$\begin{aligned}\lim_{T\to\infty}\frac{A_n(T)}{A(T)} &= \lim_{T\to\infty}\frac{D_n(T)}{A(T)}\\ &= \lim_{T\to\infty}\frac{D_n(T)}{A(T)}\lim_{T\to\infty}\frac{A(T)}{D(T)}\\ &= \lim_{T\to\infty}\frac{D_n(T)}{D(T)}.\end{aligned}$$

But,

$$\lim_{T\to\infty}\frac{A_n(T)}{A(T)} = a_n \qquad \lim_{T\to\infty}\frac{D_n(T)}{D(T)} = d_n.$$

(See Gross and Harris, p 235)

Hence, the probability of n customers in the system prior to an arrival is equal to the probability of n customers just after a departure. From this it follows that the expected number of customers, just prior to an arrival and just after a departure are equal. Finally, since the M/G/1 queue arrival times are Poisson distributed, $L^{(D)}$, the expected number of customers in the system after a departure, is equivalent to $L^{(A)}$, the expected number of customers in the system prior to an arrival, which is equivalent to L, the expected number of customers at any time.

There are, of course, many types of single server queues with Poisson input.

For an M/M/1 queue: $\mathrm{var}(s)=1/\mu^2$, and

$$L = \rho + \frac{\rho^2}{1-\rho} = \frac{\lambda}{\mu - \lambda}.$$

For an M/D/1 queue: var(s)=0, and

$$L = \rho + \frac{\rho^2}{2(1-\rho)} = \frac{2\rho - \rho^2}{2(1-\rho)}.$$

The M/G/1 results are, however, more convenient for use in general evaluations and calculations as will be performed in chapter 3.

## A Multi-server Queue

Deriving expressions for multi-server queues is much more difficult, and, for that reason, only the M/M/s queue (i.e. the queue with exponential interarrival times, exponential service times, and s parallel servers with a single queue) will be attacked here. To begin, the probability that there will be n customers waiting at time t + δt is the sum of five other probabilities:

(1) the probability of n customers at time t with no customers arriving or departing during time δt;

(2) the probability of n customers at time t with both an

arrival and a departure during time $\delta t$;

(3) the probability of n+1 customers at time t with one departure and no arrivals during $\delta t$;

(4) the probability of n-1 customers at time t with one arrival and no departures during $\delta t$;

(5) the probability of any other initial number of customers at time t requiring more than one departure or arrival during $\delta t$ in order to result in n customers at time $t+\delta t$.

Since the inter-arrival times and service times of this type of queue are given as exponentially distributed random variables with parameters $\lambda$ and $\mu$, respectively (where $\lambda$ still denotes the mean arrival rate and $\mu$ the mean service rate), the number of customers arriving or being served during any time interval of length t are random variables following Poisson distributions with parameters $\lambda t$ and $\mu t$. (for proof see Freund and Walpole p. 211). That is,

$$\mathit{Prob}(n \text{ arrivals on interval } \delta t) = \frac{(\lambda \delta t)^n \, e^{-\lambda \delta t}}{n!}.$$

Similarly,

$$Prob(m \text{ departures on interval } \delta t) = \frac{(\mu \delta t)^m \, e^{-\mu \delta t}}{m!}.$$

Using the MacLaurin's series expansion of $e^{-x\delta t}$:

$$e^{-x\delta t} = 1 - x\delta t + \frac{(x\delta t)^2}{2} - \frac{(x\delta t)^3}{6} + \dots + \frac{(-1)^n (x\delta t)^n}{n!} + \dots,$$

the probability of any number of arrivals or departures during the time interval δt can be derived. For notational consistency let n denote the number of arrivals and m the number of departures.

$$Prob(n=0) = 1 - \lambda \delta t + o(\delta t),$$

$$Prob(n=1) = \lambda \delta t + o(\delta t),$$

$$Prob(n>1) = o(\delta t),$$

$$Prob(m=0) = 1 - \mu \delta t + o(\delta t),$$

$$Prob(m=1) = \mu \delta t + o(\delta t),$$

$$Prob(m>1) = o(\delta t).$$

o(δt) represents a sum of terms involving greater than first power terms of δt. More precisely, o(δt) signifies any term or sum of terms f(δt) such that

$$\lim_{\delta t \to 0} \frac{f(\delta t)}{\delta t} = 0.$$

Note that as δt becomes small the probability of multiple arrivals (n > 1) or departures (m > 1) on a single interval becomes negligible.

Now letting $P_n(t)$ denote the probability of n customers at time t and using the probabilities above the following recursive expression can be written:

$$\begin{aligned} P_n(t+\delta t) &= P_n(t)[(1-\lambda\delta t+o(\delta t))(1-\mu\delta t+o(\delta t))] \\ &+P_n(t)[(\mu\delta t+o(\delta t))(\lambda\delta t+o(\delta t))] \\ &+P_{n-1}(t)[(\lambda\delta t+o(\delta t))(1-\mu\delta t+o(\delta t))] \\ &+P_{n+1}(t)[(\mu\delta t+o(\delta t))(1-\lambda\delta t+o(\delta t))] \\ &+\sum_{-\infty}^{-2} P_{n+i}(t)[o(\delta t)]+\sum_{2}^{\infty} P_{n+i}(t)[o(\delta t]. \end{aligned}$$

This equation yields different results when n = 0, n > s, and n < s. Taking the limit as $\delta t \to 0$ (at this point all $o(\delta t)$ terms will go to 0) generates the following three equations:

$$\begin{aligned} \frac{\partial P_0(t)}{\partial t} &= -\lambda P_0(t) + \mu P_1(t), \\ \frac{\partial P_n(t)}{\partial t} &= -(\lambda+n\mu)P_n(t) + \lambda P_{n-1}(t) + (n+1)\mu P_{n+1}(t) \qquad \text{for } 1 \le n \le s-1 , \\ \frac{\partial P_n(t)}{\partial t} &= -(\lambda+s\mu)P_n(t) + \lambda P_{n-1}(t) + s\mu P_{n+1}(t) \qquad \text{for } s \le n . \end{aligned}$$

As in the single server case, a steady state solution for L is what is desired. The steady state requires that the probability of any number of customers being in the queue be constant over time. That is,

$$\frac{\partial P_n(t)}{\partial t} = 0 .$$

Thus, at equilibrium,

$$\begin{aligned}
P_1 &= \frac{\lambda}{\mu} P_0 = pP_0, \\
\lambda P_{n-1} &= (\lambda + n\mu)P_n - (n+1)\mu P_{n+1} \qquad \text{for } 1 \le n \le s-1, \\
\lambda P_{n-1} &= (\lambda + s\mu)P_n - (s)\mu P_{n+1} \qquad \text{for } s \le n.
\end{aligned}$$

In these equations, however, p can no longer be considered the traffic intensity or utilization. That label is reserved for the value $\rho = \lambda/(\mu^*s)$. Here p is a direct substitution for the value $\lambda/\mu$, and has no other real significance.

From here it can be proven by induction that

$$\begin{aligned}
P_n &= \frac{1}{n!} p^n P_0 \qquad \text{for } n \le s , \\
P_n &= \frac{s^{s-n}}{s!} p^n P_0 \qquad \text{for } s < n .
\end{aligned}$$

Using the fact that the sum of the probabilities must be 1, $P_0$ can be derived,

$$P_0 = [\sum_0^{s-1} \frac{p^i}{i!} + \frac{p^s}{s!}(\frac{s}{s-p})]^{-1}$$

All that remains then is to calculate L, the expected value of n.

$$L = E(n) = \sum_{0}^{\infty} nP_n = p + \frac{P_0 p^{s+1}}{(s-1)!(s-p)^2} .$$

This equation, as well as those derived for the single server queues can now be inserted into the original cost model for any chosen value of s, under any learning effect condition, and subsequently, the queue can be optimized. However, during the coding and testing of the algorithm described in chapter 3, it was realized that for the M/M/s queue to be operating optimally, s must equal one. The proof of this lies in comparing the cost of the M/M/s queue with mean service rate of μ and the cost of the M/M/1 queue with mean service rate sμ. The two systems have equivalent servicing capability but very different costs.

Given:

$$\mathit{Cost}\ (M/M/1, s\mu) = C_s(\mu s)^m + C_w p + \frac{C_w p^2}{s(s-p)},$$

and

$$\mathit{Cost}\ (M/M/s, \mu) = C_s s \mu^m + C_w p + \frac{C_w p^{s+1}}{(s-1)!(s-p)^2 \sum_{0}^{s-1} \frac{p^i}{i!} + p^s(s-p)} .$$

Where s is an integer greater than zero, p is the ratio of λ to μ and is bounded between zero and s by the equilibrium condition, and m is a real number less than or equal to 1.

Then:

$$Cost\ (M/M/1, s\mu) \leq Cost\ (M/M/s, \mu)$$

implies

$$0 \geq Cost\ (M/M/1, s\mu) - Cost\ (M/M/s, \mu)\ .$$

Which, in turn, implies

$$0 \geq C_s \lambda^m p^{-m}(s^m - s) + C_w p(\frac{1}{s} - 1) + \frac{C_w p}{s-p}\left[\frac{p}{s} - \frac{p^s}{(s-1)!(s-p)^2 \sum_0^{s-1} \frac{p^i}{i!} + p^s}\right].$$

Which, again, implies

$$\frac{(s-p)C_s \lambda^m (s - s^m)}{C_w p^{m+1}} \geq 1 - s + p - \frac{p^s}{(s-1)!(s-p) \sum_0^{s-1} \frac{p^i}{i!} + p^s}\ .$$

When s is one, the two queues are the same, so the two costs are obviously equivalent. When s ≥ 2, p < 1 implies s > 1 + p. Knowing this, the left-hand side must always be positive, the right hand side always negative and the single server cost is always less than that of the multi-server queue.

This proof does assume there are no restrictions on μ other than that it must be strictly greater than $\lambda$. In the shipping problem studied here, $\mu$ is adjusted by increasing manpower, upgrading and improving machinery, increasing the number of accesses (ramps, cranes, etc.) to the ship being serviced, etc. In the case of a high ratio of waiting costs to service costs, the optimal mean service rate might climb so high as to be impossible with a single server in a **real** situation (i.e. a situation where $\mu$ is bounded).

The cost model is now queue specific and ready for optimization.

# Chapter 3

# GEOMETRIC PROGRAMMING AND THE SINGLE SERVER QUEUE

Now there are complete cost models in terms of the constants--service cost per customer per server, $C_s$, waiting cost per customer per unit time, $C_w$, mean arrival rate, $\lambda$, number of servers, and variance of service times, var(t)--and dependant upon a single variable, $\rho$, the ratio of the mean arrival rate to the mean service rate.

For the M/G/1 queue:

$$TEC = C_s \mu^m + C_w L_s + C_w L_w = C_s \lambda^m \rho^{-m} + C_w \rho + \frac{C_w \rho^2}{2(1-\rho)} + \frac{C_w \lambda^2 var(t)}{2(1-\rho)} ,$$

and for the M/M/s queue, which is optimal when s = 1 (by the preceding proof, pp. 27 and 28):

$$TEC = C_s s \mu^m + C_w L_s + C_w L_w = C_s \lambda^m \rho^{-m} + C_w \rho + \frac{C_w \rho^2}{(1-\rho)} .$$

The task at hand is to minimize these cost equations with respect to the ratio of mean arrival rate to mean service rate, $\rho$.

## Geometric Programming Overview

Geometric programming (GP) is a still evolving technique for minimization of linear and nonlinear systems. The technique, which was first studied by Duffin, Peterson, and Zener in the 1960's, is derived from Cauchy's arithmetic-geometric inequality and the fact that, at optimality, the inequality becomes an equality ( a more complete history of GP is contained in Thome (1988)).

GP revolves about four basic rules, which are described as follows by Woolsey and Swanson (1969):

Rule 1: The form of the optimal solution of any posynomial GP problem is:

$$TEC^* = \prod_{obj.funct.} (\mathit{Coefficient\ of\ term\ } i/\delta_i)$$

$$\times \prod_{all\,const.} \left[ \prod_{a\,const.} (\mathit{Coefficient\ of\ term\ } j/\delta_j) \times \left( \sum_{a\,const.} \delta_j \right)^{\sum_{a\,const.} \delta_j} \right],$$

where the asterisk indicates a value at optimality, i refers to a term in the objective function, and j refers to a term in a constraint.

Rule 2: The exponent matrix is constructed as follows:

Rule 2A: The sum of the contributions to cost in the objective function is one.

$$\sum_{obj.\ funct.} \delta_i = 1 .$$

Rule 2B: For each primal variable the equations in the exponent matrix are:

$$\sum_{obj. funct.} (\delta_i \ exponent of variable at term\ i)$$
$$+ \sum_{all const.} (\delta_j \ exponent of variable at term j) = 0 .$$

Rule 3: At optimality, for each term in the objective function

$$TEC^* = \frac{term\, i}{\delta_i^*} .$$

Rule 4: At optimality, for every term, in each constraint

$$\delta_k = term k \ ( \sum_{const.} \delta_j) .$$

These four rules (the first two of which construct the dual problem while the last two demonstrate relationships at optimality that facilitate easier solution) allow direct solution of simple problems.

Degree of difficulty (hereafter, DD) is defined as the number of terms in both the objective function and the constraints, minus the number of terms in the problem, minus one. Zero DD problems are easily solved using the four basic rules. Higher DD problems, however, require modification before they can be optimized.

Ratliff (1986) developed a method to optimize nonlinear, multi-variable, unconstrained posynomials. By holding all variables but one constant, condensing all like terms, and iterating until the variable settled on a value, then moving to another variable and cycling through all the variables until they all settled, Ratliff arrived at the optimum of high DD problems.

Later, Thome (1988) used the Greening technique to optimize a class of nonlinear, single variable, unconstrained signomials. This class of problems was restricted to terms with positive coefficients, or negative coefficients and positive exponents. By putting his problems into a form such that the objective function became a constraint on an additional variable and manipulating this constraint until the negative coefficient terms could be condensed with this new variable, he was able to condense the entire problem into a zero DD problem and optimize.

A third option for using GP on high DD problems is to select from the original problem some terms with which to build a lower DD, bounded sub-problem and solve that sub-problem to obtain bounds on the actual objective function and the variables.

If the M/G/1 cost model is reformulated to the standard GP form (where the objective function consists of terms with variables, exponents, and coefficients, versus terms with polynomial expressions, exponents, and coefficients), a substitution for the awkward denominator in the waiting cost must be made, creating a constraint.

$$TEC = C_s \lambda^m \rho^{-m} + C_w \rho + C_w \rho^2 \gamma^{-1} + C_w \gamma^{-1} \lambda^2 var(t)$$

$$s.t. \quad \rho + \gamma \leq 1 \, .$$

The problem now has six terms, two variables, and three degrees of difficulty. A similar reformulation can be found for the M/M/1 queue, with four terms, two variables, and one DD. The constraint eliminates the possibility of using either Thome's or Ratliff's method, so, perhaps bounding is the only option left (using GP) to evaluate this problem.

## Distorting GP to Solve the Single Server Queue

Up until now, condensation of terms has been performed only on like terms (i.e. on terms with the same sign exponent and the same sign coefficient). This is because condensation of like terms yields a like term, while condensation of dissimilar terms yields a term which is not wholly predictable. Single server cost models, however, can be reformulated so that condensation of the objective function yields a single term, for which the signs of the exponent and coefficient are predictable.

To arrive at a zero DD, single server cost problem, *all* the terms of the objective function must be condensed (since there are two variables and two terms in the constraint). That is, the problem must appear as follows:

$$TEC = k\rho^{\omega 1}\gamma^{\omega 2}$$
$$s.t. \quad \rho + \gamma \leq 1 \ .$$

For the above problem to balance and be bounded, both $\omega 1$ and $\omega 2$ must be less than zero. So, a form of the objective function must be found which condenses to the form above such that $\omega 1$ and $\omega 2$ are always negative in the solution space $0 < \rho$, $0 < \gamma$.

Through necessity, ingenuity, trial and error, etc., a form of the M/G/1 cost model fulfilling those requirements was found. Since

$$L = \frac{\rho^2}{2(1-\rho)} + \rho + \frac{var(t)\lambda^2}{2(1-\rho)} = \frac{\rho}{2} + \frac{1}{2\rho(1-\rho)} - \frac{1}{2\rho} - \frac{1}{2} + \frac{var(t)\lambda^2}{2(1-\rho)}$$
$$= .5\rho + .5\rho^{-1}\gamma^{-1} - .5\rho^{-1} - .5 + \frac{var(t)\lambda^2}{2(1-\rho)} \ ,$$

the objective function can be reformulated as

$$TEC = C_s\lambda^m\rho^{-m} + .5C_w\rho^{-1}\gamma^{-1} - .5C_w\rho^{-1} + .5C_w\rho + .5C_w\gamma^{-1}var(t)\lambda^2 - .5C_w\ .$$

Ignoring the constant, moving the signomial term to the left hand side, and condensing yields:

$$\left(\frac{TEC}{\delta_0}\right)^{\delta_0}\left(\frac{C_w}{2\rho\delta_1}\right)^{\delta_1} = \left(\frac{C_s\lambda^m}{\rho^m\delta_2}\right)^{\delta_2}\left(\frac{C_w}{2\rho\gamma\delta_3}\right)^{\delta_3}\left(\frac{C_w\rho}{2\delta_4}\right)^{\delta_4}\left(\frac{c_w var(t)\lambda^2}{2\gamma\delta_5}\right)^{\delta_5}$$

$$\begin{aligned} \textit{where}\quad \delta_0 &= \frac{TEC}{TEC+.5C_w\rho^{-1}} & \delta_1 &= \frac{.5C_w\rho^{-1}}{TEC+.5C_w\rho^{-1}} \\ \delta_2 &= \frac{C_s\lambda^m\rho^{-m}}{TEC+.5C_w\rho^{-1}} & \delta_3 &= \frac{.5C_w\gamma^{-1}\rho^{-1}}{TEC+.5C_w\rho^{-1}} \\ \delta_4 &= \frac{.5C_w\rho}{TEC+.5C_w\rho^{-1}} & \delta_5 &= \frac{.5C_w var(t)\lambda^2\gamma^{-1}}{TEC+.5C_w\rho^{-1}}\ . \end{aligned}$$

Then,

$$\begin{aligned} TEC &= k\rho^{\omega 1}\gamma^{\omega 2} \\ s.t.\quad \rho + \gamma &\le 1\ , \end{aligned} \tag{3.1}$$

$$\begin{aligned} \textit{where}\quad \omega 1 &= \frac{-m\delta_2 + \delta_4 + \delta_1 - \delta_3}{\delta_0} = \frac{-mC_s\lambda^m\rho^{-m} + .5C_w\left[\frac{-\rho^2 + (\rho-1)}{1-\rho}\right]}{TEC} \\ \omega 2 &= \frac{-(\delta_3 + \delta_5)}{\delta_0} = \frac{-(.5c_w\gamma^{-1}\rho^{-1} + .5C_w var(t)\lambda^2\gamma^{-1})}{TEC} \\ k &= \delta_0^{\delta_0}\left(\frac{2\delta_1}{C_w}\right)^{\frac{\delta_1}{\delta_0}}\left(\frac{C_s\lambda^m}{\delta_2}\right)^{\frac{\delta_2}{\delta_0}}\left(\frac{C_w}{2\delta_3}\right)^{\frac{\delta_3}{\delta_0}}\left(\frac{C_w}{2\delta_4}\right)^{\frac{\delta_4}{\delta_0}}\left(\frac{C_w var(t)\lambda^2}{2\delta_5}\right)^{\frac{\delta_5}{\delta_0}}\ . \end{aligned}$$

These definitions of $\omega 1$ and $\omega 2$ are obviously negative wherever $\rho$

and $\gamma$ are positive (recalling that if $\gamma$ is positive then $\rho$ is $< 1$), so the condensation was successful.

Now the four GP rules can be used to solve the problem directly.

By rule 2:

$$\begin{array}{llll} d_1 & & & = 1 \\ \omega 1 d_1 & & + d_3 & = 0 \\ \omega 2 d_1 & + d_2 & & = 0\ . \end{array}$$

$$So,\ \ \delta_1 = 1,\ \ \delta_2 = -\omega 1,\ \ \delta_3 = -\omega 2\ .$$

Then by rule 1:

$$TEC^* = k d_2^{-d_2} d_3^{-d_3} (d_2 + d_3)^{d_2 + d_3} = k(-\omega 2)^{\omega 2}(-\omega 1)^{\omega 1}(-\omega 2 - \omega 1)^{-\omega 2 - \omega 1}. \quad (3.2)$$

But by rule 4:

$$\begin{aligned} d_2 &= -\omega 2 = \gamma(d_2 + d_3) = -\gamma(\omega 1 + \omega 2) \\ d_3 &= -\omega 1 = \rho(d_2 + d_3) = -\rho(\omega 1 + \omega 2)\ . \end{aligned}$$

So,

$$\gamma = \frac{-\omega 2}{-\omega 1}\rho\ .$$

Substituting this for $\gamma$ in the original equation [3.1]:

$$TEC^* = k\rho^{\omega 1 + \omega 2}(-\omega 2)^{\omega 2}(-\omega 1)^{-\omega 2}\ . \quad (3.3)$$

Solving eq.s (3.2) and (3.3) together:

$$k(-\omega 2)^{\omega 2}(-\omega 1)^{\omega 1}(-\omega 2-\omega 1)^{-\omega 2-\omega 1} = k\rho^{\omega 1+\omega 2}(-\omega 2)^{\omega 2}(-\omega 1)^{-\omega 2}$$

$$\rho = \frac{\omega 1}{\omega 1 + \omega 2} .$$

Iteration through these equations, from any initial $\rho$, appears to converge quadratically to the optimum just as with Thome's and Ratliff's methods. Once $\rho^*$ has been found, $\mu^* = \lambda/\rho^*$ follows directly and the optimal mean service rate is determined.

The flow chart for this algorithm (appendix A) attempts to control round-off error by avoiding division. That is, since the value of $\rho$ derived finally from the above equations can be reduced to a ratio of terms from the objective function, the program simply defines five variables to be the five terms of the objective function, two variables to be the numerators of the $\omega$'s, and the new value of $\rho$ to be the ratio of the first numerator to the sum of the numerators. In this way, propagation of round-off error is avoided, while the final result for each iteration remains the same.

A similar derivation has been performed on the M/M/1 queue with the following results:

$$L = \frac{\rho}{1-\rho} = \frac{1}{\rho(1-\rho)} - \frac{1}{\rho} - 1$$

$$TEC = C_s\lambda^m\rho^{-m} + \frac{C_w}{\rho\gamma} - \frac{C_w}{\rho} - C_w \qquad s.t.\ \ \rho + \gamma \le 1$$

$$TEC = k\rho^{\omega1}\gamma^{\omega2}$$

$$\text{where} \quad \omega1 = \frac{-mC_s\lambda^m\rho^{-m} - C_w(\frac{1}{1-p})}{TEC} \qquad \omega2 = \frac{-C_w\rho^{-1}\gamma^{-1}}{TEC}$$

$$\rho = \frac{\omega1}{\omega1 + \omega2} \ .$$

Again, $\omega1$ and $\omega2$ are negative wherever $\gamma$ and $\rho$ are positive so the condensation is successful and will converge on the minimum. The flow chart for this algorithm (appendix A) also uses the method discussed above to avoid round-off error.

Sample problems, for both algorithms, are offered in chapter 4.

## Chapter 4

# SAMPLE RUNS OF THE ALGORITHM AND EXPLORATION OF THE QUEUE MODEL

It is not the intent of this paper to present more than empirical proofs of either the geometric programming algorithms' optimality or their rate of convergence. To demonstrate both the success of the optimization method and interesting aspects of the two models, five example problems for each of the two queue types, M/G/1 and M/M/1, (see Table 4.1) have been chosen.

Each of these models was run with nine different starting points on the open interval (0,1) ($\rho$ is a probability, and as such is restricted to the closed interval [0,1]; the endpoints, zero and one, lead to division by zero error, so the open interval was used).

The iteration criterion was the absolute value of the difference of the two most recent $\rho$'s calculated. If this absolute value exceeded $10^{-5}$, the iteration would continue.

Verification of the algorithms' results, in the form of spreadsheets of total expected cost on the interval $0 < \rho < 1$, is found in appendix B.

Table 4.1: Example Problems Used for Analysis

| Problem | Mean Arrival Rate $\lambda$ | Cost of Waiting $C_w$ | Cost of Service $C_s$ | Learning Rate m | Variance var(t) |
|---|---|---|---|---|---|
| M/M/1 | | | | | |
| A | 2 | 5 | 10 | 1 | |
| B | 2 | 5 | 10 | .87 | |
| C | 1 | 20 | .5 | .81 | |
| D | 1 | .5 | 20 | .81 | |
| E | 2 | 1 | 3 | .93 | |
| M/G/1 | | | | | |
| A | 1 | 4 | 4 | .95 | .7 |
| B | 4 | 1 | 20 | .97 | .7 |
| C | 4 | 20 | 1 | .97 | .7 |
| D | 2 | 1 | 3 | .93 | 0 |
| E | 2 | 1 | 3 | .93 | .1 |

## The M/M/1 Examples

Since the M/M/1 queue is actually a special case of the M/G/1, these problems will be discussed first. The five M/M/1 problems were chosen to illustrate the effects of learning on the total cost and the responses (in cost) to extreme differences between service cost per customer and waiting cost per customer.

The first problem (M/M/1A in table 4.1) is one with no learning effect and is modelled by the following TEC equation:

$$TEC = \frac{20}{\rho} + \frac{5\rho}{1-\rho} .$$

This problem is simple enough to solve with calculus:

$$\frac{\delta TEC}{\delta \rho} = \frac{-20}{\rho^2} + \frac{5}{1-\rho} + \frac{5\rho}{(1-\rho)^2} = 0$$

implies

$$1.5\rho^2 - 4\rho + 2 = 0 .$$

Then, by the quadratic formula:

$$\rho = \frac{2}{3} .$$

This was verified by the algorithm (or vice versa), as seen in table

4.2. Convergence (to five decimals in ρ and eight decimals in the total expected cost), for all nine starting values, was achieved in ten to twelve iterations.

**Table 4.2: Results of Problem M/M/1A**

| Initial ρ | Final ρ | μ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00001 | .666669 | 2.99999 | 40.0000000013 | 12 |
| .001 | .666669 | 2.99999 | 40.0000000013 | 12 |
| 1 | .666660 | 3.00003 | 40.0000000092 | 11 |
| .25 | .666666 | 3.00002 | 40.0000000059 | 11 |
| .5 | .666664 | 3.00001 | 40.0000000013 | 11 |
| .75 | .666662 | 3.00002 | 40.0000000041 | 10 |
| .9 | .666671 | 2.99998 | 40.0000000046 | 11 |
| .999 | .666674 | 2.99997 | 40.0000000114 | 11 |
| .99999 | .666664 | 3.00001 | 40.0000000013 | 12 |

Problem M/M/1B adds an 87% learning rate to the first problem resulting in the following TEC equation:

$$TEC = \frac{34.82}{\rho^{-.8}} + \frac{5\rho}{1-\rho} .$$

Intuitively, learning effects should decrease total costs, as explained in chapter 1. The GP results reflect this (see Table 4.3), converging, in seven or eight iterations from all starting points, on a TEC nearly eighteen percent less than the optimal total expected cost without learning.

**Table 4.3: Results of Problem M/M/1B**

| Initial ρ | Final ρ | μ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00001 | .613828 | 3.25824 | 33.6744500266 | 8 |
| .001 | .613829 | 3.25824 | 33.6744500250 | 8 |
| .1 | .613832 | 3.25822 | 33.6744500221 | 7 |
| .25 | .613828 | 3.25824 | 33.6744500271 | 7 |
| .5 | .613831 | 3.25822 | 33.6744500228 | 7 |
| .75 | .613839 | 3.25818 | 33.6744500251 | 7 |
| .9 | .613832 | 3.25822 | 33.6744500219 | 8 |
| .999 | .613832 | 3.25823 | 33.6744500228 | 8 |
| .99999 | .613831 | 3.25823 | 33.6744500228 | 8 |

Problems M/M/1C and M/M/1D were chosen to demonstrate the stronger effect on total cost of service cost per customer than waiting cost per customer.

M/M/1C has an extremely high waiting cost (relative to service cost) of 20 monetary units per customer per unit time, resulting in the following cost equation:

$$TEC = \frac{1}{2\rho^{-.7}} + \frac{20\rho}{1-\rho} .$$

and yet the optimal total cost is only 4.67 (see Table 4.4).

**Table 4.4: Results of Problem M/M/1C**

| initial $\rho$ | Final $\rho$ | $\mu$ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00001 | .083760 | 11.97949 | 4.6652327717 | 56 |
| .001 | .083479 | 11.97903 | 4.6652326570 | 55 |
| .1 | .083608 | 11.96061 | 4.6652325825 | 36 |
| .25 | .083613 | 11.95982 | 4.6652327779 | 46 |
| .5 | .083604 | 11.96109 | 4.6652324717 | 49 |
| .75 | .083611 | 11.96015 | 4.6652326940 | 49 |
| .9 | .083613 | 11.95982 | 4.6652327784 | 49 |
| .999 | .083604 | 11.96109 | 4.6652324714 | 50 |
| .99999 | .083604 | 11.96109 | 4.6652324716 | 50 |

M/M/1D has a service cost of 20 monetary units per customer and total cost equation as below:

$$TEC = \frac{20}{\rho^{-.7}} + \frac{\rho}{2(1-\rho)} .$$

Which leads to a optimal total cost of approximately 25.22 (see Table 4.5). This indicates that the optimal TEC is much more sensitive to the

**Table 4.5: Results of Problem M/M/1D**

| Initial ρ | Final ρ | μ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00001 | .837471 | 1.19407 | 25.2202700662 | 25 |
| .001 | .837361 | 1.19409 | 25.2202700670 | 22 |
| .1 | .837472 | 1.19407 | 25.2202700664 | 24 |
| .25 | .837472 | 1.19407 | 25.2202700672 | 24 |
| .5 | .837471 | 1.19407 | 25.2202700659 | 24 |
| .75 | .837471 | 1.19407 | 25.2202700661 | 22 |
| .9 | .837461 | 1.19409 | 25.2202700665 | 22 |
| .999 | .837471 | 1.19407 | 25.2202700659 | 25 |
| .99999 | .837471 | 1.19407 | 25.2202700659 | 25 |

service cost per customer than to waiting cost per customer.

The final example of this type problem, M/M/1E, may be viewed as a typical problem reflecting the original shipping model discussed. In this example, the waiting cost per customer is one third the service cost per customer. This reflects the nature of the shipping dock since the ships (customers) are owned by the dock operator, implying the waiting cost is due only to the ships' lost hours of productivity. The ninety percent learning rate is reasonable in such a labor intensive operation, and, if the unit of time is considered a day or work week, two arrivals during this time is appropriate (per week if the ships are large, per day if small).

Given the parameters as in table 4.1, the resultant model is of the form:

$$TEC = \frac{5.41}{\rho^{-.85}} + \frac{\rho}{1-\rho} .$$

As illustrated in table 4.6, the optimal ratio of mean arrival rate to mean service rate was found to be approximately .68, with a minimum total expected cost of operation of approximately 10.04 monetary units per time unit.

**Table 4.6: Results of Problem M/M/1E**

| Initial $\rho$ | Final $\rho$ | $\mu$ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00002 | .687801 | 2.90782 | 10.0433680145 | 11 |
| .001 | .687802 | 2.90781 | 10.0433680153 | 10 |
| .1 | .687798 | 2.90783 | 10.0433680134 | 12 |
| .25 | .687798 | 2.90783 | 10.0433680133 | 12 |
| .5 | .687790 | 2.90787 | 10.0433680140 | 11 |
| .75 | .687789 | 2.90787 | 10.0433680144 | 10 |
| .9 | .687792 | 2.90786 | 10.0433680133 | 12 |
| .999 | .687790 | 2.90787 | 10.0433680140 | 12 |
| .99999 | .687790 | 2.90787 | 10.0433680140 | 12 |

## The M/G/1 Examples

The M/G/1 problems were chosen to demonstrate the effects of interarrival time variance (previously referred to as var(t)) on the optimal total expected cost, and again, the strong dependence of the optimal total cost on the service cost per customer.

M/G/1A is of the form:

$$TEC = \frac{4}{\rho^{-.95}} + \frac{4(\rho^2 + .7)}{2(1-p)} .$$

This model was chosen because of the equality of the service and waiting costs per customer. The optimal ratio of mean arrival rate to mean service rate is approximately .49 (see table 4.7). This result is interesting to compare to the results of problems M/G/1B and M/G/1C (see tables 4.8 and 4.9, respectively).

Example M/G/1B is similar to M/M/1D in that the service cost per customer is considerably greater than the waiting cost per customer. The cost equation is of the form:

$$TEC = \frac{76.7}{\rho^{-.97}} + \frac{\rho + .7}{2(1-\rho)} .$$

Table 4.7: Results of Problem M/G/1A

| Initial ρ | Final ρ | μ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00001 | .486953 | 2.05359 | 13.4118065116 | 10 |
| .001 | .486952 | 2.05356 | 13.4118065120 | 10 |
| .1 | .486953 | 2.05358 | 13.4118065116 | 11 |
| .25 | .486961 | 2.05355 | 13.4118065129 | 10 |
| .5 | .486952 | 2.05359 | 13.4118065119 | 8 |
| .75 | .486950 | 2.05360 | 13.4118065132 | 10 |
| .9 | .486958 | 2.05356 | 13.4118065116 | 11 |
| .999 | .486959 | 2.05356 | 13.4118065117 | 11 |
| .99999 | .486959 | 2.05356 | 13.4118065117 | 11 |

Table 4.8: Results of Problem M/G/1B

| Initial ρ | Final ρ | μ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00001 | .771701 | 5.18330 | 122.0847497261 | 192 |
| .001 | .771698 | 5.18337 | 122.0847497247 | 193 |
| .1 | .771708 | 5.18331 | 122.0847497246 | 192 |
| .25 | .771698 | 5.18337 | 122.0847497258 | 189 |
| .5 | .771708 | 5.18330 | 122.0847497254 | 182 |
| .75 | .771708 | 5.18330 | 122.0847497254 | 144 |
| .9 | .771708 | 5.18331 | 122.0847497253 | 179 |
| .999 | .771708 | 5.18331 | 122.0847497250 | 193 |
| .99999 | .771708 | 5.18331 | 122.0847497252 | 193 |

Table 4.9: Results of Problem M/G/1C

| Initial ρ | Final ρ | μ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00001 | .13603 | 29.40557 | 157.4004681453 | 11 |
| .001 | .13602 | 29.40815 | 157.4004681084 | 10 |
| .1 | .13603 | 29.40634 | 157.4004680913 | 11 |
| .25 | .13602 | 29.40648 | 157.4004680855 | 11 |
| .5 | .13603 | 29.40577 | 157.4004681280 | 11 |
| .75 | .13603 | 29.40557 | 157.4004681452 | 11 |
| .9 | .13063 | 29.40564 | 157.4004681389 | 11 |
| .999 | .13603 | 29.40575 | 157.4004681295 | 11 |
| .99999 | .13603 | 29.40575 | 157.4004681294 | 11 |

Models M/G/1C and M/M/1C are similar in their high waiting costs per customer relative to service costs per customer. Model M/G/1C has a total expected cost equation of:

$$TEC = \frac{3.84}{\rho^{-.97}} + \frac{20(\rho + .7)}{2(1-\rho)} .$$

The results of these two models illustrate the same effects seen in the M/M/1 examples: when the service cost per customer is high relative to the waiting cost per customer, the optimal ratio of mean arrival rate to mean service rate is high (In the case of M/G/B, $\rho \approx .77$ and for M/M/1D, $\rho \approx .84$); when the service cost per customer is equal, or nearly equal, to the waiting cost per customer, the optimal ratio will tend to one half (as in M/G/1A); and when the service cost is relatively small, the optimal ratio will be small (For M/G/1C, $\rho \approx .14$ and for M/M/1C, $\rho \approx .08$).

Problems M/G/1D and M/G/1E are two other versions of the typical shipping example introduced as M/M/1E. M/G/1D is a deterministic model of the form:

$$TEC = \frac{5.60}{\rho^{-.9}} + \frac{\rho^2}{2(1-\rho)} .$$

M/G/1E is the probabilistic version of the problem, assuming a variance of .1, so that the model becomes:

$$TEC = \frac{5.60}{\rho^{-.9}} + \frac{\rho^2+.1}{2(1-\rho)} .$$

As might be expected, the optimal total expected cost of the deterministic model (see Table 4.10) is less than that of the probabilistic model (see Table 4.11). The exponential service model of this problem (M/M/1E) is even slightly more expensive because the variance of exponential service is the inverse of the square of the mean; in this case slightly less than .12.

Beyond the interesting aspects of each of these ten particular models, it is important to note the convergence of the algorithms. Not

Table 4.10: Results of Problem M/G/1D

| Initial ρ | Final ρ | μ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00001 | .752333 | 2.65840 | 9.1273515668 | 13 |
| .001 | .752335 | 2.65839 | 9.1273515661 | 17 |
| .1 | .752333 | 2.65840 | 9.1273515671 | 17 |
| .25 | .752333 | 2.65840 | 9.1273515669 | 17 |
| .5 | .752335 | 2.65839 | 9.1273515662 | 17 |
| .75 | .752344 | 2.65836 | 9.1273515668 | 10 |
| .9 | .752342 | 2.65837 | 9.1273515662 | 17 |
| .999 | .752335 | 2.65839 | 9.1273515662 | 18 |
| .99999 | .752335 | 2.65839 | 9.1273515662 | 18 |

Table 4.11: Results of Problem M/G/1E

| Initial ρ | Final ρ | μ | Total Exp. Cost | Iterations |
|---|---|---|---|---|
| .00001 | .719739 | 2.77879 | 9.8840006268 | 18 |
| .001 | .719738 | 2.77878 | 9.8840006267 | 20 |
| .1 | .719740 | 2.77882 | 9.8840006264 | 21 |
| .25 | .719728 | 2.77878 | 9.8840006273 | 20 |
| .5 | .719728 | 2.77883 | 9.8840006275 | 19 |
| .75 | .719729 | 2.77883 | 9.8840006275 | 16 |
| .9 | .719738 | 2.77879 | 9.8840006275 | 20 |
| .999 | .719738 | 2.77879 | 9.8840006267 | 21 |
| .9999 | .719738 | 2.77879 | 9.8840006267 | 21 |

only was the optimal cost calculated to extreme precision but, this result was achieved, with only two exceptions, in less than 25 iterations.

## Chapter 5

# CONCLUDING STATEMENTS AND RECOMMENDATIONS FOR FURTHER RESEARCH

This paper has utilized a method of geometric programming to approximate the optimal service rate of two classes of queues: the M/G/1 and M/M/s queues with linear waiting costs and server learning effects. It should be noted that no claim of global optimality has been made for this algorithm, or, as of yet, for any geometric condensation algorithm.

This is somewhat of a divergence from the current avenues of study in geometric programming. Most recent work has centered on solution, via condensation of like terms, of unconstrained problems. Here a class of constrained problems, which proved solvable using condensation of unlike terms, was introduced. This method may prove applicable to more unconstrained, single variable problems with polynomial denominators (as the queuing problems were before the substitution and the resulting constraint).

This model and algorithm are applicable to the hypothetical shipping problem initially discussed; however, it is much more common that the mean service rate be constrained (it was assumed here that it was not). A simple method for approximation of the optimum of a bounded problem such as this could be as follows:

1. Solve the unbounded problem with the geometric programming condensation method of this paper.
2. If the resultant mean service rate is not within the bounds then one of the bounds is the estimated optimum.
3. If the resultant mean service rate is within the bounds then it is the estimated optimum.

This will hold for the single server case but not for the multi-server case. That is, it may prove less costly to use multiple servers if the mean service rate is bounded. Analysis of when the multi-server queue is more economical is certainly worthy of more research.

A proof was presented for single server optimality of the M/M/s queue, which simplified the optimization method significantly. Intuition might assert that the M/G/s is also subject to single server optimality. A straight forward expression for the expected number of customers in the M/G/s queue system was not discovered in the research for this paper, but through stochastic methods similar to those of Stidham

(1968), single server optimality may be provable. Such proof would certainly extend the value of the GP condensation presented here.

The sensitivity of the optimal TEC to the service cost per customer also deserves further investigation. Sensitivity analysis with regard to changes in the service cost per customer and the waiting cost per customer is possible with this GP method (referring back to eq. 3.1, k and $\omega 1$ are dependent on the service cost per customer and k, $\omega 1$, and $\omega 2$ are dependent upon the waiting cost per customer), but intricate.

As a final note on queue design: the literature search for this paper indicated that design is a nearly dead topic and all interest is in queue control. As was, hopefully, demonstrated with this paper, there is much more to explore in the field of queue design, be it other methods of optimization, analysis of other queue types, or even queuing networks.

## REFERENCES CITED

Andress, F. J. 1954. The learning curve as a production tool. Harvard Business Review 32 (1):87-97.

Bell, C. E. 1971. Characterization and computation of optimal policies for operating an M/G/1 queuing system with removable server. Operations Research 19 (1):208-218.

Bell, C. E. 1980. Optimal operation of an M/M/2 queue with removable servers. Operations Research 28(5):1189-1204.

Chelst, K., A. Z. Tilles, and J. S. Pipis. 1981. A coal unloader:a finite queuing system with breakdowns. Interfaces 11(5):12-25.

Cooper, R. B. 1981. Introducing queuing theory. New York: Elsevier North Holland, Inc.

Crabill, T. B. 1972. Optimal control of a service facility with variable exponential service times and constant arrival rate. Management Science 18(9):560-566

Duffin, R., E. Peterson,and C. Zener. 1967. Geometric Programming. New York: John Wiley & Sons.

Freund, J. E. and R. E. Walpole. 1964. Mathematical Statistics. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Gross, D. and C. M. Harris. 1974. Fundamentals of Queuing Theory. New York: John Wiley & Sons.

Heyman, D. P. 1968. Optimal operating policies for M/G/1 queuing systems. Operations Research 16(2):362-372.

Heyman, D. P. and K. T. Marshall. 1968. Bounds on the optimal operating policy for a class of single-server queues. Operations Research 16(6):1138-1146.

Hillier, F. S. 1963. Economic models for industrial waiting line problems. Management Science 10(1):119-130.

Hillier, F. S. 1964. The application of waiting line theory to industrial problems. Journal of Industrial Engineering 15:3-8.

Hillier, F. S. and G. J. Lieberman. 1986. Introduction to Operations Research. 4th. ed. Oakland, CA: Holden-day, Inc.

Mangelsdorf, T. M. 1959. Waiting line theory applied to manufacturing problems. Chapter 5 in Analyses of industrial operations. Ed.s E. H. Bowman and R. B. Fetter. Homewood, IL: Richard D. Irwin

Parzen, E. 1962. Stochastic Processes. Oakland, CA: Holden-Day, Inc.

Ratliff, R. M. 1986. A generalized condensation algorithm for the solution of unconstrained balanced posynomial problems using geometric programming. Master's thesis, Colorado School of Mines, Golden, CO.

Schiller, D. H. and M. M. Lavin. 1956. The determination of requirements for warehouse dock facilities. Operations Research 4:231-243.

Shelton, J. R. 1960. Solution methods for waiting line problems. Journal of Industrial Engineering 12:293-303.

Stidham Jr., S. 1968. Static decision models for queuing systems with nonlinear waiting costs. Technical Reports 9[Nonr 225(89)], 112[Nonr 225(53)], and 1(NSF GL-2925). Stanford University, Palo Alto, CA.

Stidham Jr., S. 1970. On the optimality of single server queuing systems. Operations Research 18(4):708-732.

Szarkowicz, D. S. and T. W. Knowles. 1975. Optimal control of an M/M/s queuing systems. Operations Research 23(4):644-660.

Thome, J. J. 1988. An algorithm for solving a class of nonlinear, unconstrained, signomial economic models using the greening technique. Doctoral thesis, Colorado School of Mines, Golden, CO.

Van Voorhis, W. R. 1956. Waiting-line theory as a management tool. Operations Research 4:221-231.

Winston, W. L. 1987. Operations research: Applications and algorithms. Boston, MA: Duxbury Press.

Woolsey, R. E. D. and H. S. Swanson. 1969. Operations research for immediate application: A quick and dirty manual. New York: Harper and Row Publishers.

Yadin, M. and P. Naor. 1967. On queuing systems with variable service capacities. Naval Research Logistics Quarterly 14(1):43-54.

Yadin, M. and P. Naor. 1963. Queuing systems with a removable service station. Operational Research Quarterly 14(4):393-405.

Yelle, L. E. 1981. The learning curve: A historical review and comprehensive survey. Management Science 10:302-328

# APPENDIX A: FLOW CHARTS OF GEOMETRIC PROGRAMMING CONDENSATION ALGORITHMS

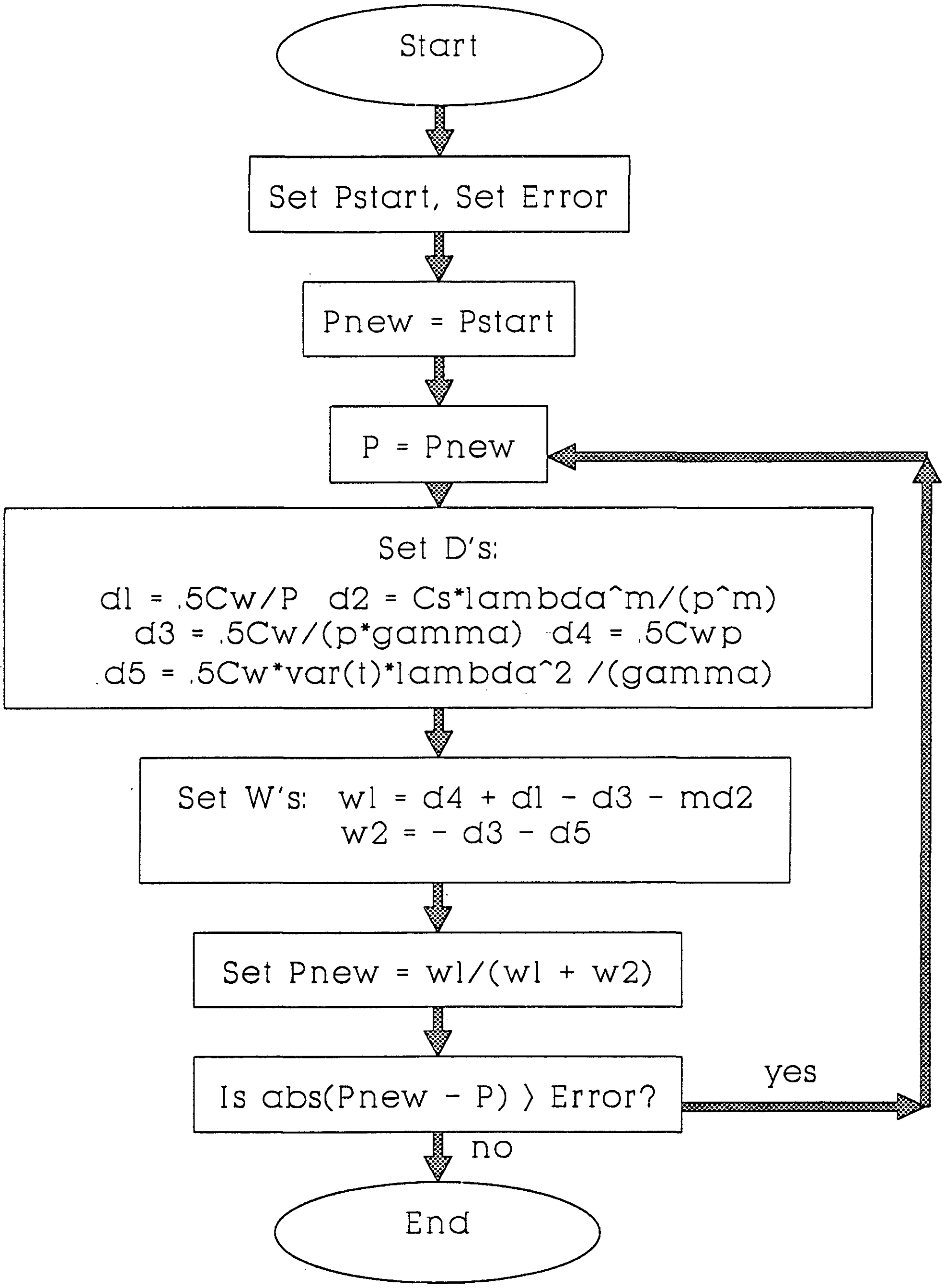

**Figure A.1: Flow chart for GP condensation of M/G/1 Queue**

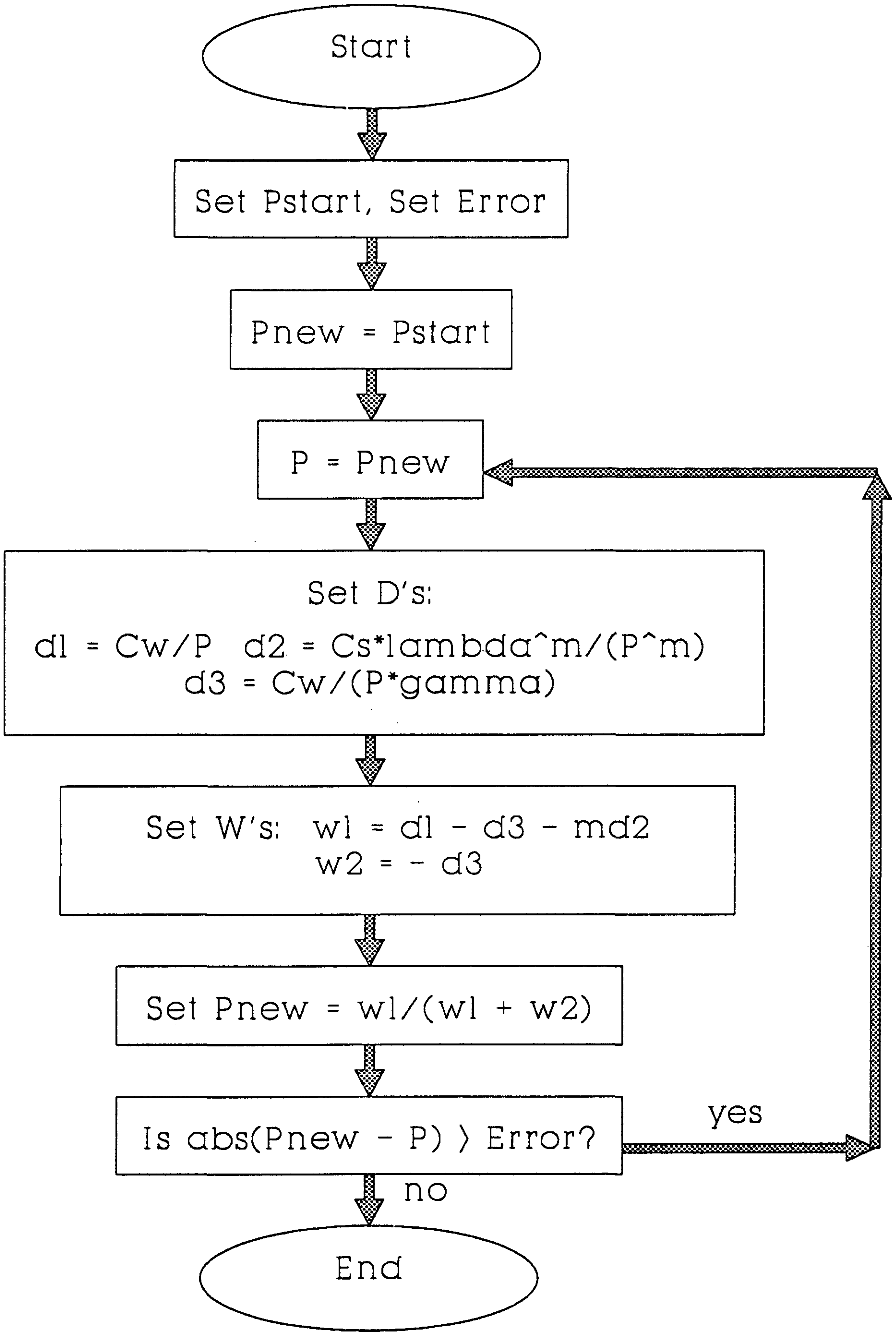

Figure A.2: Flow chart for GP condensation of M/M/1 Queue

# APPENDIX B: SPREADSHEET VERIFICATION OF GP CONDENSATION ALGORITHM RESULTS

**Table B.1: Total Expected Cost of Problem M/M/1A**

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 200.56 | 20.00000 |
| 0.2 | 101.25 | 10.00000 |
| 0.3 | 68.81 | 6.66667 |
| 0.4 | 53.33 | 5.00000 |
| 0.5 | 45 | 4.00000 |
| 0.6 | 40.83 | 3.33333 |
| 0.7 | 40.24 | 2.85714 |
| 0.8 | 45 | 2.50000 |
| 0.9 | 67.22 | 2.22222 |

| ρ | TEC | μ |
|---|---|---|
| 0.61 | 40.6074 | 3.27869 |
| 0.62 | 40.4160 | 3.22581 |
| 0.63 | 40.2596 | 3.17460 |
| 0.64 | 40.1389 | 3.12500 |
| 0.65 | 40.0550 | 3.07692 |
| 0.66 | 40.0089 | 3.03030 |
| 0.67 | 40.0023 | 2.98507 |
| 0.68 | 40.0368 | 2.94118 |
| 0.69 | 40.1145 | 2.89855 |

| ρ | TEC | μ |
|---|---|---|
| 0.661 | 40.00645 | 3.02572 |
| 0.662 | 40.00438 | 3.02115 |
| 0.663 | 40.00271 | 3.01659 |
| 0.664 | 40.00143 | 3.01205 |
| 0.665 | 40.00056 | 3.00752 |
| 0.666 | 40.00009 | 3.00300 |
| 0.667 | 40.00002 | 2.99850 |
| 0.668 | 40.00036 | 2.99401 |
| 0.669 | 40.00111 | 2.98954 |

| ρ | TEC | μ |
|---|---|---|
| 0.6661 | 40.00006 | 3.00255 |
| 0.6662 | 40.00004 | 3.00210 |
| 0.6663 | 40.00003 | 3.00165 |
| 0.6664 | 40.00001 | 3.00120 |
| 0.6665 | 40.00001 | 3.00075 |
| 0.6666 | 40.00000 | 3.00030 |
| 0.6667 | 40.00000 | 2.99985 |
| 0.6668 | 40.00000 | 2.99940 |
| 0.6669 | 40.00001 | 2.99895 |

Optimum is approx. $\rho = .6667$, TEC = 40, and $\mu = 2.9999$.

**Table B.2: Total Expected Cost of Problem M/M/1B**

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 110.41 | 20.00000 |
| 0.2 | 64.35 | 10.00000 |
| 0.3 | 47.76 | 6.66667 |
| 0.4 | 39.57 | 5.00000 |
| 0.5 | 35.31 | 4.00000 |
| 0.6 | 33.70 | 3.33333 |
| 0.7 | 34.83 | 2.85714 |
| 0.8 | 40.81 | 2.50000 |
| 0.9 | 63.94 | 2.22222 |

| ρ | TEC | μ |
|---|---|---|
| 0.61 | 33.67644 | 3.27869 |
| 0.62 | 33.67966 | 3.22581 |
| 0.63 | 33.71067 | 3.17460 |
| 0.64 | 33.77059 | 3.12500 |
| 0.65 | 33.8607 | 3.07692 |
| 0.66 | 33.98254 | 3.03030 |
| 0.67 | 34.13786 | 2.98507 |
| 0.68 | 34.32874 | 2.94118 |
| 0.69 | 34.55755 | 2.89855 |

| ρ | TEC | μ |
|---|---|---|
| 0.611 | 33.67554 | 3.27332 |
| 0.612 | 33.67491 | 3.26797 |
| 0.613 | 33.67454 | 3.26264 |
| 0.614 | 33.67445 | 3.25733 |
| 0.615 | 33.67464 | 3.25203 |
| 0.616 | 33.67509 | 3.24675 |
| 0.617 | 33.67582 | 3.24149 |
| 0.618 | 33.67682 | 3.23625 |
| 0.619 | 33.6781 | 3.23102 |

| ρ | TEC | μ |
|---|---|---|
| 0.6131 | 33.67452 | 3.26211 |
| 0.6132 | 33.6745 | 3.26158 |
| 0.6133 | 33.67449 | 3.26105 |
| 0.6134 | 33.67448 | 3.26052 |
| 0.6135 | 33.67447 | 3.25998 |
| 0.6136 | 33.67446 | 3.25945 |
| 0.6137 | 33.67445 | 3.25892 |
| 0.6138 | 33.67445 | 3.25839 |
| 0.6139 | 33.67445 | 3.25786 |

Optimum is approx. ρ = .6138, TEC = 33.67445, and μ = 3.25839.

**Table B.3: Total Expected Cost of Problem M/M/1C**

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 4.73 | 10 |
| 0.2 | 6.54 | 5 |
| 0.3 | 9.73 | 3.33 |
| 0.4 | 14.28 | 2.5 |
| 0.5 | 20.81 | 2 |
| 0.6 | 30.71 | 1.67 |
| 0.7 | 47.31 | 1.43 |
| 0.8 | 80.58 | 1.25 |
| 0.9 | 180.54 | 1.11 |

| ρ | TEC | μ |
|---|---|---|
| 0.01 | 12.76145 | 100 |
| 0.02 | 8.139401 | 50 |
| 0.03 | 6.43939 | 33.33333 |
| 0.04 | 5.592468 | 25 |
| 0.05 | 5.123537 | 20 |
| 0.06 | 4.859739 | 16.66667 |
| 0.07 | 4.722008 | 14.28571 |
| 0.08 | 4.668722 | 12.5 |
| 0.09 | 4.675763 | 11.11111 |

| ρ | TEC | μ |
|---|---|---|
| 0.081 | 4.6670 | 12.34568 |
| 0.082 | 4.6659 | 12.19512 |
| 0.083 | 4.6653 | 12.04819 |
| 0.084 | 4.6653 | 11.90476 |
| 0.085 | 4.6658 | 11.76471 |
| 0.086 | 4.6668 | 11.62791 |
| 0.087 | 4.6683 | 11.49425 |
| 0.088 | 4.6703 | 11.36364 |
| 0.089 | 4.6728 | 11.23596 |

| ρ | TEC | μ |
|---|---|---|
| 0.0831 | 4.665285 | 12.03369 |
| 0.0832 | 4.665263 | 12.01923 |
| 0.0833 | 4.665248 | 12.0048 |
| 0.0834 | 4.665237 | 11.99041 |
| 0.0835 | 4.665232 | 11.97605 |
| 0.0836 | 4.665232 | 11.96172 |
| 0.0837 | 4.665238 | 11.94743 |
| 0.0838 | 4.665249 | 11.93317 |
| 0.0839 | 4.665265 | 11.91895 |

Optimum is approx. $\rho = .08355$, TEC = 4.665232, and $\mu = 11.9689$.

**Table B.4: Total Expected Cost of Problem M/M/1D**

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 100.29 | 10 |
| 0.2 | 61.83 | 5 |
| 0.3 | 46.67 | 3.333333 |
| 0.4 | 38.32 | 2.5 |
| 0.5 | 32.99 | 2 |
| 0.6 | 29.35 | 1.666667 |
| 0.7 | 26.84 | 1.428571 |
| 0.8 | 25.38 | 1.25 |
| 0.9 | 26.03 | 1.111111 |

| ρ | TEC | μ |
|---|---|---|
| 0.81 | 25.31035 | 1.234568 |
| 0.82 | 25.25832 | 1.219512 |
| 0.83 | 25.22756 | 1.204819 |
| 0.84 | 25.22115 | 1.190476 |
| 0.85 | 25.24307 | 1.176471 |
| 0.86 | 25.29844 | 1.162791 |
| 0.87 | 25.39402 | 1.149425 |
| 0.88 | 25.53885 | 1.136364 |
| 0.89 | 25.74532 | 1.123596 |

| ρ | TEC | μ |
|---|---|---|
| 0.831 | 25.22576 | 1.20337 |
| 0.832 | 25.22421 | 1.20192 |
| 0.833 | 25.22292 | 1.20048 |
| 0.834 | 25.22187 | 1.19904 |
| 0.835 | 25.22109 | 1.19761 |
| 0.836 | 25.22056 | 1.19617 |
| 0.837 | 25.22030 | 1.19474 |
| 0.838 | 25.22031 | 1.19332 |
| 0.839 | 25.22059 | 1.19190 |

| ρ | TEC | μ |
|---|---|---|
| 0.8371 | 25.22029 | 1.19460 |
| 0.8372 | 25.22028 | 1.19445 |
| 0.8373 | 25.22027 | 1.19431 |
| 0.8374 | 25.22027 | 1.19417 |
| 0.8375 | 25.22027 | 1.19403 |
| 0.8376 | 25.22027 | 1.19388 |
| 0.8377 | 25.22028 | 1.19374 |
| 0.8378 | 25.22029 | 1.19360 |
| 0.8379 | 25.22030 | 1.19346 |

Optimum is approx. ρ = .8375, TEC = 25.2202702, and μ = 1.19403.

Table B.5: Total Expected Cost of Problem M/M/1E

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 44.58 | 20 |
| 0.2 | 24.08 | 10 |
| 0.3 | 16.97 | 6.666667 |
| 0.4 | 13.44 | 5 |
| 0.5 | 11.45 | 4 |
| 0.6 | 10.37 | 3.333333 |
| 0.7 | 10.05 | 2.857143 |
| 0.8 | 10.84 | 2.5 |
| 0.9 | 15.16 | 2.222222 |

| ρ | TEC | μ |
|---|---|---|
| 0.61 | 10.29887 | 3.278689 |
| 0.62 | 10.23945 | 3.225806 |
| 0.63 | 10.18751 | 3.174603 |
| 0.64 | 10.14317 | 3.125 |
| 0.65 | 10.10662 | 3.076923 |
| 0.66 | 10.07807 | 3.030303 |
| 0.67 | 10.05781 | 2.985075 |
| 0.68 | 10.04619 | 2.941176 |
| 0.69 | 10.04360 | 2.898551 |

| ρ | TEC | μ |
|---|---|---|
| 0.681 | 10.04551 | 2.936858 |
| 0.682 | 10.04493 | 2.932551 |
| 0.683 | 10.04444 | 2.928258 |
| 0.684 | 10.04404 | 2.923977 |
| 0.685 | 10.04373 | 2.919708 |
| 0.686 | 10.04352 | 2.915452 |
| 0.687 | 10.04340 | 2.911208 |
| 0.688 | 10.04337 | 2.906977 |
| 0.689 | 10.04344 | 2.902758 |

| ρ | TEC | μ |
|---|---|---|
| 0.6871 | 10.043391 | 2.91078 |
| 0.6872 | 10.043385 | 2.91036 |
| 0.6873 | 10.043380 | 2.90994 |
| 0.6874 | 10.043375 | 2.90951 |
| 0.6875 | 10.043372 | 2.90909 |
| 0.6876 | 10.043370 | 2.90867 |
| 0.6877 | 10.043368 | 2.90825 |
| 0.6878 | 10.043368 | 2.90782 |
| 0.6879 | 10.043369 | 2.90739 |

Optimum is approx. $\rho = .6878$, TEC = 10.04336801, and $\mu = 2.907822$.

**Table B.6: Total Expected Cost of Problem M/G/1A**

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 36.02330 | 10.000 |
| 0.2 | 20.51908 | 5.000 |
| 0.3 | 15.71282 | 3.333 |
| 0.4 | 13.84540 | 2.500 |
| 0.5 | 13.42110 | 2.000 |
| 0.6 | 14.13249 | 1.667 |
| 0.7 | 16.30672 | 1.429 |
| 0.8 | 21.52251 | 1.250 |
| 0.9 | 38.21179 | 1.111 |

| ρ | TEC | μ |
|---|---|---|
| 0.41 | 13.74853 | 2.439 |
| 0.42 | 13.66475 | 2.381 |
| 0.43 | 13.59359 | 2.326 |
| 0.44 | 13.53462 | 2.273 |
| 0.45 | 13.48749 | 2.222 |
| 0.46 | 13.45190 | 2.174 |
| 0.47 | 13.42761 | 2.128 |
| 0.48 | 13.41446 | 2.083 |
| 0.49 | 13.41231 | 2.041 |

| ρ | TEC | μ |
|---|---|---|
| 0.481 | 13.413751 | 2.079 |
| 0.482 | 13.413152 | 2.075 |
| 0.483 | 13.412664 | 2.070 |
| 0.484 | 13.412285 | 2.066 |
| 0.485 | 13.412016 | 2.062 |
| 0.486 | 13.411857 | 2.058 |
| 0.487 | 13.411807 | 2.053 |
| 0.488 | 13.411866 | 2.049 |
| 0.489 | 13.412035 | 2.045 |

| ρ | TEC | μ |
|---|---|---|
| 0.4866 | 13.41181344 | 2.055 |
| 0.4867 | 13.41181009 | 2.055 |
| 0.4868 | 13.41180784 | 2.054 |
| 0.4869 | 13.41180668 | 2.054 |
| 0.4870 | 13.41180661 | 2.053 |
| 0.4871 | 13.41180765 | 2.053 |
| 0.4872 | 13.41180977 | 2.053 |
| 0.4873 | 13.41181299 | 2.052 |
| 0.4874 | 13.41181731 | 2.052 |

Optimum is approx. $\rho = .4870$, TEC = 13.41180661, and $\mu = 2.053$.

### Table B.7: Total Expected Cost of Problem M/G/1B

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 671.58100 | 40.000 |
| 0.2 | 351.58166 | 20.000 |
| 0.3 | 242.63705 | 13.333 |
| 0.4 | 188.11685 | 10.000 |
| 0.5 | 156.15007 | 8.000 |
| 0.6 | 136.31719 | 6.667 |
| 0.7 | 124.93088 | 5.714 |
| 0.8 | 122.66808 | 5.000 |
| 0.9 | 143.45050 | 4.444 |

| ρ | TEC | μ |
|---|---|---|
| 0.71 | 124.23498 | 5.634 |
| 0.72 | 123.62715 | 5.556 |
| 0.73 | 123.11141 | 5.479 |
| 0.74 | 122.69274 | 5.405 |
| 0.75 | 122.37721 | 5.333 |
| 0.76 | 122.17219 | 5.263 |
| 0.77 | 122.08666 | 5.195 |
| 0.78 | 122.13148 | 5.128 |
| 0.79 | 122.21345 | 5.063 |

| ρ | TEC | μ |
|---|---|---|
| 0.771 | 122.08508 | 5.188 |
| 0.772 | 122.08480 | 5.181 |
| 0.773 | 122.08587 | 5.175 |
| 0.774 | 122.08826 | 5.168 |
| 0.775 | 122.09201 | 5.161 |
| 0.776 | 122.09712 | 5.155 |
| 0.777 | 122.10361 | 5.148 |
| 0.778 | 122.11149 | 5.141 |
| 0.779 | 122.12078 | 5.135 |

| ρ | TEC | μ |
|---|---|---|
| 0.7711 | 122.084990 | 5.187 |
| 0.7712 | 122.084917 | 5.187 |
| 0.7713 | 122.084857 | 5.186 |
| 0.7714 | 122.084811 | 5.185 |
| 0.7715 | 122.084777 | 5.185 |
| 0.7716 | 122.084757 | 5.184 |
| 0.7717 | 122.084750 | 5.183 |
| 0.7718 | 122.084756 | 5.183 |
| 0.7719 | 122.084775 | 5.182 |

Optimum is approx. ρ = .7717, TEC = 122.084750, and μ = 5.183.

### Table B.8: Total Expected Cost of Problem M/G/1C

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 159.82 | 40.000 |
| 0.2 | 161.72 | 20.000 |
| 0.3 | 179.00 | 13.333 |
| 0.4 | 206.25 | 10.000 |
| 0.5 | 246.21 | 8.000 |
| 0.6 | 307.06 | 6.667 |
| 0.7 | 408.90 | 5.714 |
| 0.8 | 612.61 | 5.000 |
| 0.9 | 1223.13 | 4.444 |

| ρ | TEC | μ |
|---|---|---|
| 0.11 | 158.562 | 36.364 |
| 0.12 | 157.809 | 33.333 |
| 0.13 | 157.454 | 30.769 |
| 0.14 | 157.423 | 28.571 |
| 0.15 | 157.659 | 26.667 |
| 0.16 | 158.122 | 25.000 |
| 0.17 | 158.780 | 23.529 |
| 0.18 | 159.611 | 22.222 |
| 0.19 | 160.783 | 21.053 |

| ρ | TEC | μ |
|---|---|---|
| 0.131 | 157.4377 | 30.534 |
| 0.132 | 157.4242 | 30.303 |
| 0.133 | 157.4138 | 30.075 |
| 0.134 | 157.4064 | 29.851 |
| 0.135 | 157.4020 | 29.630 |
| 0.136 | 157.4005 | 29.412 |
| 0.137 | 157.4018 | 29.197 |
| 0.138 | 157.4060 | 28.986 |
| 0.139 | 157.4129 | 28.777 |

| ρ | TEC | μ |
|---|---|---|
| 0.1356 | 157.40072 | 29.499 |
| 0.1357 | 157.40062 | 29.477 |
| 0.1358 | 157.40054 | 29.455 |
| 0.1359 | 157.40049 | 29.433 |
| 0.136 | 157.40047 | 29.412 |
| 0.1361 | 157.40048 | 29.390 |
| 0.1362 | 157.40051 | 29.369 |
| 0.1363 | 157.40058 | 29.347 |
| 0.1364 | 157.40067 | 29.326 |

Optimum is approx. ρ = .1360, TEC = 157.40047, and μ = 29.412.

**Table B.9: Total Expected Cost of Problem M/G/1D**

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 44.57 | 20.000 |
| 0.2 | 24.05 | 10.000 |
| 0.3 | 16.91 | 6.667 |
| 0.4 | 13.30 | 5.000 |
| 0.5 | 11.20 | 4.000 |
| 0.6 | 9.92 | 3.333 |
| 0.7 | 9.23 | 2.857 |
| 0.8 | 9.24 | 2.500 |
| 0.9 | 11.11 | 2.222 |

| ρ | TEC | μ |
|---|---|---|
| 0.71 | 9.19845 | 2.817 |
| 0.72 | 9.16972 | 2.778 |
| 0.73 | 9.14803 | 2.740 |
| 0.74 | 9.13381 | 2.703 |
| 0.75 | 9.12759 | 2.667 |
| 0.76 | 9.12998 | 2.632 |
| 0.77 | 9.14174 | 2.597 |
| 0.78 | 9.16378 | 2.564 |
| 0.79 | 9.18697 | 2.532 |

| ρ | TEC | μ |
|---|---|---|
| 0.751 | 9.12743 | 2.663 |
| 0.752 | 9.12736 | 2.660 |
| 0.753 | 9.12737 | 2.656 |
| 0.754 | 9.12747 | 2.653 |
| 0.755 | 9.12766 | 2.649 |
| 0.756 | 9.12795 | 2.646 |
| 0.757 | 9.12832 | 2.642 |
| 0.758 | 9.12878 | 2.639 |
| 0.759 | 9.12933 | 2.635 |

| ρ | TEC | μ |
|---|---|---|
| 0.7521 | 9.1273541 | 2.659 |
| 0.7522 | 9.1273524 | 2.659 |
| 0.7523 | 9.1273516 | 2.659 |
| 0.7524 | 9.1273517 | 2.658 |
| 0.7525 | 9.1273527 | 2.658 |
| 0.7526 | 9.1273546 | 2.657 |
| 0.7527 | 9.1273573 | 2.657 |
| 0.7528 | 9.1273609 | 2.657 |
| 0.7529 | 9.1273654 | 2.656 |

Optimum is approx. ρ = .7523, TEC = 9.1273516, and μ = 2.659.

**Table B.10: Total Expected Cost of Problem M/G/1E**

| ρ | TEC | μ |
|---|---|---|
| 0.1 | 44.796 | 20.000 |
| 0.2 | 24.305 | 10.000 |
| 0.3 | 17.194 | 6.667 |
| 0.4 | 13.637 | 5.000 |
| 0.5 | 11.597 | 4.000 |
| 0.6 | 10.416 | 3.333 |
| 0.7 | 9.901 | 2.857 |
| 0.8 | 10.243 | 2.500 |
| 0.9 | 13.105 | 2.222 |

| ρ | TEC | μ |
|---|---|---|
| 0.71 | 9.88810 | 2.817 |
| 0.72 | 9.88400 | 2.778 |
| 0.73 | 9.88877 | 2.740 |
| 0.74 | 9.90305 | 2.703 |
| 0.75 | 9.92759 | 2.667 |
| 0.76 | 9.96331 | 2.632 |
| 0.77 | 10.01130 | 2.597 |
| 0.78 | 10.07287 | 2.564 |
| 0.79 | 10.14957 | 2.532 |

| ρ | TEC | μ |
|---|---|---|
| 0.711 | 9.887312 | 2.813 |
| 0.712 | 9.886602 | 2.809 |
| 0.713 | 9.885977 | 2.805 |
| 0.714 | 9.885437 | 2.801 |
| 0.715 | 9.884981 | 2.797 |
| 0.716 | 9.884612 | 2.793 |
| 0.717 | 9.884329 | 2.789 |
| 0.718 | 9.884133 | 2.786 |
| 0.719 | 9.884024 | 2.782 |

| ρ | TEC | μ |
|---|---|---|
| 0.7191 | 9.8840184 | 2.781 |
| 0.7192 | 9.8840132 | 2.781 |
| 0.7193 | 9.8840089 | 2.780 |
| 0.7194 | 9.8840055 | 2.780 |
| 0.7195 | 9.8840030 | 2.780 |
| 0.7196 | 9.8840014 | 2.779 |
| 0.7197 | 9.8840007 | 2.779 |
| 0.7198 | 9.8840008 | 2.779 |
| 0.7199 | 9.8840018 | 2.778 |

Optimum is approx. $\rho = .7197$, TEC = 9.8840007 and $\mu = 2.779$.