OPTIMAL REGULATION
OF
FINITE MARKOV CHAINS

by

John Michael Howl , M.A.

A thesis submitted for the degree
of Doctor of Philosophy and for
the Diploma of Imperial College

December 1977

Department of Electrical Engineering,
Imperial College,
University of London.

J.M. Howl

Optimal Regulation of Finite Markov Chains

ABSTRACT

This thesis investigates the problem of choosing optimal feedback control laws for Markov chains and semi-Markov chains with controllable transition mechanisms. Except in certain special cases the optimal control law cannot be determined analytically and it is necessary to make use of numerical procedures for optimization. The principal questions of interest concerning such procedures are (i) the conditions under which they may be used, and (ii) their computational efficiency.

The thesis surveys existing standard optimization procedures and proposes new algorithms with certain computational advantages. The most important of the new algorithms is a policy-iteration algorithm in which the control law is iteratively improved by a convergent sequence of single-component changes. The other new algorithm is an accelerated version of a successive-approximations procedure, in which the acceleration factor varies from iteration to iteration. In the development and analysis of these new procedures considerable use is made of the concept of equivalence between semi-Markov chains, and it is shown that use of the equivalence concept makes possible the extension of many results concerning Markov chains to the more general semi-Markov case.

The special transition structure of chains of the birth-death type are shown to permit a certain amount of simplification in the optimization algorithms. In addition, it is shown that for such chains it is possible to determine optimal quantized control laws by applying a modified form of the new policy-iteration algorithm to a certain embedded semi-Markov chain. The problem of state-space truncation for unrestricted birth-death chains is also investigated.

The thesis concludes with the computational study of a specific optimization problem of the birth-death type, in which the control switching cost is a significant component of the total cost.

ACKNOWLEDGEMENTS

I am indebted to Professor J.H. Westcott for giving me the opportunity to register for a part-time higher degree and for encouraging me to work in the general field of stochastic control; to Professor A. Tustin, whose suggestion that I should take an interest in traffic control problems led indirectly to the work described in this thesis; and, above all, to my supervisor, Professor D.Q. Mayne, for his interest in the project and for his active encouragement throughout its course.

I am very grateful to Mrs. Christine Lourdan for typing this thesis with her usual efficiency and accuracy.

Finally, I should like to acknowledge the unflagging support I have received from my wife, Felicity, whose patience in the face of a receding completion date has been remarkable.

J.M.H.

CONTENTS

CHAPTER 1

INTRODUCTION

The growth of man's knowledge and understanding of the world is characterized in the main by the gradual accumulation of factual knowledge and the gradual evolution of ideas. From time to time, however, the development of a subject is marked by a new discovery or concept which in retrospect is seen to have been of crucial importance. Such a step forward occurred in optimization theory twenty years ago when Richard Bellman published his theory of multistage decision processes in the book entitled "Dynamic Programming" (1957). The concept of dynamic programming has been extremely fruitful in a wide variety of optimization problems, particularly in the fields of economics, operations research, control engineering, and statistical decision theory.

The dynamic programming technique is principally of value when the decision process to be optimized has sufficient structure to permit some simplification of the basic optimality equations - the so-called functional equations of dynamic programming. One class of such processes is the class of Markovian decision processes, introduced by Bellman(1957) and the subject of intensive study during the past fifteen years. It has been found that a Markovian decision process is the appropriate mathematical model in a wide variety of sequential decision problems in the fields of operations research and management science; in particular, problems concerning the control of queueing systems, the control of material stocks (or "inventory control", in U.S. terminology), the dynamic scheduling of resources, and many related problems, can often be modelled in this way. In addition, since a Markovian decision process is a special type of controllable stochastic process, such proccsses are of interest in stochastic control theory as models for physical systems with noisy dynamics.

Roughly speaking, a Markovian decision process is a Markov process whose transition mechanism is directly dependent on the value of an externally controllable variable called the control action or decision. The main problem of interest with such processes is to choose a sequence of decisions that will result in good, preferably optimal, process behaviour. Since the process is stochastic the sequence of decisions must be generated by feedback: that is, by making each decision a function of the available information about the current state of the process. The set of rules by which the decisions are related to the feedback information is called a feedback control policy; and a control policy which minimizes (maximizes) some suitably-defined cost (return) function is an optimal control policy.

This thesis is concerned with procedures for deriving optimal control policies for Markovian decision processes having a finite number of possible states and a finite number of possible decisions. The restriction to a finite state set is a natural one in many operations research applications and at the same time permits the utilization of the classical theory of Markov chains in the analysis of the optimization problem. For example, Markov chain theory tells us that, under certain minor restrictions on the transition mechanism, a chain will tend to move towards a natural statistical equilibrium in which the average proportion of time spent in each state is easily computed from the transition properties of the chain. For such processes an optimization problem of major interest is the problem of choosing a control policy which will result in the minimum (maximum) average cost (return) per unit time when the process is in statistical equilibrium. In the language of control theory this is an optimal regulation problem, and it is this type of problem in particular with which the thesis is primarily concerned.

1.1 An optimal regulation problem

As an example of an optimal regulation problem which can be formulated in terms of a finite-state Markovian decision model, consider the simple first-order control system shown in Fig.(1).

As usual, the aim is to make the integrator output y(t) follow the reference input $y_R(t)$ as closely as possible. The inputs to the integrator are the error-driven control action u(t) and the Gaussian white noise input w(t), assumed to have zero mean and autocorrelation function $\sigma^2 \delta(\tau)$, where $\delta(\tau)$ denotes the unit delta function. If $y_R(t) = y_o$ = constant, the system is a regulator which in the absence of the noise w(t) would maintain the output y(t) at the required constant level y_o. In the presence of w(t) the regulator settles down to a statistical equilibrium in which y(t) fluctuates about the value y_o with a variance of $\sigma^2/_{2K}$. This equilibrium output variance can clearly be made as small as desired by increasing the control gain K. But the equilibrium variance of the control action u(t) is $\frac{\sigma^2 K}{2}$ so that large values of K result in large control efforts which are normally undesirable. Let us therefore choose K so as to minimize the quadratic cost function,

$$L(K) = E\left[e^2(t)\right] + \alpha_o . E\left[u^2(t)\right]$$

where $E\left[\;\right]$ denotes expected value and α_o is a weighting factor.

The resulting problem is an example of the well known stochastic regulator problem (see, for example, Kwakernaak and Sivan(1972)) and the solution in this case is that the optimal value of K is $\alpha_o^{-\frac{1}{2}}$, the corresponding value of L being $\alpha_o^{\frac{1}{2}} \sigma^2$.

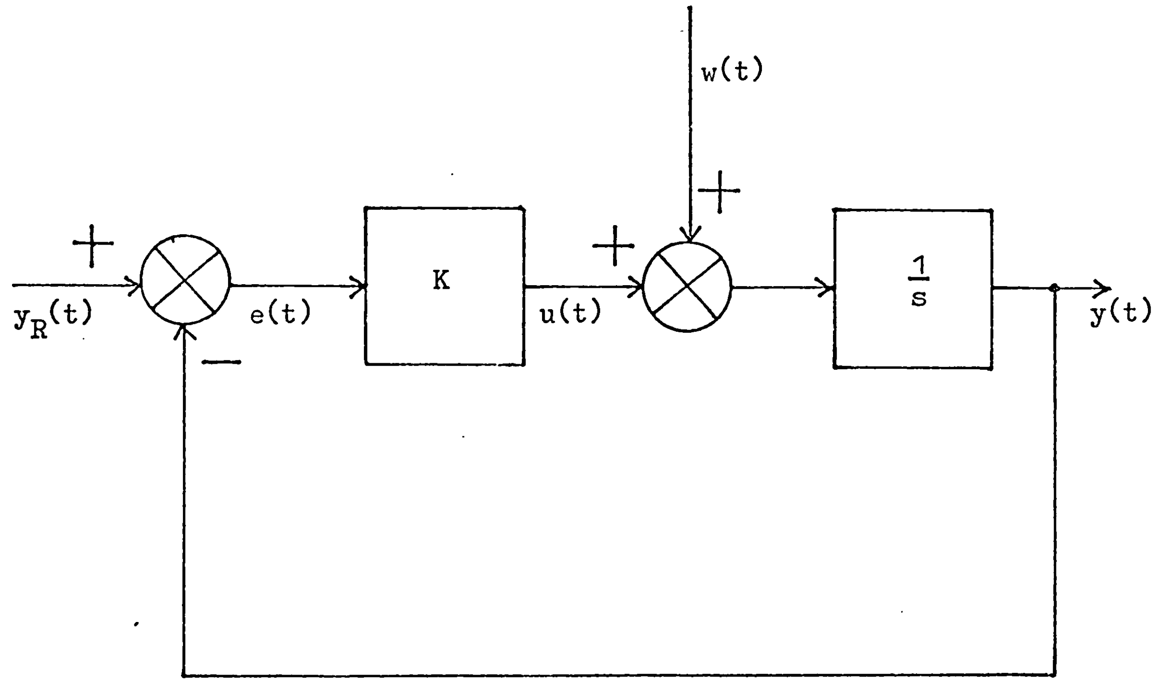


Fig. (1) A first-order linear control system

Suppose now that the error e(t) is quantized, so that the error input to the controller, $e_q(t)$, is given by the uniform quantization law:

$$\begin{aligned} e_q(t) &= n\,q_o \quad , \quad (n-\tfrac{1}{2})\,q_o \leqslant e(t) < (n+\tfrac{1}{2})\,q_o \ , \\ &\qquad\qquad \text{for } n = 0, \pm 1, \pm 2, \ldots, \pm N-1 \\ &= N\,q_o \quad , \quad e(t) \geqslant (N-\tfrac{1}{2})\,q_o \\ &= -N\,q_o \quad , \quad e(t) < -(N-\tfrac{1}{2})\,q_o \end{aligned}$$

Because the relation between $e_q(t)$ and e(t) is non-linear it is no longer true that the optimal linear control law (ie. a control law of the form $u = K\,e_q$) is the best control law available. In fact, it is now worthwhile to look for a more general optimal control law of the form $u = f(e_q)$ where f is some non-linear function. The optimal regulation problem would then be to determine f so as to minimize the cost function L.

This optimal regulation problem can be formulated as a Markovian decision problem. For if t_1, t_2 ... are the instants at which the quantized error $e_q(t)$ switches to a new value, the sequence $(e_q(t_1), e_q(t_2), \ldots.)$ is a Markov chain whose state set consists of the (2N + 1) possible values of $e_q(t)$. The equilibrium properties of this Markov chain depend on the form of the control law f; by expressing the cost function L in terms of these equilibrium properties and then using one of the optimization procedures outlined in this thesis it is possible to determine the optimal form of f. It should be emphasized that finite-state Markov chain $(e_q(t_1), e_q(t_2), \ldots.)$ arising in this example provides an exact representation of the process to be controlled; it is not an approximation resulting from simplification of the original model of the process. Unfortunately this approach to control law optimization is not easily generalized to quantized linear regulators of higher order.

The above application is in the field of control engineering; later we shall consider in detail an application in the operations research field - the problem of choosing a control strategy for a queueing system with an adjustable service rate.

1.2 Background

The concept of a Markovian decision problem (or Markov programming problem, as it is also termed) was introduced by Bellman(1957) as an optimization problem amenable to numerical solution by the method of dynamic programming. The first major step in the development of the subject was taken by Howard(1960) who showed that finite-state Markovian decision problems can be classified into four main categories:

(i) Regulation problems, in which the aim is to minimize the average cost per unit time incurred by a Markov chain in statistical equilibrium;

(ii) Discounted regulation problems, in which the aim is essentially as in (i) but future costs are discounted exponentially in time so that the expected value of the accumulated future costs is finite;

(iii) Finite-time problems, in which the cost function is the expected total cost accumulated over a finite time interval; and

(iv) Transient-cost problems, in which the chain has an absorbing state and the cost function is the expected cost incurred before the absorbing state is entered.

By making use of the equilibrium properties of finite Markov chains Howard derived his so-called policy-iteration algorithm for the solution of optimization problems in the first of the above categories. Howard's algorithm has been widely used in applications and is the basis from which similar algorithms, due to Hastings$^{(1968)}$ and Schweitzer$^{(1971a)}$, have been developed. The basic computational disadvantage of Howard's algorithm and its derivatives is that each iteration involves the solution of N simultaneous equations where N is the number of states in the system - a major difficulty when N is very large.

A second, more direct, type of optimization procedure for Markov regulation problems (i.e. problems in category (i) above) is to solve the standard dynamic programming equations recursively until all transient terms in the solution have died away. This direct approach was first suggested by Eaton and Zadeh$^{(1962)}$ for problems of the transient-cost type (category (iv) above) and later modified by White$^{(1963)}$ for use in Markov regulation problems. White's successive-approximations algorithm has the advantage that it is computationally simple; on the other hand the rate of convergence of the algorithm depends on the dynamic characteristics of the system which is being optimized and in some cases convergence can be very slow. Further contributions related to White's algorithm have been made by Odoni$^{(1969)}$, who derived an improved stopping criterion for the algorithm, and MacQueen$^{(1966)}$, who developed a successive-approximations algorithm for discounted regulation problems (category (ii) above).

A third approach to the Markov regulation problem is to formulate it as a linear programming problem, that is, the optimization of a linear cost function subject to a set of linear constraints. This approach is due to Manne$^{(1960)}$ and was investigated further by

Derman[1962], and by Wolfe and Dantzig[1962], who showed by using the linear programming formulation of the problem that the optimal control law, assumed to be a deterministic rule for specifying the decision to be used in any state, cannot be further improved by randomization.

Extensions of the theory have been in three main directions, each of which we now briefly indicate. In 1954 Levy[1954] generalized the concept of a continuous-time Markov chain by introducing processes which switch from state to state in the same way as Markov chains but in which the time spent in each state is a general rather than an exponential, random variable. Levy called such processes semi-Markov processes and their properties have since been thoroughly investigated by W.L. Smith[1955], R. Pyke[1961a,1961b] and E. Çinlar[1969a,1969b]. A natural development was the extension of Bellman's notion of a Markovian decision problem to that of a semi-Markovian decision problem in which the process to be controlled is a semi-Markov process. The extension was suggested by Jewell[1963] who showed how such problems might be solved by the use of a modified version of Howard's policy-iteration algorithm. Similar extensions have also been proposed by Schweitzer[1969] and de Cani[1964]. Jewell's work was followed by semi-Markov generalizations of White's algorithm, by Schweitzer[1971b] and of Manne's linear programming algorithm by Osaki and Mine[1968]. The introduction of the semi-Markov concept has been of major importance to the development of the subject since it has vastly widened the range of potential applications of the theory. And, as we shall show, the semi-Markov concept helps to unify and clarify some of the results previously derived for pure Markov processes.

The second extension of the theory, which is mainly of mathematical rather than engineering interest, is the consideration of

Markovian decision processes with infinite state sets. Markov regulation problems in which the state set is countably infinite have been studied by Veinott$^{(1966)}$, Derman$^{(1966)}$, and by Haussman$^{(1971)}$, who demonstrated the existence of stationary optimal control laws for semi-Markov regulation problems with countably infinite state sets. The most general case, when the state space is non-countable, has been investigated by Ross$^{(1968a,1968b)}$ who gives sufficient conditions for the existence of a stationary optimal control law. Of course computation of the optimal control law for such a problem by a finite-state optimization algorithm necessitates the use of a finite-state approximation to the original infinite-state process. Methods for constructing such finite-state approximations have been proposed by Fox$^{(1971,1973)}$.

The remaining topic to have been studied in some detail is the control of partially-observable Markovian decision processes; that is to say, processes in which complete knowledge of the current state is not always available to the controller. In such systems the controller must make the best use possible of whatever information is available. Usually the available information consists of a set of observations related, perhaps stochastically, to the past and present motion of the process; the set of data available as inputs to the controller is called the information pattern for the system. The case most readily amenable to analysis is when the controller has what is known as perfect recall, which means the information available at any time includes the information available at all earlier times; the information pattern is then said to be "classical". A general discussion of some aspects of the control of partially-observable stochastic systems has been given by Witsenhausen$^{(1971)}$. In the field of Markovian decision processes, Aström$^{(1965)}$ showed that for a finite-time problem the optimal control scheme satisfies the so-called

Separation Principle: that is, the optimal scheme consists of an estimator which generates a probability distribution for the states, conditional on the available information, followed by a controller whose input is the distribution generated by the estimator. Similar results were obtained at about the same time by Aoki[(1965)], and the results were later extended to the countable-state case by Sarawagi and Yoshikawa[(1970)]. The essential argument in all the above work is that the partially-observable problem is equivalent to a completely-observable problem in which the "states" are the possible outputs of the estimator. Since the "state" set is no longer discrete the actual computation of the optimal control law is difficult. However by using the fact that the optimal expected total cost is a convex, piecewise-linear function of the estimator output, Smallwood and Sondik[(1973)] have developed an elegant procedure for determining the optimal control law for a finite-time problem. Unfortunately the method does not apply to regulation (i.e. infinite-time) problems. All of the above work deals with discrete-time Markovian decision problems; at the time of writing the more general semi-Markov case does not appear to have been studied, although Rudemo[(1973,1975)] has published some results for pure Markov chains in continuous time.

It is clear from the above brief review that a substantial amount of effort has been put into the development of this branch of stochastic control theory. Nevertheless in applications the difficulty has remained that in the actual computation of the optimal control law the standard optimization algorithms can be very expensive in computational resources when the system to be optimized has a large state set. It is this difficulty which originally motivated the work described in this thesis.

1.3 Outline of the thesis

Chapter 2 reviews the basic properties of discrete-state Markov processes in both discrete and continous time. The concept of a semi-Markov chain is then introduced and the equilibrium properties of such chains are summarized. We next introduce the important notion of equivalence between two semi-Markov chains. This idea, of which considerable use is made later, is based on the fact that the observed sample paths of a semi-Markov chain do not uniquely define the law of motion of the chain: there is, in fact, a whole equivalence class of chains with the same set of possible sample paths. Finally in this Chapter, we consider the equilibrium behaviour of the cost function when an additive cost structure is imposed on a Markov or semi-Markov chain. The growth of the expected total cost is asymptotically linear, the rate of growth depending on the equilibrium probability distribution for the states of the chain. The optimal regulation problem consists of choosing a control law so that the resulting equilibrium distribution results in the smallest possible value for this asymptotic rate of growth.

The optimal regulation problem is defined in detail in Chapter 3 and the existing standard optimization algorithms are then reviewed. As has been mentioned, there are three basic types of algorithm: (i) policy-iteration algorithms, in which an initial trial control law is systematically improved by an appropriate form of Bellman's "approximation in policy-space" technique(1957); (ii) successive-approximations algorithms, in which the non-linear optimality equations are solved by simple Jacobi iteration; and (iii) linear programming algorithms, in which the optimization problem is formulated as a linear programming problem and then solved by a linear programming procedure.

When the number of states is large the standard algorithms require considerable computational resources (core storage and central processor time). In Chapter 4 we present some new optimization algorithms which are, in general, likely to be more efficient than the standard methods. The first of the new methods is a modification of the standard policy-iteration algorithm in which multi-state changes in the control law are replaced by successive single-state changes. Two related policy-iteration algorithms, in which the single-state improvements to the control law are themselves optimal, are also discussed. We give a detailed comparison of the various policy-iteration algorithms; in particular, we demonstrate the convergence of the new algorithms, compare the computational effort required, and show that their performance is not adversely affected when the chain to be optimized possesses transient states.

We next consider algorithms of the successive-approximations type. We give a new convergence proof for White's original algorithm, which uses a contraction mapping argument and also makes use of some properties of inhomogeneous Markov chains. As we then show, the proof suggests a natural generalization of White's algorithm to the semi-Markov case, achieved by invoking the concept of equivalent chains mentioned earlier. Finally, we examine the possibility of accelerating the successive-approximations method by means of over-relaxation. It turns out that some degree of acceleration is feasible provided that a variable acceleration factor is used; this is a useful result since the standard successive-approximations procedure converges very slowly for certain classes of problem.

Many queueing systems are appropriately modelled by Markov processes in which transitions are possible only between adjacent states. Such processes are called birth-death processes and in Chapter 5 we consider optimal regulation of this type of process. The special structure of the birth-death process permits a certain amount of simplification in the optimization algorithms, in particular the Howard policy-iteration algorithm. A problem that arises in the control of queueing systems is that the natural choice of state space for the system is countably infinite so that the finite-state optimization algorithms cannot be used without some form of truncation of the state space. We show that by introducing a certain embedded Markov chain it is possible to truncate the state space without distorting the properties of the process. A related problem is that of choosing an optimal quantized control law for a Markovian decision problem. In general this is a very difficult problem; it can, in fact, be formulated as a partially-observable control problem with a non-classical information pattern - and such problems are notoriously difficult to solve. However, in the case of a birth-death process, it is possible to show that use of a modified version of our single-state policy-iteration algorithm will generate the globally-optimal quantized control law for the process. The last part of Chapter 5 is devoted to this topic.

Finally, in Chapter 6, we present some numerical results for a specific optimal regulation problem. The problem is the optimal regulation of a simple queueing system in which the number of service channels is variable but there is, in addition to the usual customer delay costs and open service-channel costs, a cost associated with any change in the number of active channels. The object of this numerical investigation is to compare the performances resulting from each of the following approaches to optimization of the system:

(i) Determine the control law which is optimal in the absence of switching costs and add in the cost contribution due to switching after the optimization.

(ii) For some sensibly chosen quantization of the state space, determine the quantized control law which is optimal in the absence of switching costs. A quantized control law results in less frequent changes in the number of active channels and hence in a lower switching cost contribution.

(iii) Determine the control law which is optimal in the presence of switching costs. This approach involves redefinition of the state space of the system in order that the switching costs can be properly incorporated in a separable cost function. The resulting control law exhibits a hysteresis-like characteristic in which the number of channels active for a given queue length depends on whether the queue is growing or shrinking.

The first two approaches result in sub-optimal control laws: the control law produced in (iii) is optimal. The question we have sought to resolve is this: is the performance of the optimal system sufficiently better than that of the sub-optimal systems to justify the extra computational effort needed to determine the truly optimal control law? In addition we have compared the performances of the main optimization algorithms in part (iii). above.

1.4 Contributions of the thesis

The work described in this thesis lies in the general field of system optimization and control; in particular it deals with the optimization of Markovian decision processes. The main contributions to this field, believed to be original, are:

(1) The development and use of the concept of equivalence for semi-Markov chains with additive costs. (Chapter 2)

(2) The development of new optimization algorithms of the policy-iteration type and an investigation of their properties. (Chapter 4)

(3) A new proof of the convergence of the successive-approximations algorithm for discrete-time chains, leading to the development of (a) a generalized semi-Markov

version* of the algorithm, and (b) an accelerated version of the algorithm. (Chapter 4)

(4) A study of the application of Markovian decision theory to birth-death processes, including proposals for handling processes with a countably-infinite state space and for optimimizing the performance of quantized systems. (Chapter 5)

(5) A numerical study of a specific optimal regulation problem of considerable practical significance. (Chapter 6)

*The semi-Markov version is not new: it was first proposed by Schweitzer(1971b) in 1971. The arguments presented here, leading to the development of the algorithm, are new.

CHAPTER 2

MARKOV AND SEMI-MARKOV CHAINS

2.1 Introduction

As is well known, the analysis of the behaviour of a deterministic dynamic system is usually simplified by suitably defining a state for the system and then analysing the motion of the state. The essential feature of this so-called state representation is that, given the present state, the future motion of the system is independent of its past history. An analogous situation holds for stochastic systems, that is, systems in which the motion is wholly or partially influenced by random effects. In such cases the law of motion is probabilistic rather than deterministic and the appropriate mathematical model is a stochastic process. As in the deterministic case, it is in principle possible to introduce a state for the system having the property that the future motion, given the present state of the system, is (stochastically) independent of the past history. The stochastic process representing the motion of the state is then a Markov process and analysis of the system's behaviour is then reduced to analysis of the behaviour of a specific Markov process. For this reason the concept of a controllable Markov process plays a key role in stochastic control theory.

In this chapter a brief outline is presented of the relevant theory of Markov processes and semi-Markov processes with particular emphasis on finite-state processes. The concept of equivalence between regular semi-Markov chains is introduced and the long-run behaviour of Markov chains with additive costs is then described.

2.2 Markov Processes

Given a probability space $\{\Omega, \mathcal{F}, \mathcal{P}\}$ the indexed collection of random variables $\{(X_t : \Omega \rightarrow \mathcal{X}) : t \in \mathcal{T}\}$ is called a stochastic process with state space $\mathcal{X}$ and parameter set (or index set) $\mathcal{T}$. In most applications t is the time at which X_t is observed and usually the only cases of interest are (i) $\mathcal{T} = \mathbb{Z}_+ \triangleq \{0,1,2,\ldots ,\}$ in which case $\{X_t\}$ is a discrete-time process, and (ii) $\mathcal{T} = \mathbb{R}_+ \triangleq [0,\infty)$, in which case $\{X_t\}$ is a continuous-time process. For fixed t, X_t is a random variable taking values in the state space $\mathcal{X}$; thus X_t is the state of the process at time t. The state space $\mathcal{X}$ may be discrete (ie. finite or countable) or continuous (for example, k dimensional Euclidean space, $\mathbb{R}^k$); but in this thesis we are concerned largely with finite-state processes, that is, processes with a finite number of possible states. We may then without loss of generality take $\mathcal{X} = \mathbb{N}_N \triangleq \{1,2,\ldots,N\}$. Once $\mathcal{X}$ and $\mathcal{T}$ have been specified the process $\{(X_t: \Omega \to \mathcal{X}): t \in \mathcal{T}\}$ is denoted by the abbreviation $\{X_t\}$.

A stochastic process $\{(X_t: \Omega \to \mathcal{X}): t \in \mathcal{T}\}$ is said to be a Markov process if $\mathcal{T}$ is an infinite set and if, for every integer k, for every set of times $\{t_i \in \mathcal{T} : i = 1,2,\ldots,k+1\}$ ordered so that $t_1 < t_2 < \ldots < t_k < t_{k+1}$, for every set of states $\{x_i \in \mathcal{X} : i = 1,2,\ldots,k+1\}$, and for every event E in $\mathcal{X}$,

$$P\left[X_{t_{k+1}} \in E \,\middle|\, X_{t_k} = x_k,\ X_{t_{k-1}} = x_{k-1},\ldots,\ X_{t_1} = x_1\right]$$
$$= P\left[X_{t_{k+1}} \in E \,\middle|\, X_{t_k} = x_k\right] \qquad \ldots.(2.1)$$

Property (2.1) is called the Markov property. It asserts that, given the "present state", X_{t_k}, (interpreting t_k as the "present time"), the future behaviour of the process is stochastically independent of every past value of the process.

A Markov process with a discrete state space is called a Markov chain. Such a chain may be finite, in which case we may take $\mathcal{X} = \mathbb{N}_N$, or countable in which case we may take $\mathcal{X} = \mathbb{N}$.

In a Markov chain the random variables X_t are discrete and we can therefore work with probability mass distributions defined on the product sets $\mathcal{X}^k$, $k = 1,2,\ldots$.

For the Markov chain $\{X_t\}$ we define, for every $k \in \mathbb{N}$,

$$p_{i_1,\ldots,i_k}(t_1,\ldots,t_k) \triangleq P\left[X_{t_1} = i_1,\ldots,X_{t_k} = i_k\right]$$

and, for every $r \in \{1,2,\ldots,k\}$,

$$p_{i_{r+1},\ldots,i_k \mid i_1,\ldots,i_r}(t_1,\ldots,t_k)$$
$$P\left[X_{t_{r+1}} = i_{r+1},\ldots,X_{t_k} = i_k \,\middle|\, X_{t_1} = i_1,\ldots,X_{t_r} = i_r\right]$$

Then the Markov property (2.1) can be written in the specialized form

$$\boxed{p_{i_{k+1} \mid i_1,\ldots i_k}(t_1,\ldots,t_{k+1}) = p_{i_{k+1} \mid i_k}(t_k,t_{k+1})} \qquad \ldots.(2.2)$$

from which it follows that

$$p_{i_1,\ldots,i_{k+1}}(t_1,\ldots,t_{k+1}) = p_{i_1}(t_1) \prod_{r=1}^{k} p_{i_{r+1} \mid i_r}(t_r,t_{r+1}) \qquad \ldots.(2.3)$$

Thus if $\{X_t\}$ is a Markov chain, its finite-order distributions are uniquely determined by

(i) the initial distribution $p_{i_1}(t_1)$

and (ii) the condition distributions $p_{i_{r+1} \mid i_r}(t_r,t_{r+1})$, $r = 1,2,\ldots$

Equation (2.2) is an assertion that the Markov property holds at the specified (ie. fixed) time t_k. It turns out in the subsequent development that we sometimes need the Markov property to hold, not at a fixed time t_k, but at some random time T. Let $\{X_t : t \in \mathcal{T}\}$ be a Markov chain and let T be a random variable, with values in $\mathcal{T}$, defined on the same probability space. The random variable T is said to be a stopping time for the chain $\{X_t\}$ iff for every $t \in \mathcal{T}$ the event $\{T > t\}$ is independent of the posterior behaviour of the chain, $\{X_s : s > t\}$. Roughly speaking, if the value of T can be determined by observation of the chain $\{X_t\}$ then T is a stopping time if its value is determined by $\{X_s : s \leqslant t\}$ for some $t \leqslant T$. For example "the time of the first occurrence of the event $X_t = i$ " is a stopping time; "t_o seconds before the first occurrence of $X_t = i$ "is not.

It may be shown (for example, Chung(1967)) that the Markov property holds at any stopping time: that is, if T is a stopping time for X_t and if t is a positive time such that $T' \triangleq T + t \in \mathcal{T}$, then for $t_1, \ldots, t_{k-1}$ all less than T,

$$\boxed{p_{j|i_1,\ldots,i_{k-1},i}\,(t_1,\ldots,t_{k-1}, T, T') = p_{j|i}\,(T, T')} \qquad \ldots.(2.4)$$

Since any fixed time $t_k \in \mathcal{T}$ is clearly a stopping time, property (2.4) is a more general one than (2.2). It is known as the strong Markov property and it may be regarded as the defining characteristic of a Markov chain.

We now proceed to review those properties of finite Markov chains that will be needed in the sequel.

2.2.1 Finite Markov chains in discrete time

In this section we consider Markov processes for which $\mathcal{X} = \mathbb{N}_N$ and $\mathcal{T} = \mathbb{Z}_+$. As we have seen, such processes are completely characterised by the conditional probabilities connecting successive times of interest. We therefore introduce the one-step transition probabilities

$$p_{ij}(t) \triangleq P\left[X_{t+1} = j \,\middle|\, X_t = i\right] \quad , \ i,j \in \mathcal{X}$$

and, more generally, the k-step transition probabilities

$$p_{ij}^{(k)}(t) \triangleq P\left[X_{t+k} = j \,\middle|\, X_t = i\right] \quad , \; i,j \in \mathcal{X}$$

An immediate consequence of the Markov property is that the $p_{ij}^{(k)}(t)$ must satisfy the Chapman-Kolmogorov relation :-

$$p_{im}^{(k+l)}(t) = \sum_{j} p_{ij}^{(k)}(t) \quad p_{jm}^{(l)}(t+k) \qquad \ldots.(2.5)$$

for every $i,j,m \in \mathcal{X}$ and every $t,k,l \in \mathcal{T}$.

In particular, with $l = 1$ equation (2.5) is a recurrence relation for generating the $p_{ij}^{(k)}(t)$ from the $p_{ij}(t)$.

In this thesis we are concerned only with Markov chains in which the transition mechanism does not vary with time, that is

$$p_{ij}(t+t_o) = p_{ij}(t) \quad , \quad \begin{array}{l} \forall i,j \in \mathcal{X} \\ \forall t,t_o \in \mathcal{T} \end{array}$$

Such a chain is said to be homogeneous (in time) and we denote its transition probabilities simply by $p_{ij}^{(k)}$.

We now introduce the state probability vector

$$\underline{p}_t \triangleq \mathrm{Col}\left[p_1(t), \ldots, p_N(t)\right]$$

where $\quad p_i(t) \triangleq P\left[X_t = i\right] \quad , \; i \in \mathcal{X}$, and the (1-step) transition probability matrix

$$P \triangleq \left[p_{ij}\right]_{N \times N}$$

Then, by the Markov property,

$$\boxed{\underline{p}_t^T = \underline{p}_{t-1}^T \; P} \qquad \ldots.(2.6)$$

and hence

$$\boxed{\underline{p}_t^T = \underline{p}_o^T \; P^t} \qquad \ldots.(2.7)$$

where P^t denotes the t^{th} power of P.

Equation (2.7) permits the state distribution at any time t to be computed in terms of the initial state distribution and the transition probability matrix. Furthermore if $P^{(k)}$ is the N x N matrix of k-step transition probabilities then the matrix form of (2.5) is

$$\boxed{P^{(k+1)} = P^{(k)}P^{(1)}} \qquad \text{....(2.8)}$$

from which it follows that $P^{(k)} = P^k$.

The transition probability matrix P is a stochastic matrix; that is, a square matrix with real, non-negative elements and unit row sums. The properties of such matrices are well established (see, for example, Çinlar(1975) or Seneta(1973)): in particular, it can be shown that for any stochastic matrix P

(i) P^k is stochastic for every positive integer k ;

(ii) the eigenvalues, λ , of P all lie on the closed unit disc $\{\lambda : |\lambda| \leqslant 1\}$;

and (iii) $\lambda = 1$ is always an eigenvalue of P.

It is clear that the long run ($t \to \infty$) behaviour of P^t and hence, via (2.7), of $\underline{p}_t^T$ will depend on the eigenvalues of P with unit modulus and the associated eigenvectors. The form of the eigenvalue spectrum on the unit circle $|\lambda| = 1$ is directly related to the availability of communication paths between the various states of the chain, a question which we now consider briefly.

Let R_i denote the event that, for at least one integer $k > o$, $X_{t+k} = i$ and let $r_i \triangleq P\left[R_i \middle| X_t = i\right]$. Then the state i is recurrent if $r_i = 1$ and transient if $r_i < 1$. Suppose that i is recurrent and that $T_i(k)$ is the time interval between the k^{th} and $(k+1)^{th}$ occupations of state i. Then $T_i(k)$ is a random variable with distribution independent of k (by the strong Markov property the chain "restarts" at every visit to i). The expectation $t_i \triangleq E\left[T_i\right]$ is called the mean recurrence time of state i. In a finite chain t_i is finite for every recurrent state i. If the only possible values of T_i are k, 2k, 3k,... for some $k > 1$, the state i is periodic; otherwise i is ergodic.

In particular, if $p_{ii} = 1$ so that T_i is always 1 then state i is absorbing.

Now it may be shown (for example, Chung(1967)) that the state space $\mathcal{X}$ of any finite Markov chain is the union of two disjoint subsets,

$$\mathcal{X} = \mathcal{X}_T \cup \mathcal{X}_R ,$$

where $\mathcal{X}_T$ consists of all the transient states ($\mathcal{X}_T$ may be empty),

$\mathcal{X}_R$ consists of all the recurrent states,

and no state in $\mathcal{X}_T$ is accessible from any state in $\mathcal{X}_R$. Furthermore, $\mathcal{X}_R$ may be uniquely partitioned into closed sets in each of which all states intercommunicate and are of the same type and period. (The set of states $\mathcal{X}_A$ is closed if $P\left[X_{t+k} \in \mathcal{X}_A \middle| X_t \in \mathcal{X}_A\right] = 1$ for every $k > 0$. Two states i and j intercommunicate if, for some k and l, $p_{ij}^{(k)} > 0$ and $p_{ji}^{(l)} > 0$.) The implication is that every finite chain consists of one or more recurrent subchains together, possibly, with some transient states and ultimately $\{X_t\}$ will be absorbed into one or other of the subchains.

The long-run behaviour of $\{X_t\}$ clearly depends on the number and nature of the subchains it possesses. Throughout this thesis our attention is confined to chains possessing a single ergodic subchain (as well as, possibly, some transient states). Such a chain is said to be regular - in which case we also say that P is regular - and possesses the following properties:

R.1 There exists a unique stationary probability distribution

$\underline{\pi} \triangleq \text{Col}\,(\pi_1, \ldots, \pi_N)$ satisfying

$$\boxed{\underline{\pi}^T = \underline{\pi}^T P} \qquad \ldots.(2.9)$$

and, furthermore,

$$\pi_i = \frac{1}{t_i} , \quad i \in \mathcal{X}_R$$

$$= 0 , \quad i \in \mathcal{X}_T$$

Clearly if $\underline{p}_o = \underline{\pi}$, then from (2.7) $\underline{p}_t = \underline{\pi}$ for every time t and so $\{X_t\}$ is stochastically stationary.

<u>R.2</u> The chain is asymptotically stable in distribution: that is

$$\boxed{\lim_{t \to \infty} P^t = \underline{e}\,\underline{\pi}^T} \qquad \text{....(2.10)}$$

where $\underline{e} \triangleq$ Col(1,1,...,1), so that, for every $\underline{p}_o$,

$$\lim_{t \to \infty} \underline{p}_t^T = \lim_{t \to \infty} \underline{p}_o^T P^t$$

$$= \underline{\pi}^T$$

Thus a regular chain always settles down to an equilibrium behaviour, which is independent of the initial state of the chain, and which is governed by the stationary distribution, $\underline{\pi}$.

<u>R.3</u> The chain is strictly ergodic: that is, if $I_i : \mathcal{X} \to \{0,1\}$ is the indicator function for state i then

$$\lim_{t \to \infty} \frac{1}{t+1} \sum_{k=0}^{t} I_i(X_k) = \pi_i \text{ , almost surely.}$$

This means that for almost all sample paths of the chain $\{X_t\}$ the long-run proportion of time in which state i is occupied is equal to π_i.

As we have already remarked, the long-run behaviour of a finite chain is governed by the unit-modulus eigenvalues of its P-matrix. In particular, the chain will be regular iff the principal eigenvalue, $\lambda = 1$, of its P-matrix is simple and is the only eigenvalue of unit modulus. In this case the state distribution $\underline{p}_t$ can be resolved into steady-state and transient components.

Let $P^\infty \triangleq \lim_{t \to \infty} P^t = \underline{e}\,\underline{\pi}^T$ (using (2.10))

Then

$$P\,P^\infty = P^\infty P = (P^\infty)^2 = P^\infty \qquad \text{....(2.11)}$$

so that, if $\tilde{P} \triangleq P - P^{\infty}$,

$$P^{\infty}\tilde{P} = \tilde{P}P^{\infty} = 0 \qquad \text{....(2.12)}$$

whence

$$\boxed{P^t = P^{\infty} + \tilde{P}^t} \qquad \text{....(2.13)}$$

Note incidentally that $\lim\limits_{t\to\infty} \tilde{P}^t = 0$.

Using (2.13) in (2.7) we obtain

$$\underline{p}_t^T = \underline{\pi}^T + \underline{p}_o^T \tilde{P}^t \qquad \text{....(2.14)}$$

The rate at which the transient term $\underline{p}_o^T \tilde{P}^t$ decays is governed by λ_2, the eigenvalue of P with second largest modulus (since if the spectrum of P is $\{1, \lambda_2, \lambda_3, ...\}$ that of $\tilde{P}$ is $\{0, \lambda_2, \lambda_3, ...\}$).

2.2.2 Finite Markov chains in continuous time

We next consider Markov processes for which $\mathcal{X} = \mathbb{N}_N$ and $\mathcal{T} = \mathbb{R}_+$, ie. continuous-time Markov chains. The sample paths of such chains are random step functions in which instantaneous jumps between states are made at randomly occurring times. As before we confine our attention to homogeneous chains in which the transition probabilities are invariant with respect to translations in time.

Corresponding to the k-step transition probabilities in discrete time, we now introduce transition probability functions

$$p_{ij}(\tau) \triangleq P\left[X_{t+\tau} = j \,\middle|\, X_t = i\right], \quad i,j \in \mathcal{X}$$

with $t, \tau \in \mathcal{T}$.

The transition function matrix

$$P(\tau) \triangleq \left[p_{ij}(\tau)\right]_{N \times N}$$

must satisfy the continuous-time Chapman-Kolmogorov relation : -

$$\boxed{P(\tau_1 + \tau_2) = P(\tau_1)P(\tau_2)}, \quad \tau_1, \tau_2 \in \mathcal{T} \qquad \text{....(2.15)}$$

The time interval between entry into a (non-absorbing) state and subsequent exit from the state is called the sojourn time (or holding time) in that state. If $S_i^{(k)}$ is the k^{th} sojourn time in state i it follows from the strong Markov property that the random variables $S_i^{(1)}, S_i^{(2)}, \ldots$ are independent and identically distributed. However to maintain the Markov property at every point in $\mathcal{T}$ more is needed: the sojourn times must be exponential random variables. That is, the distribution function $F_i : \mathbb{R}_+ \to [0,1]$ of S_i must have the form (see, for example, Çinlar$^{(1975)}$)

$$F_i(t) = 1 - e^{-\mu_i t} \quad , \quad t \geqslant 0 \qquad \ldots.(2.16)$$

for each $i \in \mathcal{X}$.

It follows that for small Δt, regardless of the entry time into state i ,

$$P\left[X_{t+\Delta t} \neq i \,\middle|\, X_t = i\right] = \mu_i \Delta t + o(\Delta t) \qquad \ldots.(2.17)$$

Thus with probability $\mu_i \Delta t$ the chain will leave state i in the small interval $(t, t+\Delta t)$, and by the Markov property the destination can depend only on the state i. Define the next-jump probabilities, r_{ij}, by

$$r_{ij} \triangleq P\left[X_{t+\Delta t} = j \,\middle|\, X_t = i,\; X_{t+\Delta t} \neq i\right] \quad , \; i,j \qquad \ldots.(2.18)$$

i.e. r_{ij} is the probability, given the occurrence of a jump out of state i, that the destination will be j. Clearly $r_{ii} = 0$, $\forall i \in \mathcal{X}$. (Note that this argument holds only for non-absorbing states. If state i_o is absorbing, then F_{i_o} and r_{i_oj} are not defined. In such a case it is convenient to allow "pseudo-jumps" from i_o into itself: we can then take F_{i_o} to be exponential with an arbitrarily-chosen parameter μ_{i_o} and the next-jump probabilities for i_o by

$$\begin{aligned} r_{i_oj} &= 0 \quad , \quad j \neq i_o \\ &= 1 \quad , \quad j = i_o \end{aligned}$$

It is then of course no longer true that $r_{ii} = 0$, $\forall i \in \mathcal{X}$)

It is now convenient to introduce transition intensities, q_{ij}, defined by

$$\begin{aligned} q_{ij} &\triangleq \mu_i \, r_{ij} && , \; j \neq i \\ &\triangleq -\mu_i \sum_{k \neq i} r_{ik} && , \; j = i \end{aligned} \qquad \text{....(2.19)}$$

Then from (2.17) and (2.18) it follows that

$$\begin{aligned} p_{ij}(\Delta t) &= q_{ij} \, \Delta t + o\,(\Delta t) \;, && j \neq i \\ &= 1 + q_{ii} \, \Delta t + o\,(\Delta t) \;, && j = i \end{aligned} \qquad \text{....(2.20)}$$

The intensity matrix

$$Q \triangleq \left[q_{ij} \right]_{N \times N} \qquad \text{....(2.21)}$$

is the continuous-time counterpart of the transition probability matrix P of a discrete-time chain. By considering an appropriate form of the Chapman-Kolmogorov equation it may be shown that the transition function matrix $P(\tau)$ satisfies the differential equation

$$\dot{P}(\tau) = P(\tau)\, Q \qquad \text{....(2.22)}$$

with initial condition $P(0) = I$.

The solution of (2.22) is

$$\boxed{P(\tau) = \exp\,(Q\tau)} \qquad \text{....(2.23)}$$

and so the state probability distribution at any time $t \in \mathcal{T}$ is given by

$$\boxed{\underline{p}_t^T = \underline{p}_o^T \exp\,(Qt)} \qquad \text{....(2.24)}$$

Now, setting $\tau = 1$ in (2.23) yields the matrix $P(1) = \exp\,(Q)$. Thus if $\lambda_1, \lambda_2, \ldots, \lambda_M$ are the eigenvalues of Q and $\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_M$ are the corresponding eigenvalues of P(1), we have

$$\bar{\lambda}_i = e^{\lambda_i} \qquad , \; i = 1,2,..,M$$

whence

$$\lambda_i = \log \bar{\lambda}_i \quad , \quad i = 1,2,\ldots,M$$

But P(1) is a stochastic matrix and hence its eigenvalues are all on the unit disc $|\lambda| \leqslant 1$; we therefore conclude that the eigenvalues of Q are all in the closed half-plane $\mathrm{Re}(\lambda) \leqslant 0$, and that corresponding to the principal eigenvalue $\lambda_1 = 1$ of P(1), Q has a principal eigenvalue $\lambda_1 = 0$.

The classification of states for discrete-time chains carries over to continuous time (with the exception that periodic states can no longer arise). Furthermore, if the chain is regular, the following properties hold : -

R.4 There is a unique stationary distribution $\underline{\sigma}$ which satisfies the equation

$$\boxed{\underline{\sigma}^T Q = \underline{0}^T} \qquad \ldots.(2.25)$$

R.5 The chain is asymptotically stable in distribution: that is

$$\boxed{\lim_{t\to\infty} \exp(Qt) = \underline{e}\,\underline{\sigma}^T} \qquad \ldots.(2.26)$$

so that, for every $\underline{p}_0$,

$$\lim_{t\to\infty} \underline{p}_t = \underline{\sigma}^T$$

as in the discrete-time case.

R.6 The chain is strictly ergodic: that is

$$\lim_{t\to\infty} \frac{1}{T}\int_0^T I_i(X_t)\,dt = \sigma_i \,, \qquad \text{almost surely}$$

where $I_i : \mathcal{X} \to \{0,1\}$ is the indicator function for state i.

A continuous-time finite chain is regular iff the principal eigenvalue, $\lambda_1 = 0$, of Q is simple. In such a case we have, in analogy to (2.13),

$$P(\tau) = P^{\infty} + \tilde{P}(\tau) \qquad \ldots.(2.27)$$

where

$$\lim_{\tau \to \infty} \tilde{P}(\tau) = 0$$

and

$$\tilde{P}(\tau)\, P^{\infty} = P^{\infty}\, \tilde{P}(\tau) = 0$$

2.2.3 Finite semi-Markov chains

The matrix $R \triangleq \left[r_{ij}\right]_{N \times N}$ of next-jump probabilities (see (2.18)) may be regarded as the transition probability matrix of a discrete-time Markov chain $\{(\overline{X}_t : \Omega \rightarrow N_N) : t \in \mathbb{Z}_+\}$. If we identify the sequence of points 0,1,2,... in $\mathbb{Z}_+$ with the sequence of random times $0, T_1, T_2, \ldots$ in $\mathbb{R}_+$ at which the continuous-time chain $\{X_t\}$ changes state then $\{\overline{X}_t\}$ is said to be embedded in $\mathbb{R}_+$ and is called the embedded chain of the original continuous-time chain $\{X_t\}$. It has the special property that the diagonal elements r_{ii} of its transition probability matrix R are all either zero (non-absorbing states) or one (absorbing states).

The closed subchains of $\{\overline{X}_t\}$ correspond to the closed subchains of $\{X_t\}$, so that if the latter is regular then so is $\{\overline{X}_t\}$ and a unique equilibrium distribution will exist for each chain. These two equilibrium distributions will not in general be identical since that of $\{X_t\}$ will depend on the sojourn-time distributions whereas that of $\{\overline{X}_t\}$ will not. In fact if $\underline{\sigma}$ and $\underline{\pi}$ are the equilibrium distributions of $\{X_t\}$ and $\{\overline{X}_t\}$ respectively, we have (see Appendix)

$$\boxed{\sigma_i = \frac{\pi_i \tau_i}{\sum_j \pi_j \tau_j}} \quad , \quad \forall i \in \mathcal{X} \qquad \ldots.(2.28)$$

where $\tau_i \triangleq E\left[S_i\right] = \mu_i^{-1}$, the mean sojourn time in state i.

Note that $\sigma_i = 0$ iff $\pi_i = 0$ (we ignore the possibility of so-called ephemeral states for which $\tau_i = 0$) in which case state i is transient. Otherwise i is ergodic and then σ_i and π_i give (almost surely) the long-run occupancy of state i as a relative duration and as a relative count, respectively.

The concept of an embedded chain is clearly a general one. Thus, for example, if $\{X_t\}$ is a discrete-time chain we may introduce an embedded chain $\{\overline{X}_t\}$ whose index set $\mathcal{T}$ is the set of times $t \in \mathbb{Z}_+$ for which $X_t \neq X_{t-1}$. The sojourn times S_i are then discrete random variables whose distributions F_i are geometric. The equilibrium probabilities of $\{X_t\}$ and $\{\overline{X}_t\}$ in the regular case are again related by (2.28).

We thus have an interpretation of any finite Markov chain $\{X_t : t \in \mathcal{T}\}$ as a discrete-time chain $\{\overline{X}_t : t \in \mathbb{Z}_+\}$ embedded in the index set $\mathcal{T}$ in such a way that the sojourn-time distributions F_i are all exponential (if $\mathcal{T} = \mathbb{R}_+$) or geometric (if $\mathcal{T} = \mathbb{Z}_+$). Suppose however that the sojourn-time distributions, while still dependent only on the current state, were not exponential (continuous $\mathcal{T}$) or geometric (discrete $\mathcal{T}$). The process $\{X_t\}$ would then no longer be Markov in $\mathcal{T}$ but the Markov property would still hold at the jump times $t_1, t_2, \ldots$ at which $\{X_t\}$ changes state. Such processes are called semi-Markov chains (or semi-Markov processes) and they are of considerable interest for the following reasons (a) They constitute a general class of processes of which Markov chains in discrete and continuous time are special cases; (b) certain more general stochastic processes called semi-regenerative processes (see, for example, Çinlar(1975)), which arise in queueing theory, always possess an embedded semi-Markov chain; and (c) semi-Markov chains are closely related to renewal processes (see, for example, Cox(1962)) and so provide a link between two distinct branches of applied probability theory. We now briefly review the main properties of this class of process when the number of states is finite. The theory is mainly due to Levy$^{(1954)}$, W.L. Smith$^{(1955)}$ and Pyke$^{(1961a, 1961b)}$, with further developments by Çinlar$^{(1969a, 1969b)}$ and Teugels$^{(1968)}$. An excellent survey of the theory has recently been given by Çinlar$^{(1975)}$.

The joint process $\{(\overline{X}_n : \Omega \to \mathcal{N}_N), (T_n : \Omega \to \mathcal{T}) : n \in \mathbb{Z}_+\}$ is called a (finite) <u>Markov renewal process</u> with state space $\mathcal{N}_N$,

if (i) $0 = T_0 \leqslant T_1 \leqslant T_2 \leqslant \ldots$

and (ii) for all $n \in \mathbb{Z}_+$, $j \in \mathcal{N}_N$, $t \in \mathcal{T}$,

$$P\left[\overline{X}_{n+1} = j \,,\ \Delta T_{n+1} \leqslant t \,\middle|\, \overline{X}_0, \ldots, \overline{X}_n \,;\, T_0, \ldots, T_n\right]$$

$$= P\left[\overline{X}_{n+1} = j \,,\ \Delta T_{n+1} \leqslant t \,\middle|\, \overline{X}_n\right] \qquad \ldots.(2.29)$$

where $\Delta T_n \triangleq T_n - T_{n-1}$

Assume the process $\{\overline{X}_n, T_n\}$ is homogeneous in n and define the <u>transition functions</u>

$$F_{ij}(t) \triangleq P\left[\overline{X}_{n+1} = j \,,\ \Delta T_{n+1} \leqslant t \,\middle|\, \overline{X}_n = i\right],$$

$$i,j \in \mathcal{N}_N,\ t \in \mathcal{T} \qquad \ldots.(2.30)$$

(To exclude the possibility of ephemeral states, assume that $F_{ij}(o) = 0 \,,\ \forall\, i,j \in \mathcal{N}_N$)

The matrix $F(t) \triangleq \left[F_{ij}(t)\right]_{N \times N}$ is called the <u>semi-Markov kernel</u> of the process $\{\overline{X}_n, T_n\}$; from it we derive the following quantities:

(i)
$$p_{ij} \triangleq P\left[\overline{X}_{n+1} = j \,\middle|\, \overline{X}_n = i\right]$$

$$= \lim_{t \to \infty} F_{ij}(t) \quad , \quad i,j \in \mathcal{N}_N,\ t \in \mathcal{T} \qquad \ldots.(2.31)$$

(ii)
$$G_{ij}(t) \triangleq P\left[\Delta T_{n+1} \leqslant t \,\middle|\, \overline{X}_n = i \,,\ \overline{X}_{n+1} = j\right]$$

$$= \frac{F_{ij}(t)}{p_{ij}} \quad , \quad \text{if } p_{ij} \neq 0$$

$$\triangleq 1 \quad , \quad \text{if } p_{ij} = 0 \qquad \ldots.(2.32)$$

(iii) $$H_i(t) \triangleq P\left[\Delta T_{n+1} \leqslant t \,\middle|\, \overline{X}_n = i\right] \ , \ i \in \mathbb{N}_N,\ t \in \mathbb{R}_+ \qquad \ldots.(2.33)$$

The process $\{\overline{X}_n, T_n\}$ is best interpreted as a Markov chain $\{\overline{X}_n : n \in \mathbb{Z}_+\}$, with transition probability matrix $P \triangleq [p_{ij}]_{N \times N}$, embedded in the index set $\mathcal{T}$ by the mapping $n \mapsto T_n(\omega)$.

Given $(\overline{X}_n = i,\ T_n = t)$ we can suppose the transition to $(\overline{X}_{n+1},\ T_{n+1})$ to be generated

either (a) by the selection of $\overline{X}_{n+1}$ from the conditional state distribution p_{ij}, $j = 1,\ldots,N$, followed by the selection of ΔT_{n+1} from the conditional sojourn-time distribution $G_{ij}(t)$, where j is the value of $\overline{X}_{n+1}$;

or (b) by the selection of ΔT_{n+1} from the unconditional sojourn-time distribution $H_i(t)$, followed by the selection of $\overline{X}_{n+1}$ from the time-dependent conditional state distribution $p_t(i,j)$ defined by

$$p_t(i,j) \triangleq P\left[\overline{X}_{n+1} = j \,\middle|\, \overline{X}_n = i\ ;\ \Delta T_{n+1} = t\right] \ ,$$
$$i,j \in \mathbb{N}_N,\ t \in \mathcal{T} \ . \qquad \ldots.(2.34)$$

There are some obvious relations between the functions F_{ij}, G_{ij}, H_i and p_t defined above which we now list :

$$F_{ij}(t) = p_{ij}\, G_{ij}(t) \ , \quad \forall i,j \in \mathbb{N}_N,\ \forall t \in \mathbb{R}_+ \qquad \ldots.(2.35)$$

$$F_{ij}(t) = \int_0^t p_t(i,j)\, dH_i(t) \ , \quad \forall i,j \in \mathbb{N}_N,\ \forall t \in \mathbb{R}_+ \qquad \ldots.(2.36)$$

$$H_i(t) = \sum_j F_{ij}(t) \ , \quad \forall i \in \mathbb{N}_N,\ \forall t \in \mathbb{R}_+ \qquad \ldots.(2.37)$$

Now for any $t \in \mathcal{T}$ let the random variable $N_t^{(i)} : \Omega \to \mathbb{Z}_+$ denote the number of transitions into state i in the interval $(0,t]$, and let $\overline{N}_t \triangleq \sum_{i \in \mathbb{N}_N} N_t^{(i)}$. Then the integer-valued process $\{\overline{N}_t : t \in \mathcal{T}\}$ counts the total number of transitions in the interval

$(0,t]$, and is called the renewal counting process associated with the process $\{\overline{X}_n, T_n\}$.

The stochastic process $\{(X_t: \Omega \rightarrow \mathbb{N}_N): t \in \mathcal{T}\}$ defined by setting

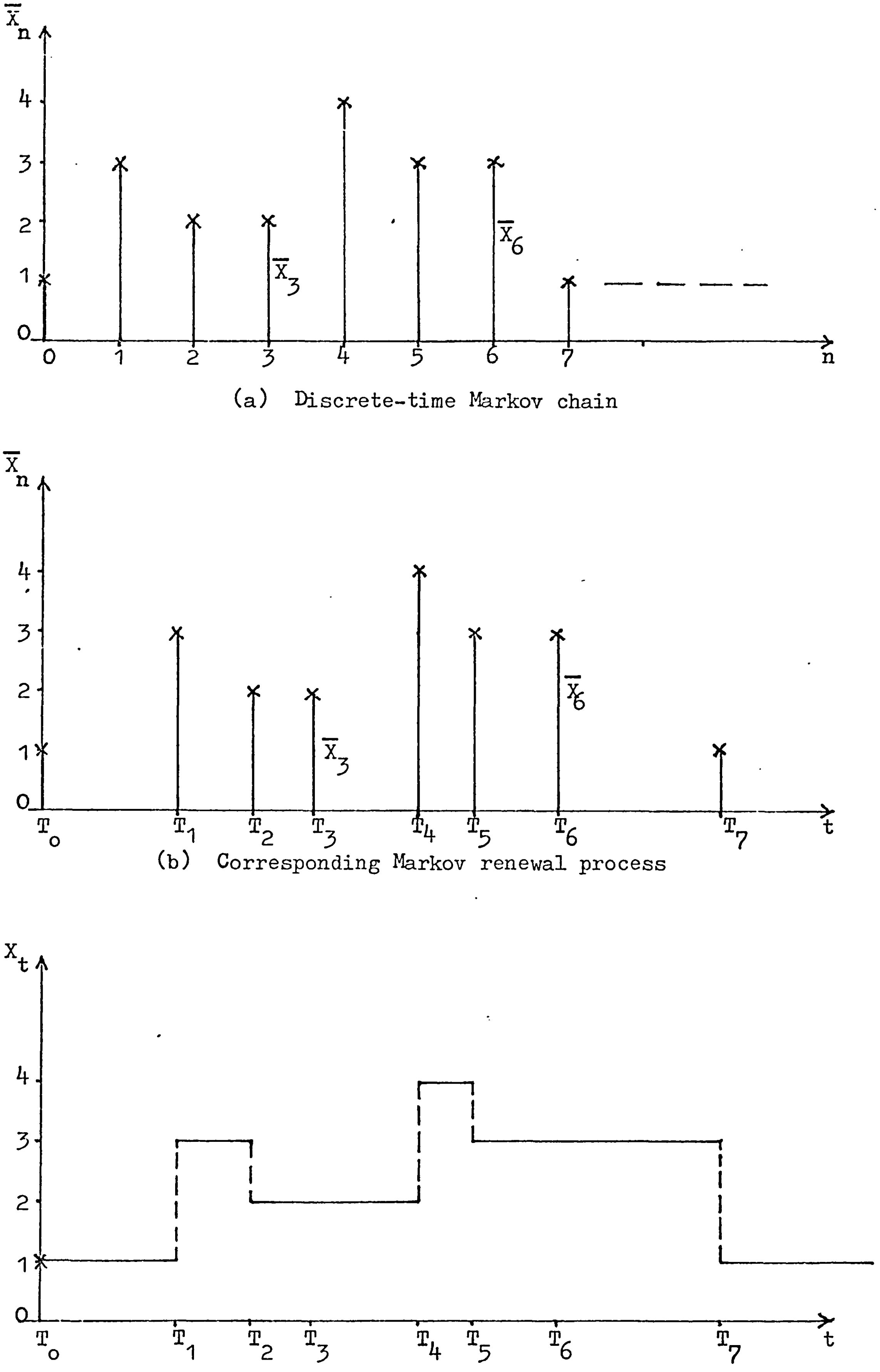
$$\boxed{X_t \triangleq \overline{X}_{N_t}} \quad , \forall t \in \mathcal{T} \qquad \text{....(2.38)}$$

is called the semi-Markov chain associated with the process $\{\overline{X}_n, T_n\}$, and $\{\overline{X}_n\}$ is its embedded Markov chain. The relationship between these processes when $\mathcal{T} = \mathbb{R}_+$ is illustrated in Fig.(2).

Notice that in the above definition of a semi-Markov process, the embedded chain $\{\overline{X}_n\}$ is allowed to make self-transitions, ie. the diagonal elements p_{ii} of the transition probability matrix are not necessarily zero. In applications, however, it is usually natural to work with semi-Markov processes in which the embedded chain defines changes in state, so that necessarily $p_{ii} = 0, \forall i \in \mathbb{N}_N$. (But see Section 2.2.4). Note also that in some applications the time increment ΔT_{n+1} is statistically independent of the destination state $\overline{X}_{n+1}$, so that $G_{ij} = H_i, \forall j \in \mathbb{N}_N$, but this constraint does not simplify the analysis of the process.

The random times $T_1, T_2, \ldots.$ are stopping times for the semi-Markov chain $\{X_t\}$ and the Markov property holds at each T_n; unless, however, the sojourn-time distributions H_i are all exponential (for $\mathcal{T} = \mathbb{R}_+$) or geometric (for $\mathcal{T} = \mathbb{Z}_+$) the Markov property does not hold for $\{X_t\}$ at points between the times T_n.

The classification of states for a semi-Markov chain $\{X_t\}$ follows that of its embedded chain $\{\overline{X}_n\}$. Thus state i is recurrent in $\{X_t\}$ iff it is recurrent in $\{\overline{X}_n\}$. Furthermore, except in rather special circumstances (see Çinlar$^{(1975)}$), the recurrent states in $\{X_t\}$ can be assumed to be ergodic. As before, a chain possessing a single ergodic subchain is said to be regular and for such semi-Markov



(a) Discrete-time Markov chain

(b) Corresponding Markov renewal process

(c) Corresponding semi-Markov chain

Fig. (2)

chains we have the following long-run properties (cf. R.1 - R.3 in section 2.2.1 and R.4 - R.6 in section 2.2.2).

Suppose that, for the recurrent state j ,

$$K_{ij}(t) = P\left[N_t^{(j)} > 0 \,\middle|\, \overline{X}_o = i\right] , \; i,j \in \mathbb{N}_N, \; t \in \mathcal{T}$$

Then $K_{ij} : \mathcal{T} \to [0,1]$ is the distribution function of the so-called <u>first passage time</u>, T_{ij}, from state i to the recurrent state j.

Let

$$\begin{aligned} \mu_{ij} &\triangleq E\left[T_{ij}\right] \\ &= \int_o^{\infty} t \, dK_{ij}(t) \\ &= \text{mean first passage time from i to j} \end{aligned}$$

and let

$$\begin{aligned} \tau_i &\triangleq E\left[\Delta T_{n+1} \,\middle|\, \overline{X}_n = i\right] \\ &= \int_o^{\infty} t \, dH_i(t) \qquad \qquad \ldots.(2.39) \\ &= \text{mean sojourn time in state i} \end{aligned}$$

Assuming that $\tau_i < \infty$, $\forall i \in \mathbb{N}_N$, we then have

<u>R.7</u> If $\underline{\pi} = \mathrm{Col}(\pi_1, \ldots, \pi_N)$ is the unique stationary distribution for the P-matrix of the embedded chain $\{\overline{X}_n\}$, then the <u>mean recurrence time</u>, μ_{jj}, of any ergodic state j is given by (see Appendix)

$$\boxed{\mu_{jj} = \frac{1}{\pi_j} \sum_{i \in \mathcal{X}_R} \pi_i \tau_i} \quad , \forall j \in \mathcal{X}_R \qquad \ldots.(2.40)$$

where $\mathcal{X}_R$ is the set of recurrent states in $\mathbb{N}_N$.

<u>R.8</u> For any recurrent state j ,

$$\lim_{t \to \infty} P\left[X_t = j \,\middle|\, X_o = i\right] = \frac{\tau_j}{\mu_{jj}} \qquad \ldots.(2.41)$$

This result, which is demonstrated in the Appendix, shows that $\{X_t\}$ has a long-run distribution over $\mathcal{X}_R$ which is independent of the initial state X_o. For, defining

$$\sigma_j \triangleq \lim_{t \to \infty} P\left[X_t = j \,\middle|\, X_o = i\right]$$

and using (2.40) in (2.41), we obtain

$$\sigma_j = \frac{\pi_j \tau_j}{\sum_{i \in \mathcal{X}_R} \pi_i \tau_i} , \quad \forall j \in \mathcal{X}_R \qquad \text{....(2.42)}$$

and since $\sum_{j \in \mathcal{X}_R} \sigma_j = 1$ it follows that the probabilities $\sigma_j : j \in \mathcal{X}_R$ form a distribution over $\mathcal{X}_R$. In fact, since $\pi_j = \sigma_j = 0$ for any transient state j in N_N, we can write (2.42) in the form

$$\boxed{\sigma_j = \frac{\pi_j \tau_j}{\sum_{i \in N_N} \pi_i \tau_i}} , \quad \forall j \in N_N \qquad \text{....(2.43)}$$

This is a key result for semi-Markov chains; it shows that the equilibrium probabilities σ_j depend on the semi-Markov kernel F(t) only through the means τ_j of the sojourn-time distributions H_j and the stationary probabilities π_j of the embedded Markov chain. Equation (2.28) is a special case of (2.43).

<u>R.9</u> The semi-Markov chain $\{X_t\}$ is strictly ergodic: that is, if $I_i : N_N \to \{0,1\}$ is the indicator function of state i,

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T I_i(X_t)\, dt = \sigma_i , \text{ almost surely}$$

This result confirms the intuitively reasonable idea that (on almost every sample path) the long-run proportion of time spent in an ergodic state i should be the ratio of the mean sojourn time in i to the mean recurrence time of i.

R. 10 The equilibrium mean sojourn time,

$$\bar{\tau} \triangleq \lim_{n \to \infty} E\left[\Delta T_n\right] ,$$

for $\{X_t\}$ is given by

$$\boxed{\bar{\tau} = \sum_{i \in N_N} \pi_i \tau_i} \qquad \ldots\ldots(2.44)$$

since
$$E\left[\Delta T_n\right] = E\left[E\left[\Delta T_n \,\middle|\, \bar{X}_{n-1}\right]\right]$$
$$= \sum_{i \in N_N} \tau_i \, P\left[\bar{X}_{n-1} = i\right]$$

2.2.4 Equivalent chains

From the preceding Sections we conclude that finite semi-Markov chains constitute a fairly general class of finite-state process which includes Markov chains in discrete and continuous time as special sub-classes. To be specific, the semi-Markov chain $\{(X_t : \Omega \to N_N) : t \in \mathcal{T}\}$ is

(i) a discrete-time Markov chain if $\mathcal{T} = \mathcal{Z}_+$ and

either (a) every sojourn time ΔT_n is equal to unity so that

$T_n = n, \ \forall n \in \mathcal{Z}_+$,

or (b) the sojourn-time distributions H_i are all geometric, ie. $H_i(t) = 1 - \alpha_i^t$ (We may then interpret α_i as the self-transition probability, p_{ii}, of state i of some underlying Markov chain with $T_n = n, \ \forall n \in \mathcal{Z}_+$.);

(ii) a continuous-time Markov chain if $\mathcal{T} = R_+$ and the sojourn-time distributions are all exponential, ie. $H_i(t) = 1 - e^{-\mu_i t}$.

(It is perhaps worth remarking here that the associated Markov renewal process $\{\bar{X}_n , T_n\}$ is a generalization of an ordinary renewal process, since if the state space of $\{\overline{X}_n\}$ contains only one point $(N = 1)$ then $\{\overline{X}_n, T_n\}$ is an ordinary renewal process.)

We have previously defined a general semi-Markov chain $\{X_t\}$ in terms of an associated Markov renewal process $\{\overline{X}_n, T_n\}$. The sample-paths of $\{X_t\}$ are then completely determined by the semi-Markov kernel F of the underlying process $\{\overline{X}_n, T_n\}$. However, if self-transitions are permitted in $\{\overline{X}_n\}$ the converse is not true: the sample-path behaviour of $\{X_t\}$ is not sufficient to determine the kernel F uniquely, since self-transitions in $\{\overline{X}_n\}$ are hidden in the observed process $\{X_t\}$.

Consider two Markov renewal processes $\{\overline{X}_n, T_n\}$ and $\{\overline{X}_n', T_n'\}$, with the same index set $\mathcal{T}$, the same state space $\mathbb{N}_N$, and with semi-Markov kernels F and F'. Let us say that $\{\overline{X}_n, T_n\}$ and $\{\overline{X}_n', T_n'\}$ are <u>equivalent</u> - or, F and F' are <u>equivalent</u> - iff they generate statistically identical semi-Markov chains, $\{X_t\}$ and $\{X_t'\}$. As a trivial example, all ordinary renewal processes are equivalent, since in each case the associated semi-Markov chain exhibits no transitions.

Now if the semi-Markov kernels F and F' are equivalent then the associated semi-Markov processes $\{X_t\}$ and $\{X_t'\}$

(a) possess the same communication structure: that is if P and P' are the transition probability matrices of the embedded chains $\{\overline{X}_n\}$ and $\{\overline{X}_n'\}$,

$$p_{ij} > 0 \iff p_{ij}' > 0 \ , \ \forall i, \ \forall j \neq i$$

and (b) possess the same long-run properties; in particular if $\{X_t\}$ - and hence also $\{X_t'\}$ - is regular, the equilibrium state distributions are identical: that is

$$\sigma_i = \sigma_i' \ , \ \forall i \in \mathbb{N}_N$$

In optimal regulation problems it is only the communication structure and the long-run properties that are significant. It is therefore useful to widen the notion of equivalence and to say that two regular semi-Markov kernels F and F' (defined on the same state space and the same index set) are weakly equivalent if their associated semi-Markov processes possess the same communication structure and identical equilibrium state distributions. This weak type of equivalence is useful because the equilibrium properties of a regular chain depend only on the stationary probabilities, π_i, and the mean sojourn times, τ_i, and not on the precise form of its semi-Markov kernel F. Thus any changes to F that leave the products $\pi_i \tau_i$ unchanged will, by (2.43), produce a kernel F' that is weakly equivalent to F (provided that the changes do not alter the communication structure).

(Equivalence has been defined as a relation between semi-Markov kernels. In what follows the term "equivalent chains" means chains with equivalent kernels.)

Given any non-trivial regular chain $\{X_t\}$ it is possible to construct an equivalent canonical chain $\{X_t^o\}$ by suitably modifying the semi-Markov kernel, F, of $\{X_t\}$. (By a non-trivial chain we mean one whose recurrent subchain does not consist of a single absorbing state, so that $p_{ii} < 1$, $\forall i \in \mathbb{N}_N$.)

Let $\underline{\pi} \triangleq \text{Col}(\pi_1, \ldots, \pi_N)$ ....(2.45)

$\underline{\sigma} \triangleq \text{Col}(\sigma_1, \ldots, \sigma_N)$ ....(2.46)

$\underline{\tau} \triangleq \text{Col}(\tau_1, \ldots, \tau_N)$ ....(2.47)

The transformation to the canonical semi-Markov kernel F^o changes $\underline{\pi}$ and $\underline{\tau}$ but leaves $\underline{\sigma}$ unchanged. It works by replacing the embedded chain $\{\overline{X}_n\}$ by a new embedded chain $\{\overline{X}_n^o\}$ in which there are no self-transitions of the form $i \to i$. (Recall that the embedded chain of a continuous-time Markov chain is of this type.) Thus if $P^o \triangleq \left[p_{ij}^o\right]_{N \times N}$ is the transition probability matrix of $\{\overline{X}_n^o\}$, we choose, for all $i, j \in \mathbb{N}_N$,

$$p^o_{ij} = 0 \quad , \quad j = i$$

$$= \frac{p_{ij}}{1 - p_{ii}} \quad , \quad j \neq i \qquad \text{....(2.48)}$$

so that the p^o_{ij} are (for $j \neq i$) the next-jump probabilities associated with $\{\overline{X}_n\}$.

Now embed $\{\overline{X}^o_n\}$ in $\mathcal{T}$ via the Markov renewal process $\{\overline{X}^o_n, T^o_n\}$, where T^o_n is the time of the n^{th} change of state in the original semi-Markov process $\{X_t\}$: the associated semi-Markov chain $\{X^o_t\}$ is then equivalent to $\{X_t\}$. The relationship between $\{\overline{X}_n, T_n\}$ and $\{\overline{X}^o_n, T^o_n\}$ is illustrated in Fig.(3).

Although the chains $\{X_t\}$ and $\{X^o_t\}$ have the same sample-path properties, their semi-Markov kernels F and F^o are different. In fact, from (2.48) we have

$$\boxed{P^o = I - \Phi(I - P)} \qquad \text{....(2.49)}$$

where

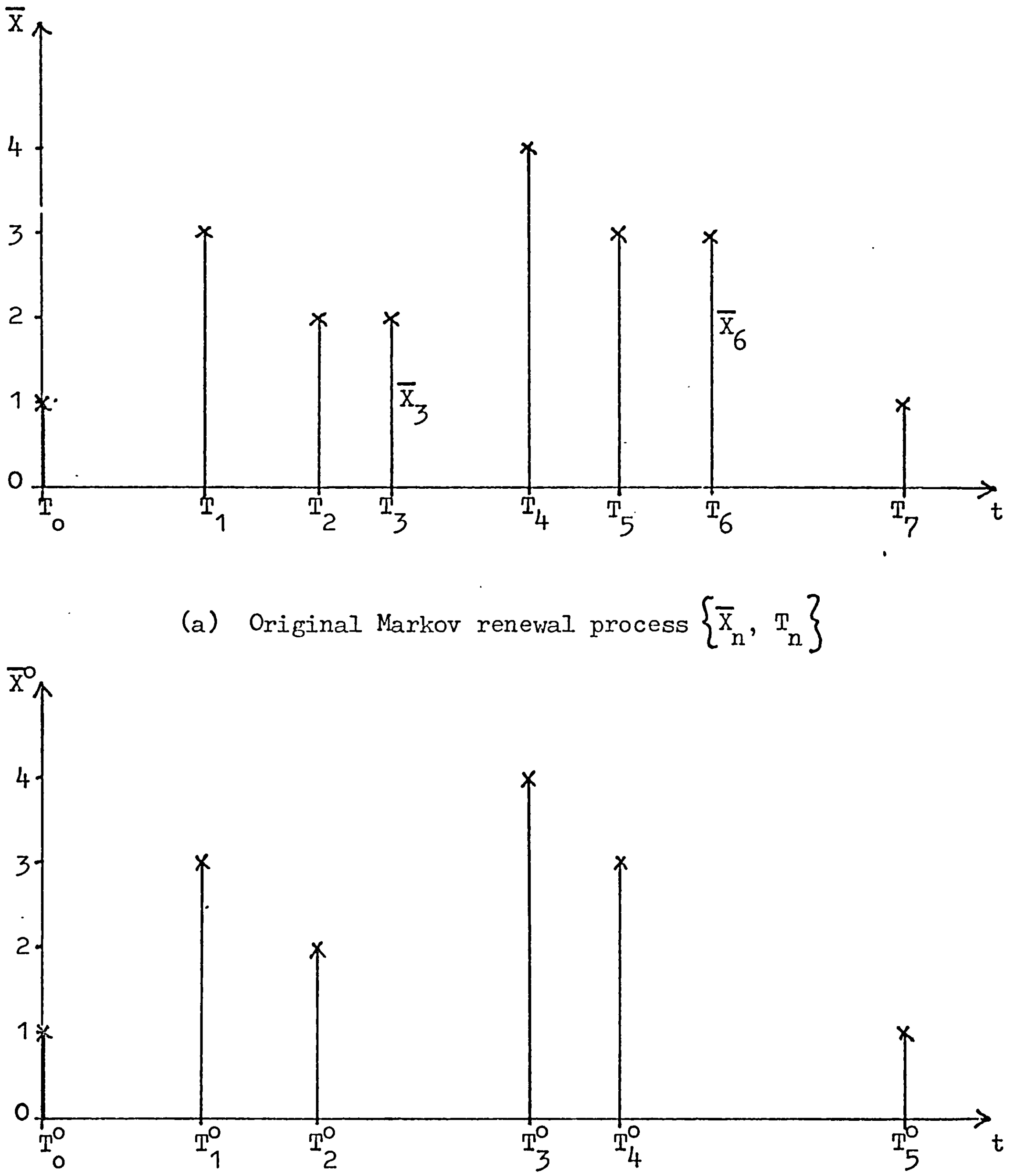
$$\Phi \triangleq \text{diag}(\phi_1, \ldots, \phi_N)$$

with

$$\phi_i \triangleq (1 - p_{ii})^{-1} \quad , \quad \forall i \in \mathbb{N}_N$$

Note that the matrix Φ is non-singular.

Except when P has a special structure, we can say that if P is regular then so will P^o be. For, from (2.49), the rank of $(I - P^o)$ equals the rank of $(I - P)$, which implies that the principal eigenvalue, $\lambda = 1$, has the same multiplicity in P^o as in P. Thus if P is regular P^o will have a simple eigenvalue $\lambda = 1$, and provided there are no other eigenvalues of P^o on the unit circle $|\lambda| = 1$, P^o will be regular. If P^o does have some unit-modulus eigenvalues other than $\lambda = 1$, then (see, for example, Chung(1967)) every recurrent state in $\{X^o_t\}$ is periodic with the same period and the matrix P^o is said to be periodic. (Properties R.1 - R.3 in Section 2.2.1



(a) Original Markov renewal process $\left\{\overline{X}_n, T_n\right\}$

(b) Canonical Markov renewal process $\left\{\overline{X}^o_n , \overline{T}_n\right\}$

Fig. (3)

then hold in modified form.)

We shall pay no further attention to periodic chains, since (a) results for regular chains are easily extended to the periodic case, and (b) we shall normally use the transformation $\{X_t\} \rightarrow \{X_t^o\}$ in conjunction with a second transformation $\{X_t^o\} \rightarrow \{X_t^*\}$ which will restore regularity.

The stationary distribution , $\underline{\pi}^o$, of P^o satisfies

$$(\underline{\pi}^o)^T = (\underline{\pi}^o)^T P^o \quad ,$$

ie., using (2.49),

$$(\underline{\pi}^o)^T \Phi (I - P) = \underline{o}^T \qquad \ldots.(2.50)$$

But since $\underline{\pi}^T$ is the unique stationary distribution of P, and since Φ is non-singular, the only solutions of (2.50) are

$$\boxed{(\underline{\pi}^o)^T = c \, \underline{\pi}^T \Phi^{-1}} \quad , \; c \in R \qquad \ldots.(2.51)$$

and to make $\underline{\pi}^o$ a probability distribution, we require that

$$c = (\underline{\pi}^T \Phi^{-1} \underline{e})^{-1} \qquad \ldots.(2.52)$$

Now consider the new sojourn-time distributions, H_i^o, i = 1,...,N. These are not, in general, simply related to the original distributions H_i; we can however, by means of a simple renewal argument,relate the new mean sojourn times τ_i^o to the original times τ_i. We have, at the stopping time $T_m^o = T_n$,

$$E\left[\Delta T_{m+1}^o \,\middle|\, \overline{X}_m^o = i\right] = E\left[\Delta T_{m+1}^o \,\middle|\, \overline{X}_n = i\right]$$

$$= E\left[E\left[\Delta T_{m+1}^o \,\middle|\, \overline{X}_n = i, \overline{X}_{n+1}\right]\right]$$

$$= P\left[\overline{X}_{n+1} \neq i \,\middle|\, \overline{X}_n = i\right] E\left[\Delta T_{m+1}^o \,\middle|\, \overline{X}_n = i \; , \; \overline{X}_{n+1} \neq i\right]$$

$$+ P\left[\overline{X}_{n+1} = i \,\middle|\, \overline{X}_n = i\right] E\left[\Delta T_{m+1}^o \,\middle|\, \overline{X}_n = i, \overline{X}_{n+1} = i\right]$$

$$= (1 - p_{ii}) \; E\left[\Delta T_{n+1} \,\middle|\, \overline{X}_n = i\right] \quad +$$

$$+ \; p_{ii} \left(E\left[\Delta T_{n+1} \middle| \bar{X}_n = i\right] + E\left[\Delta T^o_{m+1} \middle| \bar{X}^o_m = i\right] \right),$$

where, in deriving the second term on the right-hand side, we have used the fact that $(\bar{X}_{n+1} = i, T_{n+1})$ is a <u>regeneration point</u> for the process $\{\bar{X}_n, T_n\}$. The above relation expressed in terms of τ_i and τ^o_i is

$$\tau^o_i = (1 - p_{ii})\tau_i + p_{ii}(\tau_i + \tau^o_i) \qquad \text{....(2.53)}$$

and since this holds for each i in $\mathbb{N}_N$ we have

$$\boxed{\underline{\tau}^o = \Phi \, \underline{\tau}} \qquad \text{....(2.54)}$$

If we now use (2.51) and (2.54) in (2.43) we find that, for any j in $\mathbb{N}_N$,

$$\sigma^o_j = \frac{\pi^o_j \tau^o_j}{(\underline{\pi}^o)^T \underline{\tau}^o}$$

$$= \frac{c\,\pi_j \tau_j}{c\,\underline{\pi}^T \underline{\tau}}$$

so that $\underline{\sigma}^o = \underline{\sigma}$, as required for equivalence. Also, from (2.44), the new equilibrium sojourn time $\bar{\tau}^o$ is given by

$$\bar{\tau}^o = (\underline{\pi}^o)^T \underline{\tau}^o = c\bar{\tau} \qquad \text{....(2.55)}$$

where c is given by (2.52).

Since the equilibrium distribution $\underline{\sigma}^o$ of the canonical chain $\{X^o_t\}$ depends on the sojourn-time distributions H^o_i only via the vector $\underline{\tau}^o$ of mean sojourn times, any change in the H_i which leaves $\underline{\tau}^o$ unchanged will result in a chain weakly equivalent to $\{X^o_t\}$ and hence to the original chain $\{X_t\}$. In particular, if $\mathcal{T} = \mathbb{R}_+$ there exists a continuous-time Markov chain with sojourn-time distributions $H^*_i(t) = 1 - e^{-t/\tau^o_i}$, $\forall i \in \mathbb{N}_N$ which is weakly equivalent to $\{X^o_t\}$; and if $\mathcal{T} = \mathbb{Z}_+$ there exists a discrete-time Markov chain with geometric sojourn-time distributions $H_i^*(t) = 1 - (1 - (\tau_i^o)^{-1})^t$, $\forall i \in \mathbb{N}_N$ which is weakly equivalent to $\{X_t^o\}$ - <u>provided that</u> $\underline{\min_i (\tau_i^o) \geqslant 1}$ (for otherwise we would have at least one τ_i^o less than the time for one transition in the equivalent Markov chain, which is impossible.)

Given a canonical chain $\{X_t^o\}$ we can construct an equivalent chain $\{X_t^*\}$ having <u>the same mean sojourn time for each state</u>.

Let

$$\tau_{\min} \triangleq \min_{i \in \mathbb{N}_N} \left[\tau_i^o\right] \qquad \text{....(2.56)}$$

and, for some fixed k $\in (0,1]$, let

$$\boxed{\tau_o \triangleq k\tau_{\min}} \qquad \text{....(2.57)}$$

(If the chain $\{X_t^o\}$ is discrete (ie. $\mathcal{T} = \mathbb{Z}_+$) it is advantageous to choose k so that $\tau_o \in \mathcal{T}$; see comment after equation (2.71))

Then reduce all mean sojourn times to τ_o by introducing fictitious self-transitions i → i with probabilities p_{ii}^*, i = 1,...,N, such that, as in (2.54) ,

$$\boxed{\underline{\tau}^o = \Phi^* \underline{\tau}^*} \qquad \text{....(2.58)}$$

where

$$\Phi^* \triangleq \text{diag}\,(\phi_1^*,\ldots,\phi_N^*)$$

with

$$\phi_i^* \triangleq (1 - p_{ii}^*)^{-1} \ , \ \forall i \in \mathbb{N}_N$$

and where, to ensure that $\underline{\tau}^* \triangleq \text{Col}(\tau_1^*,\ldots,\tau_N^*) = \tau_o \underline{e}$ we must have

$$\tau_i^o = \phi_i^* \tau_o \ , \ \forall i \in \mathbb{N}_N \qquad \text{....(2.59)}$$

so that

$$p_{ii}^* = 1 - \frac{\tau_o}{\tau_i^o} \ , \ \forall i \in \mathbb{N}_N \qquad \text{....(2.60)}$$

and, as in (2.48), the off-diagonal elements of the new transition probability matrix P^* must satisfy

$$p^o_{ij} = \frac{p^*_{ij}}{1 - p^*_{ii}} , \quad j \neq i \qquad \text{....(2.61)}$$

Thus P^* must satisfy (see 2.49)

$$P^o = I - \Phi^*(I - P^*) \qquad \text{....(2.62)}$$

and so, since Φ^* is non-singular,

$$\boxed{P^* = I - (\Phi^*)^{-1} (I - P^o)} \qquad \text{....(2.63)}$$

The semi-Markov chain $\{X^*_t\}$ with transition probability matrix P^* given by (2.63) has, by (2.58 - 2.60), a mean sojourn time of τ_o in each state i, and it is easy to verify that for $\{X^*_t\}$ we have the equilibrium properties

$$(\underline{\pi}^*)^T = (c^*)^{-1} (\underline{\pi}^o)^T \Phi^* \qquad \text{....(2.64)}$$

where $$c^* = (\underline{\pi}^o)^T \Phi^* \underline{e} \qquad \text{....(2.65)}$$

and

$$\boxed{\underline{\sigma}^* = \underline{\sigma}^o} \qquad \text{....(2.66)}$$

so that $\{X^*_t\}$ is equivalent to $\{X^o_t\}$ and hence to any semi-Markov chain $\{X_t\}$ equivalent to $\{X^o_t\}$.

Note, that on using (2.58) in (2.65) with $\underline{\tau}^* = \tau_o \underline{e}$, we have

$$c^* = \frac{\bar{\tau}^o}{\tau_o} \qquad \text{....(2.67)}$$

so that, from (2.64),

$$\pi^*_i = \left(\frac{\tau^o_i}{\bar{\tau}^o}\right) \pi^o_i \quad , \forall i \in N_N \qquad \text{....(2.68)}$$

a result which is intuitively acceptable as the relationship between the stationary probabilities of the embedded chains of $\{X^*_t\}$ and $\{X^o_t\}$.

There are three final points to be made in this Section:

(i) It is clear that equivalence as defined in the present context is a proper equivalence relation on the class of finite semi-Markov kernels defined on the same index set and same state space. Each equivalence class contains a canonical kernel F^o characterized by zero self-transition probabilities; and, as we have shown above, each equivalence class also contains a kernel F^* which generates equal mean sojourn times for all states.

(ii) Suppose that equivalent to a given canonical chain $\{X_t^o\}$ we construct, by the procedure described above, a chain $\{X_t^*\}$ with mean sojourn times all equal to τ_o. The detailed properties of $\{X_t^*\}$ are determined by its semi-Markov kernel F^* and this is not simply related to F^o. However, we can construct a semi-Markov kernel F^+ which is weakly equivalent to F^*, and hence to F^o, by choosing

$$\begin{aligned} G_{ij}^+(t) &= 0 \quad , \quad t < \tau_o \\ &= 1 \quad , \quad t \geqslant \tau_o \end{aligned} \qquad \ldots.(2.69)$$

for every i, j $\in \mathbb{N}_N$, so that (almost) all sojourn times are equal to τ_o.

If we also choose

$$p_{ij}^+ = p_{ij}^* \quad , \quad \forall i,j \in \mathbb{N}_N \qquad \ldots.(2.70)$$

and then use (see (2.35))

$$F_{ij}^+(t) = p_{ij}^+ \, G_{ij}^+(t) \quad , \quad \forall i,j \in \mathbb{N}_N, \forall t \in \mathbb{R}_+ \qquad \ldots.(2.71)$$

we shall create a chain $\{X_t^+\}$, with semi-Markov kernel $F^+(t) \triangleq \left[F_{ij}^+(t)\right]_{N \times N}$, which is weakly equivalent to $\{X_t^*\}$ but whose transitions are regularly spaced in $\mathcal{T}$. (N.B. The reason for choosing k in equation (2.57) so that $\tau_o \in \mathcal{T}$ should now be clear: it is only possible to choose the G_{ij}^+ according to (2.69) if $\tau_o \in \mathcal{T}$)

(iii) The successive transformations $\{X_t\} \rightarrow \{X_t^o\}$ and $\{X_t^o\} \rightarrow \{X_t^*\}$ can be combined by using (2.49) in (2.63) to obtain

$$P^* = I - \Phi'(I - P) \qquad \text{....(2.72)}$$

where

$$\Phi' \triangleq \left(\Phi^*\right)^{-1} \qquad \text{....(2.73)}$$

It is easy to verify that

$$\Phi' = \text{diag}(\phi'_1, \ldots, \phi'_N) \qquad \text{....(2.74)}$$

where

$$\phi'_i = \frac{\tau_o}{\tau_i} \quad , \quad \forall i \in \mathbb{N}_N \qquad \text{....(2.75)}$$

and that P^* is a stochastic matrix.

The concept of equivalence defined in this Section turns out to be a useful one in the study of the optimal regulation problem, in which it is average equilibrium properties rather than detailed sample-path properties that are of principal interest.

2.3 Markov chains with costs

In this thesis we are concerned with Markov chains which represent dynamic systems with which are associated certain operating costs or running costs. The general problem considered is that of minimizing the average operating cost per unit time over a long period of time by controlling, where possible, the transition probabilities of the system. A precise statement of the problem is given in Section 3.2 of Chapter 3 . In this Section we define the cost structure to be considered and, since it is the long-run behaviour of the system which is of primary interest, we examine the asymptotic form of the total incurred cost as the operating time increases indefinitely.

2.3.1 Discrete-time chains with costs

Consider a homogeneous, regular, finite, discrete-time chain, $\{(X_t : \Omega \rightarrow \mathbb{N}_N) : t \in \mathbb{Z}_+\}$, with transition probability matrix P.

With each one-step transition $X_t \rightarrow X_{t+1}$ let there be associated a cost $c(X_t, X_{t+1})$ whose value depends only on X_t and X_{t+1}; more precisely, let there be a bounded function $c : \mathbb{N}_N^2 \rightarrow \mathbb{R}$ such that if $X_t = i$, $X_{t+1} = j$ the cost of the transition is $c(i,j)$, for all $(i,j) \in \mathbb{N}_N^2$.

Define, for each $i \in \mathbb{N}_N$,

$$\alpha_i \triangleq E\left[c(X_t, X_{t+1}) \,\middle|\, X_t = i\right] \qquad \text{....(2.76)}$$

Then

$$\alpha_i = \sum_{j \in \mathbb{N}_N} p_{ij}\, c(i,j) \quad , \quad \forall i \in \mathbb{N}_N \qquad \text{....(2.77)}$$

and since the costs are bounded all the α_i will be finite. α_i is the expected one-step cost from state i.

Define also, for each $i \in \mathbb{N}_N$, the conditional expectation

$$v_i(n) \triangleq E\left[\sum_{t=k}^{k+n-1} c(X_t, X_{t+1}) \,\middle|\, X_k = i\right] \qquad \text{....(2.78)}$$

(Because $\{X_t\}$ is homogeneous the expectation is independent of k.)

Then (see, for example, Howard(1960)) the $v_i(n)$ satisfy the recurrence relations

$$v_i(n) = \alpha_i + \sum_{j \in \mathbb{N}_N} p_{ij}\, v_j(n-1) \quad , \quad \forall i \in \mathbb{N}_N \qquad \text{....(2.79)}$$

or, in vector form,

$$\boxed{\underline{v}(n) = \underline{\alpha} + P\,\underline{v}(n-1)} \qquad \text{....(2.80)}$$

where

$$\begin{cases} \underline{\alpha} \triangleq \text{Col}\,(\alpha_1, \ldots, \alpha_N) \\ \underline{v}(n) \triangleq \text{Col}\,(v_1(n), \ldots, v_N(n)) \end{cases}$$

From (2.80) we deduce immediately that, with $\underline{\Delta v}_n \triangleq \underline{v}(n) - \underline{v}(n-1)$,

$$\underline{\Delta v}_n = P^{n-1}\,\underline{\Delta v}_1 \qquad \text{....(2.81)}$$

and assuming, as we always shall do, that $\underline{v}(o) = \underline{0}$ (no terminal costs) we find from (2.80) that $\underline{\Delta v}_1 = \underline{\alpha}$, so that

$$\boxed{\underline{\Delta v}_n = P^{n-1}\underline{\alpha}} \qquad \text{....(2.82)}$$

Now since the chain $\{X_t\}$ is regular, equation (2.13) holds so that (2.82) can be written as

$$\underline{\Delta v}_n = \underline{e}\,\underline{\pi}^T\underline{\alpha} + \tilde{P}^{n-1}\underline{\alpha} \qquad \text{....(2.83)}$$

whence

$$\boxed{\lim_{n\to\infty} \underline{\Delta v}_n = \bar{\alpha}\,\underline{e}} \qquad \text{....(2.84)}$$

where

$$\boxed{\bar{\alpha} \triangleq \underline{\pi}^T\underline{\alpha}} \qquad \text{....(2.85)}$$

Note that $\underline{\pi}$ is the stationary distribution of $\{X_t\}$ so that $\bar{\alpha}$ is the equilibrium one-step expected cost for $\{X_t\}$. Equation (2.84) states that, whatever the state X_t, for sufficiently large n the expected value of the cost increment $c(X_{t+n-1}, X_{t+n})$ is $\bar{\alpha}$. It follows from (2.83) that $\underline{v}(n)$ is asymptotically linear in n : that is

$$\lim_{n\to\infty}\left[\underline{v}(n) - n\bar{\alpha}\,\underline{e}\right] = \underline{\omega} ,$$

which we write as

$$\boxed{\underline{v}(n) \sim n\bar{\alpha}\,\underline{e} + \underline{\omega}} , \qquad \text{....(2.86)}$$

where, to ensure that (2.80) is satisfied, the constant vector $\underline{\omega}$ must satisfy

$$(I - P)\,\underline{\omega} = \underline{\alpha} - \bar{\alpha}\,\underline{e} \qquad \text{....(2.87)}$$

Since $\underline{e}$ is a null-vector of $(I - P)$, $\underline{\omega}$ is not completely determined by (2.87). But we also have, from (2.80), that

$$\underline{\pi}^T\,\underline{v}(n) = n\bar{\alpha} + \underline{\pi}^T\,\underline{v}(o) ,$$

which, on using (2.86), gives

$$\underline{\pi}^T \underline{\omega} = \underline{\pi}^T \underline{v}(0) = 0 \qquad \text{....(2.88)}$$

The unique solution to (2.87) and (2.88) is

$$\boxed{\underline{\omega} = (I - \tilde{P})^{-1} \underline{\alpha} - \bar{\alpha}\underline{e}} \qquad \text{....(2.89)}$$

Note that the inverse $(I - \tilde{P})^{-1}$ always exists when $\{X_t\}$ is regular.

We shall refer to $\bar{\alpha}$ as the cost rate of the chain $\{X_t\}$ and $\underline{\omega}$ as the corresponding value vector. It plays an important role in some of the optimization algorithms to be discussed in Chapter 4.

2.3.2 Continuous-time chains with costs

Now consider a homogeneous, regular, continuous-time chain $\{(X_t: \Omega \to N_N): t \in R_+\}$ with transition intensity matrix Q. Suppose that the transition times of $\{X_t\}$ are $T_1, T_2, T_3, \ldots$ so that, if $\{\bar{X}_n\}$ is the embedded chain, $X_t = \bar{X}_n$ when $t \in [T_n, T_{n+1})$. Associate with $\{X_t\}$ a running cost of $c(X_t)$ per unit time; more precisely, let there be a cost function $c: N_N \times R_+ \to R$ such that if $X_{T_n} = i$ and $\Delta T_{n+1} > t$, the cost incurred between time T_n and time $T_n + t$ is $c(i).t$. Also associate with $\{X_t\}$ a transition cost $d(X_{T_n}, X_{T_{n+1}})$, that is, a function $d : N_N^2 \to R$ such that if $X_{T_n} = i$ and $X_{T_{n+1}} = j$ the transition cost incurred is $d(i,j)$. (Since $j \neq i$, the values of $d(i,i)$ are of no significance.)

Define, for each $i \in N_N$,

$$\beta_i \triangleq \lim_{\Delta t \to 0} E\left[\frac{c(X_{t+\Delta t}) - c(X_t)}{\Delta t} \,\middle|\, X_t = i\right]$$

Then (see Howard(1960))

$$\beta_i = c(i) + \sum_{j \neq i} q_{ij}\, d(i,j) \quad , \quad \forall i \in N_N \qquad \text{....(2.90)}$$

β_i is the expected cost rate in state i.

Define also the conditional expectation

$$v_i(t) \triangleq E\left[\left\{\int_{t_o}^{t_o+t} c(X_t)\,dt + \underbrace{\sum_{k \ni}}_{T_{k-1},T_k \in (t_o,t_o+t)} d(X_{T_{k-1}},X_{T_k})\right\} \middle| X_{t_o} = i\right] \qquad \text{....(2.91)}$$

(The expectation is independent of t_o by homogeneity.)

Then (see Howard$^{(1960)}$) the $v_i(t)$ satisfy the differential equations

$$\dot{v}_i(t) = \beta_i + \sum_{j \in \mathbb{N}_N} q_{ij}\, v_j(t) \;, \quad \forall i \in \mathbb{N}_N \qquad \text{....(2.92)}$$

or, in vector form,

$$\boxed{\underline{\dot{v}}(t) = \underline{\beta} + Q\,\underline{v}(t)} \qquad \text{....(2.93)}$$

where
$$\left\{\begin{array}{l} \underline{\beta} \triangleq \text{Col}\,(\beta_1,\ldots,\beta_N) \\ \underline{v}(t) \triangleq \text{Col}\,(v_1(t),\ldots,v_N(t)) \end{array}\right.$$

Equation (2.93) is clearly the continuous-time analogue of equation (2.80), and as in the discrete-time case we can show that $\underline{v}(t)$ is asymptotically linear in t :

$$\boxed{\underline{v}(t) \sim t\,\bar{\beta}\,\underline{e} + \underline{\omega}} \qquad \text{....(2.94)}$$

where

$$\boxed{\bar{\beta} \triangleq \underline{\sigma}^T \underline{\beta}} \qquad \text{....(2.95)}$$

and

$$\boxed{\underline{\omega} = (\underline{e}\,\underline{\sigma}^T - I)\,(\underline{e}\,\underline{\sigma}^T + Q)^{-1}\,\underline{\beta}} \;, \qquad \text{....(2.96)}$$

$\underline{\sigma}$ being the stationary distribution of $\{X_t\}$.

2.3.3 Semi-Markov chains with costs

We now come to the general case. Consider a regular semi-Markov chain $\{(X_t : \Omega \to \mathbb{N}_N) : t \in \mathcal{T}\}$ with semi-Markov kernel F. Using the notation of Section 2.2.3 and letting $\Delta t \triangleq t - T_n$ for all $t \in [T_n, T_{n+1})$, suppose that with each transition $X_{T_n} \to X_{T_{n+1}}$ there is associated a cost which accumulates, as Δt increases from 0 to ΔT_{n+1}, according to the cost function $c : \Delta t \mapsto c(\Delta t \, ; X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1})$. We assume that c is a non-decreasing function of Δt and that $c(0; X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}) = 0$.

Now define the transition cost

$$C(X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}) \triangleq c(\Delta T_{n+1}; X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1})$$

and, for each $i \in \mathbb{N}_N$,

$$\gamma_i \triangleq E\left[C(X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}) \,\middle|\, X_{T_n} = i\right] \qquad \text{....(2.97)}$$

Then $\underline{\gamma} \triangleq \mathrm{Col}(\gamma_1, \ldots, \gamma_N)$ is the vector of <u>expected one-step costs</u> of the chain $\{X_t\}$.

Also define, for each $i \in \mathbb{N}_N$,

$$v_i(n) \triangleq E\left[\sum_{k=k_o}^{k_o+n-1} C(X_{T_k}, X_{T_{k+1}}, \Delta T_{k+1}) \,\middle|\, X_{T_{k_o}} = i\right] \qquad \text{....(2.98)}$$

and

$$t_i(n) \triangleq E\left[(T_{k_o+n} - T_{k_o}) \,\middle|\, X_{T_{k_o}} = i\right] \qquad \text{....(2.99)}$$

Then (see Appendix) the vectors

$\underline{v}(n) \triangleq \mathrm{Col}(v_1(n), \ldots, v_N(n))$ and $\underline{t}(n) \triangleq \mathrm{Col}(t_1(n), \ldots, t_N(n))$

satisfy the recurrence relations

$$\underline{v}(n) = \underline{\gamma} + P\,\underline{v}(n-1) \qquad \text{....(2.100)}$$

and

$$\underline{t}(n) = \underline{\tau} + P\,\underline{t}(n-1) \qquad \text{....(2.101)}$$

with initial conditions $\underline{v}(o) = \underline{t}(o) = \underline{0}$.

It follows, by an argument parallel to that following (2.80), that when $\{X_t\}$ is regular

$$\lim_{n \to \infty} \underline{\Delta v}_n = \bar{\gamma}\, \underline{e} \qquad \text{....(2.102)}$$

and

$$\lim_{n \to \infty} \underline{\Delta t}_n = \bar{\tau}\, \underline{e} \qquad \text{....(2.103)}$$

where

$$\left\{ \begin{array}{ll} \bar{\gamma} \triangleq \underline{\pi}^T \underline{\gamma} & \text{(equilibrium mean cost/transition)} \\ \bar{\tau} \triangleq \underline{\pi}^T \underline{\tau} & \text{(equilibrium mean time/transition)} \end{array} \right.$$

and

It follows that, for any $i \in \mathbb{N}_N$,

$$\lim_{n \to \infty} \left[\frac{v_i(n) - v_i(n-1)}{t_i(n) - t_i(n-1)} \right] = \bar{c} \qquad \text{....(2.104)}$$

where

$$\boxed{\bar{c} \triangleq \frac{\bar{\gamma}}{\bar{\tau}}} \qquad \text{....(2.105)}$$

It seems reasonable to speak of $\bar{c}$ as the <u>cost rate</u> associated with the chain $\{X_t\}$.

In fact $\underline{v}(n)$ and $\underline{t}(n)$ are asymptotically linear in n :

$$\underline{v}(n) \sim n \bar{\gamma}\, \underline{e} + \underline{\zeta} \qquad \text{....(2.106)}$$

$$\underline{t}(n) \sim n \bar{\tau}\, \underline{e} + \underline{\xi} \qquad \text{....(2.107)}$$

where $\underline{\zeta}$ and $\underline{\xi}$ are fixed vectors.

Multiply (2.107) by $\bar{c}$ and subtract from (2.106): we obtain

$$\boxed{\underline{v}(n) \sim \bar{c}\, \underline{t}(n) + \underline{\delta}} \qquad \text{....(2.108)}$$

where $\underline{\delta} \triangleq \underline{\zeta} - \bar{c}\, \underline{\xi}$.

Thus the expected total cost over n transitions increases, for sufficiently large n, linearly with the expected total time for the transitions, the rate of increase being $\bar{c}$.

Furthermore, on multiplying (2.101) by $\bar{c}$ and subtracting from (2.100), we obtain

$$\underline{v}(n) - \bar{c}\,\underline{t}(n) = \underline{\gamma} - \bar{c}\,\underline{\tau} + P\left[\underline{v}(n-1) - \bar{c}\,\underline{t}(n-1)\right] \qquad \text{....(2.109)}$$

Now let $n \rightarrow \infty$ and use (2.108) : there results

$$\boxed{\underline{\delta} = \underline{\gamma} - \bar{c}\,\underline{\tau} + P\,\underline{\delta}} \qquad \text{....(2.110)}$$

This equation is the semi-Markov generalization of (2.87), and as before we shall call $\underline{\delta}$ the value-vector. As with (2.87) equation (2.110) determines the cost rate $\bar{c}$ uniquely but the value-vector $\underline{\delta}$ only to within an additive constant vector $k\,\underline{e}$.

Equation (2.108) says nothing about the behaviour of the expected total cost when the time to go is some fixed time $t \in \mathcal{T}$. We do, however, have the following result due to Ross$^{(1969, 1970)}$.

Let, for each $i \in \mathbb{N}_N$, and each $t \in \mathcal{T}$,

$$v_i'(t) \triangleq v_i(M_t) + E\left[\int_{T_{k_0+M_t}}^{T_{k_0}+t} c(t; X_{T_{M_t}}, X_{T_{m_t+1}}, \Delta T_{M_t+1})\,dt \,\middle|\, X_{T_{k_0}} = i\right]$$

where M_t is the number of transitions in $(T_{k_0}, T_{k_0}+t]$; that is, $v_i'(t)$ is the expected total cost accumulated between T_{k_0} and $T_{k_0}+t$, given that state i was entered at time T_{k_0}.

Then (see Appendix)

$$\boxed{\lim_{t \to \infty} \left[\frac{v_i'(t)}{t}\right] = \bar{c}} \quad , \quad \forall i \in \mathbb{N}_N \qquad \text{....(2.111)}$$

provided that $\{X_t\}$ is regular, with a finite mean sojourn time in each state.

This result permits us to interpret $\bar{c}$ as the long-run average cost per unit time of the chain $\{X_t\}$.

2.3.4 Equivalence with costs

We have seen in the previous section that if $\{X_t\}$ is a regular semi-Markov chain with additive costs, the cost rate $\bar{c}$ of $\{X_t\}$ is given by (see (2.105))

$$\boxed{\bar{c} = \frac{\underline{\pi}^T \underline{\gamma}}{\underline{\pi}^T \underline{\tau}}} \qquad \text{....(2.112)}$$

where $\underline{\pi}$ is the stationary distribution of P (the transition probability matrix of the embedded chain), $\underline{\gamma}$ is the vector of expected one-step costs, and $\underline{\tau}$ is the vector of mean sojourn times.

Suppose now that the equivalence transformation defined by (2.49) is applied to $\{X_t\}$ to produce a canonical chain $\{X_t^o\}$, with mean one-step cost vector $\underline{\gamma}^o$, mean sojourn-time vector $\underline{\tau}^o$, and cost rate $\bar{c}^o$.

By a renewal argument exactly parallel to that leading to (2.54), we find that

$$\boxed{\underline{\gamma}^o = \Phi \underline{\gamma}} \qquad \text{....(2.113)}$$

and hence, using (2.51), (2.54), (2.113) in (2.112),

$$\boxed{\bar{c}^o = \bar{c}} \qquad \text{....(2.114)}$$

Thus the cost rate $\bar{c}$ is invariant under the equivalence transformation $\{X_t\} \to \{X_t^o\}$; and it is clear that the same is true of the transformation $\{X_t^o\} \to \{X_t^*\}$ defined by (2.63), ie.

$$\boxed{\bar{c}^* = \bar{c}} \qquad \text{....(2.115)}$$

where $\bar{c}^*$ is the cost rate associated with the equal-sojourn-time chain, $\{X_t^*\}$, equivalent to $\{X_t\}$. So, given a regular semi-Markov chain $\{X_t\}$ with cost rate $\bar{c}$, there exists an equivalent chain with the same cost rate but with equal mean sojourn times in all states. Furthermore, since $\bar{c}$ is the mean cost per unit time, and τ_o is the time per transition, of the equal-sojourn-time chain $\{X_t^*\}$, we can interpret $\bar{c}\,\tau_o$ as the mean cost per transition of the embedded Markov chain, $\{\overline{X}_n^*\}$, of $\{X_t^*\}$. Thus any regular semi-Markov chain $\{X_t\}$ with mean cost rate $\bar{c}$ has associated with it a discrete-time Markov chain, $\{X_n^*\}$, whose mean cost rate (measured as a cost/transition) is $\bar{c}\,\tau_o$. We shall make use of this important result in the next Chapter.

Note, finally, that the results presented in Section 2.3 rest on the additive nature of the cost structure that has been assumed: the total cost accumulated over n transitions is the sum of the n individual transition costs. Concrete results are difficult to obtain for any other type of cost structure, but fortunately in most applications the transition costs are additive.

CHAPTER 3

THE OPTIMAL REGULATION PROBLEM

3.1 Introduction

We now consider Markov and semi-Markov chains with costs, whose transition mechanisms can be modified by the application of control signals. If knowledge of the current state is available, suitable control signals can be generated by a feedback scheme, and if the number of possible control actions is finite the feedback scheme can usually be designed to minimize the cost rate of the chain. The problem of choosing such a feedback scheme is called the optimal regulation problem and in this Chapter we formulate the problem more precisely and then review existing methods for its solution.

3.2 Optimal Regulation

It will be convenient to develop the ideas in terms of semi-Markov chains and then to show where the extra structure of pure Markov chains leads to simplifications and/or stronger results. As hitherto, attention is restricted to finite chains.

3.2.1 Controllable chains

A semi-Markov chain $\{(X_t: \Omega \to \mathbb{N}_N) : t \in \mathcal{T}\}$ with semi-Markov kernel F is said to be a controllable semi-Markov chain (CSMC) if the elements of F depend on the value of some scalar control variable u. More precisely, $\{X_t\}$ is a CSMC if there exists a set of controls $\mathcal{U}$, such that for each $(i,j) \in \mathbb{N}_N^2$ we can define the function F_{ij} : $\mathcal{T} \times \mathcal{U} \to \mathbb{R}_+$ by

$$F_{ij}(t;u) \triangleq P\left[\overline{X}_{n+1} = j,\ \Delta T_{n+1} \leqslant t \,\middle|\, \overline{X}_n = i\ ;\ u\right] \quad \ldots.(3.1)$$

The operational interpretation of (3.1) is as follows. At any transition time T_n a value of u is selected from the set $\mathcal{U}$ and applied as a control input until the next transition time T_{n+1}; call this value u_n. Then $\{X_t\}$ makes a transition governed by the semi-Markov kernel $F(t; u_n)$, arriving at state $\overline{X}_{n+1}$ at time T_{n+1}. A new control value u_{n+1} is then selected and applied throughout the next transition $\overline{X}_{n+1} \longrightarrow \overline{X}_{n+2}$; and so on.

In this thesis we shall always assume that the control set is finite , ie. that

$$\mathcal{U} = \left\{ u^1, u^2, \ldots\ldots, u^k \right\}$$

and we shall refer to the elements u^i as control actions.

We shall further assume, without loss of generality, that for each state i the subset $\mathcal{U}_i \subset \mathcal{U}$ of feasible control actions in state i is $\mathcal{U}$ itself, ie. that the choice of control action is not restricted by the state currently occupied.

A fixed sequence of controls $(u_o, u_1, u_2, \ldots\ldots)$, chosen before the starting time T_o of the process, is called a control schedule. With such a schedule the choice of control on, say, the interval $[T_k, T_{k+1})$ is predetermined : there is no feedback of information about the past and present motion of the chain $\{X_t : t \leqslant T_k\}$ to the process of selecting u_k. In the majority of practical control problems, however, such information is available for feedback purposes and can therefore be used to implement a control policy in which each control, u_k, is made a function of the currently available information about the motion of the chain. The problem of choosing a satisfactory control policy is, in engineering terms, that of designing a suitable on-line controller whose inputs are data concerning past and present behaviour of the process; in mathematical terms, the aim is to find a suitable sequence of mappings from the space of available data histories into the control set $\mathcal{U}$.

Throughout this thesis attention is confined to so-called *completely observable* (or *perfectly observable*) chains (see, for example, Mayne(1967)), in which the current state $\overline{X}_n$ is known without ambiguity at every stopping time T_n. By the strong Markov property, the behaviour of such a chain after time T_n, given the value of $\overline{X}_n$, is independent of the past history of the chain, in which case the control action u_n may, without loss of generality, be taken to be a function of $\overline{X}_n$ alone. Such a function is called a *control law*.

Let $f_n : \mathbb{N}_N \to \mathcal{U}$ be the control law for the chain $\{X_t\}$ at the stopping time T_n, so that if $\overline{X}_n = i$ then $u_n = f_n(i)$. The sequence of control laws

$$(f)_n \triangleq (f_o, f_1, \ldots, f_{n-1})$$

is called a *control policy* for the chain $\{X_t\}$ on the interval $[0, T_n)$. Once $(f)_n$ is determined the controlled chain $\{X_t\}$ becomes an ordinary (but, in general, non-homogeneous) semi-Markov chain on the given interval.

If the objective is to control the chain over an indefinitely long period the control policy will be an infinite sequence of control laws. In particular, an infinite sequence of identical control laws, $(f) \triangleq (f, f, f, \ldots.)$, is called a *stationary policy*. When a stationary control policy is used the resulting CSMC is homogeneous in time and has a well-defined long-run behaviour. In particular, if the policy (f) is such that the controlled chain is regular there will be an equilibrium state distribution which is independent of the initial state of the chain and hence a unique equilibrium cost rate associated with the chain.

A controllable chain which, for every possible stationary policy (f), is regular and has finite mean sojourn times will be called a *totally regular* chain. Although we shall confine our attention to such chains it should be pointed out that most of the optimization methods discussed in the thesis are, with suitable modification, applicable to chains which may be non-regular under certain policies (see, for example, Denardo and Fox(1968)). There is, in any case, a wide range of practical applications in which a totally regular CSMC is the appropriate system model.

When associating a cost with a controllable chain it is natural to allow the cost for any transition $X_{T_n} \longrightarrow X_{T_{n+1}}$ to depend on the control action u_n as well as on the states X_{T_n} , $X_{T_{n+1}}$ and the sojourn time ΔT_{n+1}. In order to allow for control costs we assume a cost function of the form $c : \Delta t \mapsto c(\Delta t ; X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}, u_n)$ instead of that used in Section 2.3.3, so that the transition cost becomes $C(X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}, u_n)$. Then, generalizing (2.97), we define, for each $i \in \mathbb{N}_N$, each $u \in \mathcal{U}$,

$$\gamma_i^u \triangleq E\left[C(X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}, u_n) \,\middle|\, X_{T_n} = i , u_n = u\right] \qquad \text{....(3.2)}$$

γ_i^u is the expected one-step cost from state i under control action u. Similarly, τ_i^u will denote the expected sojourn time in state i under control u.

Consider a totally regular CSMC, $\{X_t\}$, controlled according to the stationary policy (f). Then if $X_{T_n} = i$ we shall have $u_n = u = f(i)$ and the corresponding transition functions defined by (3.1) will be $F_{ij}(t ; f(i))$, $\forall j \in \mathbb{N}_N$. The matrix

$$F^f(t) \triangleq \left[F_{ij}(t ; f(i))\right]_{N \times N} \qquad \text{....(3.3)}$$

is the closed-loop semi-Markov kernel under the stationary policy (f). From it, we can determine the closed-loop transition probability matrix P^f by (2.31) and the vector of closed-loop mean sojourn times $\underline{\tau}^f$ by (2.37) and (2.39). Similarly, the expected one-step cost from state i under the stationary policy (f) will be $\gamma_i^{f(i)}$, and the vector

$$\underline{\gamma}^f \triangleq \mathrm{Col}\,(\gamma_1^{f(1)},\ldots,\gamma_N^{f(N)}) \qquad \ldots.(3.4)$$

is thus the vector of expected one-step costs under (f).

Since the chain $\{X_t\}$ is totally regular the asymptotic relation (2.108) will hold for every stationary policy (f) ; that is, for large n,

$$\boxed{\underline{v}^f(n) \sim \bar{c}^f\ \underline{t}^f(n) + \underline{\delta}^f} \qquad \ldots.(3.5)$$

where the superscript f denotes quantities determined under the policy (f) : in particular

$$\boxed{\bar{c}^f \triangleq \frac{(\underline{\pi}^f)^T \underline{\gamma}^f}{(\underline{\pi}^f)^T \underline{\tau}^f}} \qquad \ldots.(3.6)$$

is the cost rate under (f).

3.2.2 Optimization of the control policy

The problem with which this thesis is concerned is as follows: determine a stationary control policy which will minimize the average operating cost per unit time of a given controllable semi-Markov chain which is expected to operate for an indefinitely long time. More precisely, given a totally regular CSMC $\{(X_t : \Omega \to \mathbb{N}_N) : t \in \mathcal{T}\}$ with a finite control set $\mathcal{U}$, find a control law $f^o : \mathbb{N}_N \to \mathcal{U}$ such that, if $f : \mathbb{N}_N \to \mathcal{U}$ is any other feasible control law, the equilibrium cost rate under the stationary policy (f) = (f,f,...) is not less than that under the stationary policy $(f^o) = (f^o, f^o, \ldots)$, ie. such that $\bar{c}^{f^o} \leqslant \bar{c}^f$ for all feasible f. If such an f^o can be found it is called an optimal control law and (f^o) is the corresponding optimal stationary policy.

Note that since the state space and the control set are both finite, the number of possible control laws (and hence the number of stationary policies) is finite (in fact K^N, where K is the number of possible control actions). It follows, since $\{X_t\}$ is totally regular, that at least one optimal stationary policy must exist. Uniqueness, on the other hand, is not guaranteed.

At this point there are two questions to consider. First, if the class of control policies is enlarged to include non-stationary policies, does there exist a stationary policy which is optimal in this larger class ? This question has been answered in the affirmative for pure Markov chains by Blackwell(1962) and as we shall show in the next Section the result is also applicable to semi-Markov chains. Secondly, if we allow randomized control laws (that is, control laws which map the state space into the space of probability distributions on $\mathcal{U}$ instead of into $\mathcal{U}$ itself) may we thereby achieve a lower value of $\bar{c}$? The answer is no (see Wolfe/Dantzig(1962) and Osaki/Mine(1968)): there exists at least one pure (non-randomized) control law such that the corresponding stationary policy is optimal in the wider class of randomized policies.

Thus in order to design a controller which will cause the CSMC $\{X_t\}$ to operate at minimal average cost per unit time under equilibrium conditions we need to determine an optimal control law f^o : $\mathbb{N}_N \rightarrow \mathcal{U}$. The problem of finding such a function is the optimal regulation problem.

Note on terminology : In the operations research literature it is usual to refer to the elements of $\mathcal{U}$ as possible decisions ; a controllable semi-Markov chain is called a semi-Markov decision process and the problem of finding an optimal f is called a semi-Markov programming problem. Our terminology, which extends that of Aström(1965), seems more appropriate in a control engineering context.

3.2.3 Equivalent regulation problems

Consider a totally regular CSMC $\{(X_t : \Omega \to \mathbb{N}_N) : t \in \mathcal{T}\}$ with control set $\mathcal{U}$ and transition cost function c. Under a given stationary policy (f), the chain $\{X_t\}$ has a stationary distribution $\underline{\pi}^f$, a mean sojourn-time vector $\underline{\tau}^f$, and a mean one-step cost vector $\underline{\gamma}^f$, from which we may determine the equilibrium state distribution $\underline{\sigma}^f$ by (2.43) and the equilibrium cost rate $\bar{c}^f$ by (2.112).

By using (2.43) in (2.112) we obtain an alternative expression for the equilibrium cost rate:

$$\boxed{\bar{c}^f = (\underline{\sigma}^f)^T \underline{\beta}^f} \qquad \text{....(3.7)}$$

where

$$\underline{\beta}^f \triangleq \text{Col}(\beta_1^f, \ldots, \beta_N^f)$$

with

$$\beta_i^f \triangleq \left(\frac{\gamma_i^f}{\tau_i^f}\right) \quad , \quad \forall i \in \mathbb{N}_N$$

The component β_i^f is the ratio of the expected transition cost to the expected transition time for state i (under the given policy (f)), and so if i is recurrent ($\sigma_i^f > 0$) we can interpret β_i^f as the long-run average cost rate associated with state i.

The result contained in (2.94 - 5) of Section 2.3.2 is the special case of (3.7) which arises when $\mathcal{T} = \mathbb{R}_+$, the sojourn-time distributions are all exponential, and the cost function c is linear in t, ie. when $\{X_t\}$ is a continuous-time Markov chain. In this particular case, β_i^f has the stronger interpretation as the instantaneous rate at which the expected cost grows in state i.

Similarly if $\mathcal{T} = \mathbb{Z}_+$ and all the sojourn times are unity (under the given policy (f)) then $\underline{\tau}^f = \underline{e}$ in (3.6) and so we get $\bar{c}^f = (\underline{\pi}^f)^T \underline{\gamma}^f$ which is just the result contained in (2.85 - 6) for discrete-time Markov chains.

In viewing Markov chains as special cases of semi-Markov chains we are led to the interesting question : is it possible to re-formulate the optimal regulation problem for a controllable semi-Markov chain as an "equivalent" pure Markov regulation problem ? To make sense of the question let us first define two totally regular controllable semi-Markov chains $\{{}_1X_t\}$ and $\{{}_2X_t\}$ to be <u>totally equivalent</u> iff

(i) $\{{}_1X_t\}$ and $\{{}_2X_t\}$ possess the same finite state space $\mathcal{N}_N$, the same index set $\mathcal{T}$, and the same finite control set $\mathcal{U}$;

(ii) for every stationary control policy (f), $\{{}_1X_t\}$ and $\{{}_2X_t\}$ are weakly equivalent in the sense defined in Section 2.2.4.

It is clear that a policy (f^o) which is optimal for $\{{}_1X_t\}$ is also optimal for $\{{}_2X_t\}$, and vice versa. Thus $\{{}_1X_t\}$ and $\{{}_2X_t\}$ possess the same optimal policies, with the same associated minimum cost rate $\bar{c}^{f^o}$. The optimal regulation problems for $\{{}_1X_t\}$ and $\{{}_2X_t\}$ are then <u>equivalent</u> in the sense that any solution to one of the problems is a solution to the other.

The answer to the above question is thus that by carrying out appropriate equivalence transformations of the type defined in Section 2.2.4, it is possible to create a controllable discrete-time Markov chain which is totally equivalent to any given totally regular CSMC, and hence to optimize the cost rate of the Markov chain rather than the original CSMC.

The required transformation is obtained as follows.

For each u $\in \mathcal{U}$, let

$$(\underline{\tau}^o)^u = \Phi^u \, \underline{\tau}^u \qquad \text{....(3.8)}$$

where

$$\Phi^u \triangleq \operatorname{diag}(\phi_1^u, \ldots, \phi_N^u)$$

with

$$\phi_i^u \triangleq (1 - p_{ii}^u)^{-1} \quad , \quad \forall i \in \mathcal{N}_N$$

Then find
$$\tau_{\min} \triangleq \min_{u} \min_{i} \left[(\tau_i^o)^u \right] \qquad \ldots.(3.9)$$

and set
$$\tau_o = K\tau_{\min} \qquad \ldots.(3.10)$$

with $K \in (0,1]$.

With τ_o determined we can then , for any policy (f), transform $\{X_t\}$ to a corresponding chain $\{X_t^*\}$ having all mean sojourn times equal to τ_o by using the transformation defined by (2.72); that is, by choosing
$$(P^*)^f = I - \Phi^f (I - P^f) \qquad \ldots.(3.11)$$

where
$$\Phi^f \triangleq \mathrm{diag}(\phi_1^f, \ldots, \phi_N^f) \qquad \ldots.(3.12)$$

with
$$\phi_i^f = \frac{\tau_o}{\tau_i^{f(i)}} \quad , \quad \forall i \in \mathbb{N}_N \qquad \ldots.(3.13)$$

From (2.113), the transformation (3.11) gives
$$(\underline{\gamma}^*)^f = \Phi^f \underline{\gamma}^f \qquad \ldots.(3.14)$$

for the vector of expected one-step costs associated with $\{X_t^*\}$; and we have chosen Φ^f so that
$$(\underline{\tau}^*)^f = \Phi^f \underline{\tau}^f = \tau_o \underline{e} \qquad \ldots.(3.15)$$

Thus, on using (3.15) in (3.6), the equilibrium cost rate, $(\bar{c}^*)^f$, of the equal-sojourn-time chain under the policy (f) is given by
$$\boxed{(\bar{c}^*)^f = \frac{1}{\tau_o} \left[(\underline{\pi}^*)^f \right]^T (\underline{\gamma}^*)^f} \qquad \ldots.(3.16)$$

where $(\underline{\pi}^*)^f$ is the stationary distribution of $(P^*)^f$.

Finally, bearing in mind comment (ii) at the end of Section 2.2 , there exists a discrete-time Markov chain with index set $\mathcal{T} = \{0, \tau_o, 2\tau_o, \ldots\}$, transition probability matrix $(P^*)^f$, and expected one-step cost vector $(\underline{\gamma}^*)^f$, whose equilibrium mean cost per unit time is given by (3.16), i.e. whose equilibrium mean cost per transition is $\left[(\underline{\pi}^*)^f\right]^T (\underline{\gamma}^*)^f$. We shall make use of this idea in the next Chapter.

The transformation (3.11) is well-defined for any feasible stationary policy (f); therefore by applying it to $\{X_t\}$ for <u>every</u> (f) we convert the original CSMC to a totally equivalent controlled Markov chain with the same optimal performance. It is important to note, however, that when making use of the concept of total equivalence in the design of optimization algorithms it is not necessary to carry out transformation (3.11) for every feasible stationary policy, but only for those policies, say $(f)_o$, $(f)_1$, $(f)_2$,...., arising as iterates in the course of the optimization. Furthermore if $(f)_n$ differs from $(f)_{n-1}$ only in state i, then $(P^*)^{f_n}$ will differ from $(P^*)^{f_{n-1}}$ only in the i^th row and hence is easily updated. We take advantage of this property in the successive-approximations algorithm described in Chapter 4 .

3.2.4 Related problems

Before describing existing methods for solving the optimal regulation problem we mention here two closely related control problems. In the first of these, the so-called <u>discounted-cost problem</u>, future accrued costs are discounted at a constant rate so that the expected total cost accumulated over an infinite operating period remains finite; then the optimal control problem is to find a policy which minimizes this cost. Under the conditions that we are assuming in this thesis (finite state space, total regularity) it may be shown (see Ross(1969)) that the discounted-cost problem "tends to" the regulation problem as the discount rate tends to zero, in the sense that for sufficiently small discount rate a policy which is optimal for the discounted-cost problem is also optimal for the regulation problem.

The second related problem is the so-called transient-cost problem, in which the only recurrent state is a single absorbing state, say i_o, and the optimal control problem is to find a policy which minimizes the total expected cost of travelling to the target state. This problem has been treated by Howard(1960), Eaton and Zadeh(1962), and, in particular, by Kushner and Kleinman(1968,1971). We mention it here because some of the algorithms suggested for the optimal regulation problem are adaptations of those used for the transient cost problem.

3.3 Existing methods of optimization

We now consider procedures for solving the optimal regulation problem. The problem has attracted considerable attention during the past 15 years and several methods have been developed for computing the optimal control law. Of these, the main ones are the policy-iteration method due to Howard(1960) and Jewell(1963), the successive-approximations method due to White(1963), and the linear programming method due to Manne(1960) and Osaki and Mine(1968).

3.3.1 Policy-iteration methods

Much of the interest in controllable Markov chains as system models has stemmed from Howard's pioneering work in this field. His policy-iteration algorithm for the discrete-time Markov regulation problem rests on the following argument. Let $\{(X_t : \Omega \to N_N): t \in \mathcal{Z}_+\}$ be a totally regular controllable Markov chain with finite control set $\mathcal{U}$ and bounded cost function $c : N_N^2 \times \mathcal{U} \to R$, with value $c(X_t, X_{t+1}, u_t)$ for the transition $X_t \xrightarrow{u_t} X_{t+1}$.

Define, for each $i \in N_N$, the optimal expected n-step cost from state i ,

$$V_i(n) \quad \underset{\substack{u_k \in \mathcal{U} \\ \vdots \\ u_{k+n-1} \in \mathcal{U}}}{\text{Min}} \; E\left[\sum_{t=k}^{k+n-1} c(X_t,\, X_{t+1},\, u_t) \,\middle|\, X_k = i\right] \qquad \text{....(3.17)}$$

Then $\frac{1}{n} V_i(n)$ is clearly the optimal expected value of the mean cost per transition over the n transitions from $X_k = i$.

A dynamic programming argument shows that the $V_i(n)$ satisfy the non-linear recurrence relations:

$$V_i(n) = \underset{u \in \mathcal{U}}{\text{Min}}\left[\alpha_i^u + \sum_{j \in \mathbb{N}_N} p_{ij}^u \; V_j(n-1)\right], \quad \forall i \in \mathbb{N}_N \qquad \text{....(3.18)}$$

where

$$\alpha_i^u \triangleq E\left[c(X_t, X_{t+1}, u_t) \,\middle|\, X_t = i\right] \;, \quad \forall i \in \mathbb{N}_N \qquad \text{....(3.19)}$$

Now recursive solution of equations (3.18) will yield a sequence $(f_o, f_1, f_2, \ldots.)$ of control laws, namely, those control laws which minimize the successive right-hand sides of (3.18). In fact the sequence $(f_o, f_1, f_2, \ldots.)$ is an optimal control policy for the chain, though not in general a stationary one. However, it has been shown by Bellman$^{(1957)}$ that the sequence (f_n) tends, with increasing n, to a fixed control law; more precisely, there exists an n_o and an f^o such that for all $n \geqslant n_o$, $f_n = f^o$. Thus the policy $(f_o, f_1, f_2 \ldots.)$ is asymptotically stationary and it follows immediately that $V_i(n)$ is asymptotically linear in n. (Take n_o as a new time origin.) Furthermore, the stationary policy $(f^o) = (f^o, f^o, f^o \ldots)$ will clearly yield the same equilibrium cost rate , $\bar{\alpha}$, as the above non-stationary policy and no other stationary policy can yield a lower cost rate. We therefor seek the control law f^o which minimizes the right-hand side of (3.18) when n is large. But under the stationary policy (f^o) the asymptotic relation (2.86) holds; so if we subtract $n\bar{\alpha}$ from each side of (3.18) and let $n \to \infty$ we shall obtain

$$\boxed{\omega_i + \bar{\alpha} = \underset{u \in \mathcal{U}}{\mathrm{Min}} \left[\alpha_i^u + \sum_j p_{ij}^u \, \omega_j \right]} \quad , \forall i \in \mathbb{N}_N \qquad \ldots.(3.20)$$

This is a set of N non-linear simultaneous equations for the cost rate $\bar{\alpha}$ and value-vector $\underline{\omega}$ associated with an optimal policy.

Now let us define a control law f^o by

$$\forall i \in \mathbb{N}_N : \; f^o(i) = \mathrm{Arg.} \underset{u \in \mathcal{U}}{\min} \left[\alpha_i^u + \sum_j p_{ij}^u \, \omega_j \right] \qquad \ldots.(3.21)$$

Then the right-hand side of (3.20) can be written

$$\left[\alpha_i^{f^o(i)} + \sum_j p_{ij}^{f^o(i)} \, \omega_j \right]$$

and so (3.20) takes the form

$$\underline{\omega} + \bar{\alpha}\,\underline{e} = \underline{\alpha}^{f^o} + P^{f^o}\,\underline{\omega} \qquad \ldots.(3.22)$$

which is equation (2.87) under the stationary policy (f^o). Since $\bar{\alpha}$ and $\underline{\omega}$ are, by assumption, the solution to (2.87) under an optimal policy, it follows that (f^o) is an optimal policy.

Howard's procedure consists of iterating between (3.21) and (3.22) until an optimal policy is found, ie. until (3.21) and (3.22) are satisfied simultaneously (so that (3.20) is then satisfied). In algorithmic form the procedure is :-

(1) Choose an initial control law, f.

(2) "Value-determination" : With the given f solve (3.22) for $\bar{\alpha}$ and $\underline{\omega}$.

(3) "Policy-improvement" : With the given $\bar{\alpha}$ and $\underline{\omega}$ determine a new f by (3.21). If this differs from the previous f return

to step (2); otherwise the iteration process has converged, the latest f is an optimal control law, and the latest $\bar{\alpha}$ is the optimal cost rate.

Howard(1960) showed that the successive values of $\bar{\alpha}$ converge monotonically to the globally optimal value $\bar{\alpha}^o$. The number of iterations required is necessarily finite and in practice is usually very small compared with the number (K^N) of feasible control laws. A minor point is that, as we have seen, the solution $(\bar{\alpha}, \underline{\omega})$ of (3.22) is determined only to within an additive constant vector, k $\underline{e}$, in $\underline{\omega}$, until we impose some additional constraint on $\underline{\omega}$ such as $\omega_1 = 0$. A rigorous proof that any optimal policy for a totally regular chain must indeed satisfy (3.21) and (3.22) is given by Ross(1969).

The policy-iteration algorithm described above has been modified by Howard(1960) for application to the continuous-time Markov regulation problem. A much more important development is the extension of the Howard algorithm to cover the semi-Markov case. This has been achieved independently by Jewell(1963), Schweitzer(1969), and de Cani(1964). As we shall now show, the Jewell and Schweitzer algorithms can be developed rather elegantly from the Howard algorithm by application of appropriate equivalence transformations.

We first determine the effect of the equivalence transformation (3.11) on the value-vector $\underline{\delta}$ of a regular semi-Markov chain. From (2.110) we know that if f is any feasible control law $\underline{\delta}$ satisfies the relation

$$(I - P^f)\,\underline{\delta} = \underline{\gamma}^f - \bar{c}\,\underline{\tau}^f \qquad \ldots.(3.23)$$

and, correspondingly, the value-vector $\underline{\delta}^*$ of the equivalent CSMC, under the same policy f, must satisfy

$$(I - (P^*)^f)\,\underline{\delta}^* = (\underline{\gamma}^*)^f - (\bar{c}^*)^f\,(\underline{\tau}^*)^f \qquad \ldots.(3.24)$$

Now using (3.11), (3.14) and (3.15) in (3.24), we obtain

$$\Phi^f (I - P^f) \underline{\delta}^* = \Phi^f \underline{\gamma}^f - \Phi^f \underline{\tau}^f (\bar{c}^*)^f \qquad \ldots.(3.25)$$

and, since Φ^f is non-singular, we have, on using (2.115),

$$(I - P^f) \underline{\delta}^* = \underline{\gamma}^f - \bar{c}\, \underline{\tau}^f \qquad \ldots.(3.26)$$

Thus $\underline{\delta}^*$ satisfies the same equation as $\underline{\delta}$ and we may take $\underline{\delta}^* = \underline{\delta}$.

Now consider the discrete-time Markov regulation problem of minimizing the equilibrium mean cost rate $\bar{c}^*$ of the equal-sojourn-time chain $\{X_t^*\}$ resulting from the equivalence transformation (3.11). On identifying $\underline{\alpha}$ with $\underline{\gamma}^*$ and $\underline{\omega}$ with $\underline{\delta}^*$ in equation (3.20), and using the fact that the equilibrium mean cost per transition, $\bar{\alpha}$, is given, via (3.16), by $\bar{\alpha} = \tau_o \bar{c}^*$, we find that the optimal cost rate satisfies the equations

$$\delta_i^* + \tau_o \bar{c}^* = \min_{u \in \mathcal{U}} \left[(\gamma_i^*)^u + \sum_j (p_{ij}^*)^u \delta_j^* \right], \quad \forall i \in N_N \qquad \ldots.(3.27)$$

or, what is equivalent,

$$\operatorname*{Min}_{u \in \mathcal{U}} \left[(\gamma_i^*)^u + \sum_j (p_{ij}^*)^u \delta_j^* - \delta_i^* - \tau_o \bar{c}^* \right] = 0, \quad \forall i \in N_N \qquad \ldots.(3.27)$$

Now use (3.11) - (3.15) to express the conditions (3.27) in terms of the properties of the original semi-Markov chain $\{X_t\}$. The result is

$$\operatorname*{Min}_{u \in \mathcal{U}} \left[\tau_o \left\{ \frac{\gamma_i^u}{\tau_i^u} + \frac{1}{\tau_i^u} \sum_j p_{ij}^u \delta_j - \frac{\delta_i}{\tau_i^u} - \bar{c} \right\} \right] = 0, \quad \forall i \in N_N \qquad \ldots.(3.28)$$

or, since $\tau_i^u > 0$, $\forall i$, $\forall u$, and $\tau_o > 0$,

$$\operatorname*{Min}_{u \in \mathcal{U}} \left[\gamma_i^u + \sum_j p_{ij}^u \delta_j - \delta_i - \bar{c}\, \tau_i^u \right] = 0, \quad \forall i \in N_N \qquad \ldots.(3.29)$$

$$\Phi^f (I - P^f)\, \underline{\delta}^* = \Phi^f \underline{\gamma}^f - \Phi^f \underline{\tau}^f (\bar{c}^*)^f \qquad \ldots.(3.25)$$

and, since Φ^f is non-singular, we have, on using (2.115),

$$(I - P^f)\, \underline{\delta}^* = \underline{\gamma}^f - \bar{c}\, \underline{\tau}^f \qquad \ldots.(3.26)$$

Thus $\underline{\delta}^*$ satisfies the same equation as $\underline{\delta}$ and we may take $\underline{\delta}^* = \underline{\delta}$.

Now consider the discrete-time Markov regulation problem of minimizing the equilibrium mean cost rate $\bar{c}^*$ of the equal-sojourn-time chain $\{X^*_t\}$ resulting from the equivalence transformation (3.11). On identifying $\underline{\alpha}$ with $\underline{\gamma}^*$ and $\underline{\omega}$ with $\underline{\delta}^*$ in equation (3.20), and using the fact that the equilibrium mean cost per transition, $\bar{\alpha}$, is given, via (3.16), by $\bar{\alpha} = \tau_o \bar{c}^*$, we find that the optimal cost rate satisfies the equations

$$\delta^*_i + \tau_o \bar{c}^* = \min_{u \in \mathcal{U}} \left[(\gamma^*_i)^u + \sum_j (p^*_{ij})^u \delta^*_j \right], \quad \forall i \in \mathbb{N}_N \qquad \ldots.(3.27)$$

or, what is equivalent,

$$\underset{u \in \mathcal{U}}{\text{Min}} \left[(\gamma^*_i)^u + \sum_j (p^*_{ij})^u \delta^*_j - \delta^*_i - \tau_o \bar{c}^* \right] = 0, \quad \forall i \in \mathbb{N}_N \qquad \ldots.(3.27)$$

Now use (3.11) - (3.15) to express the conditions (3.27) in terms of the properties of the original semi-Markov chain $\{X_t\}$. The result is

$$\underset{u \in \mathcal{U}}{\text{Min}} \left[\tau_o \left\{ \frac{\gamma^u_i}{\tau^u_i} + \frac{1}{\tau^u_i} \sum_j p^u_{ij} \delta_j - \frac{\delta_i}{\tau^u_i} - \bar{c} \right\} \right] = 0, \quad \forall i \in \mathbb{N}_N \qquad \ldots.(3.28)$$

or, since $\tau^u_i > 0$, $\forall i$, $\forall u$, and $\tau_o > 0$,

$$\underset{u \in \mathcal{U}}{\text{Min}} \left[\gamma^u_i + \sum_j p^u_{ij} \delta_j - \delta_i - \bar{c}\, \tau^u_i \right] = 0, \quad \forall i \in \mathbb{N}_N \qquad \ldots.(3.29)$$

From (3.28) we deduce that

$$\bar{c} = \underset{u \in \mathcal{U}}{\text{Min}} \left[\frac{1}{\tau_i^u} \left\{ \gamma_i^u + \sum_j p_{ij}^u \delta_j - \delta_i \right\} \right], \; \forall i \in N_N \qquad \text{....(3.30)}$$

and from (3.29) we have

$$\delta_i = \underset{u \in \mathcal{U}}{\text{Min}} \left[\gamma_i^u + \sum_j p_{ij}^u \delta_j - \bar{c}\, \tau_i^u \right], \; \forall i \in N_N \qquad \text{....(3.31)}$$

Thus by the same argument as for the pure Markov case we can define an optimal control law f^o either by

$$\forall i \in N_N: \quad f^o(i) = \text{Arg.} \underset{u \in \mathcal{U}}{\text{min}} \left[\frac{1}{\tau_i^u} \left\{ \gamma_i^u + \sum_j p_{ij}^u \delta_j - \delta_i \right\} \right] \qquad \text{....(3.32)}$$

or by

$$\forall i \in N_N: \quad f^o(i) = \text{Arg.} \underset{u \in \mathcal{U}}{\text{min}} \left[\gamma_i^u + \sum_j p_{ij}^u \delta_j - \bar{c}\, \tau_i^u \right] \qquad \text{....(3.33)}$$

where, in each case, $(\bar{c}, \underline{\delta})$ satisfies the value-determination equation

$$\underline{\delta} = \underline{\gamma}^{f^o} + P^{f^o} \underline{\delta} - \bar{c}\, \underline{\tau}^{f^o} \qquad \text{....(3.34)}$$

The Jewell policy-iteration algorithm is based on iteration between (3.34) and (3.33), and the Schweitzer algorithm uses iteration between (3.34) and (3.32). As our derivation of (3.30) and (3.31) has shown, both algorithms are equivalent to applying the Howard algorithm to pure Markov regulation problem generated from the original semi-Markov regulation problem by the equivalence transformation (3.11). This fact may be used to prove convergence of either algorithm, though more direct proofs are available.

One feature of the policy-iteration algorithms defined above is that it is possible to compute, after each iteration, upper and lower bounds on the true optimal cost rate, $\bar{c}^o$, and hence to monitor the progress of the algorithm. These bounds were first derived by Hastings(1971).

3.3.2 Successive-approximation methods

A disadvantage of the policy-iteration methods of Section 3.3.1 is that each iteration cycle involves the solution of the linear N-vector equation (3.23) (or its Markov equivalent, (3.22)), which can be computationally expensive if N is large. In the transient-cost problem (see Section 3.2.4), the successive-approximation method of Eaton and Zadeh(1962), or the Kushner/Kleinman variants(1968,1971) of it, are often less expensive in total computer resources than the corresponding policy-iteration method. It is natural therefore to enquire whether the Eaton/Zadeh method can be adapted to the optimal regulation problem. Such an adaptation has been developed by White(1963) for the discrete-time Markov regulation problem.

White's successive-approximations algorithm is based on the following result. Suppose that there is a state, call it state 1, that is recurrent under every feasible control policy. Then define, for each $i \in \mathcal{N}_N$, the sequences $V_i(n)$ and $v_i(n)$ by the recurrence relations

$$\left\{\begin{array}{lcl} V_i(n) & = & \underset{u \in \mathcal{U}}{\text{Min}} \left[\alpha_i^u + \sum\limits_{j \in \mathcal{N}_N} p_{ij}^u \; v_j(n-1) \right] \qquad \ldots.(3.35) \\ \\ v_i(n) & = & V_i(n) - V_1(n) \qquad \ldots.(3.36) \end{array}\right.$$

with $v_i(0)$, $i = 1,2,\ldots,N$, arbitrary but specified.

Then, with $\bar{\alpha}$ and $\underline{\omega}$ satisfying condition (3.22), and with $\omega_1 = 0$,

$$\lim_{n \to \infty} V_1(n) = \bar{\alpha} \qquad \ldots.(3.37)$$

and

$$\lim_{n \to \infty} v_i(n) = \omega_i \, , \quad \forall i \in \mathcal{N}_N \qquad \ldots.(3.38)$$

That is to say, the algorithm defined by the iteration of (3.35) and (3.36) converges to the (unique) solution of the Howard equations (3.20). It is also clear that the sequence of control laws, obtained by the successive minimizations of the right-hand side of (3.35), converges to the optimal control law given by (3.21).

The advantage of White's algorithm is of course that it is computationally straightforward; the main disadvantage is that, in common with many iterative methods for solution of simultaneous equations, convergence may be slow. In fact the rate of convergence will depend on the detailed forms of the various closed-loop transition probability matrices P^f that are generated by the algorithm.

As with the policy-iteration algorithms of Section 3.3.1, it is possible to monitor the rate of convergence of White's algorithm. Define, for each $n \in \mathbb{Z}_+$,

$$\alpha_U(n) \triangleq \underset{i \in N_N}{\text{Max}} \left[v_i(n) - \omega_i(n-1) \right] \qquad \ldots.(3.39)$$

$$\alpha_L(n) \triangleq \underset{i \in N_N}{\text{Min}} \left[v_i(n) - \omega_i(n-1) \right] \qquad \ldots.(3.40)$$

Then $\alpha_U(n) \downarrow \bar{\alpha}$ and $\alpha_L(n) \uparrow \bar{\alpha}$. Thus if we take $\frac{1}{2}\left[\alpha_U(n) + \alpha_L(n)\right]$ as an estimate of $\bar{\alpha}$, the magnitude of the error is bounded by $\frac{1}{2}\left[\alpha_U(n) - \alpha_L(n)\right]$, and the algorithm can be terminated when this has fallen to a specified level. These bounds on $\bar{\alpha}$ are due to Odoni$^{(1969)}$.

The successive-approximations algorithm described above has no obvious counterpart for the continuous-time regulation problem, nor is it easily adapted for application to the semi-Markov regulation problem. However, as we shall show in Chapter 4, it is possible, by invoking the concept of total equivalence (see Section 3.2.3), to develop successive-approximation methods for the general semi-Markov regulation problem.

3.3.3 Linear-programming methods

At about the same time as Howard's development of the basic policy-iteration method for the discrete-time regulation problem, Manne$^{(1960)}$ showed that the problem can be formulated as a linear-programming (LP) problem. This idea is interesting since there exist highly-developed and efficient computer programs for solving LP problems. The formulation of the discrete-time regulation problem as an LP problem is as follows.

Suppose that the control set for the totally regular controllable chain $\{(X_t : \Omega \to \mathbb{N}_N) : t \in \mathbb{Z}_+\}$ is $\mathcal{U} = \{u^k : k \in \mathbb{N}_K\}$, and define for each $i \in \mathbb{N}_N$, each $k \in \mathbb{N}_K$,

$$d_{ik} \triangleq P\left[u_t = u^k \;\middle|\; X_t = i\right] \qquad \ldots.(3.41)$$

Then a stationary randomized control policy (d) is a set $\{d_{ik} : i \in \mathbb{N}_N,\ k \in \mathbb{N}_K\}$ where all the d_{ik} are non-negative and also

$$\sum_{k \in \mathbb{N}_K} d_{ik} = 1 , \quad \forall i \in \mathbb{N}_N \qquad \ldots.(3.42)$$

Under the stationary policy (d) the one-step expected cost α_i^d is given, for each $i \in \mathbb{N}_N$, by

$$\begin{aligned} \alpha_i^d &\triangleq E\left[c(X_t, X_{t+1}, u_t) \;\middle|\; X_t = i\right] \\ &= E\left[E\left[c(X_t, X_{t+1}, u_t) \;\middle|\; u_t ; X_t = i\right] \;\middle|\; X_t = i\right] \\ &= \sum_{k \in \mathbb{N}_K} d_{ik}\, \alpha_i^{u^k} \end{aligned} \qquad \ldots.(3.43)$$

It then follows that $\overline{\alpha}^d$, the cost rate under policy (d) is given by

$$\begin{aligned} \overline{\alpha}^d &\triangleq \sum_{i \in \mathbb{N}_N} \pi_i^d\, \alpha_i^d \\ &= \sum_i \sum_k \alpha_i^{u^k}\, \pi_i^d\, d_{ik} \end{aligned} \qquad \ldots.(3.44)$$

If we now introduce the variables

$$\boxed{x_{ik} \triangleq \pi_i^d \; d_{ik}} \;, \quad \forall i \in N_N, \quad \forall k \in N_K, \qquad \text{....(3.45)}$$

and, for notational convenience, denote $\alpha_i^{u^k}$ by α_{ik}, we can re-write (3.44) as

$$\boxed{\bar{\alpha}^d = \sum_i \sum_k \alpha_{ik} \, x_{ik}} \qquad \text{....(3.46)}$$

Furthermore, from their definition the x_{ik} must satisfy the relations

$$\left\{ \begin{array}{ll} x_{ik} \geqslant 0 \;, \quad \forall i \in N_N, \; \forall k \in N_K & \text{....(3.47)} \\ \sum_i \sum_k x_{ik} = 1 & \text{....(3.48)} \\ \sum_k x_{jk} - \sum_i \sum_k x_{ik} \, p_{ij}^{u^k} = 0 \;, \quad \forall j \in N_N & \text{....(3.49)} \end{array} \right.$$

The optimal regulation problem is now to minimize the linear function (3.46) of the NK variables x_{ik} subject to the linear constraints (3.47) - (3.49). This is a standard LP problem in canonical form which may be solved by, for example, the simplex method. Once the solution has been found, the stationary distribution under the optimal policy (d^o) is given by

$$\pi_i^{d^o} = \sum_k x_{ik} \;, \quad \forall i \in N_N \qquad \text{....(3.50)}$$

and then optimal control law d^o is given by

$$d_{ik}^o = \frac{x_{ik}}{\pi_i^{d^o}} \;, \quad \forall i \in N_N, \; \forall k \in N_K \qquad \text{....(3.51)}$$

This LP formulation of the optimal regulation problem is of theoretical interest since it may be shown (see Wolfe/Dantzig$^{(1962)}$) that for each i only one of the d^o_{ik} is non-zero (and hence equal to 1). This means that the optimal stationary policy (d^o) is deterministic, as asserted at the end of section 3.2.2.

By again introducing the control probabilities d_{ik} it is a straightforward matter to formulate the continuous-time regulation problem as an LP problem. However, in the semi-Markov case the transformation, which is due to Osaki and Mine$^{(1968)}$, is a little more involved. Suppose now that $\{(X_t : \Omega \to N_N) : t \in \mathcal{T}\}$ is a totally regular CSMC with control set $\mathcal{U} = \{u^k : k \in N_K\}$, and again introduce the probabilities d_{ik} defined by (3.41). Then, under the stationary policy (d), the expected one-step cost, γ^d_i, and expected one-step sojourn time, τ^d_i, are given respectively by

$$\gamma^d_i = \sum_k d_{ik}\, \gamma^{u^k}_i \qquad \text{....(3.52)}$$

and

$$\tau^d_i = \sum_k d_{ik}\, \tau^{u^k}_i \qquad \text{....(3.53)}$$

for each $i \in N_N$. (As before, γ^u_i and τ^u_i are the expected one-step cost and expected sojourn time from state i under control action u.)

Furthermore, the equilibrium cost-rate under (d) is, from (3.6), given by

$$\boxed{\bar{c}^d = \frac{\sum_i \sum_k \gamma_{ik}\, x_{ik}}{\sum_i \sum_k \tau_{ik}\, x_{ik}}} \qquad \text{....(3.54)}$$

where the x_{ik} are defined by (3.45) and, again for notational convenience, $\gamma_{ik} \triangleq \gamma^{u^k}_i$, $\tau_{ik} \triangleq \tau^{u^k}_i$.

The semi-Markov regulation problem is now to minimize the objective function (3.54) subject to the linear constraints (3.47) - (3.49). This is an example of a so-called fractional programming problem (see Charnes/Cooper$^{(1961)}$): it may be transformed to an equivalent LP problem as follows.

Let

$$\mu_{ik} \triangleq \frac{\gamma_{ik}}{\tau_{ik}} \quad , \quad \forall i \in N_N, \ \forall k \in N_K$$

and introduce the variables

$$y_{ik} \triangleq \frac{\tau_{ik}\, x_{ik}}{\sum\limits_{j} \sum\limits_{k} \tau_{jk}\, x_{jk}} \quad , \quad \forall i \in N_N, \ \forall k \in N_K \qquad \ldots.(3.55)$$

and

$$y \triangleq \frac{1}{\sum\limits_{j} \sum\limits_{k} \tau_{jk}\, x_{jk}} \qquad \ldots.(3.56)$$

Then, from (3.54) we have

$$\boxed{\bar{c}^{d} = \sum_{i} \sum_{k} \mu_{ik}\, y_{ik}} \qquad \ldots.(3.57)$$

and from (3.36) - (3.38) the y_{ik} must satisfy

$$\left\{ \begin{array}{ll} y_{ik} \geqslant 0 \ , \ \forall i \in N_N, \ \forall k \in N_K & \ldots.(3.58) \\ \sum\limits_{i} \sum\limits_{k} \left(\frac{y_{ik}}{\tau_{ik}}\right) = y & \ldots.(3.59) \\ \sum\limits_{k} \left(\frac{y_{jk}}{\tau_{jk}}\right) - \sum\limits_{i} \sum\limits_{k} \left(\frac{y_{ik}}{\tau_{ik}}\right) p_{ij}^{k} = 0 \ , \ \forall j \in N_N & \ldots.(3.60) \end{array} \right.$$

Furthermore, it is also clearly necessary that

$$\sum_i \sum_k y_{ik} = 1 \qquad \ldots.(3.61)$$

We now have an LP problem in canonical form, namely: minimize the linear function (3.57) of the NK variables y_{ik}, subject to the linear constraints (3.58) - (3.61). Furthermore one can show (see Osaki/Mine(1968)) (a) that the constraint (3.59) is redundant and hence so is the variable y ; and (b) that as in the discrete-time Markov case, there exists an optimal control law, d^o, for the above problem which is deterministic, ie. such that $d^o_{ik} = 0$ or 1 for each $i \in \mathbb{N}_N$ and each $k \in \mathbb{N}_K$.

Finally, there is an important point to be noted about the relation between the LP formulations considered in this section and the policy-iteration algorithms of Section 3.3.1. The LP problem defined by (3.57) - (3.61) has the form:-

$\underline{\underline{P}}$: Minimize $\underline{\mu}^T \underline{y}$

subject to

$$\begin{cases} A\,\underline{y} = \underline{b} \\ \underline{y} \geqslant \underline{0} \end{cases}$$

where $\underline{y}$ is a variable NK-vector, $\underline{\mu}$ and $\underline{b}$ are fixed NK-vectors, and A is a fixed (NK x (N + 1)) matrix.

As is well known (see, for example, Trustrum(1971)), problem P has associated with it a <u>dual</u> problem having the form:-

$\underline{\underline{D}}$: Maximize $\underline{b}^T \underline{z}$

subject to

$$\begin{cases} A^T \underline{z} \leqslant \underline{\mu} \\ \underline{z} \text{ unconstrained in sign} \end{cases}$$

where $\underline{z}$ is a variable (N + 1) - vector.

Furthermore, since P has a solution then so has D, and $(\underline{b}^T \underline{z}^o) = (\underline{\mu}^T \underline{y}^o)$, where $\underline{y}^o$ and $\underline{z}^o$ are the solutions of P and D respectively. So if we identify z_i with δ_i, i = 1,...,N, and z_{N+1} with $\bar{c}$, and use the appropriate A and $\underline{b}$ we obtain the following statement of the dual problem, D :-

$$\left\| \begin{array}{l} \text{Maximize } \bar{c} \\ \text{subject to} \\ \delta_i \leq \gamma_i^{u^k} - \bar{c}\tau_i^{u^k} + \sum_{j \in N_N} p_{ij}^{u^k} \delta_j \, , \quad \begin{cases} \forall i \in N_N \\ \forall k \in N_K \end{cases} \end{array} \right. \qquad \ldots.(3.62)$$

This LP problem is clearly equivalent to the problem : -

$$\left\| \begin{array}{l} \text{Maximize } \bar{c} \\ \text{subject to} \\ \boxed{\delta_i \leq \underset{u \in \mathcal{U}}{\text{Min}} \left[\gamma_i^u - \bar{c}\tau_i^u + \sum_{j \in N_N} p_{ij}^u \delta_j \right]} \, , \forall i \in N_N \end{array} \right. \qquad \ldots.(3.63)$$

One can show that the maximal $\bar{c}$ (ie. the minimal $\bar{c}$ in the primal problem) is achieved at the vertex of the feasible region defined by (3.63), ie. that the optimal $\bar{c}$ satisfies the equality constraints :-

$$\boxed{\delta_i = \underset{u \in \mathcal{U}}{\text{Min}} \left[\gamma_i^u - \bar{c}\tau_i^u + \sum_{j \in N_N} p_{ij}^u \delta_j \right]} \quad , \forall i \in N_N \qquad \ldots.(3.64)$$

Reference to Section 3.3.1 shows that it is precisely this set of equations which the Howard/Jewell algorithm is designed to solve for the optimal $\bar{c}$. We therefore have an alternative view of the Howard/Jewell algorithm, namely, as an algorithm for solving the dual of the LP program defined by (3.57) - (3.61). With such an interpretation, however, it is not obvious that the minimizing arguments of the right-hand side of (3.64) define the optimal control law for the chain.

CHAPTER 4

NEW OPTIMIZATION ALGORITHMS

4.1 Introduction

Although the optimization procedures reviewed in the previous chapter are satisfactory in many applications, the fact remains that when the number of states is large (say $\geqslant 100$) the solution of the optimal regulation problem demands considerable computational effort; this is particularly true when the number of possible control alternatives in each state is also large. The search for efficient optimization algorithms has therefore continued and in this chapter we propose some new algorithms which, at least in certain circumstances, are computationally more efficient than the standard methods. As before, we first consider policy-iteration algorithms and then look at successive-approximation methods.

4.2 New policy-iteration methods

The basic Howard/Jewell policy-iteration algorithm for the semi-Markov regulation problem requires the solution of N simultaneous linear algebraic equations after each policy-improvement cycle and furthermore the policy-improvement procedure itself uses a value-vector which is not updated until the end of the cycle. The first attempt to improve on the basic Howard/Jewell algorithm was proposed by Hastings$^{(1968)}$ who suggested modifying the value-vector $\underline{\delta}$ at each step in the policy-improvement cycle. The difference between the Hastings algorithm and the Jewell algorithm is best demonstrated by the flow-chart shown in Fig.(4).

It can be seen that, if we borrow some appropriately descriptive terminology from the field of linear iterative analysis (see, for example, Varga$^{(1962)}$), the Hastings routine is a "Gauss-Seidel" version of the Jewell routine, which we can think of as the basic

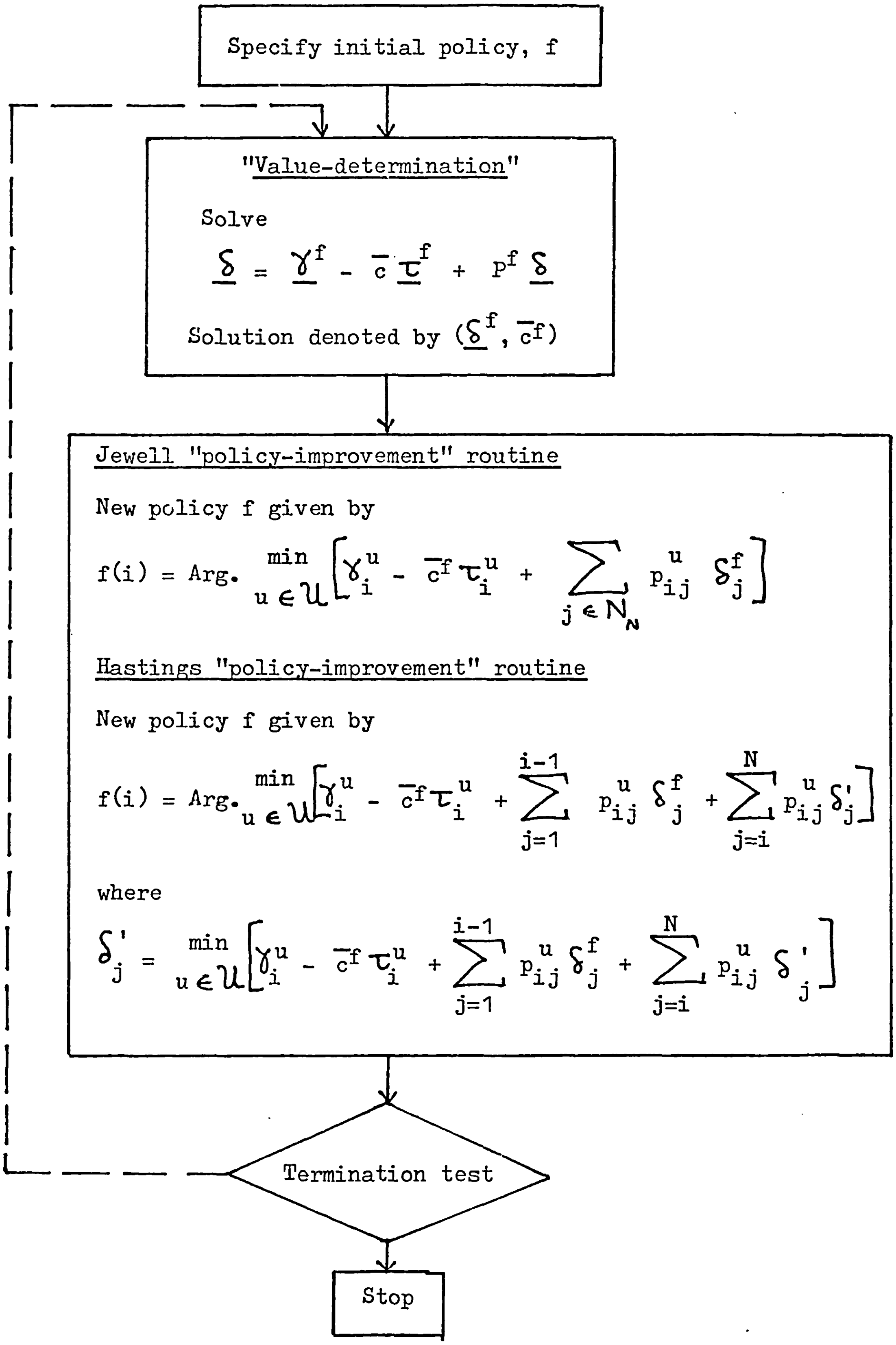

Specify initial policy, f
"Value-determination"
Solve
$\underline{\delta} = \underline{\gamma}^f - \bar{c}\,\underline{\tau}^f + P^f \underline{\delta}$
Solution denoted by $(\underline{\delta}^f, \bar{c}^f)$
Jewell "policy-improvement" routine
New policy f given by
$f(i) = \text{Arg.} \min_{u \in \mathcal{U}} \left[\gamma_i^u - \bar{c}^f \tau_i^u + \sum_{j \in N_N} p_{ij}^u \, \delta_j^f \right]$
Hastings "policy-improvement" routine
New policy f given by
$f(i) = \text{Arg.} \min_{u \in \mathcal{U}} \left[\gamma_i^u - \bar{c}^f \tau_i^u + \sum_{j=1}^{i-1} p_{ij}^u \, \delta_j^f + \sum_{j=i}^{N} p_{ij}^u \, \delta'_j \right]$
where
$\delta'_j = \min_{u \in \mathcal{U}} \left[\gamma_i^u - \bar{c}^f \tau_i^u + \sum_{j=1}^{i-1} p_{ij}^u \, \delta_j^f + \sum_{j=i}^{N} p_{ij}^u \, \delta'_j \right]$
Termination test
Stop

Fig. (4)

"Jacobi" version of the procedure. It is possible to show (see Hastings(1968)) that the Hastings algorithm converges to an optimal policy provided that the controlled semi-Markov chain is totally regular. There is, however, no guarantee that convergence will be more rapid than with the Jewell algorithm, though limited computational experience suggests that with certain types of P-matrix the Hastings version does converge more rapidly.

As a further modification to the basic Jewell policy-improvement routine, Schweitzer(1971a) has proposed a procedure in which the Hastings policy-improvement routine is used iteratively before each return to the value-determination stage of the optimization cycle. It is not claimed that this procedure is computationally superior to the Hastings algorithm.

4.2.1 A revised policy-iteration algorithm

We now outline a rather different modification of the Howard/Jewell algorithm, in which the policy-improvement is carried out one state at a time, the value-vector $\underline{\delta}$ being re-computed after each single-state policy improvement. That is, the policy-improvement and value-determination operations are interleaved, with the result that a properly updated value-vector is always used in the policy-improvement stage.

Consider the value-determination equation in the basic flow chart (Fig.(4)):

$$\underline{\delta} = \underline{\gamma} - \bar{c}\,\underline{\tau} + P\,\underline{\delta} \qquad \text{....(4.1)}$$

or,

$$(I - P)\,\underline{\delta} + \bar{c}\,\underline{\tau} = \underline{\gamma} \qquad \text{....(4.2)}$$

If P is regular the corresponding $\underline{\pi}$-vector is unique and so pre-multiplication of (4.2) by $\underline{\pi}^T$ yields the unique solution $\bar{c} = \dfrac{\underline{\pi}^T\underline{\gamma}}{\underline{\pi}^T\underline{\tau}}$ for the cost rate, as required. On the other hand since

$I - P$ is of rank $(N - 1)$ (for P regular) the vector $\underline{\delta}$ is not uniquely determined by (4.2). However, if we set $\delta_1 = 0$ we can write (4.2) in the form

$$R\,\underline{v} = \underline{\gamma} \qquad \text{....(4.3)}$$

where

$$\underline{v} \triangleq \text{Col}\,(\bar{c},\ \delta_2, \delta_3, \ldots, \delta_N)$$

and

$$R \triangleq \left[(I - P)(I - \underline{e}_1\,\underline{e}_1^T) + \underline{\tau}\,\underline{e}_1^T \right]$$

(R is the matrix $(I - P)$ with its first column replaced by $\underline{\tau}$)

It is easy to see that R is non-singular if P is regular, so that the unique solution to (4.3) is

$$\underline{v} = R^{-1}\,\underline{\gamma} \qquad \text{....(4.4)}$$

Incidentally by equating the first rows of the identity $R^{-1}R = I$ we find that

$$\underline{e}_1^T\ R^{-1} = \left(\frac{1}{\bar{\tau}}\right)\underline{\pi}^T \qquad \text{....(4.5)}$$

where, as usual,

$$\bar{\tau} \triangleq \underline{\pi}^T\,\underline{\tau}$$

We make use of this property later.

Now consider two control laws, f and f', which differ only in state i, i.e.

$$\begin{cases} f'(j) = f(j) & , \quad j \neq i \\ \neq f(j) & , \quad j = i \end{cases}$$

The corresponding R-matrices will differ only in their i^{th} rows and so we can write

$$R^{f'} = R^{f} + \underline{e}_i\,\underline{a}_i^T \qquad \text{....(4.6)}$$

where $\underline{a}_i^T$ is the difference between the i^{th} rows of $R^{f'}$ and R^f.

Suppose that $(R^f)^{-1}$ is known: then, by (4.4), $\underline{v}^f = (R^f)^{-1}\underline{\gamma}^f$. We now make use of the Sherman-Morrison matrix inversion lemma$^{(1949)}$ to relate $(R^{f'})^{-1}$ to $(R^f)^{-1}$. The lemma states that if A is a non-singular n x n matrix, and the n x n matrix $A' \triangleq A + \underline{b}\,\underline{c}^T$ is also invertible, then

$$\boxed{(A')^{-1} = A^{-1}\left[I - \lambda\,\underline{b}\,\underline{c}^T A^{-1}\right]} \qquad \text{....(4.7)}$$

where

$$\lambda \triangleq (1 + \underline{c}^T A^{-1}\,\underline{b})^{-1}$$

Applying this result to (4.6) we obtain

$$(R^{f'})^{-1} = (R^f)^{-1}\left[I - \lambda_i\,\underline{e}_i\,\underline{a}_i^T\,(R^f)^{-1}\right] \qquad \text{....(4.8)}$$

with

$$\lambda_i = \left[1 + \underline{a}_i^T\,(R^f)^{-1}\,\underline{e}_i\right]^{-1} \qquad \text{....(4.9)}$$

or, alternatively, on re-arranging (4.8),

$$(R^{f'})^{-1} = (R^f)^{-1} - \lambda_i(\underline{r}_i\,\underline{s}_i^T) \qquad \text{....(4.10)}$$

where
$$\begin{cases} \underline{r}_i \triangleq (R^f)^{-1}\,\underline{e}_i \\ \underline{s}^T \triangleq \underline{a}_i^T(R^f)^{-1} \end{cases}$$

and
$$\lambda_i = (1 + \underline{a}_i^T\,\underline{r}_i)^{-1}$$

Thus if the control law is changed only in state i, the inverse of the R-matrix can be updated by simply adding the dyad $\left[-\lambda_i\underline{r}_i\,\underline{s}_i^T\right]$ to the original inverse. We can then use (4.4) to obtain the updated $\underline{v}$-vector.

We can now construct an optimization algorithm which makes use of the above updating procedure, as shown in Fig.(5).

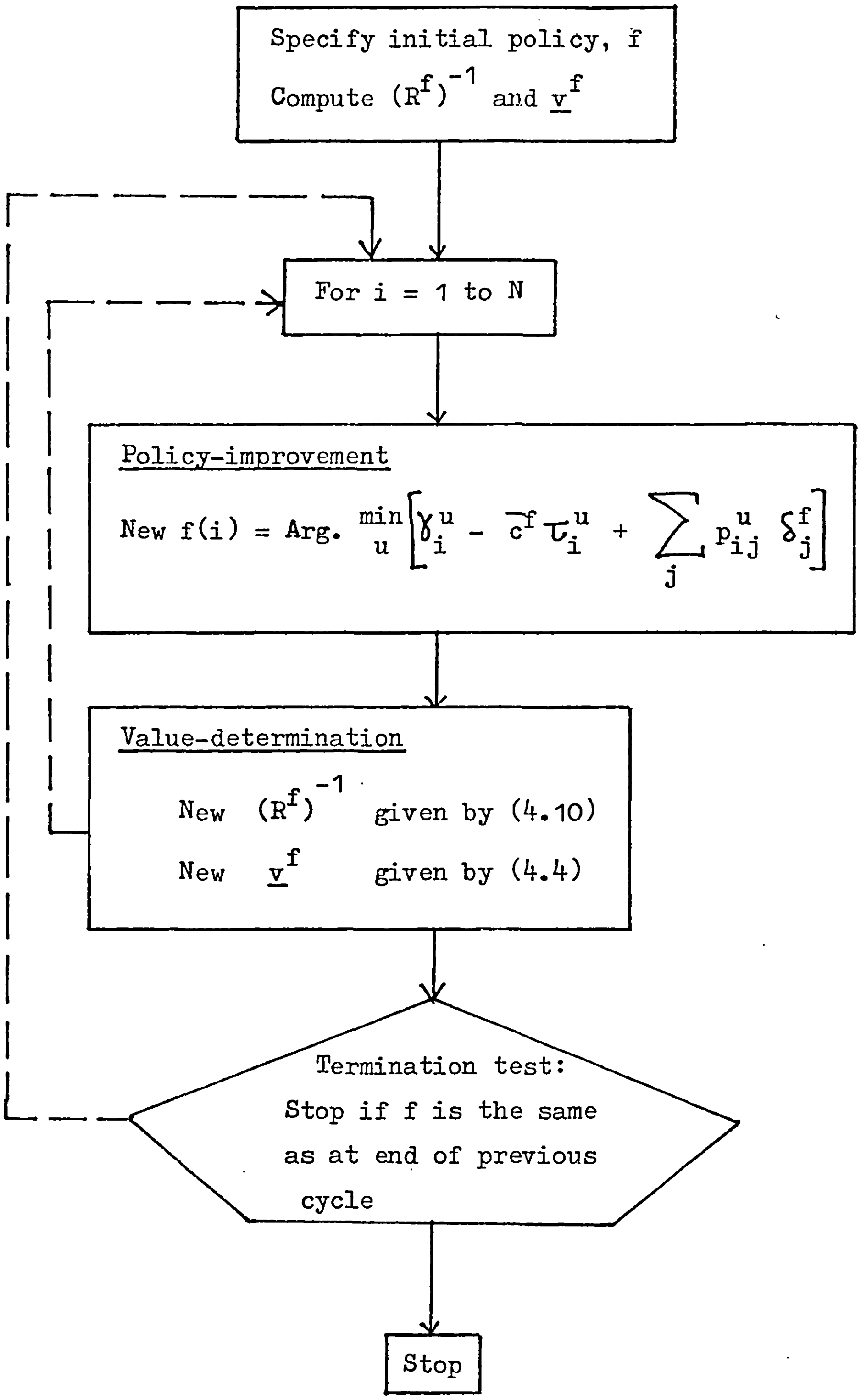

Specify initial policy, f
Compute $(R^f)^{-1}$ and $\underline{v}^f$
For i = 1 to N
Policy-improvement
New f(i) = Arg. $\min_u \left[\gamma_i^u - \bar{c}^f \tau_i^u + \sum_j p_{ij}^u \delta_j^f \right]$
Value-determination
New $(R^f)^{-1}$ given by (4.10)
New $\underline{v}^f$ given by (4.4)
Termination test:
Stop if f is the same as at end of previous cycle
Stop

Fig. (5)

From the above flow-chart it can be seen that the value-determination operation of the Howard/Jewell algorithm (and its variants) has now been split up into N stages (for an N-state problem), with the result that the policy-improvement operation in any state always makes use of the best available values of $\bar{c}$ and $\underline{\delta}$. This is in contrast with the previous algorithms in which the best available values of $\bar{c}$ and $\underline{\delta}$ are used only when i = 1.

Convergence of the above algorithm - which we shall call revised policy-iteration (RPI) - to a globally optimal policy may be proved by the following argument which derives from Howard's original proof of convergence for the basic algorithm.

For any two stationary policies, (f) and (f'), define the test quantities

$$\forall i \in \mathbb{N}_N: \ \xi_i(f',f) \triangleq \gamma_i^{f'(i)} - \bar{c}^f \tau_i^{f'(i)} + \sum_{j \in \mathbb{N}_N} p_{ij}^{f'(i)} \delta_j^f \qquad \text{....(4.11)}$$

and also the following :

$$\forall i \in \mathbb{N}_N: \quad \Delta\xi_i(f',f) \triangleq \xi_i(f',f) - \xi_i(f,f) \qquad \text{....(4.12)}$$

$$\underline{\Delta\xi} \triangleq \text{Col}\,(\Delta\xi_1,\ldots,\Delta\xi_N) \qquad \text{....(4.13)}$$

$$\Delta\bar{c}(f',f) \triangleq \bar{c}^{f'} - \bar{c}^f \qquad \text{....(4.14)}$$

$$\forall i \in \mathbb{N}_N: \ \Delta\delta_i(f',f) \triangleq \delta_i^{f'} - \delta_i^f \qquad \text{....(4.15)}$$

$$\underline{\Delta\delta} \triangleq \text{Col}\,(\Delta\delta_1,\ldots,\Delta\delta_N) \qquad \text{....(4.16)}$$

Then, as is easily verified, $(\underline{\Delta\delta}, \Delta\bar{c})$ satisfies the equation

$$\underline{\Delta\delta} = \underline{\Delta\xi} - \Delta\bar{c}\,\underline{\tau}^{f'} + P^{f'}\,\underline{\Delta\delta} \qquad \text{....(4.17)}$$

from which, on pre-multiplication by $(\underline{\pi}^{f'})^T$, we deduce that

$$\Delta\bar{c} = \frac{(\underline{\pi}^{f'})^T \underline{\Delta\xi}}{(\underline{\pi}^{f'})^T \underline{\tau}^{f'}}$$

$$= \frac{(\underline{\pi}^{f'})^T \underline{\Delta\xi}}{\bar{\tau}^{f'}} \qquad \text{....(4.18)}$$

where, as usual, we have written $\bar{\tau}^{f'}$ for the equilibrium mean sojourn time $(\underline{\pi}^{f'})^T \underline{\tau}^{f'}$.

Then, introducing the scaled probability distribution

$$\underline{\theta}^{f'} \triangleq \left(\frac{1}{\bar{\tau}^{f'}}\right)\underline{\pi}^{f'}$$

we can compute the reduction in mean cost rate by

$$\boxed{\Delta\bar{c}(f',f) = (\underline{\theta}^{f'})^T \underline{\Delta\xi}(f',f)} \qquad \text{....(4.19)}$$

The coefficients of the $\Delta\xi_i$ in this linear functional are all non-negative. It follows that

$$\Delta\bar{c} < 0 \implies \exists i \in N_N : \Delta\xi_i < 0$$

and, since an optimal control exists, we have

$$\text{f non-optimal} \implies \exists f' : \Delta\bar{c} < 0$$

$$\implies \exists f' \; \exists i \in N_N : \Delta\xi_i < 0$$

Now from (4.11) and (4.12) $\Delta\xi_i$ depends on f' only through f'(i). It therefore follows that

$$\text{f non-optimal} \implies \exists i \in N_N \left[\exists f' : f'(j) = f(j), \forall j \neq i \right] : \Delta\xi_i < 0$$

or, equivalently,

$$\forall i \in N_N \left[\forall f' : f'(j) = f(j), \forall j \neq i \right] : \Delta\xi_i \geq 0$$

$$\implies \text{f optimal}$$

The lefthand side of the above implication is precisely the stopping condition for the revised policy-iteration algorithm. We have therefore shown that the algorithm converges to a globally optimal policy by a sequence of single-state policy changes. Furthermore, convergence is clearly monotonic in $\bar{c}$.

It does not seem possible to show that the RPI algorithm always converges more rapidly than the basic Howard algorithm and its variants. The chief advantages offered by the new algorithm are: (i) the computational effort associated with the value-determination operation in each optimization cycle is now proportional to the number of states in which the control law is changed - in a large but highly-structured problem this number may be very much less than the total number of states, with a consequent substantial reduction in computing effort; and (ii) the values of $\underline{\delta}$ and $\bar{c}$ used in the test quantities $\Delta\xi_i$, defined by (4.11) and (4.12) and used for policy-improvement, are continuously updated as the policy-improvement routine steps sequentially through the states of the chain - in contrast to the Howard algorithm, in which the values of $\underline{\delta}$ and $\bar{c}$ used are always those available at the end of the previous optimization cycle. The significance of this second feature is discussed in more detail in Section 4.2.4. (Footnote: If instead of using the "Jewell" test vector $\underline{\Delta\xi}$, defined by (4.11) - (4.13), we use the "Schweitzer" test vector $\underline{\Delta\eta}$ defined by

$$\forall i \in N_N: \eta_i(f',f) \triangleq \frac{1}{\tau_i^{f'(i)}} \left[\gamma_i^{f'(i)} + \sum_i p_{ij}^{f'(i)} \delta_j^f - \delta_i^f \right]$$

$$\forall i \in N_N: \Delta\eta_i(f',f) \triangleq \eta_i(f',f) - \eta_i(f,f)$$

$$\underline{\Delta\eta} \triangleq \mathrm{Col}\,(\Delta\eta_1, \Delta\eta_2, \ldots, \Delta\eta_N)$$

then it is easily shown that

$$\forall i \in N_N: \quad \Delta\xi_i(f',f) = \tau_i^{f'(i)} \, \Delta\eta_i(f',f)$$

and hence that

(a) $\Delta\xi_i(f',f) < 0 \iff \Delta\eta_i(f',f) < 0$

and

(b) $$\boxed{\Delta\bar{c}(f',f) = (\underline{\sigma}^{f'})^T \, \underline{\Delta\eta}(f',f)} \qquad \ldots.(4.19a)$$

which is perhaps a more elegant form of (4.19))

4.2.2 An accelerated policy-iteration algorithm

The policy-improvement routine in the RPI algorithm minimizes, in each state i, the test quantity ξ_i with respect to u = f'(i), and by (4.12) the resulting u also minimizes $\Delta\xi_i$. There is however no guarantee that the resultant improvement in the cost rate $\bar{c}$ is the best obtainable with respect to changes in f(i). For if f' differs from f only in state i, then from (4.11) and (4.12) we have $\Delta\xi_j(f',f) = 0$, $j \neq i$ and so, from (4.19) the difference between $\bar{c}^{f'}$ and $\bar{c}^{f}$ is given by

$$\Delta\bar{c}(f',f) = \theta_i^{f'} \, \Delta\xi_i(f',f) \qquad \ldots.(4.20)$$

so that minimization of $\Delta\xi_i$ does not necessarily correspond to minimization of $\Delta\bar{c}$. Indeed, if state i is transient under the policy f' then $\theta_i^{f'}$ is zero, and so $\Delta\bar{c}$ will be zero even though $\Delta\xi_i(f',f)$ may be non-zero. The problem of transient states is discussed in Section 4.2.5. Even if i is recurrent under each of several improved policies, say f', f", f'", ..., it is clear from (4.20) that minimization of $\Delta\bar{c}$ is not necessarily achieved by minimizing $\Delta\xi_i$.

Suppose we wish to achieve the greatest possible reduction in $\bar{c}$ at each single-state policy improvement. Such an optimal improvement is achieved in the following accelerated policy-iteration (API) algorithm.

From (4.5) we have

$$\left(\frac{1}{\overline{\tau}^{f'}}\right)\pi_i^{f'} = \underline{e}_1^T(R^{f'})^{-1}\,\underline{e}_i \qquad \text{....(4.21)}$$

and from (4.8)

$$(R^{f'})^{-1} = \left[I - \lambda_i(R^f)^{-1}\,\underline{e}_i\,\underline{a}_i^T\right](R^f)^{-1} \qquad \text{....(4.22)}$$

so that

$$\theta_i^{f'} \triangleq \left(\frac{1}{\overline{\tau}^{f'}}\right)\pi_i^{f'}$$

$$= \underline{e}_1^T\left[I - \lambda_i\,\underline{r}_i\,\underline{a}_i^T\right]\underline{r}_i\ . \qquad \text{....(4.23)}$$

where $\quad \underline{r}_i \triangleq (R^f)^{-1}\,\underline{e}_i \qquad \text{....(4.24)}$

Thus in equation (4.20) the scaled probability $\theta_i^{f'}$ can be computed by (4.23).

We can now construct an accelerated policy-iteration algorithm in which the policy-improvement stage minimizes $\Delta\overline{c}$ rather than $\Delta\xi_i$. The flow chart is as in Fig.(6).

Proof of convergence of this API algorithm is as for the RPI algorithm. As before, $\overline{c}$ decreases monotonically to its minimal value, but we now have the additional property that each single-state policy improvement achieves the maximum possible reduction in $\overline{c}$. On the other hand, if the control set $\mathcal{U}$ contains K alternative control actions, there are (K- 1) additional inner products (one for each trial θ_i) to be computed at each policy-improvement. As in the RPI algorithm, $(R^f)^{-1}$ and $\underline{v}^f$ need only be recomputed in those states for which the control law is changed.

Incidentally, in the special case where the control set $\mathcal{U}$ contains only two elements, so that the optimization proceeds by a sequence of binary choices, minimization of $\Delta\xi_i(f',f)$ is equivalent to

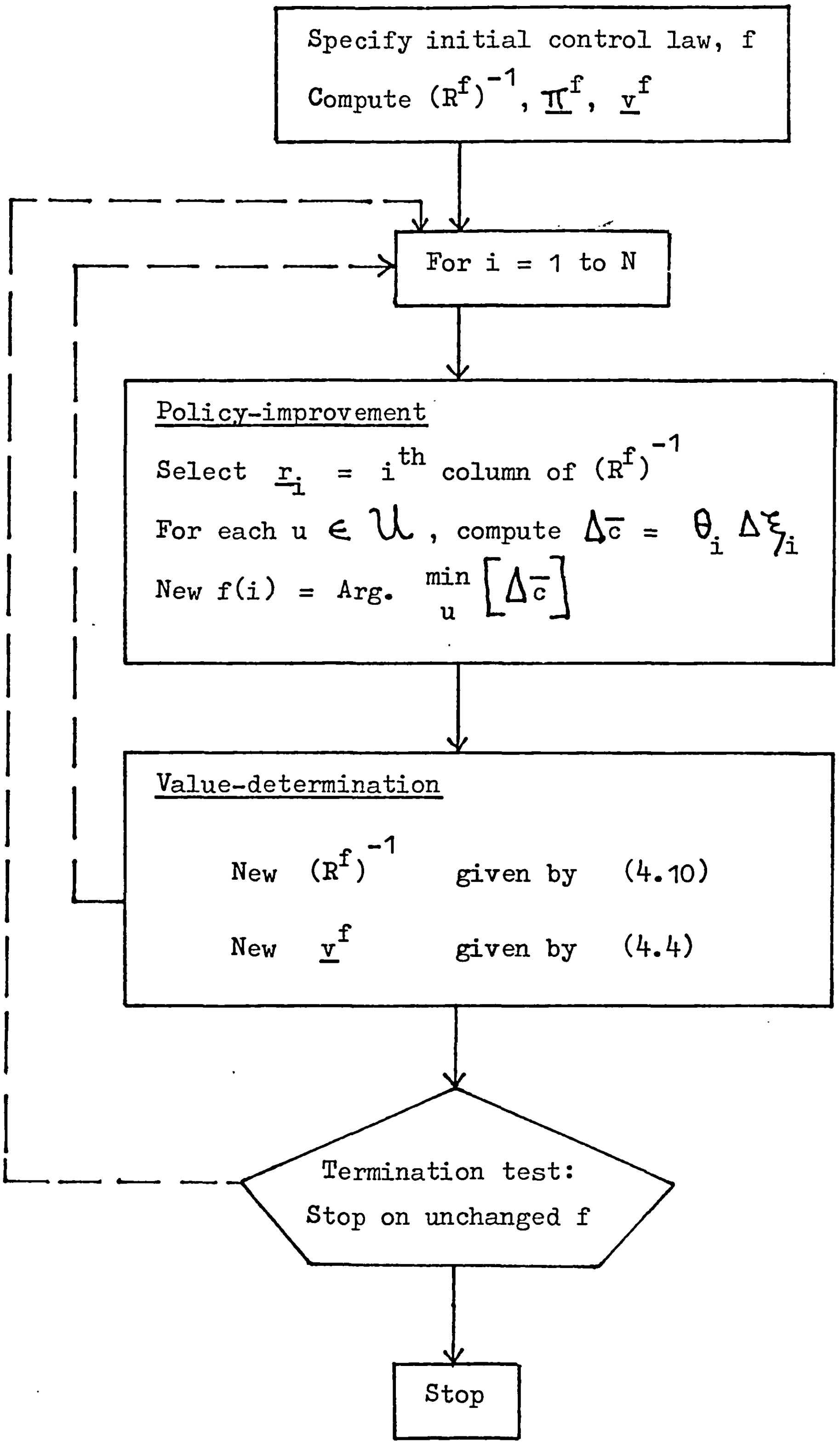

Specify initial control law, f
Compute $(R^f)^{-1}$, $\underline{\pi}^f$, $\underline{v}^f$
For i = 1 to N
Policy-improvement
Select $\underline{r}_i$ = i^{th} column of $(R^f)^{-1}$
For each $u \in \mathcal{U}$, compute $\Delta\bar{c} = \theta_i \Delta\xi_i$
New f(i) = Arg. $\min_u [\Delta\bar{c}]$
Value-determination
New $(R^f)^{-1}$ given by (4.10)
New $\underline{v}^f$ given by (4.4)
Termination test:
Stop on unchanged f
Stop

Fig. (6)

minimization of $\Delta\bar{c}(f',f)$. It is therefore unnecessary to compute the θ_i in this type of problem.

4.2.3 A direct policy-iteration algorithm

Finally, in this group of algorithms we outline an alternative to the API algorithm in which $\bar{c}$ is computed directly from (3.6). Changes in $\bar{c}$ due to single-state policy changes are computed by evaluating directly the effect of a row change in the P-matrix on the corresponding stationary distribution, $\underline{\pi}$.

Suppose that P^f and $P^{f'}$ are the transition probability matrices of a totally regular CSMC under the stationary policies (f) and (f'), and let

$$\Delta P \triangleq P^{f'} - P^f \qquad \ldots.(4.25)$$

Since the chain is totally regular, the corresponding stationary distributions, $\underline{\pi}^f$ and $\underline{\pi}^{f'}$, exist and satisfy the relations

$$(\underline{\pi}^f)^T (I - P^f) = \underline{0}^T \qquad \ldots.(4.26)$$

$$(\underline{\pi}^{f'})^T (I - P^{f'}) = \underline{0}^T \qquad \ldots.(4.27)$$

Subtracting (4.26) from (4.27), we obtain

$$(\underline{\pi}^{f'})^T (I - P^{f'}) = (\underline{\pi}^f)^T (I - P^f) \qquad \ldots.(4.28)$$

Now the matrix $(I - P^{f'})$ is of rank (N - 1) and hence singular. However, let us separate out the principal dyad of P^f by writing

$$\tilde{P}^f \triangleq P^f - \underline{e}(\underline{\pi}^f)^T \qquad \ldots.(4.29)$$

Then, on using (4.25) and (4.29) in (4.28), we have

$$(\underline{\pi}^{f'})^T (I - \tilde{P}^f - \Delta P) = (\underline{\pi}^f)^T (I - \tilde{P}^f) \qquad \ldots.(4.30)$$

and furthermore the matrix $(I - \tilde{P}^f - \Delta P)$ is non-singular. To see this, note first that

$$M \triangleq (I - \tilde{P}^f - \Delta P) = I - P^{f'} + \underline{e}(\underline{\pi}^f)^T$$

Now,

(i) M singular $\Longrightarrow \exists \underline{v} \neq 0 : M\underline{v} = \underline{0}$

(ii) $\Longrightarrow \exists \underline{v} \neq \underline{0} : (I - P^{f'})\,\underline{v} = -\left[(\underline{\pi}^{f})^{T}\,\underline{v}\right]\underline{e}$

(iii) $\Longrightarrow \exists \underline{v} \neq \underline{0} : (\underline{\pi}^{f'})^{T}(I - P^{f'})\,\underline{v} = -\left[(\underline{\pi}^{f})^{T}\,\underline{v}\right]$

(iv) $\Longrightarrow \exists \underline{v} \neq \underline{0} : (\underline{\pi}^{f})^{T}\,\underline{v} = 0$

Combining conditions (ii) and (iv) we find that

$$\text{M singular} \Longrightarrow \exists \underline{v} \neq \underline{0} : \begin{cases} (\underline{\pi}^{f})^{T}\,\underline{v} = 0 & \text{(iv)} \\ \underline{\text{and}}\quad (I - P^{f'})\,\underline{v} = \underline{0} & \text{(v)} \end{cases}$$

But since $P^{f'}$ is regular the only solutions to (v) are $\underline{v} = \underline{0}$ and $\underline{v} = \underline{e}$; and since $\underline{\pi}^{f} \geqslant \underline{0}$, the solution $\underline{v} = \underline{e}$ cannot satisfy (iv). It follows that $\underline{v} = \underline{0}$ is the only solution to $M\underline{v} = \underline{0}$ and hence that M is non-singular.

Thus (4.30) can be written

$$(\underline{\pi}^{f'})^{T} = (\underline{\pi}^{f})^{T}(I - \tilde{P}^{f})(I - \tilde{P}^{f} - \Delta P)^{-1} \qquad \text{....(4.31)}$$

or, alternatively, since $(I - \tilde{P}^{f})$ is also non-singular

$$\boxed{(\underline{\pi}^{f'})^{T} = (\underline{\pi}^{f})^{T}\left[I - \Delta P\,(I - \tilde{P}^{f})^{-1}\right]^{-1}} \qquad \text{....(4.32)}$$

Now suppose that f' differs from f only in state i ; then $P^{f'}$ will differ from P^{f} only in the i^{th} row, and we can write

$$\Delta P = \underline{e}_i\,\underline{a}_i^{T}$$

If we also introduce

$$T \triangleq (I - \tilde{P}^{f})^{-1} \qquad \text{....(4.33)}$$

and

$$\underline{b}_i^{T} \triangleq \underline{a}_i^{T}\,T \qquad \text{....(4.34)}$$

then (4.32) becomes

$$(\underline{\pi}^{f'})^T = (\underline{\pi}^f)^T \left[I - \underline{e}_i\, \underline{b}_i^T\right]^{-1} \qquad \text{....(4.35)}$$

which, on use of the Sherman-Morrison lemma (equation (4.7)), finally becomes

$$\boxed{(\underline{\pi}^{f'})^T = (\underline{\pi}^f)^T \left[I + \lambda_i \underline{e}_i\, \underline{b}_i^T\right]} \qquad \text{....(4.36)}$$

where

$$\lambda_i = (1 - \underline{b}_i^T \underline{e}_i)^{-1} \qquad \text{....(4.37)}$$

Thus the stationary distribution $\underline{\pi}^f$ can be updated by (4.36) for any single-state change in the control law f. In principle we could use (4.36) in conjunction with a single-state policy-improvement routine to optimize the cost rate $\bar{c}$. However, as we shall now show, there are two features of such an algorithm that are capable of improvement. In the first place, the computation of $\Delta\bar{c}$ for each possible control alternative $f'(i)$, in state i involves the evaluation of the two inner products $(\underline{\pi}^{f'})^T \underline{\gamma}^{f'}$ and $(\underline{\pi}^{f'})^T \underline{\tau}^{f'}$; by working with a suitably scaled stationary distribution it is possible to compute $\bar{c}^{f'}$ and hence $\Delta\bar{c}$ by a single inner product evaluation. Secondly, the updating of the matrix T required after any single-state change in f involves a rank-2 modification of the Sherman-Morrison type; by using a slightly modified matrix it is possible to perform the updating by a rank-1 formula with a consequent reduction in computational effort.

Consider the matrix

$$W \triangleq \left[I - P^f + \underline{\tau}^f \underline{p}^T\right] \qquad \text{....(4.38)}$$

where $\underline{p}$ is an arbitrary fixed probability vector. By the same argument as that following (4.30), W is non-singular if P^f is regular; thus

$$\boxed{E \triangleq W^{-1}} \qquad \text{....(4.39)}$$

exists, and

$$EW = I \qquad \ldots.(4.40)$$

$$WE = I \qquad \ldots.(4.41)$$

Post-multiplying (4.40) by $\underline{e}$ gives

$$\boxed{E\,\underline{\tau}^f = \underline{e}} \qquad \ldots.(4.42)$$

so that, using (4.40) again, $E\left[I - P^f\right] = I - \underline{e}\,\underline{p}^T$.

Pre-multiplying (4.41) by $(\underline{\pi}^f)^T$ gives

$$\boxed{\underline{p}^T E = \left(\frac{1}{\bar{\tau}^f}\right) (\underline{\pi}^f)^T} \qquad \ldots.(4.43)$$

where $\bar{\tau}^f = (\underline{\pi}^f)^T \underline{\tau}^f$

= equilibrium mean sojourn time under (f).

Thus the scaled stationary distribution,

$$\underline{\theta}^f \triangleq \left(\frac{1}{\bar{\tau}^f}\right) \underline{\pi}^f \qquad \ldots.(4.44)$$

can be determined by taking a fixed linear combination of the rows of E. From (3.6), the cost rate under policy (f) is given by single inner product

$$\boxed{\bar{c}^f = (\underline{\theta}^f)^T \underline{\gamma}^f} \qquad \ldots.(4.45)$$

Furthermore if E' is the E-matrix associated with the control law f', then

$$(E')^{-1} = I - P^{f'} + \underline{\tau}^{f'}\,\underline{p}^T$$

$$= E^{-1} - \Delta P + \underline{\Delta\tau}\,\underline{p}^T$$

where

$$\underline{\Delta\tau} \triangleq \underline{\tau}^{f'} - \underline{\tau}^f$$

Thus

$$E' = E\left[I - (\Delta P - \underline{\Delta\tau}\,\underline{p}^T)\,E\right]^{-1} \qquad \ldots.(4.46)$$

If f' differs from f only in state i, then

$$\Delta P - \underline{\Delta\tau}\,\underline{p}^T = \underline{e}_i(\underline{a}_i^T - \Delta\tau_i\underline{p}^T)$$

so introducing

$$\underline{d}_i^T \triangleq (\underline{a}_i^T - \Delta\tau_i\underline{p}^T)\,E \qquad \ldots.(4.47)$$

$$= \underline{a}_i^T E - \Delta\tau_i(\underline{\theta}^f)^T \qquad \ldots.(4.48)$$

and using the Sherman-Morrison lemma, we have

$$\boxed{E' = E\left[I + \lambda_i\,\underline{e}_i\,\underline{d}_i^T\right]} \qquad \ldots.(4.49)$$

where $\lambda_i = (1 - \underline{d}_i^T\,\underline{e}_i)^{-1}$

Finally, writing

$$\Delta E = E' - E \qquad \ldots.(4.50)$$

$$\underline{\Delta\theta} = \underline{\theta}^{f'} - \underline{\theta}^f \qquad \ldots.(4.51)$$

we have, from (4.49), (4.43) and (4.44),

$$\Delta E = \lambda_i\underline{E}_i\underline{d}_i^T \quad , \qquad \ldots.(4.52)$$

where $\underline{E}_i$ is the i^{th} column of E,

and $$\underline{\Delta\theta}^T = \lambda_i\,\theta_i^f\,\underline{d}_i^T \qquad \ldots.(4.53)$$

Also, since $\overline{c}^f = \underline{p}^T E\underline{\gamma}^f$, we can write

$$\Delta\overline{c} = \underline{p}^T\,\underline{\Delta w} \qquad \ldots.(4.54)$$

where $$\underline{w}^f \triangleq E\underline{\gamma}^f$$

and $$\underline{\Delta w} = \underline{w}^{f'} - \underline{w}^f \qquad \ldots.(4.55)$$

We thus have a matrix E from which the scaled stationary distribution $\underline{\theta}$ is easily determined and which can be updated by the rank-1 formula (4.52) when a single-state change is made to the control law f.

Note that subject to the conditions $\underline{p} \geqslant 0$, $\underline{p}^T \underline{e} = 1$, the choice of $\underline{p}$ is arbitrary. We shall choose $\underline{p} = \underline{e}_1$, in which case $(\underline{\theta}^f)^T$ is the first row of E. Note also that the vector $\underline{w}$ is easily updated by the rank-1 formula

$$\underline{w}^{f'} = \left[I + \lambda_i \underline{E}_i (\underline{a}_i^T - \Delta\tau_i \underline{p}^T)\right] \left[\underline{w}^f + \Delta\gamma_i \underline{E}_i\right] \qquad \text{....(4.56)}$$

We are thus led to the direct policy-iteration algorithm shown in Fig.(7).

Proof of convergence is as before. The cost rate $\bar{c}$ decreases monotonically and, as with the API algorithm, the maximum possible reduction in $\Delta\bar{c}$ is achieved at each single-state policy change.

4.2.4 Comparison of policy-iteration algorithms

The development of the above new policy-iteration algorithms was motivated by the search for improved computational efficiency in the solution of the optimal regulation problem. We shall now therefore attempt to compare those features of the various algorithms discussed above which potentially influence their computational efficiency.

(i) Howard/Jewell algorithm

(a) Policy-iteration requires evaluation of the inner product $\sum_j p_{ij}^u \delta_j$ for each control alternative u in each state i. Thus assuming that the number of states is N and the number of possible control actions is K, the total number of multiplications required for policy-improvement is approximately KN^2.

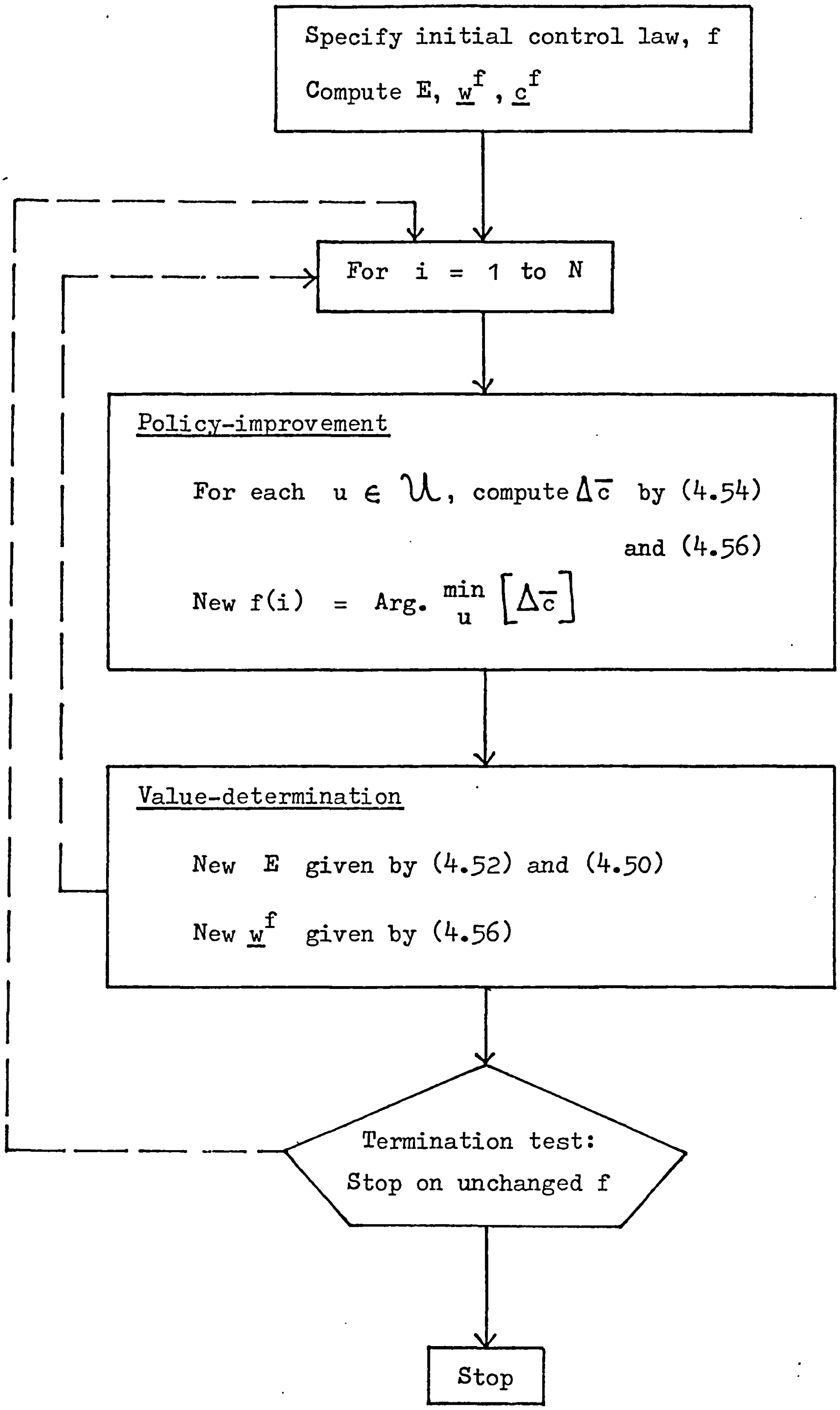

Specify initial control law, f
Compute E, $\underline{w}^f$, $\underline{c}^f$
For i = 1 to N
Policy-improvement
For each u $\in \mathcal{U}$, compute $\Delta\bar{c}$ by (4.54)
and (4.56)
New f(i) = Arg. $\min\limits_{u} \left[\Delta\bar{c}\right]$
Value-determination
New E given by (4.52) and (4.50)
New $\underline{w}^f$ given by (4.56)
Termination test:
Stop on unchanged f
Stop

Fig. (7)

(b) Value-determination requires the solution of N simultaneous linear equations, equivalent to N rank-1 modifications to the matrix $(R^f)^{-1}$. If, say, Gaussian elimination is used for the solution the number of multiplications required is approximately $N^3/3$.

(c) Thus the total number of multiplications required per major iteration cycle is approximately $(K + \frac{1}{3}N)\,N^2$.

(d) The values of $\underline{\delta}$ and $\overline{c}$ used in the policy-improvement stage get progressively more out of date through a cycle as i increases from 1 to N.

(e) The reduction in $\overline{c}$ achieved by each single-state policy-improvement is not necessarily the maximum attainable.

(ii) Hastings algorithm

Essentially the same properties as the Howard/Jewell algorithm.

(iii) Revised policy-iteration (RPI)

(a) Policy-improvement requires the evaluation of the inner product $\sum_j p^u_{ij} \; \delta_j$ for each alternative u in each state i, so that the total number of multiplications required for policy improvement is approximately KN^2.

(b) Value-determination requires N_1 rank-1 modifications to $(R^f)^{-1}$ where $N_1 \leqslant N$ is the number of states in which the control law has changed in the optimization cycle; the number of multiplications required is approximately N_1N^2.

(c) Thus the total number of multiplications required per major iteration cycle is approximately $(K + N_1)\,N^2$.

(d) The values of $\underline{\delta}$ and $\overline{c}$ used in the policy-improvement stage are always the best available.

(e) The reduction in $\bar{c}$ achieved by each policy-improvement is not necessarily optimal.

(iv) Accelerated policy-iteration (API)

(a) Policy-improvement requires the evaluation of two inner products (one for $\Delta\xi_i$, one for θ_i) for each alternative u in each state i. Value-determination is as in the RPI. Thus the total number of multiplications required per major iteration cycle is approximately $(2K + N_1) N^2$.

(b) The reduction in $\bar{c}$ at each single-state policy-improvement is always the maximum attainable.

(v) Direct policy-iteration (DPI)

This has essentially the same properties as the API, the only difference being that the two inner product evaluations per alternative arise in the computation of trial values of $\underline{w}^{f'}$ by (4.56) for use in (4.54).

It can be seen from the above comparison that the new algorithms offer hope of more rapid convergence than the basic algorithm, together with a significantly more efficient major iteration cycle (less computational effort) when the number of control alternatives K is much less than the number of states N and changes in the control law are confined to relatively few states. This will be true for example when the set $\mathcal{X}_R$ of recurrent states is, for all control laws, a small subset of the whole state space $\mathcal{N}_N$. An additional minor advantage of the new algorithms is that since the procedure is identical for every state there is no need to define an iteration cycle as a cycle ending in state N: the procedure has converged as soon as the N most recent policy-improvement stages have left the control law f unchanged. Thus by bringing the termination test within the state-incrementing loop in the flow charts of Figs.(4), (5) and (6) the possibility (inherent in the basic algorithm) of carrying out up to N redundant policy-improvement stages is eliminated.

We have implied, in the above comparison and in the description of the RPI algorithm, that the use of continually updated values for $\underline{\varsigma}$ and $\bar{c}$ in the test quantities $\Delta\xi_i$ offers some advantage over the basic Howard/Jewell/Schweitzer procedure. To see the justification for such an assertion, consider a complete policy-improvement cycle of the Jewell algorithm which results in a change of policy from (f_A) at the beginning of the cycle to (f_B) at the end of the cycle as a result of changes in control action in states $i_1, i_2, \ldots, i_n$. Now define the intermediate policy (f_r) by

$$\begin{cases} f_r(i) = f_A(i) \quad , \quad i = i_r \\ = f_B(i) \quad , \quad i \neq i_r \end{cases}$$

for any state $i_r \in \{ i_1, \ldots, i_n \}$ in which $f_B(i) \neq f_A(i)$

Now using (4.44) in (4.18) we have, in general,

$$\Delta\bar{c}\,(f',f) = (\underline{\theta}^{f'})^T \,\underline{\Delta\xi}\,(f',f) \qquad \ldots.(4.57)$$

and, in particular, for the policies defined above

$$\Delta\bar{c}\,(f_B,f_A) = (\underline{\theta}^{f_B})^T \,\underline{\Delta\xi}\,(f_B,f_A) \qquad \ldots.(4.58)$$

$$\Delta\bar{c}\,(f_r,f_A) = (\underline{\theta}^{f_r})^T \,\underline{\Delta\xi}\,(f_r,f_A) \qquad \ldots.(4.59)$$

Thus

$$\Delta\bar{c}\,(f_B,f_r) = \Delta\bar{c}\,(f_B,f_A) - \Delta\bar{c}\,(f_r,f_A)$$

$$= \sum_{i=1}^{N} \left[\theta_i^{f_B} \,\Delta\xi_i(f_B,f_A) - \theta_i^{f_r}\,\Delta\xi_i(f_r,f_A) \right] \qquad \ldots.(4.60)$$

However, by the definition of f_r ,

$$\Delta\xi_i(f_r,f_A) = \Delta\xi_i(f_B,f_A) \quad , \quad i \neq i_r$$

$$= 0 \quad , \quad i = i_r$$

and so (4.60) gives

$$\Delta\bar{c}(f_B,f_r) = \theta_{i_r}^{f_B} \; \Delta\xi_{i_r}(f_B,f_A) + \sum_i{}' \; (\theta_i^{f_B} - \theta_i^{f_r}) \; \Delta\xi_i(f_B,f_A) \qquad \ldots.(4.61)$$

where the primed summation is over all states except i_r.

Now the scaled distributions $\underline{\theta}^{f_B}$ and $\underline{\theta}^{f_r}$ are necessarily non-negative; and, because f_B if the control law that minimizes $\Delta\xi_i(f,f_A)$ for each state i, the test vector $\underline{\Delta\xi}(f_B,f_A)$ is non-positive. Thus the first term on the right of (4.61) is non-positive. However, the change in control in state i_r from $f_r(i)$ (= $f_A(i)$) to $f_B(i)$ may change the scaled probability distribution $\underline{\theta}$ in such a way that the second term on the right of (4.61) is positive and larger in magnitude than the first term. In such a case we would then have $\Delta\bar{c}\,(f_B,f_r) > 0$. This means that, even though by assumption the test quantity $\Delta\xi_{i_r}(f_B,f_A)$ is strictly negative (for otherwise the policy-improvement routine would leave $f_A(i_r)$ unchanged), f_r is a better control law than f_B. In other words, given the changes from $f_A(i)$ to $f_B(i)$ in all states other than i_r, the change from $f_A(i_r)$ to $f_B(i_r)$ will actually degrade the performance of the system (increase the cost rate $\bar{c}$).

The possible occurrence of such counter-improvements is avoided in the RPI algorithm, in which for any single-state policy change $f(i) \rightarrow f'(i)$ the change in cost rate $\Delta\bar{c}(f',f)$ is, via (4.19), guaranteed to be non-positive if $\Delta\xi_i(f',f)$ is non-positive.

As an example of a Howard/Jewell policy-improvement cycle in which the above effect occurs, consider the following regulation problem :

$$\mathcal{X} = N_2 = \{1,2\}$$
$$\mathcal{U} = \{u^1,u^2\}$$

Parameters	Under u^1	Under u^2
γ_1	1	$-4/3$
γ_2	1	$1/3$
τ_1	1	1
τ_2	1	1
p_{11} p_{12}	0.25 0.75	0.75 0.25
p_{21} p_{22}	0.75 0.25	0.25 0.75

$$\text{Policy } (f_A) = \begin{pmatrix}1\\1\end{pmatrix}$$

Then $\left\{\begin{array}{l} \bar{c}^A = 1 \\ \underline{\delta}^A = \underline{0} \end{array}\right.$

and so Jewell policy-improvement gives : -

$$f_B(1) = \text{Arg.} \min_u (\gamma_1^u - \bar{c}^A \tau_1^u) = u^2$$

$$f_B(2) = \text{Arg.} \min_u (\gamma_2^u - \bar{c}^A \tau_2^u) = u^2$$

$$\text{Thus } (f_B) = \begin{pmatrix}2\\2\end{pmatrix}$$

and so

$$\underline{\pi}^B = \begin{bmatrix}0.5\\0.5\end{bmatrix}$$

whence $$\bar{c}^B = \frac{(\underline{\pi}^B)^T \underline{\gamma}^B}{\bar{\tau}^B} = \boxed{-\frac{1}{2}}$$

However, policy-improvement in state 1 alone would lead to the policy

$$(f') = \begin{pmatrix}u^2\\u^1\end{pmatrix} ,$$

for which

$$\underline{\pi}' = \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix}$$

whence
$$\bar{c}' = \frac{(\underline{\pi}')^T \underline{\gamma}'}{\bar{\tau}'} = \boxed{-\frac{3}{4}}$$

Thus, given the policy change in state 1, the change in state 2 indicated by the Jewell policy-improvement test actually increases the cost rate from $-\frac{3}{4}$ to $-\frac{1}{2}$. With the RPI, on the other hand, the change in f(2), from u^1 to u^2, would be seen to increase $\bar{c}$ and would not be carried out; at the end of the iteration cycle the policy would be $(f') = \begin{pmatrix} u^2 \\ u^1 \end{pmatrix}$ and the cost rate $\bar{c}^{f'}$ would be $-\frac{3}{4}$.

Suppose however that the states are re-ordered, so that the RPI algorithm tests state 2 first and then state 1. It can easily be verified that the policy (f'') resulting from such a cycle of single-state policy improvements would then be $(f'') = \begin{pmatrix} u^2 \\ u^2 \end{pmatrix}$, so that the cost rate $\bar{c}''$ at the end of the cycle would be $-\frac{1}{2}$ as in the Jewell iteration cycle.

The above example shows that, in contrast with the Howard/Jewell algorithm, the single-step algorithms will, in general, give different one-cycle cost reductions $\Delta\bar{c}$ for different orderings of the states of the chain. Except in certain special cases (see below) there is usually little point in trying to optimize the state ordering before using a single-step algorithm, since (a) the optimal state ordering will change from cycle to cycle, and (b) improvement in the cost reduction $\Delta\bar{c}$ achievable in a single optimization cycle does not necessarily guarantee improvement in the overall convergence properties of the algorithm.

It is perhaps worth pointing out that in the special case where the control set $\mathcal{U}$ contains only two elements (ie. a binary choice of control in each state) there is always at least one state ordering for which the cost reduction achieved by a single cycle of the RPI algorithm is at least as good as the reduction achievable by a single Jewell iteration cycle. For if $\bar{c}^J$ is the cost rate achieved by the Jewell cycle, and if $\{i_1, i_2, \ldots, i_k\} \triangleq \mathcal{X}_J (\subset \mathcal{X})$ is the set of states in which the Jewell algorithm changes the control action, we can certainly order the states in $\mathcal{X}$ so that the states in $\mathcal{X}_J$ are the first k states examined by the RPT algorithm. Furthermore there is an ordered subset of states in $\mathcal{X}_J$, say $(i_{\alpha_1}, i_{\alpha_2}, \ldots, i_{\alpha_m})$, such that the RPI algorithm can achieve a cost rate $\bar{c}^R \leqslant \bar{c}^J$ by single-state improvements first in i_{α_1}, then in i_{α_2}, and so on to i_{α_m}. For <u>either</u> the RPI algorithm makes changes in all the states in $\mathcal{X}_J$ (i.e. $m = k$), in which case $\bar{c}^R = \bar{c}^J$; <u>or</u> the RPI algorithm makes changes in at most $m < k$ states, in which case $\bar{c}^R \leqslant \bar{c}^J$ (for otherwise further one-state improvement is possible in at least one of the remaining $k-m$ states in $\mathcal{X}_J$).

The argument fails when $\mathcal{U}$ contains more than 2 elements since a control law f' is not then uniquely specified by listing the states in which it differs from some reference control law f. In practice however, re-ordering of the states has very much the same effect in the general case as when $\mathcal{U}$ is binary.

4.2.5 <u>Transient states</u>

As we have seen (Chapter 2), the state space $\mathcal{X}$ of a regular semi-Markov chain is the union of two disjoint subsets, $\mathcal{X}_T$ and $\mathcal{X}_R$, of which the first is the set of all the transient states of the chain and the second is a closed set of intercommunicating recurrent states. In a totally regular controllable chain the subsets $\mathcal{X}_T$ and $\mathcal{X}_R$ will in general depend on the choice of stationary control policy, so that in the policy-iteration algorithms considered in this Chapter policy changes may be made which move states from $\mathcal{X}_T$ to $\mathcal{X}_R$ and vice versa.

Now in a single-state policy improvement the cost reduction in any state i is given by equation (4.20) :

$$\Delta\bar{c}(f',f) \quad = \quad \theta_i^{f'} \; \Delta\xi_i(f',f)$$

where $\theta_i^{f'}$ is the scaled equilibrium probability of state i under the control law f. We also know that

$$\begin{cases} \theta_i > 0 & \Longleftrightarrow \quad i \in \mathcal{X}_R \\ \theta_i = 0 & \Longleftrightarrow \quad i \in \mathcal{X}_T \end{cases}$$

so that the cost reduction $\Delta\bar{c}(f',f)$ can be non-zero only if state i is recurrent under f'. This raises the following question: in the algorithms (API and DPI) based on optimal reduction in $\bar{c}$ at each iteration, is it possible for the algorithm to halt prematurely because changes in the control law will be confined to states which are recurrent under the changed control law ? This could happen, for example, in the following situation :

$$\begin{tikzcd}
 & f' = \begin{bmatrix} f^o(1) \\ f\ (2) \end{bmatrix} \arrow[dr, "(ii)"] & \\
f = \begin{bmatrix} f(1) \\ f(2) \end{bmatrix} \arrow[ur, "(i)"] \arrow[dr, "(i)"'] & & f^o = \begin{bmatrix} f^o(1) \\ f^o(2) \end{bmatrix} \\
 & f'' = \begin{bmatrix} f\ (1) \\ f^o(2) \end{bmatrix} \arrow[ur, "(ii)"'] &
\end{tikzcd}$$

with $\bar{c}^{f^o} < \bar{c}^{f}$. If state 1 is transient under f' and state 2 is transient under f'', (but both 1 and 2 are recurrent under f and under f^o), the control law f^o cannot be reached from f by two successive policy improvements of the API/DPI type, since the first step, f → f' or f → f'', will not be taken.

Fortunately such a situation cannot arise since it is forbidden by the following lemma.

Lemma: Let f be a feasible stationary control policy for a totally regular CSMC, and let $\mathcal{X}_T^f$ and $\mathcal{X}_R^f$ be respectively the sets of states transient under f and recurrent under f. Then for any single-state policy change f → f' in state i,

$$\text{(a)} \quad i \in \mathcal{X}_T^f \quad \Longrightarrow \quad \mathcal{X}_T^{f'} = \mathcal{X}_T^f \quad (\text{and } \mathcal{X}_R^{f'} = \mathcal{X}_R^f)$$

$$\text{(b)} \quad i \in \mathcal{X}_R^f \quad \Longrightarrow \quad i \in \mathcal{X}_R^{f'}$$

To prove this lemma, first order the states of the chain so that the states recurrent under f precede the states transient under f. The transition probability matrix P^f then has the canonical form (see, for example, Seneta$^{(1973)}$):

$$P^f = \left[\begin{array}{c|c} P_R^f & 0 \\ \hline P_{TR}^f & P_T^f \end{array}\right]$$

where

$$\left\{\begin{array}{l} P_R^f \text{ represents one-step transitions within } \mathcal{X}_R^f \\ P_T^f \text{ represents one-step transitions within } \mathcal{X}_T^f \\ P_{TR}^f \text{ represents one-step transitions from } \mathcal{X}_T^f \text{ to } \mathcal{X}_R^f \end{array}\right.$$

Now suppose that the control is changed from f to f' in state $i \in \mathcal{X}_T^f$; this will change a row of the submatrix $\left[\begin{array}{c|c} P_{TR}^f & P_T^f \end{array}\right]$. There will be no change in P_R^f : hence $\mathcal{X}_R^f$ will remain a closed set of recurrent states. But in a totally regular chain there cannot be more than one recurrent subchain. It follows that no state in $\mathcal{X}_T^f$ can become recurrent as a result of the change f → f', and hence that $\mathcal{X}_T^{f'} = \mathcal{X}_T^f$.

To prove part (b) of the lemma, consider a change from f to f' in a recurrent state $i \in \mathcal{X}_R^f$. If the change $f \rightarrow f'$ made i transient under f' we would have $i \in \mathcal{X}_T^{f'}$. But under the reverse change $f' \rightarrow f$ in the now transient state i we must, by part (a), have $\mathcal{X}_T^f = \mathcal{X}_T^{f'}$ so that $i \in \mathcal{X}_T^f$; this contradicts the assumption that $i \in \mathcal{X}_R^f$.

The important part of the lemma is (b), which asserts that the situation depicted in the left-hand side of the above diagram* cannot occur : a single-state policy change in a recurrent state i cannot make i transient under the new policy.

Now consider optimization by single-state policy improvement. We have

$$\text{f non-optimal} \quad \Longrightarrow \quad \exists f^o : \Delta\bar{c}(f^o,f) < 0$$

$$\Longrightarrow \quad \exists f^o : \sum_R \theta_i^{f^o} \Delta\xi_i(f^o,f) < 0$$

where $\sum_R$ denotes summation over all states in $\mathcal{X}_R^{f^o}$. Now by part (a) of the lemma, we cannot have $\mathcal{X}_R^{f^o} \subset \mathcal{X}_T^f$; so at least one state $i \in \mathcal{X}_R^{f^o}$ must be recurrent under f. Then the single-state policy change $f \rightarrow f'$, where

$$\begin{cases} f'(j) \triangleq f^o(j) \quad , \quad j = i \\ \quad\quad\;\; \triangleq f(j) \quad , \quad j \neq i \quad , \end{cases}$$

will, by part (b) of the lemma, leave state i recurrent under f'. So we shall have

$$\begin{cases} \theta_i^{f'} > 0 \\ \Delta\xi_i(f',f) = \Delta\xi_i(f^o,f) < 0 \end{cases}$$

and hence

$$\Delta\bar{c}(f',f) = \theta_i^{f'} \Delta\xi_i(f',f) < 0$$

* p. 112

Thus if f is non-optimal there is at least one state, recurrent under f, in which a single-state policy change can produce a reduction in the cost rate $\bar{c}$, ie. at least one state in which API/DPI iteration can continue.

There are two final points to be made on this topic. Note, in the first place, that part (b) of the lemma does not assert that $\mathcal{X}_R^{f'} = \mathcal{X}_R^{f}$ if a change $f \rightarrow f'$ is made in a recurrent state; it is quite possible for a change $f \rightarrow f'$ in the recurrent state i to move states other than i from $\mathcal{X}_R$ to $\mathcal{X}_T$ and vice versa. For example, the control change in state 1 represented by the incidence matrices

$$\begin{bmatrix} * & * & \\ * & * & \\ * & & * \end{bmatrix} \longrightarrow \begin{bmatrix} * & & * \\ * & * & \\ * & & * \end{bmatrix}$$

leaves state 1 recurrent, but changes state 2 from a recurrent state to a transient state and state 3 from a transient state to a recurrent state. In general, then, control changes in recurrent states may change the communication structure of the chain.

Secondly, policy-iteration algorithms based on minimization of the test quantities $\Delta\xi_i$ (such as Howard, Jewell, RPI) will minimize the relative values ξ_i of the transient states as well as the average cost rate $\bar{c}$ (see Howard(1960)). As has been pointed out by Schweitzer(1969), a policy which minimizes $\bar{c}$ does not necessarily satisfy the functional equations (3.31), (3.34). In Schweitzer's terminology, f^o is functional-optimal if it satisfies equations (3.31), (3.34), and is minimal-cost if it minimizes $\bar{c}$. The set of functional-optimal policies is a subset of the set of minimal-cost policies; in fact, just the subset of minimal-cost policies for which the transient state costs are also minimized. Schweitzer has shown that the Howard/Jewell algorithm always converges to a functional-optimal policy; it follows immediately that the same is true of our RPI algorithm. On the other hand, the API/DPI algorithms will converge to a minimal-cost policy which is not necessarily functional-optimal. Of course, by (2.108), a functional-optimal policy will minimize the expected total cost from <u>any</u> state (transient or recurrent); but the contribution to the expected total cost of any transient cost δ_i becomes less and less significant as the operating time increases.

4.2.6 <u>Relation between the three single-state algorithms</u>

It is perhaps worth pointing out that the three single-state policy iteration algorithms considered in this Chapter are very closely related. The DPI algorithm is, in fact, easily modified to yield the test quantities needed in the other algorithms. For equation (4.55) may be written

$$E^{-1}\ \underline{w}^f = \underline{\gamma}^f$$

or, on using (4.39),

$$\left[I - P^f + \underline{\tau}^f\ \underline{p}^T \right] \underline{w}^f = \underline{\gamma}^f \qquad \text{....(4.62)}$$

ie.

$$(I - P^f)\ \underline{w}^f + (\underline{p}^T\ \underline{w}^f)\ \underline{\tau}^f = \underline{\gamma}^f \qquad \text{....(4.63)}$$

Multiplication of (4.63) by $(\underline{\pi}^f)^T$ gives

$$(\underline{p}^T\ \underline{w}^f) = \frac{(\underline{\pi}^f)^T \underline{\gamma}^f}{(\underline{\pi}^f)^T\ \underline{\tau}^f} = \bar{c}^f$$

so that (4.63) can be written

$$(I - P^f)\ \underline{w}^f + \bar{c}^f\ \underline{\tau}^f = \underline{\gamma}^f \qquad \text{....(4.64)}$$

Comparison of (4.62) with (4.2) shows that $\underline{w}^f$ is a value-vector: in fact, the unique value-vector satisfying $\underline{p}^T \cdot \underline{w}^f = \bar{c}^f$. We could therefore use $\underline{w}^f$ to compute the test quantities $\Delta\xi_i$ required in the RPI and API algorithms. Conversely the matrix R in equation (4.3) is very closely related to the special case of E^{-1} which results from choosing $\underline{p} = \underline{e}_1$.

4.3 New successive-approximation methods

As we saw in Section 3.3.2 of the previous Chapter, a computationally convenient (though not necessarily more efficient) alternative to the policy-iteration method of optimization is the successive-approximations method developed by White(1963) for the discrete-time Markov regulation problem. A naive extension of White's method to the semi-Markov case does not work since the resulting algorithm does not always converge. In this section we develop an effective semi-Markov version of White's algorithm and also consider the possibility of using accelerated-convergence algorithms analogous to those suggested by Kushner and Kleinman(1968,1971) for the transient-cost problem. Before doing so, it will be useful to demonstrate the convergence of White's basic algorithm by means of a contraction mapping argument.

4.3.1 The White contraction mapping

In what follows, $\|\underline{x}\|$ denotes the l_∞ - norm of the real n-vector $\underline{x}$ and $\|A\|$ denotes the corresponding subordinate norm of the real n x n matrix A. Recall that the mapping $T : R^n \to R^n$ is a contraction mapping with respect to the norm $\| \;\|$ on R^n iff

$$\forall \underline{x}, \underline{y} \in R^n : \quad \|T(\underline{x}) - T(\underline{y})\| \leqslant \alpha \|\underline{x} - \underline{y}\|$$

for some $\alpha \in [0,1)$.

If T is a contraction the equation $\underline{x} = T(\underline{x})$ has a unique solution $\underline{x}^o$, called the fixed point of T, and furthermore the iteration $\underline{x}_n = T(\underline{x}_{n-1})$ converges to $\underline{x}^o$. More generally, the iteration $\underline{x}_n = T(\underline{x}_{n-1})$ converges to $\underline{x}^o$ if for some finite r the mapping $T^r : R^n \to R^n$ is a contraction (in which case T is called an r-stage contraction).

The affine mapping $T : R^n \to R^n$ defined by

$$T(\underline{x}) \triangleq A\,\underline{x} + \underline{b} \qquad \text{....(4.65)}$$

where $A \in R^{n \times n}$, $\underline{b} \in R^n$, is a contraction if $\|A\| < 1$, since

$$\begin{aligned} \|T(\underline{x}) - T(\underline{y})\| &= \|A(\underline{x} - \underline{y})\| \\ &\leqslant \|A\| \,.\, \|\underline{x} - \underline{y}\| \end{aligned}$$

Now suppose that we have some control set $\mathcal{U}$ and that the first row of $\left[A \,\vdots\, \underline{b}\right]$ is a function of $f_1 \in \mathcal{U}$, the second row of $\left[A \,\vdots\, \underline{b}\right]$ is a function of $f_2 \in \mathcal{U}$, etc. The complete matrix $\left[A \,\vdots\, \underline{b}\right]$ is then determined by the sequence $f \triangleq (f_1,\dots,f_n)$ and we denote it by $\left[A^f \,\vdots\, \underline{b}^f\right]$.

For every $f \in \mathcal{U}^n$, define the mapping $T^f : R^n \to R^n$ by

$$T^f(\underline{x}) \triangleq A^f\,\underline{x} + \underline{b}^f \qquad \text{....(4.66)}$$

Then, as before, T^f is a contraction if $\|A^f\| < 1$.

Now consider the non-linear mapping $\hat{T} : R^n \to R^n$ defined by

$$\left[\hat{T}(\underline{x})\right]_i \triangleq \underset{f_i \in \mathcal{U}}{\mathrm{Min}} \left[A^{f_i}\,\underline{x} + \underline{b}^{f_i}\right]_i \qquad \text{....(4.67)}$$

for $i = 1,2,\dots,n$.

For brevity we write equations (4.67) in the symbolic form

$$\hat{T}(\underline{x}) \triangleq \underset{f}{\mathrm{Min}} \left[T^f\,(\underline{x})\right] \qquad \text{....(4.68)}$$

Now if T^f is a contraction for every $f \in \mathcal{U}^n$ then $\hat{T}$ is also a contraction.

For we have

$$\begin{aligned} \hat{T}(\underline{x}) - \hat{T}(\underline{y}) &= \underset{f}{\mathrm{Min}}\left[T^f(\underline{x})\right] - \underset{f}{\mathrm{Min}}\left[T^f(\underline{y})\right] \\ &= T^{f_x}(\underline{x}) - T^{f_y}(\underline{y}) \end{aligned}$$

where $\begin{cases} f_x & \text{minimizes} \quad T^f(\underline{x}) \\ f_y & \text{minimizes} \quad T^f(\underline{y}) \end{cases}$

But $$T^{f_x}(\underline{x}) \leqslant T^{f_y}(\underline{x})$$

so that
$$\hat{T}(\underline{x}) - \hat{T}(\underline{y}) \leqslant T^{f_y}(\underline{x}) - T^{f_y}(\underline{y})$$

$$= A^{f_y}(\underline{x} - \underline{y}) \qquad \ldots.(a)$$

By a similar argument

$$\hat{T}(\underline{x}) - \hat{T}(\underline{y}) \geqslant A^{f_x}(\underline{x} - \underline{y}) \qquad \ldots.(b)$$

so that, from (a) and (b),

$$A^{f_y}(\underline{x} - \underline{y}) \geqslant \hat{T}(\underline{x}) - \hat{T}(\underline{y}) \geqslant A^{f_x}(\underline{x} - \underline{y})$$

Now the component of $\hat{T}(\underline{x}) - \hat{T}(\underline{y})$ with maximum modulus is either zero - in which case $\|\hat{T}(\underline{x}) - \hat{T}(\underline{y})\| = 0$; or positive - in which case it is bounded above by the corresponding component of $A^{f_y}(\underline{x} - \underline{y})$, whose modulus does not exceed $\|A^{f_y}(\underline{x} - \underline{y})\|$; or negative - in which case it is bounded below by the corresponding component of $A^{f_x}(\underline{x} - \underline{y})$, whose modulus does not exceed $\|A^{f_x}(\underline{x} - \underline{y})\|$. We conclude immediately that

$$\|\hat{T}(\underline{x}) - \hat{T}(\underline{y})\| \leqslant \|A^{f_x}(\underline{x} - \underline{y})\| \vee \|A^{f_y}(\underline{x} - \underline{y}\|$$

$$\leqslant (\|A^{f_x}\| \vee \|A^{f_y}\|)\ \|\underline{x} - \underline{y}\|$$

so that
$$\|\hat{T}(\underline{x}) - \hat{T}(\underline{y})\| \leqslant \beta\ \|\underline{x} - \underline{y}\| \quad ,$$

where
$$\beta \triangleq \underset{f}{\mathrm{Max}}\ \|A^f\|$$

and, by hypothesis, $\beta < 1$.

As a generalization of the above idea suppose that not all the matrices A^f are contraction operators but that instead they satisfy the weaker condition

$$\exists r\ \left[\forall (f^1, f^2, \ldots, f^r) \in (\mathcal{U}^n)^r\right]: \ \|A^{f^r} A^{f^{r-1}} \ldots . A^{f^2} A^{f^1}\| < 1 \qquad \ldots.(4.69)$$

Let

$$\begin{cases} \underline{x}_{r-1} \triangleq \hat{T}^{r-1}(\underline{x}) \\ \underline{y}_{r-1} \triangleq \hat{T}^{r-1}(\underline{y}) \end{cases}$$

and

$$\begin{cases} f_x^{r-1} \triangleq \underset{f}{\text{Arg. min}} \; T^f(\underline{x}_{r-1}) \\ f_y^{r-1} \triangleq \underset{f}{\text{Arg. min}} \; T^f(\underline{y}_{r-1}) \end{cases}$$

Then

$$\begin{aligned} \hat{T}^r(\underline{x}) - \hat{T}^r(\underline{y}) &= \hat{T}(\underline{x}_{r-1}) - \hat{T}(\underline{y}_{r-1}) \\ &= T^{f_x^{r-1}}(\underline{x}_{r-1}) - T^{f_y^{r-1}}(\underline{y}_{r-1}) \\ &\leqslant T^{f_y^{r-1}}(\underline{x}_{r-1}) - T^{f_y^{r-1}}(\underline{y}_{r-1}) \\ &= A^{f_y^{r-1}}(\underline{x}_{r-1} - \underline{y}_{r-1}) \\ &= A^{f_y^{r-1}}\left\{\hat{T}^{r-1}(\underline{x}) - \hat{T}^{r-1}(\underline{y})\right\} \qquad \ldots(a) \end{aligned}$$

A similar argument gives

$$\hat{T}^r(\underline{x}) - \hat{T}^r(\underline{y}) \geqslant A^{f_x^{r-1}}\left\{\hat{T}^{r-1}(\underline{x}) - \hat{T}^{r-1}(\underline{y})\right\} \ldots(b)$$

Then (4.70) follows by recursion on (a) and (b), with

$$\alpha = \underset{f^1,\ldots,f^r}{\text{Max}} \left\| A^{f^r} A^{f^{r-1}} \ldots\ldots A^{f^1} \right\|$$

Then by an extension* of the above argument it is straightforward to show that $\hat{T}$ is an r-stage contraction, ie. that $\hat{T}^r$ is a contraction:

$$\forall \underline{x}, \underline{y} \in R^n : \quad \left\| \hat{T}^r(\underline{x}) - \hat{T}^r(\underline{y}) \right\| \leqslant \alpha \left\| \underline{x} - \underline{y} \right\| \qquad \text{....(4.70)}$$

for some $\alpha \in [0,1)$.

Consider now the valu-determination equation (equation 2.110) for a controllable semi-Markov chain. If the control law is f the equation is

$$\underline{\delta} = \underline{\gamma}^f - \bar{c}\,\underline{\tau}^f + P^f \underline{\delta} \qquad \text{....(4.71)}$$

Let us ensure a unique solution to (4.71) by adjoining the constraint (see Section 4.2.6)

$$\underline{p}^T \underline{\delta} = \bar{c} \qquad \text{....(4.72)}$$

where $\underline{p}$ is an arbitrary but fixed probability vector.

Then (4.71) becomes

$$\underline{\delta} = (P^f - \underline{\tau}^f \underline{p}^T)\,\underline{\delta} + \underline{\gamma}^f \qquad \text{....(4.73)}$$

or,

$$\underline{\delta} = T^f(\underline{\delta}) \qquad \text{....(4.74)}$$

where $T^f : R^N \to R^N$ is the mapping defined by

$$T^f(\underline{x}) \triangleq (P^f - \underline{\tau}^f \underline{p}^T)\,\underline{x} + \underline{\gamma}^f \qquad \text{....(4.75)}$$

Clearly (4.75) is the particular case of (4.66) obtained by taking $n = N$, $A^f = (P^f - \underline{\tau}^f \underline{p}^T)$ and $\underline{b}^f = \underline{\gamma}^f$. Thus the solution $\underline{\delta}^o$ to (4.73) is the unique fixed point of the mapping T^f, and if T^f is a contraction the iteration

$$\underline{\delta}_n = T^f(\underline{\delta}_{n-1}) \qquad \text{....(4.76)}$$

* see facing page

will converge to $\underline{\delta}^{o}$. More generally (4.76) will converge to $\underline{\delta}^{o}$ if T is an r-stage contraction for some finite r.

Now in the particular case of a discrete-time Markov chain we have $\underline{\tau}^{f} = \underline{e}$, so that $A^{f} = (P^{f} - \underline{e}\ \underline{p}^{T})$.

Also,

$$(A^{f})^{2} = (P^{f} - \underline{e}\ \underline{p}^{T})\ P^{f} = A^{f}\ P^{f}$$

and, by induction,

$$(A^{f})^{n} = A^{f}(P^{f})^{n-1} \qquad \text{....(4.77)}$$

Thus if P^{f} is regular, so that

$$\lim_{n \to \infty} (P^{f})^{n} = \underline{e}(\underline{\pi}^{f})^{T} ,$$

we have

$$\lim_{n \to \infty} (A^{f})^{n} = A^{f} . \ \underline{e}(\underline{\pi}^{f})^{T}$$

$$= (P^{f} - \underline{e}\ \underline{p}^{T})\ \underline{e}(\underline{\pi}^{f})^{T}$$

$$= 0$$

It follows that T^{f} is an r-stage contraction for some choice of r, and hence that the iteration (4.76) will converge to $\underline{\delta}^{o}$.

Now consider the non-linear equation

$$\underline{\delta} = \hat{T}(\underline{\delta}) \qquad \text{....(4.78)}$$

where $\hat{T}$ is defined by (4.68), with T^{f} given by (4.75). The component equations of (4.78) are precisely the optimality equations, (3.31), whose solution we are seeking. If $\hat{T}$ is a contraction the required solution can be obtained by the iterative procedure

$$\underline{\delta}_{n} = \hat{T}(\underline{\delta}_{n-1}) \qquad \text{....(4.79)}$$

White's successive-approximations algorithm for the discrete-time Markov regulation problem is equivalent to iterative use of (4.79). To prove convergence of the algorithm we must therefore show that $\hat{T}$ is a contraction or at least that $\hat{T}^r$ is a contraction for some finite r. Since we are dealing with a particular case of a mapping of the form (4.68), a sufficient condition for $\hat{T}^r$ to be a contraction is condition (4.69), with, in this case, $A^{f^i} = P^{f^i} - \underline{e}\,\underline{p}^T$ for $i = 1,2,\ldots,r$.

But, as is easily verified, (4.77) generalizes to

$$A^{f^r} A^{f^{r-1}} \ldots A^{f^2} A^{f^1} = A^{f^r} (P^{f^{r-1}} \ldots P^{f^2} P^{f^1}) \qquad \ldots.(4.80)$$

In order to proceed we now invoke a rather elegant theorem due to Wolfowitz(1963) which in turn is based on some results of Hajnal(1958) on inhomogeneous products of stochastic matrices. If P is a finite stochastic matrix and if

$$\lambda(P) \triangleq 1 - \min_{i_1,i_2} \sum_j p_{i_1 j} \wedge p_{i_2 j}$$

then P is said to be a <u>scrambling matrix</u> if $\lambda(P) < 1$. The scrambling property, $\lambda(P) < 1$, implies that for every pair of distinct states i_1, i_2 there exists at least one state j (possibly i_1 or i_2 itself) accessible in one step from both i_1 and i_2. It may be shown that the set of scrambling matrices of a given order is a proper subset of the corresponding set of regular matrices.

<u>Wolfowitz' theorem</u> may for present purposes be stated in the following form :

If $\mathcal{F} \triangleq \left\{P^{(1)}, P^{(2)},\ldots,P^{(k)}\right\}$ is a finite family of stochastic matrices of the same order such that

<u>W</u> : for every positive integer n the inhomogeneous product $P_{(n)} \triangleq P_n P_{n-1} \ldots P_2 P_1$, $\forall P_i \in \mathcal{F}$, is a regular matrix,

then the following weak ergodicity property holds : -

$$\forall \varepsilon > 0 \; \exists n_o(\varepsilon) \; \forall n \geqslant n_o \; \forall P_{(n)} : \; \delta(P_{(n)}) \leqslant \varepsilon \qquad \text{....(4.81)}$$

where for any stochastic matrix P the parameter $\delta(P)$ is defined by

$$\delta(P) \triangleq \max_{j} \max_{i_1, i_2} \left| p_{i_1 j} - p_{i_2 j} \right| \qquad \text{....(4.82)}$$

Furthermore, a sufficient condtion for property W to hold is that every $P^{(i)} \in \mathcal{J}$ be a scrambling matrix.

Roughly speaking, the theorem asserts that for sufficiently large n any product $P_{(n)}$, whose factors are scrambling matrices drawn from a finite set, is a stochastic matrix with almost identical rows. In a sense the theorem can be regarded as a generalization of the asymptotic stability property R.2 (equation 2.10) for regular homogeneous chains. Now if property W holds then $P_{(n)}$ is a regular matrix and therefore possesses a unique stationary distribution $\underline{\pi}_n$. We can then write

$$P_{(n)} = \underline{e}\,\underline{\pi}_{(n)}^T + \tilde{P}_{(n)} \qquad \text{....(4.83)}$$

where $\tilde{P}_{(n)}$ is a differential matrix. As we shall now show, $\tilde{P}_{(n)}$ is a contraction operator for sufficiently large n.

First note that if P is a regular stochastic matrix with stationary distribution $\underline{\pi}$, then

$$\underline{\pi}^T = \underline{\pi}^T P$$

$$= \sum_i \pi_i \, \underline{p}_i^T \qquad \text{....(4.84)}$$

where $\underline{p}_i^T$ denotes the i^{th} row of P.

The right-hand side of (4.84) is a convex combination of the rows of P : thus $\underline{\pi}^T$ belongs to the convex hull of the rows of P. The following then holds :

$$\forall i \in N_N : \quad \left\| \underline{p}_i^T - \underline{\pi}^T \right\| \leqslant \max_j \left\| \underline{p}_i^T - \underline{p}_j^T \right\| \qquad \text{....(4.85)}$$

(Proof:

$$\begin{aligned}
\left\| \underline{p}_i^T - \underline{\pi}^T \right\| &= \max_j \left| p_{ij} - \pi_j \right| \\
&= \max_j \left| p_{ij} - \sum_k \pi_k \, p_{kj} \right| \\
&= \max_j \left| \sum_k \pi_k (p_{ij} - p_{kj}) \right| \quad \left\{ \text{since } \sum_k \pi_k = 1 \right. \\
&\leqslant \max_j \sum_k \pi_k \left| p_{ij} - p_{kj} \right| \quad \left\{ \text{since } \pi_k \geqslant 0 \right. \\
&\leqslant \sum_k \pi_k \max_j \left| p_{ij} - p_{kj} \right| \\
&= \sum_k \pi_k \left\| \underline{p}_i^T - \underline{p}_k^T \right\| \\
&\leqslant \max_k \left\| \underline{p}_i^T - \underline{p}_k^T \right\| \quad \text{by convexity} \;)
\end{aligned}$$

Thus, if $\tilde{P} \triangleq P - \underline{e}\,\underline{\pi}^T$, and $\underline{\tilde{p}}_i^T$ denotes the i th row of $\tilde{P}$, we have

$$\begin{aligned}
\max_i \left\| \underline{\tilde{p}}_i^T \right\| &= \max_i \left\| \underline{p}_i^T - \underline{\pi}^T \right\| \\
&\leqslant \max_i \max_j \left\| \underline{p}_i^T - \underline{p}_j^T \right\| \quad , \text{ by (4.85)} \\
&= \max_{i,j} \max_k \left| p_{ik} - p_{jk} \right| \\
&= \delta(P) \quad , \text{ by (4.82)}
\end{aligned}$$

Then

$$\left\| \tilde{P} \right\| = \max_i \sum_k \left| \tilde{p}_{ik} \right|$$

$$\leqslant \max_i \sum_k \max_k \left| \tilde{p}_{ik} \right|$$

$$= \max_i \quad N \, . \left\| \underline{\tilde{p}}_i^T \right\|$$

$$\leqslant N \, . \, \delta(P)$$

so that

$$\delta(P) \leqslant \varepsilon \quad \Longrightarrow \quad \| \tilde{P} \| \quad \leqslant \quad N\varepsilon \qquad \text{....(4.86)}$$

It follows that the weak ergodicity property (4.81) asserted by Wolfowitz' theorem may be re-stated : -

$$\forall \varepsilon > 0 \quad \exists n_o(\varepsilon) \; \forall n \geqslant n_o \; \forall P_{(n)} \quad : \quad \| \tilde{P}_{(n)} \| \leqslant \varepsilon$$

....(4.87)

where $\tilde{P}_{(n)}$ is given by (4.83). The contraction property follows immediately.

Returning now to the successive-approximations iteration, (4.79), suppose that for each feasible control law f the transition probability matrix P^f is a scrambling matrix. Then in analogy to (4.77) we can write (4.80) as

$$A^{f^r} A^{f^{r-1}} \ldots A^{f^2} A^{f^1} = A^{f^r} P_{(r-1)} \qquad \text{....(4.88)}$$

where

$$P_{(r-1)} \triangleq P^{f^{r-1}} \ldots\ldots P^{f^2} P^{f^1}$$

is an inhomogeneous product of the type to which Wolfowitz' theorem is applicable.

So

$$A^{f^r} \ldots A^{f^2} A^{f^1} = A^{f^r} P_{(r-1)}$$

$$= (P^{f^r} - \underline{e} \, \underline{p}^T) \, (\underline{e} \, \underline{\pi}^T_{(r-1)} + \tilde{P}_{(r-1)})$$

$$= A^{f^r} \tilde{P}_{(r-1)}$$

$$\longrightarrow 0 \quad , \text{as } r \longrightarrow \infty$$

since, by (4.87), $\tilde{P}_{(r-1)}$ converges element-wise to the zero matrix as r increases.

It follows immediately that for sufficiently large r the matrix $A^{f^r} \dots A^{f^2} A^{f^1}$ is a contraction operator on R^N. Hence by (4.69) and (4.70) the successive-approximations algorithm based on (4.79) is convergent.

The requirement, in our statement of Wolfowitz' theorem, that every matrix in the family $\mathcal{F} = \left\{P^{(1)}, \dots, P^{(k)}\right\}$ be a scrambling matrix, is in fact unnecessarily restrictive, and may be relaxed in either of the following ways.

(1) If the family $\mathcal{F}$ is such that

<u>P.1</u> : for some fixed $n_1 \geqslant 1$, every product

$$P_{(n_1)} \triangleq P_{n_1} P_{n_1-1} \dots P_2 P_1 \quad , \quad \forall P_i \in \mathcal{F} ,$$

is a scrambling matrix,

then the weak ergodicity property (4.81), and hence (4.87), will hold.

Thus a sufficient condition for the successive-approximations scheme to converge is that there exist an n_1 such that for every feasible sequence of control laws $(f^1, f^2, \dots, f^{n_1})$ the product $(P^{f^{n_1}} P^{f^{n_1-1}} \dots P^{f^2} P^{f^1})$ is a scrambling matrix.

As an example of a controllable chain in which, although the individual P-matrices, P^f, are not scrambling, condition P.1 is satisfied, consider a controllable birth-death process (Çinlar(1975)), in which transitions are possible only between adjacent states. For such a process, P^f is tri-diagonal for every feasible f, and hence (for $N \geqslant 4$) not scrambling. However, it is easy to see that condition P.1 will hold with $n_1 \geqslant \frac{1}{2}(N - 1)$, where N is the number of states.

It is not, in general, possible to express condition P.1 in terms of equivalent conditions on the individual members of $\mathcal{F}$.

(2) If the family $\mathcal{F}$ is such that

<u>P.2</u> : each $P^{(i)} \in \mathcal{F}$ is regular and has a strictly positive principal diagonal,

then property W of Wolfowitz' theorem holds (see Seneta$^{(1973)}$) and hence so does the weak ergodicity property (4.81).

Thus an alternative sufficient condition for the successive-approximations algorithm to converge is that each of th^e^ feasible transition probability matrices P^f have a strictly positive principal diagonal. (Since we are dealing with totally regular processes the regularity of each P^f is assured.) This is a most important result since, as we shall show in Section 4.3.2 it is always possible to transform an optimal regulation problem into an equivalent problem in which every feasible P^f has a strictly positive principal diagonal. (Footnote: A stochastic matrix with a strictly positive principal diagonal is said to be <u>normed</u>.)

We have presented the above convergence proof in detail because although it is less direct than White's own proof$^{(1963)}$ it has the following advantages:

(i) It establishes convergence under less stringent conditions than those required by White. He requires that for some n_1 every product $P_{(n_1)} = P_{n_1} P_{n_1-1} \cdots P_1$ (where each P_i is a feasible transition probability matrix) be a Markov matrix (a stochastic matrix with a strictly positive column). In fact, more is required: there must be a state, say m, such that for some n_1 the m^{th} column of every product $P_{(n_1)}$ be strictly positive. We require only that every product $P_{(n_1)}$ be scrambling; or, alternatively, that each feasible P^f has a strictly positive principal diagonal.

(ii) By emphasizing the role of the matrices A^f in determining the contractive properties of the successive- approximations transformation $\hat{T}$, the proof offers a hint as to how the method might be extended to the semi-Markov case. We now proceed to consider this extension.

4.3.2 Successive-approximations in the semi-Markov problem

White's algorithm is designed to solve the discrete-time Markov regulation problem. As we have shown, the convergence of the algorithm rests on the contractive nature of the matrices $A^f \triangleq \left[P^f - \underline{e}\,\underline{p}^T\right]$ arising in the iterative solution of equation (4.78). In the more general case of semi-Markov regulation, the vector $\underline{e}$ of unit sojourn times is replaced by the vector $\underline{\tau}^f$ of mean sojourn times under the current policy; that is, we must work with matrices $A^f \triangleq \left[P^f - \underline{\tau}^f\,\underline{p}^T\right]$. The crucial properties (4.77) and (4.80) then no longer hold, $\hat{T}$ is no longer necessarily a contraction mapping, and (4.79) is not guaranteed to converge.

However, as we have seen (Section 3.2.3), any semi-Markov regulation problem is equivalent to some discrete-time Markov regulation problem. This observation leads directly to an appropriate extension of the White algorithm to the semi-Markov case.

Consider a totally regular CSMC, $\{X_t\}$, and the corresponding equal-sojourn-time chain, $\{X_t^*\}$, derived via the transformation defined by equations (3.8) - (3.15). For any feasible control law f, let $(P^*)^f$, $(\underline{\tau}^*)^f$ and $(\underline{\gamma}^*)^f$ be respectively the transition probability matrix, the mean sojourn-time vector and the mean one-step cost vector, of the chain $\{X_t^*\}$.

Then

$$(\underline{\tau}^*)^f = \tau_0\,\underline{e} \qquad \text{....(4.89)}$$

where τ_0 is chosen according to equations (3.8) - (3.10). Because of the way in which the equivalence transformation is defined, (4.89) holds for every feasible control law f; hence, for every feasible f, $\{X_t^*\}$ is weakly equivalent to a pure Markov chain with index set $\mathcal{T} = \{0, \tau_0, 2\tau_0, ...\}$. The original semi-Markov regulation problem has been converted to a pure Markov regulation problem to which White's successive-approximations algorithm is applicable.

The rest is straightforward. For the transformed problem, equation (4.71) becomes

$$\underline{\delta}^* = (\underline{\gamma}^*)^f - \bar{c}^*(\underline{\tau}^*)^f + (P^*)^f \underline{\delta}^* \qquad ....(4.90)$$

and if we now add the constraint

$$\underline{p}^T \underline{\delta}^* = \tau_0 \bar{c}^* \qquad ....(4.91)$$

equation (4.90) can be written

$$\underline{\delta}^* = T^f(\underline{\delta}^*) \qquad ....(4.92)$$

where, now, the mapping T^f is defined by

$$T^f(\underline{x}) \triangleq \left[(P^*)^f - \underline{e}\,\underline{p}^T\right]\underline{x} + (\underline{\gamma}^*)^f \qquad ....(4.93)$$

As before, the non-linear mapping $\hat{T}$ defined by (4.68) will be an r-stage contraction for some $r < \infty$ provided that (4.69) holds with, now, $A^f \triangleq \left[(P^*)^f - \underline{e}\,\underline{p}^T\right]$. But, for any f, P^f and $(P^*)^f$ possess the same communication structure (ie. same incidence matrix). Thus if in the original semi-Markov problem the matrices P^f satisfy the weak ergodicity condition then the transformed successive-approximations algorithm

$$\underline{\delta}^*_n = \hat{T}(\underline{\delta}^*_{n-1}) \qquad ....(4.94)$$

will converge to the unique solution of the transformed semi-Markov regulation problem - which, as we have seen (Section 3.2.3), is the solution to the original (untransformed) problem.

It remains to write (4.94) in terms of the characteristic parameters of the original CSMC. In component form, (4.94) is

$$(\delta^*_i)_n = \underset{u\in\mathcal{U}}{\text{Min}}\left[(P^* - \underline{e}\,\underline{p}^T)\,\underline{\delta}^*_{n-1} + \underline{\gamma}^*\right]_i \quad , \quad \forall i \in N_N \qquad \text{....(4.95)}$$

which, on using (3.11) - (3.15) and (4.91), becomes

$$\begin{aligned}(\delta^*_i)_n &= \underset{u\in\mathcal{U}}{\text{Min}}\left[\left\{I - \Phi(I-P)\right\}\underline{\delta}^*_{n-1} - \tau_o\,\bar{c}^*_{n-1}\,\underline{e} + \Phi\,\underline{\gamma}\right]_i , \forall i \in N_N \\ &= \underset{u\in\mathcal{U}}{\text{Min}}\left[\underline{\delta}^*_{n-1} + \Phi\left\{(P-I)\,\underline{\delta}^*_{n-1} - \bar{c}^*_{n-1}\,\underline{\tau} + \underline{\gamma}\right\}\right]_i , \forall i \in N_N \\ &= (\delta^*_i)_{n-1} + \underset{u\in\mathcal{U}}{\text{Min}}\left[\Phi\left\{(P-I)\,\underline{\delta}^*_{n-1} - \bar{c}^*_{n-1}\,\underline{\tau} + \underline{\gamma}\right\}\right]_i , \forall i \in N_N\end{aligned}$$

Finally, dropping the * and using (3.12), (3.13), we obtain the required iteration equations : -

$$\boxed{(\delta_i)_n = (\delta_i)_{n-1} + \underset{u\in\mathcal{U}}{\text{Min}}\left[\frac{\tau_o}{\tau^u_i}\left\{\gamma^u_i + \sum_j p^u_{ij}(\delta_j)_{n-1} - \bar{c}_{n-1}\tau^u_i - (\delta)_{in-1}\right\}\right], \quad \forall i \in N_N} \qquad \text{....(4.96)}$$

Provided that the matrices P^f satisfy the weak ergodicity condition, iteration of equations (4.96) will yield a value-vector $\underline{\delta}$ satisfying $\underline{p}^T\,\underline{\delta} = \tau_o\,\bar{c}$ where $\bar{c}$ is the minimal equilibrium mean cost rate.

Comments:

(i) The iterative algorithm defined by (4.96) has been developed as a natural generalization of White's algorithm to the case of semi-Markov regulation. Thus, as we should expect, equations (4.96) reduce to White's equations when the state transitions are uniformly spaced in time (ie. when $\tau^u_i = \tau_o$, $\forall i$, u).

(ii) The rate of convergence of (4.96) will depend, inter alia, on the magnitude of the equivalent sojourn time, τ_o, defined by equation (3.10). In general it will pay to make τ_o as large as possible: thus we would normally take $K = 1$ in (3.10).

Note however that if $K < 1$ then, by equations (3.9) - (3.13), the equivalent transition probability matrix $(P^*)^f$ will possess a strictly positive principal diagonal (even though P^f may not) for every feasible control law f. Thus by choosing K to be slightly less than unity we can guarantee that property P.2 of Section 4.3.1 will hold for the transformed matrices $(P^*)^f$ and hence, by Wolfowitz' theorem, that the successive-approximations algorithm will always converge when applied to the transformed regulation problem (provided only that all the P^f, and hence all the $(P^*)^f$ are regular - as they will be for any totally regular chain).

(iii) A closely-related iterative procedure has been proposed by Schweitzer(1971b) whose argument is based on the idea that a semi-Markov chain is (in a sense not explicitly defined) equivalent to a continuous-time pure Markov chain. We have presented the above development, partly as independent corroboration of Schweitzer's conclusions, and partly because the arguments on which it rests are (at least in our view) rather more compelling than those put forward by Schweitzer.

4.3.3 An accelerated-convergence algorithm

In two related papers by Kushner and Kleinman(1968,1971) on the transient-cost problem (for Markov chains with an absorbing state), Kushner and Kleinman point out that some of the convergence results of linear iterative analysis are relevant to the non-linear iterations arising in the successive-approximations method for the transient-cost problem.

As is well known, the affine mapping

$$T_o : R^n \rightarrow R^n \quad , \quad \underline{x} \longmapsto A_o\underline{x} + \underline{b} \qquad \text{....(4.97)}$$

has the fixed point $\underline{x}^o = (I - A_o)^{-1}\underline{b}$, provided that the matrix A_o does not have a unit eigenvalue. Furthermore, the iteration $\underline{x}_n = T_o(\underline{x}_{n-1})$ converges to $\underline{x}^o$ iff the <u>spectral radius</u>, $\sigma(A_o)$, of A_o is strictly less than unity, and the rate of convergence increases with decreasing $\sigma(A_o)$. Now A_o can be split into three terms:

$$A_o = L + D + U \qquad \text{....(4.98)}$$

where

$$\left\{ \begin{array}{ll} L & \text{is strictly lower triangular} \\ D & \text{is diagonal} \\ U & \text{is strictly upper triangular} \end{array} \right.$$

Then, provided $(I - D)$ is non-singular, each of the mappings

$$T_1 : \underline{x} \longmapsto A_1\underline{x} + (I - D)^{-1}\underline{b} \qquad \text{...(4.97a)}$$

$$T_2 : \underline{x} \longmapsto A_2\underline{x} + (I - L)^{-1}\underline{b} \qquad \text{...(4.97b)}$$

$$T_3 : \underline{x} \longmapsto A_3\underline{x} + (I - D - L)^{-1}\underline{b} \qquad \text{...(4.97c)}$$

where

$$A_1 \triangleq (I - D)^{-1}(L + U) \qquad \text{...(4.98a)}$$

$$A_2 \triangleq (I - L)^{-1}(D + U) \qquad \text{...(4.98b)}$$

$$A_3 \triangleq (I - D - L)^{-1}U \qquad \text{...(4.98c)}$$

has the same fixed point as the mapping T_o. The iteration $\underline{x}_n = T_i(\underline{x}_{n-1})$ will converge iff $\sigma(A_i) < 1$. Iterations using T_o or T_1 are called Jacobi iterations, while those using T_2 or T_3 are called Gauss-Seidel iterations.

If A_o is a non-negative matrix, the following results concerning the spectral radii of A_o, A_1, A_2 and A_3 are available in a theorem due to Stein and Rosenberg(1948) :

$$\text{(i)} \quad \sigma(A_o) = 1 \quad \Longrightarrow \quad \sigma(A_1) = \sigma(A_2) = \sigma(A_3) = 1$$

$$\text{(ii)} \quad \sigma(A_o) < 1 \quad \Longrightarrow \quad \sigma(A_2) < \sigma(A_o) \quad \underline{\text{and}}\ \sigma(A_3) < \sigma(A_1)$$

provided that A_o is irreducible (see Seneta(1973)). If, in addition, the diagonal elements of A_o belong to $[0,1)$, with at least one $a_{ii} > 0$, then also

$$\text{(iii)} \quad \sigma(A_o) < 1 \qquad \sigma(A_1) < \sigma(A_o) \quad \underline{\text{and}}\ \sigma(A_3) < \sigma(A_2)$$

Thus under the appropriate conditions a Gauss-Seidel iterative scheme is more rapidly convergent than the corresponding Jacobi iterative scheme. In their first paper(1968) on the transient-cost problem, Kushner and Kleinman show that a "Gauss-Seidel" successive-approximations method based on the use of (4.97b) converges more rapidly than the corresponding "Jacobi" method based on the use of (4.97). In their second paper(1971) they consider accelerated methods based on so-called over-relaxation iterative methods of the form $\underline{x}_n = T'(\underline{x}_{n-1})$, in which T' is one of the mappings T'_o, T'_1, T'_2, T'_3, where T'_o, T'_1, T'_2, T'_3 are given by (4.97), (4.97a), (4.97b), (4.97c) after replacing A_o by $\left[\omega A_o + (1 - \omega)I\right]$ with $\omega \in (1,2)$.

The Kushner/Kleinman results suggest that the convergence of the successive-approximations algorithm (4.94) will be accelerated if we can reduce the spectral radius of each of the matrices $A^f \triangleq \left[(P^*)^f - \underline{e}\,\underline{p}^T\right]$ by transforming from the "Jacobi" form of (4.94) to the corresponding "Gauss-Seidel" form. Unfortunately, however, the matrices A^f are not in general non-negative and the Stein-Rosenberg theorem is no longer applicable. Thus even though, for any feasible f, $\sigma(A^f) < 1$ (since $(A^f)^n \xrightarrow[n]{} 0$), we cannot conclude that the corresponding Gauss-Seidel matrix will have a spectral radius less that $\sigma(A^f)$; indeed, the new spectral radius may even be greater than one. Thus the only obvious way of accelerating the convergence of the basic algorithm (4.94) is by the use of over-relaxation in conjunction with the "Jacobi" matrix A^f.

In the affine mapping T_o defined by (4.97) replace the matrix A_o by the corresponding <u>accelerated-Jacobi</u> matrix

$$A(\omega) \triangleq \left[\omega A_o + (1-\omega) I\right] \qquad \text{....(4.99)}$$

where $\omega \in R_+$; and replace $\underline{b}$ by $\omega\underline{b}$. The resultant mapping, T_o', will have the same fixed point as T_o (provided that ω is chosen so that $A(\omega)$ does not have a unit eigenvalue). From (4.99), the eigenvalues λ_1', λ_2',..., of $A(\omega)$ are related to the eigenvalues λ_1, λ_2,..., of A_o by

$$\lambda_i' = \omega\lambda_i + (1-\omega) \qquad \text{....(4.100)}$$

and so, for given ω, the spectral radius of $A(\omega)$ is given by

$$\sigma(A(\omega)) = \underset{i}{\mathrm{Max}} \left|\omega\lambda_i + (1-\omega)\right| \qquad \text{....(4.101)}$$

Now the linear iteration $\underline{x}_n = T_o'(\underline{x}_{n-1})$ will converge more rapidly than the iteration $\underline{x}_n = T_o(\underline{x}_{n-1})$ if we can find an ω such that $\sigma(A(\omega)) < \sigma(A_o)$ (assumed < 1). Clearly such an ω must differ from unity since, by (4.99), $A(1) = A_o$. <u>Under-relaxation</u> corresponds to $\omega < 1$, <u>over-relaxation</u> to $\omega > 1$; which procedure is used will depend on the location of the eigenvalues of the original matrix A_o.

Now consider the application of the transformation (4.99) to the successive-approximations algorithm (4.94). At each iteration the matrix A_o will have the form $P - \underline{e}\,\underline{p}^T$, where $\underline{p}^T$ is a fixed probability vector and P is a regular stochastic matrix whose elements depend on the current control law.

Lemma : Let the eigenvalues of the regular stochastic matrix P be $\lambda_1(=1), \lambda_2, \lambda_3, \ldots, \lambda_n$. Then the matrix $A = P - \alpha\,\underline{e}\,\underline{p}^T$, where α is a real number and $\underline{p}^T$ is a probability vector, has eigenvalues $(1 - \alpha)$, $\lambda_2, \lambda_3, \ldots, \lambda_n$.

Proof : We have

$$A\underline{e} = \left[P - \alpha\,\underline{e}\,\underline{p}^T\right]\underline{e} = P\,\underline{e} - \alpha\,(\underline{p}^T\underline{e})\,\underline{e}$$

$$= (1 - \alpha)\,\underline{e}$$

so that $\underline{e}$ is an eigenvector of A, with associated eigenvalue $(1 - \alpha)$.

For any $\lambda_i \neq \lambda_1$ let $\underline{q}_i$ be an associated eigenvector of P. Then, for any complex number μ_i

$$A(\underline{q}_i + \mu_i\,\underline{e}) = \left[P - \alpha\,\underline{e}\,\underline{p}^T\right](\underline{q}_i + \mu_i\,\underline{e})$$

$$= P\,\underline{q}_i - \alpha(\underline{p}^T\,\underline{q}_i)\underline{e} + \mu_i\,P\underline{e} - \alpha\mu_i\,\underline{e}$$

$$= \lambda_i\,\underline{q}_i + \left[\mu_i - \alpha\mu_i - \alpha(\underline{p}^T\,\underline{q}_i)\right]\underline{e}$$

So if μ_i is chosen so that

$$\mu_i = \left\{\frac{\alpha(\underline{p}^T\,\underline{q}_i)}{1 - \alpha - \lambda_i}\right\}$$

then

$$A(\underline{q}_i + \mu_i\,\underline{e}) = \lambda_i(\underline{q}_i + \mu_i\,\underline{e})$$

so that λ_i is an eigenvalue of A, with an associated eigenvector

$(\underline{q}_i + \mu_i \underline{e})$.

(Note : In the particular case when $\alpha = 1 - \lambda_i$ the associated eigenvector is $\underline{e}$ and the eigenvalue $(1 - \alpha)$ is multiple.)

Now since P is a regular stochastic matrix the only eigenvalue of P with unit modulus is the principal eigenvalue, $\lambda_1 = 1$; the remaining eigenvalues λ_i all lie in the interior of the unit disc, $|\lambda| \leqslant 1$, in the complex plane. It follows from the lemma that for $\alpha \in (0,2)$ the eigenvalues of A all lie in the interior of the unit disc and hence that the spectral radius of A is strictly less than one. In fact, provided that $|1 - \alpha| \leqslant \max_{i \neq 1} |\lambda_i|$, $\sigma(A)$ will be independent of the value of α, being equal to $|\lambda_2|$, the modulus of the subdominant eigenvalue of P. In particular this is the case when $\alpha = 1$, as in our successive-approximations algorithm.

We should expect an accelerated algorithm to result if, at each iteration, the matrix $A_o = \left[P - \underline{e}\,\underline{p}^T\right]$ is replaced by the corresponding accelerated-Jacobi matrix $A(\omega)$ defined, for some suitably chosen relaxation factor ω, by (4.99). If the eigenvalues of P are known then so, by the above lemma, are those of A_o and it is possible in principle to determine the optimal value of ω, ie. the value of ω which minimizes $\sigma(A(\omega))$. In practice, of course, the eigenvalues of P are not known and a suitable value of ω must be estimated and, if necessary, improved by trial and error. We now show how a reasonable estimate of ω may be made in some cases.

We have seen that, apart from the special eigenvalue $(1 - \alpha)$, the eigenvalues of $A = \left[P - \alpha\,\underline{e}\,\underline{p}^T\right]$ are eigenvalues of P. Now the eigenvalues of P all lie on the largest Gershgorin disc(1962) associated with P : that is, the region G_p in the λ-plane defined by

$$G_p \triangleq \left\{\lambda : \left|\lambda - p_o\right| \leqslant 1 - p_o\right\} \qquad \ldots\ldots(4.102)$$

where p_o is the smallest diagonal element of P. Thus with α suitably chosen (ie. so that the eigenvalue $\lambda_1' = (1 - \alpha)$ belongs to G_p), G_p will contain all the eigenvalues of A. The situation is illustrated in Fig.(8) for the case

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.4 & 0.4 \end{bmatrix} \quad ; \quad \alpha = 1$$

In this example the eigenvalue of A with largest modulus is $\lambda = 0.8$: thus $\sigma(A) = 0.8$. If we now use over-relaxation ($\omega > 1$) to shift the eigenvalues to the left, we can produce an accelerated-Jacobi matrix $A(\omega)$ whose spectral radius is less that $\sigma(A)$. For example a relaxation factor of $\omega = \frac{5}{3}$ will reduce the spectral radius from 0.8 to 0.67.

Generally speaking we can say that if $p_o << 1$ G_p will cover most of the unit disc $\{\lambda : |\lambda| \leqslant 1\}$ and it is unlikely that the spectral radius of A can be significantly reduced by over-relaxation. If on the other hand p_o is not too small (say $\geqslant 0.2$) G_p will be mainly in the right half-plane and it is then likely that over-relaxation can reduce the spectral radius of A. We now consider briefly how the relaxation factor ω should be chosen. In doing so it is convenient to work with the parameter

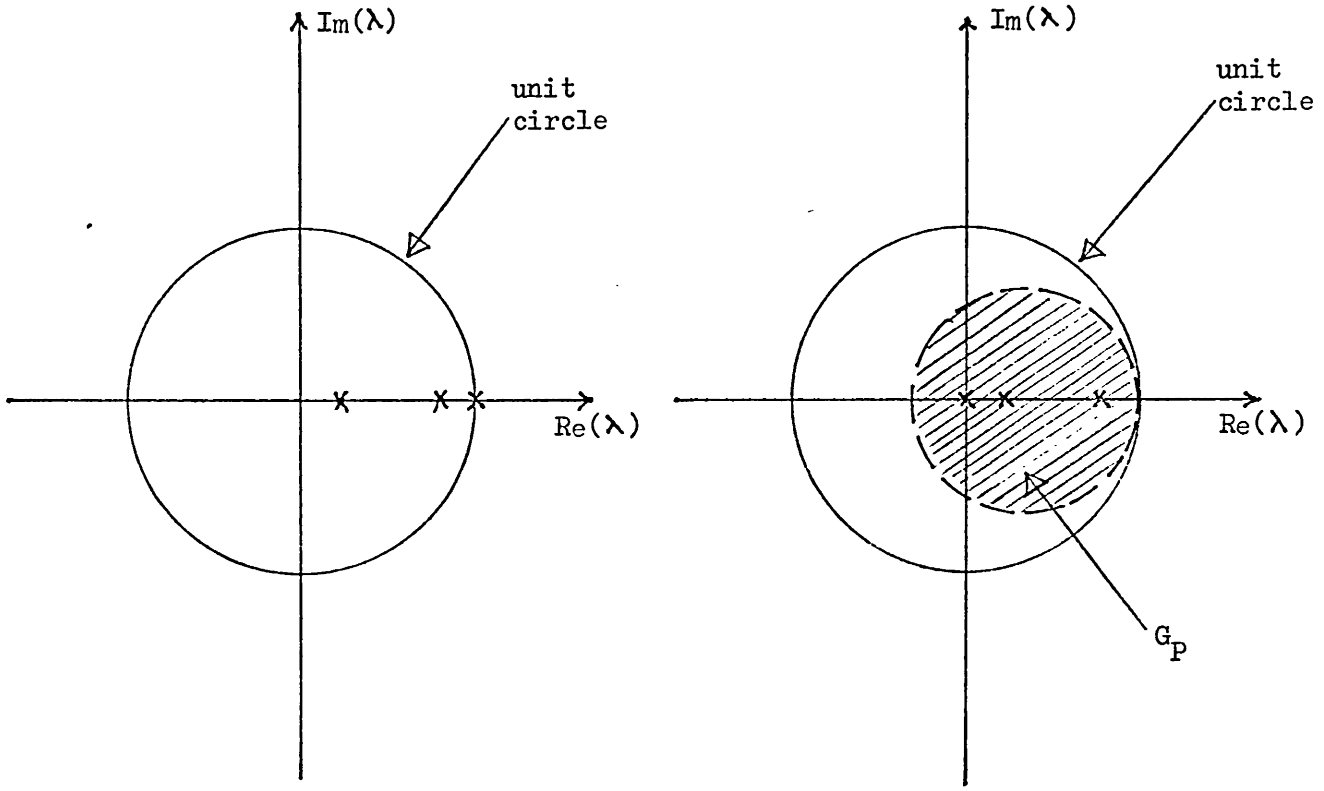
$$\beta \triangleq \omega^{-1} - 1$$

in terms of which equation (4.99) and (4.100) have the forms

$$A(\beta) = (1 + \beta)^{-1} (A_o + \beta I) \qquad \text{....(4.103)}$$

$$\lambda_i' = \left\{ \frac{\lambda_i + \beta}{1 + \beta} \right\} \qquad \text{....(4.104)}$$

We now have the following fact available to us (see, for example,



(a) Eigenvalues of P

(b) Eigenvalues of $A = \left[P - \underline{e}\,\underline{p}^T\right]$

Fig. (8)

Isaacson and Keller$^{(1966)}$) : if all the eigenvalues of P (and hence of A) are <u>real</u>, the value of β for which $\sigma(A(\beta))$ is a minimum is

$$\beta_o = -\frac{1}{2}(\lambda_{max} + \lambda_{min})$$

where λ_{max} and λ_{min} are the maximal and minimal eigenvalues of A.

In practice λ_{max} and λ_{min} will not be known. However, we know that they belong to the largest Gershgorin disc G_p and hence that

$$\begin{cases} \lambda_{max} & \leqslant & 1 \\ \lambda_{min} & \geqslant & 2\,p_o - 1 \end{cases}$$

If then we assume that $\lambda_{max} = 1 - \varepsilon$, $\lambda_{min} = (2p_o - 1) + \varepsilon$, we obtain the estimate

$$\beta_o \simeq -p_o \qquad \ldots.(4.105)$$

ie. the optimal relaxation factor when all the eigenvalues are real is $\omega_o \simeq (1 - p_o)^{-1}$. Thus in the above example the value $\omega = \frac{5}{3}$ is in fact the optimal value of ω . It must be emphasized that the estimate (4.105) can be very different from the true optimum if λ_{max} and λ_{min} are not symmetrically disposed about the centre of G_p as they are in the example.

Two further comments are in order here :

(i) The estimate $\beta_o \simeq -p_o$ is based on the assumption that the eigenvalues of P are all real. This will be so if, for example, P is symmetric or — of more importance in practice — if P is tri-diagonal (as it will be for a controllable birth-death process).

(ii) The accelerated matrix $P(\beta) = (1 + \beta)^{-1}\left[P + \beta I\right]$, with $\beta = -p_o$, is a stochastic matrix in which the diagonal element $p_{i_1 i_1} \triangleq \min_i (p_{ii}) = p_o$ is replaced by 0, but whose incidence matrix is otherwise the same as that of P. This implies that even if P has some complex eigenvalues, use of the acceleration factor $\omega = (1 - p_o)^{-1}$ will not take any eigenvalue of P outside the unit disc in the complex plane – though in such a case there may well be no reduction in spectral radius. Furthermore, if ω is chosen slightly greater than $- p_o$ so that $\omega < (1 - p_o)^{-1}$, the accelerated matrix $P(\beta)$ will have the same incidence matrix as P ; thus if P is regular (scrambling/normed/etc.) $P(\beta)$ will be regular (scrambling/normed/etc.).

Returning to the problem of accelerating the successive approximations algorithm (4.94), we now propose an algorithm based on the use of the acceleration factor $\omega = 1 + p_o$, a value which is between 1 and the "optimal" value $(1 - p_o)^{-1}$ and which is close to the latter when $p_o \ll 1$ (as it usually will be in practice).

Using the symbolic notation defined in (4.68) equation (4.94) may be written in the more explicit form

$$\underline{\delta}^*_n = \underset{f}{\text{Min}}\left[A^f \, \underline{\delta}^*_{n-1} + (\underline{\gamma}^*)^f\right] \qquad \text{....(4.106)}$$

where

$$A^f = \left[(P^*)^f - \underline{e}\,\underline{p}^T\right]$$

Equation (4.106) defines the unaccelerated form of the successive-approximations scheme. Now consider the following scheme : -

Set $\underline{\delta}_o$ = arbitrary real N-vector and, for $n \geqslant 1$,

$$\left\{ \begin{array}{ll} \tilde{\underline{\delta}}_n = \underset{f}{\text{Min}}\left\{\left[(P^*)^f - \underline{e}\,\underline{p}^T\right] \underline{\delta}_{n-1} + (\underline{\gamma}^*)^f\right\} & \text{....(4.107)} \\ \underline{\delta}_n = \omega_n \tilde{\underline{\delta}}_n + (1 - \omega_n)\,\underline{\delta}_{n-1} & \text{....(4.108)} \end{array} \right.$$

where

$$\omega_n = 1 + p_n \qquad \text{....(4.109)}$$

with
$$p_n \triangleq \underset{i}{\text{Min}} \; (p^*_{ii})^{f^n} \qquad \ldots.(4.110)$$

where f^n is the control law which minimizes the right-hand side of (4.107).

Equations (4.107 - 8) define an accelerated iteration on $\underline{\delta}_n$ with a variable acceleration parameter ω_n. To see this write (4.107) in the form

$$\underline{\tilde{\delta}}_n = \left[(P^*)^{f^n} - \underline{e}\,\underline{p}^T\right] \underline{\delta}_{n-1} + (\underline{\gamma}^*)^{f^n} \qquad \ldots.(4.111)$$

and substitute this into (4.108) to get

$$\underline{\delta}_n = \left[\tilde{P}^n - \omega_n\,\underline{e}\,\underline{p}^T\right] \underline{\delta}_{n-1} + \underline{\tilde{\gamma}}^n \qquad \ldots.(4.112)$$

where

$$\tilde{P}^n \triangleq \left[I - \omega_n(I - (P^*)^{f^n})\right] \qquad \ldots.(4.113)$$

and

$$\underline{\tilde{\gamma}}^n \triangleq \omega_n(\underline{\gamma}^*)^{f^n} \qquad \ldots.(4.114)$$

The matrix $\tilde{P}^n$ is the accelerated version of $(P^*)^{f^n}$; note that with ω_n chosen as in (4.109 - 10) $\tilde{P}^n$ will be a stochastic matrix with the same incidence matrix as $(P^*)^{f^n}$ and, it is to be hoped, with a smaller spectral radius than $(P^*)^{f^n}$.

If $\underline{\delta}$ is a solution to the non-linear difference equations (4.107 - 8) then, as is easily seen,

$$\underline{\delta} = \underset{f}{\text{Min}}\left\{\left[(P^*)^f - \underline{e}\,\underline{p}^T\right]\underline{\delta} + (\underline{\gamma}^*)^f\right\} \qquad \ldots.(4.115)$$

ie. $\underline{\delta}$ is indeed the value-vector for the given optimization problem, since (4.115) is precisely the equation $\underline{\delta} = \hat{T}(\underline{\delta})$ with $\hat{T}$ defined by (4.68) and (4.93).

Recall that the vector $\underline{p}$ is an arbitrary probability N-vector. We now show that in the particular case $\underline{p} = \left(\frac{1}{N}\right)\underline{e}$ the above accelerated successive-approximations (ASA) algorithm is guaranteed to converge for any $\underline{\delta}_o$.

If f^o is the optimal control law then

$$\underline{\delta} = A^{f^o} \underline{\delta} + (\underline{\gamma}^*)^{f^o}$$

where, as usual, $A^f \triangleq \left[(P^*)^f - \underline{e}\,\underline{p}^T\right]$. It follows that, for any ω,

$$\underline{\delta} = \omega\left[A^{f^o} \underline{\delta} + (\underline{\gamma}^*)^{f^o}\right] + (1 - \omega)\,\underline{\delta}$$

In particular,

$$\begin{aligned}\underline{\delta} &= \omega_n\left[A^{f^o} \underline{\delta} + (\underline{\gamma}^*)^{f^o}\right] + (1 - \omega_n)\,\underline{\delta} \\ &\leqslant \omega_n\left[A^{f^n} \underline{\delta} + (\underline{\gamma}^*)^{f^n}\right] + (1 - \omega_n)\,\underline{\delta}\end{aligned}$$

by the minimizing properties of f^o.

But, at stage n,

$$\underline{\delta}_n = \omega_n\left[A^{f^n} \underline{\delta}_{n-1} + (\underline{\gamma}^*)^{f^n}\right] + (1 - \omega_n)\,\underline{\delta}_{n-1}$$

so that, subtracting,

$$\underline{\delta} - \underline{\delta}_n \leqslant \omega_n A^{f^n} (\underline{\delta} - \underline{\delta}_{n-1}) + (1 - \omega_n)(\underline{\delta} - \underline{\delta}_{n-1})$$

ie.
$$\underline{\varepsilon}_n \geqslant \tilde{A}^n \underline{\varepsilon}_{n-1} \qquad \text{....(4.116)}$$

where
$$\underline{\varepsilon}_n \triangleq \underline{\delta}_n - \underline{\delta} \qquad \text{....(4.117)}$$

and
$$\tilde{A}^n \triangleq \left[\omega_n A^{f^n} + (1 - \omega_n)\, I\right] \qquad \text{....(4.118)}$$

$$= \left[\tilde{P}^n - \omega_n \underline{e}\,\underline{p}^T\right] \qquad \text{....(4.119)}$$

Iterating on $\underline{\varepsilon}$,

$$\underline{\varepsilon}_n \geqslant \left[\tilde{A}^n \tilde{A}^{n-1} \ldots \tilde{A}^2 \tilde{A}^1\right] \underline{\varepsilon}_o \qquad \text{....(4.120)}$$

We now require a generalization of (4.80), which because of its intrinsic interest, we state as a

<u>Lemma</u>: Let A_i, $i = 1,2,\ldots$ be an infinite sequence of matrices, each of the form $A_i = \left[P_i - \omega_i \underline{e}\, \underline{p}^T\right]$, where P_i is stochastic and $\omega_i \in (0,2)$.

Then for any $m \geqslant 1$

$$\lim_{n \to \infty} \left[A_{n+m} A_{n+m-1} \cdots A_2 A_1 - A_{n+m} A_{n+m-1} \cdots A_{m+1} P_m P_{m-1} \cdots P_1\right] = 0$$

<u>Proof</u>: We have

$$A_{n+m} \cdots A_2 A_1 = A_{n+m} \cdots A_2 (P_1 - \omega_1 \underline{e}\, \underline{p}^T)$$

$$= A_{n+m} \cdots A_2 P_1 - \omega_1 \prod_{i=2}^{n+m} (1-\omega_i)\, \underline{e}\, \underline{p}^T$$

since $\underline{e}$ is an eigenvector of A_i with corresponding eigenvalue $(1-\omega_i)$.

But $\omega_i \in (0,2)$ and so $|(1-\omega_i)| < 1$; it follows that

$$\prod_{i=2}^{n+m} (1 - \omega_i) \xrightarrow[n]{} 0$$

and hence that

$$\left[A_{n+m} \cdots A_2 A_1\right] \xrightarrow[n]{} \left[A_{n+m} \cdots A_2 P_1\right]$$

Now use induction: if the result holds for $(m-1)$ then

$$\left[A_{n+m} \cdots A_1\right] = \left[A_{n+m} \cdots A_m P_{m-1} \cdots P_1\right] + E_{n,m}$$

where $E_{n,m} \xrightarrow[n]{} 0$.

Therefore

$$\left[A_{n+m} \cdots A_1\right] = \left[A_{n+m} \cdots A_{m+1} (P_m - \omega_m \underline{e}\, \underline{p}^T) P_{m-1} \cdots P_1\right] + E_{n,m}$$

$$= \left[A_{n+m}\ldots A_{m+1}\; P_m\ldots . P_1\right]$$

$$- \omega_m \prod_{i=m+1}^{n+m} (1 - \omega_i)\, \underline{e}\, \underline{p}^T \left[P_{m-1}\ldots . P_1\right]$$

$$+ \quad E_{n,m}$$

$$= \left[A_{n+m}\ldots . A_{m+1}\; P_m\ldots . P_1\right] + E'_{m,n} + E_{n,m}$$

and $E'_{m,n} \xrightarrow[n]{} 0$, since $\left\| \underline{e}\, \underline{p}^T\, P_{m-1}\ldots P_1 \right\| \leqslant 1.$

Applying the lemma to the matrix product in (4.120) we get

$$\left[\tilde{A}^{n+m}\ldots . \tilde{A}^1\right] \xrightarrow[n]{} \left[\tilde{A}^{n+m}\ldots . \tilde{A}^{m+1}\; \tilde{P}^m\ldots . \tilde{P}^1\right]$$

But each $\tilde{P}^i$ is a normed, regular stochastic matrix and so the weak ergodicity property holds : that is

$$\tilde{P}^m\ldots . \tilde{P}^1 \xrightarrow[m]{} \underline{e}\, \underline{\pi}^T_{(m)}$$

So

$$\left[\tilde{A}^{n+m}\ldots \tilde{A}^{m+1}\; \tilde{P}^m\ldots \tilde{P}^1\right] \xrightarrow[m]{} \left[\tilde{A}^{n+m}\ldots \tilde{A}^{m+1}\right] \underline{e}\, \underline{\pi}^T_{(m)}$$

$$= \prod_{i=m+1}^{n+m} (1 - \omega_i)\, \underline{e}\, \underline{\pi}^T_{(m)}$$

$$\xrightarrow[n]{} \quad 0$$

We have shown that

$$\lim_{n \to \infty} \left[\tilde{A}^n \ldots\ldots \tilde{A}^1\right] = 0 \qquad \ldots .(4.121)$$

and hence, from (4.120), that

$$\lim_{n \to \infty} \underline{\varepsilon}_n \geqslant \underline{o}$$

ie. $$\lim_{n \to \infty} \underline{\delta}_n \geqslant \underline{\delta} \qquad \ldots .(4.122)$$

Thus $\underline{\delta}_n$ is bounded below (component-wise) by $\underline{\delta}$ for all n sufficiently large.

Furthermore, reversing the argument leading to (4.120),

$$\underline{\delta}_n - \underline{\delta} \leqslant \omega_n A^{f^o}(\underline{\delta}_{n-1} - \underline{\delta}) + (1 - \omega_n)(\underline{\delta}_{n-1} - \underline{\delta})$$

ie. $$\underline{\varepsilon}_n \leqslant \left[I - \omega_n (I - A^{f^o})\right] \underline{\varepsilon}_{n-1} \qquad \text{....(4.123)}$$

Multiplying by the stationary distribution, $\underline{\pi}_o^T$, of $(P^*)^{f^o}$ we get

$$(\underline{\pi}_o^T \underline{\varepsilon}_n) \leqslant (\underline{\pi}_o^T \underline{\varepsilon}_{n-1}) - \omega_n(\underline{p}^T \underline{\varepsilon}_{n-1}) \qquad \text{....(4.124)}$$

But, from (4.122), $\underline{\varepsilon}_{n-1} \geqslant \underline{0}$ for all n sufficiently large; so since we are taking $\underline{p} = \left(\frac{1}{N}\right)\underline{e}$ we shall have $(\underline{p}^T \underline{\varepsilon}_{n-1}) > 0$ unless $\underline{\varepsilon}_{n-1} = \underline{0}$. Then, for all n sufficiently large, (4.124) gives

$$(\underline{\pi}_o^T \underline{\varepsilon}_n) < (\underline{\pi}_o^T \underline{\varepsilon}_{n-1})$$

which, since the error vectors are (for $n \to \infty$) non-negative, implies that $(\underline{\pi}_o^T \underline{\varepsilon}_n) \xrightarrow[n]{} 0$. This in turn, using (4.124) again, implies that

$$\lim_{n \to \infty} (\underline{p}^T \underline{\varepsilon}_{n-1}) \leqslant 0$$

But, by (4.122), $\underline{\varepsilon}_{n-1} \geqslant 0$ for sufficiently large n. Therefore, since $\underline{p} = \left(\frac{1}{N}\right) \underline{e} > \underline{o}$, it follows that

$$\lim_{n \to \infty} \underline{\varepsilon}_n = \underline{o}$$

ie. that $$\lim_{n \to \infty} \underline{\delta}_n = \underline{\delta}$$

Comments :

(i) We have shown that acceleration of the standard successive-approximations algorithm, by the use of over-relaxation with a variable acceleration factor, is feasible. The accelerated algorithm is defined by equations (4.107) - (4.110).

(ii) What we have not shown is that the accelerated (ASA) algorithm will always converge at least as rapidly as the standard (SA) algorithm. However, since the transformation matrices used in the ASA algorithm have, in general, smaller spectral radii than those used in the SA aıgorithm, it is reasonable to expect the ASA algorithm to be more rapidly convergent. This expectation is borne out in practice.

(iii) Although we have proved convergence only for the case $\underline{p} = \left(\frac{1}{N}\right)\underline{e}$, it should be clear from the proof that it is sufficient to have $\underline{p} > \underline{0}$. Furthermore, numerical experience suggests that even this condition may not be necessary, though we have been unable to dispense with it.

(iv) Any attempt to use over-relaxation with a constant relaxation factor fails, because in order to guarantee that the matrix

$$\left[I - \omega \left\{ I - (P^*)^f \right\} \right]$$

remains a normed stochastic matrix for all control laws f, we must use $\omega \leqslant 1 + p_o$, where

$$p_o = \underset{u}{\text{Min}} \ \underset{i}{\text{Min}} \ (p^*_{ii})^u$$

But it is easily shown that $p_o = 0$ and hence that $\omega \leqslant 1$. Thus we cannot use over-relaxation. (In fact, this statement is equivalent to the conclusion that we cannot take $K > 1$ in equation (3.10) when transforming the original semi-Markov regulation problem to the equivalent Markov regulation problem.)

CHAPTER 5

OPTIMAL REGULATION OF GENERALIZED BIRTH-DEATH PROCESSES

5.1 Introduction

A birth-death process is a Markov chain, usually with state space $\mathcal{X} = \mathbb{N}$, in which transitions are possible only between adjacent states. Thus if $\{x_t\}$ is a birth-death process with embedded chain $\{\overline{x}_t\}$, the transition probability matrix, P, of $\{\overline{x}_t\}$ is such that

$$\forall i \in \mathbb{N} : p_{ij} = o \ , \ |i - j| > 1 \qquad \text{....(5.1)}$$

The term originated from the use of such chains to model the dynamics of biological populations subject to randomly-occurring births and deaths. In discrete time (ie. when $\mathcal{T} = \mathbb{Z}_+$), a birth-death process is often called a random walk with a barrier. We shall use the term generalized birth-death process (GBDP) to denote a semi-Markov chain whose embedded chain has a transition probability matrix satisfying (5.1). By a controllable GBDP we shall mean a controllable semi-Markov chain which is a GBDP for every feasible control law.

Such processes are of considerable interest in queueing theory since they serve as useful models for a wide variety of queueing and congestion systems. In particular, so-called Markov/Markov queues (see, for example, Gross and Harris$^{(1974)}$) with state-dependent service rate are appropriately modelled by controllable GBDP's. As an example, consider an M/M/u/N queueing system (see Gross/Harris —— —(1974) for the standard nomenclature for classification of queues) in which the number, u_t, of open service channels at time t can be either 1 or 2. Suppose that the mean arrival rate is λ , the mean service rate per open channel is μ, and that $\mu > \lambda$. Then if $u_t = 1$ for all t the queue is stable in the sense that when $N = \infty$ (no upper bound to the permitted queue length) the queue length possesses a well-defined stationary distribution whose properties are parametrized by the so-called traffic intensity, $\rho \triangleq \lambda/\mu$. Let N_t denote the number of items in the system at time t and let $X_t \triangleq N_t + 1$. Then, as is well known (see, for example, Gross and Harris$^{(1974)}$, $\{X_t\}$ is a continuous-time Markov chain with state space $\mathcal{X} = \mathbb{N}_{N+1}$. In particular, if u_t is made the following function of the current state X_t :

$$\begin{cases} u_t = 1 \ , & X_t \leqslant i_o \\ = 2 \ , & X_t > i_o \end{cases} \qquad \ldots .(5.2)$$

then the embedded chain of $\{X_t\}$ will have the transition probability matrix : -

$$P = \begin{array}{c} 1 \\ 2 \\ \vdots \\ i_o \\ i_o+1 \\ \vdots \\ \vdots \\ N+1 \end{array} \left[\begin{array}{ccccccccc} 0 & 1 & & & & & & & \\ \mu_1 & 0 & \lambda_1 & & & & & & \\ & \mu_1 & 0 & \lambda_1 & & & & & \\ & & & \mu_1 \ddots & 0 & \lambda_1 & & & \\ \hline & & & & \mu_2 & 0 & \lambda_2 & & \\ & & & & & \mu_2 & 0 & \lambda_2 & \\ & & & & & & & \ddots & \\ & & & & & & & 1 & 0 \end{array}\right]$$

where $\lambda_1 = (\frac{\lambda}{\lambda+\mu})$; $\mu_1 = (\frac{\mu}{\lambda+\mu})$; $\lambda_2 = (\frac{\lambda}{\lambda+2\mu})$; $\mu_2 = (\frac{2\mu}{\lambda+2\mu})$.

Note that P has the tri-diagonal structure characteristic of birth-death processes. Note also that the function (5.2) is an example of a feasible control law for the controllable GBDP $\{X_t\}$ with control set $\mathcal{U} = \{1,2\}$. It is clear that an optimal regulation problem appears in a natural way if we associate with $\{X_t\}$ a state cost, representing the cost per unit time incurred by each queued item, plus a control cost, representing the cost per unit time of providing an open service channel. A discrete-time version of this optimal regulation problem has been studied by Brosh$^{(1970)}$, and the more general optimal regulation problem in which $\mathcal{U} = \{0,1,2,\ldots,K\}$ has been analysed in detail by Crabill$^{(1972)}$.

In this chapter, we first consider the possibility of simplifying the algorithms described in Chapters 3 and 4 when the chain to be optimized is a controllable GBDP. Next, we examine the problem of truncating the state space of an infinite-state GBDP so that optimization may be performed by one of the finite-state optimization algorithms of Chapters 3 and 4; this is an important consideration in the application of the algorithms to queueing systems since in many such systems the state space is infinite, ie. $\mathcal{X} = \mathbb{Z}_+$. Finally, we show how to determine globally-optimal quantized control laws for controllable GBDP's. In Chapter 6 we shall consider the application of our ideas to a specific optimal regulation problem.

5.2 Simplifications arising from the birth-death structure

Consider a controllable GBDP $\{(X_t : \Omega \rightarrow N_N) : t \in R_+\}$ with control set $\mathcal{U} = \{u^1,\ldots,u^k\}$. Under any feasible control law f, the transition probability matrix of the canonical embedded chain for $\{X_t\}$ has the form (assuming that no state is absorbing) :-

$$P^f = \begin{bmatrix} 0 & 1 & & & & & \\ 1-\alpha_2^f & 0 & \alpha_2^f & & & & \\ & 1-\alpha_3^f & 0 & \alpha_3^f & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1-\alpha_{N-1}^f & 0 & \alpha_{N-1}^f \\ & & & & 1 & 0 \end{bmatrix} \qquad \ldots.(5.3)$$

where

$$\alpha_i^f \triangleq p_{i,i+1}^{f(i)} \quad , \quad i = 2,3,\dots,N-1 \qquad \dots.(5.4)$$

In what follows it is assumed that $\{X_t\}$ is in canonical form (see Section 2.2.4) so that the transition probability matrices associated with $\{X_t\}$ will always be of the form (5.3).

The first simplification arising from the above structure is that it is easy to determine by inspection the states which are transient under f. Thus

(i) if $\alpha_i^f \in (0,1)$, $i = 2,3,\dots,N-1$, then all states intercommunicate and hence belong to a single recurrent class : P^f is regular and possesses a unique, strictly positive, stationary distribution;

(ii) if $\alpha_j^f = 0$ and all other $\alpha_i^f \in (0,1)$, then states $1,2,\dots,j$ are recurrent and the remaining states are transient;

(iii) if $\alpha_j^f = 1$ and all other $\alpha_i^f \in (0,1)$, then states $j,j+1,\dots,N$ are recurrent and the remaining states are transient;

(iv) if $\alpha_j^f = 0$, $\alpha_k^f = 1$ for some $k < j$, and all other $\alpha_i^f \in (0,1)$, then states $j,j+1,\dots,k$ are recurrent and the remaining states are transient;

(v) if $\alpha_j^f = 0$, $\alpha_k^f = 1$ for some $k > j$, and all other $\alpha_i^f \in (0,1)$, then the states $1,2,\dots,j$ form one recurrent class, states $k,k+1,\dots,N$ form a second recurrent class, and the states $j+1, j+2,\dots,k-1$ are transient : P^f is no longer regular.

In this chapter we continue to restrict our attention to totally regular chains, that is, chains which are regular under every feasible control law f. If case (ii), (iii) or (iv) holds for every feasible f in a given optimal regulation problem, the state space can be reduced to the set $\mathcal{X}_R$ of states recurrent under all f.

A second simplification is that the stationary distribution, $\underline{\pi}^f$, of P^f is very easily determined. In fact, introducing $\alpha_1^f \triangleq 1$,

$\alpha_N^f \triangleq 0$, the components of $\underline{\pi}^f$ are related by

$$\pi_{i+1}^f = \left(\frac{\alpha_i^f}{1 - \alpha_{i+1}^f}\right) \pi_i^f \quad , \quad i = 1,2,\ldots,N-1 \qquad \ldots.(5.5)$$

so that if f is changed only in state i_o, the ratios (π_{i+1}^f/π_i^f) will be unchanged except at $i = i_o - 1$ and $i = i_o$. As we shall see, it is this fact which permits state quantization to be introduced without destroying the convexity of the optimal regulation problem.

When we come to consider the effect of the special birth-death structure on the performance of the optimization algorithms of Chapters 3 and 4, we note that P^f is a sparse matrix in which the only independent parameters are $\alpha_2^f, \alpha_3^f, \ldots, \alpha_{N-1}^f$. Thus instead of storing P^f as a 2-dimensional array of size N^2 it is only necessary to store a 1-dimensional array of size $(N-2)$. Briefly, the behaviour of the chain under the control law f is characterized by (i) the parameter vector $\underline{\alpha}^f \triangleq (\alpha_2^f, \ldots, \alpha_{N-1}^f)$, (ii) the mean sojourn time vector, $\underline{\tau}^f$, and (iii) the vector, $\underline{\gamma}^f$, of mean one-step costs.

In the Howard policy-iteration algorithm and its variants (Section 3.3.1) it is necessary to solve a set of N simultaneous linear equations once per iteration. As we have seen, this is normally done by Gaussian elimination, involving an operation count of approximately $\frac{N^3}{3}$. When P^f is tri-diagonal the equations can be solved much more efficiently by using the following procedure :

(1) Determine the stationary distribution $\underline{\pi}^f$ using equations (5.5) and the usual normalizing condition $(\underline{\pi}^f)^T \underline{e} = 1$. The operation count is approximately $2N_R$, where N_R is the number of recurrent states.

(2) Evaluate the corresponding cost rate $\bar{c}$ by equation (3.6). The operation count is again approximately $2N_R$.

(3) Solve the system

$$\underline{\delta} = \underline{\gamma}^f - \bar{c}\,\underline{\tau}^f + P^f\,\underline{\delta}$$

in the form

$$(I - P^f)\,\underline{\delta} = \underline{\gamma}^f - \bar{c}\,\underline{\tau}^f$$

where now the right-hand side is a known vector. (As usual, to obtain a unique $\underline{\delta}$ we set, for example, $\delta_N = 0$). The coefficient matrix $(I - P^f)$ is tri-diagonal and so the system can be solved by the so-called Thomas algorithm (see, for example, Williams(1972)), for which the operation count is approximately $5N$.

The operation count for this procedure is thus $\leqslant 10N$, an enormous improvement (for large N) on the figure of $\frac{N^3}{3}$ for the standard method using Gaussian elimination.

On the other hand very little simplification is possible when the modified policy-iteration algorithms of Chapter 4 are applied to a birth-death optimization problem. The reason is that the methods are based on the use of the inverse matrix, $E \triangleq \left[I - P^f + \underline{e}\,\underline{p}^T\right]$, in which the tri-diagonal structure of P^f is completely hidden. For example, in the DPI algorithm the most efficient way of implementing the calculation of $\Delta\bar{c}$ via equations (4.54) and (4.55) is, instead of using the E-matrix as in (4.56), to compute the new $\underline{w}$ vector directly via steps (1), (2) and (3) listed above for the Howard/Jewell algorithm. (This is possible since $\underline{w}$ is the unique value vector $\underline{\delta}$ satisfying $\underline{p}^T\,\underline{\delta} = \bar{c}$.) The operation count for the DPI algorithm is thus $\leqslant 10N$ for each single-step policy improvement, compared with N^2 operations per step in the general case.

Because the value vector is updated only once per optimization cycle in the Howard/Jewell algorithm, as against once per single-step policy improvement in the DPI algorithm, the latter can actually be less efficient than the former when the P-matrix is tri-diagonal.

In the various successive-approximations algorithms the tri-diagonal structure of P^f confers benefits in both speed and storage requirements. The speed improvement results from the reduction in the number of multiplications required for the product $P\,\underline{\delta}_{n-1}$ from N^2 in the general case to $3N$ in the tri-diagonal case.

5.3 Truncation of the state space for the $M/M/k/\infty$ queue

In many applications of queueing theory there is no restriction on the length of the queue that may form at the entrance to the service facility. Thus if a system is to be modelled by, say, an $M/M/u/N$ queueing model, it is necessary to set $N = \infty$. Then if the number of items in the system at time t is N_t, the process $\{N_t\}$ is a continuous-time Markov chain with an infinite state space. This presents a major difficulty in that optimization of such a system by any of the algorithms of Chapters 3 and 4 is not possible unless the state space can be reduced to a finite set such as $\mathbb{N}_N$. The usual way round this difficulty is to approximate the $M/M/k/\infty$ model by the corresponding $M/M/k/N$ model for some sufficiently large N; that is, to use a model in which the arrival rate to the system drops to zero whenever the number of items in the system exceeds N. If N is such that under all feasible control laws $P\left[N_t > N\right] \simeq 0$ then the behaviour of the two models should be almost identical – a hypothesis that can be checked by varying N.

As we shall now show, an alternative approach is possible in which the $M/M/k/\infty$ model is replaced by an exactly equivalent controllable finite-state GBDP. The basic idea is to introduce an embedded semi-Markov process in which there is a reflecting barrier at

$i = N$ as well as at $i = 0$.

5.3.1 The controllable M/M/k/∞ queue

The basic system with which we are concerned is the following controllable queueing system, denoted by CQS 1 .

CQS 1 : | Single Poisson arrival stream : arrival rate, λ
Exponential service channels : service rate, μ
Number of open service channels, $k \in \mathcal{U} = \{0,1,2,\dots,K\}$
Queue discipline : FIFO (ie. arrival order)
System capacity, $N = \infty$

Let N_t be the number of items in the system at time t ; then $\{N_t\}$ is a continuous-time Markov chain with state space $\mathcal{X} = \mathbb{Z}_+$. Associate with $\{N_t\}$ a set of expected one-step costs γ_i^k , $i \in \mathcal{X}$, $k \in \mathcal{U}$, defined according to (3.2) for some specified cost function c; the costs γ_i^k are assumed to be non-negative and non-decreasing in i and k. The problem of finding an optimal control law $f^o: \mathcal{X} \to \mathcal{U}$ for the controllable birth-death process $\{N_t\}$ is then an optimal regulation problem with a countably infinite state space. The problem is properly posed, in the sense that an optimal control law exists for $\{N_t\}$, if the following conditions are satisfied (see Lippman$^{(1973)}$):

(i) For some $k \in \mathcal{U}$, $k\mu > \lambda$. (This condition ensures that there is sufficient service capacity to maintain stability.)

(ii) For some positive constant C and some positive integer m, $\gamma_i^k \leqslant C(i \vee 1)^m$, $\forall k \in \mathcal{U}$. (This polynomial bound on the one-step costs ensures that the mean cost rate $\bar{c}$ is finite when the system is properly regulated.)

5.3.2 The equivalent semi-Markov chain

Assume now that the above conditions hold, and consider the associated optimal regulation problem in which the set of allowable control laws is the restricted set

$$\mathcal{F}_N \triangleq \left\{ f : (\forall i \geqslant N : f(i) = k),\ k \geqslant k_* \right\} \qquad \ldots.(5.6)$$

where k_* is the minimum value of $k \in \mathcal{U}$ such that $k\mu > \lambda$.

For every $f \in \mathcal{F}_N$ the mean cost rate $\overline{c}^f$ is finite; and since $\mathcal{F}_N$ is finite the existence of $\underset{f \in \mathcal{F}_N}{\mathrm{Min}} (\overline{c}^f)$ is guaranteed. To determine a control law which is optimal over $\mathcal{F}_N$, we introduce the <u>embedded semi-Markov chain</u>, $\{M_t\}$, defined by setting

$$\begin{cases} M_t \triangleq N_t \quad , \quad N_t \leqslant N \\ \triangleq N \quad , \quad N_t > N \end{cases} \qquad \ldots.(5.7)$$

Thus $\{M_t\}$ is an $(N+1)$-state process in which the top state, $i = N$, is occupied whenever $N_t \geqslant N$. Now with $f \in \mathcal{F}_N$ the traffic intensity for $\{N_t\}$ is less than one whenever $N_t \geqslant N$; it follows (see, for example, Cinlar[(1975)]) that the state $i = N - 1$ is positive-recurrent and hence that the mean first-passage time from $i = N$ to $i = N - 1$ is finite. In turn, this implies that the transition $(M_t = N) \longrightarrow (M_t = N - 1)$ is certain and that the mean sojourn time in state N for the process $\{M_t\}$ is finite. Note however that $\{M_t\}$ is not a pure Markov process: in states $0,1,2,\ldots,N-1$ the behaviour of $\{M_t\}$ is identical to that of $\{N_t\}$ and the sojourn time distributions are therefore exponential, but in state N this is no longer the case.

As usual the equilibrium properties of the semi-Markov chain $\{M_t\}$ are characterized by (i) the transition probabilities $\overline{p}_{ij}^k$ of the embedded Markov chain, $\{\overline{M}_n\}$; (ii) the mean sojourn times, $\overline{\tau}_i^k$; and (iii) the mean one-step costs, $\overline{\gamma}_i^k$. We now determine these quantities.

(i) <u>Transition probabilities, $\overline{p}_{ij}^k$</u>

For $0 < i < N$, we have, by (5.7),

$$\begin{aligned}
\bar{p}_{ij}^{k} &\triangleq P\left[\bar{M}_{n+1} = j \;\middle|\; \bar{M}_n = i \, , \, k_n = k\right] \\
&= \left(\frac{\mu k}{\lambda + \mu k}\right) , \qquad j = i - 1 \\
&= \left(\frac{\lambda}{\lambda + \mu k}\right) , \qquad j = i + 1 \\
&= \quad 0 \qquad\quad , \qquad \text{otherwise}
\end{aligned}$$

Also, the transitions $(\bar{M}_n = 0) \longrightarrow (\bar{M}_{n+1} = 1)$ and $(\bar{M}_n = N) \longrightarrow (\bar{M}_{n+1} = N - 1)$ are certain. Thus under any feasible control law f, the transition probability matrix $\bar{P}^f \triangleq \left[p_{ij}^{f(i)}\right]$ has the form

$$\bar{P}^f = \left[\begin{array}{ccccccc}
0 & 1 & 0 & & & & \\
\mu_1 & 0 & \lambda_1 & & & & \\
 & \mu_2 & 0 & \lambda_2 & & & \\
 & & \mu_3 & 0 & \lambda_3 & & \\
 & & & & \ddots & & \\
 & & & \mu_{N-1} & 0 & \lambda_{N-1} \\
 & & & 0 & 1 & 0
\end{array}\right] \qquad \ldots.(5.8)$$

where $\lambda_i \triangleq \left(\frac{\lambda}{\lambda + \mu k_i}\right)$, $\mu_i \triangleq \left(\frac{\mu k_i}{\lambda + \mu k_i}\right)$, and $k_i \triangleq f(i)$.

(ii) <u>Mean sojourn times, $\bar{\tau}_i^k$</u>

Consider first the original process $\{N_t\}$, whose sojourn time distributions are all exponential. In state 0 the event counting process is the Poisson arrival process which has a mean rate of λ. In any other state i the event counting process is the total Poisson process generated by arrivals and departures: this has a mean rate $(\lambda + \mu k_i)$ where $k_i = f(i)$.

Thus, using (5.7),

$$\begin{aligned} \bar{\tau}_i^k &= \left(\frac{1}{\lambda}\right) , \quad i = 0 \\ &= \left(\frac{1}{\lambda + \mu k}\right) , \quad 0 < i < N \end{aligned} \qquad \text{....(5.9)}$$

Now for all $i \geqslant N$ and any $f \in \mathcal{F}_N$, let

$$\text{and} \quad \begin{cases} \tau_i^f = \text{Mean sojourn time of } \{N_t\} \text{ in state } i \\ \bar{\tau}_i^f = \text{Mean first passage time from } i \text{ to } i-1 \\ \qquad\qquad \text{for the process } \{N_t\} \end{cases}$$

Then, by a renewal argument,

$$\bar{\tau}_i^f = \tau_i^f + p_{i,i+1}^f \left(\bar{\tau}_{i+1}^f + \bar{\tau}_i^f \right) \qquad \text{....(5.10)}$$

since on exit from state i the process $\{N_t\}$ either enters state $i - 1$, in which case $\bar{\tau}_i^f = \tau_i^f$, or (with probability $p_{i,i+1}^f$) it enters state $i + 1$, in which case a further time $\bar{\tau}_{i+1}^f$ is required for the passage to state i, followed by a still further time $\bar{\tau}_i^f$ for the subsequent passage to state $i - 1$.

Since $f(i) = \text{constant} = f(N)$ for all $i \geqslant N$ the process $\{N_t\}$ is homogeneous in $i(\geqslant N)$ and so $\bar{\tau}_{i+1}^f = \bar{\tau}_i^f$. Using this fact in (5.10) we obtain

$$\bar{\tau}_i^f = \frac{\tau_i^f}{1 - 2 p_{i,i+1}^f} \qquad \text{....(5.11)}$$

But, with $f(N) = k_N$, we have

$$\tau_i^f = \left(\frac{1}{\lambda + \mu k_N}\right)$$

$$p_{i,i+1}^f = \left(\frac{\lambda}{\lambda + \mu k_N}\right)$$

and so, for all $i \geqslant N$,

$$\overline{\tau}_i^f = \left(\frac{1}{\mu k_N - \lambda}\right) \qquad \text{....(5.12)}$$

But $\overline{\tau}_N^f$ is just the mean sojourn time $\overline{\tau}_N^{k_N}$ of the process $\{M_t\}$ in state N.

Thus, using (5.9) and (5.12) ,

$$\begin{cases} \overline{\tau}_i^k = \left(\frac{1}{\lambda}\right) \quad , \quad i = o \\ \quad\;\; = \left(\frac{1}{\lambda + \mu k}\right) \; , \; 0 < i < N \\ \overline{\tau}_N^{k_N} = \left(\frac{1}{\mu k_N - \lambda}\right) \end{cases} \qquad \text{....(5.13)}$$

provided that $k_N > \frac{\lambda}{\mu}$, so that $f \in \mathcal{F}_N$.

(iii) <u>Mean one-step costs, $\overline{\gamma}_i^k$</u>

If, for all $i \geqslant N$,

and $\begin{cases} \gamma_i^f = \text{Mean one-step cost in state } i \text{ of } N_t \\ \overline{\gamma}_i^f = \text{Mean cost accumulated by } \{N_t\} \text{ in the passage from } i \text{ to } i-1 \end{cases}$

then, by an analogue of the above argument,

$$\overline{\gamma}_i^f = \gamma_i^f + p_{i,i+1}^f \left(\overline{\gamma}_{i+1}^f + \overline{\gamma}_i^f\right) \qquad \text{....(5.14)}$$

It is not now possible to argue that $\overline{\gamma}_{i+1}^f = \overline{\gamma}_i^f$ since, in general, γ_i^f is not independent of i. However, since $p_{i,i+1}^f = \left(\frac{\lambda}{\lambda + \mu k_N}\right)$ for all $i \geqslant N$, we can write (5.14) as

$$(-\lambda_N)\overline{\gamma}_{i+1}^f + (1 - \lambda_N)\overline{\gamma}_i^f = \gamma_i^f \qquad \text{....(5.15)}$$

where $\lambda_N \triangleq \left(\frac{\lambda}{\lambda + \mu k_N}\right)$. This is a linear, constant coefficient, first-order difference equation for $\overline{\gamma}_i^f$ whose solution can be determined once γ_i^f is specified.

For example, suppose that γ_i^k is the sum of a linear state cost and a general control cost :

$$\gamma_i^k = (C\,i + D_k)\,\tau_i^k \qquad \text{....(5.16)}$$

Then, since $\tau_i^k = \tau_i^{k_N} = \left(\dfrac{1}{\lambda + \mu k_N}\right)$ for all i $\geqslant$ N,

$$\gamma_i^f = \left(\frac{1}{\lambda + \mu k_N}\right)(C\,i + D_{k_N}) \qquad \text{....(5.17)}$$

With γ_i^f given by (5.17) the solution of (5.15) is

$$\overline{\gamma}_i^f = \left(\frac{1}{\lambda + \mu k_N}\right)\left[\frac{C\,i + D_{k_N}}{(1 - 2\lambda_N)} + \frac{C\,\lambda_N}{(1 - 2\lambda_N)^2}\right] + A\left(\frac{1 - \lambda_N}{\lambda_N}\right)^i \qquad \text{....(5.18)}$$

where A is a constant.

But any solution to (5.15) must satisfy the condition

$$\lim_{\lambda_N \to 0} \overline{\gamma}_i^f = \gamma_i^f$$

and so, in (5.18), A = 0.

Finally, substituting for λ_N and taking i = N, the mean one-step cost in state N for the process $\{M_t\}$ is given by

$$\overline{\gamma}_N^{k_N} = \overline{\gamma}_N^f = \frac{CN + D_{k_N}}{\mu k_N - \lambda} + \frac{C\,\lambda}{(\mu k_N - \lambda)^2} \qquad \text{....(5.19)}$$

In general, since $M_t = N_t$ when $N_t \leqslant N$ the complete specification of one-step costs for $\{M_t\}$ is

$$\begin{cases} \overline{\gamma}_i^k = \gamma_i^k \quad , & i < N \\ \overline{\gamma}_N^{k_N} = \overline{\gamma}_N^f \end{cases} \qquad \text{....(5.20)}$$

where $\overline{\gamma}_N^f$ is the solution of the equation (5.15) at the point $i = N$.

For any control law $f \in \mathcal{F}_N$, the $(N+1)$-state semi-Markov chain $\{M_t\}$ will have the same mean cost rate $\overline{c}^f$ as the infinite-state Markov chain $\{N_t\}$; and so optimization of the performance of $\{N_t\}$ over the restricted class $\mathcal{F}_N$ is equivalent to optimization of the performance of $\{M_t\}$.

<u>Comments</u> :

(i) We have shown that it is possible to truncate the state space of the M/M/k/∞ queueing system in a way which properly incorporates the contributions to the cost rate generated in states above the truncation level N. Since no approximations are made, the cost rate $\overline{c}$ computed from the truncated model is the true cost rate of the original system. This is in contrast with the standard method of truncation, in which the arrival rate λ is assumed to fall to zero whenever the number of items in the system reaches N.

(ii) Of course a control law which is optimal over the class $\mathcal{F}_N$ is not in general optimal over the class, $\mathcal{F} = \{f : \mathbb{Z}_+ \to \mathcal{U}\}$, of all possible control laws for $\{N_t\}$. For example, if N = 1 the embedded process $\{M_t\}$ will have only two states and the corresponding optimal control law will generate, at most, two different values. However, if the one-step costs γ_i^k associated with $\{N_t\}$ are such that, for all k, $\gamma_i^k \to \infty$ as $i \to \infty$, then for sufficiently large N optimality over $\mathcal{F}_N$ is equivalent to optimality over $\mathcal{F}$. This fact follows from a result due to Crabill(1972) which states that under the above condition on the γ_i^k the control law f^o which is optimal over $\mathcal{F}$ is what Crabill calls "simply connected": that is, there exist K states $(0 \leqslant)\ldots$ $i_1 \leqslant i_2 \leqslant \ldots i_{K-1} \leqslant i_K$, such that

$$\begin{aligned} f^o(i) &= 0 , && i < i_1 \\ &= 1 , && i_1 \leqslant i < i_2 \\ &\vdots && \\ &= K-1 , && i_{K-1} \leqslant i < i_K \\ &= K , && i \geqslant i_K \end{aligned}$$

Thus there is some state i_K in and above which it is optimal to use the control action K. If in our truncation procedure we choose $N \geqslant i_K$ it is clear that the resulting optimization over $\mathcal{F}_N$ will yield the above optimal control law f^o. Of course i_K is not known a priori and in practice a check must be made that the chosen value of N is sufficiently large – for example, by increasing N and re-optimizing.

(iii) The results of this section apply, with only trivial modification, to the slightly more general form of the CQS 1 system in which $\mathcal{U} = \{k^1, k^2, \ldots, k^K\}$ where $k^1, k^2, \ldots, k^K$ are real numbers such that $0 \leqslant k^1 < k^2 < \ldots < k^K$. More generally, the method of state truncation proposed here, based on the introduction of an appropriate finite-state semi-Markov chain, is applicable in principle to any infinite-state controllable GBDP for which the optimal regulation problem is properly posed.

5.4 Quantization of the state space

In the previous section we considered the problem of optimizing the performance of the M/M/k/∞ queue with respect to a set of control laws $\mathcal{F}_N$ in which the control action is constrained to be constant over a given subset $\{i : i \geqslant N\}$ of the state space. A natural extension is to consider those control laws in which the control action is constant on each of the subsets $\{i : i \geqslant N\}$ and $\{i : i < N\}$. More generally we may partition the state space of any GBDP into a finite number of subsets and consider only those control laws in which the control action is constant on each subset. Quantization of the state space generates such a partition and in this section we show how state quantization may be handled. The basic idea is again to use an embedded chain with an appropriately defined state space.

5.4.1 The embedded semi-Markov chain for a partitioned state space

Consider for the moment a general semi-Markov process $\{X_t\}$ with state space $\mathcal{X} = \mathbb{N}_N$. Let $\{S_i : i = 1,2,\ldots,M\}$ be a partition of $\mathcal{X}$ such that S_1 contains N_1 states, S_2 contains N_2 states, etc., and, if necessary, re-label the states so that

$$\begin{aligned} S_1 &= \{1,2,\ldots,N_1\} \\ S_2 &= \{N_1+1,\ N_1+2,\ldots,N_1+N_2\} \\ &\text{etc.} \end{aligned}$$

For any two states in $\mathcal{X}$ write $i \equiv j$ if i and j belong to the same subset S, and $i \not\equiv j$ otherwise.

Denote the transition probability matrix, the mean sojourn time vector, and the mean one-step cost vector, of $\{X_t\}$ by P, $\underline{\tau}$ and $\underline{\gamma}$, respectively. Then, provided that P is regular, the equilibrium mean cost rate of the process $\{X_t\}$ is

$$\bar{c}_x = \frac{\underline{\pi}^T \underline{\gamma}}{\underline{\pi}^T \underline{\tau}}$$

where $\underline{\pi}^T$ is the stationary distribution of P.

We now introduce an embedded semi-Markov chain $\{Y_t\}$, related to

$\{X_t\}$ in the following way. Let $\{\overline{X}_n, T_n\}$ and $\{\overline{Y}_n, U_n\}$ be the Markov renewal processes (see Section 2.2.3) underlying $\{X_t\}$ and $\{Y_t\}$ respectively, and for each positive integer i let n_i be the value of n for which $\overline{X}_n \not\equiv \overline{X}_{n-1}$ for the i^{th} time. Then $\{\overline{Y}_n, U_n\}$ is defined by

$$\begin{cases} U_i \triangleq 0 & , \quad i = 0 \\ \quad\; \triangleq T_{n_i} & , \quad i > 0 \end{cases} \qquad \ldots.(5.21)$$

$$\begin{cases} \overline{Y}_i \triangleq \overline{X}_0 & , \quad i = 0 \\ \quad\; \triangleq \overline{X}_{n_i} & , \quad i > 0 \end{cases} \qquad \ldots.(5.22)$$

and as usual the semi-Markov process $\{Y_t\}$ is related to $\{\overline{Y}_n, U_n\}$ by

$$Y_t \triangleq \overline{Y}_{\overline{N}_t} \qquad \ldots.(5.23)$$

where $\{\overline{N}_t\}$ is the renewal counting process for $\{\overline{Y}_n, U_n\}$.

The process $\{Y_t\}$ changes state only when $\{X_t\}$ changes subset, and the new state of $\{Y_t\}$ is the state at which $\{X_t\}$ enters the new subset. The relationship between $\{X_t\}$ and $\{Y_t\}$ is illustrated in Fig.(9)

We now consider the properties of the embedded chain $\{Y_t\}$. Note first that $\{Y_t\}$ has the same state space as $\{X_t\}$, since $\overline{Y}_0 = \overline{X}_0$. Let $R \triangleq \left[r_{ij}\right]_{N \times N}$ be the transition probability matrix of the process $\{Y_t\}$ (strictly, of $\{\overline{Y}_n\}$). Then immediately

$$\begin{aligned} r_{ij} &= P\left[\overline{Y}_{n+1} = j \,\middle|\, \overline{Y}_n = i\right] \\ &= 0 \quad , \quad \text{if } j \equiv i \end{aligned} \qquad \ldots.(5.24)$$

and, by a simple renewal argument,

$$r_{ij} = p_{ij} + \sum_{k \equiv i} p_{ik}\, r_{kj} \quad , \quad \text{if } j \not\equiv i \qquad \ldots.(5.25)$$

since the Y-transition i → j involves either the single X-transition

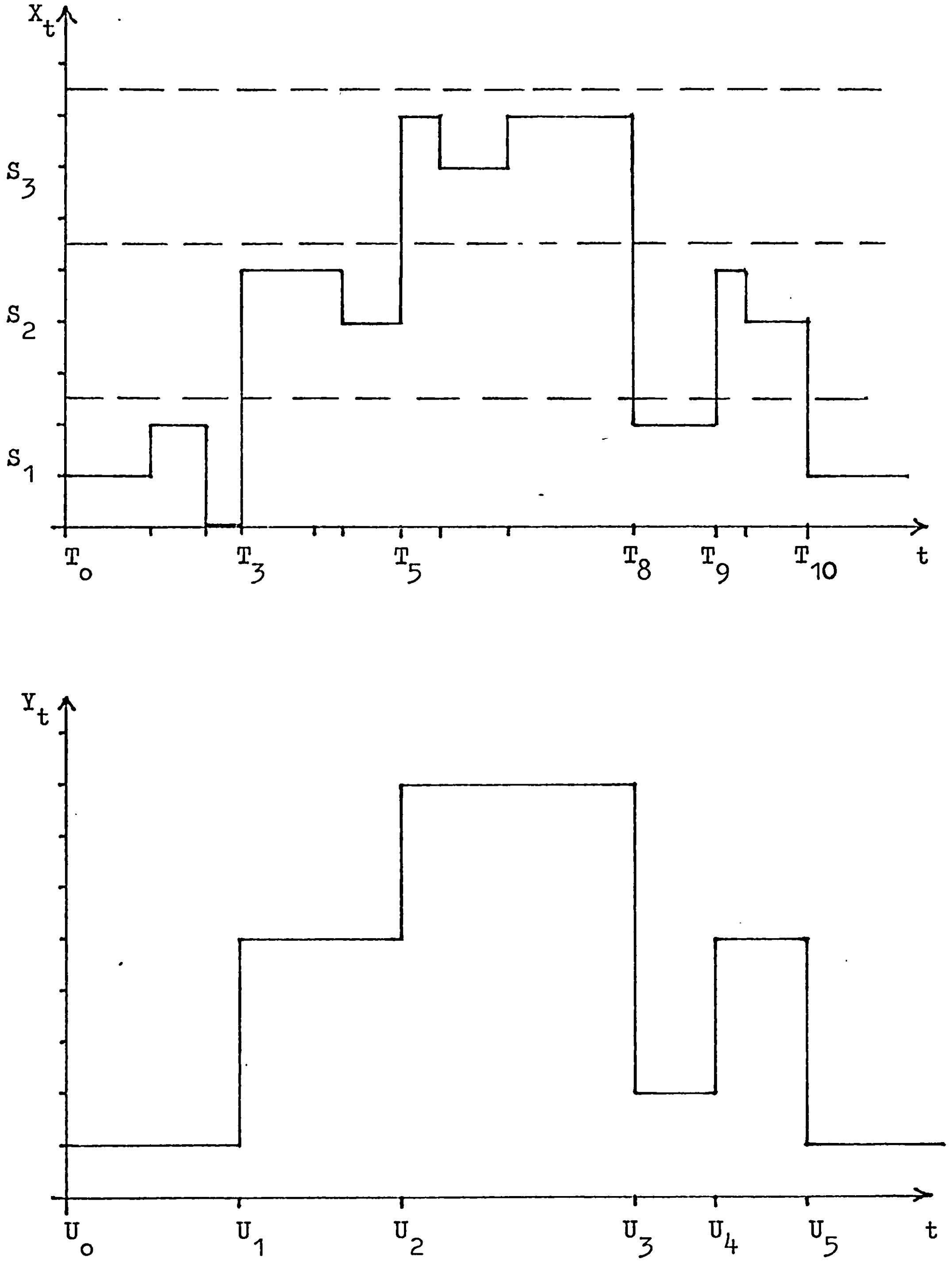


Fig. (9) Relationship between $\{X_t\}$ and $\{Y_t\}$

$i \rightarrow j$ or an X-transition $i \rightarrow k \equiv i$ followed by one or more further transitions.

Now define

$$\begin{aligned} q_{ij} &\triangleq p_{ij} \quad , \quad j \equiv i \\ &\triangleq 0 \quad , \quad j \not\equiv i \end{aligned} \qquad \text{....(5.26)}$$

Then (5.24) and (5.25) can be combined in the form

$$r_{ij} = (p_{ij} - q_{ij}) + \sum_{k \in \mathcal{X}} q_{ik} \, r_{kj} \quad , \quad \forall i, j \in \mathcal{X} \qquad \text{....(5.27)}$$

ie.

$$(I - Q)\, R = P - Q \qquad \text{....(5.28)}$$

where $Q \triangleq \left[q_{ij}\right]_{N \times N}$. Note that the matrix Q has a block diagonal structure in which the block sizes correspond to the sizes of the subsets $S_1, S_2, \ldots$ of $\mathcal{X}$.

We are assuming that $\{X_t\}$ is regular, that is, that it possesses a single recurrent subchain, with state set $\mathcal{X}_R$, say. Provided that the partition $\{S_i : i = 1,2,\ldots,M\}$ is such that $\mathcal{X}_R$ is not a subset of any single S_i the matrix $I - Q$ will be non-singular. Equation (5.28) can then be written

$$\boxed{R = (I - Q)^{-1} (P - Q)} \qquad \text{....(5.29)}$$

Properties of R

(i) It is easily verified that R is a stochastic matrix. We have

(a) $$P - Q \geqslant 0$$

and

$$(I - Q)^{-1} = \sum_{i=o}^{\infty} Q^i \geqslant 0$$

so that $$R \geqslant 0$$

(b) $$\begin{aligned} R\,\underline{e} &= (I - Q)^{-1} (P - Q)\,\underline{e} \\ &= (I - Q)^{-1} (I - Q)\,\underline{e} \quad , \quad \text{since } P\,\underline{e} = \underline{e} \end{aligned}$$

so that $$R\,\underline{e} = \underline{e}$$

(ii) Furthermore if P is regular, with stationary distribution $\underline{\pi}$, we have

$$\underline{\pi}^T = \underline{\pi}^T P$$

Subtracting $\underline{\pi}^T Q$ from each side :

$$\underline{\pi}^T(I - Q) = \underline{\pi}^T(P - Q)$$

ie. $$\underline{\pi}_R^T = \underline{\pi}_R^T R$$

where

$$\underline{\pi}_R^T = k\,\underline{\pi}^T(I - Q) \qquad \text{....(5.30)}$$

Since $\underline{\pi}^T(I - Q) = \underline{\pi}^T(P - Q) \geqslant \underline{0}^T$, the vector $\underline{\pi}_R^T$ is a probability distribution if k is suitably chosen. Thus R possesses the unique stationary distribution $\underline{\pi}_R^T$ (unique because $\underline{\pi}^T$ is unique and (I - Q) is non-singular), and hence $\{\bar{Y}_n\}$ possesses only one recurrent subchain. R is therefore regular or periodic; but the latter possibility may be ignored since R is equivalent (in the sense of Section 2.2.4) to the matrix $R = I - \varepsilon(I - R)$ which, for $\varepsilon \in (0,1)$, has a strictly positive diagonal (and cannot therefore be periodic). So regularity of P implies regularity of R.

As has been pointed out by Smith(1971), although the state space of $\{Y_t\}$ is the same as that of $\{X_t\}$, the set of recurrent states of $\{Y_t\}$ will consist of only those recurrent states of $\{X_t\}$ which are accessible in one step from states in a different subset. Call a state j an <u>entry state</u> for the subset S if $p_{ij} > 0$ for at least one $i \not\leq j$. Then the recurrent states of $\{Y_t\}$ are the recurrent entry states of $\{X_t\}$. Use is made of this fact in formulating the optimal regulation problem for birth-death processes with state quantization.

<u>Cost rate of the embedded chain $\{Y_t\}$</u>

Denote by $\tilde{\tau}_i$ the mean sojourn time in state i of the semi-Markov chain $\{Y_t\}$. Then, by applying the usual renewal argument to $\{X_t\}$,

$$\tilde{\tau}_i = \tau_i + \sum_{j \equiv i} p_{ij} \, \tilde{\tau}_j$$

$$= \tau_i + \sum_{j \in \mathcal{X}} q_{ij} \, \tilde{\tau}_j \qquad \text{....(5.31)}$$

ie.

$$\boxed{\underline{\tilde{\tau}} = (I - Q)^{-1} \underline{\tau}} \qquad \text{....(5.32)}$$

where $\underline{\tau}$ and $\underline{\tilde{\tau}}$ are the mean sojourn time vectors of $\{X_t\}$ and $\{Y_t\}$.

A precisely analogous argument shows that

$$\boxed{\underline{\tilde{\gamma}} = (I - Q)^{-1} \underline{\gamma}} \qquad \text{....(5.33)}$$

where $\underline{\gamma}$ and $\underline{\tilde{\gamma}}$ are the mean one-step cost vectors of $\{X_t\}$ and $\{Y_t\}$.

Then, assuming that $\{X_t\}$ and hence $\{Y_t\}$ are regular, the equilibrium mean cost rate $\bar{c}_y$ of $\{Y_t\}$ is given by

$$\bar{c}_y = \frac{\underline{\pi}_R^T \, \underline{\tilde{\gamma}}}{\underline{\pi}_R^T \, \underline{\tilde{\tau}}} \qquad \text{....(5.34)}$$

$$= \frac{k \, \underline{\pi}^T (I - Q)(I - Q)^{-1} \underline{\gamma}}{k \, \underline{\pi}^T (I - Q) \, (I - Q)^{-1} \underline{\tau}} \quad \text{, by (5.30),(5.32),(5.33)}$$

$$= \frac{\underline{\pi}^T \underline{\gamma}}{\underline{\pi}^T \underline{\tau}}$$

ie. $$\bar{c}_y = \bar{c}_x \qquad \text{....(5.35)}$$

Thus the cost rate $\bar{c}_x$ of the original chain $\{X_t\}$ can be computed as the cost rate $\bar{c}_y$ of the embedded chain $\{Y_t\}$, ie. by using (5.34) and then (5.35).

Optimal regulation of the embedded chain $\{Y_t\}$

If $\{X_t\}$ is a totally regular controllable chain then $\{X_t\}$ and hence $\{Y_t\}$ will be regular for every control law f which is feasible for both $\{X_t\}$ and $\{Y_t\}$. Note however that a control law for $\{Y_t\}$ will generate a sequence of control actions each of which remains constant while $\{Y_t\}$ remains constant, that is, while $\{X_t\}$ remains within a single state subset S; furthermore the control action while $\{X_t\} \in S$ will, in general, depend on the state at which S is entered and hence may vary from one S-occupation to the next. Thus a given control law for $\{Y_t\}$ does not in general induce an equivalent control law for $\{X_t\}$; nor can a general control law for $\{X_t\}$ be represented by an equivalent control law for $\{Y_t\}$. However, each control law for $\{X_t\}$ belonging to the set of quantized control laws, $\mathcal{F}_Q \triangleq \{f : (f(i) = f_\alpha, \forall i \in S_\alpha), \alpha = 1,2,\ldots,M\}$ is equivalent to a quantized control law for $\{Y_t\}$. For each such control law, the cost rate $\bar{c}_x$ can be determined by evaluating $\bar{c}_y$ and then using (5.35). In this way it is, in principle, possible to determine the optimal quantized control law for the controllable chain $\{X_t\}$.

It is of course possible to treat the optimization of $\bar{c}_y$, with respect to unrestricted (ie. unquantized) control laws for $\{Y_t\}$, as an optimal regulation problem in its own right. The resulting optimal control law will not in general, have a representation as an equivalent control law (ie. as a map from $\mathcal{X}$ to $\mathcal{U}$) for the process $\{X_t\}$.

5.4.2 The quantized birth-death process

Now consider the specific case when $\{X_t\}$ is a generalised birth-death process with state space $\mathcal{X} = N_N$. Let $\{S_i : i = 1,2,\ldots,M\}$ be a natural partition of $\mathcal{X}$ such that

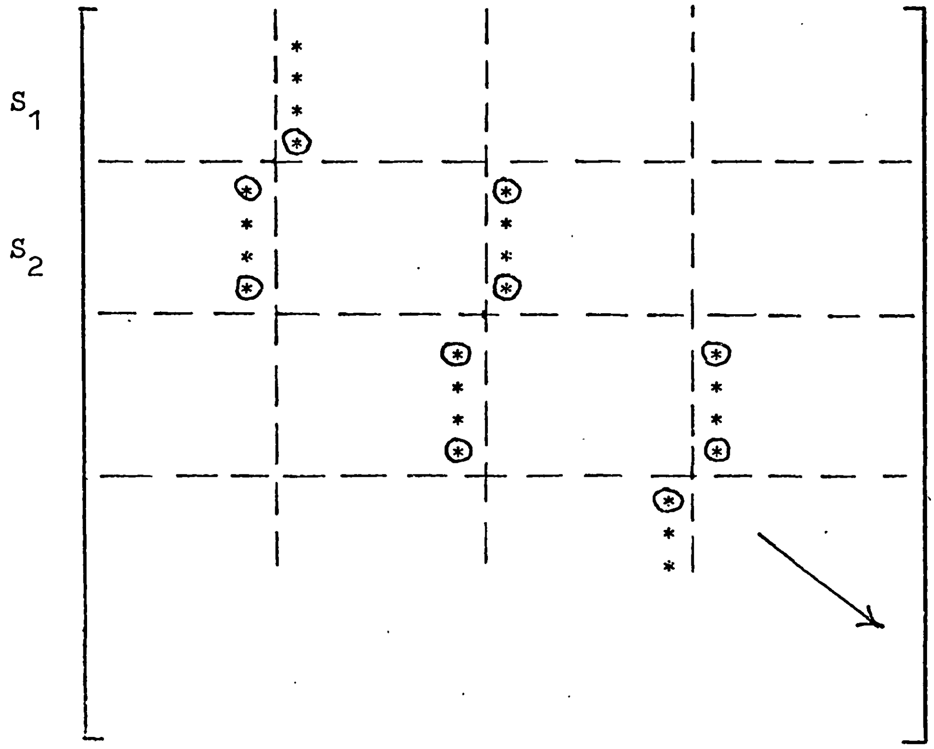
$$\begin{cases} S_1 = \{1, 2, \ldots, N_1\} \\ S_2 = \{N_1 + 1,\ N_1 + 2, \ldots, N_1 + N_2\} \\ \text{etc.} \end{cases}$$

Then, for any feasible f, P^f has the form given by (5.3) and transitions are possible only between adjacent states. It follows that the entry states of $\{X_t\}$ are :

Subset	Entry states
S_1	$i = N_1$
S_2	$i = N_1 + 1;\ i = N_1 + N_2$
S_3	$i = N_1 + N_2 + 1\ ;\ i = N_1 + N_2 + N_3$
$\vdots$	
S_M	$i = \sum_{1}^{M-1} N_\alpha + 1$

Thus if all the states of $\{X_t\}$ are recurrent the embedded process $\{Y_t\}$ defined by (5.23) possesses a total of $2M-2$ recurrent states, viz. the states listed above. Furthermore, using (5.29), the transition probability matrix of $\{Y_t\}$ has the form

R =



where * = non-zero transition probability (ringed for each recurrent state of $\{Y_t\}$).

Note incidentally that (I - Q) in (5.29) is of block diagonal form; hence R is easily determined by appropriate partitioning. Furthermore in the present case each block of (I - Q) is tri-diagonal and so use may be made of the Thomas algorithm in the calculation of R, $\underline{\tilde{\tau}}$ and $\underline{\tilde{\gamma}}$ via (5.29), (5.32) and (5.33).

Now if f is a quantized control law, ie. f $\in \mathcal{F}_Q$, we have seen that the cost rate $\bar{c}_x$ may be evaluated via (5.34) and (5.35), ie. by working with the embedded chain $\{Y_t\}$ and its cost rate $\bar{c}_y$. But the stationary distribution $\underline{\pi}_R$ in (5.34) is non-zero only over the (2M-2) recurrent states of $\{Y_t\}$: the transient states of $\{Y_t\}$ contribute nothing to $\bar{c}_y$ and may be jettisoned. (It follows that the only components of $\underline{\tilde{\tau}}$ and $\underline{\tilde{\gamma}}$ that need be evaluated are those associated with recurrent states of $\{Y_t\}$.)

Thus for any quantized control law the equilibrium mean cost rate of the N-state controllable GBDP $\{X_t\}$ is equal to the equilibrium mean cost rate of the (2M-2)-state embedded chain $\{Y_t\}$. The problem of determining the optimal quantized control law is reduced from an N-state regulation problem to a (2M-2)-state problem, where M is the number of quantum sets.

5.4.3 Convexity of the optimal regulation problem for quantized birth-death processes

We have seen (see Section 3.3.3) that the optimal regulation problem can be formulated as a linear programming problem. As is well known, any such problem is convex and so the objective function possesses no local minima apart from the global minimum (minima). It is this fact which ensures that the optimization methods of Chapters 3 and 4 will always converge to the globally minimal value of the cost rate $\bar{c}$. If however the problem is to determine the optimal quantized control law, an additional set of non-convex constraints must be satisfied; the problem is then no longer convex and the existing optimization algorithms are no longer guaranteed to find a globally optimal control law.

For the discrete-time Markov regulation problem the non-convex constraints arise as follows. With d_{ik} and x_{ik} defined by (3.41) and (3.45), quantized control requires that for each quantum set, S , we must have

$$d_{ik} = d_{jk} \quad , \quad \forall i \equiv j \; , \quad \forall k \qquad \text{....(5.36)}$$

ie.

$$\frac{x_{ik}}{\pi_i} = \frac{x_{jk}}{\pi_j} \quad , \quad \forall i \equiv j \; , \quad \forall k \qquad \text{....(5.37)}$$

ie. on using (3.50) ,

$$\frac{x_{ik}}{\sum\limits_k x_{ik}} = \frac{x_{jk}}{\sum\limits_k x_{jk}} \quad , \quad \forall i \equiv j \, , \quad \forall k \qquad \text{....(5.38)}$$

The equality constraints (5.38) are seen to be quadratic in the variables x_{ik} and hence non-convex. Minimization of the function $\overline{\alpha}^{d}$, defined by (3.46), subject to the constraints (5.38) is thus a non-convex programming problem.

However, as we shall now show, in the particular case when the controllable chain $\{X_t\}$ is a generalized birth-death process, quantization of the state space does not destroy the convexity of the problem and our one-step policy-iteration algorithms will yield a globally optimal control law.

Let $\{X_t\}$ be a controllable GBDP and let S_α , $\alpha = 1,2,\ldots,M$, be the quantum sets of states of $\{X_t\}$. Denote the set of feasible control laws for $\{X_t\}$ by $\mathfrak{F}$, and the set of feasible quantized control laws for $\{X_t\}$ by $\mathfrak{F}_Q$. Then for any two control laws f,f' we have, by equation (4.18) ,

$$\Delta\bar{c}(f',\check{f}) \;=\; \frac{1}{\bar{\tau}^{f'}}\left[(\underline{\pi}^{f'})^{T}\,\underline{\Delta\xi}\,(f',f)\right] \qquad \text{....(5.39)}$$

$$=\; K\sum_{i\,\in\,\mathcal{X}}\pi_i^{f'}\;\Delta\xi_i(f',f) \qquad \text{....(5.40)}$$

where $K = \left(\frac{1}{\bar{\tau}^{f'}}\right) > 0$.

But assuming that each quantum set is recurrent we can write, for each $i \in \mathcal{X}$,

$$\pi_i^{f'} \;=\; P_{\alpha}^{f'}\,.\,\pi_{i|\alpha}^{f'} \quad , \quad = 1,2,\ldots,M \qquad \text{....(5.41)}$$

where $P_{\alpha}^{f'}$ is the equilibrium probability under f' that $X_t \in S_{\alpha}$ and $\pi_{i|\alpha}^{f'}$ is the equilibrium conditional probability under f' that $X_t = i$ given $X_t \in S_{\alpha}$.

Thus (5.40) can be written

$$\Delta\bar{c}(f',f) \;=\; K\sum_{\alpha=1}^{M}P_{\alpha}^{f'}\sum_{i\in S_{\alpha}}\pi_{i|\alpha}^{f'}\;\Delta\xi_i(f',f)$$

$$=\; K\sum_{\alpha=1}^{M}P_{\alpha}^{f'}\;\overline{\Delta\xi_{\alpha}}^{f'} \qquad \text{....(5.42)}$$

where

$$\overline{\Delta\xi_{\alpha}}^{f'} \;\triangleq\; \sum_{i\in S_{\alpha}}\pi_{i|\alpha}^{f'}\;\Delta\xi_i(f',f) \qquad \text{....(5.43)}$$

We next show that the value of the averaged test quantity $\overline{\Delta\xi_{\alpha}}^{f'}$ depends only on control law changes in the quantum set S_{α} , ie. only on the restriction, f_{α}' , of f to S_{α} . Suppose that $S_{\alpha} = \{I,I+1,\ldots,\ldots J-1,J\}$ and consider the embedded chain $\{Z_t\}$ related to $\{X_t\}$ as follows:

$$\left\{\begin{array}{llll} Z_t & \triangleq I-1 & , & X_t < I \\ & \triangleq X_t & , & X_t \in S_\alpha \\ & \triangleq J+1 & , & X_t > J \end{array}\right.$$

Since $\{Z_t\}$ and $\{X_t\}$ have the same sample paths on the quantum set S_α they will have identical stationary conditional distributions on S_α. But the canonical transition probability matrix of $\{Z_t\}$ has the form when the current control law for $\{X_t\}$ is f :-

$$P_Z^f = \begin{array}{c|ccccccc|} I-1 & 0 & 1 & 0 & & & & \\ & 1-\alpha_I^f & 0 & \alpha_I^f & & & & \\ & & 1-\alpha_{I+1}^f & 0 & \alpha_{I+1}^f & & & \\ & & & & \ddots & & & \\ & & & & & 1-\alpha_J^f & 0 & \alpha_J^f \\ J+1 & & & & & 0 & 1 & 0 \end{array}$$

where, as usual, $\alpha_I^f \triangleq p_{I,I+1}^f = P\left[\overline{X}_{n+1} = I+1 \,\middle|\, \overline{X}_n = I \,;\, f(I)\right]$

The first row of P_Z^f is independent of f ; the second row depends on f(I) ; the third row on f(I+1) ; and so on. Thus P_Z^f depends only on f(I), f(I+1)....,f(J), ie. only on the restriction of f to S_α. But the stationary distribution $\underline{\pi}_Z$ of P_Z^f is uniquely determined by P_Z^f. It follows that $\underline{\pi}_Z^f$, and hence the stationary conditional distribution of $\{X_t\}$ on S_α, is determined by the restriction f_α. Thus (5.42) may be written

$$\Delta\bar{c}(f',f) = K \sum_{\alpha=1}^{M} p_\alpha^{f'} \, \overline{\Delta\xi_\alpha}^{f'_\alpha} \qquad \text{....(5.44)}$$

Now by an argument analogous to that following equation (4.19),

$$\text{f non-optimal} \Longrightarrow \exists\alpha\left[\exists f' : f'_\beta = f_\beta\, ,\ \forall\beta \neq \alpha\right] : \overline{\Delta\xi_\alpha}^{f'_\alpha} < 0 \qquad \text{....(5.45)}$$

from which it follows that

$$\forall\alpha\left[\forall f' \in \mathcal{F} : f'_\beta = f_\beta\ , \forall\beta \neq \alpha\right] : \overline{\Delta\xi_\alpha}^{f'_\alpha} \geqslant 0 \Longrightarrow f' \text{ optimal over } \mathcal{F} \qquad \text{....(5.46)}$$

If in (5.46) the choice of control law is restricted to the set of feasible quantized control laws, $\mathcal{F}_Q$, the statement remains true. Thus policy iteration algorithms based on the use of the test, quantities $\overline{\Delta\xi_\alpha}^{f'_\alpha}$ defined by (5.43) will always converge to a globally optimal quantized control law.

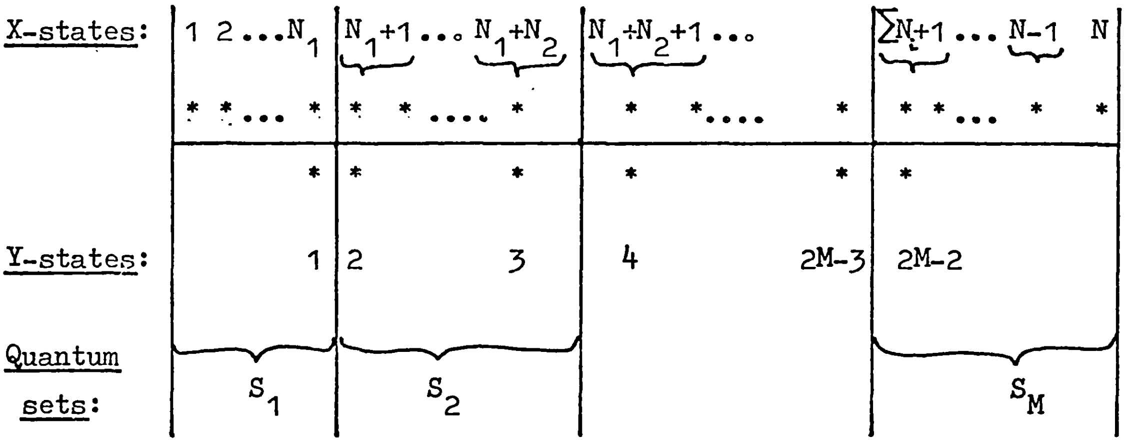
Comments

(i) We have shown that the single-state policy iteration algorithms described in Chapter 4 may be used in modified form to determine the optimal quantized control law for any controllable GBDP. The modification consists of changing the control law on one quantum set at a time instead of in one state at a time.

(ii) As we have seen, quantized control of any birth-death process $\{X_t\}$ is equivalent to quantized control of an associated process $\{Y_t\}$ with a smaller set of recurrent states. It is usually more efficient to work with the Y-process when carrying out the optimization, since the quantum sets of $\{Y_t\}$ contain, at most, two recurrent states.

(iii) Equations (5.42) and (5.43) show that evaluation of $\Delta\bar{c}(f',f)$ for any new quantized control law f' requires a knowledge of $\pi^{f'}_{i|\alpha}$, the conditional state probabilities under the new law f'. This means that the standard Howard/Jewell algorithm (see Section 3.3.1) cannot be used for optimization – nor, indeed, can the successive-approximations method (Section 3.3.2) or the linear programming method (Section 3.3.3). We must use policy-iteration with the control law f being updated one quantum set at a time.

(iv) The single-state policy-iteration algorithm most easily adapted to the quantized control problem is the direct policy-iteration (DPI) algorithm of Section 4.2.3. The essential modification is that, using the embedded Y-process, the control law must be changed in <u>both</u> Y-states in any quantum set simultaneously. Label the recurrent states of Y_t as follows : -



Then if f' differs from f_α only on the quantum set S_α , we shall have, with i = 2α - 2 ,

$$\Delta P - \underline{\Delta\tau}\ \underline{p}^T = \begin{bmatrix} \underline{e}_i & \underline{e}_{i+1} \end{bmatrix} \left\{ \begin{bmatrix} \underline{a}_i^T \\ \underline{a}_{i+1}^T \end{bmatrix} - \begin{bmatrix} \Delta\tau_i & 0 \\ 0 & \Delta\tau_{i+1} \end{bmatrix} \underline{p}^T \right\}$$

so that (4.48) and (4.49) must be replaced by

$$D_i = \begin{bmatrix} \underline{a}_i^T \\ \underline{a}_{i+1}^T \end{bmatrix} E - \begin{bmatrix} \Delta\tau_i & 0 \\ 0 & \Delta\tau_{i+1} \end{bmatrix} (\underline{\theta}^f)^T \qquad \ldots.(5.47)$$

and

$$E' = E\left[I + H_i \Lambda_i D_i \right] \qquad \ldots.(5.48)$$

where $H_i \triangleq \begin{bmatrix} \underline{e}_i & \underline{e}_{i+1} \end{bmatrix}$

and $\Lambda_i \triangleq \left[I - D_i H_i \right]^{-1}$ ....(5.49)

Equation (5.48) gives the required updating of the E-matrix when the controls are changed in the quantum set S_α. The corresponding change $\Delta\overline{c}$ is then determined via appropriately modified versions of (4.52) - (4.55). Note that each updating of E requires the inversion of the 2 x 2 matrix $\left[I - D_i \, H_i\right]$. The test quantities $\overline{\Delta\xi_\alpha}$ need not be determined explicitly.

(v) In order to determine an optimal quantized control law for the $M/M/k/\infty$ queueing system it is first necessary to truncate the state space to $\mathcal{N}_N$ by the method described in Section 5.3; the quantizing partition is then imposed on $\mathcal{N}_N$. A numerical example is given in Chapter 6.

CHAPTER 6

OPTIMAL REGULATION OF AN $M/M/k/\infty$ QUEUE WITH SWITCHING COSTS

6.1 Introduction

In this chapter we consider the application of some of the previous ideas to a specific optimal regulation problem. The problem in question is that of regulating the job mix of a computer whose load is a variable mixture of batch jobs and time-sharing jobs. Control is exerted by varying the proportion of central processor power allocated to the batch load; the optimization problem is to determine an on-line allocation algorithm which will minimize the long-run average value of some suitably defined operating cost per unit time. Under certain assumptions, the system can be modelled as a controllable birth-death process in which the state is the number of batch jobs currently in the system. The operating cost of the system has four components : (a) a state-dependent cost, representing the costs of delays to the batch job stream; (b) a control-dependent cost, representing the effect of reduced processor power on the time-sharing response time; (c) a second control-dependent cost, representing lost traffic due to saturation of the time-sharing system; (d) a switching cost, representing the adverse effect of time lost incurred whenever the processor power is re-allocated.

It is important to include the switching component (d) in the cost function, since the amount of production time lost due to frequent re-allocation of central processor resources can be significant. Unfortunately, however, it is not possible to define a separable (additive) cost function incorporating control switching costs unless the state space of the system is suitably enlarged. The redefined state space may be very much larger than the original state space, with a consequent substantial increase in the size of the optimal regulation problem when switching costs are included. It is therefore tempting to look for sub-optimal solutions in which the number of re-allocations per unit time (and hence the mean switching cost) is kept down to a reasonable level by using a quantized control law. Our aim in the work described in this chapter has been to see how far the performance of the system is improved (i) by the use of quantization as a method of reducing switching costs, and (ii) by the use of a control law which is optimal when the cost function incorporates the switching cost component. The approach we have used is as follows:

(A) With switching costs excluded, solve the optimal regulation problem with no state quantization. To the resulting optimal cost rate add the appropriate switching cost contribution and so obtain a figure for the overall (non-optimal) cost rate. (Section 6.4)

(B) Again with switching costs excluded, solve the optimal regulation problem with a "reasonable" choice of state quantization. As before, the resultant overall (sub-optimal) cost rate can be determined by adding in the appropriate switching cost contribution after the optimization. (Section 6.5)

(C) With switching costs included, solve the full optimal regulation problem by working with the appropriately re-defined state space. The resultant overall cost rate is optimal. (Section 6.6)

The results for the three approaches are compared and discussed in Section 6.7.

In the final part of this chapter we attempt to draw some general conclusions concerning the investigations described in this thesis, and also make some suggestions as to how the work might be extended.

6.2 Description of the system

The system under investigation is a large computer complex which provides (a) a batch processing service and,concurrently, (b) a time-sharing service to a large number of remote user terminals. Batch jobs are input to the computer system via a local batch job entry (BJE) terminal, or via a data communications link from another computer, or via the time-sharing system itself. Time-sharing jobs are submitted from the remote user terminals via a switching network and multiplexor. The computer itself is a multiprocessor system in which several central processors share a common memory. The number of central processors allocated to processing the batch job stream can be altered at any time; feedback control of this dynamic allocation process is implemented by means of a resource allocation controller whose input is the size of the batch queue and whose output is the currently-required division of resources between batch processing and the time-sharing system. The general scheme is as shown in Fig.(10).

The multiprocessor system It is assumed that the multiprocessor system consists of K identical processors sharing a common memory. Each processor can handle one job (batch or time-shared) at a time, so that when the system is fully loaded there are K jobs being processed at any one time. The number, k, of processors allocated to batch processing is the control variable in our optimal regulation problem.

The batch-processing load The batch jobs arriving to join the batch input queue are regarded as a single Poisson arrival stream and the processing times of the jobs are assumed to have a negative-exponential distribution. Jobs are processed in arrival order, ie. first-in, first out.

Arrival stream	:	Poisson, mean rate λ_B
Job processing times	:	Neg - exp., mean rate μ_B
Queue discipline	:	Arrival order (FIFO)

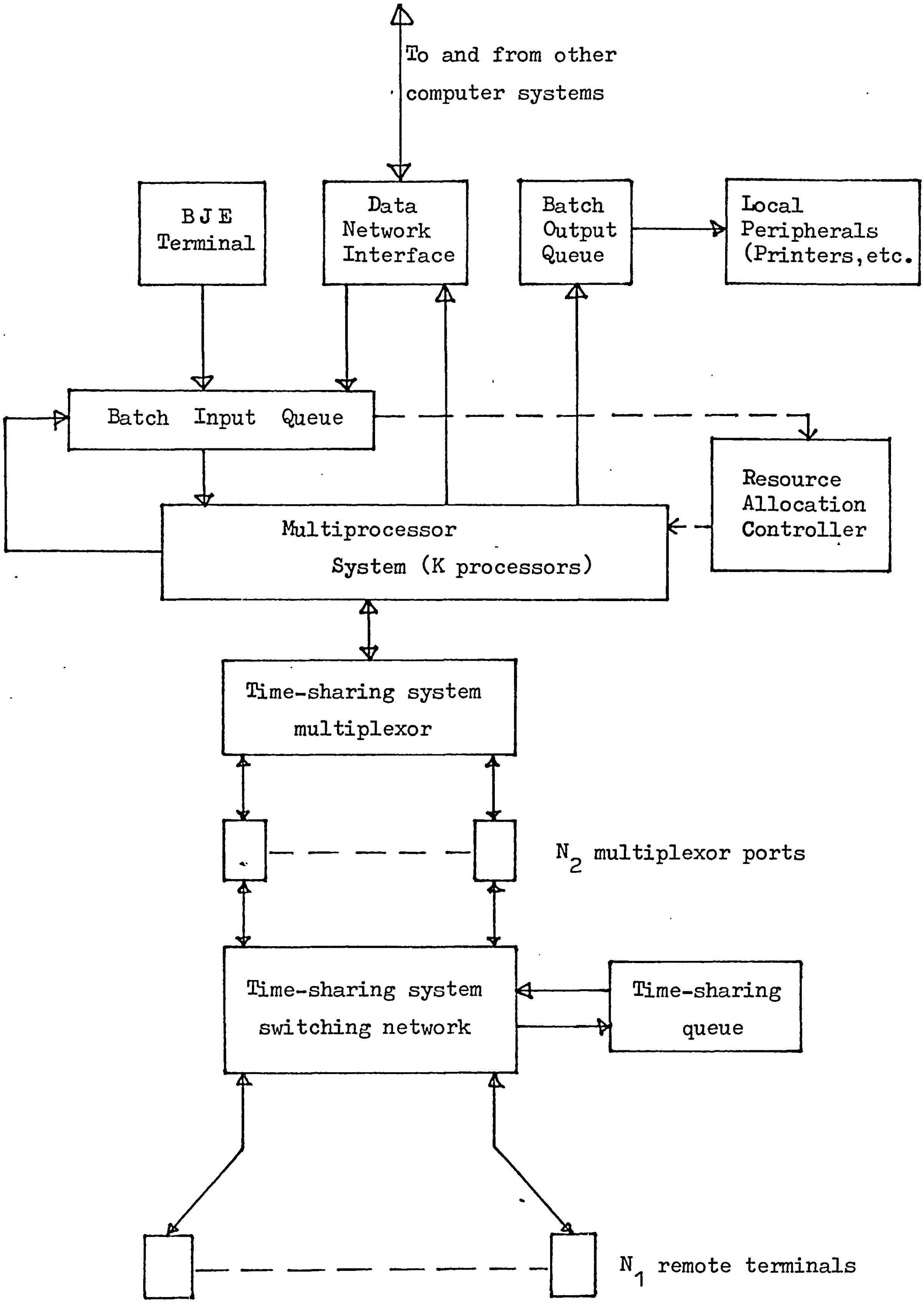


Fig. (10) Schematic of the computer system

The time-sharing load The switching network between the N_1 remote terminals and the N_2 multiplexor ports provides full availability: that is to say, there is a route from every terminal to every port. It is assumed that $N_1 \gg N_2$ so that the switching network concentrates the traffic from the terminals. A job submitted from one of the remote terminals is routed to one of the currently free ports, or, if all ports are active, is held in a FIFO queue until a port becomes available. [It should be emphasized that in the present context a time-sharing job means a single task requiring the use of a central processor, such as compiling a program, running a program, processing a file, etc. At the completion of each such job the associated port is assumed to be released and subsequent jobs from the same terminal will in general be associated with different ports.] The length of the time-sharing queue is restricted to N_2: if an attempt is made to submit a job when the time-sharing queue is full the system returns a message announcing that there is saturation and the job is then presumed to be lost. Arrivals to the system are assumed to be Poisson, at a rate which is independent of the degree of congestion in the system. The job processing times are again negative-exponential, and since the processor is time-shared between the currently active ports the effective mean service rate at any time is inversely proportional to the number of currently active ports.

Arrival stream	:	Poisson, total mean rate λ_T
Job processing times	:	Neg-exp., mean rate μ_T.
No. of servers	:	N_2
Max. queue length	:	N_2
Queue discipline	:	Arrival order (FIFO)

System operating costs As already mentioned, there are four components to the total system operating cost; we now discuss these in turn.

(a) State-dependent cost The state of the system is (for the present) taken to be the number of batch jobs in the system (including those currently being processed). For any time t, let X_t denote the number of batch jobs in the system at time t and let k_t denote the number of processors devoted to batch processing at time t. Then the number of batch jobs awaiting service at time t is 0, if $X_t \leqslant k_t$, or $(X_t - k_t)$, if $X_t > k_t$. Each waiting job accumulates waiting time at unit rate until it is serviced; thus the rate at which the total waiting time (summed over all waiting jobs) accumulates is equal to Max $(0, X_t - k_t)$. We shall assume that the state-dependent cost is proportional to this rate, ie. that

$$C_1(X_t, k_t) \quad = \quad c_1 \text{ Max } (0, X_t - k_t) \qquad \ldots.(6.1)$$

where c_1 is a constant.

(b) Control-delay cost It is assumed that the dynamics of the time-sharing load are fast compared with those of the batch-processing load; more precisely, we assume that $\lambda_T >> \lambda_B$ and $\mu_T >> \mu_B$. This means that between any two consecutive batch events (arrivals or departures) there will, with high probability, be a large number of time-share events (arrivals and/or departures). The control variable k_t is to be a function of the state X_t only, and hence its value cannot change between batch events. Whenever k_t does change to a new value we can assume, because of the relatively high rate at which time-share events occur, that the state of the time-sharing load (ie. number of ports active + size of time-sharing queue) reaches its new statistical equilibrium very rapidly and that this equilibrium is maintained until k_t again changes value. The equilibrium properties when $k_t = k$ ($k \in 0, 1, 2, \ldots K$) are determined as follows.

Essentially we have a $M/M/N_2/N_2$ queueing system in which the number

* p. 183 follows p. 184

We now argue that it is desirable for $m_{\bar{k}}$ to be close to N_2, the total number of multiplexor ports. For if $m_{\bar{k}} << N_2$ the average number of idle ports will be large which means inefficient use of the time-sharing system ; and if $m_{\bar{k}} >> N_2$ the response time (ie. the total time spent in the system by a time-sharing job) will tend to be unacceptably large. The mean response time is in fact $(m_{\bar{k}}/\lambda_E)$ where λ_E is the effective mean arrival rate for time-sharing jobs and is given by $\lambda_E = \lambda_T(1 - P_{2N_2})$. Thus, denoting the mean response time by $t_{\bar{k}}$ and using (6.2) and (6.3), we have

$$t_{\bar{k}} = \frac{\rho_{\bar{k}}}{\lambda_T}\left[\left(\frac{1}{1-\rho_{\bar{k}}}\right) - \left(\frac{2N_2\,\rho_{\bar{k}}^{2N_2}}{1-\rho_{\bar{k}}^{2N_2}}\right)\right] \qquad \text{....(6.4)}$$

Thus both $m_{\bar{k}}$ and $t_{\bar{k}}$ depend in a known way on the traffic intensity $\rho_{\bar{k}}$. In particular when $\rho_{\bar{k}} = 1$ equation (6.3) reduces to

$$m_{\bar{k}} = N_2 \qquad \text{....(6.3a)}$$

and equation (6.4) reduces to

$$\begin{aligned} t_{\bar{k}} &= \left(\frac{N_2 + \frac{1}{2}}{\lambda_T}\right) \\ &\simeq \frac{N_2}{\lambda_T}, \quad N_2 >> 1 \end{aligned} \qquad \text{....(6.4a)}$$

so that $m_{\bar{k}} \simeq \lambda_T t_{\bar{k}}$ when $\rho_{\bar{k}} = 1$.

Since N_2 is taken to be the desirable value of $m_{\bar{k}}$ we shall assume that the control-delay cost rate has the form

$$C_2(k_t) = c_2.(N_2 - m_{\bar{k}_t})^2 \qquad \text{....(6.5)}$$

where c_2 is a constant and $\bar{k}_t \triangleq K - k_t$ is the number of processors of available servers is the number, N_2, of multiplexor ports and the queue size is also limited to N_2. The mean arrival rate is λ_T, and if $\bar{k} \triangleq K-k$ is the number of processors currently devoted to time-sharing, the mean departure rate (ie. the mean rate at which completed time-share jobs leave the system) is $\bar{k}\mu_T$. This assumes that there are always at least $\bar{k}$ time-sharing jobs in the system; this is likely to be the case provided that $N_2 >> K$.

Let Y_t denote the number of time-sharing jobs in the system at time t, and let $P_i \triangleq P\left[Y_t = i\right]$ denote the equilibrium probability that $Y_t = i$. The equilibrium birth-death equations are then

$$\lambda_T \; P_i \;=\; \bar{k}\mu_T \; P_{i+1} \quad , \quad i = 0,1,\ldots,2N_2+1$$

so that

$$P_i \;=\; \left[\frac{1-\rho_{\bar{k}}}{1-\rho_{\bar{k}}^{2N_2+1}}\right] \rho_{\bar{k}}^{\,i} \quad , \; i = 0,1,\ldots,2N_2 \qquad \ldots.(6.2)$$

where $\rho_{\bar{k}} \triangleq \left\{\frac{\lambda_T}{\bar{k}\mu_T}\right\}$, the traffic intensity in the time-sharing system when $\bar{k}$ processors are devoted to time-sharing.

[NB. In the special case when $\rho_{\bar{k}} = 1$ the above equilibrium distribution becomes the uniform distribution

$$P_i \;=\; \frac{1}{2N_2+1} \quad , \quad i = 0,1,2,\ldots,2N_2 \qquad \ldots.(6.2a)$$]

From (6.2) or (6.2a) we can compute the equilibrium mean value, $m_{\bar{k}} \triangleq E\left[Y_t\right]$, of the number of time-sharing jobs in the system. The result is

$$m_{\bar{k}} \;=\; \left\{\frac{\rho_{\bar{k}}}{1-\rho_{\bar{k}}}\right\} \;-\; (2N_2+1)\left\{\frac{\rho_{\bar{k}}^{\,2N_2+1}}{1-\rho_{\bar{k}}^{\,2N_2+1}}\right\} \qquad \ldots.(6.3)$$

devoted to time-sharing at time t.

(c) Control-loss cost The quadratic loss function (6.5) is symmetrical about the point $m_{\bar{k}} = N_2$. In practice, however, the over-busy situation, $m_{\bar{k}} > N_2$, is less desirable than the under-busy situation, $m_{\bar{k}} < N_2$, because the rate at which jobs are lost to the system, due to saturation, increases with $m_{\bar{k}}$. The probability that a job is lost to the system is the probability P_{2N_2} that the system is full, and from (6.2) this is given by

$$P_{2N_2} = \left[\frac{1 - \rho_{\bar{k}}}{1 - \rho_{\bar{k}}^{2N_2+1}} \right] \rho_{\bar{k}}^{2N_2} \qquad \ldots.(6.6)$$

The mean rate at which jobs are lost is $\lambda_T P_{2N_2}$; we therefore assume a cost-rate associated with lost jobs of the form

$$C_3(k_t) = c_3 . \lambda_T P_{2N_2} \qquad \ldots.(6.7)$$

where c_3 is a constant and P_{2N_2} is given by (6.6).

(d) Switching cost There is a system overhead associated with each change in k, the number of processors devoted to batch processing. If k is reduced, then one or more of the batch jobs currently being processed will have to be "frozen" (contents of registers, states of flags, etc. must be stored) until a processor again becomes available. If k is increased, the time-shared jobs will have to be re-arranged for the reduced number of processors available for time-sharing. For simplicity we shall assume that the switching cost function is symmetrical: if at time t the value of k changes from k_{t-} to k_t there is incurred an instantaneous cost

$$C_4(k_t , k_{t-}) = c_4 \left| k_t - k_{t-} \right| \qquad \ldots.(6.8)$$

where c_4 is a constant.

Note that the value of C_4 depends not only on the current value of the control, k_t, but also on the immediately preceding value, k_{t-}.

6.3 Mathematical model of the system

In the system specified in the previous section the batch job arrival process is Poisson and the job processing times are negative-exponential. As a consequence, the process $\{X_t : t \in \mathcal{R}_+\}$ is a continuous-time birth-death process whose state space $\mathcal{X}$, if there is no limit on the number of waiting batch jobs, is the set $\mathcal{Z}_+$ of non-negative integers. By utilising the truncation procedure introduced in Section 5.3 of the previous chapter we now represent the system by a finite-state controllable birth-death process. In order to be able to do so we must assume that the total number of processors K is such that $K\mu_B > \lambda_B$ so that the restricted set of control laws, $\mathcal{F}_N$, defined by (5.6) is non-void. With this assumption we now define a controllable GBDP, $\{M_t\}$, by setting

$$\begin{cases} M_t \triangleq X_t \quad , \quad X_t \leqslant N \\ \quad\;\; \triangleq N \quad , \quad X_t > N \end{cases} \qquad \ldots.(6.9)$$

where, as already stated, X_t is the number of batch jobs in the system at time t.

The characteristic parameters of the semi-Markov chain $\{M_t\}$ are obtained as follows.

(i) Transition probability matrix

This is given by equation (5.8), with $\lambda = \lambda_B$, $\mu = \mu_B$, and $f(i)$ = number of processors allocated to batch processing when $M_t = i$, under the control law f.

(ii) Mean sojourn times

These are given by equation (5.13), again with $\lambda = \lambda_B$, $\mu = \mu_B$.

(iii) Mean one-step costs

For the optimal regulation problem to be properly posed the mean one-step costs must not depend on past values of state and/or control. This means (see equation (3.2)) that the transition cost function $C(X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}, u_n)$ must be independent of past control values — a condition which is not satisfied by the switching-cost component C_4 (equation (6.8)) in the present problem. For the prcsent, therefore, we set $c_4 = 0$ so that there is no switching cost; the resulting transition cost function for the process $\{X_t\}$ is defined by

$$C(X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}, u_n) \triangleq \left[C_1(X_{T_n}, u_n) + C_2(u_n) + C_3(u_n)\right] \Delta T_{n+1} \qquad \ldots.(6.10)$$

where C_1, C_2, C_3 are defined by (6.1), (6.5), (6.7). Then, using (3.2), the mean one-step cost for the process $\{X_t\}$, from the state $X_{T_n} = i$ under the control $u_n = k$, is given by

$$\begin{aligned}\gamma_i^k &= \left[C_1(i,k) + C_2(k) + C_3(k)\right] E\left[\Delta T_{n+1} \middle| X_{T_n} = i , u_n = k\right] \\ &= C_i^k \, \tau_i^k \qquad \ldots.(6.11)\end{aligned}$$

$$\text{where} \quad C_i^k \triangleq C_1(i,k) + C_2(k) + C_3(k) \qquad \ldots.(6.12)$$

Reference to (6.1) shows that the function C_i^k has the form

$$C_i^k = c_1 \operatorname{Max}(0, i - k) + C_2(k) + C_3(k) \qquad \ldots.(6.13)$$

so that γ_i^k is of the form specified by (5.16). Thus the mean one-step costs, $\bar{\gamma}_i^k$, of the embedded chain $\{M_t\}$ are given by equation (5.20), with $\bar{\gamma}_N^f$ given by (5.19).

The controllable GBDP $\{M_t\}$, with transition probabilities, mean sojourn times, and mean one-step costs defined as indicated above, has an equilibrium mean cost rate $\bar{c}^f$ which depends on the choice of control law f. Minimization of $\bar{c}^f$ over the set $\mathcal{F}_N$ of feasible control laws is an optimal regulation problem of the type considered in the earlier chapters of this thesis.

Parameter values

The following values were used in the numerical computations :

Batch arrival rate, λ_B	=	0.1 $\sec^{-1}$
Batch service rate, μ_B	=	0.06 $\sec^{-1}$
Time-share arrival rate, λ_T	=	2 $\sec^{-1}$
Time-share service rate, μ_T	=	1 $\sec^{-1}$
Number of processors, K	=	3
Number of time-share ports, N_2	=	20
Batch queue truncation level, N	=	9

Comments

(i) There is clearly no loss in generality in taking $c_1 = 1$. In the optimization computations we have therefore taken $c_1 = 1$, $c_4 = 0$ (no switching costs) and examined the effect of changes in c_2 and c_3 on the optimal control law and optimal cost rate.

(ii) Note that with the above values $K\mu_B > \lambda_B$ so that the truncation procedure is valid.

6.4 Method A : no quantization

The optimal regulation problem specified in the previous section has been solved for various values of the cost coefficients c_2 and c_3, using the direct policy-iteration (DPI) algorithm described in Section 4.2.3. The results are given in Tables 1 - 3 below.

Suppose now that the system has been optimized by the above procedure. We now ask : what would be the additional contribution to the (hitherto optimal) mean cost rate $\bar{c}^f$ if the switching-cost coefficient were changed to a non-zero value ? The answer will depend on the mean

TABLE 1

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE ZERO
METHOD A : NO QUANTIZATION

$c_2 = 0.005$

		$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
OPTIMAL COST RATE =		2.693	5.325	14.703
	STATE	CONTROL	CONTROL	CONTROL
	0	1	1	1
OPTIMAL	1	1	1	1
	2	3	3	1
CONTROL	3	3	3	1
	4	3	3	3
LAW	5	3	3	3
	6	3	3	3
	7	3	3	3
	8	3	3	3
	9	3	3	3
	STATE	PROBABILITY	PROBABILITY	PROBABILITY
	0	0.106	0.106	0.032
OPTIMAL	1	0.283	0.283	0.086
	2	0.275	0.275	0.143
STATIONARY	3	0.153	0.153	0.238
	4	0.085	0.085	0.231
DISTRIBUTION	5	0.047	0.047	0.129
	6	0.026	0.026	0.071
	7	0.015	0.015	0.040
	8	0.008	0.008	0.022
	9	0.002	0.002	0.008

TABLE 2

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE ZERO

METHOD A : NO QUANTIZATION

$c_2 = 0.01$

		$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
OPTIMAL COST RATE =		4.205	6.837	16.150
	STATE	CONTROL	CONTROL	CONTROL
	0	1	1	1
OPTIMAL	1	1	1	1
	2	3	3	1
CONTROL	3	3	3	1
	4	3	3	3
LAW	5	3	3	3
	6	3	3	3
	7	3	3	3
	8	3	3	3
	9	3	3	3
	STATE	PROBABILITY	PROBABILITY	PROBABILITY
	0	0.106	0.106	0.032
OPTIMAL	1	0.283	0.283	0.086
	2	0.275	0.275	0.143
STATIONARY	3	0.153	0.153	0.238
	4	0.085	0.085	0.231
DISTRIBUTION	5	0.047	0.047	0.129
	6	0.026	0.026	0.071
	7	0.015	0.015	0.040
	8	0.008	0.008	0.022
	9	0.002	0.002	0.008

TABLE 3

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE ZERO

METHOD A : NO QUANTIZATION

$c_2 = 0.02$

		$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
OPTIMAL COST RATE =		7.228	9.789	19.045
	STATE	CONTROL	CONTROL	CONTROL
OPTIMAL CONTROL LAW	0	1	1	1
	1	1	1	1
	2	3	1	1
	3	3	3	1
	4	3	3	3
	5	3	3	3
	6	3	3	3
	7	3	3	3
	8	3	3	3
	9	3	3	3
	STATE	PROBABILITY	PROBABILITY	PROBABILITY
OPTIMAL STATIONARY DISTRIBUTION	0	0.106	0.057	0.032
	1	0.283	0.151	0.086
	2	0.275	0.252	0.143
	3	0.153	0.245	0.238
	4	0.085	0.136	0.231
	5	0.047	0.076	0.129
	6	0.026	0.042	0.071
	7	0.015	0.023	0.040
	8	0.008	0.013	0.022
	9	0.002	0.005	0.008

rate at which the control variable k changes and this in turn will depend on the equilibrium state distribution of the system.

The additional cost rate component due to switching costs may be computed as follows. Given a control law f, let $\Delta\gamma_i^f$ denote the *expected one-step switching cost* for a transition out of state i. Since, under f, the control action is f(i) in state i, the switching cost for a transition i → j is, by (6.8), $c_4 \left| f(j) - f(i) \right|$. It follows immediately that

$$\Delta\gamma_i^f = \sum_j p_{ij}^{f(i)} \, c_4 \left| f(j) - f(i) \right| \qquad \text{....(6.14)}$$

The additional cost rate due to switching costs is then computed by (3.6) in the usual way, so that

$$\Delta\bar{c}^f = \frac{(\underline{\pi}^f)^T \, \underline{\Delta\gamma}^f}{(\underline{\pi}^f)^T \, \underline{\tau}^f} \qquad \text{....(6.15)}$$

where $\Delta\bar{c}^f$ is the required switching-cost component, and $\underline{\Delta\gamma}^f \triangleq \mathrm{Col}\,(\Delta\gamma_0^f, \ldots, \Delta\gamma_N^f)$. The mean one-step switching costs $\Delta\gamma_i^f$ are very easily computed in the present problem since in (6.14) the p_{ij}^f are non-zero only for $j = i + 1$ and $j = i - 1$.

Note that $\underline{\Delta\gamma}^f$ cannot be computed unless the control law f is already known. It is this fact that precludes the use of the present model for the optimization of f when switching costs are present. (It is also, of course, necessary to determine $\underline{\pi}^f$ explicitly, for use in (6.15).)

Switching-cost components $\Delta\bar{c}^f$ have been computed for each of the optimal control laws f listed in Tables 1 - 3, for various values of the cost coefficient c_4. The results are given in Tables 4 - 6 below.

6.5 Method B : fixed quantization

We now consider the possibility of achieving a lower overall cost

TABLES 4 - 6

TOTAL COST RATES FOR A-OPTIMAL SYSTEMS WHEN SWITCHING COSTS ARE INCLUDED

Table 4 : $c_4 = 25$

	$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
$c_2 = 0.005$	6.203	8.835	17.623
$c_2 = 0.01$	7.715	10.347	19.070
$c_2 = 0.02$	10.738	12.904	21.965

Table 5 : $c_4 = 50$

	$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
$c_2 = 0.005$	9.713	12.345	20.543
$c_2 = 0.01$	11.225	13.857	21.990
$c_2 = 0.02$	14.248	16.019	24.885

Table 6 : $c_4 = 100$

	$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
$c_2 = 0.005$	16.733	19.365	26.383
$c_2 = 0.01$	18.245	20.877	27.830
$c_2 = 0.02$	21.268	22.249	30.725

rate by using a quantized control law. Our argument is as follows: by optimizing f over a set of quantized control laws the optimal cost rate (with $c_4 = 0$) will not be as low as in the optimal unquantized case; however, the equilibrium mean switching rate should be significantly lower in the quantized case and hence so should the switching costs when c_4 is non-zero. As a consequence we might expect the overall cost rate (including switching cost component) to be lower in the quantized case than in the unquantized case — at least for sufficiciently large values of c_4.

Two different quantized versions of the present optimal regulation problem have been solved, each for various values of c_2 and c_3, using the modified DPI algorithm described at the end of Chapter 5. The quantizing partitions used were as follows : -

$$(\mathcal{X} = \{0,1,2,\ldots,8,9\} \quad \text{, as before)}$$

Case B.1

Quantum subset	1	2	3	4	5	6
States	0	1,2	3,4	5,6	7,8	9

Case B.2

Quantum subset	1	2	3	4
States	0	1,2,3,4	5,6,7,8	9

Comments :

(i) In both cases we have isolated the boundary states as single-state subsets : state 0 because the condition $M_t = 0$ is clearly a special case, and state 9 because the condition $M_t = 9$ in the truncated chain represents the condition $X_t \geqslant 9$ in the original, untruncated GBDP.

(ii) In both cases the quantization of the interior of $\mathcal{X}$ is uniform ;

although this simplifies the programming of the optimization algorithm it is not, of course, an essential feature of the procedure.

(iii) The embedded chain $\{Y_t\}$ defined by (5.21) - (5.23) (with, in this case, $\{M_t\}$ as the unquantized process) is, in case B.1, as big as $\{M_t\}$ itself. However, in case B.2, the state space of the embedded chain $\{Y_t\}$ is the reduced set $\{0, 1, 4, 5, 8, 9\}$ and it is necessary to use equations (5.29), (5.32) and (5.33) to compute the parameters of $\{Y_t\}$.

The results of the quantized optimization are given in Tables 7 - 9 (case B.1) and Tables 10 - 12 (case B.2) shown below.

As in case A, it is possible to compute the additional component of cost due to switching costs. Again we use (6.14) and (6.15), applied in this case either to the quantized process $\{M_t\}$ or to the equivalent embedded process $\{Y_t\}$. The results are given below in Tables 13 - 15 (case B.1) and Tables 16 - 18 (case B.2).

6.6 Method C : variable quantization

The control laws determined by methods A and B above are sub-optimal when the switching costs are non-zero. As we have already remarked, when the system state is defined in the natural way (ie. as the number of batch jobs in the system) it is not possible to include the switching cost component in the cost function to be optimized. This is because, with the natural definition of system state, the switching cost per step is associated with two consecutive control actions rather than a single control action. In order to include the switching cost in a separable cost function it is necessary to redefine the state in such a way that it incorporates the immediately past control action. We therefore now define W_t, the system state at time t, as the ordered pair (X_t , u_{t-}), where as before X_t is the number of batch jobs in the system at time t (ie. the natural system state) and u_{t-} is the control

TABLE 7

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE ZERO
METHOD B1 : 1 - 2 - 2 - 2 - 2 - 1 QUANTIZATION

$c_2 = 0.005$

		$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
OPTIMAL COST RATE =		2.939	5.382	14.728
	STATE	CONTROL	CONTROL	CONTROL
	0	1	1	1
OPTIMAL	1	3	1	1
	2	3	1	1
CONTROL	3	3	3	3
	4	3	3	3
LAW	5	3	3	3
	6	3	3	3
	7	3	3	3
	8	3	3	3
	9	3	3	3
	STATE	PROBABILITY	PROBABILITY	PROBABILITY
	0	0.223	0.057	0.057
OPTIMAL	1	0.347	0.151	0.151
	2	0.193	0.252	0.252
STATIONARY	3	0.107	0.245	0.245
	4	0.060	0.136	0.136
DISTRIBUTION	5	0.033	0.076	0.076
	6	0.018	0.042	0.042
	7	0.010	0.023	0.023
	8	0.006	0.013	0.013
	9	0.003	0.005	0.005

TABLE 8

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE ZERO
METHOD B1 : 1 - 2 - 2 - 2 - 2 - 1 QUANTIZATION

c_2 = 0.01

		c_3 = 1	c_3 = 4	c_3 = 16
OPTIMAL COST RATE =		4.515	6.851	16.197
	STATE	CONTROL	CONTROL	CONTROL
	0	1	1	1
OPTIMAL	1	1	1	1
	2	1	1	1
CONTROL	3	3	3	3
	4	3	3	3
LAW	5	3	3	3
	6	3	3	3
	7	3	3	3
	8	3	3	3
	9	3	3	3
	STATE	PROBABILITY	PROBABILITY	PROBABILITY
	0	0.057	0.057	0.057
OPTIMAL	1	0.151	0.151	0.151
	2	0.252	0.252	0.252
PROBABILITY	3	0.245	0.245	0.245
	4	0.136	0.136	0.136
DISTRIBUTION	5	0.076	0.076	0.076
	6	0.042	0.042	0.042
	7	0.023	0.023	0.023
	8	0.013	0.013	0.013
	9	0.005	0.005	0.005

TABLE 9

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE ZERO

METHOD B1 : 1 - 2 - 2 - 2 - 2 - 1 QUANTIZATION

$c_2 = 0.02$

		$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
OPTIMAL COST RATE =		7.542	9.789	19.135
	STATE	CONTROL	CONTROL	CONTROL
OPTIMAL CONTROL LAW	0	1	1	1
	1	1	1	1
	2	1	1	1
	3	3	3	3
	4	3	3	3
	5	3	3	3
	6	3	3	3
	7	3	3	3
	8	3	3	3
	9	3	3	3
	STATE	PROBABILITY	PROBABILITY	PROBABILITY
OPTIMAL STATIONARY DISTRIBUTION	0	0.057	0.057	0.057
	1	0.151	0.151	0.151
	2	0.252	0.252	0.252
	3	0.245	0.245	0.245
	4	0.136	0.136	0.136
	5	0.076	0.076	0.076
	6	0.042	0.042	0.042
	7	0.023	0.023	0.023
	8	0.013	0.013	0.013
	9	0.005	0.005	0.005

TABLE 10

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE ZERO

METHOD B2 : 1 - 4 - 4 - 1 QUANTIZATION

$c_2 = 0.005$

		$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
OPTIMAL COST RATE =		2.939	6.272	15.111
	STATE	CONTROL	CONTROL	CONTROL
OPTIMAL	0	1	1	1
	1	3	3	1
CONTROL	4	3	3	1
	5	3	3	3
LAW	8	3	3	3
	9	3	3	3
	STATE	PROBABILITY	PROBABILITY	PROBABILITY
OPTIMAL	0	0.453	0.453	0.053
	1	0.453	0.453	0.053
STATIONARY	4	0.043	0.043	0.408
	5	0.043	0.043	0.408
DISTRIBUTION	8	0.004	0.004	0.039
	9	0.004	0.004	0.039

Note : The stationary distributions shown here refer to the equivalent embedded chain $\{Y_t\}$.

TABLE 11

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE ZERO
METHOD B2 : 1 - 4 - 4 - 1 QUANTIZATION

$c_2 = 0.01$

		$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
OPTIMAL COST RATE =		4.552	7.885	16.547
	STATE	CONTROL	CONTROL	CONTROL
OPTIMAL	0	1	1	1
	1	3	3	1
CONTROL	4	3	3	1
	5	3	3	3
LAW	8	3	3	3
	9	3	3	3
	STATE	PROBABILITY	PROBABILITY	PROBABILITY
OPTIMAL	0	0.453	0.453	0.053
	1	0.453	0.453	0.053
STATIONARY	4	0.043	0.043	0.408
	5	0.043	0.043	0.408
DISTRIBUTION	8	0.004	0.004	0.039
	9	0.004	0.004	0.039

Note : The stationary distributions shown here refer to the equivalent embedded chain $\{Y_t\}$.

TABLE 12

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE ZERO
METHOD B2 : 1 - 4 - 4 - 1 QUANTIZATION

$c_2 = 0.02$

		$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
OPTIMAL COST RATE =		7.779	10.981	19.418
	STATE	CONTROL	CONTROL	CONTROL
OPTIMAL	0	1	1	1
	1	3	1	1
CONTROL	4	3	1	1
	5	3	3	3
LAW	8	3	3	3
	9	3	3	3
	STATE	PROBABILITY	PROBABILITY	PROBABILITY
OPTIMAL	0	0.453	0.053	0.053
	1	0.453	0.053	0.053
STATIONARY	4	0.043	0.408	0.408
	5	0.043	0.408	0.408
DISTRIBUTION	8	0.004	0.039	0.039
	9	0.004	0.039	0.039

Note : The stationary distributions shown here refer to the equivalent embedded chain $\{Y_t\}$.

TABLES 13 - 15

TOTAL COST RATES FOR B1-OPTIMAL SYSTEMS WHEN SWITCHING COSTS ARE INCLUDED

Table 13 : $c_4 = 25$

	$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
$c_2 = 0.005$	7.384	8.497	17.843
$c_2 = 0.01$	7.630	9.966	19.312
$c_2 = 0.02$	10.657	12.904	22.250

Table 14 : $c_4 = 50$

	$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
$c_2 = 0.005$	11.829	11.612	20.958
$c_2 = 0.01$	10.745	13.081	22.427
$c_2 = 0.02$	13.772	16.019	25.365

Table 15 : $c_4 = 100$

	$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
$c_2 = 0.005$	20.719	17.842	27.188
$c_2 = 0.01$	16.975	19.311	28.657
$c_2 = 0.02$	20.002	22.249	31.595

TABLES 16 - 18

TOTAL COST RATES FOR B2-OPTIMAL SYSTEMS WHEN SWITCHING COSTS ARE INCLUDED

Table 16 : $c_4 = 25$

	$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
$c_2 = 0.005$	7.384	10.717	17.924
$c_2 = 0.01$	8.997	12.330	19.360
$c_2 = 0.02$	12.224	13.794	22.231

Table 17 : $c_4 = 50$

	$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
$c_2 = 0.005$	11.829	15.162	20.736
$c_2 = 0.01$	13.442	16.775	22.172
$c_2 = 0.02$	16.669	16.606	25.043

Table 18 : $c_4 = 100$

	$c_3 = 1$	$c_3 = 4$	$c_3 = 16$
$c_2 = 0.005$	20.719	24.052	26.361
$c_2 = 0.01$	22.332	25.665	27.797
$c_2 = 0.02$	25.559	22.231	30.668

action in the interval immediately preceding the most recent transition of the process $\{X_t\}$.

We shall call the process $\{W_t\}$, where

$$W_t \triangleq (X_t , u_{t-}) \qquad \text{....(6.16)}$$

the augmented chain.

The transition cost function for the augmented chain $\{W_t\}$ now has the required separable form, even when switching costs are included, since the function C_4 defined in (6.8) is now a function only of the current state and current control. With switching costs included we now define the transition cost function for the augmented chain $\{W_t\}$ by

$$\begin{aligned} C(W_{T_n}, W_{T_{n+1}}, \Delta T_{n+1}, u_n) \triangleq \Big[C_1(X_{T_n}, u_n) \\ + C_2(u_n) + C_3(u_n) \Big] \Delta T_{n+1} \\ + C_4(u_n, u_{n-1}) \end{aligned} \qquad \text{....(6.17)}$$

where C_1, C_2, C_3 and C_4 are defined by (6.1), (6.5), (6.7) and (6.8) respectively, and X_{T_n} is the first component of W_{T_n}.

Then the expected one-step cost for the augmented chain $\{W_t\}$, from the state $W_{T_n} = (i,h)$ under the control $u_n = k$, is given by

$$\gamma^k_{i,h} = C^k_i \tau^k_i + c_4 \left| k - h \right| \qquad \text{....(6.18)}$$

where C^k_i is given by (6.13). The second term, of course, represents the switching cost.

As in Methods A and B it is necessary to truncate the state space by replacing the component X_t in (6.16) by the finite component M_t, defined by (6.9). The result is a truncated version of the augmented chain, say $\{V_t\}$, in which

$$V_t \triangleq (M_t, u_{t-}) \qquad \text{....(6.19)}$$

The essential parameters of $\{V_t\}$ are easily determined.

Define the following quantities :

(i) Transition probabilities of $\{V_t\}$

$$p^k_{(i,h),(j,l)} \triangleq P\left[V_{T_{n+1}} = (j,l) \,\middle|\, V_{T_n} = (i,h) \, , \, u_n = k\right]$$

(ii) Mean sojourn times of $\{V_t\}$

$$\tau^k_{i,h} \triangleq E\left[\Delta T_{n+1} \,\middle|\, V_{T_n} = (i,h) \, , \, u_n = k\right]$$

(iii) Mean one-step costs of $\{V_t\}$

$$\gamma^k_{i,h} \triangleq E\left[C(V_{T_n}, V_{T_{n+1}}, \Delta T_{n+1}, u_n) \,\middle|\, V_{T_n} = (i,h) \, , u_n = k\right]$$

Then it is straightforward to show that

$$p^k_{(i,h),(j,l)} = p^k_{ij} \, \delta_{kl} \qquad \text{....(6.20)}$$

where p^k_{ij} is the relevant transition probability for $\{M_t\}$ and δ_{kl} is the Kronecker delta ;

that

$$\tau^k_{i,h} = \bar{\tau}^k_i \qquad \text{....(6.21)}$$

where $\bar{\tau}^k_i$ is the relevant mean sojourn time of $\{M_t\}$;

and that

$$\gamma^k_{i,h} = \bar{\gamma}^k_i + c_4 \left| k - h \right| \qquad \text{....(6.22)}$$

where $\bar{\gamma}^k_i$ is the relevant mean one-step cost of $\{M_t\}$.

A control law for $\{V_t\}$ is a map from the new state space $(\mathcal{X} \times \mathcal{U})$

to the control set $\mathcal{U}$. Once such a control law f has been specified, the equilibrium mean cost rate (assuming that $\{V_t\}$ is a totally regular chain) is given by

$$\bar{c}_v^f = \frac{\sum_{i,h} \pi^f_{i,h} \gamma^f_{i,h}}{\sum_{i,h} \pi^f_{i,h} \tau^f_{i,h}} \qquad \text{....(6.23)}$$

where the $\tau^f_{i,h}$ and $\gamma^f_{i,h}$ are given by (6.21) and (6.22), and the $\pi^f_{i,h}$ are the unique stationary probabilities for $\{V_t\}$ under the control law f.

As already discussed in Chapter 3 the minimization of $\bar{c}_v^f$ with respect to f is a properly defined optimal regulation problem only if the controllable chain $\{V_t\}$ is totally regular, ie. only if $\{V_t^f\}$ is regular for every feasible control law f. Unfortunately, unless the set of feasible f is suitably restricted this will not be the case in the present problem. For if f is allowed to be any function from the augmented state space ($\mathcal{X} \times \mathcal{U}$) to the control set $\mathcal{U}$, there is always one control law for which $\{V_t\}$ possesses more than one recurrent subchain and hence is not regular. The control law in question is the following :

$$\forall (i,h) \in \mathcal{X} \times \mathcal{U} \ : \ f(i,h) = h \qquad \text{....(6.24)}$$

ie. always make the current control action the same as the immediately past control action, regardless of the current number of batch jobs in the system.

With this control law, if the initial state is (M_o,h_o) all subsequent states will be of the form (M_t,h_o). Thus, as inspection of the state transition diagram (Fig.(11)) shows, there will be K (= 4) recurrent subchains, one for each value of h_o. In fact there are many control laws for which $\{V_t\}$ is not regular and it is not easy to list

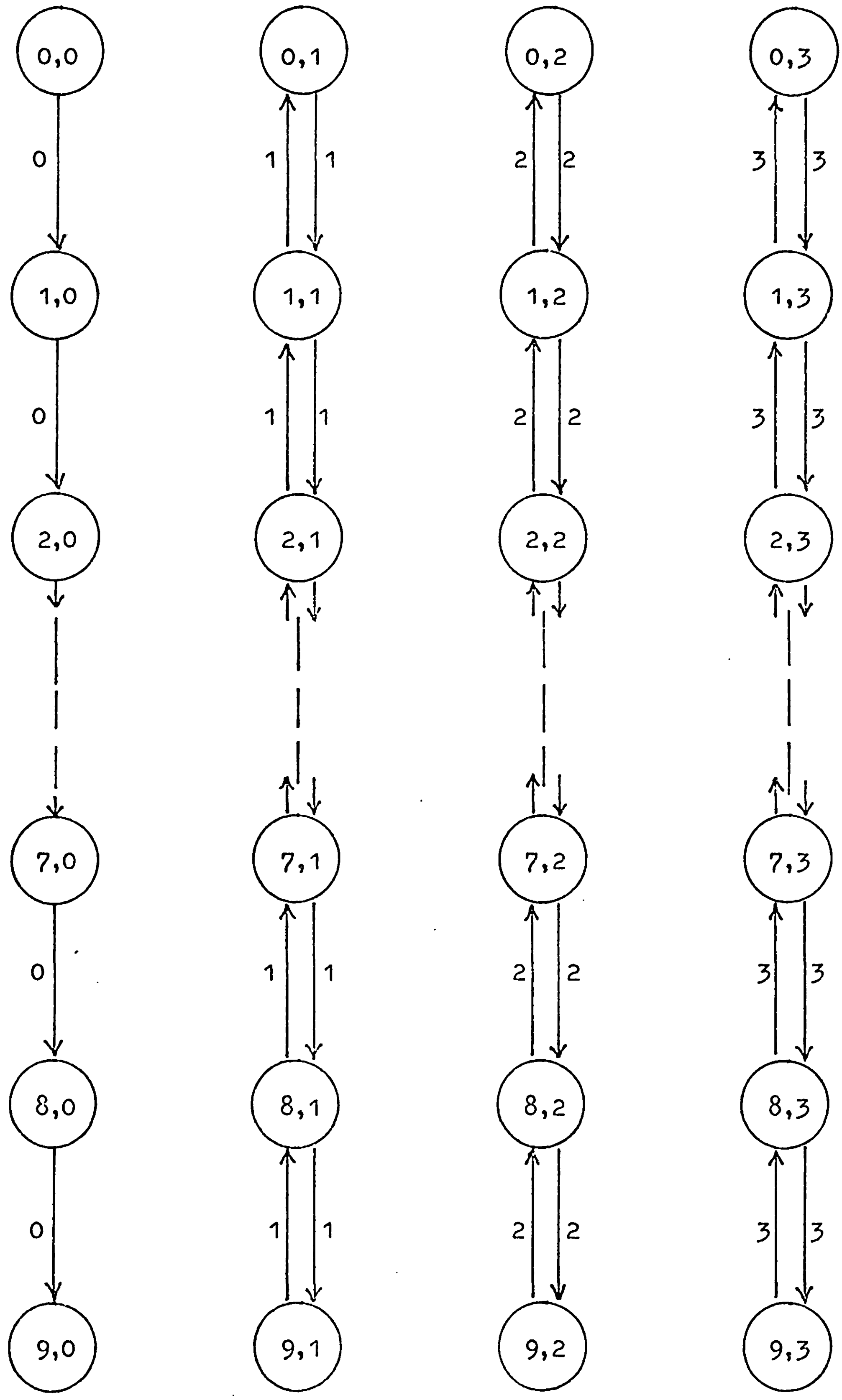


Fig.(11) State transition diagram - 4 recurrent chains them all systematically. If an attempt is made to use the DPI algorithm to optimize f, there will be a high probability (at least for certain choices of initial control law) that a "policy improvement" step will lead to a non-regular f : in such a case the E-matrix becomes singular and the algorithm explodes.

Fortunately, a simple restriction on the set of feasible control laws will restore total regularity. We make use of the following sufficient condition for total regularity : if $\{V_t\}$ possesses a single state (i,h) accessible from all other states under all feasible control laws, f, then (i,h) is recurrent for every f, and the set of states accessible from (i,h) constitutes the single recurrent subset of the state space of $\{V_t\}$.

This condition is met if we impose the constraint

$$f(N,h) = 3 \quad , \quad \forall h \in \mathcal{U} \qquad \ldots.(6.25)$$

ie. restrict the control value at the upper boundary i = N to the single value K = 3, regardless of the immediately preceding control action. The state transition diagram for an example of a control law satisfying (6.25) is shown in Fig.(12) where it can be seen that state (8,3) is necessarily recurrent since it is accessible from every other state. The choice f(N,h) = 3 is a natural constraint in view of the results for methods A and B. (Incidentally the apparently equivalent constraint f(O,h) = 1 does not yield a totally regular chain.)

The well-defined optimal regulation problem resulting from the use of (6.25) has been solved for various values of c_2, c_3 and c_4 using the DPI algorithm of Chapter 4. The results are given in Tables 19 - 24.

6.7 Discussion of results

Generally speaking, there are no major surprises in the results obtained, but there are several specific points which are worthy of

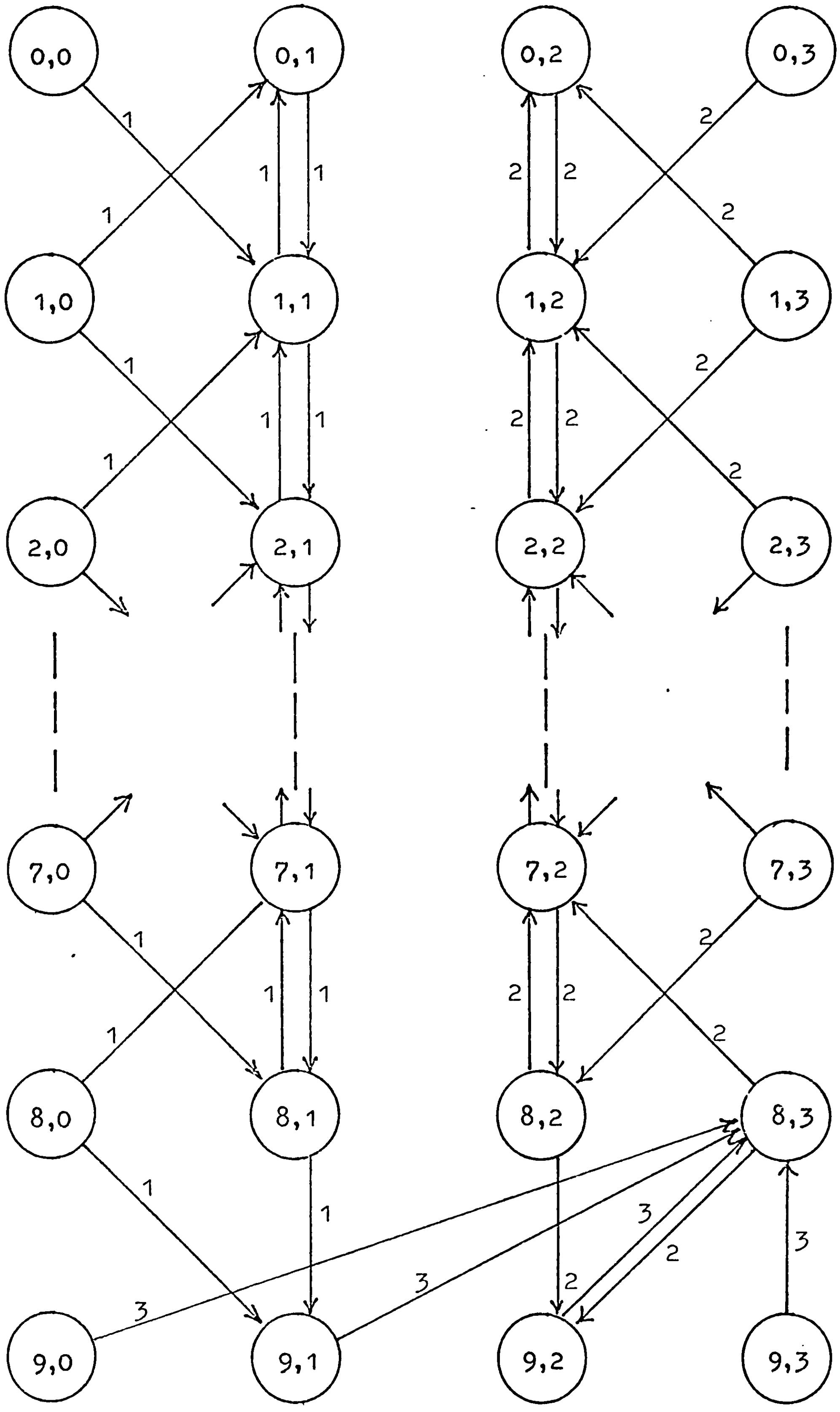


Fig.(12) State transition diagram - 1 recurrent chain

TABLE 19

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE NON-ZERO
METHOD C : VARIABLE QUANTIZATION

$c_2 = 0.005$; $c_3 = 4$; $c_4 = 50$

OPTIMAL COST RATE = 7.237

OPTIMAL CONTROL LAW

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	1	1	1	3	3	3	3	3	3	3
1	1	1	1	1	1	3	3	3	3	3
2	1	1	1	3	3	3	3	3	3	3
3	1	3	3	3	3	3	3	3	3	3

OPTIMAL STATIONARY DISTRIBUTION

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0.043	0.116	0.104	0.085	0.053	0.033	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0.033	0.052	0.080	0.096	0.105	0.077	0.061	0.034	0.019	0.009

TABLE 20

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE NON-ZERO
METHOD C : VARIABLE QUANTIZATION

$c_2 = 0.01$; $c_3 = 4$; $c_4 = 50$

OPTIMAL COST RATE = 8.723

OPTIMAL CONTROL LAW

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	1	1	1	3	3	3	3	3	3	3
1	1	1	1	1	1	3	3	3	3	3
2	1	1	1	3	3	3	3	3	3	3
3	1	3	3	3	3	3	3	3	3	3

OPTIMAL STATIONARY DISTRIBUTION

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0.043	0.116	0.104	0.085	0.053	0.033	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0.033	0.052	0.080	0.096	0.105	0.077	0.061	0.034	0.019	0.009

TABLE 21

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE NON-ZERO

METHOD C : VARIABLE QUANTIZATION

$c_2 = 0.02$; $c_3 = 4$; $c_4 = 50$

OPTIMAL COST RATE = 11.694

OPTIMAL CONTROL LAW

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	1	1	1	3	3	3	3	3	3	3
1	1	1	1	1	1	3	3	3	3	3
2	1	1	1	3	3	3	3	3	3	3
3	1	3	3	3	3	3	3	3	3	3

OPTIMAL STATIONARY DISTRIBUTION

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0.043	0.116	0.104	0.085	0.053	0.033	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0.033	0.052	0.080	0.096	0.105	0.077	0.061	0.034	0.019	0.009

TABLE 22

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE NON-ZERO

METHOD C : VARIABLE QUANTIZATION

$c_2 = 0.005$; $c_3 = 4$; $c_4 = 100$

OPTIMAL COST RATE = 8.344

OPTIMAL CONTROL LAW

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	1	1	1	3	3	3	3	3	3	3
1	1	1	1	1	1	1	3	3	3	3
2	1	1	1	3	3	3	3	3	3	3
3	1	3	3	3	3	3	3	3	3	3

OPTIMAL STATIONARY DISTRIBUTION

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0.037	0.100	0.094	0.085	0.069	0.043	0.027	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0.027	0.042	0.066	0.079	0.086	0.090	0.065	0.051	0.028	0.011

TABLE 23

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE NON-ZERO
METHOD C : VARIABLE QUANTIZATION

$c_2 = 0.01$; $c_3 = 4$; $c_4 = 100$

OPTIMAL COST RATE = 9.817

OPTIMAL CONTROL LAW

i= h=	0	1	2	3	4	5	6	7	8	9
0	1	1	1	3	3	3	3	3	3	3
1	1	1	1	1	1	1	1	3	3	3
2	1	1	1	3	3	3	3	3	3	3
3	1	3	3	3	3	3	3	3	3	3

OPTIMAL STATIONARY DISTRIBUTION

i = h =	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0.033	0.088	0.085	0.080	0.072	0.059	0.037	0.023	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0.023	0.036	0.056	0.067	0.073	0.076	0.078	0.056	0.044	0.014

TABLE 24

OPTIMAL SYSTEM WHEN SWITCHING COSTS ARE NON-ZERO
METHOD C : VARIABLE QUANTIZATION

$c_2 = 0.02$; $c_3 = 4$; $c_4 = 100$

OPTIMAL COST RATE = 12.751

OPTIMAL CONTROL LAW

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	1	1	1	3	3	3	3	3	3	3
1	1	1	1	1	1	1	1	3	3	3
2	1	1	1	3	3	3	3	3	3	3
3	1	3	3	3	3	3	3	3	3	3

OPTIMAL STATIONARY DISTRIBUTION

h = \ i =	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0.033	0.088	0.085	0.080	0.072	0.059	0.037	0.023	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0.023	0.036	0.056	0.067	0.073	0.076	0.078	0.056	0.044	0.014

comment.

Method A

(i) As expected, the optimal control law f^o is always monotone in the state: that is, $f^o(i_2) \geqslant f^o(i_1)$ for every pair of states i_1, i_2 such that $i_2 > i_1$. That the optimal control law for a simple, infinite-state, birth-death process is of this form has been demonstrated by Crabill(1972).

(ii) For the chosen range of values for the cost parameters c_2 and c_3, the controls $k = 0$ and $k = 2$ are absent from the optimal control laws. Their absence is a consequence of the particular cost structure associated with our system. Thus for the given range of values of c_2 and c_3 the control $k = 1$ is actually cheaper to use than $k = 0$ and hence is always chosen in preference to the latter. Similarly, with the given cost parameters the control-cost penalty for using $k = 3$ rather than $k = 2$ is relatively small whereas the corresponding state-cost benefit is large.

(iii) The stationary distributions have the expected form, the maximum probability being located near the control-switching boundary. Below this boundary X_t will tend to increase until k is switched from 1 to 3, and above this boundary X_t will tend to decrease until k is switched from 3 to 1.

(iv) With the given values of c_4 the switching cost component is a significant part of the total cost. For this reason we should expect the control laws yielded by Method A to be non-optimal when c_4 is non-zero.

Method B1

(i) Again the controls $k = 0$ and $k = 2$ are absent from the optimal control laws, for the same reasons as before.

(ii) The performance of the B1-optimal systems are very little worse than (in some cases, the same as) the performance of the corresponding A-optimal systems. The reason is that the A-optimal control laws need only a very small change (in some cases, no change) to convert them to control laws which are feasible as quantized control laws for the given quantization pattern.

(iii) Because of (ii) there is very little reduction in switching cost when the control is quantized.

Method B2

(i) Once again the controls k = 0 and k = 2 are absent from the optimal control laws.

(ii) Because the number of quantum subsets is now relatively small, the choice of control laws is now highly constrained. As a result the performance of the B2-optimal systems is significantly worse than that of the corresponding A-optimal systems.

(iii) The switching cost component is in many cases, larger than for the corresponding A-optimal or B1-optimal system. The reason is that the switching boundary is, in these cases, forced down to $i = 0/i = 1$, where the rate of switching from k = 1 to k = 3 is increased. Only in those cases where the switching boundary is forced up to $i = 4/i = 5$ is the switching cost component lower than in the corresponding A-optimal and B1-optimal systems.

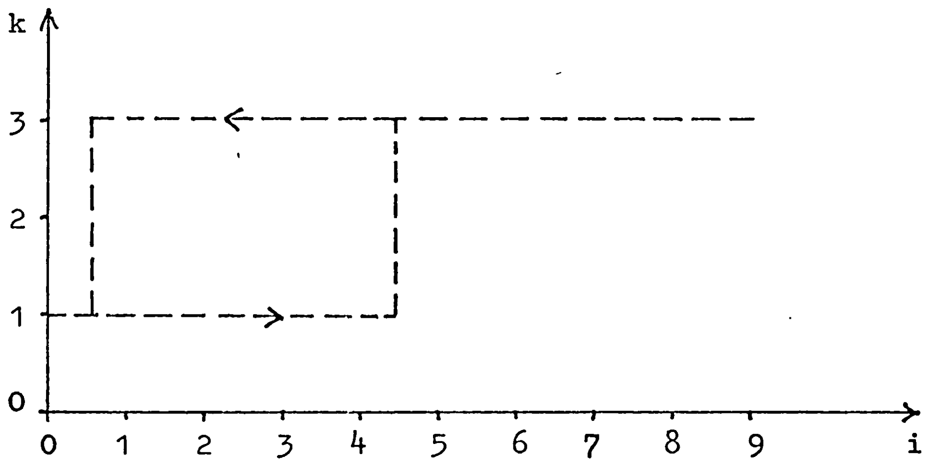
Method C

(i) For all the results shown in Tables 19 - 24, the initial control law was:

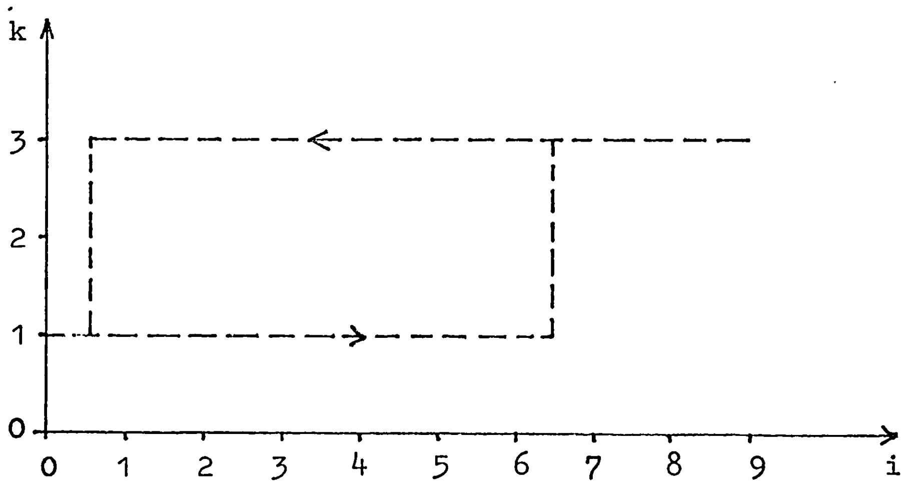
$$\begin{cases} f(i,h) = 1 , & i \leqslant 2 , \text{ all } h \\ = 3 , & i > 2 , \text{ all } h. \end{cases}$$

Under this control law all states of the form (i,0) or (i,2) are transient. As we have shown (Chapter 4) no change in control in such a state can (a) result in a reduced cost rate $\bar{c}$, or (b) cause the state to become recurrent. Thus the controls in these states are left unchanged by the DPI optimization algorithm.

(ii) The optimization procedure results, in each case, in a hysteresis type of control, in which the value k = 1 tends to be maintained when the state X_t is increasing and the value k = 3 tends to be maintained when X_t is decreasing. The width of the hysteresis loop increases with increasing c_4. For example, for Table 20 the optimal control law can be represented by the diagram:



and for Table 23 the optimal control law can be represented by:



(iii) The set of recurrent states in each case is readily seen to be closely related to the form of the control law.

(iv) It is clear that, by permitting the use of hysteresis-type control laws, optimization of the system by Method C ensures that the mean control-switching rate is kept low with the result that the overall cost rate is substantially lower than is attainable by the other methods.

A comparison of the performance achieved by the various methods is shown in Table 25. The average optimization times (on a CDC Cyber 174 computer) were as follows : -

Method A	:	0.42 sec
Method B1	:	0.40 sec
Method B2	:	0.28 sec
Method C	:	2.64 sec .

Although the much enlarged state space needed for Method C clearly results in a substantially more expensive optimization procedure, the results in Table 25 suggest that Method C is nevertheless well worth using when the switching costs are likely to be significant.

Comparison of algorithms

In order to obtain a practical estimate of the relative computational efficiencies of the DPI algorithm and the standard Howard/Jewell algorithm, the optimization problems in case C were also solved using the latter algorithm. To ensure that the results obtained were not specific to the birth-death optimization problem, no account was taken of the special structure of the P-matrix in either the DPI program or the Howard/Jewell program. The results, for the same set of optimization runs, were as follows :

	Mean optimization time	Mean number of optimization cycles
DPI method	2.64 sec	3.2
Howard/Jewell method	2.93 sec	3.5

TABLE 25

COMPARISON OF COST RATES FOR
DIFFERENT OPTIMIZATION METHODS

$c_3 = 4$

		METHOD A	METHOD B1	METHOD B2	METHOD C
$c_2 = 0.005$	$c_4 = 0$	5.325	5.382	6.272	5.325
	$c_4 = 50$	12.345	11.612	15.162	7.237
	$c_4 = 100$	19.365	17.842	24.052	8.334
$c_2 = 0.01$	$c_4 = 0$	6.837	6.851	7.885	6.837
	$c_4 = 50$	13.857	13.081	16.775	8.723
	$c_4 = 100$	20.877	19.311	25.665	9.817
$c_2 = 0.02$	$c_4 = 0$	9.789	9.789	10.981	9.789
	$c_4 = 50$	16.019	16.019	16.606	11.694
	$c_4 = 100$	22.249	22.249	22.231	12.751

It can be seen that, for the 40-state problem investigated here, the DPI algorithm is approximately 10% faster than the Howard/Jewell algorithm. Since, as the analysis in Section 4.2.4 shows, the time per optimization cycle increases at least quadratically with the number of states N, the DPI algorithm can be expected to show a substantial time advantage over the Howard/Jewell algorithm when N is very large.

6.8 Concluding remarks

In this thesis we have presented a critical review of the main algorithms currently used for the numerical optimization of controllable semi-Markov chains. The major difficulty with the main algorithms is that the amount of computational effort involved increases rapidly with the number of states in the chain. In an attempt to mitigate this difficulty we have developed new algorithms of the policy-iteration type and have shown that these new algorithms may be expected to be more efficient in many cases than the standard procedures.

We have also investigated optimization algorithms of the successive-approximations type, in which, although the amount of computational effort may not increase rapidly with the number of states in the chain, the computation time depends strongly on the dynamics of the chain and may consequently be very long for certain problems. Here again we have shown how it may sometimes be possible, by suitable modification of the standard procedure, to speed up the computation substantially.

In the special case when the controllable chain is a birth-death process we have shown that it is possible to determine the optimal control law even when the state space is quantized. In such a case the standard algorithms do not work and it is necessary to make use of one of the new single-step algorithms.

Two major problems to which no reference has been made in this thesis, but which merit further investigation, are the following :

(i) Optimal regulation of partially-observable semi-Markov chains (with both classical and non-classical information patterns). This is a difficult problem for which, at the time of writing, no substantial results have been obtained.

(ii) Optimal regulation of multi-dimensional birth-death chains (vector chains) in which the state space is quantized. Such problems are of interest because a vector birth-death chain can in principle serve as a general (though approximate) model for a wide class of non-linear stochastic systems.

LIST OF PRINCIPAL SYMBOLS

$\mathbb{N}$	the set of natural numbers 1,2,3,...
$\mathbb{N}_N$	the first N natural numbers 1,2,3,...,N
$\mathbb{Z}$	the set of integers
$\mathbb{Z}_+$	the set of non-negative integers
$\mathbb{R}$	the set of real numbers
$\mathbb{R}_+$	the set of non-negative real numbers
I	unit matrix
$\underline{x}$	column vector
$\underline{x}^T$	row vector : transpose of $\underline{x}$
$\underline{e}$	vector whose elements are all unity
$\underline{e}_i$	vector whose i^{th} element is unity, the rest zero
Ω	sample space
$\mathcal{X}$	state space of a stochastic process
$\mathcal{T}$	index set of a stochastic process
$\mathcal{U}$	control set of a controllable stochastic process
$P[A]$	probability of the event A
$E[X]$	expectation of the random variable X
$\{X_t\}$	Markov chain or semi-Markov chain
$\{\bar{X}_n\}$	embedded Markov chain
p_{ij}	one-step transition probability
P	matrix of one-step transition probabilities
F	semi-Markov kernel
$\underline{\pi}$	vector of stationary probabilities
$\underline{\tau}$	vector of mean sojourn times
$\underline{\gamma}$	vector of expected one-step costs
$\underline{v}(n)$	vector of expected n-step costs
$\underline{\delta}$	value vector

$\bar{\tau}$	equilibrium mean sojourn time
$\bar{\gamma}$	equilibrium mean one-step cost
$\bar{c}$	equilibrium mean cost rate
u_t	control action at time t
f_t	control law at time t
f	stationary control law
f^o	optimal stationary control law
$\wedge$	minimum (glb)
$\vee$	maximum (lub)

REFERENCES

AOKI, M., "Optimal control of partially observable Markovian control systems", J. Franklin Inst., 280, 1965, pp.367-386.

ASTROM, K.J., "Optimal control of Markov processes with incomplete state information", J. Math. Anal. & App., 10, 1965, pp.174-204.

BELLMAN, R., "Dynamic Programming", Princeton Univ. Press, 1957, Chapter XI.

BLACKWELL, D., "Discrete dynamic programming", Ann. Math. Stat., 33, 1962, pp.719-726.

BROSH, I., "The policy space structure of Markovian systems with two types of service", Manage. Sci., 16, 1970, pp.607-621.

de CANI, J.S., "A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity", Manage. Sci., 10, 1964, pp.716-733.

CHARNES, A. and COOPER,W.W., "Management models and industrial applications of linear programming", Wiley, 1961.

CHUNG, K.L., "Markov chains with stationary transition probabilities", 2nd Edition, Springer-Verlag, 1967.

ÇINLAR, E., "Markov renewal theory", Adv. Appl. Prob., 1, 1969a, pp.123-187.

ÇINLAR, E., "On semi-Markov processes on arbitrary spaces", Proc. Camb. Phil. Soc., 66, 1969b, pp.381-392.

ÇINLAR, E., "Markov renewal theory: a survey ", Management Sci., 21, 1975, pp.727-752.

ÇINLAR, E., "Introduction to Stochastic Processes", Prentice Hall, 1975.

COX, D.R., "Renewal Theory", Methuen, 1962.

CRABILL, T.B., "Optimal control of a service facility with variable exponential service time and constant arrival rate", Manage. Sci., 18, 1972, pp.560-566.

DENARDO, E.V. and FOX, B.L., "Multichain Markov renewal programs", SIAM J. Appl. Maths, 16, 1968, pp.468-487.

DERMAN, C., "On sequential decisions and Markov chains", Manage. Sci., 9, 1962, pp.16-24.

DERMAN, C., "Denumerable state Markovian decision processes - average cost criterion", Ann. Math. Stat., 37, 1966, pp.1545-53.

EATON, J.H. and ZADEH, L.A., "Optimal pursuit strategies in discrete state probabilistic systems", A.S.M.E. J. Basic Eng., 84, 1962, pp.23-29.

FOX, B.L., "Finite-state approximations to denumerable-state dynamic programs", J. Math, Anal. & App., 34, 1971, pp.665-670.

FOX, B.L., "Discretizing dynamic programs", J. Opt. Theory & App., 11, 1973, pp.228-234.

GROSS, D. and HARRIS, C.M., "Fundamentals of Queueing Theory", Wiley, 1974.

HAJNAL, J., "Weak ergodicity in non-homogeneous Markov chains", Proc. Camb. Phil. Soc., 54, 1958, pp.233-246.

HASTINGS, N.A.J., "Some notes on dynamic programming and replacement", Opnl. Res. Quart., 19, 1968, pp.453-464.

HASTINGS, N.A.J., "Bounds on the gain of a Markov decision process", Opns. Res., 19, 1971, pp.240-244.

HAUSSMAN, U.G., "On the optimal long-run control of Markov renewal processes", J. Math. Anal. & App., 36, 1971, pp.123-40.

HOWARD, R.A., "Dynamic programming and Markov processes", Wiley/MIT Press, 1960.

ISAACSON, E. and KELLER, H.B., "Analysis of Numerical Methods", Wiley, 1966.

JEWELL, W., "Markov renewal programming I and II ", Opns. Res., 2, 1963, pp.938-971.

KUSHNER, H. and KLEINMAN, A., "Numerical methods for the solution of the degenerate non-linear elliptic equations arising in optimal stochastic control theory", IEEE Trans. Automat. Contr., AC-13, 1968, pp.344-353.

KUSHNER, H. and KLEINMAN, A., "Accelerated procedures for the solution of discrete Markov control problems", IEEE Trans. Automat. Contr., AC-16, 1971, pp.147-152.

KWAKERNAAK, H. and SIVAN, R., "Linear optimal control systems", Wiley, 1972.

LEVY, P., "Processus semi-Markoviens", Proc. Int. Congress Math. (Amsterdam), 3, 1954, pp.416-426.

LIPPMAN, S.A., "Semi-Markov decision processes with unbounded rewards", Manage. Sci., 19, 1973, pp.717-731.

MACQUEEN, J.B., "A modified dynamic programming method for Markovian decision problems", J. Math. Anal. and App., 14, 1966, pp.38-43.

MANNE, A.S., "Linear programming and sequential decisions", Manage. Sci., 6, 1960, pp.259-267.

MAYNE, D.Q., "Recent approaches to the control of stochastic processes", J. Inst. Maths. & App., 3, 1967, pp.46-59.

ODONI, A.R., "On finding the maximal gain for Markov decision processes", Opns. Res., 17, 1969, pp.857-860.

OSAKI, S. and MINE, H., "Linear programming algorithms for semi-Markovian decision processes", J. Math. Anal. Applics., 22, 1968, pp.356-381.

PYKE, R., "Markov renewal processes: definitions and preliminary properties", Ann. Math. Stat., 32, 1961a, pp.1231-1242.

PYKE, R., "Markov renewal processes with finitely many states", Ann. Math. Stat., 32, 1961b, pp.1243-1259.

ROSS, S.M., "Non-discounted denumerable Markovian decision models", Ann. Math. Stat., 39, 1968a, pp.412-423.

ROSS, S.M., "Arbitrary-state Markovian decision processes", Ann. Math. Stat., 39, 1968b, pp.2118-2122.

ROSS, S.M., "Average cost semi-Markov decision processes", Operations Res. Centre, Berkeley, Univ. of California, Report No. ORC 69-27, 1969.

ROSS, S.M., "Applied Probability Models with Optimization Applications", Holden-Day, 1970.

RUDEMO, M., "State estimation for partially-observed Markov chains" J. Math. Anal. & App., 44, 1973: pp.581-611.

RUDEMO, M., "Prediction and smoothing for partially-observed Markov chains", J. Math. Anal. & App., 49, 1975, pp.1-23.

SARAWAGI, Y. and YOSHIKAWA, T., "Discrete-time Markovian decision processes with incomplete state observation", Ann. Math. Stat., 41, 1970, pp.78-86.

SCHWEITZER, P.J., "Perturbation theory and undiscounted Markov renewal programming", Opns. Res., 17, 1969, pp.716-727.

SCHWEITZER, P.J., "Multiple policy improvements in undiscounted Markov renewal programming", Opns. Res., 19, 1971a, pp.784-793.

SCHWEITZER, P.J., "Iterative solution of the functional equations of undiscounted Markov renewal programming", J. Math. Anal. & App., 34, 1971b, pp.495-501.

SENETA, E., "Non-negative matrices", George Allen and Unwin, 1973.

SHERMAN, J. and MORRISON, W.J., "Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix", Ann. Math. Stat., 20, 1949, p.621.

SMALLWOOD, R.D. and SONDIK, E.J., "The optimal control of partially-observable Markov processes over a finite horizon", Opns. Res., 21, 1973, pp.1071-1088.

SMITH, J.L., "Markov decisions on a partitioned state space", IEEE Trans. Systems, Man., Cyb., 1, 1971, pp.55-60.

SMITH, W.L., "Regenerative stochastic processes", Proc. Roy. Soc., Ser. A, 232, 1955, pp.6-31.

STEIN, P. and ROSENBERG, R.L., "On the solution of linear simultaneous equations by iteration", J. London Math. Soc., 23, 1948, pp.111-118.

TEUGELS, J.L., "Exponential ergodicity in Markov renewal processes", J. Appl. Prob., 5, 1968, pp.387-400.

TRUSTRUM, K., "Linear Programming", Routledge, Kegan Paul, 1971.

VARGA, R., "Matrix Iterative Analysis", Prentice-Hall, 1962.

VEINOTT, A.F., "On finding optimal policies in discrete dynamic programming with no discounting", Ann. Math. Stat., 37, 1966, pp.1284-94.

WHITE, D.J., "Dynamic programming, Markov chains, and the method of successive approximations", J. Math. Anal. & App., 6, 1963, pp.373-376.

WILLIAMS, P.W., "Numerical Computation", Nelson, 1972.

WITSENHAUSEN, H., "Separation of estimation and control for discrete time systems", Proc. IEEE, 59, 1971, pp.1557-1565.

WOLFE, P. and DANTZIG, G.B., "Linear programming in a Markov chain", Opns. Res., 10, 1962, pp.702-710.

WOLFOWITZ, J., "Products of indecomposable, aperiodic, stochastic matrices", Proc. Amer. Math. Soc., 14, 1963, pp.733-7.

APPENDIX

SOME PROPERTIES OF SEMI-MARKOV CHAINS

I. Equilibrium Probabilities

Let $\{X_n : n \in \mathbb{Z}_+\}$ be a sequence of non-negative, independent, identically-distributed random variables. Such a sequence is a discrete-parameter stochastic process known as a renewal process. We may think of $\{X_n\}$ as the sequence of inter-event times for some underlying point process. Associated with the renewal process $\{X_n\}$ are the stochastic processes $\{S_n : n \in \mathbb{Z}_+\}$ and $\{N_t : t \in \mathbb{R}_+\}$, defined by

$$S_n \triangleq 0 \quad , \quad n = 0$$

$$\triangleq \sum_{i=1}^{n} X_i \quad , \quad n > 0 \qquad \text{....(A.1)}$$

and

$$N_t \triangleq \sup\{n : S_n \leqslant t\} \qquad \text{....(A.2)}$$

The process $\{N_t\}$ is called the renewal counting process for $\{X_n\}$; and S_n is the time of the n^{th} renewal generated by $\{X_n\}$. The fundamental relationship between $\{S_n\}$ and $\{N_t\}$ is

$$[N_t \geqslant n] \Longleftrightarrow [S_n \leqslant t] \qquad \text{....(A.3)}$$

Let μ be the common expectation of the random variables X_n (μ exists since the X_n are non-negative) and assume that μ is finite. From the strong law of large numbers, S_n/n converges almost surely to μ and so the event $S_n \leqslant t$ only finitely often. It follows from (A.3) that, for any finite t, N_t is finite.

The function $m : t \mapsto E[N_t]$ is called the renewal function associated with the processes $\{X_n\}$, $\{S_n\}$ and $\{N_t\}$. It may be shown that if F is the common distribution function of the random variables X_n the renewal function m satisfies the so-called renewal equation

$$m(t) = F(t) + \int_0^t m(t-x)\, dF(x) \qquad \text{....(A.4)}$$

from which the *Laplace transform, $\overline{m}(s)$, of m(t) is given by

$$\overline{m}(s) = \frac{\overline{F}(s)}{1 - \overline{F}(s)} \qquad \text{....(A.5)}$$

where $\overline{F}(s)$ is the Laplace transform of F(t).

More generally, if h(t) is a known function of t, the integral equation

$$w(t) = h(t) + \int_0^t w(t-x)\, dF(x) \qquad \text{....(A.6)}$$

is called a renewal-type equation, and its solution is given in terms of the Laplace transforms $\overline{w}(s)$ and $\overline{h}(s)$ by

$$\overline{w}(s) = \frac{\overline{h}(s)}{1 - \overline{F}(s)} \qquad \text{....(A.7)}$$

or, using (A.5),

$$\overline{w}(s) = \overline{h}(s) + \overline{h}(s)\ \overline{m}(s) \qquad \text{....(A.8)}$$

Inversion of (A.8) yields

$$w(t) = h(t) + \int_0^t h(t-x)\, dm(x) \qquad \text{....(A.9)}$$

In order to derive the equilibrium probabilities of a regular semi-Markov chain, a result known as the key renewal theorem (see, for example, Ross(1970)) is needed. A simplified statement of this theorem is as follows :

If h is a non-negative, non-increasing function of t such that $\int_0^\infty h(t)dt < \infty$, and if the distribution function F is not

* defined by $\int_0^\infty e^{-st} dm(t)$

lattice (ie. the points of increase in F are not restricted to the lattice set $\{t = nT : n \in \mathbb{Z}_+, T \text{ fixed}\}$), then

$$\lim_{t \to \infty} \int_0^t h(t-x)\, dm(x) = \frac{1}{\mu} \int_0^\infty h(t)\, dt \qquad \text{....(A.10)}$$

Now consider a stochastic process $\{(X_t : \Omega \to \mathbb{Z}_+) : t \in \mathcal{T}\}$ with the following property: there exists a time T_1 such that, for every positive integer k, for every sequence of values $(x_1, x_2, \ldots, x_k) \in \mathbb{Z}_+^k$, for every sequence of times $(t_1, t_2, \ldots, t_k) \in \mathcal{T}^k$, and for each $i \in \mathbb{Z}_+$,

$$P\left[X_{t_1+T_1} \leqslant x_1, X_{t_2+T_1} \leqslant x_2, \ldots, X_{t_k+T_1} \leqslant x_k \,\middle|\, X_{T_1} = i\right]$$

$$= P\left[X_{t_1} \leqslant x_1, X_{t_2} \leqslant x_2, \ldots, X_{t_k} \leqslant x_k \,\middle|\, X_0 = i\right] \qquad \text{....(A.11)}$$

that is to say, there exists a time T_1 such that the continuation of the process beyond $t = T_1$ is a probabilistic replica of the whole process starting at $t = 0$.

It follows immediately that if such a T_1 exists then so also do later times $T_2, T_3, \ldots,$ having the same property as T_1.

A stochastic process with the property (A.11) is called a <u>regenerative process</u>, and the times $T_1, T_2, \ldots,$ are called <u>regeneration times</u> for the process. Clearly the sequence of time intervals $(T_1 - T_0), (T_2 - T_1), (T_3 - T_2), \ldots,$ (where $T_0 \triangleq 0$), is a renewal process. The segment of the process $\{X_t\}$ on the interval $[T_{i-1}, T_i)$ is called the <u>i^{th} cycle</u> of $\{X_t\}$.

Now define the indicator variable

$$\begin{cases} Y_j(t) \triangleq 1, & X_t = j \text{ and } T_1 > t \\ \triangleq 0, & \text{otherwise} \end{cases}$$

Then $\int_0^\infty Y_j(t)\,dt$ is the amount of time for which $X_t = j$ during the first cycle of $\{X_t\}$. Furthermore, if

$$q_j(t) \triangleq \int_t^\infty P\left[X_t = j \,\middle|\, T_1 = s\right] dF(s) \qquad \text{....(A.12)}$$

then

$$E\left[\int_0^\infty Y_j(t)\;dt\right] = \int_0^\infty E\left[Y_j\,(t)\right] dt$$

$$= \int_0^\infty P\left[X_t = j \,\middle|\, T_1 > t\right] dt \qquad \text{....(A.13)}$$

and

$$P\left[X_t = j\,, T_1 > t\right] = \int_0^\infty P\left[X_t = j,\ T_1 > t \,\middle|\, T_1 = s\right] dF(s)$$

$$= \int_t^\infty P\left[X_t = j \,\middle|\, T_1 = s\right] dF(s)$$

$$= q_j(t) \qquad \text{....(A.14)}$$

Note that since $\{X_t\}$ is regenerative $E\left[\int_0^\infty Y_j(t)\;dt\right]$ is the expected time spent in state j during <u>any</u> cycle of $\{X_t\}$.

Now define, for each $t \in \mathcal{T}$ and each $j \in \mathbb{Z}_+$,

$$P_j(t) \triangleq P\left[X_t = j\right] \quad ;$$

then

$$P_j(t) = \int_0^\infty P\left[X_t = j \,\middle|\, T_1 = s\right] dF(s)$$

$$= \int_0^t P_j(t-s)\;dF(s) + \int_t^\infty P\left[X_t = j \,\middle|\, T_1 = s\right] dF(s)$$

$$= \int_0^t P_j(t-s)\;dF(s) + q_j(t) \qquad \text{....(A.15)}$$

ie. $P_j(t)$ satisfies a renewal-type equation (see A.6) whose solution (see A.9) is

$$P_j(t) = q_j(t) + \int_0^t q_j(t - x) \, dm(x) \qquad \text{....(A.16)}$$

But $q_j(t)$ is a non-negative, non-increasing function of t such that (provided $E\left[T_1\right] < \infty$) $\int_0^\infty q_j(t) \, dt < \infty$; it follows by the key renewal theorem that (again, provided that $E\left[T_1\right] < \infty$)

$$\lim_{t \to \infty} P_j(t) = \frac{\int_0^\infty q_j(t) \, dt}{E\left[T_1\right]}$$

ie. on using (A.13) and (A.14)

$$\lim_{t \to \infty} P_j(t) = \frac{E\left[\text{Time in state j during one cycle}\right]}{E\left[\text{Duration of one cycle}\right]}$$

$$\text{....(A.17)}$$

This result is true for any regenerative process. More generally, for a so-called delayed regenerative process, in which the time at which the first cycle of the process starts is $T_0 > 0$, the result (A.17) remains true but the expected duration of a cycle is then measured by $E\left[T_1 - T_0\right]$.

Now if $\{X_t\}$ is a finite-state, regular semi-Markov chain in which the state j is recurrent, with mean recurrence time μ_{jj}, we can regard $\{X_t\}$ as a delayed regenerative process in which the regeneration times T_1, T_2,..., are the times of entry into state j. Then by (A.17), for any initial state i,

$$\lim_{t \to \infty} P\left[X_t = j \,\middle|\, X_0 = i\right] = \frac{\tau_j}{\mu_{jj}} \qquad \text{....(A.18)}$$

where τ_j is the mean sojourn time in state j.

ie. $P_j(t)$ satisfies a renewal-type equation (see A.6) whose solution (see A.9) is

$$P_j(t) = q_j(t) + \int_o^t q_j(t - x) \; dm(x) \qquad \text{....(A.16)}$$

But $q_j(t)$ is a non-negative, non-increasing function of t such that (provided $E\left[T_1\right] < \infty$) $\int_o^\infty q_j(t)\, dt < \infty$; it follows by the key renewal theorem that (again, provided that $E\left[T_1\right] < \infty$)

$$\lim_{t \to \infty} P_j(t) = \frac{\int_o^\infty q_j(t)\, dt}{E\left[T_1\right]}$$

ie. on using (A.13) and (A.14)

$$\lim_{t \to \infty} P_j(t) = \frac{E\left[\text{Time in state j during one cycle}\right]}{E\left[\text{Duration of one cycle}\right]} \qquad \text{....(A.17)}$$

This result is true for any regenerative process. More generally, for a so-called delayed regenerative process, in which the time at which the first cycle of the process starts is $T_o > 0$, the result (A.17) remains true but the expected duration of a cycle is then measured by $E\left[T_1 - T_o\right]$.

Now if $\{X_t\}$ is a finite-state, regular semi-Markov chain in which the state j is recurrent, with mean recurrence time μ_{jj}, we can regard $\{X_t\}$ as a delayed regenerative process in which the regeneration times T_1, T_2,..., are the times of entry into state j. Then by (A.17), for any initial state i,

$$\lim_{t \to \infty} P\left[X_t = j \,\middle|\, X_o = i\right] = \frac{\tau_j}{\mu_{jj}} \qquad \text{....(A.18)}$$

where τ_j is the mean sojourn time in state j.

This is property R.8 (equation 2.41) in Chapter 2.

In order to evaluate μ_{jj} note first that, if T_{ij} is the first passage time from i to j and $\mu_{ij} = E\left[T_{ij}\right]$, then, denoting the embedded Markov chain by $\{\overline{X}_n\}$ as usual,

$$\mu_{ij} = \sum_{k} E\left[T_{ij} \,\middle|\, \overline{X}_o = i\,,\ \overline{X}_1 = k\right] p_{ik}$$

$$= \sum_{k\neq j} p_{ik}(\eta_{ik} + \mu_{kj}) + p_{ij}\,\eta_{ij} \qquad \text{....(A.19)}$$

where

$$\eta_{ij} \triangleq \int_o^{\infty} t\ dF_{ij}(t)$$

$$= E\left[\text{transition time from i to j}\right]$$

Now

$$\sum_{k} p_{ik}\,\eta_{ik} = \tau_i$$

so that, from (A.19),

$$\mu_{ij} = \tau_i + \sum_{k\neq j} p_{ik}\,\mu_{ik} \qquad \text{....(A.20)}$$

Multiply both sides of (A.20) by the stationary probability π_i and sum over i : we obtain

$$\sum_{i} \pi_i\,\mu_{ij} = \sum_{i} \pi_i\,\tau_i + \sum_{i} \pi_i \sum_{k\neq j} p_{ik}\,\mu_{kj}$$

$$= \sum_{i} \pi_i\,\tau_i + \sum_{k\neq j} \pi_k\,\mu_{kj}\ ,$$

since $\sum_{i} \pi_i\, p_{ik} = \pi_k$.

Finally, subtracting $\sum_{k\neq j} \pi_k\,\mu_{kj}$ from each side,

$$\pi_j \mu_{jj} = \sum_i \pi_i \tau_i$$

ie.

$$\mu_{jj} = \frac{\sum_i \pi_i \tau_i}{\pi_j} \qquad \text{....(A.21)}$$

This is property R.7 (equation 2.40) in Chapter 2. The equilibrium probabilities σ_i for the semi-Markov chain $\{X_t\}$ are obtained immediately by using (A.21) in (A.18):

$$\boxed{\sigma_j = \frac{\pi_j \tau_j}{\sum_i \pi_i \tau_i}} \qquad \text{....(A.22)}$$

II. Semi-Markov Chains with Costs

Let $\{(X_t : \Omega \to \mathbb{N}_N) : t \in \mathcal{T}\}$ be a semi-Markov chain with semi-Markov kernel F. Associate with each transition $X_{T_n} \to X_{T_{n+1}}$ the transition cost

$$C(X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1})$$

and, for each $i \in \mathbb{N}_N$, define

$$\gamma_i \triangleq E\left[C(X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}) \,\middle|\, X_{T_n} = i\right] \qquad \text{....(A.23)}$$

$$\underline{\gamma} \triangleq \text{Col}(\gamma_1, \gamma_2, \ldots, \gamma_N) \qquad \text{....(A.24)}$$

$$v_i(n) \triangleq E\left[\sum_{k=k_o}^{k_o+n-1} C(X_{T_k}, X_{T_{k+1}}, \Delta T_{k+1}) \,\middle|\, X_{T_{k_o}} = i\right] \qquad \text{....(A.25)}$$

$$\underline{v}(n) \triangleq \text{Col}(\underline{v}_1(n), \ldots, v_N(n)) \qquad \text{....(A.26)}$$

Then, separating out the first term in (A.25) ,

$$v_i(n) = \gamma_i + E\left[\sum_{k=k_o+1}^{k_o+n-1} C(X_{T_k}, X_{T_{k+1}}, \Delta T_{k+1}) \,\middle|\, X_{T_{k_o}} = i\right]$$

$$= \gamma_i + \underset{X_{T_{k_o+1}}}{E}\left[E\left[\sum_{k=k_o+1}^{k_o+n-1} C(X_{T_k}, X_{T_{k+1}}, \Delta T_{k+1}) \,\middle|\, X_{T_{k_o}} = i, X_{T_{k_o+1}}\right]\right]$$

$$= \gamma_i + \underset{X_{T_{k_o+1}}}{E}\left[v_{X_{T_{k_o+1}}}(n-1) \,\middle|\, X_{T_{k_o}} = i\right]$$

$$= \gamma_i + \sum_j p_{ij}\, v_j(n-1) \qquad \text{....(A.27)}$$

ie. using (A.24) and (A.26)

$$\boxed{\underline{v}(n) = \underline{\gamma} + P\,\underline{v}(n-1)} \qquad \text{....(A.28)}$$

Similarly, if

$$t_i(n) \triangleq E\left[(T_{k_o+n} - T_{k_o}) \,\middle|\, X_{T_{k_o}} = i\right] \qquad \text{....(A.29)}$$

and

$$\underline{t}(n) \triangleq \mathrm{Col}(t_1(n),\ldots,t_N(n)) \qquad \text{....(A.30)}$$

then taking the transition cost in (A.25) to be

$$C(X_{T_n}, X_{T_{n+1}}, \Delta T_{n+1}) = \Delta T_{n+1}$$

immediately yields, as the corresponding version of (A.28) ,

$$\boxed{\underline{t}(n) = \underline{\tau} + P\,\underline{t}(n-1)} \qquad \text{....(A.31)}$$

III. The Equilibrium Mean Cost Rate

Let $C_i(t)$ be the total cost accumulated between T_{k_o} and $T_{k_o} + t$, given that the recurrent state i was entered at time T_{k_o}, and let $v_i'(t) \triangleq E\left[C_i(t)\right]$.

Then (see Ross(1970) - theorem 3.16)

$$\lim_{t \to \infty} \left[\frac{v_i'(t)}{t}\right] = \frac{E\left[\Delta C_i\right]}{E\left[\Delta T_i\right]} \qquad \text{....(A.33)}$$

where $\begin{cases} \Delta C_i = \text{cost incurred on cycle starting in state i} \\ \Delta T_i = \text{duration of a cycle starting in state i} \end{cases}$

But $E\left[\Delta T_i\right]$ is just the mean recurrence time of state i, and so (by A.21)

$$E\left[\Delta T_i\right] = \frac{\sum_i \pi_i \tau_i}{\pi_i} \qquad \text{....(A.34)}$$

If now we interpret the mean one step costs γ_i as the mean sojourn times of a fictitious semi-Markov chain with the same embedded chain, $\{\overline{X}_n\}$, as $\{X_t\}$, we then have that $E\left[\Delta C_i\right]$ is the mean recurrence time of state i for the fictitious chain, and hence that

$$E\left[\Delta C_i\right] = \frac{\sum_i \pi_i \gamma_i}{\pi_i} \qquad \text{....(A.35)}$$

Then using (A.34), (A.35) in (A.33),

$$\boxed{\lim_{t \to \infty} \left[\frac{v_i'(t)}{t}\right] = \frac{\sum_i \pi_i \gamma_i}{\sum_i \pi_i \tau_i}} \qquad \text{....(A.36)}$$

(If not all the γ_i are positive it is necessary to modify the argument slightly by considering $(\gamma_i + \gamma)$ to be recurrence times, where γ is some constant such that $(\gamma_i + \gamma) > 0$ for each recurrent state i.)