OPTIMAL DYNAMIC INVESTMENT POLICIES FOR PUBLIC FACILITIES:

THE TRANSPORTATION CASE

By

JOSE ENRIQUE FERNANDEZ LARRAÑAGA

Ingeniero Civil, Universidad Católica de Chile

(1969)

Diplomado en Administracion de Empresas,
Universidad de Madrid
(1970)

S.M., Massachusetts Institute of Technology
(1978)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June, 1979

Signature of Author.....
Department of Civil Engineering
May 21, 1979

Certified by.....
Thesis Supervisors

Accepted by.....
Chairman, Departmental Committee on Graduate Students of the Department of Civil Engineering.

OPTIMAL DYNAMIC INVESTMENT POLICIES FOR PUBLIC FACILITIES: THE TRANSPORTATION CASE

By

JOSE ENRIQUE FERNANDEZ LARRAÑAGA

Submitted to the Department of Civil Engineering on May 21, 1979, in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

ABSTRACT

Public facilities are characterized in this study by two attributes, quality and capacity, whose values are assumed to change over time due to natural factors, use and investments. It is also assumed that the users of the facility obtain a private benefit from the use of the facility, which is a function of the characteristics mentioned above and the total number of users. The objective is to find and analyze dynamic investment policies for quality and capacity that maximize the present value of the net social benefits derived from the operation of the public facility over a planning period [0,T].

With this objective, dynamic models are developed using optimal control theory formulations which consider the investments in quality and capacity as control variables. Alternative assumptions are used with respect to the continuity or discretness of the investments and the dependence or independence between the demand for the facility and its characteristics.

The models formulated are solved using different results of optimal control theory. Necessary and sufficient conditions for optimality are obtained in each case and economic interpretations are given. These conditions describe optimal dynamic investment rules not previously reported in the literature.

Thesis Supervisor: Fred Moavenzadeh
Title: Professor of Civil Engineering

Thesis Supervisor: Terry Friesz
Title: Assistant Professor of Civil Engineering

ACKNOWLEDGMENTS

I wish to thank my thesis committee chairman, Professor Fred Moavenzadeh, for his advice, guidance and support throughout the course of my graduate studies at M.I.T. His help was always invaluable and is greatly appreciated.

Professor Terry Friesz, co-chairman of the thesis committee, supervised this research. He taught me many things in the course of its development, but most of all, provided the indispensable encouragement when I needed it and always rewarded me with his friendship.

Further thanks go to my thesis committee members, Professor Ann Friedlaender, and Professor Alain Kornhauser who read the draft and made helpful comments. All the above mentioned constituted an ideal thesis committee, leaving me free to work at my own pace, but ready to help when necessary.

My fellow students in the Egypt project, especially Sergio Jara Diaz and Francisco Tourreilles, contributed to make my work more pleasant and were always willing to give a hand or provide moral support in the low moments. Patricia Vargas typed the draft and the final version of this thesis. Her patience and efficiency were invaluable in the final moments when everything tends to become hectic.

Finally, those to whom this thesis really belongs. My parents that ever put their sons as first priority and scarificed so much to make this possible. My wife Margarita for her constant support and understanding and my three sons, Jose Manuel, Jose Ignacio and Juan Jose. They contributed with the husband and father time that I should have dedicated to them during all this time, but that was invested in this thesis, and never made me feel guilty. Receive all of them my love and recognition.

TABLE OF CONTENTS

	Page
TITLE	1
ABSTRACT	2
ACKNOWLEDGEMENTS	3
TABLE OF CONTENTS	4
LIST OF FIGURES	8
LIST OF TABLES	9
I. INTRODUCTION	10
1. Scope of the Study	10
2. Methodology	13
3. Organization and Plan of the Study	14
II. OPTIMAL CONTROL MODELS	17
1. Introduction	17
2. Elements of Control Models	18
3. Necessary Conditions for Optimality	23
3.1 Continuous Systems with Final State Constraints and Free Terminal Time	24
3.2 Linear Systems and Singular Controls	30
3.3 Systems with Discontinuities in the State Variables and System Equations	40
4. Economic Interpretation of the Adjoint Variables and the Hamiltonian	45
5. Sufficient Conditions for Optimality	48
III. OPTIMUM POLICIES FOR INVESTMENTS IN QUALITY	50
1. Introduction	50
2. A Mathematical Model for Optimal Investments in Quality. Case of External Demand	51
2.1 Necessary Conditions for Optimality	54
2.2 Economic Interpretation of the Necessary Conditions	56

TABLE OF CONTENTS (continued)

	Page
3. A Mathematical Model for Optimum Investments in Quality. Internal Demand Case	67
3.1 Necessary Conditions for Optimality	68
3.2 Economic Interpretation of the Necessary Conditions	70
4. Extensions to the Case of Multiple User Types	84
5. Sufficient Conditions for Optimality	91
IV. OPTIMUM POLICIES FOR INVESTMENTS IN CAPACITY, CONTINUOUS CASE	95
1. Introduction	95
2. A Mathematical Model for Optimal Investments in Capacity	96
3. Necessary Conditions for Optimality: The Functional Form of Optimal Policies	101
3.1 Bang-Bang Controls	103
3.2 Singular Controls	105
3.3 Dynamic Optimum Policies	112
4. Sufficient Conditions for Optimality	116
5. Special Cases of Interest	120
5.1 Constant Returns to Scale in Capacity Construction	123
5.2 Decreasing Returns to Scale in Capacity Construction	130
5.3 Increasing Returns to Scale Case	137
V. OPTIMAL INVESTMENTS IN CAPACITY AND QUALITY. DISCRETE CASE	141
1. Introduction	141
2. A Mathematical Model for Optimal Staging of Capacity and Quality	142
3. Necessary Conditions for Optimality	144
4. Economic Interpretation. Optimal Investment Rules	146

TABLE OF CONTENTS (continued)

	Page
5. Numerical Solutions and Special Cases	153
VI. INFLUENCE OF DEMAND-QUALITY INTERRELATIONSHIPS ON OPTIMAL POLICIES OF STAGE CONSTRUCTION FOR TRANSPORTATION FACILITIES	157
1. Introduction	157
2. The Model	159
3. Solution of the Necessary Conditions	162
4. Sufficient Conditions	166
5. Final Remarks and a Numerical Example	176
VII. A MODEL OF OPTIMAL TRANSPORT MAINTENANCE WITH DEMAND RESPONSIVENESS	183
1. Introduction	183
2. Description of the Model	185
3. The Optimal Maintenance Policy: Necessary Conditions and Economic Interpretations	191
3.1 Bang-bang Policy	193
3.2 Singular Controls	198
4. Solution in a Particular Case	206
4.1 Singular Case	207
4.2 Bang-bang Case	209
5. An Algorithm for Determining Switching Times	224
APPENDIX A: Integration Constants	227
Case 1. $\Delta > 0$	227

TABLE OF CONTENTS (continued

	Page
Case 2. $\Delta < 0$	228
Case 3. $\Delta = 0$	230
APPENDIX B: Sufficient Conditions	233
VIII. SUMMARY OF RESULTS AND CONCLUSIONS	236
REFERENCES	240

LIST OF FIGURES

		Page
4.1	Operating Cost Function C (k,q)	98
4.2	Capacity Production Function f (k)	99
4.3	Bang-Bang Policy	106
4.4	Optimum Policies in (q,k)	113
4.5	Optimal Policies. Constant Returns to Scale Case	126
4.6	Capacity Production Costs Function Decreasing Returns Case	134
4.7	Optimal Policies. Decreasing Returns to Scale Case	135
4.8	Optimal Policies. Increasing Returns to Scale Case	139
5.1	Operating Cost Function, Before and After the Discrete Investment is Made	150
6.1	Graphical Representation of the Demand-Quality Decision Rule for the Case of Decreasing Marginal Benefits	173 173
7.1	Natural Deterioration Process	187
7.2	Bang-Bang Maintenance Policy	197
7.3	Maintenance Policy with a Singular Arc	205
7.4	Exponential Behavior of the System for the Case $\Delta > 0$	215
7.5	Oscilatory Behavior for the Case $\Delta < 0$	216

LIST OF TABLES

	Page
6.1 Values of t^* for Different Values of s_2 and U.	182

I. INTRODUCTION

1. Scope of the Study

As the title reveals the objective of this work is the study of dynamic optimal investment policies in public facilities, with special consideration given to the transportation case. However, this statement probably does not adequately convey an understanding of the analyses attempted in the several chapters that follow. In order to provide a better idea of the scope of the study we will give here a brief explanation of what it is meant by each of the key words appearing in the title.

The word "dynamic" has been extensively used in the economic literature without always implying the same meaning and rather often implying vague attributes of the analysis performed. In the words of Professor Samuelson*, "we damm another man's theory by terming it static, and advertise our own by calling it dynamic." Thus, it seems appropriate to agree with Professor Marglin** in that, in view of the loaded nature of the magic word "dynamic" it seems incumbent upon anyone who would employ it to demonstrate that he intends something more by its use that the distinction between bad and good. However, one feels a certain dissatisfaction with his subsequent definition of the word dynamic as merely a reference which"is supposed to convey the idea that time enters in an essential way". This point of view

*See P.A. Samuelson, [1947], pp. 311

**See S.A. Marglin, [1963], pp. 1

is still too unprecise to explain what is meant by the word dynamic in the title of this work.

We use the word dynamic to refer both to the systems under study and to the policies proposed in order to influence their behavior. The systems that we study are dynamic in the sense that their characteristics at any given time t will depend on the initial conditions at a certain time t_0 and the history of the policies applied over influencing them throughout the planning horizon or period of analysis $[t_0,t]$. As we will see in Chapter III the evolution of these systems will be governed by differential equations that specify the rate of change in the values of the variables that represent the main characteristics of the system. These rates of change will be influenced by the application of different policies, which will therefore, to some extent, determine the evolution of the system. To these policies we also apply the adjective dynamic.

We claim also that we will focus on "optimal" policies. The word optimal is obviously not justified in an absolute sense but we use it to mean that, given a model specification, with all the assumptions and simplifications that any modelling effort in general requires, an optimization methodology will be applied in order to obtain values for the independent policy variables used in the specification of the model. Thus, given a sound specification, the main characteristics of policies that optimize the behavior of the system, with respect to a prespecified performance index, should be obtained.

This study focuses on the analysis of investment policies in public facilities, for which we take as a special case transportation and construct illustrations and examples of the theory in terms of transportation. Although the same, or similar, models to those presented in the following chapters could be applied to other public facilities with similar economic characteristics to transportation, such as those related to power generation, communications or public recreation, a strong bias in the author's personal interests has influenced the decision of concentrating on transportation facilities.

We will assume that the facilities considered are provided and managed by some public authority. We will also assume that the users of these facilities obtain a certain utility and perceive a certain cost from the use of the facility. The cost perceived will determine what we call the level of service provided by the facility. If the cost is high the level of service will be low and vice versa. On the other hand, we will assume that a facility can be characterized at any time t by the values of two variables, one representing the capacity of the facility and the other its quality. By capacity we refer to the ability to accommodate a certain number of users with an arbitrary prespecified level of service. If the capacity is increased, eitehr the number of users accommodated could be increased, maintaining the same level of service will be improved if the number of users does not change. This trade-off is a consequence of the existence of congestion externalities in the consumption of the facility, a public good. The variable quality will represent those characteristics of the facility that do not affect its capacity

but influences the level of service perceived by the users. In general, the combination of the facility and its users will define the system over which our analysis will concentrate.

The word investment is used in this work to represent any expenditure, made by the public authority that manages the facility, with the objective of modifying the capacity or quality of the facility. An investment policy will be represented by a series of outlays indexed by time within the period of analysis [t,T].

2. Methodology

The features of dynamic investment decisions that we have briefly described are difficult to handle with the usual linear or non-linear programming techniques commonly used in static optimization. However, modern control theory, as we shall see in the following chapters, provides a natural framework for the analysis of the type of problems in which we are interested. All the models that we use, in order to analyze dynamic investment policies under different circumstances, are formulated as optimal control models. Actually, one of our objectives throughout this study has been to investigate the potentialities of this technique for the analysis of the type of problems described. Different model formulations that make use of special results of optimal control theory have been utilized in order to handle special characteristics of the problems studied. The treatment presented is in this sense new and has not been attempted before in the economics or transportation literature.

When an optimization technique is applied to any problem the main task is to find the necessary and sufficient conditions that characterize the optimal solution. If the problem at hand has a simple structure, these conditions can sometimes be solved analytically in terms of the unknowns with respect to which the optimization is being performed. Even if this is not the case, such conditions are still the fundamental base for the development of algorithms that could provide numerical solutions in special cases. Moreover, even though these optimality conditions may not be solvable analytically they are of value in themselves. In economic problems, like those treated in this study, their careful interpretation can provide important insights about the structure and characteristics of the optimum solutions. It should be remembered that many times models are developed not to provide solutions which are followed to the letter, but to provide additional information that together with all other pieces of information available helps to improve the process of decision making.

3. Organization and Plan of the Study

The remainder of this study is comprised of three main parts. The first one corresponds to Chapter II in which the main results of control theory to be used in the following chapters are set forth. The principal elements of control models are presented and necessary conditions for optimality are derived in a heuristic way for different dynamic models formulations to be used later. At the end of the chapter, a useful sufficiency theorem is stated.

The second part, comprised of Chapters III to V, is mainly dedicated to the theoretical analysis of the characteristics of optimal investments in quality and capacity under different general assumptions. In Chapter III optimal investments in quality are studied. Quality is considered there as a continuous variable whose evolution over time is defined by a general deterioration function. In the first part of the chapter, it is assumed that demand is externally specified and independent of the quality of the facility. In the second part this assumption is relaxed by the introduction of a dynamic equation that links demand to quality. The third part extends the models studied to the consideration of different types of users. Finally, sufficient conditions for optimality are derived at the end of the chapter.

Chapter IV is devoted to the analysis of optimal investments in capacity. A dynamic model is set forth assuming general construction and operating cost functions and taking capacity as a continuous variable. Optimal dynamic investment policies are derived and given economic interpretations. In the last part of the chapter the results obtained are applied to different special cases of interest which have been considered in the economic literature previously.

Chapter V presents a model formulation in which quality and capacity are taken simultaneously as decision variables. Capacity is considered an absolutely discrete variable that can take only certain prespecified values. Quality is assumed to be a piece-wise

continuous variable that can manifest discontinuities at the times that capacity is changed. Optimal staging policies for quality and capacity are obtained and given economic interpretations.

The third part of the study is comprised by Chapters VI and VII. Here a more applied approach is taken in order to derive decision rules or solution algorithms in special cases. Chapter VI studies the influence of demand-quality interrelations in the time staging of transport facilities. An optimal staging rule is derived and given economic interpretation. Sufficiency conditions under which the rule proposed is optimal are analyzed. Finally, a numerical example is developed in order to compare the results given by the rule proposed with those obtained from the application of rules available in the literature.

Chapter VII shows how to use the models developed in Chapter II in order to obtain solutions in a special case. With this purpose, the problem of determining optimal maintenance policies for a road is studied. Linear functional forms are assumed for the dynamics of quality and demand and the corresponding optimal maintenance policies are obtained. Then a method to obtain numerical solutions for a particular case is developed and a numerical algorithm is proposed.

Finally, in Chapter VIII a summary of the main conclusions is presented and suggestions for further research are provided.

II. OPTIMAL CONTROL MODELS

1. Introduction

The principal aim of this chapter is to present and describe the main results of control theory which are relevant for the later chapters. Some of these results, such as those related to singular controls (analyzed in Section 3.2) and to model formulations that allow discontinuities in the state variables and system equations (Section 3.3) on which Chapters 4 to 7 heavily rely, correspond to rather special results that may be difficult for the unfamiliar reader to find in the literature. Nevertheless, the presentation here is basically heuristic; the reader who wishes to see rigorous proofs of the results presented should consult the control theory texts referenced in the bibliography.

We begin by describing the main elements of control models in Section 2. In Section 3 we present the model formulations used in later chapters and the necessary conditions corresponding to their optimum solutions. In Section 4, we make use of some special results in order to give a general economic interpretation for the adjoint variables and the Hamiltonian. Finally, in Section 5 we present without any proof the Arrow theorem that is used in later chapters to derive sufficiency conditions for optimality.

The notation used throughout this chapter assumes that all vectors are column vectors, with exception of the gradients of any function which are assumed to be row vectors. Thenk if two vectors x and y exist in the same space R^n, x'y or y'x will denote the cartesian product, unless x represents a gradient of some function,

in which case will write xy or y'x'.

2. Elements of Control Models

Control theory deals with dynamic systems. Its objective is to find ways to optimize the evolution of a system over a certain period of time $[t_o, T]$ according to a given pre-specified criterion. Any system, be it physical, economic, or other, can in general be described at a given time t in terms of a set of variables of interest $y(t) = (y_1(t),.....,y_N(t))$. If all of these variables were out of our control (e.g. we cannot set the values of any of them) we would have a completely uncontrollable system from our point of view. The most we can hope for with respect to such a system is to develop a descriptive model of its behavior. The movement of celestial bodies could be a good example of this case. However, most of the systems that engineers and economists deal with are not of this type. In general, certain attributes of the system, represented by some of the variables y_i, can be controlled and through the interrelations of these with the rest of the variables, the behavior of the whole system can be influenced. There are still cases in which it doesn't matter that we can control the values of selected variables, the system is not controllable in a certain sense. The notion of controllability is a very important one in the study of dynamic systems and precise mathematical statements have been developed to define it. Nevertheless, we will not go into them here, given that we do not make any explicit use of them later. The reader interested in the topic can

consult introductory books in dynamic systems. All that we use in later chapters is the idea that the result of the application of the criterion used to evaluate the system can be influenced, by the manipulation of the controllable variables, and therefore the behavior of the system can be optimized with respect to this criterion.

We will denote by $V(t) = (V_1(t),\ldots,V_m(t))$, with $m < N$, the set of variables y_i that we can manipulate, which will receive the name of "control variables". The rest of the variables y_i will be represented by the vector $x(t) = (x_1(t),\ldots,x_n(t))$, with $n = N-m$, and will be called "state variables". We will have then

$$y(t) = (x(t),V(t)), \quad y \in R^N,\ x \in R^n,\ V \in R^m,$$
$$\forall\ t \in [t_o,T] \qquad\qquad (2\text{-}1)$$

where R^r denotes the space of r-dimensional vectors.

Thus, the first task in the specification of a control model is to select a set of variables $y(t)$ that can adequately describe the system of interest at any time t within $[t_o,T]$. The second step is to classify these variables into "controls" and "states". Sometimes this classification can be obvious from the characteristics of the variables involved. However, in itself and from the point of view of the model it is an arbitrary decision and will depend on the objectives of the analysis.

The next task is to define a model which indicates how the values of the state variables $x(t)$ evolve with time. In all contin-

uous-time control models it is considered that evolution of the system of interest can be described by a system of ordinary differential equations

$$\dot{x}(t) = dx(t)/dt = f(x(t),V(t),t), \quad t \in [t_0,T] \qquad ,(2\text{-}2)$$
$$x(0) = x_0$$

where in this case the function $f: R^{n+m+1} \rightarrow R^n$ provides a dynamic description of the system. Given the value of the states and the controls at a certain time t, (2-2) gives us the instantaneous rate of change in the value of the state variables. Also, if the function f is valid for all t in $[t_0,T]$ and we know the values of the state variables, x_0, at the initial time, the whole path x(.) followed by the state variables can be obtained through the integration of (2-2), provided that the values of the controls are specified for all t in $[t_0,T]$ and the following conditions are satisfied (see Athans and Falb [1966]).

1. The functions $f_i(x,V,t)$, $\partial f_i(x,V,t)/\partial x_j$ and $\partial f_i(x,V,t)/\partial t, (i,j = 1,\ldots,n)$ are continuous in $[t_0,T]$.
2. V(t) is a piecewise continuous function mapping from $[t_0,T]$ into R^m.

Therefore, it is not necessary that the controls V(t) be continuous over all $[t_0,T]$. Only the weaker condition of piecewise continuity is required. This is a general characteristic of all

continuous time control models. In most cases, the values permitted are also subject to constraints of the form

$$V(t) \in \Omega(t), \quad t \in [t_o,T] \qquad ,(2\text{-}3)$$

where Ω is a subset of R^m. It may also happen that all possible values of the states are not permitted, a requirement which can be expressed in a similar fashion as

$$x(t) \in X(t), \quad t \in [t_o,T] \qquad ,(2\text{-}4)$$

where X is a subset of R^n. The sets Ω and X are called the set of admissible controls and admissible states respectively. An important special case of (2-4) is

$$\psi[x(T)] = 0, \quad x(t) \in R^n \quad \forall\, t \neq T \qquad ,(2\text{-}5)$$

indicating that the final state x(T) is constrained to those values defined by $\psi = 0$, but the state at all other times is unrestricted.

The object of control theory is to choose the control function V(.) in order to optimize a stated objective function or measure of performance. The performance index is assumed to be of the form

$$J = k(x(T),T) + \int_{t_o}^{T} L(x(t),V(t),t)dt \qquad ,(2\text{-}6)$$

where k is a terminal payoff, that is assumed to be function of the

value taken by the state variables at the final time T and of the value of T in itself. L is an instantaneous performance index, evaluated at each time t in $[t_0,T]$, which is a function of the values taken by the states and the controls at the time and also of the value of t. Therefore, the value of J will depend on the values taken by the controls through the whole period $[t_0,T]$ and the specific path followed by the state variables x(t) during the same period. This path is defined by (2-2) for a given control function V(.). The inclusion of k in (2-6) allows one to give a special weight to the values taken by the state variables at time T.

We will assume in general that the functions L and k satisfy the following conditions:

1. The functions $L(x,V,t)/\partial x$ and $\partial L(x,V,t)/\partial t$ are continuous in the interval for which (2-6) is defined.
2. The functions k, $\partial k/\partial x$, $\partial k/\partial t$, $\partial x/(\partial x \partial t)$, $\partial k^2/\partial x^2$ and $\partial k^2/\partial t^2$ are also continuous.

For the analysis of economic systems, both k and L will represent benefits or costs depending on the case. In the analyses presented in subsequent chapters, J will always represent total benefits perceived from the operation of the system during the period $[t_0,T]$ and therefore the problem will be formulated as

$$\text{Max. } J; \text{ s.t. (2-2), (2-3) and (2-4).} \qquad (2\text{-}7)$$

An important element of any control model is what is called

the Hamiltonian function, which is defined as

$$H(t) \equiv L\ (x(t),V(t),t) + \lambda(t)f(x(t),V(t),t),\ t \in [t_0,T]\ , \quad (2\text{-}8)$$

where L and f are the functions defined in (2-6) and (2-2) and $\lambda(t)$: $[t_0,T] \rightarrow R^n$. The Hamiltonian plays a role in control models similar to that of Lagrangian in programming models; consequently, we can think of a λ as a dynamic generalization of the Lagrangian multiplier. These dynamic multipliers which receive the name of adjoint variables will be explicitly defined in the following sections, when we develop necessary conditions for optimality. A general economic interpretation will be also provided in Section 4. The Hamiltonian function, though mainly defined for notational convenience, can also be shown to have a general economic interpretation.

3. Necessary Conditions for Optimality

In this section we will develop necessary conditions for optimality for those model formulations used in later chapters. As we said before, the approach will be heuristic and we refer the reader to the relevant formal proofs in the literature. Our aim is to give an intuitive feeling for why the results presented hold and to motivate their later use. With this purpose we will make the derivations using only variational techniques.

3.1 Continuous Systems with Final State Constraints and Free Terminal Time.*

We will assume here that the continuity assumptions formulated in Section 2 for the functions $L(x,V,t)$ and $f(x,V,t)$ hold for all t in $[t_o,T]$. Our problem will be formulated as

$$\text{Max. } J = k(x(T),T) + \int_{t_o}^{T} L(x(t),V(t),t)dt \qquad ,(3\text{-}1)$$

subject to:

$$\dot{x} = f(x(t),V(t),t) \qquad ,(3\text{-}2)$$

$$\psi(x(T),T) = 0, \quad \psi : R^{n+1} \rightarrow R^{r} \qquad ,(3\text{-}3)$$

where (3-3) defines r general conditions that the state variables have to satisfy at time T. We will consider that this final time is unspecified.

The main idea is to introduce two sets of multipliers ν and $\lambda(t)$ that allow us to adjoin the equations (3-2) and (3-3) to the performance index (3-1), creating a function similar to the Lagrangian used in static optimization, and then to analyze the variations of this function around an optimal solution. Let ν be a vector representing the r multipliers associated with the r equations (3-3). Given that these equations are static conditions at time T, ν will actually be a vector of normal Lagrangian multipliers that take into

---

(*) The developments of this section are based on the work of J.V. Breakwell, [1959], as described in Bryson and Ho [1975].

account the influence of variations in the constraints (3-3) on the optimum value of the performance index. Let λ be a vector of n multipliers each of them associated with one of the dynamic equations (3-2). Since these equations can be interpreted as an infinity of static constraints indexed by t, these multipliers must be functions of time and will therefore be time-varying analogs of Lagrangian multipliers.

Adjoining the constraints (3-3) and the system differential equations (3-2) to the performance index by means of the multipliers ν and λ (t) we obtain

$$J = [k + \nu'\psi] + \int_{t_o}^{T} \{L(x,V,t) + \lambda'[f(x,V,t) - \dot{x}]\}dt. \qquad (3\text{-}5)$$

Now paraphrasing the theory of Lagrangian multipliers, it follows that in order for V(.) and x(.) to be an optimal control and an optimal trajectory, the variations dJ of (3-5) around the optimal solution must be equal to zero.

The differential of (3-5), taking into account differential changes of x, V, t_o and T can be written as

$$dJ = [\Phi_t(T) + L(T)dT + \Phi_x(T)dx] - L(t_o)dt_o + \int_{t_o}^{T} (H_x\delta x + H_V\delta V - \lambda'\delta\dot{x})dt \qquad ,(3\text{-}6)$$

where we have used the definition of H(t) given in (2-8) and the following notation:

$$\Phi(T) = k(x(T),T) + \nu'\psi(x(T),T) \qquad ,(3\text{-}7)$$

$$\Phi_t = \partial\Phi/\partial t, \qquad \Phi_x = \partial\Phi/\partial x$$

$$H_x = \partial H/\partial x, \qquad H_V = \partial H/\partial V$$

and δx, the variation in x, means "for time held fixed." Therefore dx, the total differential in x, may be written for any time t

$$dx(t) = \delta x(t) + \dot{x}(t)dt \qquad ,(3\text{-}8)$$

Now, integrating the term $-\lambda'\delta\dot{x}$ by parts in (3-6) we obtain

$$-\int_{t_o}^{T} (\lambda'\delta\dot{x})dt = \lambda'(t_o)\delta x(t_o) - \lambda'(T)\delta x(T) + \int_{t_o}^{T} (\dot{\lambda}'\delta x)dt \qquad ,(3\text{-}9)$$

Thus, if we make use of (3-9), to replace the third term of the integral in (3-6) and of the following relations obtained from (3-8)

$$\delta x(T) = dx(T) - \dot{x}(T)dT$$

$$\delta x(t_o) = dx(t_o) - \dot{x}(t_o)dt \ ,$$

we can write dJ as

$$
\begin{aligned}
dJ = {} & [\Phi_t(T) + L(T) + \lambda'(T)\dot{x}(T)]dt \\
& + [\Phi_x(T) - \lambda'(T)]\, dx(T) + \lambda'(t_o)dx(t_o) \\
& -[L(t_o) + \lambda'(t_o)\dot{x}(t_o)]\, dt_o \\
& + \int_{t_o}^{T} [(H_x + \dot{\lambda}')\delta x + H_v \delta V]dt
\end{aligned}
\qquad ,(3\text{-}10)
$$

We have therefore the variation of J expressed in terms of variations of the variables $x(t_o)$, $x(T)$, $x(t)$, $V(t)$, t_o and T. If any of these variables is given, its value will be fixed and the corresponding variation will be zero causing the term which it is multiplying to disappear (e.g. if x_o and t_o are given, then the third and fourth terms in the right hand side of (3-10) will disappear). If all mentioned variables are assumed free (their values are not externally specified as data), at an optimum solution the value of dJ must be equal to zero for all possible values of dt_o, dT, $dx(t_o)$, $dx(T)$, $\delta x(.)$ and $\delta V(.)$. This implies that the coefficients of all these variations in (3-10) must be zero, otherwise we could always find a set of variations for which $dJ > 0$. This leads to the following necessary conditions:

$$
\dot{\lambda} = -H_x = -L_x - \lambda' f_x, \quad \forall\ t \in [t_o, T] \qquad ,(3\text{-}11)
$$

which are called adjoint equations and must satisfy the boundary conditions:

$$
\lambda(T) = \Phi_x(T) = k_x(T) + \nu'\psi_x(T) \qquad ,(3\text{-}12)
$$

usually called transversality conditions.

If the final time T is free, a necessary condition for an optimum value of this variable can be derived by setting the coefficient of dT in (3-10) equal to zero to obtain:

$$\Phi_t(T) = -H(T) = -L(T) - \lambda'(T)f(T). \qquad (3\text{-}13)$$

Similarly, if t_o and $x(t_o)$ are not specified, necessary conditions for optimality in these variables are given by:

$$H(t_o) = L(t_o) + \lambda'(t_o)f(t_o) = 0 \qquad (3\text{-}14)$$

$$\lambda(t_o) = 0 \ . \qquad (3\text{-}15)$$

Obviously, if t_o and $x(t_o)$ are given then dt_o and $dx(t_o)$ are identically zero and therefore $\lambda(t_o)$ and $H(t_o)$ can take any value.

Finally, if no constraints exist for V(t), the variations δV can also be arbitrary (within the restrictions imposed by the piecewise continuity characteristic that we required in Section 2) and therefore its coefficient in (3-10) must also vanish giving:

$$H_V(x^*,\lambda^*,V,t) = 0, \quad \forall\ t \in [t_o,T] \qquad ,(3\text{-}16)$$

where the * means that x and λ satisfy the equations (3-2) and (3-11, 3-12) respectively. Notice that by using the definition of the Hamiltonian, we can rewrite (3-2) as

$$\dot{x} = H_{\lambda}(x,\lambda,V,t), \quad \forall\ t \in [t_o,T] \qquad ,(3\text{-}17)$$

which is in a sense symmetrical with respect to condition (3-11). Condition (3-16) implies that at an optimum solution, the Hamiltonian obtains an extremum value with respect to the control V for all t in $[t_o,T]$. Pontryagin's maximum principle guarantees in addition that this extremum must correspond to a maximum value of the Hamiltonian with respect to the control (Pontryagin et.al. [1964]).

Until now we have assumed that there are no constraints on the values which the state variables and the controls can take. If we introduce control constraints of the form:

$$V(t) \in \Omega, \quad \forall\ t \in [t_o,T]$$

where Ω is a convex set and $\Omega \subset R^m$, then the variations $\delta V(t)$ in (3-10) are no longer arbitrary. For instance, if the optimal control is located over a boundary of Ω, only variations of V toward the interior of Ω can be considered. In this case the condition $dJ = 0$ at an optimum solution must be replaced by the condition $dJ \leqq 0$, given that we have a maximization formulation in (3-1). Because all the other variables considered are unconstrained their variations are still arbitrary and therefore their coefficients in (3-10) must vanish as before. Thus, the necessary conditions (3-11) to (3-15) are still valid. As a result of that we can write:

$$dJ = \int_{t_o}^{T} (H_V \delta V)\, dt \leqq 0 \qquad ,(3\text{-}18)$$

at an optimum solution. This condition will be satisfied for all admissible variations δV if

$$H_V^* \delta V \leq 0, \quad \forall\ t \in [t_o, T] \qquad ,(3\text{-}19)$$

$$\delta V = (V - V^*), \quad V \in \Omega$$

which is a necessary condition for the maximization of the function H with respect to the variable V over the convex set Ω at each time $t \in [t_o, T]$ (see Luenberger, [1973]). Therefore, the maximization of the Hamiltonian as a necessary condition for optimality also carries over to this more general case.

3.2 Linear Systems and Singular Controls*

In this section we will analyze a very important special case that appears when both the dynamic equations and the integrand of the objective function are linear in the controls. It is easy to see that then the Hamiltonian will also be linear in the controls and can be written as:

$$\begin{aligned} H &= L(x,V,t) + \lambda' f(x,V,t) \\ &= G(x,\lambda,t) + V'F(x,\lambda,t); \quad G: R^{2n+1} \to R^1, \ F: R^{2n+1} \to R^m, \end{aligned} \qquad (3\text{-}20)$$

where G and F can be in the general case non-linear functions of x but

(*) This section is mainly based on the treatments for linear control problems presented in Athans and Falb [1966] and Bryson and Ho [1975].

independent of V.

Obviously, all the necessary conditions developed in the preceding section are valid for this special case. In particular, at an optimal solution the Hamiltonian must attain a maximum with respect to V, for each t in $[t_o,T]$. Nevertheless, a problem appears now in satisfying this necessary condition. We know that linear functions never attain a maximum in R^m with respect to those variables for which the slope of the corresponding hyperplane is different from zero and for those which slope is zero any value in R^1 corresponds to a maximum. In order to avoid notational complications we will assume in the subsequent discussion that our problem has a single control. The concepts are the same for problems with many control variables. Then, the variable V will exist in R^1, the function f will be a mapping from R^{n+2} to R^n and the functions G and F will be mappings from R^{n+2} to R^1. Thus, if no additional constraints are defined for the states and/or the control, the necessary condition requiring the maximization of the Hamiltonian does not provide any useful information to characterize the optimal control. Actually, if the functions L and f (and therefore also G and F) are simultaneously linear in the states and the control (with F independent of x), the control problem will be completely linear and a maximum does not exist, unless constraints are imposed on x and/or V.

In this and subsequent chapters we only deal with control constraints whose general form will be:

$$V(t) \in \Omega, \quad \forall\, t \in [t_o,T] \tag{3-21}$$
$$\Omega = \{V(t): m(t) \leqq V(t) \leqq M(t), \quad t \in [t_o,T]\}$$

Then, the condition of maximizing the Hamiltonian subject to the constraint $V(t) \in \Omega$ gives the following functional form for the optimal control $V^*(.)$.

$$V^*(t) = \begin{cases} M(t) & \text{,if } F(t) > 0 \\ m(t) & \text{,if } F(t) < 0 \\ \text{undetermined,} & \text{if } F(t) = 0 \end{cases} \qquad (3\text{-}22)$$

If the Hamiltonian is completely linear in the states and the control, it is reasonable to expect that the maximum solution to our problem will always require the control variable to be at one point or another on the boundary of the feasible region Ω. In general, one or more changes in control, from one point on the boundary to another point on the boundary, will occur during the time of operation of the system. The times, t, at which thecontrol switches are identified by the condition $F(t) = 0$ and will correspond to only a countable set in $[t_0,T]$. The optimal controls thus defined receive the name of "bang-bang" controls. In later chapters we will also call bang-bang those portions of an optimal control history during which the control obtains the value of one of the boundaries of its feasible region.

Nevertheless, if the Hamiltonian is non-linear in the state variable x, or presents cross terms in the state and control, the value of $F(t)$ can vanish identically over a finite interval of time in $[t_0,T]$ and then (3-22) does not provide a complete definition of $V^*(t)$ along $[t_0,T]$. The portions of the optimal trajectory of the system for which $F(t) = 0$ are called "singular arcs". In that case, we

must manipulate the other necessary conditions provided in Section 3.1 in order to determine the optimal value of the control, along the singular arc, which will receive the name of singular control.

Given that the dynamic equation and the performance index are linear in the control, we can write:

$$f = a(x,t) + Vb(x,t) \tag{3-23}$$

with, $a : R^{n+1} \to R^n$; $b : R^{n+1} \to R^n$

$$L = c(x,t) + Vd(x,t) \tag{3-24}$$

with, $c : R^{n+1} \to R^1$.

Then the expression of the Hamiltonian becomes:

$$H = (c + \lambda'a) + V(d + \lambda'b) \tag{3-25}$$

where according to our previous notation in (3-20)

$$G = c + \lambda'a; \quad F = d + \lambda'b$$

Also the gradient of the Hamiltonian will be:

$$H_V = F = d + \lambda'b \tag{3-26}$$

If the gradient of the Hamiltonian vanishes identically on a singular arc, its value during this period will be constant and equal to zero and therefore all its time derivatives must also vanish during the

same period. We will use this property in order to derive an expression for the singular control, V_s. Thus, the first necessary condition for a singular arc will be:

$$F = d + \lambda' b = 0, \quad \text{or, } \lambda' b = -d, \quad \forall\, t \in (t_1, t_2], \tag{3-27}$$

where (t_1, t_2) is a sub-interval of $[t_o, T]$. Then, if we take the first time derivative of F, we will have:

$$\dot{F} = d_x \dot{x} + d_t + b' \dot{\lambda} + \lambda' (b_x \dot{x} + b_t) = 0. \tag{3-28}$$

However, from necessary condition (3-12) we obtain that:

$$\dot{\lambda} = -H_x = -c_x' - a_x' \lambda - V(d_x' + b_x' \lambda). \tag{3-29}$$

and using (3-23) and (3-29) to eliminate $\dot{x}$ and $\dot{\lambda}$ from (3-28) we can write F as:

$$\dot{F} = d_x a - c_x b + d_t + \lambda' (b_x a - a_x b + b_t) = 0 \tag{3-30}$$

where, b_x is the Jacobian matrix of b with respect to x, a_x is the Jacobian of a with respect to x, c_x is the gradient of c with respect to x and d_x is the gradient of d with respect to x. Note that still the control V_s does not appear explicityly in (3-30). If the reader follows the derivation of (3-30) from (3-28) he will notice that the reason is that it is multiplied by a factor that is identically zero. This is in general a characteristic of the first

derivative of the Hamiltonian along a singular arc. Therefore, (3-30), although it constitutes a new necessary condition for the existence of a singular control, does not provide an explicit expression for it. New derivatives must be taken in order to obtain that expression, if it exists. Nevertheless, before doing that let us use expression (3-30) in order to show some assertions made earlier with respect to the completely linear case.

If the problem is completely linear in the states and the control, the functions b and d must be independent of the states x (otherwise cross-terms in the states and control would appear, and the problem would not be completely linear). In addition, the Jacobian a_x and the gradient c_x will only be functions of t (because a and c are linear functions of x). Therefore we will have:

$$d_x'a = 0; \qquad b_x a = \{0\}$$

and $c_x b$, $a_x b$, d_t and b_t are only functions of time. Therefore (3-30) can be written as:

$$\Theta(t) + \lambda'\eta(t) = 0 \tag{3-31}$$

where

$$\Theta(t) = -c_x b + d_t; \qquad \eta(t) = \{-a_x b + b_t\}.$$

From (3-27) we also have that λ' will be a function of t only along the singular arc and therefore (3-31) will be an equation in t that in general will be satisfied for a countable number of points t in $[t_o, T]$. Therefore no singular arc or singular control exists

in that case and the solution must be purely bang-bang.

Going back to the general case we can take a new time derivative of F to obtain:

$$\begin{aligned}\ddot{F} &= a'd_{xx}\dot{x} + d_{xt}a + d_x a_x\dot{x} + d_x a_t \\ &+ b'c_{xx}\dot{x} + c_{xt}b + c_x b_x\dot{x} + c_x b_t \\ &+ d_{tt} + \dot{\lambda}'(b_x a - a_x b + b_t) \\ &+ \Big(\sum_1^n \lambda_i a'b_{ixx}\Big)\dot{x} - \Big(\sum_1^n \lambda_i b'a_{ixx}\Big)\dot{x} \\ &+ \lambda'(b_x a_x\dot{x} - a_x b_x\dot{x} + b_{xt}a + b_x a_t \\ &- a_{xt}b - a_x b_t + b_{tt}) = 0\end{aligned} \tag{3-32}$$

where d_{xx}, c_{xx}, b_{ixx} and a_{ixx} are the Hessian matricies of the functions d, c, b_i and a_i respectively. If we rearrange terms in (3-32) we can write:

$$\ddot{F} = e'\dot{x} + \dot{\lambda}'g + h = 0 \tag{3-33}$$

where
$$\begin{aligned}e'(x,\lambda,t) &= a'd_{xx} + b'_x c_{xx} + c_x b_x \\ &+ \Big(\sum_1^n \lambda_i a'b_{ixx}\Big) - \Big(\sum_1^n \lambda_i b'a_{ixx}\Big) \\ &+ \lambda'(b_x a_x - a_x b_x)\end{aligned}$$

$$g(x,t), = b_x a - a_x b + b_t$$

$$h(x,\lambda,t) = d_{xt}a + d_x a_t + c_{xt}b + c_x b_t + d_{tt}$$
$$+ \lambda' (b_{xt}a + b_x a_t - a_{xt}b - a_x b_t + b_{tt}).$$

Upon using (3-23) and (3-29) to eliminate $\dot{x}$ and $\dot{\lambda}$ from (3-33) we have:

$$\ddot{F} = \phi_2(x,\lambda,t) + V_s\psi_2(x,\lambda,t) = 0 \qquad (3\text{-}34)$$

where $\phi_2 = e'a + h - (c_x - \lambda' a_x)g$

$\psi_2 = e'b - (d_x + \lambda' b_x)\, g.$

In general, the coefficient of V_s, ψ_2, will not be zero for all $t \in (t_1, t_2)$. When $\psi_2 \neq 0$, we can solve equation (3-34) for V_s, to find that

$$V_s = \phi_2(x,\lambda,t)/\psi_2(x,\lambda,t). \qquad (3\text{-}35)$$

If, on the other hand $\psi_2 = 0$, then equation (3-34) reduces to

$$\ddot{F} = \phi_2(x,\lambda,t) = 0 \qquad (3\text{-}36)$$

and we must take new derivatives of F until for some ith derivative we find a function $\psi_k \neq 0$ that allows us to obtain an explicit expression for the singular control V_s:

$$V_s = \phi_k\,(x,\lambda,t)/\psi_k(x,\lambda,t) \qquad ,(3\text{-}37)$$

The reader can appreciate the notational difficulties involved in the computation of the higher derivatives in terms of the original functions a, b, c and d.

It is important to note that if an explicit expression for V_s is obtained at the kth derivative of F with respect to time, the equations

$$\phi_i (x,\lambda,t) = 0, \quad i = 0, 1,\ldots,(k-1) \qquad ,(3\text{-}38)$$

constitute k necessary conditions* for the existence of the singular control V_s. In the case of economic problems, these necessary conditions can provide economic interpretations for the optimum policies along a singular arc, as we will see in Chapters 4 and 7.

In Section 3.1 we saw that a necessary condition for an optimum solution to control problems is the maximization of the Hamiltonian with respect to the control variables. As we have just seen in this section, an expression for a singular control is obtained from the condition:

$$H_v = F = 0, \quad \forall\ t \in (t_1, t_2) \qquad ,(3\text{-}39)$$

which only constitutes a first order necessary condition for the maximization of the Hamiltonian along a singular arc. A second order necessary condition for maximization is in general provided by the concavity condition:

(*) Note that $\phi = 0$ and $\phi_1 = 0$ are defined in (3-27) and (3-30) respectively

$$H_{VV} \leqq 0 \ . \qquad (3\text{-}40)$$

For singular arcs, $H_{VV} \equiv 0$, so condition (3-40) yields no useful information. A more useful condition for this case was derived by Tait [1965]; Kelley, Kopp, and Moyer [1966]; and Robbins [1965]. Its derivation using variational techniques is far from being straight forward and can be found in Bryson and Ho [1975], so we will only give here the statement that will be used in the following chapters. For a maximization problem with a single control variable the condition can be stated as:

$$(-1)^{k/2} \ \frac{\partial}{\partial V}[(d/dt)^k H_V] \leqq 0 \qquad ,(3\text{-}41)$$

which in terms of our notation in the present section can be written as:

$$(-1)^{k/2} \psi_k(x,\lambda,t) \leqq 0 \qquad ,(3\text{-}42)$$

where k is the order of the time derivative of F at which $\psi_i \neq 0$ for the first time.

Thus, for problems that are linear in the controls but present non-linearities in the states, or cross-terms in the states and controls, the functional form of the optimal control V^* can be written as:

$$V^*(t) = \begin{cases} M(t), \text{ if } F(t) > 0 \\ m(t), \text{ if } F(t) < 0 \\ \phi_k/\psi_k, \text{if } F(t) = 0 \end{cases} . \qquad (3\text{-}43)$$

It is worth noting that the singular control is also constrained to be within Ω (see, 3-21). If the problem is completely linear in the states and controls, the condition $F(t) = 0$ will be attained at most at a countable number of times in $[t_o,T]$ and the expression of the optimal control can be simplified to:

$$V^*(t) = \begin{cases} M(t), \text{ if } F(t) > 0 \\ m(t), \text{ if } F(t) < 0 \end{cases} . \qquad (3\text{-}44)$$

3.3 Systems with Discontinuities in the State Variables and System Equations

Some models that we will analyze in this thesis (see, Chapters 5 and 6) present discontinuities in the state variables as well as discontinuities of the system equations at interior points in $[t_o,T]$. Furthermore, the performance index on the constraints may be functions of the state and/or time at several discrete points in $[t_o,T]$. In that case we can partition the interval $[t_o,T]$ in N sub-intervals (t_{i-1},t_i), $(i = 1,....N)$ where $t_N = T$ and t_i $(i = 1,...,(N-1))$ are the above mentioned interior points of discontinuity. Now the continuity assumptions formulated in Section 2 for the functions $L(x,V,t)$ and $f(x,V,t)$ will hold within each sub-interval (t_{i-1},t_i) and the problem can be formulated as:

$$\text{Max: } J = \sum_{i=0}^{N} k_i(x(t_i^-),\ x(t_i^+),\ t_i)$$

$$+ \sum_{i=1}^{N} \int_{t_{i-1}^+}^{t_i^-} L_i(x(t),V(t),t)dt \qquad ,(3\text{-}45)$$

subject to:

$$\dot{x} = f_i(x,V,t); \quad t_{i-1} < t < t_i, \quad i = 1,\ldots,N \qquad ,(3\text{-}46)$$

$$\psi_j(x(t_j^-),x(t_j^+),t_j) = 0, \quad j = 0,\ldots,N \qquad ,(3\text{-}47)$$

where (3-47) defines general conditions that the state variables have to satisfy at the points t_j, $j = 1,\ldots,N$. The notation t_i^- and t_i^+ is used to represent the moments just before and after $t = t_i$.

In this case the set of variables that we wish to optimize is $(x(t);V(t);\ t_i, i = 0,\ldots,N)$. Necessary conditions for a maximum of J defined in (3-45), with respect to these variables can be derived by adjoining (3-46) and (3-47) to (3-45) by Lagrangian multiplier functions $\lambda(t)$ and constant Lagrangian multipliers ν_i respectively:

$$J = \sum_{i=0}^{N} [k_i + \nu_i' \psi_i]$$

$$+ \sum_{i=1}^{N} \int_{t_{i-1}^+}^{t_i^-} \{L_i(x,V,t) + \lambda' [f_i(x,V,t) - \dot{x}]\}\, dt\ . \qquad (3\text{-}48)$$

Now we can use the same variational procedure of Section 3.1 to obtain necessary conditions for optimality. It consists in producing variations of the independent variables around an optimal solution and analyzing the corresponding variations dJ for the

Lagrangian represented in (3-48). As in Section 3.1, to simplify the notation we will define:

$$\phi_i \equiv k_i(x(t_i^-),x(t_i^+),t_i) + \nu_i' \psi_i(x(t_i^-),x(t_i^+),t_i), \qquad (3\text{-}49)$$

$$H_i \equiv L_i(x,V,t) + \lambda' f_i(x,V,t) \qquad ,(3\text{-}50)$$

where H_i is the Hamiltonian for the interval $[t_{i-1}^+,t_i^-]$. Then, the first variation of J, taking into account variations of $x(t)$, $x(t_i^-)$, $x(t_i^+)$, $V(t)$ and t_i, $(i = 0,\ldots,N)$, can be written as:

$$\begin{aligned} dJ = & \sum_{i=0}^{N} [\phi_t^i \, dt_i + \phi_x^{i-} \, dx(t_i^-) + \phi_x^{i+} \, dx(t_i^+)] \\ & + \sum_{i=1}^{N} \{[H_i(t_i^-) - \lambda'\dot{x}(t_i^-)]dt_i - [H_i(t_{i-1}^+) - \lambda'\dot{x}(t_{i-1}^+)]dt_{i-1} \\ & + \sum_{i=1}^{N} \int_{t_{i-1}^+}^{t_i^-} [H_x^i \, \delta x + H_V^i \, \delta V - \lambda'\delta\dot{x}] \, dt \qquad ,(3\text{-}51) \end{aligned}$$

where we have used the following notation

$$\phi_t^i = (\partial\phi_i/\partial t_i)$$

$$\phi_x^{i-} = [\partial\phi_i/\partial x(t_i^-)], \qquad \phi_x^{i+} = [\partial\phi_i/\partial x(t_i^+)]$$

$$H_x^i = (\partial H_i/\partial x), \qquad H_V^i = (\partial H_i/\partial V) \; .$$

Now, if we use the intergration by parts:

$$- \int_{t_{i-1}^+}^{t_i^-} (\lambda' \dot{\delta x})\, dt = \lambda'(t_{i-1}^+)\, \delta x(t_{i-1}^+) - \lambda'(t_i^-)\, \delta x(t_i^-) + \int_{t_{i-1}^+}^{t_i^-} (\dot{\lambda}' \delta x)\, dt \qquad ,(3\text{-}52)$$

and the relations:

$$\begin{aligned} dx(t_i^-) &= \delta x(t_i^-) + \dot{x}(t_i^-)\, dt_i \\ dx(t_i^+) &= \delta x(t_i^+) + \dot{x}(t_i^+)\, dt_i \end{aligned} \qquad ,(3\text{-}53)$$

in expression (3-51) and regroup terms, we have:

$$\begin{aligned} dJ = &\sum_{i=0}^{N} [\phi_t^i + H_i(t_i^-) - H_{i+1}(t_i^+)]\, dt_i \\ &+ \sum_{i=1}^{N} [\phi_x^{i-} - \lambda'(t_i^-)]\, dx(t_i^-) + \sum_{i=0}^{N-1} [\phi_x^{i+} + \lambda'(t_i^+)]\, dx(t_i^+) \\ &+ \sum_{i=1}^{N} \int_{t_{i-1}^+}^{t_i^-} [\dot{\lambda}' + H_x^i)\delta x + H_V^i \delta V]\, dt \quad . \end{aligned} \qquad (3\text{-}54)$$

As in Section 3.1, the first variation of the Lagragian, dJ, has to vanish at an optimum solution, for arbitrary variations of the independent variables, if no path constraints exist for the state variables x(t) and the control variables V(t). This requires that the coefficients of dt_i, $dx(t_i^-)$, $dx(t_i^+)$, δx and δV in (3-54) be equal to zero, which provides the following necessary optimality

conditions:

$$\dot{\lambda}' = -H_x^i; \quad t_{i-1}^+ < t < t_i^-, \ (i = 1,\ldots,N) \qquad ,(3\text{-}55)$$

$$\lambda'(t_i^-) = \phi_x^{i-} \equiv k_x^{i-} + \nu_i' \psi_x^{i-}; \quad (i = 1,\ldots,N) \qquad ,(3\text{-}56)$$

$$\lambda'(t_i^+) = \phi_x^{i+} \equiv k_x^{i+} + \nu_i' \psi_x^{i+}; \quad (i = 0,\ldots,N-1) \qquad ,(3\text{-}57)$$

$$\phi_t^i \equiv k_t^i + \nu_i' \psi_t^i = H_{i+1}(t_i^+) - H_i(t_i^-); \quad (i = 0,\ldots,N), \qquad (3\text{-}58)$$

with

$$H_o = H_{N+1} \equiv 0$$

$$H_V^i = 0; \ t_{i-1}^+ < t < t_i^-, \ (i=1,\ldots,N) \ . \qquad (3\text{-}59)$$

We must also choose ν_i in such a way to satisfy the constraints $\psi_j = 0$ in (3-47). Actually, additional arguments must be made here in order to justify the vanishing of the Hamiltonian gradient H_V^i expressed in (3-59) since, no matter that no path constraints exist for V(t) inside $[t_{i-1}^+, t_i^-]$, δV is not completely arbitrary and must produce variations $dx(t_i^-)$, $dx(t_i^+)$ and dt_i consistent with (see, Bryson and Ho, [1975]).

$$d\psi_j = \psi_t^j \, dt_j + \psi_x^{j-} \, dx(t_j^-) + \psi_x^{j+} \, dx(t_j^+) = 0 \ . \qquad (3\text{-}60)$$

Equation (3-55) defines the adjoint equations, for the adjoint variables λ, along each sub-interval $[t_{i-1}^+, t_i^-]$. Equations (3-56) and

(3-57) define transversality conditions that the adjoint variables must satisfy, at the beginning and end of each sub-interval, in order that the values of $x(t_{i-1}^{+})$ and $x(t_i^{-})$ be optimal. Equation (3-59) corresponds to the familiar Hamiltonian maximization condition and equation (3-58) provides a set of transversality conditions for the Hamiltonian, that must be satisfied for optimal solutions of the times t_i ($i = 0,\ldots,N$). Obviously, if some variable considered here as independent is externally specified, the corresponding variation in (3-54) will be zero and the necessary condition attached to it will disappear. For instance, if $x(t_0^{+})$ is specified, we have $dx(t_0^{+}) = 0$ in (3-54), and equation (3-57), with $i = 0$, is not required. Similarly, if t_i is specified, the corresponding relation in (3-58) is not required since $dt_i = 0$ in (3-54).

Finally, note that equations (3-54), (3-57) and (3-58) imply discontinuities of the adjoint variables and the Hamiltonian at each point t_i.

4. Economic Interpretation of the Adjoint Variables and the Hamiltonian

In this section, we will take advantage of some results obtained in Section 3, in order to provide general economic interpretations for two fundamental elements of any control problem: the adjoint variables and the Hamiltonian function.

Let us use the system formulation of Section 3.1 and let us consider the first order variation dJ of the performance index given by equation (3-10). Then, if we consider an optimum solution $x^{*}(t)$,

$\lambda^*(t)$, $V^*(t)$ that satifies the necessary conditions (3-2), (3-3), (3-11), (3-12) and (3-16) for a value of T satisfying (3-13), we can write the variation of the corresponding performance index value J^* produced by variations in the initial conditions $t_o, x(t_o)$ as

$$dJ^* = \lambda'(t_o)dx(t_o) - H(t_o)\, dt_o \quad . \tag{4-1}$$

From this we obtain:

$$\lambda^*(t) = \frac{\partial J^*}{\partial x(t_o)} \qquad H^*(t_o) = -\frac{\partial J^*}{\partial t_o} \quad .$$

However, because the above is true for arbitrary t_o we can write

$$\lambda^*(t) = \frac{\partial J^*}{\partial x(t)} \tag{4-2}$$

$$H^*(t) = -\frac{\partial J^*}{\partial t} \tag{4-3}$$

Actually, if we use the principle of optimality (Bellman, [1957]) we have that "any portion of an optimal trajectory $J^*(t)$, $x^*(t)$ is also an optimal trajectory"(see Athans and Falb, [1966]). Therefore, we can partition our initial interval $[t_o,T]$ in two sub-intervals $[t_o,T]$, $[t,T]$ and rewrite the performance index of Section 3.1 as

$$J = k(x,T) + \int_{t_o}^{t} L(x,V,t)dt + \int_{t}^{T} L(x,V,t)\, dt \, . \tag{4-4}$$

Then, if we consider the sub-problem of optimizing J along the sub-interval [t,T], an optimum solution to it has to be coincident with the portion [t,T] of an optimal solution for the problem defined over the whole interval $[t_o T]$. Thus, the variation of J^* with respect to t and x(t) will be:

$$dJ^* = \lambda'(t)dx(t) + H(t)dt, \quad t \in [t_o,T] \quad . \qquad (4\text{-}5)$$

We must note however that (4-2) and (4-3) are true only if the performance index is evaluated along an optimal solution; otherwise expressions (4-1) or (4-5) are not valid.

From (4-2) we have that each adjoint variable represents the change experienced in the value of the objective function as a consequence of a change in the corresponding state variable, around its optimum value, at time t. Therefore, the adjoint variables can be interpreted as shadow prices for unitary values of the state variables, around an optimal solution. They are, actually, dynamic shadow prices, functions of time. In mathematical and economic terms the adjoint variables are, as we have suggested before, dynamic generalizations of the concept of Lagrangian multipliers. On the other hand, we have from (4-3) that the Hamiltonian, evaluated along an optimal solution path, gives us the change in the value of the objective function per unit of time. If our performance index represents the total benefits (or total costs) associated with the operation of the system, along the period $[t_o,T]$, the value of the Hamiltonian, at time t, represents the total marginal benefit (or

marginal cost), per unit of time, at time t.

5. Sufficient Conditions for Optimality

In Section 3 we have developed necessary conditions for optimal control problems which are known in the literature as the Pontryagin maximum principle. Nevertheless, these conditions are not, in general, sufficient for optimality. In the mathematical literature few and only rather special results were available until Mangasarian [1966] proved a rather general sufficiency theorem in which he was dealing with a non-linear system, state and control variable constraints and a fixed time interval. In the maximization case, when there are no state space constraints, his result was, essentially, that the Pontryagin necessary conditions plus concavity of the Hamiltonian with respect to the state and control variables, were sufficient for optimality.

The Mangasarian concavity condition is rather strong and in many economic problems his theorem does not apply. A very interesting generalization of the Mangasarian result was however proposed by Arrow [1968]. A precise statement and a rigorous proof of the Arrow sufficiency theorem has been given only recently by Seierstad and Sydsaeter [1977]. For the type of systems described in Section 3.1, when the interval $[t_0,T]$ is fixed and the initial conditions $x(0) = x_0$ are specified, the theorem can be expressed as follows: "Suppose $(x^*(t),V^*(t))$ is an admissible pair satisfying all the necessary conditions for optimality. Then, if $H^*(x,\lambda(t),t)$,

as defined below, is concave in x, we have that $(x^*(t),V^*(t))$ is an optimal solution to the problem"

$$H^*(x,\lambda,t) = \underset{V \in \Omega}{\text{Max}}\ H(x,V,\lambda,t) \qquad (5\text{-}1)$$

III. OPTIMUM POLICIES FOR INVESTMENTS IN QUALITY

1. Introduction

We can argue that a transportation facility can be characterized in general by two attributes: quality and capacity. The concept of capacity is easy to understand and has been the one that has received more attention in the economic literature. In the next chapter we will develop a mathematical model for optimal dynamic investment policies in capacity and different cases of interest will be studied in detail. The quality attribute, nevertheless, has not received much attention and therefore models that study optimum policies with respect to it are almost non-existent. The development of such models and the analysis of the characteristics of the optimum policies derived from them will be our main objective in this chapter. Special emphasis will be given to the economic interpretation of the results obtained.

By quality of a facility we will mean those characteristics that are not related to its capcity, but which affect the utility and/or cost, and therefore the benefit, derived from its use. In the case of a road the characteristics will be: the roughness of the surface, the radius of the curves, the grade of the road, etc. In general, quality will have to be represented by an index that adequately represents the characteristics of interest. In the case of a road such an index is represented by the present serviceability index (PSI) defined by AASHO [1962] or the virtual length of the road (Miquel. S, [1972]).

We will assume in this chapter that the quality of a facility can be represented by a continuous variable through all the period

of analysis [0,T]. The case in which discontinuities can appear at certain interior times in [0,T] will be analyzed in Chapter V.

2. A Mathematical Model for Optimal Investments in Quality. Case of External Demand

Different factors affect the quality of a facility over the course of its economic life. Natural factors and normal use are the principal causes of quality deterioration. On the other hand, this deterioration can be alleviated or remedied if maintenance or repair is undertaken. Finally, quality increases can be obtained if enough money is spent in improving those characteristics that determine the level of quality of the facility. In this chapter we will use the word "maintenance" in a generic way to refer to any of the activities that influence the quality of a facility in a continuous way.

We will assume that we have a transportation facility with a fixed capacity k which serves homogeneous users. Each of these users obtains a utility U(t) and perceives a cost C(s,q) each time that they he/she uses the facility (e.g. from each trip performed). We consider that the utility is determined by factors external to the model, though its value will in general be a function of time. This seems a realistic assumption for all kinds of trips, except for the pure recreational ones. In any case it only constitutes a convenient assumption allowing a more simplified analysis and can be dropped without major consequences. The operating cost function C(s,q) corresponds to an average variable cost function, which includes all expenses of user supplied inputs, and it is assumed to depend on the quality of

the facility s and the number of users q, having as a parameter the fixed value of capacity k. We will make the following assumptions with respect to the function C

$$C_s < 0,\ C_q \geqq 0$$

$$C_{sq} = 0,\ C_{qq} \geqq 0,\ C_{ss} \geqq 0. \qquad (2\text{-}1)$$

The conditions (2-1) are statements that additional quality will always decrease operating costs; additional traffic, holding capacity and quality constant, will increase operating costs because of congestion; no interrelations exist between quality and congestion. Congestion only depends on the relative values of level of traffic and capacity. Finally, C is a convex function of traffic and quality.

Therefore, the net private benefit obtained by an individual user as a consequence of using the facility can be represented at any time t by:

$$B(t) = U(t) - C\ (s(t),q(t)) \qquad (2\text{-}2)$$

We will let V(t) represent the amount of money spent on maintenance at time t, which will be our control variable.

Our focus will be on the determination of optimum maintenance policies from a public or social point of view. Therefore, our objective will be to maximize the present value of the private benefits minus the public costs of maintenance, through the life of the facility. Thus, our objective function can be represented as:

$$J(V(t)) = \int_0^T [B(t)q(t) - V(t)]\exp(-\rho t)dt, \qquad (2\text{-}3)$$

where T is the economic life and we assume that there is no residual economic value for the facility at time T. At that time a discrete predetermined upgrading of the facility will be performed or it will be destroyed and a completely new facility put in place. Therefore, no salvage value is associated with time T.

Let the change in quality of the facility be represented by the following differential equation:

$$\dot{s}(t) = f\,(s(t),\, q(t),\, V(t), t) \qquad (2\text{-}4)$$
$$s(o) = s_o$$

where s(t) is an index representing the qualtiy of the facility at time t, q(t) is the number of users at time t and V(t) is the amount of money spent on maintenance at time t. Note that $\dot{s}(t)$ represents the first derivative of s with respect to the independent variable time. Expression (2-4) means that the change of quality of the facility, per unit of time, depends upon the level of quality, the number of users and the amount spent in maintenance, at the time t considered.

In addition we will assume that the amount of money that can be spent in maintenance at each time t is constrained by:

$$m(t) \leqq V(t) \leqq M(t), \quad \forall\ t \in [0,T], \qquad (2\text{-}5)$$

2.1 Necessary Conditions for Optimality

If we consider that the demand for using the facility $q(t)$ is given, for each time t in the period $[0,T]$, then the maximization of (2-3) subject to (2-4) and (2-5) will determine the optimum maintenance policy. The problem is an optimal control problem of the type discussed in Section 3.1 of Chapter II, where the only state variable is the quality of the facility $s(t)$ and the only control is the amount spent in maintenance $V(t)$. The Hamiltonian is in this case equal to:

$$H(t) = \{[U(t) - C(s(t),q(t))]q(t) - V(t)\}\exp(-\rho t) + \lambda(t)f(s(t),q(t), V(t),t), \qquad (2\text{-}6)$$

where $\lambda(t)$ is the adjoint variable that must satisfy the adjoint equation:

$$\dot{\lambda}(t) = -(\partial H/\partial s) = (\partial C/\partial s)\, q\exp(-\rho t) - \lambda(\partial f/\partial s) \qquad (2\text{-}7)$$

where for simplicity we have eliminated the arguments of all the variables.

Given that the "penalty function" at time T has a null value (salvage value equal zero) the transversality condition for λ at time T will be

$$\lambda(T) = 0, \qquad (2\text{-}8)$$

The necessary conditions for a maximum of $J(V(t))$ over $V(t)$ state that there must exist a function $\lambda(t)$ that satisfies the adjoint equation (2-7) and the transversality condition (2-8) and that the optimum control $V^*(t)$ must be such that the Hamiltonian (2-6) is maximized for all t in [0,T], i.e.

$$H(s^*,\lambda^*,V^*,t) \geqq H(s^*,\lambda^*,V,t), \quad \forall\ t \in [0,T], \qquad (2\text{-}9)$$
$$(V \in \Omega)$$

where in our case we have from (2-5)

$$\Omega = \{V(t) : m(t) \leqq V(t) \leqq M(t), \quad \forall\ t \in [0,T]\}, \qquad (2\text{-}10)$$

and the * in s and λ means that these variables satisfy (2-4) and (2-7,8) respectively. Expressions (2-4), (2-5), (2-7), (2-8) and (2-9) constitute a complete set of necessary conditions for our problem.

It is easy to see that the gradient of the Hamiltonian with respect to the control variable V is given by

$$H_V = \lambda f_V - \exp(-\rho t), \qquad (2\text{-}11)$$

and in order to characterize the value V^* that satisfies (2-9) we can use the following theorem from non-linear programming (see Luenberger [1973]): If V^* is a local maximum of the function H over the convex set Ω, defined by (2-10), then:

$$H_V(V^*) \quad (V-V^*) \leqq 0, \quad \forall\ V \in \Omega \quad . \tag{2-12}$$

If in addition, the Hamiltonian is concave in V over the constraint set Ω, the control V^* defined by (2-12) will be a global maximum of H. Given that Ω, defined in (2-10), is an unidimensional convex closed set it is easy to see that (2-12) is equivalent to the conditions

$$H_V(V^*) = 0, \qquad \text{if} \quad m < V^* < M$$

$$H_V(V^*) \leqq 0, \qquad \text{if} \quad V^* = m$$

$$H_V(V^*) \geqq 0, \qquad \text{if} \quad V^* = M$$

which, using the gradient definition (2-11), can be written as

$$\lambda f_V = \exp(-\rho t), \qquad \text{if} \quad m < V^* < M \tag{2-13}$$

$$\lambda f_V \leqq \exp(-\rho t), \qquad \text{if} \quad V^* = m \tag{2-14}$$

$$\lambda f_V \geqq \exp(-\rho t), \qquad \text{if} \quad V^* = M \quad . \tag{2-15}$$

These three relations give an expression for the optimum control V^* in all possible cases to be found.

2.2 Economic Interpretation of the Necessary Conditions

In this section we will provide economic interpretations of the necessary conditions for optimality presented in Section 2.1. A fundamental element of these conditions is the adjoint variable λ

that appears in the expressions for the optimal control V^* and is defined by the adjoint equation (2-7) and the transversality condition (2-8).

Expression (2-7) corresponds to a first order ordinary linear differential equation whose solution can be written as:

$$\lambda(t) = \exp(-\int^t f_s dz)\{-\int^T C_s q \exp(-\rho x)\exp(\int^x f_s dz)dx + \lambda(T)\exp(\int^T f_s dz)\} \qquad (2\text{-}16)$$

where we use the following notation

$$C_s = (\partial C/\partial s), \text{and } f_s = (\partial f/\partial s) \quad .$$

Upon rearranging (2-16) and using (2-8) we can write $\lambda(t)$ as:

$$\lambda(t) = -\int_t^T \{[C_s \exp(\int_t^x f_s dz)]q(x)\exp(-\rho x)\}dx \qquad (2\text{-}17)$$

According to (2-1) C_s is always negative and given that $q(t)$ cannot be negative it is obvious that the integrand in (2-17) will be always negative. Therefore $\lambda(t)$ will be positive for all t in $[0,T]$ and its value will increase as T increases. This fact will have important consequences for the optimal policy V^*. To obtain an economic interpretation of $\lambda(t)$ we must give an interpretation to each element in the integrand of expression (2-17).

To begin we can note that f_s represents the rate of deterioration*

(*) Here we use "deterioration" in a general way. It can mean "improvement" if an increase in quality, reduces the rate of deterioration of the facility.

per unit of time, per unit of quality.

$$f_s = (\partial \dot{s}/\partial s) \ .$$

If we change s(t) by one unit, at time t, the rate of deterioration will change by f_s.

Proposition 2.2.1 If x and t are two different times such that $x \in [0,T]$, $t \in [0,T]$ and $x > t$, and f_s is a continuous function of time, then

$$g\ (x) = \exp(\int_t^x f_s(z)dz)$$

is the "equivalent value" (or residual value), at time x, of one unit of quality implemented at time t.

In order to demonstrate this proposition let us consider that the interval (t,x) is divided in n finite time differences Δt

$$\Delta t = \frac{x-t}{n}$$

Then, if we implement one unit of quality in the facility at time t, it is obvious that after Δt its value will become

$$g\ (t + \Delta t) = 1 + f_s(t)\ \Delta t$$

given that as we saw above f_s represents the rate of deterioration per unit of time, per unit of quality. Similarly after 2 Δt the

value of the unit of quality will transform to

$$g(t + 2\Delta t) = g(t + \Delta t)(1 + f_s(t + \Delta t)\Delta t)$$
$$= (1 + f_s(t)\Delta t)(1 + f_s(t + \Delta t)\Delta t)$$

Following the same recursive procedure it is easy to find that after $n\Delta t$ we can write

$$g(x) = \prod_{i=0}^{n-1} [1 + f_s(t + i\Delta t)\Delta t]$$

If we now take ℓn on both sides of this expression we obtain

$$\ell n\ g(x) = \sum_{i=0}^{n-1} \{\ell n[1 + f_s(z_i)\Delta z)]^{(1/\Delta z)}\}\Delta z, \qquad (*)$$

where we have used the change of variable

$$z_i = t + i\Delta t, \quad \text{with } \Delta z = z_i - z_{i-1} = \Delta t$$

and we have also multiplied and divided each term of the sum by the finite difference Δz.

Now we can take the limit in (*) when Δt goes to zero or n goes to infinity. Note that the equal sign in expression (*) is only strictly correct when we take this limit and therefore transform t to a continuous variable. Before that, the right hand side of (*) constitutes only a discrete approximation to the value of $\ell n\ g(x)$.

Using the following results from basic calculus:

- the limit of a sum is equal to the sum of the limits
- the limit of a product is equal to the product of the limits
- the limit of ℓn is equal to the ℓn of the limit

and the well-known limit

$$\lim_{h \to 0} (1 + ah)^{(1/h)} = \exp(a),$$

we obtain from (*) that

$$\ell n\ g(x) = \lim_{\substack{\Delta z \to 0 \\ n \to \infty}} \sum_{i=0}^{n-1} f_s(z_i)\Delta z$$

and applying the definition of the definite integral we obtain

$$\ell n\ g(x) = \int_t^x f_s(z)dz$$

from where we easily get the desired result

$$g(x) = \exp\left(\int_t^x f_s(z)dz\right)$$

In order to further clarify this concept we will analyze two examples.

Example 1: $\dot{s} = -\alpha(t) + f_1(q,V,t)$

In this case the facility suffers a deterioration of $\alpha(t)$ units of quality per unit of time, as a consequence of natural factors, and a change of f_1 units of quality, as a result of use and maintenance activities. The rate of deterioration $\dot{s}$ is independent of the level of quality. Clearly we will have

$$f_s = 0, \quad \text{and} \quad \exp\left(\int_t^x f_s dz\right) = 1 .$$

Given that a change of quality does not have any effect on the rate of deterioration, the equivalent value, at any time x (x > t), of one additional unit of quality implemented at time t will be always one.

Example 2: $\dot{s} = -\alpha s(t) + f_1(q,V,t)$

Here we have changed the characteristics of the deterioration produced by natural factors. The natural deterioration per unit of time is now equal to a constant percentage of the level of quality in the facility at that time. Thus

$$f_s = -\alpha, \quad \text{and} \quad \exp\left(\int_t^x f_s d\right) = \exp[-\alpha(x-t)]$$

Here, one additional unit of quality at time t will increase the rate of deterioration of the facility by α units per unit of time and therefore the equivalent value, at time x, of one additional unit of quality implemented at time t will be lower than one and equal $\exp[-\alpha(x-t)]$.

Then, if we return to (2-17) we have that

$$C_s \exp(\int_t^x f_s dz)$$

represents the reduction in operating costs experienced by each user, at time x, as a consequence of the implementation of one additional unit of quality at time t ($t < x$) and therefore, $\lambda(t)$ is the present value, at time $t = 0$, of the cost reductions experienced by all the users of the facility, during the period [t,T] as a consequence of the implementation of one additional unit of quality at time t. Or in more general terms, we can say that $\lambda(t)$ represents the present value at time $t = 0$, of the total benefits perceived during the period [t,T] as a consequence of an improvement of one unit of quality in the facility at time t. The total benefits are calculated without considering the cost of implementing the additional unit of quality at t.

Now from (2-6) we have that the expression of the Hamiltonian will be:

$$H(t) = [(U - C)\, q - V]\exp(-\rho t) + \lambda f, \qquad (2\text{-}6)$$

where for simplicity we have elminiated the arguments of all the variables. The first term on the right hand side of (2-6) represents the present value, at time $t = 0$, of the social benefits produced by the facility at time t. It is a result of the sum of the private benefits minus the social costs of maintenance at that time. The second term is the product of the change of quality experienced by the facility at time t (as a consequence of the level of quality, the use and the maintenance experienced by the facility at t) and the

present value associated with a change of one unit of quality at that time.

Therefore, the Hamiltonian gives us the present value of the net social benefits associated with the decisions taken with respect to the operation of the facility at time t. If for instance we decide not to do maintenance at time t, i.e. V(t) = 0, we will have

$$H(t) = (U - C)q \exp(-\rho t) + \lambda f(V=0)$$

where $f(V=0)$ represents the deterioration experienced by the facility at time t as a consequence of natural factors and the use of the facility at that time. This deterioration (produced at time t) will cause an additional cost $\lambda f(V=0)$ in the operation of the facility during period [t,T]. If on the other hand we decide to invest in maintenance an amount $V(t) = V_c$ such that no deterioration is experienced at time t, $f(t) = 0$, and we will have

$$H(t) = [(U - C)q - V_c]\exp(-\rho t)$$

From (2-13) we have that interior optimal values of V in Ω will satisfy the necessary condition

$$\lambda f_V = \exp(-\rho t), \qquad m < V^* < M$$

while corner optimum solutions will satisfy (from 2-14 and 2-15)

$$\lambda f_V \leqq \exp(-\rho t), \qquad V^* = m$$

$$\lambda f_V \geqq \exp(-\rho t), \qquad V^* = M$$

where f_V is the reduction of deterioration, or the quality improvement, produced at time t, by the investment of one additonal dollar in maintenance at that time and λf_V is the present value, at t = 0, of the total benefits associated with an improvement of f_V units of quality in the facility at time t. The quantity $\exp(-\rho t)$ represents the present value, at t = 0, of one dollar invested in maintenance at time t. Therefore, the optimal policy says that maintenance should be performed at a level such that the present value of the marginal benefits of maintenance are equal to the present value of the corresponding marginal costs. If the marginal benefits are higher than the marginal cost, for all values of V in Ω, the maximum possible level of maintenance should be provided at t. On the otner hand, if the marginal benefits are lower than the marginal cost, for all vaues of V in Ω, the minimum possible level of maintenance should be provided at t. This policy will lead to a global maximization of the Hamiltonian only if it is concave in V over Ω. In other words, if the second derivative H_{VV} is non-positive. Differentiating (2-11) with respect to V we obtain $H_{VV} = \lambda f_{VV}$ and given that λ is positive for all t in [0,T], the Hamiltonian will be concave if f is concave in V or, there exists constant or decreasing returns to scale in the production of quality through maintenance.

For the case of interior solutions, expression (2-13) gives us only an implicit expression for V^*. In order to obtain an explicit

expression we would need to specify the form of the deterioration function f. Nevertheless, if condition (2-13) holds for some finite time interval within [0,T], then we can obtain an explicit expression for $\dot{V}^*$ which will give us the dynamic characteristics of such optimum policy. In particular if (2-13) holds for a finite period of time, within this period, we will have

$$\dot{H}_V = \dot{\lambda} f_V + \lambda f_{VV} \dot{V} + \rho \exp(-\rho t) = 0, \tag{2-18}$$

where we have used the expression of the gradient of the Hamiltonian from (2-11) and we have assumed that $f_{Vs} = f_{Vq} = f_{Vt} = 0$. In other words, the marginal effectiveness of maintenance f_V is only a function of the amount spent in maintenance at each time t. If we now use the adjoint equation (2-7) and the necessary condition (2-13) to eliminate $\dot{\lambda}$ and λ respectively from (2-18) and then divide the resulting expression by the positive value $\exp(-\rho t)$, we obtain

$$C_s q f_V - f_s + \rho + (f_{VV}/f_V)\dot{V} = 0, \tag{2-19}$$

and then, if $f_{VV} < 0$, or decreasing returns to scale exist in the production of quality through maintenance, we can divide (2-19) by f_{VV} and obtain

$$\dot{V}^* = (f_V/f_{VV})(f_s - \rho - C_s q f_V) \tag{2-20}$$

Therefore we will have

$$\dot{V}^* > 0, \quad \text{if} \quad -C_s q f_V < \rho - f_s, \; f_V > 0$$
$$\dot{V}^* < 0, \quad \text{if} \quad -C_s q f_V > \rho - f_s, \; f_V > 0$$
$$\dot{V}^* = 0, \quad \text{if} \quad -C_s q f_V = \rho - f_s, \text{ or } f_V = 0$$

where $-C_s q f_V$ is the value of the operating costs reductions experienced by all the users of the facility at time t, as a consequence of one dollar spent in maintenance at that time and $(\rho - f_s)$ is what is usually called "effective discount rate" (see Arora and Lele, [1970]). A stationary solution with V^* constant will only be obtained if

$$q(t) = -(\rho - f_s)/C_s f_V.$$

In particular, if f_s, C_s and f_V are not explicit functions of time, q(t) must be constant.

It is interesting to remember here that from (2-17) we concluded that $\lambda(t)$ will in general increase when T increases and vice-versa. This fact together with any of the equations (2-13), (2-14), or (2-15) implies that for an optimum maintenance policy, the amount spent in maintenance will increase if we decide to use the same facility for a longer period of time and vice-versa. This to some extent implies a trade-off between maintenance and replacement policies.

Finally, it is important to note that the case presented in this section, in which we consider the demand q(t) as externally specified and fixed for the purposes of the analysis, is not a mere simplification, but actually corresponds to some real situations. A typical example would be that of a mine that uses a private transportation facility

in order to take its production out of the extraction site. In that case the demand q(t) will be determined by the production plan of the mine and therefore will be external to any model of management of the facility.

3. A Mathematical Model for Optimum Investments in Quality. Internal Demand Case.

In the preceeding section we considered a model in which the demand was externally specified and not affected by the parameters of the model. Therefore, the benefits associated with the operation of the facility were affected by the maintenance policy only through the supply side of the problem. Actually, this influence occurred through changes in the operating costs experienced by the given users and the investment costs in maintenance necessary to produce this change. In this section we will consider a model formulation in which the maintenance policy will affect both the supply and demand sides of the problem. For this, we will assume that the number of users of the facility at each time t is a function of the history of the quality of the facility during the period [0,T], $t < T$.

Let the change in demand for the facility be represented by the following differential equation

$$\dot{q} = \ell\,(s(t),t), \qquad q(0) = q_0 \,. \tag{3-1}$$

Expression (3-1) means that the demand for using the facility will change, per unit of time, as a function of the quality of the

quality of the facility and some external factors that can be expressed explicitly as functions of time. The demand for the facility will therefore no longer be an external condition, but will be internally determined by the model. Now we will have two state variables: quality of the facility and number of users; the same control as before will be utilized namely the mainenance expenditures.

3.1 Necessary Conditions for Optimality

Now an optimum maintenance policy will be obtained by the maximization of $J(V(t))$ subject to (2-4), (2-5) and the new condition (3-1). The Hamiltonian becomes in this case

$$H(t) = \{[U(t) - C(s(t),q(t))]q(t) - V(t)\}\exp(-\rho t) + \lambda(t)f(s(t),q(t), V(t),t) + \nu(t)\ell(s(t),t), \quad (3\text{-}2)$$

where $\lambda(t)$ and $\nu(t)$ are adjoint variables that must satisfy the adjoint equations

$$\dot{\lambda}(t) = -(\partial H/\partial s) = C_s q\exp(-\rho t) - \lambda f_s - \nu \ell_s \quad (3\text{-}3)$$

$$\dot{\nu}(t) = -(\partial H/\partial q) = [-(U - C) + C_q q]\exp(-\rho t) - \lambda f_q . \quad (3\text{-}4)$$

Again, given that no economic value is attached to the quality of the facility at time T the transversality condition for λ at T will be

$$\lambda(T) = 0 \quad (3\text{-}5)$$

Nevertheless, in this case, some value may be assigned to the size of the demand at time T. The rationale for this is that if the facility considered is going to be replaced, at time T, by a new facility, which will perform basically the same function as the old one, the benefits derived from the use of the new facility at T will depend on q(T). Let us assume that the value attached to this demand at time T is represented by a function $\Psi(q(T),T)$ that will be added to the formulation of our objective function (2-3). Then, the transversality condition for ν at time T will be given by

$$\nu(T) = (\partial\Psi/\partial q)_T = \nu_T \ . \qquad (3\text{-}6)$$

The necessary conditions for a maximum of J(V) over Ω state that there must exist functions $\lambda(t)$ and $\nu(t)$ that satisfy the adjoint equations (3-3), (3-4) and the transversality conditions (3-5), (3-6); and the value of the optimum control $V^*(t)$ must be chosen such that the Hamiltonian (3-2) is maximized for all t in [0,T]:

$$H(s^*,q^*,\lambda^*,\nu^*,V^*,t) \geq H(s^*,q^*,\lambda^*,\nu^*,V,t), \quad \forall\ t \in [0,T] \quad (V \in \Omega) \qquad (3\text{-}7)$$

where Ω is defined in (2-10) and the superscript "*" over the state and adjoint variables means that they satisfy the corresponding state and adjoint equations. The expressions (2-4), (2-5), (3-1), (3-3) to (3-6) and (3-7) constitute in this case a complete set of necessary conditions.

Given that the traffic generation function ℓ defined by (3-1) is not a function of the control V, the gradient of the Hamiltonian with respect to this variable will have the same expression as in the case of external demand (see, 2-11):

$$H_V = \lambda f_V - \exp(-\rho t) \ . \qquad (3\text{-}8)$$

Therefore, given that the constraint set Ω over which the Hamiltonian must be maximized is also unchanged, the same expressions will be also obtained for the optimality conditions specifying the control V. Interior solutions in Ω will be given by:

$$\lambda f_V = \exp(-\rho t), \qquad m < V^* < M \qquad (3\text{-}9)$$

and in the case of corner solutions we will have

$$\lambda f_V \leqq \exp(-\rho t), \qquad \text{if} \quad V^* = m \qquad (3\text{-}10)$$

$$\lambda f_V \geqq \exp(-\rho t), \qquad \text{if} \quad V^* = M \qquad (3\text{-}11)$$

Nevertheless, it is important to note that the adjoint variable λ is here different from that associated with the case of external demand (compare expression (2-7) and (3-3).

3.2 Economic Interpretation of the Necessary Conditions

As we saw in Section 2.2, for the case of external demand, the

economic interpretation of the necessary conditions relies heavily on the meaning of the adjoint variables. The main difference between the necessary conditions for the cases of external and internal demand lies in the expressions of the adjoint equations. Therefore, we would expect important changes in the economic interpretation and role of the adjoint variables. We consider C_s, C_q, f_s, f_q and ℓ_s to be expressed as functions of time, and rewrite (3-3) and (3-4) as

$$\dot{\lambda}(t) = - \lambda f_s + [C_s q \exp(-\rho t) - \nu \ell_s] \qquad (3\text{-}12)$$

$$\dot{\nu}(t) = [(C + C_q q) - U]\exp(-\rho t) - \lambda f_q \ . \qquad (3\text{-}13)$$

We then observe that expression (3-12) corresponds to a first order ordinary linear differential equation whose solution can be written as

$$\lambda(t) = \exp(-\int^t f_s dz) \left\{ - \int_t^T C_s q \exp(-\rho x) \exp(\int^x f_s dz(dx) + \right.$$

$$\left. \int_t^T \nu \ell_s \exp(\int^x f_s dz)\, dx + \lambda(T)\exp(\int^T f_s dz) \right\}$$

which, upon rearranging some terms and using (3-5), can be expressed as

$$\lambda(t) = - \int_t^T \{[C_s \exp(\int_t^x f_s dz)]\exp(-\rho x)\}dx +$$

$$\int_t^T \nu(x)\ell_s(x)\exp(\int_t^x f_s dz)dx \ . \qquad (3\text{-}14)$$

Expression (3-13) can be integrated directly, using (3-6) to obtain

$$\nu(t) = \int_t^T [U(x) - C(s(x),q(x))]\exp(-\rho x)dx$$

$$+ \int_t^T - C_q q(x)\exp(-\rho x)dx + \int_t^T \lambda(x) f_q(x)dx$$

$$+ \nu(T), \qquad (3\text{-}15)$$

The interpretation of λ and ν is complicated in this case by the fact that they are interdependent. We cannot give an interpretation for λ without knowing the meaning of ν and vice versa. We must therefore give a simultaneous coherent interpretation for both. A way to proceed is the following: we obtained in Section 2.2 an interpretation of λ for the case of external demand. We can begin assuming then that the same basic general interpretation holds here and proceed to use it in order to obtain an interpretation for ν in (3-15). Then we can go back to (3-14) and check that our initial interpretation of λ is still correct.

Recall that our general interpretation for λ in Section 2.2 says that it represents "the present value at time t = 0 of the total benefits perceived during the period [t,T] as a consequence of an improvement of one unit of quality in the facility at time t". Now we can proceed to the interpretation of (3-15). Note that $C_q q(x)$ is the cost of the congestion externalities produced by the introduction of one additional user to the facility at time x ($t \leq x \leq T$). It is equal to the difference between average and marginal operating costs associated with the facility at time x. Note that $f_q(x)$ represents

the change in quality (or deterioration) produced by an additional user at time x. We will in general assume $f_q < 0$. Therefore, $\lambda f_q(x)$ will be the present value, at t = 0 of the loss in benefit (or cost) associated with the deterioration of facility quality produced by an additional user generated at time x. Note that this term includes the effect that the deterioration produced by the additional user at time x has on all the users of the facility over the period [x,T].

The first term of expression (3-15) corresponds to the present value of the private net benefit perceived by a user of the facility during the period [t,T]. The second term is the present value of the additional costs incurred by all the users of the facility during the period [t,T] as a consequence of the additional congestion produced by the introduction of an additional user during that period. The third term is the present value, at time t = 0, of the total social cost associated with the deterioration produced during the period [t,T] by the introduction of an additional user at time t. The second and third terms are therefore the externalities, associated with congestion and deterioration respectively, produced by an additional user during the period [t,T]. Finally, from (3-6) we have that the last term is the value associated with an additional user at time T.

Therefore, we can say that $\nu(t)$ represents the present value, at time t = 0, of the social net benefit produced during the period [t,T] by the introduction of an additional user at time t. In other words it is the shadow price of demand at time t.

Given that $\nu(t)$ is the shadow price of demand, the sum of the integrands of the second and third terms at any time x gives us the present value of the amount that, according to an optimum pricing policy, each user of the facility should be charged at time x. If we think for instance of a road, the term $\lambda f_q(x)$ can be very important for heavy weight trucks that produce considerable deterioration of the infrastructure. We will show later that the model presented here can be easily disaggregated to consider different types of users.

Now, we can go back to check the interpretation of $\lambda(t)$ in (3-14). The first term there is exactly the same obtained in (2-17), the interpretation of which is "the present value, at time t = 0, of the cost reduction experienced by all the users of the facility, during the period [t,T], as a consequence of one additional unit of quality at time t". In addition we have that

$\exp(\int_t^x f_s dz)$ is (see Section 2.2) the equivalent value, at time x, of one additional unit of quality implemented at time t

$\ell_s(x)$ is the number of additional users generated at time x by a change of one unit of quality in the facility at that time. We will in general assume $\ell_s > 0$

$\ell_s(x)\exp(\int_t^x f_s dz)$ is then the number of additional users generated at time x by an improvement of an additional unit of quality in the facility at time t

Note also that the second term of (3-14) is then the present value, at time t = 0, of the social net benefit associated with the new traffic generated during the period [t,T] as a consequence of the implementation of an additional unit of facility quality at time t.

The general interpretation of $\lambda(t)$ is then exactly that with which we began and that we repeat here: "$\lambda(t)$ represents the present value, at time t = 0, of the total benefits perceived during the period [t,T] as a consequence of an improvement of one unit of quality in the facility at time t". The difference between this case (internal demand) and that of Section 2 (external demand) is that now $\lambda(t)$ also includes benefits (or costs) produced as a consequence of new traffic generated by the quality improvement at time t. These benefits (or costs) did not appear in Section 2 because in that case the demand was externally given and was not influenced by changes in quality of the facility.

Therefore, the adjoint variables λ and ν correspond exactly to "shadow prices" associated with unitary values of the corresponding state variables s and q. They are actually dynamic shadow prices since they are functions of time. Consequently, it is logical to expect them to serve as indicators for the implementation of optimum investment (in the case of λ) and pricing (in the case of ν) policies.

According to expression (3-14) for the present case of internal demand, the value of $\lambda(t)$ will be affected by the values that $\nu(x)$ takes in the period [t,T]. As we saw in Section 2 the first term on the right hand side of expression (3-14) is always positive. The value of the second term could be negative if $\nu(x)$ takes negative

values in [t,T]. If that is the case, and assuming the same q(x), $x \in [t,T]$, the value of $\lambda(t)$ would be lower than that corresponding to the case in which demand is assumed external. The explanation for this is that, given that quality influences demand and given a case in which new users produce big externalities with absolute values in excess to those of their private benefits, the new users generated by an improvement of quality will make less attractive the reduction in operating costs C_s produced by the same quality improvement. If the second term in (3-14) is positive, the value of $\lambda(t)$ will be higher here than in the case of external demand, because of the positive social benefits attached to the new users generated by an improvement in quality.

With respect to $\nu(t)$ we have four terms in the right hand side of expression (3-15). We will assume that the first term, representing the net private benefits experienced by each individual user of the facility, will be always positive, because if at some time x the operating cost C(s,q) is higher than the utility U(x) an individual will not use the facility. The second and third terms representing the externalities produced by each user will obviously be negative. The last term requires some further analysis.

If we assume that the value given to the amount of demand at time T, q(T), is equal to the benefits attached to the use of the new facility (by these q(T) users) for an infinite period of time, beginning at T, we have that

$$\Psi\ (q(T),T) = \int_T^{\infty}[U(x) - C'(s(x),q(x))]q(T)\exp(-\rho x)dx$$

where $C'(s,q)$ is the function that gives the operating costs of the new facility. By using (3-6) we also obtain for these circumstances

$$\nu(T) = \int_T^{\infty}[U - C') - C'_q q(T)]\exp(-\rho x)dx \qquad ,(3\text{-}16)$$

where C' is again equal to the operating cost experienced by each user of the new facility and $C'_q q(T)$ is the congestion externality produced over $q(T)$ users of the new facility.

Therefore, the higher the quality and the capacity with which the new facility will be provided the lower will be C' and $C'_q q$ and the higher will be the value of $\nu(T)$ and consequently of $\nu(t)$ and $\lambda(t)$. From (3-15) we can see that $\nu(t)$ could become negative if the externalities represented by the second and third terms are high enough. In general, an increase in the value of these externalities will decrease the values of both ν and λ.

The changes in the values of $\nu(t)$ and $\lambda(t)$ when the life of the facility is varied will be given by the derivatives of these variables with respect to T. Differentiating (3-14) and (3-15) with respect to T and using the boundary conditions (3-6) and (3-16) we can easily obtain:

$$(\partial\lambda(t)/\partial T) = -[C_s \exp(\int_t^T f_s dz)]q(T)\exp(-\rho T) + \nu(T)\ell_s(T)\exp(\int_t^T f_s dz) + \int_t^T (\partial\nu(x)/\partial T)\ell_s(x)\exp(\int_t^x f_s dz)dx \qquad (3\text{-}17)$$

and

$$\partial\nu(t)/\partial T) = [(C' + C'_q q(T)) - (C + C_q q(T))]\exp(-\sigma T) + \int_t^T (\partial\lambda(x)/\partial T) f_q(x)dx \qquad (3\text{-}18)$$

In Section 2, when demand was considered external to the model, we easily concluded that as the life of the facility is increased, the shadow price of quality $\lambda(t)$ increased for any t in [0,T]. Here it is impossible to reach a definitive conclusion like that. This is due to the dynamic interactions between $\nu(t)$ and $\lambda(t)$ manifested by the last terms of expressions (3-17) and (3-18). If we assume $\nu(T)$ positive then the first two terms of (3-17) will be positive, but to know the sign of the third term we need to know $\partial\nu(x)/\partial T$ over the interval [t,T]. Also if we assume that the new facility to be provided at time T constitutes a general improvement over the old one (higher quality and more capacity) $C'(s'(T),q(T)) < C(s(T),q(T))$ and $C'_q < C_q$, then the first term of expression (3-18) which represents the difference between the marginal costs on the new and old facilities at time T, will be clearly negative. However, to know the sign of the second term, we need to know the value of $\partial\lambda(x)/\partial T$ over the interval [t,T].

Under the reasonable assumption that the new facility to be provided at time T constitutes a general improvement over the old one, the value of $\partial v(t)\partial T$ cannot be positive for any important interval of time tin [t,T]. Given that $f_q(x)$ is negative, such a situation would require that $\partial\lambda(x)/\partial T$ be negative over an important period of time in [t,T]. This is highly improbable since according to equation (3-17), in order for $\partial\lambda(x)/\partial T$ to be negative over an important period of time, it is necessary that $\partial v(x)/\partial T$ be negative over an important period of time in [t,T], which is a contradiction. Therefore, we would expect $\partial v(t)\partial T$ to be negative in general, making the sign of $\partial\lambda(t)/\partial T$ uncertain and highly dependent on the magnitude of each of the terms involved in equation (3-17).

From (3-2) we have that the Hamiltonian in the present case is given by

$$H(t) = [U - C)q - V]\exp(-\rho t) + \lambda f + v\mathfrak{L}$$

where for simplicity we have eliminated the arguments of all the variables. It is easy to see that H(t) has here the same general interpretation as in Section 2.2. It is the present value of the net social benefits associated with the decisions taken with respect to operation of the facility at time t.

Given that the general interpretation of $\lambda(t)$ is the same here as in Section 2.2 and that the optimality conditions for V have also the same functional form, the interpretation of the optimum policy is again that maintenance should be performed at a level

such that the present value of the marginal benefits of maintenance be equal to the present value of the corresponding marginal cost at each time t. If this is not possible for any value of $V \in \Omega$, then the optimum policy is to perform the maximum or the minimum maintenance possible depending on whether the marginal benefits are higher or lower than the respective marginal costs. Nevertheless, the marginal benefit to which we refer here is different than that in the case of Section 2.2. It is evaluated through the use of $\lambda(t)$ defined by expression (3-14) instead of (2-17). Moreover, it is obvious from our analysis of $\lambda(t)$ and $\nu(t)$ that then the optimum policies corresponding to expressions (3-9) to (3-11) will be different than those derived from the corresponding expressions in Section 2.1. As we saw before, in this case $\lambda(t)$ will depend on the value of $\nu(t)$ that represents the net social value of a new user at time t, a quantity for which an important role is played by the externalities generated.

Let us for instance assume the same demand $q(x)$ for the period $[t,T]$ as in the case of Section 2. Using this demand schedule we may simultaneously evaluate (3-14) and (3-15) in general obtaining a value of $\lambda(t)$ that leads to a different optimum policy at time t than obtained for the same situation in Section 2. If the assumed demand $q(x)$ is too low compared with the capacity of the facility and the users considered do not produce too much deterioration, $\nu(x)$ will be positive in the interval $[t,T]$ and the value of $\lambda(t)$ will be higher than that obtained in Section 2. Therefore, for the same deterioration function, more maintenance will be justified

in the present case in order to increase the quality of the facility and attract new users since those users have a positive social value ($\nu(t) > 0$). This means that the number of users assumed, $q(x)$, is lower than it should be in order to obtain the maximum possible benefit out of the operation of the facility. The converse would happen if the assumed demand $q(x)$ is too high with respect to the capacity of the facility and/or the users assumed produce too much deterioration. In order to obtain the same optimal policy for both cases it is necessary that $\nu(x) = 0$, $\forall\ x \in [t,T]$, which means that the schedule assumed which will be called $\bar{q}(x)$, is such that at all $x \in [t,T]$ the externalities produced by the marginal user are equal to the private benefits obtained by him plus the net social benefit of a new user at time T, $\nu(T)$. If this is the case, $\bar{q}(x)$ represents the demand schedule that generates the maximum possible amount of benefits out of the operation of the given facility (represented by the operating cost function $C(s,q)$ and the deterioration function f). This does not mean that this schedule would be the one corresponding to all the solutions obtained from the application of our model. The reason for this is that in a particular case the solution of the model is constrained by the following conditions not considered in our analysis above:

- Initial demand at $t = 0$, $q(0) = q_0$
- Demand dynamics, $\ell(s,t)$
- Control constraints, $V \in \Omega$.

There will be, in general, one value of q(o) (q(0) equal to $\bar{q}(x=0)$ and a lower bound $\underline{M}$ for M in Ω which allows one to reach $\bar{q}(x)$ for the facility considered. If $q(o) \neq \bar{q}(x=0)$ or the value of M in Ω is lower than $\underline{M}$, the model will do what it can to get a solution that generates a demand as close as possible to $\bar{q}(x)$ but it will never exactly reach that level for all $x \in [0,T]$. Note that if $\nu(t) \neq 0$ the model will try to optimize the maintenance policy taking into account its consequences after T. Therefore the optimum policy in such a case will be decidely different than if we considered that no facility will be provided after T ($\nu(T) = 0$). The solution obtained for $\nu(T) \neq 0$ will clearly be a sub-optimum if applied to the case $\nu(T) = 0$. In other words, the optimum policy for the case in which we assume that a new "better" facility will be provided after T will in general generate more users than what would be the optimum if no facility were to be provided after T. This is easy to check with the results of our previous analysis of $\lambda(t)$ and $\nu(t)$ derived from equations (3-14) and (3-15). There we saw that an increase in the value of $\nu(T)$ (an increase in the quality and capacity of the facility to be provided after T) will cause an increase of $\lambda(t)$, $\forall\ t \in [0,T]$ (this increase will be more important as $t \to T$) and therefore, according to expressions (39) to (42), an increase in the optimum amount of maintenance for all t.

Contrary to the case of Section 2.2., nothing definitive can be said here, in general, with respect to the influence of a change of T on the optimum maintenance policy. Since the analysis is complicated by the introduction of externalities and the existance of a

new facility after T, no general conclusion can be reached regarding the affect of changes in T. In particular, both positive and negative values of $\partial\lambda/\partial T$ appear possible, depending on the specific case analyzed.

It is interesting to note that the model presented in this section implicitly introduces a pricing policy defined by the adjoint variable $\nu(t)$, which as we saw in (3-14) influences the value of $\lambda(t)$ which represents the attractiveness of quality improvements of the facility and therefore determines the optimum maintenance policy $V(t)$. This is a consequence of the explicit consideration of interrelationships between quality and demand. The quality variable is used by the model to manipulate the cost function $C(s,q)$ in such a way that it implicitly includes an optimum pricing policy that ensures the optimum possible utilization of the facility during the period of analysis. The maintenance policy $V^*(t)$ is then the control variable that brings about a quality level at each time such that the relevant criterion function is extremized.

Finally, we make note that the same comments about the concavity of the Hamiltonian with respect to the control, made in Section 2.2, are valid here. In other words, the optimality conditions analyzed will correspond to a global maximum value of $H(t)$ if decreasing or constant returns to scale exist in the production of quality over the whole set Ω.

We may also note that for the internal demand case λ can be negative for some t. If that happens, the optimum solution will be $V^* = m$ because the only optimality relation that can hold then for V

is (3-10), given that f_V is always positive. This is an obvious result, given that λ is the marginal social value of quality. If this value is negative, at some point in time, we should try to produce the minimum possible increase of facility quality. Given that any amount spent in maintenance produces a positive change in quality, $f_V > 0$, it is obvious that we should spend the minimum amount possible in maintenance.

4. Extensions to the Case of Multiple User Types

In preceding sections we assumed a homogeneous type of facility user. Nevertheless, we usually have in real world cases that public facilities give service to different types of users simultaneously. In highways we have cars and trucks, in airports a great variety of aircraft types can be distinguished, etc. Each of these users experiences different utilities and operating costs and produce different congestion and deterioration externalities. In order to treat this more general case we have to modify the objective function used before in the following way:

$$J(V(t)) = \int_0^T \{\sum_{i=1}^{n} [U_i(t) - C_i(s,q)]q_i(t) - V(t)\}\exp(-\rho t)dt + \Psi(q_i(T),T), \qquad (4\text{-}1)$$

where U_i and C_i represent the utility and operating costs perceived by type i users and $q_i(t)$ is the number of type i users at time t.

The quantity n denotes the number of user types or categories. We in addition have defined:

$$q = \sum_{i=1}^{n} \mu_i q_i$$

with q equal to the total number of "equivalent users" over the facility at time t, and μ_i is the number of equivalent users that would produce the same congestion effect as one type i user. The equivalent user concept used here is completely analogous to that of "equivalent vehicle" used in highway transportation to represent congestion effects. It is also important to remember that we are here using the term "user" to identify an operating unit over the facility; therefore in the case of a road, our users will correspond to the vehicles operating over the road.

In this case, the number of state variables will increase to n + 1, with n of them corresponding to the demands for the n types of users that we are differentiating and one for the quality of the facilty as before. The evolution of these state variables will be governed by the following dynamics:

$$\dot{s} = f(s,\bar{q},V,t), \quad s(o) = s_o \tag{4-2}$$

$$\dot{q}_i = \ell_i(s,t), \quad q_i(o) = q_{io}, \ (i=1,\ldots,n), \tag{4-3}$$

$$\bar{q} = \sum_{i=1}^{n} \beta_i q_i$$

where β_i is the deterioration, in terms of quality units, produced by one user type i, per unit of time.

Finally, we have as before, the control constraint:

$$m(t) \leqq V(t) \leqq M(t), \tag{4-4}$$

The Hamiltonian corresponding to the control problem (4-1) to (4-4) can be written as:

$$H(t) = [\sum_{i=1}^{n} (U_i - C_i)q_i - V]\exp(-\rho t) + \lambda f + \sum_{i=1}^{n} \nu_i \ell_i, \tag{4-5}$$

and the adjoint equations become:

$$\dot{\lambda} = -\frac{\partial H}{\partial s} = \sum_{i=1}^{n} (\partial C_i/\partial s)q_i \exp(-\rho t) - \lambda f_s - \sum_{i=1}^{n} \nu_i \ell_{is}, \tag{4-6}$$

$$\dot{\nu}_k = -\frac{\partial H}{\partial q_k} = [-(U_k - C_k) + \mu_k(\sum_{i=1}^{n} C_{iq} q_i)]\exp(-\rho t) - \lambda \beta_k f_{\bar{q}}, \qquad (k = 1,\ldots,n) \tag{4-7}$$

In a similar way to that used in preceding sections we can find the following expressions for the adjoint variables λ and ν_k

$$\lambda(t) = -\int_t^T \{[\sum_{i=1}^{n} C_{is}\exp(\int_t^x f_s dz(q_i(x)]\exp(-\rho x)\}dx$$

$$+ \int_t^T [\sum_{i=1}^{n} \nu_i(x)\ell_{is}(x)\exp(\int_t^x f_s dz)]dx, \qquad (4\text{-}8)$$

$$\nu_k(t) = \int_t^T [U_k(x) - C_k(x)]\exp(-\rho x)dx +$$

$$- \int_t^T \{[\mu_k(\sum_{i=1}^{n} C_{iq}q_i(x))]\exp(-\rho x)\}dx$$

$$+ \int_t^T \lambda(x)\beta_k f_q dx + \nu_k(T). \qquad (4\text{-}9)$$

In this case ν_k is the social value of an additional type k user generated at time t. The first term in (4-9) represents the present value of the private benefit perceived by each type k user during the period [t,T]. The second term is the present value of the congestion externalities produced by a user type k over all the users of the facility during the same period. The third term represents the present value of the deterioration externalities produced by one type k user during the period [t,T]. This deterioration externality is proportional to the value β_k and includes, through the multiplication by the value of λ, the cost increases experienced by all the users of the facility during the period [t,T] as a consequence of the deterioration produced by one type k user during the same period. The sum of the integrands of the second and third terms is the amount that should be charged to each type k user for using the facility at time x if an optimum pricing policy were implemented. The interpretation of $\lambda(t)$ is here the same as in preceding sections. The first

term corresponds to the operating cost reductions experienced by all the users of the facility during the period [t,T] as a consequence of the implementation of an additional unit of facility quality at time t. The second term represents the social value of all the traffic generated during the period [t,T] as a consequence of one unit of quality implemented at time t.

As before, optimum maintenance policies corresponding to interior values of V(t) in $\Omega(t)$ will be given by:

$$\exp(-\rho t) = \lambda(t) \frac{\partial f}{\partial V} . \qquad (4\text{-}10)$$

An interesting interpretation of this optimum policy can be given here in terms of welfare economics. For the system defined by the facility of interest and its users; the quality of the facility is a public good. Any user of the facility has the same potential consumption of each unit of quality provided. The production function for the public good "quality" is given at each time t by:

$$\dot{s} = f(s,\bar{q},V,T)$$

and therefore $(\partial f/\partial V)_t$ is the marginal rate of transformation of maintenance into quality of the facility at time t. The amount of maintenance V is here expressed in dollars which will be our numeraire private good. We can now write (4-10) as:

$$\frac{\partial V}{\partial f} \exp(-\rho t) = \lambda(t). \qquad (4\text{-}11)$$

The left hand side of (4-11) can then be interpreted as the present value of the marginal rate of transformation between the public good quality and our numeraire private good dollars. Let us go back now to $\lambda(t)$ in (4-8). There

$$-C_{is}\exp(\int_t^x f_s dz)$$

is the reduction in operating costs experienced by one type i user at time x, $x \in [t,T]$, as a consequence of the implementation of an additional unit of facility quality at time t. It will therefore represent how much user i is willing to sacrifice of the private good, dollars at time x, to pay for one more unit of the public good, quality provided at time t. This quantity can be then interpreted as an individualized price for user type i or marginal rate of substitution at time x, $MRS_i(x)$, between the public good, one unit of quality provided at time t, and the private good, dollars at time x. Let us now assume first that demand is independent of quality, $\ell_{is} = 0$, $\forall i$. Then, the second term of $\lambda(t)$ in (4-8) disappears and the condition for optimality in the provision of the public good quality, given by (4-11), can be written as:

$$MRT = \int_t^T \sum_{i=1}^{n} MRS_i(x)q_i(x)\exp(-\rho x)dx. \qquad (4\text{-}12)$$

Expression (4-12) is an obvious generalization, for the dynamic case, of the well-known rule of welfare economics (see, Varian, [1968]):

$$MRT = \sum_i MRS_i . \qquad (4\text{-}13)$$

In (4-12) since the provision of one unit of quality at time t has effects on the operating costs expereinced by the users of the facility throughout the whole period [t,T], the static marginal rates of substitution, $MRS_i(x)$, are added both over all the users of the facility at each time x and over all the times x in the period [t,T]. The inclusion of the actualization factor, $\exp(-\rho x)$, brings this sum to present value at time t = 0, as is the case for MRT in (4-11).

Nevertheless, rule (4-12) assumes that no further effects are produced, over the public system analyzed as a consequence of the provision of one unit of public good, that those that make the users of the system enjoy the consumption of the additional unit of public good. In other words, no externalities are assumed in the production of quality. This assumption holds perfectly for the case just analyzed in which demand is independent of quality. However, if $\ell_{is} \neq 0$ an obvious externality appears in the production of quality. The production of one additional unit of quality at time t produces an increase in the number of each type of user of the facility during the period [t,T], which in turn produces congestion and deterioration externalities. These externalities are taken care of by the second term of $\lambda(t)$ in (4-8). As we know, $\nu_i(x)$ is the social value obtained from the generation (or production) of a new type i user at time x, $x \in [t,T]$. Then

$$SV_i = \int_t^T \nu_i(x)\ell_{is}(x)\exp\left(\int_t^x f_s\,dz\right)dx$$

is the social value attached to the type i users generated during the period [t,T] as a consequence of the production of one additional unit of quality at time t. Given that $\nu_i(x)$ is expressed in terms of present value at time t = 0, SV_i is also expressed in present value. Now (4-11) can be put as:

$$MRT = \int_t^T \sum_{i=1}^{n} MRS_i(x)q_i(x)\exp(-\rho t)\,dx + \sum_{i=1}^{n} SV_i$$

where the last sum represents the total social value of the externalities generated by the provision of one unit of quality at time t. As with all other expressions, it is articulated in present value at t = 0. A similar interpretation of the optimum policy can be given for the cases of corner solutions for $V^*(t)$ in $\Omega(t)$.

5. Sufficient Conditions for Optimality

In the preceding sections we have anlayzed investment policies $V^*(t)$, $t \in [0,T]$, that satisfy the necessary conditions for optimality specified by the Pontryagin maximum principle. Here we will try to find some additional conditions that, when taken together with the necessary conditions already analyzed, form a set of sufficient conditions for optimality. In other words, we will analyze the circumstances under which the Pontryagin conditions produce the optimum

value of $J(V(t))$ for which we are looking. It is obviously enough to analyze the most general case of internal demand, given that the situation in which demand is external can be treated as a special case.

In order to produce sufficient conditions, we will use the Arrow theorem for optimal control problems (see Section 5 of Chapter II). Applied to our case it says that: the policies $[V^*(t), s^*(t), q^*(t)]$ obtained from the necessary conditions, developed in Section 3.1, will lead to a maximum of $J(V(t))$ if $H^*(s,q,\lambda,\nu,t)$ is concave in s and q, for all $t \in [0,T]$", where

$$H^*(s,q,\lambda,\nu,t) = \underset{V}{\text{Max}}\ H(s,q,\lambda,\nu,V,t),\ V \in \Omega \qquad (5\text{-}1)$$

We can distinguish three cases depending on the expression of the optimal control V^*.

(a) Interior Solution, $m(t) < V^*(t) < M(t)$

In this case we can write H^* as

$$H^* = (U - C)q \exp(-\rho t) + \nu\ell + \lambda f^* - V^*(s,q,\lambda,t)\exp(-\rho t), \qquad (5\text{-}2)$$

where $V^*(s,q,\lambda,t)$ is obtained from $\lambda f_V = \exp(-\rho t)$ and $f^* = f^*(s,q,\lambda,t)$ is the deterioration function f in which we have replaced the control variable V by expression of the optimal control $V^*(s,q,\lambda,t)$.

In order for H^* in (5-2) to be concave we need that its Hessian $\hat{H}^*$ be negative definite or semi-definite.

$$\hat{H}^* = \begin{bmatrix} H^*_{qq} & H^*_{qs} \\ H^*_{sq} & H^*_{ss} \end{bmatrix} \tag{5-3}$$

with

$$\begin{aligned} H^*_{qq} &= -(2C_q + C_{qq}q)\exp(-\rho t) + \lambda f^*_{qq} - V^*_{qq}\exp(-\rho t) \\ H^*_{sq} &= -C_s\exp(-\rho t) + \lambda f^*_{qs} - V^*_{qs}\exp(-\rho t) \\ H^*_{ss} &= -C_{ss}q\exp(-\rho t) + \lambda f^*_{ss} + \nu\ell_{ss} - V^*_{ss}\exp(-\rho t) \end{aligned} \tag{5-4}$$

where in the expression for H^*_{sq} we have used the fact that $C_{qs} = 0$ (see expression 2-1).

If f_V is only a function of V or, in other words, if the effectiveness of maintenance is independent of s and q ($f_{Vq} = f_{Vs} = 0$), then

$$V^*_{qq} = V^*_{qs} = V^*_{ss} = 0 \tag{5-5}$$

and

$$f^*_{qq} = f_{qq}, \quad f^*_{qs} = f_{qs}, \quad f^*_{ss} = f_{ss}$$

and the expression for $\hat{H}^*$ simplifies correspondingly.

(b) Corner Solution with $V^* = m(t)$

Now the expression for H^* will be

$$H^* = (U - C)q \exp(-\rho t) + \nu\ell + \lambda f^* - m(t)\exp(-\rho t) \qquad (5\text{-}6)$$

and the components of the Hessian $\hat{H}^*$ will become

$$H^*_{qq} = (2C_q + C_{qq}q)\exp(-\rho t) + \lambda f_{qq}$$

$$H^*_{qs} = - C_s\exp(-\rho t) + \lambda f_{qs} \qquad (5\text{-}7)$$

$$H^*_{ss} = - C_{ss}q \exp(-\rho t) + \lambda f_{ss} + \nu\ell_{ss}$$

In this case, given that $V^* = m(t)$, conditions (5-5) are obviously satisfied.

(c) Corner Solution with $V^* = M(t)$. In this case

$$H^* = (U - C)q \exp(-\rho t) + \nu\ell + \lambda f^* - M(t)\exp(-\rho t) \qquad (5\text{-}8)$$

and the expression for the Hessian $\hat{H}^*$ will be the same as in case (b), given that (5-6) and (5-8) have the same functional form.

In any particular case, in order for the Hessian $\hat{H}^*$ to be negative semi-definite we must have (see Simmons,[1975])

$$H_{qq} < 0, \quad \text{and} \quad \begin{bmatrix} H^*_{qq} & H^*_{qs} \\ H^*_{sq} & H^*_{ss} \end{bmatrix} \geqq 0 \qquad (5\text{-}9)$$

IV. OPTIMUM POLICIES FOR INVESTMENTS IN CAPACITY, CONTINUOUS CASE

1. Introduction

The analysis of optimum policies of capacity investments for transportation facilities has been undertaken by different authors in the economic literature. Mohring (1962), Stroz (1964) and Keeler et al (1975) developed models that relate optimal pricing and investment decisions and produced important insights about the nature of optimum policies. Nevertheless, these models correspond to static formulations of the problem and fail to produce explicit expressions for the optimum investment function. Therefore, the study of dynamic characteristics of the optimum policies is impossible with such models.

In this section we develop a simply dynamic model for capacity investments. Following the authors mentioned above, we consider capacity as a continous variable. The objective is to find explicit expressions for the optimum investment policies and to use them in order to analyze the temporal or dynamic characteristics of these policies in different cases. As has been traditionally argued, the assumption of a continuous capacity function is not very realistic for many public facilities, especially those in the transportation sector which present important plant indivisibilities. Nevertheless, it has been shown that the models built under this assumption can still provide a good deal of insight about the problem, if conveniently analyzed. In any event, we will comment later about the implications

of this continuity assumption and in a subsequent chapter we will develop a different dynamic model, one that explicitly considers indivisibilities in the capacity function. The model developed here will help to complement the results obtained there.

2. A Mathematical Model for Optimal Investments in Capacity

We will assume that we have a transportation facility that serves homogeneous users. Each of these users obtains a utility U(t) and perceives a cost C (k,q) from each trip performed over the facility. The utility depends only on external factors to our model and can be expressed as a function of time. This seems a realistic assumption for all kinds of trips except for the purely recreational ones. In any case, it only constitutes a convenient assumption in order to simplify the formulation of the model and can be dropped without major consequences other than some increase in the complexity of the associated analysis. The operating cost function C (k,q) corresponds to an average variable cost function, which includes all expenses of user-supplied inputs, and it is assumed to depend on the capacity of the facility k and the number of users q. Both variables will, in turn, be functions of time. For the case of purely recreational trips, the utility could be also expressed as U (k,q), but as was mentioned above we will only consider the case U(t) for simplicity. To begin we will only make very general assumptions about the function C (k,q). In particular:

$$C_q \geq 0, \qquad C_k \leq 0$$

$$C_{qq} \geq 0, \qquad C_{kk} \geq 0, \qquad C_{qk} \leq 0 \tag{2-1}$$

The operating cost function will therefore look as in Figure 4.1 Additional traffic, holding capacity constant, will increase operating costs; additional capacity, at the same level of traffic, will on the other hand decrease operating costs. Both effects could be zero at low levels of traffic relative to capacity. In addition, (2-1) states that the value of C_q will increase with q when k is held constant; contrarily $|C_k|$ will decrease as k increases, reaching a value of zero at high capacity values (free flow situation); finally, C_q will decrease as capacity is added, reaching also a limit value of zero for the free flow situation.

Let us now define an investment function f, that gives the amount of investment I necessary to obtain different levels of capacity k.

$$I = f(k) \tag{2-2}$$

We will consider I as a continuous function of the capacity k which, in turn, will be considered continuous. Different special cases for the function f are presented in Figure 4.2. If we differentiate (2-2) with respect to time we obtain:

$$i(t) \equiv \dot{I}(t) = f_k \dot{k}(t) \tag{2-3}$$

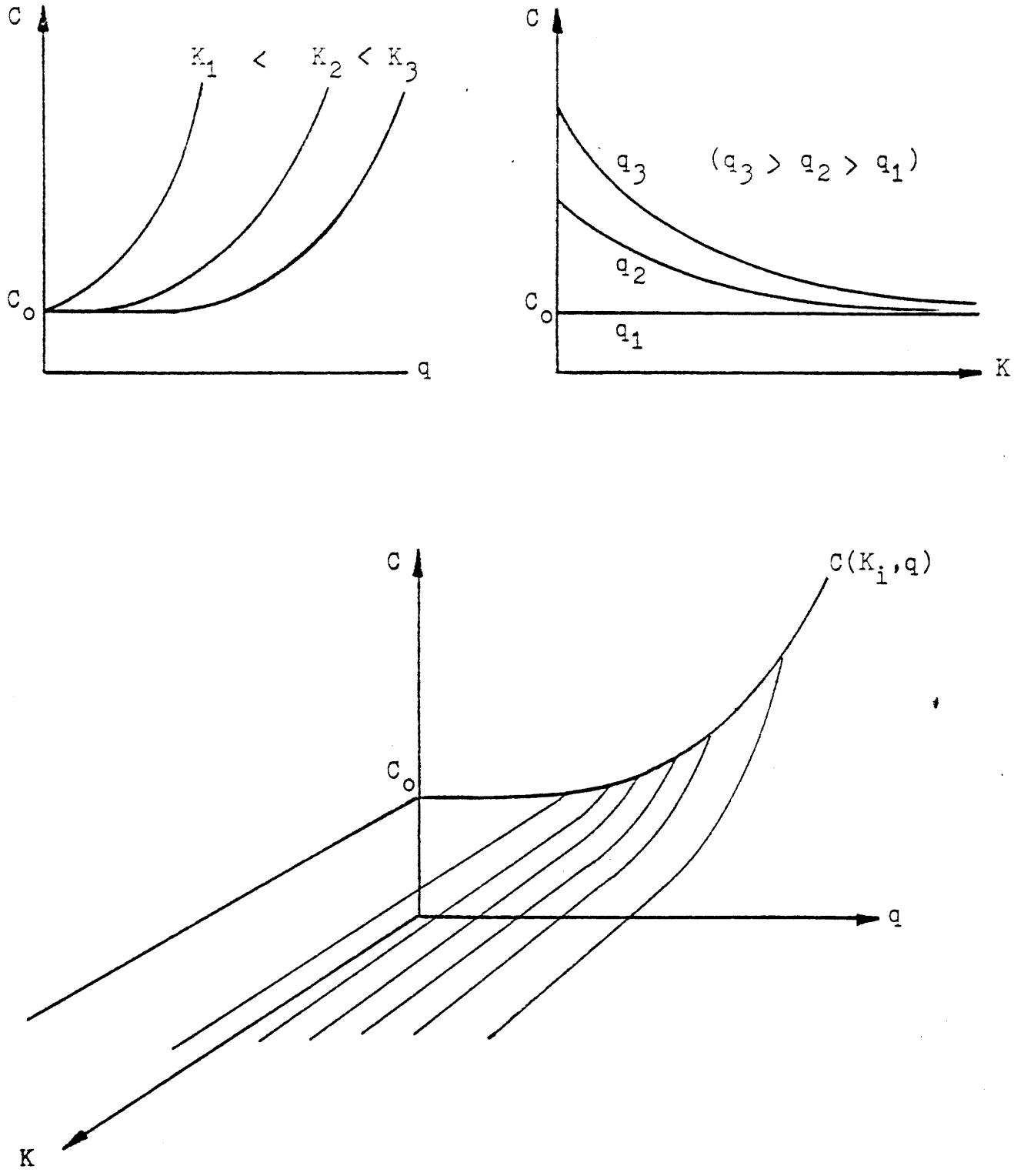


FIGURE 4.1. Operating Cost Function C(K,q).

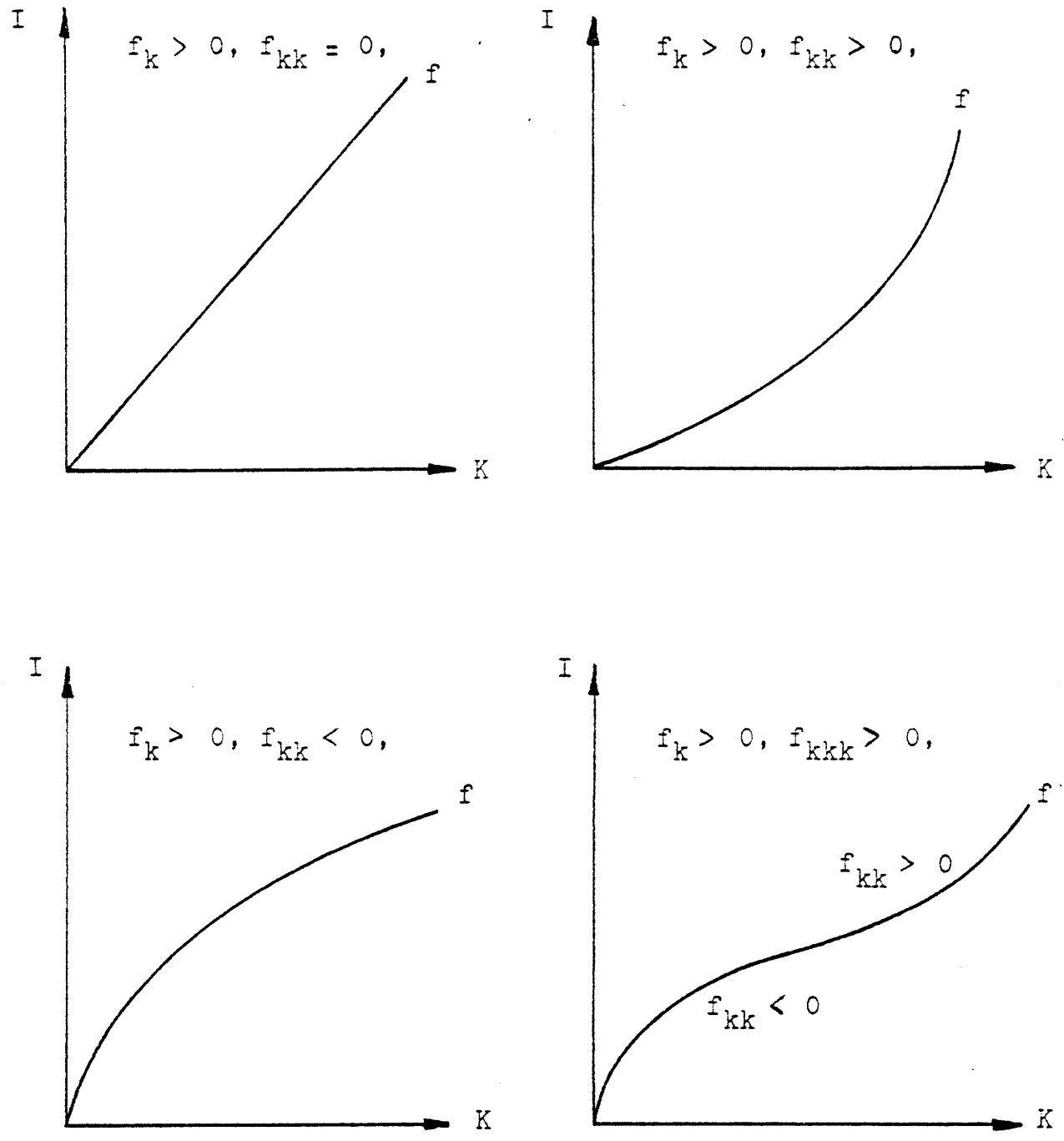


FIGURE 4.2 Capacity Production Function f(k).

where $\dot{k}(t)$ is equal to the change in capacity produced at time t, f_k is the marginal cost of capacity, and i(t) is the amount of dollars spent on capacity at time t.

Two state variables will characterize the transport system of interest: the capacity of the facility at time t, k(t); and the number of users at time t, q(t). As a control variable we will choose, the change in capacity of the facility at time t which we will denote by V(t). With this notation we can now express the expenditures in capacity, per unit of time, using (2-3), as:

$$i(t) = f_k\, V(t) \tag{2-4}$$

We wish to find a function $V^*(t)$ that maximizes the present value of the net social benefits of the system over the period of analysis [0,T]. Therefore the objective function will be:

$$J(V(t)) = \int_0^T \{[U(t) - C(k,q)]q(t) - f_k V(t)\}\exp(-\rho t)dt + \Psi(k(T))\exp(-\rho T) \tag{2-5}$$

where ρ is the appropriate interest rate and Ψ is the residual value of the facility at time T, that we assume to be a function of the capacity of the facility at that time. According to our definition of V(t) above, the evolution of the capacity of the facility over time will be determined by the simple differential equation

$$\dot{k}(t) = V(t), \qquad k(0) = k_0 \tag{2-6}$$

Finally, we will assume that there is a limit M in the amount of capacity that we can provide per unit of time and that, once in place, capacity cannot be removed (dis-investment is impossible). Therefore, we will have the following constraints for our control variable:

$$0 \leq V(t) \leq M, \ \forall \ t \in [0,T] \qquad (2\text{-}7)$$

3. Necessary Conditions for Optimality: the functional form of optimal policies.

The problem of maximizing the objective function J(V(t)) defined in Equation (2-5), subject to the capacity dynamics (2-6) and the control constraints (2-7) constitutes an optimal control problem with fixed terminal time and no state space constraints. In order to specify the necessary conditions for a maximum of J(V(t)) we need to define the Hamiltonian function:

$$H(t) = \{[U(t) - C(k,q)]q(t) - f_k V(t)\}\exp(-\rho t) + \lambda(t)V(t), \qquad (3\text{-}1)$$

where $\lambda(t)$ is the adjoint variable corresponding to the state variable k(t).

The necessary conditions for a maximum of J(V(t)) over V(t) state that there must exist a function $\lambda(t)$ that satisfies the adjoint equation

$$\dot{\lambda} = - (\partial H/\partial k) = [C_k q + f_{kk} V]\exp(-\rho t) \qquad (3\text{-}2)$$

with

$$\lambda(T) = (\partial \Psi/\partial k)_T \exp(-\rho T) = \lambda_T \qquad (3\text{-}3)$$

and that the optimum control $V^*(t)$ must be such that the value of the Hamiltonian (-) is maximized for all t in [0,T]. This last requirement can be denoted as

$$H(k^*,\lambda^*,V^*,t) \geq H\ (k^*,\lambda^*,V,t), \quad \forall\ t \in [0,T] \qquad (3\text{-}4)$$
$$(V \in \Omega)$$

where Ω is defined by (2-7) and the * in k and λ mean that these variables satisfy (2-6) and (3-2) respectively. Expressions (2-6), (2-7), (3-2), (3-3) and (3-4) constitute a complete set of necessary conditions for our problem.

It is easy to see that the gradient of the Hamiltonian with respect to the control variable V is given by

$$H_V = \lambda - f_k \exp(-\rho t) \qquad (3\text{-}5)$$

and given that the Hamiltonian is a linear function of V, the optimum function $V^*(t)$ will be given by

$$V^*(t) = \begin{cases} M & , \text{ if } H_V^* > 0 \\ 0 & , \text{ if } H_V^* < 0 \\ \text{undetermined}, & \text{if } H_V^* = 0 \end{cases} \tag{3-6}$$

3.1 Bang-Bang Controls

We observe that expressions (3-5) and (3-6) imply that $V^*(t)$ is a well defined function of $k^*(t)$, $\lambda^*(t)$ and t as long as the gradient $H_V^*(t)$ is non-zero. In that case we have

$$V^*(t) = \begin{cases} M, \text{ if } \lambda^*(t) > f_k^* \exp(-\rho t) \\ 0, \text{ if } \lambda^*(t) < f_k^* \exp(-\rho t) \end{cases} \tag{3-7}$$

and the optimum control is called "bang-bang". In order to give an economic interpretation of the controls defined in (3-7) we can use the following equality that holds along an optimum solution (see Chapter II; Section 4.)

$$\lambda^*(t) = \frac{\partial J}{\partial k^*}(k^*, t) \tag{3-8}$$

Expression (3-8) says that $\lambda^*(t)$ represents the shadow price of capacity at time t. A more insightful view of $\lambda^*(t)$ can be obtained using the adjoint Equation (3-2). If we integrate (3-2) using (3-3) we obtain the following expression for $\lambda(t)$:

$$\lambda(t) = \int_t^T \{-C_k q(x) - (\partial i/\partial k)\}\exp(-\rho x)dx + \lambda_T, \qquad (3\text{-}9)$$

where we have used the relation

$$f_{kk}V = \partial(f_k V)/\partial K = \partial i/\partial k = i_k \qquad (3\text{-}10)$$

In expression (3-9) we have that $-C_k q(x)$ is equal to the total operating cost reductions experienced by the users of the facility at time x ($x > t$) as a consequence of an additional unit of capacity provided at time t, and i_k is the effect that an additional unit of capacity provided at time t has over the cost of providing capacity at time x. The value of i_k takes into account the changes in the cost of providing capacity in the future (after t) caused by one additional unit of capacity provided at time t. Only if there are constant returns to scale in the facility construction will this term always be zero within the interval [0,T], because then the marginal cost of capacity f_k is constant for all k. Nevertheless, if f presents decreasing returns to scale ($f_{kk} > 0$), i_k will be positive as long as $V > 0$. On the other hand, if f presents increasing returns to scale ($f_{kk} < 0$), i_k will be negative for all the periods with $V > 0$. Therefore, there exists a clear interrelation between present and future investment decisions when $f_{kk} \neq 0$. For $f_{kk} > 0$, future investments, after t, will decrease the value of $\lambda(t)$, and vice versa for $f_{kk} < 0$. It appears then, that more investment will be justified in general when f presents increasing returns than when it presents decreasing

returns to scale. In Section 5, this will be clearly shown to be the case. The value of λ_T in (3-9) represents, according to (3-3), the present value of the marginal value of capacity at time T.

The bang-bang policy defined by (3-7) says that capacity should be provided, at a rate of M units of capacity per unit of time, as long as the shadow price of capacity remains higher than the marginal cost of capacity. When the converse happens, capacity should be held constant. If the gradient function $H_V^*(t)$ vanishes only at a countable number of times, within the interval [0,T], our optimal control problem is called "normal" and the optimum policy $V^*(t)$ is "bang-bang." The value of $V^*(t)$ switches from one boundary of Ω (defined by (2-7)) to the opposite one, at certain well-defined times t_s given by:

$$\lambda^*(t_s) = f_k^* \exp(-\rho t_s), \quad t_s \in [0,T], \ (S = 1,2,\ldots,N) \qquad (3\text{-}11)$$

Figure 4.3 illustrates a function H_V^* and the corresponding $V^*(t)$ in this case. H_V^* is then the so-called "switching function". However, given that our objective function (2-5) is non-linear in the state variable k, there exists the possibility that H_V^* vanishes identically along some finite periods of time in [0,T] (see Chapter II, Section 3.2). The optimum controls during those periods, if they exist, are called "singular" and will be analyzed in the next section.

3.2 Singular Controls

In the preceding section we assumed that the switching function H_V

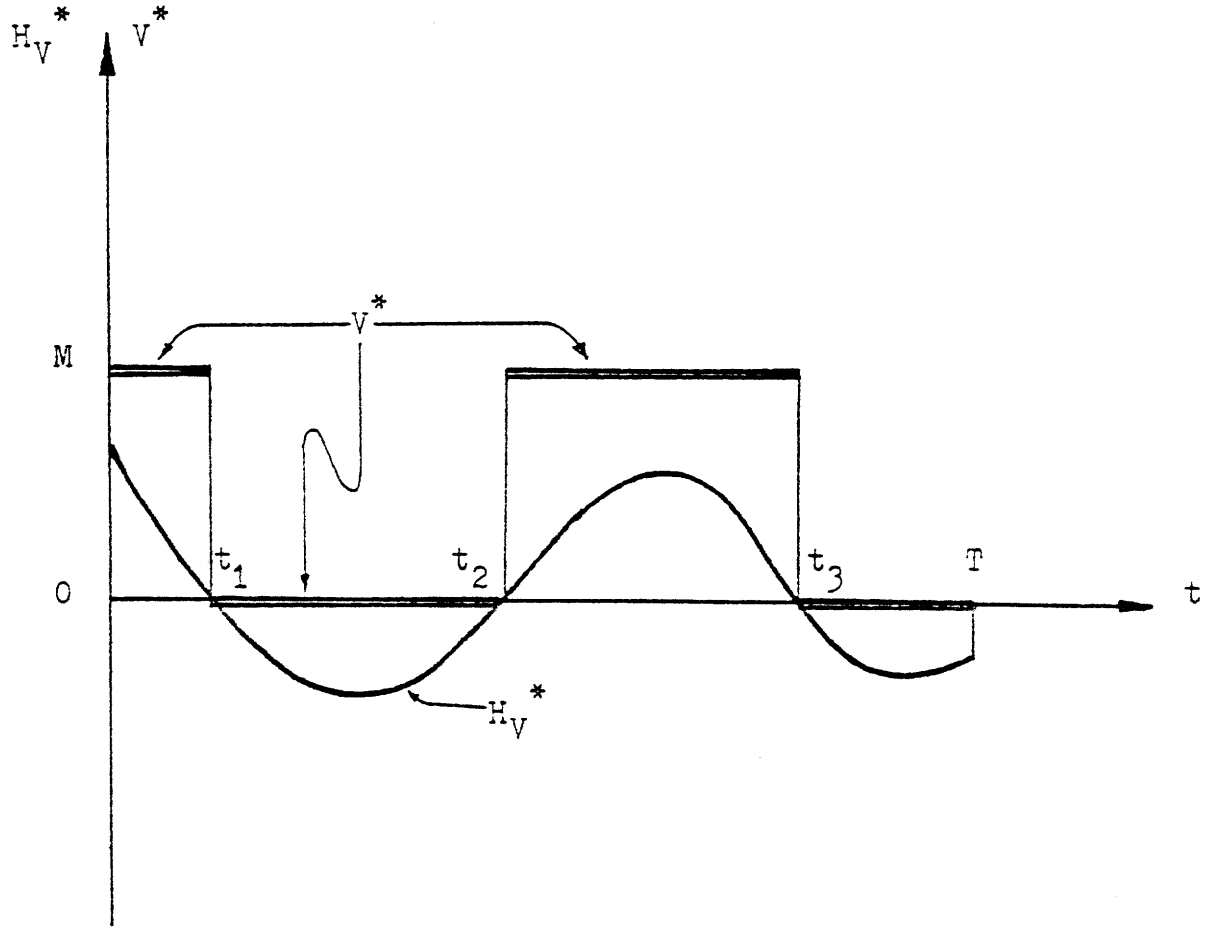


FIGURE 4.3 Bang-bang Policy.

vanishes only at a countable number of times in the period [0,T]. In this section we will analyze the possibility that the gradient H_V vanishes identically over one or more finite periods of time or sub-intervals in [0,T]. Then we will say that we have a singular optimal control problem and the periods for which $H_V = 0$ are called singularity intervals or singular arcs. As we noted in (3-6) the necessary condition (3-4) do not provide in this case enough information in order to define $V^*(t)$ along a singular arc. In the absence of such information, we must manipulate the other necessary conditions in an effort to determine a well-defined expression for $V_s^*(t)$, which receives the name of singular control.

Singular controls can be in general determined making use of the following observation (see Chapter II, Section 3.2): "If the gradient H_V of the Hamiltonian vanishes identically along a singular arc, then the time derivatives of H_V must remain equal to zero during the same period." From (3-5) and (3-6) we have that at a singular arc

$$H_V = \lambda - f_k \exp(-\rho t) = 0 \quad . \tag{3-12}$$

Thus upon taking the time derivative of (3-12) we obtain

$$\dot{H}_V = \dot{\lambda} - f_{kk}\dot{k} \exp(-\rho t) + \rho f_k \exp(-\rho t) = 0 \tag{3-13}$$

which, making use of the necessary conditions for k and λ expressed by the capacity dynamics (2-6) and the adjoint equations (3-2) and dividing by the positive value exp (-ρt), becomes

$$C_k q + \rho f_k = 0 \quad . \tag{3-14}$$

Now, given that (3-14) must hold along the singular arc we can take a new time derivative of this expression to obtain

$$C_{kk}\dot{k}q + C_{kq}\dot{q}q + C_k\dot{q} + \rho f_{kk}\dot{k} = 0 \tag{3-15}$$

and using again the necessary condition (2-6) we get from (3-15) the following expression for the singular control

$$V_s^* = -\dot{q}\{(C_{kq}q + C_k)/(C_{kk}q + f_{kk}\rho)\} \quad . \tag{3-16}$$

In addition, the singular control must obviously satisfy the control constraint

$$0 \leqq V_s^* \leqq M \quad . \tag{3-17}$$

Given our assumption (2-1) about the cost function C, we can easily see that V_s^* in (3-16) will automatically satisfy the non-genativity constraint for all the periods of increasing demand ($\dot{q} > 0$) if

$$f_{kk} > -C_{kk}(q/\rho) \tag{3-18}$$

where the right hand side of (3-18) is always non-positive. As we will see later (3-18) constitutes an additional necessary condition

for singular arcs.

It is important to note that (3-12) and (3-14) constitute necessary conditions for the existence of the singular control V_s^* given by (3-16). In other words, V_s^* can be an optimum policy only when condition (3-12) holds. In that case the use of a control V_s^* will maintain the solution path on the singular arc, along which (3-14) is also satisfied. Equation (3-12) tells us that, for all the values of t along a singular arc, the application of V_s^* will produce a level of capacity $k^*(t)$ such that the marginal benefit of capacity $\lambda^*(t)\exp(\rho t)$ is equal to the marginal cost of capacity f_k^*. In addition, if we put (3-14) as

$$-C_k q(t) = \rho f_k \quad , \quad (3\text{-}19)$$

we have that, along a singular arc, capacity is provided in such a way that, at each time t, the marginal reductions in operating costs, produced by the last unit of capacity, are equal to the rental value of the cost of providing that unit of capacity. We can think of (3-19) as the Marglin naive static rule (Marglin, 1963) applied to the marginal unit of capacity. In that case (3-19) says that the next unit of capacity will be provided at a time t such that the present value at t of a perpetual stream of benefits at the immediate rate $-C_k q(t)$ equals the construction cost of that unit

$$f_k = -[C_k q(t)/\rho] \quad . \quad (3\text{-}20)$$

The construction cost of the next unit is the marginal cost of capacity f_k. This rule applies to all the units of capacity provided along the singular arc.

At the beginning of Section 3 we saw that one of the necessary conditions that an optimum control must satisfy is the maximization of the Hamiltonian required by (3-4). It is easy to see that the control V_s^* specified by (3-16) does not necessarily satisfy that condition. Actually, we derived V_s^* from the fact that along a singular arc $H_V = 0$, which is only a first order condition for the maximization of the Hamiltonian with respect to the control. Therefore, we still need a second order condition in order to ensure that V_s^* corresponds to a maximum of H(t) and not a minimum or an inflexion point. This second order condition is provided, in general, by the requirement that $H_{VV}^* > 0$ (see Byson and Ho, 1975). In the linear case, for non-singular arcs (bang-bang controls), the fact that for all $V \in \Omega$, $H_V^* (V-V^*) < 0$, is a sufficient condition for the maximization of the Hamiltonian (see Luenberger, 1973). However, for singular arcs we have both

$$H_V^* \equiv 0 \qquad \text{and} \qquad H_{VV}^* \equiv 0 \tag{3-21}$$

and therefore an additional necessary condition must be applied. This condition, developed by Tait [1965], Robins [1965] and Kelley et al [1966], states that in order for V_s^* to produce a maximum of J(V(t)) it is necessary that in addition to all the necessary conditions already developed [(2-6), (2-7), (3-2), (3-3), (3-12) and (3-14)] the

following condition must be satisfied in our case

$$(\ddot{H}_V)_V = \frac{\partial}{\partial V} \{(d/dt)^2 H_V\} \geq 0 \quad . \tag{3-22}$$

For our problem it is easy to check that this reduces to

$$(\ddot{H}_V)_V = C_{kk} q + f_{kk} \rho \geq 0 \quad . \tag{3-23}$$

Given our assumptions (2-1) about the function C, and given that q(t) is by definition non-negative, condition (3-23) will be always satisfied for all f functions with decreasing or constant returns to scale ($f_{kk} \geq 0$). It could be violated nevertheless for functions that present strong increasing returns to scale if

$$f_{kk} < -C_{kk}(q/c).$$

In that case singular control given by (3-16) would not correspond to an optimal solution to our problem. It is also interesting to note that (3-23) can be put in the form

$$\frac{\partial}{\partial k} (-C_k q - f_k \rho) < 0 \tag{3-24}$$

which is a second order condition, with respect to k, for rule (3-19). The term in parenthesis in expression (3-24) is the net benefit, at time t, of providing an additional unit of capacity at that time.

Therefore condition (3-23) ensures that condition (3-19) leads to a static maximization of benefits at time t.

3.3 Dynamic Optimum Policies

In the preceding sections we have used the necessary conditions provided by the Pontryagin maximum principle in order to derive expressions for the optimum controls for our problem. Two cases were identified and analyzed: bang-bang and singular controls. In practice, optimum policies will, in general, involve a combination of both. In that case, the optimum controls will be defined by:

$$V^*(t) = \begin{cases} \overline{M} & , \text{ if } \lambda^* > f_k \exp(-\rho t) \\ -\dot{q}\,\{C_{kq}q + C_k)/(C_{kk}q + f_{kk}\rho)\}, & \text{if } \lambda^* = f_k \exp(-\rho t) \\ 0 & , \text{ if } \lambda^* < f_k \exp(-\rho t) \end{cases} \quad (3\text{-}25)$$

A dynamic optimum policy can in general be represented by a path in the positive quadrant of the space of state variables (q,k). In that space, bang-bang arcs with $V^* = 0$ are represented by horizontal lines, k = constant, given that along them $\dot{k} = 0$ (see Figure 4.4 *a*). For bang-bang arcs with $V^* = M$ we have:

$$\dot{k} = M \Rightarrow k(t) = k_s + M(t - t_s), \quad (3\text{-}26)$$

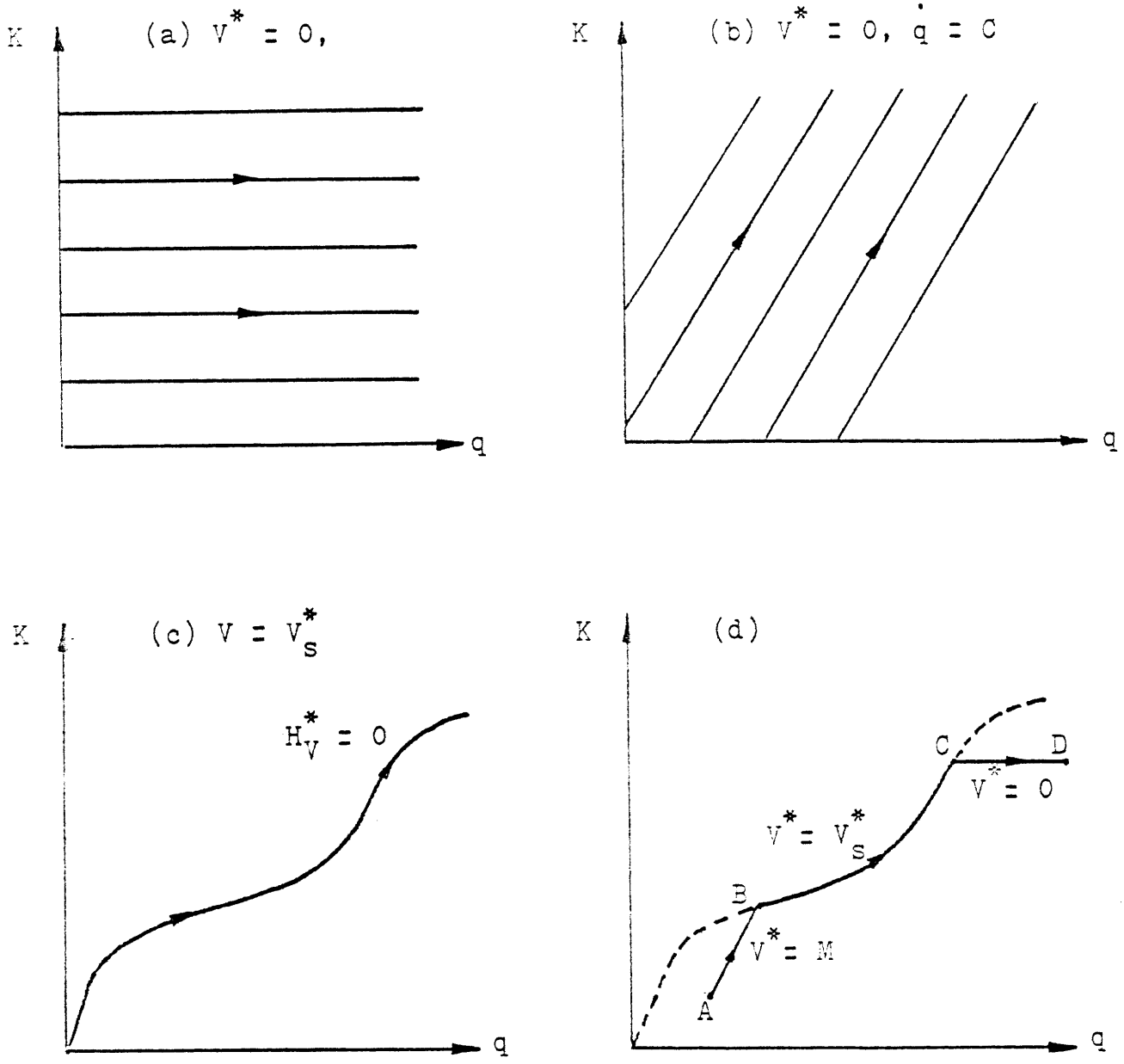


FIGURE 4.4 Optimal Policies in (q,K)

(a) Bang-bang arcs with $V^* = 0$,
(b) Bang-bang arcs with $V^* = M$, $\dot{q} = C$,
(c) Singular arc, $H_V^* = 0$, $V^* = V_s$,
(d) Optimum Policy representation.

where t_s is the switching time at which the application of $V^* = M$ begins and k_s is the capacity at that time. Using (3-26) to eliminate the variable t from the demand schedule $q = q(t)$ we obtain the following equation for arcs with $V^* = M$

$$q = q((k-k_s)/M + t_s \tag{3-27}$$

Applying the chain rule it is easy to see that the slope of this curve is given by

$$\frac{dk}{dq} = \left(\frac{dq}{dt}\frac{dt}{dk}\right)^{-1} = (M/\dot{q}) \tag{3-28}$$

which is always positive for $\dot{q} > 0$. If $\dot{q} = 0$ the bang-bang arcs defined by (3-27) become vertical straight lines in (q,k). Given a constant value of M the slope (3-28) will change with the value of $\dot{q}$, decreasing when the latter increases and vice versa. In Figure 4b we represent a family of these bang-bang arcs for a case with $\dot{q}$ constant. Finally, an expression for the singular arc in (q,k) can be obtained from the necessary condition (3-14).

$$C_k\, q + \rho f_k = 0 .$$

For any particular case, this is an equation in q and k that defines a unique line in (q,k). It specifies the points, or states of the system, for which the condition (3-12), that defines the singular arc, holds. Given condition (3-23) this line will always

have a positive slope for $q > 0$. A singular arc is represented in Figure 4c. It divides the space (q,k) in two half-spaces; states of the system represeted by points located below the singular arc correspond to situations with capacity values lower than optimum and vice versa for points located above the singular arc. In that case, the bang-bang arcs represented in Figures 4a and 4b provide optimum paths to get onto the singular arc. On the other hand, we can see from (3-3) and (3-5) that the final state of the system, at time T, can only be over the singular arc, if

$$\Psi_k(T)\exp(-\rho T) = f_k\exp(-\rho T) \ . \qquad (3\text{-}29)$$

This condition will always be satisfied for $T = \infty$, but if T is finite, it will in general be violated. In particular, if the residual value of the facility Ψ is equal to zero we will have for any finite value of T

$$\lambda_T < f_k\exp(-\rho T) \ , \ \ f_k > 0.$$

If (3-29) does not hold we will have to leave the singular arc at some time t before T, in order to satisfy the final conditions (also called transversality conditions) of the problem. Again in this case, the bang-bang arcs provide optimum paths to follow. We know from Section 3.1 and 3.2 that an optimal solution can only be formed by bang-bang and/or singular arcs. Therefore, the optimum solution to our problem will involve, in general, an initial bang-bang arc

to get onto the singular arc, the use of such an arc for as long as possible and a final bang-bang arc in order to meet the final conditions of the problem. The value of V^* along the initial and final bang-bang arcs will depend on the location of the points (q,k) that represent the initial state of the system (q_0,k_0) and the optimum final state $(q_T,k^*(T))$. In Figure 4d we represented a typical optimum policy. Optimum policies for some special cases of interest will be analyzed in detail in Section 5.

4. Sufficient Conditions for Optimality

In Section 3 we have found investment policies $V^*(t)$, $t \in [o,T]$, that satisfy the necessary conditions for optimality specified by the Pontryagin maximum principle. In this section we will specify the circumstances under which those policies produce a maximum value of the objective function $J(V(t))$. In other words we will study the sufficiency conditions for our optimization problem. With this purpose we will make use of the Arrow sufficiency theorem for optimal control problems (see Chapter II, Section 5).

The Arrow sufficiency theorem, applied to our problem, says that: the policies $[V^*(t),k^*(t)]$ obtained in Section 3 will lead to a maximum of $J(V(t))$ in (2-5) if $H^*(k,\lambda,t)$ is concave in k, for all $t \in [0,T]$, where

$$H^*(k,\lambda,t) = \underset{V \in \Omega}{\text{Max}}\ H(k,\lambda,V,t) \qquad (4\text{-}1)$$

In our case we have three different expressions for $H^*(k,\lambda,t)$ depending on the expression of the optimal control V^*. These are:

(a) Bang-Bang control with $V^* = 0$

$$H^* = (U - C)\, q \exp(-\rho t) \tag{4-2}$$

(b) Bang-Bang control with $V^* = M$

$$H^* = (U - C)\, q \exp(-\rho t) + M[\lambda(t) - f_k \exp(-\rho t)] \tag{4-3}$$

(c) Singular control, $V^* = V_s$

$$H^* = (U - C)\, q \exp(-\rho t) + V_s[\lambda(t) - f_k \exp(-\rho t)] \tag{4-4}$$

Remember that in this case $\lambda = f_k \exp(-\rho t)$ for all t along the singular arc.

The function H^* will be concave in k if its second derivative with respect to this variable in non-positive. For case (a) we have from (4-2) that:

$$H^*_{kk} = - C_{kk} q \exp(-\rho t) \tag{4-5}$$

and therefore H^* will always be concave for this case given that, according to the assumptions (2-1), C_{kk} will be always non-negative. In other words, for bang-bang policies with $V^* = 0$ the necessary conditions provided by the maximum principle are also sufficient for optimality.

Now, for case (b) we have from (4-3) that

$$H^*_{kk} = - (C_{kk}q + f_{kkk}M)\exp(-\rho t) \qquad (4\text{-}6)$$

and H^* will be concave here only if

$$C_{kk}q + f_{kkk}M \geqq 0. \qquad (4\text{-}7)$$

Given assumptions (2-1), this condition will always be satisfied for capacity production functions f with a non-negative third order derivative f_{kkk}. Therefore for those functions the necessary conditions will also be sufficient for optimality. Nevertheless, if f_{kkk} is negative then (4-7) defines additional conditions that the policy obtained from the necessary conditions must satisfy in order to be optimal. In particular, (4-7) imposes the following constraint on the value of M:

$$M \leqq - q\ (C_{kk}/f_{kkk}) \qquad (4\text{-}8)$$

where the right hand side is positive because $q \geqq 0$, $C_{kk} \geqq 0$ and $f_{kkk} < 0$. Now, given that we are analyzing the case $V^* = M$, for which $H^*_V > 0$, it is obvious that the value of the Hamiltonian will increase as V^* increases. Nevertheless, we can only increase V^* until it reaches the value $-q\ (C_{kk}/f_{kkk})$, because beyond that value, the necessary conditions are no longer sufficient for optimality. Therefore, for the case in which $f_{kkk} < 0$, the maximum possible value of the Hamiltonian corresponding to a control that satisfies the

sufficiency conditions is obtained by choosing

$$V^* = M \quad , \text{ if } M \leq -q(C_{kk}/f_{kkk}), \tag{4-9a}$$

$$V^* = -q(C_{kk}/f_{kkk}), \text{ if } M > -q(C_{kk}/f_{kkk}), \tag{4-9b}$$

which provides the expression for the optimal control in the case $H_V^* > 0$, when $f_{kkk} < 0$.

Finally for case (c) we can obtain from (4-4)

$$H_{kk}^* = -(C_{kk}q + V_s f_{kkk})\exp(-\rho t) \quad ,(4\text{-}10)$$

and H^* will be concave if

$$C_{kk}q + V_s f_{kkk} \geq 0 \quad ,(4\text{-}11)$$

Therefore, we have a similar situation to case (b) in which if $f_{kkk} < 0$ the optimal control will be given by

$$V^* = V_s \quad , \quad \text{if } V_s \leq -q(C_{kk}/f_{kkk}), \tag{4-12a}$$

$$V^* = -q(C_{kk}/f_{kkk}), \quad \text{if } V_s > -q(C_{kk}/f_{kkk}), \tag{4-12b}$$

where V_s is given by expression (3-16)

5. Special Cases of Interest

The discussion in previous sections has been largely abstract and theoretical. We have tried to maintain in preceeding sections the highest possible level of generality in our analysis. Some general assumptions about the operating costs function C were made at the beginning in order to keep certain realism in the analysis. No additional conditions about C were later necessary to guarantee sufficiency. With respect to the construction cost function f, we began the analysis without special assumptions other than continuity and differentiability. Later, it was shown that f should satisfy condition (4-7) in order to guarantee sufficiency in the bang-bang case with $V^* = M$ and condition (3-23) for singular paths to be candidates to produce optimum solutions. Finally, no conditions at all were necessary with respect to the demand function q(t) other than the obvious requirement that the total number of users of the facility be non-negative for each t in [0,T]. In particular, the rate of change of demand $\dot{q}$ can be positive, negative, zero or any combination of them along the period [0,T].

In this section, to illustrate the results obtained, we will apply them to some simple special cases. We will in particular assume that the operating cost function C is homegeneous of degree zero in q and k. This is equivalent to assuming that the individual operating costs are dependent only on the volume-capacity ratio of the facility. This assumption is commonly used in the transportation economics literature and therefore is interesting to apply our results to this special case. In analytic terms it implies, using Euler's theorem on

homogeneous functions (see Allen, [1971])

$$C_k k + C_q q = 0 \tag{5-1}$$

If we differentiate C with respect to time and use (5-1) we can easily obtain

$$\frac{dC}{dt} = C_k k\left(\frac{\dot{k}}{k} - \frac{\dot{q}}{q}\right) \tag{5-2}$$

and taking the derivative of (5-1) with respect to k we can also obtain

$$C_{kk} k = -\left(C_k + C_{kq} q\right). \tag{5-3}$$

Then, if we use (5-3) in (3-16) we get

$$V_s^* = \dot{q}\left\{C_{kk} k / C_{kk} q + f_{kk} \rho\right\} \tag{5-4}$$

which gives an expression for the optimum singular control in the homogeneous case. To understand the implications of V_s^* in this case, we can use (5-1) in (3-14) to obtain

$$(C_q q) q = \rho(k f_k) \tag{5-5}$$

which must hold for all t along a singular arc. Given that $C_q q$ is equal to the congestion costs produced by each user of the facility at time t, expression (5-5) says that, along a singular arc, the

capacity k provided at each time t should be such that the congestion costs, produced by each user, times the numer of users of the facility be equal to the rental value of the facility when a unitary cost of capacity equal to the marginal cost of capacity is used. If each user of the facility is charged a toll such that the total cost perceived by him is equal to short-term marginal cost,(5-5) leads to the well-known result that says that, if constant returns to scale exist in capacity construction, the total income collected from congestion tolls will just cover the total rental costs of the facility or

$$(C_q q)\, q = \rho f(k), \quad f(k) = \gamma k. \tag{5-6}$$

This condition must be true for each time t along a singular arc. If there are decreasing returns, for all possible values of k, $f_{kk} > 0$, and then marginal cost tolls will yield an operating surplus, per unit of time, along a singular arc. With increasing returns, on the other hand, a deficit per unit of time will appear. Nevertheless, we must remember that for this last case V_s^* will be a candidate for an optimum policy only if (3-23) is satisfied, or in other words, the increasing returns characteristic is not too strong. In any case V_s^* given by (5-4) constitutes an explicit expression for the implementation of the policy stated by (5-5). Such an explicit expression has not previously been available in the economic literature. A more general expression that does not depend on the assumption of homogeneity for the operating cost function C is given by (3-16).

5.1 Constant Returns to Scale in Capacity Construction

In this case we can represent the capacity production function by

$$f = \gamma k, \qquad \text{with: } f_k = \gamma,\ f_{kk} = 0 \tag{5-7}$$

Therefore, we can write from (3-5) and (3-2):

$$H_V = \lambda - \gamma \exp(-\rho t) \tag{5-8}$$

$$\dot{\lambda} = C_k q \exp(-\rho t) \ . \tag{5-9}$$

We will assume here that the planning period is $[0,\infty]$; therefore from (3-3) we obtain

$$\lambda_T = \lambda(\infty) = 0 \tag{5-10}$$

and also, that demand is not decreasing, or $\dot{q} \geqq 0$, $\forall\ t \in [0,\infty]$.

Using (5-7) in (5-4) we can easily get that the expression for the singular control V_s^* becomes in this case

$$V_s^* = k\ (\dot{q}/q) \ . \tag{5-11}$$

If we now use (2-6) in (5-11) we get that the equation of the singular arc in the space (k,q) is given by

$$\frac{dk}{k} = \frac{dq}{q} \tag{5-12}$$

Therefore, if constant returns to scale exist in capacity construction and we can assume that C is homogenous of degree zero in k and q, once we get onto a singular arc the optimum policy is to increase capacity in the same percentage as increases in demand.

Equation (5-12) corresponds to a straight line going through the origin of the space (k,q), (see Figure 4.5a). The equation of this straight line can also be obtained directly from (3-14) or (5-5) which represent the necessary conditions to stay on a singular arc. It is interesting to note that (5-11) implies:

$$(\dot{k}/k) = (\dot{q}/q)$$

which when introduced into (5-2) gives $\dot{C} = 0$. Therefore, the operating cost will remain constant along the singular arc.

Let us assume now that the operating cost function C is given by

$$C(k,q) = \gamma + \beta(q/k)^n, \qquad n \geq 2, \qquad \beta > 0. \tag{5-13}$$

This is an obvious extension of those cost functions used in practice by the U.S. Federal Highway Administration (FHWA) (see Comsis, 1972). Actually for each fixed value of k, (5-13) transforms into one of the FHWA functions. From (5-13) we can now get

$$C_k = - n\beta(q^n/k^{n+1})$$

$$C_{kk} = n(n+1)\beta(q^n/k^{n+2}) \tag{5-14}$$

$$C_q = n\beta(q^{n-1}/k^n)$$

$$C_{qq} = n(n-1)\beta(q^{n-2}/k^n) \qquad ,(5\text{-}14)$$

$$C_{qk} = n^2\beta(q^{n-1}/k^{n+1}).$$

It is easy to check that conditions (2-1), (4-7) and (3-23) are satisfied in this case; therefore, bang-bang and singular arcs that satisfy all the necessary conditions will maximize J(V), and combinations of both will be possible.

Now, from (5-5), (5-7) and (5-14) we can get the following explicit expression for the equation of the singular arc represented in Figure 4.5a

$$k = \theta q, \qquad \theta = (n\beta/\gamma\rho)^{(n+1)^{-1}} \qquad (5\text{-}15)$$

Here θ is directly proportional to the parameter β describing the influence of the congestion effects on the operating cost C and inversely proportional to the rental value of the constant marginal cost of construction $\rho\gamma$. If the rental value of capacity increases, less capacity will be justified for each level of q and vice versa. On the other hand, the optimum level of capacity for each q will increase if the cost of congestion β increases.

The straight line $\overline{OS}$ in Figure 5a describes a circumstance for which the condition $\lambda(t) = \gamma\exp(-\rho t)$ is satisfied. At the initial time t = 0, if the initial point (q_o, k_o) is below $\overline{OS}$ we will have less

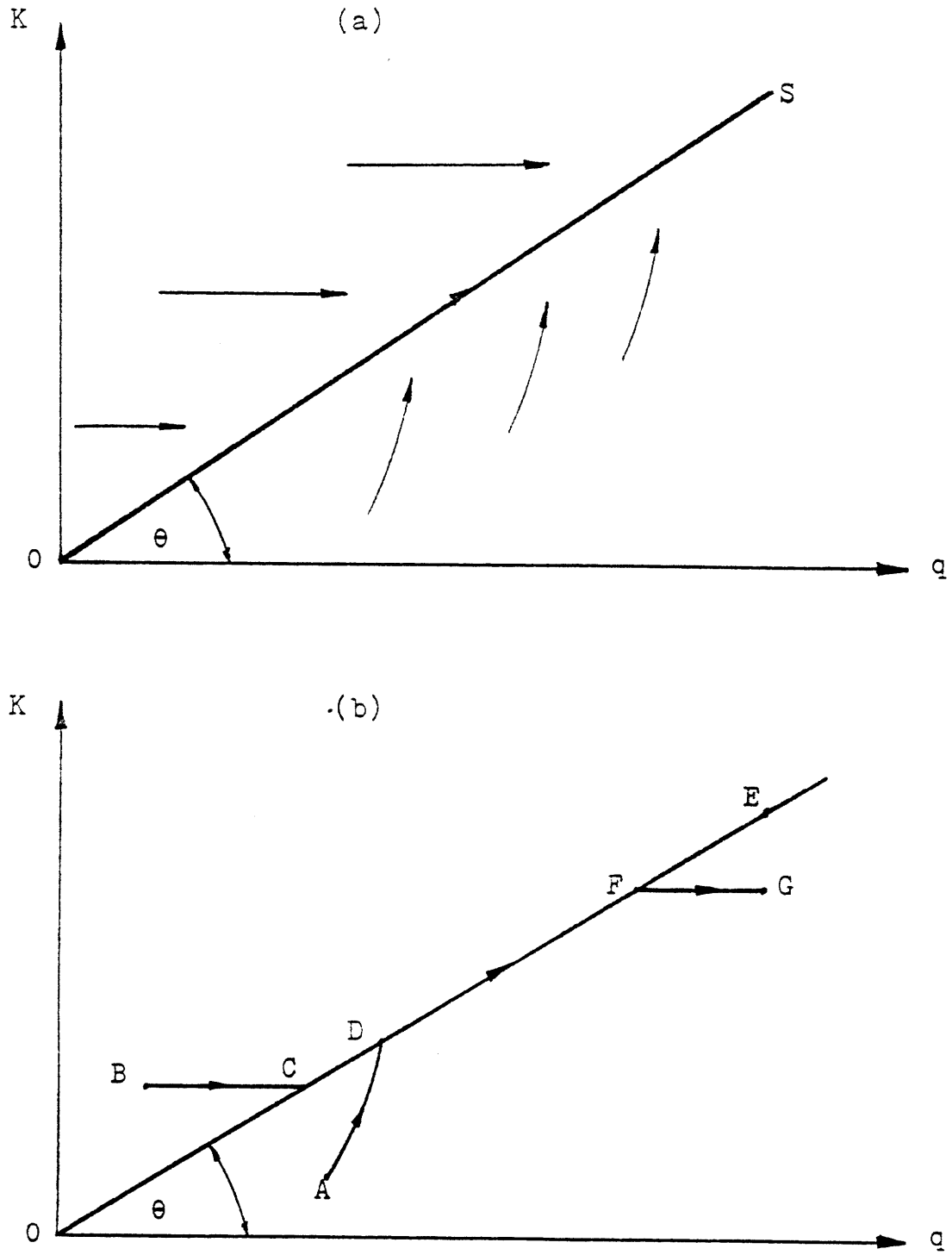


FIGURE 4.5 Optimal Policies. Constant Returns to Scale Case.
(a) Singular arc,
(b) Optimal Policies.

than the optimum capacity and therefore $\lambda(0) > \lambda$. On the other hand, if (q_o, k_o) is above $\overline{OS}$ too much capacity will be in place and $\lambda(0) < \lambda$.

In Figure 4.5b we have represented some typical cases of optimum singular policies. These are:

Case 1 (A-D-E)

$$\text{arc } \overline{AD}: \quad H_V^* > 0, \qquad V^* = M, \text{(bang-bang)}$$

$$\text{arc } \overline{DE}: \quad H_V^* = 0, \qquad V_s^* = k\,(\dot{q}/q), \text{(singular)}$$

The terminal point of this policy is (q_E, k_E) at $t = \infty$, assuming that the value of $q(t)$ is bounded, with $q(\infty) = q_E$. Note that we do not need in this case to leave the singular path to meet the final boundary conditions because

$$\lambda(\infty) = \gamma \exp(-\rho\infty) = 0.$$

Therefore if we consider an infinite horizon, the terminal point will actually belong to the singular arc. In order to get onto the singular arc we must first undertake capacity construction at a rate M during $\overline{AD}$. To determine the point D we can use the relation

$$k_D - k_A = M\,(t_D - t_o), \qquad k_A = k(0)$$

and the equation of the singular arc evaluated at D:

$$k_D = \theta q_D$$

where θ is given by (5-15).

Case 2. (B-C-E)

arc $\overline{BC}$: $H_V^* < 0$, $V^* = 0$ (bang-bang)

arc $\overline{CE}$: $H_V^* = 0$, $V_s^* = k(\dot{q}/q)$. (singular)

Here we have assumed the same terminal condition $q(\infty) = q_E$, but we began with initial conditions (k_B, q_B) that correspond to a point above the singular path. Therefore, to get onto the singular arc we must wait until demand increases to q_C holding capacity constant during the period $\overline{BC}$. We can determine the point C in a way similar to that used for D in Case 1.

Case 3. (A-D-F-G)

arc $\overline{AD}$: $H_V^* > 0$, $V^* = M$ (bang-bang)

arc $\overline{DF}$: $H_V^* = 0$, $V_s^* = k(\dot{q}/q)$ (singular)

arc $\overline{FG}$: $H_V^* < 0$, $V^* = 0$ (bang-bang)

This case has the same initial conditions A as Case 1 but it corresponds to a finite planning horizon $[0,T]$ with $q(T) = q_G$ and we in

addition assume that the final boundary condition for λ (or the transversality condition) is such that

$$\lambda(T) = \lambda_G < \gamma \exp(-\rho T).$$

Therefore we must leave the singular arc at a point F in order to meet the final conditions. The point of exit F is given in this case by

$$\lambda(t_F) = \gamma \exp(-\rho t_F)$$

which upon using (5-9) and the final value λ_T, can be written, by directly integrating equation (5-9) as

$$\int_{t_F}^{T} -C_k q \exp(-\rho t)\, dt + \lambda_T = \gamma \exp(-\rho t_F) .$$

Substituting the value of C_k from (5-14) and using the fact that capacity is constant after F, this expression becomes

$$n\beta K_F^{-(n+1)} \int_{t_F}^{T} q^{n+1} \exp(-\rho t) dt + \lambda_T = \exp(-\rho t_F) .$$

Using this equation and the equation of the singular arc evaluated at F:

$$k_F = \theta q_F$$

where θ is defined in (5-2), we can get both k_F and t_F.

5.2 Decreasing Returns to Scale in Capacity Construction.

Here we will assume a capacity production function of the form

$$f = \gamma' k + \delta k^m \ , \ m > 1 \ . \tag{5-16}$$

The associated derivatives of interest are

$$\begin{aligned} f_k &= \gamma' + \delta m \ k^{m-1} \\ f_{kk} &= \delta m(m-1)k^{m-2} \\ f_{kkk} &= \delta m(m-1)(m-2) \ k^{m-3} \ . \end{aligned} \tag{5-17}$$

The linear term $\gamma' K$ was included in (5-16) in order to allow marginal costs different than zero for $k = 0$. This will permit more general and natural comparisons with the constant return case. Using (5-17) in (3-2) and (3-5) we can obtain

$$H_v = \lambda - (\gamma' + \delta m \ k^{m-1}) \exp(-\rho t) \tag{5-18}$$

$$\dot{\lambda} = (C_k q + \delta m(m-1) \ k^{m-2} V) \exp(-\rho t) \ . \tag{5-19}$$

From (5-4) we have that the expression for optimum singular controls, when $C(k,q)$ is homogeneous of degree zero, is

$$V_s^* = \dot{q} \ \{C_{kk} k/(C_{kk} q + \rho f_{kk})\} \ .$$

Thus, using (5-14) and (5-17) to replace C_{kk} and f_{kk} we get the following explicit expression for V_s^* in terms of k,q and the parameters of the problem

$$V_s^* = k\ (\dot{q}/q)\mu \tag{5-20}$$

with $$\mu = n(n+1)\ \beta/[n(n+1)\beta + m(m-1)\ \rho\delta\ (k^{m+n}/q^{n+1})] \quad . \tag{5-21}$$

where it is obvious that μ will be lower than one for all m greater than one.

Using expression (5-5), the equation of a singular arc when C(k,q) is homogenous of degree zero, and the expressions for C_q and f_k from (5-14) and (5-17), we can obtain the following equation for the singular arc

$$n\beta q^{n+1} - \rho\gamma' k^{n+1} - \rho\delta\ mk^{m+n} = 0 \tag{5-22}$$

which is a curve going through the origin of (q,k). Unfortunately we cannot obtain for this case an explicit expression k as in the constant returns case. An explicit expression is obtained only if we eliminate the linear term of f in (5-16), $\gamma' = 0$; in that case we have

$$k = \bar{\theta}\ q^{(n+1)/(n+m)} \tag{5-23}$$

with

$$\bar{\theta} = (n\beta/m\rho\delta)^{(n+m)^{-1}} \quad .$$

Note that for $m > 1$ (5-23) describes a concave function of q going through the origin of (k,q).

The value of μ can be expressed as a function of k and the parameters of the problem if we use (5-22) to eliminate q from (5-21); in that case we obtain

$$\mu = n(n+1)\beta/[n(n+1)\beta + m(m-1)\rho\delta n\beta(\rho\gamma' k^{-(m-1)} + \rho\delta m)^{-1}] . \quad (5\text{-}24)$$

Therefore we can say that the value of μ will decrease as k increases, and therefore q increases (see 5-22), along the singular arc. Now, from (5-20) and the fact that μ will always be lower than one for $m > 1$ we can write, for all t on the singular arc:

$$(\dot{k}/k) < (\dot{q}/q) \quad \text{or} \quad (dk/k) < (dq/q) \quad (5\text{-}25)$$

Using the first of these relations in (5-2) we obtain that $\dot{C} > 0$, which means that for the decreasing returns to scale case the optimum investments in capacity along a singular arc will produce an increasing operating cost function through time as long as $q > 0$. From the second relation in (5-25) we get that the optimum percentage increase of capacity per unit of time will be lower than the corresponding percentage increase of demand. In addition, from the fact that μ decreases with q, the difference will increase as q increases and

therefore the singular arc will be concave in q as shown in Figure 4.7a From (5-22) it is obvious that it will go through the origin of (q,k).

In order to obtain an expression for (dk/dq) we can differentiate (5-22) with respect to q to obtain

$$(dk/dq) = n(n+1)\beta q^n/[\rho\gamma'(n+1)k^n + \rho\delta m(m+n)k^{(m+n-1)}] . \quad (5\text{-}26)$$

Using (5-22) again to eliminate q, and after some simple algebraic manipulations we obtain

$$\frac{dk}{dq} = \frac{n(n+1)\beta[\rho\gamma'/n\beta + (\rho\delta m/n\beta)k^{m-1}]^{n/(n+1)}}{\rho\gamma'(n+1) + \rho\delta m(m+n)k^{m-1}} \quad (5\text{-}27)$$

which is always positive. If $m > 1$ we will have, for $k = 0$,

$$(dk/dq)_{k=0} = \theta' = (n\beta/\rho\gamma')^{(n+1)^{-1}} . \quad (5\text{-}28)$$

As we saw before (dk/dq) decreases as q increases. If γ' in (5-16) has the same value as γ in (5-7) we have that the initial marginal cost of capacity, at $k = 0$, for the decreasing returns case is equal to the constant marginal cost corresponding to the constant returns case. The production cost functions are as shown in Figure 4.6. Then from (5-15) and (5-28) we have that the singular arc, for the constant returns case, coincides with the tangent to the singular arc corresponding to the decreasing returns case, at the origin. In other words $\theta' = \theta$, as it is shown in Figure 4.7a. The singular arc for

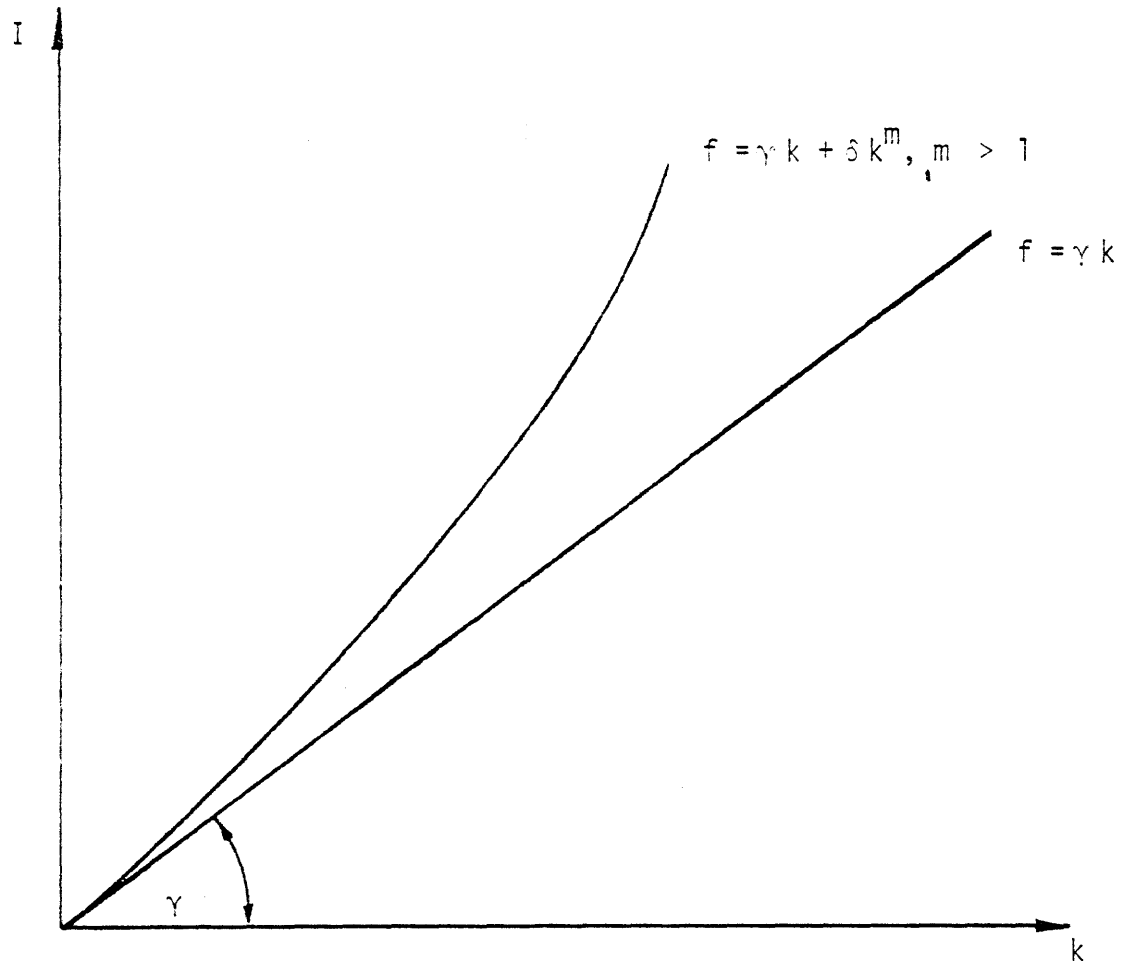
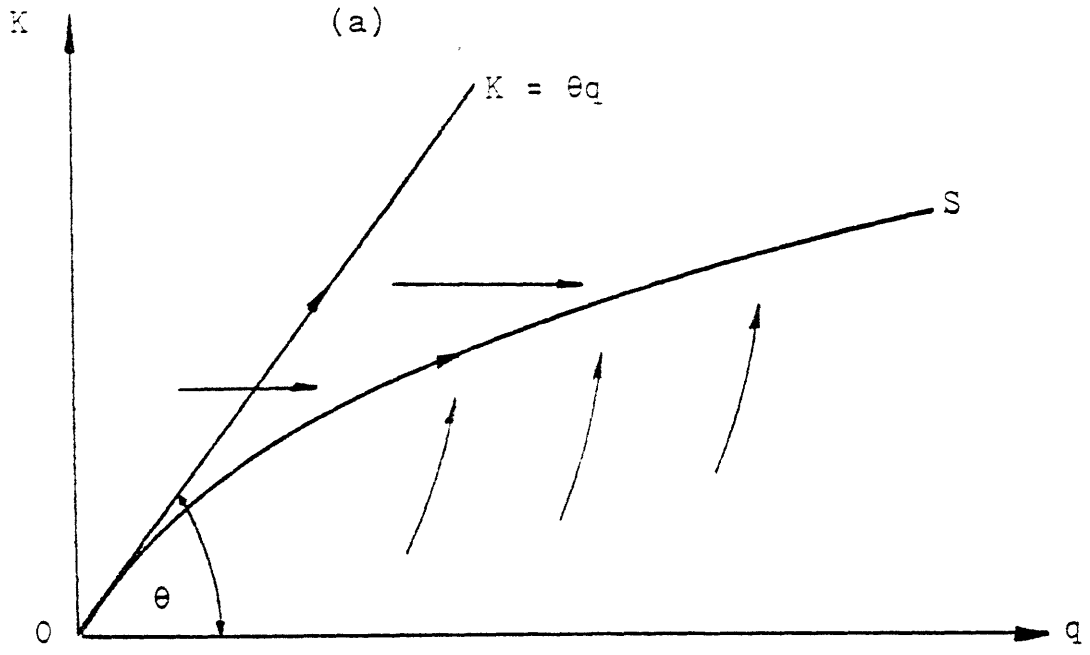


FIGURE 4.6 Capacity Production Costs Function Decreasing Returns Case



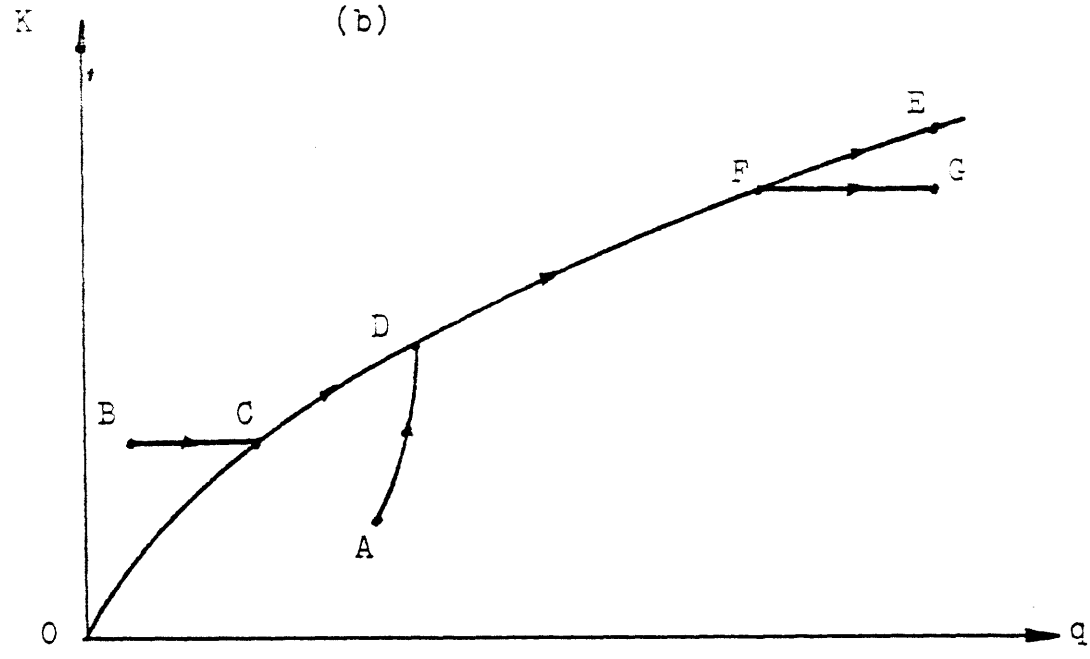


FIGURE 4.7 Optimal Policies. Decreasing Returns to Scale Case.
(a) Singular arc,
(b) Optimal Policies.

the decreasing returns case, that is represented by the arc $\overline{OS}$, lies below the straight line $k = \theta q$ for all q and is concave. Therefore, the optimum amount of capacity for the decreasing returns case, is lower, for all q, than the corresponding optimum amount for the constant returns case. This difference increases with q, for a given m, and also increases with m for each q. Some typical examples of singular policies, similar to those analyzed for the constant returns case, are depicted in Figure 4.7b.

It is easy to show that the second order condition (3-23) for singular arcs is satisfied for all $m > 1$ for the present circumstances. Nevertheless, the convexity condition (4-7) for bang-bang arcs will be satisfied for any values of the variables and parameters of the problem only if $m > 2$. If $1 < m < 2$, then (4-7) will impose additional conditions to those already analyzed. If we substitute into (4-7) the expressions of C_{kk} and f_{kkk} from (5-14) and (5-17) we obtain

$$\gamma' m(m-1)(2-m)k^{m-3} \leq [n(n+1)\beta/M](q^{n+1}/k^{n+2}), \qquad (5\text{-}29)$$

where $0 < (2-m) < 1$.

As we saw in Section 4. , this expression can be seen as a condition for M if we write it as

$$M \leq [n(n+1)\beta/\gamma' m(m-1)(2-m)](q^{n+1}/k^{m+n-1}), \qquad (5\text{-}30)$$

which gives an upper bound for the value of M, and therefore $V^*(t)$,

for each point (q,k) situated below the singular arc. If the value of M violates condition (5-30) for a point below the singular arc, then the optimum investment policy at that point becomes (see Section 4):

$$V^* = [n(n+)\beta/\gamma' m(m-1)(2-m)](q^{n+1}/k^{m+n-1}); \qquad (5\text{-}31)$$

this policy should be maintained until the singular arc is reached. Expressions (5-30) and (5-31) are also applicable to the singular control if we replace M by V_s.

5.3 Increasing Returns to Scale Case

This case is largely symmetrical to the one analyzed in the preceding section. Nevertheless, the second order conditions are different. We will simply assume here that the function f is given by

$$f = \varepsilon\, k^m,\ 0 < m < 1 \qquad (5\text{-}32)$$

$$\begin{aligned} f_k &= \varepsilon\, m\, k^{m-1} \\ f_{kk} &= \varepsilon\, m(m-1)k^{m-2} \\ f_{kkk} &= \varepsilon\, m(m-1)(m-2)\, k^{m-3} \end{aligned} \qquad (5\text{-}33)$$

We can find the equation for the singular arc using expressions (5-5), (5-14) and (5-33) to obtain after simple algebraic manipulations:

$$k = \tau q^{(n+1)/(n+m)}, \quad \text{with } \tau = (n\beta/\rho\varepsilon m)^{(n+m)^{-1}} \tag{5-34}$$

which represents a convex function of q going through the origin of (q,k). A representation of (5-34) is given in Figure4.8a by the arc $\overline{OS}$.

The expression for the singular controls is obtained from (5-4), (5-14) and (5-33). In fact, we find that

$$V_s^* = k(\dot{q}/q)\mu' \tag{5-35}$$

$$\text{with } \mu' = n(n+1)\beta/[n(n+1)\beta + \rho\varepsilon m(m-1)\ (k^{m+n}/q^{n+1})] \tag{5-36}$$

If we now use (5-34) to eliminate k and q from (5-36) we can obtain the following expression for μ' as a function only of the degrees of the functions C and f:

$$\mu' = (n+1)/(n+m). \tag{5-37}$$

Here it is obvious that μ'will be positive and higher than one for all m between zero and one. In addition μ' will increase as m decreases or, in other words, the economies of scale increase. Therefore, we can write for all points along the singular arc:

$$(\dot{k}/k) > (\dot{q}/q) \quad \text{or} \quad (dk/k) > (dq/q) \tag{5-38}$$

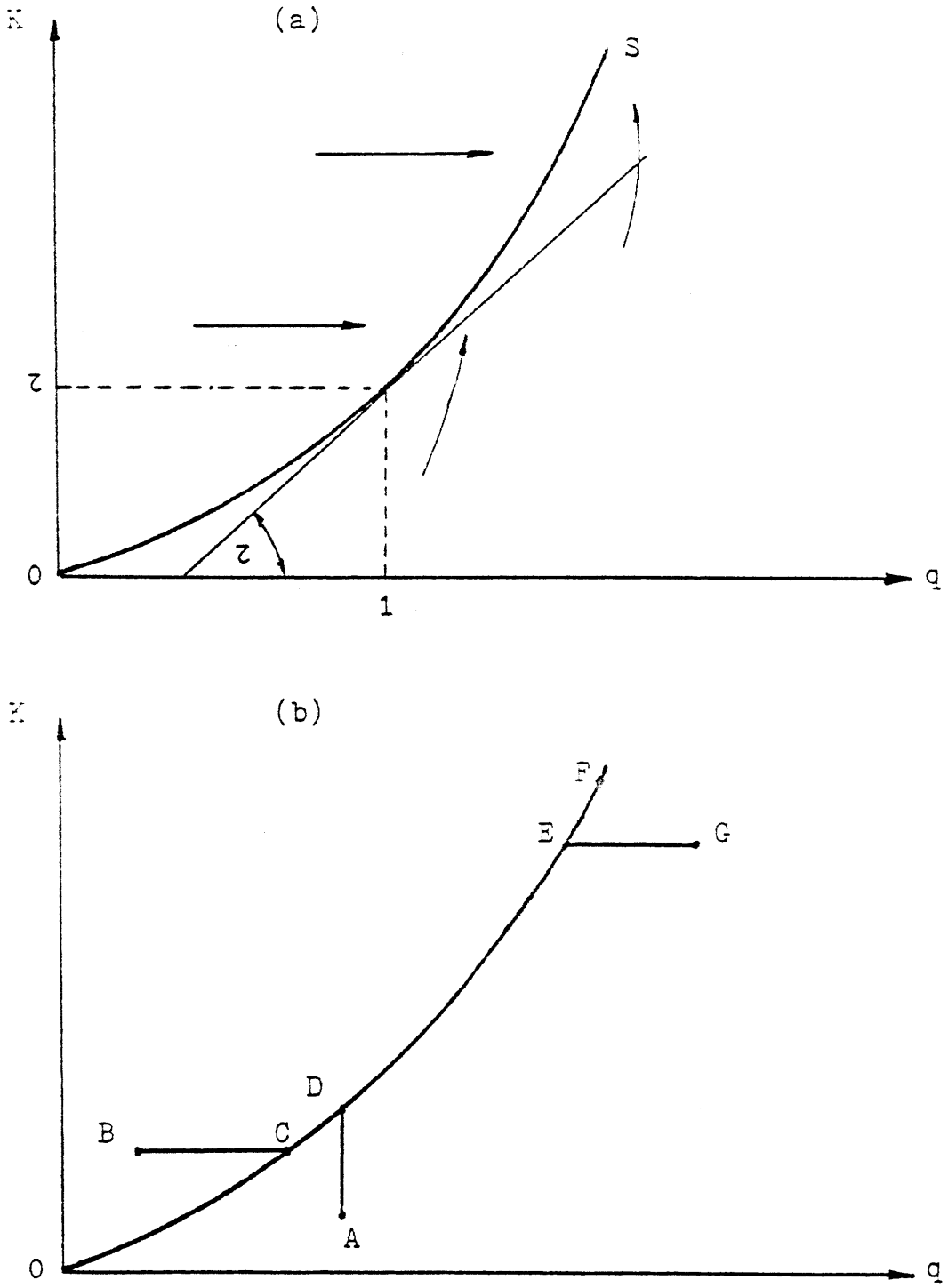


FIGURE 4.8 Optimal Policies. Increasing Returns to Scale Case.
(a) Singular arc,
(b) Optimal Policies.

From the first of these relations we obtain, using (5-2) that $\dot{C} < 0$. Therefore, for the increasing returns to scale case, the optimum investments in capacity, along a singular arc, will produce a decreasing operating cost function through time, as long as $\dot{q} > 0$. From the second expression in (5-38) we deduce that the optimum percentage increase of capacity, per unit of time, will be higher than the corresponding percentage increase of demand. Some typical singular policies are presented in Figure 4.8b.

It is easy to check in this case that the sufficient conditions (4-7) and (4-11) for bang-bang arcs with $V^* = M$, and singular arcs respectively are satisfied for all possible combinations of positive values of q and k. This is because $f_{kkk} > 0$, as we can check in (5-33). The second order necessary condition (3-23) for singular arcs, can be expressed, using (5-14) and (5-33) as

$$n(n+1)\beta(q^{n+1}/k^{n+2}) + \rho\varepsilon\, m(m-1)k^{m-2} > 0. \qquad (5\text{-}39)$$

Upon rearranging terms this can be written as

$$k < \eta q^{(n+1)/(n+m)}$$

with

$$\eta = \{[(n+1)/(1-m)]\ (n\beta/\rho\varepsilon m)\}^{(n+m)^{-1}} \qquad (5\text{-}40)$$

Given that the value of m must be between zero and one, it is obvious that η will be higher than τ in (5-34). Therefore, condition (5-40) will be satisfied for any point along the singular arc.

V. OPTIMAL INVESTMENTS IN CAPACITY AND QUALITY. DISCRETE CASE

1. Introduction

As was mentioned in the preceding chapter, investments in transportation facilities can present important indivisibilities. This is especially relevant for the capacity variable, feasible values of which are usually contained in a finite countable set. Thus, the capacity of a road can only be increased adding new lanes, the capacity of a port adding new loading sites, etc., where the feasible increases are obviously discrete. Therefore, the decisions about capacity provision for individual transportation facilities are in practice transformed into time staging decisions. In that case, different discrete levels of capacity are defined and the problem is to find the optimum time at which a jump from one level of capacity to the next one must be made. In practice, a given level of quality is also associated with each level of capacity, which defines a set of feasible states for the facility. The jump from one state to another defines a project whose implenentation time must be decided.

In this chapter we will present a mathematical model that treats capacity as a discrete variable, feasible levels of which are externally specified, and also allows for discrete jumps in quality at the times at which capacity is changed. Nevertheless, we will consider quality as a piece wise continuous function which value is internally determined by the model at all times t with $t \in [0,T]$. The necessary conditions for optimality will be developed and from them the characteristics of optimum policies and their economic interpretation will be derived.

2. A Mathematical Model for Optimal Staging of Capacity and Quality

As before we will assume that we have a transportation facility with homogeneous users, who perceive a utility U(t) from each trip performed over the facility. In addition, let us assume that N possible levels of capacity have been defined for the facility, each of them associated with an operating cost function $C_i(s,q)$ like that defined in Chapter III. The same assumptions made there about the characteristics of this cost function and the quality variable s are valid here.

Therefore, the net private benefit obtained by an individual user from each trip performed over the facility, when the capacity is at level i is

$$B_i(t) = U(t) - C_i(s(t),q(t)) \quad ,i = 1....N \qquad (2\text{-}1)$$

Let us use the notation t_i (i = 0, 1,....,N-1) to refer to the times at which an increase of capacity is provided and t_i^- and t_i^+ to refer to the moments just before and after t_i. We will call V(t) the amount of money spent on maintenance* at each time t with $t \in [t_i^+, t_{i+1}^-]$ (i = 0,1,....,N-1) and $I_i(s(t_i^+),s(t_i^-))$ the investment necessary, at time t_i, to go from a capacity level i to a capacity level (i + 1).

---

* In this chapter maintenance will refer to only those activities that decrease or eliminate deterioration of the facility.

This investment will be a function of the quality of the facility just before and after t_i. In other words, discrete jumps of quality are allowed at each time t_i at which capacity is increased. Then, the cost of the fixed increase in capacity will depend on the amount of quality improvement provided at the same time.

Our objective in this case will be to choose a set of times $\{t_i\}$ $(i = 1,\ldots,N-1)$, a function $V(t)$ for each interval $[t_i^+, t_{i+1}^-]$ and the values of $s(t_i^+)$ $(i = 1,\ldots,N-1)$ in such a way that the social net benefit will be maximized. We will assume that the initial time t_0 and the final time $t_N = \infty$ are given. Then, the objective function can be written as:

$$J(V_i;t_1,\ldots,t_{N-1};\ s_0^+,\ldots,s_{N-1}^+) =$$

$$\sum_{i=1}^{N} \int_{t_{(i-1)}^+}^{t_i^-} \{[U(t) - C_i(s(t),q(t))]q(t) - V(t)\}\exp(-\rho t)dt$$

$$- \sum_{i=0}^{N-1} I_i(s(t_i^+),s(t_i^-))\exp(-\rho t_i) \qquad ,(2\text{-}2)$$

In order to simplify notation we will define

$$\phi = - \sum_{i=0}^{N-1} I_i(s(t_i^+),s(t_i^-)\exp(-\rho t_i),$$

$$s_i^+ = s(t_i^+), \qquad s_i^- = s(t_i^-) \qquad ,(2\text{-}3)$$

$$I_{is+} = (\partial I_i/\partial s_i^+) \qquad I_{is-} = (\partial I_i/\partial s_i^-)$$

As in Chapter 3 we will in addition assume that the change of quality in the facility, per unit of time, for all $t \neq t_i$ $(i = 0, 1,$

....,N) can be represented by differential equations of the form

$$\dot{s} = f_i(s(t),q(t),V(t),t),\ t \in [t_{i-1}^{+},t_i^{-}],\ (i = 1,\ldots,N), \quad (2\text{-}4)$$

and that the amount of money that can be spent in maintenance at each t is constrained by

$$m(t) \leqq V(t) \leqq M(t)\ . \quad (2\text{-}5)$$

3. Necessary Conditions for Optimality

If we consider that the demand for using the facility q(t) is given for each time t in the period $[t_o,t_N]$, then the maximization of the objective function (2-2), subject to the quality dynamics (2-4) and the control constraints (2-5) is an optimal control problem of the type presented in Section 3.3, of Chapter III. The Hamiltonian is in this case defined by:

$$H_i = \{[U - C_i(s,q)]q - V\}\exp(-\rho t) + \lambda f_i,$$
$$t \in [t_{i-1}^{+},t_i^{-}],\ (i = 1,\ldots,N) \quad (3\text{-}1)$$

and in order for a set $(V(t),t_1,\ldots,t_{N-1},\ s_o^{+},\ s_1^{+},\ldots,s_{N-1}^{+})^*$ to be optimal, the following necessary conditions must be satisfied.

$$\dot{\lambda} = -(\partial H_i/\partial s) = C_{is}q\exp(-\rho t) - \lambda f_{is},$$
$$t \in [t_{i-1}^{+}.t_i^{-}],\ (i = 1,\ldots N) \quad (3\text{-}2)$$

with:

$$\lambda(t_i^-) = (\partial\phi/\partial s(t_i^-)) = I_{is^-}\exp(-\rho t_i),\ (i = 1,\ldots N) \qquad (3\text{-}3)$$

$$\lambda(t_i^+) = (\partial\phi/\partial s(t_i^+)) = I_{is^+}\exp(-\rho t_i),\ (i = 1,\ldots N) \qquad (3\text{-}4)$$

and

$$(\partial\phi/\partial t_i) = H_{(i+1)}(t_i^+) - H_i(t_i^-),\ (i = 1,\ldots,N\text{-}1). \qquad (3\text{-}5)$$

By using the definitions of ϕ in (2-3) and H_i in (3-1) this becomes

$$\begin{aligned}\rho I_i\exp(-\rho t_i) = {} & [C_i(t_i^-) - C_{i+1}(t_i^+)]q(t_i)\exp(-\rho t_i) \\ & + [\lambda(t_i^+)f_{i+1}(t_i^+) - V(t_i^+)\exp(-\rho t_i)] \\ & - [\lambda(t_i^-)f_i(t_i^-) - V(t_i^-)\exp(-\rho t_i)], \\ & (1 = 1,\ldots.,N\text{-}1)\end{aligned} \qquad (3\text{-}6)$$

where we have assumed that the variable $q(t)$ is continuous for all $t \in [t_0,t_N]$ though its derivative with respect to time can be discontinuous at some points within the same interval.

Finally, the value of $V(t)$ within each interval (t_i^+,t_{i+1}^-) has to be chosen in such a way that the corresponding value of the Hamiltonian is maximized:

$$\begin{aligned}H_i(s^*,\lambda^*,V^*,t) \geq H_i(s^*,\lambda^*,V,t),\ \forall\ t \in [t_{i-1}^+, t_i^-] \\ \forall\ V \in \Omega \qquad (i = 1,\ldots.,N)\end{aligned} \qquad (3\text{-}7)$$

where Ω is defined as usual by (2-5).

For each interval $[t_{i-1}^{+}, t_i^{-}]$, the problem of selecting an optimum maintenance policy $V^*(t)$ corresponds to the case analyzed in Section 2 of Chapter III; therefore, V^* must satisfy (see expressions (2-13) to (2-15) of Chapter III the following:

$$\left.\begin{array}{l} \lambda\, f_{iV} = \exp(-\rho t), \text{ if } m < V^* < M, \\ \lambda\, f_{iV} \leqq \exp(-\rho t), \text{ if } V^* = m \\ \lambda\, f_{iV} \geqq \exp(-\rho t), \text{ if } V^* = M \end{array}\right\} \forall\, t \in [t_{i-1}^{+}, t_i^{-}]\,, \quad (i = 1,\ldots,N) \qquad (3\text{-}8)$$

4. Economic Interpretation. Optimal Investment Rules

Expressions (2-4), (2-5), (3-2) to (3-4), (3-6) and (3-8) constitute a complete set of necessary conditions for the problem defined in Section 2.

As in Section 2.2 of Chapter III we can integrate the adjoint equation (3-2) for each interval $[t_{i-1}^{+}, t_i^{-}]$, using the boundary value of $\lambda(t_i^{-})$ provided by the transversality condition (3-3), to obtain

$$\lambda(t) = -\int_t^{t_i^-} \{[C_{is}\exp(\int_t^x f_{is}\,dz)]q(x)\exp(-\rho x)\}dx + I_{is-}\exp(\int_t^{t_i^-} f_{is}\,dz)\exp(-\rho t_i) \qquad (4\text{-}1)$$

with

$$t \in [t_{i-1}^{+}, t_i^{-}], \quad x \in [t, t_i^{-}], \quad (i = 1,\ldots,N).$$

The first term in this expression for $\lambda(t)$ was interpreted in Section 2.2 of Chapter 3 as the present value, at time t, of the cost reductions experienced by all the users of the facility, during the period $[t,t_i^-]$, as a consequence of the implementation of one additional unit of quality at time t. The second term represents the reduction in construction costs at time t_i due to an additional unit of quality provided at time t. This second term appears because in the definition of I_i we have assumed that quality is an additive variable. In other words, the cost of providing a level of quality s_i^+, at time t_i^+, depends on the level of quality s_i^- at time t_i^-. In many practical situations this is not the case and the I_i depends only on the value of s_i^+. Then, $I_{is^-} = 0$ and the second term of expression (4-1) disappears. We will assume this latter case for our subsequent analysis. Therefore, as usual, $\lambda(t)$ is the shadow price of quality at time t.

Expression (3-4) can be used to determine the optimum level of quality s_i^+ to be provided at time t_i^+. Using (4-1) we can write

$$I_{(i-1)s^+}\exp(-\rho t_{i-1}) =$$

$$-\int_{t_{i-1}^+}^{t_i^-} \{[C_{is}\exp(\int_{t_{i-1}^+}^{x} f_{is}dz)]q(x)\exp(-\rho x)\}dx \quad , \quad (4\text{-}2)$$

where the left hand side is the present value of the marginal cost of quality for project (i-1) and the right hand side is equal to the present value of the marginal benefits of one unit of quality implemented at time t_{i-1}. This benefit is a consequence of the

operating cost reductions experienced by all the users of the facility during the period $[t_{i-1}^+, t_i^-]$, and its value will increase as the number of users in the period $q(x)$ increases or the length of the period increases. We will expect in general that decreasing returns exist in the production of quality (at least after a certain level of quality has been reached) and therefore, for a given construction date t_{i-1}, the left hand side of (4-2) will increase with s. On the other hand we expect that the marginal reductions of operating costs, produced by quality increases, C_{is}, will be constant or non-increasing with s. As a consequence, the right hand side of (4-2) will be non-increasing with s. In that case, the net marginal benefit per unit of quality implemented at t_{i-1}^+ will be decreasing with s and relation (4-2) will lead to a value of s_{i-1}^+ that maximizes our objective function J in (2-2).

The necessary conditions (3-6) give us N-1 equations for the determination of the interior upgrading times t_i. Using our assumption that I_i is independent of the quality of the facility at t_i^-, which leads to $\lambda(t_i^-) = 0$ due to (3-3) and $V^*(t_i^-) = m(t_i^-)$ due to (3-8), we can rewrite (3-6) as

$$\rho I_i + [V^*(t_i^+) - m(t_i^-)] - \lambda^*(t_i^+) f_{i+1}^*(t_i^+) \exp(\rho t_i) = [C_i(t_i^-) - C_{i+1}(t_i^+)] q(t_i) \qquad (4\text{-}3)$$

The first term on the left hand side of this expression is the rental value of investment I_i and the second term is the difference between the optimal amounts spent in maintenance, per unit of time,

just after and before the investment I_i is made. Note that if an optimum maintenance policy is performed within each interval between investments, this term will be positive or zero. The third term is the social value of the optimal deterioration of the upgraded facility at time t_i^+ (just after the investment is made). The value of $\lambda^*(t_i^+)$ is here given by

$$\lambda^*(t_i^+) = - \int_{t_i^+}^{t_{i+1}^-} \{[C^*_{(i+1)s} \exp(\int_{t_i^+}^{x} f^*_{(i+1)s} dz)] q(x) \exp(-\rho t)\} dx, \quad (4\text{-}4)$$

where the "*" in this case means that the optimum maintenance policy specified by conditions (3-8) is performed during the period $[t_i^+, t_{i+1}^-]$. Given that we are assuming that only maintenance activities (not improvements of quality) are performed during each period between investments, $f^*_{i+1}(t_i^+)$ will always be negative or zero and therefore the entire third term will be negative or zero, because $\lambda^*(t_i^+)$ is in this case always positive. The actualization factor $\exp(\rho t_i)$ transforms the value of $\lambda^*(t_i^+)$, that is expressed in present value at $t = t_0$, to current value at $t = t_i$. All other terms in equation (4-3) are also expressed in current value at t_i. Therefore, the left hand side of (4-3) is the amount that we save if we postpone our investment I_i by one unit of time, or in other words, it is the marginal benefit of postponement per unit of time.

The right hand side of (4-3) is equal to the difference between the total operating costs of facility users, just before and after the investment I_i is made. If the investment I_i increases both the quality and the capacity of the facility, this term will obviously be

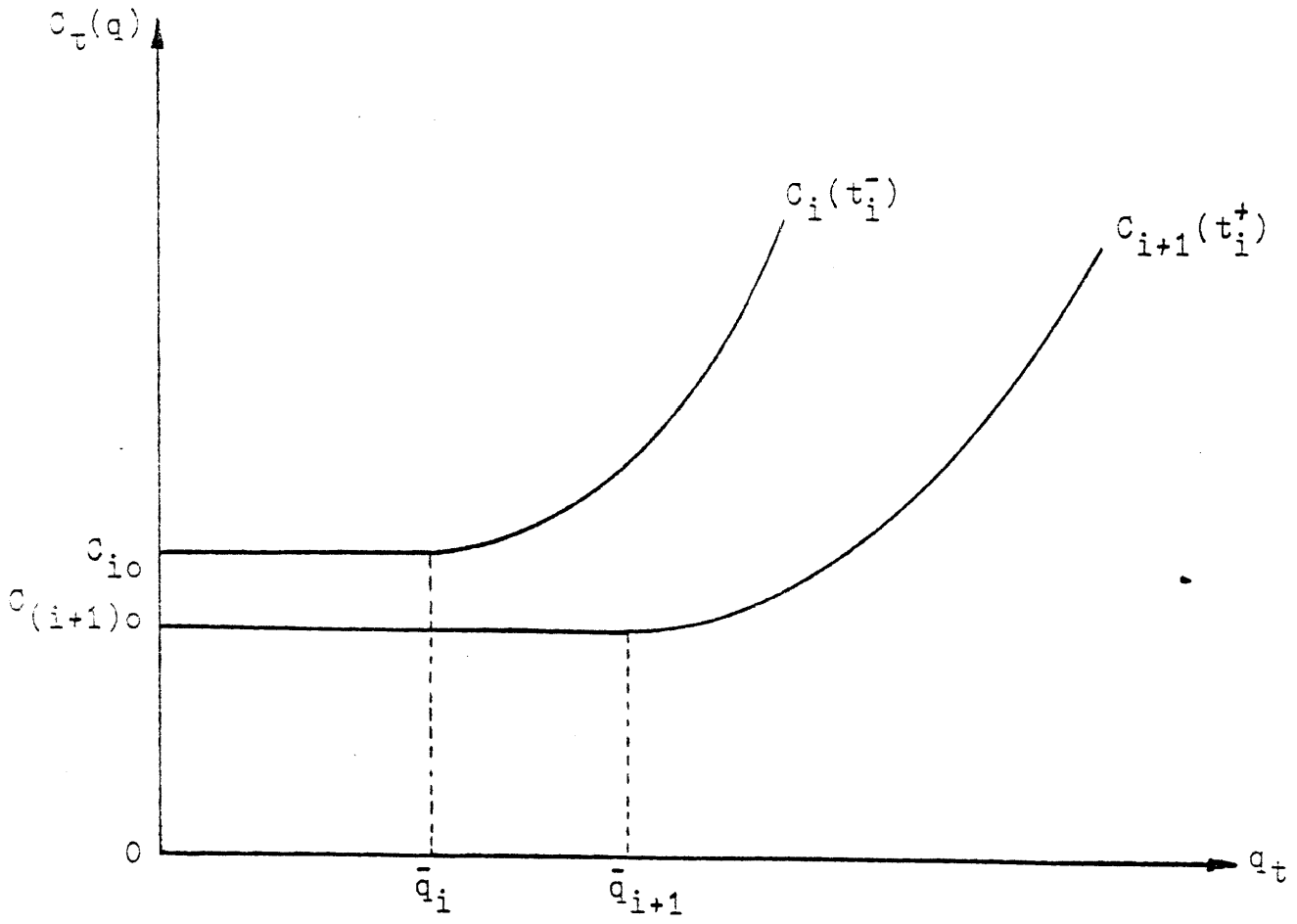


FIGURE 5.1 Operating Cost Function, before and after the discrete investment is made.

positive for any positive value of q(t) (see Figure 5.1) In this case, the free flow cost after the investment, $C_{(i+1)o}$, will be lower than before the investment, C_{io}, given the increase in quality, and the congestion costs C_q will also decrease, given the increase of capacity. Therefore, the difference between operating costs in (4-3) will be positive and increasing with q. If the quality of the facility were the same just before and after the investment, then the free flow costs would be the same in both cases and a difference between operating costs would only appear when congestion appears. On the other hand, if congestion does not exist over the facility and free flow conditions prevail, the difference between unit operating costs will be constant with respect to q ($q < \bar{q}_i$, in Figure 5.1) but still the difference between total operating costs will increase with q if $C_{io} > C_{(i+1)o}$. Thus, the right hand side of (4-3) corresponds to the total savings in operating costs that would be obtained per unit of time if the investment I_i were undertaken. In other words, it is equal to the marginal cost derived from the postponement of the investment decision per unit of time.

Then, we can state rule (4-3) as follows: the optimum time t_i^*, for upgrading the facility, comes when the marginal benefit of postponement per unit of time is equal to the marginal cost of postponement per unit of time. This statement corresponds to Marglin's navie rule (Marglin, [1967]) for dynamic investments. It already appeared in Chapter IV, underlying the optimum investment policies in continuous capacity along a singular arc. Remember that the corres-

ponding optimal policy was interpreted there (see 3-19 in Chapter IV) as the same rule applied to the marginal unit of capacity.

We make note that rule (4-3) does not apply to the determination of time t_o at which the initial investment I_o must be undertaken, because $C_o(t_o^-)$ and $m(t_o^-)$ do not exist. For that case, using (3-5) and the fact that $H_o \equiv 0$, we obtain

$$\rho I_o + V^*(t_o^+) - \lambda^*(t_o^+) f_1^*(t_o^+) \exp(\rho t_o) = [U(t_o^+) - C_1(t_o^+) q(t_o) \quad . \tag{4-5}$$

The main difference in this case is that $U(t_o^+)$ replaces to $C_i(t_i^-)$ of equation (4-3). Therefore, we need an estimation of the utility $U(t_o^+)$ in order to determine the optimum value t_o^*.

Both (4-3) and (4-5) are marginal rules saying basically that the investment should be undertaken when the net marginal social benefit of postponement becomes zero. In order for this rule to lead to a maximum value of the objective function, and not a minimum, we need that a second order condition be satisfied. This second order condition is obviously that the above mentioned net marginal benefit must be monotonically non-increasing with t_i. It is easy to see that this will be in general the case if $q(t\ )$ increases with t .

Finally, if we assume that the values of t_i^* and s_i^{+*} have been already determined, the problem of finding optimal maintenance policies V^* for each period $[t_{i-1}^+, t_i^-]$ is identical to that analyzed in Section 2 of Chapter 3, as it is shown by the necessary conditions (3-8).

5. Numerical Solutions and Special Cases

The necessary conditions developed in the preceding section can be used to obtain numerical solutions simultaneously for the set of unknowns $(V(t),t_1,\ldots,t_{N-1}, s_0^+,s_1^+,\ldots,s_{N-1}^+)^*$. Nevertheless, finding solutions to such problems can be quite involved (see Bryson and Ho, [1975]). The difficulty of the problem comes from the fact that the optimum values of $V(t),t_i$ and s_i^+ are all interrelated. For instance, we can see from (4-3) that the optimum values of t_i^* depend on the quality s_i^+ (because $I_i,C_{i+1}(t_i^+)$ and $f_{i+1}(t_i^+)$ are all functions of s_i^+)and the optimum maintenance policy $V^*(t_i^+)$. But from (4-2) we have that the optimum value of s_i^+ will in turn depend on the value of t_i^*, and from (3-8) and (4-1) we can see that the value of $V^*(t)$ will also depend on the values of t_i^*. Obviously, the problem will increase in complexity as the number of investments that we want to analyze increases.

If we make assumptions that break the above mentioned interrelations important simplifications can be obtained. For instance, if we assume that for each investment I_i not only the increase in capacity, but also the quality s_i^+ is externally determined and in addition a maintenance policy $\hat{V}_i(t)$ is specified for each interval, such that the deterioration f_{i+1} will be zero or negligible in terms of its influence on the operating costs, then (4-3) becomes a pure staging rule:

$$\rho I_i + \hat{V}_{i+1}(t_i^+) - \hat{V}_i(t_i^-) = [C_i(t_i^-) - C_{i+1}(t_i^+)]q(t_i) \ . \qquad (5\text{-}1)$$

This simple rule has the great attractiveness that it only depends on the known functions $\hat{V}_i$, C_i and the observable quantity $q(t_i)$. Thus, for its application no predictions about the future values of the variables involved are needed at all if we make the weak assumption that $q(t_i)$ is not decreasing with time. We can actually observe the value taken at each time t by the independent variable q(t) and calculate the values involved in (5-1). The decision to undertake the investment I_i should be made when a value of q(t) is observed such that (5-1) holds. Of course, if we have an estimation of the values that q(t) will take throughout the whole planning period $[t_0,t_N]$, then we can calculate all the values t_i^* $(i = 1,\ldots,N-1)$ at time t_0 by means of (5-1).

Now, if we assume that we know the length of the life interval $[t_{i-1}^+,t_i^-]$ for project I_{i-1} and the maintenance policy is externally specified, we can easily calculate the optimum value of s_i^+ by means of (4-2), given that we have an estimate for the demand q(t) throughout the period $[t_{i-1}^+,t_i^-]$. Actually if C_{is} and f_{is} are constants independent of the value of s, a fact which implies that the operating cost reductions per unit of quality are independent of the level of quality and that the natural deterioration per unit of quality is fixed, then it is not necessary to assume that the maintenance policy is externally specified, because in that case the optimum value of s_i^+ is independent of V(t) (the right hand side of equation (4-2) can be integrated without knowing the values that s takes inside the interval $[t_{i-1}^+,t_i^-]$). The values of s_i^{+*} and $V^*(t)$ can be calculated sequencially.

Of course, interrelationships between t_i^* and s_i^{+*} can be heuristically considered if we begin with a set of values of s_i^+, calculate the set $\{t_i\}$ and go back to check if the initial values assumed for s_i^+ are optimum. The values of s_i^+ obtained from the second step, if different than the initial ones could then be used to reinitiate the process. An iterative procedure of this type could eventually converge to the optimum values that we are seeking, although there is no guarantee that such would occur.

The model formulation analyzed in this chapter suffers from two main limitations. The first one is a consequence of the discrete characteristic of the capacity investments. If economies of scale exist in the construction of capacity it could be possible that a sequence $\{I_i\}$ that groups together some intermediate investments could be better than another one that considers the maximum disaggregation possible. Thus, the savings obtained, in terms of construction costs, when going directly from capacity level i to capacity level i + 2, instead of passing through an intermediate stage i + 1, could more than compensate the reduced fitness of capacity to acutal flow. The problem is that given the absolute discretness of the capacity variable, the economies of scale charateristics of capacity construction cannot be built into the model. The model can give an answer if a specific sequence of investments is externally proposed, but it cannot internally determine the optimal sequence. If we want to find such an optimal sequence, the model should be applied to all possible sequences and the results be compared.

The second limitation is that in practice the demand for the

use of the facility can be a function of the operational characteristics of the facility. If that is the case, $q(t)$ cannot be externally estimated but must rather be internally determined by the model. An additional dynamic equation describing the relationships between demand and facility characteristics should be introduced. This relationship is studied in a limited way in the next chapter where we analyze the influence of such considerations on pure time staging optimal rules.

VI. INFLUENCE OF DEMAND-QUALITY INTERRELATIONSHIPS ON OPTIMAL POLICIES OF STAGE CONSTRUCTION FOR TRANSPORTATION FACILITIES

1.0 Introduction

The need for dynamic investment strategies in transportation is perhaps clearest and best illustrated in the case of developing countries. Developing countries are in general characterized by poorly developed intrastructures. This is particularly true of the transportation sector, where sparsely connected networks with numerous links providing sub-standard levels of service are frequently encountered. Furthermore, it is often argued that in such circumstances investments in transportation, especially highways, are required in order to foster the growth of hinterlands and bring about inter-regional equity. These attitudes not withstanding, it is not at all clear what the level of service, or as we shall call it, the quality, of transportation facilities should be and how discontinuous changes in quality, such as the replacement or upgrading of highways, should be staged over time.

Since transportation volumes in underdeveloped countries are relatively low on inter-city links and rural roads and are only expected to grow as a consequence of the development process, it is generally not advisable to implement high volume, high quality facilities immediately; this conventional wisdom is further underscored by the fact that there is usually an overall scarcity of public investment funds in such developing nations. Because of these facts multiple stage development policies have long been advocated for underdeveloped countries.

Marglin (1967) in a classic study of public investment policy appears to be one of the first to deal with the type of dynamic investment problems described above in a general way. His "naive static rule"

states that "the optimal contruction date, t_o, of a project arrives when the present value of a perpetual stream of benefits, at the instantaneous rate corresponding to t_o, equals the construction cost of the project for the first time." Beenhakker and Daskin (1973), though apparently unaware of the work of Marglin, used the naive static rule to derive time staging formulas for transportation facilities, corresponding to different assumptions with respect to the way in which demand increases as a function of time. De Neufville (1969) used dynamic programming to solve basically the same problem. Finally, Venezia (1977), also using a dynamic programming approach, derived staging decision rules for the case of uncertain demands. Venezia's main result may be considered a generalization of Marglin's naive rule for stochastic demands. All of the aforementioned efforts make the rather strong assumption that the transportation demand is independent of the quality of the facility provided, and in some cases, e.g. Venezia (1977), only the uncongested case is analyzed.

The interrelationship between transportation level of service, or quality, and the development of socio-economic activities has been widely recognized; see e.g. Manheim (1978). In the case of developing countries they can be especially important, as the numerical example presented later in this chapter dramatically illustrates. Such interrelationships imply that present levels of service (quality) not only affect the benefits and costs accruing to current users, but also influence the characteristics of future demand through their effect on the location and development of new activities within the area of influence of the facility considered.

The primary objective of this chapter is to investigate the influence of the previously mentioned demand-quality interrelationships on time staging decision rules for transportation facilities. To this end the case of upgrading a road will be considered. We will assume that a lump investment which improves the quality and/or capacity, of the road in a discrete fashion has been externally defined and the main question to be answered is when should the investment be undertaken. In order to analyze this problem a dynamic optimization model will be stated and solved using the results of optimal control theory, presented in Section 3.3 of Chapter II.

2.0 The Model

In this section an optimum decision rule for the time staging problem with respect to upgrading a road will be derived. We assume that our system can be adequately described using two variables, the quality of the road s(t), which will be considered here as a control variable, and transportation demand q(t), considered as a state variable. The variable s(t) could in practice be represented by an index that takes into account the different factors that determine the quality of a road from the point of view of users, such as total length, width, alignment, type of surfact, etc. (practical ways of handling the variable s in real applications are proposed in section 5). The variable q(t) corresponds to the number of users per unit of time.

In order to represent the dynamic interrelationship between the two variables s(t) and q(t) the following linear differential equation is utilized:

$$\dot{q}(t) = a(t)s(t) + b(t) \; ; \; q(o) = q_o \qquad (2\text{-}1)$$

Where a(t) is the rate of change in demand, at time t, per unit of quality; and b(t) represents those effects which are independent of the quality of the facility and which influence the rate of change of demand. Thus, we are assuming that transportation demand will be determined by the interaction of our control variable s(t) with the time-varying parameter a(t) and some external factors, outside our control, represented by the non-service rate of change of demand, or "natural" rate of growth of demand, b(t). In general s(t) will be a function of different factors such as the natural rate of deterioration of the road, intensity of use, maintenance policy and discrete investments (see Chapters III and VII). However, for the purpose of this chapter, it will suffice to assume that within each of the two stages considered, before and after upgrading of the road, the quality will be constant and equal to that existing at the beginning of each stage. This perspective assumes that the appropriate maintenance policy has been performed during each stage to ensure the constant quality level assumed to exist throughout the stage. Thus our control variable must satisfy the following constraint:

$$s(t) \in S = \{s(t): s(t) = s_1, \forall\, t \in [o, t^{*-}];$$

$$s(t) = s_2, \forall\, t \in [t^{*+}, T]\} \qquad (2\text{-}2)$$

Where t* is the time at which the road is upgraded, T is the fixed terminal time of the planning period, and the s_i are constant road

qualities during the i^{th} stage. We consider, for expository purposes, only two stages: stage 1 corresponding to the time interval $[0,t^{*-})$ and stage 2 corresponding to the interval $(t^{*-}, T]$. The arguments given below may be generalized to any finite number of such stages. It should be noted that (2-2) defines a functional form for our control variable, information which will play a key role in the analysis which follows.

We are interested in choosing t^* such that the present value of the net benefits produced during the planning period $[0,T]$ is maximized. Consequently, we denote our objective as:

$$\text{Maximize } J = \int_0^{t^{*-}} [U - C_1(s(t),q(t))]q(t) \exp(-\rho t)dt$$

$$-I(t^*)\exp(-\rho t^*)$$

$$+ \int_{t^{*+}}^{T} [U - C_2(s(t),q(t))]q(t)\exp(-\rho t)dt$$

$$+ \Psi(q(T),T) \qquad (2\text{-}3)$$

Where U is the average user utility per trip which will in general be a function of t; $C_i(s,q)$ is the average cost of operating a vehicle over the road during stage i, which we consider a function of the quality of the road and the number of users; $I(t^*)$ is the amount of resources necessary to upgrade the road at time t^*; $\Psi(q(T),T)$ salvage or residual value of the road at the end of the planning period; and ρ is the appropriate constant interest rate. The quantities U, C, I and Ψ are assumed to be expressed in terms of some common numeraire, presumably dollars.

The problem defined by (2-1), (2-2) and (2-3) is an optimal control problem where both the objective function J and the constraints (2-2), are functions of one discrete interior point in time $t^* \in [0,T]$. Therefore, in order to find an optimum solution we can make use of the necessary conditions for optimality developed in section 3.3 of Chapter II for these kind of problems.

3.0 Solution of the Necessary Conditions

The Hamiltonian function for the problem defined in Section 2.0 is:

$$H(t)=\begin{cases} H_1(t) = (U-C_1)q \exp(-\rho t) + p(as+b), \ \forall\, t \in [o,t^{*-}] \\ \\ H_2(t) = (U-C_2)q \exp(-\rho t) + p(as+b), \ \forall\, t \in [t^{*+},T] \end{cases} \tag{3-1}$$

Where for simplicity in notation we have eliminated the arguments of all variables. The variable p, the adjoint variable, is a function of time and its interpretation is given later in the discussion. The symbols t^{*-} and t^{*+} refer to, respectively, the instants just before and just after t^*.

Necessary conditions for an optimal solution of our problem can now be expressed as follows: (See Chapter II, Section 3.3)

$$p = \begin{cases} -\partial H_1/\partial q, \ \forall\, t \in [o,t^{*-}] \\ \\ -\partial H_2/\partial q, \ \forall\, t \in [t^{*+},T] \end{cases} \tag{3-2}$$

$$p(t^*-) = p(t^*+) = p(t^*),\ p(T) = \left(\frac{\partial \Psi}{\partial q}\right)_T \tag{3-3}$$

$$\frac{\partial \phi}{\partial t^*} + H_1(t^*-) - H_2(t^*+) = 0, \tag{3-4}$$

where in our case $\phi = -I(t^*) \exp(-\rho t^*)$.

Following the standard terminology we refer to equations (3-2) as the adjoint equations and the boundary conditions (3-3) as the transversality conditions. Equation (3-4) determines the extremal staging time t^*. In general, as we saw in Chapter II, in order to determine the functional form of the extremal controls, the Hamiltonian must be maximized with respect to them on the interval $[o,T]$. In our case this additional necessary condition is redundant in light of constraint (2-2) which already defined the functional form for $s(t)$. That is to say, constraint (2-2) replaces in the present problem the usual maximization of the Hamiltonian as a necessary condition.

We now use (3-4) to find the extremal staging time t^*. Upon performing the differentiation denoted by $\partial\phi/\partial t^*$ and using (3-1), condition (3-4) may be rewritten as:

$$-I_{t^*} + \rho\, I(t^*) = [C_1(t^*) - C_2(t^*)]q(t^*)$$

$$+ \pi\,(t^*)a(t^*)(s_2 - s_1), \tag{3-5}$$

where I_{t^*} is the rate of change of the upgrading cost at the extremal

staging time t*; note that I_{t*} can be either positive or negative. Furthermore, the change of variable:

$$\pi(t) = p(t) \exp(\rho t)$$

has been made in (3-5). To obtain result (3-5) we have made use of the continuity of U, q, p, a and b. The continuity of these quantities may be expressed as:

$$\begin{aligned} U(t^*-) &= U(t^*+) = U(t^*) \\ q(t^*-) &= q(t^*+) = q(t^*) \\ p(t^*-) &= p(t^*+) = p(t^*) \\ a(t^*-) &= a(t^*+) = a(t^*) \\ b(t^*-) &= b(t^*+) = b(t^*) \end{aligned} \qquad (3\text{-}7)$$

It is easy to see that if the upgrading cost is constant over time ($I_{t*} = o$) and demand is independent of the quality of the facility ($a(t) = o$), expression (3-5) may be reduced to:

$$\rho\, I(t^*) = [C_1(t^*) - C_2(t^*)]q(t^*), \qquad (3\text{-}8)$$

where the left hand side is the rental value of the investment needed to upgrade the facility from s_1 to s_2 and therefore represents the marginal benefit, per unit of time, obtained from postponement of the investment decision. The term on the right-hand side corresponds to the marginal cost per unit of time, resulting from postponement of the investment and is a consequence of the operating cost reductions that would be obtained per unit of time if the upgrading of the road were

performed, but which are foregone due to postponement. Therefore, (3-8) says that the upgrading of the facility should be undertaken when the marginal cost of postponement becomes equal to the marginal benefit of postponement, per unit of time. However, because we have assumed that demand is independent of the quality of the facility, the marginal cost of postponement only considers the operating costs reductions that would be experienced, per unit of time, by the current users of the facility, if the upgrading were undertaken. This is exactly the rule proposed by Beenhakker and Danskin (1973), and extended by Venezia (1977) for the case of stochastic demands. The second term on the right hand side of (3-5), which appears if $a(t) \neq 0$, represents the consequence of explicity considering the interrelationship between quality and demand. We will turn now to its interpretation.

We saw in section 4 of Chapter II that the adjoint variables represent in general, dynamic shadow prices for the corresponding state variables. In this particular case, $p(t)$ is the shadow price of demand at time t. Thus, $p(t)$ evaluated along an optimal path will describe how much the objective function would be altered if we change the demand for transportation by one unit during the period $[t,T]$. It should be noted that our objective function is expressed in terms of present value as is the adjoint variable $p(t)$. Therefore, $\pi(t)$ will represent the transformation of $p(t)$ to current value at time t. Such a transformation has the advantage of expressing all the important variables of (3-5) in terms of current values.

Expressions (2-3) and (3-9) allow us to write the following expression for $\pi(t^*)$:

$$\pi(t^*) = \exp(\rho t^*) \int_{t^*+}^{T} [U - (C_2 + C_{2q}{}^* q^*)] \exp(-\rho t)dt + \left(\frac{\partial \Psi}{\partial q^*}\right)_T \qquad (3\text{-}10)$$

Where the integrand in this expression corresponds to the present value of the difference between utility and the social marginal cost due to a trip at time t; the notation $C_{2q^*} = \partial C_2/\partial q^*$ is used in (3-10). We are now able to identify the second term on the right side of (3-5) as the present value of the total benefit (which may be negative) obtained as a consequence of the new traffic generated at time t* from the change of facility quality produced by the investment $I(t^*)$. We can see that the sign of this additional term will depend on the sign of $\pi(t)$ if we make the reasonable assumption that $a(t) > 0$, $\forall\ t \in [0,T]$. Moreover, $\pi(t^*)$ will be positive when congestion is not high and utility U remains greater than the social marginal cost during the period $[t^*,T]$. Nevertheless, in some cases it could happen that the road improvement generates so much new traffic that at a certain time during the period $[t^*,T]$ the value of the social marginal cost becomes higher than the average utility U. In that case a negative value of $\Pi(t^*)$ can be obtained.

4.0 Sufficient Conditions

Before discussing the implications of our demand-quality investment rule (3-5) in greater detail, we must first analyze the second order or sufficiency conditions for the optimization model posed in

Section 2.0. That is to say, we want to know under what circumstances the extremal solutions obtained in Section 3.0 in fact lead to a maximum of objective function (2-3). For simplicity we will assume in the following analysis that construction cost I is constant with respect to time as is the discount rate ρ.

Since the decision rule (3-5) is stated as a marginal condition at time t^*, it will lead to a maximum of (2-3) if the benefits per unit of time generated by the new project are increasing at t^*; it will lead to a minimum if the marginal benefits per unit of time are decreasing. In analytical terms, if we use the notation:

$$B(t) = [C_1(t) - C_2(t)]q(t) + \pi(t)a(t)(s_2 - s_1), \tag{4-1}$$

then (3-5) will lead to a maximum if:

$$\dot{B}(t^*) > 0. \tag{4-2}$$

Differentiating (4-1) with respect to time and rearranging terms leads to:

$$\dot{B}(t) = \dot{q}(MC_1 - MC_2) + (s_2 - s_1)(a\,\dot{\pi} + \dot{a}\,\pi), \tag{4-3}$$

where:

$$MC_i = C_i + C_{iq}q. \tag{4-4}$$

is the marginal social cost produced by an additional user at time t, if the facility is at stage i. Expressions (3-1) and (3-2) may be used to write:

$$\dot{p}(t^*) = -[U - MC_2(t^*)] \exp(-\rho t^*). \tag{4-5}$$

Using (3-6), this last expression becomes:

$$\dot{\pi}(t^*) = \rho\, \pi(t^*) - [U - MC_2(t^*)]. \tag{4-6}$$

Also (2-1) gives:

$$\dot{q}(t^*) = a(t^*)s_2 + b(t^*). \tag{4-7}$$

We may now use the expressions for $\dot{\pi}(t^*)$ and $\dot{q}(t^*)$, equations (4-6) and (4-7), in conjunction with (4-1) and (4-2) to write the following sufficiency condition:

$$\dot{B}(t)/a(t) = (s_2 + b/a)(MC_1 - MC_2) + (s_2 - s_1)[(U - MC_2) - \pi(\rho + a')] > 0, \tag{4-8}$$

Where all variables are evaluated at t^* and:

$$a' = \dot{a}/a$$

is the proportionate rate of change of $a(t)$. The expression (4-8) will have the same sign as $\dot{B}(t)$ provided $a(t) > 0$, $\forall\ t \in [0,T]$. If the investment $I(t^*)$ produces an enhancement of facility quality, without changing capacity, we will have that the average operating cost C_i will be reduced for all q but $C_{1q} = C_{2q}$, given that capacity is the same in both cases. If $I(t^*)$ both increases quality and capacity, then, in addition to a reduction in the average operating costs, we will have $C_{2q} < C_{1q}$ for all q. Therefore, we will in general have:

$$MC_1(t^*) > MC_2(t^*). \tag{4-10}$$

In addition, if we assume that $a(t^*) > 0$ and $\dot{q}(t^*) > 0$, then

$$(s_2 + b/a) > 0$$

Consequently, the sufficiency condition (4-8) can be written as:

$$MC_1 - MC_2 > \Omega[(U - MC_2) - \pi(\rho + a')], \tag{4-11}$$

where:

$$0 < \Omega = \frac{(s_2 - s_1)}{(s_2 + b/a)} < 1; \tag{4-12}$$

it is assumed in (4-11) and (4-12) that all variables are evaluated at the extremal staging time t^*.

When congestion does not occur the following identity of course holds:

$$C_i = MC_i \tag{4-13}$$

Under the assumption that (4-13) holds, (4-6) becomes:

$$\dot{\pi} = \rho\, \pi - G_2 \tag{4-14}$$

Where G_2 represents individual gains obtained for times $t > t^*$ and is written:

$$G_2 = U - C_2. \tag{4-15}$$

Thus, a complete specification of a sufficiency condition when there is no congestion requires that one solve the linear differential equation (4-14) which will be subject to the boundary condition:

$$\pi\,(T) = \left(\frac{\partial\, \Psi}{\partial\, q}\right)_T \exp(\rho T) \tag{4-16}$$

which is obtained form (3-3) and (3-6). Boundary condition (4-16) makes it clear that in order to determine the value of $\pi(t)$ needed for our sufficiency condition we must assume an expression for the residual or salvage value of the road. We assume the following:

$$\Psi(T) = \int_{T}^{\infty} G_2 q(T)\exp(-\rho t)dt \qquad (4\text{-}17)$$

which states that the salvage value of the road will be equal to the present value of an infinite stream of benefits, starting at time T, with a stationary value equal to that obtained at time T. In other words, we assume that the demand will become stationary at a value q(T). To simplify the analysis we will further assume that the benefit measure G_2, defined by (4-15), is a constant with respect to time*. In that case:

$$\Psi(T) = \frac{G_2}{\rho} q(T)\exp(\rho T). \qquad (4\text{-}18)$$

It follows immediately that a general solution to (4-14), for the assumptions indicated, is give by:

$$\pi = \frac{G_2}{\rho}. \qquad (4\text{-}19)$$

Therefore, the sufficiency condition (4-11) becomes

$$(C_1 - C_2) > -\Omega \frac{a'}{\rho} G_2. \qquad (4\text{-}20)$$

---

*Simple modifications in the integration of equation (4-14) could allow the consideration of time dependent variables.

Clearly, if we assume $G_2 > 0$, then (4-20) will be satisfied for any non-negative value of $a'(t)$ since we expect operating cost reductions brought about by the upgrading of the facility to cause $(C_1 - C_2) > 0$. In the event a' is negative, then we write the sufficiency condition as:

$$(C_1 - C_2) > \Omega \frac{\| a' \|}{\rho} G_2 \qquad (4\text{-}21)$$

Since our sufficinecy expressions depend on $\dot{B}(t^*) > 0$, violation of (4-21) implies, for the case of no congestion, that $\dot{B}(t)$ is non-positive for $t = t^*$. If this is so, application of the decision rule (3-5) will lead to a minimum rather than a maximum, as Figure 6.1 illustrates. In such a circumstance the increase in marginal benefits derived from the reduction in costs, $q(t)(C_1 - C_2)$, will not be enough to compensate for the decrease in the marginal benefits produced per unit of time by the new traffic generated as a consequence of constructing the new facility. The benefits corresponding to the new traffic generated by the new facility per unit of time are:

$$\beta(t) = \pi(t)a(t)(s_2 - s_1).$$

Given that $\dot{\pi} = 0$ (see (4-19)) for the case analyzed, we have that:

$$\dot{\beta}(t) = \pi \dot{a}(t)(s_2 - s_1)$$

where $\pi = B_2/\rho > 0$ and $s_2 > s_1$. Therefore $\dot{\beta} < 0 \; \forall \; t \in [0,T]$ since

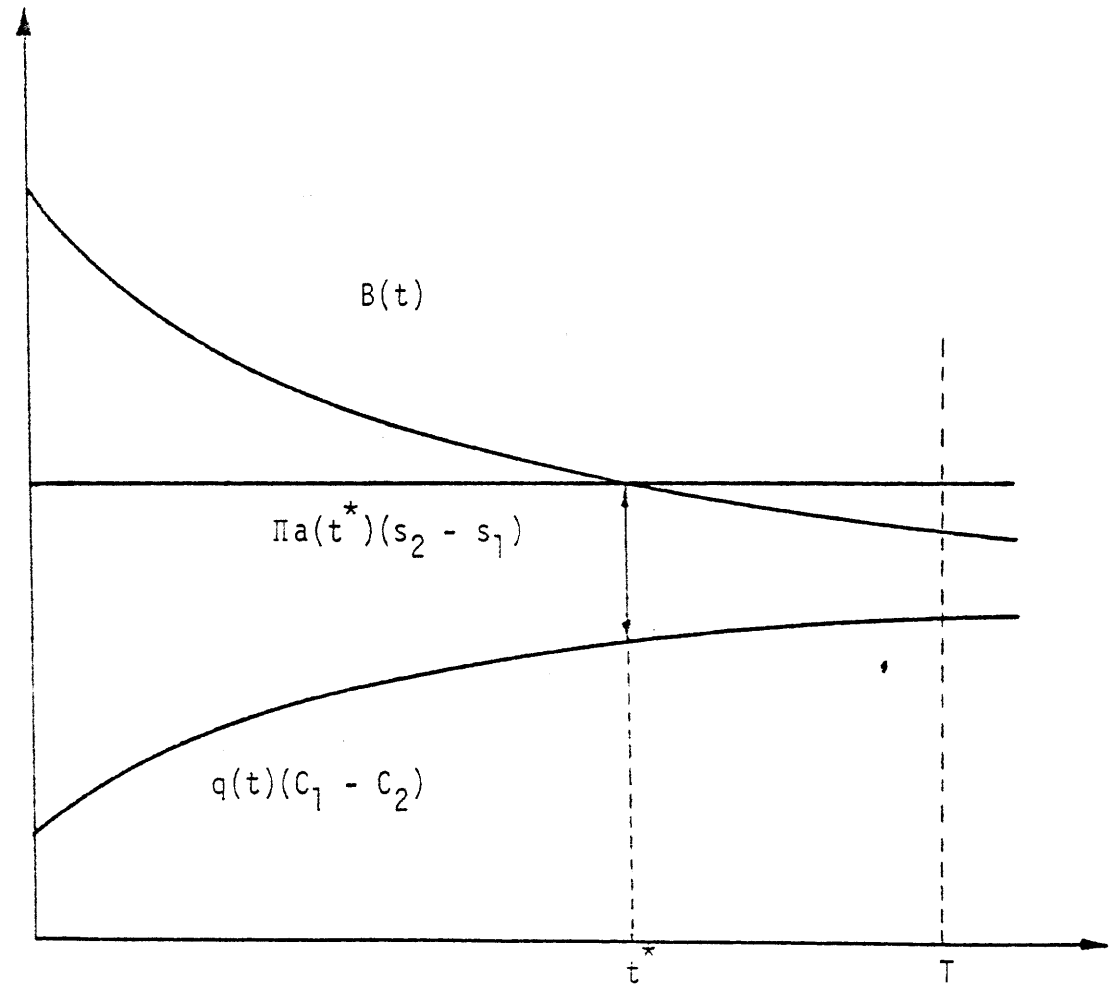


FIGURE 6.1 Graphical Representation of the Demand-Quality Decision Rule for the Case of Decreasing Marginal Beneifts

we have also assumed that $a(t) > 0$ and are considering $a' < 0$. Therefore the marginal benefits produced per unit of time by the new traffic generated are in fact decreasing and

$$|\dot{B}(t)| > \dot{q}(C_1 - C_2).$$

In the case that $\dot{q}(t)$, as well as a', is negative, both the marginal benefits derived from the reduction in costs and the marginal benefits associated with new traffic generated will be decreasing over time.

It is clear that when the circumstances illustrated by Figure 6.1 and described previously in the text occur, a non-interior or corner maximum for our objective function will be obtained at $t = 0$. This implies that, in effect, a "postponement" in the decision of implementing the new facility has occured which will cause a loss of $(B(t) - \rho I)$ marginal benefits per unit of time during the period $[0,t^*]$. Thus, the optimal construction time is $t = 0$ if the present value of net benefits produced during the period $[0,t^*]$ is superior to the present value of the losses obtained after t^*. Otherwise, the facility should never be constructed. In analytical terms the condition for construction at $t = 0$ will be:

$$\int_0^T (B(t) - \rho I)\exp(-\rho t)dt + \int_T^\infty [q(T)(C_1 - C_2) - \rho I]\exp(-\rho t)dt > 0, \quad (4\text{-}21)$$

where $q(T)$ is given by:

$$q(T) = q(0) + \int_0^T [a(t)s_2 + b(t]dt. \qquad (4\text{-}22)$$

Throughout this discussion of the case of no congestion the fact that the multiplier $\pi(t)$ has been a constant on $[0,T]$ has played an important role. We conclude this section with a discussion of the role of $\pi(t)$, which in general will be time varying, for the case of congestion effects.

According to (3-10) the value of $\pi(t)$ gives us an indication of the desirability of generating new traffic at time t. If congestion externalities do not exist, the benefits derived from any new traffic generated will depend simply on the difference between utility and operating costs "perceived" by the new users generated. If congestion is, however, an important factor those benefits will depend on the difference between utility and marginal cost "produced" by the new users. This means that we must reduce the benefits obtained by the new users in the amount of the increase in operating costs that they produce to the other road users. Therefore, congestion will in general reduce the desirability of new traffic and, "ceteris paribus", the desirability of the new investment that generates this traffic. This fact is represented in our model by the reduction in the value of $\pi(t^*)$ when $MC_2(t)$ increases (see again (3-10)). In some limiting cases, the congestion generated by new traffic could be of such magnitude that the marginal cost $MC_2(t)$ becomes higher than utility U over some critical portion of the period $[t^*,T]$, producing as a consequence a negative value of $\pi(t)$. This may cause an investment that is justified from the point of view of cost reductions to current

users to become infeasible if we take into account the congestion produced by the newly generated traffic. Such cases can be expected to correspond to relatively low values of U and are, therefore, likely to be most efficiently dealt with through adjustments in pricing policy.

5.0 Final Remarks and a Numerical Example

The model presented in Section 2.0 has allowed us to show that the explicit introduction of interrelationships between quality of serivice and demand has important consequences with respect to the derivation of optimal investment rules for transportation facilities. In the case of development projects, where congestion is not an important factor, the optimal investment rule (3-5) indicates that investments should be undertaken sooner than indicated by decision rules which take into account only the benefits derived from cost reductions to current users. In some extreme cases rule (3-5) will justify investments that would never be justified by these other rules. On the other hand, expansion investments will be penalized if the new traffic generated by them produces too much congestion in the future, causing their desirability to be less than that perceived by alternative decision rules.

Decision rule (3-5) can be considered as a generalization of Marglin's naive rule. The rule has the appeal of being formulated as a marginal condition at the staging time t^*. Therefore, its application only requires the knowledge at time t^* of the decision variables and key parameters plus some general assumptions about sufficiency.

However, not surprisinlgy, every bit of additional information has its price. In order to apply rule (3-5) we not only need to know the flow volume $q(t^*)$, but also an estimate of the new traffic that would be generated by the facility. This seems a reasonable price to pay to obtain a decision rule which will more completely capture the phenomena important to the time staging of transportation investments.

The discussion in previous sections has been largely abstract and theoretical. A better grasp of the ideas and implications of our findings may be obtained with a numerical example. With this purpose in mind and in order to compare our results with those obtained by other authors, we will solve the same problem orginally presented by Beenhakker and Daskin (1973) and subsequently analyzed by Venezia (1977). Beenhakker and Daskin (1973) found that the optimal sequence for upgrading an existing road in Iran involves two stages: the initial construction of a primitive facility and its subsequent widening and paving after 16 years. The relevant data are as follows:

(1) The cost functions of operating the road at stage $j = 1,2$ are given by:

$$C_1(t) = d_1 + c_1 q(t) = 553 + 80.60q(t)$$

$$C_2(t) = d_2 + c_2 q(t) = 3720 + 28.15q(t) \qquad (5\text{-}1)$$

where d_i represents fixed maintenance costs and the c_i are operating plus variable maintenance costs. The variable t denotes the number of years from the beginning of the analysis or planning period.

(2) The costs of upgrading the road are given by:

$$I(t) = I = 59{,}900.$$

(3) The discount factor, ρ, is 10%.

(4) Traffic grows at a fixed 7.5% per year, before and after the investment is performed, and the initial flow at t = 0 is 55 vehicles per day (VPD). Thus demand is described by:

$$q(t) = \gamma\ (\tau)^{t} \tag{5-2}$$

where:

$$q(0) = \gamma = 55 \text{ VPD.}$$

$$\tau = 1 + r$$

$$r = 0.075.$$

Consequently, the following dynamic description obtains:

$$\dot{q}(t) = \gamma\ (\ell n\ \tau)\ \exp(t \ell n\ \tau) = 3.978\ \exp(.072t). \tag{5-3}$$

We will assume that (5-3) gives us the evolution of demand before the investment is made, but contrary to Beenhakker and Daskin (1973) and Venezia (1977) we will assume that this rate of growth is affected by the investment I.

In order to carry out our analysis we use the following dynamics to represent the system:

$$\dot{q}(t) = [\gamma\ (\ell n\ \tau)\exp(t\ell n\ \tau)]s_1,\ \forall\ t\ \varepsilon\ [0,t^{*-}]$$

$$\dot{q}(t) = [\gamma\ (\ell n\ \tau)\exp(t\ell n\ \tau)]s_2,\ \forall\ t\ \varepsilon\ [t^{*+},T]. \tag{5-4}$$

Thus, in terms of the notation of Section 4:

$$a(t) = \gamma\ (\ell n\ \tau)\exp(t\ell n\ \tau) = 3.978\ \exp(0.72t),$$
$$\forall\ t\ \varepsilon\ [0,T]. \tag{5-5}$$

To be consistent with Beenhakker and Daskin (1973) we set $s_1 = 1.0$. We will consider different values of s_2 in order to investigate the sensitivity of the optimum investment date to the interrelationship between demand and quality. Specifically we will consider the set of values:

$$s_2 = (1.0,\ 1.2,\ 1.4,\ 1.6,\ 1.8,\ 2.0)$$

Since Beenhakker and Daskin (1973) and Venezia (1977) consider a cost minimization objective, they do not need to use values for the utility, U, that an average traveller obtains from use of the road. It is obvious that the restricted cost minimization formulation does not make sense in our case given the assumed interrelationship between demand and quality. Such an objective for our problem will lead

to the obvious optimal solution: do not invest. We will, therefore, assume different values of U ranging from a minimum value equal to the vehicle operating costs before the investment is performed to an arbitrary upper bound. We are assuming that if the utility obtained as a consequence of using the road is lower than the corresponding direct operating cost, flow should be zero. The values of U which will be considered are given by:

$$U = (61.2,\ 70.0,\ 80.6,\ 100,\ 130,\ 160).$$

It is easy to see that the expression for our decision rule (3-5) will be slightly modified in this case due to the presence of fixed maintenance cost d_i, to:

$$\rho I + (d_2 - d_1) = (C_1 - C_2)q(t^*) + \pi(t^*)a(t^*)(s_2 - s_1), \quad (5\text{-}6)$$

where:

$$\pi(t^*) = (U - C_2)/\rho$$

We can solve (5-6) for t^*, in this case, obtaining:

$$t^* = \frac{1}{\ell n\ \tau}\ \ell n \left\{ \frac{\rho I + (d_2 - d_1)}{\gamma\ [(C_1 - C_2) + (\ell n\ \tau)\ \pi\ (s_2 - s_1)]} \right\} \quad (5\text{-}7)$$

In Table 6.1 we present the different values of t^* obtained for

the different assumptions about the values of U and s_2. The first row of Table 1 shows the solution obtained by Beenhakker and Daskin (1973) for all the values of U; the results of this row correspond to what we would have expected given that $s_2 = s_1 = 1$, or no influence of quality over demand exists. For all other cases, substantial reductions over that value for t^* previously reported are obtained. The last row of the table shows the value of $\pi(t^*)$ that indicates the desirability of new flow for different values of U. Obviously the desirability increases as U increases. According to (4-19) second order conditions for an optimum are satisfied in this case since:

$$r > 0 \Rightarrow \tau > \lambda \Rightarrow a'(t) = \lambda n\, \tau > 0, \forall\, t \in [0,T].$$

Although the data for U and s_2 used in Table 1 is hypothetical, the ramifications of considering demand-quality interrelationships are obvious - for certain circumstances one may make gross errors in predicting the optimal staging time if such interrelationships are ignored.

TABLE 6.1 VALUES OF t^* FOR DIFFERENT VALUES OF s_2 and U.

s_2 \ U	61.2	70.0	80.6	100	130	160
1	16	16	16	16	16	16
1.2	14.8	14.5	14.1	13.5	12.5	11.7
1.4	13.7	13.1	12.5	11.4	9.8	8.4
1.6	12.6	11.9	11.0	9.5	7.5	5.8
1.8	11.7	10.7	9.7	7.9	5.6	3.6
2	10.8	9.7	8.5	6.4	3.8	1.6
π	330.5	418.5	524.5	718.5	1018.5	1318.5

VII. A MODEL OF OPTIMAL TRANSPORT MAINTENANCE WITH DEMAND RESPONSIVENESS

1.0 Introduction

In this chapter we formulate and solve a simple dynamic model to determine optimal maintenance policies for transport facilities. The formulation corresponds to a special case of the general models presented in Chapter III for the determination of quality investments. An example is worked out in detail and an algorithm for obtaining numerical solutions is proposed. Finally a sufficiency argument is presented.

Dynamic maintenance models have been presented before in the economic and management literature for the case of machines utilized in private industry. Näslund [1966] discusses the history of the problem of maintenance of machines, including various solution techniques, and is the first to propose that the problem be formulated as a dynamic optimization problem which may be solved by application of the Pontryagain maximum principle. Näslund [1966] outlines how the maximum principle may be utilized to obtain a solution. Later Thompson [1968] and Arora and Lele [1970] developed detailed solutions for control models of optimal machine maintenance. Finally, Bensoussan, et al. [1974] presents a summary of these control formulations. Though similar models can be developed to determine optimal maintenance policies for public facilities (e.g., transport infrastructure), little attention has been given to such modeling approaches in the economic and transportation literature. This can be explained in part by the fact that

the above mentioned models developed for the case of machines in the private sector possess some important short-comings that prevent their direct application to the analysis of public facilities. In particular: (1) the deterioration produced as a result of the intensity of use of the machine/facility is not considered; and (2) the potential for a good maintenance policy to reduce operating costs experienced by present users of the machine/facility is not explicitly articulated. By virtue of this latter type of savings, the number of future users may be expected to increase. It is thereby clearly necessary that an optimal transport maintenance policy reflect consideration not only of the present number of users, but also of the effect the facility will have in terms of generating additional users. This interelationship between quality of the facility and demand generated for its use is not considered in the models mentioned above. Recently Büttler and Shortreed [1978] have presented a dynamic investment planning model for the case of road transport. Their formulation does not provide the economic insights obtained in this chapter and relies on highly specialized assumptions concerning benefits and costs; most importantly their dynamical description does not explicitly consider the interaction of demand and quality.

In the remainder of the chapter the discussion will center around a dynamic road maintenance model. However, as our presentation will make clear, the same type of model could be applied to other forms of transport infrastructure and equipment.

2.0 Description of the Model

We choose to characterize the transport system of interest, an abstract road, in terms of two state variables: the "quality" of the road at time t, denoted by S(t); and the number of users at time t, denoted by q(t). The quality S could be represented by the present serviceability index (PSI), defined by AASHO [1962], if we are dealing with a paved road or by a roughness index for lower standard roads.

We will assume that S and q are interdependent variables and that their evolution over time is defined by the following system of differential equations:

$$\dot{S}(t) = -\alpha S(t) - \beta q(t) + \gamma V(t), \; S(0) = S_0 \qquad (2\text{-}1)$$

$$\dot{q}(t) = \alpha S(t) + b, \; q(0) = q_0 \qquad (2\text{-}2)$$

where

$V(t)$ = rate of maintenance expenses, the control variable

$S(t)$ = quality of the road at time t, a state variable

$q(t)$ = number of road users at time t, a state variable

α = parameter reflecting the natural rate at which the quality of the road deteriorates, i.e., $\alpha S(t)$ is the instantaneous rate of deterioration, at time t, regardless of travel or maintenance.

β = parameter reflecting the deterioration produced by each user

γ = parameter reflecting the rate of increase in road quality per dollar spent on maintenance per unit time

α = parameter reflecting the rate of change in the number of road users per unit of time, as a consequence of a unit change in road quality.

δ = parameter reflecting that portion of the rate of growth of transportation demand not influenced by changes in road quality.

Note that equations (2-1) and (2-2) constitute a pair of coupled differential equations; they relate the value of the input or control variable V(t) to the outputs or state variables S(t) and q(t). The formulation given by equations (2-1) and (2-2) represents specific assumptions about the dynamic behavior of the roadway. Equation (2-1) assumes that the quality of the facility changes over time primarily as a result of three separate causes: natural factors, use, and maintenance. The rate of natural deterioration is assumed to be proportional to the quality of the facility; this assumption produces a negative exponential deterioration process of the form depicted in Figure 7.1 when the facility is abandoned and only natural factors have an influence on quality. This is the usual assumption with respect to physical equipment and facilities (see Arora and Lele [1970], and Bensoussan et al., [1974]). In addition, (2-1) assumes that each user of the facility produces a constant deterioration β per unit of time of use. Finally, it is assumed that each dollar spent in maintenance produces an improvement γ in the quality of the facility. This linear relation between the rate of change in quality and maintenance expenditures amounts to assuming that constant returns to scale exist in the production of quality of the facility. Though this is not likely to be the case over an infinite range of values of V(t) it will generally be a good assumption over a limited range of expenditures in maintenance $m \leq V(t) \leq M$, as is the case in many production processes. The constant returns to scale assumption is common

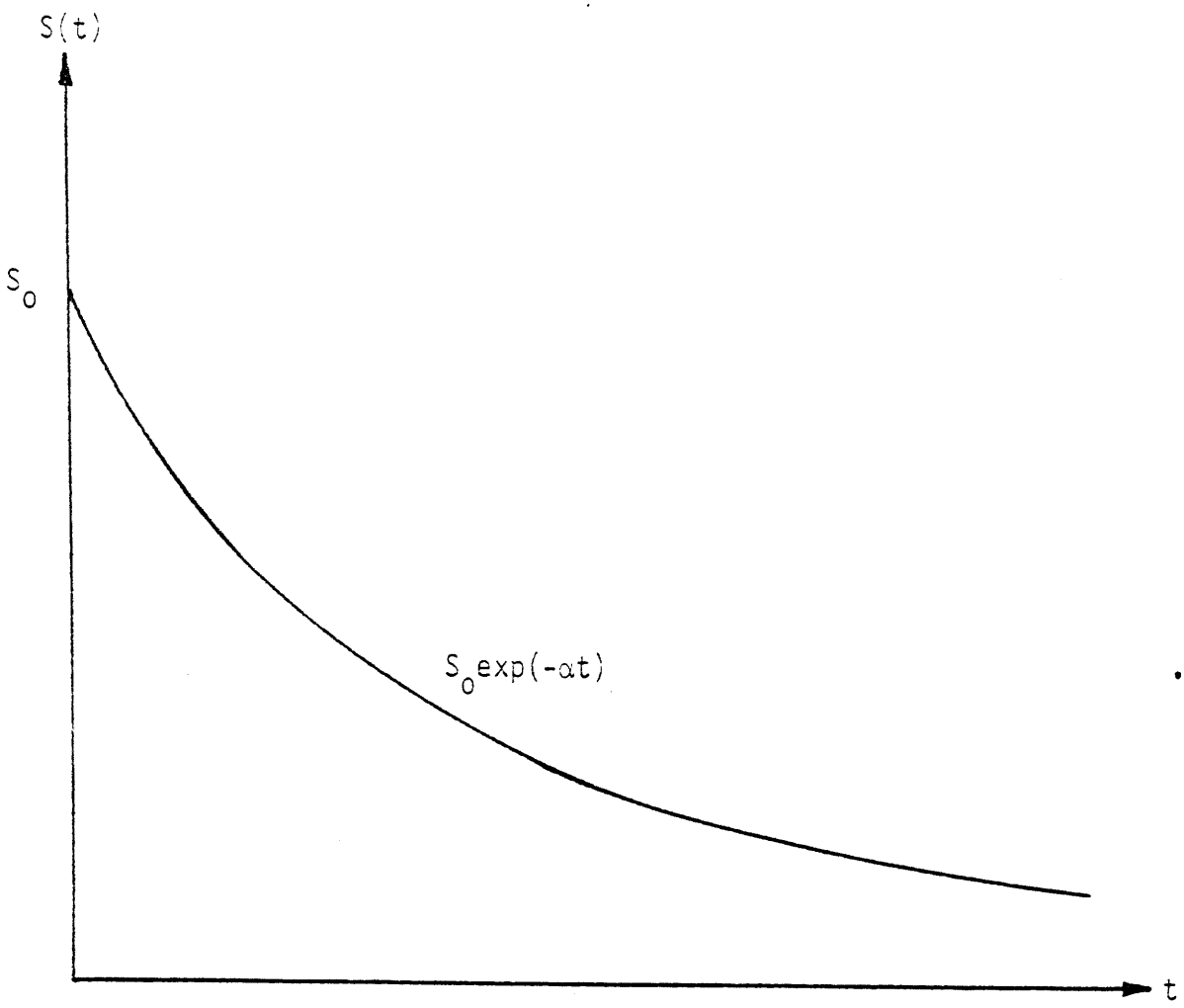


Figure 7.1: Natural Deterioration Process

to all maintenance models presented in the economic literature (see Thompson [1968], Arora and Lele [1970], and Bensoussan et al., [1974]).

In the interest of realism we make an additional assumption concerning the effects of maintenance. If we call S_0 the quality of the new facility at time t=0, then we want to have

$$S(t) \leq S_0, \ \forall \ t \in [0,T].$$

In other words, we assume that pure maintenance cannot drive the quality of the facility above its original value when new. This implies that V(t) obeys the following condition:

$$V(t) \leq \bar{V}(t) = \gamma^{-1}(\alpha S_0 + \beta q(t)). \qquad (2\text{-}3)$$

If V(t) is larger than $\bar{V}(t)$ the excess maintenance expenditure $V(t) - \bar{V}(t)$ has a null effect on the quality of the facility.

Equation (2-2) provides the link between quality of the facility and demand, creating a maintenance model with demand responsiveness. It assumes that demand is the consequence of two factors: 1) some external development pattern that is outside our control, represented by a natural rate of growth of demand (which may be negative), and 2) the quality of the road. Equation (2-2) may be placed in the alternative form

$$\dot{q}(t) = \alpha(S(t) - \bar{S}) + \acute{b}, \qquad q(o) = q_0, \qquad (2\text{-}4)$$

where $\bar{S} \leq S_0$. If $S(t) = \bar{S}$, demand grows at the natural rate $\acute{b}$. If $S(t) \geq \bar{S}$ or $S(t) \leq \bar{S}$, the natural rate of growth is respectively

increased or decreased by an amount α per unit change of quality. It is clear upon inspection that (2-4) may be reduced to (2-2) by the obvious transformation $b = (\acute{b} - \alpha\bar{S})$. Hence, it will suffice to consider only the dynamical description (2-2) in subsequent analyses. For the most general circumstance the parameters α and b of equation (2-2) as well as α β and γ of equation (2-1) would be functions of time. It will suffice for our purpose to consider these as constants. Of course, the initial values S_0 and q_0, defined in equations (2-1) and (2-2) respectively, are also known constants.

We will further assume that there is a utility U(t) attached to the use of the road which is the same for all users and that there is an associated operating cost C(S,q). The per user utility may be considered to be determined entirely by factors exogenous to the model so that it is written as a function of t only. Operating cost will, however, generally depend on road quality and the number of users; it is consequently written as a function of S and q. We wish to maximize the present value of net benefits derived from operation of the road over a fixed planning horizon T. Thus, the objective of interest is:

$$\text{MAXIMIZE } J = \int_0^T \{[U(t) - C(S,q)]q(t) - V(t)\} \exp(-\rho t)dt, \quad (2\text{-}5)$$

where ρ is a constant discount rate. In (2-5) the utility, cost and investment functions are of course assumed to be expressed in terms of a common numeraire, presumably dollars. Hence (2-5) actually represents the maximization of net benefits measured as dollars.

The final element necessary to specify the model is the assumption that maintenance expenditures will be bounded from above and below. We express this as:

$$m(t) \leq V(t) \leq M(t) \qquad \forall(t)\varepsilon[0,T] \qquad (2\text{-}6)$$

where m(t) and M(t) are respectively the lower and upper bounds on maintenance expenditures at time t. The lower bound m(t) will be determined by the fixed factors of maintenance production at time t; the upper bound M(t) will be the budget constraint at time t. Values of M(t) should obviously correspond to reasonable maintenance expenditures. A range of such reasonable values for M(t) can be obtained from equation (2-3).

3.0 The Optimal Maintenance Policy: Necessary Conditions and Economic Interpretations

The problem of maximizing the objective function J, defined in equation (2-5), subject to the growth dynamics (2-1) and (2-2), as well as the limitation on maintenance expenditures (2-6), constitutes an optimal control problem with fixed terminal time and no state space constraints. Necessary conditions for such problems were described in Section 3.1 of Chapter II.

Solution of our optimal control problem begins by specifying the Hamiltonian function:

$$H(t) = \{[U(t) - C(S,q)]q(t) - V(t)\}\exp(-\rho t) \qquad (3\text{-}1)$$

$$+ P_1(t)[-\alpha S(t) - \beta q(t) + \gamma V(t)] + P_2(t)[\alpha S(t) + b].$$

The adjoint variables $P_i(t)$ must satisfy

$$\dot{P}_1(t) = -(\partial H/\partial S) \qquad \dot{P}_2(t) = -(\partial H/\partial q)$$

which take the form

$$\dot{P}_1 = C_S q \exp(-\rho t) + \alpha P_1 - \alpha P_2 \qquad (3\text{-}2)$$

$$\dot{P}_2 = [U - (C + C_q q)]\exp(-\rho t) + \beta P_1 \qquad (3\text{-}3)$$

In (3-2) and (3-3) the arguments of all variables have been eliminated for simplicity. The subscripts S and q denote partial derivatives with

respect to those variables. The adjoint variables must also satisfy the following boundary conditions which are a result of the more general transversality conditions for optimal control problems

$$P_1(T) = P_2(T) = 0. \tag{3-4}$$

In addition the maximum principle requires that we seek values of the control V(t) that maximize the Hamiltonian. That is, we seek controls V^* such that

$$H(S^*, q^*, P_1^*, P_2^*, V^*, t) \geq H(S^*, q^*, P_1^*, P_2^*, V, t), \quad \forall\, V \in \Omega \tag{3-5}$$

where $\Omega = [V:\ m(t) \leq V(t) \leq M(t), \forall t]$ and the superscript "*" means that the corresponding variables satisfy the appropriate necessary conditions.

In order to apply the maximum principle it is expedient to rewrite the Hamiltonian as

$$H = (U - C)q \exp(-\rho t) - P_1(\alpha S + \beta q) + P_2(\alpha S + b) + [\gamma P_1 - \exp(-\rho t)]V. \tag{3-6}$$

From this last expression it is easy to see that the gradient of the Hamiltonian with respect to the control variable V is given by

$$H_V = \gamma P_1 - \exp(-\rho t). \tag{3-7}$$

Moreover, since the Hamiltonian is a linear function of V, the extremal control will obey:

$$V^*(t) = \begin{cases} m(t) & \text{, if } P_1(t) < \frac{1}{\gamma} \exp(-\rho t) \\ M(t) & \text{, if } P_1(t) > \frac{1}{\gamma} \exp(-\rho t) \\ \text{undetermined, if } P_1(t) = \frac{1}{\gamma} \exp(-\rho t). & \end{cases} \qquad (3\text{-}8)$$

3.1 Bang-bang Policy

We observe that expressions (3-7) and (3-8) imply that $V^*(t)$ is a well defined function of $P_1^*(t)$, γ and t as long as the gradient H_V^* is non-zero. If the gradient function $H_V^*(t)$ vanishes only at a countable number of times within the interval [0,T] our optimal control problem is called "normal" and the optimum policy $V^*(t)$ is "bang-bang" (see Section 3.2 of Chapter II). The value of $V^*(t)$ switches from one boundary of its constraint set to another at certain well defined times given by

$$H_V^*(t) = F_S(t) = P_1(t) - [\gamma \exp(\rho t)]^{-1} = 0 \qquad (3\text{-}9)$$

Generally, $F_S(t)$ is referred to as the "switching function".

To interpret the maintenance policy described by (3-8), we must give an interpretation of the adjoint variable $P_1(t)$. This interpretation is provided by the following identity which holds at optimality (see Section 4 of Chapter II).

$$P_1(t) = \frac{\partial J^*(S^*;t)}{\partial S^*}, \quad \forall t \in [0,T] \qquad (3\text{-}10)$$

where the superscript "*" now denotes the fact that J is evaluated along the optimal solution path. This identity suggests that we interpret $P_1(t)$ as a dynamic shadow price of quality. As such $P_1(t)$ represents the additional benefit, in present value, obtained from a unit increase in road quality at time t. On the other hand, $[\gamma \cdot \exp(\rho t)]^{-1}$ is equal to the present value of the amount we should spend on maintenance to obtain a unit increase in road quality at time t.

The (bang-bang) policy states, therefore, that maintenance should be performed to the extent that the present value of the marginal benefit produced by one additional dollar spent in maintenance at time t, $\gamma P_1(t)$, is higher than the present value of that dollar, $\exp(-\rho t)$. Given that we have assumed constant returns to scale in the production of quality through maintenance for values of V(t) in Ω, the marginal benefit $\gamma P_1(t)$ will be a constant for each t, independent of V(t).

Therefore, we must spend the maximum amount available M(t) as long as the marginal benefit $\gamma P_1(t)$ is higher than the marginal cost $\exp(-\rho t)$. If the marginal cost becomes higher than the marginal benefit we should spend the minimum possible m(t). The case of equality of the marginal benefit and the marginal cost can be neglected here, given that it occurs only at a countable number of times in [0,T].

Further insight can be gained from the interpretation of equations (3-2) and (3-3) that describe the evolution of the adjoint variables $P_1(t)$ and $P_2(t)$ during the period of analysis. This interpreation

requires that we first make note that a general interpretation of $P_2(t)$ can be obtained from the fact that along an optimal solution path

$$P_2(t) = \frac{\partial J^*(q^*,t)}{\partial q^*}, \quad \forall t \in [0,T]. \tag{3-11}$$

As such $P_2(t)$ represents the additional benefit, in present value, obtained from an additional user generated at time t. That is, $P_2(t)$ may be thought of as a dynamic shadow price of demand. Equations (3-2) and (3-3) correspond to a system of coupled first order ordinary differential equations in P_1 and P_2. The solution of this system can be expressed, using (3-4), as

$$P_1(t) = -\int_t^T C_S \exp[-\alpha(x-t)]q(x)\exp(-\rho x)\,dx + \int_t^T P_2(x)\,\alpha \exp[-\alpha(x-t)]dx. \tag{3-12}$$

$$P_2(t) = \int_t^T [U(x) - C(x)]\exp(-\rho x)dx - \int_t^T C_q q(x)\exp(-\rho x)dx - \int_t^T P_1(x)\beta\,dx. \tag{3-13}$$

The first term of (3-12) corresponds to the present value of the direct benefit (operating cost reductions), produced during the period [t,T], by a unit enhancement of facility quality (through maintenance expenditure) at time t. Note that $\exp[-\alpha(x-t)]$ is the equivalent value at time x of a unit enhancement of facility quality at time t. The second term of (3-12) is equal to the present value of the benefit attached to the

new traffic generated during the period [t,T] as a consequence of a unit enhancement of facility quality at time t. In expression (3-13) the first term is the present value of the private benefit perceived by an individual user of the facility during the period [t,T]. The second term in (3-13) is equal to the present value of the externalities of congestion produced during the period [t,T] as the result of an additional user generated at time t. Finally, the third term in (3-13) represents the externalities of deterioration produced during the period [t,T] by an additional user generated at time t. This last term takes into account the fact that the deterioration produced by one user affects the operating costs perceived by all other users of the facility. The sum of the integrands of the second and third terms of (3-12) is therefore equal to the value of the externalities produced by an additional user at time t. This sum is equal to the value of the toll that each user of the facility at time t should be charged if an optimal pricing policy were applied.

Therefore, even though the bang-bang maintenance policy is expressed only in terms of $P_1(t)$, we can see that the value of the marginal benefit attached to a new user of the facility generated at time t, $P_2(t)$, also plays a fundamental role in the interpretation and potential implementation of the policy. That role is diminished if we assume $\alpha = 0$. Nevertheless, even then $P_2(t)$ still tells us how much the value of the objective function J will increase or decrease for a unit change in the number of users of the facility.

The bang-bang policy is illustrated in Figure 7.2 for the case of a control set Ω with the respective constant upper and lower bounds M and m.

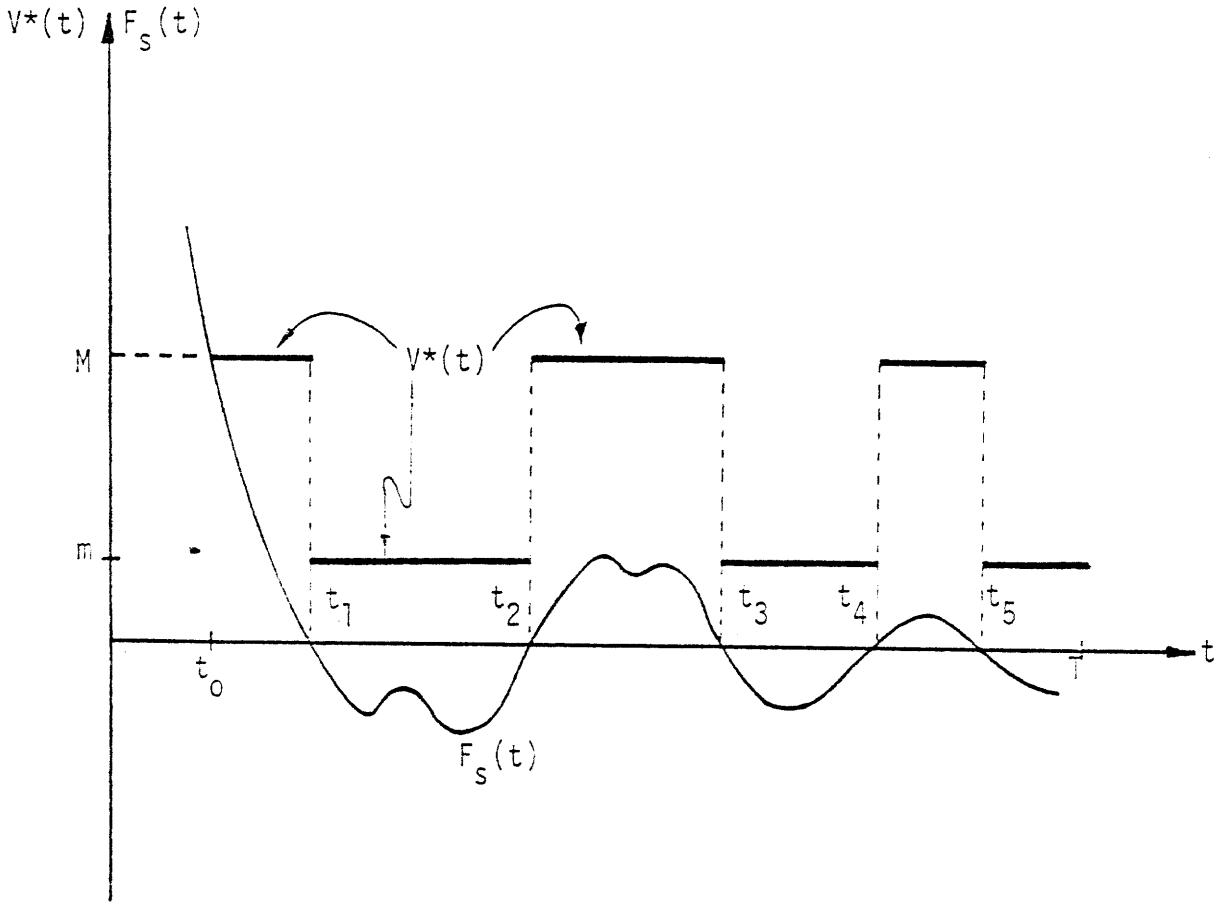


Figure 7.2 Bang-Bang Maintenance Policy

3.2 Singular Controls

In the preceeding section we assumed that the gradient H_V^* (or switching function F_s) vanishes only at a countable number of times in the period [0,T]. In this section we will analyze the possibility that the gradient H_V^* vanishes identically over one or more finite periods of time or subintervals in [0,T]. In such a case we say that we have a singular optimal control problem and the periods for which $H_V^* = 0$ are called singular intervals or singular arcs. As we noted in (3-8), the necessary condition (3-5) does not provide enough information in this case to define $V^*(t)$ along a singular arc. In the absence of such information, we must manipulate the other necessary conditions in an effort to determine a well defined expression for the control on the singular arc, denoted as $V_s^*(t)$.

Singular controls can in general be determined by making use of the following observation: if the gradient H_V of the Hamiltonian vanishes identically along a singular arc, then the time derivatives of H_V^* must remain equal to zero during the same period. From (3-7) and (3-8) we have that on a singular arc(see Section 3.2 of Chapter II)

$$H_V^* = \gamma P_1 - \exp(-\rho t) = 0. \qquad (3\text{-}14)$$

Upon taking the derivative of (3-14) with respect to time and using the adjoint equation (3-2) to eliminate $\dot{P}_1$, we obtain

$$\dot{H}_V^* = (\gamma C_S q + \rho)\exp(-\rho t) + \gamma \alpha P_1 - \gamma \alpha P_2 = 0. \qquad (3\text{-}15)$$

Now if we again use (3-14) to eliminate P_1 from this last expression, we can write

$$\dot{H}_V^* = (\gamma C_S q + \alpha + \rho)\exp(-\rho t) - \gamma \alpha P_2 = 0. \qquad (3\text{-}16)$$

An economic interpretation of this condition for a singular arc can be obtained if we rewrite it as

$$\alpha P_2 \exp(\rho t) - C_S q = r/\gamma, \quad r = (\alpha + \rho) \qquad (3\text{-}17)$$

The first term on the left hand side of (3-17) is equal to the present value, at time t, of the benefits derived from the generation of α new users at this time as a consequence of a unit improvement of facility quality. The second term on the left hand side of (3-17) is the total savings in operating costs, perceived by all users of the facility, at time t, as a consequence of a unit improvement of facility quality at that time. The right hand side of (3-17) is expressed in terms of the "effective discount rate" $r = \alpha + \rho$, and γ^{-1}, which is equal to the maintenance investment necessary to obtain a unit improvement of facility quality at time t. Therefore the right hand side of (3-17) is equal to the rental value of this maintenance investment, using r as the interest rate.

As a further step we take the second derivative of (3-14) with respect to time. In doing this we will assume that $C_{Sq} = 0$, or, in other words, the quality S that we are considering is not capacity related and therefore has no effect on the congestion produced in the facility

for each level of use. This assumption is for convenience only; if it were dropped the only consequence would be a modification of the expression obtained for the singular control. In particular, this assumption does not affect the economic interpretation obtained from (3-17). Using this assumption we obtain

$$\ddot{H}_V^* = (-\rho\ \gamma\ C_S q - \rho\ \alpha - \rho^2 + \gamma\ C_{SS}\dot{S}q + \gamma\ C_S\dot{q})\exp(-\rho t) - \gamma\ \alpha\ \dot{P}_2 = 0.$$

By using (2-1),(2-2),(3-3) and (3-14) to eliminate $\dot{S}$, $\dot{q}$, $\dot{P}_2$ and P_1, this last expression becomes

$$\ddot{H}_V^* \exp(\rho t) = C_{SS}\ \gamma\ q(-\alpha s - \beta q + \gamma V) + \gamma\alpha(U - C - C_q q)$$

$$+ C_S\gamma(\alpha S + b - \rho q) - \alpha\beta - \rho(\alpha + \rho) = 0. \qquad (3\text{-}18)$$

If the operating cost function C is nonlinear in S and therefore C_{SS} is different from zero, we obtain from (3-18) an expression for the singular control in terms of the values of the state variables at time t, and the parameters of the problem. That expression is

$$V_S^* = (\gamma^2 q C_{SS})^{-1}[C_{SS}\ \gamma\ q(\alpha S + \beta q) - \gamma\alpha(U - C - C_q q)$$

$$+ (C_S\ \gamma(\rho q - \alpha S + b) + \alpha\beta + \rho(\alpha + \rho)]. \qquad (3\text{-}19)$$

Obviously, the singular control V_S^* must also obey the control constraints (2-6). However, optimal singular controls must satisfy an additional necessary condition. For a maximization problem with a single control variable (recall V is a scalar) the condition can be stated as (see

Section 3.2 of Chapter II

$$(-1)^m \frac{\partial}{\partial V} [(\frac{d}{dt})^{2m} H_V^*] \leq 0 \qquad (3\text{-}20)$$

where m is an integer. Note that in our case m = 1. By using (3-18) it is easy to see that (3-20) becomes

$$\gamma^2 q \, C_{SS} \geq 0. \qquad (3\text{-}21)$$

Since q is always positive, (3-21) implies that in order for V_S^* in (3-19) to be an optimal control, C_{SS} must be positive. This means that no optimal singular control exists if C_{SS} is negative, or in other words if C is non-convex in S.

Now we can analyze the case in which C is linear in S and therefore $C_{SS} = 0$. Then condition (3-18) does not provide an expression for V_S^* and we must take a new derivative with respect to time, which after using (2-1) to replace $\dot{S}$, can be written as

$$\ddot{H}_V^* \exp(\rho t) = -\dot{q} \, \gamma h = 0; \; h = 2\alpha C_q + \alpha C_{qq} q + \rho \, C_S. \qquad (3\text{-}22)$$

It is clear that there are only two possibilities of satisfying (3-22): $\dot{q} = 0$ or $h = 0$, given that $\gamma \neq 0$. Let us assume first $h = 0$, $\dot{q} \neq 0$. Then, given that we are on a singular arc, we must have

$$\dot{h} = \dot{q}\alpha(3C_{qq} + C_{qqq} \, q) = 0. \qquad (3\text{-}23)$$

It can now be seen by inspection that if $\dot{q} \neq 0$ we can continue taking time derivatives without obtaining any expression for the singluar

control. Therefore a singular control will exist in this case only if $\dot{q} = 0$ which corresponds to an equilibrium solution for the problem. From (2-2) we see that $\dot{q} = 0$ implies that

$$S = -(b/a) \tag{3-24}$$

which in turn implies that $\dot{S} = 0$. Moreover, if we assume that S cannot be negative, (3-24) only makes sense if b is negative, or in other words the demand dynamical description is such that demand decreases when $S = 0$. Using (3-24) and (2-1) we obtain

$$V_S^* = \gamma^{-1} (\beta q - \alpha b/a). \tag{3-25}$$

In order to obtain the equilibrium value of q in a particular case we can make use of expression (3-18) with $C_{SS} = 0$, which constitutes a necessary condition for the existence of V_S^*. If one utilizes expressions for C_q and C_S in (3-18) associated with a particular cost function $C(\cdot)$ together with the equilibrium value of S obtained from (3-24), the result will be an equation in q whose solution, if it exists, will provide the value of the equilibrium demand. It is important to note that the singular controls defined by (3-19) and (3-25) will both correspond to policies whose economic interpretation is that obtained from (3-17), since that expression constitutes a necessary condition for the existence of such controls.

Using (2-2) to replace $\dot{q}$ in (3-22) and taking a new derivative with respect to time it is easy to obtain

$$\frac{\partial}{\partial V}\left[\left(\frac{d}{dt}\right)^4 H_V^*\right] = -\gamma^2\,\alpha(2\alpha C_q + \alpha C_{qq}q + \rho C_S)\exp(-\rho t). \quad (3\text{-}26)$$

The additional necessary condition (3-20) now corresponds to the case $m = 2$ and can be expressed as

$$-\gamma^2\,\alpha(2\alpha C_q + \alpha C_{qq}q + \rho C_S)\exp(-\rho t) \leq 0.$$

which in turn requires that

$$\alpha(2C_q + C_{qq}q) \geq -\rho C_S. \quad (3\text{-}27)$$

We can write this last expression as

$$\alpha\,\frac{\partial}{\partial q}\,(C + C_q q) \geq -\rho C_S. \quad (3\text{-}28)$$

The term in parenthesis in (3-28) is the social cost at time t of introducing a new user into the facility and the right hand side of (3-28) is the rental value of the operating cost reductions experienced by each user of the facility at time t when quality is improved by one unit. The equilibrium value of q must satisfy (3-27), otherwise V_S^* given by (3-25) cannot be an optimal policy.

In practice, the existence of initial and final conditions that the variables of the problem must satisfy will not allow the application of singular controls over the whole period [0,T]. A singular arc, can be represented as a trajectory in the space of the state variables

(S,q). In general the initial values (S_o, q_o) will define a starting point of this trajectory and we will need to make use of the maximum or minimum values of V(t) in order to get on to the singular arc. On the otherhand, along a singular arc condition (3-14) must always hold. If the transversality conditions for the adjoint variables are such that $\gamma P_1(T)$ is not equal to $\exp(-\rho t)$, then the final point of an optimal trajectory cannot be over the singular arc, where the optimal control is V_s^*. Thus we will have to again use the maximum or minimum values of V(t) in order to meet the final condition of the problem. In our case it is easy to see from the transversality conditions (3-4) that the final point of an optimal singular policy will be over a singular arc only if the period of analysis is $[0,\infty]$. Therefore, possible optimal singular policies will in general involve a combination of bang-bang and singular arcs.

Thus, for singular policies the optimal control will in general have the form:

$$V^*(t) = \begin{cases} m(t), & \text{if} \quad P_1(t) < \frac{1}{\gamma} \exp(-\rho t) \\ V_s^*(t), & \text{if} \quad P_1(t) = \frac{1}{\gamma} \exp(-\rho t) \\ M(t), & \text{if} \quad P_1(t) > \frac{1}{\gamma} \exp(-\rho t). \end{cases} \qquad (3\text{-}29)$$

A representation of such a policy is given in Figure 7.3.

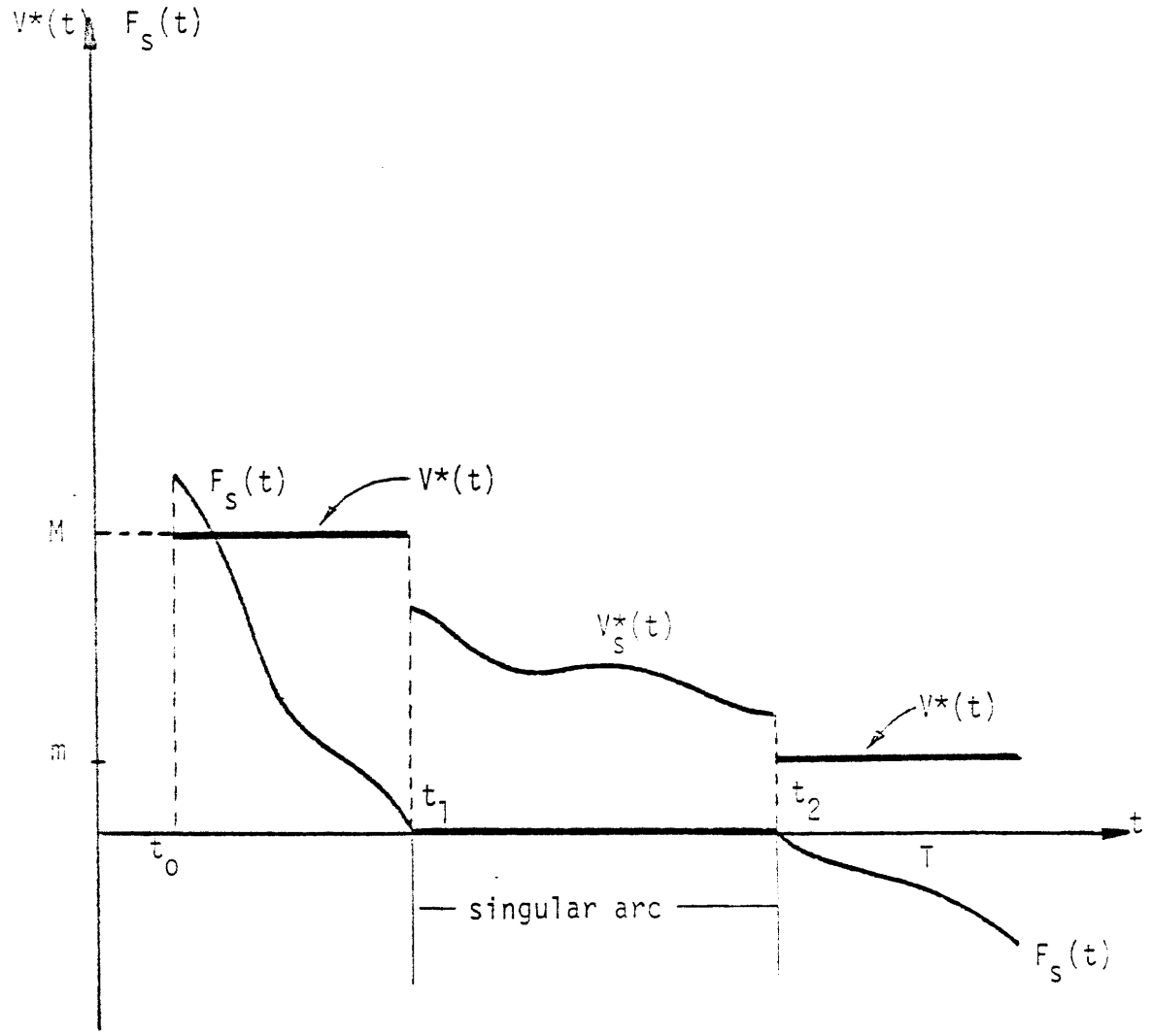


Figure 7.3 Maintenance Policy with a Singular Arc.

4.0 Solution in a Particular Case.

In order to illustrate explicitly how the necessary conditions, presented in Section 3, are used to obtain solutions in a particular case, we make the following assumptions regarding the functional form for the utility and operating cost function:

$$U = \text{constant}$$
$$C(S,q) = c - \varepsilon S + \delta q^2 \qquad (4\text{-}1)$$

where c, ε, and δ are constants. In addition we assume that the feasible region for maintenance expenditures, Ω, is invariant for all t in $[0,T]$, i.e.,

$$m \leq V(t) \leq M \qquad \forall t \varepsilon [0,T]$$

where m and M are constant minimum and maximum maintenance expenditures.

Assumption (4-1) describes a circumstance where the user obtains a constant gross benefit from utilizing the system with costs that depend on the quality of the facility and the number of users. The first term of the operating cost function is the cost perceived by one user when the quality of the road is equal to an arbitrary reference value $S = 0$ and free flow conditions exist over the road. The second term of the operating cost function requires a unit increase in facility quality to bring about an operating cost decrease of ε monetary units. Finally, the last term of the operating cost function considers the effect of congestion on individual operating costs; it requires operating cost to increase as the number of users increases. The operating cost functions

commonly used in practice are those proposed by the U.S. Federal Highway Administration (COMSIS[1972]) and are of the form $C(q) = c + \delta q^n$. Although n is taken to be 4 for the case of urban highways, the value of n is to a great extent arbitrary as long as the resulting function is increasing and convex in q. We have chosen in (4-1) a value of n = 2, that is more appropriate for intercity than for urban situations. The constant c and δ are empircally determined parameters for each road which depend on its length, speed limit and geometric design, including number of Lanes (COMSIS[1972]). In (4-1) we have also included the term $-\varepsilon S$ to take into account the influence of road quality on operating costs. Note that we obtain the partial derivatives

$$C_S = -\varepsilon, \; C_q = 2\delta q, C_{qq} = 2\gamma \qquad (4\text{-}2)$$

immediately from (4-1).

4.1 Singular Case

In order to identify the characteristics of the optimal solution, we must first analyze the existence of singular controls. From (4-1) it is easy to check that $C_{Sq} = C_{SS} = 0$. Therefore we can make use of the conditions developed in Section 3 for such characteristics of the user cost function. Thus, using (4-1), (4-2) and (3-24) in (3-18) we obtain

$$3\alpha\delta q^2 - \varepsilon\rho q - A = 0 \qquad (4\text{-}3)$$

where

$$A = \alpha(U-c) - \varepsilon\beta - (\alpha\beta + \rho\alpha + \rho^2)/\gamma.$$

Upon solving (4-3) for q we obtain

$$q = [\varepsilon\rho \pm (\varepsilon^2\rho^2 + 12\alpha\delta A)^{\frac{1}{2}}]/6\alpha\delta. \tag{4-4}$$

Therefore, one condition for the existence of a singular control for our particular problem is

$$\varepsilon^2\rho^2 \geq -12\ \alpha\delta A \tag{4-5}$$

since a real stationary demand does not otherwise exist. A second condition is given by (3-27) which specializes to the form

$$q \geq (\rho\varepsilon/6\alpha\delta) \tag{4-6}$$

and which eliminates from consideration the solution of (4-4) given by the minus sign of the radical. Therefore, if (4-5) is satisfied, a singular control exists and, from (3-25), is given by

$$V_S^* = \gamma^{-1}(\beta q_e - \alpha b/\alpha) \tag{4-7}$$

where

$$q_e = [\varepsilon\rho + (\varepsilon^2\rho^2 + 12\alpha\delta A)^{\frac{1}{2}}]/6\alpha\delta \tag{4-8}$$

$$S_e = -b/\alpha.$$

The quantities q_e and S_e are respectively the equilibrium values of demand and quality obtained along the singular arc when V_S^* is applied. Obviously a singular control does not exist in this case if V_S^* given by (4-7) is higher than M or lower than m.

Since the initial condition (S_0, q_0) will in general be different than the equilibrium values (S_e, q_e) given by (4-8), we will need to use bang-bang controls in order to get on to the singular arc (which in this case is represented by a point in the space (S,q)). In addition, given that $P_1(T) = 0$ (see (3-4)), it is easy to see from (3-8) that the optimal control will be $V^* = m$ for values of t close to T, with T finite. In order to obtain the values of t for which the optimal control changes from bang-bang to singular and conversely, we must integrate the adjoint equations and thereby find the values of t for which condition (3-14) is satisfied. This involves the solution of a two point boundary value problem essentially identical to that analyzed below in Sections 4.2 and 5.0 for the pure bang-bang case.

4.2 Bang-bang Case

If singular controls do not exist, the optimal solution must be purely bang-bang. In that case in order to completely specify the optimal policy $V^*(t)$ we must find the countable number of times for which the gradient of the Hamiltonian H_V^* vanishes. This is equivalent to finding the solutions of the switching function $F_S(t)$ given by (3-9).

Using expressions (4-2) in (3-2) and (3-3) leads to

$$\dot{P}_1 = -\varepsilon q \exp(-\rho t) + \alpha P_1 - \alpha P_2 \qquad (4\text{-}9)$$

$$\dot{P}_2 = (3\delta q^2 - \varepsilon S - B)\exp(-\rho t) + \beta P_1 \qquad (4\text{-}10)$$

where

$$B = U - C \tag{4-11}$$

and of course (3-4) still holds, that is

$$P_1(T) = P_2(T) = 0. \tag{4-12}$$

The system described by (4-9) and (4-10) may be uncoupled to yield

$$\ddot{P}_1 - \alpha\dot{P}_1 + \alpha\beta P_1 = F(t) \tag{4-13}$$

where

$$F(t) = (\alpha B - \varepsilon b + \rho\varepsilon q - 3\alpha\delta q^2)\exp(-\rho t). \tag{4-14}$$

In order to solve (4-13) for $P_1(t)$ we need to find an expression for $q(t)$. In order to do this we must solve equations (2-1) and (2-2) for $S(t)$ and $q(t)$, a step which is complicated by the fact we do not have an exact expression for $V(t)$ before the values of the switching times, i.e., the roots of equation (3-9), are known. Nonetheless, we know from Section 3 that the optimal $V(t)$ will be a piecewise continuous function with values m or M, except for a countable number of points corresponding to the switching times. The derivative of $V(t)$ with respect to time will be equal to zero for all $t\varepsilon[0,T]$ not corresponding to a switching time. This information is adequate for finding general expressions for $S(t)$ and $q(t)$.

We restate equations (2-1) and (2-2) as

$$\left.\begin{aligned} \dot{S} &= -\alpha S - \beta q + \gamma \tilde{V} \\ \dot{q} &= \alpha S + \zeta \end{aligned}\right\} \qquad (4\text{-}15)$$

where $\tilde{V}$ is the unknown optimal function $V(t)$ and $d\tilde{V}/dt = 0$. System (4-15) may then be uncoupled to yield

$$\ddot{S} + \alpha\dot{S} + \beta\alpha S = -\beta\zeta, \qquad (4\text{-}16)$$

a differential equation which is valid everywhere except possibly at the countable number of switching times. The solution of this differential equation depends on the roots of its auxiliary or indicial equation which may be expressed as,

$$m_1 = -\frac{\alpha}{2} + \sqrt{\frac{\alpha^2}{4} - \beta\alpha} \quad \text{and} \quad m_2 = -\frac{\alpha}{2} - \sqrt{\frac{\alpha^2}{4} - \beta\alpha}. \qquad (4\text{-}17)$$

Of course, the precise nature of solutions to (4-10) will depend on the discriminant in (4-17) which we write as,

$$\Delta = \frac{\alpha^2}{4} - \beta\alpha. \qquad (4\text{-}18)$$

The discriminant Δ is directly related to the intrinsic characteristics of the dynamics of the problem. The first term in (4-18) is related to the natural deterioration process; the second term is related to the deterioration process associated with facility utilization, which we call use deterioration. In fact, the value of ζ is the result of a

comparison between the second order effects of the natural and use deterioration processes when an improvement of one unit of quality in the facility occurs. The first order effect of a unit increase in the value of quality on the use deterioration process is an increase in q given by $\partial q = \alpha$. The second order effect is that this increase in users $\partial q = \alpha$ produces a use deterioration of the facility given by $-\beta\alpha$. For the same unit change of quality, the first order effect on the natural deterioration process will be a change $\partial S = -\alpha$; the second order effect on natural deterioration will be a change $\partial S = \alpha^2$. These results can be obtained directly from an analysis of (4-15) considering periods of time of unit length. The sign of Δ can be used to determine which of the two processes, natural deterioration or use deterioration, dominates.

It should be noted that the auxilliary or indicial equation associated with (4-7), the differential equation which determines the adjoint variable P_1, has roots

$$r_1 = \frac{\alpha}{2} + \sqrt{\frac{\alpha^2}{4} - \beta a} \quad \text{and} \quad r_2 = \frac{\alpha}{2} - \sqrt{\frac{\alpha^2}{4} - \beta a}. \qquad (4\text{-}19)$$

Clearly the discriminant of (4-19) is identical to that of (4-17), namely Δ as defined by (4-18). Thus it will suffice in the remainder of the analysis to consider the three cases $\Delta > 0$, $\Delta = 0$ and $\Delta < 0$. Of course, any solution for the adjoint variable P_1 based on a solution of (4-15) is, like S and q, valid everywhere except at the countable number of switching times.

Case 1. $\Delta > 0$. In this case the roots m_1 and m_2 are real and unequal; the solution of (4-10) becomes,

$$S = C_1\exp(m_1 t) + C_2\exp(m_2 t) - \frac{b}{a} . \qquad (4\text{-}20)$$

Simple integration of the expression for q leads to

$$q = C_3\exp(m_1 t) + C_4\exp(m_2 t) + C_5 \qquad (4\text{-}21)$$

where

$$C_3 = \frac{aC_1}{m_1} , \quad C_4 = \frac{aC_2}{m_2} .$$

Equation (4-20) and (4-21) give the values of the state variables S(t) and q(t) for each interval between switching points. The constants of integration C_1, C_2 and C_5 must be calculated for each interval using the corresponding boundary condition for S and q (see Section 5). In the present case, $\Delta > 0$, the natural deterioration process dominates the use deterioration process, and, therefore, the state of the system denoted as (S(t),q(t)), is explained by monotonic exponential functions. A typical example of the evolution of the system under these circumstances is given in Figure 7.4 for the case of one switching.

During the first period, that is when V(t) = M, the quality S(t) remains almost constant due to the influence of a high level of maintenance. When the switch to V(t) = m occurs, quality begins to decrease faster and q(t) experiences a second order change (an alteration of curvature); S(t) will tend to a stable position whose magnitude will depend on the magnitude of m. Also q(t) will eventually decrease, continue increasing or tend to level off depending on the amount of decrease experienced by S(t).

Case 2. $\Delta < 0$. In this case the roots m_1 and m_2 are complex conjugates and we write the solutions for S and q as

$$S = \exp(-\tfrac{\alpha}{2}t)[C_1\cos(\sqrt{|\Delta|}t) + C_2\sin(\sqrt{|\Delta|}t)] - \frac{b}{a}$$

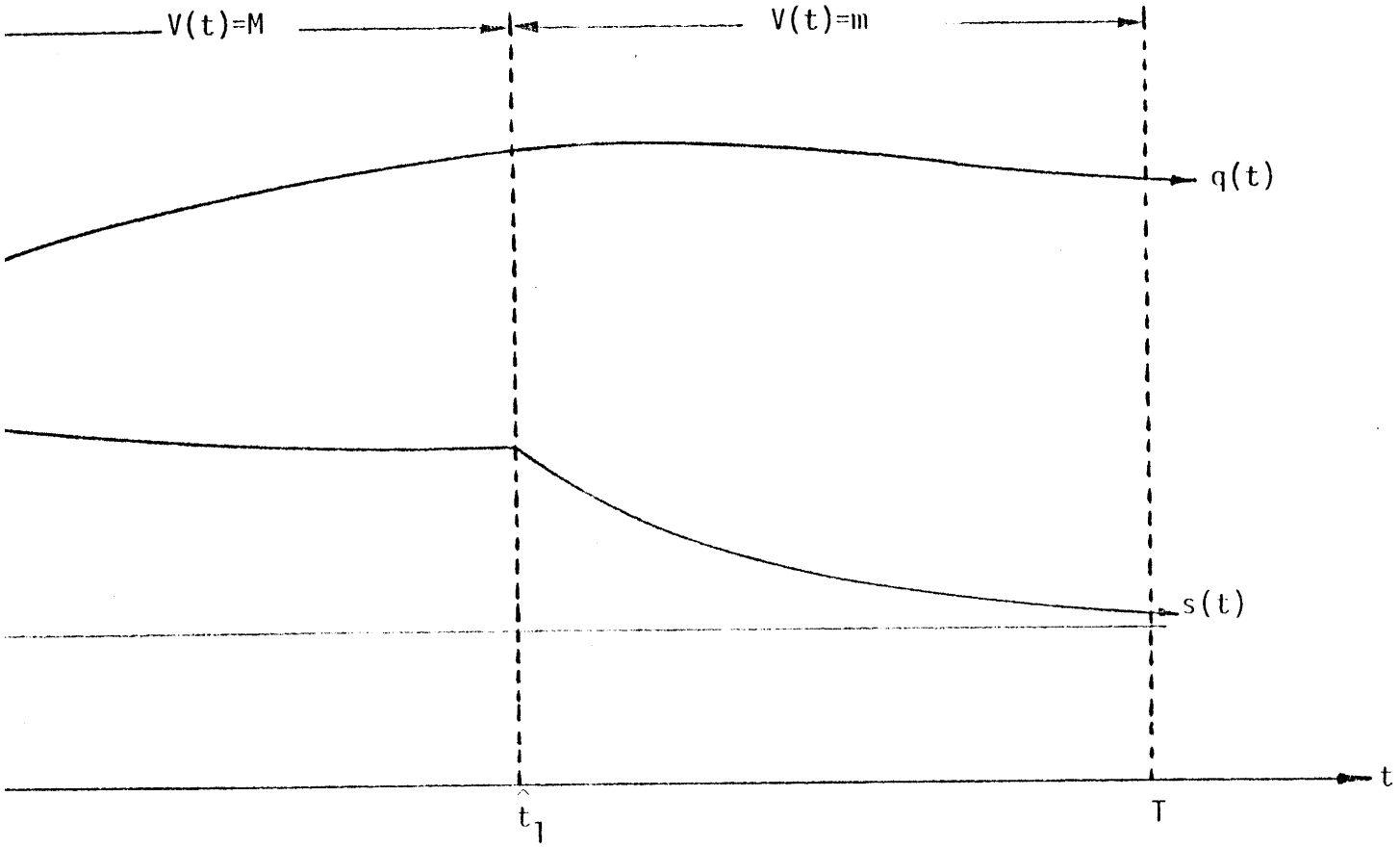
$$q = \exp(-\tfrac{\alpha}{2}t)[C_3\sin(\sqrt{|\Delta|}t) - C_4\cos(\sqrt{|\Delta|}t)] + C_5 \quad (4\text{-}22)$$

where

$$C_3 = \frac{2a(2C_1\sqrt{|\Delta|} - \alpha C_2)}{\alpha^2 + 4|\Delta|}$$

$$C_4 = \frac{2a(\alpha C_1 + 2\sqrt{|\Delta|}\,C_2)}{\alpha^2 + 4|\Delta|}. \quad (4\text{-}23)$$

Here, the use deterioration process dominates the natural deterioration process and the evolution of the system becomes oscilatory. This behavior can be easily explained as follows: given the high value of $\beta\alpha$,



ure 7.4. Exponential behavior of the system for the Case $\Lambda > 0$.

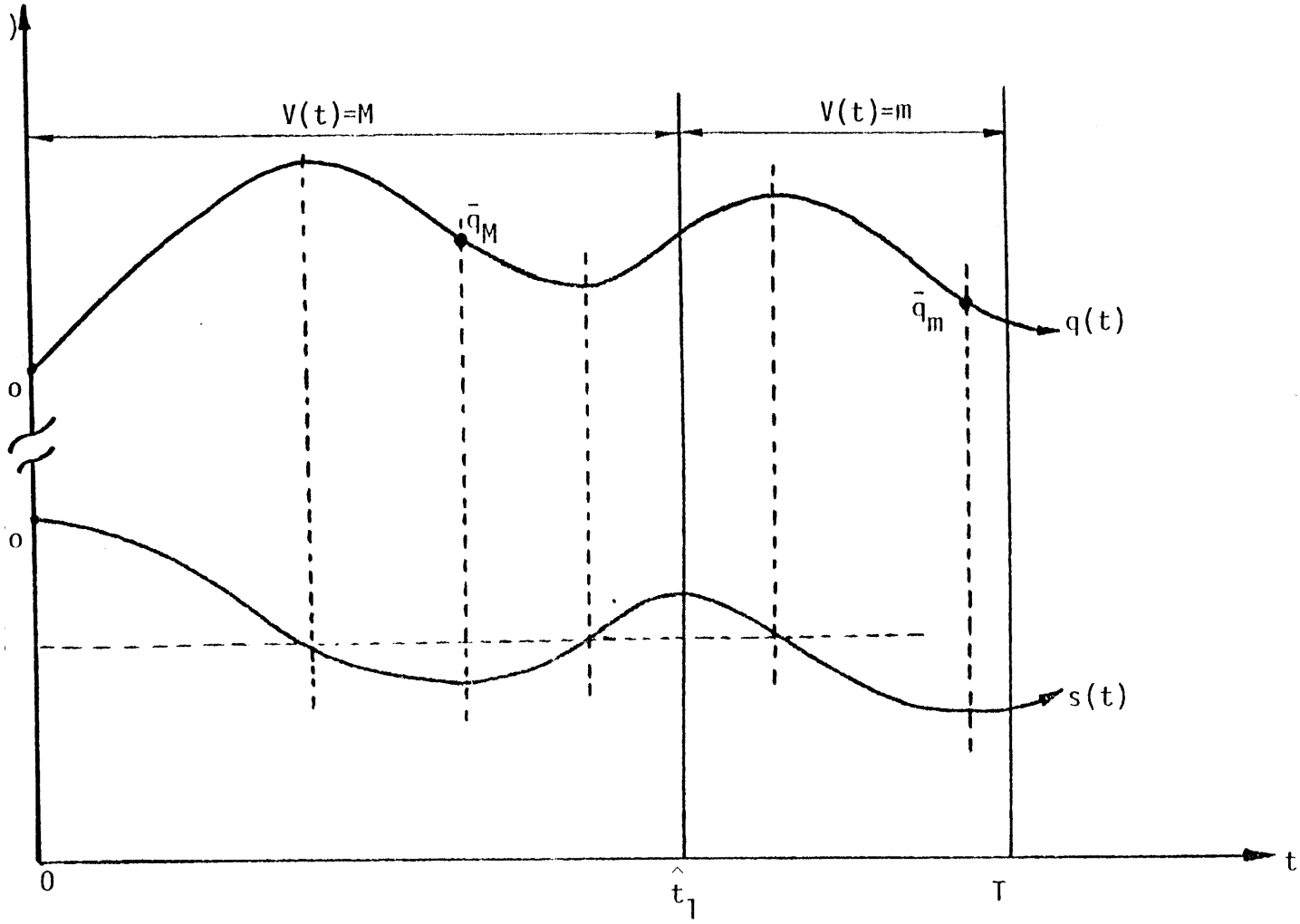


Figure 7.5 Oscilatory behavior for the case $\Lambda<0$

the deterioration produced by the users attracted as a consequence of improvements of quality (or lack of deterioration) due to maintenance can become higher than the quality improvement that generated the new users. In such a case, quality will reach a maximum and then decrease until the reduction of quality brings about a reduction of users such that quality can begin to increase again. An example of the evolution of the system when $\Delta < 0$ is given in Figure 7.5 for the case of one switching. In Figure 7.5, $\bar{S}$ represents the value of $S(t)$ for which $\dot{q}(t) = 0$; $\bar{q}_M$ is the value of $q(t)$ for which $\dot{S} = 0$ when $V(t) = M$; and $\bar{q}_m$ corresponds to $\dot{S} = 0$ for $V(t) = m$; that is

$$\bar{q}_M = \frac{\alpha S + \gamma M}{\beta} \; , \; \bar{q}_m = \frac{-\alpha S + \gamma m}{\beta} \; .$$

Case 3. $\Delta = 0$. This case is of little practical importance due to the fact it is not likely to occur. Nevertheless we present results analogous to those obtained for the other two cases for the sake of completeness. In this case the roots m_1 and m_2 are real and equal with value $-\alpha/2$. Consequently the following solutions obtain

$$\begin{aligned} S &= C_1 \exp(-\tfrac{\alpha}{2}t) + C_2 t \exp(-\tfrac{\alpha}{2}t) \\ q &= C_3 \exp(-\tfrac{\alpha}{2}t) + C_4 t \exp(-\tfrac{\alpha}{2}t) + C_5 \end{aligned} \qquad (4\text{-}24)$$

where

$$\begin{aligned} C_3 &= -\frac{2a}{\alpha}\left(C_1 - \frac{2C_2}{\alpha}\right) \qquad (4\text{-}25) \\ C_4 &= -\frac{2aC_2}{\alpha} \; . \end{aligned}$$

The switching functions denoted by F_s for the three cases analyzed are given by the following expressions:

Case 1. $\Delta > 0$.

$$F_s = C_6\exp(r_1 t) + C_7\exp(r_2 t) + C_{14}\exp[(m_1-\rho)t] + C_{15}\exp[(m_2-\rho)t] - C_{16}\exp[(2m_1-\rho)t] - C_{17}\exp[(2m_2-\rho)t] - C_{18}\exp[(m_1+m_2-\rho)t] + (C_{19} - \frac{1}{\gamma})\exp(-\rho t). \quad (4\text{-}26)$$

Case 2. $\Delta < 0$.

$$F_s = C_6\exp(\tfrac{\alpha}{2}t)\sin(\sqrt{|\Delta|}\,t + C_7) + C_{13}\exp[-(\tfrac{\alpha}{2}+\rho)t]\sin(\sqrt{|\Delta|}\,t - C_9) + C_{14}\exp[-(\tfrac{\alpha}{2}+\rho)t]\cos(\sqrt{|\Delta|}\,t - C_9) + C_{15}\exp[-(\alpha+\rho)t]\sin^2(\sqrt{|\Delta|}\,t - C_9) + C_{16}\exp[-(\alpha+\rho)t]\sin(\sqrt{|\Delta|}\,t - C_9)\cos(\sqrt{|\Delta|}\,t - C_9) + C_{17}\exp[-(\alpha+\rho)t]\cos^2(\sqrt{|\Delta|}\,t - C_9) + (C_{18} - \frac{1}{\gamma})\exp(-\rho t). \quad (4\text{-}27)$$

Case 3. $\Delta = 0$.

$$F_s = C_6\exp(\tfrac{\alpha}{2}t) + C_7 t\exp(\tfrac{\alpha}{2}t) + C_{14}\exp[-(\tfrac{\alpha}{2}+\rho)t] + C_{15}t\exp[-(\tfrac{\alpha}{2}+\rho)t] + C_{16}\exp[-(\alpha+\rho)t] + C_{17}t\exp[-(\alpha+\rho)t] + C_{18}t^2\exp[-(\alpha+\rho)t] + (C_{19} - \frac{1}{\gamma})\exp(-\rho t) \quad (4\text{-}28)$$

Detailed expressions for all the constants involved, in addition to the expressions for $F(t)$ and $P_1(t)$ that were used to derive F_s in each case are given in Appendix A. That appendix illustrates that a complete specification of F_s in general requires determination of values for C_1, C_2, C_5,

C_6, and C_7 since all other constants will then be defined. A procedure for doing this is presented in Section 5.

For Case 1, where $\Delta > 0$, an upper bound for the number of switchings can be established. The switching function (4-26) is a weighted sum of exponentials. Pontryagin, et al. [1962, p.120] proved that the number of zeros of a sum of n real exponential functions of a given variable is at most (n-1). Therefore, the switching function (4-26) will exhibit at most seven switchings. It is worth noting that for Case 1 the maximal number of switchings will depend on cost and demand characteristics. For instance, if no congestion costs are taken into account, i.e., if $\delta = 0$, we have that $C_{16} = C_{17} = C_{18} = 0$, and therefore the maximal number of switchings is four since (4-26) is reduced to the sum of five real exponentials. The same result is obtained if we assume that demand q is independent of the level of service, i.e. that $\alpha = 0$.

We might alternatively assume that operating cost is constant and independent of S and q, i.e. that $\epsilon = \delta = 0$, so that $C_{14} = C_{15} = C_{16} = C_{17} = C_{18} = 0$. Under this assumption expression (4-26) simplifies to

$$F_s = C_6 \exp(r_1 t) + C_7 \exp(r_2 t) + (C_{19} - \frac{1}{\gamma}) \exp(-\rho t), \ \Delta > 0 . \quad (4\text{-}29)$$

Expression (4-29) demonstrates that, in Case 1, for the assumption of constant operating cost the maximal number of switchings is two. The assumption of constant operating cost allows expressions (4-27) and (4-28) to be rewritten respectively as:

$$F_s = C_6\exp(\tfrac{\alpha}{2}t)\sin(\sqrt{|\Delta|}\,t+C_7) + (C_{18} - \tfrac{1}{\gamma})\exp(-\rho t),\ \Delta < 0. \qquad (4\text{-}30)$$

$$F_s = C_6\exp(\tfrac{\alpha}{2}t) + C_7 t\exp(\tfrac{\alpha}{2}t) + (C_{19} - \tfrac{1}{\gamma})\exp(-\rho t),\ \Delta = 0. \qquad (4\text{-}31)$$

We cannot, without additional information, state the maximal number of switchings admitted by the functions (4-30) and (4-31); we know only that this maximal number is finite for a fixed planning horizon. This model formulation, the model resulting from the assumption of constant operating cost, is the simplest that we can articulate; it has meaning only in the case when demand is dependent on service, i.e., when the parameter $\alpha \neq 0$ in equation (2-2), for benefits are then affected by maintenance policy only through the influence that quality of service has on the number of road users, who experience a constant individual cost which is independent of the state variables S and q. This is to be contrasted to the more general case where benefits are influenced by changes in the operating cost brought about by changes in the state variables; that is to say, the more general case exhibits both a demand and supply sensitivity. It should also be noted that if C(S,q) = C, a constant, and $\alpha = 0$, the system is completely uncontrollable and the maintenance policy V(t) has no effect on benefits, since it cannot impact either the supply or the demand side of the system. This characteristic of the model is made clear by noting that since $\alpha = \varepsilon = 0$ equation (4-9) with boundary condition from (4-12) leads to a shadow price of quality $P_1(t)$ equal to zero for all $t\varepsilon[0,T]$; consequently, the optimal maintenance policy given

by (3-8) will be $V(t) = m$ for all $t \in [0,T]$ since $[\gamma \exp(\rho t)]^{-1}$ will always be positive.

It is worth noting that the boundary condition $P_1(T) = 0$ implies that the optimum maintenance policy will always have a value $V(t) = m$ in a neighborhood of T (see (3-8)). This is a direct consequence of the fact that no conditions involving $S(T)$ (or $q(T)$), are specified in the model. In other words, we do not care about the final value of $S(t)$ as long as the total benefits over the period $[0,T]$ are maximized. This is a good assumption as long as the period $[0,T]$ comprises the whole economic life of the road analyzed. The time T can then be considered as the moment at which an investment that replaces the old road by a new one is made. If a period shorter than the whole economic life is considered a residual value representing unrealized benefits or salvage value should be included in the objective function. This residual value will naturally depend on the value $S(T)$. In that case the maximum principle says that the final value of $P_1(t)$, the value at $t = T$, must be equal to the derivative of the residual value function with respect to $S(t)$ evaluated at T. In this case the possibility that $P_1(T) > [\gamma \exp(\rho t)]^{-1}$ exists which would require that $V(t) = M$ in a neighborhood of T.

Sufficiency conditions for the maintenance policies presented in this section are given in Appendix B.

knowledge of $V^*(0)$ (optimal policy at $t = 0$) and to know this we need to know $P_1(0)$; b) condition 5 involves knowledge of $q(T)$ which we can only obtain after we know the whole history of $q(t)$ and therefore the switching times. In order to overcome these difficulties the following iterative procedure is proposed.

Step 1. Guess an initial value for the optimal maintenance policy $V^*(0)$ and a final value for the number of users $q(T)$. In order to facilitate these initial guessings, we can make use of the following information obtained from our analysis in Section 3:

- $V^*(0)$ can only take on the values m or M.
- $$V^*(0) = \begin{cases} m, \text{ if } P_1(0) < \gamma^{-1} \\ M, \text{ if } P_1(0) > \gamma^{-1}. \end{cases}$$
- The set of possible values of $q(T)$ is bounded due to the following:

$$V(t) = m, \forall t \in [0,T] \text{ implies } q(T) = q_\ell$$

$$V(t) = M, \forall t \in [0,T] \text{ implies } q(T) = q_u,$$

where q_ℓ is a lower bound and q_u an upper bound for $q(T)$. In addition we know that in our case $V(t) = m$ for t in a neighborhood of T, and therefore $q(T) < q_u(T)$. Obviously we also have $q_\ell(T) \geq 0$. Thus the following procedure is indicated:

a) Calculate $q_\ell(T)$ making use of conditions 1, 2, and 3 in (5-1) and chosing $V(t) = m$, $\forall t \in [0,T]$. Similarly calculate $q_u(T)$.

b) Choose $V(0) = 0$ and $q(T)$ such that $q_\ell(T) < q(T) < q_u(T)$.

c) Now we can calculate all the integration constants involved in the switching function F_s, valid for all $t < \hat{t}_1$, where $\hat{t}_1$ is the first switching point*. Therefore we can calculate $P_1(0)$.

d) Check now if $P_1(0) < \gamma^{-1}$ holds. If yes we have a compatible set of initial values V(o) and q(T). Otherwise change the value of V(o) (and q(T) if necessary) until a compatible initial set is found.

In all the calculations involving S, q, and P_1 we make use of the solutions and expressions for constants developed in Section 4.

Step 2. Find the smallest root of F_s; this will give us the first switching point $\hat{t}_1$. Given that the state variables must be continuous and that they are a function of V(t) (which is not continuous), we must force continuity at each switching point. This requires recalculating the constants of integration C_1, C_2, and C_5 for each solution interval. That is, we will have different values for these constants for each interval of time between successive switching points. In order to recalculate the constants, we make use of the following conditions

$$S(\hat{t}_i^-) = S(\hat{t}_i^+), \qquad q(\hat{t}_i^-) = q(\hat{t}_i^+) \tag{5-2}$$

$$\dot{S}(\hat{t}_i^+) = -\alpha S(\hat{t}_i^-) - \beta q(\hat{t}_i^-) + \gamma V(\hat{t}_i^+)$$

where $\hat{t}_i^-$ is the moment just before and $\hat{t}_i$ the moment just after the switching time $\hat{t}_i$. Let $q'(T)$ be the value of q obtained for t = T.

*See related comment in Step 2 below.

5.0 An Algorithm for Determining Switching Times

As we saw in Section 4, to obtain the values of the switching times in a particular case we need to calculate the values of five constants: C_1, C_2, C_5, C_6, C_7. In order to calculate these constants we make use of the following boundary conditions:

$$\begin{aligned} &1)\ S(0) = S_0 \\ &2)\ q(0) = q_0 \\ &3)\ \dot{S}(0) = -\alpha S_0 - \beta q_0 + \gamma V(0) \\ &4)\ p_1(T) = 0 \\ &5)\ \dot{p}_1(T) = \varepsilon q(T)\exp(-\rho T) \end{aligned} \qquad (5\text{-}1)$$

where the last condition is obtained from the adjoint equation for $P_1(t)$ (see (4-9)) using the boundary conditions (4-12).

The calculation of the five integration constants that we need to specify an optimal policy for the particular case described in Section 4, given the boundary conditions listed above in (5-1), constitutes a two-point boundary value problem. Finding solutions to this kind of problem generally requires the use of numerical methods that involve iterative procedures. An iterative procedure that makes use of the special characteristics of our problem will be described here. The method proposed could be classified within the category of neighboring extremal methods (Bryson and Ho [1975])

If we analyze our boundary conditions (5-1) we can immediately see that two main difficulties appear with respect to the use of them in the calculation of our integration constants: a) condition 3 involves

Successive switching points are found by applying these continuity conditions and solving $F_S = 0$ repeatedly for the smallest root greater than the previously considered switching point.

Step 3. Compare the calculated value $q'(T)$ with $q(T)$, the value guessed in Step 1. In particular execute the following logic:

a) Calculate $q'(T) - q(T) = \Delta q$. If $\Delta q \leq \eta$ stop, where η is a preset tolerance. If $\Delta q > \eta$ continue.

b) Let $q(T) = q(T) + \tau \Delta q$ with $\tau \leq 1$ and return to Step 1.

The values of τ must be chosen at each iteration in such a way that convergence is assured; $|\tau| = 1$ can produce overshooting and a value of $|\tau|$ too small will slow the convergence process.

Note that because we use throughout the algorithm the general solutions obtained in Section 4 from the application of the necessary conditions provided by the maximum principle, a solution obtained during any iteration satisfies the system equations (2-1) and (2-2), the adjoint equations and boundary conditions (3-2), (3-3) and (3-4) and condition (3-5) requiring maximization of the Hamiltonian. The algorithm then tries to find a value of $\dot{P}_1(T)$ for the last boundary condition in (5-1) that is compatible with the other four conditions in (5-1). A special case arises if we assume that the operating cost function is independent of the quality variable S, that is if

$$C = c + \delta q^2, \quad \varepsilon = 0.$$

Under this circumstance the boundary condition 5 in (5-1) becomes $\dot{P}_1(T) = 0$, and the optimum policy can be found in one iteration of the algorithm, since no guessing of $q(T)$ is necessary. This can provide an alternative way for guessing an approximate initial value for $q(T)$ in a general problem with $\varepsilon \neq 0$.

APPENDIX A: Integration Constants

(Chapter VII)

CASE 1. $\Delta > 0$

Using (4-20) and (4-21), the forcing function $F(t)$ in (4-14) may be written as

$$\begin{aligned} F(t) = {} & C_8\exp[(m_1-\rho)t] + C_9\exp[(m_2-\rho)t] \\ & -C_{10}\exp[(2m_1-\rho)t] - C_{11}\exp[(2m_2-\rho)t] \\ & -C_{12}\exp[(m_1+m_2-\rho)t] + C_{13}\exp(-\rho t) \end{aligned} \qquad \text{(A-1)}$$

where

$$\begin{aligned} C_8 &= \rho\varepsilon C_3 - 6a\delta C_5 C_3 \\ C_9 &= \rho\varepsilon C_4 - 6a\delta C_4 C_5 \\ C_{10} &= 3a\delta C_3^2 \quad , \quad C_{11} = 3a\delta C_4^2 \\ C_{12} &= 6a\delta C_3 C_4 \\ C_{13} &= aB - \varepsilon b + \rho\varepsilon C_5 . \end{aligned}$$

From this result we are led to a general solution for equation (4-13) of the following form since the roots r_1 and r_2 are real and unequal:

$$\begin{aligned} P_1 = {} & C_6\exp(r_1 t) + C_7\exp(r_2 t) + C_{14}\exp[(m_1-\rho)t] \\ & + C_{15}\exp[(m_2-\rho)t] - C_{16}\exp[(2m_1-\rho)t] - C_{17}\exp[(2m_2-\rho)t] \\ & - C_{18}\exp[(m_1+m_2-\rho)t] + C_{19}\exp(-\rho t) \end{aligned} \qquad \text{(A-2)}$$

where

$$C_{14} = \frac{\rho\varepsilon C_3 - 6a\delta C_5 C_3}{(m_1-\rho)^2 - \alpha(m_1-\rho) + \beta a}$$

$$C_{15} = \frac{\rho\varepsilon C_4 - 6a\delta C_4 C_5}{(m_2-\rho)^2 - \alpha(m_2-\rho) + \beta a}$$

$$C_{16} = \frac{3a\delta C_3^2}{(2m_1-\rho)^2 - \alpha(2m_1-\rho) + \beta a}$$

$$C_{17} = \frac{3a\delta C_4^2}{(2m_2-\rho)^2 - \alpha(2m_2-\rho) + \beta a}$$

$$C_{18} = \frac{6a\delta C_3 C_4}{(m_1+m_2-\rho)^2 - \alpha(m_1+m_2-\rho) + \beta a}$$

$$C_{19} = \frac{aB - \varepsilon b + \rho\varepsilon C_5}{\rho^2 + \alpha\rho + \beta a}$$

and from the solution for the state variables

$$C_3 = \frac{aC_1}{m_1} \text{ and } C_4 = \frac{aC_2}{m_2}.$$

CASE 2. $\Delta < 0$

Now using (4-22) and (4-23) the forcing function $F(t)$ in (4-14) may be written as

$$\begin{aligned} F(t) = {} & C_{10}\exp[-(\tfrac{\alpha}{2} + \rho)t] \sin(\sqrt{|\Delta|}\, t - C_9) \\ & + C_{11}\exp[-(\alpha+\rho)t] \sin^2(\sqrt{|\Delta|}\, t - C_9) \\ & + C_{12}\exp(-\rho t) \end{aligned} \qquad \text{(A-3)}$$

where

$$C_{10} = (\rho\varepsilon C_8 - 6a\delta C_5 C_8), \quad C_{11} = -3a\delta C_8^2$$

$$C_{12} = (a\beta - \varepsilon b + \rho\varepsilon C_5 - 3a\delta C_5^2)$$

$$C_8 = \sqrt{C_3^2 + C_4^2}, \qquad \tan C_9 = \frac{C_4}{C_3}.$$

From this result we are led to a general solution of (4-13) of the following form since the roots r_1 and r_2 are imaginary and unequal:

$$P_1 = C_6 \exp(\tfrac{\alpha}{2}t)\sin(\sqrt{|\Delta|}\,t + C_7) + C_{13}\exp[-(\tfrac{\alpha}{2}+\rho)t] \qquad (A\text{-}4)$$

$$\sin(\sqrt{|\Delta|}\,t - C_9) + C_{14}\exp[-(\tfrac{\alpha}{2}+\rho)t]\cos(\sqrt{|\Delta|}\,t - C_9)$$

$$C_{15}\exp[-(\alpha+\rho)t]\sin^2(\sqrt{|\Delta|}\,t - C_9) + C_{16}\exp[-(\alpha+\rho)t]$$

$$\sin(\sqrt{|\Delta|}\,t - C_9)\cos(\sqrt{|\Delta|}\,t - C_9) + C_{17}\exp[-(\alpha+\rho)t]$$

$$\cos^2(\sqrt{|\Delta|}\,t - C_9) + C_{18}\exp(-\sigma t)$$

with

$$C_{13} = \frac{K_1(\rho\varepsilon C_8 - 6a\delta C_5 C_8)}{K_1^2 + K_2^2}$$

$$C_{14} = \frac{K_2(\rho\varepsilon C_8 - 6a\delta C_5 C_8)}{K_1^2 + K_2^2}$$

$$C_{15} = -\frac{3a\delta C_8^2}{K_3 + 2|\Delta|}\left[\frac{2K_4^2 + (K_3 - 2|\Delta|)^2 + 2|\Delta|(K_3 - 2|\Delta|)}{4K_4^2 + (K_3 - 2|\Delta|)^2}\right]$$

$$C_{17} = -\frac{3a\delta C_8^2}{K_3 + 2|\Delta|}\left[\frac{2K_4^2 - 2|\Delta|(K_3 - 2|\Delta|)}{4K_4^2 + (K_3 - 2|\Delta|)^2}\right]$$

$$C_{16} = -\frac{1}{K_4}(3a\delta C_8^2 + C_{15}K_3 + 2|\Delta|C_{17})$$

$$C_{18} = \frac{(aB - \varepsilon b + \rho\varepsilon C_5 - 3a\delta C_5^2)}{\rho^2 + \alpha\rho + a\beta}$$

$$K_1 = [(\tfrac{\alpha}{2} + \rho)^2 + \alpha(\tfrac{\alpha}{2} + \rho) + a\beta - |\Delta|]$$

$$K_2 = [2\sqrt{|\Delta|}\ (\tfrac{\alpha}{2} + \rho) + \alpha\sqrt{|\Delta|}]$$

$$K_3 = [(\alpha + \rho)^2 + \alpha(\alpha + \rho) - 2|\Delta| + a\beta]$$

$$K_4 = [2\sqrt{|\Delta|}\ (\alpha + \rho) + \alpha\sqrt{|\Delta|}]$$

and from the solution for the state variables

$$C_3 = \frac{2a(2C_1\sqrt{|\Delta|} - \alpha C_2)}{\alpha^2 + 4|\Delta|}$$

$$C_4 = \frac{2a(\alpha C_1 + 2\sqrt{|\Delta|}\, C_2)}{\alpha^2 + 4|\Delta|}$$

CASE 3. $\Delta = 0$.

Using (4-24) and (4-25), the forcing function $F(t)$ in (4-14) may be written as

$$\begin{aligned} F(t) = {} & C_8\exp[-(\tfrac{\alpha}{2} + \rho)t] + C_9 t\exp[-(\tfrac{\alpha}{2} + \rho)t] \\ & -C_{10}\exp[-(\alpha+\rho)t] - C_{11} t\exp[-(\alpha+\rho)t] \\ & -C_{12}t^2\exp[-(\alpha+\rho)t] + C_{13}\exp(-\rho t) \end{aligned} \qquad (A\text{-}5)$$

where

$$C_8 = [\rho\varepsilon C_3 - 6a\delta C_3C_5]$$

$$C_9 = [\rho\varepsilon C_4 - 6a\delta C_4C_5]$$

$$C_{10} = 3a\delta C_3^2, \quad C_{11} = 6a\delta C_3C_4$$

$$C_{12} = 3a\delta C_4^2$$

$$C_{13} = (aB - \varepsilon b + \rho\varepsilon C_5 - 3a\delta C_5^2).$$

From this result we are led to a general solution for equation (4-13) of the following form since the roots r_1 and r_2 are real and equal with value $\alpha/2$:

$$\begin{aligned} P_1 = {} & C_6\exp(\tfrac{\alpha}{2}t) + C_7 t\exp(\tfrac{\alpha}{2}t) + C_{14}\exp[-(\tfrac{\alpha}{2}+\rho)t] \\ & + C_{15}t\exp[-(\tfrac{\alpha}{2}+\rho)t] + C_{16}\exp[-(\alpha+\rho)t] \\ & + C_{17}t\exp[-(\alpha+\rho)t] + C_{18}t^2\exp[-(\alpha+\rho)t] \\ & + C_{19}\exp(-\rho t) \end{aligned} \qquad (A\text{-}6)$$

where

$$C_{14} = \frac{1}{K_1}(\rho\varepsilon C_3 - 6a\delta C_5C_3) + \frac{2}{K_1^2}(\alpha+\rho)(\rho\varepsilon C_4 - 6a\delta C_4C_5)$$

$$C_{15} = \frac{1}{K_1}(\rho\varepsilon C_4 - 6a\delta C_4C_5)$$

$$C_{16} = \frac{1}{K_2^3}[6a\delta C_4^2(3\alpha+2\rho)^2] - \frac{1}{K_2^2}[6a\delta C_3C_4(3\alpha+2\rho)]$$

$$+ \frac{1}{K_2^2}\,6a\delta C_4^2 \qquad - \frac{1}{K_2}(3a\delta C_3^2)$$

$$C_{17} = \frac{1}{K_2^2}\left[6a\delta C_4^2(3\alpha+2\rho)\right] - \frac{1}{K_2}(6a\delta C_3 C_4)$$

$$C_{18} = -\frac{1}{K_2}(3a\delta C_4^2)$$

$$C_{19} = \frac{(aB - \varepsilon b + \rho\varepsilon C_5 - 3a\delta C_5^2)}{(\rho^2 + \alpha\rho + \beta a)}$$

$$K_1 = \left[\left(\frac{\alpha}{2} + \rho\right)^2 + \alpha\left(\frac{\alpha}{2} + \rho\right) + \beta a\right]$$

$$K_2 = \left[(\alpha + \rho)^2 + \alpha(\alpha + \rho) + \beta a\right]$$

and from the solution for the state variables

$$C_3 = -\frac{2a}{\alpha}\left(C_1 - \frac{2C_2}{\alpha}\right)$$

$$C_4 = -\frac{2aC_2}{\alpha} .$$

APPENDIX B: Sufficient Conditions

Solutions of the form described in Section 4 satisfy all the conditions established by the maximum principle. Nevertheless, these are only necessary conditions and therefore any solution derived from them is only a candidate for optimality. In this section we will analyze the circumstances which cause the policies described in Section 4 to produce a maximum value of J(V(t)) in (2-5). With this purpose in mind, we will use the Arrow sufficiency theorem (see Section 5 of Chapter II)

The Arrow sufficiency theorem, applied to our problem, says that a policy $[V^*(t), S^*(t), q^*(t)]$, obtained from the necessary conditions provided by the maximum principle, will produce a global maximum of J(V(t)) in (2-5) if $H^*(S, q, P_1, P_2, t)$ is concave in the state variables S and q for all $t \in [0,T]$, where

$$H^*(S, q, P_1, P_2, t) = \underset{V \in \Omega}{\text{Max}}\ H(S, q, P_1, P_2, V, t),\ t \in [0,T]. \tag{B-1}$$

The results of Section 4 allow us to distinguish three different cases which lead to distinct expressions for H^*:

a) Bang-Bang case with $V^* = M$.

$$H^* = (U-C)q \exp(-\rho t) - P_1(\alpha S + \beta q) + P_2(\alpha S + b)$$

$$+ [\gamma P_1 - \exp(-\rho t)]M \tag{B-2}$$

b) Bang-Bang case with $V^* = m$.

$$H^* = (U-C)q \exp(-\rho t) - P_1(\alpha S + \beta q) + P_2(\alpha S + b)$$

$$+ [\gamma P_1 - \exp(-\rho t)]m \qquad \text{(B-3)}$$

c) Singular case, $V^* = V_S^*$.

$$H^* = (U-C)q \exp(-\rho t) - P_1(\alpha S + \beta q) + P_2(\alpha S + b) \qquad \text{(B-4)}$$

Note that in this last case $\gamma P_1 = \exp(-\rho t)$. To determine the conditions for which H^* is concave we must examine the quadratic form

$$[q,S]\ H''\ \begin{bmatrix} q \\ S \end{bmatrix} \qquad \text{(B-5)}$$

Where H'' is the so-called Hessian matrix associated with H^*. For the example problem of Section 4 we have

$$H'' = \exp(-\rho t) \begin{bmatrix} -6\delta q & \varepsilon \\ \varepsilon & 0. \end{bmatrix} \qquad \text{(B-6)}$$

For H^* to be concave the quadratic form (B-5) must be negative definite. Thus we require for the example problem of Section 4

$$-6\delta q^3 + 2\varepsilon Sq \quad < 0. \qquad \text{(B-7)}$$

Since $q > 0$ this becomes

$$-3\delta q^2 + \varepsilon S \qquad < 0. \tag{B-8}$$

Note that (4-1) describes the cost function for the example problem and can be restated as

$$\frac{C - c}{\delta} + \frac{\varepsilon}{\delta} S = q^2. \tag{B-9}$$

If $C - c \geq 0$ then clearly

$$\frac{\varepsilon}{\delta} S \leq q^2. \tag{B-10}$$

and also

$$\frac{1}{3} \frac{\varepsilon}{\delta} S < q^2. \tag{B-11}$$

This last expression is exactly that obtained from (B-8) in solving for q^2. Thus if total costs are always greater than or equal to the costs which are incurred independent of any quality or congestion effects H^* is concave and the Pontryagin necessary conditions are sufficient.

There are clearly other circumstances under which (B-8) holds and for which the Pontryagin necessary are sufficient; the essential point is that (B-8) is satisfied in virtually all cases of practical importance.

VIII. SUMMARY OF RESULTS AND CONCLUSIONS

As mentioned in the introduction, the objective of this work was the study of dynamic optimal investment policies in public facilities with special consideration given to the transportation case.

After presenting in Chapter II the methodology and the mathematical results to be used in the rest of the study, Chapter III to V were dedicated to developing dynamic models for the analysis of optimal investments in quality and capacity for both the cases of continuous and discrete quality/capacity variables. This analysis led to dynamic optimal investment policies which were given economic interpretations. The policies derived are general and do not depend on particular forms of the functions involved in the specification of the models. The results obtained served to provide important new insights about the structure and characteristics of optimal investment policies.

In Chapter III optimal investment policies in quality were derived both for the case in which demand is assumed externally specified and for the case when it is related to the quality of the facility. The introduction of this interrelationship in the specification of the model, which had proven impossible before in all the static investment models used in the literature, was easily accomplished through the definition of a dynamic equation for the changes in demand over time.

In Chapter IV a dynamic model for continuous investments in capacity was developed and solved. Explicit expressions for the optimal

value of capacity investments were obtained using general statements of the construction and operating cost functions involved. These expressions had not been previously cited in the literature. The dynamic characteristics of optimal investment policies were also obtained for the first time and new sufficiency conditions were developed. At the end of the Chapter, the application of the results obtained was illustrated for different special cases of interest.

In Chapter V a general dynamic model was proposed to handle investment discontinuities. The general structure and characteristics of optimal investment policies in capacity and quality were obtained under this circumstance and given economic interpretation. It was shown that the solution of the general problem in a particular case requires the use of iterative numerical methods. However, practical marginal rules were obtained for special cases.

Chapters VI and VII were dedicated to the development of particular model formulations in order to study some special cases of interest. In ChapterVI we analyzed the implications of explicitly considering the interrelationship between level of service and demand on the time staging of optimal investment decisions. The analysis lead to a new time staging rule that considers the interrelationships mentioned above and contains as a special case the naive static rule previously proposed in the literature. A numerical example presented at the end of the chapter was used to illustrate the dramatic effect that demand quality interrelationships can have on optimal time staging policies.

Finally, Chapter VII was dedicated to showing how the theoretical results of Chapter III can be used to obtain optimal maintenance policies for a road in a special case. Explicit analytic expressions were obtained for the optimal solution in terms of the parameters of the problem and an algorithm to find numerical solutions was proposed.

Throughout the development of this work it has been shown that modern control theory constitutes a powerful and useful tool for the analysis of dynamic investment policies. As we have seen it allows the development of more realistic models and the obtainment of results impossible to get from static formulations. An apparent drawback of the application of this technique has been the relative difficulty confronted in the interpretation of the results obtained. Nevertheless, this should not be surprising given that we are dealing with more complete and general results than previously adhered to in the literature and also from the fact that we are not commonly used used to thinking in dynamic terms. If a static formulation corresponds to a simplification of the system under analysis, it is obvious that the results obtained will be, though less general and complete, easier to apprehend. We have also shown that special results of the theory like those refering to singular controls and to model specifications with discontinuities in the state variables and dynamic equations can be fruitfully used in order to obtain more insightfull results or to analyze problems with special structure. These approahces had not been used before in the economic literature; it is obvious that a wide field is open for future research. Other model formulations with alternative dynamics and objective function specifications other than

those used here could be developed. Some interesting extensions could be the analysis of multimodel systems and multiobjective specifications. Formulations for the explicit study of dynamic pricing policies could be also tried. In addition, the developments of numerical methods for the solution of special models of interest could be undertaken. Hopefully, the analyses presented here will serve as a motivation for further study of dynamic investment policies from a microeconomic point of view.

REFERENCES

Allen, R.G.D., (1971), Mathematical Analysis for Economists. St. Martin's Press, New York.

Arora, S.R. and P.T. Lele, (1970), A Note on Optimal Maintenance Policy and Sale Date of a Machine, Management Science, Vol. 17, No. 3, pp. 170-173.

Arrow, K.J., (1968), Applications of Control Theory to Economic Growth, in G.B. Dantzig and A.F. Veinott, Jr., eds. Mathematics of the Decision Sciences. American Mathematical Society, Providence, R.I.

Athans, M. and P.L. Falb, (1966), Optimal Control. McGraw-Hill, New York.

Beenhakker, H. and Danskin J., (1973), Economies of Stage Construction for Transport Facilities, Transportation Research, Vol. 7, pp. 163-178.

Bellman, R., (1957), Dynamic Programming. Princeton University Press, Princeton, New Jersey.

Bensoussan, A., G.E. Hurst, and B. Näslund, (1974), Management Applications of Modern Control Theory. North-Holland/American Elsevier, Amsterdam/New York.

Breakwell, J.V.,(1959), The Optimization of Trajectories. SIAM Journal, Volume 7.

Bryson, A.E., and Yu-Chi Ho, (1975), Applied Optimal Control. Wiley, New York.

Büttler, H.J., and J.H. Shortreed, (1978), Investment Planning of a Road Link. Transportation Research, Volume 12, pp. 357-366

COMSIS Corporation, (1972), Traffic Assignment: Methods-Application-Products, prepared for Urban Planning Division, Office of Planning, Federal Highway Administration.

DeNeufville, R.L., (1969), Optimal Highway Staging by Dynamic Programming, Cambridge, M.I.T., Department of Civil Engineering.

Highway Research Board, (1962), The AASHO Road Test. Special Report 61E, Publication No. 954, National Research Council, Washington, D.C.

Keeler, T.E., Small, K.A., Cluff, G.S., and Finke, J.K., (1975), Optimal Peakload Pricing, Investment, and Service Levels on Urban Expressways. Working Paper No. 253, Institute of Urban and Regional Development, University of California, Berkeley.

Kelley, H.J., R.E. Kopp, and A.G. Moyer, (1966), Singular Extremals, Topics in Optimization, G. Leitmann (ed.), Volume II, Chapter 3. Academic Press, New York.

Luenberger, D.G., (1973), Introduction to Linear and Non-linear Programming. Addison-Wesley Publishing Company, Reading, Massachusetts.

Mangasarian, O.L., (1966), Sufficient Conditions for the Optimal Control of Non-Linear Systems. SIAM Journal on Control, IV, pp. 139-152.

Manheim, M.L., (1979), Fundamentals of Transportation System Analysis. In print, M.I.T. Press, Cambridge.

Marglin, S.A., (1963), Approaches to Dynamic Investment Planning. North-Holland Publishing Company, Amsterdam.

Miquel, S., (1972), Caminos. Dept. Obras Civiles, Universidad de Chile, Santiago.

Mohring, H., and Harwitz, (1962), Highway Benefits: An Analytical Framework. Northwestern University, Evanston, Illinois.

Näslund, B., (1966), Simultaneous Determination of Optimal Repair Policy and Service Life. The Swedish Journal of Economics, Vol. LXVIII., pp. 63-73.

Pontryagin, L.S., V.A. Boltyanski, R.V. Gankrelidge, and E.F. Mishenko, (1962), The Mathematical Theory of Optimal Processes. Wiley, New York.

Samuelson, P.A., (1947), Foundations of Economic Analysis. Eighth Printing, Atheneum, New York.

Seierstad, A., and K. Sydsaeter, (1977), Sufficient Conditions in Optimal Control Theory. International Economic Review, Vol. 18, No. 2, pp. 367-391.

Stroz, R., (1965), Urban Transportation Parables, in Julius Margolis, ed., The Public Economy of Urban Communities, Johns Hopkins University Press, pp. 452-465.

Tait, K., (1965), Singular Problems in Optimal Control. Ph.D. Thesis, Harvard University, Cambridge.

Thompson, G.L., (1968), Optimal Maintenance Policies and Sale Date of a Machine. Management Science, Volume 14, No. 9, pp. 543-550.

Varian, H.R., (1978), Microeconomic Analysis. W.W. Norton and Company, Inc., New York.

Venezia, I., (1977), Optimal Policies of Stage Construction for Transportation Facilities Under Uncertainty. Transp. Res. 11, pp. 377-383.