PERTURBATION THEORY AND MARKOVIAN DECISION PROCESSES

by

PAUL JEROME SCHWEITZER

B.S. Physics, B.S. Mathematics
Massachusetts Institute of Technology

(1961)

SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF

SCIENCE

at the

MASSACHUSETTS INSTITUTE OF

TECHNOLOGY

May 1965

Signature of Author . . . . . .
Department of Physics, May 25, 1965

Certified by . . . . . .
Thesis Supervisor

Accepted by . . . . . .
Chairman, Departmental Committee
on Graduate Students

PERTURBATION THEORY AND MARKOVIAN DECISION PROCESSES

by

PAUL JEROME SCHWEITZER

Submitted to the Department of Physics on May 20, 1965
in partial fulfillment of the requirements for the degree of
Doctor of Science

ABSTRACT

The Howard-Jewell algorithm for programming over a Markov-renewal process is analyzed in terms of a perturbation theory formalism which describes how the stationary distribution changes when the transition probabilities change. The policy improvement technique is derived from this new viewpoint. The relative values may be interpreted as partial derivatives of the gain rate with respect to policy.

The value equations are shown to be solvable, with the relative values unique up to one additive constant, if and only if the underlying Markov chain is irreducible. The policy iteration algorithm is shown not to cycle, thus guaranteeing convergence.

A discussion of the existence, uniqueness, and characterization of the solution to the functional equation of dynamic programming is given. Emphasis is placed upon the value-maximization of transient states.

The fundamental matrix is developed as a useful tool for doing perturbation theory, describing first-passage properties of semi-Markov processes, and for dealing with semi-Markov processes with rewards.

Thesis Supervisor: Dr. Philip M. Morse
Title: Professor of Physics

ACKNOWLEDGEMENTS

The author owes a deep debt of gratitude to Professor Ronald A. Howard, whose book "Dynamic Programming and Markov Processes" and lectures opened his eyes to the mathematics of decision-making, and whose insight and encouragement helped him through several difficult passages.

Professors Philip M. Morse and John D. C. Little deserve a sincere thank you for many hours of useful discussion and for critical reading of the manuscript.

Financial support by the National Science Foundation made the author's graduate studies possible, and is gratefully acknowledged.

Several computations were carried out at the M.I.T. Computation Center.

The cheerfulness and excellent secretarial work of Mrs. Nobu McNeill and Mrs. Karen Murley greatly simplified the chore of manuscript preparation.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF FIGURES

CHAPTER 1

INTRODUCTION

1A. Formulation of Problem

We consider a system which can be in any one of a multitude of states. In each state a decision must be made as to which of a multitude of alternative strategies is to be used. Once the decision is made, the system earns a random reward and makes a random transition, after some random holding time, to another state. The distributions of rewards, holding times, and terminal states depend only on the initial state and decision.

A choice of strategy in each state determines a policy. Once the policy is specified, the description of system behavior via the embedded chain of transitions is a Markovian one. The description in continuous time is that of a Markov renewal process[1] or semi-Markov process.

The basic problem of programming over such a Markov renewal process is the determination of the stationary policy which has the largest expected reward per unit time, if the rewards are undiscounted, or the largest total expected reward if future rewards are discounted. Jewell calls this Markov renewal programming.

A second problem is the determination of the optimal time-dependent policy if the system is to earn rewards for a specified length

of time, after which a terminal reward ("scrap value") is paid depending upon the state at the time of termination. This can usually be accomplished by the iterative technique of dynamic programming known as the principle of optimality (Ref. 2, Pg. 83):

An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

A third problem is the investigation of conditions under which the optimal time-dependent policy approaches the optimal stationary policy as the time duration of the process becomes very large.

1B. Literature Survey

The earliest reference to Markovian reward processes known to the author occurs in a remark of Smith[3] that just as the transition probability matrix P operating to the left give the probabilities for the process, so too does that P matrix operating to the right give the reward behavior (values).

The first problem, that of programming over a Markov chain was first posed by Bellman[4], who established the concept of gain rate and also derived equation (1.1) under the assumption that $P > 0$.

The major breakthrough in this field is due to Howard whose doctoral thesis and lucid book[5] gave a detailed discussion of the concept of gain rate and gave the first presentation of finite algorithms for finding optimal stationary policies for both discounted and non-discounted rewards, by improvement in policy-space. These held for systems making one transition per unit time (Markov chains) and for Markov processes in continuous time with exponentially distributed holding times.

The extension of Howard's policy-iteration algorithm to Markov renewal processes was made simultaneously and independently by Jewell[1], Howard[6], de Cani[7] and the author, all of whom noted

that merely the mean holding time is required. Jewell has a lucid description of semi-Markov processes, with references to the pioneering work by Pyke. The reader is assumed to be familiar with both Howard's[5, 6] and Jewell's work[1].

An alternate formulation by linear programming is due to Manne[8] and Wolfe and Dantzig[9]. Wagner[10] used linear programming to show that optimal policies are pure, never randomized. Blackwell[11] and Jewell have investigated the case of near-vanishing interest rate, nearly infinite number of transitions, and nearly infinite time durations. Derman has looked at randomized strategies. A good bibliography of recent papers containing these references and others is found in Reference 1. In addition, Jewell has investigated[12] the case of ties--policies with equal expected reward per unit time--and has shown[13] that the relative values, although inadequate for finding the policy with largest total expected reward, are adequate for finding the policy with minimum variance of the reward per unit time.

Markov processes in continuous time with rewards are probably best handled by the techniques of optimal control theory, for example, by Pontryagin's maximum principle.

The second problem, called value-iteration by Howard, is found scattered throughout the literature, for example in studies by Bellman, Glicksberg and Gross[14] and by Iglehart[15] on the asymptotic

behavior of the optimal inventory equation. A recent paper by Derman and Klein[16] showed that the finite-horizon problem can be set up via linear programming, and moreover the transition probabilities need not be time-independent.

Very little work has been done on the third problem, except for some situations involving discounting. White[17] has given a result equivalent to the theorem that if some state m is accessible from all others after u transitions, regardless of policy, then the vector of total rewards for the optimal time-dependent policy and the vector of total rewards for the optimal stationary policy differ asymptotically by a vector all of whose components are identical (see corollary to Theorem 10.5) as the number of remaining steps goes to infinity.

A new aspect of the problem, a selection of policies if the transition probabilities are not known exactly but rather have distributions placed upon them, is solved by Bayesian techniques developed at M.I.T. by Cozzolino, Gonzalez-Zubieta, and Miller[18] and by Martin[19].

1C. Present Results

This thesis continues the investigation into the structure of Markovian decision processes. While many general results about convergence and asymptotic behavior are derived, no special structure--such as Howard's formulation of the replacement problem--is looked at.

The topics dealt with herein fall in three distinct classes. First we resolve some purely mathematical questions which were not explicitly raised in previous treatments of Markov renewal programming by Howard and Jewell. For example, we give necessary and sufficient conditions for the value-determination equations to be solvable, that is, possess a non-vanishing system determinant. We show that the Howard-Jewell algorithm for policy iteration in the nondiscounted case converges, supplying a proof that cycling is impossible which has been omitted so far. The relation of this algorithm to linear programming is touched upon.

We show that the relative values are involved in policy improvement in a purely mathematical fashion with no need of an interpretation as a limiting intercept. They may be interpreted as partial derivatives of the gain rate with respect to policy. The policy improvement routine may then be interpreted as technique for finding a new policy by chosing a ray in policy space in whose direction the directional derivative is

largest. This is the explanation for the test quantities. It then becomes clearer why the values, properties of one policy, are useful for policy improvement. In addition, the mathematical properties of the functional equation

$$v_i = \max_{1 \le k \le n_i} \left[q_i^k - g T_i^k + \sum_{j=1}^{N} p_{ij}^k v_j \right] \qquad 1 \le i \le N \qquad (1.1)$$

are studied. A sufficient condition for existence of a solution is given. So are uniqueness proofs: g is unique, as the largest possible gain rate, and the v_i are unique up to an additive constant. We also continue Howard's discussion of the maximization of the relative values of transient states, and prove rigorously that this is so: the v_i are "as large as possible".

Again in the formal vein, expressions for the absolute values are obtained and compared with previous results of Jewell[1]. A simple proof of the theorem that a pure (not randomized) policy achieves the highest possible gain rate is presented. We show that the relative values for a policy are equal if and only if the expected reward per transition is, for all states, proportional to the expected holding time.

A decomposition of a Markov chain into steady-state and transient components, $p^n = p^\infty + E^n \quad n \ge 1$, and a factorization of the singular kernel $I - P = (I-p^\infty)(I-E) = (I-p^\infty)Z^{-1}$ into singular and regular

components are presented. These are used over and over and deserve mastery by operations researchers interested in stationary Markov chains. Many additional properties of the fundamental matrix are discussed.

The properties of the operator $T : E_N \rightarrow E_N$ defined by

$$(Tf)_i = \max_{1 \leq k \leq n_i} \left[b_i^k + \sum_{j=1}^{N} p_{ij}^k f_j \right] \qquad 1 \leq i \leq N \qquad (1.2)$$

where

$$\max_{1 \leq k \leq n_i} \left[b_i^k \right] = 0 \qquad 1 \leq i \leq N \qquad (1.3)$$

are discussed, and four important results,

$$\| Tf \| \leq \| f \|$$

$$\| Tf - Th \| \leq \| f - h \|$$

$$T^n \underline{f} \rightarrow c \underline{1} \qquad \text{as } n \rightarrow \infty$$

$$T\underline{f} = \underline{f} \qquad \text{if and only if } \underline{f} = c \underline{1}$$

are derived, the last two holding in general only if all policies are ergodic.

While the dry mathematical format used in resolving these technical questions lacks the motivation and intuitive appeal of Howard's arguments, it has the advantage of clearly separating the purely mathematical consequences from the postulates.

Second, we develop a new approach to Markov renewal programming. We use a perturbation technique (calculus of variations) which provides an alternative to dynamic programming for deriving the algorithm for gain maximization. The values and test quantities appear in a different way than previously, and the alternate approach supplies fresh insight.

Formulas for $\frac{\partial \pi_i}{\partial p_{mn}}$ and $\frac{\partial g}{\partial p_{mn}}$ are presented and group-theoretic implications of this perturbation approach are touched upon. A method of steepest ascent is suggest for finding optimal parameter settings, for example the values of the trigger level and reorder amount in an (s, S) inventory system.

The fundamental matrix plays an important role in this perturbation technique, and is shown to be closely connected with the ergodic behavior of the chain. It is developed here as a useful tool.

Third, as many of our results as possible are stated for the case of a continuum of states as well as for the N-state case. A unified description is made possible by use of an abstract operator notation and by appeal to some results of functional analysis, especially the Fredholm theory of integral operators.

The continuum generalization was not carried out for its own sake--even though the abstract operator notation is easier to read than more explicit expressions, and even though the inventory equation assumes a continuum of states. It was undertaken because of the observation that

many Markovian decision problems occur with thousands of states where these states may most naturally be thought of as occupying a lattice in some multidimensional Euclidean space. Furthermore, the rewards and holding times vary slowly from state to state and the relative values will also. It then seems desirable to (1) approximate the thousands of simultaneous linear equations for the values by a single integral equation with continuous kernel (2) quantize the kernel of the integral equation into a kernel of finite rank, hopefully, by a clever choice of basis functions, a much smaller rank than the original rank of several thousand of the coefficient matrix for the simultaneous linear equations.

It is part (1) which involves a continuum of states. If the lattice points are dense enough, the integral equation is a good approximation to the simultaneous linear equations. Hence the method should work best with problems involving many states.

Such a situation was encountered in solving a single-server queueing problem with a queue capacity of $m = 8$, and two classes of customers with Poisson arrival rates λ_1, λ_2, mean service time T_1, T_2, and waiting cost-time C_1, C_2. A state for the imbedded chain of epochs where the server has to decide which class to service next is described by the vector (n_1, n_2) where n_i = number on queue of type i. Since $n_1 + n_2 \leq m = 8$, there are $\frac{(m+1)(m+2)}{2} = 45$ states.

The optimal policy (minimum cost per unit time) can be found, once all λ's, T's and C's are specified numerically, by the policy iteration algorithm. One finds that the relative values $v(n_1, n_2)$ for the 45 states vary smoothly, when plotted on a two dimensional grid. Indeed a 2^{nd} or 3^{rd} degree polynomial in n_1 and n_2 least squares fit yields $v(n_1, n_2)$ to one per cent. This remains true even though the traffic intensity factor $\rho = \lambda_1 T_1 + \lambda_2 T_2$ was varied from 10^{-3} to 20. The smoothness is encouraging because the variation in ρ changes the frequency of occurrence of the various states from 99.9 per cent occupancy in (0, 0), (0, 1), (1, 0) when $\rho = 10^{-3}$ to 99.9 per cent occupancy of the states (0, 8) and (1, 7) when $\rho = 20$. The optimal policy for the finite queue capacity case turns out to be the same as for the infinite queue capacity case: service first the class whose C_i/T_i ratio is highest.[20]

The important observations to be made about this problem are the following:

(1) the 45 state points lie naturally in a lattice in two dimensional space. Similarly inventory problems when quantized lead to lattice-state-spaces.

(2) the 45 states can be numbered in an <u>arbitrary</u> one-dimensional fashion when the policy iteration algorithm is used. This is wasteful because valuable geometric insight is lost when the results come off the computer in a one-dimensional array.

(3) if m were much larger than 8, we would have hundreds of states and equations, and brute force numerical solution by the Howard-Jewell algorithm is out of the question. But due to the simple lattice structure, we would expect a continuum approximation to be reasonable.

Howard's taxicab problem, (Ref. 5, Chapter 5) in which the driver must decide at every street-corner in each of the three cities which strategy to follow is another example of a natural lattice ordering. The states may be considered as points in a two-dimensional region--namely a map of the three cities. The continuum-state version is obtained when the driver must decide at every point upon a strategy to follow. Many pursuit and search problems (minimum expected time to capture or detection) have similar structures.

Stated more formally, our objection to the usual Howard-Jewell algorithm is that if the states are numbered in a one-dimensional sequence, any natural ordering or smoothness of behavior is lost. In order to have a topological structure (ordered states), a lattice or continuum of state approach is needed.

Since the continuum approach is somewhat unfamiliar, explicit results concerning the chain structure, eigenvalue spectrum, and so forth have been collected together to provide some insight.

An alternate way of grouping the present results is in terms of the three problems stated in Section 1B.

(1) The Markov renewal programming algorithm and functional equation (1.1) are discussed in great detail, as described above.

(2) Since no special reward structures are postulated in this thesis, nothing can be said about the form of the policy obtained by solution of the value-iteration equations.

(3) Several results have been obtained about the asymptotic behavior of value-iteration. We are able to prove Howard's conjecture, that the solution $v_i(n)$ of the equation

$$v_i(n+1) = \max_{1 \le k \le n_i} \left[q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j(n) \right] \qquad 1 \le i \le N \qquad (1.3)$$

has the asymptotic behavior, if all policies are ergodic,

$$\lim_{n \to \infty} \left[v_i(n) - ng - v_i \right] = c \qquad 1 \le i \le N \qquad (1.4)$$

where v_i and g are given by (1.1).

From this follows both White's theorem and also convergence of the policy to (one of) the optimal stationary policy. Thus a quantized inventory problem, for example, will have order amounts and trigger levels which become independent of time if the process is many transitions from termination.

Equation (1.4) implies that the gain rate of the optimal time-dependent policy and the gain rate of the optimal time-dependent policy agree. Hence Howard's restriction to stationary policies is justified.

In the discounted case, convergence of the total expected reward for the optimal time-dependent policy to that of the optimal stationary policy is proved, along with policy convergence, as the remaining number of transitions becomes large. This justifies Howard's restriction to stationary policies. In addition, the case of vanishing interest rate is investigated.

CHAPTER 2

THE MODEL

2A. Introduction

In this chapter the semi-Markov model with alternatives is presented. The N-state case and the continuum-state case are presented separately for pedagogical reasons, although may be treated simultaneously by use of the Stieljies integral.

A constant interest rate is assumed, so that if a discount factor a is defined as the present value of a dollar received one time period hence, then a^t is the present value of a dollar received at a time t in the future. The equations for the discounted case $(0 \leq a < 1)$ and the undiscounted case $(a = 1)$ are derived simultaneously.

2B. The N-State Semi-Markov Model: Value-Iteration Equations

Consider a system which has N states labeled 1, 2, ... N. At the instant of entering state i, one of n_i alternative strategies must by selected. If the k^{th} alternative is selected ($1 \leq k \leq n_i$), then there is a probability $p_{ij}^{\ k}(t)dt$ that the system leaves state i for the first time after a duration of length t to t+dt and enters state j ($1 \leq j \leq N$). We assume that $p_{ij}^{\ k}(t)$ has no impulse component at the origin.

Of major interest are the probability $p_{ij}^{\ k}$ of a transition from i to j, the product $T_{ij}^{\ k}$ of $p_{ij}^{\ k}$ with the mean time for a transition from i to j, and the mean holding time $T_i^{\ k}$ in state i. These are defined by

$$p_{ij}^{\ k} = \int_0^\infty dt\, p_{ij}^{\ k}(t) \tag{2.1}$$

$$T_{ij}^{\ k} = \int_0^\infty dt\, p_{ij}^{\ k}(t)t \tag{2.2}$$

$$T_i^{\ k} = \sum_{j=1}^{N} \int_0^\infty dt\, p_{ij}^{\ k}(t) = \sum_{j=1}^{N} T_{ij}^{\ k} \tag{2.3}$$

$$1 \leq i \leq N \quad 1 \leq k \leq n_i$$

we assume that these are all finite and that probability is conserved:

$$\sum_{j=1}^{N} p_{ij}^{\ k} = 1 \qquad 1 \leq i \leq N \quad 1 \leq k \leq n_i \tag{2.4}$$

The finiteness of the mean holding time T_i guarantees that a transition out of state i is certain to occur if one waits long enough.

Indeed by the Chebyshev inequality,

$$\Pr(\text{holding time in } i \geq t) = \sum_{j=1}^{N} \int_{t}^{\infty} dt'\, p_{ij}(t')$$

$$\leq \sum_{j=1}^{N} \int_{t}^{\infty} dt'\, \frac{t'}{t}\, p_{ij}(t')$$

$$\leq \sum_{j=1}^{N} \int_{0}^{\infty} dt'\, \frac{t'\, p_{ij}(t')}{t}$$

$$= \frac{T_i}{t}$$

So called "virtual transitions" from a state i back to itself are permitted in this model, since we nowhere assume that $p_{ii} = 0$.

As Jewell has pointed out, a convenient way to think of the transition mechanism out of i if the k^{th} alternative is used is to imagine first that the final state j is chosen with probability p_{ij}^{k} and second that the holding time t is chosen with probability density $p_{ij}^{k}(t)/p_{ij}^{k}$. The imbedded chain of transitions is Markovian, hence the name semi-Markov process.

Any reward structure (lump payments at the beginning or end of the sojourn in i, or at a rate $r_{ij}^{k}(t)$ dependent upon the duration t so far) is permitted. We merely need to know

$q_i^k(t, a)$ = immediate expected reward if process terminates at time t after the system enters state i, and if k^{th} alternative is used. $i \leq i \leq N$, $1 \leq k \leq n_i$. (2.5)

and

$$q_i^k(a) = \lim_{t \to \infty} q_i^k(t, a) \tag{2.6}$$

These are the earnings (discounted via a) only while in state i and do not include the earnings in any state j to which a transition was made from i.

We assume that the $q_i^k(t, a)$ are finite and the limit in (2.6) exists.

A simple example is the Markov case,

$$p_{ij}^k(t) = p_{ij}^k \, \delta(t-1)$$

where transitions occur precisely one unit of time apart. If the system is in i and is headed for j, we assume rewards are earned at a rate r_{ij}. If there is no discounting (a=1) and if the terminal value (scrap value) of being in state i is $v_i(0)$, then

$$q_i^k(t, a=1) = \begin{cases} \sum_{j=1}^{N} p_{ij} \, r_{ij} \, t + v_i(0) & t < 1 \\ \sum_{j=1}^{N} p_{ij} \, r_{ij} & t \geq 1 \end{cases} \tag{2.7}$$

$$q_i^k(a=1) = \sum_{j=1}^{N} p_{ij} \, r_{ij}$$

It is a general result that the scrap values $v_i(0)$ do not enter $q_i^k(a)$ since a transition out of i is certain to occur if one waits long enough.

We will use the notation $q_i^k = q_i^k(a=1)$ for the undiscounted case.

A specification of the alternative chosen for each of the N states is called a pure policy. If the k_i^{th} alternative is chosen in state i, then the N-vector $\underline{k} = (k_1, k_2, \ldots k_N)$ is called the decision vector. There are NN pure policies, where

$$NN = \prod_{i=1}^{N} n_i$$

Capital letters A, B, C, ... will be used to denote pure policies. Thus policy A means $\underline{k}_A = (k_{A_1}, k_{A_2}, \ldots k_{A_N})$ is given. Randomized policies and alternatives will be discussed later.

Once a policy A has been selected, one can speak of the expected return $v_i^A(t)$ in the next t using policy A and starting in state i. It satisfies the equation

$$v_i^A(t, a) = q_i^A(t, a) + \sum_{j=1}^{N} \int_0^t dt' \, p_{ij}^A(t') a^{t'} v_j^A(t-t', a) \qquad (2.8)$$

$$1 \leq i \leq N \quad t \geq 0$$

where the first term describes expected earnings in state i and the

second term describes expected earnings thereafter starting from a state j reached from i after a holding time t'.

In the Markov case, $p_{ij}(t) = p_{ij}\,\delta(t-1)$, and one speaks of the expected return $v_i^A(n, a)$ in n transitions, starting from i and using policy A. It satisfies the equation(Ref. 5, eq. 7.2)

$$v_i^A(n+1, a) = q_i^A(a) + a \sum_{j=1}^{N} p_{ij}\, v_j^A(n, a) \qquad (2.9)$$

$$1 \leq i \leq N \qquad n \geq 0$$

where $q_i^A(a)$ is the expected discounted reward in one transition, excluding the scrap values $v_i^A(0)$. The scrap values enter via

$$v_i^A(0, a) = v_i^A(0) \qquad 1 \leq i \leq N \qquad (2.10)$$

It is understood in (2.9-2.10) that $p_{ij}^A(t)$ and $q_i^A(t, a)$ depend only on k_i^A and not on the entire decision vector $\underline{k}^A$.

Equations (2.8-2.9) hold for a stationary policy A, that is, a policy A independent of time. If we believe that a time-dependent policy gives sufficiently higher reward to merit consideration, we could find the optimal time-dependent policy by the following considerations.

If $v_i(t, a)$ (or $v_i(n, a)$) denotes the maximum possible expected return, if the system has just entered state i and if the process will terminate after duration t (or after n transitions) then the principle of optimality leads to

$$v_i(t, a) = \max_{1 \leq k \leq n_i} \left[q_i^k(t, a) + \sum_{j=1}^{N} \int_0^t dt'\, p_{ij}^k(t')\, a^{t'}\, v_j(t-t') \right] \qquad (2.11)$$

$$1 \leq i \leq N \qquad t \geq 0$$

and

$$v_i(n+1, a) = \max_{1 \le k \le n_i} \left[q_i^k(a) + a \sum_{j=1}^{N} p_{ij}^k \, v_j(n, a) \right] \qquad (2.12)$$

$$1 \le i \le N \qquad n \ge 0$$

for the semi-Markov and Markov cases, respectively.

Equations (2.8, 2.9, 2.11, 2.12) are called value-iteration equations by Howard because they are sequential.

As Jewell[1] has pointed out, the a in Equation (2.9) can be interpreted as the probability of not terminating after a transition, while in (2.8) the quantity ln 1/a can be interpreted as the probability per unit time of terminating the process.

2C. N-State Semi-Markov Model: Asymptotic Behavior

We conjecture (see below) that for a fixed policy A,

$$v_i^A(t, a) \to \begin{cases} v_i^A(a) & a < 1 \quad (2.13a) \\ g_i^A t + v_i^A & a = 1 \quad (2.13b) \end{cases}$$

as $t \to \infty$

$$v_i^A(n, a) \to \begin{cases} v_i^A(a) & a < 1 \quad (2.13c) \\ g_i^A n + v_i^A & a = 1 \quad (2.13d) \end{cases}$$

as $n \to \infty$

where the total rewards are finite in the discounted cases (a, c) and diverge linearly with time in the undiscounted cases (b, d). The g_i^A in the undiscounted cases are called gains using policy A and are interpretable as the expected reward per unit time in the semi-Markov case b and as expected reward per transition in the Markov case d. In this thesis only the single-chain case is considered so that the gain rate g_i^A is assumed to be independent of i.

The v_i^A are called values or absolute values and, in b and d, are interpretable as asymptotic intercepts.

Howard has shown by z-transform that (2.13c, d) are correct in the Markov case, and that g_i^A is independent of i if p^A is ergodic.

Equation (2.13b) can be deduced from (2.11) by Laplace transform: see Section 5J. Equation (2.13a) can also be proved by Laplace transforms.

Similarly we conjecture (see below) that for the optimal time-dependent policy,

$$v_i(t, a) \to \begin{cases} v_i(a) & a < 1 \quad (2.14a) \\ g_i t + v_i & a = 1 \quad (2.14b) \end{cases}$$

as $t \to \infty$

$$v_i(n, a) \to \begin{cases} v_i(a) & a < 1 \quad (2.14c) \\ ng_i + v_i & a = 1 \quad (2.14d) \end{cases}$$

as $n \to \infty$

Equation (2.14c) is proved in Chapter 3, Equation (2.14d) is proved--and again g_i^A is independent of i--if all NN policies are ergodic in Chapter 10. At present, nothing is known about the correctness of (2.14a, b).

If (2.13-2.14) are inserted into (2.8, 2.9, 2.11, 2.12), one obtains the following infinite horizon equations for $1 \leq i \leq N$:

1) fixed policy, discounted Markov case

$$v_i^A(a) = q_i^A(a) + a \sum_{j=1}^{N} p_{ij}^A v_j^A(a) \quad (2.15a)$$

2) fixed policy, discounted semi-Markov case

$$v_i^A(a) = q_i^A(a) + \sum_{j=1}^{N} \int_0^\infty dt'\, p_{ij}^A(t') a^{t'} v_j^A(a)$$

$$= q_i^A(a) + \sum_{j=1}^{N} \tilde{p}_{ij}^A(\ln 1/a)\, v_j^A(a) \qquad (2.15b)$$

3) fixed policy, undiscounted Markov case

$$v_i^A = q_i^A - g^A + \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad (2.15c)$$

4) fixed policy, undiscounted semi-Markov case

$$v_i^A = q_i^A - g^A T_i^A + \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad (2.15d)$$

5) optimal time-dependent policy, discounted Markov case

$$v_i(a) = \max_{1 \le k \le n_i} \left[q_i^k(a) + a \sum_{j=1}^{N} p_{ij}^k v_j(a) \right] \qquad (2.15e)$$

6) optimal time-dependent policy, discounted semi-Markov case

$$v_i(a) = \max_{1 \le k \le n_i} \left[q_i^k(a) + \sum_{j=1}^{N} \int_0^\infty dt'\, p_{ij}^k(t')\, a^{t'} v_j(a) \right]$$

$$= \max_{1 \le k \le n_i} \left[q_i^k(a) + \sum_{j=1}^{N} \tilde{p}_{ij}^k(\ln 1/a)\, v_j(a) \right] \qquad (2.15f)$$

7) optimal time-dependent policy, undiscounted Markov case

$$v_i = \max_{1 \le k \le n_i} \left[q_i^k - g + \sum_{j=1}^{N} p_{ij}^k v_j \right] \tag{2.15g}$$

8) optimal time-dependent policy, undiscounted semi-Markov case

$$v_i = \max_{1 \le k \le n_i} \left[q_i^k - g T_i^k + \sum_{j=1}^{N} p_{ij}^k v_j \right] \tag{2.15h}$$

where the tilde stands for Laplace transform:

$$\tilde{f}(s) = \int_0^\infty dt\, f(t) e^{-st} \tag{2.16}$$

Some sleight-of-hand has gone into these derivations and must be justified. For example, in the undiscounted case, the second term in (2.8) is

$$\sum_{j=1}^{N} \int_0^t dt'\, p_{ij}^A(t')\, v_j(t-t')$$

For large t, one breaks the t' integral up into two pieces, $(0, T_0)$ and (T_0, t) where $t - T_0 >> 0$ and $T_0 >> 0$.

In the first piece, $t' \epsilon (0, T_0)$ and $t - t' >> 0$ so that the asymptotic form (2.13b) is justified. In the second piece, (2.13b) is not justified. However, the integral will vanish as T_0 and t approach ∞, provided v(t) grows no faster than linearly in t, because the finiteness of T_{ij}^A implies that

$$\int_t^\infty dt' \, p_{ij}{}^{A}(t') \, t' \to 0$$

as $t \to \infty$.

Equations (2.15a, b, c, d) are called the value-determination equations by Howard, and equations (2.15e, f, g, h) will be called functional equations. In the discounted case, the v_i are uniquely determined from the equations. In the undiscounted case, equations (2.15c, d, g, h) determine the v_i only up to an additive constant: if v_i is a solution, so is $v_i + c$. One arbitrarily picks the constant (a convenient choice is to set $v_N = 0$) and then deals with the so-called relative values.

Solution of the Value-Determination Equations: A Preview

1) Since $ap_{ij}^{A} \geq 0$ while $\sum_{j=1}^{N} ap_{ij}^{A} < 1$, the series

$$(I - aP^{A})^{-1} = I + \sum_{n=1}^{\infty} a^{n}(P^{A})^{n} \tag{2.17}$$

converges. The solution to (2.15a) is therefore

$$v_i^{A}(a) = \sum_{j=1}^{N} (I - aP^{A})^{-1}_{ij} q_j^{A}(a) \qquad 1 \leq i \leq N \tag{2.18a}$$

2) Similarly the solution to (2.15b) is

$$v_i^{A}(a) = \sum_{j=1}^{N} \left[I - \tilde{P}^{A}(\ln 1/a)\right]^{-1}_{ij} q_j^{A}(a) \qquad 1 \leq i \leq N \tag{2.18b}$$

3) Equation (2.15c) is subsumed under (2.15d) by setting $T_i^{A} = 1$.

4) The solution to (2.15d) is given in Section 5H and also by Theorem 5.9.

5) We show in Section 3C that the solution to (2.15e) is

$$v_i(a) = \max_{\text{all NN policies } B} \left[\sum_{j=1}^{N} (I - aP^{B})^{-1}_{ij} q_j^{B}(a)\right] \qquad 1 \leq i \leq N \tag{2.18c}$$

6) The solution to (2.15f) can, by the same arguments, be shown to be

$$v_i(a) = \max_{\text{all NN policies } B} \left[\sum_{j=1}^{N} \left[I - \tilde{P}^B (\ln 1/a) \right]^{-1}_{ij} q_j^B(a) \right] \quad (2.18d)$$

$$1 \leq i \leq N$$

7) Equation (2.15g) is subsumed under (2.15h) by setting $T_i^k = 1$ for all i and k.

8) The existence, uniqueness and characterization of the solution to (2.15h) is discussed in Chapter 8.

Optimal Stationary Policies: A Preview

Equations (2.15a, b, c, d) hold for any fixed policy A with an infinite horizon. If we are restricted to stationary policies, it is of interest to learn which stationary policies A are optimal.

In the discounted case, we would like, for each state i, to pick A to maximize the right hand side of (2.18a) or (2.18b). This can be done: the same policy A which maximizes for one i maximizes for all i. Furthermore, the $v_i(a)$ as well as the policy A, which one obtains by this procedure agree with those given by (2.18c) and (2.18d); the optimal stationary policy for the discounted Markov or semi-Markov cases agrees with the (stationary) policy which achieves the values for the optimal non-stationary policy, and the values agree as well. This justifies the restriction to stationary policies in the discounted infinite-horizon case.

Similarly in the undiscounted Markov case, the gain rate of the optimal time-dependent policy agrees with the maximum, overall NN policies, of the gain rate of the pure policies. The relative values also agree up to an additive constant. This justifies the restriction to optimal stationary policies in the undiscounted case.

Vanishing Interest Rate

Another way to handle the case of infinite horizon is to let the discount factor a approach 1. This is done in Section 8I, where we are able to relate (2.15e) to (2.15g) and (2.15f) to (2.15h).

Warning

The v's and g's in (2.15a-h) all have different numerical values and should not be considered equivalent. For example, the g is (2.15g) is the maximum expected reward per transition (π, q) while the g in (2.15h) is the maximum expected reward per unit time $(\pi, q)/(\pi, T)$.

2D. The Continuum-State Model

We consider the situation where the state of the system is given by a vector x which lies in a state space Ω which is a region in some finite-dimensional Euelidean space. We assume that Ω is bounded and, therefore, has finite volume V:

$$V = \int_{\Omega} dx \quad < \infty \tag{2.19}$$

For each state x there is a set of alternatives S(x) from which one must be selected. If $\kappa \in S(x)$ is selected, $p^{\kappa}(x,y,t)\,dt$ is the probability that the transition from x occurs between t and t + dt after entering x, and goes to a region of volume dy about $y \in \Omega$.

If

$$p^{\kappa}(x,y) = \int_0^{\infty} dt \; p^{\kappa}(x,y,t) \tag{2.20}$$

$$\pi^{\kappa}(x,y) = \int_0^{\infty} dt \; p^{\kappa}(x,y,t)\, t \quad > 0 \tag{2.21}$$

$$\pi'^{\kappa}(x) = \int_{\Omega} dy \int_0^{\infty} dt \; p^{\kappa}(x,y,t)\, t = \int_{\Omega} dy \; \pi^{\kappa}(x,y) \tag{2.22}$$

are all finite,

$$\int dy\ P^{K}(x,y) = 1 \qquad x \in \Omega \qquad K \in S(x) \tag{2.23}$$

and if $q_i^K(t,a)$ is replaced by $q^K(x,t,a)$,

the analogy between the N-state and continuum state models is complete. For example, (2.8) becomes

$$V^A(x,t,a) = q^A(x,t,a) + \int_\Omega dy \int_0^t dt'\ P^K(x,y,t') a^{t'} V^A(y,t-t',a) \tag{2.24}$$

$$x \in \Omega \qquad t \geq 0$$

while (2.9) becomes

$$V^A(x,n+1,a) = q^A(x,a) + a \int_\Omega dy\ P^A(x,y)\ V^A(y,n,a) \tag{2.25}$$

$$x \in \Omega \qquad n \geq 0$$

Equation (2.15g-h) can be similarly transcribed. In particular, (2.15e) becomes

$$v(x,a) = \sup_{K \in S(x)} \left[q^K(x,a) + a \int_\Omega dy\ P^K(x,y)\ v(y,a) \right] \tag{2.26}$$

$$x \in \Omega$$

and may be thought of as the optimal inventory equation with discounting (x = amount of inventory) in which restrictions have been placed on the allowed actions such that (2.19) holds: the inventory x lies in a bounded interval $[-L, L]$.

A policy $\underline{K} = \{K(x)\}$ is a choice of alternative $K(x) \in S(x)$ for each $x \in \Omega$. We assume that for each policy, $P^K(x,y)$ is L_2:

$$\int_\Omega dx \int_\Omega dy \; |P^K(x,y)|^2 = \|P\|^2 < \infty \tag{2.27}$$

This restriction is not too serious since most transition matrices in practice are not only L_2 but are piecewise continuous as well.

The L_2 assumption is made for two reasons. First, if the functions $v^A(x)$, $q^A(x)$ and so on are L_2,

$$\int dx \; |v^A(x)|^2 < \infty \tag{2.28}$$

the existence of the integrals in (2.24 - 2.26) is guaranteed almost everywhere (for all x except a set of measure zero). Thus $v^A(x)$ will be finite for almost all x. Second, it permits the Fredholm Theory of integral operators to be brought to bear to discuss the eigenvalue spectrum and ergodicity of the $P(x,y)$'s.

The restriction (2.19) is actually no restriction at all since any infinite region can be mapped in a one-to-one fashion into a bounded region. For example, the mapping $x' = \tan^{-1} x$ takes $(-\infty, \infty)$

into $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. However, such a mapping might destroy the square integrability of the transition probability so that (2.28) is a true restriction.

It may also be argued that (2.19) is not unreasonable since all physical quantities (the components of x) assume bounded values. In the inventory problem, for example, it is highly unlikely that the inventory will ever be allowed to exceed $\$ 10^{10}$, or that the backlog will ever exceed $\$ 10^{10}$.

We will assume that the supremum in (2.26) and the other variants of (2.15e-h) are actually attained so that "sup" can be replaced by "max"; sufficient conditions for this would be that S(x) contains only finitely many elements or that S(x) is compact and the brackets in (2.26) are continuous in k for each x. This assumption can be justified mathematically since it simplifies the theory, and also physically, since an unattained supremum does not define an optimal policy (only an infinite collection of near-optimal policies) and our major concern is for finding policies.

2E. The N-State Process in the Continuum Formalism

An N-state Markov chain with transition probability matrix $P=[P_{ij}]$ can always be considered [21] as a special case of a continuum-state Markov chain in which "the system is in state i" becomes transcribed to "the system point lies in cell i" where cell i is the interval $[i-1, i]$. More concretely, if we consider the continuum process with the kernel

of finite rank, rank N, given by P' ,

$$P'(x,y) = \sum_{ij=1}^{N} P_{ij}\, \phi_i(x)\, \phi_j(y) \qquad 1 \le x,y \le N \tag{2.29}$$

where

$$\phi_i(x) = \begin{cases} 1 & i-1 \le x \le i \\ 0 & \text{otherwise} \end{cases} \tag{2.30}$$

P^1 is a Markov chain with state space $\Omega = [0, N]$. Then the iterates of P^1 satisfy

$$\begin{aligned} (P')^{n+1} &= \int_{\Omega} dz\, (P')^n(x,z)\, P'(z,y) \\ &= \sum_{ij=1}^{N} [P^n]_{ij}\, \phi_i(x)\, \phi_j(y) \end{aligned} \tag{2.31}$$

which may be proved by induction, using the orthonormality property

$$\int_{\Omega} dx\ \phi_i(x)\, \phi_j(x) = \delta_{ij} \qquad 1 \le i,j \le N \tag{2.32}$$

Equation (2.31) shows that the n-step transition probabilities P^n can be obtained from $(P^1)^n$ and conversely.

In addition, if $d_i^{(0)}$ is any initial distribution,

$$d_i^{(0)} \geq 0 \qquad \sum_{i=1}^{N} d_i^{(0)} = 1 \tag{2.33}$$

then the n-step distribution is

$$d_i^{(n)} = \sum_{j=1}^{N} d_j^{(0)} [P^n]_{ji} \qquad 1 \leq i \leq N \tag{2.34}$$

while a corresponding initial distribution

$$d^{(0)}(x) = \sum_{i=1}^{N} d_i \, \phi_i(x) \qquad 0 \leq x \leq N \tag{2.35}$$

leads to an n-step distribution

$$d^{(n)}(x) = \int_\Omega dy\, d^{(0)}(y)\, (P')^n(y,x)$$

$$= \sum_{j=1}^{N} d^{(n)}_j\, \phi_j(x) \qquad 0 \le x \le N \tag{2.36}$$

Hence the n-step distributions given by (2.35) and (2.36) agree. $d^{(n)}(x)$ can be recovered from $d^{(n)}_i$ by (2.36) while $d^{(n)}_i$ can be recovered from $d^{(n)}(x)$ by

$$d^{(n)}_i = \int_0^N dx\, d^{(n)}(x)\, \phi_i(x) \qquad 1 \le i \le N \tag{2.37}$$

Thus any iterate or probability in the N-state case can be obtained from the corresponding iterate or probability in the continuum case. In addition, the L_2 norms will agree. If P^n and $(P^1)^n$ are given by (2.31), and $d^{(n)}_i$ and $d^{(n)}(x)$ by (2.37), then

$$\|P^n\| = \|(P')^n\| \qquad n \ge 1$$

$$\|d^{(n)}_i\| = \|d^{(n)}(x)\| \qquad n \ge 0 \tag{2.38}$$

Thus the N-state case is a specialization of the continuum case to kernels of finite rank. All theorems which we prove for the continuum case must hold true for the N-state case as well.

This result should come as no surprise since it is well-known that kernels of finite rank lead to simultaneous linear equations.

Generalizations

1. This trick by replacing a state i by an interval enables us to replace a process with a continuum of states plus finitely many discrete states by a process with only a continuum of states.

2. Conversely, any L_2 kernel $P(x,y)$ can be approximated as closely as desired by a kernel of finitely many box functions [28, Pg. 158]. That is, equation (2.29) will be approximately true. The interpretation would be that if x lies in the i^{th} cell (the cell in which ϕ_i is non-zero), then we say the system is in state i. These cells need not be the same sizes: we could have a finer division of cells (note the cells form a partition of Ω, and may be multi-dimensional) in areas of greater interest or in areas where things are changing more rapidly.

If (2.29) and (2.32) hold, then we can think of (2.29) as a usual expansion in an orthonormal family of functions (Fourier series) and obtain the coefficients by

$$P_{ij} = \int_{\Omega} dx \int_{\Omega} dy \quad P'(x,y)\ \phi_i(x)\ \phi_j(y) \tag{2.39}$$

Since $P' \geq 0$ and the box functions $\phi_i \geq 0$, the P_{ij} are ≥ 0.

3. More generally we can expand $P^1(x,y)$ via (2.29) in terms of any orthonormal set, for example, reliability problems

in terms of sums of exponentials. Equation (2.39) will still be valid but it will no longer be true, in general, that $P_{ij} \geq 0$. Equations (2.31 - 2.36) will be approximately valid, but the interpretation as an N-state approximation to a continuum process will be lost: it will no longer be possible to speak of "the system is in state i (cell i)" because the ϕ_i may be overlapping or negative and no longer define cells.

Nevertheless, this method of approximating continuum problems by finite systems of linear equations is potentially much more powerful than the brute-force partition of Ω discussed in 2, because one can tailor the family of functions ϕ_i to match the kernel P(x, y) and perhaps get adequate approximations with N small.

2F. A Unified Notation

A unified notation is presented in Appendix A which permits simultaneous treatment of the N-state and continuum-state cases. The symbol f, or occasionally $\underline{f}$ for emphasis, will stand for an abstract vector with components f_i in the N-state and f(x) in the continuum-state case. Similar, the symbol M will stand for an abstract linear operator with components M_{ij}, or M(x, y). I is the identity with components δ_{ij} or $\delta(x-y)$.

In this notation, equations (2.8) and (2.24) can be consolidated to read

$$\underline{v}^A(t,a) = \underline{q}^A(t,a) + \int_0^t dt' \; P^A(t') \, a^{t'} \, \underline{v}^A(t-t', a)$$

Equations (2.9) and (2.25) become

$$\underline{v}^A(n+1,a) = \underline{q}^A(a) + a\, p^A\, \underline{v}^A(n,a)$$

Equations (2.15a, b, c, d) and their continuum generalizations become

$$v^A(a) = q^A(a) + a\, \tilde{p}^{AA} v(a)$$

$$v^A(a) = q^A(a) + \tilde{p}^A\left(\ln \tfrac{1}{a}\right) v^A(a)$$

$$v^A = q^A - g^A \underline{1} + p^A v^A$$

$$v^A = q^A - g^A \pi^A + p^A v^A$$

We have here introduced a "one" vector $\underline{1}$, all of whose components are unity:

$$1_i = 1 \qquad 1 \le i \le N$$

$$1(x) = 1 \qquad x \in \Omega$$

The $\underline{1}$ vector notation is convenient in several places:

a) $v_i(n) \to ng + v_i$

becomes

$$v(n) \to ng\,\underline{1} + \underline{v}$$

b) $\sum_j P_{ij} = 1$ all i and $\int_\Omega dy\ P(x,y) = 1$ all x

become

$$P\underline{1} = \underline{1}$$

c) The normalization of a distribution $\pi \geq 0$

$$\sum_i \pi_i = 1 \qquad \int_\Omega dx\ \pi(x) = 1$$

becomes

$$(\pi, 1) = 1$$

d) Operating with the P^∞ kernel

$$P^\infty{}_{ij} = \pi_j \qquad P^\infty(x,y) = \pi(y)$$

becomes

$$P^\infty \underline{f} = (\pi, f)\,\underline{1}$$

$$f P^\infty = (f, 1)\,\underline{\pi}$$

CHAPTER 3

DISCOUNTING

3A. Introduction

Markovian decision process with discounted rewards are well-understood (see for example Ref. 2, Chapter IV). This chapter rederives some of the typical results in order to familiarize the reader with our notation and with techniques which will prove useful later.

Only the continuum case is treated in detail here, similar arguments being applicable to the N-state case. The case of vanishing interest rate is treated in Section 8I. Semi-Markov processes are not treated.

Our results are contained in Theorems 3.1 and 3.2. They show that the total expected reward for the optimal infinite-horizon time-dependent policy is finite and can be approached arbitrarily closely by a stationary policy.

3B. Convergence and Fixed Point Theorems

The iterative equation of dynamic programming has been derived in Chapter 2:

$$v_{n+1}(x) = \max_{k \in S(x)} \left[q^k(x) + \alpha \int_{\Omega} dy\, p^k(x,y) v_n(y) \right]$$
$$x \in \Omega$$

where the dependence of $q^k(x)$ on α has been suppressed. Due to the presence of the discount factor α, we are able to handle the supremum case as easily as the case where the maximum is achieved: only the continuum case will be treated, the N-state case occurring as specialization where $p(xy)$ is of rank N. In this chapter the L_∞ norm is used throughout:

$$\| f \| = \sup_{x \in \Omega} |f(x)|$$

Theorem 3.1

Let

$$v_{n+1}(x) = \sup_{k \in S(x)} \left[q^k(x) + \alpha \int_{\Omega} dy\, p^k(x,y) v_n(y) \right] \qquad (3.1)$$
$$x \in \Omega$$

$$0 \leq \alpha < 1 \qquad (3.2)$$

$$\left| \sup_{k \in S(x)} q^k(x) \right| \leq Q \qquad \text{all } x \in \Omega \qquad (3.3)$$

$$\| v_0 \| < \infty \qquad (3.4)$$

$p^k(xy)$ and $v_0(x)$ obey sufficient conditions of measurability that $v_n(x)$ exist and be measurable for all $n \geq 0$ and $x \in \Omega$. Then

i) v_n is uniformly bounded:

$$\| v_n \| \leq \frac{Q}{1-\alpha} + \| v_0 \| \tag{3.5}$$

ii) $v_n(x) \rightarrow v_\infty(x)$ uniformly in x as $n \rightarrow \infty$ (3.6)

iii) $$\| v_\infty \| \leq \frac{Q}{1-\alpha} \tag{3.7}$$

iv) $v_\infty(x)$ satisfies the functional equation

$$v_\infty(x) = \sup_{k \in S(x)} \left[q^k(x) + \alpha \int_\Omega dy \; p^k(x,y) v_\infty(y) \right] \qquad x \in \Omega \tag{3.8}$$

v) $v_\infty(x)$ is independent of $v_0(x)$.

Proof

i) Insertion of $v_n(x) \geq - \| v_n \|$ and $v_n(x) \leq \| v_n \|$ into (3.1) leads to

$$\sup_{k \in S(x)} \left[q^k(x) - \alpha \| v_n \| \right] \leq v_{n+1}(x) \leq \sup_{k \in S(x)} \left[q^k(x) + \alpha \| v_n \| \right]$$

$$- Q - \alpha \| v_n \| \leq v_{n+1}(x) \leq Q + \alpha \| v_n \|$$

$$\| v_{n+1} \| \leq Q + \alpha \| v_n \|$$

$$\| v_{n+1} \| \leq \sum_{k=0}^{n} \alpha^k Q + \alpha^{n+1} \|v(0)\| \leq \frac{Q}{1-\alpha} + \|v(0)\| \quad (3.9)$$

which proves (3.5).

Using

$$\inf (A - B) \leq \sup A - \sup B \leq \sup (A - B) \quad (3.10)$$

where all infimum and supremum go over the same set, we get

$$\inf_{k \in S(x)} \int dy\, p^k(x,y) \left[v_n(y) - v_{n-1}(y) \right] \leq v_{n+1}(x) - v_n(x) \leq \sup_{k \in S(x)} \int dy\, p^k(x,y) \left[v_n(y) - v_{n-1}(y) \right]$$

$$-\alpha \| v_n - v_{n-1} \| \leq v_{n+1}(x) - v_n(x) \leq \alpha \| v_n - v_{n-1} \|$$

$$\| v_{n+1} - v_n \| \leq \alpha \| v_n - v_{n-1} \| \qquad n \geq 1 \quad (3.11)$$

Also, $v_1 - v_0$ is bounded, since v_1 and v_0 are bounded:

$$\| v_1 - v_0 \| \leq \| v_1 \| + \| v_0 \| < \infty$$

Then the series

$$v_0(x) + \sum_{n=1}^{\infty} \left[v_{n+1}(x) - v_n(x) \right]$$

is majorized by the geometrically convergent series

$$\| v_0 \| + \sum_{n=1}^{\infty} \| v_{n+1} - v_n \|$$

and must have a limit $v_\infty(x)$ for each x:

$$v_\infty(x) = \lim_{n \to \infty} v_n(x) \qquad x \in \Omega \tag{3.12}$$

To show that the limit in (3.12) is uniform in x, as well as pointwise, we sum (3.11) over m successive values of n to obtain

$$-\sum_{k=1}^{m} \alpha^k \|v_n - v_{n-1}\| \leq v_{n+m}(x) - v_n(x) \leq \sum_{k=1}^{m} \alpha^k \|v_n - v_{n-1}\|$$

$$|v_{n+m}(x) - v_n(x)| \leq \frac{\alpha}{1-\alpha} \|v_n - v_{n-1}\| \leq \frac{\alpha^n}{1-\alpha} \|v_1 - v_0\|$$

Let $m \to \infty$ to obtain

$$|v_\infty(x) - v_n(x)| \leq \frac{\alpha^n}{1-\alpha} \|v_1 - v_0\| \tag{3.13}$$

$$x \in \Omega$$

which shows uniform convergence, and also shows that $\|v_\infty - v_n\| \to 0$ at least geometrically fast. This completes the proof of (ii).

By combination of (3.9) and (3.13) we obtain

$$\|v_\infty\| \leq \|v_n\| + \frac{\alpha^n}{1-\alpha} \|v_1 - v_0\|$$

$$\leq \frac{Q}{1-\alpha} + \alpha^n \|v_0\| + \frac{\alpha^n}{1-\alpha} \|v_1 - v_0\|$$

If $n \to \infty$, (iii) is obtained.

Insertion of

$$v_\infty(x) - \frac{\alpha^n}{1-\alpha} \left\| v_1 - v_0 \right\| \leq v_n(x) \leq v_\infty(x) + \frac{\alpha^n}{1-\alpha} \left\| v_1 - v_0 \right\|$$

into (3.1) produces

$$\left| v_{n+1}(x) - \sup_{k \in S(x)} \left[q^k(x) + \alpha \int_\Omega dy\, p^k(x,y) v_\infty(y) \right] \right| \leq \frac{\alpha^{n+1}}{1-\alpha} \left\| v_1 - v_0 \right\| \qquad x \in \Omega$$

Letting $n \to \infty$, we obtain

$$v_\infty(x) = \sup_{k \in S(x)} \left[q^k(x) + \alpha \int_\Omega dy\, p^k(x,y) v_\infty(y) \right] \qquad x \in \Omega$$

This proves (iv).

The proof of (v) follows from the third part of the next theorem. QED

Theorem 3.2 (Existence and Uniqueness)

Consider the functional equation

$$f(x) = \sup_{k \in S(x)} \left[q^k(x) + \alpha \int_\Omega dy\, p^k(x,y) f(y) \right] \qquad x \in \Omega \tag{3.14}$$

where

$$\left| \sup_{k \in S(x)} q^k(x) \right| \leq Q \qquad x \in \Omega \tag{3.15}$$

$$0 \leq \alpha < 1$$

and we require f be finite everywhere:

$$\|f\| < \infty$$

Then

i) a solution f exists

ii) f is unique

iii) $f = \sup_B (I - \alpha p^B)^{-1} q^B$

Proof

Existence follows from part iv of Theorem 3.1. To show uniqueness, let f and g denote two solutions to (3.14). Subtraction and use of (3.10) leads to

$$\alpha \inf_{k \in S(x)} \int dy\, p^k(x,y)\left[f(y)-g(y)\right] \leq f(x)-g(x) \leq \alpha \sup_{k \in S(x)} \int dy\, p^k(x,y)\left[f(y)-g(y)\right]$$

$$\| f - g \| \leq \alpha \| f - g \|$$

Since $\alpha \neq 1$, this implies $\| f - g \| = 0$ so that f and g agree everywhere. This proves ii. The proof of iii now follows. Result iii is an alternate method of proving ii.

It follows from (3.14) that for all policies B,

$$f(x) \geq q^B(x) + \alpha p^B f(x) \qquad \text{all } x \in \Omega .$$

On the other hand, given $\epsilon > 0$ there is a policy A such that

$$f(x) \leq q^A(x) + \alpha p^A f(x) + \epsilon \qquad \text{all } x \in \Omega .$$

An n-fold iteration of these equations yields

$$f(x) \geq q^B(x) + \sum_{m=1}^{n} (\alpha p^B)^m q^B(x) + (\alpha p^B)^{n+1} f(x) \qquad (3.16)$$

all $x \in \Omega$

$$f(x) \leq q^A(x) + \sum_{m=1}^{n} (\alpha p^A)^m q^A(x) + \frac{1-\alpha^{n+1}}{1-\alpha} \epsilon + (\alpha p^A)^{n+1} f(x) \qquad (3.17)$$

all $x \in \Omega$

But $\| (\alpha p)^n f \| \to 0$ as $n \to \infty$ because $\| p^n f \| \leq \| f \| < \infty$

for $n \geq 1$. Consequently, the last term in (3.16-3.17) disappears as $n \to \infty$. The series $q + \sum_{m=1}^{n} (\alpha p)^m q$ is convergent in L_∞ norm:

$$\| q + \sum_{m=1}^{n} (\alpha p)^m q \| \leq \| q \| + \sum_{m=1}^{n} \alpha^n \| p^m q \| \leq \frac{1-\alpha^{n+1}}{1-\alpha} \| q \|$$

as $n \to \infty$, to $(I - \alpha p)^{-1} q$.

Equations (3.16) and (3.17) become, as $n \to \infty$,

$$f(x) \geq (I - \alpha p^B)^{-1} q^B (x) \qquad \text{all } x \in \Omega \qquad (3.18)$$

all policies B

$$f(x) \leq (I - \alpha p^A)^{-1} q^A (x) + \frac{\epsilon}{1-\alpha} \qquad \text{all } x \in \Omega \qquad (3.19)$$

Equations (3.18-3.19) complete the proof of (iii):

$$f(x) = \sup_B \left[(I - \alpha p^B)^{-1} q^B (x) \right] \qquad (3.20a)$$

QED

Remarks

(1) The policy B which achieves the maximum (if the supremum is achieved) on the right side of (3.20) for some value of x need not be unique.

(2) If a policy B achieves the maximum on the right side of (3.20), it can be taken to be independent of x. This occurs because we are trying to maximize the reward for all states.

(3) In the N-state case, equation (3.20a) becomes

$$f_i = \max_{\text{NN policies } B} \left[\sum_{j=1}^{N} (I - \alpha p^B)^{-1}_{ij} q_j^B \right] \quad 1 \leq i \leq N \qquad (3.20b)$$

3C. Stationary Policies

A stationary policy B will have a total expected reward $\left[(I - \alpha p^B)^{-1} q^B\right](x)$ in an infinite number of transitions starting from state x. Equation (3.20) therefore indicates that the optimal total reward using a time-dependent policy is the supremum over all stationary policies of the total reward using a stationary policy.

More concretely, (3.18-3.19) show that there always exists a stationary policy A whose total expected reward differs from the total expected reward of the time-dependent policy by as little as desired. This conclusion is due to Blackwell[11]. It justifies the restriction by Howard and Jewell to consideration, in the infinite horizon discounted case, of only stationary policies.

In the finite horizon discounted case, equation (3.6) shows that $v_n(x)$, the total reward from the optimal time-dependent policy approaches f(x), the total reward from the optimal stationary policy as $n \to \infty$. Hence stationary policies may be adequate for large but finite n. It is intuitive that result v of Theorem 3.1 holds: as the number of remaining transitions becomes very large, the terminal values become irrelevant if there is discounting.

3D. Policy Convergence

In the N-state case, policy convergence as well as value convergence can be demonstrated. Let

$$f_i = \max_{1 \le k \le n_i} \left[q_i^k + \alpha \sum_{j=1}^{N} p_{ij}^k f_j \right] \qquad 1 \le i \le N \tag{3.21}$$

$$v_i(n+1) = \max_{1 \le k \le n_i} \left[q_i^k + \alpha \sum_{j=1}^{N} p_{ij}^k v_j(n) \right] \quad 1 \le i \le N \tag{3.22}$$

We know that $\underline{f}$ is unique and $v_i(n) \to f_i$ as $n \to \infty$. Let $k_i(n+1)$ denote the alternative chosen on the right side of (3.19), i.e., the decision in state i at the time n+1. Let A_i be the set of integers k, $1 \le k \le n_i$, which achieve the maximum on the right side of (3.21). Then we claim that for all sufficiently large,

$$k_i(n+1) \subseteq A_i \tag{3.23}$$

Proof

Insertion of (3.13):

$$f_i - \frac{\alpha^n}{1-\alpha} \| v_1 - v_0 \| \le v_i(n) \le f_i + \frac{\alpha^n}{1-\alpha} \| v_1 - v_0 \|$$

into both sides of (3.22) leads to

$$f_i - \frac{\alpha^{n+1}}{1-\alpha} \|v_1 - v_0\| \leq v_i(n+1) \leq q_i^{k_i(n+1)} + \sum_{j=1}^{N} p_{ij}^{k_i(n+1)} f_i + \frac{\alpha^{n+1}}{1-\alpha} \|v_1 - v_0\|$$

$$f_i - q_i^{k_i(n+1)} - \sum_{j=1}^{N} p_{ij}^{k_i(n+1)} f_j \leq \frac{2\alpha^{n+1}}{1-\alpha} \|v_1 - v_0\| \qquad (3.24)$$

Let

$$d_i = f_i - \max_{\substack{1 \leq k \leq n_i \\ k \notin A_i}} \left[q_i^k + \alpha \sum_{j=1}^{N} p_{ij}^k f_j \right]$$

According to (3.21), $d_i > 0$.

If $k_i(n+1) \notin A_i$, then (3.24) implies

$$0 < d_i \leq \frac{2\alpha^{n+1}}{1-\alpha} \|v_1 - v_0\|$$

and is a contradiction if n is sufficiently large. Therefore, $k_i(n+1) \in A_i$ if n is sufficiently large that

$$\frac{2\alpha^{n+1}}{1-\alpha} \|v_1 - v_0\| < d_i \qquad (3.25)$$

QED

3E. Further Developments

The above sections show that even in the continuum case, strong results can be stated about the asymptotic behavior of the value-iteration equations provided (3.3) and (3.4) hold.

At this point, two avenues of development are possible. One is to make explicit assumptions about the form of the rewards and probability densities, and then investigate the structure of the optimal policy. For example, under what conditions will the optimal policy for an inventory system be s, S? Investigation along these lines very rapidly becomes specialized and is not undertaken here.

The second approach is to relax (3.3) and (3.4) and see what can be said about the value-iteration procedure. For example, will $v_n(x)$ have a limit $v_\infty(x)$? These assumptions that the scrap value $v_0(x)$ and maximum possible immediate rewards $\sup_{k \in S(x)} q^k(x)$ from any state x are bounded seem to be quite reasonable and no realistic reason seems to exist for dropping them. The usual difficulty with the inventory equation, for example, which requires such a painful analysis[15] to show that $v_\infty(x)$ exists is the assumption of a constant purchase cost per item with no restriction on the amount that could be ordered. This assumption invalidates (3.3) and could be argued to be unrealistic: all physical quantities assume bounded values and in particular, purchase orders or backlogs above 10^{10} dollars seem unlikely.

The simplicity of the results in 3B-3E follow from the presence of the discount factor α. For example, no attention has been paid to the chain structure of the process.

In the case of no discounting, $\alpha = 1$, the description of the process becomes more complicated. The chain structure and periodicities or lack of them play important roles in determining the accumulation of rewards.

Chapters 4 and 5 develop the tools needed in the remainder of this thesis for dealing with the undiscounted case. Chapter 4 discusses the chain structure while Chapter 5 discusses the fundamental matrix which plays an important role in describing the steady-state behavior.

CHAPTER 4

CHAIN STRUCTURE, ERGODIC BEHAVIOR, AND SPECTRAL PROPERTIES OF MARKOV CHAINS

4A. Introduction

The present chapter attempts to relate the chain structure of a Markov chain (number of irreducible disjoint sets of states) with its eigenvalue spectrum. That relations exist might be suspected from the fact that the stationary distribution π of a Markov chain P satisfies $\pi P = \pi$ and therefore is a left eigenvector of P with eigenvalue unity. Carried further this approach simultaneously yields all the subchains of P and all of the eigenvectors with eigenvalue 1. An extensive discussion of the chain structure is given. Necessary and sufficient conditions for a Markov chain to be ergodic are discussed. The fundamental matrix Z, which will prove invaluable later, is obtained by a new method which seems well-motivated. The method involves the decomposition of P into steady-state and transient components, and the factorization of I-P.

The collection of theorems contained here is an attempt to give the reader an intuitive feeling for Markov chains with a continuum of states. It consists of a systematic exposition of techniques and results spread throughout the literature and hopefully will be useful for reference purposes.

The results hold for both N-state case and the continuum-state case, and the reader is urged to relate the theorems to his own experience in handling finite-state Markov chains. The analogies are quite extensive between the finite and continuum-state cases.

In the remainder of this thesis, we will confine ourselves to Markov chains with a single irreducible set of states. The many-chain case is included here for completeness and may be skipped if desired: Sections 4G and 4H are not needed.

The assumption which enables us to get these results rather painlessly is that P is L_2:

$$\int_{\Omega} dx \int_{\Omega} dy \, |P(xy)|^2 < \infty$$

It permits the functional analysis approach to Hilbert space to be employed, with all the elaborate machinery for eigenvalue spectrum analysis and so on. The assumption is partly justified because it enables us to easily get an overall picture of continuous-state Markov chains. However, the principle justification is that it is not unreasonable. In many modeled situations, transition probabilities are piecewise continuous or are normal or exponential--or at any rate have tails which fall off fast enough--and the assumption is met. One somewhat unexpected consequence of the L_2 assumption is that there can be only finitely many sub-chains.

We shall be somewhat careless in our vocabulary and use "all x" or "almost all x" to mean all but a set x of measure zero. This allows us to omit some dull and uninspired reasoning. The L_2 norm is used in this chapter.

We will say π is a "unique distribution" if $\pi P = \pi$ and any other left eigenvector of P at eigenvalue 1 is a multiple of π.

4B. Norms of Probability Vectors and Markov Operators

In the N-state case, the real vector $f = (f_1, \ldots, f_N)$ is called a probability vector if

$$f_i \geq 0 \quad 1 \leq i \leq N \qquad \sum_{i=1}^{N} f_i = 1$$

The real N x N matrix P is called a Markov operator if

$$p_{ij} \geq 0 \quad 1 \leq i, j \leq N \qquad \sum_{j=1}^{N} p_{ij} = 1 \quad 1 \leq i \leq N$$

These possess L_2 norms which are bounded as described in

Theorem 4.1 N-State Case

Let f be a probability vector and P a Markov operator. Then

$$\frac{1}{\sqrt{N}} \leq \|f\| \leq 1 \tag{4.1}$$

$\|f\| = 1$ if and only if there is some integer j, $1 \leq j \leq N$, such (4.2)

that

$$f_i = \delta_{ij} \qquad 1 \leq i \leq N$$

$\|f\| = \frac{1}{\sqrt{N}}$ if and only if $f_i = \frac{1}{N}$ $\quad 1 \leq i \leq N$ (4.3)

$$1 \leq \|P\| \leq \sqrt{N} \tag{4.4}$$

$\|P\| = 1$ if and only if $p_{ij} = \frac{1}{N}$ $\quad 1 \leq i, j \leq N$ (4.5)

$\|P\| = \sqrt{N}$ if and only if for each i, there is one j for which $p_{ij} = 1$. (4.6)

Proof

4.1-4.3

$$0 \leq f_i \leq 1 \qquad f_i^2 \leq f_i$$

$$\|f\|^2 = \sum_i f_i^2 \leq \sum_i f_i = 1 \quad \text{so} \quad \|f\| \leq 1$$

Equality if and only if $f_i^2 = f_i$ for all i, whence $f_i = 0$ or 1. Also,

$$1 = (f, 1) = |(f, 1)| \leq \|f\| \, \|1\| = \|f\| \sqrt{N}$$

Thus $f \geq \frac{1}{\sqrt{N}}$ with equality if and only if $\underline{f}$ is parallel to $\underline{1}$.

4.4-4.6

$$0 \leq \sum_{i,j=1}^{N} (p_{ij} - \frac{1}{N})^2 = \sum_{ij=1}^{N} (p_{ij})^2 - 1 = \|P\|^2 - 1$$

whence $\|P\| \geq 1$ with equality if and only if $p_{ij} = \frac{1}{N}$ for all i and j.

Also,

$$\|P\|^2 = \sum_{ij=1}^{N} (p_{ij})^2 \leq \sum_{ij=1}^{N} p_{ij} = N$$

whence $\|P\| \leq \sqrt{N}$ with equality if and only if $(p_{ij})^2 = p_{ij}$ for all i and j. QED

In the continuum case, a real L_2 function f(x), $x \in \Omega$, is called a probability vector if

$f(x) \geq 0$ almost all $x \in \Omega$

$$(f, 1) = \int dx\, f(x) = 1$$

A real L_2 kernel $p(x, y)$ is called a Markov operator if

$$p(x, y) \geq 0 \qquad \text{almost all } x, y \in \Omega$$

$$\int_{\Omega} dy\, p(x, y) = 1 \qquad \text{almost all } x \in \Omega$$

These are bounded as described in

Theorem 4.2 (continuum case)

Let f by a probability vector and P a Markov operator defined on a state-space Ω with volume V given by (2.19). Then

$$\frac{1}{\sqrt{V}} \leq \|f\| < \infty \tag{4.7}$$

$$\|f\| = \frac{1}{\sqrt{V}} \text{ if and only if } f(x) = \frac{1}{V} \text{ for} \tag{4.8}$$

almost all $x \in \Omega$

$$1 \leq \|P\| < \infty \tag{4.9}$$

$$\|P\| = 1 \text{ if and only if } p(x, y) = \frac{1}{V} \text{ for} \tag{4.10}$$

almost all $x, y \in \Omega$

Proof

4.7-4.8

$$1 = (f, 1) = |(f, 1)| \leq \|f\| \, \|1\| = \|f\| \sqrt{V}$$

whence $\|f\| \geq 1/V$ with equality if and only if f is proportional to 1 almost everywhere.

$\|f\|$ can get arbitrarily close to ∞ as seen by the choice

$$f(x) = \begin{cases} 1/a & x \in a \text{ subset of } \Omega \text{ of volume } a < V \\ 0 & \text{elsewhere} \end{cases}$$

for which $\|f\| = 1/a$ and ranges from $1/\sqrt{V}$ to ∞ as a ranges from V down to 0.

4.9-4.10

$$0 \leq \int_{\Omega \times \Omega} dx \int dy \left[p(x, y) - \frac{1}{V} \right]^2 = \|P\|^2 - 1$$

whence $\|P\| \geq 1$ with equality if and only if $p(x, y) = 1/V$ almost everywhere. By letting P come arbitrarily close to a lattice process (i.e., impulse components), $\|P\|$ can be made arbitrarily large. QED

An Application

Using (4.4), it is possible to give a simple proof of the existence of the limit as m goes to infinity of

$$S_m = \frac{P + P^2 + \ldots + P^m}{m} \tag{4.11}$$

in the N-state case.

Theorem 4.3

Let P by an N-state Markov operator, and S_m be defined by (4.11). Then

$$S = \lim_{m \to \infty} S_m$$

exists (element by element convergence).

Proof

Since $\|P^j\| \leq \sqrt{N}$ for $j = 1, 2, \ldots$, it follows from (4.11) that

$$\left\| S_m \right\| \leq \frac{\left\| P \right\| + \left\| P^2 \right\| + \ldots + \left\| P^m \right\|}{m} \leq \sqrt{N}$$

so that S_m is uniformly bounded and possesses a cluster point S. That is, a sequence $\{n_i\}$ exists for which $S_{n_i} \to S$.

It follows from (4.11) both that $S_{m+1} \to S_m P = P S_m$ and that $S_{m+1} - S_m \to 0$, so that $S = P S = S P$.

To show that S is unique, we suppose that T is another cluster point of S_m. Again $T = P T = T P$ from which follows $T = S_m T = T S_m$. Letting m run through $\{n_i\}$, we obtain $T = S T = T S$. Interchange of S and T leads to $S = T$. S_m therefore has one cluster point and the limit exists. QED

Corollary

$S = S P = P S$

$S = S^2$

4C. The Spectral Properties of Markov Operators

The reader is asked to turn to Appendix B for definitions and discussion of the terms eigenvalue, eigenvector, geometric multiplicity, algebraic multiplicity spectral radius, Fredholm determinant, and so forth. It is assumed that P is L_2 so that the Fredholm theory can be brought to bear. We recall that $P \geq 0$ and $P\,\underline{1} = \underline{1}$. Both Chapter 4 and Appendix B treat the N-state case and continuum case simultaneously.

Theorem 4.4

Let λ be an eigenvalue of P. Then $|\lambda| \leq 1$.

Proof (continuous case only)

If λ is an eigenvalue, then at least one L_2 left eigenvector e with eigenvalue λ must exist.

$$\lambda\, e(x) = \int dy\; e(y)\, p(y,x) \qquad \text{almost all } x \in \Omega$$

$$|\lambda|\;|e(x)| = \left|\int dy\; e(y)\, p(y,x)\right| \leq \int dy |e(y)|\, p(y,x) \tag{4.12}$$

Integrate over all x to get

$$\lambda\;(|e|, 1) \leq \int dy\, |e(y)| = (|e|, 1) \tag{4.13}$$

Since e is non-trivial, $(|e|, 1) > 0$ and (4.13) implies $|\lambda| \leq 1$.

QED

Corollary

If e(x) is a left eigenvector of P with eigenvalue λ, and $|\lambda| = 1$, then $|e(x)|$ is a left eigenvector of P with eigenvalue 1.

Proof

(4.13) must be an identity, whence (4.12) must hold for almost all x. QED

An Example

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$\lambda = 1 \qquad e = [1 \quad 1]$ = left eigenvector

$\lambda = -1 \qquad f = [1 \quad -1]$ = left eigenvector

Then $|f_i| = e_i$.

Theorem 4.5

$\lambda = 1$ is an eigenvalue for P

Proof

$P\underline{1} = \underline{1}$

Thus 1 is an eigenvalue of P, with $\underline{1}$ as a right eigenvector. QED

Corollary 1

The spectral radius of P is 1.

Proof

Follows from Theorems 4.4 - 4.5. QED

According to Theorem 4.5, P must have one or more left eigenvalues at $\lambda = 1$. These will all be denoted by "π", and satisfy $\pi P = \pi$.

Corollary 2

If $\pi(x)$ is a left eigenvector of P with eigenvalue 1, then so is $|\pi(x)|$.

Proof

See Corollary to Theorem 4.4. QED

Theorem 4.6

Let P have r linearly independent left eigenvectors $\pi_1, \ldots \pi_r$ at $\lambda = 1$. Then with no loss of generality, these can all be taken as real and non-negative: $\pi_i(x) \geq 0$. $1 \leq i \leq r$

Proof

Since P is real, the equation $\pi P = \pi$ implies that both the real and imaginary parts of π are left eigenvector with $\lambda = 1$. From the 2r real left eigenvectors

$$\frac{\pi_j(x) + [\pi_j(x)]^*}{2} \quad , \quad \frac{\pi_j(x) - [\pi_j(x)]^*}{2i} \qquad 1 \leq j \leq r$$

one can extract r linearly independent ones. Hence there is no loss of generality in taking $\pi_1, \ldots \pi_r$ as all real.

According to Corollary 2 of Theorem 4.5, if $\pi_1, \ldots \pi_r$ are r real left eigenvectors of P at 1, then so are $|\pi_j(x)|$, $1 \leq j \leq r$. From the 2r real, non-negative left eigenvectors

$$\pi_j(x) \pm |\pi_j(x)| \qquad 1 \leq j \leq r$$

one can extract r which are linearly independent. QED

Corollary

The π's can be normalized to a probability distribution: $(\pi, 1) = 1$.

We will later show that if P has r left eigenvectors $\pi_1, \ldots \pi_r$ at 1, then P contains exactly r disjoint irreducible sets of states. By an appropriate choice, π_j is the stationary distribution for the j^{th} irreducible set of states.

According to (B1), the number r of left eigenvectors of P at $\lambda = 1$ can be bounded:

$$r \leq \|P\|^2 \tag{4.14}$$

This also implies that an L_2 P operator has only finitely many irreducible sets of states. The bound in (4.14) is the best possible, since the N-state case $p_{ij} = \delta_{ij}$ $1 \leq i, j \leq N$ has $r = N$ eigenvectors $\pi_1, \ldots \pi_N$ at $\lambda = 1$: $(\pi_i)_j = \delta_{ij}$ and also satisfies (see (4.4)) $\|P\| = \sqrt{N}$. This example confirms our assertion that r is the number of chains.

As the example to Theorem 4.4 demonstrates, eigenvalues of P on the unit circle other than 1 correspond to periodic states. This property will not be discussed further except for our frequent usage of the corollary to Theorem 4.4.

The eigenvalue spectrum of P can be described in terms of the results summarized in Appendix B:

(1) All eigenvalues of P lie in or on the unit circle. $\lambda = 1$ is an eigenvalue of P.

(2) Each eigenvalue corresponds to at least one left (and at least one right) L_2 eigenvector. The number of eigenvectors at an eigenvalue $\lambda \neq 0$ is finite, and by (B1), equal to or less than $\|P\|^2 / |\lambda|^2$. The algebraic multiplicity of any eigenvalue $\lambda \neq 0$ is also finite.

(3) The number of eigenvalues of P is either finite (as in the N-state case) or denumerably infinite. The number of eigenvalues within any annulus $0 < a \leq |\lambda| \leq b$ is finite. The eigenvalues, if infinite in number, have $\lambda = 0$ as their only cluster point. The eigenvalues can be numbered $\lambda_1, \lambda_2, \lambda_3, \ldots$, each repeated as often as its geometric or algebraic multiplicity, with $|\lambda_{n+1}| \leq |\lambda_n|$. Any closed region of the λ plane which excludes the origin has finitely many eigenvectors. In particular, there are only a finite number of eigenvalues on the unit circle.

(4) $\lambda = 0$ can be an eigenvalue, although it need not be as the N-state case $P = I$ illustrates. If $\lambda = 0$ is an eigenvalue, it can correspond to a finite number, say k ($k \geq 1$) eigenvectors, as the following case, where P is a $(k+1)x(k+1)$ matrix, illustrates:

$$p_{ij} = \delta_{j,1} \qquad 1 \leq i,\ j \leq k+1$$

$$^{m}_{\underline{e}} P = \underline{0} \qquad 1 \leq m \leq k$$

$$(^{m}e)_i = \begin{cases} 1 & i = 1 \\ -1 & i = m+1 \\ 0 & \text{all other } i,\ 1 \leq i \leq k+1 \end{cases}$$

On the other hand, if $\lambda = 0$ is an eigenvalue, it could also correspond to a non-denumerably infinite number of eigenvectors, as the following example illustrates:

$$p(x, y) = 1 \qquad 0 \leq x,\ y \leq 1$$

$$^{\alpha}e(x) = \begin{cases} 1/\alpha & 0 \leq x \leq \alpha \\ -1/\alpha & 1 - \alpha \leq x \leq 1 \\ 0 & \text{all other } x\ \ 0 \leq x \leq 1 \end{cases}$$

$$^{\alpha}_{\underline{e}} P = 0$$

so that as α ranges over the open interval $(0, 1/2)$ the $^{\alpha}e$ form a non-denumerably infinite set of L_2 eigenvectors with eigenvalue 0.

(5) If λ is an eigenvalue, so is λ^*. Indeed if $P\underline{e} = \lambda \underline{e}$, then $P\underline{e}^* = \lambda^* \underline{e}^*$.

4D. Chain Structure of Markov Operators

We turn next to a systematic discussion of the chain structure of Markov processes, and show how the chain structure is related to the eigenvectors of P with eigenvalue 1.

We say that a set of states S is closed if, for almost all $x \in S$,

$$\int_S dy\ p(x, y) = 1$$

Stated verbally, the probability of leaving S once inside is zero.

The entire state space Ω is a closed set of states. A more useful way of finding closed sets of states is given by

Theorem 4.7

Let π by a left eigenvector of P with eigenvalue 1. According to Theorem 4.6, we can assume that π is real and $\pi(x) \geq 0$. Let

$$S = \left\{ x \,\middle|\, \pi(x) > 0 \right\}$$

Then S is a closed set of states.

Proof

Let $f(x) = \int_S dy\ p(xy)$

Then integration of

$$\pi(x) = \int_\Omega dy\ \pi(y)\ p(y, x) \qquad \text{almost all } x \in \Omega$$

over $x \in S$ leads to

$$(\pi, 1) = \int dy\, \pi(y) f(y) = (\pi, f) \quad .$$

But by its definition $0 \leq f(x) \leq 1$. The last equation therefore implies that $f(x) = 1$ for almost all $x \in S$. QED

We say that a set R is irreducible if it is closed and if no proper subset is closed. (We say that A is a proper subset of B if $A \subsetneq B$ and if the measure of A is strictly less than the measure of B.)

An example

Let S be a closed set of states with a stationary distribution $\pi \geq 0$ defined on it. If the subset of S on which $\pi = 0$ has positive measure, then S is not irreducible.

Proof

According to Theorem 4.7 we can delete the set of states with $\pi(x) = 0$ and still have a closed set.

QED

The moral is that "transient states" must be removed to get an irreducible set of states. We say that a set S contains a unique irreducible set of states if (1) it contains at least one irreducible set of states and (2) if S contains two irreducible sets of states, then these must be identical up to a set of zero measure.

Theorem 4.8

Let a set S contain two irreducible sets of states R_1 and R_2. Then either R_1 and R_2 are disjoint (except for a set of zero measure) or identical (except for a set of zero measure).

Proof

Either R_1 and R_2 are disjoint, and the theorem is proved, or they overlap on a set $R_{12} = R_1 \cap R_2$ of positive measure. We wish to show that in the latter case, R_1 and R_2 coincide: $R_1 = R_2 = R_{12}$. Assuming the contrary, R_{12} and say $R_1 - R_{12}$ are disjoint sets of states. Moreover, R_{12} is not a closed set of states for it is a proper subset of R_1 and R_{12} being closed would contradict the irreducibility of R_1. Therefore, there is a set of x's in R_{12} of positive measure for which

$$\int_{R_1 - R_{12}} dy\, p(x, y) > 0$$

However, these x's are in R_{12}, hence R_2 and the above inequation states that $R_1 - R_{12}$ is accessible from R_{12}. Since $R_1 - R_{12}$ is disjoint from R_2, this would contradict the irreducibility of R_2. QED

Theorem 4.9

If π is any stationary distribution defined on some irreducible set R, then $\pi(x) > 0$ for almost all $x \in R$.

Proof

By Theorem 4.7, the set

$$R' = \{x \mid \pi(x) > 0\} \qquad R' \subseteq R$$

is closed. Hence R is irreducible only if R and R' differ by a set of zero measure. QED

Theorem 4.10

Let $\pi(x)$ be a real left eigenvector of P with eigenvalue 1, taking on both positive and negative values. Then the sets S_+ and S_-

$$S_+ = \{x \mid \pi(x) > 0\}$$
$$S_- = \{x \mid \pi(x) < 0\}$$

are closed sets of states, and possess, respectively, the stationary distributions

$$\pi_+(x) = \begin{cases} \pi(x) & x \in S_+ \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_-(x) = \begin{cases} -\pi(x) & x \in S_- \\ 0 & \text{otherwise} \end{cases}$$

with the properties

$$\pi(x) = \pi_+(x) - \pi_-(x) \tag{4.15}$$

$$\pi_+ P = \pi_+ \qquad \pi_- P = \pi_- \tag{4.16}$$

Proof

If $\pi(x)$ is a left eigenvector, so is $|\pi(x)|$. So are $|\pi(x)| \pm \pi(x)$. The definitions

$$\pi_+(x) = \frac{|\pi(x)| + \pi(x)}{2}$$
$$\pi_-(x) = \frac{|\pi(x)| - \pi(x)}{2}$$

are equivalent to the ones given above and demonstrate (4.15). Since they are both left eigenvectors, (4.16) holds. S_+ and S_- are closed by appeal to Theorem 4.7. QED

An example

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \pi = [2 \quad -1]$$

$$\pi_+ = [2 \quad 0] \qquad \pi_- = [0 \quad 1] \qquad \pi_{\pm} P = \pi_{\pm}$$

S_+ = state 1, and is closed.

S_- = state 2, and is closed.

The example shows that if π takes both signs, then the state space can be split into two two disjoint closed sets of states, those for which $\pi > 0$ and those for which $\pi < 0$.

Theorem 4.11

Ω contains only finitely many, indeed at most $\|P\|^2$, disjoint closed sets of states.

Proof

Each closed set S_i of states has a stationary distribution π_i. Also $\pi_i P = \pi_i$ since π_i has all its mass on S_i. The π_i are disjoint thus linearly independent. By (4.14), there are at most $\|P\|^2$ π's, hence at most $\|P\|^2$ disjoint closed sets of states. QED

Corollary

Ω contains only finitely many, indeed at most $\|P\|^2$, disjoint irreducible sets of states.

Proof

Irreducible sets of states are closed.

QED

Theorem 4.12 (Converse to Theorem 4.9)

Let S be a closed set of states, which we can consider as a new state-space for a new Markov chain using P. Suppose S has only one stationary vector π and that $\pi(x) > 0$ for all $x \in S$. Then S is irreducible.

Proof

If S is not irreducible, it has a proper subset S' which is closed. Let π' be a stationary vector for S'. Define $\pi(x)$ by

$$\pi(x) = \begin{cases} \pi'(x) & x \in S' \\ 0 & x \in S-S' \end{cases}$$

Then $\pi P = \pi$ whence π is the unique stationary vector for S. But this contradicts $\pi > 0$. QED

Theorem 4.13

Let $S \subseteq \Omega$ be closed and possess at least two stationary distributions π_1, π_2. With no loss of generality $\pi_1(x) \geq 0$, $\pi_2(x) \geq 0$, $(\pi_1, 1) = (\pi_2, 1) = 1$. Then S can be split into two disjoint closed sets of states. These are S_+ and S_- given by

$$S_+ = \{x \mid \pi_2(x) - \pi_1(x) > 0\}$$

$$S_- = \{x \mid \pi_2(x) - \pi_1(x) < 0\}$$

Proof

$\pi_2 - \pi_1$ cannot have the same sign for all $x \in S$, for $\pi_2(x) - \pi_1(x) \geq 0$ $x \in S$ would imply

$$0 = 1 - 1 = \int_S dx \left[\pi_2(x) - \pi_1(x)\right] \geq 0$$

whence $\pi_2 = \pi_1$ almost everywhere. Hence $\pi_2 - \pi_1$ changes sign. It is a left eigenvector of P at 1. Now invoke Theorem 4.10. QED

A Technique for Finding Disjoint Closed Sets of States

Given any closed set S, such as $S = \Omega$, it either has a unique stationary distribution π or two or more such π's. In the latter case, Theorem (4.13) shows that S can be split into two disjoint closed sets of states. We apply the same splitting procedure to each of the closed sets of states with two or more π's, and keep getting more or more disjoint closed sets of states. According to Theorem(4.11), this procedure must stop after finitely many splits since there will be at most $\|P\|^2$ disjoint closed sets of states. Convergence is achieved when each of the closed set of states has a unique π.

Theorem 4.14

Let $S \subseteq \Omega$ be closed and possess a unique stationary distribution $\pi \geq 0$. Then S contains an irreducible set of states, namely those states for which $\pi > 0$.

Proof

Take the subset S' of S on which $\pi(x) > 0$. Then S' is closed and possesses a stationary vector $\pi'(x)$. We claim that $\pi'(x)$ is unique:

$\pi'(x) = \pi(x)$. For $\pi(x)$ is one stationary distribution for S', and any other stationary distribution π'' of S' (i.e., $\pi''(x) = 0$ if $x \notin S'$) will be a stationary distribution of S, hence $\pi'' = \pi$.

Appeal to Theorem 4.12, since S' is closed and $\pi > 0$ on S', yields the result that S' is irreducible. QED

Remarks

This theorem enables us to locate irreducible sets of states.

Theorem 4.15 will show that S possesses only one irreducible set of states.

Corollary

Ω contains at least one irreducible set of states.

Proof

The splitting technique described above this theorem produces disjoint closed sets of states, each with a unique π vector. Then apply this theorem to each of the disjoint closed sets of states with unique π to get an irreducible set of states for each.

QED

Theorem 4.15

A set of states S has a unique stationary distribution π (that is, the geometric multiplicity of $\lambda = 1$ is 1) if and only if it contains a unique irreducible set of states.

Proof

a) Let S have a unique π vector. Then it either contains

no, one, or two or more irreducible sets of states. The above corollary eliminates the first possibility while Theorem 4.13 eliminates the second possibility.

b) Let S contain a unique irreducible set of states. Then either S has either 1 or two or more π vectors. The latter case cannot hold for Theorem 4.13 would then imply that S contains two or more closed sets of states. QED

Theorem 4.16

Let R be an irreducible set of states. Then R possesses a unique left eigenvector π, $\pi \geq 0$, $(\pi, 1) = 1$. Furthermore $\pi > 0$ almost everywhere on R.

Proof

$\pi > 0$ follows from Theorem 4.9. Uniqueness of π follows from Theorem 4.13 and the irreducibility of R. QED

Another criterion for irreducibility is given by

Theorem 4.17

Suppose an integer n exists such that $P^n > 0$. Then the state space Ω is irreducible and $\pi > 0$ almost everywhere on Ω.

Proof

We know that Ω contains at least one irreducible set R. If $\pi \geq 0$ is a stationary distribution on R, then $\pi P = \pi$ and $\pi P^n = \pi$. Written out, this says

$$\int_R dy\, \pi(y)\, P^n(y, x) = \pi(x) \qquad \text{all } x \in \Omega$$

The left side is positive for all x, while the right side is positive only for $x \in R$, whence $R = \Omega$. The above equation becomes $\pi(x) > 0$ everywhere. QED

Decomposition of Ω Into Irreducible Sets of States

Theorem 4.14 provides a method by which the irreducible set of states can be obtained from any closed set of states with a unique π vector. But the technique described above Theorem 4.14 enabled us to decompose Ω into say r disjoint closed sets of states $S_1, S_2, \ldots S_r$. Applying Theorems 4.14-4.15 to each of these we can decompose S_k into an irreducible set of states Ω_k plus a remaining set of states. Lumping the remaining set of states together into a set Ω_{r+1}, we have achieved the decomposition:

i) $\Omega = \Omega_1 + \ldots + \Omega_{r+1}$ $\quad$ Ω_i are disjoint

ii) $\Omega_1, \ldots \Omega_r$ are irreducible with Ω_k possessing a stationary distribution ${}^k\pi > 0$.

iii) Ω_{r+1} is not closed.

As before, $r \leq \|P\|^2$. r gives the number of "chains" within the process.

We let

$$ {}^{ij}P(x,y) = \begin{cases} P(x,y) & x \in \Omega_i, \quad y \in \Omega_j \\ 0 & \text{otherwise} \end{cases} $$

$$ 1 \leq i, j \leq r+1 $$

$$^{i}P(x,y) = {}^{ii}P(x,y) \qquad 1 \leq i \leq r+1$$

Since Ω_i is closed, $1 \leq i \leq r$, then if $x \in \Omega_i$, $P(x,y)$ vanishes unless $y \in \Omega_i$. Thus $^{ij}P(x,y) = 0 \quad i \neq j \quad 1 \leq i \leq r$. Schematically we have

$$P = \begin{bmatrix} {}^{1}P & 0 & 0 \cdots & 0 & 0 \\ 0 & {}^{2}P & 0 \cdots & 0 & 0 \\ 0 & 0 & {}^{3}P \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & {}^{r}P & 0 \\ {}^{r+1,1}P & {}^{r+1,2}P & {}^{r+1,3}P & {}^{r+1,r}P & {}^{r+1}P \end{bmatrix} \tag{4.17}$$

where the i^{th} row refers to Ω_i.

Since Ω_{r+1} is not a closed set of states,

$$\int_{\Omega_{r+1}} dy \; {}^{r+1}P(x,y) < 1$$

for a set $x \in \Omega_{r+1}$ of positive measure. Then Ω_{r+1} is a transient set of states with

$$\left\| \left({}^{r+1}P \right)^{n} \right\| \to 0$$

In particular, $\lambda = 1$ is not an eigenvalue for ${}^{r+1}P$ and ${}^{r+1}P$ does not possess a stationary distribution. The operator $(I - {}^{r+1}P)^{-1}$ will exist since 1 is not eigenvalue for ${}^{r+1}P$. Indeed all eigenvalues of ${}^{r+1}P$ will be strictly inside the unit circle, so that the expansion (see Theorem B 2)

$$(I - {}^{r+1}P)^{-1} = I + {}^{r+1}P + ({}^{r+1}P)^2 + \ldots \qquad (4.18)$$

converges in L_2 norm.

4E. Ergodicity

We say that P is ergodic if there exists an L_2 vector $\pi(x)$(or π_i) such that

$$\lim_{n \to \infty} \| P^n - P^\infty \| = 0 \tag{4.19}$$

where
$$\begin{cases} P^\infty_{ij} = \pi_j & 1 \leq ij \leq N \\ P^\infty(x,y) = \pi(y) & \text{all } x, y \in \Omega \end{cases} \tag{4.20}$$

Theorem 4.18

If P is ergodic, then the following properties hold:

P^∞ is L_2 and unique (4.21)

$$P^m P^\infty = P^\infty P^m = P^\infty \qquad m \geq 0 \tag{4.22}$$

$$(P^\infty)^2 = P^\infty = (P^\infty)^m \qquad m \geq 1 \tag{4.23}$$

P^∞ is real (4.24)

$P^\infty \geq 0$ almost everywhere (4.25)

$$P^\infty \underline{1} = \underline{1} \tag{4.26}$$

$$(\pi, 1) = 1 \tag{4.27}$$

$$\pi P = \pi \tag{4.28}$$

$\pi(x) \geq 0$ almost all x and $\pi_i \geq 0 \quad 1 \leq i \leq N$ (4.29)

$$\pi P^\infty = \pi \tag{4.30}$$

Proof

(4.21): $\| P^\infty \| = \| 1 \| \, \| \pi \| < \infty$

$$= \begin{cases} \sqrt{N} \, \| \pi \| & \text{N-state case} \\ \sqrt{V} \, \| \pi \| & \text{continuum case} \end{cases}$$

(4.22): $\| P^m P^\infty - P^\infty \| \leq \| P^m P^\infty - P^m P^n \| + \| P^m P^n - P^\infty \|$

$\leq \| P^m \| \, \| P^\infty - P^n \| + \| P^{m+n} - P^\infty \| \to 0$

(4.23): $\| (P^\infty)^2 - P^\infty \| = \| (P^\infty)^2 - P^\infty P^n \| \leq \| P^\infty \| \, \| P^\infty - P^n \| \to 0$

(4.24): Let $P^\infty = A + iB$ where A and B are real and L_2. Then P^n is real so that

$$\| B \|^2 \leq \| P^n - A \|^2 + \| B \|^2 = \| P^n - P^\infty \|^2 \to 0$$

(4.25): Since $P^n \geq 0$,

$$\iint_{P^\infty < 0} dx\, dy\, |P^\infty|^2 \leq \iint_{P^\infty < 0} dx\, dy\, |P^n - P^\infty|^2 \leq \| P^n - P^\infty \|^2 \to 0$$

(4.26): Since $P^n \underline{1} = \underline{1}$,

$$\| P^\infty \underline{1} - \underline{1} \| \leq \| (P^\infty - P^n)\underline{1} \| + \| P^n \underline{1} - \underline{1} \| \leq \| P^\infty - P^n \| \, \| 1 \| \to 0$$

(4.27): follows from (4.26).

(4.28): follows from $P^\infty P = P^\infty$.

(4.29): follows from $P^\infty \geq 0$.

(4.30) follows from $(P^\infty)^2 = P^\infty$. QED

The results show that π is a real, non-negative stationary distribution of P and justifies the notation (π as a left eigenvector) in (4.20). π turns out to be the unique stationary distribution, for if $\pi' P = \pi'$, then $\pi' = \pi' P^n \to \pi' P^\infty = (\pi', 1)\pi$ so that any distribution π' is parallel to π. The equations

$$P^\infty \underline{f} = (\pi, f)\,\underline{1} \qquad (4.31)$$

$$\underline{f} P^\infty = (f, 1)\underline{\pi}$$

for any vector $\underline{f}$ are important, since they indicate that operating with P^{∞} performs an "averaging".

We also note the property

$$\mathrm{tr}\,(P^{\infty})^n = \mathrm{tr}\,P^{\infty} = 1 \qquad n = 1 \tag{4.32}$$

The Error Operator

Since $P^n \to P^{\infty}$, it is convenient to define an "error" operator

$$E = P - P^{\infty} \tag{4.33}$$

The equations $P^{\infty}P = P\ P^{\infty} = P^{\infty} = (P^{\infty})^2$ imply that

$$EP^{\infty} = P^{\infty}E = 0 \tag{4.34}$$

i.e., $E\,\underline{1} = 0$ and $\pi E = 0$.

If the equation

$$P = P^{\infty} + E$$

is raised to the n^{th} power, the cross-terms all vanish, and one obtains

$$P^n = P^{\infty} + E^n \qquad n \geq 1 \tag{4.35}$$

Hence just as E measures the discrepancy between P and P^{∞}, so does E^n measure the discrepancy between P^n and P^{∞}.

Since $\|E^n\| = \|P^n - P^{\infty}\| \to 0$, it follows that all eigenvalues of E are strictly less than one in magnitude. Therefore E is regular at unity and $(I - E)^{-1}$ exists. The operator Z

$$Z = (I - E)^{-1} = (I - P + P^{\infty})^{-1} \tag{4.36}$$

$$Z = I + \sum_{n=1}^{\infty} E^n = I + \sum_{n=1}^{\infty} (P^n - P^{\infty})$$

is called the fundamental matrix and will play an important role in our future discussion. As(B 10) indicates, the series

$$I + \sum_{n=1}^{\infty} E^n = (I - E)^{-1}$$

converges in L_2 norm if P is ergodic, and furthermore, since E is L_2, Z - I is L_2.

An alternate proof of the existence of $(I - E)^{-1}$ which does not explicitly use $\|E^n\| \to 0$ is obtained by noting that if $(I - E)^{-1}$ fails to exist, a left eigenvector f exists for E with eigenvalue 1:

$$f = fE = f(P - P^{\infty}) = fP - (f, 1)\pi \tag{4.37}$$

Taking dot products with $\underline{1}$ yields

$$(f, 1) = (fP1) - (f, 1)(\pi, 1) = (f, 1) - (f, 1) = 0$$

whence f = fP.

But we know that the only stationary distribution for P is π (i.e., the geometric multiplicity of $\lambda = 1$ for P is one) whence $f = \pi$, contradicting (f, 1) = 0.

Since P^{∞} and E are orthogonal,

$$P^{\infty}E = EP^{\infty} = 0$$

it follows that

$$(I-zP)^{-1} = \left[(I-zP^{\infty})(I-zE)\right]^{-1} = (I-zE)^{-1}(I-zP^{\infty})^{-1} \tag{4.38}$$

By (B 7), this becomes

$$(I-zP)^{-1} = (I-zP^{\infty})^{-1} - I + (I-zE)^{-1} \tag{4.39}$$

$$= \frac{z}{1-z} P^{\infty} + (I-zE)^{-1} \tag{4.40}$$

Equation (4.40) indicates that the eigenvalue spectrum of P consists precisely of the spectrum of P^{∞} (namely 1) plus the spectrum of E (inside the unit circle). That is $\lambda \neq 0$ is an eigenvalue of P if and only if it is an eigenvalue of P^{∞} or E. This argument is extended to the eigenvectors in Theorem 4.23.

Finally the factorization

$$I - P = (I-P^{\infty})(I-E) = (I-P^{\infty})Z^{-1} \tag{4.41}$$

$$= (I-E)(I-P^{\infty}) = Z^{-1}(I-P^{\infty}) \tag{4.42}$$

of I-P, which possesses no inverse since $P\,\underline{1} = \underline{1}$, into the product of an invertible operator I-E (with inverse the fundamental matrix Z) and a simple non-invertible operator $I-P^{\infty}$ again hint at the usefulness of the concepts of the error operator and fundamental matrix.

The factorization in (4.41-4.42) and the decomposition $P = P^{\infty} + E$ of a kernel (P) at an eigenvalue (1) into an operator (P^{∞}) of finite rank singular at the eigenvalue and an operator (E) which is regular are both well known to workers in integral equations (28, pp. 167-172) but apparently are not widely-used in Markov process theory. One can look upon $P = P^{\infty} + E$ as a splitting of the P operator into projections onto two perpendicular subspaces ($P^{\infty}E = EP^{\infty} = 0$). P^{∞} would be a "steady-state" operator while E is a "transient" operator.

Theorem 4.19

If P is ergodic, then

i) the geometric multiplicity of $\lambda = 1$ for P is 1

ii) the algebraic multiplicity of $\lambda = 1$ for P is 1

iii) all other eigenvalues of P are strictly less than one in magnitude.

Conversely, if (i) and (iii) hold, then P is ergodic.

Proof

i) Let P be ergodic. If π' is any left eigenvector of P at 1, then $\pi' = \pi'P = \pi'P^n - \pi'P^\infty = (\pi', 1)\pi$ so that π is the only left eigenvector at 1. Hence the geometric multiplicity at 1 is 1.

ii) According to (B 6), $\delta(z;P) = \delta(z;P^\infty)\delta(z;E)$. Since E is regular at 1, the order of the zero of $\delta(E;P)$ at $z = 1$ is that of the zero of $\delta(z; P^\infty)$. But

$$f(z;P^\infty) = \int_0^z du \sum_{n=1}^{\infty} u^n \, tr(P^\infty)^{n+1} = \int_0^z du \, \frac{u}{1-u} = -\ln(1-z) - z$$

by use of (4.32). Then

$$\delta(z;P^\infty) = (1-z)e^z$$

indicating a simple zero at $z = 1$.

iii) If λ is an eigenvalue of P, then it has a left eigenvector f, $Pf = \lambda f$, $P^n f = \lambda^n f$. As $n \to \infty$, the left approaches $P^\infty f$ while the right has no limit unless either $\lambda = 1$ or $|\lambda| < 1$.

To prove the converse, we suppose that (i) and (iii) hold. Let π be the unique left eigenvector of P, normalized to $\pi \geq 0$, $(\pi, 1) = 1$. If P^∞ and E are defined by (4.20) and (4.33), then (4.34) holds and

$$P^n = P^\infty + E^n \qquad n \geq 1$$

We will succeed in proving that P is ergodic if we can show that all eigenvalues of E are strictly less than 1 in magnitude, for Theorem B1 would then imply that

$$\| P^n - P^\infty \| = \| E^n \| \to 0.$$

If $\lambda \neq 0$ is an eigenvalue of E with right eigenvector f, then

$$\lambda f = Ef = (P - P^\infty)f$$

$$= Pf - (\pi, f)\, \underline{1}$$

Take the scalar product with π to get $\lambda(\pi, f) = 0$ so that $Pf = \lambda f$. Hence either $\lambda = 1$ or $|\lambda| < 1$. The first case is impossible since it would imply $\underline{f} = \underline{1}$ contradicting $(\pi, f) = 0$. QED

Corollary

Let the algebraic multiplicity of $\lambda = 1$ be 1, and all other eigenvalues of P strictly inside the unit circle. Then P is ergodic.

Proof

Since the geometric multiplicity of an eigenvalue is bounded below by 1 and above by the algebraic multiplicity, the geometric multiplicity of $(\lambda = 1)$ for P is 1. Now invoke the second half of the theorem.

Note that our definition of ergodicity allows transient states. The state space Ω consists of an irreducible set R on which $\pi(x) > 0$ and a set of transient states Ω-R. The condition that the geometric multiplicity of $\lambda = 1$ is 1 guarantees that there is only one irreducible set, while the assumption that there are no other eigenvalues on the unit circle prevents R from being periodic.

Another way of preventing periodicities is given by the following theorem, and provides a test for ergodicity which is more useful than investigating the eigenvalue spectrum of P.

Theorem 4.20

Suppose an integer n exists such that $P^n > 0$ almost everywhere. Then P is ergodic and $\pi > 0$ almost everywhere (i.e., the irreducible set is Ω).

Proof

Invoking a theorem by Zaanen on positive operators (Ref. 22, pp. 496), P^n has an eigenvalue λ_0:

i) λ_0 real, $\lambda_0 > 0$, λ_0 has geometric multiplicity 1 for P^n.

ii) all other eigenvalues of P^n are strictly less than λ_0 in magnitude.

iii) the left eigenvector π can be taken real and positive almost everywhere.

But we know that the spectral radius of P^n is 1, with 1 an eigenvalue. Hence,

iv) 1 has geometric multiplicity 1 for P^n.

v) all other eigenvalues of P^n are strictly less than 1 in magnitude.

But if f is an eigenvector of P with eigenvalue λ, then f is also an eigenvector of P with eigenvalue λ^n. Results iv and v then imply that

vi) $\lambda = 1$ has geometric multiplicity 1 for P^n

vii) $\pi = 0$ is the left eigenvector of P at 1. Results vi and vii, together with Theorem 4.19, imply ergodicity.

QED

Remark

This theorem is reminiscent of the N-state case, since the communicability without periodicity of all the states guarantees ergodicity.

4F. The Single Chain Case

If Ω has only a single irreducible set of states, with unique left eigenvector π for P at 1, $\pi \geq 0$, $(\pi, 1) = 1$, then much of the analysis in 4E remains valid.

We again define E and P^∞ by

$$P^\infty(x, y) = \pi(y)$$

$$E = P - P^\infty$$

It still follows that $P P^\infty = P^\infty P = (P^\infty)^2 = P^\infty$, $\pi E = 0$, $E \underline{1} = 0$, $P^\infty \underline{1} = \underline{1}$, $\pi P = \pi P^\infty = \pi$ and so forth. Most important are the results

$$P^\infty E = E P^\infty = 0$$

$$P = P^\infty + E$$

$$P^n = P^\infty + E^n \qquad n \geq 1$$

$$(I - zP) = (I - zE)(I - zP^\infty).$$

The geometric multiplicity of $\lambda = 1$ for P is unity (Theorem 4.15) whence the proof by (4.37) again shows that the fundamental matrix

$$Z = (I - E)^{-1} = (I - P + P^\infty)^{-1} \tag{4.43}$$

exists.

The only results which fail to hold are those involving the limiting behavior. In general, E^n will not approach 0 (E will have an eigenvalue on the unit circle if P is periodic) and P^n will not approach P^∞. In particular $I + \sum_{n=1}^{\infty} E^n$ will fail to converge even though $(I - E)^{-1}$ exists.

4G. The Stable Case

In conformity with engineering parlance, we say a Markov operator P is stable if it possesses a bounded limit as time becomes infinite. That is, P is stable if there exists an L_2 operator P^∞ such that

$$\lim_{n\to\infty} \left\| P^n - P^\infty \right\| = 0 \tag{4.44}$$

Ergodicity is a special case of stability, where the additional requirement that P^∞ be of rank 1 is made.

The error operator E is again defined by

$$E = P - P^\infty$$

Just as in 4E, the following properties hold:

P^∞ is unique

$P^m P^\infty = P^\infty P^m = P^\infty \qquad m \geq 0$

$(P^\infty)^2 = (P^\infty) = (P^\infty)^m \qquad m \geq 1$

P^∞ is real $\qquad P^\infty \geq 0$

if $\pi P = \pi$, then $\qquad \pi P^\infty = \pi$

if $Pf = f$, then $\qquad P^\infty f = f$

in particular, $P^\infty \underline{1} = \underline{1}$

Also,

$$P^\infty E = E P^\infty = 0$$

$$P^n = P^\infty + E^n \qquad n \geq 1$$

$$\|E^n\| \to 0$$

All eigenvalues of E are inside unit circle. Since E is regular at 1, the fundamental matrix

$$Z = (I - E)^{-1} = (I - P + P^\infty)^{-1} \tag{4.45}$$

$$= I + \sum_{n=1}^{\infty} E^n$$

$$= I + \sum_{n=1}^{\infty} (P^n - P^\infty)$$

can still be defined.

Theorem 4.21

P is stable if and only if it has no eigenvalues on the unit circle except at $\lambda = 1$. If P is stable, then P^∞ is of some finite rank r, where $r \leq \|P\|^2$. r is both the algebraic and geometric multiplicity of $\lambda = 1$ for P. Also

$$P^\infty(x, y) = \sum_{i=1}^{r} \phi_i(x)\, \pi_i(y) \qquad \text{almost all } x, y \tag{4.46}$$

where the π_i are the left (and the ϕ_i the right) linearly independent eigenvectors of P at 1.

Proof

If P is stable with eigenvalue λ and right eigenvector f, then $Pf = \lambda f$, $\lambda^n f = P^n f \to P^\infty f$. Hence either $\lambda = 1$ or $|\lambda| < 1$. Since $PP^\infty = P^\infty$, $P^\infty(x, y)$ when considered as a function of x with y held constant (note $P^\infty(x, y)$ is L_2 in x alone for almost all y if $\|P^\infty\| < \infty$) is a right eigenvector of P with eigenvalue 1. By letting y vary we obtain various eigenvectors, but P^∞ has only finitely many, whence P^∞ must be of finite rank.

Thus (4.46) holds, where the r functions $\pi_1(x) \ldots \pi_r(x)$ are L_2 linearly independent, and left eigenvectors of P:

$$\pi_i P = \pi_i \qquad i = 1, \ldots r \tag{4.47}$$

Similarly $P^\infty(x, y)$ considered as a function of y is a right eigenvector of P. Thus the r functions $\phi_1, \phi_2, \ldots \phi_r$ are linearly independent, L_2, and right eigenvectors of P:

$$P\phi_i = \phi_i \tag{4.48}$$

Equation (4.47) shows that the geometric multiplicity of $\lambda = 1$ for P is at least r. To show it cannot exceed r, we note that if $Pf = f$, then $f = P^\infty f$ so f must be a linear combination of the r ϕ_i's.

Since $(P^\infty)^2 = P^\infty$, the linear independence of the ϕ_i, and of the π_i, requires that

$$(\pi_i, \phi_j) = \delta_{ij} \qquad 1 \le i, j \le r \tag{4.49}$$

Consequently $\mathrm{tr}(P^{\infty})^n = \mathrm{tr}\, P^{\infty} = \sum_{i=1}^{r} (\pi_i, \phi_i)$

$= r$

Then just as in the proof of Theorem (4.19)

$$f(z; P^{\infty}) = -r \ln (1-z) - r z$$

$$\delta(z; P^{\infty}) = (1-z)^r \, e^{r z}$$

$$\delta(z;P) = \delta(z; P^{\infty})\delta(z;E) = (1-z)^r \delta(z; E)\, e^{r z}$$

Since E is regular at 1, $\delta(z=1, E) \neq 0$ and the algebraic multiplicity of P at 1 is r.

$r \leq \|P\|^2$ follows from (4.14).

Conversely, suppose P has no eigenvalues on the unit circle except at unity. Since the (say r) left and right eigenvectors of P at unity may be biorthogonalized (see Theorem 4.25ii)

$$\pi_i P = \pi_i \qquad P\phi_i = \phi_i \qquad 1 \leq i \leq r$$

$$(\pi_i, \phi_j) = \delta_{ij} \qquad 1 \leq i, j \leq r$$

we can define P^{∞} by (4.46) with $PP^{\infty} = P^{\infty}P = P^{\infty}$ and the properties: $(P^{\infty})^2 = P^{\infty}$. If $E = P - P^{\infty}$, then $\pi_i E = 0$ $1 \leq i \leq r$ so that $P^{\infty}E = 0$. Similarly $EP^{\infty} = 0$. Also E is regular at $\lambda = 1$ for if $fE = f$ then $f(P-P^{\infty}) = f$ and postmultiplication by P^{∞} leads to $0 = fP^{\infty}$ whence $fP = f$. This would imply that f is a linear combination of the π_i, in which case the results $\pi_i E = 0$ and $fE = f$ are contradictory.

Then the factorization in (B 6) indicates that the spectrum of E is the spectrum of P(λ = 1 plus other eigenvalues inside the unit circle) excluding $\lambda = 1$. All eigenvalues of E are inside the unit circle, $\|E^n\| \to 0$, and the equation

$$P^n = P^\infty + E^n$$

implies that P is stable. QED

Stability and the State Space Decomposition

If the decomposition in (4.17) is employed, then it is possible to show that P is stable if and only if each of its r disjoint irreducible sets of states is ergodic.

Indeed if P is stable, (4.17) leads to

$$P^\infty = \begin{bmatrix} {}^1P^\infty & 0 & 0 \cdots 0 & 0 \\ 0 & {}^2P^\infty & 0 \cdots 0 & 0 \\ 0 & 0 & {}^3P^\infty \cdots 0 & 0 \\ \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \\ 0 & 0 & 0 \cdots {}^rP^\infty & 0 \\ A_1 & A_2 & \cdots A_r & 0 \end{bmatrix} \qquad (4.50)$$

where

$${}^kP^\infty(x, y) = {}^k\pi(y) \qquad 1 \leq k \leq r$$

$^{k}\pi$ = unique stationary distribution for ^{k}P,

$(^{k}\pi, 1) = 1$, $^{k}\pi > 0$ on Ω_k, $^{k}\pi = 0$ elsewhere

$$A_k(x, y) = f_k(x)\pi_k(y) \qquad 1 \leq k \leq r \qquad x \in \Omega_{r+1}$$

$$f_k(x) = \int_{\Omega_k (\text{or } \Omega)} dy \, (I - {}^{r+1}P)^{-1} \; {}^{r+1,k}P(x, y)$$

By insertion of (4.18), $f_k(x)$ = probability of eventually reaching Ω_k, conditioned on starting at x. $0 \leq f_k(x) \leq 1$ for $1 \leq k \leq r$.

If r functions $^{k}\phi(x)$ are defined by

$$^{k}\phi(x) = \begin{cases} 1 & x \in \Omega_k \\ 0 & x \in \Omega_j \quad 1 \leq j \leq r, \; j \neq k \\ f_k(x) & x \in \Omega_{r+1} \end{cases}$$

for $1 \leq k \leq r$, then (4.19) may be rewritten as

$$P^{\infty}(x, y) = \sum_{k=1}^{r} {}^{k}\phi(x) \, {}^{k}\pi(y) \qquad x, y \in \Omega \tag{4.51}$$

in confirmation of (4.46). The quantities $^{k}\phi$ satisfy

$$0 \leq {}^{k}\phi(x) \leq 1 \tag{4.52}$$

$^{k}\phi(x)$ = probability of reaching Ω_k, conditioned on starting at $x \in \Omega$. (4.53)

Then (4.51) has the interpretation that the steady-state probability of being at y, starting at x, is the summation over all chains of the

probability of reaching that chain from x times the probability of being at y conditioned on starting in that chain.

In particular, for any $x \in \Omega$, $P^{\infty}(x, y) = 0$ for any $y \in \Omega_{r+1}$. Thus, Ω_{r+1} is a transient set of states.

In addition,

$$\sum_{k=1}^{r} {}_k\phi(x) = 1 \qquad x \in \Omega$$

so that the probability of reaching some irreducible set of states is one.

Warning

While the biorthogonality $(\pi_i, \phi_j) = \delta_{ij}$ and even the choice $\pi_i(y) \geq 0$, $(\pi_i, 1) = 1$ holds for (4.46), the interpretation of $\phi_i(x)$ as the probability of reaching the i^{th} chain does not hold unless the ϕ_i and π_i are picked cleverly, with π_i being the stationary distribution for the i^{th} chain.

For example, the process

$$P = P^{\infty} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

has $\Omega_k = \{\text{state } k\}$ $k = 1, 2$

$$P_{ij}^{\infty} = \sum_{k=1}^{2} {}_k\phi_i \, {}_k\pi_i$$

where

$${}_k\phi_i = {}_k\pi_i = \delta_{ik} \qquad 1 \leq i, k \leq 2$$

On the other hand, the less clever choice

$$P^{\infty}_{ij} = \sum_{k=1}^{2} {}^{k}\phi'_{i} \; {}^{k}\pi'_{j}$$

where ${}^{1}\phi' = (-1 \quad 4)$

${}^{2}\phi' = (2 \quad -3)$

${}^{1}\pi' = (.6 \quad .4)$

${}^{2}\pi' = (.8 \quad .2)$

has ${}^{k}\pi' \geq 0$ but does not meet $0 \leq {}^{k}\phi_{i} \leq 1$. Clearly the failure to separate the closed sets into irreducible closed sets is responsible.

4H. The General Case

In the most general case of a Markov chain with L_2 transition matrix P--the ergodic, single chain and stable cases being specializations--the decomposition in (4.17) still holds.

Since Ω_k is irreducible, $1 \leq k \leq r$, ${}^kP$ has a unique left eigenvector ${}^k\pi(x)$ ($=0$ $x \notin \Omega_k$, >0 $x \in \Omega_k$) with normalization ${}^k\pi \geq 0$, $({}^k\pi, 1) = 1$. Just as in Section 4F we may let

$$ {}^kP^{\infty}(x,y) = {}^k\pi(y) \qquad 1 \leq k \leq r $$

$$ {}^kE = {}^kP - {}^kP^{\infty} $$

Again we define

$$ A_k(x,y) = f_k(x)\,{}^k\pi(y) \qquad 1 \leq k \leq r\,,\ x \in \Omega_{r+1}\,,\ y \in \Omega_k $$

where

$$ f_k(x) = \int_{\Omega_k} dy\, (I - {}^{r+1}P)^{-1}\ {}^{r+1,k}P(x,y) $$

$$ = \text{probability of reaching } \Omega_k \text{ from } x $$

$$ 0 \leq f_k(x) \leq 1 $$

We define P^{∞} schematically by

$$ P^{\infty} = \begin{bmatrix} {}^1P^{\infty} & 0 & 0 \ldots 0 & 0 \\ 0 & {}^2P^{\infty} & 0 \ldots 0 & 0 \\ 0 & 0 & 0 \ldots . & . \\ . & . & . \quad . & . \\ 0 & 0 & 0 \ldots {}^rP^{\infty} & 0 \\ A_1 & A_2 & A_3 \ldots A_r & 0 \end{bmatrix} \tag{4.54} $$

and $E = P - P^{\infty}$.

It follows from

$$ {}^{k}P^{\infty} P^{j} = {}^{j}P\, {}^{k}P^{\infty} = {}^{k}P^{\infty} \delta_{jk} \qquad 1 \leq j,\ k \leq r $$

$$ ({}^{r+1,k}P\, {}^{k}P^{\infty} + {}^{r+1}PA_{k})(x,y) $$

$$ = ({}^{r+1,k}P\, \underline{1} + {}^{r+1}P(I - {}^{r+1}P)^{-1}\ {}^{r+1,k}P\, \underline{1})(x)\ {}^{k}\pi(y) $$

$$ = ((I - {}^{r+1}P^{-1}\ {}^{r+1,k}P\, \underline{1})(x)\ {}^{k}\pi(y) = A_{k}(x,y) $$

$$ 1 \leq k \leq r $$

and so forth that

$$ PP^{\infty} = P^{\infty}P = (P^{\infty})^{2} = P^{\infty} \tag{4.55} $$

In addition, (proofs omitted)

$$ P^{\infty} \geq 0 \tag{4.56} $$

$$ P^{\infty}\, \underline{1} = \underline{1} \qquad \left(\text{since } \sum_{k=1}^{r} A_{k}\, \underline{1} = \underline{1}\right) \tag{4.57} $$

P^{∞} is singular at 1 (since $(P^{\infty})^{2} = P^{\infty}$) and (4.58)

regular everywhere else

P^{∞} is of finite rank r: (4.59)

$$ P^{\infty}(x,y) = \sum_{k=1}^{r} {}^{k}\phi(x)\ {}^{k}\pi(y) \tag{4.60} $$

where

$$ {}^{k}\phi(x) = \begin{cases} 1 & x \in \Omega_{k} \\ 0 & x \in \Omega_{j} \quad j \neq k \quad 1 \leq j \leq r \\ f_{k}(x) & x \in \Omega_{r+1} \end{cases} $$

= probability of reaching Ω_k, starting from x.

$$0 \leq {}^{k}\phi(x) \leq 1 \tag{4.61}$$

$$({}^{j}\pi, {}^{k}\phi) = \delta_{jk} \qquad 1 \leq j, k \leq r \tag{4.62}$$

(since $(P^{\infty})^2 = P^{\infty}$)

$$({}^{j}\pi)P = ({}^{j}\pi)P^{\infty} = {}^{j}\pi \qquad 1 \leq j \leq r \tag{4.63}$$

$$P({}^{j}\phi) = P^{\infty}({}^{j}\phi) = {}^{j}\phi \qquad 1 \leq j \leq r \tag{4.64}$$

$$({}^{j}\pi)E = 0 \qquad 1 \leq j \leq r \quad \text{(from 4.63)} \tag{4.65}$$

$$E({}^{j}\phi) = 0 \qquad 1 \leq j \leq r \quad \text{(from 4.64)} \tag{4.66}$$

$$EP^{\infty} = P^{\infty}E = 0 \qquad \text{(from (4.65-4.66))} \tag{4.67}$$

$$P^{n} = P^{\infty} + E^{n} \qquad n \geq 1 \tag{4.68}$$

The fundamental matrix is again defined by

$$Z = (I - E)^{-1} = (I - P + P^{\infty})^{-1} \tag{4.69}$$

Theorem 4.22

Z exists.

Proof

If the inverse in (4.58) fails to exist, then 1 is an eigenvalue of $P - P^{\infty}$, with corresponding right eigenvector say h: $(P-P^{\infty})h = h$.

Decomposition of h into the r + 1 subspaces $\Omega_1, \ldots \Omega_{r+1}$: $h = ({}^{1}h, {}^{2}h, \ldots {}^{r+1}h)$ leads to the r + 1 equations

$${}^{k}E\,{}^{k}h = {}^{k}h \qquad 1 \leq k \leq r$$

$$\sum_{k=1}^{r} ({}^{r+1,k}P - A_k)\,{}^{k}h + {}^{r+1}P\,{}^{r+1}h = {}^{r+1}h$$

Since ${}^{k}E$ is regular at 1, the first r equations imply

$${}^{k}h = 0 \qquad 1 \leq k \leq r$$

The $r+1^{st}$ equation becomes ${}^{r+1}P\,{}^{r+1}h = {}^{r+1}h$ and since ${}^{r+1}P$ is regular at 1 implies ${}^{r+1}h = 0$. Thus $h = 0$ and $P - P^{\infty}$ is regular at $\lambda = 1$. QED

Corollary

$(I - zE)^{-1}$ is analytic in z near $z = 1$.

Proof

A small neighborhood about $z = 1$ is regular for E. In this neighborhood $(I - zE)^{-1}$ exists and is infinitely differentiable. QED

Theorem 4.23 Eigenfunctions of P and E. [28, pp 167-172]

i) If $\lambda \neq 1$ or 0, $Pf = \lambda f$ if and only if $Ef = \lambda f$; $P^{\infty} f = 0$.

ii) If $\lambda = 1$, $Pf = f$ if and only if $P^{\infty} f = f$. Similarly for left eigenvectors.

Proof

If $\lambda \neq 1$, $\lambda f = Pf = P^{\infty} f + Ef$ implies that $\lambda P^{\infty} f = P^{\infty} f$ whence $P^{\infty} f = 0$ and $Ef = \lambda f$.

If $\lambda \neq 1, 0$ then $Ef = \lambda f$ implies that $0 = P^{\infty} Ef = \lambda P^{\infty} f$ whence $P^{\infty} f = 0$ and $\lambda f = (P^{\infty} + E)f = Pf$

If $\lambda = 1$, $Pf = f$ implies $f = (P^{\infty} + E)f = (I-E)^{-1} P^{\infty} f = Z P^{\infty} f = P^{\infty} f$ (see (4.70))

If $\lambda = 1$, $P^{\infty} f = f$ implies $Ef = EP^{\infty} f = 0$ so that $f = P^{\infty} f + Ef = Pf$. QED

This theorem shows that just as E and P^{∞} pick up the eigenvalues of P, so too do they pick up the eigenvectors.

Theorem 4.24 Properties of the Fundamental Matrix

$$ZP^{\infty} = P^{\infty}Z = P^{\infty} \qquad \text{(Z commutes with } P^{\infty}\text{)} \tag{4.70}$$

$$ZP = PZ = Z - I + P^{\infty} \qquad \text{(Z commutes with P)} \tag{4.71}$$

$$Z\,\underline{1} = \underline{1} \tag{4.72}$$

$$Z(I - P) = (I - P)Z = I - P^{\infty} \tag{4.73}$$

$$(I - P) = (I - P^{\infty})Z^{-1} = Z^{-1}(I - P^{\infty}) \quad \text{factorization} \tag{4.74}$$

Proof

Since $Z = (I - P + P^{\infty})^{-1}$,

$$Z = I + Z(P - P^{\infty}) \tag{4.75}$$

whence $ZP^{\infty} = P^{\infty}$. Similarly $P^{\infty}Z = P^{\infty}$ follows from $Z = I + (P - P^{\infty})Z$.

From (4.75) follows

$$ZP = Z - I + ZP^{\infty} = Z - I + P^{\infty}$$

Similarly $PZ = Z - I + P^{\infty}$

$$Z\,\underline{1} = Z\left[(I - P + P^{\infty})\,\underline{1}\right] = ZZ^{-1}\,\underline{1} = \underline{1}$$

Equation (4.73) follows from (4.71)

$$I - P = (I - P^{\infty})(I - E) = (I - P^{\infty})Z^{-1}$$

$$= (I - E)(I - P^{\infty}) = Z^{-1}(I - P^{\infty}) \qquad \text{QED}$$

Our general theorem about spectral properties of P is now given in terms of the decomposition given above.

Theorem 4.25

i) The algebraic and geometric multiplicities of P at $\lambda = 1$ are identical, say r, and agree with the number of disjoint irreducible sets of states of P and with the rank of P^{∞}.

ii) The r left eigenvectors of P at 1 are ${}^{k}\pi$, $1 \leq k \leq r$ and the r right eigenvectors are ${}^{k}\phi$ $1 \leq k \leq r$ defined above. They may be biorthogonalized: $({}^{i}\pi, {}^{j}\phi) = \delta_{ij}$ $1 \leq i, j \leq r$.

iii) The resolvent $(I - zP)^{-1}$ has a simple pole at z = 1:

$$(I - zP)^{-1} = (I - zP^{\infty})^{-1} - I + (I - zE)^{-1}$$

$$= \frac{zP^{\infty}}{1-z} + (I - zE)^{-1} \tag{4.76}$$

$$= \frac{zP^{\infty}}{1-z} + Z + O(1 - z) \tag{4.77}$$

iv) The number of chains r is finite and satisfies (4.14):

$$r \leq \| P \|^{2}$$

The bound is tight in the N-state case if P = I.

Proof

i) According to Theorem 4.23 ii, the geometric multiplicity of P and P^{∞} at $\lambda = 1$ agree. The latter has geometric multiplicity r (${}^{j}\phi$ $1 \leq j \leq r$ are the r right eigenvectors of P and P^{∞} at 1, and ${}^{j}\pi$ the r left eigenvectors of P and P^{∞} at 1), hence the geometric multiplicity of 1 for P is r.

To calculate the algebraic multiplicity of $\lambda = 1$ for P we use (B 6) to obtain

$$\delta(z; P) = \delta(z; P^{\infty})\ \delta(z; E)$$

But $\mathrm{tr}\,(P^{\infty})^n = r$ so that, just as in Theorem 4.21,

$$\delta(z; P^{\infty}) = (1-z)^r e^{rz}$$

Since E is regular at 1, $\delta(z; E) \neq 0$ and $\delta(z; P)$ has an r-fold zero at $z = 1$. Thus the algebraic multiplicity of $\lambda = 1$ for P is r. This completes the proof of (i).

(ii) is proved by (4.63-4.64), with the biorthogonalization in (4.62). An alternate proof is by Theorem 4.23ii.

(iii) is proved by use of (B 7). Note that by the Corollary to Theorem 4.22, $(I - zE)^{-1}$ is analytic at $z=1$. Its value at $z = 1$ is $(I - E)^{-1} = Z$, whence

$$(I - zE)^{-1} = Z + O(z-1) \qquad \text{QED}$$

Example 1

$$P = \begin{bmatrix} .8 & .2 & 0 & 0 & 0 \\ .6 & .4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & .5 & .5 & 0 & 0 \end{bmatrix}$$

$r = 2$ chains : $\Omega_1 = (1,2)$ ${}^1\pi = (.75, .25, 0, 0, 0)$

${}^1\phi = (1, 1, 0, 0, .5)$

$\Omega_2 = (3,4)$ ${}^2\pi = (0, 0, .5, .5, 0)$

${}^2\phi = (0, 0, 1, 1, .5)$

$\Omega_3 = (5)$ = transient state

Can verify that $({}^i\pi, {}^j\phi) = \delta_{ij}$ $1 \leq i, j \leq 2$. And that ${}^i\pi$ and ${}^i\phi$ are the only vectors for which ${}^i\pi P = {}^i\pi$ and $P{}^i\phi = {}^i\phi$ $1 \leq i \leq 2$. Hence geometric multiplicity of $\lambda=1$ for P is $r = 2$.

$$\det (I - zP) = (z - 1)^2 z(z - .2)(z + 1)$$

so geometric multiplicity of $\lambda = 1$ is $r = 2$.

$$P^{\infty}_{ij} = \sum_{k=1}^{2} ({}^k\phi)_i ({}^k\pi)_j \text{ implies rank of } P^{\infty} \text{ is } r = 2.$$

$$P^{\infty} = \begin{bmatrix} .75 & .25 & 0 & 0 & 0 \\ .75 & .25 & 0 & 0 & 0 \\ 0 & 0 & .5 & .5 & 0 \\ 0 & 0 & .5 & .5 & 0 \\ .375 & .125 & .25 & .25 & 0 \end{bmatrix}$$

Example 2

The N-state process with $P = I$ has $\det (I-zP) = (1-z)^N$ confirming that the order of the zero of the Fredholm denominator, here N,

gives the number of irreducible sets of states (sub-chains). In addition

$$(I - zP)^{-1} = \frac{I}{1-z} = \frac{P^{\infty}}{1-z}$$

confirming both that the resolvent $(I - zP)^{-1}$ has a simple pole at $z = 1$ and that the coefficient multiplying $1/1-z$ is P^{∞}.

As the theorem and examples show, the number of chains r can be quickly obtained by calculating the algebraic multiplicity at $\lambda = 1$, i.e., det $(I - zP)$.

If we recall how $S = \lim S_m$ was defined by (4.11), we see that $P^{\infty} = S$. This is no accident, for if S exists, then the equation $PS = P$ implies that the columns of S are right eigenvectors of P at $\lambda = 1$ while the equation $SP = S$ implies that the rows of S are left eigenvectors of P at $\lambda = 1$. Therefore S is of finite rank:

$$S(x, y) = \sum_{i,j=1}^{r} {}^{i}\phi(x)\, T_{ij}\, {}^{j}\pi(y) \qquad x, y \in \Omega$$

Since the ϕ's and π's are biorthogonal,

$$T_{ij} = ({}^{i}\pi,\ S\,{}^{j}\phi) = \lim_{m\to\infty} ({}^{i}\pi,\ S_m({}^{j}\phi))$$

But $({}^{i}\pi,\ S_m\,{}^{i}\phi) = ({}^{i}\pi,\ {}^{j}\phi) = \delta_{ij}$, whence $T = I$ and

$$S = P^{\infty} \tag{4.78}$$

CHAPTER 5

THE FUNDAMENTAL MATRIX

5A. Introduction

If P is stable, $\|P^n - P^\infty\| \to 0$ and the fundamental matrix Z is defined by (4.45) as

$$Z = (I - P + P^\infty)^{-1} \tag{5.1}$$

$$Z = I + \sum_{n=1}^{\infty} (P^n - P^\infty) \tag{5.2}$$

convergent in norm

Even in the general case where P is not stable, Z is given by (5.1) where P^∞ is defined by (4.54). The properties of P^∞ are given by (4.55 - 4.68).

Most of the discussion in this chapter is limited to the case where P has a unique irreducible set of states, i.e., a unique stationary distribution $\pi = \pi P$ which we will always normalize to $\pi \geq 0, (\pi, 1) = 1$. In this case, P^∞ is of rank 1: $P^\infty(x,y) = \pi(y)$. The ergodic case, it will be recalled, is a specialization of this case in which periodicities are not allowed.

The arguments in this chapter apply equally well to the N-state and continuum-state cases.

In this chapter we discuss the eigenvalue spectrum of Z and the usefulness of Z for solving equations containing the kernel I - P. The fundamental matrix contains within it a virtually complete description of the transient and first passage properties of an N-state semi-Markov chain.

The chapter concludes with the derivation of expressions, involving Z, for the absolute (not relative) values of Markov and semi-Markov processes with stationary policies.

In Chapter 6, particularly (6.10), we will show that the fundamental matrix is useful for treating perturbations of a Markov chain containing a unique irreducible set of states, and is intimately connected with the ergodic behavior of the chain.

Another advantage of the fundamental matrix is that it can be obtained numerically for large N-state Markov chains with a unique irreducible set of states ($N >> 1$) by matrix inversion. The most straightforward scheme is to solve

$$\underline{\pi}\,(I - P) = \underline{0} \qquad \sum_i \pi_i = 1 \tag{5.3}$$

for π, thus P^*. Then inversion of $(I - P + P^*)$ produces Z. This method involves two inversions of N x N matrices and can be carried out on a digital computer even if N is several hundred. An alternate method, presented in Section 6E, involves only one matrix inversion. Once the numerical values for the Z_{ij} are known, of course, all of the formulas (5.32, 5.40, 5.41, 5.127, and so forth) can be immediately applied.

By contrast, the generating function approach discussed in [5], although providing important theoretical insight into the transient behavior and so forth, is totally impractible for large systems, due to the difficulty of locating the (possibly complex) zeros of $\det(I - zP)$ and the resultant difficulty in performing a partial fraction expansion of $(I - zP)^{-1}$. Flow graph analysis of Markov processes also founders on the

shoals of dimensionality.

The above remarks indicate that the fundamental matrix is useful both theoretically and computationally for providing information about the steady-state and transient behavior of semi-Markov chains and of semi-Markovian reward processes. The fundamental matrix deserves an important position in the analyst's bag of tools for treatment of Markov chains.

5B. Properties of the Fundamental Matrix

The following properties of Z are recalled from Theorem 4.24:

First Z commutes with P and P^{∞} :

$$Z P^{\infty} = P^{\infty} Z = P^{\infty} \tag{5.4}$$

$$Z P = P Z = Z - I + P^{\infty} \tag{5.5}$$

Second, the factorization

$$(I - P) Z = Z (I - P) = I - P^{\infty} \tag{5.6}$$

which we will use often.

Third, the rows of Z sum to unity:

$$Z \underline{1} = \underline{1} \tag{5.7}$$

However, as the example in Section 6C shows, Z is not a stochastic matrix: it may possess negative matrix elements.

Fourth, as (5.1) shows, Z is always invertible.

5C. Eigenvalue Spectrum of Z

Theorem 5.1

a) 1 is always an eigenvalue of Z

b) 0 is never an eigenvalue of Z

c) if e(f) is a left (right) eigenvector of P with eigenvalue $\lambda \neq 0$, then e(f) is a left (right) eigenvector of P with eigenvalue

$$\begin{cases} \dfrac{1}{1-\lambda} & \lambda \neq 1 \\ 1 & \lambda = 1 \end{cases}$$

d) if e(f) is a left (right) eigenvector of Z with eigenvalue $\lambda (\neq 0)$, then e(f) is a left (right) eigenvector of P with eigenvalue $1 - \frac{1}{\lambda}$ if $\lambda \neq 1$; if $\lambda = 1$, e(f) lies in the left (right) nullspace of $P - P^{\infty}$

e) all eigenvalues λ of Z lie in the region

$$\mathrm{re}\ \lambda \geq \frac{1}{2} \tag{5.8}$$

Proof

a) follows from (5.7)

b) Z is invertible, hence $Zf = \underline{0}$ implies $f = Z^{-1}0 = \underline{0}$

c) suppose $Pf = \lambda f$

If $\lambda = 1$, Theorem 4.23 ii implies that $P^{\infty} f = f$

so that

$$Z^{-1} f = (I - P + P^{\infty}) f = f$$
$$f = Zf$$

If $\lambda \neq 1$, Theorem 4.23 i implies that

$$P^{\infty} f = 0$$

so that

$$Z^{-1}f = (I - P + P^{\infty})f = (1 - \lambda)f$$

and

$$Zf = \frac{1}{1-\lambda} f$$

A similar proof holds for left eigenvectors.

d) Suppose $Zf = \lambda f$. Multiply by P^{∞} and use (5.4) to obtain $(1 - \lambda)P^{\infty} f = 0$. If $\lambda \neq 1$, then $P^{\infty} f = 0$, and, since 0 is not an eigenvalue of Z,

$$f = \lambda Z^{-1}f = \lambda(I - P + P^{\infty})f = \lambda(I - P)f$$

$$Pf = (1 - \frac{1}{\lambda})f$$

If $\lambda = 1$, then $Zf = f$, $f = Z^{-1}f = (I - P + P^{\infty})f$ so that $(P - P^{\infty})f = 0$.

A similar proof holds for left eigenvectors.

e) If λ is an eigenvalue of Z, a right eigenvector f must exist for Z with λ as its eigenvalue. Invoke part d to conclude that either $\lambda = 1$ or $1 - \frac{1}{\lambda}$ is an eigenvalue of P (or both). But $\lambda = 1$ clearly satisfies (5.8).

Since all eigenvalues of P are not greater than unity in magnitude (Theorem 4.4), the second possibility satisfies

$$\left|1 - \frac{1}{\lambda}\right| \leq 1 \tag{5.9}$$

which can be shown to be equivalent to (5.8)

QED

In particular, (c) states that if $\pi P = \pi$, then

$$\pi Z = \pi \tag{5.10}$$

so that the steady state vector π of an ergodic chain P is always a left eigenvector of Z with eigenvalue 1.

5D. Recovery of P From Z

Theorem 5.2

Let P have a unique irreducible set of states and let P^{-1} exist. Then P can be recovered from Z.

Proof

Since $\lambda=1$ is an eigenvalue of Z, the equation f = fZ has at least one solution. We show now that it has precisely one solution with f proportional to the unique π vector for P.

$$f = fZ \qquad f = fZ^{-1} = f(I - P + P^{\infty})$$
$$fP = fP^{\infty} = (f,1)\pi$$

But $\pi P = \pi$ implies $\pi = \pi P^{-1}$ and therefore

$$f = (fP)P^{-1} = (f,1)\pi P^{-1} = (f,1)\pi$$

Consequently the unique left eigenvector f of Z at 1, after a normalization to (f,1) = 1, gives the π vector and then $P^{\infty}(x,y) = \pi(y)$. Then P can be recovered from Z via $P = -Z^{-1} + P^{\infty} + I$

QED

In general, P cannot be recovered from Z if P^{-1} fails to exist. For example, the set of P matrices with all rows identical (these satisfy det P = 0, so P^{-1} fails to exist) all satisfy $P = P^{\infty}$, so that Z = I. Thus, P cannot be recovered if the only information given is that Z = I.

5E. The Kernel I - P

The fundamental matrix provides a powerful tool for solution of equations involving the kernel I - P. The reason for this is the convenient technique of kernel factorization derived in Chapter 4. If P is stable, then

$$I - P = (I - P^{\infty})(I - E) = (I - P^{\infty})Z^{-1} \tag{5.11}$$

$$I - P = (I - E)(I - P^{\infty}) = Z^{-1}(I - P^{\infty}) \tag{5.12}$$

Our results concerning the solutions of such equations are contained in the following two theorems which are closely related to the Fredholm alternative theorem[28 , Pg. 172.]

Theorem 5.3

The equation

$$(I - P)U = A \tag{5.13}$$

has a solution U if and only if

$$P^{\infty} A = 0 \tag{5.14}$$

If the consistency condition in (5.14) is met, then the most general

solution to (5.13) is

$$U = Z A + P^{\infty} M \tag{5.15}$$

where $M(= P^{\infty} U)$ is arbitrary.

The equation

$$W(I - P) = B \tag{5.16}$$

has a solution W if only if

$$B P^{\infty} = 0 \tag{5.17}$$

If the consistency condition in (5.17) is met, then the most general solution to (5.16) is

$$W = B Z + N P^{\infty} \tag{5.18}$$

where $N(= WP^{\infty})$ is arbitrary.

Proof

The necessity of (5.14) and (5.17) follow from (5.13) and (5.16). If (5.13) has a solution, then (5.12) implies that the solution must be

$$U = Z A + P^{\infty} U$$

If (5.16) has a solution, then (5.11) implies that the solution

must be

$$W = BZ + WP^{\infty}$$

Conversely, to show the sufficiency of (5.14) and (5.17), we compute, by (5.6),

$$(I-P)(ZA+P^{\infty}M) = (I-P^{\infty})A = A$$
$$(BZ+NP^{\infty})(I-P) = B(I-P^{\infty}) = B$$

QED

Corollary

If P is a Markov chain with a unique irreducible set of states, then

$$(I-P)\,\underline{x} = \underline{a} \qquad (\pi, a) = 0 \tag{5.19}$$

implies that

$$\underline{x} = Z\,\underline{a} + c\,\underline{1} \qquad c = (\pi, x) \tag{5.20}$$

The equation

$$\underline{y}\,(I-P) = \underline{b} \qquad (b, 1) = 0 \tag{5.21}$$

implies that

$$\underline{y} = \underline{b}\, Z + d\, \underline{\pi} \qquad d = (y, I) \tag{5.22}$$

The constants c and d cannot be determined from (5.19) and (5.21).

Theorem 5.4

The equations

$$(I - P) X = A \tag{5.23}$$

$$X (I - P) = B \tag{5.24}$$

are solvable for X if and only if

$$P^{\infty} A = 0 \tag{5.25 a}$$

$$B P^{\infty} = 0 \tag{5.25 b}$$

$$A(I-P) = (I-P) B \tag{5.25 c}$$

If these consistency constraints are met, then the most general solution to (5.23 - 5.24) is

$$X = Z A + P^{\infty} B Z + P^{\infty} M P^{\infty} \tag{5.26}$$

$$= B Z + Z A P^{\infty} + P^{\infty} M P^{\infty} \tag{5.27}$$

where $M\ (= P^{\infty} \times P^{\infty})$ is arbitrary.

Proof

Suppose (5.23) - 5.24) have a solution X. Then (5.25 a,b,c) follow directly. To find X, we may apply (5.15) to (5.23) to obtain

$$X = ZA + P^{\infty} X \tag{5.28}$$

If we apply (5.18) to (5.24), we obtain

$$X = BZ + X P^{\infty} \tag{5.29}$$

Insertion of (5.29) on the right of (5.28) yields (5.26), while insertion of (5.28) on the right of (5.29) yields (5.27).

Conversely, if (5.25 a,b,c) hold, it is possible to show that (5.26) and (5.27) both satisfy (5.23 - 5.24):

$$\begin{aligned}
(I-P)(ZA+P^{\infty}BZ+P^{\infty}MP^{\infty}) &= (I-P)ZA \\
&= (I-P^{\infty})A = A \\
(ZA+P^{\infty}BZ+P^{\infty}MP^{\infty})(I-P) &= Z(I-P)B+P^{\infty}BZ(I-P) \\
&= (I-P^{\infty})B+P^{\infty}B(I-P^{\infty}) \\
(I-P)(BZ+ZAP^{\infty}+P^{\infty}MP^{\infty} &= B-P^{\infty}B+P^{\infty}B = B \\
&= A(I-P)Z+(I-P)ZAP^{\infty} \\
&= A(I-P^{\infty})+(I-P^{\infty})AP^{\infty} \\
&= A-AP^{\infty}+AP^{\infty} = A \\
(BZ+ZAP^{\infty}+P^{\infty}MP^{\infty})(I-P) &= BZ(I-P) \\
&= B(I-P^{\infty}) = B
\end{aligned}$$

QED

5F. Mean First Passage Times for Semi-Markov Chains

We give an example of the usefulness of Theorem 5.3 by deriving expressions, in terms of the $\mathbb{Z}$ matrix, for the mean time u_{ij} and mean square time $u^{(2)}_{ij}$ until the system, starting from state i, lands in state j for the first time. We assume in 5F that we have an N-state Markov chain whose irreducible set of states consists of all N-states. This implies that $\pi_j > 0 \quad 1 \leq j \leq N$ so that all N-states are recurrent. (Note that ergodicity is not assumed: the process could be periodic).

We denote by T_i the mean holding time in state i, by $T^{(2)}_i$ the mean square holding time in state i, and by T_{ij} the quantity P_{ij} multiplied by the mean time for a transition from i to j.

The matrices u and $u^{(2)}$ satisfy consistency equations which may be obtained by the following considerations. The mean time to land in j from i is the mean time for the first transition plus the mean time for remaining transitions, if these prove necessary. Consequently,

$$u_{ij} = \pi_i + \sum_{\substack{K=1 \\ K \neq j}}^{N} P_{iK} u_{Kj} \qquad (5.30)$$

$$1 \leq i, j \leq N$$

Similarly, the mean square time is the expected value of square of the sum of the time for the first transition plus the time for remaining transitions, if these prove necessary. Consequently,

$$U^{(2)}_{ij} = \pi_i^{(2)} + \sum_{\substack{k=1 \\ k \neq j}}^{N} \left[P_{ik} U^{(2)}_{kj} + 2 \pi_{ik} U_{kj} \right] \qquad (5.31)$$

$$1 \leq i, j \leq N$$

Theorem 5.5

$$U_{ij} = \sum_{k=1}^{N} (Z_{ik} - Z_{jk}) \pi_k + \frac{\langle \pi \rangle}{\pi_j} (\delta_{ij} + Z_{jj} - Z_{ij}) \qquad (5.32)$$

$$1 \leq i, j \leq N$$

where

$$\langle \pi \rangle = \sum_{i=1}^{N} \pi_i \pi_i = (\pi, \pi) \qquad (5.33)$$

= mean time per transition

In particular,

$$U_{jj} = \frac{\langle \pi \rangle}{\pi_j} \qquad 1 \leq j \leq N \qquad (5.34)$$

Proof

Rewrite (5.30) as

$$U_{ij} = \pi_i + (PU)_{ij} - P_{ij} U_{jj} \qquad (5.35)$$

Multiply by π_i and sum over i to get, using $\pi P = \pi$

$$U_{jj} = \frac{\langle \pi \rangle}{\pi_j}$$

which proves (5.34).

Rewrite (5.35) as

$$(I - P)\,U = F \tag{5.36}$$

where

$$F_{ij} = \pi_i - \frac{P_{ij}\langle\pi\rangle}{\pi_j} \qquad 1 \le i,j \le N \tag{5.37}$$

Invoke (5.15) --after checking that $\sum_{i=1}^{N} \pi_i F_{ij} = 0$ --to obtain

$$U_{ij} = (ZF)_{ij} + (P^{\infty}U)_{ij}$$

The second term on the right is independent of i and can be eliminated by our knowledge of U_{jj}:

$$U_{ij} = (ZF)_{ij} - (ZF)_{jj} + U_{jj} \tag{5.38}$$

Since $ZP = Z - I + P^{\infty}$,

$$(ZF)_{ij} = \sum_{k=1}^{N} Z_{ik}\pi_k - \frac{\langle\pi\rangle}{\pi_j}\left(Z_{ij} - \delta_{ij} + \pi_j\right) \tag{5.39}$$

Insertion of (5.39) into (5.38) yields (5.32).

QED

Theorem 5.6

$$U_{ij}^{(2)} = (\delta_{ij} - Z_{ij} + Z_{jj}) U_{jj}^{(2)} + \sum_{m=1}^{N} (Z_{im} - Z_{jm}) \tau_m^{(2)} \tag{5.40}$$

$$+ 2 \left[(Z \Pi U)_{ij} - (Z \Pi U)_{jj} \right]$$

$$- 2 \left[(Z \Pi)_{ij} - (Z \Pi)_{jj} \right] U_{jj}$$

$$1 \le i, j \le N$$

where

$$U_{jj}^{(2)} = \frac{1}{\pi_j} \left[\langle \tau^{(2)} \rangle + 2 \sum_{i=1}^{N} \sum_{\substack{K=1 \\ K \neq j}}^{N} \pi_i \tau_{iK} U_{Kj} \right] \tag{5.41}$$

$$1 \le j \le N$$

$$\langle \tau^{(2)} \rangle = \sum_{i=1}^{N} \pi_i \tau_i^{(2)} \tag{5.42}$$

= mean square time per transition

Proof

Rewrite (5.31) as

$$(I - P) U^{(2)}_{ij} = F_{ij} - P_{ij} U^{(2)}_{jj} \tag{5.43}$$

where

$$F_{ij} = \pi_i^{(2)} + 2 \sum_{\substack{K=1 \\ K \neq j}}^{N} \pi_{iK} U_{Kj} \tag{5.44}$$

Multiplication of (5.43) by π_i and summation over i, using $\pi(I-P)=0$, yields

$$U^{(2)}_{jj} = \frac{1}{\pi_j} \sum_{i=1}^{N} \pi_i F_{ij} \tag{5.45}$$

which demonstrates (5.41).

Invoke (5.15)--now that (5.14) is satisfied--to obtain the solution to (5.43):

$$U^{(2)}_{ij} = \sum_{K=1}^{N} Z_{iK} \left[F_{Kj} - P_{Kj} U^{(2)}_{jj} \right] + \left[P^{\infty} U^{(2)} \right]_{ij} \tag{5.46}$$

The last term on the right is independent of i, and may be eliminated by knowledge of $u^{(2)}_{jj}$. If we also use $ZP = Z - I + P^{\infty}$, equation (5.46) becomes

$$U^{(2)}_{ij} = \sum_{K=1}^{N} (Z_{iK} - Z_{jK}) F_{Kj} - (Z_{ij} - Z_{jj} - \delta_{ij} + 1) U^{(2)}_{jj} + U^{(2)}_{jj}$$

from which (5.40) follows.

QED

In a similar fashion higher moments of the first passage times can be expressed in terms of the fundamental matrix. The expressions become very clumsy and are not presented here.

Note that (5.34) can be given a very simple interpretation via the usual renewal theory arguments.

If all the transition times become unity:

$$\pi_i \to 1 \qquad \pi_i^{(2)} \to 1 \qquad 1 \le i \le N$$

$$\langle \pi \rangle \to 1 \qquad \langle \pi^{(2)} \rangle \to 1 \qquad \pi_{ij} \to P_{ij} \quad 1 \le i,j \le N$$

then $U_{ij} \to N_{ij}$ and $U^{(2)}_{ij} \to N^{(2)}_{ij}$ where N_{ij} and $N^{(2)}_{ij}$ are the first two moments of the number of transitions until the system, starting in state i, lands in state j.

Theorems 5.5 and 5.6 now can be collected into

Theorem 5.7

$$N_{ij} = \frac{\delta_{ij} + Z_{jj} - Z_{ij}}{\pi_j} \qquad 1 \le i,j \le N \qquad (5.47)$$

$$N^{(2)}_{ij} = \frac{2 Z_{jj}}{(\pi_j)^2} (\delta_{ij} + Z_{jj} - Z_{ij}) \qquad (5.48)$$

$$- \frac{1}{\pi_j} [\delta_{ij} + 3 (Z_{jj} - Z_{ij}) + 2 (Z^2_{ij} - Z^2_{jj})]$$

$$1 \le i, j \le N$$

Proof

i) Since $Z \underline{1} = \underline{1}$, (5.32) goes into

$$N_{ij} = \frac{\delta_{ij} + Z_{jj} - Z_{ij}}{\pi_j}$$

which demonstrates (5.47).

ii) Equation (5.41) goes into

$$N^{(2)}_{jj} = \frac{1}{\pi_j} \left\{ 1 + 2 \sum_{i=1}^{N} \left[\sum_{K=1}^{N} \pi_i P_{iK} N_{Kj} - \pi_i P_{ij} N_{jj} \right] \right\}$$

$$= \frac{1}{\pi_j} \left\{ 1 + 2 \left[\sum_{K=1}^{N} \pi_K N_{Kj} - \pi_j N_{jj} \right] \right\}$$

$$= \frac{1}{\pi_j} \left\{ -1 + 2 \sum_{K=1}^{N} \pi_K \frac{(\delta_{Kj} + Z_{jj} - Z_{Kj})}{\pi_j} \right\}$$

Use $\pi Z = \pi$ $\qquad (\pi, 1) = 1$

$$= \frac{1}{\pi_j} \left\{ -1 + 2 \frac{(\pi_j + Z_{jj} - \pi_j)}{\pi_j} \right\}$$

$$N^{(2)}_{jj} = \frac{1}{\pi_j}\left\{-1 + \frac{2 Z_{jj}}{\pi_j}\right\} \tag{5.49}$$

Equation (5.40) becomes

$$N^{(2)}_{ij} = (\delta_{ij} - Z_{ij} + Z_{jj}) N^{(2)}_{jj} + 0 + 2\left[(ZPN)_{ij} - (ZPN)_{jj}\right] \tag{5.50}$$

$$- 2\left[(ZP)_{ij} - (ZP)_{jj}\right] N_{jj}$$

Since $ZP = Z - I + P^{\infty}$,

$$(ZP)_{ij} - (ZP)_{jj} = Z_{ij} - Z_{jj} - \delta_{ij} + 1 \tag{5.51}$$

$$(ZPN)_{ij} = \frac{1}{\pi_j}\left[(Z - I + P^{\infty})(I + Z_{jj} - Z)\right]_{ij}$$

$$= \frac{1}{\pi_j}\left[(Z - I + P^{\infty})_{ij} + Z_{jj} - (Z^2 - Z + P^{\infty})_{ij}\right]$$

$$= \frac{1}{\pi_j}\left[-\delta_{ij} + 2 Z_{ij} - Z^2{}_{ij} + Z_{jj}\right]$$

so that

$$(ZPN)_{ij} - (ZPN)_{jj} = \frac{1}{\pi_j}\left[1 - \delta_{ij} + z(z_{ij} - z_{jj}) - z_{ij}^2 + z^2_{jj}\right] \tag{5.52}$$

Insertion of (5.49), (5.51) and (5.52) into (5.50) yields (5.48).

QED

Formulas (5.47) and (5.48) were derived by Kemeny and Snell (23, Chapters 3, 4, 5 and 6) from the equations

$$N_{ij} = 1 + \sum_{\substack{k=1 \\ k \neq j}}^{N} P_{ik} N_{kj} \tag{5.53}$$

$$N^{(2)}_{ij} = 1 + \sum_{\substack{k=1 \\ k \neq j}}^{N} \left[P_{ik} N^{(2)}_{kj} + 2 P_{ik} N_{kj}\right] \tag{5.54}$$

An Application

An N-state ergodic Markov chain is under observation and the off-diagonal mean first passage times N_{ij} $i \neq j$ are known from measurements. What are N_{ii}, π_i, and p_{ij}? Assume $\pi_j > 0$, $1 \leq j \leq N$.

Solution: Let M be given,

$$M_{ij} = \begin{cases} N_{ij} & i \neq j \\ 0 & i = j \end{cases} \qquad 1 \leq i, j \leq N$$

It follows from (5.47) that

$$\sum_{j=1}^{N} M_{ij}\pi_j = \sum_{\substack{j=1 \\ j\neq i}}^{N} N_{ij}\pi_j = \sum_{j=1}^{N} (\delta_{ij} + z_{jj} - z_{ij}) - N_{ii}\pi_i$$

$$= 1 + \mathrm{tr}\, Z - 1 - 1$$

$$= \mathrm{tr}\, Z - 1 \qquad 1 \le i \le N$$

so that

$$\pi_j = \sum_{k=1}^{N} [M^{-1}]_{jk} (\mathrm{tr}\, Z - 1)$$

The (trZ-1) can be eliminated by recalling that π is normalized to be a probability vector. Therefore, π can be calculated via

$$\pi_j = \frac{\sum_{k=1}^{N} [M^{-1}]_{jk}}{\sum_{r,s=1}^{N} [M^{-1}]_{r,s}} \qquad 1 \le j \le N \tag{5.55}$$

and N_{jj} can be calculated via

$$N_{jj} = \frac{1}{\pi_j} \qquad 1 \le j \le N$$

The matrix D

$$D_{ij} = \delta_{ij} - N_{ij}\,\pi_j \qquad 1 \le i,j \le N$$

(note $D_{ii} = 0$) can now be computed and so can the vector f,

$$f_j = \sum_{i=1}^{N} \pi_i D_{ij} \qquad 1 \le j \le N$$

But (5.47) implies that

$$D_{ij} = Z_{ij} - Z_{jj}$$

and since $\pi Z = \pi$, $Z_{jj} = \pi_j - f_j$. Therefore, Z can be calculated via

$$Z_{ij} = D_{ij} + \pi_j - f_j \qquad 1 \le i,j \le N$$

and P can be calculated via

$$P_{ij} = -\left[Z^{-1}\right]_{ij} + \delta_{ij}\,\pi_j \qquad 1 \le i,j \le N$$

An Example

If $$N_{ij} = \begin{bmatrix} ? & 10/3 \\ 5 & ? \end{bmatrix}$$

then

$$M = \begin{bmatrix} 0 & 10/3 \\ 5 & 0 \end{bmatrix} \qquad M^{-1} = \begin{bmatrix} 0 & 1/5 \\ 3/10 & 0 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 1/5 & 3/10 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 2/5 & 3/5 \end{bmatrix} = \begin{bmatrix} \frac{1}{N_{11}} & \frac{1}{N_{22}} \end{bmatrix}$$

$$D = \begin{bmatrix} 0 & -2 \\ -2 & 0 \end{bmatrix} \qquad f = \begin{bmatrix} -\frac{6}{5} & -\frac{4}{5} \end{bmatrix}$$

$$Z = \frac{1}{5}\begin{bmatrix} 8 & -3 \\ -2 & 7 \end{bmatrix} \qquad Z^{-1} = \frac{1}{10}\begin{bmatrix} 7 & 3 \\ 2 & 8 \end{bmatrix}$$

$$P = -Z^{-1} + I + P^{\infty} = \begin{bmatrix} .7 & .3 \\ .2 & .8 \end{bmatrix}$$

An alternate method of finding the P_{ij}'s consists of holding i fixed and considering

$$\sum_{K=1}^{N} P_{iK} N_{Kj} = N_{ij} - 1 \qquad 1 \le j \le N \qquad j \neq i$$

$$\sum_{K=1}^{N} P_{iK} = 1$$

as a set of N simultaneous linear equations for the N unknowns P_{ij} with $1 \le j \le N$. This method requires inversion of N distinct NxN matrices, one matrix (say R(i)) for each i. The first method, which exploits the fundamental matrix, requires inversion of 2 NxN matrices and involves much less computation if N is large.

The two methods are actually closely related, for R(i) differs from M only by having the i^{th} column of M replaced by a column of 1's.

The above technique assumed that M^{-1} exists. This is indeed so and is proved in the following

lemma

Let P be ergodic.

$$\pi_j > 0 \qquad 1 \le j \le N$$

$$M_{ij} = \frac{-Z_{ij} + Z_{jj}}{\pi_j} \qquad 1 \le i,j \le N$$

Then M^{-1} exists.

Proof

Write M = AB where

$$A_{ij} = Z_{ij} - Z_{jj} \qquad B_{ij} = -\delta_{ij}/\pi_j \qquad 1 \le i,j \le N$$

B^{-1} exists, so that M^{-1} exists if and only if A^{-1} exists. If A^{-1} fails to exist, there is a vector $\underline{e}$ such that $\underline{e}A = \underline{0}$. This may be rewritten as

$$\underline{e}Z = (\underline{e},1)\underline{h}$$

where $h_i = Z_{ii}$. Since $Z\underline{1} = \underline{1}$, the above becomes

$$(\underline{e},\underline{1}) = (\underline{e},Z\underline{1}) = (\underline{e},\underline{1})(\underline{h},\underline{1}) = (\underline{e},\underline{1}) \quad (\mathrm{tr}Z)$$

so that either $(\underline{e},\underline{1}) = 0$, in which case $\underline{e}Z = \underline{0}$, $\underline{e} = \underline{0}Z^{-1} = \underline{0}$ and A^{-1} exists, or trZ = 1. In the latter case,

$$\sum_{\substack{j=1 \\ j \ne i}}^{N} N_{ij}\pi_j = 0 \qquad 1 \le i \le N$$

This is impossible since $N_{ij} \geq 1$ and $\pi_j > 0$ for all i and j. Thus, A^{-1} exists.

QED

5G Transient Processes

If j is a transient state of an ergodic Markov chain P, then $\pi_j = 0$ and

$$Z_{ij} = \delta_{ij} + \sum_{n=1}^{\infty} (P^n)_{ij} \qquad \pi_j = 0 \tag{5.56}$$

may be interpretated as the expected number of occurrences of state j, for an infinite duration process starting in state i.

This interpretation shows that the fundamental matrix contains information about the transient behavior of the process. Since Z contains both transient and recurrent (first passage) information, it provides an amazingly compact description of the intrinsic structure of the Markov chain.

We turn next to showing how the fundamental matrix is useful for describing semi-Markov chains with a reward structure. That this is so is not too surprising since the values $\underline{V}$ and steady state $\underline{\pi}$'s act as dual or adjoint variables to each other (pointed out by Smith (3)), the former appearing in expressions of the form $P\underline{v}$, the latter in expressions of the form $\underline{\pi}P$.

5H. Semi-Markov Chains with Rewards

The gain and relative values v defined in Chapter 2 by the equation

$$\underline{v} = \underline{q} - g\underline{\pi} + P\underline{v} \tag{5.57}$$

can be obtained by appeal to (5.19 - 5.20), since the form

$$(I-P)\underline{v} = \underline{q} - g\underline{\pi} \tag{5.58}$$

shows the appearance of the kernel (I - P). If P has a unique irreducible set of states, (5.14) states that

$$(\pi,\ q - g\pi) \tag{5.59}$$

is a necessary and sufficient condition for (5.58) to be solvable. The gain rate must, therefore, be

$$g = \frac{(\pi, q)}{(\pi, \pi)} \tag{5.60}$$

Equation (5.20) becomes [1]

$$\underline{v} = Z(\underline{q} - g\underline{\pi}) + c\,\underline{1} \qquad c = (\pi, v) \tag{5.61}$$

where the unknown scalar c is the arbitrary constant which may be added to the relative values. The most convenient choice is c = 0, corresponding to

$$v = Z(q - g\pi) \qquad (\pi, v) = 0 \tag{5.62}$$

We shall call this choice the natural convention for relative values.

The convention $v_N = 0$ is more convenient for numerical computation, and is equivalent to setting

$$c = -\sum_{j=1}^{N} Z_{Nj}(q_j - g\pi_j) \tag{5.63}$$

Regardless of convention chosen, equation (5.61) shows that the fundamental matrix enters the description of relative values. Jewell has pointed out [17] that if P is ergodic,

$$\begin{aligned} Z(q - g\pi) &= q - g\pi + \sum_{n=1}^{\infty} (P^n - P^\infty)(q - g\pi) \\ &= \sum_{n=0}^{\infty} P^n (q - g\pi) \end{aligned} \tag{5.64}$$

may be interpreted as a cumulative sum of the discrepancy $q - g\pi$ between the actual reward q per transition and the expected reward per transition gT, averaged with respect to the deviation $P^n - P^\infty$ of the process at the n^{th} step from its statistical equilibrium. It is this averaging over deviations which causes the Z matrix to enter, and helps explain why $v_i - v_j$, the difference in deviations, measures the difference in total earnings between states i and j.

It is of interest to know when the relative values of all states are equal. This question is resolved by

Theorem 5.8

$\underline{V}$ is proportional to $\underline{1}$ if and only if $\underline{q}$ is proportional to $\underline{T}$.

Proof

If $\underline{V} = a\underline{1}$, then (5.61) implies c = a,

$$0 = Z(q - gT)$$

$$q - gT = Z^{-1}\underline{0} = \underline{0}$$

so that q is proportional to T.

Conversely, if q is proportional to T, (5.60) shows that the proportionality constant is g. Then

$$q = gT \qquad q - gT = 0$$

$$V = Z(q - gT) + c\underline{1} = c\underline{1}$$

QED

In the Markov case, $\underline{T} = \underline{1}$ so that the relative values are constant if and only if the immediate expected rewards are constant.

In the continuum case, the theorem may be phrased as v(x) = a for almost all X if and only if q(x) = gT(x) for almost all X.

Scaling

If the values are given by the natural convention

$$v = Z(q - gT)$$

then under the transformation

$$q \to a q + b \pi$$

$$\pi \to \pi$$

$$P \to P$$

the gain, values, stationary distribution and fundamental matrix change according to

$$\pi \to \pi \qquad Z \to Z$$

$$g \to \frac{(\pi, a q + b \pi)}{(\pi, \pi)} = a g + b$$

$$v \to Z (a q + b \pi - (a g + b) \pi) = a v$$

Periodic and Multichain Processes

The assumption of ergodicity is, by (5.61), sufficient to guarantee the existence of the relative values, i.e., the solvability of equation (5.57). However, it is not necessary since the value equations are solvable for the periodic case, such as

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

provided there is only one irreducible set of states. In such a case $\underline{v}(n) - ng\ \underline{1}$ possesses no limit as $n \to \infty$. Therefore, the relative

values can no longer be interpreted as the asymptote in (2.13 b,d). However, the relative values can be interpreted as the time average of $\underline{v}(n) - ng\,\underline{1}$, as (5.86) illustrates.

General conclusions about the solvability of

$$(I - P)v = \underline{q} - g\underline{1} \tag{5.65}$$

are contained in the following

Theorem 5.9

a) If P has only one irreducible set of states, (5.65) is always solvable: g is unique and v is unique up to one additive constant

$$g = \frac{(\pi, q)}{(\pi, \pi)} \tag{5.66}$$

$$v = Z(q - g\pi) + c\,\underline{1} \tag{5.67}$$

where π is the unique stationary distribution,

$$p^{\infty}{}_{ij} = \pi_j \qquad p^{\infty}(x,y) = \pi(y) \tag{5.68}$$

$$Z = [I - P + P^{\infty}]^{-1} \tag{5.69}$$

b) If P has several irreducible sets of states, (5.65) is solvable if and only if all of the irreducible sets of states have the same gain rate. That is, if ${}^k\pi$ and ${}^k g$ are the stationary distribution and gain rate for the $k^{\underline{th}}$ chain,

$$ {}^k g = \frac{({}^k\pi, q)}{({}^k\pi, \mathbf{1})} \qquad 1 \le k \le r \tag{5.70} $$

then (5.65) is solvable if and only if

$$ {}^k g = g \qquad 1 \le k \le r \tag{5.71} $$

If P^∞ is defined by (4.60),

$$ P^\infty(x,y) = \sum_{k=1}^{r} \phi^k(x)\, {}^k\pi(y) \tag{5.72} $$

and the fundamental matrix by

$$ Z = [I - P + P^\infty]^{-1} \tag{5.73} $$

then g is unique (5.71) and v is unique up to one constant in each irreducible set of states:

$$ v = Z(q - g\mathbf{1}) + \sum_{k=1}^{r} \phi^k c_k \tag{5.74} $$

where

$$ c_k = ({}^k\pi, v) \tag{5.75} $$

c) (5.65) is solvable with g unique and v unique up to <u>one</u> additive constant if and only if P has only one irreducible set of states.

<u>Proof</u>

a) See arguments by (5.59-5.61).

b) According to Theorem (5.3), especially (5.14), equation (5.65) is solvable if and only if

$$P^{\infty}(q - g^{T}) = 0 \tag{5.76}$$

Insertion of (5.72) leads to

$$\sum_{k=1}^{r} ({}^{k}g - g)({}^{k}\pi, T)\,{}^{k}\phi(x) = 0 \tag{5.77}$$

Since $({}^{k}\pi, T) > 0$ and the ϕ's are linearly independent, we conclude that (5.71) is necessary and sufficient for solvability.

The solution is given by (5.15) as

$$V = Z(q - g^{T}) + P^{\infty}v \tag{5.78}$$

which proves (5.74).

c) Follows from a) and b).

QED

5I. Absolute Values for a Fixed Markov Chain

Let P be stable and the n-step return v(n) defined by

$$\underline{v}(n+1) = \underline{q} + P\underline{v}(n) \qquad n \geq 0 \tag{5.79}$$

The gain rate for this process is

$$\underline{g} = P^{\infty}\underline{q} \tag{5.80}$$

The absolute values Va are defined by

$$\underline{va} = \lim_{n\to\infty} [\underline{v}(n) - n\underline{g}] \tag{5.81}$$

It follows from (5.79) that

$$\underline{v}(n+1) = \sum_{k=0}^{n} P^{k}\underline{q} + P^{n+1}\underline{v}(0)$$

so that

$$\underline{v}(n+1) - (n+1)\underline{g} = -\underline{g} + P^{n+1}\underline{v}(0) + \underline{q} + \sum_{k=1}^{n}(P^{k} - P^{\infty})\underline{q}$$

Letting $n\to\infty$, this becomes (1)

$$\underline{va} = -\underline{g} + P^{\infty}\underline{v}(0) + Z\underline{q} \tag{5.82}$$

In the ergodic case, (5.82) becomes

$$\underline{g} = (\pi, q)\,\underline{1} = g\,\underline{1} \qquad g = (\pi, q) \tag{5.83}$$

$$\underline{v_a} = Z(q - g\,\underline{1}) + (\pi, v(0))\,\underline{1} \tag{5.84}$$

We note that

$$(\underline{\pi}, \underline{v_a}) = (\pi, \underline{v}(0)) \tag{5.85}$$

so that if the average scrap value $(\pi, v(0)) = 0$, then the absolute values and the natural convention for relative values agree.

In general the absolute values will not exist if P is not stable. For example the periodic process

$$q = [1 \quad 3] \qquad v(0) = [0 \quad 0] \qquad P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

has

$$\pi = \left[\tfrac{1}{2} \quad \tfrac{1}{2}\right] \qquad g = (\pi, q) = 2$$

$$v_1(n) = 2n - \tfrac{1}{2} + \tfrac{1}{2}(-1)^n \qquad n \geq 0$$

$$v_2(n) = 2n + \tfrac{1}{2} - \tfrac{1}{2}(-1)^n \tag{5.86}$$

We note first that the linear divergence as 2n justifies the terminology gain rate of $g = 2$. Second $\underline{v}(n) - n\underline{g}$ possesses no limit as $n \to \infty$, so the absolute values are undefined. Third, appeal to Theorem 5.9 leads to the result

$$[rv_1 \quad rv_2] = [-1 \quad 0] + c \quad [1 \quad 1] \tag{5.87}$$

for the relative values. Comparison of (5.86) and (5.87) shows that the relative values differ only by a constant (here c = 1/2) from the time average of $\underline{v}(n) - n\underline{g}$. It is only in this last sense that the relative values can be interpreted when the policy is not stable.

5J. Absolute Values for a Fixed Semi-Markov Chain

We wish to find the limiting behavior as $t \to \infty$ of the quantities $V_i(t)$ and $V(x,t)$ for the N-state case and continuum case, respectively, defined by

$$V_i(t) = q_i(t) + \sum_{j=1}^{N} \int_0^t dt' \; p_{ij}(t') \, V_j(t-t') \qquad 1 \le i \le N \quad t \ge 0 \tag{5.88}$$

$$V(x,t) = q(x,t) + \int_\Omega dy \int_0^t dt' \; p(x,y,t') \, V(y,t-t') \qquad x \in \Omega \quad t \ge 0 \tag{5.89}$$

These may be condensed into vector notation as

$$\underline{V}(t) = \underline{q}(t) + \int_0^t dt' \; P(t') \, \underline{V}(t-t') \qquad t \ge 0 \tag{5.90}$$

and treated simultaneously. We assume that the transition matrix P

$$P = \int_0^\infty dt \; p(t)$$

of the imbedded Markov chain is ergodic, with steady-state distribution π.

We also assume that

$$\underline{q} = \lim_{t \to \infty} \underline{q}(t) \tag{5.91}$$

exists and is finite for all components of the vector, and furthermore that the area $\underline{\eta}$

$$\underline{\eta} = \int_0^\infty dt \left[\underline{q}(t) - \underline{q}\right] \tag{5.92}$$

under the curve $\underline{q}(t) - \underline{q}$ is finite for all components.

We will use the Laplace transform method to find the asymptotic behavior. The Laplace transform $\tilde{W}(s)$ of a time function w(t) is indicated by a tilde:

$$\tilde{w}(s) = \int_0^\infty dt\, w(t) e^{-st} = \mathcal{L}\{w(t)\}$$ re s sufficiently large.

Since $\tilde{P}(0) = P$ exists, we know that $\tilde{P}(s)$ exists for re $s \geq 0$. In addition, since $\int dy |P(xys)| < 1$ for re $s > 0$, the method of Theorem 4.4 shows that all eigenvalues of $\tilde{P}(s)$ are strictly less than unity in magnitude. With due caution for $P(xyt)$ having a lattice distribution of t's this statement holds true as well for almost s with re s = 0. It then follows ($\tilde{P}(s)$ is L_2 since P is L_2) that $[I - \tilde{P}(s)]^{-1}$ exists for all s with re $s > 0$, and almost all s with re s = 0.

According to (5.92), the Laplace transform of q(t) - q exists at s = 0, hence for re $s > 0$. Then the Laplace transform of

$$q(t) = q + q(t) - q \,,$$

$$\tilde{q}(s) = \frac{q}{s} + \mathcal{L}[q(t) - q] \tag{5.93}$$

exists for re s > 0. Indeed

$$\tilde{q}(s) = \frac{q}{s} + \eta + O(s) \tag{5.94}$$

Equation (5.90) can be transformed into

$$\tilde{v}(s) = \tilde{q}(s) + \tilde{P}(s)\tilde{v}(s) \qquad \text{re } s > 0 \tag{5.95}$$

$$\tilde{v}(s) = [I - \tilde{P}(s)]^{-1} \tilde{q}(s) \qquad \text{re } s > 0 \tag{5.96}$$

Since $I - \tilde{P}(0) = I - P$ has no inverse, equation (5.96) breaks down at s = 0. We conjecture that the Laurent series

$$[I - \tilde{P}(s)]^{-1} = \frac{A_{-K}}{s^K} + \frac{A_{-K+1}}{s^{K-1}} + \ldots + \frac{A_{-1}}{s} + A_0 + \ldots \tag{5.97}$$

$$K \geq 1 \qquad A_{-K} \not\equiv 0$$

converges (for almost all x,y) for s lying in some deleted neighborhood of the origin (in which all future manipulations will be carried out), and can then show that k = 1.

Insertion of (5.93) and (5.97) into (5.96) produces

$$\tilde{\underline{v}}(s) = \frac{A_{-1}\underline{q}}{s^2} + \frac{A_{-1}\underline{\eta} + A_0\underline{q}}{s} + O(s)$$

so that if the gain rate $\underline{g}$ is defined by

$$\underline{g} = A_{-1}\underline{q} \tag{5.98}$$

then the absolute values are given by

$$\underline{va} = \lim_{t\to\infty}\left[\underline{v}(t) - \underline{g}\,t\right] = \lim_{s\to 0} s\left[\tilde{\underline{v}}(s) - \frac{A_{-1}\underline{q}}{s^2}\right] \tag{5.99}$$

$$\underline{va} = A_{-1}\underline{\eta} + A_0\underline{q}$$

We now know that $\underline{v}(t)$ diverges linearly with t with slope $\underline{g}$, and will carry out the above program to find $\underline{g}$ and the absolute value $\underline{va}$.

Jewell has also worked out expressions for the gain and absolute values of an ergodic semi-Markov chain[1]. Why then are we going through this laborious transform analysis, for which the crucial assumption, the existence of (5.97), cannot be justified? First because Jewell's expressions for the absolute values involve the mean and mean square first passage times U_{ij}, $U_{jj}^{(2)}$ which do not exist for the continuum case. The Laplace transform approach allows unified treatment of the N-state and continuum state cases. Second, there is the pleasure of doing the calculation from a new viewpoint (thereby gaining a deeper appreciation of the structure of the chain) as well as obtaining an

independent check on Jewell's results. Third it gives us an opportunity to use Theorem 5.3 and 5.4 and demonstrate their power.

Fourth, the theorem can be proved rigorously in the N-state case, and presumably generalized to the continuum-state case. For suppose all holding times vanish exponentially fast:

$$P_{ij}(t) \le Ae^{-ct} \qquad c>0 \quad \text{all } i,j$$

Then $\tilde{P}_{ij}(s)$ exist for re $s \ge -c$, and will be analytic in s in the disc $|s| < c$. The remarks in the first part of the proof hold and the theorem is valid.

Finally, our expressions for the absolute values will involve the fundamental matrix (not the U_{ij}) and are both more convenient for numerical work and indicative of the important role that Z plays in determining the reward structure.

Some notation is needed before A_{-1} and A_o can be evaluated. First we define operators T(x,y) and $T^{(2)}(xy)$ (T_{ij} and $T^{(2)}_{ij}$ in the N-state case) by

$$\tilde{P}(s) = \int_0^\infty dt\, P(t)\left[I - s\Pi + \frac{s^2t^2}{2} + O(s^3)\right]$$

$$= P - s\Pi + s^2\frac{\Pi^{(2)}}{2} + O(s^3) \qquad \text{re } s \ge 0 \qquad (5.100)$$

$T = T^{(1)}$ and $T^{(2)}$ are identical in meaning to the T and $T^{(2)}$ introduced in 5F: $T^{(n)}(xy)$ is P(xy) times the n^{th} moment of the time for a transition from x to y.

We define vectors

$$\underline{\tau}(x) = \int_{\Omega} dy\, \tau(x,y) \qquad \underline{\tau} = \tau \underline{1} \tag{5.101}$$

= mean holding time at x

$$\underline{\tau}^{(2)}(x) = \int_{\Omega} dy\, \tau^{(2)}(x,y) \qquad \underline{\tau}^{(2)} = \tau \underline{1} \tag{5.102}$$

= mean square holding time at x.

Since P is ergodic, it will be convenient to use the notation

$$\langle M \rangle = (\underline{\pi}, M\underline{1}) \tag{5.103}$$

$$\langle \underline{f} \rangle = (\underline{\pi}, \underline{f}) \tag{5.104}$$

where M is any square matrix (or kernel) and f is any vector. For example, in the N-state case (5.103) becomes

$$\langle M \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_i M_{ij}$$

This definition of "averaging" is useful because it possesses

the properties

$$P^{\infty} M P^{\infty} = \langle M \rangle P^{\infty} \qquad \text{any M} \qquad (5.104)$$

$$P^{\infty} \underline{f} = \langle f \rangle \underline{1} \qquad \text{any f} \qquad (5.105)$$

Because of (5.101-5.102), there is no ambiguity whether $\langle T \rangle$ refers to the average of the matrix T or vector $\underline{T}$. Note also that $\langle T \rangle$ and $\langle T^{(2)} \rangle$ are mean and mean-square holding times.

The derivation of A_{-1}, A_0, $\underline{g}$ and $\underline{va}$ which now follows is frankly heuristic, since the convergence of (5.97) is assumed. We will see that a sufficient condition for convergence in the N-state case is that all of the $\tilde{P}_{ij}(s)$ are analytic in a small vicinity of the origin. Unfortunately this is not necessarily true. The analyticity of the $\tilde{P}_{ij}(s)$ is known only for re $s > 0$ unless the $P_{ij}(t)$ all decay at least exponentially fast for large t.

Theorem 5.10

Let P be ergodic and all η's be finite. Then

$$[I - \tilde{P}(s)]^{-1} = \frac{P^{\infty}}{\langle \pi \rangle s} + A_0 + O(s) \qquad (5.106)$$

where

$$A_0 = Z - \frac{Z\Pi P^\infty}{\langle \pi \rangle} - \frac{P^\infty \Pi Z}{\langle \pi \rangle} + [\langle A_0 \rangle + 1] P^\infty \tag{5.107}$$

$$\langle A_0 \rangle = \frac{\frac{1}{2} \langle \pi^{(2)} \rangle + \langle \pi Z \pi \rangle}{[\langle \pi \rangle]^2} - 1 \tag{5.108}$$

The gain and absolute values are given by

$$\underline{g} = g \underline{1} \tag{5.109}$$

where

$$g = \frac{(\pi, q)}{(\pi, \pi)} = \frac{\langle q \rangle}{\langle \pi \rangle} \tag{5.110}$$

$$\underline{va} = Z(q - g\pi) + (\pi, va) \underline{1} \tag{5.111}$$

where

$$(\pi, va) = \frac{\langle \eta \rangle - \langle \pi' Z q \rangle}{\langle \pi \rangle} + [\langle A_0 \rangle + 1] \langle q \rangle \tag{5.112}$$

Equation (5.109) shows that in the ergodic case the gain rate is the same for all states.

We note that the relative values Z(q - gT) in $\underline{va}$ appear in a very natural way and that an explicit expression (5.112) is given to determine

the additive constant which converts relative values to absolute values.

Note also that the relative values require knowledge of only the mean holding time vector $\underline{T}$, while the absolute values require knowledge of the entire matrix T_{ij}. Similarly the relative values require only $q(\infty) = q$ while the absolute values require also the areas η under $q(t) - q$.

Equation (5.112) can be rewritten by use of

$$\langle \pi' Z \pi \rangle = \langle \pi Z \underline{\pi} \rangle \; ; \; \langle q \rangle = g^{\langle \pi \rangle}$$

as

$$(\pi, Va) = \frac{\langle \underline{\eta} \rangle + \frac{1}{2}\langle \underline{\pi}^{(2)} \rangle g}{\langle \pi \rangle} - \frac{\langle \pi Z (\underline{q} - g\underline{\pi}) \rangle}{\langle \pi \rangle}$$

From this equation and (5.111) we see that $\underline{q}$ never enters the expression for absolute values alone; it is always the combination $Z(q - gT)$ which appears.

Proof

i) Motivation for the Laurent series (5.97) in the N-state case comes from Cramer's rule, where

$$\left[I - \tilde{P}(s) \right]^{-1}_{ij} = \frac{\text{cofactor } \left[I - \tilde{P}(s) \right]_{ji}}{\det \left[I - \tilde{P}(s) \right]} \tag{5.113}$$

If each of the N^2 $\tilde{P}_{ij}(s)$ are analytic near the origin, then the numerator and denominator are finite sums of finite products of analytic functions and are also analytic near the origin. Since the zeros of a

(not-trivially zero) analytic function are of finite order, say ℓ (29), we can write $\det[I-\tilde{P}(s)] = s^{\ell} h(s)$ where h(s) is analytic near the origin and $h(0) \neq 0$, thereby deriving (5.97).

Presumably this argument can be extended to the continuum case by using the Fredholm expression instead of (5.113).

ii) Motivation for k = 1 comes from the Tauberian theorem which deduces from

$$\tilde{v}(s) = \frac{A_{-k}\, \ell}{s^{k+1}} + O\left(\frac{1}{s^{k}}\right)$$

that v(t) grows as t^k. We suspect the growth is linear, so that k = 1.

iii) More motivation for k = 1 comes from the N-state case, where we can show that $\ell = 1$. Let

$$f(s) = \det[I-\tilde{P}(s)] = \det[I-P+s\Pi+O(s^2)]$$
$$= s\, f'(0) + O(s^2)$$

since det (I - P) = 0. We show that $f'(0) \neq 0$ by expanding f'(0) as the sum of N determinants, where the i^{th} determinant corresponds to differentiating only the i^{th} row of I - P(s) and setting s = 0. Expanding the i^{th} determinant on its i^{th} row leads to

$$f'(0) = \sum_{i=1}^{N} \sum_{j=1}^{N} \Pi_{ij} \left[\text{cofactor of } (I-P)_{ij}\right] \qquad (5.114)$$

Let C_{ij} = cofactor of $(I - P)_{ji}$. By a general result of matrix theory

$$(I-P)\,C = C\,(I-P) = I \det (I-P) = 0$$

so that $C = PC = P^n C$ and $C = CP = CP^n$. Letting $n \to \infty$ and recalling that P is ergodic, these lead to

$$C = P^\infty C = CP^\infty = P^\infty C P^\infty = \langle C \rangle P^\infty \tag{5.115}$$

Then

$$\text{cofactor of } (I-P)_{ij} = C_{ji} = \langle C \rangle \pi_i \tag{5.116}$$

The independence of this cofactor of j has previously been noted in the literature [30]. We note also that $\langle C \rangle \neq 0$. If $\langle C \rangle$ were 0, then all cofactors of I - P would vanish and det M, given by (9.3-9.6), would vanish, violating our assumption of the existence of a unique π vector.

Insertion of (5.116) into (5.114) yields

$$f'(0) = \langle C \rangle \langle \pi \rangle \neq 0$$

so that the system determinant has a first order zero at s = 0. Presumably this argument can be extended to discussion of the order to the zero of the Fredholm denominator at s = 0 in the continuum state case.

iv) A proof of k = 1 valid for the continuum case is obtained by inserting (5.97) and (5.100) into

$$\left[I - \tilde{P}(s)\right]\left[I - \tilde{P}(s)\right]^{-1} = I = \left[I - \tilde{P}(s)\right]^{-1}\left[I - \tilde{P}(s)\right] \tag{5.117}$$

and equating coefficients of $1/s^k$ and $1/s^{k-1}$. One obtains

$$(I-P)\,A_{-k} = 0 = A_{-k}(I-P) \tag{5.118}$$

$$(I-P)\,A_{-k+1} + \Pi A_{-k} = I\,\delta_{k1} = A_{-k+1}(I-P) + A_{-k}\Pi \tag{5.119}$$

Equation (5.118) implies that

$$A_{-k} = P^{\infty} A_{-k} = A_{-k} P^{\infty} = P^{\infty} A_{-k} P^{\infty} = \langle A_{-k} \rangle P^{\infty} \tag{5.120}$$

Equation (5.119), after premultiplication by P^{∞}, becomes

$$\langle \Pi \rangle \langle A_{-k} \rangle P^{\infty} = P^{\infty} \delta_{k1} \tag{5.121}$$

Now $\langle A_{-k} \rangle \neq 0$ else (5.120) imply that the leading term A_{-k} in (5.97) vanish. Also $\langle T \rangle$ is the mean holding time and is positive. Equation (5.121) therefore implies that k = 1 and that $\langle A_{-k} \rangle = 1/\langle \Pi \rangle$. Therefore

$$A_{-1} = \frac{P^{\infty}}{\langle \Pi \rangle} \tag{5.122}$$

v) Evaluation of A_0 may be accomplished by rewriting (5.119) as

$$(I-P)\,A_0 = I - \frac{\Pi P^\infty}{\langle \Pi \rangle} \tag{5.123}$$

$$A_0(I-P) = I - \frac{P^\infty \Pi}{\langle \Pi \rangle} \tag{5.124}$$

It is easy to check that (5.25 a,b,c) hold. Theorem 5.4 may be invoked to yield the expression

$$A_0 = Z - \frac{Z\Pi P^\infty}{\langle \Pi \rangle} + P^\infty - \frac{P^\infty \Pi Z}{\langle \Pi \rangle} + \langle A_0 \rangle P^\infty$$

which confirms (5.107).

vi) Evaluation of $\langle A_0 \rangle$ requires the coefficient of s in (5.117):

$$(I-P)\,A_1 + \Pi A_0 - \frac{\Pi^{(2)} A_{-1}}{2} = 0$$

$$P^\infty \Pi A_0 P^\infty = P^\infty \frac{\Pi^{(2)} A_{-1} P^\infty}{2} = \frac{\langle \Pi^{(2)} \rangle P^\infty}{2\langle \Pi \rangle} \tag{5.125}$$

According to (5.107),

$$P^\infty \Pi A_0 P^\infty = \langle \Pi \rangle P^\infty - \frac{P^\infty \langle \Pi Z \Pi \rangle}{\langle \Pi \rangle} - P^\infty \langle \Pi \rangle + [\langle A_0 \rangle + 1]\langle \Pi \rangle P^\infty$$

Comparison with (5.125) and cancellation of P^∞ yields (5.108).

Insertion of (5.122) into (5.98) leads to (5.109), while insertion

of (5.122) and (5.107) into (5.99) leads to

$$\nabla \underline{a} = \frac{(\pi, \eta)}{\langle \pi \rangle} \underline{1} + Z \underline{q} - \frac{Z \pi \langle q \rangle \underline{1}}{\langle \pi \rangle} - \frac{P^{\infty} \pi Z q}{\langle \pi \rangle}$$

$$+ [\langle A_0 \rangle + 1] \langle q \rangle \underline{1}$$

$$= Z (\underline{q} - g \underline{\pi}) + c \underline{1} \tag{5.126}$$

where

$$c = \frac{\langle \eta \rangle}{\langle \pi \rangle} - \frac{\langle \pi Z q \rangle}{\langle \pi \rangle} + [\langle A_0 \rangle + 1] \langle q \rangle \tag{5.127}$$

Dotting (5.126) with π leads to

$$(\pi, \nabla a) = c \tag{5.128}$$

Equations (5.126-5.128) complete the proof of the theorem.

QED

5K. Comparison with Jewell's Result

Jewell has derived an alternate expression[1] for the absolute values in the N-state case:

$$va_i = q_i + \sum_{j=1}^{N} \left\{ q_j \left[\frac{u_{jj}^{(2)}}{2(u_{jj})^2} - \frac{u_{ij}}{u_{jj}} \right] + \frac{a_j}{u_{jj}} \right\} \qquad (5.129)$$

where $a_i = -\eta_i$ and the u's are the first passage moments discussion in 5F. This expression assumes all states are recurrent since it is undefined if any $u_{jj} = 0$.

If (5.32) and (5.41) are inserted into Jewell's expression, it reduces to (5.112). Some very heavy algebra is required; the proof is omitted.

5L. Reduction of the Semi-Markov Case to the Markov Case

For a Markov chain we insert

$$p_{ij}(t) = p_{ij}\,\delta(t-1) \tag{5.130}$$

into the above to obtain

$$\pi = P \qquad \underline{\pi} = \underline{1} \qquad \langle \pi \rangle = 1$$

$$\pi^{(2)} = P \qquad \langle \pi^{(2)} \rangle = 1$$

$$\langle A_0 \rangle = \tfrac{1}{2} + \langle PZP \rangle - 1 = -\tfrac{1}{2} + \langle P^{\infty} \rangle = \tfrac{1}{2} \tag{5.131}$$

$$(\pi, v_a) = (\pi, \eta) - (\pi, PZq) + \tfrac{3}{2}(\pi, q)$$

$$= (\pi, \eta) + \tfrac{1}{2}(\pi, q)$$

We assume that the process earns at a rate r_{ij} if in state i headed towards state j, so that (2.7) becomes

$$q_i(t) = \begin{cases} \sum_{j=1}^{N} p_{ij}[r_{ij}t] + v_i(0) & t < 1 \\ \sum_{j=1}^{N} p_{ij}\, r_{ij} & t \geq 1 \end{cases}$$

$$1 \leq i \leq N$$

Then

$$q_i = \lim_{t\to\infty} q_i(t) = \sum_{j=1}^{N} p_{ij} r_{ij}$$

$$\eta_i = \int_0^\infty dt \left[q_i(t) - q_i\right]$$

$$= \int_0^1 dt \left[q_i(t-1) + v_i(0)\right] \qquad (5.132)$$

$$= v_i(0) - \tfrac{1}{2} q_i$$

Insertion of (5.132) into (5.131) yields

$$(\pi, v(a)) = (\pi, v(0))$$

which agrees with (5.85). Hence the semi-Markov model has gain and absolute values which reduce to those for the Markov model when (5.130) holds.

CHAPTER 6

PERTURBATION THEORY AND MARKOV CHAINS

6A. Introduction

Let A and B denote Markov chains each with one irreducible set of states, with transition matrices P_A and P_B. If P_B is "close" to P_A, we would expect that the limiting distribution π_B and fundamental matrix Z_B for B are close to those, π_A and Z_A, of A. In this chapter we show that this is so. Perturbation expressions are derived showing the rate of change of π and Z with respect to the matrix P.

The only assumption needed below is that P_A and P_B are L_2 and have but one sub-chain. The proofs hold for the N-state case and also for the continuum case. The recurrent chains of A and B need not overlap.

The L_2 norm is used in this chapter.

6B Finite Changes in Policy

It turns out to be convenient to measure the distance between P_A and P_B by the matrix

$$U = U_{AB} = (P_B - P_A) Z_A \tag{6.1}$$

The following three theorems show how π_B and Z_B can be expressed in terms of π_A, Z_B and U.

Theorem 6.1

$(I-U)^{-1}$ exists.

Proof

Since the rows of $Z_B^{-1} = I - P_B + P_B^{\infty}$ sum to unity, it follows that $Z_B^{-1} P_B^{\infty} = P_B^{\infty}$ and $Z_B^{-1} P_A^{\infty} = P_A^{\infty}$. Then

$$\begin{aligned} I - U &= [I - P_A + P_A^{\infty} - (P_B - P_A)] Z_A \\ &= [I - P_B + P_B^{\infty} + (P_A^{\infty} - P_B^{\infty})] Z_A \\ &= [Z_B^{-1} + P_A^{\infty} - P_B^{\infty}] Z_A \\ &= Z_B^{-1} [I + P_A^{\infty} - P_B^{\infty}] Z_A \end{aligned}$$

Since $(I + D)^{-1} = I - D$ if $D^2 = 0$, the last equation implies that

$$(I-U)^{-1} = Z_A^{-1}\left[I - P_A^{\infty} + P_B^{\infty}\right] Z_B \qquad (6.2)$$

QED

Theorem 6.2

$$\pi_B = \pi_A (I - U)^{-1} \qquad (6.3)$$

Proof

Since $\pi_A Z_A^{-1} = \pi_A$ and $\pi_A(I - P_A^{\infty}) = 0$, equation (6.2) implies that

$$\pi_A(I-U)^{-1} = \pi_A P_B^{\infty} Z_B = \pi_B Z_B = \pi_B$$

QED

Corollary

$$P_B^{\infty} = P_A^{\infty} (I-U)^{-1}$$

Theorem 6.3

$$Z_B = Z_A(I - U)^{-1} - P_A^{\infty}(I - U)^{-1} U Z_A (I - U)^{-1} \qquad (6.4)$$

Proof

Premultiply (6.2) by $\left[I + P_A^{\infty} - P_B^{\infty}\right] Z_A$ to obtain

$$Z_B = \left[I + P_A^{\infty} - P_B^{\infty}\right] Z_A (I-U)^{-1}$$

Since $P_B^{\infty} = P_A^{\infty} + P_A^{\infty}(I - U)^{-1}U$, this last equation is equivalent to (6.4).

QED

6C. Perturbation Series

If $P_B - P_A$ is sufficiently small that all eigenvalues of U are strictly less than unity in magnitude -- for example, if

$$\| P_B - P_A \| < \frac{1}{\| Z_A \|}$$

is satisfied -- then $(I - U)^{-1}$ can be expanded in powers of U. Equations (6.3) and (6.4) become

$$\pi_B = \pi_A + \sum_{n=1}^{\infty} \pi_A U^n \tag{6.5}$$

$$Z_B = Z_A + \sum_{n=1}^{\infty} \left[Z_A U^n - \sum_{K=1}^{n} P_A^{\infty} U^K Z_A U^{n-K} \right] \tag{6.6}$$

These series may be thought of as the usual Rayleigh perturbation expansion in powers of the smallness parameter U, the n^{th} term in the series being of order U^n. The series show that $\pi_B \to \pi_A$ and $Z_B \to Z_A$ as $P_B \to P_A$. Bounds such as

$$\left\| \pi_B - \pi_A - \sum_{K=1}^{m} \pi_A U^K \right\| = \| \pi_A (I-U)^{-1} U^{m+1} \| \leq \| \pi_B \| \| U^{m+1} \| \tag{6.7}$$

can be used to estimate the order to which the perturbation series expansions need be carried.

An Example

Consider the ergodic Markov chain A with transition matrix

$$P_A = \begin{bmatrix} .9 & .1 \\ .3 & .7 \end{bmatrix}$$

Then solution of $\pi = \pi P \quad (\pi, 1) = 1$ yields

$$\pi_A = [.75 \quad .25]$$

$$P_A^{\infty} = \begin{bmatrix} .75 & .25 \\ .75 & .25 \end{bmatrix}$$

$$I - P_A + P_A^{\infty} = \begin{bmatrix} .85 & .15 \\ .45 & .55 \end{bmatrix}$$

$$Z_A = [I - P_A + P_A^{\infty}]^{-1} = \begin{bmatrix} 1.375 & -.375 \\ -1.125 & 2.125 \end{bmatrix}$$

We note that the rows of Z_A sum to 1, that $\pi_A Z_A = \pi_A$ and that Z can have negative matrix elements.

Suppose there is a new ergodic Markov chain B with transition

matrix

$$P_B = \begin{bmatrix} .85 & .15 \\ .34 & .66 \end{bmatrix}$$

so that P_B does not differ greatly from P_A:

$$P_B - P_A = \begin{bmatrix} -.05 & .05 \\ .04 & -.04 \end{bmatrix}$$

$$U = (P_B - P_A) Z_A = \begin{bmatrix} -.125 & .125 \\ .100 & -.100 \end{bmatrix}$$

The two eigenvalues λ_1, λ_2 of U are $\lambda_1 = 0$ (since the rows of U sum to zero) and $\lambda_2 = \text{tr } U - \lambda_1 = -.225$. Since these are both less than 1 in magnitude, equation (6.5) provides a convergent series for calculating π_B. The k^{th} partial sum for $(\pi_B)_1$,

$$(\pi_B)_1^{(K)} = \sum_{n=0}^{K} \pi_A U^K$$

is tabulated, along with the deviation from the exact value $(\pi_B)_1 = .6939$ obtained by solving $\pi_B(I-P_B) = 0 \quad (\pi_B, 1) = 1$.

k	$(\pi_B)_1^{(k)}$	error
0	.7500	.0561
1	.6812	-.0127
2	.6967	.0028
3	.6932	-.0007
	↓	
	.6939	

The ratio of errors approaches the dominant eigenvalue, here $\lambda_2 = -.225$, of U. The series therefore converges quickly in this case.

The steady state distribution π_B could alternately be obtained via

$$I - U = \begin{bmatrix} 1.125 & -.125 \\ -.1 & 1.1 \end{bmatrix}$$

$$(I-U)^{-1} = \begin{bmatrix} .8979 & .1021 \\ .0816 & .9184 \end{bmatrix}$$

$$\pi_B = \pi_A (I-U)^{-1} = [.6939 \quad .3062]$$

The perturbation series avoids the matrix inversion needed here. Note that the rows of I - U and $(I - U)^{-1}$ sum to 1. These are general results. Since the rows of $\mathbf{Z}$ and P sum to unity, it follows from (6.1) that the rows of U sum to zero

$$U\underline{1} = \underline{0}$$

Then the rows of I-U sum to 1

$$(I-U)\,\underline{1} = \underline{1}$$

and operating on this with $(I - U)^{-1}$ shows that the rows of $(I - U)^{-1}$ sum to 1:

$$(I-U)^{-1}\,\underline{1} = \underline{1}$$

6D. Partial Derivatives

Let $P_A \to P$ and $P_B \to P + \delta P$, where δP is small and satisfied the three requirements

$$\delta P \geq -P$$

$$P\underline{1} = \underline{0}$$

$P + \delta P$ has only one irreducible set of states.

Then for the N-state case (6.5) and (6.6) become

$$\pi_i(P+\delta P) = \pi_i(P) + \sum_{mn=1}^{N} \frac{\partial \pi_i}{\partial p_{mn}} \delta p_{mn} + O((\delta P)^2) \tag{6.8}$$

$$Z_{ij}(P+\delta P) = Z_{ij}(P) + \sum_{mn=1}^{N} \frac{\partial Z_{ij}}{\partial p_{mn}} \delta p_{mn} + O((\delta P)^2)$$

$$1 \leq i,j \leq N \tag{6.9}$$

where

$$\frac{\partial \pi_i}{\partial p_{mn}} = \pi_m Z_{ni} \tag{6.10}$$

$$\frac{\partial Z_{ij}}{\partial P_{mn}} = Z_{im} Z_{nj} - \pi_m (Z^2)_{nj} \qquad (6.11)$$

$$1 \le i, j, m, n \le N$$

Equations (6.10) and (6.11) show how the π vector and fundamental matrix depend upon the elements in the P matrix. In particular, (6.10) shows that the fundamental matrix is closely linked to the ergodic behavior of the Markov chain.

Equations analogous to (6.10 - 6.11) but involving functional derivatives [32] instead of partial derivatives can be derived in the continuum case to show how and Z change when P(xy) is changed.

The higher order terms in (6.8-6.9) can be obtained from (6.5-6.6).

Equations (6.10 - 6.11) are not truly partial derivatives because of the constraint $\delta \underline{\pi} \underline{1} = \underline{0}$ on the allowed variation. In particular, taking the derivative of (6.10) will <u>not</u> give the second order correction in (6.8).

6E. An Application: Finding the Fundamental Matrix With Only One Matrix Inversion

The method presented in Section 5A for obtaining the Z matrix involves two matrix inversions and is therefore very clumsy if N is large. It would be far better to have a scheme for computing Z which involves only one matrix inversion. Such a scheme can be devised by use of (6.4).

If for P_A we pick the ergodic matrix $(P_{A_{ij}}) = (\delta_{j,N})$ then $P_A = P_A^{\infty}$ and $Z_A = I$. Equation (6.4) becomes

$$Z_{B\,ij} = (I-U)^{-1}_{ij} - \left[(I-U)^{-1} U (I-U)^{-1}\right]_{Nj} \qquad (6.12)$$

$$1 \le i,j \le N$$

where

$$U_{ij} = P_{B\,ij} - \delta_{j,N} \qquad (6.13)$$

Equations (6.12 - 6.13) allow the fundamental matrix Z_B to be computed for any one-chain Markov process P_B with only a single matrix inversion, that of I - U.

6F. A Special Case

Suppose A and B are N-state ergodic Markov chains which differ only in one state, say the k^{th} state. Then the matrix $(P_B - P_A)$ vanishes except for its k^{th} row, and the same holds true for $U = (P_B - P_A)Z_A$. One then finds that

$$U^2 = U_{KK} U \tag{6.14}$$

$$(I - U)^{-1} = I + \frac{U}{1 - U_{KK}} \tag{6.15}$$

Equation (6.3) becomes

$$\pi_{Bi} = \pi_{Ai} + \frac{\pi_{AK} U_{Ki}}{1 - U_{KK}} \qquad 1 \leq i \leq N \tag{6.16}$$

which exhibits a very simple structure.

Note here that the rows of U must sum to zero if π_B is to be a probability vector. This is a general property of U which follows from (6.1) and the fact that the rows of Z_A and P sum to unity.

6G. Group Properties

Since an ergodic chain C can be reached from an ergodic chain A either directly or via an intermediate ergodic chain B, U_{AB} and U_{BC} should be related to U_{AC}. Indeed

$$\pi_A(I-U_{AC})^{-1} = \pi_C = \pi_B(I-U_{BC})^{-1} = \pi_A(I-U_{AB})^{-1}(I-U_{BC})^{-1}$$

The suggested identity

$$(I-U_{AC})^{-1} = (I-U_{AB})^{-1}(I-U_{BC})^{-1} \tag{6.17}$$

$$I-U_{AC} = (I-U_{BC})(I-U_{AB})$$

can be established from (6.2):

$$\begin{aligned}
&(I-U_{AB})^{-1}(I-U_{BC})^{-1} \\
&= Z_A^{-1}[I-P_A^{\infty}+P_B^{\infty}]Z_B Z_B^{-1}[I-P_B^{\infty}+P_C^{\infty}]Z_C \\
&= Z_A^{-1}[I-P_A^{\infty}+P_C^{\infty}]Z_C \\
&= (I-U_{AC})^{-1}
\end{aligned}$$

By setting C = A and using $U_{AA} = 0$, equation (6.17) becomes

$$(I - U_{AB})^{-1} = I - U_{BA}$$

$$U_{BA} = - U_{AB} (I - U_{AB})^{-1} \qquad (6.18)$$

Equations (6.3), (6.17) and (6.18) show that the $(I-U)^{-1}$'s possess group properties and act as "generators" for transformations from one ergodic Markov chain to another. The study of the $(I-U)^{-1}$'s, their spectral properties and transformation is suggested as an area for future research.

6H. Randomized Markov Chains

Let A and B denote two stationary Markov chains, each with one irreducible set of states. Then the stationary Markov chain with transition matrix

$$P(\lambda) = (1-\lambda) P_A + \lambda P_B \qquad 0 < \lambda < 1 \tag{6.19}$$

may be thought of as a randomization of A and B such that in each state, the transition mechanism is chosen as P_A with probability $1-\lambda$ and as P_B with probability λ .

It turns out that $P(\lambda)$ will also have exactly one irreducible set of states, namely the union of the irreducible set of states for A and B. Hence $P(\lambda)$ has a unique stationary distribution $\pi(\lambda)$.

Theorem 6.4

Let P_A and P_B each have exactly one irreducible set of states, say R_A and R_B respectively. Then $P(\lambda)$ given by (6.19) also has exactly one irreducible set of states, namely $R_A \cup R_B$.

Proof:

i) $P(\lambda)$ has at least one irreducible set of states.

To prove it cannot have to or more, we assume the contrary and obtain a contradiction. If R_1 and R_2 are two disjoint irreducible sets of states for $P(\lambda)$, then for $x \in R_1$,

$$\int_{\Omega - R_1} dy \left[\lambda P_A(x,y) + (1-\lambda) P_B(x,y)\right] = 0 \qquad x \in R_1$$

$$0 = \int_{\Omega - R_1} dy \; P_A(x,y) \qquad x \in R_1$$

$$\int_{R_1} dy \; P_A(x,y) = 1 \qquad x \in R_1$$

whence R_1 is a closed set of states for P_A. Similarly R_2 is also a closed set of states for P_A. But this contradicts the irreducibility of P_A.

ii) Let R denote the one irreducible set of states for $P(\lambda)$. It is easy to show that $R_A \cup R_B$ is a closed set of states for $P(\lambda)$, whence

$$R \subseteq R_A \cup R_B$$

Either R agrees with $R_A \cup R_B$, and the theorem is proved, or it is a proper subset. In the latter case

$$R = S_A \cup S_B$$

where S_A and S_B are the subsets of R_A and R_B which contribute to R:

$$S_A = R \cap R_A \qquad S_B = R \cap R_B$$

With no loss of generality we may take S_A as a proper subset of R_A.

Since $R_A - S_A$ is not in R,

$$\int_{R_A - S_A} dy \; [\lambda P_A(x,y) + (1-\lambda) P_B(x,y)] = 0 \qquad x \in R$$

$$\int_{R_A - S_A} dy \; P_A(x,y) = 0 \qquad x \in R$$

In particular, pick $x \in S_A$ to get

$$\int_{S_A} dy \; P_A(x,y) = 1 \qquad x \in S_A$$

whence S_A is closed. This contradicts the irreducibility of R_A.

QED.

If we invoke Theorems (6.2) and (6.3), but with P_B replaced by $P(\lambda)$, we conclude that $P(\lambda)$ has a stationary distribution $\pi(\lambda)$ and fundamental matrix $Z(\lambda)$ given by

$$\pi(\lambda) = \pi_A \left[I - \lambda (P_B - P_A) Z_A \right]^{-1} = \pi_A \left[I - \lambda U_{AB} \right]^{-1} \tag{6.20}$$

$$Z(\lambda) = Z_A (I - \lambda U_{AB})^{-1} - \lambda P_A^{\infty} (I - \lambda U_{AB})^{-1} U_{AB} Z_A (I - \lambda U_{AB})^{-1} \tag{6.21}$$

These are analytic functions of λ for $0 \leq \lambda \leq 1$ since $(I - \lambda U)^{-1}$ must exist (and therefore is analytic) for λ in that range.

Perturbation Series

If $|\lambda|$ is sufficiently small -- for example, if $|\lambda| < 1/\|U_{AB}\|$ -- that all eigenvalues of U_{AB} are strictly less than $\frac{1}{|\lambda|}$ in magnitude, then $(I-\lambda U_{AB})^{-1}$ may be expanded in powers of λ. Equations (6.20 - 6.21) become

$$\pi(\lambda) = \pi_A + \sum_{n=1}^{\infty} \lambda^n \pi_A U^n \tag{6.22}$$

$$Z(\lambda) = Z_A + \sum_{n=1}^{\infty} \lambda^n \left[Z_A U^n - \sum_{K=1}^{n} P_A^{\infty} U^K Z_A U^{n-K} \right] \tag{6.23}$$

In particular we note that an expansion in powers of λ is precisely an expansion in powers of U, i.e. in powers of $P_B - P_A$.

The Rayleigh Method

An alternate derivation of (6.22 - 6.23) by the Rayleigh method of equating coefficients of λ^n will prove instructive.

To derive (6.22), we assume convergence near $\lambda = 0$ of the series

$$\pi(\lambda) = \pi_0 + \sum_{n=1}^{\infty} \lambda^n \pi_n \tag{6.24}$$

The normalization $(\pi(\lambda),1) = 1$ (all λ) leads to the constraints $(\pi_n,1) = \delta_{no}$. If $\pi(\lambda)$ is the unique left eigenvector of $P(\lambda)$, then

$$\pi(\lambda)\left[P_A + \lambda(P_B - P_A)\right] = \pi(\lambda)\, P(\lambda) = \pi(\lambda)$$

Equating the coefficient of λ^n on both sides yields

$$\pi_0 P_A = \pi_0$$

$$\pi_n P_A + \pi_{n-1}(P_B - P_A) = \pi_n \qquad n \geq 1$$

The first, after appeal to Theorem 5.4, becomes

$$\pi_0(I - P_A) = 0 \qquad \pi_0 = \pi_0 P_A^{\infty} = (\pi_0, 1)\,\pi_A = \pi_A$$

The others, after appeal to (5.21), become

$$\pi_n(I - P_A) = \pi_{n-1}(P_B - P_A)$$

$$\pi_n = \pi_{n-1}(P_B - P_A) Z_A + \pi_n P_A^{\infty}$$

$$= \pi_{n-1} U_{AB} + (\pi_n, 1)\,\pi_A$$

$$= \pi_{n-1} U_{AB} \qquad n \geq 1$$

It follows from

$$\pi_0 = \pi_A \qquad \pi_n = \pi_{n-1} U \qquad n \geq 1 \tag{6.25}$$

that

$$\pi_n = \pi_A U^n \qquad n \geq 0 \tag{6.26}$$

which completes the derivation of (6.22).

Similarly $P^\infty(\lambda) = \sum_{n=0}^{\infty} \lambda^n P_n^\infty$ where

$$P_n^\infty = P_A^\infty U^n \quad .$$

To derive (6.23), we assume convergence near $\lambda = 0$ of the series

$$Z(\lambda) = \sum_{n=0}^{\infty} \lambda^n Z_n \tag{6.27}$$

The condition $Z(\lambda)\,\underline{1} = \underline{1}$ leads to $Z_n\,\underline{1} = \delta_{n0}\,\underline{1}$.

If $Z(\lambda)$ is indeed the inverse to $[I - P(\lambda) + P^\infty(\lambda)]$, then

$$Z(\lambda)[I - P(\lambda)] = I - Z(\lambda) P^\infty(\lambda) = I - P^\infty(\lambda)$$

Equating coefficients of λ^n yields

$$Z_0(I - P_A) = I - P_A^\infty$$

$$Z_n(I - P_A) - Z_{n-1}(P_B - P_A) = -P_A^\infty U^n \qquad n \geq 1$$

The first implies, by Theorem 5.4, that

$$Z_0 = (I - P_A^\infty) Z_A + Z_0 P_A^\infty = Z_A - P_A^\infty + Z_0 P_A^\infty$$

The second implies that

$$Z_n = Z_{n-1}(P_B - P_A) Z_A - P_A^\infty U^n Z_A + Z_n P_A^\infty \qquad n \geq 1$$

The condition $Z_n \underline{1} = \underline{1} \delta_{n0}$ leads to $Z_n P_A^\infty = P_A^\infty \delta_{n0}$ so that

$$Z_0 = Z_A \tag{6.28}$$

$$Z_n = Z_{n-1} U_{AB} - P_A^\infty U^n Z_A \qquad n \geq 1 \tag{6.29}$$

from which (6.23) follows.

The disadvantage of the Rayleigh technique, as opposed to the method employed in 6B, is that the convergence of the power series in λ must be investigated.

If the power series in (6.24) and (6.27) are used for actual numerical work, if λ is small, or if the series obtained by setting $\lambda = 1$:

$$\pi_B = \pi_A + \pi_1 + \pi_2 + \pi_3 + \cdots \tag{6.30}$$

$$P_B^{\infty} = P_A^{\infty} + P_1^{\infty} + P_2^{\infty} + \cdots \tag{6.31}$$

$$Z_B = Z_A + Z_1 + Z_2 + \cdots \tag{6.32}$$

are used if $\|P_B - P_A\|$ is small, then it is convenient to calculate the coefficients recursively by

$$\pi_{n+1} = \pi_n U \tag{6.33}$$

$$P_{n+1}^{\infty} = P_n^{\infty} U \tag{6.34}$$

$$Z_{n+1} = Z_n U - P_n Z_A \, . \tag{6.35}$$

6I. Symmetry Properties

All formulas given above should remain unchanged under the transformation A $\rightarrow$ B, B $\rightarrow$ A, $\lambda \rightarrow 1-\lambda$.

Equation (6.3) goes into

$$\pi_A = \pi_B (I - U_{BA})^{-1}$$

and is essentially unchanged, since (6.18) holds.

Equation (6.18) goes into

$$(I - U_{BA})^{-1} = I - U_{AB}$$

and is essentially unchanged.

Equation (6.20) goes into

$$\pi_A [I - \lambda U_{AB}]^{-1} = \pi_B [I - (1-\lambda) U_{BA}]^{-1} \qquad (6.36)$$

and remains true because the right side is given by

$$\pi_B [I + (1-\lambda) U_{AB} (I - U_{AB})^{-1}]^{-1}$$

$$= \pi_B [(I - U_{AB})^{-1} - \lambda U_{AB} (I - U_{AB})^{-1}]^{-1}$$

$$= \pi_B [(I - \lambda U_{AB}) (I - U_{AB})^{-1}]^{-1}$$

$$= \pi_B (I - U_{AB}) (I - \lambda U_{AB})^{-1}$$

$$= \pi_A (I - \lambda U_{AB})^{-1}$$

The identity

$$\left[I - (1-\lambda) U_{BA}\right]^{-1} = (I - U_{AB})(I - \lambda U_{AB})^{-1} \tag{6.37}$$

proved immediately above expresses a relation among the U's which would otherwise have gone unnoticed. In general, the group and symmetry properties of the U's lead to many intriguing identities.

6J. Another Look at the U Matrix

The property $U\underline{1} = \underline{0}$ leads to $UP_A^{\infty} = 0$ so that

$$I - P_B = I - P_A - U_{AB} Z_A^{-1} = I - P_A - U_{AB}(I - P_A + P_A^{\infty})$$

or

$$I - P_B = (I - U_{AB})(I - P_A) \tag{6.38}$$

Equation (6.38) is another example of kernel factorization, in which a singular kernel has been broken into the product of an invertible and a singular kernel. Replacement of P_B by $P(\lambda)$ leads to

$$I - P(\lambda) = (I - \lambda U_{AB})(I - P_A) \tag{6.39}$$

Equations (6.38) and (6.39) show that (I - U) serves as a "correction" to carry one (I - P) into another.

In addition, U can be given by

$$U = (P_B - P_A) Z_A = (P_B^{\infty} - P_A^{\infty} + Z_A^{-1} - Z_B^{-1}) Z_A$$

Since the rows of U sum to zero, $UP_A^{\infty} = UP_B^{\infty} = 0$, so that

$$U^2 = U(I - Z_B^{-1} Z_A)$$

and in general

$$U^{n+1} = U(I - Z_B^{-1} Z_A)^n \qquad n \geq 0 \tag{6.40}$$

Equation (6.40) indicates that U measures the difference between Z_A and Z_B.

6K. The Multichain Case

The perturbation approach developed above falls through if A and B are multi-chained. The mathematical breakdown occurs at the very beginning. Theorem 6.1 fails to hold since usually $Z_B^{-1} P_A^{\infty} \neq P_A^{\infty}$ in the general multichain case. That is, the rows of P_A^{∞} are not identical unless P_A has only one irreducible set of states.

One intuitive explanation for the breakdown is that P_A and P_B may have different numbers of chains, so that the chain structure of $P(\lambda)$ is discontinuous at $\lambda = 0$. This discontinuity in the $\pi(\lambda)$'s invalidates the perturbation approach.

CHAPTER 7

PERTURBATION THEORY AND PROGRAMMING OVER A MARKOV CHAIN

7A. Introduction

In this chapter the formalism for perturbation theory developed in Chapter 6 will be applied to the problem of programming over a Markov chain. We will be able to give a new derivation of the Howard-Jewell algorithm for policy iteration, and to give a new interpretation to the relative values as partial derivatives of the gain rate with respect to the matrix elements p_{ij}.

A geometric interpretation of policy iteration, as a maximizer of the directional derivative of the gain rate, is presented. This will lead to a discussion of randomized policies.

Various formulas for partial derivatives of g are given, and some interrelationships among test quantities are derived.

We assume that each of the Markov chains under consideration has a unique irreducible set of states. This guarantees the existence of a unique stationary distribution π and of the fundamental matrix Z.

7B. A New Derivation of the Policy Iteration Algorithm

i. The Policy Improvement Technique in the Markov Case

If a Markov chain A has a unique stationary vector $\underline{\pi}^A$ and an immediate expected reward vector $\underline{q}^A$, then it has a gain rate (expected reward per transition) g^A given by

$$g^A = (\pi^A, q^A) \tag{7.1}$$

In order to determine whether a new chain B has a higher gain rate, we compute the difference in gain rates,

$$\begin{aligned} g^B - g^A &= (\pi^B, q^B) - (\pi^A, q^A) \\ &= (\pi^B, q^B) - (\pi^B(I - U^{AB}), q^A) \\ g^B - g^A &= (\pi^B, \Gamma^{AB}) \end{aligned} \tag{7.2}$$

where

$$\Gamma^{AB} = q^B - q^A + (p^B - p^A)Z^A q^A \tag{7.3}$$

Since $\Gamma^{AA} = 0$, the policy B chosen in each state i by maximizing $(q^B - q^A + (p^B - p^A)Z^A q^A)_i$ satisfies $\Gamma_i^{AB} \geq 0$ so that $g^B \geq g^A$. If a quantity

$$\underline{v}^A = Z^A q^A + c\,\underline{1} \tag{7.4}$$

is introduced, the policy improvement technique can be characterized by the statement that it chooses B to maximize, in each state, the test quantity,

$$q^B + p^B v^A$$

This is precisely Howard's policy improvement technique developed in Chapter 4 of (Ref. 5), because it is possible to show that our v^A, defined

by (7.4), satisfies Howard's value determination equation:

$$\underline{v}^A = \underline{q}^A - g^A \underline{1} + p^A \underline{v}^A \tag{7.5}$$

The proof is to premultiply (7.4) by $I - p^A$ to obtain

$$\begin{aligned}(I - p^A)\underline{v}^A &= (I - p^A)Z^A \underline{q}^A \\ &= (I - p^{A^\infty})\underline{q}^A \\ &= \underline{q}^A - g^A \underline{1}\end{aligned}$$

ii. The Policy Improvement Technique in the Semi-Markov Case

If a semi-Markov chain A has a stationary vector $\underline{\pi}^A$, an immediate expected reward vector $\underline{q}^A$, and a mean holding time vector $\underline{T}^A$, then it has a gain rate (expected reward per unit time) g^A given by

$$g^A = \frac{(\pi^A, q^A)}{(\pi^A, T^A)} \tag{7.5}$$

The Markov case is obtained by setting $\underline{T}^A = \underline{1}$. In order to determine whether the gain rate g^B of a new policy B is greater than that of A, we compute the difference $g^B - g^A$ via

$$(\pi^B, T^B)(g^B - g^A) = (\pi^B, q^B - g^A T^B) \tag{7.6}$$

Since $g^B - g^A$ vanishes if $q^B - q^A = 0$, $T^B - T^A = 0$ and $p^B - p^A = 0$, we are motivated to introduce these three differences into the right of (7.6). Insertion of

$$q^B = q^A + (q^B - q^A)$$

$$T^B = T^A + (T^B - T^A)$$

converts the right side of (7.6) into

$$(\pi^B, q^B - q^A - g^A(T^B - T^A)) + (\pi^B, q^A - g^A T^A) \tag{7.7}$$

The last term on the right side of (7.7) would vanish if $p^B = p^A$, for that would imply $\pi^B = \pi^A$, and we know that

$$(\pi^A, q^A - g^A T^A) = 0 \tag{7.8}$$

We therefore seek to exhibit the factor $p^B - p^A$ in the last term. The goal can be achieved by insertion of

$$\pi^B = \pi^A + \pi^B U^{AB} \tag{7.9}$$

where

$$U^{AB} = (p^B - p^A) Z^A \tag{7.10}$$

into the last term.

One then obtains the expression

$$g^B - g^A = \frac{(\pi^B, \gamma^{AB})}{(\pi^B, T^B)} \tag{7.11}$$

where

$$\gamma^{AB} = q^B - q^A - g^A(T^B - T^A) + (p^B - p^A) Z^A (q^A - g^A T^A) \tag{7.12}$$

Since $\gamma^{AA} = 0$, it follows that if B is chosen to maximize, in each state, γ^{AB}, then $\gamma^{AB} \geq 0$.

Then (7.11) would show that $g^B \geq g^A$. If a vector

$$\underline{v}^A = Z^A(\underline{q}^A - g^A T^A) + c\,\underline{1} \tag{7.13}$$

is introduced, then the policy improvement technique can be characterized by the statement that it chooses B to maximize, in each state, the test quantity

$$TQ1 = q^B - g^A T^B + p^B v^A \tag{7.14}$$

The v^A's defined by (7.13) will be shown to satisfy the equation

$$v^A = q^A - g^A T^A + p^A v^A \tag{7.15}$$

and therefore agree with what Howard and Jewell call the relative values:

$$\begin{aligned}(I - p^A)v^A &= (I - p^A)Z^A(q^A - g^A T^A) \\ &= (I - p^{A^\infty})(q^A - g^A T^A) \\ &= q^A - g^A T^A\end{aligned}$$

since (7.8) implies that

$$p^{A^\infty}(q^A - g^A T^A) = 0.$$

Equations (7.12) and (7.15) imply that

$$\frac{\gamma_i^{AB}}{T_i^B} = \frac{(q^B - g^A T^B + (p^B - I)v^A)_i}{T_i^B} = \frac{(q^B + (p^B - I)\, v^A)_i}{T_i^B} - g^A \tag{7.16}$$

Consequently, if policy B is chosen to maximize the alternative test quantity TQ2,

$$TQ2_i = \frac{(q^B + (p^B - I)\, v^A)_i}{T_i^B} \tag{7.17}$$

then

$$TQ2_i = \frac{\gamma_i^{AB}}{T_i^B} + g^A \geq g^A + \frac{\gamma_i^{AA}}{T_i^A} = g^A$$

and

$$\gamma_i^{AB} = T_i^B (TQ2_i - g^A) \geq 0 \qquad (7.18)$$

Thus maximization of TQ2 also succeeds in producing a policy B with $g^B \geq g^A$.

The test quantities TQ1 and TQ2 are called value-oriented and gain-oriented, respectively, by Howard (Ref. 6, page 44) because of their dimensions.

The arbitrary constant c in (7.4) and (7.13) merely shifts the values and TQ1 by c, and does not enter TQ2 at all. It therefore does not affect the policy improvement technique. The usual convention is to adjust c such that $v_N = 0$. This completes the derivation of the Howard-Jewell algorithm[1,6] for gain maximization over a semi-Markov chain.

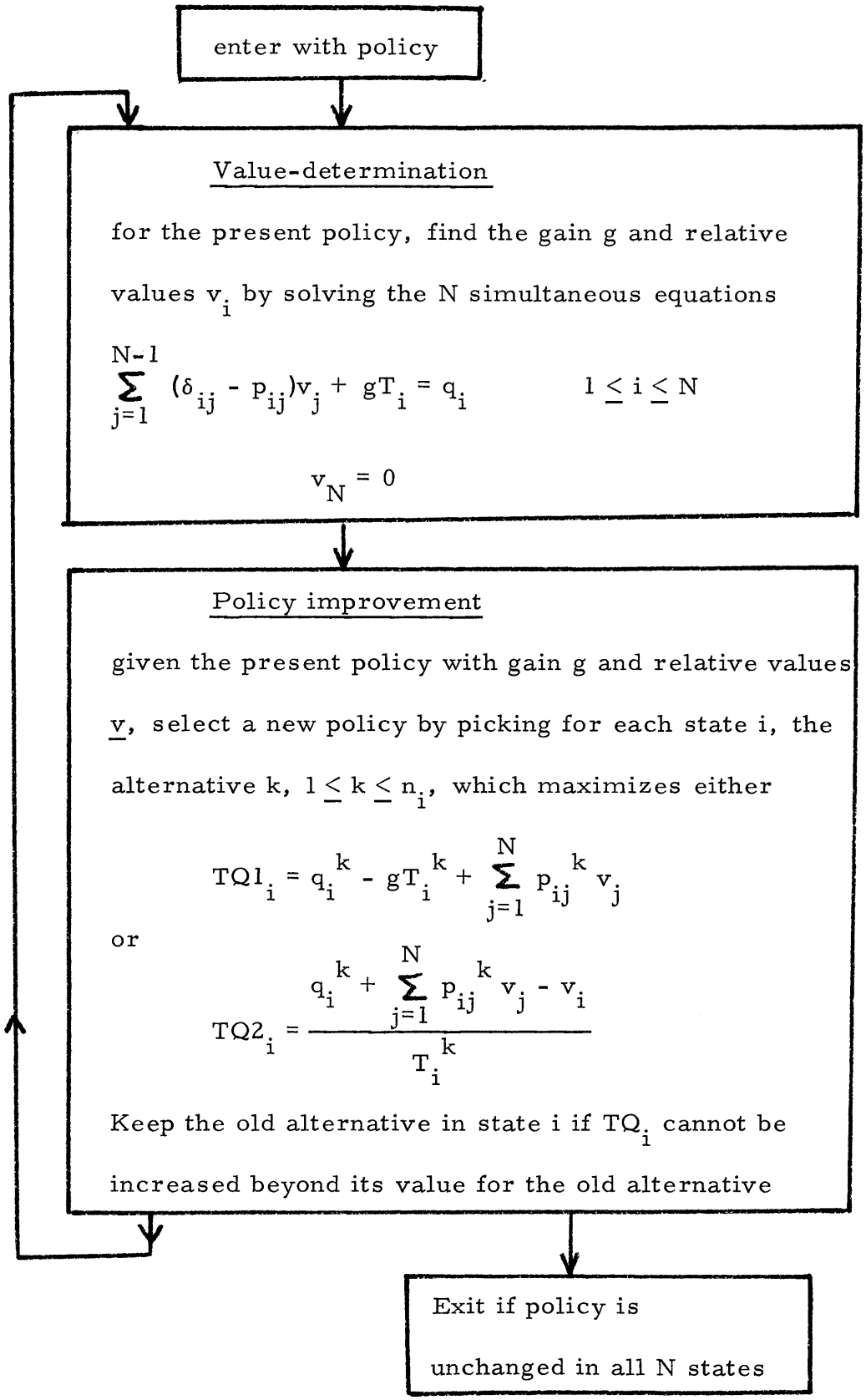

Figure 1. The Policy Iteration Algorithm

We already know that the gain rate is non-decreasing if this algorithm is used, and we will show in Chapter 8 that it converges if all policies each have only one irreducible set of states.

The "advantage" of the above derivation of the policy improvement algorithm is that the relative values enter in a purely mathematical--although very natural--fashion. No interpretation for the v's as asymptotic intercepts, as in (2.13b) is needed.

7C. A New Interpretation of the Relative Values

The above derivation of the policy-iteration algorithm by perturbation techniques stripped the v's of intuitive interpretation. In this section, perturbation techniques will be used to clothe the relative values with a new interpretation.

Inspection of (7.13) reveals that a convenient choice of additive constant is $C = 0$. This implies that

$$v = Z(q - gT) \tag{7.19}$$

and, as in (5.61), is called the natural convention for relative values. It corresponds to insisting upon the normalization

$$(\pi, v) = 0 \tag{7.20}$$

because $(\pi, v) = (\pi Z, q-gT) = (\pi, q-gT) = 0$.

If P is a small perturbation in an ergodic Markov chain P, chosen subject to the three requirements in 6D, then insertion of

$$\pi(P + \delta P) = \pi(P) + \pi(P)\delta PZ + O(\delta P)^2$$

into

$$g(P + \delta P) = \frac{(\pi(P + \delta P), q)}{(\pi(P + \delta P), T)}$$

leads to the result

$$g(P + \delta P) = g(P) + \frac{(\pi(P), \delta PZ(q-g(P)T))}{(\pi(P), T)} + O(\delta P)^2 \tag{7.21}$$

which could alternatively have been derived from (7.11) by letting $P_A \to P$, $q_A \to q$, $T_A \to T$, $P_B \to P + \delta P$, $q_B \to q$ and $T_B \to T$.

This may be written more concretely as

$$g(P + \delta P) = g(P) + \sum_{m,n=1}^{N} \frac{\partial g(P)}{\partial p_{mn}} \delta p_{mn} + O(\delta P)^2 \qquad (7.22)$$

where

$$\frac{\partial g}{\partial p_{ij}} = \frac{\pi_i v_j}{(\pi, T)} \qquad \text{(natural convention)} \qquad (7.23)$$

Equation (7.23) provides a new interpretation of the v_i's. They act as partial derivatives of the gain rate with respect to the matrix elements p_{ij}. This new interpretation is not totally unexpected since the problem of maximizing g can be set up via linear programming (Ref. 6, pages 47-52) with the v's appearing as dual variables to the π's. It is well-known that the dual variables (shadow prices) function as partial derivatives.

Previous definitions of the values, either formally by (7.19) or as a limiting asymptote, define the v's as properties of a fixed policy. It was not clear why the values are useful for policy improvement. Equation (7.23) provides the explanation.

7D. Ranking the States

Suppose i is a recurrent state, $\pi_i > 0$, and that $v_m > v_n$. Then a small change in the transition matrix which increases p_{im} and decreases p_{in} by the same amount will, according to (7.23), increase the gain rate. Stated verbally, if $v_m > v_n$ then any change in the probabilistic structure which enhances the "drift" towards m at the expense of that towards n will increase the gain rate. In this sense the relative values rank the states, since they indicate desirable modifications in the structure.

This use of the v's for ranking the states complements Howards: for a fixed policy, $v_m - v_n$ is the amount a rational person would be willing to pay to start in m rather than n.

7E. Gain Maximization by Parameter Variation

Suppose the policy chosen depended on a controllable parameter c, so that p_{ij}, q_i and T_i all depended on c. This parameter could be for example the trigger level in an s, S inventory scheme or the production rate of a machine. Of considerable interest is learning how to adjust the parameter to maximize the gain rate.

For large systems the analytical maximization of g(c) is near-impossible. But a perturbation approach is possible. If c is changed by a small amount δc, then according to (7.11) the gain rate is changed by $\delta g \doteq \frac{\partial g}{\partial c}\,\delta c$ where

$$\frac{\partial g(c)}{\partial c} = \frac{\sum_{i=1}^{N} \pi_i(c)\left[\frac{\partial q_i(c)}{\partial c} - g(c)\,\frac{\partial T_i(c)}{\partial c} + \sum_{j=1}^{N} \frac{\partial p_{ij}(c)}{\partial c} v_j(c)\right]}{\sum_{k=1}^{N} \pi_k(c)\, T_k(c)} \qquad (7.24)$$

Equation (7.24) is useful for numerical gain maximization by hill-climbing methods. One sets $c = c_0$ and finds the gain $g(c_0)$, values $v(c_0)$ and steady state vector $\pi(c_0)$ for this choice of c by solution of the value equations on a computer. $\frac{\partial g}{\partial c}(c_0)$ is then computed by (7.24) and used to determine whether to raise or lower the parameter setting.

7F. A Geometric Interpretation of Policy Iteration

The gain rate g for an N-state process is a function of $N^2 + N$ independent variables, namely the N(N - 1) independent p_{ij}'s, the N q_i's and the N T_i's. It therefore may be considered as a function defined on some region R in E_{N^2+N}.

A policy A, with p_{ij}^{A}, q_i^{A} and T_i^{A}, is a point $\underline{X^A}$ in R. The NN possible policies are NN points sprinkled in R.

A randomization of two policies A and B is given by

$$p_{ij}(\lambda) = (1 - \lambda)p_{ij}^{A} + \lambda p_{ij}^{B} \qquad 0 \leq \lambda \leq 1 \tag{7.25a}$$

$$q_i(\lambda) = (1 - \lambda)q_i^{A} + \lambda q_i^{B} \tag{7.25b}$$

$$T_i(\lambda) = (1 - \lambda)T_i^{A} + \lambda T_i^{B} \tag{7.25c}$$

These may be summarized as

$$X(\lambda) = (1 - \lambda)X^{A} + \lambda X^{B} \tag{7.25d}$$

which states that all randomizations of A and B are points on the line segment in R connecting X^A and X^B.

The randomized policy has a gain rate

$$g(\lambda) = g^{AB}(\lambda) = \frac{(\pi(\lambda), q(\lambda))}{(\pi(\lambda), T(\lambda))} \tag{7.26}$$

If (7.11) is used but with policy B replaced by the randomized policy of (7.25), then

$$g^{AB}(\lambda) - g^{A} = \lambda \frac{(\pi(\lambda), \gamma^{AB})}{(\pi(\lambda), T(\lambda))} \tag{7.27}$$

Division by λ and a limiting process produces

$$\left.\frac{dg^{AB}(\lambda)}{d\lambda}\right|_{\lambda=0} = \frac{(\pi^{A}, \gamma^{AB})}{(\pi^{A}, T^{A})} \tag{7.28}$$

Now $\pi^{A} \geq 0$ and the policy iteration algorithm picks B to maximize γ^{AB}. The policy iteration algorithm can be given the following geometric interpretation: there are NN points (policies) sprinkled in R. The algorithm, from any current point (policy) A, computes the directional derivative, evaluated at A, towards each of the remaining NN - 1 points (policies) B. The point B in whose direction the directional derivative at A is largest becomes the new policy, unless $\gamma^{AB} \leq 0$ for all B.

It is this interpretation of $\left.\frac{dg^{AB}(\lambda)}{d\lambda}\right|_{0}$ as the directional derivative of g at X^{A} in the direction $X^{B} - X^{A}$ which provides an explanation to Bellman's objection (Ref. 31, page 304) that "It is not intuitive that a decision dictated by gain and values of one policy (i. e., evaluated at $\lambda = 0$) will yield a policy with larger gain." The gain and values of X^{A} enter precisely because they determine the directional derivative of g at X^{A}.

It is important to note that the surfaces g = constant in R are not, in general, planes. It is not generally true that g is monotonic

if one proceeds along a line segment in R. This can be shown by the example in 8F where $g^{AB}(\lambda)$ is not monotonic in λ. In particular the fact that $\left.\frac{dg^{AB}(\lambda)}{d\lambda}\right|_{\lambda=0}$ is positive did not lead to $g^B > g^A$. In general $g(\lambda)$ is a rational function of λ of degree N, so that not much can be said about it. This follows from (7.26) since the $\pi(\lambda)$ may be obtained by application of Cramer's rule to (6.20) and are the ratio of a $(N-1)^{st}$ degree polynomial in λ (the cofactor) to a N^{th} degree polynomial in λ (the determinant), the latter of which cancels out in (7.26).

In one special case more detailed knowledge about the behavior of $g(\lambda)$ is available. Suppose policies A and B differ in only one state, say state k. Equation (6.16) goes into

$$\pi(\lambda)_i = \pi_i^A + \frac{\lambda \pi_k^A U_{ki}^{AB}}{1 - \lambda U_{kk}^{AB}} \qquad 1 \leq i \leq N \qquad (7.29)$$

and (7.27) then implies that $g(\lambda)$ is a linear fractional function of λ. Therefore $g(\lambda)$ is either strictly constant or strictly monotonic in λ.

7G. Relations Among the Test Quantities

In the Markov case, (7.2) implies that

$$g^B - g^A = (\pi^B, \Gamma^{AB})$$

Interchange of A and B sends $g^B - g^A$ into its negative, so that

$$(\pi^B, \Gamma^{AB}) = -(\pi^A, \Gamma^{BA}) = -\left(\pi^B(I - U^{AB}), \Gamma^{BA}\right)$$

The suggested identity

$$\Gamma^{AB} = -(I - U^{AB})\Gamma^{BA} \tag{7.30}$$

can be proved by summation of

$$\Gamma^{AB} = q^B - q^A + U^{AB}q^A$$

and

$$\begin{aligned}\Gamma^{BA} &= q^A - q^B + U^{BA}q^B \\ &= q^A - q^B - U^{AB}(I - U^{AB})^{-1} q^B \\ &= q^A - q^B - U^{AB}(I - U^{BA})q^B\end{aligned}$$

One obtains

$$\Gamma^{AB} + \Gamma^{BA} = U^{AB}\left[q^A - q^B + U_{BA}q_B\right] = U^{AB}\Gamma^{BA}$$

as desired.

Equation (7.31) implies that if $\underline{\Gamma}^{BA} = \underline{0}$, then $\underline{\Gamma}^{AB} = \underline{0}$ also.

A relation among the Γ's for three policies can be obtained by subtraction of

$$\Gamma^{AC} = (p^C - p^A)\, v^A + q^C - q^A$$

from the sum of

$$\Gamma^{AB} = (p^B - p^A)\, v^A + q^B - q^A$$

and

$$\Gamma^{BC} = (p^C - p^B)\, v^B + q^C - q^B$$

One obtains

$$\Gamma^{AB} + \Gamma^{BC} - \Gamma^{AC} = (p^C - p^B)(v^B - v^A)$$
$$= U^{BC} Z^{B-1} (v^B - v^A) \qquad (7.31)$$

Subtraction of the value equations

$$v^B = q^B - g^B \underline{1} + p^B v^B$$
$$v^A = q^A - g^A \underline{1} + p^A v^A$$

leads to

$$v^B - v^A = \Gamma^{AB} - (g^B - g^A)\underline{1} + p^B (v^B - v^A)$$

with solution

$$v^B - v^A = Z^B \Gamma^{AB} + c\,\underline{1} \qquad (7.32)$$

Insertion of (7.32) in (7.31) leads to the desired relationship,

$$(I - U^{BC})\Gamma^{AB} + \Gamma^{BC} = \Gamma^{AC} \qquad (7.33)$$

If c = A, (7.33) reduces to (7.30).

Formulas analogous to (7.30) and (7.33) for the semi-Markov case (the γ's) can be derived by similar techniques.

7H. A New Interpretation of the M Matrix

The M matrix is defined by (9.3) and has the property

$$g = \sum_{j=1}^{N} M^{-1}{}_{Nj}\, q_j$$

which implies that

$$M^{-1}{}_{Nj} = \frac{\pi_j}{(\pi, T)} \tag{7.34}$$

As A and B become close, (7.11-7.12) imply that

$$\delta g = \sum_{i=1}^{N} \left[\frac{\partial g}{\partial q_i}\, \delta q_i + \frac{\partial g}{\partial T_i}\, \delta T_i + \sum_{j=1}^{N} \frac{\partial g}{\partial p_i}\, \delta p_{ij} \right] \tag{7.35}$$

where

$$\frac{\partial g}{\partial q_i} = M^{-1}{}_{Nj} \tag{7.36}$$

$$\frac{\partial g}{\partial T_i} = -g\, M^{-1}{}_{Nj} \tag{7.37}$$

$$\frac{\partial g}{\partial p_{ij}} = M^{-1}{}_{Nj}\, v_j \tag{7.38}$$

In a similar fashion it can be shown that the other rows of the M^{-1} matrix give various partial derivatives of $v_i - v_N$.

CHAPTER 8

THE POLICY ITERATION ALGORITHM AND THE FUNCTIONAL EQUATION OF DYNAMIC PROGRAMMING

8A. Introduction

In Chapter 7, a policy iteration algorithm for finding the stationary policy with largest gain rate g was developed. In this chapter we discuss the convergence of the algorithm and describe its relation to the functional equation

$$v_i^* = \max_{1 \le k \le n_i} \left[q_i^k - g^* T_i^k + \sum_{j=1}^{N} p_{ij}^k v_j^* \right] \qquad (8.1)$$

$$1 \le i \le N$$

We assume throughout the chapter that each of the NN policies has a unique irreducible set of states.

We use the notation that R^A is the recurrent chain of a N state Markov chain P^A with stationary distribution π^A. That is,

$$R^A = \{ i \mid \pi_i^A > 0 \}$$

The transient states of A are the states i with $\pi_i^A = 0$.

R_A is the uniquely irreducible set of states for the process, and the transient states consist of all remaining states.

8B. Convergence of the Policy Iteration Algorithm

Theorem 8.1

Let all NN policies be ergodic. Suppose the policy iteration algorithm proceeds from a policy A to a different policy B, but with $g^B = g^A$. Then

(a) $R^B = R^A$ (The recurrent chains of A and B agree.)

$$\pi_i^B = \pi_i^A \qquad 1 \leq i \leq N$$

(b) A and B must have the same transient states. A and B each must have at least one transient state.

(c) by a choice of additive constant in the relative values,

$$v^A \leq v^B \tag{8.2}$$

with equality for all $i \in R^A$ and strict inequality for at least one transient state of A.

Proof

a) Subtraction of the value equations

$$v^B = q^B - g^B T^B + P^B v^B$$

$$v^A = q^A - g^A T^A + P^A v^A$$

for the two policies yields

$$v^B - v^A = \gamma^{AB} - (g^B - g^A)T^B + P^B(v^B - v^A) \tag{8.3}$$

$$= \gamma^{AB} + P^B(v^B - v^A) \tag{8.4}$$

where

$$\gamma^{AB} = q^B - q^A - g^A(T^B - T^A) + (p^B - p^A)v^A \tag{8.5}$$

Recalling the policy improvement technique by which B was chosen to supercede A,

$$\gamma_i^{AB} \geq 0 \qquad 1 \leq i \leq N \tag{8.6}$$

Dotting (8.3) with π^B gives another derivation of the equation

$$g^B - g^A = \frac{(\pi^B, \gamma^{AB})}{(\pi^B, T^B)} \tag{8.7}$$

Since $g^B - g^A = 0$, $\pi^B \geq 0$, $\gamma^{AB} \geq 0$ and $(\pi^B, T^B) > 0$, equation (8.7) implies that

$$\gamma_i^{AB} = 0 \qquad i \in R^B \tag{8.8}$$

Recalling the policy improvement technique,

$$k_i^B = k_i^A \qquad i \in R^B \tag{8.9}$$

$$p_{ij}^B = p_{ij}^A \qquad i \in R^B \quad 1 \leq j \leq N \tag{8.10}$$

Multiply (8.10) by π_i^B and sum over all $i \in R^B$. The sum can be extended to the i's with $i \notin R^B$ since these terms do not contribute. One obtains

$$\pi_j^B = \sum_{i=1}^{N} \pi_i^B \, p_{ij}^B = \sum_{i=1}^{N} \pi_i^B \, p_{ij}^A \qquad 1 \leq j \leq N$$

But p^A, since ergodic, has a unique left eigenvector π^A with eigenvalue unity. The above equation $\pi^B = \pi^B p^A$ therefore implies that

$$\pi_j^B = \pi_j^A \qquad 1 \leq j \leq N \tag{8.11}$$

so that the transient states of A and B are identical, and so are the recurrent states. Thus

$$R^B = R^A \tag{8.12}$$

b) There must be at least one L, say L_0, for which $(\gamma^{AB})_{L_0} > 0$. Otherwise $\gamma^{AB} = 0$ and B = A. Then L_0 is a transient state of B by (8.8), and is a transient state of A by (8.11).

c) The solution to (8.4) is

$$v^B - v^A = Z^B \gamma^{AB} + (\pi^B, v^B - v^A)\,\underline{1} \tag{8.13}$$

By proper choice of the additive constant in v_B, the last term on the right of (8.13) can be dropped, so that

$$v^B - v^A = Z^B \gamma^{AB} \tag{8.14}$$

If the formula

$$(Z^B)_{ij} = \delta_{ij} + \sum_{n=1}^{\infty} \left[(p^{Bn})_{ij} - \pi_j^B \right] \tag{8.15}$$

is inserted for Z^B, and if $(\pi^B, \gamma^{AB}) = 0$ is used, equation (8.14) becomes

$$(v^B - v^A)_i = \gamma_i^{AB} + \sum_{n=1}^{\infty} \sum_{j=1}^{N} (p^B)_{ij}^n \gamma_j^{AB} \qquad (8.16)$$

$$1 \leq i \leq N$$

Since $\gamma^{AB} \geq 0$ and $(p^B)^n \geq 0$ this implies

$$v^B - v^A \geq \gamma^{AB} \geq 0 \qquad (8.17)$$

In particular,

$$(v^B - v^A)_{L_0} \geq (\gamma^{AB})_{L_0} > 0 \qquad (8.18)$$

If $i \in R^B$, $(p^B)_{ij}^n = 0$ unless $j \in R^B$, and $j \in R^B$ implies $\gamma_j^{AB} = 0$.

Equation (8.16) becomes

$$(v^B - v^A)_i = 0 \qquad i \in R^B \qquad (8.19)$$

Equations (8.17, 8.18, 8.19) complete the proof of (c). QED

Stated verbally, (c) says that if the gain is not improved, then the relative value of at least one transient state must be improved.

Corollary 1

Let all NN policies be ergodic.

Let g^A achieve the maximum gain g of all NN policies.

Let $(\pi^A)_i > 0$ for all i.

Then $\gamma^{AB} \leq 0$. The policy iteration algorithm, if it reaches A, must converge on A.

Proof

Consider a policy improvement starting from A. Either a new policy B is found or the algorithm converges on A. We can show that the first alternative is impossible as follows.

We know by gain monotonicity that $g^B \geq g^A$ and also that $g^B \leq g^A$ since g^A achieves the highest possible gain. Therefore $g^B = g^A$. Appeal to part (b) of Theorem 10.1 to conclude that A has at least one transient state. This contradicts $\pi^A > 0$. QED

Corollary 2

The policy iteration algorithm cannot cycle.

Proof

Suppose the algorithm cycled on policies A B .. W A. By the monotonicity of gain, $g^A \leq g^B \leq .. \leq g^W \leq g^A$, so that $g^A = g^B = .. = g^W$. Invoke Theorem 10.1 to conclude A has at least one transient state. $v^A \leq v^B \leq .. v^W \leq v^A$; with equality for all recurrent states of A and strict inequality in $v^A \leq v^A$. for at least one transient state, which is a contradiction. QED

Corollary 3

The policy iteration algorithm must converge in a finite number of steps.

Proof

Corollary 2 shows that the algorithm cannot return to the same policy. Hence it must converge before exhausting all NN policies. QED

Corollary 4

Let all NN policies be ergodic. Let the policy iteration

algorithm converge to policy A. Then (a) g^A is the largest gain rate among all NN policies (b) policy A satisfies the functional equation

$$v_i^A = \max_{1 \le k \le n_i} \left[q_i^k - g^A T_i^k + \sum_{j=1}^{N} p_{ij}^k v_i^A \right] \qquad (8.20)$$

$$1 \le i \le N$$

Proof

(a) If the algorithm converges to A, $\gamma^{AB} \le 0$. Then using $\pi^B \ge 0$,

$$g^B = g^A + \frac{(\pi^B, \gamma^{AB})}{(\pi^B, T^B)} \le g^A$$

for any pure policy B.

(b) The property $\gamma^{AB} \le 0$, together with $\gamma^{AA} = 0$, implies that

$$\max_B \gamma^{AB} = 0$$

$$0 = \max_{1 \le k \le n_i} \left[q_i^k - q_i^A - g^A (T_i^k - T_i^A) + \sum_{j=1}^{N} (p_{ij}^k - p_{ij}^A) v_j^A \right] \qquad (8.21)$$

$$1 \le i \le N$$

By the value equation for policy A, $q^A - g^A T^A + p^A v^A = v^A$ so that (8.21) is equivalent to (8.20). QED

Theorem 8.1 and its four corollaries were proved under the assumption that all NN pure policies are ergodic, One might suspect, however, that the policy iteration algorithm converges under the weaker assumption that the NN pure policies each have a unique irreducible set of states (that is, the geometric multiplicity of $\lambda = 1$ is unity). That is, periodic chains are permitted, but only one closed set of states is allowed. This conjecture is correct and is formally stated as

Corollary 5

Theorem 8.1 and its four corollaries remain true if the ergodicity assumption is relaxed to the simpler assumption that each of the NN pure policies has a unique irreducible set of states.

Proof

Theorem 5.9 guarantees the solubility of the value-determination equations.

For any policy p^A with unique stationary vector $\pi^A \geq 0$, $(\pi^A, 1) = 1$, the recurrent states R^A are still defined as those i for which $\pi_i^A > 0$. The equation $\pi^A p^A = \pi^A$ still implies that if $i \in R^A$ then $p_{ij}^A = 0$ unless $j \in R^A$. Hence R^A is a closed set of states.

The only proof in Theorem 8.1 and its corollaries which needs revision is that of part (c) of Theorem 8.1. The new arguments proceed as follows.

If $i \in R^B$, then $\gamma_i^{AB} = 0$ and $p_{ij}^B = 0$ unless $j \in R^B$. Then (8.4) becomes

$$(v^B - v^A)_i = \sum_{j \in R^B} (p^B)_{ij}(v^B - v^A)_j \qquad i \in R^B$$

so that p^B, considered as a chain only on the recurrent states R^B, has $v^B - v^A$ as a right eigenvector with eigenvalue 1. But R^B is irreducible; therefore, the geometric multiplicity of $\lambda = 1$ is unity, hence $v^B - v^A$ must be a multiple of the unique right eigenvector $\underline{1}$:

$$(v^B - v^A)_i = c \qquad i \in R^B$$

and by choice of additive constant in v^B:

$$(v^B - v^A)_i = 0 \qquad i \in R^B \tag{8.22}$$

Let $S = \left\{ i \middle| (v^B - v^A)_i = (v^B - v^A)_{min} \right\}$

Equation (8.4) implies that

$$(v^B - v^A)_i = \gamma_i^{AB} + p^B(v^B - v^A)_i \geq \gamma_i^{AB} + (v^B - v^A)_{min}$$

Pick $i \in S$ to conclude $\gamma_i^{AB} = 0$ and $p_{ij}^B = 0$ unless $j \in S$.

Hence S is a closed set of states for p^B. Either S and R^B are disjoint or they overlap. They cannot be disjoint, for that would contradict the assumption that p^B has only one irreducible set of states. Therefore S and R^B overlap and (8.22) implies that

$$(v^B - v^A)_{min} = 0$$

Consequently

$$(v^B - v^A)_i \geq 0$$

with equality for all $i \in R^B$ and strict inequality for at least one i, namely $i = L_0$. This completes the proof of Theorem 8.1c. QED

8C. Discussion of Convergence

1) Previous descriptions[1, 6, 7] of the policy iteration algorithm showed that the gain was non-decreasing as the policy improvement part proceeded from one policy to another, but never completed the proof that the algorithm converges by showing that cycling at a fixed value of gain was impossible.

Since the policy iteration algorithm is related to linear programming (where only one component of the decision vector is changed at a time), which can cycle, it might have been feared that the policy iteration algorithm could cycle. Fortunately, the rule of not changing alternatives in a state if a positive improvement in test quantity for that state is impossible prevents cycling.

8D. Value-Maximization for Transient States

Theorem 8.1(c) shows that the policy iteration algorithm does not, in general, converge immediately upon locating the policy with highest gain rate. Instead, iterations continue from one policy to another, with all recurrent chains identical, with relative values identical for all recurrent states, and with an improvement in relative value for at least one transient state. Thus the "route" by which the system passes from a transient to recurrent state is improved.

Howard, in the Appendix to reference (5), reached the same conclusion for the Markov case, and used it to describe his baseball example. In fact, if j is a transient state of B, then our Z_{ij}^{B} and Howard's U_{ij} agree both represent the mean number of occurrences, starting from state i, of state j.

We can loosely state the above property by saying that the policy iteration algorithm does not merely maximize the gain rate. It also maximizes the relative (to the values of the recurrent states) values of the transient states as well. This statement is justified by the following theorem, which shows that if the policy iteration algorithm converges to policy A, no other policy can have larger relative values.

Theorem 8.2

Let each of the NN pure policies have a unique irreducible set of states, so that the policy iteration algorithm converges, say to policy A, where g^A is the maximum of the NN gain rates. Also v^A satisfies

$$v_i^A = \max_{\text{all NN B's}} \left[q_i^B - g^A T_i^B + (p^B v^A)_i \right] \qquad (8.20)$$

$$1 \leq i \leq N$$

Then

$$v_i^A = \max_{\substack{B \\ g^B = g^A}} \left[Z^B (q^B - g^B T^B)_i + (\pi^B, v^A) \right] \qquad (8.23)$$

$$1 \leq i \leq N$$

Proof

For any policy B with $g^B = g^A$,

$$v^A = q^B - g^B T^B + p^B v^A \quad .$$

Define $v^B = Z^B(q^B - g^B T^B)$. Then

$$v^B = q^B - g^B T^B + p^B v^B \quad .$$

Subtracting,

$$v^A - v^B \geq p^B (v^A - v^B)$$

$$v^A - v^B \geq (p^B)^m (v^A - v^B) \qquad m \geq 1$$

$$v^A - v^B \geq \frac{p^B + \ldots + (p^B)^m}{m} (v^A - v^B) = S_m (v^A - v^B)$$

Letting $m \to \infty$, $S_m \to S$ by Theorem (4.3) and $S = p^{B^\infty}$ by (4.78). Then

$$v^A \geq v^B + p^{B^\infty}(v^A - v^B) = v^B + (\pi^B, v^A)\,\underline{1}$$

since $p^{B^\infty} v^B = p^{B^\infty}(q^B - g^B T^B) = 0$. Thus

$$v_i^A \geq \max_{\substack{B \\ g^B = g}} \left[Z^B (q^B - g^B T^B)_i + (\pi^B, v^A) \right]$$

The maximum is achieved when $B = A$. QED

Corollary

Let B be any policy with $g^B = g^A$. Then be appropriate choice of additive constant in v^B,

$$v^B \leq v^A \tag{8.24a}$$

$$v_i^B = v_i^A \qquad i \in R^B \tag{8.24b}$$

Proof

Let $\underline{v}^B = Z^B(q^B - g^B T^B) + (\pi^B, v^A)\,\underline{1}$

with the property $(\pi^B, v^B) = (\pi^B, v^A)$. Then (8.23) implies that $v^A \geq v^B$, which proves (8.24a). Multiply (8.24a) by $\pi^B \geq 0$ to obtain

$$(\pi^B, v^A) = (\pi^B, v^B) \leq (\pi^B, v^A)$$

This must be an equality, whence

$$v_i^B = v_i^A \qquad i \in R^B$$

which confirms (8.24b). QED

Discussion

(1) (8.23) shows that, except for the additive constant (π^B, v^A), the relative value v_i^A of the optimal policy is the maximum, over all policies B which achieve the maximum gain rate g^A, of the relative value $Z^B(q^B - g^B T^B)_i$.

(2) The policy which achieves the maximum in (8.23) is independent of i.

(3) If only one state, say state N, is recurrent for all gain-optimal policies:

$$\pi_i^B = \delta_{i,N} \qquad 1 \leq i \leq N$$

all B with $g^B = g^A$

then the choice of additive constant $v_N^A = 0$ converts (8.23) to

$$v_i^A = \max_{\substack{B \\ g^B = g^A}} \left[Z^B(q^B - g^B T^B) \right]_i \qquad 1 \leq i \leq N \tag{8.25}$$

In particular, multiplication by π^A turns (8.25) into

$$v_N^A = (\pi^A, v^A) = (\pi^A, Z^A(q^A - g^A T^A)) = 0$$

so that (8.25) is consistent with $v_N^A = 0$.

(4) The corollary shows that the values v^A for the policy A to which the policy iteration algorithm converges are largest in the sense that any other policy B with the maximum gain rate has relative values satisfying $v^B \leq v^A$, with equality for some components.

(5) The statement is sometimes loosely made that the set S of all stationary policies which achieve the maximum gain rate can be located by first finding any policy A to which the policy iteration algorithm converges, and then identifying S as the set of all policies B with $\underset{\sim}{\gamma}^{AB} = \underline{0}$.

This is incorrect. The set of policies B with $\gamma^{AB} = 0$ not only achieves g^A but also satisfies the functional equation (8.20), hence is in general a subset of S. It is usually difficult to locate all policies B with maximum gain rate, because some of these may satisfy $\gamma_i^{AB} < 0$ for some transient states i of B.

8E. The Functional Equation of Dynamic Programming

We now discuss the existence and uniqueness of the solution to the functional equation

$$v_i^* = \max_{1 \le k \le n_i} \left[q_i^k - g^* T_i^k + \sum_{j=1}^{N} p_{ij}^k v_j^* \right] \qquad (8.26)$$

$$1 \le i \le N$$

Theorem 8.3

Let each of the NN pure policies have only one irreducible set of states. Then

(a) a solution (g^*, v_i^*) exists to (8.26)

(b) g^* is unique:

$$g^* = \max_B g^B \qquad (8.27)$$

where the maximization is taken over all NN pure policies.

(c) $\underline{v}$ is unique up to an arbitrary additive multiple of $\underline{1}$.

Proof

(a) Corollary 3 of Theorem 8.1 shows that the policy iteration algorithm converges, say to policy A. Then Corollary 4 shows that (g^A, v_i^A) satisfy (8.26).

(b) Let $\underline{k}$ denote any of the NN pure policies. Then (8.26) implies

$$v_i^* \ge g_i^k - g^* T_i^k + \sum_{j=1}^{N} p_{ij}^k v_j^* \qquad 1 \le i \le N$$

Multiplication of $\pi_i^k \geq 0$, summation over i and cancellation of (π^k, v^*) yields

$$g^* \geq \frac{(\pi^k, q^k)}{(\pi^k, T^k)} = g^k$$

This, combined with $g^* = g^A$, yields (8.27).

(c) If $(g^*, \underline{v}^*)$ is a solution to (8.26), it is easy to show that so is $(g^*, \underline{v}^* + c\,\underline{1})$ for any scalar c. Conversely, if $(g^*, \underline{v}^*)$ and $(g^{**}, \underline{v}^{**})$ are two solutions to (8.26) we will now show that $x = v^{**} - v^*$ is a multiple of $\underline{1}$.

Since $g^{**} = g^*$, the functional equations become

$$v_i^* = \max_{1 \leq k \leq n_i} \left[q^k - g^* T^k + p^k v^* \right]_i \qquad (8.28)$$

$$1 \leq i \leq N$$

$$v_i^{**} = \max_{1 \leq k \leq n_i} \left[q^k - g^* T^k + p^k v^{**} \right]_i \qquad (8.29)$$

Define b by

$$b_i^k = q_i^k - g^* T_i^k + \sum_{j=1}^{N} p_{ij}^k v_j^{**} - v_j^{**} \qquad (8.30)$$

$$1 \leq k \leq n_i$$

$$1 \leq i \leq N$$

with the property, according to (8.29),

$$\max_{1 \le k \le n_i} b_i^k = 0 \qquad 1 \le i \le N \tag{8.31}$$

If 8.30 is solved for $q_i^k - g^* T_i^k$, and this result put into (8.28), one obtains

$$x_i = \max_{1 \le k \le n_i} \left[b_i^k + \sum_{j=1}^{N} p_{ij}^k x_j \right] \tag{8.32}$$

If Theorem 10.2 is invoked to deal with (8.31-8.32) we conclude

$\underline{x} = c\,\underline{1}$. QED

Theorem 8.4

Let policy A achieve g given by (8.27). Let all states of A be recurrent, $\pi^A > 0$. Then A satisfies (8.26). That is, (8.20) holds.

Proof

According to Corollary 1 of Theorem (8.1), the policy iteration algorithm, starting from A, converges on A. Invoke Corollary 4 to get the desired result. QED

Corollary 1 Bellman's Theorem[4]

Suppose all policies satisfy $p_{ij} \ge d > 0$. Let policy A achieve g^* given by (8.27). Then policy A satisfies (8.20).

Proof

The condition $p_{ij} \ge d > 0$ ensures that $\pi^A > 0$. Then invoke the above theorem. QED

8F. Randomized Policies

We show in this section that (8.26) possesses a solution (indeed the same solution) if randomized policies are permitted. Furthermore the maximum gain rate achievable by a stationary randomized strategy agrees with the maximum gain rate achievable by one of the NN pure policies. Hence no generality has been lost be restricting ourselves to pure policies.

A stationary randomized policy is described by a matrix

f_{ik} $1 \leq i \leq N$ $1 \leq k \leq n_i$ where

$$f_{ik} \geq 0 \qquad \sum_{k=1}^{n_i} f_{ik} = 1 \qquad 1 \leq i \leq N \tag{8.33}$$

f_{ik} is the probability that alternative k is used in state i.

A randomized policy f_{ik} has a mean holding time $T_i(f)$ and mean immediate expected reward $q_i(f)$ in state i given by

$$T_i(f) = \sum_{k=1}^{n_i} f_{ik} T_i^k \qquad q_i(f) = \sum_{k=1}^{n_i} f_{ik} q_i^k \tag{8.34}$$

$$1 \leq i \leq N$$

The transition probability matrix $p_{ij}(f)$ is given by

$$p_{ij}(f) = \sum_{k=1}^{n_i} f_{ik} p_{ij}^k \qquad 1 \leq i, j \leq N \tag{8.35}$$

If each of the NN pure policies has a unique irreducible set of states, then this will be true for p(f) for any f. Hence a stationary distribution $\pi(f)$ will exist for any f. The gain rate g(f) for the randomized policy is given by

$$g(f) = \frac{(\pi(f), q(f))}{(\pi(f), T(f))} \tag{8.36}$$

Theorem 8.5

Let each of the NN pure policies have a unique irreducible set of states. Then the functional equation

$$v_i^* = \max_{f_{ik}} \left\{ \sum_{k=1}^{n_i} f_{ik} \left[q_i^k - g^* T_i^k + \sum_{j=1}^{N} p_{ij}^k v_j^* \right] \right\} \tag{8.37}$$

$$f_{ik} \geq 0 \qquad \sum_{k=1}^{n_i} f_{ik} = 1 \qquad\qquad 1 \leq i \leq N$$

has a solution. Indeed (8.37) is equivalent to (8.26).

Proof

Define v_i^* and g^* by (8.26), which we know has a solution. Since the maximum possible weighted average of a set of numbers is the largest number in the set,

$$v_i^* = \max_{1 \leq k \leq n_i} \left[q_i^k - g^* T_i^k + \sum_{j=1}^{N} p_{ij}^k v_j^* \right]$$

$$= \text{right hand side of (8.37)}$$

which proves (8.37). QED

Corollary

Let each of the NN pure policies have a unique irreducible set of states. Let g^{**} denote the maximum gain rate achievable by a stationary randomized policy:

$$g^{**} = \max_{f} g(f) \tag{8.38}$$

where g(f) is given by (8.36). (The supremum is actually achieved since g(f) is a continuous function of f defined on a closed and bounded domain for f.)

Then g^{**} agrees with the maximum gain rate achievable by the NN pure policies. That is,

$$g^{**} = g^{*} \tag{8.39}$$

where g^{*} is given by (8.27).

Proof

According to (8.37),

$$v_i^{*} \geq q_i(f) - g^{*} T_I(f) + \sum_{j=1}^{N} p_{ij}(f)\, v_j^{*}$$

$$1 \leq i \leq N$$

Multiply by $\pi_i(f) \geq 0$ and sum over i to obtain

$$g^{*} \geq \frac{(\pi(f), q(f))}{(\pi(f), T(f))} = g(f)$$

$$g^{*} \geq g^{**}$$

On the other hand, by letting f run over the NN pure policies, (8.38) implies $g^{**} \geq g^{*}$. Together these imply (8.39). QED

This corollary was originally proved by Wagner[10], who set up the right side of (8.38) as a linear programming problem. We obtained it as a by-product of the functional equation.

It is important to note the crucial feature that the different rows for p can be chosen completely independently. If a choice of alternative in one state constrains the choice of alternative in another state, then it is no longer generally true that a pure policy has equal or higher gain rate than a randomized policy.

To demonstrate this property, we consider the case of two policies A and B given by

$$p^{A} = p^{A^{\infty}} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}; \quad q^{A} = [0 \quad 0]; \quad g^{A} = 0$$

$$p^{B} = p^{B^{\infty}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}; \quad q^{B} = [-1 \quad 1]; \quad g^{B} = -1$$

If $\lambda (0 \leq \lambda \leq 1)$ is the probability, in either state, that policy B is used, then

$$p(\lambda) = (1-\lambda)p^{A} + \lambda p^{B} = \begin{bmatrix} \lambda & 1-\lambda \\ \lambda & 1-\lambda \end{bmatrix}; \quad \pi(\lambda) = [\lambda \quad 1-\lambda]$$

$$q(\lambda) = (1-\lambda)q^{A} + \lambda q^{B}$$

$$g(\lambda) = (\pi(\lambda), q(\lambda)) = \lambda(1 - 2\lambda)$$

A plot of $g(\lambda)$ reveals that $g(\lambda)$ is monotone increasing for $0 \leq \lambda \leq .25$ from $g^A = 0$ to a maximum of $1/8$, and monotone decreasing for $.25 \leq \lambda \leq 1$ from $1/8$ to a minimum of $g^B = -1$.

Thus $g(\lambda)$, the gain resulting from a randomization of A and B, is not monotone. Its maximum does <u>not</u> occur at $\lambda = 0$ or 1, pure policies. Even more striking, it is not true that $g(\lambda)$ lies between $0 = \max(g^A, g^B)$ and $-1 = \min(g^A, g^B)$. A randomized strategy $\lambda = .25$ has higher gain rate than do the pure policies $\lambda = 0, 1$.

8G. The Supremum Case

In order to demonstrate the epsilonic acrobatics which are required if one has only suprema and not maxima, we prove the following theorem:

Theorem 8.6

Let any policy $\underline{k} = k(x)$ $(k(x) \in S(x)$ for all $x \in \Omega)$ have a unique π vector. If $v^*(x)$ and g^* satisfy

$$v^*(x) = \sup_{k \in S(x)} \left[q^k(x) - g^* T^k(x) + \int_\Omega dy\, p^k(x,y) v^*(y) \right] \qquad (8.40)$$

$$x \in \Omega$$

then $g^* = \sup_{\text{all } \underline{k}} g^{\underline{k}}$ (8.41)

Proof

$$v^* \geq q^k - g^* T^k + p^k v^* \qquad \text{any } \underline{k}$$

Multiply by $\pi^k \geq 0$ and integrate over x to get

$$g^* \geq \frac{(\pi^k, q^k)}{(\pi^k, T^k)} = g^k \qquad \text{any } \underline{k} \qquad (8.42)$$

On the other hand, given any $\epsilon > 0$ there is a policy $k^*(x)$ such that

$$v^*(x) \leq q^*(x) - g^* T^*(x) + \int_\Omega dy\, p^*(x,y) v^*(y) + \epsilon$$

$$\text{all } x \in \Omega$$

Multiply by $\pi^*(x) \geq 0$ and integrate over all x to obtain

$$g^* \leq \frac{(\pi^*, q^*)}{(\pi^*, T^*)} + \epsilon \qquad \text{for some } k > (x) \qquad (8.43)$$

Equation (8.42) and (8.43) complete the proof of (8.41).

QED

Corollary

$$g^* = \sup_{\text{all } f} g(f)$$

That is, g^* is the supremum over all randomized strategies as well as the supremum over all pure strategies.

Proof

Same as the proof of the corollary to Theorem 8.5.

8H. Characterization of the Solutions to the Functional Equation

Suppose each of the NN policies possess a unique irreducible set of states, so that (8.26) possesses a solution. Then the set of all policies which satisfy (8.26) can be achieved in the following way.

Theorem 8.7

Let a policy A satisfy (8.26), i.e., let (8.20) hold. Then policy B is another solution of (8.26) if and only if $\gamma^{AB} = \underline{0}$.

Proof

a) If A and B are both solutions, we may subtract the two equations

$$v^A = q^A - g^A T^A + p^A v^A$$

$$v^B = q^B - g^B T^B + p^B v^B$$

and insert $g^B = g^A$, $v^B = v^A + c\,\underline{1}$ to obtain

$$0 = q^B - q^A - g^A(T^B - T^A) + (p^B - P^A)v^A = \gamma^{AB}$$

b) If A is a solution and $\gamma^{AB} = 0$, then (8.7) shows that $g^B = g^A$. Also

$$v^A = \max_k (q^k - g^A T^k + p^k v^A) \tag{8.44}$$

$$= q^A - g^A T^A + p^A v^A$$

$$= q^B - g^A T^A + p^B v^A + \gamma^{AB}$$

$$= q^B - g^B T^B + p^B v^A \tag{8.45}$$

(8.45) shows that the choice $k = k^B$ in (8.44) achieves the maximum.

QED

Theorem 8.8

Policy A satisfies (8.26) if and only if $\gamma^{AB} \leq 0$ for all B.

Proof

If $\gamma^{AB} \leq 0$, then $\max_B \gamma^{AB} = \gamma^{AA} = 0$ from which (8.26) follows.

If A satisfies (8.26), then (8.20) holds so that

$$q^A - g^A T^A + p^A v^A = \max_k (q^k - g^A T^k + p^k v^A)$$

whence $\max_k \gamma^{Ak} = 0$ and $\gamma^{Ak} \leq 0$. QED

Theorem 8.9

Let $K_i^* = \{k \mid 1 \leq k \leq n_i;$ k achieves the maximum on the right side of (8.26) $\}$ (8.46)

Then any policy $(\underline{k}^*) = (k_1^*, \ldots k_N^*)$ with $k_i^* \in K_i^*$ has gain and values which satisfy (8.20)

Proof

By definition,

$$v_i^* = (q^* - g^* T^* + p^* v^*)$$

which confirms that g^* and v^* are the gain and values of k^*. QED

Theorem 8.10 (Converse)

Let policy $k^A = (k_1^A, \ldots k_N^A)$ produce gain and values which satisfy (8.20). Then

$$k_i^A \in \mathcal{K}_i^* \qquad 1 \leq i \leq N$$

Proof

Comparison of (8.20) and (8.26), and appeal to Theorem 8.3c, yields the conclusion

$$\underline{v}^A = \underline{v}^* + c\,\underline{1}$$

If k^A maximizes the right side of (8.20), it also maximizes the right side of (8.26) and $k_i^A \in \mathcal{K}_i^*$. QED

A convenient way to find all solutions to (8.26) is provided by Theorems 8.9-8.10.

First the policy-iteration algorithm is used to find a single policy A which satisfies (8.20). Then $g^A = g^*$ and $v^A = v^* + c\,\underline{1}$, and (8.46) can be used to find all alternatives in state i, $\mathcal{K}_i^*$ which are optimal. These alternatives are precisely the ones for which the test quantity used in the policy improvement technique achieves its maximum value.

8I. Vanishing Interest Rate

In this section, we investigate the behavior of $\underline{v}(a)$, the maximum expected reward using a stationary policy, as the discount factor a goes to unity [11]. $\underline{v}(a)$ is given by (2.18c) in the Markov case and (2.18d) in the semi-Markov case. We show that as $a \to 1$, the equations with discounting approach the functional equations (2.15g) and 2.15h) without discounting, respectively. Hence the policy chosen in the discounted case will maximize the expected reward per transition or expected reward per unit time as the interest rate goes to zero.

We assume that the expected reward per transition $q_i^K(a)$ is twice differentiable in a so that

$$q_i^K(a) = q_i^K + (a-1)\, q_i^{K\prime}(1) + O((1-a)^2) \qquad (8.47)$$

holds for all i and K. In particular, the finiteness of $q_i^{K\prime}(1)$ is related to the finiteness of the area η_i^K under the curve $[q_i^K(t) - q_i^K(\infty)]$ if the rewards come in via some (holding-time-dependent) rate: it can be shown that

$$q_i^{K\prime}(1) = -\eta_i^K \,. \qquad (8.48)$$

In addition we assume that each of the NN policies has a single irreducible set of states, i.e. a unique π vector. Z^B will denote the fundamental matrix for policy B.

The discount factor a satisfies $0 < a < 1$ and will be allowed to approach 1 from below.

Theorem 8.11 (Markov Case)

Let $V_i(a)$ be given by (2.18a)

$$V_i(a) = \max_{\text{NN policies } B} \left[(I - a P^B)^{-1} q^B(a) \right]_i \qquad (8.49)$$

$$1 \le i \le N$$

or alternatively by (2.15e):

$$V_i(a) = \max_{1 \le K \le n_i} \left[q_i^K(a) + a \sum_{j=1}^{N} p_{ij}^K V_j(a) \right] \qquad (8.50)$$

$$1 \le i \le N$$

$$0 < a < 1$$

Define the expected reward per transition and relative values (note convention) for a policy B by

$$g^B = (\pi^B, q^B) \qquad (8.51)$$

$$V_i^B = \left[Z^B q^B \right]_i - (\pi^B, q^{B\prime(1)}) \qquad (8.52)$$

$$1 \le i \le N$$

Then

i) As $a \to 1$, the policy C which achieves the maximum on the right side of (8.49) satisfies

$$V_i(a) = \frac{g^C}{1-a} + v^C{}_i + o(1-a) \tag{8.53}$$

ii)

$$g^C = \max_{\text{NN policies } B} [g^B] \tag{8.54}$$

iii)

$$v^C{}_i = \max_{\substack{B \\ g^B = g^C}} [v^B{}_i] \qquad 1 \le i \le N \tag{8.55}$$

iv) v^C satisfies the functional equation (2.15g):

$$v^C_i = \max_{1 \le K \le n_i} \left[q_i^K + \sum_{j=1}^{N} p_{ij}^K v^C_j \right] \tag{8.56}$$

$$1 \le i \le N$$

Proof: Insert (4.77):

$$(I - aP^B)^{-1} = \frac{P^{B\infty}}{1-a} + (I - aE^B)^{-1}$$

$$= \frac{P^{B\infty}}{1-a} + (I - E^B)^{-1} + O(1-a)$$

$$= \frac{P^{B\infty}}{1-a} + Z^B + O(1-a)$$

and (8.47) into (8.49) to obtain

$$V_i(a) = \max_B \left[\frac{g^B}{1-a} + V^B{}_i + O(1-a) \right]$$

As $a \to 1$, the policy C is chosen first to maximize the divergent term and second, if there are any ties in gain rate, to maximize V^B. This proves (8.53-8.55). Insertion of (8.53) into (8.50), use of

$$\frac{1}{1-a} g^C = g^C + \frac{a}{1-a} g^C$$

and a limiting argument as $a \to 1$, noting that g^C and V^C are independent of a, leads to (8.56).

QED

Theorem 8.12 (Semi-Markov Case)

Let $V_i(a)$ be given by (2.18d):

$$V_i(a) = \max_{\text{NN policies } B} \left(\left[I - \tilde{P}^B(\ln \gamma_a) \right]^{-1} q^B(a) \right)_i \qquad (8.57)$$

$$1 \le i \le N$$

or alternately by (2.15f):

$$V_i(a) = \max_{1 \le K \le n_i} \left[q_i^K(a) + \sum_{j=1}^{N} \tilde{p}_{ij}^K(\ln \gamma_a) V_j(a) \right] \qquad (8.58)$$

$$1 \le i \le N$$

$$0 < a < 1$$

Define the expected reward per unit time and relative values (note convention) for a policy B by

$$g^B = \frac{(\pi^B, q^B)}{(\pi^B, \pi^B)} \qquad (8.59)$$

$$V_i^B = Z^B (q^B - g^B \pi^B)_i + d^B \qquad (8.60)$$

$$d = \frac{(\pi, \Pi Z q)}{(\pi, \pi)} + [\langle A_0 \rangle + 1](\pi, q)$$

$$- \frac{(\pi, q'(1))}{(\pi, \pi)} - \frac{g}{2} \tag{8.61}$$

where $\langle A_0 \rangle$ is given by (5.108), T and $T^{(2)}$ by Theorem 5.10.

Then the policy C which maximizes the right hand side of (8.57) for a near 1 satisfies

i)

$$V_i(a) = \frac{g^C}{1-a} + V^C_i + O(1-a) \tag{8.62}$$

ii)

$$g^C = \max_{\text{NN policies } B} [g^B] \tag{8.63}$$

iii)

$$V^C_i = \max_{\substack{B \\ g^B = g^C}} [V^B_i] \qquad 1 \le i \le N \tag{8.64}$$

iv) V^C satisfies the functional equation (2.15h):

$$V^C_i = \max_{1 \le K \le n_i} \left[q_i^K - g^C \pi_i^K + \sum_{j=1}^{N} p_{ij}^K V^C_j \right] \tag{8.65}$$

Proof:

By appeal to Theorem 5.10,

$$\left[I - \tilde{P}(s)\right]^{-1} = \frac{P^{\infty}}{(\pi,\pi)\, s} + Z - \frac{Z \pi P^{\infty}}{(\pi,\pi)}$$

$$- \frac{P^{\infty} \pi Z}{(\pi,\pi)} + \left[\langle A_0 \rangle + 1\right] P^{\infty} + O(s)$$

Also, when $s = \ln \gamma_a = -\ln (1-(1-a))$,

$$\frac{1}{s} = \frac{1}{1-a} - \frac{1}{2} + O(1-a)$$

so that (8.57) and (8.47) imply

$$v(a)_i = \max_B \left[\frac{g^B}{1-a} + v^B_i + O(1-a) + O(\ln \gamma_a) \right]$$

Maximization of the divergent term and then, if there are ties in gain, the relative values, leads to (8.62-8.64). Insertion of (8.62) into (8.58) and use of

$$\tilde{P}(\ln \gamma_a) = P - (\ln \tfrac{1}{a})\, \pi + O\left(\ln^2 \tfrac{1}{a}\right)$$

$$= P - (1-a)\, \pi + O((1-a)^2)$$

leads to (8.65) as $a \to 1$.

QED

CHAPTER 9

POLICY ITERATION CHANGING ONLY ONE ALTERNATIVE PER CYCLE

9A. Introduction

The simplicity of equations (6.15) and (6.16) when policies A and B differ in only one state leads us to suspect that the gain and values change in a simple fashion if the alternative is changed in only one state. This, indeed turns out to be so. The equations (9.15) and (9.16) which exhibit the change in gain, values and M^{-1} are so simple, and even more important, involve so little storage in a computer fast memory, that this method of policy improvement in only one state per iteration is suggested as a practical technique for dynamic programming over Markov chains with very many states.

To date the policy-iteration algorithm has not been used for large problems, say more than 100 states, because of the difficulty of solving 100 simultaneous equations. The method proposed here bypasses the solution of simultaneous equations. It allows computer solution of problems involving up to five thousand states, with arbitrarily many alternatives permitted in each state. It is therefore a strong competitor to linear programming for gain maximization of a Markovian reward process.

The statement is sometimes loosely made that the policy iteration algorithm is merely a modified form of linear programming--with the gain rate as objective function--in which several vectors can be brought into the basis simultaneously. This is not quite correct since linear programming may converge once the maximal gain is found while the policy iteration algorithm, according to 8D, keeps iterating until the relative values of the transient states have been maximized as well.

The proposed scheme of changing only one alternative per iteration is therefore distinct, although closely related, to linear programming. In particular, Equation (9.16) is reminiscent of the change in simplex tableau when a new vector is brought into the basis.

9B. Changing the Decision in State α

The value determination equations for a policy A

$$v_i^A = q_i^A - g^A T_i^A + \sum_{j=1}^{N} p_{ij}^A v_j^A \qquad 1 \leq i \leq N \qquad (9.1)$$

$$v_N = 0$$

can be summarized as

$$M^A w^A = q^A \qquad (9.2)$$

where

$$M_{ij}^A = \begin{cases} \delta_{ij} - p_{ij}^A & 1 \leq i \leq N \quad 1 \leq j \leq N-1 \\ T_i & 1 \leq i \leq N \quad j = N \end{cases} \qquad (9.3)$$

$$w_i^A = \begin{cases} v_i^A - v_N^A & 1 \leq i \leq N-1 \\ g^A & i = N \end{cases} \qquad (9.4)$$

The solution to (9.2) can be written as

$$w_i^A = \sum_{j=1}^{N} (M^A)_{ij}^{-1} q_j^A \qquad 1 \leq i \leq N \qquad (9.5)$$

for which the gain g^A and relative values $v_i^A - v_N^A$ may be extracted. As we have shown in Theorem 5.9, M^{A-1} exists if $\lambda = 1$ has geometric multiplicity 1 for P^A, i.e., if P^A has one subchain.

In particular, if we compare

$$g^A = \sum_{j=1}^{N} (M^A)^{-1}_{Nj} q_j^A$$

with

$$g^A = \frac{(\pi^A, q^A)}{(\pi^A, T^A)}$$

and pick off the coefficients of the q's, we obtain

$$(M^A)^{-1}_{Nj} = \frac{\pi_j^A}{(\pi, T)} \qquad 1 \leq j \leq N$$

Consequently, if $(M^A)^{-1}$ is known, the π's can be obtained as a by-product via

$$\pi_i^A = \frac{(M^A)^{-1}_{Ni}}{\sum_{j=1}^{N} (M^A)^{-1}_{Nj}} \qquad 1 \leq i \leq N \qquad (9.6)$$

Suppose policy B differs from policy A in only one state, say state α. Let

$$\delta q = q_\alpha^B - q_\alpha^A \qquad (9.7)$$

$$x_i = \begin{cases} -p_{\alpha i}^B + p_{\alpha i}^A & 1 \leq i \leq N-1 \\ T_\alpha^B - T_\alpha^A & i = N \end{cases} \qquad (9.8)$$

Then

$$q_i^B - q_i^A = \begin{cases} 0 & i \neq \alpha \\ \delta q & i = \alpha \end{cases} \tag{9.9}$$

$$p_{ij}^B - p_{ij}^A = \begin{cases} 0 & i \neq \alpha \quad 1 \leq j \leq N \\ -x_i & i = \alpha \quad 1 \leq j \leq N-1 \\ \sum_{\kappa=1}^{N-1} x_\kappa & i = \alpha \quad j = N \end{cases} \tag{9.10}$$

$$T_i^B - T_i^A = \begin{cases} 0 & i \neq \alpha \\ x_N & i = \alpha \end{cases} \tag{9.11}$$

The change $\delta M = M^B - M^A$ in the M matrix is given by

$$(\delta M)_{ij} = \begin{cases} 0 & i \neq \alpha \quad 1 \leq j \leq N \\ x_j & i = \alpha \quad 1 \leq j \leq N \end{cases} \tag{9.12}$$

Let $U = (\delta M)(M^A)^{-1}$ (9.13)

δM, hence U, vanishes except for its α^{th} row, so that

$$U^2 = U_{\alpha\alpha} U$$

$$(I + U)^{-1} = I - \frac{U}{1 + U_{\alpha\alpha}} \tag{9.14}$$

Then

$$(M^B)^{-1} = (M^A + \delta M)^{-1} = \left((I + U)M^A\right)^{-1}$$

$$= (M^A)^{-1}(I + U)^{-1} = (M^A)^{-I} - \frac{(M^A)^{-1}U}{1 + U_{\alpha\alpha}} \tag{9.15}$$

$$(M^B)^{-1}_{ij} = (M^A)^{-1}_{ij} - \frac{(M^A)^{-1}_{i\alpha}\, U_{\alpha j}}{1 + U_{\alpha\alpha}} \qquad 1 \leq i,\ j \leq N$$

Equation (9.15) shows how M^{-1} is updated if the policy changes in state α.

According to (9.5) we have

$$w^B = (M^B)^{-1} q^B = \left[(M^A)^{-1} - \frac{(M^A)^{-1} U}{1 + U_{\alpha\alpha}}\right]\left[q^B - q^A + q^A\right]$$

so that

$$w_i^B = w_i^A = \frac{(M^A)_{i\alpha}}{1 + U_{\alpha\alpha}}\,\delta q_\alpha - \sum_{j=1}^{N} U_{\alpha j}\, q_j^A \qquad (9.16)$$

$$1 \leq i \leq N$$

Equation (9.16) shows how the gain and relative values are updated if the policy changes in state α.

We note that if $(M^B)^{-1}$ and $(M^A)^{-1}$ are assumed to exist, then $(I - U)^{-1}$ must exist, so that $U_{\alpha\alpha} \neq -1$.

9C. The Proposed Algorithm for Policy Improvement One State Per Cycle

We propose a new form of policy iteration, defined as follows:

(a) policy improvement is the same as ever, except that only one state is improved per iteration

(b) The value-determination operation for the new policy is accomplished by (9.15-9.16), and not by solution of simultaneous equations.

The flow diagram is given in Figure 2.

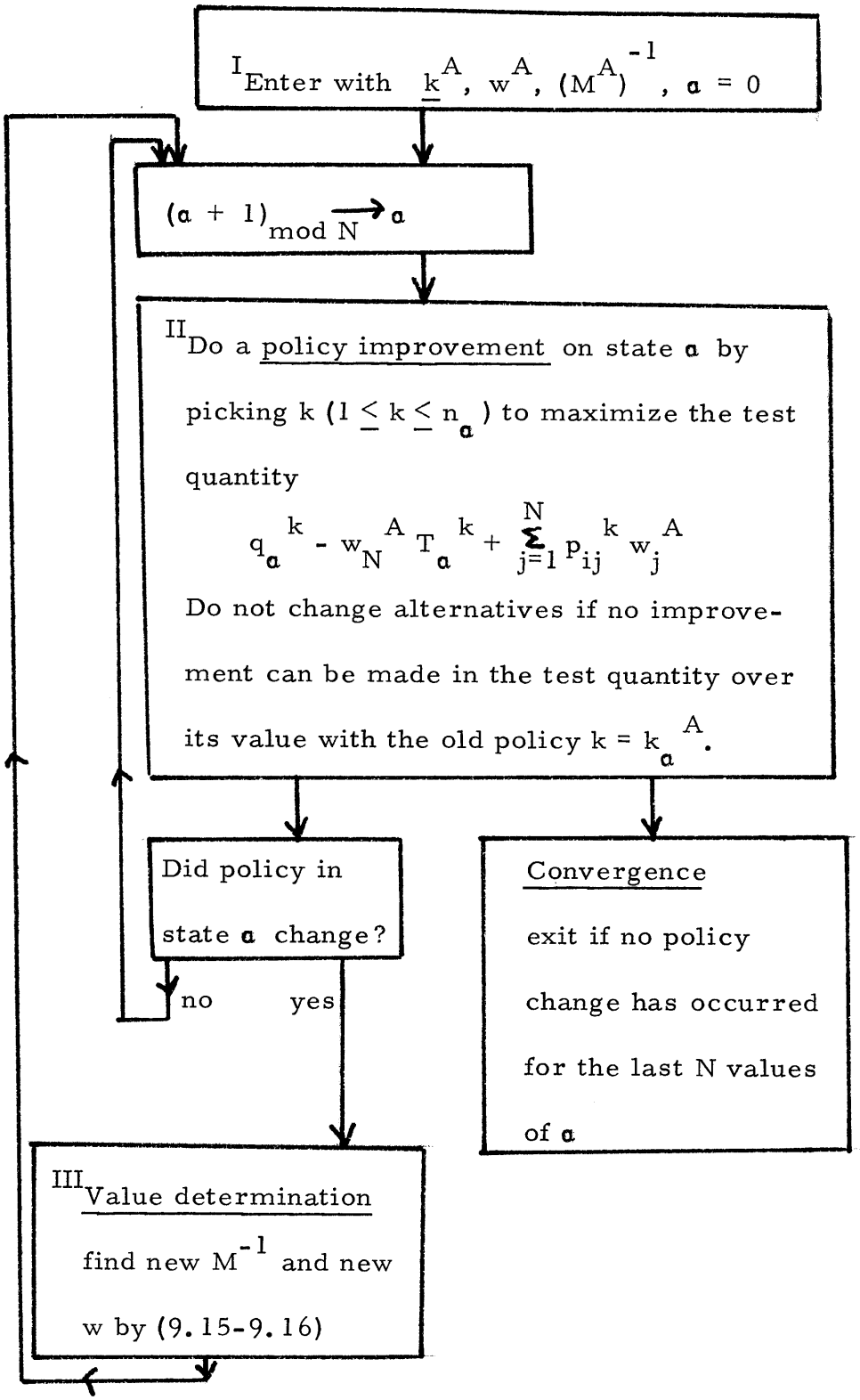

Figure 2. The Policy Iteration Algorithm Changing One Alternative Per Iteration.

The test quantity for policy improvement is

$$q_a^k - g^A T_a^k + \sum_{j=1}^{N-1} p_{ij}^k (v_j^A - v_N^A)$$

Note that the value determination operation is bypassed if there is no policy improvement in state a. The proof of convergence of this algorithm parallels the proof in Theorem 8.1.

The most important feature about this algorithm is that no simultaneous equations need be solved. Instead (9.15-9.16) are used for value evaluation.

No time penalty is paid for this scheme, which updates the values after policy improvement in each state. The reason is that the value updating by (9.15-9.16) takes $\propto N^2$ computations. A pass through all N states therefore takes $\propto N^3$ computations, which is comparable to the time spent in solving N simultaneous linear equations in the usual method where improvements are made in all N states. Provided tape handling time is equal, both methods require comparable time for value determination.

9D. Initialization

The vector gain and values w^A and inverse matrix $(M^A)^{-1}$ of the initial policy must be available (box I) before the algorithm in Figure 2 can be used. These can be determined as follows.

We start the computer program with the artificial initial policy A given by

$$q_i^A = 0 \qquad 1 \le i \le N \tag{9.17}$$

$$T_i^A = 1 \qquad 1 \le i \le N \tag{9.18}$$

$$p_{ij}^A = \delta_{j,N} \qquad 1 \le i, j \le N \tag{9.19}$$

State N is a trapping state so that p^A is ergodic. Then

$$M_{ij}^A = \begin{cases} \delta_{ij} & 1 \le i \le N \quad 1 \le j \le N-1 \\ -1 & 1 \le i \le N \quad j=N \end{cases} \tag{9.20}$$

$$(M^A)^{-1} = M^A \tag{9.21}$$

It follows from (9.17) that

$$g^A = 0 \tag{9.22}$$

and then, since $\underline{q}^A$ is proportional to $\underline{T}^A$, v^A is constant. Since $v_N^A = 0$,

$$v_i^A = 0 \qquad 1 \le i \le N \tag{9.23}$$

Equations (9.22) and (9.23) together state that

$$w_i^A = 0 \qquad 1 \leq i \leq N \qquad (9.24)$$

Thus equations (9.21) and (9.24) give the w vector and M^{-1} matrix for the initial policy.

We write a subprogram POLCH, mnemonic for "policy change", which is called with a desired policy B $(k_1^B, \ldots, K_N^B)$ where $1 \leq k_i^B \leq n_i$ for $1 \leq i \leq N$. POLCH will go through N passes. The r^{th} pass replaces the r^{th} row of p_{ij} and the r^{th} elements of q_i and T_i for the initial policy A given by (9.17-9.21) by those of alternative k_r^B and then calls the value-determination routine (box III) to update M^{-1} and w. The r^{th} pass therefore replaces the A policy in state r by the B policy. Each pass of POLCH is a <u>specified</u> policy change, followed by a policy evaluation, as opposed to a policy <u>improvement</u>, followed by a policy evaluation.

When all N passes of POLCH are completed, the A policy has disappeared completely and the B policy has taken its place. w^B and $(M^B)^{-1}$ are now available.

The POLCH subroutine can be used in two ways. First it is useful for initialization. Suppose a feasible policy B is selected, for example, which, in each state i maximizes the immediate expected reward q_i^k or the immediate expected reward rate q_i^k/T_i^k. If POLCH is called with the specified policy B, then w^B and $(M^B)^{-1}$ are computed. These can then be used to enter box I of Figure 2.

Second, the POLCH is useful to prevent round-off errors which would otherwise degrade the accuracy of w and M^{-1} if the algorithm in Figure 2 went through many cycles. The idea is to use POLCH, after every 20 or 30 iterations of the algorithm, to evaluate w and M^{-1} from scratch by calling it with k^B set to the present policy. Then we return to the algorithm.

9E. Storage Requirements

We will show in this section that roughly 5N cells of core storage are needed for the above algorithm. This implies that problems involving up to 4000 or 5000 states can be solved on a 32,000 word digital computer.

We assume that two auxiliary tapes are available for storage. Tape 1 contains all the transition data for the NN policies, stored with the state variable i varying mostly slowly and alternative variable k varying from 1 to n_i. For fixed i and K the N+2 numbers p_{ij}^k, q_i^k, T_i^k are stored with j going from 1 to N. Tape 2 contains the inverse matrix M^{-1} for the current policy, stored by rows, then the policy vector $\underline{k}$ for the current policy, finally the $\underline{q}$ vector for the current policy.

The subprograms for II, III and POLCH will be written as separate links so that their storage requirements can be discussed independently. Since POLCH does hardly anything more than call III N times, its storage requirements need not be discussed.

Box II, the policy improvement technique, needs roughly 4N cells of core storage. These would comprise

(i) N cells for $\underline{w}^A$

(ii) N+2 cells for $p_{\alpha j}^k$, q_α^k, T_α^k read in from tape 1 one K at a time

(iii) N+2 cells for $p_{\alpha j}$, q_j, T_j for the best alternative found so far, as k is increasing

(iv) N+2 cells for storage of $p_{\alpha j}{}^{A}$, $q_j{}^{A}$, T_j once k reaches $k_\alpha{}^{A}$

After the best alternative $k_\alpha{}^{B}$ is located, N+2 cells are needed to hold $\delta M_{\alpha j}$, $q_\alpha{}^{B}-q_\alpha{}^{A}$, $T_\alpha{}^{B}-T_\alpha{}^{A}$. These can be either ii, iii, or iv. The new k_α is put onto Tape 2.

Box III, the policy evaluation technique, proceeds in two passes. The first pass needs 5N cells of storage which are used as follows.

(i) N+2 cells for $\delta M_{\alpha j}$, $q_\alpha{}^{B}-q_\alpha{}^{A}$, $T_\alpha{}^{B}-T_\alpha{}^{A}$ which have been computed in Box II.

Box III will read in $(M^{A})^{-1}$ one row at a time. Hence

(ii) N cells for storing $(M^{A})^{-1}_{ij}$ $1 \leq j \leq N$ (i fixed). The i^{th} row of $(M^{A})^{-1}$ is used to compute a partial contribution to all N of the numbers U_ℓ

$$U_\ell = \left[\delta M\,(M^{A})^{-1}\right]_{\alpha\ell} = U_{\alpha\ell} \tag{9.25}$$

(iii) N cells for storing $\underline{U}$

(iv) N cells for storing $Y_i = (M^{A^{-1}})_{i\alpha}$ (9.26)

The second pass reads the $\underline{w}^{A}$ and $\underline{q}^{A}$ vectors from tape 2 into the N cells $\delta M_{\alpha j}$ in (i), and the N cells in (ii), and computes the scalar d,

$$d = q_\alpha{}^{B} - q_\alpha{}^{A} - \sum_{j=1}^{N} U_j q_j{}^{A}$$

It then computes the new w vector by (9.16):

$$w_i^B = w_i^A + \frac{Y_i d}{1 + U_{\alpha\alpha}}$$

and stores it back on tape 2. The $\underline{q}$ vector on tape 2 is updated via the addition of $q_\alpha^B - q_\alpha^A$ to its α^{th} component.

The second pass then reads $(M^A)^{-1}$ from tape 2 into core, one row i at a time--these N numbers can be put into (i)--and computes the new i^{th} row of M^{-1} by (9.15):

$$(M^B)^{-1}_{ij} = (M^A)^{-1}_{ij} - \frac{Y_i U_j}{1 + U_\alpha} \qquad 1 \leq j \leq N$$

These are returned to tape 2. The quantities

$$C_i = (M^B)^{-1}_{Ni}$$

are saved in (ii) during this updating of M^{-1} and are used after pass 2 is otherwise completed, to compute the steady state vector π^B for the new policy by (9.6):

$$\pi_i^B = \frac{C_i}{\sum_{j=1}^{N} C_j}$$

These are printed and discarded, hence require no storage.

9F. Conclusion

A modified form of the policy iteration algorithm has been presented in which only one state has its policy improved per cycle. The convergence properties of the new algorithm do not differ from the old, and the computation time is expected to be comparable.

However, the new version has several computational advantages rendering it far superior to the original algorithm. The solution of simultaneous equations, infeasible due to storage and accuracy limitations for large systems, has been eliminated. Storage limitations no longer exist: problems with 4000 or 5000 states may be treated, just as in linear programming. Accuracy limitations no longer exist: periodic appeal to POLCH eliminates roundoff errors.

A similar treatment holds for the discounted case as well as the undiscounted one, but will not be presented here.

CHAPTER 10

VALUE-ITERATION

10A. Introduction

In Chapter 2 we made the assumption that if all NN policies are ergodic, then the solution to

$$v_i(n+1) = \max_{1 \le k \le n_i} \left[q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j(n) \right] \tag{10.1}$$

has the asymptotic form

$$v_i(n) \to ng^* + v_i^* \tag{10.2}$$

In this chapter we prove that this is so and also show that the policy converges for large n to the optimal stationary policy. This justifies the restriction in the Howard-Jewell policy-iteration algorithm to stationary policies.

The technical problems associated with these proofs are connected with the Π operator discussed in 10C. Its major properties are given by (10.17-10.18) and Theorems 10.2 and 10.3.

In the case where the NN policies are not all ergodic but where it is merely known that each of them has a single irreducible set of states, equation (10.2) fails to hold. However $\|e(n)\| \le \|e(0)\|$ for all n still remains true, so that

$$\| v(n) - ng^* - v^* \| = \|e(n)\| \le \|e(0)\|$$

is finite for all n. Thus g^* is still a sort of asymptotic gain rate and the restriction to stationary policies a reasonable one for the policy-iteration algorithm.[33]

The L_∞ norm is used in this chapter.

10B. The Error Vector

If (10.2) holds, then g* and the v_i^* satisfy the functional equation

$$v_i^* = \max_{1 \le k \le n_i} \left(q_i^k - g^* + \sum_{j=1}^{N} p_{ij}^k v_j^* \right) \tag{10.3}$$

But we know from theorem (8.3) that if all NN policies are ergodic, (10.3) has a solution with g* unique and $\underline{v}^*$ unique up to an arbitrary additive multiple of $\underline{1}$. If this additive multiple is fixed, so v* is fixed, we can define the error vector

$$e_i(n) = v_i(n) - ng^* - v_i^* \qquad 1 \le i \le N \tag{10.4}$$

which is completely specified once $\underline{v}(0)$ is given.

We will justify our assumption by proving the following convergence theorem:

Theorem 10.1

Let all NN policies be ergodic. Then

(i) $\underline{L} = \lim_{n \to \infty} \underline{e}(n)$ exists

(ii) L_i is independent of i $\quad 1 \le i \le N$

Before proceeding to the proof of the theorem, some preliminary results must be presented. The L_∞ norm will be used in this chapter.

By insertion of

$$\underline{v}(n) = ng^* \underline{1} + \underline{v}^* + \underline{e}(n) \tag{10.5}$$

into (10.1), the iteration equation for $\underline{e}$ is obtained:

$$e_i(n+1) = \max_{1 \le k \le n_i} \left[b_i^k + \sum_{j=1}^{N} p_{ij}^k e_j(n) \right] \quad 1 \le i \le N \tag{10.6}$$

where

$$b_i^k = q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j^* - g^* - v_i^* \tag{10.7}$$

According to(10.3),

$$\max_{1 \le k \le n_i} \left[b_i^k \right] = 0 \qquad 1 \le i \le N \tag{10.8}$$

We will need the notation (see 8.46)

$$K_i^* = \left\{ k \mid b_i^k = 0 \right\} \qquad 1 \le i \le N \tag{10.9}$$

so that any policy $\underline{k} = (k_1, \ldots k_N)$ with $k_i \in K_i^*$ is a solution to (10.3).

10C. The T Operator

The T operator is a non-linear operator T: $E_N \rightarrow E_N$

defined by

$$(Tf)_i = \max_{1 \le k \le n_i} \left(b_i^k + \sum_{j=1}^{N} p_{ij}^k f_j \right) \qquad 1 \le i \le N \tag{10.10}$$

and

$$\max_{1 \le k \le n_i} b_i^k = 0 \qquad 1 \le i \le N \tag{10.11}$$

We summarize (10.6) as $e(n+1) = Te(n)$, or $e(n) = T^n e(0)$.

The T operator has the following properties:

$$(Tf)_{max} \le f_{max} \tag{10.12}$$

$$(Tf)_{min} \le f_{min} \tag{10.13}$$

if $\underline{f} \le \underline{g}$, then $T\underline{f} \le T\underline{g}$ (10.14)

$$T(c\,\underline{1}) = c\,\underline{1} \qquad \text{any scalar } c \tag{10.15}$$

$$T(\underline{f} + c\,\underline{1}) = T\underline{f} + c\,\underline{1} \qquad \text{any scalar } c \tag{10.16}$$

$$\| Tf \| \le \| f \| \tag{10.17}$$

$$\| Tf - Th \| \le \| f - h \| \tag{10.18}$$

Equations (10.12-10.14) are proved by insertion of $f_j \le f_{max}$, $f_j \ge f_{min}$ and $f_j \le g_j$, respectively, into (10.10). Equations (10.15-10.16) follow from $P\underline{1} = \underline{1}$. To prove (10.17) we insert $f_j \le \| f \|$, $f_j \ge - \| f \|$ into (10.10) and use (10.11) to obtain

$$- \| f \| \le (Tf)_i \le \| f \|$$

Equation (10.18) is obtained by use of the inequation

$$\min (A - B) \leq \max A - \max B \leq \max (A-B) \qquad (10.19)$$

where all maxima and minima go over the same set, so that

$$\min_{1\leq k\leq n_i} \sum_{j=1}^{N} p_{ij}^{k}(f_j-h_j) \leq Tf_i-Th_i \leq \max_{1\leq k\leq n_i} \sum_{j=1}^{N} p_{ij}^{k}(f_j-h_j)$$

Insertion of $f_j-h_j \leq \| f - h \|$ on the right and $f_j-h_j \geq - \| f - h \|$ on the left produces the desired result,

$$- \| f - h \| \leq Tf_i - Th_i \leq \| f - h \|$$

Equation (10.17) shows that T is a bounded operator, with operator norm of 1, while (10.18) shows that T is continuous.

The operator also satisfies the following fixed point theorem.

Theorem 10.2

Let T be defined by (10.10-10.11), with all NN policies ergodic. Then $Tf = f$ if and only if $\underline{f} = c\underline{1}$.

Proof

a) If $\underline{f} = c\underline{1}$, then (10.15) shows that $T\underline{f} = \underline{f}$.

b) Suppose $Tf = f$. Either $f_{max} = f_{min}$, and the theorem is proved, or $f_{max} > f_{min}$. In the latter case define the two disjoint sets of states A and B,

$$A = \{ i \mid f_i = f_{min} \} \qquad B = \{ i \mid f_i = f_{max} \}$$

Since $f_i \leq f_{max}$ and $b_i^k \leq 0$, it follows from (10.10) that if $i \in B$, then the policy k_i^{**} which achieves the maximization in $(Tf)_i$ has the properties $b_i^{k_i} = 0$ and $p_{ij}^{k_i**} = 0$ unless $j \in B$. For otherwise $f_{max} = (Tf)_i < f_{max}$.

Consequently any (p_{ij}) in which alternative k_i^{**} is chosen in all states $i \in B$ has B as a closed set of states.

On the other hand, for any $i \in A$, the choice $k_i \in K_i^*$ has $b_i^{k_i} = 0$ so that

$$f_{min} = f_i = Tf_i \geq \sum_{j=1}^{N} p_{ij}^{k_i} f_j \geq f_{min}$$

The last inequality must be an equality, whence $p_{ij}^{k_i}$ is zero unless $j \in A$.

Consequently any (p_{ij}) in which alternative $k_i \in K_i^*$ is chosen in all $i \in A$ has A as a closed set of states.

The policy $\underline{k} = (k_1, \ldots k_N)$ with k_i chosen as described above for $i \in A$ and $i \in B$, and k_i arbitrary for other i, has A and B as two disjoint closed sets of states. This contradicts the assumption that all NN policies are ergodic. Thus $f_{min} < f_{max}$ is impossible. QED

Corollary

Theorem 10.2 holds without the ergodicity assumption. It suffices to assume that each of the NN pure policies has a unique left eigenvector π with eigenvalue 1, i.e., only one subchain.

Proof

The only modification needed in the above proof occurs in the last paragraph. If A and B are disjoint closed sets of states, then each has a π vector and the assumption is violated. QED

10D. Proof of Theorem 10.1

Theorem (10.1) can be restated in the following form.

Theorem 10.3

Let $e(n+1) = Te(n) = T^{n+1}e(0)$ where T is given by (10.10-10.11) and where all NN policies are ergodic. Let $\|e(0)\| < \infty$. Then $\underline{L} = \lim_{n \to \infty} \underline{e}(n)$ exists.

L_i is independent of i $\quad 1 \leq i \leq N$

Proof

(i) By use of (10.17) we have

$$\|e(n+1)\| = \|Te(n)\| \leq \|e(n)\|$$

$$\|e(n)\| \leq \|e(n-1)\| \leq \ldots \quad \|e(1)\| \leq \|e(0)\| < \infty$$

so that each of the e(n), considered as a point in E^N, lies in a cube of side $2\|e(0)\|$ centered at the origin.

Invoking the Bolzano-Weierstrass theorem, $\underline{e}(n)$ possesses at least one cluster point $\underline{c}$. That is, there is a sequence $\{n_\ell\}$ and a function $N_0(\delta)$ such that

$$\|e(n_\ell) - c\| \leq \delta \qquad \text{all } \ell \geq N_0(\delta) \tag{10.20}$$

Invoke (10.18) to obtain

$$\|e(n_\ell + 1) - Tc\| \leq \|e(n_\ell) - c\| \leq \delta \qquad \text{all } \ell \geq N_0(\delta)$$

so that if c is a cluster point, so is Tc. By induction, so is $T^n c$ for any $n \geq 0$.

(ii) Let $m_i = \liminf e_i(n)$, $M_i = \limsup e_i(n)$ $\quad 1 \leq i \leq N$

These are finite since $|e_i(n)| \leq \|e(0)\|$. Then

$$m_{min} \leq m_i \leq c_i \leq M_i \leq M_{max} \tag{10.21}$$

We will now show that

$$c_{max} = M_{max} \tag{10.22}$$

$$c_{min} = M_{min} \tag{10.23}$$

$$e(n_\ell + u) = T^u e(n_\ell)$$

$$e(n_\ell) \leq c + \delta \underline{1} \qquad \ell \geq N_0(\delta)$$

$$e_i(n_\ell + u) \leq (T^u c)_i + \delta \leq c_{max} + \delta \qquad \ell \geq N_0(\delta)$$

These are infinitely many u's for which $e_i(n_\ell + u) \geq M_i - \delta$, so that $M_i \leq c_{max} + 2\delta$. The $\delta \geq 0$ is arbitrary and may be deleted, with the result $M_{max} \leq c_{max}$. But (10.21) implies that $c_{max} \leq M_{max}$. This completes the proof of (10.22).

$$e(n_\ell + u) = T^u e(n_\ell)$$

$$e(n_\ell) \geq c - \delta \underline{1} \qquad \ell \geq N_0(\delta)$$

$$e_i(n_\ell + u) \geq (T^u c)_i - \delta \geq c_{min} - \delta \qquad \ell \geq N_0(\delta)$$

There are infinitely many u's for which

$$e_i(n_\ell + u) \leq m_i + \delta, \quad \text{so that} \quad m_i \geq c_{min} - 2\delta$$

The $\delta \geq 0$ is arbitrary and may be deleted with the result $m_{min} \geq c_{min}$. Comparison with the conclusion from (10.21) that $m_{min} \leq c_{min}$ yields (10.23).

Equations (10.22) and (10.23) hold whenever c is a cluster point of e(n). Replacement of c by $T^n c$ yields

$$(T^n c)_{min} = m_{min} \; ; \; (T^n c)_{max} = M_{max} \qquad n \geq 0 \tag{10.24}$$

(iii) Let $\underline{k}^* = (k^*_1, \ldots k^*_N)$ by any policy $k_i^* \in K_i^*$ so that $b_i^{k_i^*} = 0$. Let R* denote its positively recurrent chain

$$R^* = \{ i \mid \pi_i^* > 0 \}$$

where π^* is the limiting distribution of the (by assumption ergodic) Markov chain $P^* = (p_{ij}^*) = (p_{ij}^{k_i^*})$. We will now show that if $\underline{c}$ is any cluster point of $\underline{e}(n)$, then $c_i = m_{min}$ for all $i \in R^*$.

$$\text{For } \ell \geq N_0(\delta), \quad e(n_\ell) \geq c - \delta \underline{1} \tag{10.25}$$

$$e_i(n_\ell + u) = (T^u e(n_\ell))_i \geq (T^u c)_i - \delta \qquad 1 \leq i \leq N$$

If we picked not the optimizing policy in Tc, $T^2 c$, $T^u c$ but rather k*, then we would have

$$(T^u c)_i \geq \sum_{j=1}^{N} (P^*)_{ij}^u \; c_j$$

Set $u = n_s - n_\ell \qquad s > \ell > N_0(\delta)$ so that

$$c_i + \delta \geq e_i(n_s) \geq \sum_{j=1}^{N} (P^*)_{ij}^u \; c_j - \delta$$

Let $s \to \infty$ so that $u \to \infty$ and $(P^*)^u_{ij} \to \pi^*_j$. The above equation becomes

$$c_i + \delta \geq \sum_{j=1}^{N} \pi^*_j c_j - \delta$$

The $\delta \geq 0$ is arbitrary and may be deleted, with the result

$$c_i \geq (\pi^*, c)$$

Multiply by $\pi^*_i \geq 0$ and sum on i to obtain $(\pi^*, c) \geq (\pi^*, c)$. This must be an equality, whence

$$c_i = (\pi^*, c) \qquad i \in R^*$$

$$c_i \geq (\pi^*, c) \qquad 1 \leq i \leq N$$

$m_{min} = c_{min} = (\pi^*, c)$.

This proves (10.25).

(iv) Since $\underline{c}$ is any cluster point of $\underline{e}(n)$, the c in (10.25) can be replaced by $T^n c$ with the result

$$(T^n c)_i = m_{min} \qquad i \in R^* \quad n \geq 0 \qquad (10.26)$$

The relation

$$c_i - \delta \leq e_i(n_\ell) \leq c_i + \delta \qquad \ell \geq N_0(\delta) \quad 1 \leq i \leq N \ .$$

becomes, after operation with T^u,

$$(T^n c)_i - \delta \leq e_i(n_\ell + u) \leq (T^n c)_i + \delta \qquad \ell \geq N_0(\delta) \quad 1 \leq i \leq N$$

By appealing to (10.26), the above becomes

$$m_{min} - \delta \leq e_i(n_\ell + u) \leq m_{min} + \delta \qquad \ell \geq N_0(\delta) \qquad i \in R^*$$

or

$$\lim_{n \to \infty} e_i(n) = m_{min} \qquad i \in R^* \tag{10.27}$$

This in turn implies

$$m_{min} = m_i = M_i \qquad i \in R^* \tag{10.28}$$

(v) There are now two cases to be considered, $m_{min} < M_{max}$ and $m_{min} = M_{max}$.

(vi) Suppose $m_{min} < M_{max}$. We know from (10.28) that $i \in R^*$ implies $m_i = M_i = m_{min}$. Also, any policy p_{ij} which uses k_i^* in state i, for all $i \in R^*$, has R^* as a closed set of states, for $i \in R^*$ implies $p_{ij}^{k_i^*}$ is zero unless $j \in R^*$.

For all n sufficiently large, $e_j(n) \leq M_j + \delta$. This holds for all j since there are only finitely many (N) values of j. Then

$$e_i(n+1) = \max_{1 \leq k \leq n_i} \left[b_i^k + \sum_{j=1}^{N} p_{ij}^k \, e_j(n) \right] \leq \max_{1 \leq k \leq n_i} \left[b_i^k + \sum_{j=1}^{N} p_{ij}^k \, M_j \right] + \delta$$

$$1 \leq i \leq N$$

Let k_i^{**} be the alternative in state i which maximizes the far right hand side of the above equation. If n is chosen cleverly,

$e_i(n+1) \geq M_i - \delta$ so that, after deleting the arbitrary $\delta \geq 0$,

$$M_i \leq b_i^{k_i **} + \sum_{j=1}^{N} p_{ij}^{**} M_j \leq \sum_{j=1}^{N} p_{ij}^{**} M_j$$

Iteration of $M \leq P^{**} M$ yields

$$M_i \leq \left[(P^{**})^n M\right]_i \rightarrow \sum_{j=1}^{N} \pi_j^{**} M_j = (\pi^{**}, M)$$

and dotting this with π^{**} yields

$$(\pi^{**}, M) \leq (\pi^{**}, M), \text{ so}$$

$$M_i = (\pi^{**}, M) \qquad i \in R^{**} = \left\{ j \mid \pi_j^{**} > 0 \right\}$$

$$M_i \leq (\pi^{**}, M) \qquad 1 \leq i \leq N$$

so that

$$(\pi^{**}, M) = M_{max} \tag{10.29}$$

$$M_i = M_{max} \qquad i \in R^{**} \tag{10.30}$$

According to (10.28) and (10.30) and the assumption $m_{min} < M_{max}$, R^* and R^{**} are disjoint. We can imagine a policy $P = p_{ij}^{k_i}$ which uses the alternatives

$$k_i = \begin{cases} k_i^{**} & i \in R^{**} \\ k_i^{*} & i \in R^{*} \\ \text{arbitrary elsewhere} \end{cases}$$

This policy will have two closed sets of states, R* and R** and contradicts the assumption that all policies are ergodic. Hence the case $m_{min} < M_{max}$ is impossible.

(vii) The remaining case $m_{min} = M_{max}$ must hold. According to (10.21), $m_i = M_i$ for $1 \leq i \leq N$ so that

$$\lim_{n \to \infty} e_i(n) = m_{min} \qquad 1 \leq i \leq N$$

Since there are only N values of i, this convergence can be made uniform in i with the result

$$\| \underline{e}(n) - m_{min} \underline{1} \| \to 0 \qquad \text{as } n \to \infty$$

This completes the proof of theorem 10.3. QED

10E. Policy Convergence

Theorem 10.3 shows that the values converge, i.e., that (10.2) holds. In addition, we can prove convergence of the policy as well.

Let $N_0(\delta)$ be the function such that

$$\left| e_i(n) - m_{min} \right| \leq \delta \qquad n \geq N_0(\delta) \quad 1 \leq i \leq N \tag{10.31}$$

Let K_i^* be defined by (10.9) and let

$$d_i = \max_{1 \leq k \leq n_i} \left[b_i^k \right] < 0 \tag{10.32}$$

$$k \notin K_i^*$$

Then the following theorem can be proved.

Theorem 10.4

Let $k_i(n)$ be (one of) the alternatives which achieves the maximum in

$$e_i(n+1) = \max_{1 \leq k \leq n_i} \left[b_i^k + \sum_{j=1}^{N} p_{ij}^k e_j(n) \right] \tag{10.33}$$

Let all NN policies be ergodic.

Then for any i, $1 \leq i \leq N$,

$$k_i(n) \in K_i^* \qquad \text{if } n \geq N_0(\delta) \tag{10.34}$$

for any δ satisfying $0 < \delta < -d_i/2$.

Proof

If all NN policies are ergodic, (10.31) holds. Then

$$e_i(n+1) = b_i^{k_i(n)} + \sum_{j=1}^{N} p_{ij}^{k_i(n)} \, e_j(n) \qquad 1 \leq i \leq N$$

Pick $n \geq N_0(\delta)$ so that

$$e_j(n) \leq m_{min} + \delta \qquad 1 \leq j \leq N$$

Then $m_{min} - \delta \leq e_i(n+1)$ and consequently,

$$-2\delta \leq b_i^{k_i(n)} \qquad 1 \leq i \leq N$$

If δ satisfies (10.34) for some i,

$$\max_{k \notin K_i^*} b_i^k = d_i < b_i^{k_i(n)}$$

whence $k_i(n) \in K_i^*$. QED

10F. Remarks On Theorem 10.3

a) It follows from (10.2) that if all NN policies are ergodic, then

$$\lim_{n\to\infty} \frac{v_i(n)}{n} = g^* \qquad \text{uniformly in } i$$

Consequently g*, defined as the maximum gain rate of the NN stationary policies, is identical with the gain rate $\lim_n v(n)/n$ of the optimal time dependent policy. An even stronger statement is that $\| \underline{v}(n) - ng^*\underline{1} \| < B$ so that $\underline{v}(n)$ and $ng^*\underline{1}$ never differ by more than a finite amount. These results show that the gain rate of the optimal time-dependent policy is achieved by a stationary policy, and justifies our self-imposed restriction of consideration of only stationary policies.

b) The actual value of $\lim_{n\to\infty} e_i(n)$ is very difficult to find, since it involves the transient buildup of the policy and values to their steady-state values. Three general remarks which can be made, however, are the following.

First if $\underline{v}(0) = \underline{v}^* + c\,\underline{1}$, then $\underline{e}(0) = \underline{e}(n) = c\,\underline{1}$ for all n. Stated verbally, if the scrap values happen to be identical with the relative values of the optimal policy, then the system begins and remains in the steady-state:

$$k_i(n) \in K_i^* \qquad 1 \le i \le N \qquad n \ge 0$$

Second, if $\underline{v}(0)$ is changed to $\underline{v}(0) + c\,\underline{1}$, then $\underline{e}(n)$ is changed to $\underline{e}(n) + c\,\underline{1}$ $n = 0, 1, 2, \ldots$. Therefore, adding a constant to all the scrap values merely shifts lim v(n) by that same constant.

Third, $\underline{L} = \underline{L}(\underline{\alpha}) = \lim T^n \underline{\alpha}$ is a continuous function of the scrap value $\underline{e}(0) = \underline{\alpha}$. This follows from

$$\| T^{n+1}\underline{\alpha}' - T^{n+1}\underline{\alpha} \| \leq \| T^n \underline{\alpha}' - T^n \underline{\alpha} \| \leq \ldots \leq \| \underline{\alpha}' - \underline{\alpha} \|$$

for all n, so that $\| L(\underline{\alpha}') - L(\underline{\alpha}) \| \leq \| \underline{\alpha}' - \underline{\alpha} \|$.

c) The relations

$$\| e(n+1) \| \leq \| e(n) \|$$

$$\| e(m+1) - e(n+1) \| \leq \| e(m) - e(n) \|$$

$$e(n+1)_{max} \leq e(n)_{max}$$

$$e(n+1)_{min} \geq e(n)_{min}$$

for the error $e(n) = T^n e(0)$, all of which follow from (10.12-10.18), are not sufficient to prove convergence of $\underline{e}(n)$. A counter-example is the case N = 2, $\underline{e}(n) = \left[(-1)^n, (-1)^{n+1} \right]$.

Apparently ergodicity or something similar is needed to guarantee convergence of $\underline{e}(n)$.

The case N = 2, $n_i = 1$ for i = 1, 2, given by (5.86) is another example where the possession of a single irreducible closed set of states is not sufficient for $\underline{v}(n) - n\underline{g}^* - \underline{v}^*$ to converge. Again ergodicity seems to be needed.

10G. White's Value-Iteration Scheme

The value iteration scheme

$$v_i(n+1) = \max_{1 \le k \le n_i} \left[q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j(n) \right] \qquad 1 \le i \le N \qquad (10.35)$$

is ill-suited for numerical computation since $\underline{v}(n)$ diverges linearly in n. White has devised [17] a modification of the scheme, which overcomes the divergence by subtracting from $v_i(n)$ another term which grows like g*n.

White uses the new variables

$$w_i(n) = v_i(n) - v_N(n-1) \qquad 1 \le i \le N \qquad (10.36)$$

If all NN policies are ergodic, we know that

$$v_i(n) \rightarrow ng^* + v_i^* + c \qquad 1 \le i \le N \qquad (10.37)$$

uniformly in i

Consequently,

$$w_i(n) \rightarrow v_i^* - v_N^* + g^* \qquad 1 \le i \le N \qquad (10.38)$$

uniformly in i

The $w_i(n)$ remain bounded for all n, and approach the limits given on the right side of (10.38). Thus the gain g* and relative values $v_i^* - v_N^*$ of the optimal stationary policy may be obtained as

$$g^* = w_N(\infty) \qquad (10.39)$$

$$v_i^* - v_N^* = w_i(\infty) - g^* \qquad 1 \leq i \leq N-1 \tag{10.40}$$

If (10.36) is inserted into (10.35), we find that the w's satisfy the iterative equation

$$w_i(n+1) = \max_{1 \leq k \leq n_i} \left[q_i^k + \sum_{j=1}^{N} p_{ij}^k \, w_j(n) \right] - w_N(n) \tag{10.41}$$

$$1 \leq i \leq N \qquad n \geq 0$$

This completes the proof of

Theorem 10.5

Let all NN pure policies be ergodic. Let $w_i(n)$ be given by (10.41), with $w_i(0) = v_i(0)$ (actually any choice of w(0) will do, since v_i^* and g^* are independent of scrap values.) Then

a) $w_i(n)$ remain bounded for all n.

b) $\lim_{n \to \infty} w_i(n) = w_i(\infty)$ exists for all i, $1 \leq i \leq N$. This convergence is uniform in i.

c) The numbers v_i^*, g^* which satisfy (10.3) may be found via (10.39) and (10.40).

d) If the initial choice $w_i(0) = v_i(0)$ all i is used, then

(i) $v_i(n)$ can be recovered from $w_i(n)$ via

$$v_N(n) = \sum_{\ell=1}^{n} w_N(\ell) + v_N(0) \qquad n \geq 1 \qquad (10.42)$$

$$v_i(n) = w_i(n) + v_N(n-1) \qquad n \geq 1 \quad 1 \leq i \leq N-1 \qquad (10.43)$$

(ii) The $w_i(n)$ can be recovered from the $v_i(n)$ via (10.36).

(iii) The alternatives $k_i(n)$ which achieve the maximum on the right hand side of (10.35) are identical with the alternatives which achieve the maximum on the right hand side of (10.41).

Corollary

Let $(P^{k_1 \ldots k_u})_{ij}$ denote the probability, conditioned on starting from i and making decisions $k_1, \ldots k_u$, that the system is found in state j after u transitions.

Suppose there exists a state m, a number α, $0 < \alpha \ll 1$, and an integer $u_0 \geq 0$ such that for all states i and for all $k_1, k_2, \ldots k_{u_0+1}$,

$$(P^{k_1 \ldots k_{u_0+1}})_{im} \geq \alpha > 0 \qquad (10.44)$$

Then the $w_i(n)$ given by (10.41) converge uniformly to $w_i(\infty)$. If g* and $v_i^* - v_N^*$ are given by (10.39) and (10.40), then they satisfy (10.3).

Proof

If $k_1 = k_2 = \ldots = k_{n_0+1}$, we see from (10.44) that every pure policy has state m as a recurrent state. Hence each of the NN policies

has only one closed set of states, since m is accessible from all states. Equation (10.44) also implies that the NN pure policies are aperiodic. The NN pure policies, being a periodic and possessing only one closed set of states, must each be ergodic. Then invoke Theorem 10.5. QED

Discussion

White proved only the above corollary. He used (10.44) to show that the convergence of $w_i(n)$ is geometric.

The assumption (10.44) is far more restrictive than necessary to guarantee uniform convergence of $w_i(n)$. It is a rather gross way of guaranteeing ergodicity of the NN pure policies, at which point Theorem 10.5 can be invoked. However, it has the advantage of ensuring that the convergence is geometric.

10H. Value Iteration For the Semi-Markov Case

We could try to use the notation of error operator to find the limiting behavior of $\underline{v}(t)$ given by

$$v_i(t) = \max_{1 \le k \le n_i} \left[q_i^k(t) + \sum_{j=1}^{N} \int_0^t dt' \, p_{ij}^k(t') \, v_j(t-t') \right] \tag{10.45}$$

$$t > 0 \quad 1 \le i \le N$$

If we insert into (10.44) the equation

$$v_i(t) = g^* t + v_i^* + e_i(t) \tag{10.46}$$

where g^* and v_i^* satisfy the functional equation

$$v_i^* = \max_{1 \le k \le n_i} \left[q_i^k(\infty) - g^* T_i^k + \sum_{j=1}^{N} p_{ij}^k \, v_j^* \right] \tag{10.47}$$

$$1 \le i \le N$$

then we would like to show that $e_i(t) \to c$ (independent of i) as $t \to \infty$. Unfortunately the proof in Theorem 10.3 no longer goes through as before. We cannot show that $e_i(t)$ is uniformly bounded, much less convergent.

Consequently, no proof exists for the convergence of value-iteration in the semi-Markov case. That the method of Theorem 10.3 breaks down is not completely unexpected, for (10.47) involves merely

the limiting immediate rewards $q_i^k(\infty)$ while we know from Theorem 5.10 that the absolute values, at least for a fixed policy, are finite only if the areas η_i^k under curves $q_i^k(t) - q_i^k(\infty)$ are finite.

10I. Value-Iteration for Multi-Chain Processes

Results similar to Theorems 10.3 and 10.4 for value convergence and policy convergence can be obtained in the multi-chained case, provided one assumes all NN pure policies are stable. These have been omitted due to the lengthy arguments required for their proof.

APPENDIX A

NOTATION

State Space and Vectors

N-state case: state space is $\Omega = (1,2,\text{--}N)$ $N < \infty$

continuum case: state space is Ω, a bounded subset of some finite dimensional Euclidean space, with finite volume

$$V = \int_{\Omega} dx < \infty \tag{A1}$$

Vectors f or $\underline{f}$ have components f_i or $f(x)$

Use the L_2 norm for length of a vector,

$$\|f\| = \begin{cases} \sqrt{\sum_{i=1}^{N} |f_i|^2} \\ \sqrt{\int_{\Omega} dx\, |f(x)|^2} \end{cases} \tag{A2}$$

and use the scalar product

$$(f,g) = \begin{cases} \sum_{i=1}^{N} f_i^* g_i \\ \int_{\Omega} dx\, f^*(x) g(x) \end{cases} \tag{A3}$$

with the properties

$$\|a \underline{f}\| = a \|f\| \qquad \text{scalar } a \tag{A4}$$

$$\|f\| \geq 0$$

$$\|f\| = 0 \quad \text{if and only if} \quad \begin{cases} f_i = 0 & 1 \leq i \leq N \\ f(x) = 0 & \text{almost all } x \end{cases} \tag{A5}$$

$$\|f+g\| \leq \|f\| + \|g\| \tag{A6}$$

$$|(f,g)| \leq \|f\| \, \|g\| \qquad \text{Schwarz inequality} \tag{A7}$$

We will usually assume f is $L_2 \cdot I_n$ particularly this implies that $f(x)$ is bounded for almost all $x \in \Omega$.

Linear Operators

Define the linear operator A

$$A_{ij} \qquad 1 \leq i,j \leq N$$

$$A(x,y) \qquad x,y \in \Omega$$

by

$$(Af)_i = \sum_{j=1}^{N} A_{ij} f_j \qquad (fA)_i = \sum_{j=1}^{N} f_j^* A_{ji}$$

$$(Af)(x) = \int_\Omega dy \; A(x,y) f(y) \qquad (fA)(x) = \int_\Omega dy \; f^*(y) A(yx)$$

We consider only operators with finite L_2 norm:

$$\|A\| = \begin{cases} \sqrt{\sum_{ij=1}^{N} |A_{ij}|^2} \\ \sqrt{\int_\Omega dx \int_\Omega dy \; |A(x,y)|^2} \end{cases}$$

$\|A\| < \infty$ "A is a bounded operator"

$$(AB)_{ij} = \sum_{k=1}^{N} A_{ik} B_{kj}$$

$$(AB)(x,y) = \int_{\Omega} dw\, A(x,w) B(w,y)$$

$$A^{n+1} = A A^n = A^q A^{n+1-q} \qquad 0 \le q \le n+1$$

If A and B are bounded, then so are cA, A+B, and AB:

$$\|cA\| = |c| \, \|A\|$$

$$\|A+B\| \le \|A\| + \|B\|$$

$$\|AB\| \le \|A\| \, \|B\|$$

If f and A are both bounded, so is Af:

$$\|Af\| \le \|A\| \, \|f\|$$

Thus (Af)(x) is an L_2 vector. Also, if A is an L_2 operator, then for almost all y, A(xy) is square-integrable in x. Thus given an equation like

$$\underline{v} = Z(\underline{q} - g\underline{T})$$

where q and T are assumed to be L_2, the square integrability of P and P^∞ guarantees the square integrability of $Z - I = [I - P + P^\infty]^{-1} [P - P^\infty]$ and thus the square integrability of $\underline{v}$. Then v(x) will be finite for almost all x.

A is called a kernel of <u>finite rank</u> if

$$A(x,y) = \sum_{i=1}^{r} f_i(x)\, g_i(y)$$

almost all x,y where the integer r is finite. If r_o is the smallest such integer, that is if $f_1 - f_{r_o}$ are linear independent and $g_1 - g_{r_o}$ are linearly independent, then we say A is of rank r_o.

Equalities, Inequalities, and Limits

$A = B$ means $\|A - B\| = 0$, Thus $A_{ij} = B_{ij}$ for all $1 \le i,j \le N$ and $A(x,y) = B(x,y)$ for almost all x and y. Similarly $\underline{f} = \underline{g}$ means $\|\underline{f} - \underline{g}\| = 0$. $\underline{f} \le \underline{g}$ means $f_i \le g_i$ for $1 \le i \le N$ or $f(x) \le g(x)$ for almost all x. Similarly $A \le B$ means $A_{ij} \le B_{ij}$ and $A(x,y) \le B(x,y)$ for all ij and for almost all x,y. Thus $\underline{f} \ge 0$ means $f(x) \ge 0$ for all x but a set of zero measure.

$$\lim_{n\to\infty} A_n = B \quad \text{means} \quad \lim_{n\to\infty} \|A_n - B\| = 0$$

$$\lim_{n\to\infty} \underline{f}_n = \underline{g} \quad \text{means} \quad \lim_{n\to\infty} \|\underline{f}_n - \underline{g}\| = 0$$

These are so-called "convergence in norm" conditions.

Identity Operator and 1 Vector

Identity operator I, $I\underline{f} = \underline{f}$, has components $\delta_{ij} = I_{ij} = 1$ if $i = j$ and 0 otherwise. Formally $I(x,y) = \delta(x-y)$.

The one vector $\underline{1}$ has components

$$1_i = 1 \qquad 1 \le i \le N$$

$$1(x) = 1 \qquad x \in \Omega$$

The requirement (A1) guarantees that $\underline{1}$ is an L_2 vector:

$$\|1\| = \begin{cases} \sqrt{N} \\ \sqrt{V} \end{cases} < \infty$$

The requirement that $\underline{1}$ be L_2 is a necessity if we insist that the Markov operator P be L_2 . For $\underline{1}$ is a right eigenvector of P, $P\underline{1} = \underline{1}$ and all eigenvectors of a L_2 kernel P must be L_2 .

Markov Operators and Distributions

P is a Markov operator if $P \ge 0$ and $P\underline{1} = \underline{1}$. Written out, these say

$$P_{ij} \ge 0 \qquad \sum_{j=1}^{N} P_{ij} = 1 \qquad 1 \le i \le N$$

$$p(x,y) \ge 0 \qquad \int_{\Omega} dy \; P(x,y) = 1 \qquad x \in \Omega$$

The quantity P_{ij} is the probability of reaching state j given a transition out of state i. Similarly P(x,y) is the probability density of the terminal state y given a transition out of x.

The real vector $\underline{f}$ is called a distribution if

$$f \ge 0 \qquad (f,1) = 1$$

That is, $f_i \geq 0 \quad 1 \leq i \leq N \quad$ and $\sum_{i=1}^{N} f_i = 1$

or $f(x) \geq 0$

and $\int_{\Omega} dx f(x) = 1$. We note that

$$|(f,1)| \leq \|f\| \, \|1\|$$

so that if f has finite L_2 norm, it is absolutely integrable, i.e. L_1.

The symbol P always stands for an L_2 Markov operator. The symbol π always stands for a real, usually non-negative left eigenvector of P with eigenvalue 1

$$\pi \geq 0 \qquad \pi P = \pi$$

It is usually normalized to $(\pi, 1) = 1$. The normalization can always be carried out if π is L_2, for

$$|(\pi,1)| \leq \|\pi\| \, \|1\| < \infty$$

So that if π is L_2 it is also absolutely integrable. This is another reason why we insist on (A1).

If $P^{\infty}_{ij} = \pi_j$ or if $P^{\infty}(x,y) = \pi(y)$

then

$$\|P^{\infty}\| = \|\pi\| \, \|1\|$$

Hence P^{∞} will be L_2 since $\underline{1}$ and π both are.

The L_∞ Norm

We will occasionally use the L_∞ for a vector, defined by

$$\|f\| = \begin{cases} \max\limits_{1 \le i \le N} |f_i| \\ \sup\limits_{x \in \Omega} |f(x)| \end{cases}$$

Each chapter states explicitly whether the L_2 or L_∞ norm is used. Equation (A4-A6) hold for the L_∞ norm.

We also use the notation

$$f_{max} = \max_{1 \le i \le N} f_i \qquad f_{min} = \min_{1 \le i \le N} f_i$$

so that

$$f_{min} \le f_i \le f_{max}.$$

APPENDIX B

ELEMENTS OF THE FREDHOLM THEORY OF L_2 OPERATORS [22,28]

Let K by an L_2 linear operator. If $(K - \lambda I)^{-1}$ exists, then K is called regular at λ. If the inverse fails to exist, K is called singular at λ, and λ is called an eigenvalue of K. It can be shown that $\lambda \neq 0$ is an eigenvalue of K if and only if there exists at least one L_2 vector e such that $Ke = \lambda e$. Also $\lambda \neq 0$ is an eigenvalue of K if and only if there exists at least one L_2 vector f such that $fK = \lambda f$. The vectors e and f are called right and left eigenvectors of K with eigenvalue λ.

The number of linearly independent left eigenvectors of K with eigenvalue $\lambda \neq 0$ agree with the number of linearly independent right eigenvectors. This number is called the geometric multiplicity of λ. It is possible to show that the geometric multiplicity of $\lambda \neq 0$ is finite:

$$\text{geometric multiplicity of eigenvalue } \lambda \neq 0 \leq \frac{\|K\|^2}{|\lambda|^2} \tag{B1}$$

This bound on the geometric multiplicity in terms of the L_2 norm on K can be proved by noting that if $f_1, \ldots f_m$ are any m linearly independent right eigenvectors of K, $Pf_i = \lambda f_i \quad 1 \leq i \leq m$, then they can be made

orthonormal by the Gram-Schmidt procedure:

$$(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij} \qquad 1 \leq i,\ j \leq m$$

Finally expansion of

$$0 \leq \iint dxdy \left| \frac{K(x,y)}{\lambda} - \sum_{i=1}^{m} \mathbf{e}_i(x)\mathbf{e}_i^*(y) \right|^2$$

$$= \frac{\|K\|^2}{|\lambda|^2} - 2\,\mathrm{re} \sum_{i=1}^{m} \frac{(\mathbf{e}_i, K\mathbf{e}_i)}{\lambda} + \sum_{i,j=1}^{m} (\mathbf{e}_i, \mathbf{e}_j)^2$$

$$= \frac{\|K\|^2}{|\lambda|^2} - 2\,\mathrm{re} \sum_{i=1}^{m} \delta_{ii} + \sum_{i,j=1}^{m} (\delta_{ij})^2$$

yields (B1).

It also turns out that $\lambda \neq 0$ is an eigenvalue of K if and only if $1/\lambda$ is a zero of the <u>modified Fredholm determinant</u> $\delta(z)$ defined by

$$\delta(z) = \delta(z, K) = e^{-f(z)} \tag{B2}$$

where

$$f(z) = f(z; K) = \int_0^z du \sum_{n=1}^{\infty} u^n \, \mathrm{tr}(K^{n+1}) \tag{B3}$$

$$\mathrm{tr}\, A = \begin{cases} \displaystyle\sum_{i=1}^{N} A_{ii} & \text{N-state case} \\ \displaystyle\int_{\Omega} dx\ A(x,x) & \text{continuum case} \end{cases} \tag{B4}$$

Since $|\mathrm{tr}(AB)| \leq \|A\| \, \|B\|$ if A and B are both L_2, $|\mathrm{tr}(K^{n+1})| \leq \|K\|^{n+1}$. It follows that f(z), and therefore $\delta(z)$, are analytic functions of z for $|z| < 1/\|K\|$. Moreover, $\delta(z)$ turns out to be an entire function of z, analytic everywhere.

In the N-state case where K is an NxN matrix, the Fredholm determinant is closely related to det (I - zK), hence its connection with the eigenvalue spectrum. To see this, we recall that if $\lambda_1, \ldots \lambda_N$ are the N eigenvalues of K, then

$$\mathrm{tr}(K^m) = \sum_{i=1}^{N} (\lambda_i)^m \qquad m = 1, 2, 3, \ldots$$

so that

$$f(z) = -z \, \mathrm{tr}(K) - \sum_{i=1}^{N} \ln (1 - z\lambda_i)$$

$$\delta(z) = e^{z \, \mathrm{tr}(K)} \prod_{i=1}^{N} (1 - z\lambda_i)$$

$$\delta(z) = e^{z \, \mathrm{tr}(K)} \det (I - zK) \tag{B5}$$

If $\lambda \neq 0$ is an eigenvalue of K, then the order of the zero of $\delta(z)$ at $z = 1/\lambda$ is called the <u>algebraic multiplicity</u> of λ. Since the order of the zeros of a not-identically-zero analytic function are of finite order, the algebraic multiplicity is always of finite order. Furthermore, the geometric multiplicity is less than or equal to the

algebraic multiplicity. This fact is well-known in the N-state case, where $\det(K - zI)$ may have a r-fold zero at $z = \lambda$ yet possess only s, $1 \leq s \leq r$, left and right eigenvectors with eigenvalue λ.

We note that if K can be decomposed into the sum of two orthogonal operators,

$$K = A + B \qquad AB = BA = 0$$

then

$$\delta(z;\ A + B) = \delta(z; A)\, \delta(z; B) \tag{B6}$$

This can be proved by noting that

$$K^n = A^n + B^n \quad ; \quad \mathrm{tr}(K^n) = \mathrm{tr}(A^n) + \mathrm{tr}(B^n)$$

$$f(z;\ A + B) = f(z;A) + f(z;B)$$

from which (B6) follows.

Equation (B6) shows that the eigenvalue spectrum of K (the reciprocals of the zeros of $\delta(z;\ A + B)$) consists of the eigenvalue spectrum of A plus that of B. This could also be deduced via

$$\begin{aligned}(I - zK)^{-1} &= (I - zA)(I - zB)^{-1} \\ &= (I - zB)^{-1}(I - zA)^{-1} \\ &= (I - zB)^{-1} + (I - zA)^{-1} - I \end{aligned} \tag{B7}$$

where the identity

$$(I - C)^{-1} = I + (I - C)^{-1}C = I + C(I - C)^{-1}$$

has been used with C = A and with C = B.

Since the number of zeros of a not-identically zero analytic function like $\delta(z)$ in any finite region of the z-plane is finite, we conclude that K has only finitely many eigenvalues λ in any region of the λ-plane which excludes the origin. Each of the eigenvalues $\lambda \neq 0$ has been shown to have a finite geometric multiplicity and algebraic multiplicity. The eigenvalues are denumerably infinite. The only possible cluster point for the eigenvalues of K is the origin. Since the zeros of an analytic function are isolated, the non-zero eigenvalues of K are isolated.

The spectral radius ρ of K is defined as the magnitude of the largest eigenvalue of K:

$$\rho = \max_{\lambda} |\lambda| = \frac{1}{\min_{z} |z|} \tag{B8}$$

where the maximization is taken over all eigenvalues of K, and the minimization over all zeros of $\delta(z)$. It exists, is finite, and turns out to be given by

$$\rho = \lim_{n \to \infty} \|K^n\|^{1/n} \tag{B9}$$

All eigenvalues λ of K must lie within the circle $|\lambda| \leq \rho$ and furthermore any annulus $0 < a \leq |\lambda| \leq b$ contains only finitely many eigenvalues.

From (B9) follows the theorem that if the spectral radius of K is less than unity, namely if all eigenvalues of K are strictly

less than 1 in magnitude, then $\| K^n \| \to 0$.

Theorem B1

If $\rho < 1$, then $\| K^n \| \to 0$ as $n \to \infty$.

Proof

According to (B9), if $\epsilon > 0$ is given, then

$$\left| \| K^n \|^{1/n} - \rho \right| < \epsilon$$

for $n \geq N_0 = N_0(\epsilon)$. Pick $\epsilon = (1-\rho)/2 > 0$ to get, for $n \geq N_0(\epsilon)$,

$$\| K^n \|^{1/n} \leq \rho + \epsilon = \frac{1+\rho}{2} < 1$$

$$\| K^n \| < \left(\frac{1+\rho}{2}\right)^n \to 0 \quad \text{as } n \to \infty. \qquad \text{QED}$$

Theorem B2

If the spectral radius of E is less than unity, then $(I-E)^{-1}$ exists and

$$(I - E)^{-1} = I + E + E^2 + E^3 + \ldots \qquad (B10)$$

in the sense of convergence in L_2 norm. Furthermore,

$(I - E)^{-1} - I$ is an L_2 operator.

Proof

According to Theorem B1, $\| E^n \| \to 0$. Pick m big enough that $\| E^m \| < 1$. Let

$$U = E + E^2 + \ldots + E^m \quad ; \qquad \| U \| < \infty$$

Then

$$\left\| \sum_{n=1}^{\infty} E^n \right\| = \left\| \sum_{j=0}^{\infty} E^{jm} U \right\|$$

$$\leq \|U\| \left(1 + \sum_{j=1}^{\infty} \|E^m\|^j\right)$$

$$= \frac{\|U\|}{1 - \|E^m\|} < \infty \qquad \text{QED}$$

APPENDIX C

SUGGESTIONS FOR FUTURE RESEARCH

(1) Investigate the spectral properties and group properties of the U^{AB} discussed in 6G.

(2) Investigate the group properties of the γ^{AB} and Γ^{AB} discussed in 7G.

(3) Investigate the role of U, in (6.3), as the generator of the process. What is the infinitesimal generator? How can finite perturbations be composed of infinitesimal ones?

(4) Extend Theorem 5.10 to the continuum case.

(5) Prove 2.14a, b. Extend Theorem 10.2 and 10.3 to the continuum case.

(6) Discuss the existence and uniqueness of the solution to the functional equation (8.26) in the continuum case. Extend all of Chapter 8 to the continuum case.

(7) Extend all of the above to the multi-chain case.

(8) Numerical studies as to how best quantize a continuum process. For example, if the family $f_1(x), \ldots f_n(x)$ is given then will the choice of $c_1, \ldots c_n$ obtained by minimizing the quadratic form in $n+1$ variables

$$h(c_1, c_2, \ldots c_n, g') = \left\| \sum_{i=1}^{N} c_i(P-I)f_i + q - g'T \right\|^2$$

yield an approximate solution to

$$v = q - gT + Pv$$

in the sense that

$$v(x) \simeq \sum_{i=1}^{n} c_i f_i(x)$$

$$g \simeq g' \qquad ?$$

If so, what is a convenient basis family for say inventory problems and how large should n be?

BIBLIOGRAPHY

1. W. S. Jewell, Markov-Renewal Programming I and II, Operations Research 11, 938-971 (1963).

2. R. Bellman, Dynamic Programming, Princeton University Press, Princeton, N. J. (1957).

3. B. O. Koopman, New Mathematical Methods in Operations Research 1, 3-9 (1952), see page 9.

4. R. Bellman, A. Markovian Decision Process, Journal of Mathematics and Mechanics 6, 679-684 (1957).

5. R. A. Howard, Dynamic Programming and Markov Processes, M.I.T. Technology Press and John Wiley & Sons (1960).

6. R. A. Howard, Semi-Markovian Control Systems, M.I.T. Operations Research Center Technical Report No. 3, December 1963. Also reported in "Semi-Markovian Decision Processes," Proceedings of the 34th session of the International Statistical Institute, Vol. XL, Book 2, Toronto, Canada, 1964, pp. 625-652.

7. J. S. de Cani, A Dynamic Programming Algorithm for Imbedded Markov Chains When the Planning Horizon is at Infinity, Management Science 10, 716-733 (1964).

8. A. S. Manne, Linear Programming and Sequential Decisions, Management Science 6, 259-267 (1960).

9. P. Wolfe and G. B. Dantzig, Linear Programming in a Markov Chain, Operations Research 10, 702-710 (1962).

10. H. Wagner, On the Optimality of Pure Strategies, Management Science 6,268-269 (1960).

11. D. Blackwell, Discrete Dynamic Programming, Ann. Math. Stat. 33, 719-726 (1962).

12. W. S. Jewell, Limiting Covariance in Markov-Renewal Processes, University of California - Berkeley Operations Research Center Report ORC 64-16(RR) 22 July 1964.

13. W. S. Jewell, private communication, April 1, 1965.

14. Bellman, Glicksberg and Gross, On the Optimal Inventory Equation, Management Science 2, 83-104 (1955).

15. D. L. Iglehart, Optimality of (s, S) Policies in the Infinite Horizon Dynamic Inventory Problem, Management Science 9, 259-267 (1962).

16. C. Derman and M. Klein, Some Remarks on Finite Horizon Markovian Decision Models, Operations Research 13, 272-278 (1965).

17. D. J. White, Dynamic Programming, Markov Chains and the Method of Successive Approximations, Journal of Mathematical Analysis and Applications 6, 373-376 (1963).

18. Cozzolino, Gonzales-Zubieta and Miller, Markovian Decision Processes with Uncertain Transition Probabilities, M.I.T. Operations Research Center Technical Report No. 11, March 1965.

19. J. J. Martin Jr., Some Bayesian Decision Problems in a Markov Chain, Ph.D. Thesis, M.I.T. Civil Engineering Dept., 1965 (unpublished).

20. Cox and Smith, Queues, Methuen and Co. Ltd., London, pp. 76-85.

21. K. Yosida and S. Kakutani, Operator - Theoretical Treatment of Markoff's Process and Mean Ergodic Theorem, Annals of Mathematics 42, 188-228 (1941), especially page 226.

22. A. C. Zaanen, Linear Analysis, Interscience, New York.

23. J. G. Kemeny and J. L. Snell, Finite Markov Chains, D. Van Nostrand, Princeton, New Jersey (1960).

24. J. G. Kemeny and J. L. Snell, On Markov Chain Potentials, Annals of Mathematics Stat. 32, 709-715 (1961).

25. J. G. Kemeny and J. L. Snell, Notes on Discrete Potential Theory, Journal of Math. Analysis and Applications 3, 117-121 (1961).

26. J. G. Kemeny and J. L. Snell, Potentials for Denumerable Markov Chains, Journal of Math. Analysis and Applications 3, 196-260 (1960).

27. J. G. Kemeny, A Further Note on Discrete Potential Theory, Journal of Math. Analysis and Applications 6, 55-57 (1963).

28. F. Reisz and B. Sz.-Nagy, Functional Analysis, Frederick Ungar Publishing Co., New York (1955).

29. Knopp, Theory of Functions I, Dover Publications, New York, Theorem 4, page 90.

30. A. Singer, The Steady State Probabilities of a Markov Chain as a Function of the Transition Probabilities, Operations Research 12, 498-499 (1964).

31. R. Bellman and S. E. Dreyfus, Applied Dynamic Programming, Princeton University Press, Princeton, New Jersey (1962).

32. I. M. Gelfand and S. V. Fomin, Calculus of Variations, Prentice-Hall, Englewood Cliffs, New Jersey (1963), pp. 11-13.

33. A. Maitra, Dynamic Programming for Countable Action Spaces. The author has seen only the abstract, in Annals of Math. Stat. 36, 735 (1965).

BIOGRAPHICAL NOTE

Paul Jerome Schweitzer was born in New York City, New York on January 16, 1941 and attended Sewanhaka Central High School in Floral Park, New York, graduating in June 1957. He was awarded a Faculty Scholarship and an M. I. T. Freshman Competitive Scholarship, and entered M. I. T. in September 1957. He received the Borden Prize as first-ranking M. I. T. freshman, and was on Dean's List all eight semesters.

Mr. Schweitzer received two Bachelor of Science degrees, in Physics and Mathematics in June 1961. His graduate studies were in the M. I. T. Physics Department, and a minor in Operations Research, and were supported by four National Science Foundation Predoctoral Fellowships.

Mr. Schweitzer worked summers at Airborne Instruments Laboratory, Deer Park, New York and at M. I. T. Lincoln Laboratory, Lexington, Massachusetts. He has served as consultant to both Airborne Instruments Laboratory and Lincoln Laboratory.

Mr. Schweitzer's papers include his bachelor's thesis, with Mr. Malvin Teich, "Determination of the Total Neutron Cross-Section of Palladium Using the M. I. T. Fast Chopper", submitted to the M. I. T.

Physics Department in June 1961, and Lincoln Laboratory Project Report PA-88, February 24, 1965, "Electromagnetic Scattering From Rotationally Symmetric Perfect Conductors".