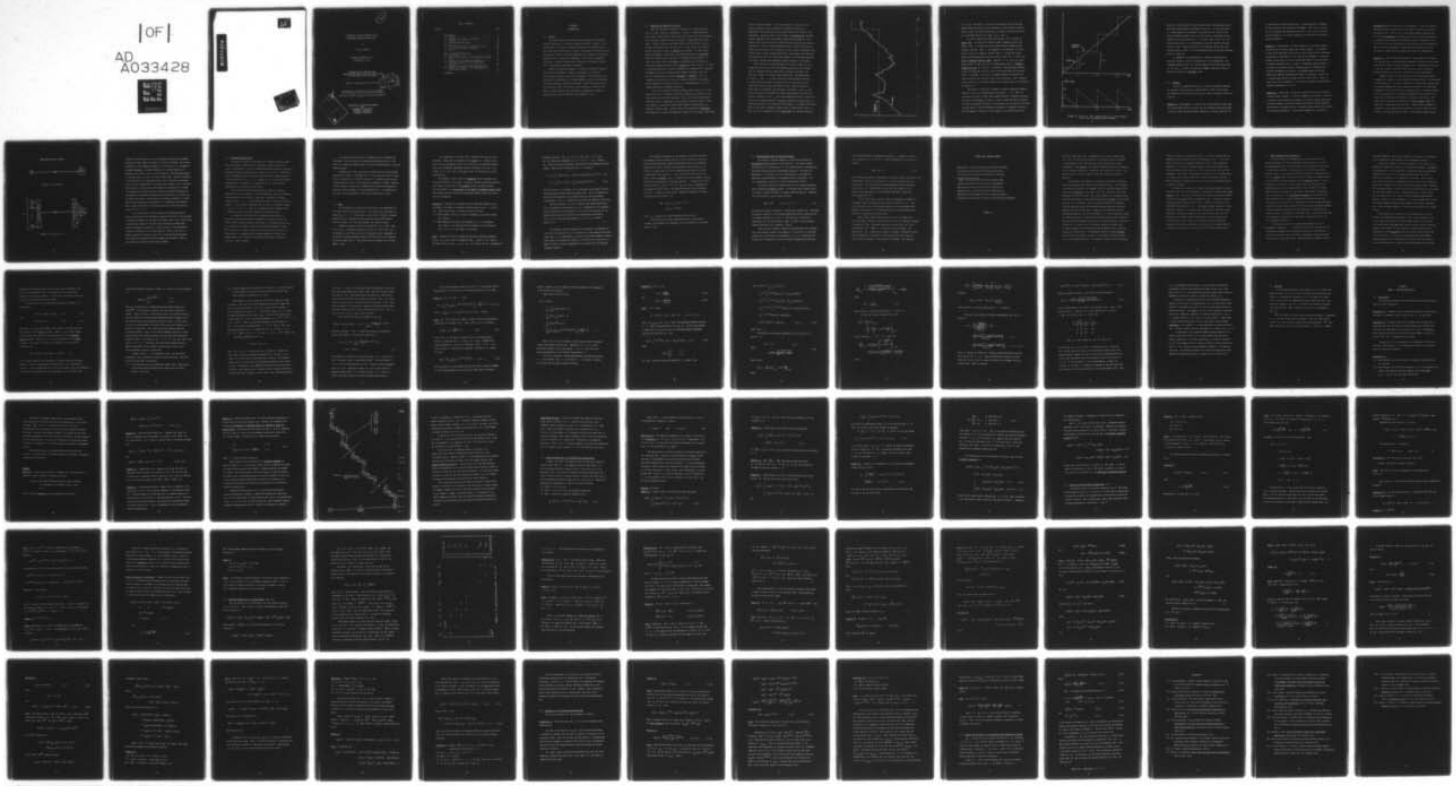
AD-A033 428 STANFORD UNIV CALIF DEPT OF STATISTICS F/G 12/2
A STOCHASTIC CAPACITY EXPANSION MODEL: NON-MODULAR TEMPORARY FA--ETC(U)
SEP 76 R S SHIPLEY N00014-75-C-0561
UNCLASSIFIED TR-178 NL



END

DATE FILMED
2-77

(12)

A STOCHASTIC CAPACITY EXPANSION MODEL: NON-MODULAR TEMPORARY FACILITIES

by

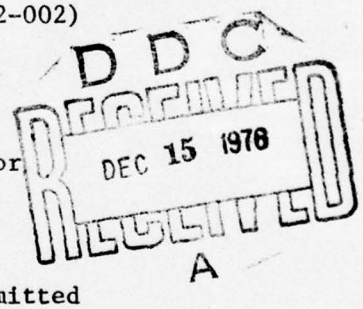
R. SCOTT SHIPLEY

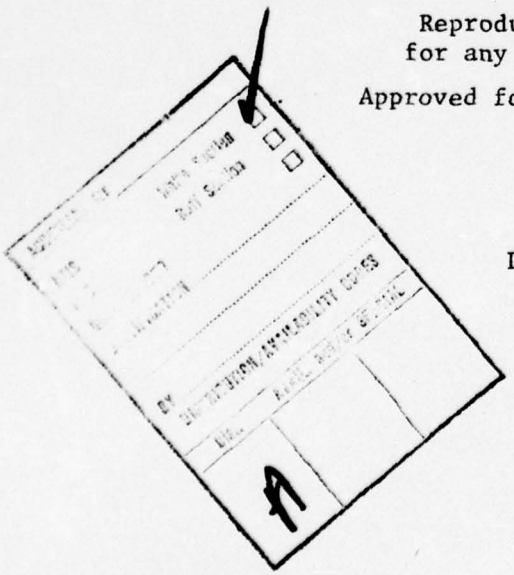
TECHNICAL REPORT NO. 178

September 27, 1976

SUPPORTED BY THE ARMY AND NAVY
UNDER CONTRACT N00014-75-C-0561 (NR-042-002)
WITH THE OFFICE OF NAVAL RESEARCH

Gerald J. Lieberman, Project Director





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

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1. Preview

This study considers optimal decision strategies with regard to capacity expansion in a stochastic demand environment. Chapter 1 is an introduction to the topic at hand and provides the preliminaries upon which later model construction is based. As indicated in the next section, there are two types of facilities which must be considered in a capacity expansion model: permanent and temporary. With regard to temporary facilities, a further classification is necessary according to whether or not the temporary facilities are modular.

Model I, presented in Chapter 2, treats the case where temporary facilities are non-modular. Model II, presented in [17], treats the case where temporary facilities are modular. For both cases, it is shown that the determination of optimal expansion sizes can be expressed as a single-variable, nearly-unconstrained, minimization problem of a very particular form. The solution to this problem is treated in [18], where a revised model is also introduced to demonstrate a number of generalizations that are possible in both Models I and II.

1.2. Permanent and Temporary Facilities

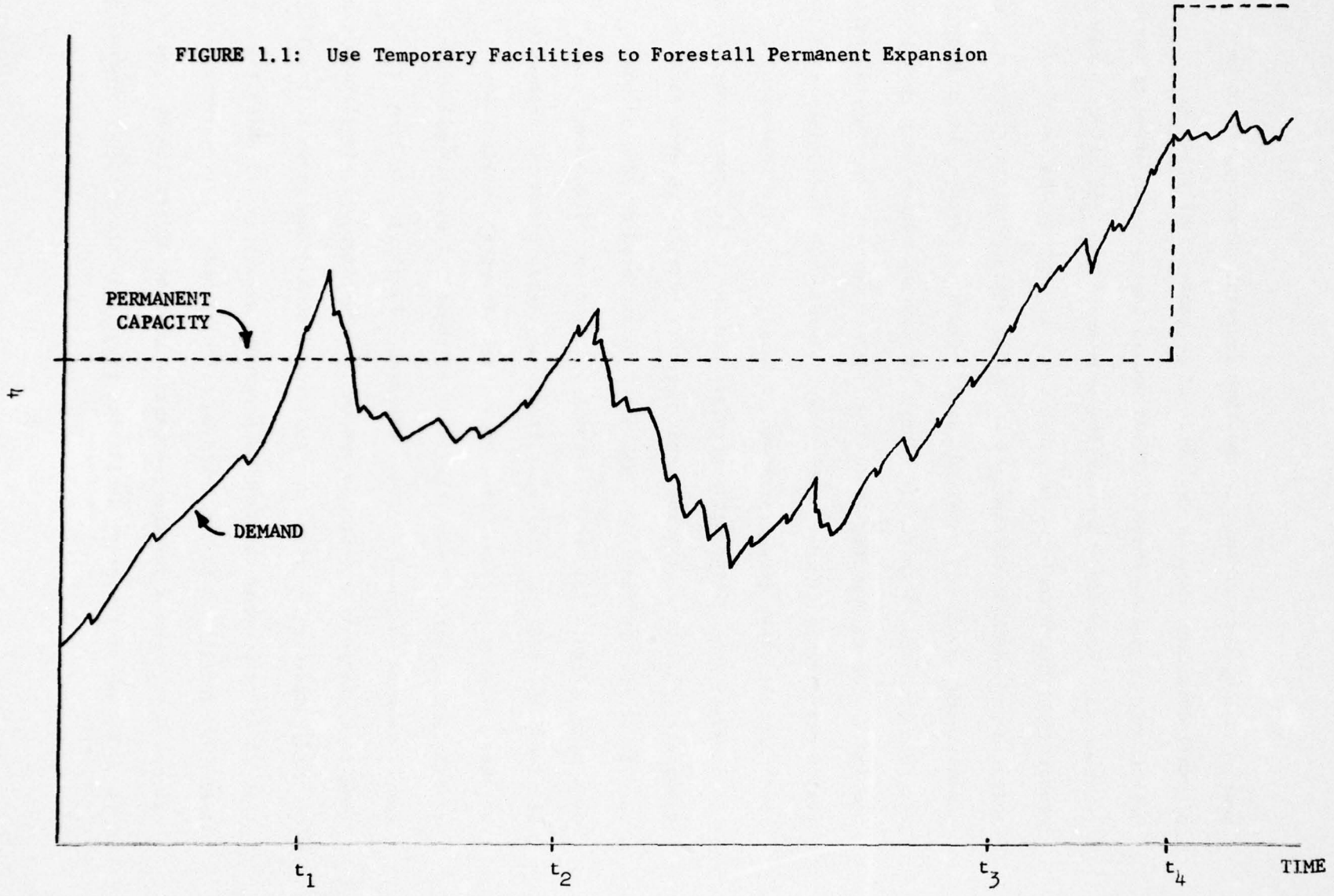
When one considers the growth in capacity of a singular system, such as a plant, pipeline, superhighway or school, a common phenomenon often arises. Specifically, when demand first exceeds the initial capacity of the system, a permanent capacity expansion is not immediately undertaken. Instead, temporary operating measures are instigated in order to satisfy excess demand over the short-run. In some cases, the temporary measure taken may be simply to backlog excess demand. In other cases, where backlogging is undesirable, the present system may be overloaded beyond its planned capacity in order to accommodate excess demand. In still other cases, where backlogging is undesirable and overloading is infeasible, temporary capacity may be rented from outside sources. In order to distinguish between the normal accommodation of demand and the emergency accommodation of excess demand through temporary measures, facilities will be categorized here as either permanent or temporary. That is, "temporary facilities" will denote the measures utilized to accommodate excess demand over the short-run prior to the implementation of a full-fledged expansion of permanent facilities.

The use of temporary facilities usually involves some sort of cost penalty over that of permanent facilities; otherwise, one might well ask which facilities are really temporary. Thus, the economic advantages of temporary facility usage might, on the surface, seem questionable. There are, however, two principal factors which promote the desirability of temporary facility usage. Firstly, there is the problem of uncertainty: when demand first exceeds the permanent capacity, is the excess a short-term

peak or a long-term trend? If the time interval of excess is to be of small duration, then an expansion of permanent facilities is undesirable, since overhead costs on unused capacity can often be very substantial. Secondly, there is the problem of capitalization: permanent capacity expansion costs often exhibit significant economies of scale, which in turn dictate substantial expansion sizes requiring large capital expenditures. Even if sustained demand growth is certain, bankruptcy is possible over the short-run if sufficient revenue cannot be generated from new facilities to meet capitalization costs. Furthermore, capitalization itself may prove difficult until excess demand is of sufficient magnitude to convince investors that permanent capacity expansion is warranted.

While each of the above problems can in itself promote the use of temporary facilities, the two taken together generate an even stronger case for temporary facility usage under an expected discounted cost criterion. Figure 1.1 provides some illustration. The demand peaks at times t_1 and t_2 are short-lived; any large permanent expansion at these times will result in a great deal of wasted capacity over a significant portion of the time interval shown. Temporary measures can be used to accommodate these excesses. Similarly, at time t_3, the long-term prospects for demand are unclear and temporary facilities can be used. However, by time t_4, the excess demand has grown to significant size so as to somehow "warrant" a permanent expansion (at which time, the temporary facilities in use would be discontinued). If temporary facilities are not used, then a permanent expansion must be undertaken at time t_1. Thus, the use of temporary facilities forestalls the permanent expansion

FIGURE 1.1: Use Temporary Facilities to Forestall Permanent Expansion



for (t_4-t_1) time units. If costs are discounted, then the discount savings alone, accruing from the prolongment of the permanent expansion capital outlay, can easily exceed any cost penalties resulting from the use of temporary facilities at times t_1, t_2 and t_3.

In order to make these ideas more precise, let k denote the spares level: the difference between present permanent capacity and demand. Then $-k$ denotes the excess demand whenever demand exceeds permanent capacity. When k is nonnegative, the permanent facilities suffice to serve all demand. However, when k is negative, excess demand exists and temporary facilities must be used. Let K denote the limit on temporary facility usage. Whenever $k = -K$ (that is, excess demand equals K) and a new unit of demand growth occurs, a permanent expansion of size $X+1$ $(X \geq K)$ is required. When permanent expansion occurs, the temporary facilities in use are discontinued and the spares level k increases to $X-K \geq 0$. If demand consists of deterministic constant growth, then the use of this type of recurrent (X,K) expansion policy results in a saw-tooth pattern for the spares level, as illustrated in Figure 1.2.

The value K serves as a "trigger", forcing a permanent expansion whenever a new unit of demand growth occurs while $k = -K$. Thus, $K+1$ represents the point at which excess demand warrants a permanent expansion. In cases of overloading, the value K may be determined based on physical constraints governing the degree of overloading that can safely be incurred. In other cases, K may be determined according to a judgment concerning the willingness of investors to participate in capitalization and the

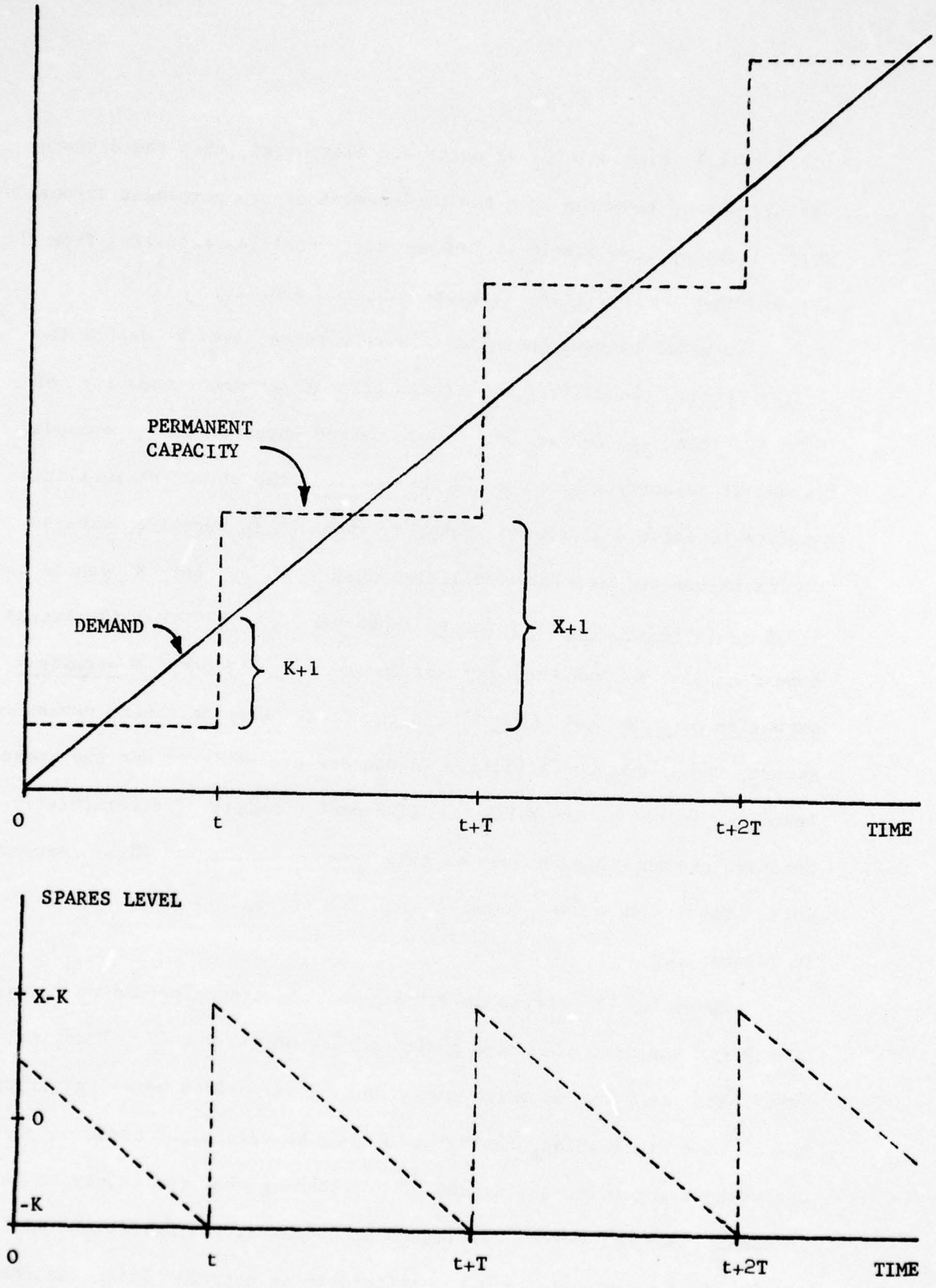


FIGURE 1.2: Recurrent (X,K) Expansion Policy for Linear Demand; Spares Level = Permanent Capacity-Demand.

prospects of generating sufficient revenue to meet capitalization costs, given that excess demand has reached the value K. In such cases where K is predetermined, the parameter of interest will be X*(K)+1, the optimal permanent expansion size to be utilized in conjunction with the temporary facilities usage limit K. In other cases, a value K* will be determined based on a cost minimization over all feasible operating policies (X,K). Given a procedure for determining X*(K), this last problem reduces to that of a cost minimization over all feasible operating policies (X*(K), K).

To summarize, the development of a capacity expansion policy typically involves at least the determination of two parameters: the permanent expansion size X+1 and the temporary facilities usage limit K. The value K answers the question of <u>when</u> to add and the value X answers the question of <u>how much</u> to add.

1.3. Examples

The use of temporary facilities to forestall permanent expansion is prevalent in many sectors at many different levels of the economy. The examples below illustrate the principal types of temporary facilities utilized.

<u>Example 1.1</u>. Backlogging. In these cases, excess demand is simply held in an unsatisfied state until such time that permanent facilities become available, either through permanent expansion or through a decrease in

the demand level actually being served. A backlog penalty is charged, per unit-time, for each demand unit backlogged. This penalty includes the cost of maintaining the backlog and also, in the case of continuous service systems (e.g., utility companies), the revenues lost during the period of backlog.

Example 1.2. Overloading. In these instances, the existing permanent facilities are overloaded beyond their design capacity. Overloading usually places additional strain on the entire system and this increased burden must be accounted for in the form of cost penalties. A specific example of this type is that of a power plant which is utilized beyond its maximal efficiency point in order to accommodate greater demand. In this case, the resulting loss in efficiency must be translated into a monetary overloading cost penalty. Overloading cannot be utilized indefinitely whenever demand growth persists; eventually, a point will be reached where additional demand cannot be safely served. Thus, an upper bound for the parameter K is dictated by physical considerations whenever overloading is in occurs.

Example 1.3. Jobletting. In this case, outside sources are utilized to increase overall capacity. A specific example of this kind is the case of a corporate division which has outgrown its in-house computer facilities. Additional computer time may be bought from either an outside supplier specializing in computer services or possibly from another division within the same parent enterprise. This specific example is an instance of

non-modular temporary facilities, (as are Examples 1.1 and 1.2), because the division can typically rent outside computer time precisely equal to the demand excess. This is also a case where permanent expansion may in fact mean replacement. If the existing division computer configuration is replaced by more-sophisticated machinery, then the expansion cost is the price of the new configuration less any salvage revenue generated by sale of the old configuration. Similarly, the expansion size is the difference in computing power between the new and old configurations.

Example 1.4. Modular Temporary Facilities. This example can be viewed as a specific case of jobletting where temporary facilities can be rented from outside sources in a fixed increment size. An instance of this type is the use of mobile "barracks" facilities (e.g., house trailers) as auxiliary classrooms to alleviate conjestion in overcrowded schools.

Another instance of this type is the use of "pair-gain" devices to augment the capacity of over-subscribed telephone cables. To illustrate, consider Figure 1.3. In order to simplify the illustration, consider only the one-way communication from a subscriber telephone to the switching equipment at the subscriber's local telephone office. As shown in Figure 1.3a, this communication is normally implemented by the dedication of a single wire-strand from the subscriber telephone to the office. This single wire is typically one strand of a telephone cable between the subscriber's neighborhood and the office. Suppose that the demand for telephones in this neighborhood increases to a point at which all wire-strands of the existing cable are used. The telephone company, being

TELEPHONE PAIR-GAIN SYSTEMS

FIGURE 1.3a: Wire-Only

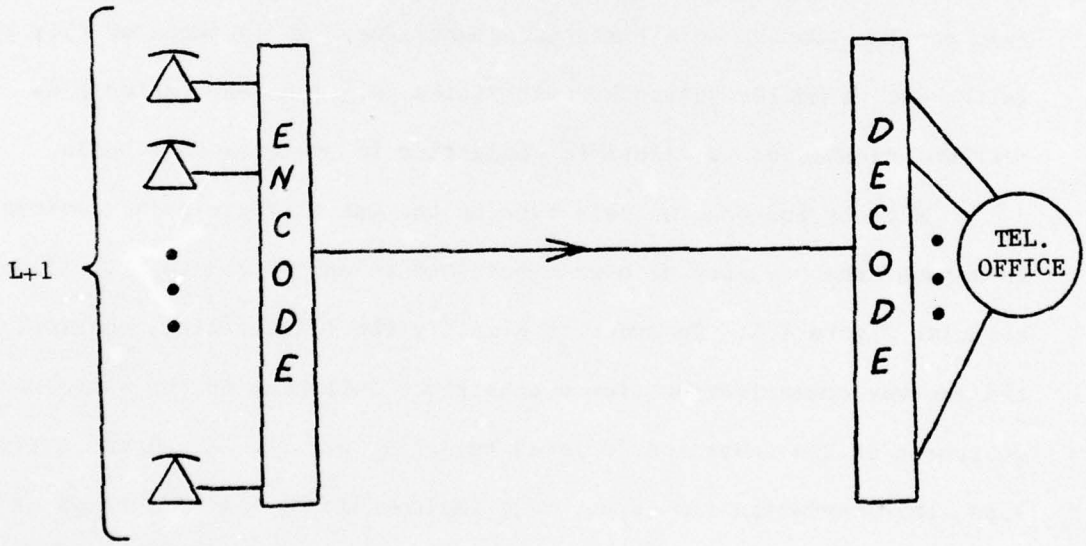


FIGURE 1.3b: Pair-Gain L:1

required to provide service to all soliciting subscribers, must somehow increase the cable capacity in order to satisfy new demands. One obvious possibility is the immediate placement of a new cable (i.e., an immediate permanent expansion). Another possibility is the implementation of multiple-party service; however, this alternative is usually unexceptable to consumers and is, in fact, not used in many areas. Another alternative to the placement of an additional cable is the utilization of pair-gain devices on the existing cable. As illustrated in Figure 1.3b, a pair-gain device consists of electronics that permit the transmission of a number of simultaneous conversations over a single wire-strand by encoding the conversations at one end and decoding the conversations at the other end. As shown in the figure, these devices are typically available in a fixed modular size allowing the transmission of L additional conversations over a single wire-strand; the "pair-gain ratio" of each such module is said to be "L to 1".

A cost characteristic which frequently distinguishes modular temporary facilities from non-modular temporary facilities is the existence of instaneous charges over and above the normal rental costs. Specifically, an installation charge is typically incurred whenever an additional module is engaged and similarly, a removal charge is incurred whenever a module is returned to the outside supplier. Thus, in the case of modular temporary facilities, there can be a sub-optimization problem with regard to how modules should be engaged and returned in order to avoid excessive installation and removal charges.

1.4. The Poisson Demand Model

In order to account for the promotion of temporary facility usage due to uncertainty, demand will be recognized as being stochastic in nature. Specifically, it will be assumed that demand increases ("arrivals") are characterized by a Poisson process [15] at rate $\lambda_1 > 0$. Similarly, it will be assumed that demand decreases ("departures") are characterized by a Poisson process at rate $\lambda_2 > 0$. Furthermore, the arrival and departure processes will be presumed to be independent of each other and also independent of the expansion policy chosen.

Given the above demand characterization, define an "event" as either an arrival or a departure. Then an equivalent, and more convenient, characterization is as follows: events occur according to a Poisson process at rate $\lambda = \lambda_1 + \lambda_2$; the probability that an event is an arrival equals $p = \lambda_1/\lambda$; and the probability that an event is a departure equals $q = \lambda_2/\lambda$ ($q = 1-p$) [15]. That is, the demand process treated here can be viewed as a random walk with exponential inter-event times.

It should be noted that the above demand characterization is pointed toward systems providing a homogeneous continuous service to a fairly stable clientele, such as a power plant or a telephone service area. It is really not intended to model actual clientele movement for cases where customers are rapidly flowing in and out, such as a job shop. In these cases, it is suggested that "customer demand" be viewed as the maximum usage level required of facilities during short intervals of time (e.g., weeks or months).

It should also be noted that the demand process is assumed time-stationary. For this reason, the system being modelled must be in the midst of a relatively stable period of either sustained growth or sustained negative growth.

Finally, notice that the arrival and departure rates are presumed constant with regard to the demand level. From a practical point of view, this is convenient since only two estimates -- a single arrival rate and a single departure rate -- are necessary to implement the model. However, in most realistic situations, it appears that the arrival and departure rates will vary according to the system demand level. In recognition of this fact, the revised model of [18] permits these rates to vary.

1.5. Costs

The optimization criterion used here will be the minimization of all future expected discounted costs; r will denote the continuous discount rate, $0 < r < 1$. Three types of costs will be allowed: permanent expansion costs, permanent facility operating charges and temporary facility charges. All costs are assumed to be time-stationary.

Permanent expansion costs will be denoted by $g(\cdot)$, where $g(X)$ is the cost of a permanent expansion of size $X+1$. Notice that $g(\cdot)$ is presumed to be a function only of the expansion size, independent of the level of existing permanent capacity and the value of the temporary facilities usage limit, K. This restriction is relaxed in the revised model of [18].

It is important to note that $g(0)$ represents the cost of a unit expansion. Thus, $g(0)$ is comprised of the fixed (i.e., "setup") expansion cost plus the marginal cost for the first unit of increased capacity. Hence, any presumption regarding a specific functional form (e.g., convexity) for g over $[0,\infty)$ will not preclude the existence of a fixed expansion cost.

With regard to other costs, the temporary facility charges are of principal interest in this study. In order to provide sufficient generality with regard to these costs, the permanent facility operating charges will be presumed to be proportional to the amount of permanent capacity used. The following theorem indicates the modelling simplifications that result from this assumption.

Theorem 1.1. Suppose that permanent facility operating charges are proportional to the amount of permanent capacity used, at rate γ per unit-time. Then an equivalent cost model is given as follows:

(i) When permanent facilities satisfy all demand, no operating charges are incurred.

(ii) When temporary facilities are necessary (i.e., $k < 0$) marginal costs equal to the temporary facility charges, less an adjustment $\gamma|k| = -\gamma k \geq 0$ per unit-time, are incurred.

Proof. Consider an arbitrary demand pattern and an arbitrary expansion policy. Let $\mathcal{D}(t)$ denote the demand at time t and let $k(t)$ denote the spares level at time t, $t \geq 0$. Let $\chi(\cdot)$ denote the 0-1 nonnegativity

indicator function: $\chi(k) = 0$ if $k < 0$ and $\chi(k) = 1$ if $k \geq 0$. Let $\bar{\chi}(\cdot)$ denote the complement of $\chi(\cdot)$: $\bar{\chi}(\cdot) = 1 - \chi(\cdot)$. Finally, let $\mathcal{F}$ denote all future discounted expansion costs and temporary facility charges. Then the total discounted cost C is given by

$$C = \mathcal{F} + \int_0^\infty \{\chi(k(t))\, \mathcal{D}(t)\, \gamma + \bar{\chi}(k(t))\, [\mathcal{D}(t) - |k(t)|]\gamma\}\, e^{-rt}\, dt \qquad (1.1)$$

$$= \mathcal{F} + \gamma \int_0^\infty \mathcal{D}(t)\, e^{-rt}\, dt - \gamma \int_0^\infty \bar{\chi}(k(t))\, |k(t)|\, e^{-rt}\, dt \quad . \qquad (1.2)$$

The first term of the integral in (1.1) represents the permanent facility operating charges incurred when the spares level $(k(t))$ is nonnegative; the second term represents the permanent facility operating charges incurred when $-k(t) > 0$ customers must be served by temporary facilities.

The first integral of (1.2) is constant upon taking expectations over the probability distribution of $\mathcal{D}(\cdot)$ (independent of the expansion policy chosen). The second integral of (1.2) can be accounted for in the expected discounted cost minimization by subtracting an adjustment $\gamma|k| = -\gamma k \geq 0$ per unit-time whenever $k(t) < 0$ (i.e., whenever $\bar{\chi}(k(t)) = 1$). □

The temporary facility charges will be assumed to be dependent on the value of the spares level k and the value of the temporary facilities usage limit K, but independent of the level of existing permanent capacity. Given these cost assumptions, Theorem 1.1 will allow the construction of Models I and II to proceed independent of the actual level of existing permanent capacity.

The presumed independence of the permanent operating costs from the permanent facility capacity may not be that restrictive, since proportional fixed operating charges which depend solely on the permanent capacity level can be included in the expansion cost function g. To illustrate, suppose that the actual permanent operating costs include a proportional fixed charge of ψ per unit-time for each unit of permanent capacity available. Let Y_0 denote the initial permanent capacity. The expected discounted cost resulting from the ψ charge on the initial capacity is then $\int_0^\infty Y_0 \psi\, e^{-rt}\, dt = \psi r^{-1} Y_0$; this cost is independent of the expansion policy chosen. The expected discounted costs resulting from the ψ charge on future capacity expansion increments can then be accounted for by defining g as

$$g(X) = g_e(X) + \int_0^\infty (X+1)\psi\, e^{-rt}\, dt$$

$$= g_e(X) + \psi r^{-1}(X+1) ,$$

where g_e represents the actual expansion function alone.

All restrictions (except time-stationarity) on the facility charges, both permanent and temporary, are also relaxed in the revised model of [18].

1.6. Related Results and the Inventory Analogy

In the area of capacity expansion, finite-horizon models for deterministic demand growth are most prevalent. The models of Manne and Veinott [13] and Erlenkotter [3] are representative of deterministic approaches employing dynamic programming techniques. The models of Bergendahl [1] and Rogers [14] are representative of deterministic approaches using mathematical programming solution techniques. [12] is a comprehensive reference on several deterministic models.

The study of Manne [11] constitutes the first capacity expansion model recognizing stochastic demand. In the Manne model, demand growth is assumed to behave according to a Weiner diffusion process. Expansion costs are given by a concave power function:

$$g(X) = aX^b, \qquad a > 0,\ 0 < b < 1 . \tag{1.3}$$

The model alsc permits backlogs at a proportional penalty rate. Utilizing the Laplace transform of the presumed demand distribution, Manne derives an integral equation in X and K. This equation is then solved numerically for all feasible pairs (X,K) in order to obtain minimum expected discounted cost parameters (X*,K*).

More recently, stochastic capacity expansion models for telephone cable and pair-gain analysis (see Example 1.4 of Section 1.3) have been reported in related papers by Freidenfelds [5], [6], and Koontz and Shipley [10]. Freidenfelds utilizes the Poisson demand model used here.

The Freidenfelds model corresponds to Example 2.5 of Chapter 2 for the special cases of $K = 0$ and $K = L$ when the expansion cost function is given by

$$g(X) = aX + b \quad . \tag{1.4}$$

By considering the Laplace transforms of first-event times (see Section 2.3), Freidenfelds derives a functional in X for the expected total discounted cost. This functional is then minimized by function iteration (see Section 4.4). Freidenfelds' results were the stimuli for the author's own investigations, the preliminary findings of which were first reported in [10].

In [10], the author presents a model corresponding to Example 2.5 of Chapter 2 when the expansion cost function is given by (1.4). Also in [10], Koontz reports steady-state results for the same problem when demand is generalized to a birth-and-death process and the minimization objective is average cost per unit-time.

Other related models can be found in the area of single-item inventory theory. Manne [11] was the first to note the similarity between the recurrent (X,K) expansion policies, as treated here, and (S,s) inventory policies. Figure 1.4 illustrates the near-analogy. The spares level (k) corresponds to inventory on-hand; a nonnegative spares level corresponds to an absence of inventory backlogs and a negative spares level corresponds to the existence of backlogs. The temporary

SINGLE-ITEM INVENTORY ANALOGY

Spares Level (k)	———	Inventory On-Hand
$k \geq 0$	———	No Backlogs
$k < 0$	———	$\|k\|$ Backlogs
Temporary Facility Usage Limit (K)	———	Backlog Limit
Arrivals	———	Item Demands
Departures	———	Item Returns
Expansion Cost (g)	———	Ordering Cost
Temporary Facility Charges	———	Backlog Penalties
Permanent Facility Charges	—— NO! ——	Inventory Holding Charges

FIGURE 1.4

facilities usage limit (K) corresponds to an inventory backlog limit (i.e., $K+1 = -s$). The arrival process in the expansion model corresponds to item demands in the inventory model, while the departure process corresponds to item returns. The capacity expansion function is analogous to the inventory ordering (or production) cost function (i.e., $X+1 = S-s$) and the temporary facility charges correspond to inventory backlog penalties.

Unfortunately, the analogy is deficient with regard to permanent facility operating charges and inventory holding charges. In the absence of intervening expansion, the permanent facility operating charges should realistically decrease as the spares level (k) increases, since fewer permanent facilities are used. But, in the absence of intervening orders, the inventory holding charges should realistically increase as the on-hand inventory (k) increases. Thus, the realistic behavior of permanent facility operating charges and inventory holding charges are precisely opposite, and the analogy fails. However, if the permanent facility operating charges are constant with regard to k, then the analogy follows for single-item inventory models (with returns) having constant inventory holding costs, independent of the level of on-hand inventory. Referring to Theorem 1.1, it follows that the inventory analogy also holds (with inventory holding costs zero) whenever the permanent facility operating charges are proportional to the permanent capacity utilized, provided that the proper adjustments are made to the temporary facility charges (i.e., the backlog penalties). To illustrate, suppose that the permanent

facility charges are proportional at rate γ per unit-time and that the temporary facility charges are also proportional at rate γ' per unit-time, for each unit of temporary capacity utilized. Then, employing Theorem 1.1, the correct inventory analogy is inventory holding costs zero and backlog penalty equal to $(\gamma'-\gamma)$ per unit-time, for each item backlogged. Thus, under the hypothesis of Theorem 1.1, the analogy holds for single-item inventory models with returns and backlogs, provided that the inventory holding costs are set to zero and the backlog penalty is properly assigned.

Although the literature for single-item inventory models is quite extensive (see [20]), the literature for single-item inventory models <u>with returns</u> is not. In fact, the literature for single-item inventory models with both returns <u>and</u> backlogs is apparently null. Stochastic inventory models permitting returns include those of Tainiter [19], Whisler [21], and Heyman and Hoadley [7], [8]. Although the demand processes of these models are similar to the Poisson model employed here, backlogs are <u>not</u> permitted. Furthermore, <u>all</u> costs are assumed proportional, with the exception of [21] where convex ordering functions are permitted over discrete time periods. Because of the simplifications promoted by proportional costs and the absence of backlogs, the techniques employed to analyse these models are quite different than those utilized here.

1.7. The Untruncated Demand Assumption

As noted earlier, the demand process is characterized as the difference between two simple independent Poisson processes. As with any mathematical model, the practitioner must proceed with caution prior to implementing the models introduced here in an actual application. However, there is one assumption made here regarding the demand process that will always differ from the actual case. Namely, the demand process is not entirely legitimate because it allows negative demands. That is, depending on the initial level of demand $\mathcal{D}(0)$ and the arrival and departure rates, there is some probability that customers will depart when there are no customers in the system! This anomaly of the process will be referred to as the untruncated assumption, reflecting the fact that the models do not truncate departures when demand becomes zero. The purpose of this section is to demonstrate that the untruncated assumption will not significantly affect the derived optimal operating parameters in most cases and to identify the few cases where it might adversely distort results. With regard to these later cases, the practitioner might well consider a revised model on the order of that introduced in [5], which provides truncation.

The reason for making the untruncated assumption is one of mathematical simplicity. Coupled with other basic assumptions, the untruncated assumption will allow model constructions that need not explicitly account for the total level of permanent capacity available and the total system demand level. Mathematically, the models can then

be characterized by a single set of equations, as opposed to a separate set of equations for each possible level of permanent capacity.

The models of Manne [11] and Freidenfelds [6] also possess this anomaly (the anomaly is not present in the inventory analogy). An argument was presented in [11] that purported to demonstrate that the untruncated assumption did not affect the optimal operating parameters at all. Briefly, the argument can be paraphrased as follows: "since no costs are incurred when demand is negative, the total costs of the untruncated model will be identical to those of the actual truncated case. Therefore, the operating parameters derived from the untruncated model will be optimal for the actual truncated case." Concerning the case at hand, observe that the spares level (k) will be positive whenever demand becomes negative. Hence, in view of Theorem 1.1, it can be assumed that no costs are incurred whenever demand becomes negative in the models considered here. Thus, the Manne argument, if valid, would also hold for the model at hand.

Unfortunately, the argument put forth by Manne has two flaws. Firstly, once demand becomes negative, there is some probability that it will remain so thereafter; for these sample paths, the total cost in the untruncated model <u>can</u> differ from the actual truncated situation. Secondly, even though total costs may be the same for many of these sample paths, total <u>discounted</u> costs (the criteria used both here and in [11]) may not be. In fact, for those sample paths where demand becomes negative, future expenditures occur later (if at all) than in the actual truncated case. Hence, the total discounted cost for these

sample paths can underestimate the actual case, and consequently, the expected total discounted cost in the models can be a conservative estimate of the actual situation. It will now be demonstrated that this estimate is quite adequate for most cases.

As shown in [6], the expectation and variance of demand $\mathcal{D}(\cdot)$ are given by

$$E[\mathcal{D}(t)] = \mathcal{D}(0) + \lambda t(p-q), \qquad t \geq 0 \tag{1.5}$$

and

$$Var[\mathcal{D}(t)] = \lambda t , \qquad t \geq 0 . \tag{1.6}$$

Thus, when $p > q$, the mean demand level increases with time; this will be referred to as the growth situation. When $p < q$, the mean demand level decreases with time; this will be referred to as the negative growth situation. The case $p = q$ will be referred to as the stable situation. The events of interest, with regard to the untruncated assumption, are

$$\mathcal{E}(t) = \{\mathcal{D}(s) < 0 \text{ for some } s \in [0,t]\}, \qquad t \geq 0 .$$

First consider $P\{\mathcal{E}(\infty)\}$, the probability of demand eventually becoming negative. Given the independent characterization (λ,p) introduced in Section 1.4 for the demand process, it should be apparent that this probability is equal to the probability that a random walk (p,q) starting at the

origin ever reaches the position $-(\mathcal{I}(0) + 1)$. From [4], this probability is

$$P\{\mathcal{E}(\infty)\} = \begin{cases} (q/p)^{\mathcal{I}(0)+1} , & p > q \\ 1 , & p \leq q \end{cases} .$$

That is, the probability of demand eventually becoming negative is $(q/p)^{\mathcal{I}(0)+1}$ for the growth situation; for other cases, the event is certain. Thus, if $\mathcal{I}(0)$ is sufficiently large in the growth situation, then the proportion of sample paths exhibiting negative demands will be small and the untruncated assumption should be viable. In fact, since no costs are incurred while the spares level is positive (Theorem 1.1), it suffices to assume that $\mathcal{I}(0)$ equals the total permanent capacity initially available. Thus, if the initial permanent capacity is sufficiently large in the growth situation, then the untruncated assumption should be viable. To illustrate, for $p = .52$ $(q = .48)$ and $\mathcal{I}(0) = 100$, $P\{\mathcal{E}(\infty)\} \approx 3.3 \times 10^{-4}$. Therefore, the untruncated assumption does not appear unreasonable for most growth situations where there is some initial permanent capacity to begin with.

Although $P\{\mathcal{E}(\infty)\} = 1$ for non-growth cases, the untruncated assumption is still reasonable for many of these situations. This fact results from two basic observations:

(i) if the discount rate is not abnormally small, then a large portion of the total expected discounted cost occurs over the finite interval $[0,t]$; and

(ii) of those sample paths contributing to the event $\mathcal{E}(t)$, only a portion of these paths will exhibit discounted costs which differ from their truncated counterparts over $[0,t]$.

Observation (i) follows from the fact that the models are time-stationary. Since the costs and the demand process do not change over time, the expected expenditures over each time interval of length t will be approximately equal, for t sufficiently large. Let Δ denote the expected discounted costs over $[0,t]$. Then the total expected discounted costs over $[0,\infty)$ can be approximated by $\Delta(1 + e^{-rt} + e^{-r2t} + \cdots) = \Delta(1 - e^{-rt})^{-1}$. Thus, the expected discounted costs incurred over $[0,t]$ will account for approximately $(1 - e^{-rt}) \times 100\%$ of the total expected discounted costs. To illustrate, for $t = 10$ and $r = .15$, $(1 - e^{-rt}) \times 100\% = 78\%$; for $t = 20$ and $r = .10$, $(1 - e^{-rt}) \times 100\% = 86\%$.

Concerning observation (ii), let

$$\tau = \min\{s : \mathcal{S}(s) < 0\} \quad .$$

Then $\mathcal{E}(t) = \{\tau \leq t\}$. As noted earlier, it suffices to assume that $\mathcal{S}(0)$ equals the initial permanent capacity level. At time τ, the spares level in the actual untruncated case is equal to the permanent capacity level; a lower bound on the permanent capacity level at time τ is $\mathcal{S}(0)$. By Theorem 1.1, no additional expenditures will occur until the spares level reaches zero; a necessary (but by no means sufficient) condition for this to happen is that at least $\mathcal{S}(0)$ additional arrivals

occur past τ. Hence, of those sample paths contributing to the event $\mathcal{E}(t)$, only those exhibiting at least $\mathcal{N}(0)$ additional arrivals during the interval $(\tau,t]$ can possibly have discounted costs which differ from the actual truncated case during the interval $[0,t]$. Let $N_a(s)$ denote the total number of arrivals during the interval $[0,s]$. Then an upper bound on the proportion of sample paths exhibiting discounted costs which differ from the actual truncated situation over $[0,t]$ is given by $U(t) = P\{\tau \leq t,\ N_a(t) - N_a(\tau) \geq \mathcal{N}(0)\}$.

Since the arrival process is Poisson at rate λp,

$$P\{N_a(t) - N_a(\tau) \geq \mathcal{N}(0) \mid \tau = s \leq t\} = \sum_{n=\mathcal{N}(0)}^{\infty} \frac{\{\lambda p(t-s)\}^n}{n!} e^{-\lambda p(t-s)} .$$

Therefore, letting f and F respectively denote the probability density and cumulative distribution functions of τ,

$$U(t) = \int_0^t f(s) \sum_{n=\mathcal{N}(0)}^{\infty} \frac{\{\lambda p(t-s)\}^n}{n!} e^{-\lambda p(t-s)}\, ds$$

$$\leq F(t) = P\{\mathcal{E}(t)\} . \qquad (1.7)$$

No attempt will be made here to actually compute $U(\cdot)$. Instead, $F(\cdot)$ will be estimated using a normal approximation. The significance of (1.7) is that the approximation for $F(\cdot)$ should be entirely adequate since, for $\mathcal{N}(0)$ sufficiently large, $U(\cdot)$ will be many orders of magnitude smaller than $F(\cdot)$ (recall also that $U(\cdot)$ itself is a rather loose upper bound on the actual relevant probabilities).

The following lemma discloses the form of F and provides further justification for the validity of the normal approximation to be used.

Lemma 1.2. Let $m = \mathcal{D}(0) + 1$. Then

$$F(t) = \sum_{j=m}^{\infty} \frac{m}{j} \binom{m}{(m+j)/2} p^{(m-j)/2} q^{(m+j)/2} \sum_{n=j}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad t \geq 0 \quad (1.8)$$

(Note: $\binom{m}{(m+j)/2} = 0$ if the parities of m and j differ.)

Proof: Let $N(t)$ denote the number of events (arrivals and departues) during the time interval $[0,t]$. Since $\{N(t), t \geq 0\}$ is Poisson,

$$P\{N(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \qquad t \geq 0 \qquad (1.9)$$

Given $N(t) = n$, we seek the probability that the demand becomes negative during the first n events. Let $P\{\mathcal{E}_j\} = P\{$demand first becomes negative at jth event$\}$. Obviously, $P\{\mathcal{E}_j\} = 0$ for $j < m$, since at least m departures are required. As shown in [4] (pages 351-352),

$$P\{\mathcal{E}_j\} = \frac{m}{j} \binom{m}{(m+j)/2} p^{(m-j)/2} q^{(m+j)/2}, \quad j \geq m \ . \qquad (1.10)$$

The derivation of (1.10) follows from the fact that in order for demand to first become negative at the jth event, there must be precisely

$(m-j)/2$ arrivals, $(m+j)/2$ departures, and the demand must be nonnegative at all previous event epochs.

Thus, from (1.9) and (1.10),

$$F(t) = P\{\mathcal{E}(t)\}$$

$$= \sum_{n=m}^{\infty} \left(\sum_{j=m}^{n} P\{\mathcal{E}_j\} \right) P\{N(t) = n\}$$

$$= \sum_{j=m}^{\infty} P\{\mathcal{E}_j\} \sum_{n=j}^{\infty} P\{N(t) = n\}$$

$$= \sum_{j=m}^{\infty} P\{\mathcal{E}_j\} \, P\{N(t) \geq j\}$$

$$= \sum_{j=m}^{\infty} \frac{m}{j} \binom{m}{(m+j)/2} p^{(m-j)/2} q^{(m+j)/2} \sum_{n=j}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad . \qquad \square$$

Notice that (1.8) is in essence of partial convolution of binomial (almost) and Poisson probabilities. Therefore, for m (i.e., $\mathcal{D}(0)$) sufficiently large, $\lambda t \geq 10$, and neither p nor q close to zero, a normal approximation to (1.8) should be applicable [2].

In order to implement a normal approximation, the mean and variance of τ are required. For $p = q$, the mean of τ is infinite. However, it is finite for the negative growth situation.

Theorem 1.3. For $p < q$,

$$E[\tau] = \frac{\mathcal{Q}(0)+1}{\lambda(q-p)} \quad , \tag{1.11}$$

and

$$Var[\tau] = \frac{\mathcal{Q}(0)+1}{\lambda^2(q-p)^3} \quad . \tag{1.12}$$

Proof. [6]. Denote

$$\tau_i = \inf\{t \geq 0 : \mathcal{Q}(t) - \mathcal{Q}(0) = i\}, \qquad i = 0, -1, -2, \ldots .$$

Then, $\tau = \tau_{-(\mathcal{Q}(0)+1)}$. Let $f_i(t)$ denote the probability density function of τ_i. Then, by conditioning on the time (s) and the type (arrival or departure) of the first event, the following integral difference equation can be written for $f_i(t)$:

$$f_i(t) = \int_0^t \lambda e^{-\lambda s}\{pf_{i-1}(t-s) + qf_{i+1}(t-s)\}ds \quad , \quad i \leq -1 \quad , \tag{1.13}$$

where

$$f_0(t) = \begin{cases} 1, & t = 0 \\ 0, & t > 0 \end{cases} \quad .$$

Let $h(i)$ denote the Laplace transform of τ_i; using (1.13),

$$h(i) = E[e^{-r\tau_i}] = \int_0^\infty f_i(t)\, e^{-rt}\, dt$$

$$= \int_0^\infty e^{-rt} \int_0^t \lambda e^{-\lambda s}\{pf_{i-1}(t-s) + qf_{i+1}(t-s)\}ds\, dt$$

$$= \lambda \int_0^\infty e^{-\lambda s} \int_s^\infty e^{-rt}\{pf_{i-1}(t-s) + qf_{i+1}(t-s)\}dt\, ds$$

$$= \lambda \int_0^\infty e^{-\lambda s}\, e^{-rs} \int_s^\infty e^{-r(t-s)}\{pf_{i-1}(t-s) + qf_{i+1}(t-s)\}dt\, ds$$

$$= \lambda \int_0^\infty e^{-(\lambda+r)s}\{ph(i-1) + qh(i+1)\}ds$$

$$= \lambda(\lambda+r)^{-1}\{ph(i-1) + qh(i+1)\}\,, \qquad i \le -1, \tag{1.14}$$

where $h(0) = 1$.

The solution to difference equations (1.14) is given in [4] and [9] as

$$h(i) = v^i\,, \qquad i \le 0$$

where

$$v = \frac{(\lambda+r) + \sqrt{(\lambda+r)^2 - 4pq\lambda^2}}{2q\lambda}\,. \tag{1.15}$$

Using (1.15),

$$E[\tau_i] = -\frac{d}{dr}\, h(i)\Big|_{r=0} = -iv^{i-1}\frac{dv}{dr}\Big|_{r=0}\,,$$

where

$$\left.\frac{dv}{dr}\right|_{r=0} = \left.\frac{1 - \dfrac{(\lambda+r)}{\sqrt{(\lambda+r)^2 - (1-(p-q)^2\lambda^2)}}}{\lambda(1-(p-q))}\right|_{r=0} = -\frac{1}{\lambda(p-q)} \quad .$$

Thus,

$$E[\tau_i] = \frac{i}{\lambda(p-q)} , \qquad i \leq 0 ,$$

which verifies (1.11) upon substituting $i = -(\ell(0) + 1)$.

To find the variance, we again use (1.15):

$$E[\tau_i^2] = \left.\frac{d^2}{dr^2} h(i)\right|_{r=0}$$

$$= i \left\{ v^{i-1} \frac{d^2 v}{dr^2} + (i-1) v^{i-2} \left(\frac{dv}{dr}\right)^2 \right\}_{r=0} ,$$

where, denoting $\epsilon = (\lambda+r)^2 - (1 - (p-q)^2\lambda^2)$,

$$\left.\frac{d^2 v}{dr^2}\right|_{r=0} = \frac{-1}{\lambda(1-(p-q))} \left\{ \frac{\sqrt{\epsilon} - \dfrac{(\lambda+r)^2}{\sqrt{\epsilon}}}{\epsilon} \right\}_{r=0}$$

$$= \frac{1}{\lambda(1-(p-q))} \frac{(1 - (p-q)^2)\lambda^2}{\{\lambda^2(p-q)^2\}^{3/2}} = \frac{1 + (p-q)}{\lambda^2(p-q)^3} \quad .$$

Hence,

$$E[\tau_i^2] = i\left\{\frac{1+(p-q)}{\lambda^2(p-q)^3} + \frac{i-1}{\lambda^2(p-q)^2}\right\} = \frac{i}{\lambda^2(p-q)^3} + \frac{i^2}{\lambda^2(p-q)^2} \quad .$$

Finally,

$$\mathrm{Var}[\tau_i] = E[\tau_i^2] - (E[\tau_i])^2 = \frac{i}{\lambda^2(p-q)^3} \quad ,$$

where verifies (1.12) upon substitution of $i = -(\mathcal{A}(0)+1)$. □

Using the above theorem, the normal approximation for $F(t)$ is given by

$$F(t) \approx \Phi\left(\frac{t - \frac{\mathcal{A}(0)+1}{\lambda(q-p)}}{\sqrt{\frac{\mathcal{A}(0)+1}{\lambda^2(q-p)^3}}}\right) = \tilde{F}(t)$$

$$= \Phi\left(\frac{\lambda t(q-p)^{3/2} - (q-p)^{1/2}(\mathcal{A}(0)+1)}{(\mathcal{A}(0)+1)^{1/2}}\right) \quad , \qquad q > p$$

$$= \Phi\left(\frac{\lambda t(1-2p)^{3/2} - (1-2p)^{1/2}(\mathcal{A}(0)+1)}{(\mathcal{A}(0)+1)^{1/2}}\right) , \qquad p < .5 \ , \ (1.16)$$

where Φ denotes the cumulation standard normal distribution function (also recall that $q = 1-p$). Using standard normal tables, the right hand side of (1.16) will not exceed .01 whenever the argument does not exceed -2.326. That is, whenever

$$\lambda t(1-2p)^{3/2} - (1-2p)^{1/2}\,(\mathcal{D}(0)+1) + 2.326\,(\mathcal{D}(0)+1)^{1/2} \leq 0 \quad . \qquad (1.17)$$

Using the quadratic formula, (1.17) will be satisfied whenever

$$\mathcal{D}(0) \geq 1 + \frac{\left[2.326 + \sqrt{5.410 + 4\lambda t(1-2p)^2}\right]^2}{4(1-2p)} . \qquad (1.18)$$

It is a simple matter to see that (1.18) increases with λ and t, and decreases with p. Table 1.1 is a small tabulation of (1.18) for $\lambda = 100$, $.35 \leq p \leq .45$, and $t = 10$ and 20.

		t	
	$\lambda = 100$	10	20
	.35	385	715
p	.40	290	520
	.45	207	336

Table 1.1: Lower Bounds on $\mathcal{D}(0)$ for $\tilde{F}(t) \leq .01$

To illustrate, for $p \geq .4$, Table 1.1 indicates that if $\mathcal{D}(0) \geq 520$ and $\lambda \leq 100$, then less than 1% of the untruncated sample points will exhibit negative demands over the first 20 years (and approximately 86% of the total discounted costs are determined in the first 20 years if $r \geq .1$). To see that Table 1.1 is not unreasonable, consider the statistic $\xi = \lambda/\mathcal{D}(0)$. ξ represents the amount of expected "churning" or "turn-over" in the system relative to the initial demand $\mathcal{D}(0)$. That

is, ξ approximates the proportion of initial demand involved with transactions (arrivals and departures) during a single year. Thus, $\mathcal{D}(0) \geq 520$ for $\lambda = 100$ requires that expected annual turnover must be somewhat less than 20% in order to insure that $\tilde{F}(20) \leq .01$.

Therefore, in nongrowth situations, the untruncated assumption appears to be viable for situations where $\mathcal{D}(0)$ is sufficiently large relative to the annual event rate λ; that is, when annual turnover is sufficiently small. This must be the case in most nongrowth situations where future permanent expansion is a realistic possibility. For, if λ is large relative to $\mathcal{D}(0)$ and $p < q$, then $E[\mathcal{D}(t)]$ tends toward 0 rather rapidly according to (1.5); in this case, future capacity contraction -- not expansion -- is the alternative requiring investigation. A similar remark holds for cases where p is small (e.g., $p < .35$).

In conclusion, the untruncated demand assumption appears to be viable in most cases where $\mathcal{D}(0)$ is sufficiently large and future capacity expansion is a realistic possibility. For growth situations ($p > q$), the assumption will be viable for nearly all applications with $\mathcal{D}(0) > 0$ (i.e., existing initial permanent capacity). For nongrowth situations, the assumption appears to be viable for applications where additional permanent expansion is a realistic future possibility.

1.8. Notation

As introduced previously in this chapter, $\mathcal{D}(t)$ will denote the demand at time $t \geq 0$, λ will denote the event rate with arrival probability p and departure probability q, k will denote the spares level, K will denote the temporary facilities usage limit, $X+1$ will denote the expansion size and g will denote the permanent expansion cost function, where $g(X)$ represents the cost for an expansion of size $X+1$.

$\mathbb{R}^m$ will denote the space of all real vectors having m components, $m \geq 1$. Only row vectors will be utilized. Given $U,V \in \mathbb{R}^m$, the dot product $U \cdot V$ will be used to denote vector multiplication. $R^{m \times n}$ will denote the space of all real matrices having m rows and n columns.

CHAPTER 2

MODEL I: THE NON-MODULAR CASE

2.1. Introduction

This chapter characterizes the untruncated constant Poisson model under two basic assumptions:

Assumption 2.1. Permanent facility operating charges are proportional to the amount of permanent capacity used, at rate γ_1 per unit-time.

Assumption 2.2. Permanent facilities are always used in preference to temporary facilities and temporary facility charges are incurred only when there is excess demand to be served (i.e., $k \leq 0$). Furthermore, these charges are functions solely of the excess demand $(-k \geq 0)$, for a given limit (K) on temporary facility usage.

Assumption 2.1 is a restatement of the hypothesis for Theorem 1.1. Therefore, using Theorem 1.1, the construction of Model I will proceed under the equivalent assumption:

Assumption 2.1'.

(i) When permanent facilities satisfy all demand, no operating costs are incurred.

(ii) When temporary facilities are necessary (i.e., $k < 0$) marginal costs equal to the temporary facility charges, less an adjustment $\gamma_1|k| = -\gamma_1 k \geq 0$ per unit-time, are incurred.

Assumption 2.2 presumes a known unique correspondence between temporary facility costs and the value of the spares level $k \leq 0$, given the usage limit $K \geq 0$. It is demonstrated in [17] that this assumption does not, in general, hold for the case of modular temporary facilities. However, the assumption appears to be viable for any situation where temporary facility costs are not a function of the past temporary facility usage pattern. For instance, referring to the inventory analogy of Chapter 1, Assumption 2.2 is always presumed with regard to backlogging charges.

These assumptions will be treated more precisely in the next section. In the meantime, the following examples are offered to illustrate situations where they hold.

EXAMPLES

Example 2.1: General Single Function. Suppose that, given the value K, temporary facility charges are given by a function

$$w_k(k,t) = \text{discounted temporary facility costs for serving } |k| \text{ customers over a length of time } t \geq 0, \quad k \leq 0.$$

In this case, the marginal costs can similarly be written as

$$\hat{w}_K(k,t) = w_K(k,t) + \int_0^t \gamma_1 k\, e^{-ru}\, du$$

$$= w_K(k,t) + \gamma_1 k\, r^{-1}(1-e^{-rt}) , \qquad k \leq 0,\ t \geq 0 .$$

Example 2.2: Piecewise Proportional Costs. Suppose that, given the value K, temporary facility charges are incurred at rate $\gamma_2^K(k)$ per unit time whenever $k \leq 0$. This is a special case of the previous example with

$$w_K(k,t) = \int_0^t \gamma_2^K(k)\, e^{-ru}\, du = \gamma_2^K(k)\, r^{-1}(1 - e^{-rt}), \quad k \leq 0,\ t \geq 0$$

and

$$\hat{w}_K(k,t) = (\gamma_2^K(k) + \gamma_1 k)\, r^{-1}(1 - e^{-rt}), \qquad k \leq 0,\ t \geq 0 .$$

Example 2.3: Proportional Costs. Suppose that, given the value K, temporary facility charges are incurred at a rate γ_2^K per unit-time, for each unit of temporary capacity, whenever $k \leq 0$. This is a special case of the previous example with $\gamma_2^K(k) = \gamma_2^K|k| = -\gamma_2^K k$, $k \leq 0$.

Example 2.4: Instantaneous Charges. Suppose that, in addition to the costs given by any of the previous examples, an instantaneous charge γ_3^K is incurred whenever an additional unit of temporary capacity is used (that is, each time that k decreases whenever $k \leq 0$). This case differs from the general modular case because the instantaneous charge is solely a function of k and is independent of the past temporary facility usage pattern.

Example 2.5: Restricted Modular Case. As noted previously, Assumption 2.2 does not hold in the general modular case. However, the assumption will hold if the utilization of temporary modules is required to track the demand pattern. Recalling that each module provides L units of temporary capacity, there is a one-to-one correspondence between the spares level (k) and the number of modules used (n) under this restriction:

$$n = \begin{cases} 0 , & k > 0 \\ \min\{i : iL \geq -k\} = \left[\frac{-k+L-1}{L}\right] , & k \leq 0 \end{cases} \qquad (2.1)$$

where $[\cdot]$ denotes integer truncation.

This situation can be represented in a transition diagram, as shown in Figure 2.1. In the diagram, spares levels are depicted vertically and spares levels using the same number of temporary modules are aligned in columns. Each circled number denotes the number of temporary modules necessary for the corresponding spares level given horizontal to the circle in the right-hand margin. Transitions occurring due to departures are denoted by solid lines, while transitions occurring due to arrivals are denoted by dashed lines.

Connection and disconnection (i.e., installation and removal) charges are denoted by triangles. Whenever all modules are fully used ($k = -nL$ for some integer n) and an arrival next occurs, two possibilities occur. If $k = -K$ ($= -NL$ for some integer N), then a permanent capacity expansion is undertaken and the N modules are disposed of; otherwise

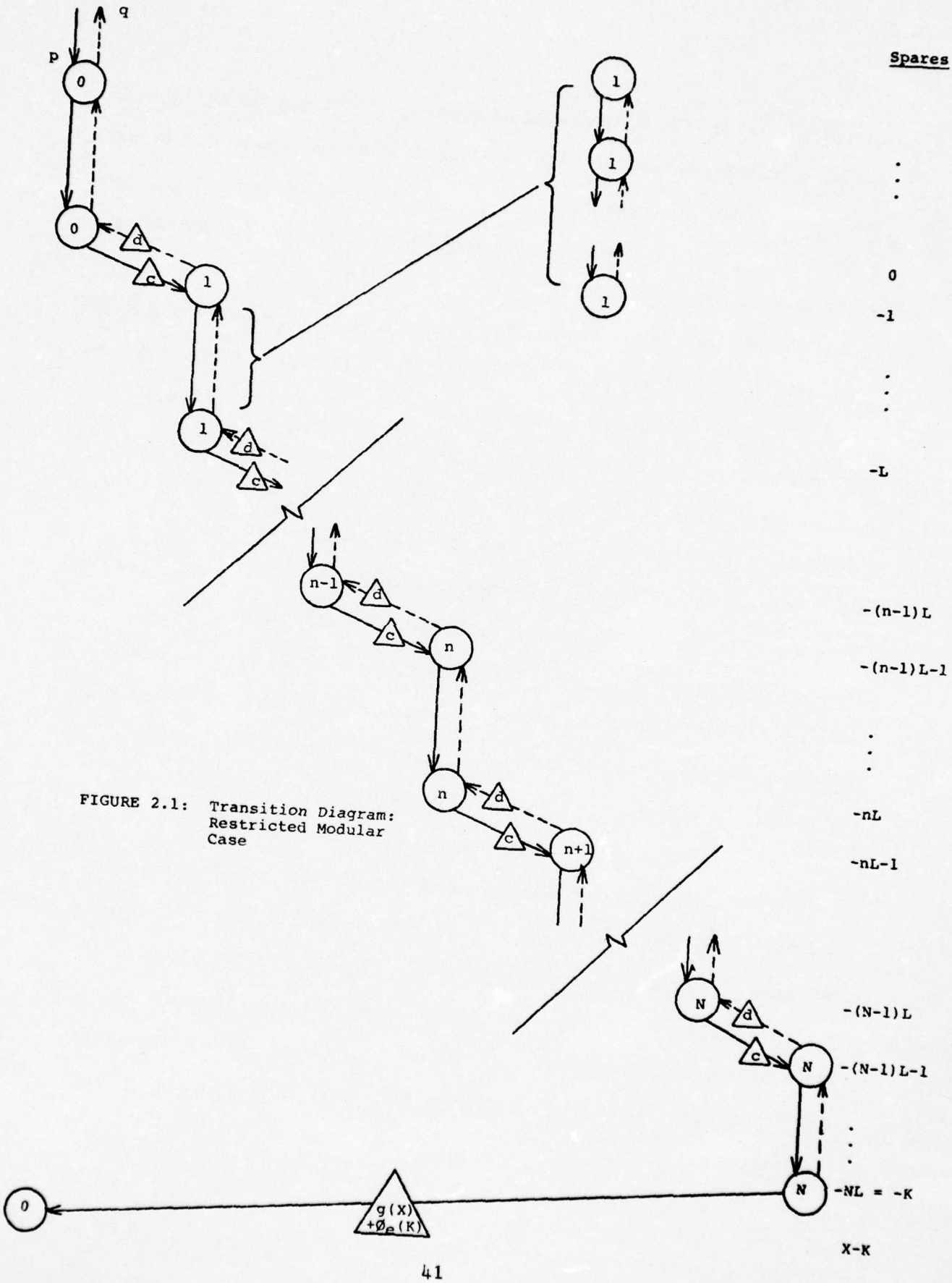


FIGURE 2.1: Transition Diagram: Restricted Modular Case

($n < N$), a new module is connected at cost c, increasing the total number of modules in use to $n+1$. Similarly, whenever a module is being used to serve a single customer (i.e., $k = -(n-1)L-1$ for some $n \geq 1$) and a departure next occurs, a module is disconnected at cost d, decreasing the number of modules in use to $n-1$.

In addition to the instantaneous connect and disconnect charges, a cost rate of π per unit-time is incurred for each module in use. Given the correspondence (2.1), adjusted cost rate $[L^{-1}(-k+L-1)]\ (\pi + \gamma_1 k)$ per unit-time is incurred whenever $k \leq 0$.

Given "free rein" on temporary module utilization, it is shown in [17] that optimal disconnections will usually not track the demand pattern. Therefore, Model I is only applicable for the modular case when the restriction of module use to the demand pattern is an imposed operating constraint. This situation can arise when temporary modules are dispatched from a central facility which services a number of systems, including the one presently being modeled. If the overall supply of modules is limited, then the restriction of module use to the demand pattern may be an imposition dictated by the central facility.

The restriction can also arise when the retention of unnecessary modules is aesthetically unpleasing to some individual or interest group. For example, suppose that mobile ("barracks") facilities were located on a school campus to augment classroom capacity. If congestion at the school temporarily declines, then the school board may have difficulty convincing its constituency that retention of the mobile facilities is optimal.

Remark-Disposal Costs. Recall that whenever the temporary facilities usage limit is reached (i.e., $k = -K$) and an arrival next occur, a permanent capacity expansion of size $X+1$ is undertaken at cost $g(X)$, $X \geq K$. When this happens, the spares level increases from $-K$ to $X-K \geq 0$ and (by Assumption 2.2) permanent facilities are again used to satisfy all demand. Thus, when an expansion occurs, there may be abnormal costs (or revenues) for "turning off" or disposing of the temporary facilities. Thus, in all of the above examples, there may be an additional cost (or revenue) $\phi_e(K)$ whenever $k = -K$ and an arrival next occurs.

2.2. Transition Equations for the Expected Discounted Costs

$C_k(X,K)$ will denote the expected total discounted cost when the initial spares level is k, the temporary facility usage limit is K, and the permanent capacity expansion size is $X+1$; $X \geq K \geq 0$, $k \geq -K$. Similarly, $F_k(K)$ will denote the expected incremental discounted costs (until the next arrival or departure) while the spares level has value k, when the temporary facilities usage limit is $K \geq 0$; $k \geq -K$. Given these designations, Assumptions 2.1' and 2.2 can now be restated:

(a) $F_k(K) = 0$, $k > 0$, $K \geq 0$;

(b) $F_k(K)$, $0 \geq k \geq -K$, $K \geq 0$, are (uniquely) known, and

(c) $F_k(K)$ includes an (expected) adjustment term

$$\int_0^\infty \gamma_1 k r^{-1}(1 - e^{-rt})\, \lambda e^{-\lambda t}\, dt = (\lambda+r)^{-1} \gamma_1 k , \qquad k \leq 0 .$$

Recall that p is the probability of arrival and $q = 1-p$ is the probability of departure. Denote:

$$\alpha = \lambda q(\lambda+r)^{-1} \qquad \text{and} \qquad \beta = \lambda p(\lambda+r)^{-1} \ . \tag{2.2}$$

Definition 2.1. The expected incremental costs $\{F_k(K)\}$ will be said to be constant in K if $F_k(K) = F_k$, $0 \geq k \geq -K+1$ (independent of K) for all $K \geq 0$. In this case, the expected incremental cost at spares level $-K$ will be denoted $F_{-K} + \phi_e(K)$.

The ramifications of the above definition, revealed in Section 2.4, are relatively minor. However, in actual practice, it appears likely that many of the examples in the previous section will often have expected incremental costs that are constant in K. Given the apparent prevalence of this property, the separate term $\phi_e(K)$ will be included in the model development so that, without loss of generality, it can be assumed that $F_{-K}(K) = F_{-K}$ when the incremental costs are constant in K. If the incremental costs are not constant in K, then $\phi_e(\cdot)$ can be taken as identically zero.

EXAMPLES (continued)

Example 2.1. Conditioning on the time of the next event gives

$$\begin{aligned} F_k(K) &= \int_0^\infty \left(w_K(k,t) + r^{-1}\, \gamma_1 k(1 - e^{-rt})\right) \lambda e^{-\lambda t}\, dt \\ &= \int_0^\infty w_K(k,t)\, \lambda e^{-\lambda t}\, dt + (\lambda+r)^{-1}\, \gamma_1 k \ , \qquad k \leq 0 \ . \end{aligned}$$

If $w_K(\cdot,\cdot) = w(\cdot,\cdot)$ for all $K \geq 0$, then the incremental costs are constant in K.

Example 2.2. Conditioning on the time of the next event gives

$$F_k(K) = \int_0^\infty (r_2^K(k) + r_1 k)\, r^{-1}(1 - e^{-rt})\, \lambda e^{-\lambda t}\, dt$$

$$= (r_2^K(k) + r_1 k)\, (\lambda+r)^{-1}, \qquad k \leq 0 .$$

If $r_2^K(k) = r_2(k)$ for all $K \geq 0$, then the incremental costs are constant in K.

Example 2.3. When $r_2^K(k) = -r_2^K k$, the previous expression gives $F_k = (r_1 - r_2^K)k\, (\lambda+r)^{-1}$, $k \leq 0$. If $r_2^K = r_2$ for all $K \geq 0$, then the incremental costs are constant in K.

Example 2.4. Conditioning on the time of the next event and the probability (p) that the next event is an arrival gives

$$F_k(K) = \int_0^\infty (w_K(k,t) + r^{-1} r_1 k(1 - e^{-rt}) + p r_3^K e^{-rt})\, \lambda e^{-\lambda t}\, dt$$

$$= \int_0^\infty w_K(k,t)\, \lambda e^{-\lambda t}\, dt + (\lambda+r)^{-1} r_1 k + \beta r_3^K, \qquad 0 \geq k > -K$$

and

$$F_{-K}(K) = \int_0^\infty w_K(-K,t)\ \lambda e^{-\lambda t}\ dt + (\lambda+r)^{-1}\ r_1 k \quad .$$

Note that the instantaneous charge r_3^K is not incurred when $k = -K$, since an arrival at that point triggers an expansion.

If $w_K(\cdot,\cdot) = w(\cdot,\cdot)$ and $r_3^K = r_3$, for all $K \geq 0$, then denote

$$F_k = \int_0^\infty w(k,t)\ \lambda e^{-\lambda t}\ dt + (\lambda+r)^{-1}\ r_1 k + \beta r_3 , \qquad k < 0 .$$

In this case, $F_k(K) = F_k$, $0 \geq k > -K$. Hence, the expected incremental costs are constant in K. Thus, the expected incremental cost when $k = -K$ will be denoted as $F_{-K} + \beta\phi_e(K)$, where the function $\phi_e(\cdot)$ includes an adjustment $-r_3$.

Example 2.5. Using the correspondence (2.1), the expected incremental usage costs are given by

$$R_k = \int_0^\infty \left([\tfrac{-k+L-1}{L}]\ \pi + r_1 k\right)\ r^{-1}(1 - e^{-rt})\ \lambda e^{-\lambda t}\ dt$$

$$= \left([\tfrac{-k+L-1}{L}]\ \pi + r_1 k\right)\ (\lambda+r)^{-1} , \qquad k \leq 0 .$$

For connect and disconnect points, conditioning on the both the time and type of the next event gives

$$F_k = \begin{cases} R_k, & \text{if } \mathrm{mod}(-k,L) \neq 0,1 \\ R_k + \alpha d, & \text{if } \mathrm{mod}(-k,L) = 1 \\ R_k + \beta c, & \text{if } \mathrm{mod}(-k,L) = 0 \end{cases} , \quad k \leq 0 \ .$$

Thus, $F_k(K) = F_k$, $0 \geq k > -K \ (= -NL)$, so the expected incremental costs are constant in K for the modular case. Hence, the expected incremental cost when $k = -K$ will be denoted $F_{-K} + \beta\phi_e(K)$, where the function $\phi_e(\cdot)$ includes an adjustment term $-c$. Note that $\phi_e(K)$ likely also includes a term Kd/L to provide for disconnection of the $N = K/L$ modules when a permanent expansion occurs.

By conditioning on the time and type of the next event, the cost transition equations are

$$C_k(X,K) = F_k(K) + \int_0^\infty \lambda e^{-\lambda t}\left(qC_{k+1}(X,K) + pC_{k-1}(X,K)\right) e^{-rt}\, dt$$

$$= F_k(K) + \alpha C_{k+1}(X,K) + \beta C_{k-1}(X,K)$$

$$= \begin{cases} \alpha C_{k+1}(X,K) + \beta C_{k-1}(X,K), & k > 0 \qquad (2.3) \\ F_k(K) + \alpha C_{k+1}(X,K) + \beta C_{k-1}(X,K), & 0 \geq k > -K \qquad (2.4) \end{cases}$$

Notice that no upper bound is imposed upon k in (2.3). This corresponds to the "untruncated demand assumption" discussed in Chapter 1. Therefore,

the comments of Chapter 1 regarding the viability of this assumption are applicable here.

When $k = -K$ and an arrival next occurs, a permanent capacity expansion of size $X+1$ $(X \geq K)$ is undertaken. Whenever this happens, the spares level increases to the value $X-K \geq 0$. Therefore, by again conditioning on the next event time and type, the boundary transition equation becomes

$$C_{-K}(X,K) = F_{-K}(K) + \int_0^\infty \lambda e^{-\lambda t}\{qC_{-K+1}(X,K) + p(\phi_e(K) + g(X) + C_{X-K}(X,K))\}\, e^{-rt}\, dt$$

$$= F_{-K}(K) + \alpha C_{-K+1}(X,K) + \beta(\phi_e(K) + g(X) + C_{X-K}(X,K)). \quad (2.5)$$

Notice that, for any scalar v, $F_{-K}(K) + \beta v$ and $\phi_e(K) - v$ may be readily substituted for $F_{-K}(K)$ and $\phi_e(K)$, respectively, in (2.5). This fact will be referred to as the Equivalent Formulation Property.

2.3. Solutions for the Transition Equations, $k > 0$

The special nature of the linear equations (2.3) is well-known and solutions are given in [4] and [9]. For the sake of completeness, the appropriate results are repeated here in both analytical and probabilistic contexts. The following lemma states some prerequisite relationships regarding the coefficients α and β.

Lemma 2.1. For α and β given by (2.2):

(i) $\alpha > 0$, $\beta > 0$,

(ii) $\alpha+\beta < 1$,

(iii) $\alpha\beta < 1/4$.

Proof. (i) follows since λ, p, q and r are all positive. (ii) follows also from positivity since $\alpha+\beta = \lambda(\lambda+r)^{-1}$. The function $p(1-p)$ attains a global maximum of $1/4$ when $p = 1/2$. Thus, $\alpha\beta = \lambda^2(\lambda+r)^{-2}\, pq < pq$ $= p(1-p) \leq 1/4$, which is (iii). □

The following theorem provides the analytical solutions to equations (2.3).

Theorem 2.2.

$$C_k(X,K) = Z^k\, C_0(X,K), \qquad k \geq 0\ , \tag{2.6}$$

where

$$Z = \frac{1 - \sqrt{1 - 4\alpha\beta}}{2\alpha}\ . \tag{2.6a}$$

Furthermore, Z is real and $\beta < Z < \alpha+\beta$.

Proof. It is easily verified that $C_k(X,K) = v^k C_0(X,K)$, $k > 0$, satisfies (2.3) if v is a root of the quadratic expression $\alpha v^2 - v + \beta = 0$. This quadratic has two roots:

$$Z = \frac{1 - \sqrt{1 - 4\alpha\beta}}{2\alpha} \quad \text{and} \quad Z' = \frac{1 + \sqrt{1 - 4\alpha\beta}}{2\alpha} \quad . \tag{2.7}$$

By Lemma 2.1, both roots are real and positive. Also,

$$\alpha+\beta < 1 \Rightarrow \beta < 1-\alpha$$

$$\Rightarrow 4\alpha\beta < 4\alpha(1-\alpha)$$

$$\Rightarrow 1-4\alpha\beta > 1-4\alpha+4\alpha^2 = (1-2\alpha)^2 = (2\alpha-1)^2$$

$$\Rightarrow \sqrt{1-4\alpha\beta} > 1-2\alpha \quad \text{and} \quad \sqrt{1-4\alpha\beta} > 2\alpha-1$$

$$\Rightarrow 1 - \sqrt{1-4\alpha\beta} < 2\alpha \quad \text{and} \quad 1 + \sqrt{1-4\alpha\beta} > 2\alpha$$

$$\Rightarrow Z < 1 \quad \text{and} \quad Z' > 1 \quad .$$

To verify that Z is the correct root for the case at hand, we rely on economic reasoning. Specifically, since no charges are incurred when $k > 0$, no costs can accrue until the first time that the spares level becomes nonpositive. Therefore, since the model is time-stationary, it must be less costly to start with a larger positive spares level; i.e.,

$C_k(X,K) \leq C_0(X,K)$, $k > 0$. Since $Z' > 1$, $C_k(X,K) \neq Z'^k\, C_0(X,K)$. Hence $C_k(X,K) = Z^k\, C_0(X,K)$, $k \geq 0$.

Regarding the lower bound on Z, we have

$$(1-2\alpha\beta)^2 = 1-4\alpha\beta + 4(\alpha\beta)^2 \Rightarrow 1-4\alpha\beta < (1-2\alpha\beta)^2 \Rightarrow \sqrt{1-4\alpha\beta} < 1-2\alpha\beta$$

$$\Rightarrow 1 - \sqrt{1-4\alpha\beta} > 2\alpha\beta \Rightarrow Z > \beta \quad .$$

The upper bound on Z follows as

$$Z = \alpha Z^2 + \beta < \alpha + \beta \quad , \qquad \text{since } Z < 1 \, . \qquad \square$$

Corollary 2.3. For $X \geq K \geq 0$, $X' \geq K' \geq 0$ and $k \geq 0$,

$$C_0(X,K) < C_0(X',K') \iff C_k(X,K) < C_k(X',K') \quad .$$

Proof. The result is a direct consequence of the theorem since $Z > \beta > 0$. $\square$

The solution (2.6) can also be obtained through a simple probabilistic argument.

Definition 2.2. Let the random variable T_i denote the first time that the net demand reaches i:

$$T_i = \min\{t \geq 0 : \mathcal{D}(t) - \mathcal{D}(0) = i\}, \quad i = 0, \pm 1, \pm 2, \ldots \, .$$

Lemma 2.4. $Z = E[e^{-rT_1}]$.

Proof. Let $v = E[e^{-rT_1}]$. Since the demand process is memoryless, $T_2 \overset{D}{=} T' + T''$ where T' and T'' are independent, $T' \overset{D}{=} T_1$ and $T'' \overset{D}{=} T_1$. Therefore,

$$E[e^{-rT_2}] = E[e^{-r(T'+T'')}] = E[e^{-rT'}]\, E[e^{-rT''}] = v^2 \quad .$$

Conditioning on the first event type and the first event time (t) yields:

$$E[e^{-rT_1} \mid \text{1st event at } t \text{ is on arrival}] = e^{-rt}$$

$$E[e^{-rT_1} \mid \text{1st event at } t \text{ is a departure}] = e^{-rt}\, E[e^{-rT_2}] = v^2\, e^{-rt} \quad .$$

Therefore, v *must satisfy*

$$v = \int_0^\infty (qv^2 + p)\, e^{-rt}\, \lambda e^{-\lambda t}\, dt = \alpha v^2 + \beta \quad .$$

Thus, v is one of the two positive real roots Z and Z' given by (2.7). As previously shown, $Z < 1$ and $Z' > 1$. Since $T_2 \geq T_1$, it follows that $v^2 = E[e^{-rT_2}] \leq E[e^{-rT_1}] = v$. Hence, $v = Z$. □

Lemma 2.5. $Z^i = E[e^{-rT_i}]$, $i > 1$.

Proof (induction on i). Since the demand process is memoryless, $T_{i+1} \overset{D}{=} T'+T''$ where T' and T'' are independent, $T' \overset{D}{=} T_1$ and $T'' \overset{D}{=} T_i$. Therefore,

$$E[e^{-rT_{i+1}}] = E[e^{-r(T'+T'')}] = E[e^{-rT_1}]\, E[e^{-rT_i}] = ZZ^i = Z^{i+1} \quad . \quad \square$$

Thus, z^i is simply the Laplace transform of T_i, evaluated at the discount rate r, for $i > 0$. In the absence of intervening permanent expansion, the excess demand $(-k)$ tracks demand. Hence, if there is no intervening permanent expansion, T_i is also the first time that the spares level decreases an amount $i > 0$ below the initial spares level. With this interpretation in mind, the alternative probabilistic proof of Theorem 2.2 can now be given.

Proof of Theorem 2.2 (alternate). Suppose that the initial spares level is $k > 0$. Permanent capacity expansions do not occur while the spares level is positive. Furthermore, no operating costs are incurred when the spares level is positive. Therefore, no costs can accrue until time T_k, when the spares level first becomes nonpositive (0). Since the model is time-stationary and the demand process is memoryless,

$$\begin{aligned} C_k(X,K) &= E[\text{costs during } [0,T_k) + \text{costs during } [T_k,\infty)] \\ &= E[\quad 0 \quad + e^{-rT_k} C_0(X,K)] \\ &= E[e^{-rT_k}] C_0(X,K) \\ &= z^k C_0(X,K), \qquad k > 0 . \end{aligned}$$ □

Let

$$\tilde{z} = \frac{1 - \sqrt{1 - 4\alpha\beta}}{2\beta} . \tag{2.8}$$

The following lemma summarizes results analogous to those previously derived for Z.

Lemma 2.6.

(i) $\tilde{Z}$ is real and $\alpha < \tilde{Z} < \alpha+\beta$,

(ii) $\tilde{Z}^i = E[e^{-rT_{-i}}]$, $i > 0$,

(iii) $\tilde{Z} = \beta^{-1}\alpha Z$.

Proof. (i) follows in a manner analogous to the first proof of Theorem 2.2, as $\tilde{Z}$ is the smallest root of the quadratic expression $\beta v^2 - v + \alpha = 0$. (ii) follows in a manner analogous to the proofs of Lemmas 2.4 and 2.5 (iii) follows by inspection of (2.6a) and (2.8). □

2.4. Recursive Computation of the Functionals $\{C_0(\cdot,K)\}$

When the spares level is nonpositive, the costs behave according to (2.4) and (2.5). Since $X \geq K$, it follows from Theorem 2.2 that (2.5) may be rewritten as

$$C_{-K}(X,K) = F_{-K}(K) + \alpha C_{-K+1}(X,K) + \beta(\phi_e(K) + g(X) + Z^{X-K} C_0(X,K)). \quad (2.9)$$

Also, $C_1(X,K) = ZC_0(X,K)$, so that the first equation of (2.4) can be rewritten as

$$C_0(X,K) = (1-\alpha Z)^{-1} F_0(K) + \beta(1-\alpha Z)^{-1} C_{-1}(X,K) \quad .$$

Let $X = \hat{X}$ and $K = \hat{K}$ be fixed. Denote $\hat{C}_k = C_k(\hat{X},\hat{K})$ and $\hat{F}_k = F_k(\hat{K})$, $0 \geq k \geq -\hat{K}$. Then (2.9) and (2.4) yield a square $(\hat{K}+1)$ nonsingular linear system, as depicted in Figure 2.2. Solving this system yields the costs $\{\hat{C}_k\}$ for the fixed values $\hat{X}$ and $\hat{K}$. By Corollary 2.3, it suffices to compute only $\hat{C}_0$.

Obviously, it is impractical to solve such systems for all possible values $(\hat{X},\hat{K})$, so some sort of parameterization is necessary. Suppose that the last equation in the system of Figure 2.2 is replaced by an equation

$$-\alpha\hat{C}_{-K+1} + \hat{C}_{-K} - \beta\hat{C}_d = \hat{F}_{-\hat{K}} + \beta\phi_e(\hat{K}) ,$$

where $\hat{C}_d$ is a dummy variable. Then the resulting linear system is independent of the value $\hat{X}$ and solving for $\hat{C}_0$ in terms of $\hat{C}_d$ will yield $\hat{C}_0 = \hat{a} + \hat{b}\hat{C}_d$. Making the substitution $\hat{C}_d = g(\hat{X}) - z^{\hat{X}-\hat{K}}\hat{C}_0$ then yields $\hat{C}_0 = (\hat{a} + \hat{b}g(\hat{X}))\,(1 - \hat{b}z^{\hat{X}-\hat{K}})^{-1}$. Since this is true for any value $\hat{X}$ chosen, it follows that $C_0(X,\hat{K}) = (\hat{a} + \hat{b}g(X))\,(1 - \hat{b}z^{X-\hat{K}})^{-1}$, for all $X \geq \hat{K}$. Hence, using the dummy substitution, the system need only be solved once for a given value $\hat{K}$ to obtain the functional $C_0(\cdot,\hat{K})$, which can then be minimized over $[\hat{K},\infty)$.

Referring to Figure 2.2, notice that the respective linear systems for $K = \hat{K}$ and $K = \hat{K}+1$ would be very similar. In fact, the coefficients in the first $\hat{K}$ rows and $\hat{K}$ columns would be identical. Hence, it seems reasonable to expect that solving for the functional $C_0(\cdot,\hat{K})$ should aid in solving for the functional $C_0(\cdot, \hat{K}+1)$. That is, it should be possible to recursively solve for the functionals $C_0(\cdot,K)$ for

$$
\begin{array}{ccccccccccc}
\hat{C}_0 & \hat{C}_{-1} & \hat{C}_{-2} & \hat{C}_{-3} & \hat{C}_{-4} & \cdot & \cdot & \cdot & \hat{C}_{-\hat{K}+2} & \hat{C}_{-\hat{K}+1} & \hat{C}_{-\hat{K}}
\end{array}
$$

$$
\begin{bmatrix}
1 & -\beta(1-\alpha z)^{-1} & & & & & & & & & \\
-\alpha & 1 & -\beta & & & & & & & & \\
 & -\alpha & 1 & -\beta & & & & & & & \\
 & & -\alpha & 1 & -\beta & & & & & & \\
 & & & -\alpha & 1 & -\beta & & & & & \\
 & & & & & \cdot & \cdot & \cdot & & & \\
 & & & & & & \cdot & \cdot & \cdot & & \\
 & & & & & & & \cdot & \cdot & \cdot & \\
 & & & & & & & & -\alpha & 1 & -\beta \\
 & & & & & & & & & -\alpha & 1 \\
-\beta z^{\hat{X}-\hat{K}} & & & & & & & & & & -\alpha
\end{bmatrix}
\begin{bmatrix} & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & -\beta \\ & & & & & & & & & & 1 \end{bmatrix}
=
\begin{bmatrix}
(1-\alpha z)^{-1}\,\hat{F}_0 \\
\hat{F}_{-1} \\
\hat{F}_{-2} \\
\hat{F}_{-3} \\
\hat{F}_{-4} \\
 \\
 \\
 \\
\hat{F}_{-K+2} \\
\hat{F}_{-K+1} \\
\hat{F}_{-\hat{K}}+\beta\{\phi_e(\hat{K})+g(\hat{X})\}
\end{bmatrix}
$$

FIGURE 2.2: Cost Transition Equations, Model I ($X = \hat{X}$, $K = \hat{K}$)

$K = 0, 1, 2, \ldots$. The derivation of that recursion is the subject of this section.

Definition 2.3. Let α and β satisfying (2.2) be given. Define the transformation $W_1 : \mathbb{R}_+ \times [0,1] \to \mathbb{R}_+$ by $W_1(\epsilon,\delta) = (1-\beta\delta)^{-1} \beta\epsilon$. Define the transformation $W_2 : [0,1] \to [0,1]$ by $W_2(\delta) = \alpha(1 - \beta\delta)^{-1}$. Define the transformation $W : \mathbb{R}_+ \times [0,1] \to \mathbb{R}_+ \times [0,1]$ by $W(\epsilon,\delta) = (W_1(\epsilon,\delta), W_2(\delta))$.

The following lemma verifies that the above transformations are well-defined.

Lemma 2.7. $W_1(0,\delta) = 0$, $W_1(\epsilon,\delta) > 0$ and $0 < W_2(\delta) < 1$, for all $\epsilon > 0$, $0 \leq \delta \leq 1$.

Proof. By Lemma 2.1, $0 \leq \delta \leq 1 \Rightarrow 1-\beta\delta \geq 1-\beta > \alpha > 0 \Rightarrow 0 < (1-\beta\delta)^{-1} < \alpha^{-1}$ $\Rightarrow 0 < \alpha(1-\beta\delta)^{-1} < 1 \Rightarrow 0 < W_2(\delta) < 1$. Note that $W_1(\epsilon,\delta) = (1-\beta\delta)^{-1} \beta\epsilon$ $= \alpha^{-1}\beta\, W_2(\delta)\epsilon$. Therefore, $0 \leq \delta \leq 1 \Rightarrow W_1(0,\delta) = 0$ and $W_1(\epsilon,\delta) > 0$, for $\epsilon > 0$. □

Since W is an into mapping, the composition mapping W^m is well defined: $W^0(\epsilon,\delta) = (\epsilon,\delta)$ and $W^m(\epsilon,\delta) = W \circ W^{m-1}(\epsilon,\delta)$, $m \geq 1$. Similarly, the composition mapping W_2^m is well-defined: $W_2^0(\delta) = \delta$ and $W_2^m(\delta) = W_2 \circ W_2^{m-1}(\delta)$, $m \geq 1$. By the previous lemma, the following transformations are also well-defined.

Definition 2.4. Let α and β satisfying (2.2) be given. Let $\mathfrak{R}_M = \mathbb{R} \times [0,1] \times \{0, 1, \ldots, M\} \times \mathbb{R}^M$. For $M = 1, 2, \ldots$, define the transformations $S : \mathfrak{R}_M \to \mathfrak{R}_M$ by

$$S(\rho,\delta,i,V) = \begin{cases} (\rho, \delta, 0, V) , & i = 0 \\ (\alpha^{-1} W_2(\delta) (\beta\rho+V_i), W_2(\delta), i-1, V), & 1 \le i \le M \end{cases}$$

where $V = (V_1, V_2, \ldots, V_M) \in \mathbb{R}^M$.

The above definition defines a separate transformation for each positive integer M. To avoid possible ambiguities, it will be assumed that dim V = M implies which transformation is being used. The composition mappings S^m are well-defined for all $m \ge 0$: $S^0(\rho,\delta,i,V) = (\rho,\delta,i,V)$ and $S^m(\rho,\delta,i,V) = S^{m-1} \circ S(\rho,\delta,i,V)$. $S_j^m(\rho,\delta,i,V)$ will denote the jth component of $S^m(\rho,\delta,i,V)$, $j = 1, 2, 3$.

Lemma 2.8. For all $\rho \in \mathbb{R}$, $\delta \in [0,1]$ and vectors V,

$$S_2^m(\rho,\delta,i,V) = W_2^m(\delta) , \qquad 0 \le m \le i \le \dim V ,$$

$$S_3^m(\rho,\delta,i,V) = i-m , \qquad 0 \le m \le i \le \dim V .$$

Proof (induction). Let $\rho \in \mathbb{R}$, $\delta \in [0,1]$, $M \in \{1, 2, \ldots\}$ and $V \in \mathbb{R}^M$ be arbitrarily chosen. For $m = 0$, the lemma is trivial and for $m = 1$, the lemma follows from Definition 2.4 since $i \ge m = 1$ and $i \le \dim V = M$. Assume, in general, that the lemma is true for some

$m-1 < M$. Denote $\rho' = S_1^{m-1}(\rho,\delta,i,V)$. Let $i \geq m$ and $i \leq M$. By the induction hypothesis,

$$S^m(\rho,\delta,i,V) = S \circ S^{m-1}(\rho,\delta,i,V)$$

$$= S(\rho', W_2^{m-1}(\delta), i-(m-1), V) \quad .$$

Since $i \geq m \Rightarrow i-(m-1) \geq 1$, it follows from Definition 2.4 that $S_2^m(\rho,\delta,i,V) = W_2 \circ S_2^{m-1}(\rho,\delta,i,V) = W_2 \circ W_2^{m-1}(\delta) = W_2^m(\delta)$ and $S_3^m(\rho,\delta,i,V) = S_3^{m-1}(\rho,\delta,i,V) - 1 = i-(m-1)-1 = i-m$. Hence, the lemma follows by induction on m. □

The transformations S have two important properties that permit a simple recursive solution of the problem at hand. These properties are given in the next two lemmas.

Lemma 2.9. Let $V = (V_1, \ldots, V_M) \in \mathbb{R}^M$ and let $V' = (0,V) \in \mathbb{R}^{M+1}$. Then

$$S_1^m(\rho,\delta,i,V') = S_1^m(\rho,\delta,i-1,V), \qquad 1 \leq m < i \leq M+1 \quad .$$

Proof (induction). For $m = 1$, let $i > m$. Then $i > 1$, so $V_i' = V_{i-1}$ and $i-1 \geq 1$. Hence, by Definition 2.4,

$$S_1(\rho,\delta,i,V') = \alpha^{-1} W_2(\delta) (\beta\rho+V_i')$$

$$= \alpha^{-1} W_2(\delta) (\beta\rho+V_{i-1}) = S_1(\rho,\delta,i-1,V) \quad .$$

In general, assume the lemma is true for some $m-1 \geq 1$. Let $i > m$. Then $i > m-1$ and so, by the induction hypothesis, $S_1^{m-1}(\rho,\delta,i,V') = \rho'$ $= S_1^{m-1}(\rho,\delta,i-1,V)$. Also, since $m-1 < i \leq M+1$ and $m-1 \leq i-1 \leq M$, Lemma 2.8 gives $S_2^{m-1}(\rho,\delta,i,V') = W_2^{m-1}(\delta) = S_2^{m-1}(\rho,\delta,i-1,V)$, $S_3^{m-1}(\rho,\delta,i,V') = i-m$, and $S_3^{m}(\rho,\delta,i-1,V) = i-m-1$. Denote $\delta' = W_2^{m-1}(\delta)$. Then,

$$S^m(\rho,\delta,i,V') = S \circ S^{m-1}(\rho,\delta,i,V') = S(\rho',\delta',i-m,V') ,$$

and

$$S^m(\rho,\delta,i-1,V) = S \circ S^{m-1}(\rho,\delta,i-1,V) = S(\rho',\delta',i-m-1,V) .$$

Since $i-m > 1$, $V'_{i-m} = V_{i-m-1}$ and (since $i-m-1 \geq 1$), Definition 2.4 gives

$$S_1^m(\rho,\delta,i,V') = \alpha^{-1} W_2(\delta') (\beta\rho' + V'_{i-m})$$

$$= \alpha^{-1} W_2(\delta') (\beta\rho' + V_{i-m-1}) = S_1^m(\rho,\delta,i-1,V) .$$

Thus, the lemma follows by induction on m. □

Lemma 2.10. For all $V = (V_1, \ldots, V_M) \in \mathbb{R}^M$,

$$S_1^m(V_M,\delta,M-1,V) = B(m,\delta,M)\cdot V , \qquad 0 \leq m \leq M-1$$

where $B(m,\delta,M) \in \mathbb{R}^M$ is unique.

Proof (induction). For $m = 0$, $B(0,\delta,M) = e_M$ (unique), where e_j denotes the jth unit vector of $\mathbb{R}^M$. In general, assume the lemma is true for some $m-1 \geq 0$, $m-1 < M-1$. Then $S_1^{m-1}(V_M,\delta,M-1,V) = B(m-1,\delta,M)\cdot V$. Denote $\delta' = S_2^{m-1}(V_M,\delta,M-1,V)$. Since $m-1 < M-1$, Lemma 2.8 gives $S_3^{m-1}(V_M,\delta,M-1,V) = M-m \geq 1$. Thus, by Definition 2.4,

$$S_1^m(V_M,\delta,M-1,V) = \alpha^{-1} W_2(\delta') (\beta B(m-1,\delta,M)\cdot V + V_{M-m})$$

$$= B(m,\delta,M)\cdot V ,$$

where (uniquely)

$$B(m,\delta,M) = \alpha^{-1} W_2(\delta') (\beta B(m-1,\delta,M) + e_{M-m}) .$$

Thus, the lemma follows by induction on m. □

Denote $F(K) = (F_0(K), F_{-1}(K), \ldots, F_{-K}(K))$. Then $F(K) \in \mathbb{R}^{K+1}$ and $F_k(K)$ is the $-(k-1)$th component of $F(K)$, $0 \geq k \geq -K$.

Theorem 2.11.

$$C_k(X,K) = \rho_k(K) + \delta_k(K)\, C_{k+1}(X,K) + \epsilon_k(K) (g(X) + \phi_e(K) + Z^{X-K} C_0(X,K)) ,$$

$$X \geq K \geq 0,\ 0 \geq k \geq -K \quad (2.10)$$

where

$$(\epsilon_k(K),\ \delta_k(K)) = W^{K+k}(\beta,\alpha)\ , \tag{2.10a}$$

and

$$\rho_k(K) = S_1^{K+k}(F_{-K}(K),\ \alpha,\ K,\ F(K))\ . \tag{2.10b}$$

Proof (induction). Denote $U(X,K) = g(X) + \phi_e(K) + Z^{X-K}\ C_0(X,K)$. For $k = -K$, $W^o(\beta,\alpha) = (\beta,\alpha)$ and $S_1^o(F_{-K}(K),\ \alpha,\ K,\ F(K)) = F_{-K}(K)$, so (2.10) agrees with (2.9). Hence, the theorem is true for $k = -K$. In general, suppose that the theorem is true for some $k-1$, $-K \le k-1 \le -1$. Then,

$$C_{k-1}(X,K) = \rho_{k-1}(K) + \delta_{k-1}(K)\ C_k(X,K) + \epsilon_{k-1}(K)\ U(X,K)\ . \tag{2.11}$$

By (2.4),

$$C_k(X,K) = F_k(K) + \alpha C_{k+1}(X,K) + \beta C_{k-1}(X,K)\ . \tag{2.12}$$

Substituting (2.11) into (2.12) gives

$$C_k(X,K) = \rho_k(K) + \delta_\rho(K)\ C_{k+1}(X,K) + \epsilon_k(K)\ U(X,K)\ ,$$

where

$$\epsilon_k(K) = (1-\beta\delta_{k-1}(K))^{-1}\ \beta\epsilon_{k-1}(K) = W_1(\epsilon_{k-1}(K),\ \delta_{k-1}(K))\ ,$$

$$\delta_k(K) = \alpha(1-\beta\delta_{k-1}(K))^{-1} = W_2(\delta_{k-1}(K))\ ,$$

and

$$\rho_k(K) = (1-\beta\delta_{k-1}(K))^{-1} (\beta\rho_{k-1}(K) + F_k(K))$$

$$= \alpha^{-1} W_2(\delta_{k-1}(K)) (\beta\rho_{k-1}(K) + F_k(K)) .$$

Hence, using the induction hypothesis,

$$(\epsilon_k(K), \delta_k(K)) = W(\epsilon_{k-1}(K), \delta_{k-1}(K))$$

$$= W \circ W^{K+k-1}(\beta,\alpha) = W^{K+k}(\beta,\alpha) ,$$

and

$$(\rho_k(K), \delta_k(K), -k, F(K)) = S(\rho_{k-1}(K), \delta_{k-1}(K), -(k-1), F(K))$$

$$= S \circ S^{K+k-1}(F_{-K}(K), \alpha, K, F(K))$$

$$= S^{K+k}(F_{-K}(K), \alpha, K, F(K)) .$$

The positivity of $(\epsilon_k(K), \delta_k(K))$ follow from Lemma 2.7. Thus, the theorem follows by induction on k. □

Theorem 2.11 provides an immediate recursion for the coefficients $\epsilon_0(\cdot)$ and $\delta_0(\cdot)$.

Corollary 2.12.

(i) $\epsilon_0(0) = \beta$; $\epsilon_0(K+1) = (1 - \beta\delta_0(K))^{-1} \beta\epsilon_0(K)$, $K \geq 0$,

(ii) $\delta_0(0) = \alpha$; $\delta_0(K+1) = (1 - \beta\delta_0(K))^{-1} \alpha$, $K \geq 0$.

Proof. $(\epsilon_0(0), \delta_0(0)) = W^0(\beta,\alpha) = (\beta,\alpha)$. For $K \geq 0$,

$$(\epsilon_0(K+1), \delta_0(K+1)) = W^{K+1}(\beta,\alpha) = W \circ W^K(\beta,\alpha) = W(\epsilon_0(K), \delta_0(K))$$

$$= ((1-\beta\delta_0(K))^{-1}\, \beta\epsilon_0(K),\ (1 - \beta\delta_0(K))^{-1}\, \alpha) \ . \qquad \square$$

Lemma 2.13.

$$\frac{1 - \delta_0(K)Z}{\epsilon_0(K)}\, Z^K = Z^{-1}, \qquad K \geq 0 \ .$$

Proof (induction). Recall that Z satisfies $-\alpha Z^2+Z-\beta = 0$, so $1-\alpha Z = \beta Z^{-1}$. Using Corollary 2.12,

$$\frac{1 - \delta_0(0)Z}{\epsilon_0(0)}\, Z^0 = \frac{1-\alpha Z}{\beta} = \frac{\beta Z^{-1}}{\beta} = Z^{-1} \ .$$

Suppose, in general that the lemma is true for $K \geq 0$. Then, using $\beta^{-1}(Z-\alpha Z^2) = 1$ and Corollary 2.12,

$$Z^{-1} = \frac{1 - \delta_0(K)Z}{\epsilon_0(K)}\, Z^K = \frac{1 - \delta_0(K)Z}{Z\epsilon_0(K)}\, Z^{K+1}$$

$$= \frac{\beta^{-1}(Z-\alpha Z^2) - \delta_0(K)Z}{Z\epsilon_0(K)}\, Z^{K+1} = \frac{1 - \alpha Z - \beta\delta_0(K)}{\beta\epsilon_0(K)}\, Z^{K+1}$$

$$= \frac{1 - (1 - \beta\delta_0(K))^{-1}\, \alpha Z}{(1-\beta\delta_0(K))^{-1}\, \beta\epsilon_0(K)}\, Z^{K+1} = \frac{1 - \delta_0(K+1)Z}{\epsilon_0(K+1)}\, Z^{K+1} \ . \qquad \square$$

Using the proceding lemma, the functional form of $C_0(\cdot,K)$ can now be obtained.

Theorem 2.14.

$$C_0(X,K) = \frac{\phi(K) + g(X)}{z^{-1} - z^X} z^K \quad , \qquad X \geq K \geq 0 \tag{2.11}$$

where

$$\phi(K) = \phi_e(K) + \frac{\rho_0(K)}{\epsilon_0(K)} \quad . \tag{2.11a}$$

Proof. By Theorem 2.11,

$$C_0(X,K) = \rho_0(K) + \delta_0(K)\ C_1(X,K) + \epsilon_0(K)(\phi_e(K)+g(X)+C_0(X,K)z^{X-K}) \ .$$

By Theorem 2.2, $C_1(X,K) = zC_0(X,K)$. Substituting this result into the above expression and collecting terms yields

$$C_0(X,K) = \frac{(\phi_e(K) + \rho_0(K)/\epsilon_0(K)) + g(X)}{(1-\delta_0(K)z)\ z^K/\epsilon_0(K) - z^X} z^K \quad ,$$

which, by Lemma 2.13, is (2.11). □

Recall that Corollary 2.11 gives a simple recursion for $\epsilon_0(\cdot)$. Hence, by (2.11a), a similar recursion for $\rho_0(\cdot)$ will provide the means for recursively computing the functionals $C_0(\cdot,K)$, K = 0, 1, 2, ... The next theorem provides the necessary recursion for $\rho_0(\cdot)$.

Theorem 2.15.

$$\rho_0(K) = B(K)\cdot F(K) \ , \qquad K \geq 0 \ , \tag{2.12}$$

where

$$B(0) = (1) \in \mathbb{R}^1$$

and

$$B(K+1) = \alpha^{-1}\, \beta\delta_0(K+1)\ (\beta^{-1},\ B(K)) \in \mathbb{R}^{K+2} \ , \quad K \geq 0 \ . \tag{2.12a}$$

Proof. By Lemma 2.10 and (2.10b), the vectors $B(K) = B(K,\alpha,K+1) \in \mathbb{R}^{K+1}$ exist and are unique, $K \geq 0$. By (2.10b), $\rho_0(0) = F_0(0)$, so $B(0) = (1)$. Recall that $F(K) \in \mathbb{R}^{K+1}$ and $F(K+1) \in \mathbb{R}^{K+2}$. Denote

$$\bar{F}(K+1) = (F_{-1}(K+1),\ \ldots,\ F_{-(K+1)}(K+1)) \in \mathbb{R}^{K+1} \ .$$

By (2.10b) and Lemma 2.9,

$$\begin{aligned} \rho_{-1}(K+1) &= S_1^K(F_{-(K+1)}(K+1),\ \alpha,\ K+1,\ F(K+1)) \\ &= S_1^K(F_{-(K+1)}(K+1),\ \alpha,\ K,\ \bar{F}(K+1)) \ . \end{aligned}$$

Since $\bar{F}(K+1) \in \mathbb{R}^{M+1}$, Lemma 2.10 gives

$$\rho_{-1}(K+1) = B(K,\alpha,K+1)\ \cdot\ \bar{F}(K+1) = B(K)\ \cdot\ \bar{F}(K+1) \ .$$

By Lemma 2.8 and (2.10a),

$$S_2^K(F_{-(K+1)}(K+1),\ \alpha,\ K+1,\ F(K+1)) = W_2^K(\alpha) = \delta_0(K)\ .$$

Hence,

$$S^K(F_{-(K+1)}(K+1),\ \alpha,\ K+1,\ F(K+1))$$

$$= (B(K)\cdot\bar{F}(K+1),\ \delta_0(K),\ 1,\ F(K+1))\ .$$

Thus, by (2.10a) and Corollary 2.12,

$$\rho_0(K+1) = S_1(B(K)\cdot\bar{F}(K+1),\ \delta_0(K),\ 1,\ F(K+1))$$

$$= \alpha^{-1}\ W_2(\delta_0(K))\ (\beta B(K)\cdot\bar{F}(K+1) + F_0(K+1))$$

$$= \alpha^{-1}\ \beta\delta_0(K+1)\ (B(K)\cdot\bar{F}(K+1) + \beta^{-1}\ F_0(K+1))$$

$$= \alpha^{-1}\ \beta\delta_0(K+1)\ (\beta^{-1},\ B(K))\cdot(F_0(K+1),\ \bar{F}(K+1))$$

$$= \alpha^{-1}\ \beta\delta_0(K+1)\ (\beta^{-1},\ B(K))\cdot F(K+1)\ .$$

Denote $\theta(K) = \alpha^{-1}\ \beta\delta_0(K)$; then, $\delta_0(K) = \beta^{-1}\ \alpha\theta(K)$. The following theorem summarizes the required recursions.

Theorem 2.15.

(i) $\theta(0) = \beta;\ \theta(K+1) = \beta(1 - \alpha\theta(K))^{-1},\ K \geq 0,$

(ii) $\epsilon_0(0) = \beta;\ \epsilon_0(K+1) = \theta(K+1)\ \epsilon_0(K),\ K \geq 0,$

(iii) $B(0) = (1);\ B(K+1) = \theta(K+1)\ (\beta^{-1},\ B(K)),\ K \geq 0.$

Proof. Note that $\beta(1 - \beta\delta_0(K))^{-1} = \beta(1 - \beta\beta^{-1}\alpha\theta(K))^{-1} = \beta(1 - \alpha\theta(K))^{-1}$. By Corollary 2.12, $\theta(0) = \alpha^{-1}\beta\delta_0(0) = \beta$ and

$$\theta(K+1) = \alpha^{-1}\beta\delta_0(K+1) = \alpha^{-1}\beta\alpha(1 - \beta\delta_0(K))^{-1}$$

$$= \beta(1 - \beta\delta_0(K))^{-1} = \beta(1 - \alpha\theta(K))^{-1}, \quad \text{for } K \geq 1 ;$$

this proves (i). By (i) and Corollary 2.12, $\epsilon_0(0) = \beta$ and

$$\epsilon_0(K+1) = (1 - \beta\delta_0(K))^{-1}\beta\epsilon_0(K) = (1-\alpha\theta(K))^{-1}\epsilon_0(K) = \theta(K+1)\epsilon_0(K) ;$$

this proves (ii). By Theorem 2.15,

$$B(K+1) = \alpha^{-1}\beta\delta_0(K+1)(\beta^{-1}, B(K)) = \theta(K+1)(\beta^{-1}, B(K)) ,$$

which proves (iii). □

By Theorem 2.14, the functional $C_0(\cdot,K)$ is completely determined by the coefficient $\phi(K)$. Hence, a recursive procedure for computing $\phi(\cdot)$ is a recursive procedure for computing the functionals. Using Theorems 2.14, 2.15 and 2.16, the desired algorithm can now be stated.

Algorithm A_1 (Compute $\phi(K)$, $K = 0, 1, 2, \ldots, \bar{K}$).

(1) $K \leftarrow 0$, $\theta \leftarrow \beta$, $\epsilon \leftarrow \beta$, $B \leftarrow (1)$.

(2) $\rho \leftarrow B \cdot F(K)$, $\phi(K) \leftarrow \epsilon^{-1}\rho + \phi_e(K)$.

(3) $K \leftarrow K+1$, $\theta \leftarrow \beta(1-\beta\theta)^{-1}$, $\epsilon \leftarrow \theta\epsilon$, $B \leftarrow \theta(\beta^{-1}, B)$.

(4) If $K > \bar{K}$, stop; otherwise, go to step (2).

Although the recursion for $\rho_0(\cdot)$ can be stated in a number of alternative ways, the vector dot product is, in general, unavoidable. However, when the expected incremental costs are both proportional and constant in K, then the dot product can be eliminated, as will now be shown.

Denote $\Delta F_k(K+1) = F_{k+1}(K+1) - F_k(K)$, $0 \geq k \geq -K$, $K \geq 0$. Denote $\Delta F(K+1) = (\Delta F_0(K+1), \ldots, \Delta F_{-K}(K+1)) \in \mathbb{R}^{K+1}$. Then, $F_{k+1}(K+1) = F_k(K) + \Delta F_k(K+1)$, $k = 0, -1, \ldots, -K$. Hence, $F(K+1) = (F_0(K+1), F(K)) + (0, \Delta F(K+1))$, $K \geq 0$.

Lemma 2.17.

$$\rho_0(K+1) = \theta(K+1)\,(\beta^{-1} F_0(K+1) + B(K)\cdot\Delta F(K+1) + \rho_0(K))\,, \quad K \geq 0. \tag{2.13}$$

Proof. By Theorem 2.16,

$$\rho_0(K+1) = B(K+1)\cdot F(K+1) = \theta(K)\,(\beta^{-1}, B(K))\cdot((F_0(K+1), F(K)) + (0, \Delta F(K+1))$$

$$= \theta(K)\,(\beta^{-1} F_0(K+1) + B(K)\cdot F(K) + B(K)\cdot\Delta F(K+1))$$

$$= \theta(K)\,(\beta^{-1} F_0(K+1) + \rho_0(K) + B(K)\cdot\Delta F(K+1))\,. \quad \square$$

Recall that, when the incremental costs are constant in K, it can be assumed that $F_k(K) = F_k$, $0 \geq k \geq -K$, $K \geq 0$ (by the introduction of the function $\phi_e(K)$). If the incremental costs are also proportional (see Example 2.3), then $\Delta F(K) = \Delta F\, 1_K$ where ΔF is a scalar constant and 1_K denotes the K-vector of all 1's. In this case, (2.13) reduces to

$$\rho_0(K+1) = \theta(K+1)\ (\beta^{-1}F_0 + \Delta F\ b(K) + \rho_0(K)), \qquad K \geq 0\ , \qquad (2.14)$$

where $b(0) = 1$, and

$$b(K) = B(K)\cdot 1_{K+1} = \theta(K)\ (\beta^{-1},\ B(K-1))\cdot 1_{K+1}$$

$$= \theta(K)\ (\beta^{-1} + B(K-1)\cdot 1_K) = \theta(K)\ (\beta^{-1} + b(K-1)), \quad K \geq 1\ . \quad (2.14a)$$

Thus, the vector dot product is eliminated when the expected incremental costs are both proportional and constant in K, giving the following algorithm.

Algorithm A_2 (Compute $\phi(K)$, $K = 0, 1, 2, \ldots, \bar{K}$, when F's are Proportional and Constant in K).

(1) $K \leftarrow 0$, $\theta \leftarrow \beta$, $\rho \leftarrow F_0$, $\epsilon \leftarrow \beta$, $b \leftarrow 1$.

(2) $\phi(K) \leftarrow \epsilon^{-1}\rho + \phi_e(K)$.

(3) $K \leftarrow K+1$, $\theta \leftarrow \beta(1-\beta\theta)^{-1}$, $\epsilon \leftarrow \theta\epsilon$, $\rho \leftarrow \theta(\beta^{-1}F_0 + \Delta F b + \rho)$, $b \leftarrow \theta(\beta^{-1}+b)$.

(4) If $K > \bar{K}$, stop; otherwise, go to step (2).

Given the recursions of this section, it is possible to derive closed-form expressions for the coefficients $\phi(\cdot)$ defining the functionals $C_0(\cdot,K)$, $K \geq 0$. The derivation is accomplished by studying the determinants of some very special tridiagonal submatrices in the linear system depicted by Figure 2.2 [16]. However, from a computation aspect, these expressions are of limited value, since they do not apparently provide any efficiencies over the recursive algorithms given here.

2.5. Assumption 2.3 with Economic Interpretation

This section discusses the final assumption of Model I:

Assumption 2.3. The coefficients $\phi(\cdot)$ of (2.11) are nonnegative and nondecreasing.

The form of the functional $C_0(\cdot,K)$ can be anticipated through a probabilistic argument that provides an economic interpretation of the coefficient $\phi(K)$. Given the economic interpretation, it is demonstrated that the above assumption in essence states that the costs of temporary facilities exceed the permanent facility operating charges that would otherwise be incurred.

Let $G_k(K)$ denote the expected discounted costs until the first expansion, starting from spares level $k \geq 0$, when K is the limit on temporary facility usage.

Lemma 2.18.

$$G_k(K) = Z^k \, G_0(K) \ , \qquad k \geq 0 \ . \tag{2.15}$$

Proof. Starting from spares level $k \geq 0$, no costs are incurred until the spares level first becomes nonpositive; the time of this event is given by T_k and $E[e^{-rT_k}] = Z^k$ by Lemma 2.4. The expected costs beyond time T_k, until the first expansion, are given by $G_0(K)$ (discounted relative to T_k). Hence

$$G_k(K) = E[G_0(K) \ e^{-rT_k}] = G_0(K) \ E[e^{-rT_k}] = G_0(K) \ Z^k \ . \qquad \square$$

When an expansion occurs, the spares level changes to $X-K \geq 0$. Hence, the inter-expansion costs are given by $G_{X-K}(K) = Z^{X-K} \, G_0(K)$.

Theorem 2.19.

$$C_0(X,K) = \frac{G_0(K) \ Z^{-(K+1)} + g(X)}{Z^{-1} - Z^X} \, Z^K \ , \qquad X \geq K \geq 0 \ . \tag{2.16}$$

Proof: Starting from spares level zero, the time until the first expansion is given by T_{K+1}. At time T_{K+1}, an expansion of size $X+1$ occurs at cost $g(X)$ and the expected costs thereafter are $C_{X-K}(X,K) = Z^{X-K} \, C_0(X,K)$ (discounted relative to T_{K+1}). Hence,

$$C_0(X,K) = G_0(K) + E[(g(X) + Z^{X-K}\, C_0(X,K))\; e^{-rT_{K+1}}]$$

$$= G_0(K) + (g(X) + Z^{X-K}\, C_0(X,K))\; E[e^{-rT_{K+1}}]$$

$$= G_0(K) + (g(X) + Z^{X-K}\, C_0(X,K))\; Z^{K+1}$$

$$= G_0(K) + g(X)\; Z^{K+1} + Z^{X+1}\, C_0(X,K)$$

$$\Rightarrow$$

$$C_0(X,K) = \frac{G_0(K) + g(X)\; Z^{K+1}}{1 - Z^{X+1}} = \frac{G_0(K)\; Z^{-(K+1)} + g(X)}{Z^{-1} - Z^{X}}\; Z^{K} \quad . \qquad \square$$

Corollary 2.20.

$$\phi(K) = G_0(K)\; Z^{-(K+1)} \quad , \qquad K \geq 0 \; . \tag{2.17}$$

Proof. This follows by comparing (2.16) and (2.11), each of which is satisfied for arbitrary g and $X \geq K \geq 0$. $\square$

Rearranging (2.17) gives $G_0(K) = \phi(K)\; Z^{K+1} = \phi(K)\; E[e^{-rT_{K+1}}]$. Similarly, using (2.15) gives $G_k(K) = \phi(K)\; Z^{K+k+1} = \phi(K)\; E[e^{-rT_{K+k+1}}]$, $k \geq 0$. When the initial spares level is $k \geq 0$, the time of the first expansion is T_{K+k+1}. Thus, in an expectation sense, $\phi(K)$ is the equivalent lump sum payment that would be incurred at the first permanent expansion time in lieu of the incremental charges over the time span previous to the expansion. Since the expected inter-expansion costs are $G_{X-K}(K) = \phi(K)\; E[e^{-rT_{X+1}}]$ and the inter-expansion time intervals are equal (in distribution) to T_{X+1}, a similar lump sum interpretation of $\phi(K)$ can be given with regard to inter-expansion costs.

Corollary 2.21. For $K \geq 0$ and $k \geq 0$,

(i) $\phi(K) \geq 0 \iff G_k(K) \geq 0$,

(ii) $\phi(K+1) \geq \phi(K) \iff G_k(K+1) \geq ZG_k(K)$,

(iii) $G_k(K+1) \geq G_k(K) \Rightarrow \phi(K+1) \geq \phi(K)$.

Proof. (i) follows directly from (2.17) and (2.15). (ii) follows from (2.17) as $\phi(K+1) - \phi(K) = (G_0(K+1) - G_0(K)Z)\, Z^{-(K+1)} = (G_k(K+1) - G_k(K)Z)\, Z^{-(k+K+1)}$. (iii) follows from (ii) since $0 < Z < 1$. □

By (i) and (iii), the coefficients $\phi(\cdot)$ will be nonnegative and nondecreasing if the expected inter-expansion costs are likewise. Since the only non-expansion costs are the expected incremental costs incurred when temporary facilities are in use, (i) states that $\phi(K)$ will be nonnegative if, in aggregate expectation, the costs of the temporary facilities used exceed the permanent facility operating charges that would otherwise be incurred. With regard to (iii), assume that the initial spares level is $k_0 \geq 0$. Using temporary facilities limit K, the time of the first expansion is T_{K+k_0+1}. Using temporary facilities limit $K+1$, the time of the first expansion is T_{K+k_0+2}. Hence, $G_k(K)$ represents the expected costs over the time interval $[0, T_{K+k_0+1}]$ and $G_k(K+1)$ represents the expected costs over the time interval $[0, T_{K+k_0+2}] \supset [0, T_{K+k_0+1}]$ (with probability 1). It is difficult to imagine practical situations where the expected costs over the time interval $[0, T_{K+k_0+1}]$ using limit K could exceed those over the greater

time interval $[0, T_{K+k_0+2}]$ using limit K+1. In fact, the next lemma indicates that there is really no optimization problem if $\phi(K+1) < \phi(K)$.

Lemma 2.22. Let $K \geq 0$. If $\phi(K+1) < \phi(K)$, then $C_0(X, K+1) < C_0(X,K)$, $X \geq K+1$.

Proof. By (2.11), for $X \geq K+1$,

$$C_0(X,K+1) = \frac{\phi(K+1) + g(X)}{z^{-1} - z^X} < \frac{\phi(K) + g(X)}{z^{-1} - z^X} = C_0(X,K) \ . \qquad \square$$

Thus, if K and K+1 are each potential limits on temporary facility usage, it suffices to compare $C_0(K,K)$ with $\{C_0(X,K+1), X \geq K+1\}$. Thus, from an optimization standpoint, Assumption 2.3 simply eliminates trivialities.

2.6. Summary and Statement of the Expansion Size Optimization Problem

Let κ denote the set of feasible values for K. To illustrate, $\kappa = \{0, 1, 2, \ldots, \bar{K}\}$ would be typical of Examples 2.1 - 2.4, where $\bar{K}$ denotes an upper bound on temporary facility usage based on physical considerations. Similarly, $\kappa = \{0, L, 2L, \ldots, \bar{N}L\}$ for the restricted modular Example 2.5, where $\bar{N}$ denotes an upper bound on temporary module usage based on physical constraints.

Given $K \in \kappa$ and an initial spares level $k_0 \geq 0$, an optimal permanent expansion size $X^*(K) + 1$ for Model I is given by

$$C_0(X^*(K), K) = \min\{C_0(X,K) : \text{integer } X \geq K\} , \tag{2.18}$$

where

$$C_0(X,K) = \frac{\phi(K) + g(X)}{z^{-1} - z^X} z^K , \qquad X \geq K ; \tag{2.19}$$

$\phi(\cdot)$ is nonnegative and nondecreasing over κ ; (2.20)

$0 < z = \frac{1 - \sqrt{1-4\alpha\beta}}{2\alpha} < 1$, given α and β satisfying (2.2) $(\alpha z^2 - z + \beta = 0)$; (2.21)

$$C_k(X,K) = z^k C_0(X,K), \qquad k \geq 0 ; \tag{2.22}$$

and

$$z^k = E[e^{-rT_k}] , \qquad k \geq 1 . \tag{2.23}$$

(2.19) follows from Theorem 2.14, (2.20) is Assumption 2.3 (discussed at length in the previous section), (2.21) and (2.22) follows from Theorem 2.2, and (2.23) follows from Lemma 2.5. The fact that $X^*(K)$ is given by (2.18), independent of $k_0 \geq 0$, follows from Corollary 2.3.

Algorithms A_1 and A_2, presented in Section 2.4, provide simple recursive procedures for determining the coefficients $\phi(\cdot)$ over κ. Once these coefficients are known, the task of determining $X^*(\cdot)$ over κ becomes the series of single-variable minimization problems given by (2.18) under conditions (2.19) - (2.23). This optimization problem is treated in [17]. By Corollary 2.3, the optimal temporary facilities usage limit K^*, and the associated optimal expansion size X^*+1, are then given by

$$C_0(X^*, K^*) = \min\{C_0(X^*(K), K) : K \in \kappa\} .$$

REFERENCES

[1] Goran Bergendahl, "Capacity Planning Models for Systems of Plant and Road Investments", Swedish Journal of Economics, Vol. 70, 1968, pp. 94-105.

[2] Sylvain Ehrenfeld and Sebastian B. Littauer, Introduction to Statistical Method, McGraw-Hill, New York, 1964.

[3] Donald Erlenkotter, "The Sequencing of Expansion Projects", Working Paper No. 166, Western Management Science Institute, University of California, Los Angeles, November 1970.

[4] William Feller, An Introduction to Probability Theory and its Applications, Vol. I, Third Edition, John Wiley and Sons, New York, 1968.

[5] John Freidenfelds, "Cable Sizing with Stochastic Demand", Proceedings of the Sixth Annual Pittsburgh Conference on Modeling and Simulation, Instrument Society of America (pub.), April 1975.

[6] John Freidenfelds, unpublished memorandum, 1974.

[7] D.P. Heyman and Bruce Hoadley, "A Two Echelon Inventory Model with Purchases, Junks, Shipments, Returns and Transhipments", ORSA/TIMS National Meeting, Las Vegas, Nevada, November 1975.

[8] D.P. Heyman, unpublished memoranda, 1974-75.

[9] Francis B. Hildebrand, Finite-Difference Equations and Simulations, Prentice-Hall, 1968.

[10] Warren L.G. Koontz and R.S. Shipley, "Application of Subscriber Pair Gain Systems in an Environment of Stochastic Demand", Proceedings of the Sixth Annual Pittsburgh Conference on Modeling and Simulation, Instrument Society of America (pub.), April 1975.

[11] Alan S. Manne, "Capacity Expansion and Probabilistic Growth", Econometrica, Vol. 29, October 1961, pp. 632-649.

[12] Alan S. Manne (ed.), Investments for Capacity Expansion: Size, Location, and Time-Phasing, The M.I.T. Press, Cambridge, Massachusetts, 1967.

[13] Alan S. Manne and Arthur F. Veinott, Jr., "Optimal Plant Size with Arbitrary Increasing Time Paths of Demand", in [12], pp. 178-190.

[14] John Scott Rogers, "A Dynamic Model for Planning Capacity Expansion: An Application to Plant Reliability in Electric Power Systems", Ph.D. Dissertation, Operations Research, Stanford University, May 1970.

[15] Sheldon M. Ross, Applied Probability Models with Optimization Applications, Holden-Day, San Francisco, 1970.

[16] Robert Scott Shipley, "Stochastic Capacity Expansion Models", Ph.D. Dissertation, Stanford University, September 1976.

[17] R. Scott Shipley, "A Stochastic Capacity Expansion Model: Modular Temporary Facilities", Stanford University, Department of Operations Research, Technical Report No. 179, September 1976.

[18] R. Scott Shipley, "Optimization of Recurrent Stochastic Capacity Expansion Models and Generalization to a Non-Recurrent Model", Stanford University, Department of Operations Research, Technical Report No. 180, September 1976.

[19] M. Tainiter, "Some Stochastic Inventory Models for Rental Situations", Management Science, Vol. 11, November 1964, pp. 316-326.

[20] Arthur F. Veinott, Jr., "The Status of Mathematical Inventory Theory", Management Science, Vol. 12, July 1966, pp. 745-777.

[21] William D. Whisler, "A Stochastic Inventory Model for Rented Equipment", Management Science, Vol. 13, May 1967, pp. 640-647.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 178	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): A Stochastic Capacity Expansion Model: Non-Modular Temporary Facilities,	5. TYPE OF REPORT & PERIOD COVERED: Technical Report,
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): R. Scott Shipley	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0561
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Dept. of Operations Research & Dept. of Statistics, Stanford University, Stanford, Calif. 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-042-002)
11. CONTROLLING OFFICE NAME AND ADDRESS: Statistics & Probability Program Code 436, Office of Naval Research, Arlington, Virginia 22217	12. REPORT DATE: September 27, 1976
	13. NUMBER OF PAGES: 30
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION IS UNLIMITED.

14 TR-178

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

BACKLOGGING, CAPACITY EXPANSION, INVENTORY THEORY, JOBLETTING, OVERLOADING, POISSON PROCESSES

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See reverse side.

11 27 Sep 76 12 83 p.

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

332580

20. Abstract

This paper considers optimal decision strategies with regard to capacity expansion in an environment where demand arrivals and departures can be characterized as independent Poisson processes. Two types of facilities are considered in the model: permanent and temporary. Permanent facilities represent the means by which demand is normally served, while temporary facilities represent the extraordinary measures taken in order to serve excess demand (prior to an expansion of permenent facilities). Examples of temporary facilities include backlogging, overloading and jobletting. This paper considers non-modular temporary facilities, the costs of which depend only on the amount of excess demand currently being served. A companion paper [17] treats the case of modular temporary facilities.

For a given limit (K) on temporary facility usage, the form of the expected discounted cost functional, parameterized in the expansion size (X+1), is derived. Recursions are given for determining these functionals over all feasible values for K. A sequel paper [18] treats the problem of minimizing the functionals in order to find optimal expansion sizes.