SOCIAL OPTIMIZATION VERSUS SELF-OPTIMIZATION IN WAITING LINES

BY

I. ADLER and P. NAOR

TECHNICAL REPORT NO. 126
SEPTEMBER 15, 1969

SUPPORTED BY THE ARMY, NAVY AND AIR FORCE
UNDER CONTRACT Nonr-225(53)(NR-042-002)
WITH THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF OPERATIONS RESEARCH
AND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

SOCIAL OPTIMIZATION VERSUS SELF-OPTIMIZATION IN WAITING LINES

by

I. Adler and P. Naor

TECHNICAL REPORT NO. 126
September 15, 1969

Supported by the Army, Navy and Air Force under
Contract Nonr-225(53) (NR-042-002)
with the Office of Naval Research

Gerald J. Lieberman, Project Director

This research was partially supported
by the Office of Naval Research under
Contract Nonr-225(89) (NR-047-061) and
National Science Foundation Grant GK-2925

DEPARTMENT OF OPERATIONS RESEARCH
and
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

ABSTRACT

A queueing model is considered where customers arriving in a Poisson stream are given the choice of either joining the waiting line or — by declining to do so — of foregoing the benefits accruing through service. The decision of each customer is based on his concrete benefit-cost analysis. Since his service time is constant and exhaustive information as to the actual state of the system is available both of the alternatives presented to the individual customer are completely deterministic and his decision is not reached under uncertainty or risk. The cost structure envisayed as well as additional assumptions give rise to a queueing model with limited waiting room apparently not previously considered in the literature. After detailed analysis of the model and blending with the cost structure it is shown that the criterion for self-optimization of the customer will not bring about social optimization, the latter being defined as the maximally feasible expected net gain per unit time accruing to the totality of customers. A number of simple and comprehensive optimization equations are derived. By marginal analysis the correctness of the simple equations is verified and their applicability is extended to models possessing more general character.

self-optimization and social optimization. However, the case of exponential service distribution does not lend itself to convenient generalizations if one intends to investigate situations where the state space is infinite and comprises all real numbers in the interval $(0,\nu)$ $(\nu > 1)$. In order to make progress toward such cases it is convenient to start with fixed and equal service times. The feasibility of attaining a higher degree of generality through the use of fixed (rather than exponentially distributed) service times contributes to the rationale underlying the approach presently taken in this communication. Finally it is useful pointing out that the "pure" queueing problem (i.e., the stochastic model only, without the assumed cost structure) proposed here, to wit, the employment of a bounded and "non-integral" waiting room, seems not to have been dealt with in the literature. It is quite feasible that such models are of use in various applications other than those envisaged here. Thus, for instance, the mixing problem in chemical engineering stated and solved by Shinnar (1967) in queueing terms may be generalized on the lines of the pure stochastic model developed in the present paper.

2. Model Characteristics and Cost Structure

In the present section we shall state the precise assumptions relating to the stochastic model as well as to cost structure. However before going into the specifics it is useful to observe and state two distinctive qualities of the general setting in which the present model is situated.

1. Since the leitmotif of this study is contrasting two

optimization procedures it is essential to define two distinct objective functions, one related to the aspirations of (non-cooperating) decision making individuals, the other concerned with the public good. There is no (mental) difficulty regarding the first position; the individual customer simply seeks to maximize his own net income. For a reasonable objective function to describe the second position one has to introduce a set of specific assumptions circumscribing the structure of public good. In the present communication we shall follow the mode employed in the preceding study (Naor (1969)) and choose the average net income accruing to the totality of customers in unit time as the objective function to be maximized. This choice of an objective function presupposes <u>one</u> of the following situations to prevail: (i) There exists essentially only one genuine (overall) optimizer and individual customers are subcontractors of decision making, as it were, who are obliged by administrative fiat to reach their conclusions on the basis of rules prescribed by the overall optimizer. (ii) Alternatively a situation is envisaged where net gains of customers are considered to be comparative and additive, and by common agreement may be redistributed. Cooperation between customers — displayed by some through refraining from joining the queue apparently against their own best interests — will produce additional net income in unit time. This will be redistributed and, eventually, the average net profit accruing to <u>each</u> customer will exceed that of self-optimizing customers within a framework of non-cooperation. A feasible instrument of control under the present circumstances is the imposition of a toll which, if wisely determined, will produce both social optimality and a reserve stock of money to be used for redistribution. Proper identification of a set of circumstances under which an overall objective function is deemed to exist is essential in an analysis of the present character.

2. In most queueing models with a built-in optimization procedure (e.g., through the agency of priority service rules or through control of the service intensity) it is assumed that all arriving customers – sooner or later – are going to be serviced. The feasible control actions (if such are envisaged at all) in most queueing models do not typically include the perempting non-admission of a customer based on a cost-benefit analysis. While a number of models were developed which included the element of potential non-service – e.g., the balking and reneging models – this contingency was presented in probability terms only; non-admission was never considered to be an instrument of economic control. In the balking and reneging models the probability of not joining (and of leaving) the queue is associated with the customer's impatience – a psychological criterion rather than an economic one. Diverting the customer from the queue without rendition of any service is then a feasible course of action in the model area under consideration. To obviate difficulties – mostly of psychological origin – which stem from the feeling that customers must get

some sort of service eventually, we may re-circumscribe the present framework in seemingly different terms: customers may be served in two alternative, distinct modes. There is the standard mode of service which can always be relied upon and which is not associated with any queue of customers; it serves as reference point against which any other mode of service (concretely just one) may be compared. The non-standard mode of service is advantageous in monetary terms as compared with the standard mode if the waiting line of customers, ahead of the new arrival, is sufficiently small. If we describe the system in such terms – advantageous non-standard service with potential queue in front of the station compared with generally disadvantageous but queue-free standard service – we really deal with a model completely equivalent to the previous situation. To sum up, for a discussion of self-optimization versus social optimization to make sense non-admission of suctomers to the service station must be a feasible control action. Customers diverted from the station may be thought of either as not receiving service at all or as being rendered a standard type of service for which it is never necessary to queue up.

After these observations and the setting of the general framework the specifics of the stochastic model and of the cost structure may be stated in the following terms:

(i) A stationary Poisson stream of customers – with parameter λ – arrives at a single service station.

(ii) The service time necessary to satisfy and dispatch a customer is a constant T; all service times are equal.

(iii) On successful completion of service, the customer is endowed with a reward R (expressible in monetary units). All customer rewards are equal.

(iv) The cost to a customer for staying in a queue (i.e., for queueing) is C monetary units in unit time. All customer costs are equal.

(v) The newly arrived customer is required to choose one of two alternatives: either (a) he joins the queue, incurs the losses associated with spending some of his time in it, and finally obtains the reward; or (b) he refuses to join the queue — an action which does not bring about any gain or loss. The choice is made by the customer on comparing the net gains associated with each of these alternatives. Two modes of decision are examined. In one mode customers are assumed to act solely in their self-interest; it is sufficient for the net gain to the individual to be non-negative in order to induce him to join the queue. In the other decision mode each individual acts on behalf of the totality of customers and he assumes every customer to act in the same spirit; this customer seeks a decision criterion by which average net income in unit time is maximized.

Model assumptions (i), (iii) and (iv) are identical with those appearing in the previous study. Assumption (v) is more inclusively formulated than its original counterpart in order to render equal states a priori in the derivations to be carried out to both self-optimization and social optimization.

Assumption (ii) is radically different from its predecessor: Fixed – rather than exponentially distributed – service times are envisaged and this causes the individual customer to be placed in a deterministic decision situation. We note, furthermore, that this assumption (in conjunction with the others, of course) gives rise to a stochastic model which is interesting per se and may be put to use in other contexts as well. The mathematical techniques which have to be employed under the present assumption (ii) are of a different quality than those useful (and sufficient) if the assumption of exponentially distributed service times is considered valid. Finally, it will be shown that several of the results to be attained here by employing assumption (ii) serve as a more advantageous point of departure for some generalization than can be expected from the original model.

3. Finite and Nonintegral Queueing Capacity

What is the decision criterion of the individual customer seeking self-optimization? Clearly he reaches his decision under conditions of certainty. Upon his arrival he views the queue ahead of him which is made up of two parts: k customers are in the waiting line and one is in service. The outstanding service time of the latter is observed to be τ $(0 \leq \tau \leq T)$. If he chooses to join the queue then – assuming throughout the discipline "first-come-first-served" – his total queueing time, from the instant of his arrival and joining to the instant of his service completion, will equal $kT + \tau + T$ (the third term being the customer's own service time. Since the decision is based on the customer's self-interest it will be considered correct if the cost of queueing does not exceed the reward. That is if, and only if, the (weak) inequality

$$R - C[(k+1)T + \tau] \geq 0 \tag{1}$$

is satisfied the newly arrived customer should join the queue.

It will be sometimes advantageous to view this from a slightly different angle. Let the occupancy or the state K of the system at any arbitrary instant be defined as the ratio of the (future) queueing time of the last customer in the line and the service time T.

$$K = \frac{kT+\tau}{T} = k + \frac{\tau}{T} \tag{2}$$

Furthermore let a dimensionless index ν_s be defined as

$$\nu_s \equiv \frac{R}{CT} \tag{3}$$

Inequality (1) is transformed into

$$K \leq \nu_s - 1 \tag{4}$$

which is interpreted in the following manner: The incoming new customer observes the state K of the system; if it does not exceed the value of ν_s-1 the customer joins the queue, otherwise the customer forgoes queueing as well as service. Now all customers act by the same criterion; hence K can take on values in the interval $[0,\nu_s]$. The maximum value ν_s will be realized under the following circumstances: an incoming customer encounters the system in a state ν_s-1, the maximum value at which the system is still accessible to new arrivals and — by the act of joining — he transforms the system into state ν_s. Whenever the system is in a state within the interval $(\nu_s-1,\nu_s]$ it is said to be inaccessible to new customers.

We note that the rule prescribing accessibility makes sense only if the inequality

$$\nu_s \geq 1 \tag{5}$$

pertains. In physical terms this means that the reward to be collected by the customer at the completion of his service must not fall short of the cost of time spent in service. If inequality (5) does not apply the proper policy is to refuse access to all customers and (possibly) disband the service station altogether.

The decision mode associated with the individual customer's self-interest has then given rise to a queueing model with Poisson arrivals, constant service times and finite queueing capacity (i.e., limited waiting room). Now the present model is different from those that have appeared in the literature on queueing theory in the following respect: In the typical model where finite queueing capacity makes its appearance the number of potentially available waiting spaces (the "size" of the waiting room) is assumed to be integral; occupancy too is considered to be an integer and changes in jumps whenever a customer departs or joins the queue. In the present model the service process changes (decreases) the occupancy continuously and uniformly, the total capacity is a real positive number (and not necessarily an integer) and arrivals — followed by absorption into the queue — bring about discrete changes. It is not difficult to verify that the specialization of capacity values to integer (without imposing any further conditions) suffices to generate what was named before the "typical model with finite queueing capacity". We observe that, if the assumption of exponentially distributed service times replaces the constant service times postulate, the "typical model" cannot be transformed into a generalized model; both capacity and actual occupancy are, of necessity, integers in this case.

Occupancy is essentially equivalent to the concept of virtual waiting time (or occupation time as it is termed sometimes) introduced by Takacs; indeed it is the ratio of occupation time to service time.

The mode of decision seeking social optimality will give rise to a queueing model of identical structure (though with one different parameter). Of course, unlike the self-optimizer, the social optimizer is not faced with a decision problem under conditions of certainty. Indeed he will have to take into account a somewhat probabilistic future, to wit, the Poisson stream of customers who will arrive at the service station. Now by the very quality of the homogeneous Poisson process the total useful information is contained in the knowledge of the arrival intensity (a parameter not relevant for the self-optimizer's decision). Hence the social optimizer, too, will at the instant of a customer's arrival, exercise control by observing the occupancy K and make the new arrival join if, and only if, the criterion

$$K \leq \nu_o - 1 \qquad (6)$$

is satisfied where ν_o is a function of both ν_s and the traffic intensity. Adherence to such a rule will generate a system possessing a structure identical with that discussed before. Poisson arrivals, constant service times, finite and nonintegral queueing capacity.

It appears then worthwhile to delve deeper into the analysis of such a system. This will be done in the sections which follow.

4. Some Basic Relations

We have then: a Poisson stream of incoming customers possessing arrival intensity λ; a single service station; each customer requires exactly T time units for the completion of its service; there is limited waiting room and the occupancy K can never exceed the constant* $\nu \geq 1$; access to the waiting line is granted to a new customer only if the occupancy does not exceed ν-1.

The traffic intensity ρ is defined as

$$\rho = \lambda T \tag{7}$$

We note that under the present model assumptions it is <u>not</u> necessary to put restrictions on the permissible values of ρ in order to obtain steady state conditions.

The state of the system at an arbitrary instant is specified either by the occupancy K or by the pair (i,t) where i is the number of customers in the queue (i.e., inclusive of the customer in service) and t is the time which has already been devoted to the customer in service

* For the specific purposes of optimization — to be discussed in a later section — the constant ν will be assigned a subscript, e.g., ν_s or ν_o.

$$i \begin{cases} = K + \frac{t}{T} = K + 1 - \frac{\tau}{T} & \text{if} \quad K > 0 \\ = 0 & \text{if} \quad K = 0 \end{cases} \tag{8}$$

$$t \begin{cases} = T - \tau & \text{if} \quad K > 0 \\ = 0 & \text{if} \quad K = 0 \end{cases} \tag{9}$$

We are concerned with the steady state regime of our system. Let p_o be defined as the probability of the service station being idle whereas $p_i(t)$ $(1 \leq i \leq [\nu] + 1,\ 0 \leq t \leq T)$ represents the probability density* pertaining to the elapsed service time t and the queue i.

Now consider the density $p_i(t)$ and, in particular, the change that is taking place $\Delta p_i(t)$ during a very small time interval Δt. Such change is associated with the difference of jump probabilities into, and out of, the state (i,t), that is $p_{i-1}(t)\lambda\Delta t - p_i(t)\lambda\Delta t$. We define

$$n = [\nu] \tag{10}$$

$$\theta = T(\nu - n) \tag{11}$$

* This is, of course, a joint density — it should be noted that one random variable (elapsed service time) is continuous while the other (queue size) is discrete. The representation in such terms, $p_i(t)$, possesses some advantage — for our present purpose — over a representation by a density associated with a single random variable, e.g., $\varphi(K)$. Potential concentrations and discontinuities (and, in fact, there is a concentration at the point $K = 0$ and a discontinuity in $\varphi(K)$ at the point $K = 1$) will be exhibited in a more natural way on utilizing the present notation.

take cognizance of the feasible values of i and of θ, and apply the idea of associating the density change within a small time duration with the difference of jump probabilities. The set of differential equations, pertaining to the present queueing model, is derived

$$\frac{dp_1(t)}{dt} = -\lambda p_1(t) \qquad (0 \leq t \leq T) \tag{12a}$$

$$\frac{dp_i(t)}{dt} = \lambda[p_{i-1}(t)-p_i(t)] \qquad (0 \leq t \leq T,\ 1 < i < n) \tag{12b}$$

$$\frac{dp_n(t)}{dt} = \lambda p_{n-1}(t) \qquad (0 \leq t \leq T-\theta) \tag{12c}$$

$$\frac{dp_n(t)}{dt} = \lambda[p_{n-1}(t)-p_n(t)] \qquad (T-\theta \leq t \leq T) \tag{12d}$$

$$\frac{dp_{n+1}(t)}{dt} = \lambda p_n(t) \qquad (T-\theta \leq t \leq T) \tag{12e}$$

Boundary conditions are established on examination of the changes that take place at times $t = 0$ and $t = \theta$

$$p_o\lambda = p_1(T) \tag{13a}$$

$$p_1(0) = p_o\lambda + p_2(T) \tag{13b}$$

$$p_i(0) = p_{i+1}(T) \qquad (1 < i \leq n) \tag{13c}$$

$$p_{n+1}(T-\theta) = 0 \tag{13d}$$

The probability of having $i\ (> 0)$ customers in the queue is given by

$$p_i = \int_0^T p_i(t)dt \qquad 1 \leq i \leq n \tag{14a}$$

$$p_{n+1} = \int_{T-\theta}^{T} p_{n+1}(t)dt \tag{14b}$$

Obviously these probabilities — together with p_o — obey

$$\sum_{i=0}^{n+1} p_i = 1 \tag{15}$$

The probability, P_c, of the service station being closed to incoming traffic may be evaluated as

$$P_c = \int_0^{T-\theta} p_n(t)dt + \int_{T-\theta}^{T} p_{n+1}(t)dt \tag{16}$$

The busy fraction — which in the present type of model is <u>not</u> identical with the traffic intensity ρ — is equal to

$$b = 1 - p_o \tag{17}$$

During the busy period the rate of discharge of customers from the service station equals T^{-1}. Hence the <u>average</u> rate of discharge — i.e., the expected number of customers leaving the service station in unit time is then evaluated as the product $(1-p_o)T^{-1}$. Now the average number of customers admitted to the service station in unit time is given by $\lambda(1-P_c)$. Within a steady state regime these two quantities must be equal. Hence after some rearrangement we obtain

$$\rho = \lambda T = \frac{1-p_o}{1-p_c} \tag{18}$$

If, as is assumed here, service times are fixed and equal then, by first principles, the average number of times a state t (disregarding i) is realized in unit time cannot depend on t. Hence the solutions $p_i(t)$ must obey the following equation

$$\sum_{i=1}^{n \text{ or } n+1} p_i(t) = (1-p_o)T^{-1} = \lambda(1-P_c) \qquad (19)$$

It is apparent that the idle fraction p_o plays an important role in the central formulas of the model. This quantity is a function of the parameters ν and ρ. Depending on the circumstances we may desire to use the obvious notation $p_o(\nu,\rho)$ or $p_o(\nu)$.

Application of (rather lengthy) standard solution methods on the set $\{12\}$ as well as combination with other equations of this section yields

$$\begin{aligned} p_o(\nu,\rho) &= \{1 + \sum_{j=1}^{n} (-1)^{j-1} \frac{[(n-j)\lambda T+\lambda\theta]^{j-1}}{(j-1)!} \lambda T e^{(n-j)\lambda T+\lambda\theta}\}^{-1} = \\ &= \{1 + \sum_{j=1}^{n} (-1)^{j-1} \frac{[(\nu-j)\rho]^{j-1}}{(j-1)!} \rho e^{(\nu-j)\rho}\}^{-1} \end{aligned} \qquad (20)$$

If n exceeds the value 1 (the alternative case is elementary) it may be shown after some further manipulation that the following result is attained

$$p_1(\nu,\rho) = p_o(\nu,\rho)(e^{\rho}-1) \qquad \text{if} \quad n > 1 \qquad (21a)$$

$$p_i(\nu,\rho) = p_o(\nu,\rho)[e^{i\rho} + \sum_{j=1}^{i-1} (-1)^j e^{(i-j)\rho} \{\frac{[(i-j)\rho]^j}{j!} + \frac{[(i-j)\rho]^{j-1}}{(j-1)!}\}] \qquad (21b)$$

$$\text{for} \quad 1 < i < n$$

Let (21) is rather interesting; formally the probabilities are given by equations which are identical with those relating to the analogous model with unlimited waiting room. These were already evaluated in the early days of queueing theory — indeed they can be found in Fry's (1928) textbook. However beyond the formal identity we must take note that the two positions diverge in three aspects at least: a) the probability p_o which appears as a multiplier in {21} is different in the two cases. b) The traffic intensity ρ must fall short of the value 1 in the infinite waiting room model; in the limited waiting room model this restriction is removed. c) In the infinite waiting room model the validity of (21b) ranges over all feasible values of i; in the present case the range of applicability of (21b) is limited to those values of i for which the station is never closed. Beyond the intrinsic usefulness of the set {21} we are made to realize — through its presentation — that basic formulas may be stable in some sense even though some model assumptions are modified — slightly or otherwise. The change brought about by the model modification manifests itself only in the variation of a key quantity, e.g., in the present case: p_o.

It may be desirable to examine the solution of the set {12} of differential equations in somewhat sharper detail. The following is a representation of $p_i(t)$.

Let two functions $z_i(t)$ and $Z_i(t)$ be recursively defined as

$$z_i(t) = e^{\lambda T} z_{i-1}(0) + \lambda[Z_{i-1}(t) - Z_{i-1}(T)] \qquad i > 2 \qquad (22)$$

$$Z_i(t) = \int_0^t z_i(t')dt' \qquad i \geq 2 \qquad (23)$$

and let the "starting function $z_2(t)$" be equal to

$$z_2(t) = e^{\lambda T} - [\lambda(T-t) + 1] \tag{24}$$

The solution $p_i(t)$ of the set {12} is then given by

$$p_1(t) = p_o \lambda e^{\lambda(T-t)} \tag{25a}$$

$$p_i(t) = p_o \lambda e^{\lambda(T-t)} z_i(t) \tag{25b}$$

for all i and t in $(1 \leq i < n,\ 0 \leq t \leq T)$ which is feasible for $\nu \geq 2$ and for i and t in $(i = n,\ T-\theta \leq t \leq T)$ in which case ν may take on any value exceeding 1. The restrictions on i, t and ν enumerated above may be physically interpreted as relating the set {25} to precisely those states in which the service station is a) busy and b) accessible. The proof of {25} is inductive and rather lengthy; it will not be presented here.

Finally in this section we put forward* an equation representing the average queue size, $q(\nu,\rho)$, in its dependence on ν and ρ

$$q(\nu,\rho) = n - p_o(\nu,\rho)\{e^{(n-1)\lambda T+\lambda\theta}(1-\lambda\theta) + \sum_{j=2}^{n} (e^{(n-j)\lambda T+\lambda\theta}[\sum_{\ell=1}^{j-1} (-1)^{\ell-1} \frac{[(n-j)\lambda T+\lambda\theta]^{\ell-1}}{(\ell-1)!}$$

$$+ (-1)^{j-1} \frac{[(n-j)\lambda T+\lambda\theta]^{j-1}}{(j-1)!} (1-\lambda\theta)])\} \tag{26}$$

* Here again no proof is furnished in the paper; we wish to state that the derivation of (26) is burdensome and apparently manipulative skill rather than depth is required.

Equation (26) continued

$$= n-p_o(\nu,\rho)\{e^{(\nu-1)\rho}(1+n\rho-\nu\rho) + \sum_{j=2}^{n}(e^{(\nu-j)\rho}[\sum_{\ell=1}^{j-1}(-1)^{\ell-1}\frac{[(\nu-j)\rho]^{\ell-1}}{(\ell-1)!}$$

$$+ (-1)^{j-1}\frac{[(\nu-j)\rho]^{j-1}}{(j-1)!}(1+n\rho-\nu\rho)])\}$$

In equation (26) sums are defined to equal zero if the lower value of the summation index exceeds the upper one, e.g., $\sum_{j=2}^{1}(\quad) = 0$. Hence the queue formula (26) is valid for <u>all</u> values of $\nu \geq 1$.

5. Optimization

The derivation of a strategy for self-optimization is rather elementary. The self-optimizing customer is aware of the quantities R, C and T. He utilizes relation (3) to compute ν_s. Upon arrival at the service station he observes the actual occupancy K. If inequality (4) is observed he reaches an affirmative decision to join. The decision is negative in the alternative case.

The impact of this strategy on the "society" of customers is that average gross gains ensue at the rate $RT^{-1}(1-p_o)$ in unit time; the resulting congestion incurs an average cost of Cq. Hence the average net gain, G, accruing to customers in unit time is given by

$$G = RT^{-1}(1-p_o) - Cq \qquad (27)$$

The quantities p_o and q in (27) are computed through the use of equations (20) and (26); the arguments ρ and ν in these equations are the observed traffic intensity λT and the chosen strategy ν_s, respectively.

Next we consider social optimization. Our point of departure is equation (27) and it is presently assumed that p_o and q are functions of ρ (an observed datum) and of a ν whose optimal value, ν_o, will have to be determined.

Now G is a differentiable (and possibly unimodal) function of ν and hence the technique of optimization that suggests itself is differentiation. After surveying the structure of p_o and of ν one is prone to think that, prima facie, differentiation would be a formidable task — technically speaking. In order to obviate the technical difficulties we proceed as follows:

Two quantities, N and D, are defined as

$$N = e^{(\nu-1)\rho}(1+n\rho-\nu\rho) + \sum_{j=2}^{n}(e^{(\nu-j)\rho}[\sum_{\ell=1}^{j-1}(-1)^{\ell-1}\frac{[(\nu-j)\rho]^{\ell-1}}{(\ell-1)!} + \qquad (28)$$

$$+ (-1)^{j-1}\frac{[(\nu-j)\rho]^{j-1}}{(j-1)!}(1+n\rho-\nu\rho)])$$

$$D = 1 + \sum_{j=1}^{n}(-1)^{j-1}\frac{[(\nu-j)\rho]^{j-1}}{(j-1)!}\rho e^{(\nu-j)\rho} \qquad (29)$$

Using this notation we may write

$$p_o = \frac{1}{D} \qquad (30)$$

$$q = n - \frac{N}{D} \qquad (31)$$

The derivatives of N and D (with respect to θ) are closely related.

$$\frac{dN}{d\theta} = -\theta K(\theta,n) \qquad (32)$$

and

$$\frac{dD}{d\theta} = TK(\theta,n) \tag{33}$$

where the function $K(\theta,n)$ is defined as

$$K(\theta,n) = \lambda^2 e^{(n-1)\rho+\lambda\theta} + \sum_{j=2}^{n} (-1)^{j-1}\lambda^2 e^{(n-j)\rho+\lambda\theta}\left(\frac{[(n-j)\rho+\lambda\theta]^{j-1}}{(j-1)!}\right.$$
$$\left. + \frac{[(n-j)\rho+\lambda\theta]^{j-2}}{(j-2)!}\right) \tag{34}$$

Again in (34) the summation is defined to yield zero if the lower value of the index exceeds the upper one. Hence $K(\theta,n)$ is defined over all feasible values of the arguments n and θ. It is not difficult to verify that it never takes on negative values.

We obtain the derivative, with respect to θ, of the net profit function (n is held constant, of course)

$$\begin{aligned}\frac{dG}{d\theta} &= \frac{d}{d\theta}\left[\frac{R(1-p_o)}{T} - Cq\right] = \frac{d}{d\theta}\left[\frac{R}{T} - \frac{R}{TD} - Cn + \frac{CN}{D}\right] \\ &= \frac{K(\theta,n)}{D^2}\left[R - C(D\theta+NT)\right] = \frac{K(\theta,n)}{D}\left[\frac{R}{D} - C(\theta + T\frac{N}{D})\right] \\ &= \frac{K(\theta,n)}{D}\{Rp_o - C[\theta + T(n-q)]\} \\ &= \frac{CTK(\theta,n)}{D}(p_o\nu_s-\nu+q)\end{aligned} \tag{35}$$

The quantity

$$D(p_o\nu_s-\nu+q) = \nu_s - \frac{D\theta}{T} - N \tag{36}$$

is a uniformly decreasing function of θ since its derivative with respect to θ is made to equal

$$\frac{d(\nu_s - \frac{D\theta}{T} - N)}{d\theta} = - \frac{D}{T} - \frac{\theta}{T}(\frac{dD}{d\theta} + \frac{T}{\theta}\frac{dN}{d\theta}) = - \frac{D}{T} \qquad (37)$$

on utilizing equations (32) and (33).

Now for sufficiently small ν expression (36) can be made positive (given that $\nu_s > 1$) and for sufficiently large ν we can always make it negative. As all other factors on the right hand side of (35) are positive we deduce that the function $\frac{dG}{d\nu}$ possesses exactly <u>one</u> zero at that value of the argument (ν or θ) at which the function $\nu_s p_o - \nu + q$ vanishes. Hence ν_o, the value which brings about social optimality, may be obtained from

$$\nu_s p_o(\nu_o,\rho) - \nu_o + q(\nu_o,\rho) = 0 \qquad (38)$$

Equation (38) is of both theoretical and practical interest. We note that the problem was originally set in terms of obtaining the derivative (with respect to θ) of the net gain function. The differentiation would have to be carried out within strips of constant n since a change in n causes θ to jump between its two extreme feasible values, 0 and T. The analysis undertaken and, in particular, the devices utilized generated optimization equation (38) in which dependence on θ is suppressed and a simple formal structure is attained.

Formula (38) is also a convenient starting point for numerical work. As formulated in this study the determination of ν_s precedes that of ν_o; hence ν_o turns out to be an <u>implicit</u> (and not particularly convenient) function of ν_s and ρ. To set up a table of numerical values one would start with given ν_o and ρ and seek the appropriate value of ν_s. This is analogous to the approach undertaken in the previous study. The numerical computation of ν_s as a function of ν_o and ρ is straightforward and presents no extraordinary practical difficulties. Furthermore the physical interpretation of such a reformulation of (38) ($\nu_s = \frac{\nu_o - q(\nu_o,\rho)}{p_o(\nu_o,\rho)}$) is not farfetched: For a queueing model of the type described here, the traffic intensity ρ and the socially optimal capacity ν_o are given; it is desired to find that capacity, ν_s, which self-optimizing customers will generate, if no regulation of traffic – financial or administrative – is imposed.

What is the optimal (maximal) rate of net gain G_o ? To derive this we return to (27) and assume that the optimal ν, i.e., ν_o, has been made the criterion of decision. We have then

$$\begin{aligned} G_o &= \frac{R(1-p_o(\nu_o,\rho))}{T} - Cq(\nu_o,\rho) \\ &= C[\nu_s(1-p_o(\nu_s,\rho)) - q(\nu_o,\rho)] \qquad (39) \\ &= C[\nu_s-\nu_o-(\nu_s p_o(\nu_o,\rho)-\nu_o+q)] = C(\nu_s-\nu_o) \end{aligned}$$

Equation (39) is both simple and informative:

First it makes one realize in immediate terms that the inequality

$$\nu_s \geq \nu_o \qquad (40)$$

must hold (where equality is realized if, and only if, $\nu_s = 1$ $(\rho > 0)$). This, of course, is one of the objectives of the present study.

Secondly we observe that the right hand side of (39) is very closely related to the regulatory toll that should be imposed on incoming customers in order to maximize average (social) net gain in unit time. Indeed, the optimal toll* S_o, is obviously given by

$$S_o = CT(\nu_s - \nu_o) \qquad (41)$$

We have then the interesting (and, on first sight, slightly strange) result that the optimal toll to be imposed on the customer is identical with the average optimal gain accumulating during one service period.

$$S_o = G_o T \qquad (42)$$

Thirdly, one is induced to pose the question whether the simple formulas attained here – such as (38), (39) and (41) – are amenable to simple physical interpretations and, possibly, to further generalizations. In the following we present the marginal analysis pertaining to social optimization. We shall show that it leads to the very same equations possessing elementary structure.

It is the social optimizer's function to select an indifference capacity $(\nu_o - 1)$ possessing the following characteristic. A customer who arrives at an instant – $t = 0$, say – at which the system possesses the occupancy $K = \nu_o - 1$ will generate identical gains to society either

* Unlike the case discussed in the previous study (where an optimal toll was one taken from a range of values) there is exactly one optimal toll value which maximizes (social) net gain in unit time.

by joining the queue or by declining to do so. Neither alternative is preferable to the other from the viewpoint of public good. We note that if the customer joins the queue the identical state ν_o-1 which would instantaneously prevail were he to balk will be regenerated[*] in exactly T time units (with probability 1). During this time access to the service station is blocked for new customers who (possibly) arrive within that interval. Exactly one customer will be discharged from the service station during the blocked period – at time $(\nu_o-n_o)T$. The queue size before and after this discharge is n_o+1 and n_o, respectively; it is easy to verify that the average queue length is ν_o. Hence, as a result of joining, the total net benefits reaped <u>during T</u> amount to $R-CT\nu_o$. However the decision to join at time $t = 0$ (and occupancy $K = \nu_o-1$) has <u>further</u> implications. It will be convenient to represent them by an instantaneous expected net gain rate $g_{join}(t;\nu_o-1)$. Since the net gain $(R-CT\nu_o)$ during the interval $(0,T)$ has already been separated out the function $g_{join}(t;\nu_o-1)$ takes on the value 0 up to time T

$$g_{join}(t;\nu_o-1) = 0 \qquad (0 \leq t \leq T) \qquad (43)$$

Beyond T the function $g_{join}(t)$ takes a course which incorporates the presently existing queue, the accumulation of new customers the discharge of serviced customers and the rewards gained by them. Clearly the instantaneous expected net gain rate tends to $G(\nu_o)$ as time t

[*] This property depends on the assumptions that customers arrive in a Poisson stream at the station.

tends to infinity

$$g_{join}(t;\nu_o-1) \underset{t\to\infty}{\to} G(\nu_o) \tag{44}$$

The alternative decision to balk at time $t = 0$ brings forth another instantaneous expected net gain rate $g_{balk}(t;\nu_o-1)$. By virtue of the characteristics stated before, joining the queue at time $t = 0$ generates a state at time $t = T$ which is identical with the state at time $t = 0$ brought about by the balking decision. Hence the following must hold

$$g_{balk}(t;\nu_o-1) = g_{join}(t+T;\nu_o-1) \tag{45}$$

and, of course, analogously to (44) we have

$$g_{balk}(t;\nu_o-1) \underset{t\to\infty}{\to} G(\nu_o) \tag{46}$$

What is the expected accumulated financial advantage $A(t)$ at t (conveniently assumed to exceed T) of balking over joining where we disregard the terms $R-C\nu_oT$ which favored joining and were separated out. Clearly $A(t)$ is given by

$$\begin{aligned} A(t) &= \int_0^t g_{balk}(t')dt' - \int_0^t g_{join}(t')dt' = \int_0^t g_{balk}(t')dt' - \int_T^t g_{join}(t')dt' \\ &= \int_0^t g_{balk}(t')dt' - \int_0^{t-T} g_{balk}(t')dt' = \int_{t-T}^t g_{balk}(t')dt' \end{aligned} \tag{47}$$

When t tends to infinity the integrand on the right hand side of (47) tends to the constant $G(\nu_o)$; hence the integral (47) — with t tending to infinite — is evaluated as

$$A(\infty) = TG(\nu_o) = TG_o \tag{48}$$

The gist of marginal analysis is that under conditions of optimality this advantage of balking over queueing (over an infinite horizon[*]) must equal the advantage $R-\nu_oCT$ of joining over balking within the interval $(0,T)$. Therefore we obtain

$$R - \nu_oCT = TG_o \tag{49}$$

But equation (49) is essentially identical with (39). The other general optimization formulas (38) and (41) may be easily derived from (39). Hence by using marginal analysis we have obtained the procedure for optimization without the "messy" computational technicalities. For actual numerical work it is, of course, still necessary to evaluate queue sizes and idle fractions through the use of formulas (28)-(31).

Marginal analysis has led us one step beyond the original model under investigation. The argument leading to (49) — and hence to (38) and (41) — remains essentially valid even if the assumption of constant and equal service times is modified. It is sufficient to assume that service times are distributed (rather than fixed) and that the class of distributions is characterized by the expected remaining service time of a customer being a strictly decreasing and continuous function of elapsed service time. This is a rather mild restriction. The minor modification that has to be introduced in the argumentation is that the

[*] We observe that the interest rate is (implicitly) assumed to equal zero; hence it does not make an appearance in the argument.

phrase "exactly after T time units" has to be replaced by "after T time units on the average" wherever it appears. The salient point is the following: whenever a situation exists such that a marginally joining customer is made to produce average net gain in unit time during the ensuing T time units on the average, equations (38), (39) and (41) must hold. We metnion in passing that, under conditions of social optimality a customer admitted to the service station in a non-marginal fashion as it were generates net gain exceeding the average.

Even if service times are distributed in a manner other than that prescribed in the preceding paragraph, equations (38), (39) and (41) may remain valid – at least in some approximative sense. Thus, for instance, let it be assumed that the service times are exponentially distributed; this is the case discussed in the previous study. Clearly the expected remaining service time of a customer is a constant rather than a strictly decreasing function as postulated before. Yet if in the equations representing the idle fraction and the queue the integer n_o is replaced by the (close) real number ν_o it can be shown that relations (38) etc. are revalidated. At the danger of being repetitious let it be restated that the analytical, as well as numerical, derivation of the optimal p_o and q may be quite a difficult task.

6. Conclusion

The program of this investigation was threefold:

The first objective was to show that the decision rule of self-optimizing customers operating within a framework of certainty and of equality (pertaining to R, C, and T) tends to overcongest a queueing system. This is the proper meaning of inequality (40). The basic reason for the divergence between social optimization and self-optimization — as expressed in inequality (40) — is the fact that the individual customer need not consider the penalties he is (possibly) inflicting upon future customers by the very act of his joining the queue. The toll levied on a marginally joining customer could be considered to represent compensation for damage, as it were, caused by the customer to future customers.

The second objective was to establish a vantage point for further generalization. This has been attained by alternating ordinary maximization (that is: caried out by differentiation) and marginal analysis. A set of formulas, simple and comprehensive — (38) (39), and (41) — has been shown to hold under conditions more general than originally specified.

Thirdly the stochastic queueing model with non-integral capacity has been developed and, possibly, this may be applicable in situations other than those possessing an optimization nationale. The structure and form of associated quantities — probabilities and expectations — may be quite interesting per se and some potential industrial applications indicate the necessity for further study.

The general subject area of this study possesses useful and interesting extensions. Some further investigations are under way.

REFERENCES

1. Fry, Thornton C., Probability and Its Engineering Uses, D. Van Nostrand Co., Inc., New York (1928).

2. Naor, P., The Regulation of Queue Size by Levying Tolls. Econometrica 37, 15-24 (1969).

3. Shinnar, R., Sizing of Storage Tanks for Off-Grade Material. Industrial and Engineering Chemistry, Process Design and Development, 6, 263-4 (1967).

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Dept. of Operations Research & Dept. of Statistics Stanford University Stanford, Calif. 94305	2a. REPORT SECURITY CLASSIFICATION 2b. GROUP

3. REPORT TITLE

SOCIAL OPTIMIZATION VERSUS SELF-OPTIMIZATION IN WAITING LINES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

TECHNICAL REPORT September 1969

5. AUTHOR(S) *(Last name, first name, initial)*

ADLER, I. and NAOR P.

6. REPORT DATE SEPTEMBER 15, 1969	7a. TOTAL NO. OF PAGES 30	7b. NO. OF REFS 3

8a. CONTRACT OR GRANT NO. Nonr-225(53)	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Report No. 126
b. PROJECT NO. NR-042-002	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* #15 ONR-225(89) NR-047-061 #4 NSF 2925
d.	

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Logistics & Mathematical Statistics Branch Office of Naval Research Washington, D. C. 20360

13. ABSTRACT

A queueing model is considered where customers arriving in a Poisson stream are given the choice of either joining the waiting line or - by declining to do so - of foregoing the benefits accruing through service. The decision of each customer is based on his concrete benefit-cost analysis. Since his service time is constant and exhaustive information as to the actual state of the system is available both of the alternatives presented to the <u>individual</u> customer are completely deterministic and his decision is not reached under uncertainty or risk. The cost structure envisayed as well as additional assumptions give rise to a queueing model with limited waiting room apparently not previously considered in the literature. After detailed analysis of the model and blending with the cost structure it is shown that the criterion for self-optimization of the customer will not bring about social optimization, the latter being defined as the maximally feasible expected net gain per unit time accruing to the totality of customers. A number of simple and comprehensive optimization equations are derived. By marginal analysis the correctness of the simple equations is verified and their applicability is extended to models possessing more general character.

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Self-Optimization						
Social-Optimization						
Constant Service Time						
General Service Time						
Optimization of Queueing						
Poisson arrival						
Queues						
Regulation of Queues						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
________________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
________________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
________________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.