ADAPTIVE CONTROL
OF
DISCRETE STATE MARKOV PROCESSES

John Spruce Riordon

A thesis submitted for the
degree of Doctor of Philosophy
in Electrical Engineering

February 1967

Centre for Computing and Automation,
Imperial College of Science and
Technology,
University of London.

TO MARSHA

ABSTRACT

An investigation is made in this thesis of the control of long duration stationary Markov processes. When the process is non-linear and/or the disturbance is non-gaussian the determination of an optimal feedback policy involves a more or less intractable variational problem. This may be avoided, however, by the quantization of both state and control as well as time, so that the process is modelled as a stochastic automaton.

A problem of particular interest is that in which the state is observable but the process dynamics and disturbance characteristics are initially unknown. The controller must then operate in an adaptive fashion, performing simultaneously the dual functions of estimation and control. A detailed study is made of the control of a repetitive Markovian (single-stage) batch process in which cost is a function of state only. A general theory of the dual control of such processes is developed, and a simple optimal dual control strategy is demonstrated.

For the more general problem in which both state and control are costed, Markovian decision theory can be used to determine a discretized optimal feedback policy for the automaton model of the process. A new iterative method of

determining the optimal continuous state feedback characteristic for processes which are themselves continuous is presented. Finally, a convergent dual control strategy is derived for the on-line multi-stage optimization of discrete long duration Markov processes. The resultant adaptive controller is quite general in nature, being able to handle initially unknown dynamics, state and control constraints, system nonlinearities, non-gaussian multiplicative disturbances of unknown distributions, and a wide variety of cost functions.

As examples of the application of these methods, several simulation studies are presented. These include control of a long sequence of batch processes, ordering of thermal-electric power generation, and control of a heat treatment process.

ACKNOWLEDGEMENTS

I should like to express my gratitude to my supervisor, Mr. D. Q. Mayne, who, through many useful discussions and constant encouragement, has greatly enhanced the progress of this work. My thanks go also to Professor J. H. Westcott, who helped me get started on the Markovian path. Much enlightenment has also been obtained from discussions with fellow research students in the Control Systems Group.

I wish to acknowledge the financial support of the Athlone Fellowships Managing Committee (Board of Trade) and of the National Research Council of Canada.

Finally, this work would not have been possible without the unfailing encouragement and cheerfulness of my wife, Marsha.

J. S. Riordon

CONTENTS

		Page
GLOSSARY OF PRINCIPAL SYMBOLS		12
CHAPTER 1:	SOME ASPECTS OF ADAPTIVE CONTROL	16
1.1	Introduction	16
1.2	The Discrete State Approach	19
1.3	Dual Control	24
1.4	Stochastic Qutomata in Adaptive and Learning Control Systems	25
1.5	Dual Control Requirements in Long Duration Processes	28
1.6	Outline of the Thesis	33
1.7	Contributions of the Thesis	40
CHAPTER 2:	MARKOV CHAINS AND CONTROL SYSTEMS	42
2.1	Introduction	42
2.2	Markov Chains	42
2.3	Markov Processes with Costs: The Markovian Decision Problem	48
2.4	Process Optimization	53
2.5	Example	58
2.6	Properties of the Transformation Matrix	64

		Page
2.7	Processes with Single-Stage Optimization	67
2.8	Process Models	73
2.9	Summary	75
CHAPTER 3:	A THEORY OF DUAL STRATEGIES SINGLE-STAGE MARKOV PROCESSES	79
3.1	The Concept of Dual Control	79
3.2	The Dual Control Requirement	83
3.3	Statistical Estimation of Process Parameters	86
3.4	Properties of Normal Likelihood Functions	91
3.5	A Realizability Problem	92
3.6	The Bayesian Approach	94
3.7	The Continuity Approximation	98
3.8	Decision Space and the Hill of Uncertainty	100
3.9	The Optimal Trajectory	101
3.10	Characteristic Vector of a Decision Process	107
3.11	The Meaning and Uses of the Characteristic Vector	108

		Page
3.12	The Inverse Problem	110
3.13	Suboptimal Trajectories	113
3.14	Realizable Control Strategies	115
3.15	Equal Rho Strategy	115
3.16	Probabilistic Strategy	120
3.17	Optimal Strategy	125
3.18	Proof of Convergence	131
3.19	Summary	134
CHAPTER 4:	COMPUTATIONAL METHODS AND RESULTS IN SINGLE-STAGE MARKOV PROCESSES	138
4.1	Introduction	138
4.2	Optimal Trajectories in (N-1) Space	139
4.3	Optimum Point in (N-1) Space	146
4.4	Doubly Constrained Optima	153
4.5	Simulated Results: A Three State System	156
4.6	Effect of Variance Estimate	167
4.7	An Example: A Fluid Mixing Process	169
4.8	Summary	179
CHAPTER 5:	ADAPTIVE ORDERING OF POWER GENERATION AS A CYCLIC MARKOV PROCESS	180
5.1	Introduction	180

		Page
5.2	The Ordering Problem	181
5.3	Optimization Technique	185
5.4	A Simulation Study: Ordering in a 2000 Mw Station	189
5.5	Adaptive Ordering	198
5.6	Conclusion	205
CHAPTER 6:	DUAL CONTROL OF MULTI-STAGE MARKOV PROCESSES	208
6.1	Introduction	208
6.2	A Multi-Stage Dual Control Algorithm	213
6.3	Updating the Estimates	218
6.4	The Optimal Feedback Transducer Characteristic	221
6.5	Adaptive Control of Continuous State Processes	226
6.6	Summary	232
CHAPTER 7:	SIMULATION RESULTS IN MULTI-STAGE MARKOV PROCESSES	234
7.1	Introduction	234
7.2	The Problem: A Heat Treatment Process	234
7.3	A Tentative Solution: A Priori Estimates	239

Page

7.4 Experimental Results: Simulated Adaptive Control 248

7.5 Summary 257

CHAPTER 8: POINTS OF DEPARTURE 259

8.1 Introduction 259

8.2 The Single-Stage Dual Strategy as a Multi-Modal Hill Climber 259

8.3 Further Investigation of the Multi-Stage Dual Strategy 260

8.4 The Economics of Generation Ordering 261

8.5 Automaton Model Relevance 262

8.6 Non-Stationary Processes 263

8.7 Processes with Uncertainty in State Measurement 263

8.8 Finite Duration Processes 264

8.9 Higher Order Processes 265

8.10 Theoretical Implications of the Discrete Formulation 268

APPENDIX 1: PROOF OF CONVERGENCE OF ITERATIVE COMPUTATION OF OPTIMAL DECISION MATRIX 272

		Page
APPENDIX 2:	"EXACT" COST ESTIMATES FROM MULTIDIMENSIONAL BETA DISTRIBUTIONS	277
APPENDIX 3:	PROPERTIES OF NORMAL LIKELIHOOD FUNCTIONS	285
APPENDIX 4:	TRANSFORMATIONS OF NOISE DISTRIBUTIONS	293
APPENDIX 5:	POWER DEMAND TRANSITION MATRICES	299
APPENDIX 6:	FORTRAN LISTINGS	302
REFERENCES		321

GLOSSARY OF PRINCIPAL SYMBOLS

B	N x L control cost matrix whose elements b_{ij} equal the cost of reaching decision state j, if the present process state is i.
C	L x N transition cost matrix, whose elements c_{ij} equal the cost associated with a probabilistic transition from decision state i to process state j.
D	N x L stochastic decision matrix whose elements d_{ij} equal the probability that if the process state is i, control will be exerted to reach decision state j.
D^*	the optimal decision matrix.
$\hat{D}^*$	the optimal decision matrix computed using the estimated transition matrix $\hat{P}$.
D(n)	decision matrix at interval n.
E(x)	expected value of random variable x.
$\underline{e}_i >$	column vector whose i^{th} element is unity, and whose other elements are zero.
$\underline{e}$	characteristic decision vector in N-1 space.
$f_i(x)$	the probability density function of the random variable x associated with process state i.
$F_i(x)$	cumulative probalility distribution of random variable x associated with process state i.

$G_i(x)$ $1 - F_i(x)$.

g the expected steady state cost per stage of a discrete ergodic Markov process with parameters B, C, P, D.

g^* the value of g in an optimally controlled process, i.e., one with parameters B, C, P, D^*.

$\underline{h}_i$ offset vector normal to i^{th} coordinate in decision space, joining the basic trajectory to a given suboptimal decision trajectory.

I identity matrix.

i,j,k state designations.

L total number of alternative control decisions.

$\underline{\ell}$ cost vector whose elements ℓ_i equal the expected cost of one stage of operation if the process state is initially i.

M L x N observation matrix whose elements m_{ij} equal the number of observed transitions from decision state i to process state j.

N number of process states.

n total number of stages of system operation observed.

n_i number of observed transitions from decision state i.

P L x N stochastic transition matrix whose elements p_{ij} equal the probability that if the present decision state is i, then the next process state will be j.

$\hat{P}$	estimated transition matrix formed from maximum likelihood estimates $\hat{p}_{ij}$.
$Q^{(i)}$	N x N probability covariance matrix of state i.
R	L x L compound transition matrix; R = PD.
s	state incurring minimum mean cost.
$\hat{s}$	state incurring minimum estimated mean cost.
u_{ij}	j^{th} control alternative available when process state is i.
u_i^*	the optimum control input when process state is i.
v_i	steady state cost difference between state i and state N in basic chain.
v_{ij}	steady state cost difference between decision state (i,j) (i.e. process state i with control alternative j) and decision state (N,s).
$\underline{w}>$	column vector all of whose elements are unity.
$\underline{z}$	N-vector whose first N-1 elements are v_i, i = 1, ... N-1, and whose N^{th} element is g.
Γ	the number of control alternatives available for each process state.
γ	convergence factor.
$\mathcal{S}_n$	disturbance signal at the n^{th} stage of operation.
η_{ij}	$b_{ij} + v_{ij}$
μ_i	mean cost of one transition from process state i when control is not costed.

$\hat{\mu}_i$	maximum likelihood estimate of μ_i
$\prod$	product operator
$\underline{\pi}$	row vector of steady state probalility of occupancy of process states.
ρ_{oi}	$\frac{\mu_i - \mu_s}{\sigma_{oi}}$
$\sum$	summation operator.
σ_{oi}^2	one sample variance of estimate of μ_i.
ϕ_i	Prob $[\mu_i = \underset{j}{\text{Min}}\{\mu_j\} \mid \hat{\mu}_1, \ldots \hat{\mu}_N]$
Ψ^{-1}	transformation matrix relating single-stage costs to multi-stage costs.
Ω	uncertainty; approximates ω
ω	the probability, condition upon past observations, that $\hat{D}^* \neq D.^*$

CHAPTER 1

SOME ASPECTS OF ADAPTIVE CONTROL

1.1 Introduction

Probably the most prominent factor influencing the development of control theory during the past decade has been the widespread use of the high speed automatic digital computer. The availability of enormous computational effort has had a profound qualitative, as well as quantitative, effect, allowing control systems engineers and theorists to deal with much more sophisticated problems than had formerly been considered. At the centre of the new technique is optimization theory, which displaces the question, "will this system work?" with the more significant one, "what is the best system?" By this change of approach the systems engineer hopes to relate the parameters of a controller to some well-defined cost criterion, and to arrive at a (mathematically) unique solution to the problem in hand.

Two broad classes of problem to which optimization techniques have been applied are:

1) the trajectory optimization problem: the

specification of a sequence of control actions relating to a dynamic system such that the expected value of a performance criterion over a known finite time interval is minimized.

2) the regulator problem: the specification of an optimal feedback policy or an optimal feedback transducer characteristic which minimizes the expected cost per unit time associated with the operation of a dynamic system over a long time interval.

In this thesis we shall treat the second problem. Specific examples, all of which will be considered later, include the control of a long sequence of batch processes, the ordering of thermal-electric power generation, and the sequential heat treatment of sections of a continuous metal slab.

The application of optimization techniques to regulator problems has yielded general and easily computed results in only a few cases, notably that of the well-known linear unconstrained system with quadratic costs and an additive gaussian disturbance. Many systems in the real world fall considerably short of this ideal. In fact the system dynamics are frequently more or less unknown. In such a case it is desirable to control in an adaptive fashion so that the control policy is modified as

information concerning the process becomes known. Much interest has grown up recently in the use of the digital computer as an adaptive controller, since it combines great speed with equal flexibility, and possesses in addition the ability to make decisions. The use of a digital computer as a control element in a closed loop system is termed "on-line control" or "direct digital control". Most of this thesis is concerned with on-line computer control, and in particular with its computational aspects.

However, it is not only uncertainty of process parameters which hinders the effective application of optimization techniques. Many processes are characterized by a variety of additional computationally ambarrassing attributes, amongst which are the following:

1) the process may be non-linear;

2) it may be stable, conditionally stable, or unstable;

3) state and control variables may be constrained;

4) disturbances may be present which are non-gaussian and multiplicative, with unknown probability distributions;

5) the nature of the problem may make the performance criterion mathematically awkward (e.g. absolute value functions).

It is the purpose of this thesis to develop practical

control algorithms, suitable for on-line use, which will handle all of the preceding factors simultaneously.

1.2 The Discrete State Approach

It need hardly be said that the optimization of a general continuous non-linear system disturbed by multiplicative noise presents a more or less intractable mathematical problem. To simplify matters, we shall assume that successive state transitions constitute a Markov process; that is,

$$p(x_{n+1} \mid x_n) = p(x_{n+1} \mid x_1, x_2, \ldots . x_n) \qquad (1.1)$$

where

$p(x_{n+1} \mid x_n)$ = probability density function of state variable x measured at the $(n+1)^{th}$ time interval, conditional upon the measured value of x at interval n.

$x_1, x_2, \ldots x_n$ = measurements of state variable x at all past time intervals 1, 2, ... n.

Moreover, by quantizing both state and control variables as well as time, we may describe state transitions in terms of a Markov chain. The overall system model is then a stochastic automation, i.e. a Markov chain whose transition

probabilities are dependent upon a set of control variables. The great advantage of this model is that transitions involved in a finite state stationary Markov chain are governed by a finite set of linear equations even if the process itself is non-linear and the disturbance non-gaussian.

If the process dynamics and the statistical distribution of the disturbance signal are known, then it is usually possible to derive a stochastic automaton model of the process. By a suitable specification of state and control costs, the problem may be cast as a Markovian decision process. Such processes have received extensive study, particularly in relation to operations research and economics problems[10-16]. Probably the best introduction to the field is given by Howard[14], who is responsible for the basic optimization algorithm. Important variants on the process include the imbedded Markov chain[15] in which transitions are considered to take place at random time intervals, and the discounted reward process[11,14] in which future income or cost associated with the process has a discounted present value.

The close relationship of Markovian decision processes to control theory has been pointed out by Bellman and Dreyfus[1] and Feldbaum[19], amongst others. Åström[38] has used the discrete Markovian framework to study systems in

which uncertainty is present in the measurement of state. He has shown that, if the system dynamics are known, the control problem may be decomposed into two separate parts: the estimation of present state from past observations, and the computation of an optimal control input given the present estimate. With the discrete formulation, the latter computation may be performed off-line, and the results stored; on-line control, even of a non-linear stochastic process, may thus be greatly simplified. A similar type of process has been studied by Kashyap[39], and has been mentioned by Lave[16] in connection with quality control problems.

In this thesis we shall consider processes whose state may be measured exactly but whose dynamics and disturbance statistics are initially unknown. The basic process is shown in fig. 1.1. The output variable, x_n, at the n^{th} interval of operation can be in one of N states i = 1, ... N (to avoid confusion we shall in future refer to x as the output variable, and reserve the term "state" to mean a quantized version of x). The process is subject to a disturbance $\xi_n(x_n, u_n)$ which may be dependent upon both output and control, e.g. multiplicative noise. The dynamics are therefore given by

$$x_{n+1} = f(x_n, u_n) + \xi_n(x_n, u_n) \qquad (1.2)$$

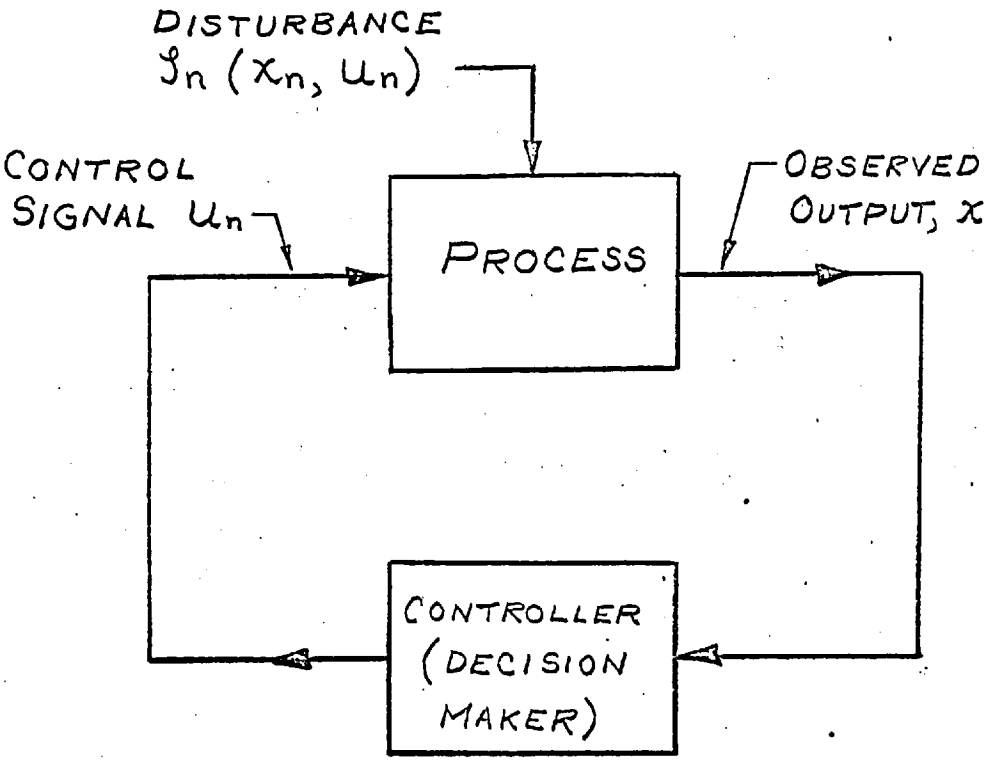


FIG. 1.1

CONTROL OF A

STOCHASTIC PROCESS

where u_n = control input at interval n

$f(x_n, u_n)$ = a function of output x_n and control u_n.

The object of control is to minimize the expected value of a performance criterion, g.[‡]

$$g = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{N} L(x_n, u_n) \qquad (1.3)$$

The parameter g is the expected cost per transition of the process, measured over a long interval, and $L(x_n, u_n)$ is a cost function associated with the process at interval n. If the parameters of (1.2) are known, then it is possible to compute an optimal control policy. By an optimal policy we mean the specification of a particular control input u_i[¤] for each state i so that g is minimized. If the process dynamics are initially unknown, it is necessary to implement some sort of on-line adaptive controller which combines the functions of estimation and control. The manner in which this combination is performed is determined by an adaptive strategy. Obviously one requirement of a successful strategy is that it must approach the optimal policy with time.[‡‡] A little thought suggests a second requirement: in doing so it must avoid, as much as possible, actions which are deemed a posteriori to have resulted in expensive

---

[‡] Equation (1.3) assumes that the process is ergodic.

[‡‡] This property will be referred to as convergence of the strategy.

operation. As an extreme example, it would be of little help to have an adaptive strategy which, by experimenting on a chemical process determined the optimal operating conditions, but destroyed the plant in so doing.

1.3 Dual Control

It is clear that a conflict arises between the requirement of convergence (a large number of trials using all combinations of control for estimation purposes), and that of minimum cost operation (trials using only the estimated optimal control policy and no other). The decision maker is called upon to explore the process while controlling it. Feldbaum[18,19] terms this operation "dual control". Although Feldbaum has provided a brilliant formal solution for discrete time, continuous state systems of finite duration, the nature of the problem makes its implementation exceedingly difficult. Sworder[21,22] has derived a dual control strategy for a linear system disturbed by Wiener noise with gaussian increments; Xirokostas and Henderson[24] have considered the dual control of a linear system preceded by a quadratic function. Related problems have been studied by Aoki[17], Rosenbrock[20], and Tou[23].

In a discrete state system with finite duration, dual strategies are feasible only when the process duration is

short (e.g. three or four time intervals), or the number of states is small (e.g., two). The reason for this is that the solution requires the study of a decision tree, each branch of which is associated with an a priori probability distribution. Silver[37] has considered some two and three state decision processes with a lifetime of three stages in this fashion. The problem has also been mentioned by Meier[33].

1.4 Stochastic Automata in Adaptive and Learning Control Systems

In stochastic processes of finite duration, dual control implies a feedback strategy which minimizes the total expected cost over a known number of intervals, despite uncertain a priori information. For processes of infinite duration the term is not well defined, however; minimization of a total is meaningless, while minimization of the mean cost per transition over an infinite period merely implies a convergent adaptive algorithm. The use of the latter criterion has given rise to a considerable variety of ingenious control structures whose merits are difficult to compare. The general term applied to the field is "learning systems". No agreed definition of

learning exists, but it is usually used to describe a wide group of extremum-seeking algorithms using the techniques of stochastic approximation, decision theory, Markov chains, and pattern recognition[29]. In this thesis no formal distinction will be made between learning and adaptation.

In discrete state processes of infinite duration, the convenient and general structure of the stochastic automaton has led to its widespread use, either as a model of the adaptive controller, or as a model of the process itself. The first algorithm suggested for the dual control of discrete state Markov processes is probably that of Pashkovskii.[35] He postulated an automaton in which the state transition probabilities, initially unknown, are affected by the choice of control signal from a discrete available set. The object of control is to determine through adaptive operation the policy which maximizes the probability, $p_{i\ell}$, of transition from each state, i, to a known state, ℓ. His solution, involving the estimation of confidence intervals for each transition probability, will be considered in more detail in section 3.15.

Basing their work on earlier research carried out by Soviet authors[25,30,31], McMurty and Fu[27] have considered a stochastic process in which the control mechanism is a reward or punishment scheme (an evaluator gives a signal 0

if the control action had a "good" result, and a signal 1 for a "bad" result) together with a simple deterministic rule for changing the control input, depending upon the last evaluation. They model a multimodal hill as a star-shaped automaton, and have shown that if one and only one arm of the star (one mode) yields a probability of punishment less than 0.5, then the system will eventually settle in that arm. In this and a subsequent paper[28] they have also treated a case in which the controller itself is probabilistic; the regions of a multimodal hill which are searched are determined by a probabilistic choice, the probabilities being updated by a linear reinforcement technique. The simple one-zero evaluator is replaced in the latter case with an index of performance evaluator (giving the height of the hill at the point searched). McLaren[26] has studied a similar type of problem, using both linear and non-linear learning reinforcement.

Another similar model, designed for on-line adaptive control of a discrete state dynamic system, has been considered by Nikolic and Fu[34]. The subjective probabilities of applying one of a discrete set of control actions are modified after each plant observation, and a randomized control strategy is used. The probability modifications depend upon whether or not the last control input used was

the estimated optimum one, and whether or not its use resulted in a decrease of the posterior estimate of the conditional mean index of performance. Convergence to the optimal policy is proved; i.e. as the number of observed transitions increases, the probability of the correct control choice being made approaches unity.

1.5 Dual Control Requirements in Long Duration Processes

It has been remarked that the diversity of adaptive control schemes for discrete state processes of long duration, and the difficulty of quantitative comparison, is partially due to the absence of a mathematical definition of the dual control requirement for such processes. Once such a definition is made, it may be incorporated into the problem as an additional constraint, thereby reducing the present need for a somewhat heuristic approach.

We shall begin by considering the concept of convergence in a dual control system. Using a decision theoretic approach, we may say that a control strategy has converged when the probability of error (the probability that the estimated optimal policy is not the true optimal policy) is zero. Moreover the error probability is a measure of convergence in that the lower it is, the closer the strategy

is to convergence. Let us now consider an ensemble of statistically identical but independent stochastic processes which are stationary and ergodic. The statistical parameters are initially unknown, and an identical learning controller is used on each process to attempt the minimization of a known performance criterion. A hypothetical plot of ensemble mean cost of the n^{th} transition, $g(n)$, and of the error probability, $\omega(n)$, is shown in fig. 1.2. Suppose that the expected cost per transition with the true optimal policy is g^{*}. Then convergence implies that for each individual process

$$\lim_{n\to\infty} \omega(n) = 0 \tag{1.4}$$

$$\lim_{n\to\infty} g(n) = g^{*} \tag{1.5}$$

Many early learning schemes, such as those of references 25 and 27 result in strategies which "dither" about the correct control input and so are not convergent in the foregoing sense. The strategies of Pashkovskii[35], McLaren[26], McMurty and Fu[28], and Nikolic and Fu[34] are all convergent, however. In none of these papers is $\omega(n)$ actually evaluated. Instead, conditions equivalent to (1.4) and (1.5) are demonstrated.

To specify the additional requirement of dual control,

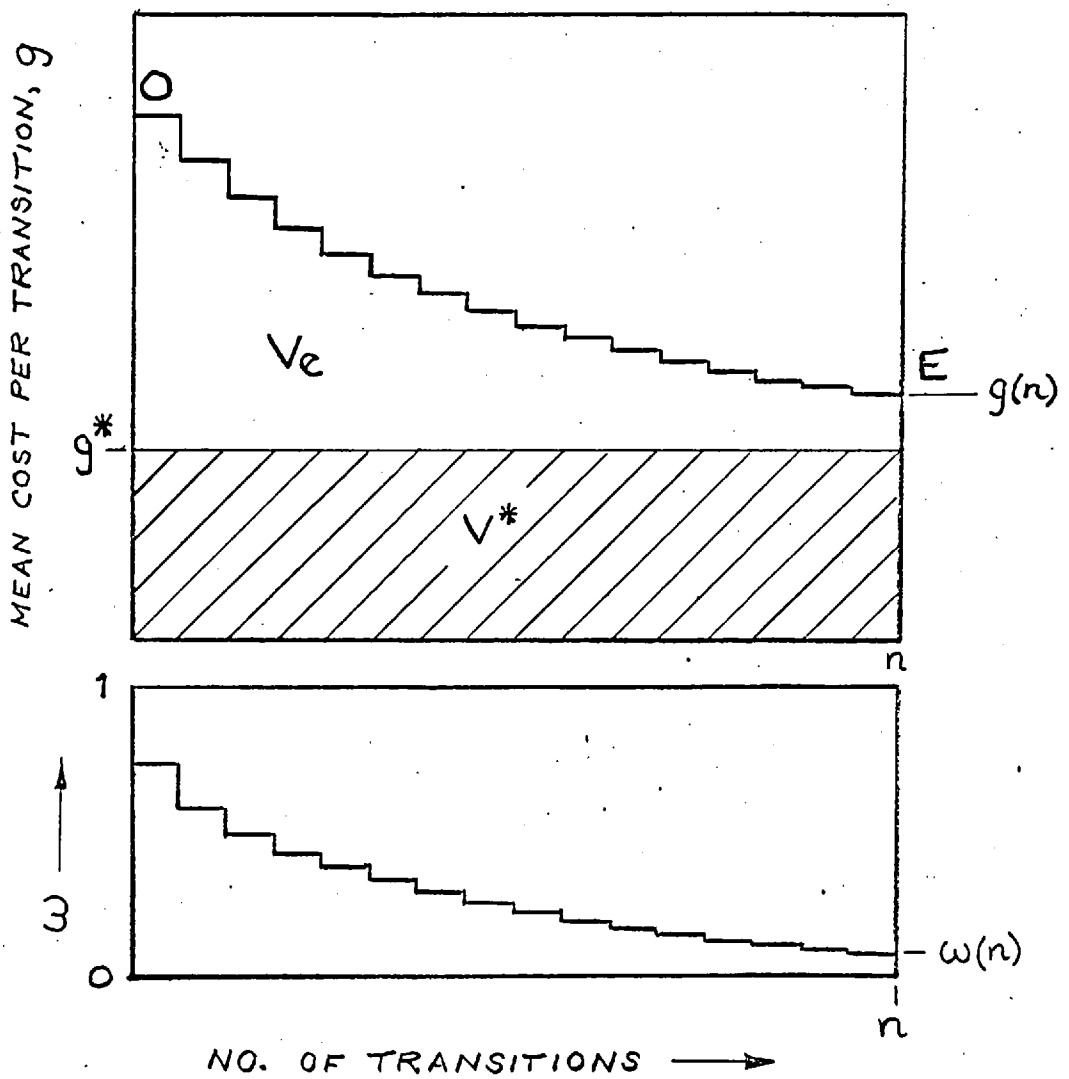


FIG. 1.2

CONVERGENCE OF A

LEARNING PROCESS

we consider the situation at the n^{th} transition, as shown in fig. 1.2. The error probability has reached a value $\omega(n) = \omega_f$, say. The total cost, V, incurred to achieve this degree of convergence, is given by the area under curve OE, which may be decomposed into two parts:

1) area V^{*}, the cost which would have accrued had the posterior estimate of the optimal control policy been used during each of the n transitions;

2) area V_e, the additional estimation cost incurred because some of the past n control decisions were non-optimal.

The estimation cost, V_e, is intimately connected with the concept of dual control. A non-optimal control action, though it increases the cost of operation, performs the service of exploring a new part of the $\{x,u\}$ space not covered by the optimal (or estimated optimal) policy. The result of each such exploration either helps confirm that the present estimate of the optimal policy is correct, or indicates that it may be incorrect. In quantitative terms, it either decreases or increases ω. We may thus regard each non-optimal control action as a purchase of information. The object of dual control is to obtain convergence, and the best dual control strategy is that which obtains a given degree of convergence for the lowest cost, i.e. it buys its information at the lowest price. Formally

we say that the ideal dual strategy is that for which, at the n^{th} transition,

the control sequence $\{u_1, \ldots. u_n\}$

$$\text{minimizes } V = E\left[\sum_{i=1}^{n} L\,(x_i, u_i)\right] \tag{1.6}$$

$$\text{subject to } \omega(n) = \omega_f \tag{1.7}$$

where the cost V is computed using the posterior estimates available after the completion of n transitions.

Equations (1.6) and (1.7) constitute the additional constraint introduced by the dual control requirement. To the author's knowledge, the inclusion of this condition has not been treated quantitatively in any previous learning control scheme for discrete state Markov processes (other than in an earlier report by the author[36]). It will be shown in this thesis that the ramifications of (1.6) and (1.7) are considerable, and lead to the synthesis of simple but very efficient dual control algorithms suitable for on-line control of a variety of physical processes.

1.6 Outline of the Thesis

Chapter 2 begins with a brief review of the terminology of Markov chains. The Markovian decision problem is introduced through a simple example; an optimization algorithm, a slight extension of the original Howard method, is presented for the case in which process parameters are known. It is shown that system operation may be expressed in terms of the alternate sequential action of a stochastic decision matrix, D, and a stochastic transition matrix, P. Given a control cost matrix, B, and a transition cost matrix, C (derived from the function $L(x,u)$), the optimal policy is represented by the decision matrix D^* which minimizes the value of g associated with the matrix PD. At the heart of the algorithm is a linear transformation which transforms the multi-stage optimization problem into an equivalent single-stage one. Two interesting and useful properties of the transformation are demonstrated, and it is shown that under certain conditions the optimization problem is inherently a single-stage one.

The problem of optimizing a repetitive single-stage batch process with initially unknown dynamics is considered in chapter 3. A simple example is discussed to give some intuitive meaning to the concept of dual control, and the

equivalence in discrete processes of dual control and sequential decision making is pointed out. Dual control requirements are presented as constraint conditions in terms of the conditional estimates of the process parameters. The computation of the error probability, ω, early in the life of the process is possible, but very involved; consequently this quantity is replaced by the uncertainty, Ω, and it is shown that both are asymptotically equal.

Solution of the dual problem by means of a Lagrange multiplier shows that the ideal strategy is non-realizable. This approach is nonetheless of great value, since it allows us to study the pattern which emerges with continued application of the ideal strategy. We may then attempt to synthesize a realizable strategy with the same asymptotic pattern, i.e. a realizable asymptotically optimal dual strategy.

For a batch process with N alternative control input, the framework within which the desired pattern may be studied is provided by the concept of an N-1 dimensional decision space, each ordinate of which is the number of past choices of a given non-optimal control action. A series of decisions, or control actions, may be regarded as a decision trajectory descending a conceptual hill of

uncertainty. Descent of the hill implies convergence (equation (1.4)), while the particular path of descent is fixed by the dual control requirement, (1.6) and (1.7). The desired pattern generated by the ideal strategy is shown to be that its decision trajectory in N-1 space is asymptotically a straight line whose direction is specified by the characteristic decision vector of the process.

There exists an infinite family of trajectories asymptotically parallel to the optimal one, each being uniquely related to it. Because each may itself be optimal under certain conditions, the strategies generating this family are termed suboptimal strategies; the performance of suboptimal strategies is seen to be very close to that of the optimal one. From an examination of suboptimal trajectories we proceed naturally to the inverse problem: "Given a strategy whose decision trajectory is asymptotically a straight line described by the characteristic decision vector, specify the conditions under which the strategy is optimal."

The foregoing theory is then used to examine the strategy of Pashkovskii, and it is shown to possess some of the properties of suboptimality for the particular problem he considered. A modified version of his strategy is presented and shown to be generally suboptimal; conditions

under which it is optimal are derived. A completely probabilistic suboptimal strategy is next presented, and compared qualitatively with that of Nikolic and Fu; the former strategy is shown to be optimal when all policies except the optimal one yield the same expected cost. Finally, a general optimal strategy is introduced, and equations of convergence are derived.

To demonstrate the performance of the optimal strategy, a detailed simulation study is presented in chapter 4. As a prelude to this, computational methods relating to the ideal strategy are derived from the theory of chapter 3. A comparison is then made between the results of the (non-realizable) ideal strategy and the (realizable) optimal strategy operating on an ensemble of one hundred three-state systems. The results show that the performance of the optimal strategy is very nearly as good as that of the ideal one, even early in the life of the process. As an application, the simulation of the control of a repetitive chemical batch process is carried out; the process is modelled as a twenty-state (i.e. twenty level) system whose input-output relationship is disturbed by multiplicative non-gaussian noise.

The optimal adaptive controller need not always use a dual strategy. To illustrate the use of a non-dual

adaptive controller and to demonstrate the application of discrete state techniques to a system with uncertain but cyclically time-varying statistical parameters, a study is made in chapter 5 of the adaptive ordering of thermal-electric power generation. While ordering decisions affect power costs, they have no effect whatever upon future demand. In Feldbaum's terminology[19] this is a "neutral" system, one in which the problems of estimation and control can be separated. The adaptive ordering of a 2000 megawatt power station is simulated; convergence to the optimal ordering policy is shown to be rapid despite initially unknown power demand statistics.

It is desirable that the results of chapter 3 relating to single-stage optimization be extended to multi-stage optimization problems (analogous to continuous systems with integral performance criteria). Such an extension is considered in chapter 6. In principle, conditions (1.6) and (1.7) can again be invoked; unfortunately, however, computation of the error probability, ω, does not seem feasible, even asymptotically. By the use of an error measure, Ω, such that $\Omega = 0$ implies that $\omega = 0$, we are able nonetheless to synthesize a multi-stage strategy which is efficient and convergent.

The number of estimated parameters necessary for the

solution of the multi-stage problem is much greater than that for the single-stage case. A method of updating these estimates at each stage of the process, suitable for on-line computation, is next presented. With this technique and the multi-stage algorithm, a discretized optimal control strategy for the automaton model of systems described by (1.2) can be implemented on-line despite initially unknown dynamics, non-gaussian multiplicative disturbances of unknown distribution, system nonlinearities, constraints, and non-quadratic cost functions.

If the output and control variables of the system to be controlled are inherently continuous in nature, then the automaton is a quantized approximation, and so is subject to quantization error. We next show that this error may be removed by what amounts to an interpolation method, providing we can make certain assumptions of continuity and differentiability. A successive approximation algorithm is presented which produces a sampled version of the optimal continuous feedback transducer characteristic for a given discrete time, continuous state Markov process and a given performance criterion. If the process parameters are initially unknown, the same task may be performed by a hierarchy of adaptive loops. As a result of process observations, the outer loops vary the automaton structure

to achieve optimum quantization. In a stationary process quantization is ultimately dispensed with in favour of a continuous feedback characteristic.

Having developed an algorithm for on-line dual control of long duration Markov processes, we are faced with the engineering questions, "Is the technique relevant?", and "Is it computationally feasible with real processes?" The author does not presume to give a definitive answer to either question. However, it is believed that the answers are affirmative; the technique is relevant and feasible for some stochastic processes. In support of this claim a numerical example is presented in chapter 7. Heat treatment of a metal slab involves the control of a distributed temperature-sensitive chemical reaction which tends to instability in some uncertain fashion at temperatures near the desired operating point. Heating costs and the variation of product value with heat treatment temperature are given, but process dynamics are unknown. It is required, first, to make an a priori econimic evaluation of a sampled data controller heating the slab section by section, and, second, to simulate adaptive control of the process.

The process model actually used (unknown to the controller) is a continuous state, discrete time conditionally stable system disturbed by noise whose amplitude is a

non-linear function of both temperature and heat input. An interesting feature of the problem is the presence of a step in the operating cost function. This represents the cost of shutting down the process if an upper temperature limit is reached; alternatively, it is the monetary value of management's displeasure with the control engineer whose adaptive system causes a process stoppage.

Simulation results show that adaptation is successful. The controller is able to maintain the temperature in the admissable range at all times. As knowledge of the process accumulates, the feedback characteristic alters to maintain the temperature near the desired operating point, despite dynamic instability in this region. Between the beginning and end of operation a cost reduction of about 50% occurs.

The possibilities of future research arising from the results of chapters 2 to 7 are discussed in the final chapter of the thesis.

1.7 Contributions of the Thesis

The principal contributions, believed to be original, which the work in this thesis makes to the theory and practice of automatic control, are the following:

1) A general theory of dual strategies for

single-stage cost minimization has been developed for discrete time, discrete state, stationary long duration Markov processes (chapter 3).

2) A realizable asymptotically dual strategy has been synthesized (chapter 3), and its performance has been shown to approach that of the ideal stratety (chapter 4).

3) A method has been presented for the computation of the optimal non-linear continuous feedback transducer characteristic for a given stationary, discrete time, continuous state Markov process (chapter 6).

4) A convergent dual control strategy has been derived for on-line multi-stage optimization of discrete long duration Markov processes (chapter 6). The resultant adaptive controller is quite general in nature, being able to handle initially unknown dynamics, state and control constraints, system nonlinearities, non-gaussian multiplicative disturbances of unknown distribution, and a wide variety of cost functions.

An attempt has been made throughout the thesis to relate theoretical results to physical applications, and to evaluate the usefulness of proposed methods by extensive numerical simulation.

CHAPTER 2

MARKOV CHAINS AND CONTROL SYSTEMS

2.1 Introduction

When both the output variable and the control signal are quantized in a sampled data process disturbed by stationary noise, the operation of the process may be described in terms of a Markov chain. In this chapter a brief introduction to the theory of Markov chains is presented, and the control of a long duration Markov process is considered as a multi-stage Markovian decision problem.

2.2 Markov Chains

A Markov chain may be defined[3,6] as a series of probabilistic trials in which the outcome of any trial depends upon the outcome of the directly preceding trial, and only upon it. To illustrate several basic properties of a Markov chain we shall consider a simple example (adapted from one presented by Howard[14]). Suppose a taxicab operates between two towns, A and B. Over a long period of operation, it has been found that when a passenger

begins his trip in town A, his destination is in town A with probability 0.8, town B with probability 0.2. Passengers from town B, however, show a somewhat greater propensity to migrate; the probability of their destination being town A is 0.4, town B, 0.6. The journeys of the taxicab may now be described in terms of a two state Markov chain in which we associate towns A and B with states 1 and 2 respectively. The dynamics of the process are described by a stochastic transition matrix, P, where

$$P = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \qquad (2.1)$$

P is assumed stationary; each element, p_{ij}, represents the probability that if the present state is i, the next state will be j. The term "stochastic" in this case refers to two properties of the matrix, viz.

$$\sum_{j=1}^{N} p_{ij} = 1 \qquad (2.2)$$

$$p_{ij} \geq 0 \qquad \forall\, i,j \qquad (2.3)$$

$$i=1,2,\ldots N$$

$$j=1,2,\ldots N$$

where N = number of states in the system.

Equation (2.2) is simply an expression of the fact that, after a transition has taken place, the system state must be one or other of the N possible states with probability unity.

Suppose that the taxicab is at present in town A. What is the probability of its being there after one transition (i.e. one journey)? After two transitions? After n transitions? To determine the answers we postulate a stochastic row vector, $\underline{\rho}(n)$, whose elements, $\rho_i(n)$, represent the probability that the system state is i at the n^{th} transition. In this case

$$\underline{\rho}(0) = [1 \quad 0]$$

It is apparent that

$$\underline{\rho}(1) = \underline{\rho}(0)P$$

$$= [1 \quad 0] \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} = [0.8 \quad 0.2]$$

and in general

$$\underline{\rho}(n) = \rho(0)\ P^n \qquad (2.4)$$

Since the elements of P are not time-dependent, we would expect that $\lim_{n\to\infty} \underline{\rho}(n)$ should approach some stationary value whose elements $\rho_i(\infty)$ represent the probability of

occupancy of state i over a very long sequence of transitions. A little arithmetic shows that P, P^2, P^3 ... has the sequence

$$\begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix}, \quad \begin{bmatrix} 0.72 & 0.28 \\ 0.56 & 0.44 \end{bmatrix}, \quad \begin{bmatrix} 0.688 & 0.312 \\ 0.624 & 0.376 \end{bmatrix} \cdots$$

and approaches a limiting value

$$\lim_{n\to\infty} P^n = \begin{bmatrix} 0.667 & 0.333 \\ 0.667 & 0.333 \end{bmatrix}$$

so that, regardless of the initial distribution, $\underline{p}(0)$, the vector $\underline{p}(n)$ approaches the limit

$$\lim_{n\to\infty} \underline{p}(n) = \underline{\pi} = [0.667 \quad 0.333]$$

Since the steady state distribution is independent of the initial distribution, matrix P is called ergodic. As an example of a non-ergodic matrix, consider

$$P_1 = \begin{bmatrix} 0.3 & 0.7 & 0.0 & 0.0 \\ 0.2 & 0.8 & 0.0 & 0.0 \\ 0.2 & 0.1 & 0.2 & 0.5 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Here, if the process starts in state 1 or 2, or if a transition from state 3 to 1 or 2 occurs, the process will remain forever in the subset, 1 and 2. If a state, j, is

reachable from another state, i, it is said to be accessible from i; two states which are mutually accessible are said to communicate. In matrix P_1, states 1 and 2 communicate with each other, and are accessible from state 3. State 3 is said to be inessential, since its steady state probability of occupancy is zero. State 4 is a trapping state, communicating only with itself. Inspection of P_1 shows that it is not ergodic, since the steady state distribution depends upon events early in the life of the process.

If every pair of states in a Markov chain communicates, then the chain is said to be irreducible. Matrix P in (2.1) is irreducible, while P_1 is not. P_1 may be decomposed into the irreducible subset, states 1 and 2, the inessential state, 3, and the (degenerate) irreducible subset, state 4.

One other type of stochastic matrix is of interest. Consider

$$P_2 = \begin{bmatrix} 0.0 & 0.0 & 0.4 & 0.6 \\ 0.0 & 0.0 & 0.7 & 0.3 \\ 0.2 & 0.8 & 0.0 & 0.0 \\ 0.9 & 0.1 & 0.0 & 0.0 \end{bmatrix}$$

It is evident that the state must alternate between the subset, 1 and 2, and the subset, 3 and 4. An examination of the eigenvalues of P_2 shows that two of them lie on the

unit circle, one at $\lambda_1 = 1$ and the other at $\lambda_2 = -1$. In general a stochastic matrix with q eigenvalues on the unit circle, only one of which lies at $\lambda = 1$, is <u>periodic</u> with period q. Matrix P_2 has period 2, i.e. the process will occupy a given subset of states at every second transition. A Markov chain with period 1, such as that described by P in (2.1) is called <u>aperiodic</u>.

An important property of every irreducible periodic Markov chain with period q is that it can be considered as the alternating sequential operation of q aperiodic Markov chains. For example, if we correctly change the labels attached to the states at each transition, we may consider the operation of P_2 as the alternation of two stochastic matrices

$$\begin{bmatrix} 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix} \Longrightarrow \begin{bmatrix} 0.2 & 0.8 \\ 0.9 & 0.1 \end{bmatrix} \Longrightarrow \begin{bmatrix} 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix} \cdots$$

States 1 and 2 States 3 and 4 States 1 and 2

Throughout this thesis we shall consider only ergodic processes, as these occur most frequently in practice. Periodic matrices will be encountered, and will be dealt with by the decomposition technique described above.

Mathematically, we may summarize several important

properties of every finite stochastic matrix, P, as follows (proofs may be found in references 3, 4, and 9)

1) There exists an eigenvalue of P which is equal to 1.

2) No eigenvalue of P exists outside the unit circle on the complex plane.

3) If the eigenvalue $\lambda = 1$ occurs singly, and no other eigenvalues occur on the unit circle, then the matrix P is aperiodic and ergodic; $\lim_{n\to\infty} P^n$ approaches a matrix E, in which the j^{th} element, e_{ij}, of every row of E is the steady state probability of occupancy of state j.

4) If the eigenvalue $\lambda = 1$ occurs singly, and (q-1) eigenvectors lie elsewhere on the unit circle, then the matrix is periodic and ergodic. Ergodicity in this case implies that

$$\lim_{n\to\infty} (P^n - P^{n+q}) = 0$$

5) $E^2 = E = EP = PE$.

2.3 Markov Processes with Costs: The Markovian Decision Problem

Let us return to the taxicab problem. Suppose that a certain expected income is associated with each type of transition as follows:

Transition	Income (arbitrary units)
A → A	5
A → B	10
B → A	9
B → B	8

We may conveniently describe this income in terms of a cost matrix, C

$$C = \begin{bmatrix} -5 & -10 \\ -9 & -8 \end{bmatrix}$$

where c_{ij} = cost (income is negative cost) of a transition from state i to state j.

If the present state is i, then the expected cost, μ_i, of the next transition, is

$$\mu_i = \sum_{j=1}^{N} p_{ij}\, c_{ij} \qquad (2.6)$$

It is apparent that in an ergodic process the expected cost per transition, g, over a long sequence of transitions is

$$g = \sum_{i=1}^{N} \pi_i\, \mu_i = \langle \underline{\pi}\ \underline{\mu} \rangle \qquad (2.7)$$

where $\langle\cdot$ denotes a row vector

$\cdot\rangle$ denotes a column vector

$\langle\cdot\ \cdot\rangle$ denotes an inner product

From (2.6) we obtain

$$\underline{\mu} = [-6.0 \quad -8.4]^T$$

From (2.5) and (2.7)

$$g = [0.667(-6.0) + 0.333(-8.4)] = -6.80$$

Thus the driver can expect an average income of 6.80 units per trip over a long period of time.

To use control terminology, we have so far described a process with dynamics, but no control input; i.e. no mechanism exists for the taxicab driver to make choices or decisions. Let us therefore postulate that in each town he has the alternative either of cruising or of going to a cab stand and waiting for a call. Moreover, let us suppose that he may, if he wishes, drive from one town to the other without a passenger. Now after each transition he can make one of the following decisions

Decision	Decision State
Go to A and cruise	1
Go to A, cab stand	2
Go to B and cruise	3
Go to B, cab stand	4

Note that while there are only two "process states" which may result from a probabilistic transition, viz. town A or town B, there are four "decision states". Again using control terminology, we may say in general that the space spanned by the process states is the (quantized) space of the output variable, x, of a process. The space spanned by the decision states is the (quantized) combination x,u, where u is the control variable. A decision state $k = (i,j)$ describes the event: present process state is i, decision has been made to apply control alternative j.

The example of table 2.2 differs slightly from the above description in that all four decision states are accessible from either process state, a situation which may occur in operations research problems and in the control of a sequence of batch processes (chapter 3), but is rarely encountered in the control of continuous dynamic processes (chapters 6 and 7).

Suppose that the probabilities and transition costs associated with the four decision states of table 2.2 are as follows:

Decision State	Prob. of transition to A	Prob. of transition to B	Cost of transition to A	Cost of transition to B
1	0.8	0.2	- 5	-10
2	0.5	0.5	- 6	-12
3	0.4	0.6	- 9	- 8
4	0.7	0.3	- 6	- 4

We may assume in addition that a certain cost is associated with each decision (cruising costs may include petrol and tire wear, for example). These decision costs (control costs) are listed below:

Process State	Cost of reaching decision state 1	2	3	4
1	2	0	4	3
2	5	3	1	0

The problem now is to determine what decisions the cab driver should make to maximize expected income (minimize cost). This type of problem, frequenctly encountered in the field of operations research, is termed a Markovian decision problem. Some variants appearing in the literature are noted in section 1.2. Before proceeding to its solution, we shall recapitulate in a generalized form. We are given the following data:

P = L x N stochastic transition matrix whose elements, p_{ij}, represent the probability that if the present decision state is i, then the next process state will be j.

C = L x N transition cost matrix whose elements, c_{ij}, represent the cost associated with a probabilistic transition from decision state i to process state j.

B = N x L control (decision) cost matrix whose elements, b_{ij}, represent the cost of reaching decision state j, if the present process state is i.

where N = total number of process states

L = total number of decision states.

Given a long duration stationary discrete process with matrices P, C, and B, and with present process state i, which of the L decision states, j = 1, 2, ... L, should now be chosen, to minimize the expected overall cost per transition?

All decisions of this type may be expressed in the framework of a decision matrix, D.

D = N x L stochastic decision matrix whose elements, d_{ij}, represent the probability that if the present process state is i, control will be exerted to reach decision state j.

The optimal policy is defined by the decision matrix D^{*} such that $g(D) \geq g(D^{*})$ for all admissable decision matrices D other than D^{*}. Though D^{*} is usually unique in practice, it need not be so.

2.4 Process Optimization

The overall process consists of the alternate sequential

operation of a probabilistic transition matrix, P, and a decision matrix, D. The N x N matrix DP describes successive transitions of the process state; it is the view of the process as seen at its output. The L x L matrix PD describes transitions of the decision state, and so is the view of the process seen by the controller. We shall call the combined operation of a probabilistic transition followed by a decision one cycle or stage of the process. If the present decision state is i, then the expected cost of the next stage of operation is

$$\ell_i = \sum_{j=1}^{N} p_{ij}(c_{ij} + \sum_{k=1}^{N} d_{jk} b_{jk}) \qquad (2.8)$$

Note that (2.8) is a generalization of (2.6).

In an ergodic process it can be shown that an optimal decision matrix D^* can be found, all elements of which are either one or zero; the optimal policy is thus regular (non-probabilistic). This being so, we could, in principle, determine the principal row eigenvector, $\underline{\pi}(PD)$, for every regular matrix D, and find the matrix D^* which minimizes g in (2.7). The difficulty with such a procedure is that there exist L^N distinct regular policies. In a system with ten process states and ten control states - only a modest size - the computational difficulties would be overwhelming. A technique other than pure search is obviously necessary.

We shall pose the problem in terms of dynamic programming[1,10]. Let $V_i(n)$ be the expected total cost associated with the process if there are n stages left before termination (the time scale runs backwards), and the present decision state is i. We may now write the recursive equation

$$V_i(n+1) = \sum_{j=1}^{N} p_{ij} \left[c_{ij} + \sum_{k=1}^{N} d_{jk}(b_{jk} + V_k(n))\right]$$

$$= \ell_i + \sum_{k=1}^{L} \left[\sum_{j=1}^{N} p_{ij} d_{jk}\right] V_k(n)$$

$$V_i(n+1) = \ell_i + \sum_{k=1}^{L} r_{ik} V_k(n) \qquad (2.9)$$

where $$r_{ik} = \sum_{j=1}^{N} p_{ij} d_{jk}$$

i.e. $$R = [r_{ik}] = PD \qquad (2.10)$$

For a process of long duration, $n \to \infty$. In such a case we know that if it is ergodic

$$\lim_{n\to\infty} [V_i(n+1) - V_i(n)] = g \qquad (2.11)$$

$$\forall i,\ i = 1, 2, \ldots. L$$

Combining (2.9) and (2.11) we obtain L steady state equations

$$g + V_i = \ell_i + \sum_{j=1}^{N} r_{ij} V_j \qquad (2.12)$$

with L + 1 unknowns, V_i, i = 1, 2, ... L, and g. Since each V_i decreases by g per stage (time running forward now), it is sufficient to know the relative values of the costs V_i. Thus we may arbitrarily set one of them, say V_L, to zero. Let us define a set of variables v_i, so that

$$v_i = V_i - V_L , \qquad i = 1, 2, \ldots L$$

$\mathbf{v}_i$ is now the relative value of starting a process in decision state i. For large n, the cost associated with n transitions starting in decision state i is ng + $\mathbf{v}_i$; starting in state j, it is ng + $\mathbf{v}_j$. While the expected cost per stage, g, is independent of starting state for n → ∞, there exists a time invariant difference $\mathbf{v}_j - \mathbf{v}_i$ which a rational person would just be willing to pay (a "one-shot" cost) to start the process in state i rather than state j.

With $\mathbf{v}_L \equiv 0$, we may put (2.12) into canonical form by defining a column vector, z, such that

$$z_i = \mathbf{v}_i , \qquad i = 1, 2, \ldots L-1$$

$$z_L = g$$

Now let

$$Rg = \begin{bmatrix} r_{11} & r_{12} \cdots\cdots & r_{1(L-1)} & -1 \\ r_{21} & r_{22} \cdots\cdots & r_{2(L-1)} & -1 \\ \vdots & \vdots & \vdots & \vdots \\ r_{(L-1)1} & r_{(L-1)2} & r_{(L-1)(L-1)} & -1 \\ r_{L1} & r_{L2} & r_{L(L-1)} & 0 \end{bmatrix} \qquad (2.13)$$

so that equation (2.12) becomes

$$\underline{z} = Rg\ \underline{z} + \underline{\ell}$$

$$\underline{z} = (I - Rg)^{-1}\ \underline{\ell}$$

$$\underline{z} = \Psi^{-1}\ \underline{\ell} \qquad (2.14)$$

where I = unit matrix

$\Psi = I - Rg$

Suppose that we begin the optimization procedure by choosing an arbitrary decision matrix, D. We may determine Rg immediately from (2.10) and (2.13) and $\underline{\ell}$ from (2.8). We then solve (2.14) for the set v_i, and the expected cost per transition, g. Now the relative cost of choosing decision state j, if the present process state is i, is

$$\eta_{ij} = b_{ij} + v_j \qquad (2.15)$$

Thus, for every process state, i, we choose a control state s = s(i) such that

$$\eta_{is} = \underset{j}{\text{Min}} \, [\eta_{ij}] \tag{2.16}$$

A new decision matrix, D_1, is now formed from row vectors $< \underline{d}_i$ where

$$\begin{aligned} d_{ij} &= 1 , \qquad j = s(i) \\ &= 0 , \qquad j \neq s(i) \end{aligned} \tag{2.17}$$

$$i = 1, 2, \ldots . N$$

$$j = 1, 2, \ldots . L$$

If $D_1 = D$, then D is the optimal decision matrix, D^*. If $D_1 \neq D$, then D is replaced by D_1, and the cycle is re-entered by the computation of a new compound matrix, R. Fig. 2.1 is a simplified flow chart of the computation of D^*. A proof of convergence of this algorithm is given in appendix 1.

2.5 Example

To illustrate the use of the optimization algorithm, we shall apply it to the taxicab example considered previously. Recall that matrices P, C, and B are

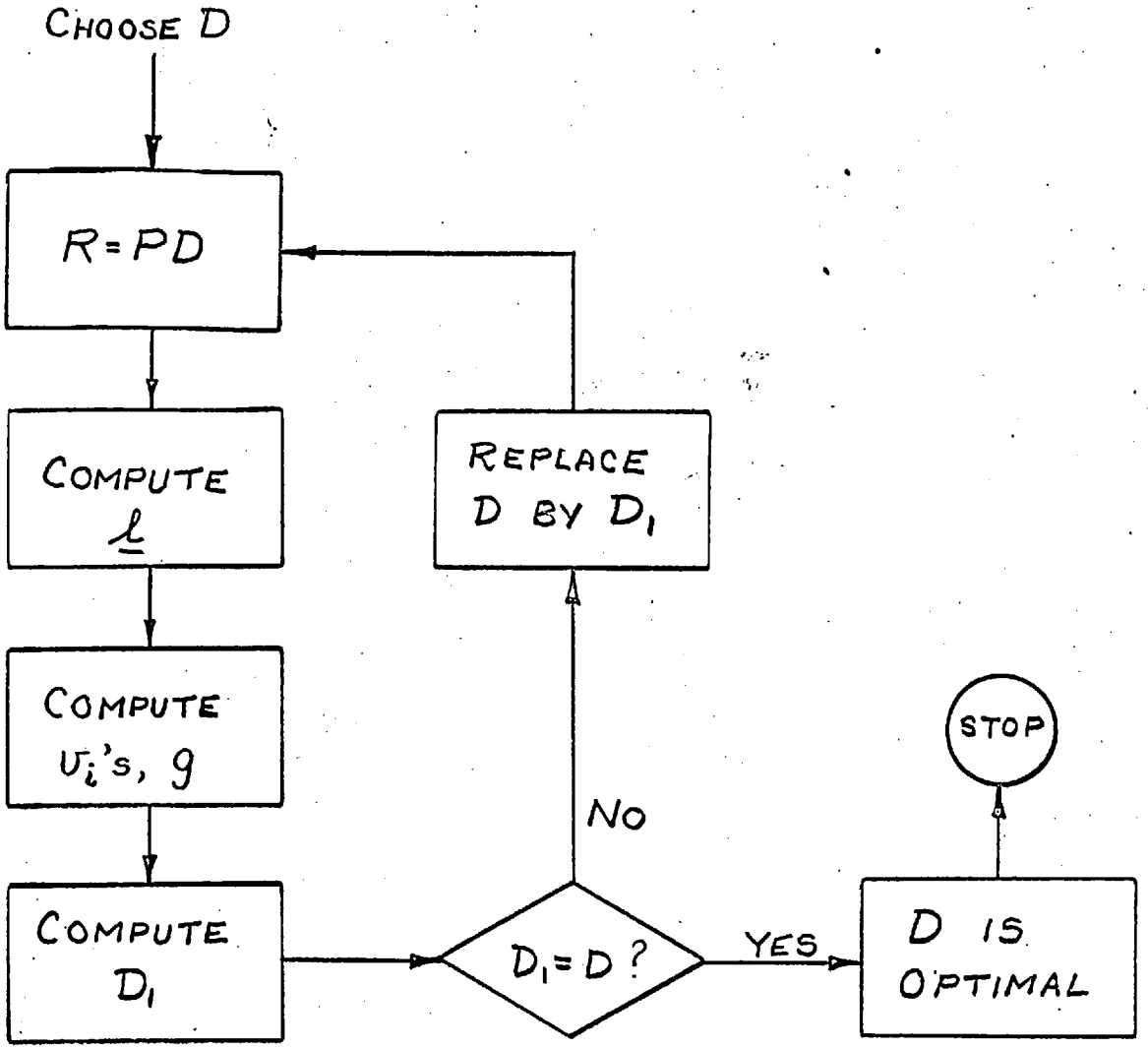


FIG. 2.1

COMPUTATION OF D^*

WITH KNOWN P

$$P = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix} \qquad C = \begin{bmatrix} -5 & -10 \\ -6 & -12 \\ -9 & -8 \\ -6 & -4 \end{bmatrix}$$

$$B = \begin{bmatrix} 2 & 0 & 4 & 3 \\ 5 & 3 & 1 & 0 \end{bmatrix}$$

Let us choose as an initial decision: if a journey takes the cab either to town A or town B, stay in that town and cruise. The corresponding decision matrix is

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

so that

$$R = PD = \begin{bmatrix} 0.8 & 0.0 & 0.2 & 0.0 \\ 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.6 & 0.0 \\ 0.7 & 0.0 & 0.3 & 0.0 \end{bmatrix}$$

Inspection of R indicates that decision states 1 and 3 communicate and form an irreducible Markov chain, while states 2 and 4 are inaccessible from all states, and so are inessential. States 1 and 3 form a <u>basic chain</u>, i.e. the smallest set of states with which (2.12) may be solved.

In considering control problems, we shall find that frequently $L > N$, and a basic chain with an $N \times N$ matrix R_b can be substituted for the L state chain described by matrix R. Note that this reduction is possible only when D is a regular matrix (all elements either one or zero).

The basic R matrix in the problem at hand is

$$R_b = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \quad \begin{matrix} \longleftarrow \text{ Decision State 1} \\ \longleftarrow \text{ Decision State 3} \end{matrix}$$

$$R_{gb} = \begin{bmatrix} 0.8 & -1 \\ 0.4 & 0 \end{bmatrix}$$

$$\Psi = I - R_{gb} = \begin{bmatrix} 0.2 & 1.0 \\ -0.4 & 1.0 \end{bmatrix}$$

$$\underline{\ell} = \begin{bmatrix} 0.8\ (-5+2) + 0.2\ (-10+1) \\ 0.4\ (-9+2) + 0.6\ (-8+1) \end{bmatrix} = \begin{bmatrix} -4.2 \\ -7.0 \end{bmatrix}$$

From (2.14) we have

$$\begin{bmatrix} v_1 \\ g \end{bmatrix} = \Psi^{-1}\,\underline{\ell} = \begin{bmatrix} 1.667 & -1.667 \\ 0.667 & 0.333 \end{bmatrix} \begin{bmatrix} -4.2 \\ -7.0 \end{bmatrix} \qquad (2.18)$$

so that

$$v_1 = 4.667$$
$$g = -5.133$$

while $v_3 = 0$ by hypothesis.

Using this policy, the driver can expect an average income of 5.133 units per journey. To test for optimality it is necessary to compute costs v_2 and v_4 associated with decision states 2 and 4. This is easily done with equation (2.12).

$$v_i = \underline{\ell}_i + \left[\sum_{j=1}^{N} r_{ij} v_j \right] - g , \quad i \text{ inessential}$$

Since i is inessential, $r_{ii} = 0$, and all right hand terms are known. Thus

$$\begin{aligned} v_2 &= 0.5(-6 + 2 + 4.667) + 0.5(-12 + 1 + 0.0) \\ &\quad - (-5.133) \\ &= -0.0333 \\ v_4 &= 0.7(-6 + 2 + 4.667) + 0.3(-4 + 1 + 0.0) \\ &\quad - (-5.133) \\ &= 4.70 \end{aligned}$$

The matrix of η values is

$$\eta = \begin{bmatrix} 6.667 & -0.0333 & 4.00 & 7.70 \\ 9.667 & 2.9667 & 1.00 & 4.70 \end{bmatrix}$$

The minimum element of row 1 is η_{12}, of row 2, η_{23}. Therefore the improved decision matrix is

$$D = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad (2.19)$$

In the revised system, decision states 1 and 4 are inessential. Re-solving equations (2.10), (2.12), (2.13), and (2.14), we obtain

$$\begin{aligned} v_1 &= 1.245 \\ v_2 &= -1.889 \\ v_3 &= 0.0 \\ v_4 &= 2.133 \\ g &= -8.556 \end{aligned}$$

$$\eta = \begin{bmatrix} 3.245 & -1.889 & 4.00 & 5.133 \\ 6.245 & 1.111 & 1.00 & 2.133 \end{bmatrix}$$

Inspection of the η matrix above indicates that the decision matrix will remain unchanged at the next iteration; therefore (2.19) is the optimum decision matrix. The driver's best policy is to stay in whichever town he finds himself at the end of a trip; if in town A, he should go to a cab stand; if in town B he should cruise. Note that the new policy, with an expected income of 8.556 per stage, is a considerable improvement on the former one.

2.6 Properties of the Transformation Matrix

The algorithm outlined in section 2.4 yields an optimum multi-stage policy. The term "multi-stage" is used to denote that the control policy optimizes system performance over many (an infinite number) stages of operation. The equivalent in continuous systems is the minimization of an integral performance criterion. The matrix Ψ^{-1} is termed the transformation matrix, since it performs the important function of transforming the multi-stage cost problem into an equivalent single-stage problem with costs η_{ij} given by (2.15). A closer examination of Ψ^{-1} is warranted, since an understanding of its properties yields considerable insight into the operation of discrete state Markov processes.

Let $\underline{e}_i >$ be a column vector of N elements, all of which are zero except the i^{th}, which is unity. Let $\underline{w} >$ be a column vector of N elements, all unity. The corresponding row vectors are $< \underline{e}_i$ and $< \underline{w}$, respectively. Now consider the matrix R_g of the N-state basic chain of a discrete Markov process. From (2.13)

$$R_g = R - R\underline{e}_N><\underline{e}_N - \underline{w}><\underline{e}_N + \underline{e}_N><\underline{e}_N$$

$$\Psi = I - R_g = I - R + R\underline{e}_N><\underline{e}_N + \underline{w}><\underline{e}_N - \underline{e}_N><\underline{e}_N$$

$$\Psi = [I + \underline{w}><\underline{e}_N - \underline{e}_N><\underline{e}_N] + [R(\underline{e}_N><\underline{e}_N - I)] \qquad (2.20)$$

We may now use the matrix inversion lemma

If $A_2 = A_1 + FGH$

then $$A_2^{-1} = A_1^{-1} - A_1F(HA_1^{-1}F + G^{-1})^{-1}HA_1^{-1} \qquad (2.21)$$

where, in this case,

$$A_2 = \Psi$$
$$A_1 = I + \underline{w}\rangle\langle\underline{e}_N - \underline{e}_N\rangle\langle\underline{e}_N$$
$$F = R$$
$$G = I$$
$$H = \underline{e}_N\rangle\langle\underline{e}_N - I$$

Application of the lemma to the inversion of A_1 (the first term in brackets in 2.20) yields

$$A_1^{-1} = I - (\underline{w}\rangle - \underline{e}_N)\langle\underline{e}_N \qquad (2.22)$$

Substitution of (2.21) into (2.20) gives

$$\Psi^{-1} = A_1^{-1} - A_1^{-1}R\,[(\underline{e}_N\rangle\langle\underline{e}_N - I)A_1^{-1}R + I]^{-1} \cdot (\underline{e}_N\rangle\langle\underline{e}_N - I)A_1^{-1} \qquad (2.23)$$

Substitution of (2.22) into (2.23) yields

$$\Psi^{-1} = [I - (\underline{w}\rangle - \underline{e}_N)\langle\underline{e}_N] \cdot \left\{I - R[(\underline{w}\rangle\langle\underline{e}_N - I)R + I]^{-1}[\underline{w}\rangle\langle\underline{e}_N - I]\right\} \qquad (2.24)$$

Let us now postmultiply Ψ^{-1} by the vector $\underline{w}\rangle$

$$\Psi^{-1}\underline{w}\rangle = [I - (\underline{w}\rangle - \underline{e}_N) < \underline{e}_N] \cdot \left\{ \underline{w}\rangle - R[(\underline{w}\rangle\langle e_N - I)R + I]^{-1} [\underline{w}\rangle\langle \underline{e}_N\underline{w}\rangle - \underline{w}\rangle] \right\}$$

Since $\langle \underline{e}_N \underline{w}\rangle = 1$, the multiplier of the inverted term becomes a null vector, and

$$\Psi^{-1}\underline{w}\rangle = [I - (\underline{w}\rangle - \underline{e}_N\rangle) < \underline{e}_N] \underline{w}\rangle$$

i.e. $$\Psi^{-1}\underline{w}\rangle = \underline{e}_N\rangle \qquad (2.25)$$

We have thus obtained the first property of $^{-1}$:

Property 1

The first N-1 rows of the transformation matrix, Ψ^{-1}, sum to zero, while the N^{th} row sums to unity.

Property 1 follows directly from (2.25). It is a useful check to ascertain that Ψ^{-1} has been computed correctly; in addition it points to the significance of the N^{th} row of Ψ^{-1}, and so leads to

Property 2

The N^{th} row of Ψ^{-1} is the principal row eigenvector of the matrix R.

To see why this is so, we recall that

$$\underline{z} = (v_1 \quad v_2 \cdots v_{N-1} \quad g]^T$$

The N^{th} scalar equation associated with the vector equation (2.14) is

$$g = \langle \underline{e}_N \Psi^{-1} \underline{\ell} \rangle$$

However we know from the generalization of (2.7) that

$$g = \langle \underline{\pi} \underline{\ell} \rangle$$

where $\langle \underline{\pi}$ is the principal row eigenvector of the matrix R associated with the basic chain. Thus

$$\langle \underline{e}_N \Psi^{-1} = \langle \underline{\pi}$$

and property 2 is proved. An example of both of these properties may be seen in equation (2.18).

Property 2 shows that the algorithm which determines the matrix D^{∞} also yields the probability distribution of states as a byproduct. This information is particularly useful in the control of continuous state systems (chapter 6) since optimum quantization levels depend upon the state distribution.

2.7 Processes with Single Stage Optimisation

While in general it is necessary to carry out multi-stage optimization to ensure an overall optimum decision policy, there exists a class of system for which single-stage optimization achieves the same result. We shall

refer to this class as batch processes. We shall define a batch process as a discrete Markov process in which, after any probabilistic transition from process state i to process state j, i, j = 1, 2, ... N, control may be applied so that the process state, j, may be changed deterministically to any other process state, k, before the next probabilistic transition takes place. Moreover, control cost is independent of which control input is applied. Note that for a batch process, L = N.

As an example consider a batch chemical process which is to be run repeatedly. Suppose that the initial concentration (initial state) of the constituents can be set by the controller. In the absence of any disturbance there is a certain final value of concentration (final state) corresponding to each initial state. However, the process is subject to a disturbance in the form of variable catalyst activity. The statistics of the disturbance are known stationary functions of initial state; thus to each initial concentration there corresponds a particular distribution of final concentrations. The cost associated with the running of a batch is a known function of initial and final states. After a batch has been run, the initial state of the next batch can be re-set at no cost.

To minimize the expected cost per stage of such a

process it is necessary simply to compute which initial state, s, incurs a minimum value of μ_i, using equation (2.6). Then, regardless of the final state of a given batch, the initial state of the next one is set to s. Note that we have achieved overall cost minimization by minimizing over only one stage of the process; this can be done because each stage is independent of the one preceding it. Evidently the optimal decision matrix is one in which all of the rows are identical, each containing N-1 zeros, and one unit element in the s^{th} column. Such a matrix is called a stochastic dyad, since it is a stochastic matrix which can be expressed as the outer product of two vectors, i.e.

$$D^* = \underline{w} > < \underline{e}_s$$

Since the dual control of sequential batch processes is to be treated in detail in chapters 3 and 4, it seems worthwhile to present here a formal proof of the relevance of single-stage optimization. We shall prove the following:

If, in an N-state Markovian decision problem,

1) all elements, b_{ij}, of the control cost matrix, B, are equal;

2) after a probabilistic transition from process state i to process state j has occurred, control may be

applied so that the process state j may be changed deterministically to any other process state, k, before the next probabilistic transition takes place; i, j, k = 1, 2, ... N; then it follows that:

1) the optimal decision matrix, D^{*}, is a stochastic dyad;

2) $(\Psi^{*})^{-1}$ may be determined explicitly without a matrix inversion;

3) optimization of one stage of the process is equivalent to multi-stage optimization.

Proof:

Using the optimization algorithm of section 2.4, we may begin by choosing D arbitrarily; let us choose a stochastic dyad

$$D = \underline{w} > < \underline{d}$$

so that every row of D is the row vector $<\underline{d}$ where

$$<\underline{d} = [d_{i1} \quad d_{i2}, \ \ldots \ d_{iN}]$$

and

$$\sum_{j=1}^{N} d_{ij} = 1$$

$$d_{ij} \geq 0 \qquad\qquad i, j = 1, 2, \ldots N$$

Since $\sum_{j=1}^{N} p_{ij} = 1,$

$$P\,\underline{w}\rangle = \underline{w}\rangle$$

so that $R = PD = P\underline{w}\rangle\langle\underline{d} = \underline{w}\rangle\langle\underline{d}$

Substitution of $R = \underline{w}\rangle\langle\underline{d}$ into (2.24) gives

$$\Psi^{-1} = [I - (\underline{w}\rangle - \underline{e}_N\rangle) < \underline{e}_N]$$

$$\cdot \left\{ I - \underline{w}\rangle\langle\underline{d}\,[(\underline{w}\rangle\langle\underline{e}_N - I)\,\underline{w}\rangle\langle\underline{d} + I]^{-1}\,[\underline{w}\rangle\langle\underline{e}_N - I] \right\}$$

The inverted term simplifies to

$$[(\underline{w}\rangle\langle\underline{e}_N\,\underline{w}\rangle - \underline{w}\rangle) < \underline{d} + I]^{-1} = [\underline{0}\rangle\langle\underline{d} + I]^{-1} = I$$

Thus

$$\Psi^{-1} = [I - (\underline{w}\rangle - \underline{e}_N) < \underline{e}_N][I - (\underline{w}\rangle\langle\underline{d})(\underline{w}\rangle\langle\underline{e}_N - I)$$

$$\Psi^{-1} = I - \underline{w}\rangle\langle\underline{e}_N + \underline{e}_N\rangle\langle\underline{d} \qquad (2.26)$$

Substituting (2.26) into (2.14) we obtain

$$\underline{z}\rangle = \underline{\ell}\rangle - \underline{w}\rangle\langle\underline{e}_N\,\underline{\ell}\rangle + \underline{e}_N\rangle\langle\underline{d}\,\underline{\ell}\rangle$$

i.e. $$v_i = \ell_i - \ell_N \qquad (2.27)$$

$$g = \sum_{j=1}^{N} d_{ij}\,\ell_j \qquad (2.28)$$

Having completed one step of the iteration, we now choose a new matrix, D_1, according to equation (2.16); for each i, we choose an s = s(i) such that

$$b_{is} + v_s = \min_j [b_{ij} + v_j] \tag{2.29}$$

Since all elements, b_{ij}, are equal by hypothesis, (2.29) reduces to

$$v_s = \min_j [v_j] \quad \forall_i,\ i = 1, 2, \ldots N; \tag{2.30}$$

Evidently the state s is the same for all states i; the new matrix D_1 is therefore a stochastic dyad

$$D_1 = \underline{w} > < \underline{e}_s \tag{2.31}$$

If a multiple minimum occurs in states r, s, t, ... then the new decision matrix may be any one of $\underline{w} > < \underline{e}_r$, $\underline{w} > < \underline{e}_s$, $\underline{w} > < \underline{e}_t$ .... , but it is still a stochastic dyad. It is easy to verify that if another iteration of the algorithm is performed, a repetition of (2.27), (2.28), and (2.31) will occur. Therefore $D_1 = D^*$, and result 1) is proved. Result 2) follows from (2.26) and (2.31), i.e.

$$(\Psi^*)^{-1} = I - \underline{w} > < \underline{e}_N + e_N > < \underline{e}_s \tag{2.32}$$

Result 3) follows from (2.27) and (2.28) which show that the multi-stage equation (2.14) degenerates for batch

processes into a set of equations dependent only upon single-stage costs.

2.8 Process Models

The basic Markovian decision problem treated in this chapter can be modified to apply to a variety of physical processes. Three types of process which will be examined in this thesis are the following:

1) Batch Processes

The properties of batch processes have been presented in the previous section. We note that the process states and decision states coincide (L = N). If the control cost matrix contains equal elements (frequently zero), then one-stage optimization can be used.

2) Dynamic Processes

In physical processes with sampled data control, the control and disturbance act simultaneously, not sequentially. A measurement of state is made at the beginning of a sampling interval, and a control effort, usually of constant magnitude, is applied during the interval. The control signal thus has no effect on the present process state, but does

affect the probability distribution of states one time interval in the future. For such a process, we may postulate the usual N proces states. Corresponding to each of these are Γ different control choices which may be applied. Thus there is a total of ΓN decision states ($L = \Gamma N$), but only a subset, numbering Γ, is accessible from each process state. The optimization of dynamic processes is done in terms of the N-state basic chain. Examples are given in chapter 7.

3) Cyclic Decision Processes

The dynamics and/or statistics of certain processes undergo cyclic changes. Each overall period or interval consists of T sub-intervals, each with its own matrices $P(t)$, $C(t)$, $B(t)$, $t = 1, 2, \ldots T$. The characteristics of each sub-interval may be either those described in 1) or 2). To control the process optimally, a sequence of T optimal decision matrices $D^{*}(1)$, $D^{*}(2)$, $\ldots D^{*}(t)$ must be computed. An example will be presented in chapter 5.

2.9 Summary

This chapter has been largely a review of results well known in the theory of Markov chains and Markovian decision

problems. The presentation, however, has emphasized the relationship of methods customarily reserved for inherently discrete problems in operations research to the control of physical processes which may be continuous in nature.

The process to be controlled is assumed to be modelled as a long duration stationary discrete Markov process. If a cost of operation is associated with each possible state transition and each possible decision or control action, then optimization implies the solution of a Markovian decision problem. The optimal decision policy is expressed in terms of a decision matrix, D^{M}, which minimizes the expected cost per stage of process operation. The discretized output variable forms a set of process states, which constitute the observations made by the controller. The latter must choose one of a discrete set of control inputs based upon the current observation. The combination of all process states with all admissible controls forms a set of decision states. Any control policy may therefore be regarded as a mapping of the N process states into a subset of N decision states, the mapping operator being the decision matrix. In the language of process control, an optimal policy means the specification of an optimal feedback transducer.

It is evident from the definition of decision states

that their number usually exceeds that of the process states. Providing we specify a deterministic feedback policy (a "pure strategy", in the language of game theory), all decision states need not be considered at once for computational purposes. Once a decision policy is chosen, all but N of the decision states become inessential in the resulting Markov chain, and the effect of the policy is described by a set of N linear equations. An iterative policy improvement scheme yields the optimal decision matrix, D^{*}.

The algorithm which determines D^{*} produces other important information about the process as well. It has been shown that the steady state probability of occupancy of each process state is contained in the transformation matrix, ψ^{-1}. In addition the η matrix provides the relative costs of actions which deviate from the optimal policy. This information is necessary to determine the optimum policy in a discrete state system; in chapter 6 we shall see that it has special importance as a gradient measure in the adaptive control of continuous state processes.

Throughout this chapter we have assumed that the matrix P is known exactly; the problem is therefore probabilistic in nature, and the solution is exact. In chapter 3 we

shall discard this comfortable assumption, and consider the statistical problem in which P is stationary but initially unknown. In anticipation of this development we have treated the single-stage optimization process in some detail. We shall see in the next chapter that it forms a good starting point for the study of the dual control of discrete state Markov processes.

CHAPTER 3

A THEORY OF DUAL STRATEGIES FOR SINGLE-STAGE MARKOV PROCESSES

3.1 The Concept of Dual Control

In this chapter we shall study the control of stationary N-state Markov batch processes, as defined in section 2.7. We shall assume that B = 0, i.e. there is no cost of control. From (2.27) and (2.30), we see that optimal control of such processes is implemented by choosing the state s such that

$$\ell_s = \underset{j}{\text{Min}} [\ell_j]$$

and setting $D^* = \underline{w} > < \underline{e}_s$.

In other words the optimal policy is, "after every transition, re-set the process state so that the next transition starts from state s."

If P is unknown, then so are the values ℓ_i. These can of course be estimated by conducting a number of trials, and observing the results. How should this be done? Suppose we observe one transition from each state, and then forever afterwards choose as the initial state the

one which yielded the minimum one-sample mean. Common sense, as well as statistical theory, rejects such a policy, since the probability of error would be high, and would never approach zero. Suppose alternatively that we conduct a large fixed number of trials from each state, so that our estimates of g are good (low variance), and the error probability is small. We might then choose the estimated minimum cost state thereafter. This strategy, while seemingly logical, is doubly undesirable: first, the cost incurred while estimating might be extremely high, and second, the probability of error would approach a positive non-zero limit once the initial estimation phase was terminated.

The drawback of both of the foregoing strategies is that they fail to integrate the simultaneous requirements of estimation and control. The initial estimation phase may be regarded as process perturbations which are followed by control actions. Because they are essentially parallel in nature, if not in time, the initial estimates make inefficient use of the information available. It is a commonplace fact of decision theory that a sequential policy in which future trials are governed by past results, is generally superior to a parallel search procedure. From a control standpoint, the initial estimation phase is an

open loop schedule, while a sequential policy is a closed loop system; information feedback is used. Properly designed, a sequential decision strategy thus performs a dual control function.

A word about nomenclature might be useful at this point. We shall use the word "policy" to denote a stationary feedback control function calculated, as in chapter 2, when P is known. A strategy is a method of adaptive control used when P is initially uncertain; if successful, the adaptive strategy coincides asymptotically with the optimal policy.

To further illustrate the concept of a dual strategy, we shall consider a simple example. A man travels to work each day by public transport. He lives in a large city, and has available to him half a dozen feasible alternate routes, using some combination of walking, bus, and underground railway; each incurs substantially the same monetary cost. The traffic situation being what it is, the man considers bus arrivals and point-to-point travelling times to be random variables with more or less unknown parameters. He wishes to find by experiment the route which incurs minimum expected travelling time. Since the experiments and the process (travelling to work) are essentially the same thing, there is no question of minimizing the number

of experiments required to reach a decision. Rather, the man wants to find the quickest route and, in so doing, avoid as far as possible having to try slow routes as part of the experiment. How will he proceed?

First of all, he will intuitively reject the open loop estimation schedule mentioned previously; it is unlikely, for instance, that he will try each route in turn for, say, a month at a time, and then make a final decision. He will probably try each route once or twice, and then concentrate more on the apparently best route, and less on the unpromising ones. He will not settle down to one particular route until he is quite confident that it is best. Even then, he will occasionally try the others, both to confirm his earlier conclusions, and to detect any non-stationarity in traffic patterns. A similar game is played quite successfully by automobile commuters.

Whether he realizes it or not, our hypothetical man is operating as an adaptive system exercising dual control. In this chapter we shall formulate a control strategy which is similarly adaptive in a stationary Markov process. As in the foregoing example, we would expect such a strategy to try all control actions in a more or less unbiased fashion early in the life of the process, but to concentrate increasingly upon the most promising one as time progresses.

At no finite time, however, would such a strategy allow one alternative to be chosen permanently to the exclusion of all others. Such a procedure implies certainty, and certainty comes only with infinite time in a stationary process, never in a non-stationary one.

3.2 The Dual Control Requirement

Let

$$\mu_i = \sum_{j=1}^{N} p_{ij}\, c_{ij} \tag{3.1}$$

State s is defined by

$$\mu_s = \text{Min } [\mu_1, \mu_2, \ldots \mu_N], \text{ assumed unique} \tag{3.2}$$

If P is unknown, then we may use the estimate $\hat{P} = \left\{\hat{p}_{ij}\right\}$ so that

$$\hat{\mu}_i = \sum_{j=1}^{N} \hat{p}_{ij}\, c_{ij} \tag{3.3}$$

State $\hat{s}$ is defined by

$$\hat{\mu}_{\hat{s}} = \text{Min } [\hat{\mu}_1, \hat{\mu}_2, \ldots \hat{\mu}_N] \tag{3.4}$$

It is possible that the estimated minimum cost state, $\hat{s}$,

is not the true minimum cost state. We therefore define the conditional probability of error, ω, as

$$\omega = \text{prob}\left[\hat{s} \neq s \;\middle|\; \text{all past observations}\right] \qquad (3.5)$$

The convergence requirement of any adaptive strategy demands that

$$\lim_{n\to\infty} \omega(n) = 0 \qquad (3.6)$$

where n = number of observed transitions.
Equation (3.6) is an estimation requirement. We shall see in section 3.4 that if transitions are observed from each state an indefinitely large number of times, the probability of error will approach zero. This condition can be achieved by a variety of schemes, such as controlling so that transitions occur from each state an equal number of times. Such a strategy would give a good estimate of P, but could hardly be called adaptive. If we are to have an adaptive strategy which is of any use at all, we must specify the control requirement as well, viz.

$$\lim_{n\to\infty} \frac{n_i}{n_s} = 0, \qquad i \neq s \qquad (3.7)$$

where n_i = number of transitions observed from state i.
Equation (3.7) specifies that after many transitions have occurred, most of the past control actions should be correct

in retrospect. Equation (3.6) and (3.7) are equivalent to (1.4) and (1.5) respectively. It might be said that the two of them specify minimum requirements for an adaptive controller. We have seen in chapter 1, though, that an additional requirement of the ideal dual strategy is that a given level of error probability must be reached at the minimum possible cost. After a large number of observed transitions the total cost incurred, V_T, is[‡]

$$V_T = \sum_{i=1}^{N} n_i \hat{\mu}_i = \sum_{i=1}^{N} [n_i \hat{\mu}_{\hat{s}} + n_i(\hat{\mu}_i - \hat{\mu}_{\hat{s}})]$$

$$V_T = n\,\hat{g}^* + \sum_{i=1}^{N} n_i(\hat{\mu}_i - \hat{\mu}_{\hat{s}}) \qquad (3.8)$$

The total cost consists then of two parts

1) $n\hat{g}^*$, the estimate of the minimum cost which could have been incurred if the estimated optimum policy had been used;

2) $\hat{V} = \sum_{i=1}^{N} n_i(\hat{\mu}_i - \hat{\mu}_{\hat{s}})$, the estimated "learning cost".

If the probability of error after n transitions is

[‡] Let $V_T = \sum_{i=1}^{N} V_i$, where $V_i = \sum_{j=1}^{N} m_{ij} c_{ij}$; ($m_{ij}$ as in (3.11))

From (3.3) and (3.11), $\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{N} m_{ij} c_{ij} = \frac{V_i}{n_i}$; (3.8) follows.

$\omega(n) = \omega_f$, then the ideal dual strategy is that for which the set $\{n_1, n_2, \ldots n_N\}$

$$\left.\begin{array}{ll} \text{minimizes} & \hat{V} = \sum_{i=1}^{N} n_i(\hat{\mu}_i - \hat{\mu}_{\hat{s}}) \\ \text{subject to} & \omega(\hat{P}, C, n_1, n_2, \ldots n_N) = \omega_f \\ \text{and} & n_i > 0, \quad i = 1, 2, \ldots N \end{array}\right\} \quad (3.9)$$

Such a strategy is not realizable, but its performance may be approached in practice, as we shall see. Before attempting to solve the constrained minimization (3.9) it is necessary to consider a means of computing the error probability.

3.3 Statistical Estimation of Process Parameters

The problem of estimating the transition probabilities, p_{ij}, of a discrete Markov process is essentially that of determining the parameters of a set of N multinomial distributions, each of order N. Strictly speaking, the estimates, $\tilde{p}_{ij}$, form a multidimensional beta distribution, and the likelihood function of μ_i is obtained by integration over a bounded hypervolume of dimension N-2. Use of this approach has been found computationally impractical for N greater than three or four. Further details are given in appendix 2.

We may take advantage, however, of the fact that the large sample distribution of $\tilde{p}_{ij}$ is multivariate normal[2,5]. Since estimation in our process is to continue indefinitely we are justified, at least asymptotically, in using such a distribution. Suppose that a total of n transitions has been observed in the past; n_i of these originated from state i, and m_{ij} occurred from state i to state j. Thus

$$n = \sum_{i=1}^{N} n_i = \sum_{i=1}^{N} \sum_{j=1}^{N} m_{ij} \qquad (3.10)$$

Elements of row i, of the stochastic matrix, P, have maximum likelihood values

$$\hat{p}_{ij} = \frac{m_{ij}}{n_i} \qquad (3.11)$$

Associated with each row, $[\hat{p}_{i1}, \ldots. \hat{p}_{iN}]$, is a covariance matrix, $Q^{(i)}$. It is symmetrical, with diagonal elements

$$q_{jj}^{(i)} = \frac{p_{ij}(1-p_{ij})}{n_i} \qquad (3.12)$$

and off-diagonal elements

$$q_{jk}^{(i)} = -\frac{p_{ij}p_{ik}}{n_i}, \quad j \neq k \qquad (3.13)$$

Note that there exist N covariance matrices $Q^{(i)}$, $Q^{(2)}$, ... $Q^{(N)}$. We shall in general consider the rows of P and C to be independent, which means that the matrices

$Q^{(i)}$ and the estimators $\tilde{\mu}_i$ are also independent. This assumption allows the piecewise constant function μ_i, $i = 1, 2, \ldots N$, to take on an arbitrary shape, e.g. it may be multimodal. If μ_i is known to be unimodal, the assumption may be modified with considerable saving in overall cost of operation, as we shall see in section 4.7.

Let $< \hat{\underline{p}}_i$ = row vector $[\hat{p}_{i1}, \ldots \hat{p}_{iN}]$ of matrix $\hat{P}$

$< \underline{c}_i$ = row vector $[c_{i1}, \ldots c_{iN}]$ of matrix C

Then $\tilde{\mu}_i$, the estimate of the expected one-stage cost of a transition from state i, is normally distributed with maximum likelihood value

$$\hat{\mu}_i = < \hat{\underline{p}}_i \quad \underline{c}_i > \tag{3.14}$$

and variance

$$\sigma_i^2 = < \underline{c}_i \; Q^{(i)} \; \underline{c}_i > \tag{3.15}$$

i.e.
$$\sigma_i^2 = \frac{1}{n_i} \Big[\sum_{j=1}^{N} p_{ij}(1-p_{ij})c_{ij}^2 - 2 \sum_{j=1}^{N-1} \Big(\sum_{k=j+1}^{N} p_{ij}p_{ik}c_{ij}c_{ik} \Big) \Big] \tag{3.16}$$

The varance, σ_i^2, is correct for any sample size, but the distribution is normal only for large n_i. If P is uncertain, then so are $Q^{(i)}$ and σ_i^2. In such a case it is necessary to approximate σ_i^2 by the estimate $\hat{\sigma}_i^2$. The latter is obtained by using $\hat{Q}^{(i)}$ in (3.15), where $\hat{Q}^{(i)}$ is determined using $\hat{p}_{ij}$ from (3.11) in place of p_{ij} in equations (3.12) and (3.13). The estimated probability density function of μ_i (the likelihood function) is given by

$$f_i(x) = \frac{1}{\sqrt{2\pi}\ \hat{\sigma}_i} \exp\left[-\frac{1}{2}\left(\frac{x-\hat{\mu}_i}{\hat{\sigma}_i}\right)^2\right] \qquad (3.17)$$

We may now consider the computation of the error probability. We note first that the foregoing equations are based on a large sample assumption. In order to make estimates early in the life of the process, we introduce a measure of the error probability, termed the uncertainty. The latter is the error probability computed using the assumption of normally distributed estimates. Both measures are, of course, asymptotically identical. Denoting the uncertainty as Ω, we see that

$$\Omega = 1 - \text{prob}\left[\mu_{\hat{s}} = \underset{i}{\text{Min}}\left\{\mu_i\right\} \,\middle|\, C,\ M\right] \qquad (3.18)$$

where

M = matrix of observations, m_{ij}

$\mu_{\hat{s}}$ = true mean cost associated with state $\hat{s}$.

In words, the situation is this: on the basis of past observations, we believe that state $\hat{s}$ is the minimum cost state. Ω is the probability, conditional on these observations, that we are wrong.

The probability, $1-\Omega$, that $\hat{s} = s$, is the integral over all x of the compound probability

$$\text{prob } [\mu_{\hat{s}} = x \text{ and all other } \mu_i\text{'s} > x]$$

If the rows of P and C are independent

$$\Omega = 1 - \int_{-\infty}^{\infty} f_{\hat{s}}(x) \prod_{\substack{i=1 \\ i\neq s}}^{N} \left[\int_{x}^{\infty} f_i(y_i)\, dy_i \right] dx$$

Let $$G_i(x) = \int_{x}^{\infty} f_i(y)\, dy \qquad (3.19)$$

Then

$$\Omega = 1 - \int_{-\infty}^{\infty} f_{\hat{s}}(x) \prod_{\substack{i=1 \\ i\neq \hat{s}}}^{N} [G_i(x)]\, dx \qquad (3.20)$$

3.4 Properties of Normal Likelihood Functions

Two properties of normal likelihood functions which are relevant to the estimation-control scheme under consideration are stated below, and proved in appendix 3. If a unique minimum, μ_s, exists, and if the estimates, $\tilde{\mu}_i$, are normally distributed with finite stationary means and variances, then

1) If across an ensemble of statistically equivalent processes the mean value of Ω at stage n is $\Omega(n)$, and a trial is carried out in each process from a particular non-optimal state i ($i \neq s$), then at the $(n+1)^{th}$ stage,

$$E[\Omega(n+1)] < E[\Omega(n)] \qquad (3.21)$$

In other words, if we pay the price of choosing a non-minimum cost state for estimation purposes, we may expect a positive return in the form of decreased uncertainty.

2) $$\lim_{n\to\infty} \Omega(n) = 0$$

if and only if

$$n \longrightarrow \infty \Longrightarrow n_i \longrightarrow \infty \quad \forall_i \qquad (3.22)$$

Property 2) is a necessary condition of convergence; it states that in a convergent process, n_i has no upper

limit with time, even though state i is non-optimal. At first sight this condition seems rather difficult to reconcile with the requirement of equation (3.7). The answer must be that the growth of n_s dominates that of n_i, $i \neq s$. We shall see later that this is so.

3.5 A Realizability Problem

Let us now return to the constrained minimisation, (3.9); taking results across an ensemble, we may use the true values μ_i rather than the estimates $\hat{\mu}_i$. Since the optimum set $\{n_1, n_2 \ldots n_N\}$ will be shown to lie within the non-negativity constraints of (3.9), we may safely ignore these. Using a Lagrange multiplier, λ, we may form the adjoined cost function, V_λ.

$$V_\lambda = \sum_{i=1}^{N} n_i(\mu_i - \mu_s) + \lambda(\Omega - \Omega_f)$$

Setting $\frac{\partial V_\lambda}{\partial n_i} = 0$ we obtain N equations

$$(\mu_i - \mu_s) + \lambda \frac{\partial \Omega}{\partial n_i} = 0 \qquad (3.23)$$

while $\partial V_\lambda / \partial \lambda = 0$ gives the original constraint condition. For $i = s$, (3.23) becomes

$$\lambda \frac{\partial \Omega}{\partial n_s} = 0$$

so that either $\lambda = 0$, or $\partial\Omega/\partial n_s = 0$ or both are zero. Examination of (3.23) for $i \neq s$ in the light of (3.21) shows that for n_i finite, $\lambda \neq 0$. Therefore we conclude that a condition of optimality is

$$\frac{\partial \Omega}{\partial n_s} = 0 \qquad (3.24)$$

The only condition under which (3.24) is always satisfied is $n_s = \infty$. Given the requirements of the problem, this is an intuitively reasonable solution. Since choice of state s incurs no estimation cost, we should choose s enough times to reduce Ω to the minimum possible value; this is free information. We then turn to the other N-1 states and begin "buying" information. This condition seems of little use, though, when we try to control a real process, since if P is uncertain, we do not know s. Which state shall we choose an infinite number of times? If we are wrong, the additional cost will be infinite!

To handle this dilemma, we shall invoke the maxim "Pretend it doesn't exist, and maybe it will go away." Specifically, our approach will be as follows: We shall consider the pattern which emerges when we begin the decision process with $n_s = \infty$. Then we shall attempt to synthesize a realizable strategy which has the same asymptotic pattern.

3.6 The Bayesian Approach

Suppose that at some stage in the process we have chosen n_s a very large number of times ($n_s \to \infty$; we do not specify how we decided which was state s), and, to further decrease Ω we turn our attention to the remaining N-1 states. For $n_s \to \infty$, $f_s(x)$ in (3.17) becomes a delta function centred on μ_s; Ω then simplifies to

$$\lim_{n_s \to \infty} \Omega = 1 - \prod_{\substack{i=1 \\ i \neq s}}^{N} G_i(\mu_s) \qquad (3.25)$$

where $G_i(\mu_s) = G_i(x) \,\big|\, x = \mu_s$.

Let

$\hat{\mu}_i(n_i)$ = maximum likelihood estimate of μ_i made after observation of n_i transitions from state i.

$\hat{\mu}_i(n_i+1 \,|\, n_i)$ = <u>predicted</u> value of the estimate $\hat{\mu}_i(n_i+1)$ made after observation of only n_i transitions from state i.

$\hat{\sigma}_i^2(n_i)$ = maximum likelihood estimate of the variance of $\hat{\mu}_i(n_i)$ made after observation of n_i transitions from state i.

$\hat{\sigma}_i^2(n_i+1|n_i)$ = predicted value of the estimate $\hat{\sigma}_i^2(n_i+1)$ made after observation of only n_i transitions from state i.

The total cost incurred in n_i transitions from state i is

$$V_i = \sum_{j=1}^{N} m_{ij}c_{ij} = n_i\hat{\mu}_i(n_i)$$

since

$$\hat{\mu}_i(n_i) = \frac{1}{n_i} \sum_{j=1}^{N} m_{ij}c_{ij}$$

where $$n_i = \sum_{j=1}^{N} m_{ij}$$

If one more transition from state i is observed, the maximum likelihood prediction of the cost of that transition is by definition $\hat{\mu}_i(n_i)$ since the process is stationary. The prediction of the estimate of mean cost after $n_i + 1$ transitions from state i have been observed is

$$\hat{\mu}_i(n_i+1|n_i) = \frac{1}{n_i+1} \text{ [Total observed cost of } n_i \text{ transitions + Predicted cost of one more transition]}$$

$$\hat{\mu}_i(n_i+1|n_i) = \frac{1}{n_i+1} [n_i\hat{\mu}_i(n_i) + \hat{\mu}_i(n_i)]$$

$$\hat{\mu}_i(n_i+1 \mid n_i) = \hat{\mu}_i(n_i) \tag{3.26}$$

Now consider the estimate of variance. Let us define

$$\hat{\sigma}_{oi}^2(n_i) = \sum_{j=1}^{N} \hat{p}_{ij}(1-\hat{p}_{ij})c_{ij}^2 - 2\sum_{j=1}^{N-1}\left(\sum_{k=j+1}^{N} \hat{p}_{ij}\hat{p}_{ik}c_{ij}c_{ik}\right)$$

where $\hat{p}_{ij} = \hat{p}_{ij}(n_i) = \frac{m_{ij}}{n_i}$.

We see from (3.16) that $\hat{\sigma}_{oi}^2(n_i)$ is the estimate (with n_i observations) of the ensemble variance of the cost of one transition from state i, and that

$$\hat{\sigma}_i^2(n_i) = \frac{\hat{\sigma}_{oi}^2(n_i)}{n_i} \tag{3.27}$$

In a stationary process σ_i^2, unlike μ_i, is non-stationary; it decreases as n_i increases, signifying that our knowledg of $\hat{\mu}_i$ becomes more precise as more observations are made. However, σ_{oi}^2 is stationary, so that

$$\hat{\sigma}_{oi}^2(n_i+1 \mid n_i) = \hat{\sigma}_{oi}^2(n_i)$$

From (3.27) we obtain

$$\hat{\sigma}_i^2(n_i+1|n_i) = \frac{1}{n_i+1}\,\sigma_{oi}^2(n_i+1|n_i)$$

$$\hat{\sigma}_i^2(n_i+1|n_i) = \frac{n_i}{n_i+1}\,\hat{\sigma}_i^2(n_i) \qquad (3.28)$$

Using (3.25), (3.26), and (3.28) we can compute the expected change in Ω if n_i is increased by one unit.

$$E\left[\frac{\Delta\Omega}{\Delta n_i}\right] = -\frac{\Delta G_i(\mu_s)}{\Delta n_i} \prod_{\substack{k=1 \\ k\neq i \\ k\neq s}}^{N} G_k(\mu_s) \qquad (3.29)$$

At a given stage in the process, we wish to select the state, j, which is expected to yield the maximum decrease in Ω per unit change in $\hat{V}$, the estimated learning cost. Let $\Delta\hat{V}_i \equiv (\hat{\mu}_i-\mu_s)\,\Delta n_i$. Then j is chosen so that

$$E\left[-\frac{\Delta\Omega}{\Delta\hat{V}_j}\right] = \underset{i}{\text{Max}}\left[\frac{1}{(\hat{\mu}_i-\mu_s)}\ \frac{\Delta G_i(\mu_s)}{\Delta n_i}\ \prod_{\substack{k=1\\k\neq i\\k\neq s}}^{N} G_k(\mu_s)\right] \qquad (3.30)$$

A control decision is then made so that the next transition is initiated from state j; suppose the transition is from j to k. The transition cost, c_{jk}, is observed, and the posterior parameters of the state j are computed, i.e. our knowledge of the process is updated.

$$\hat{\mu}_j^{\,n_j+1\,|\,n_j+1} = \frac{n_j}{n_j+1}\ \hat{\mu}_j^{\,n_j+1\,|\,n_j} + \frac{1}{n_j+1}\ c_{jk} \qquad (3.31)$$

The estimates $\hat{p}_{j1}$, ... $\hat{p}_{jN}$ are updated using (3.11). $\hat{\sigma}_{oj}^{\,2}$ is then up dated with (3.16), and $(\hat{\sigma}_j^{\,2})^{n_j+1\,|\,n_j+1}$ is computed from (3.27). The cycle is then repeated: calculation of prior estimates with (3.26) and (3.28), decision based on (3.30), and updating as outlined. In the following section, we shall consider the pattern which emerges with continued repetition of this cycle.

3.7 The Continuity Approximation

In sections 3.8-3.12, we shall consider results pertaining to an ensemble of processes for which all n_i are

large. We shall therefore use the true values, μ_i and σ_i^2, averaged across the ensemble, instead of the estimates. While n_i can of course possess only discrete values in any one process, it is useful to regard it as continuous when considering an ensemble. Suppose we have an ensemble of processes, each with the same stochastic matrix, P. In one half of the ensemble $n_i = 20$, and in the other half $n_i = 21$. We shall consider the situation to be described by a single ensemble with $n_i = 20.5$. In more general terms, let $\alpha(n_i)$ be a parameter depending upon n_i (e.g. $\alpha = G_i(\mu_s)$). At the k^{th} stage of the process let the ensemble average of α be

$$\alpha^{(k)} = \alpha(n_i)$$

At the $(k+1)^{th}$ stage state i is chosen in a fraction "b" of the members of the ensemble, so that n_i becomes n_i+1 in that fraction, and remains as n_i in the others (in the remaining fraction, 1-b, some other state is chosen). At stage (k+1) the ensemble average of α is thus

$$\alpha^{(k+1)} = b[\alpha(n_i+1)] + (1-b)\,[\alpha(n_i)]$$

$$= \alpha(n_i) + b[\alpha(n_i+1) - \alpha(n_i)]$$

$$\alpha^{(k+1)} = \alpha(n_i) + b\,\frac{\Delta\alpha}{\Delta n_i} \qquad (3.32)$$

The right hand side of (3.32) approximates a Taylor expansion of $\alpha(n_i+b)$ if $\Delta\alpha/\Delta n_i$ is small and $n_i >> 1$. For $\alpha = G_i(\mu_s)$ or $\alpha = \Omega$, the chief parameters of interest, the approximation is valid for large n_i.

3.8 Decision Space and the Hill of Uncertainty

It is useful to describe a sequence of past transitions by the set $\{n_1, n_2, \ldots n_N\}$. By postulating an N-dimensional decision space with coordinates $n_1, n_2, \ldots n_N$, we can plot the evolution of a decision process. Corresponding to every point $\underline{n}$ in the space is a value of Ω determined by equation (3.20). Successive decisions therefore represent a decision trajectory descending the hypersurface Ω, the hill of uncertainty. Descent of the hill implies convergence of the decision process, and there are as many convergent adaptive control algorithms as there are paths down the hill. The most efficient path is the one which reaches the lowest point on the hill for a given cost, or, conversely, minimizes the cost of reaching a given value of Ω. This path is of course the one generated by the ideal dual strategy.

As explained in sections 3.5 and 3.6, it is assumed that we are starting the decision process with $n_s \to \infty$. No change in Ω occurs with any further increase in n_s.

Therefore we may disregard the ordinate corresponding to n_s, and consider the projection of the N space into N-1 space at $n_s = \infty$. More generally, if the minimum occurs with multiplicity K, then the pertinent subspace has dimension N-K. In the following development, however, we may consider that K = 1 without significant loss of generality.

Since the cost function is linear, contours of constant cost V are hyperplanes defined by

$$V = \sum_{\substack{i=1 \\ i \neq s}}^{N} n_i(\mu_i - \mu_s) \qquad (3.33)$$

Contours of the Ω hill are non-linear, and the hill is concave. Fig. 3.1 shows an example of contours of V and Ω in a three state system; it is assumed that μ_1 is the minimum of $\{\mu_1, \mu_2, \mu_3\}$, and that $n_1 \to \infty$.

3.9 The Optimal Trajectory

We define an optimal trajectory in decision space as the locus of points which minimize V for every value of ξ. It is the purpose of this section to investigate the asymptotic properties of such a trajectory.

From (3.25) we have in N-1 space

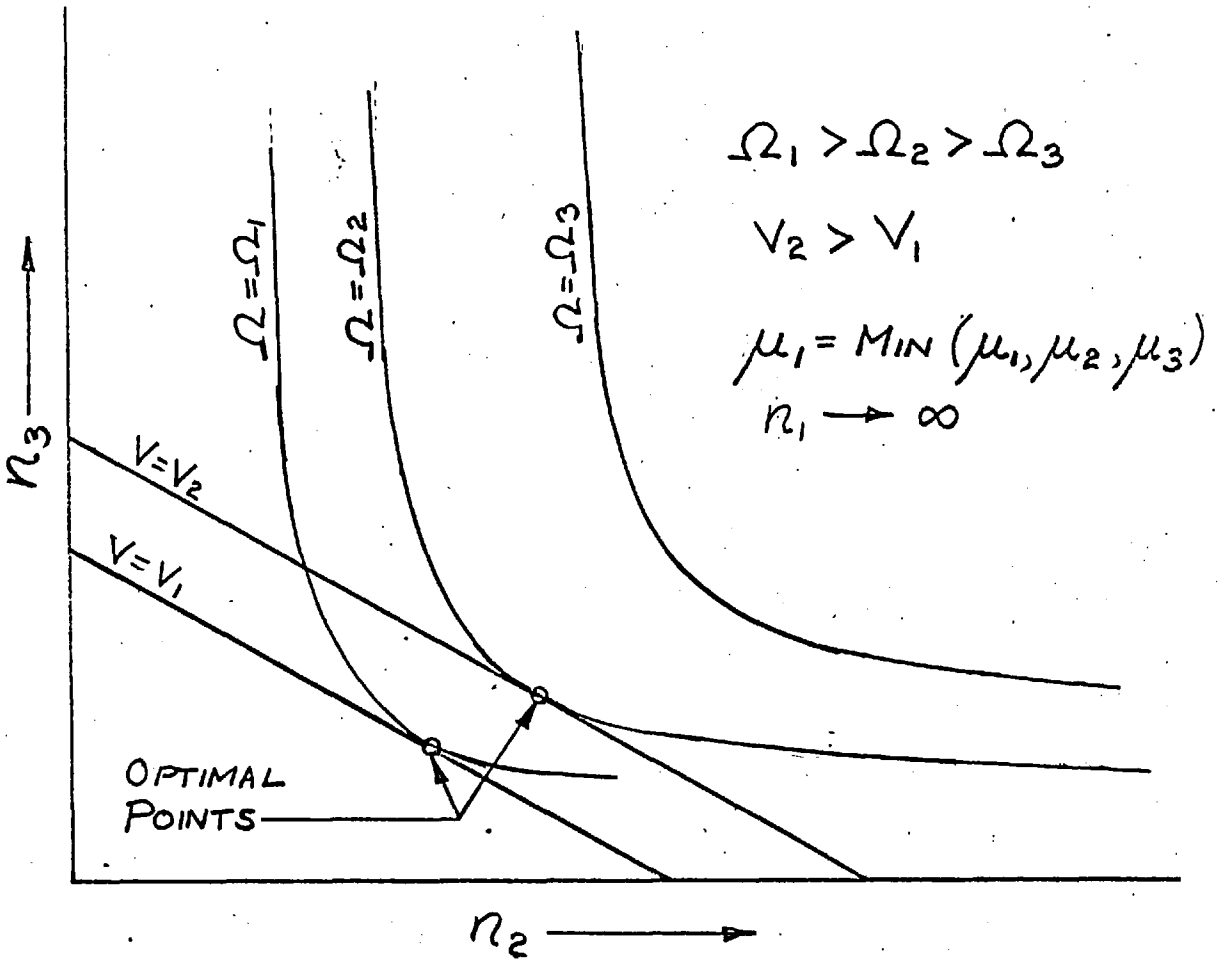


FIG. 3.1

THE HILL OF UNCERTAINTY

(THREE STATE SYSTEM)

$$\Omega = 1 - \prod_{\substack{i=1 \\ i \neq s}}^{N} [1 - F_i(\mu_s)] \qquad (3.34)$$

where $F_i(\mu_s)$ is the cumulative distribution of the normal likelihood function

$$F_i(\mu_s) = 1 - G_i(\mu_s) = \int_{-\infty}^{x_{\mu_s}} f_i(y)dy \qquad (3.35)$$

$F_i(\mu_s)$ may be expressed by the asymptotic series

$$F_i(\mu_s) = \frac{1}{\sqrt{2\pi}\,\rho_i} \exp(-\frac{\rho_i^2}{2}) \left[1 - \frac{1}{\rho_i^2} + \frac{1.3}{\rho_i^4} - \frac{1.3.5}{\rho_i^6} + \ldots\right] \qquad (3.36)$$

where $$\rho_i = \frac{\mu_i - \mu_s}{\sigma_i} . \qquad (3.37)$$

Let $$\rho_{oi} = \frac{\mu_i - \mu_s}{\sigma_{oi}}$$

so that, since $\sigma_{oi}^2 = n_i \sigma_i^2$,

$$\rho_i = n_i^{1/2} \rho_{oi} \qquad (3.38)$$

Since ρ_i increases monotonically with n_i, we can truncate (3.36) for large n_i to

$$\lim_{n_i \to \infty} F_i(\mu_s) = \frac{1}{\sqrt{2\pi}\,\rho_i} \exp(-\frac{\rho_i^2}{2}) \qquad (3.39)$$

Since $F_i(\mu_s)$ becomes small for large n_i, we may write (3.34) as

$$\lim_{\substack{n_i \to \infty \\ \forall\, i}} \Omega = \sum_{\substack{i=1 \\ i \neq s}}^{N} F_i(\mu_s) \qquad (3.40)$$

Substituting (3.39) we have

$$\lim_{\substack{n_i \to \infty \\ \forall\, i}} \Omega = \frac{1}{\sqrt{2\pi}} \sum_{\substack{i=1 \\ i \neq s}}^{N} \frac{1}{\rho_i} \exp\left(-\frac{\rho_i^2}{2}\right) \qquad (3.41)$$

Differentiation by n_i gives

$$\frac{\partial \Omega}{\partial n_i} = -\frac{1}{2\sqrt{2\pi}\rho_i} \exp\left(-\frac{\rho_i^2}{2}\right) \left[\frac{1}{n_i} + \rho_{oi}^2\right]$$

Substituting (3.38) we obtain

$$\lim_{\substack{n_i \to \infty \\ \forall\, i}} \frac{\partial \Omega}{\partial n_i} = -\frac{\rho_{oi}}{2\sqrt{2\pi n_i}} \exp\left(-\frac{\rho_i^2}{2}\right) \qquad (3.42)$$

Let us now recall the optimality requirement expressed by (3.23)

$$(\mu_i - \mu_s) + \lambda \frac{\partial \Omega}{\partial n_i} = 0 \qquad (3.23)$$

$$i = 1, 2, \ldots N$$

It follows that for any two states, i and j ($i, j \neq s$) on an optimal trajectory

$$\frac{1}{(\mu_i-\mu_s)}\ \frac{\partial\Omega}{\partial n_i} = \frac{1}{(\mu_j-\mu_s)}\ \frac{\partial\Omega}{\partial n_j} \qquad (3.43)$$

Combining (3.42), (3.43), and (3.37) we have

$$\frac{\exp(-\rho_i^2/2)}{\exp(-\rho_j^2/2)} = \left(\frac{n_i}{n_j}\right)^{\frac{1}{2}}\ \frac{\sigma_{oi}}{\sigma_{oj}} \qquad (3.44)$$

The value of the normal likelihood function at the point $x = \mu_s$ can be expressed as

$$f_i(\mu_s) = \frac{n_i^{1/2}}{\sqrt{2\pi}\ \sigma_{oi}}\ \exp(-\frac{\rho_i^2}{2}) \qquad (3.45)$$

From (3.44) and (3.45) we obtain the <u>asymptotic optimality criterion</u>

$$\frac{n_i}{n_j} = \frac{f_i(\mu_s)}{f_j(\mu_s)} \qquad i,\ j \neq s \qquad (3.46)$$

We can thus define an <u>optimal point</u> in N-1 space as one for which the ratio of any pair of ordinates is equal to the ratio of the corresponding likelihood functions evaluated at $x = \mu_s$.

It is interesting to note that the same result can be

obtained from the Bayes strategy with continuous n_i. Equation (3.30) becomes

$$-\frac{\partial \Omega}{\partial V_j} = \max_i \left| \frac{\partial \Omega}{\partial V_i} \right| \qquad (3.47)$$

where $V_i = n_i(\mu_i - \mu_s)$

Since from (3.42) $\partial \Omega / \partial V_i$ is a monotonically decreasing function of n_i, optimality implies that for all i and j other than s

$$\frac{\partial \Omega}{\partial V_i} = \frac{\partial \Omega}{\partial V_j} \qquad (3.48)$$

which is equivalent to (3.43), so that (3.46) follows.

3.10 Characteristic Vector of A Decision Process

Of great importance in the theory of optimum decision making is the asymptotic value of the ratio n_i/n_j along an optimal trajectory. From (3.44)

$$\rho_j^2 - \rho_i^2 = \log\left(\frac{n_i}{n_j}\right) + \log\left(\frac{\sigma_{oi}^2}{\sigma_{oj}^2}\right)$$

Rearranging and applying (3.38) gives

$$\rho_j - \rho_i = \frac{\log\left(\frac{n_i}{n_j}\right) + \log\left(\frac{\sigma_{oi}^2}{\sigma_{oj}^2}\right)}{n_i^{1/2}\rho_{oi} + n_j^{1/2}\rho_{oj}}$$

$$\lim_{\substack{n_i\to\infty \\ n_j\to\infty}} \rho_j - \rho_i = \lim_{\substack{n_i\to\infty \\ n_j\to\infty}} \frac{\log\left(\frac{n_i}{n_j}\right) + \log\left(\frac{\sigma_{oi}^2}{\sigma_{oj}^2}\right)}{n_i^{1/2}\rho_{oi} + n_j^{1/2}\rho_{oj}}$$

$$= 0$$

so that

$$\lim_{\substack{n_i\to\infty \\ n_j\to\infty}} \left(\frac{n_i}{n_j}\right)_{\text{optimal}} = \frac{\rho_{oj}^2}{\rho_{oi}^2} \qquad (3.49)$$

Equation (3.49) shows that for large n_i, $i = 1, 2, \ldots N$, all points on an optimal trajectory in N-1 space are colinear. The vector joining any two points on the asymptotic trajectory is a scalar multiple of a vector $\underline{e}$, which we shall call the characteristic vector of the decision process.

$$\underline{e} = [e_1, e_2, \ldots e_{s-1}, e_{s+1}, \ldots e_N]^T$$

defined by

$$\frac{e_i}{e_j} = \frac{\rho_{oj}^2}{\rho_{oi}^2} \qquad (3.50)$$

and

$$\sum_{\substack{i=1 \\ i \neq s}}^{N} e_i = 1 \qquad (3.51)$$

3.11 The Meaning and Uses of the Characteristic Vector

The elements e_i of the characteristic vector of a decision process represent the asymptotic value of the relative frequency of choice of state i ($i \neq s$) with an ideal decision strategy. Asymptotically, for every time state i is chosen, state j should be chosen (e_j/e_i) times; in a randomized decision scheme, if the probability of choosing state i is ϕ_i, then the probability of choosing

state j should be $(\frac{e_j}{e_i} \phi_i)$ for any states i and j other than state s. In a convergent process ϕ_i approaches zero as n, the number of observed transitions, approaches infinity, but the ratios (e_i/e_j) remain constant. Later in this chapter we shall see that, for a certain class of strategies each n_i $(i \neq s)$ may be expressed as a harmonic series in terms of n; in such a case the elements e_i represent multipliers which determine the relative magnitudes of equivalent terms in the set of N-1 simultaneous series.

If we look at N-1 space geometrically, we may regard the vector $\underline{e}$ as an inverted ridge in the hill of uncertainty. In chapter 4 we shall introduce algorithms which determine theoretical optimum points and trajectories in decision space. If one point on an optimum trajectory is known, then other points in the vicinity can be found simply by searching along a vector $\underline{e}$. It will be seen that this concept results in a large reduction of computing time, and makes feasible computations which would otherwise be impracticable.

Finally the characteristic vector is the bond which ties together a whole class of strategies to be considered in the remainder of this chapter.

3.12 The Inverse Problem

Consider a problem slightly more general than (3.9). Suppose we wish to minimize

$$V = \sum_{\substack{i=1 \\ i \neq s}}^{N} a_i \, n_i \qquad (3.52)$$

subject to $\Omega(\hat{P}, C, n_1, \ldots n_N) = \Omega_f$ and $n_i > 0$, $i = 1, 2, \ldots N$, where each parameter a_i is a finite positive constant coefficient. We are now costing wrong decisions (i.e. decisions which turn out a posteriori to have been incorrect) not by an amount $\mu_i - \mu_s$, but according to an arbitrary price, a_i. As in the former problem, though, no cost is attached to choice of state s. We might follow through the developments of sections 3.5-3.9 with this new problem. Equation 3.23 becomes

$$a_i + \lambda \frac{\partial \Omega}{\partial n_i} = 0$$

The optimality condition (3.43) becomes

$$\frac{1}{a_i} \frac{\partial \Omega}{\partial n_i} = \frac{1}{a_j} \frac{\partial \Omega}{\partial n_j}$$

and (3.44) is generalized to

$$\frac{\exp(-\rho_i^2/2)}{\exp(-\rho_j^2/2)} = \frac{a_i}{a_j} \left(\frac{n_i}{n_j}\right)^{1/2} \frac{\rho_{oj}}{\rho_{oi}} \qquad (3.53)$$

Now comes a discovery of interest: providing all a_i's are finite and positive, equation (3.49) remains unchanged. The new decision strategy possesses the same characteristic vector as the original one. In other words any straight line parallel to $\underline{e}$ in N-1 space is an asymptotically ideal decision strategy for some set of costs $\{a_1, a_2, \cdots a_{s-1}, 0, a_{s+1}, \cdots a_N\}$. We can thus postulate an infinite family of parallel trajectories in decision space. Immediately we are confronted with the inverse problem:

Given a straight line in N-1 decision space parallel to the vector $\underline{e}$, find the cost function for which the line is an ideal decision strategy.

Let the straight line passing through the origin and having direction $\underline{e}$ be designated as the basic trajectory. If $\underline{n}$ is any point on it (see fig. 3.2), then $\underline{n} = \alpha\underline{e}$, where α is a positive scalar. A point $\underline{n}'$ on a trajectory parallel to the first one is given by

$$\underline{n}' = \alpha\underline{e} + \underline{h}_i \qquad (3.54)$$

where $\underline{h}_i$ = offset vector with respect to the i^{th} coordinate. The components of $\underline{h}_i$, $h_{1i}, h_{2i}, \cdots h_{(s-1)i}, h_{(s+1)i}, \cdots h_{Ni}$ are defined by

$$h_{ji} = n_j' - \frac{\rho_{oi}^2}{\rho_{oj}^2} n_i \qquad i, j \neq s \qquad (3.55)$$

and $h_{ii} = 0$.

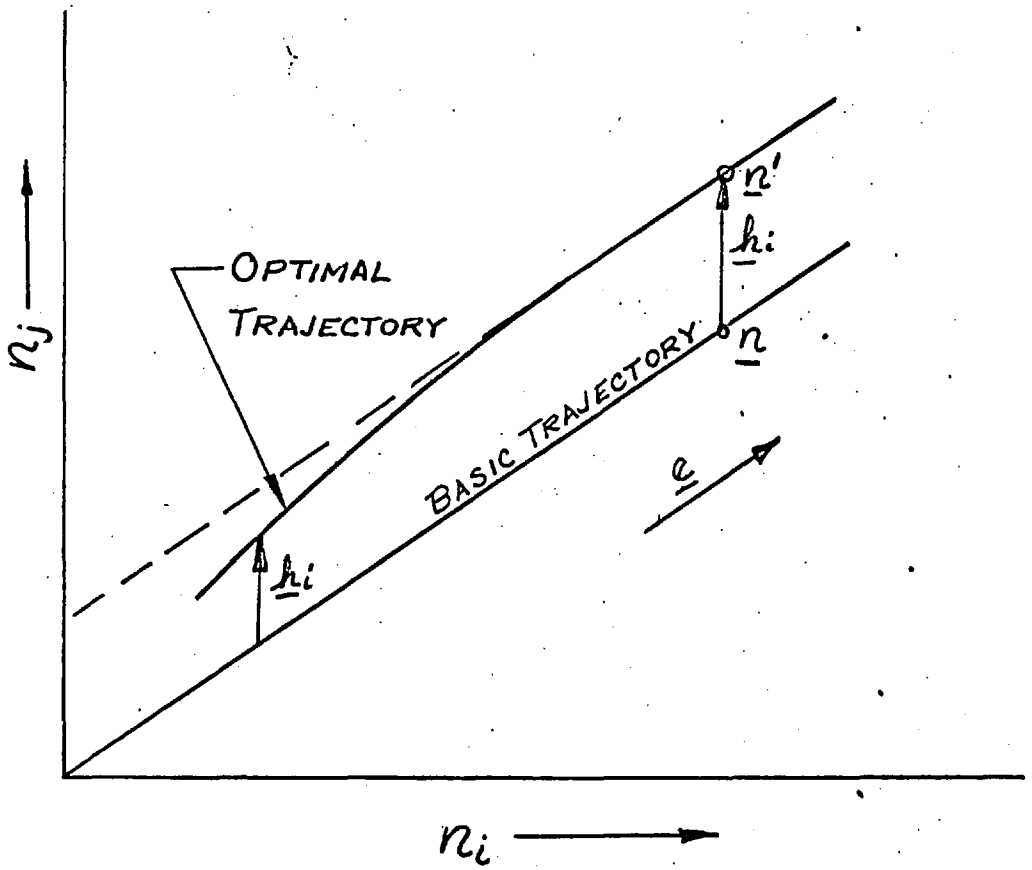


FIG. 3.2

RELATIONSHIP OF TRAJECTORIES IN DECISION SPACE

Combining (3.55) with (3.53) and recalling that $\rho_i^2 = n_i \rho_{oi}^2$, we obtain

$$\frac{\exp(-\rho_i^2/2)}{\exp(-\rho_j^2/2)} = \exp\left(\frac{\rho_{oj}^2 h_{ji}}{2}\right)$$

$$= \left(\frac{a_i}{a_j}\right)\left(\frac{\rho_{oj}}{\rho_{oi}}\right)\left[\frac{\rho_{oj}^2}{\rho_{oi}^2} + \frac{h_{ji}}{n_i}\right]^{\frac{1}{2}} \qquad (3.56)$$

For large n_i, $\underline{h}_i$ is independent of n_i, and

$$\frac{a_j}{a_i} = \left(\frac{\rho_{oj}}{\rho_{oi}}\right)^2 \exp\left(-\frac{\rho_{oj}^2 h_{ji}}{2}\right) \qquad (3.57)$$

By expressing the costs a_i and a_j in terms of the components of the offset vector, $\underline{h}_i$, and the statistical parameters of the process, equation (3.57) solves the inverse problem. It will be seen in sections 3.15 and 3.16 that this solution is useful in evaluating various proposed realizable strategies.

3.13 Suboptimal Strategies

In a given system, we shall define a <u>suboptimal strategy</u> as one which exhibits the following three properties:

1) $$\lim_{n\to\infty} \Omega = 0 \qquad (3.58)$$

$$2) \qquad \lim_{n\to\infty} \frac{\partial n_i}{\partial n_s} = 0 , \qquad i \neq s \qquad (3.59)$$

$$3) \qquad \lim_{n\to\infty} \frac{\partial n_i}{\partial n_j} = \frac{\rho_{oj}^2}{\rho_{oi}^2} , \qquad i,j \neq s \qquad (3.60)$$

Observe the meaning of these three equations. (3.58) implies convergence and (3.59) implies that the suboptimal strategy eventually becomes the optimal policy. These two equations are the minimum requirements of any adaptive strategy, as outlined in section 1.5. Equation (3.60), based on (3.49), is a new way of expressing the dual control requirement (3.9); this is the criterion which distinguishes between strategies which are merely convergent, and those whose performance approaches that of the ideal dual strategy.

The term "suboptimal" has been used since we include in (3.60) any strategy whose decision trajectory is asymptotically parallel to the optimal decision trajectory. It can be shown that, for any given value of Ω, the estimation cost associated with a suboptimal trajectory differs from the cost associated with the optimal trajectory by an amount which is independent of Ω. For any other decision scheme, this cost disadvantage depends upon Ω, and eventually becomes more and more marked as Ω decreases. Any strategy which can be shown to be suboptimal

will therefore be almost as effective in practice as the optimal one.

3.14 Realizable Control Strategies

The significance and usefulness of the inverse problem and the concept of a suboptimal strategy lie in the fact that we now have a theoretical framework which can be used to evaluate any proposed algorithm for the dual control of long duration discrete Markov processes of the batch type. In practice this means that, given the complete decision rule for such an algorithm, we can determine its trajectory in decision space, and so determine whether or not it meets the requirement of suboptimality.

Developments thus far have yielded asymptotic optimality criteria. However, nothing has been said about the awkward problem of making decisions early in the life of the process, when s is unknown and all n_i are small. In the sections which follows, we shall use the foregoing theory to examine several realizable decision strategies.

3.15 Equal Rho Strategy

We shall begin with a brief examination of a strategy

proposed by Pashkovskii[35]. As mentioned in section 1.4, he postulated an N-state stochastic process in which the transition probabilities, p_{ij}, may be affected by choice of a control signal $u_k(i)$ associated with state i, where u_k is itself a member of a discrete set $u_1, \ldots u_N$. The object of control is to choose a control input $u_k^x(i)$ for each state, i, which maximizes the one-stage probability, $p_{i\ell}$, of transition from state i to a particular known state, ℓ. If the transition $i \to \ell$ takes place, the cost is zero; for any other transition $i \to j$, $j \neq \ell$, the cost is a fixed amount c. The transition matrix, P, is initially unknown, and the process is to be repeated indefinitely.

The dual strategy Pashkovskii suggested is the following: at each stage, compute the confidence intervals associated with each probability estimate, $\widetilde{p_{i\ell}}(i, u_k)$, $k = 1, 2, \ldots N$, and choose that control u_s which maximizes the upper confidence limit of $\widetilde{p_{i\ell}}$.

Pashkovskii did not explain why he chose this strategy, nor how he chose the confidence limits. We shall present here a brief analysis of the strategy in the light of the theory of decision trajectories. Since the cost of transition to any state other than ℓ is the same, the system is essentially a two-state one. The estimate $p_{i\ell}(u_k) = p_{i\ell}^k$ has mean and variance

$$\hat{\mu}_k = \hat{p}_{i\ell}^{\kappa} \tag{3.61}$$

$$\hat{\sigma}_k^2 = \hat{p}_{i\ell}(1 - \hat{p}_{i\ell}) \tag{3.62}$$

Choosing k to maximize $p_{i\ell}^{\kappa}$ is equivalent to choosing the minimum of $1 - \mu_k$, $k = 1, \ldots N$. Let the confidence interval associated with the μ_k have bounds

$$\hat{\mu}_k - m\hat{\sigma}_\kappa \quad \text{and} \quad \mu_k + m\hat{\sigma}_\kappa$$

where, for a confidence level $\emptyset$, m is given by

$$\frac{1}{\sqrt{2\pi}} \int_{-m}^{m} \exp(-\frac{x^2}{2})\, dx = \emptyset$$

To determine $\underset{k}{\text{Min}}[1-\mu_k]$ according to the Pashkovskii strategy, we choose $k = \hat{s}$ to minimize $\hat{\mu}_k - m\hat{\sigma}_k$.

We observe that each time an alternative k is chosen, the value of $(\hat{\mu}_k - m\hat{\sigma}_k)$ will tend to change in a positive direction, since an increase in n_k decreases $\hat{\sigma}_k$. As the various alternative controls are chosen early in the process, all of the lower confidence limits will gradually increase until they reach the limit $\hat{\mu}_{\hat{s}}$. Once $\hat{\mu}_k - m\hat{\sigma}_k > \hat{\mu}_{\hat{s}}$, $k \neq \hat{s}$, alternative k is no longer chosen. Across an ensemble this situation is reached when

$$\mu_k - m\,\sigma_k = \mu_s, \qquad k \neq s$$

so that $$\rho_k = \frac{\mu_k - \mu_s}{\sigma_k} = m \qquad \forall\, k,\ k \neq s \tag{3.63}$$

Since all ρ_k's are equal, and since $\rho_k = n_k^{1/2}\rho_{ok}$, we see that across an ensemble

$$\lim_{n\to\infty} \frac{n_j}{n_k} = \frac{\rho_{ok}^2}{\rho_{oj}^2}, \qquad j,\ k \neq s \tag{3.64}$$

Equation (3.64) is not true for every individual process unless Ω tends to zero with increasing n. If m is fixed, Ω does not tend to zero, but to a limiting value, Ω_f

$$\Omega_f \simeq \frac{N-1}{\sqrt{2\pi}} \int_{-\infty}^{-m} \exp(-\frac{x^2}{2})\, dx$$

$$\Omega_f \simeq \frac{N-1}{2}(1-\phi) \tag{3.65}$$

The strategy as stated has two disadvantages: first, as we have seen, it is not convergent for any fixed value of m. Second, it is unsuitable if the cost matrix C takes on a more general form. Note that $\hat{\mu}_k$ in (3.61) is not a function of the cost matrix elements. To minimize the expected transition cost we must set

$$\hat{\mu}_k = \hat{\mu}_k(i) = \sum_{j=1}^{N} \hat{p}_{ij}^{\,k}\, c_{ij}$$

For the particular cost matrix used, $c_{i\ell} = 0$ and $c_{ij} = c$, $j \neq \ell$, so that

$$\hat{\mu}_k = (1 - \hat{p}_{i\ell}^{k})\, c \qquad (3.66)$$

and

$$\hat{\sigma}_k^{\,2} = p_{i\ell}^{k}(1 - p_i^{k})\, c^2 \qquad (3.67)$$

Maximization of (3.61) is thus equivalent to minimization of (3.66). If a general cost matrix is used, this equivalence cannot be established, and the strategy is not necessarily of any usefulness.

Pashkovskii's strategy can be generalized quite easily to overcome these difficulties. Its basic premise is simple and effective: choose from amongst the non-optimum states so that all $\hat{\rho}_i$'s remain approximately equal. Let us therefore express the strategy as follows:

Choose state $\hat{s}$ probabilistically, with probability $1-\Omega(n)$; if state s is not chosen, choose that state from the remaining N-1 states for which $\hat{\rho}_i$ is a minimum. Since an increase in n_i tends to increase $\hat{\rho}_i$, all $\hat{\rho}_i$'s tend to become equal as the decision process continues. Equation (3.60) is thus satisfied. The proof that $\Omega(n)$

does tend to zero, (3.58), and (3.59) hold, and the policy is suboptimal is given in section 3.18.

Let us consider the inverse question: when is an equal rho strategy optimal? We note that its trajectory in decision space is extremely simple; an equal rho trajectory is a basic trajectory, i.e. a straight line passing through the origin with direction $\underline{e}$ as defined by (3.50) and (3.51). Since all h_{ji}'s are zero, we see from (3.57) that

$$\frac{a_j}{a_i} = \frac{\rho_{oj}^2}{\rho_{oi}^2}$$

The strategy is optimal when

$$\frac{a_j}{a_i} = \frac{\mu_j - \mu_s}{\mu_i - \mu_s} \qquad (3.68)$$

i.e.
$$\frac{\upsilon_{oi}^2}{\mu_i - \mu_s} = \frac{\sigma_{oj}^2}{\mu_j - \mu_s}, \qquad i, j \neq s \qquad (3.69)$$

An example is a process in which the minimum mean is zero, and all other states possess variances proportional to their means.

3.16 Probabilistic Strategy

It is interesting to consider the use of a decision strategy which is wholly probabilistic. In the early

stages of this work the author postulated such a strategy on heuristic grounds, and found it quite effective. In this section we shall demonstrate that it is suboptimal in an ensemble sense; the convergence of each individual process is proved in section 3.18. The strategy is the following:

Choose each state i probabilistically, with probability ϕ_i, where ϕ_i is the conditional probability that μ_i is the minimum of the set $\{\mu_1, \mu_2, \cdots \mu_N\}$.

From (3.17) and (3.19),

$$\phi_i = \int_{-\infty}^{\infty} f_i(x) \prod_{\substack{j=1 \\ j\neq i}}^{N} G_j(x)\, dx \qquad (3.70)$$

As all n_i become large, the product $\prod_{\substack{j=1 \\ j\neq i}}^{N} G_j(x)$ tends toward a reversed step function at $x = \mu_s$, so that

$$\lim_{n\to\infty} \phi_i = \int_{-\infty}^{\mu_s} f_i(x)\, dx = F_i(\mu_s), \qquad (3.71)$$

$$i \neq s$$

To show that the strategy is suboptimal we shall solve the equation of ensemble convergence. We may write

$$\frac{d\phi_i}{dn} = \frac{d\phi_i}{dn_i} \frac{dn_i}{dn} \qquad (3.72)$$

where $n = \sum_{i=1}^{N} n_i$.

The term $d\phi_i/dn_i$ is found from (3.71) and (3.39) to be

$$\lim_{n\to\infty} \frac{d\phi_i}{dn_i} = -\frac{\rho_{oi}^2}{2}\,\phi_i, \qquad i \neq s \tag{3.73}$$

The term dn_i/dn is specified by the decision strategy. In this case

$$\frac{dn_i}{dn} = \phi_i, \qquad i \neq s \tag{3.74}$$

From (3.72), (3.73), and (3.74), we obtain the asymptotic convergence equation of the probabilistic strategy. For each state i, $i \neq s$, the result is a degenerate Riccati equation, viz.

$$\frac{d\phi_i}{dn} = -\frac{\rho_{oi}^2}{2}\,\phi_i^2 \tag{3.75}$$

The asymptotic solution of this equation has an important property: it is independent of the initial value of ϕ_i. Thus atypical results early in the process do not disturb the ultimate decision trajectory. The probability ϕ_i assumes an asymptotic value

$$\lim_{n\to\infty} \phi_i = \frac{2}{\rho_{oi}^2 n}, \qquad i \neq s \tag{3.76}$$

From (3.41) we obtain

$$\lim_{n\to\infty} \Omega = \frac{2}{n} \sum_{\substack{i=1 \\ i\neq s}}^{N} (\rho_{oi}^{-2}) \tag{3.77}$$

We now show that (3.58)-(3.60) hold, so that the policy is suboptimal across an ensemble. First, from (3.77)

$$\lim_{n\to\infty} \Omega = 0$$

Second, from (3.74) and (3.76)

$$\lim_{n\to\infty} \frac{\partial n_i}{\partial n_s} = \frac{\phi_i}{\phi_s} = 0$$

Third, from (3.74) and (3.76)

$$\lim_{n\to\infty} \frac{\partial n_i}{\partial n_j} = \frac{\phi_i}{\phi_j} = \frac{\rho_{oj}^2}{\rho_{oi}^2}, \quad i, j \neq s.$$

The probabilistic strategy is suboptimal across an ensemble. We may now ask the inverse question, when is the probabilstic strategy optimal? To answer the question, we note first that probabilistic decisions are defined by

$$\frac{\phi_i}{\phi_j} = \frac{F_i(\mu_s)}{F_j(\mu_s)} = \frac{\rho_{oj}^2}{\rho_{oi}^2}, \qquad i, j \neq s \quad (3.78)$$

From (3.39) we see that

$$\lim_{n\to\infty} \frac{F_i(\mu_s)}{F_j(\mu_s)} = \frac{\rho_{oj}}{\rho_{oi}} \left(\frac{n_j}{n_i}\right)^{\frac{1}{2}} \frac{\exp(-\rho_i^2/2)}{\exp(-\rho_j^2/2)} \quad (3.79)$$

However, the optimality criterion (3.57) demands that

$$\lim_{n\to\infty} \frac{\exp(-\rho_i^2/2)}{\exp(-\rho_j^2/2)} = \exp\left(\frac{\rho_{oj}^2 h_{ji}}{2}\right)$$

$$= \frac{a_{ji}}{a_j} \left(\frac{\rho_{oj}}{\rho_{oi}}\right)^2 \quad (3.80)$$

From (3.78), (3.79), (3.80) and (3.49) we see that

$$\left(\frac{a_i}{a_j}\right)\left(\frac{\rho_{oj}}{\rho_{oi}}\right)^2 = \frac{\rho_{oj}}{\rho_{oi}}\left(\frac{n_i}{n_j}\right)^{\frac{1}{2}} = \left(\frac{\rho_{oj}}{\rho_{oi}}\right)^2$$

Therefore

$$\frac{a_i}{a_j} = 1$$

Recalling (3.68) we see that the probabilistic strategy is optimal when

$$\mu_i = \mu_j, \qquad i, j \neq s \qquad (3.81)$$

The strategy is optimal when all mean costs are equal except that of the minimum cost state.

A discrete state dual control problem treated with a probabilistic strategy has been presented recently by Nikolic and Fu[34]. Their strategy is to update a set of subjective probabilities using a reinforcement rule which depends upon which control alternative, i, was used in the last interval, and whether $\hat{\mu}_i$ increased or decreased as a result of its use. The strategy is convergent in the sense that requirements (3.58) and (3.59) are met. However there seems to be no other connection between their subjective probabilities and the parameters ϕ_i used in the present strategy, so that the additional requirement (3.60) is

presumably not met. It would be interesting to plot the decision trajectory for an ensemble of processes using the Nikolic and Fu strategy. It seems probable that even the asymptotic trajectory in N-1 space would be influenced by search parameters chosen subjectively at the beginning of the process.

3.17 Optimal Strategy

A realizable strategy whose asymptotic decision trajectory is always optimal can be synthesized by combining the technique of probabilistic decision with the optimality criterion (3.46). The set of all states is divided into two subsets, the first containing state $\hat{s}$, and the second containing the other N-1 states. The control decision then breaks down into two stages:

1) Choose state $\hat{s}$ probabilistically, with probability $\theta_{\hat{s}}$, where

$$\begin{aligned} \theta_{\hat{s}} &= 1 - \gamma\alpha, && \gamma\alpha < 0.5 \\ &= 0.5, && \gamma\alpha \geq 0.5 \end{aligned} \qquad (3.82)$$

with $$\alpha = \sum_{\substack{i=1 \\ i \neq s}}^{N} F_i(\hat{\mu}_{\hat{s}})$$

If state $\hat{s}$ is chosen, control is applied so that the next transition is initiated from state $\hat{s}$. With $\gamma = 1$, $\theta_{\hat{s}} \simeq 1-\Omega$ for large n. The saturation function (3.82) is used to prevent θ_s from assuming very small or negative values early in the process. The effect of varying γ will be considered later.

2) if state $\hat{s}$ is not chosen, choose state j from the remaining N-1 states so that

$$\frac{\exp(-\hat{\rho}_j^2/2)}{\hat{\sigma}_{oj}\, n_j^{1/2}} = \underset{\substack{i=1,\ldots N \\ i\neq s}}{\text{Max}} \left[\frac{\exp(-\hat{\rho}_i^2/2)}{\hat{\sigma}_{oi}\, n_i^{1/2}}\right] \qquad (3.83)$$

From (3.42) it can be seen that (3.83) is an implementation of the Bayes strategy (3.48). It therefore follows from the properties of the optimal decision trajectory that (3.60) holds, i.e. the asymptotic trajectory in N-1 space is in the direction of $\underline{e}$. We shall proceed to show that the strategy also satisfies equations (3.58) and (3.59).

Note that the strategy is a hybrid one, partly probalistic and partly deterministic. We shall derive here the convergence equations for any hybrid policy which chooses state $\hat{s}$ probabilistically, and other states according to any criterion which satisfies (3.60). We begin by evaluating dn_i/dn in (3.72). From step 1) of the strategy we see that

$$\frac{dn_i}{dn} = \gamma\Omega \quad \frac{\Delta n_i}{\Delta \sum_{\substack{j=1 \\ j\neq i}}^{N} n_j}$$

where the Δ notation may be used because of the straight line trajectory in N-1 space. In view of (3.49) the above equation may be written

$$\frac{dn_i}{dn} = \frac{\gamma\Omega}{\rho_{oi}^2 \sum_{j=1\neq s}^{N} (\rho_{oj}^{-2})} \qquad (3.84)$$

Using the general cost coefficients a_i of equation (3.52), we may combine (3.39), (3.41), (3.49), and (3.53) to obtain the asymptotic relationship between ϕ_i and Ω

$$\lim_{n\to\infty} \frac{\phi_i}{\phi_j} = \frac{F_i(\mu_s)}{F_j(\mu_s)} = \frac{\rho_j}{\rho_i} \frac{\exp(-\rho_i^2/2)}{\exp(-\rho_j^2/2)}$$

$$= \left(\frac{n_j^{1/2}\rho_{oj}}{n_i^{1/2}\rho_{oi}}\right)\left(\frac{a_i}{a_j}\right)\left(\frac{n_i^{1/2}}{n_j^{1/2}}\right) \cdot \frac{\rho_{oj}}{\rho_{oi}}$$

$$= \frac{a_i}{a_j} \frac{\rho_{oj}^2}{\rho_{oi}^2}$$

so that

$$\Omega = \phi_i \frac{\rho_{oi}^2}{a_i} \sum_{j=1\neq\hat{s}}^{N} \frac{a_j}{\rho_{oj}^2} \tag{3.85}$$

We now obtain from (3.72), (3.73), (3.84), and (3.85)

$$\lim_{n\to\infty} \frac{d\phi_i}{dn} = -\left[\frac{\rho_{oi}^2}{2} \frac{\gamma}{a_i} \frac{\sum_{j=1\neq\hat{s}}^{N} (a_j \rho_{oj}^{-2})}{\sum_{j=1\neq\hat{s}}^{N} (\rho_{oj}^{-2})}\right] \phi_i^2$$

with a solution

$$\lim_{n\to\infty} \phi_i = \frac{2\, a_i}{n\gamma\rho_{oi}^2} \frac{\sum_{j=1\neq\hat{s}}^{N} (\rho_{oj}^{-2})}{\sum_{j=1\neq\hat{s}}^{N} (a_j \rho_{oj}^{-2})} \tag{3.86}$$

Ω is given by

$$\lim_{n\to\infty} \Omega = \sum_{i=1\neq\hat{s}}^{N} \phi_i = \frac{2}{n\gamma} \sum_{j=1\neq\hat{s}}^{N} (\rho_{oj}^{-2}) \tag{3.87}$$

Note that if $\gamma=1$, (3.87) is identical to the convergence equation (3.77) for probabilistic decision. The proof that (3.58) and (3.59) apply follows immediately as in the previous case. Equation (3.87) is the asymptotic

convergence equation for any hybrid strategy in which state s is chosen according to rule 1 of the optimal policy, and the other states are chosen by any rule which satisfies (3.60). The constant γ is called the convergence factor. If γ is large, convergence is fast and initial cost per transition is high; if γ is low convergence is slower, and the cost of estimation is spread over a longer period of time.

This strategy possesses another useful property: since $E[F_i(\mu_s)]$ is a monotonically decreasing function of μ_i, the expected cost per transition across an ensemble with the optimal strategy is a monotonically decreasing function of time. Similarly, the probability of choosing state s in the next transition increases monotonically with time.

To observe the manner in which n_i $(i \neq s)$ grows we may combine (3.84) and (3.87) to obtain

$$\lim_{n\to\infty} \frac{dn_i}{dn} = \frac{2}{n\rho_{oi}^2}, \qquad i \neq s$$

whence, if n_i is considered continuous, a change Δn in n gives rise to a change Δn_i in n_i where

$$\lim_{n\to\infty} \Delta n_i = \frac{2}{\rho_{oi}^2} \log(\Delta n) \qquad (i \neq s) \qquad (3.88)$$

If n_i is considered discrete, Δn_i becomes the sum of all the expected changes in n_i over Δn stages of the process (here Δn need not be small), so that

$$\lim_{n\to\infty} \Delta n_i = \frac{2}{\rho_{oi}^2} \sum_{k=n}^{n+\Delta n} \frac{1}{k} \quad (i \neq s) \qquad (3.89)$$

It is easy to show that (3.88) and (3.89) are equivalent for large n.

An interesting relationship exists between equation (3.89) and the characteristic vector, $\underline{e}$. We may re-write (3.89) as

$$\lim_{n\to\infty} \Delta n_i = \beta\, e_i \sum_{k=n}^{n+\Delta n} \frac{1}{k} \qquad (3.90)$$

where β = constant = $2 \sum_{j=1\neq\hat{s}}^{N} (\rho_{oj}^{-2})$

Thus the multipliers of the harmonic series governing the growth of n_i are proportional to the respective elements of the characteristic decision vector.

In many stochastic approximation schemes[8], step size is decreased according to a harmonic series. The reason for this choice is that the harmonic series is the fastest shrinking series of the type n^{-k} which is still divergent.

In a noisy hill-climbing scheme divergence implies an unlimited correction effort (from starting point to top of hill) if necessary. In a discrete Markov process harmonic divergence implies an unlimited estimation effort, which similarly guarantees convergence of the overall process. It is interesting that we have demonstrated from theoretical considerations the optimal properties of the harmonic series for a broad class of decision strategies.

3.18 Proof of Convergence

In the past three sections the policies dealt with have been shown to possess the suboptimal properties (3.58)-(3.60) in an ensemble sense. To show that these properties apply to every individual process, it is necessary to prove that property 2) of section 3.4 applies, i.e. that $n \to \infty$ implies $n_i \to \infty$ for all states $i = 1, 2, \ldots N$. According to the decision strategy the probability of choosing state $\hat{s}$ at any stage of the process is equal to or greater than the probability of choosing any of the other N-1 states. It therefore suffices to show that for any finite number $m = \sum_{i=1 \neq \hat{s}}^{N} n_i$, there exists a finite number n such that the probability of observing fewer than m transitions from the set of N-1 states estimated to be non-optimal ($i \neq \hat{s}$) during

a sequence of n stages can be made less than an arbitrarily small positive number, δ. Note that the estimate of which is state $\hat{s}$ may change as the process continues.

The uncertainty, Ω , takes on a succession of values Ω_k, k = 1, 2, ... n, where k is the number of stages which have elapsed; n is finite. Since Ω is infinitesimally small only when all n_i are infinite, we know that all values of Ω_k are finite positive numbers.

Let $\theta = \underset{k}{\text{Min}} \left\{ \Omega_k \right\}$

By definition $0 < \theta < 1$

Let the sequence of length n be divided into m equal intervals, each of length q = n/m, n being a multiple of m. Now if at each stage, k, state $\hat{s}$ is chosen with probability $1-\Omega_k$, then the probability, p_1, of observing no transitions from one of the other N-1 states in a sequence of length q is bounded by

$$p_1 \leq (1-\theta)^q$$

The probability, p_2, of observing at least m such transitions in the sequence of m intervals is bounded by

$$p_2 \geq [1 - (1-\theta)^q]^m$$

$$p_2 \geq (1-\epsilon)^m > 1 - m\epsilon \qquad (3.91)$$

where $\epsilon = (1-\theta)^q$.

The probability of observing fewer than m transitions from estimated non-optimal states is thus $1-p_2$; it is necessary to show that $1-p_2$ can be made less than some arbitrarily small positive number δ. From (3.91)

$$1 - p_2 < m\epsilon$$

To prove that $1-p_2 < \delta$, we substitute $\epsilon = (1-\theta)^q$ and $q = n/m$, and show that $m\epsilon < \delta$. Convergence follows if there exists a finite number n such that

$$(1-\theta)^{\frac{n}{m}} < \frac{\delta}{m} \qquad (3.92)$$

where $0 < \theta < 1$

$0 < \delta$

$0 < m < \infty$

$0 < n < \infty$

Inspection of (3.92) shows that since θ and δ are greater than zero and m is finite, a sufficiently large finite n can always be found which will ensure the inequality. Thus there is no finite limit on m, and therefore none on n_i. Since all n_i's increase without limit, all estimates, $\hat{p}_{ij}$ and $\hat{\mu}_i$, tend towards their true values, p_{ij} and μ_i respectively, and the decision strategy converges. Equations (3.58)-(3.60) therefore apply asymptotically to each individual process.

3.19 Summary

In this chapter we have considered in some detail the interesting case in which control of a repetitive single-stage discrete Markov process is to be carried out when the dynamics, as embodied in the matrix P, are initially unknown. In the presence of uncertainty concerning the optimal decision matrix, the controller must perform the dual function of estimation and control. Added to the cost of operation with the optimal policy, therefore, is an additional "learning cost" associated with the estimation necessary to determine the optimum policy. Too great an emphasis on estimation, i.e. a thorough "exploration" of the dynamics, defeats the selective purpose of control. On the other hand, a pure control strategy based on present estimates does little to reduce the uncertainty, and may prove non-convergent.

Rather than attempt to steer between these two dangers by some heuristic means, we have set up an additional criterion which the dual controller must attempt to satisfy. The ideal adaptive controller is that which minimizes the learning cost associated with every error probability. It must therefore attempt to satisfy a constrained statistical minimization at every stage of the process.

What is sought is a decision strategy which converges to the optimal decision policy in an optimal (low cost) fashion. It has been shown that the ideal dual strategy is non-realizable. Nevertheless, a study of its asymptotic properties has allowed us to construct a framework within which we may evaluate the performance of dual algorithms. We have postulated an N-1 dimensional decision space whose coordinates are the quantities n_i (number of transitions observed from state i), $i \neq s$. A series of decisions defines a decision trajectory descending a hill of uncertainty. An optimal trajectory eventually becomes a straight line in decision space; the vector defining its direction is the characteristic decision vector of the process. If we can construct a realizable strategy whose decision trajectory coincides asymptotically with that of the ideal strategy, we shall then have an asymptotically optimal strategy. Analysis shows that evan a strategy whose asymptotic trajectory is parallel to, but not coincident with, the ideal trajectory is very nearly as good as the optimal strategy; such parallel strategies are called "suboptimal". The synthesis of two different suboptimal strategies, together with an optimal strategy, has been demonstrated.

Since the constrained minimization (3.9) performed by

the controller yields an indeterminate value of n_s ($n_s \to \infty$), the designer has a free choice of one parameter. In the formulation considered, this is the convergence factor; it may be regarded as a time-scaling factor affecting the convergence rate of the uncertainty, Ω . The choice of convergence factor is somewhat analogous to the choice of step length when hill climbing with a gradient technique.

The three realizable strategies presented are each partly probabilistic in form. For policies of this type it has been shown that the quantities n_i, $i \neq s$, grow as a set of harmonic series. In this problem as in many others involving extremum-seeking in a stochastic environment, seemingly contradictory requirements are reconciled by the paradoxical quality of the harmonic series, whose ever-decreasing terms sum to infinity.

We review here the principal properties of the optimal strategy

1) Convergence

$$\lim_{n\to\infty} \Omega = 0$$

$$\lim_{n\to\infty} \frac{\partial n_i}{\partial n_s} = 0, \qquad i \neq s$$

2) Optimality

For $n \to \infty$, the set $\{n_1, \ldots n_N\}$ minimizes E[V] for every value of Ω .

3) Ensemble Monotonicity

$$E[g(n+1)] < E[g(n)]$$

It should be pointed out that because of the indeterminacy of n_s there exist many realizable strategies which will satisfy 1) and 2), each differing in the manner in which state s is chosen. The particular strategy described in this chapter is put forward for three reasons:

1) it is computationally simple;

2) g is monotonic across an ensemble;

3) it is easily described in probabilistic terms, and its convergence rate is predictable.

In probabilistic terms we may consider the current decision matrix, D(n) as being made up of identical rows $< \underline{d}_i = [d_{i1}(n), \ldots\ldots d_{iN}(n)]$ where

$$d_{i\hat{s}}(n) = 1 - \Omega \qquad (3.93)$$

$$d_{ij}(n) = \Omega e_j, \qquad j \neq \hat{s} \qquad (3.94)$$

e_j = j^{th} component of the characteristic decision vector in N-1 space.

CHAPTER 4

COMPUTATIONAL METHODS AND RESULTS IN SINGLE-STAGE MARKOV PROCESSES

4.1 Introduction

This chapter is the companion piece of Chapter 3; in it we shall verify the theory of Chapter 3 numerically, and demonstrate its application. To verify the theory we shall compare simulated results using the (realizable) optimal strategy, with those which would be achieved using the (non-realizable) ideal strategy. To compute the latter efficiently, we shall use the normal approximation of the parameter estimates, i.e. we shall approximate the error probability, ω, by the uncertainty, Ω. Given the posterior estimates $\left\{\hat{\mu}_i(n)\right\}$ and $\left\{\hat{\sigma}_{oi}^2(n)\right\}$, and given the fact that the system has reached a level of uncertainty $\Omega = \Omega_f$, the ideal strategy determines the set of choices $\left\{n_1^*, \ldots\ldots n_N^*\right\}$ which <u>would have</u> minimized the cost of reaching $\Omega = \Omega_f$ if the posterior estimates had been available a priori. The development of the computational techniques associated with the ideal strategy is therefore first considered.

The results of the simulation of an ensemble of three-state processes is then presented in detail. It is shown that the performance of the optimal strategy closely approaches that of the ideal one. As an example, the control of a twenty-level batch chemical process is simulated, and the effects of various prior assumptions are demonstrated.

4.2 Optimal Trajectories in (N-1) Space

The simplest computation is that of an optimal decision trajectory taken across an ensemble in N-1 space, given the statistical parameters of the system and assuming that $n_s \to \infty$. The trajectory may be computed as a series of steps Δn_i, using the Bayes approach of section 3.6. The computational flow diagram is shown in fig. 4.1; fig. 4.2 shows the resulting trajectory for a three-state system in which

$$P = \begin{bmatrix} 0.3 & 0.3 & 0.4 \\ 0.6 & 0.3 & 0.1 \\ 0.7 & 0.1 & 0.2 \end{bmatrix}; \quad C = \begin{bmatrix} 6 & 14 & 10 \\ 14 & 12 & 3 \\ 16 & 12 & 8 \end{bmatrix}$$

so that

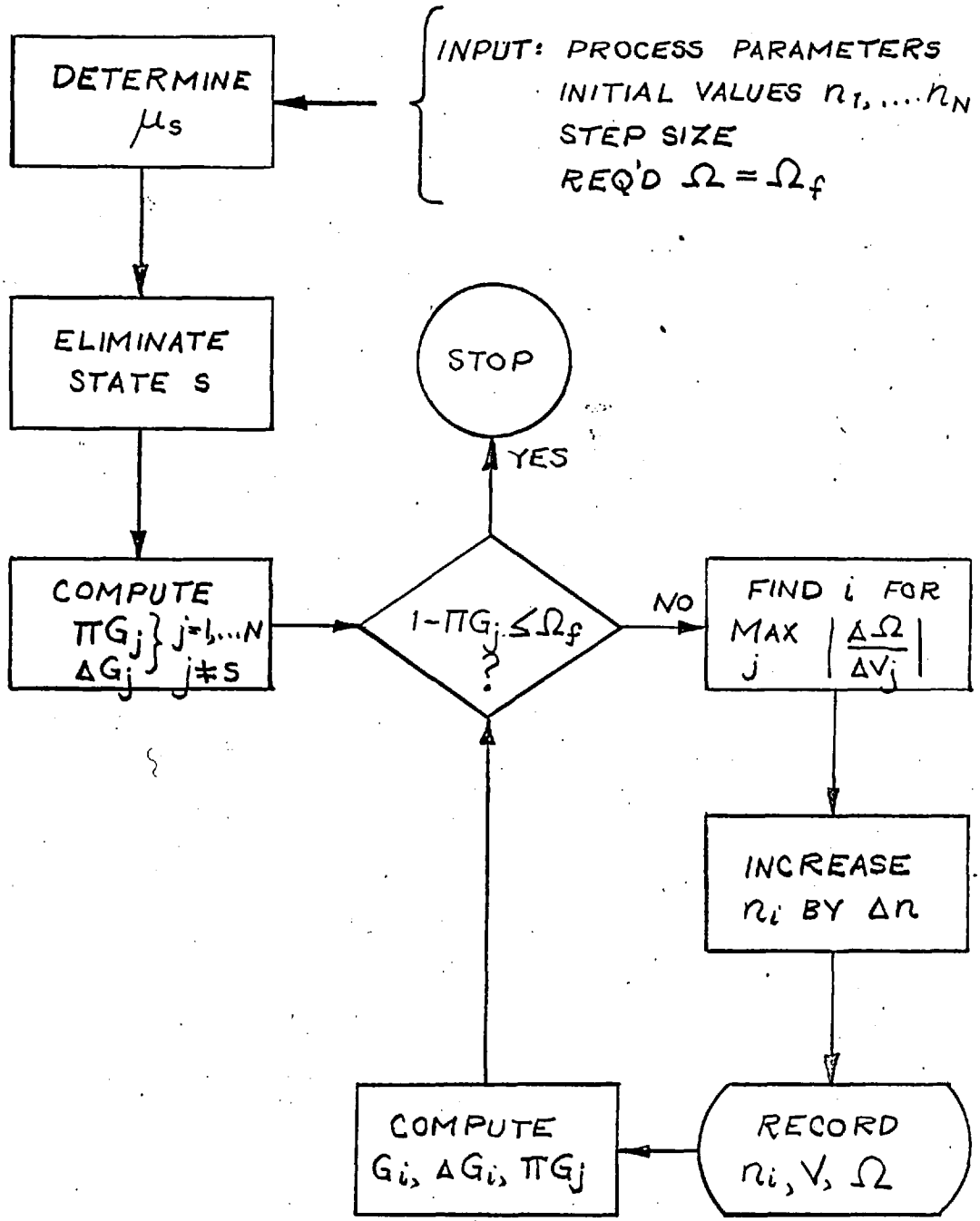


FIG. 4.1

COMPUTATION OF AN OPTIMAL DECISION TRAJECTORY IN (N-1) SPACE

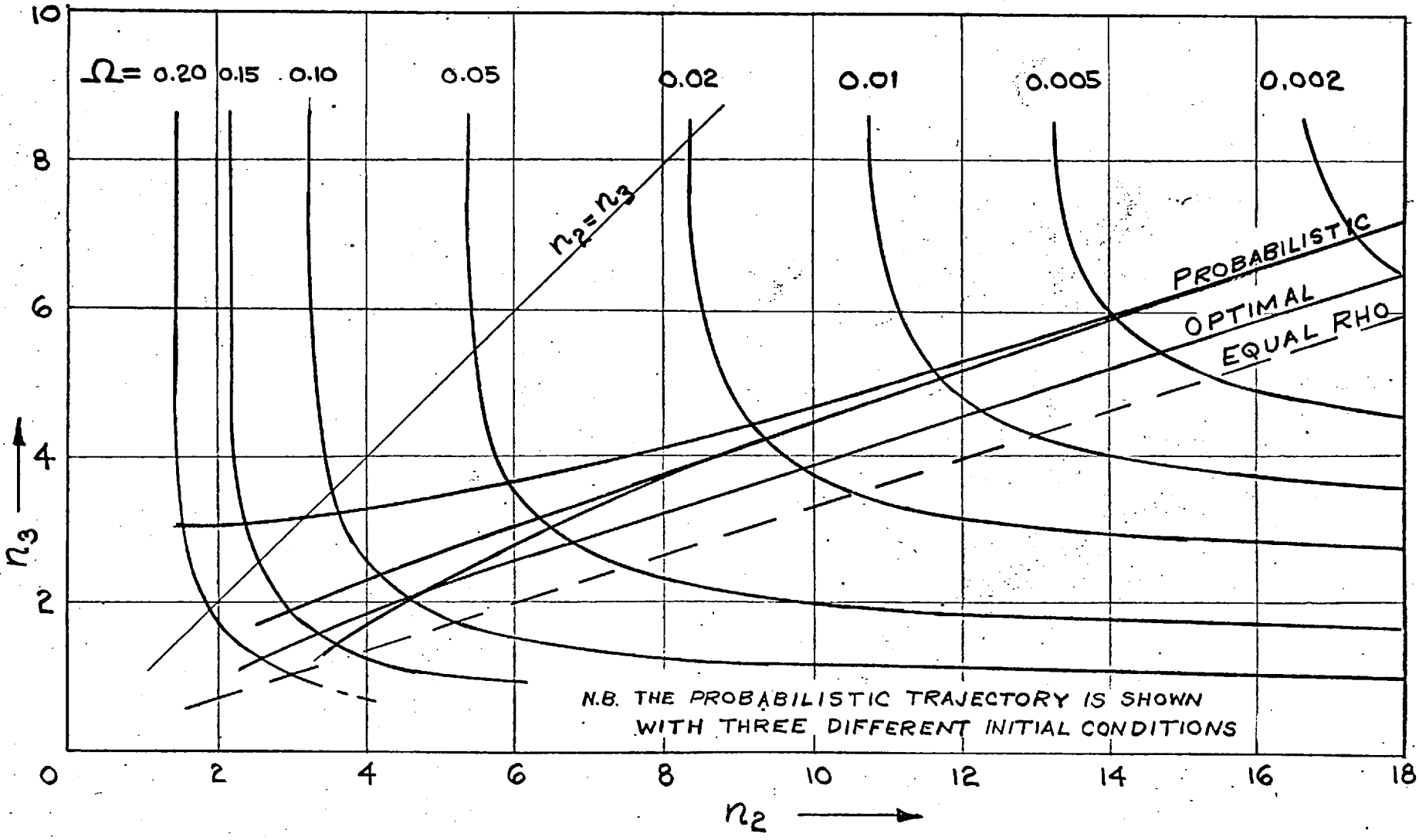


FIG. 4.2 TRAJECTORIES IN DECISION SPACE

$\mu_1 = 10.0$ $\quad\quad$ $\sigma_{o1}^2 = 9.60$ $\quad\quad$ $n_1 \to \infty$

$\mu_2 = 12.3$ $\quad\quad$ $\sigma_{o2}^2 = 10.41$

$\mu_3 = 14.0$ $\quad\quad$ $\sigma_{o3}^2 = 10.40$

The increment Δn was chosen as 0.20. To obtain an accurate hill shape, the contours of Ω were computed from equation (3.25), valid for all n_i, rather than the asymptotic version (3.41).

Suppose, as an example, that after a long run the final estimate of parameters is as given above, and that the uncertainty associated with the estimate of minimum cost state is 0.02. Had these parameters been assumed at the beginning of the decision process, what set $\{n_1, n_2, n_3\}$ would we have chosen to verify that μ_1 was in fact the minimum mean with probability 0.98, bearing in mind that the cost of estimation, V, must be minimized?

After many transitions have taken place, n_1 will be large so that σ_1^2, the variance associated with our estimate of μ_1, will be zero for practical purposes. The optimal values of n_2 and n_3 are then obtained from fig. 4.2. We observe that the optimal trajectory cuts the $\Omega = 0.02$ contour at

$$n_2 = 9.85$$
$$n_3 = 3.85$$

so that $V = [9.85(12.3 - 10.0) + 3.85(14.0 - 10.0)]$

$= 38.05$

These values are of course ensemble averages. Results in individual processes will vary, and can be guaranteed only for all $n_i \to \infty$. However in a typical process we would expect to reduce Ω to slightly less than 0.02 with $n_2 = 10$ and $n_3 = 4$.

As a comparison with the optimal trajectory generated by the ideal strategy, the probabilistic and equal rho trajectories have also been computed. Note that, as expected, all are asymptotically parallel. The characteristic vector is given by equations (3.50) and (3.51)

$$\frac{e_2}{e_3} = \frac{\rho_{o3}^2}{\rho_{o2}^2} = \frac{(\mu_3-\mu_1)^2}{\sigma_{o3}^2} \frac{\sigma_{o2}^2}{(\mu_2-\mu_1)^2}$$

$$= \frac{(4.0)^2}{10.40} \frac{10.41}{(2.3)^2} = 3.028$$

and $e_2 + e_3 = 1$

so that $$\underline{e} = \begin{bmatrix} 0.7517 \\ 0.2483 \end{bmatrix}$$

Asymptotically, each trajectory cuts the Ω contour at a constant angle determined by the cost multipliers, a_i,

for which that particular trajectory is optimal. Table 4.1 gives the equations of the Ω tangents at the point of intersection of the trajectory and the Ω contour. The equations are determined from (3.33) and from the results of sections 3.15 and 3.16. Direct measurement of the tangent angles in fig. 4.2 may be used to verify the validity of the solution of the inverse problem, as presented in section 3.12.

TABLE 4.1

OMEGA CONTOUR TANGENTS

Trajectory	Tangent Equation
Optimal	$V = k_1(2.3n_2 + 4.0n_3)$
Equal Rho	$V = k_2(n_2 + 3.028n_3)$
Probabilistic	$V = k_3(n_2 + n_3)$

k_1, k_2, k_3 are arbitrary constants

We may use fig. 4.2 to determine the estimation cost using a suboptimal strategy to obtain $\Omega = 0.02$. With an equal rho strategy the cost would be

$$V = 2.3\ (10.7) + 4\ (3.5) = 38.6$$

With a probabilistic strategy it would be

$$V = 2.3\ (9.3) + 4\ (4.3) \quad = \quad 38.6$$

As predicted, the performance of the suboptimal strategies is nearly as good as that of the optimal one. It is not true in general, though, that all suboptimal strategies yield equal costs. To compare the above figures with the performance of an adaptive strategy which is not suboptimal, we shall consider a strategy in which $n_1 \to \infty$, and n_2 and n_3 are chosen equally until $\Omega = 0.02$. The resulting trajectory has been plotted in fig. 4.2. We observe that it cuts the $\Omega = 0.02$ contour at $n_2 = n_3 = 8.3$; the cost is then

$$V = 2.3\ (8.3) + 4\ (8.3) \quad = \quad 52.3$$

Finally, as an example of how high the cost can be if we totally neglect the concept of dual control, we compute the cost incurred when all states are chosen an equal number of times. Thus

$$n_1 = n_2 = n_3$$

and

$$1 - \int_{-\infty}^{\infty} f_1(x)\ G_2(x)\ G_3(x)\ dx \quad = \quad 0.02$$

These equations can be solved by successive approximation to obtain

$$n_1 = n_2 = n_3 = 15.6$$

and $$V = 2.3\ (15.6) + 4\ (15.6) = 98.3$$

4.3 Optimum Point in (N-1) Space

Often we are not interested in the whole decision trajectory, but in a single optimum point, $\underline{n}^*(\Omega_f)$, in N-1 space. In such a case we would like to find by a quick approximation some point $\underline{n}$ on the optimal trajectory in the vicinity of $\underline{n}^*$; we can then reach the latter by following the optimal trajectory until $\Omega = \Omega_f$. To form the initial approximation two pieces of information are necessary:

1) the approximate relationship between $\underline{n}$ and Ω in the vicinity of the optimal trajectory;

2) the approximate location of the optimal trajectory in N-1 space.

Fortunately, both 1) and 2) can be computed using the theory of chapter 3. Our approach will be the following: we shall compute the approximate point of intersection ($\underline{n}'$ in fig. 4.3) of an equal rho (basic) trajectory with

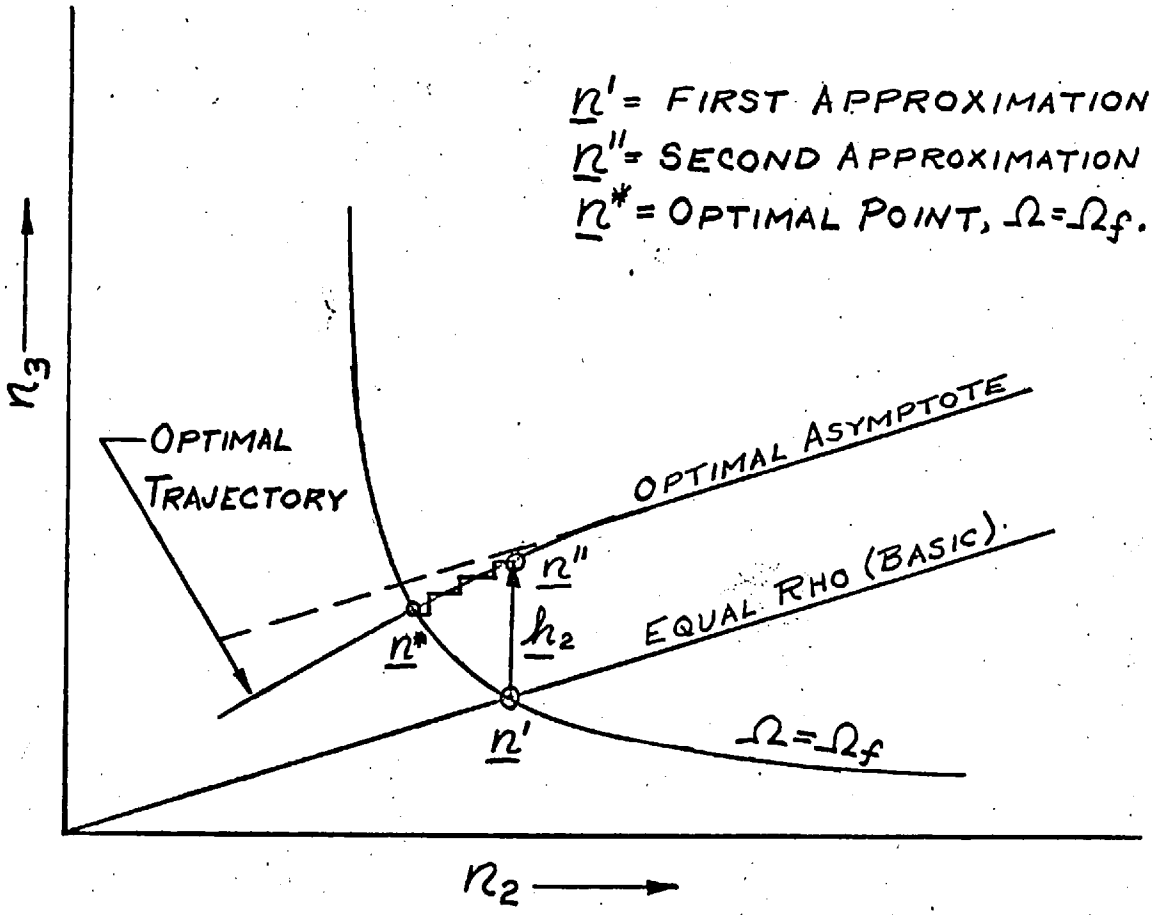


FIG. 4.3

METHOD OF DETERMINING AN OPTIMAL DECISION POINT

the $\Omega = \Omega_f$ contour; we shall then compute the offset vector $\underline{h}_i$ between the equal rho and optimal trajectories. Having determined the point $\underline{n}''$ on the optimal trajectory in the vicinity of $\underline{n}^*$, we can reach $\underline{n}^*$ by a succession of Bayes decisions, as in the previous section.

1) Relationship of $\underline{n}$ to Ω near $\underline{n}^*$

It is assumed that we are interested, for practical purposes in values of Ω between 0.005 and 0.1. For these values we introduce here the approximation

$$F_i(\mu_s) \simeq \alpha \exp(-\beta\rho_i^2) \qquad (4.1)$$

where $\alpha = 0.250$ $\qquad 1.3 < \rho_i < 2.6$

$\beta = 0.594$ $\qquad 0.005 < F_i(\mu_s) < 0.1$

Unlike the asymptotic relation (3.39), (4.1) is simply a "best fit" approximation, useful for computing purposes, but having no theoretical basis.

For $\Omega < 0.1$,

$$\Omega \simeq \sum_{i=1\neq s}^{N} F_i(\mu_s) = \alpha \sum_{i=1\neq s}^{N} \exp(-\beta\rho_i^2)$$

In an equal rho trajectory

$$\frac{n_i}{n_j} = \frac{\rho_{oj}^2}{\rho_{oi}^2}$$

so that

$$n_i \rho_{oi}^2 = \frac{\sum_{j=1\neq s}^{N} n_j}{\sum_{j=1\neq s}^{N} (\rho_{oj}^{-2})}$$

Thus

$$\Omega \simeq \alpha(N-1) \exp \left[\frac{-\beta \sum_{j=1\neq s}^{N} n_j}{\sum_{j=1\neq s}^{N} (\rho_{oj}^{-2})} \right] \qquad (4.2)$$

Each of the N-1 coordinates of $\underline{n}'$ in fig. 4.3 is given by

$$n_i' = \frac{1}{\beta \rho_{oi}^2} \log \left[\frac{\alpha(N-1)}{\Omega_f} \right] \qquad (4.3)$$

Equation (4.3) defines the point $\underline{n}'$ in fig. 4.3.

2) Location of Optimal Trajectory

Given the point $\underline{n}'$ we can determine the components h_{ji} of the offset vector $\underline{h}_i$ from (3.56)

$$h_{ji} = -\frac{1}{\rho_{oj}^2} \left[\log \left(\frac{\rho_{oi}^2}{\rho_{oj}^2} + \frac{h_{ji}}{n_i} \right) + 2 \log \left(\frac{\sigma_{oj}}{\sigma_{oi}} \right) \right] \qquad (4.4)$$

where i may be chosen arbitrarily ($1 \neq s$). Given the value of n_i ($i = 2$ in fig. 4.3), we can solve (4.4) iteratively for the components h_{ji}. Now in fig. 4.3,

$$\underline{n}'' = \underline{n}' + \underline{h}_i \tag{4.5}$$

Having determined $\underline{n}''$, which is on the optimal trajectory, we can locate $\underline{n}^{\ast}$ by a succession of Bayes decisions.

One consideration of some subtlety arises here. After a point $\underline{n}''$ has been determined, the question may arise, 'Should a given n_i be increased or decreased to approach $\underline{n}^{\ast}$?' If $\underline{n}''$ is an arbitrary point in the vicinity of $\underline{n}^{\ast}$, the answer is not obvious; a straightforward application of the Bayes approach may lead to a non-optimum point. To resolve this difficulty, we have placed $\underline{n}''$ as nearly as possible on the optimal trajectory, thereby making use of the fact that if and only if a point is on the trajectory

$$\left.\begin{aligned} \Omega < \Omega_f &\iff n_i > n_i^{\ast} \\ \text{and} \quad \Omega > \Omega_f &\iff n_i < n_i^{\ast} \end{aligned}\right\} \forall_i,\ i \neq s$$

Since reflection about $\underline{n}^{\ast}$ along the optimal trajectory inverts the Bayes criterion used to approach $\underline{n}^{\ast}$, it is useful to place $\underline{n}''$ on the same side of $\underline{n}^{\ast}$ in every

computation. We have adopted the convention of making $\Omega < \Omega_f$ at $\underline{n}''$. If initially $\Omega > \Omega_f$ then $\underline{n}''$ is increased by $\Delta \underline{n}$ to ensure that, at $\underline{n}'' + \Delta \underline{n}$, $\Omega < \Omega_f$.

$$\Delta \underline{n} = c \left[\frac{1}{\beta} \sum_{j=1 \neq s}^{N} (\rho_{oj}^{-2})(1 - \frac{\Omega_f}{\Omega(\underline{n}'')} \right] \underline{e}$$

where c is a constant ≥ 1.

$\underline{e}$ is the characteristic vector.

Fig. 4.4 is a flow diagram of the algorithm used to compute $\underline{n}^{*}(\Omega_f)$. Using the parameters in the example given previously, we find that, for $\Omega_f = 0.02$,

$$\underline{n}' = \begin{bmatrix} n_2 \\ n_3 \end{bmatrix} = \begin{bmatrix} 10.67 \\ 3.52 \end{bmatrix}$$

and $$\underline{h}_2 = \begin{bmatrix} 0.0 \\ 0.62 \end{bmatrix}$$

so that $$\underline{n}'' = \underline{n}' + \underline{h}_2 = \begin{bmatrix} 10.67 \\ 4.14 \end{bmatrix}$$

Comparing these values with fig. 4.2, we see that $\underline{n}'$ is indeed an equal rho approximation, and that $\underline{n}''$ lies on the optimal trajectory, fairly close to the final optimum point, $\underline{n}^{*} = \begin{bmatrix} 9.85 \\ 3.85 \end{bmatrix}$.

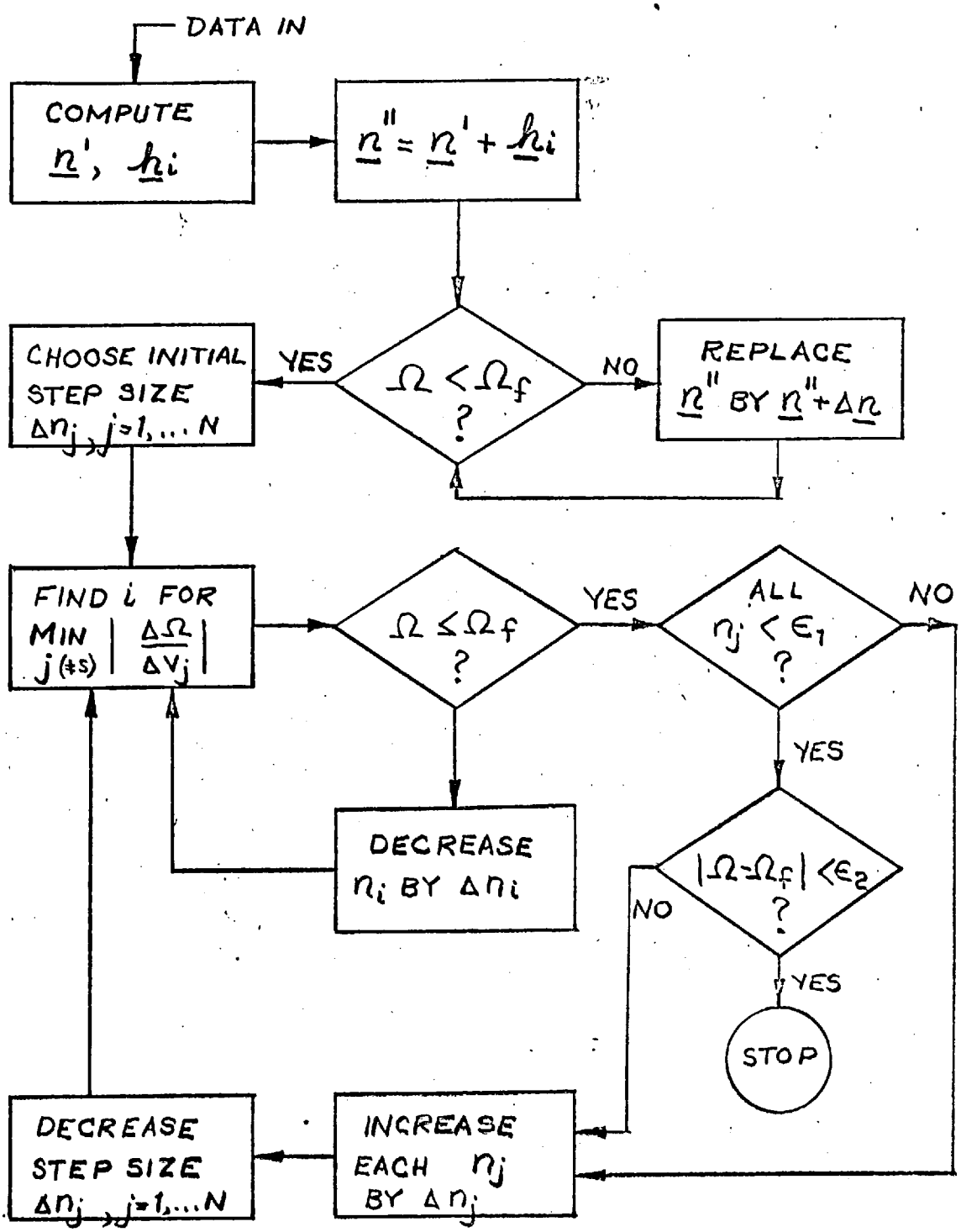


FIG. 4.4

COMPUTATION OF AN OPTIMAL POINT IN (N-1) SPACE

4.4 Doubly Constrained Optima

The foregoing algorithm allows us to compute the minimum possible ensemble average cost which need be incurred to ensure that Ω is reduced to a required level. No upper limit is placed on n, the total number of transitions which may be observed, and optimality requires that $n_s \to \infty$, where s is the minimum cost state. We have seen that realizable decision policies exist which approximate the situation for large n. Early in the life of a process, however, when n is small by definition, this theoretical optimum is obviously an unsuitable measure of the performance of a decision strategy. For limited n, then, the problem is

Choose the set $\{n_1, \ldots\ldots n_N\}$ to minimize

$$V = \sum_{i=1}^{N} n_i(\mu_i - \mu_s)$$

subject to

$$1 - \int_{-\infty}^{\infty} f_s(x) \prod_{i=1 \neq s}^{N} G_i(x)\, dx = \Omega_f$$

and

$$\sum_{i=1}^{N} n_i = n_f \qquad (4.6)$$

Notice that the problem differs from the original one only by the presence of the additional constraint (4.6). It is possible to solve it using an algorithm similar to that shown in fig. 4.1. In the latter case, however, Ω is expressed as a product, and the gradient, $(\Delta\Omega/\Delta V_i)$, is relatively easy to compute. Moreover this gradient must be re-evaluated for only one state at each step of the search in N-1 space. In the doubly constrained problem in which n_s is finite, $\Delta\Omega / \Delta V_i$ is a more complicated function, viz.

$$\frac{\Delta\Omega}{\Delta V_i} = \frac{1}{\Delta n_i(\mu_i-\mu_s)} \left[\int_{-\infty}^{\infty} f_s(n_s,x) \prod_{j=1\neq s}^{N} G_j(n_j,x)\, dx \right.$$

$$\left. - \int_{-\infty}^{\infty} f_s(n_s,x) G_i(n_i+\Delta n_i) \prod_{\substack{j=1\neq s \\ \neq i}}^{N} G_j(n_j,x)\, dx \right] \qquad (4.7)$$

where $f_i(n_i,x) = \frac{n_i^{1/2}}{\sqrt{2\pi}\,\sigma_{oi}} \exp\left[-\frac{n_i}{2}\left(\frac{x-\mu_i}{\sigma_{oi}}\right)^2\right]$

and $G_i(n_i,x) = \int_{x}^{\infty} f_i(n_i,y)\, dy$

Since each gradient calculation involves an integral evaluation, this computation is much slower than the

corresponding one given by (3.29), in which the function $G_i(x)$ can be interpolated from a look-up table in the computer memory. In addition (4.7) must be computed N times at every step of the search, rather than once as in the former algorithm. Even for modest values of N (5-10) this method of descent in N-space is excessively slow.

Fortunately, the theory of chapter 3 once again allows us to simplify the search. We know that for $n_s \to \infty$, the N-1 variables, n_1, n_2 ... n_{s-1}, n_{s+1}, ... n_N, are related by the characteristic vector, $\underline{e}$. We shall make the assumption that the characteristic vector is also valid for finite values of n_s. Thus if $\underline{n}$ is the projection in N-1 space of an arbitrary point on the optimal trajectory, and $\underline{n}^*$ is the desired optimum point where $\Omega = \Omega_f$, we assume that $\underline{n}$ and $\underline{n}^*$ are colinear, i.e.

$$\underline{n}^* = \underline{n} + a\underline{e}$$

where "a" is a scalar multiplier. Providing a suitable starting point $\underline{n}$ can be found, we need only vary "a", computing n_s from (4.6) until $\Omega = \Omega_f$. The process is carried out iteratively, with "a" approximated by

$$a = \frac{1}{\beta} \sum_{j=1 \neq s}^{N} (\rho_{oj}^{-2}) \left[1 - \frac{\Omega_f}{\Omega(\underline{n})} \right] \qquad (4.8)$$

As a starting point $\underline{n}$ for this procedure we use the singly constrained optimum point with $\Omega = \Omega_f$, computed as outlined in section 4.3 and fig. 4.4. The computational flow diagram is shown in fig. 4.5. Because the number of evaluations of Ω is drastically reduced the convergence of this algorithm is very much faster than that obtained with descent in N-space. Using the parameters given in section 4.2, we find that even for low values of n_f the computed optimum cost is within less than 0.1 % of that given by the method of descent in N-space. For $n_f = 100$, the difference is less than 0.001%. Using descent in 3 space the computation requires about 30 seconds on an IBM 7090 computer; using the simplified method the time is 2.5 seconds. As the number of states increases, this time disparity becomes even greater.

4.5 Simulated Results: A Three-State System

Having developed the necessary computational procedures in sections 4.2 to 4.4, we are now ready to compare simulated results with those of the ideal strategy. One hundred three-state systems with identical stochastic transition matrices were simulated separately on an IBM 7090 computer, and mean values of the observation matrix, M, were

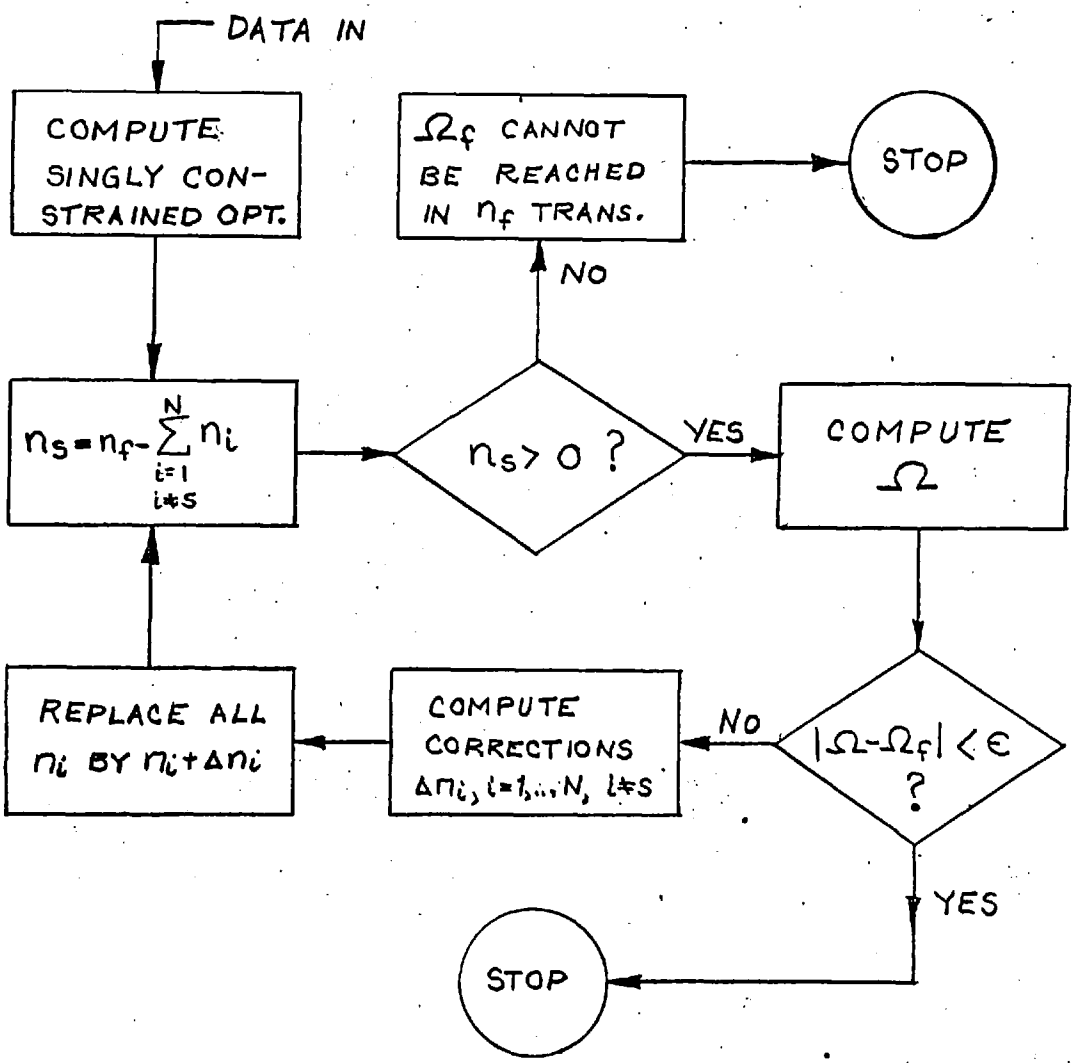


FIG. 4.5

COMPUTATION OF

DOUBLY CONSTRAINED OPTIMAL POINT

taken across the ensemble at intervals of five transitions. From these the posterior ensemble parameters $\left\{\hat{\mu}_i\right\}$, $\left\{\hat{\sigma}_{oi}^2\right\}$, and Ω (n) were computed. These posterior values were fed into the algorithm of fig. 4.5, and the doubly constrained optima, $\left\{n_i^*(\Omega)\right\}$ and $\hat{V}^*$, were computed and compared with the observed ensemble values $\left\{n_i\right\}$ and $\hat{V}$. The basic comparison procedure is shown in fig. 4.6.

To start the process it is necessary to have a prior estimate of $\left\{\sigma_{oi}^2\right\}$. Since this is a finite state system with a finite range of possible costs, the range of σ_{oi}^2 is bounded. In the total absence of any knowledge of P, the prior estimate of σ_{oi}^2 is taken as that which would result if all p_{ij}'s were equal ($p_{ij} = 1/N$). This estimate is nearly always higher than the true value of σ_{oi}^2, and results in a slightly conservative estimation strategy at the beginning of the process; i.e. the estimated minimum cost state is chosen somewhat less frequently than it would be if the true values σ_{oi}^2 were known. Such a procedure has the advantage of being less easily deceived by atypical results occurring early in the life of the process. As time progresses, an updated estimate of each σ_{oi}^2 is computed as a weighted sum of the maximum likelihood estimate and the prior estimate; the estimate of σ_{oi}^2 can be made in this way to converge to its true value.

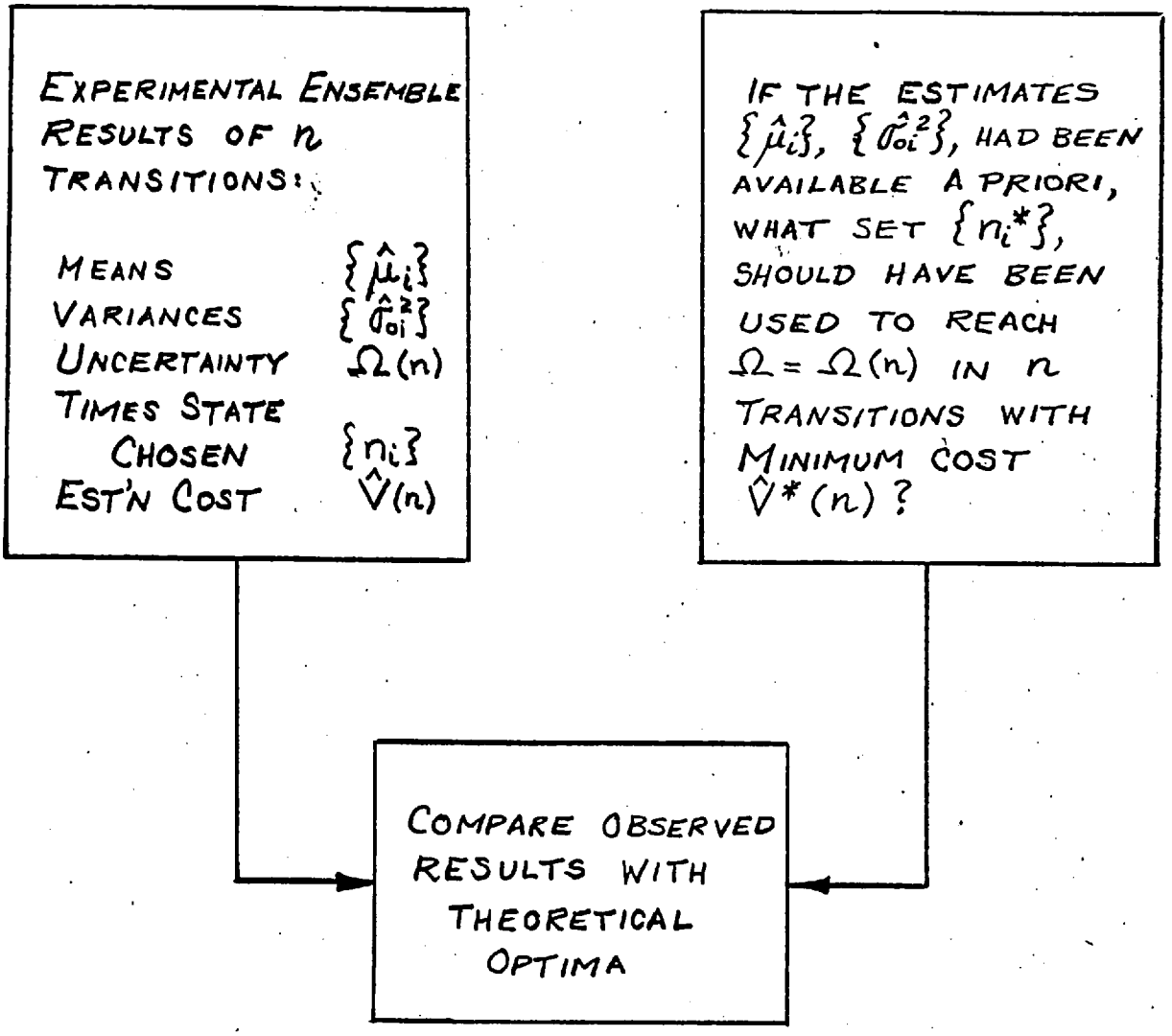


FIG. 4.6

COMPARISON OF EXPERIMENTAL SYSTEM WITH THEORETICAL OPTIMUM

As we shall see, it is often useful to compute (or guess) a prior estimate of $\{\mu_i\}$ as well. However in this case the process was started by a choice of each state once in turn, and application of the optimal dual strategy with $\gamma = 1.0$ thereafter.

The process parameters used were those of the three-state system of section 4.2. The matrices are repeated here for convenience

$$P = \begin{bmatrix} 0.3 & 0.3 & 0.4 \\ 0.6 & 0.3 & 0.1 \\ 0.7 & 0.1 & 0.2 \end{bmatrix} ; \quad C = \begin{bmatrix} 6 & 14 & 10 \\ 14 & 12 & 3 \\ 16 & 12 & 8 \end{bmatrix}$$

Only the matrix C is known to the controller. The results of an ensemble of 100 processes, each of which ran for 100 transitions, are presented in figs. 4.7-4.10.

If P is known it is apparent that the optimal decision vector $\underline{d}_i^*$ is [1 0 0]; since all states are reachable deterministically by one control action, and B = 0, each row of the optimal decision matrix, D^*, is the same.

$$D^* = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

and $\underline{\pi}(PD^*) = [1 \quad 0 \quad 0]$

so that $g^* = \langle \underline{\pi}, \underline{\mu} \rangle = 10.0$

Fig. ~~9.7~~ 4.7 shows the convergence of the mean cost per transition with time when the optimal dual strategy is used. The overall mean cost per transition is 11.69 at stage 5 and 10.33 at stage 100. The incremental cost per transition, measured over the previous five transitions, drops from 11.69 to 10.03 in the same period. Note that after only five transitions the cost is already below the value of 12.10 which would be obtained if all states were chosen equally.

The uncertainty, Ω, steadily approaches zero as n increases, as shown in fig. 4.8a. It is interesting to compare the measured value of Ω at n = 100 with that predicted by the asymptotic equation (3.87). The latter predicts $\Omega = 0.0524$; the measured value is $\Omega = 0.0384$. There are two reasons for this discrepancy. First, (3.87) applies strictly only to infinitesimal values of Ω. More important, however, is the fact that the theory upon which (3.87) is based assumes that $n_s = n_1 \to \infty$, so that all reduction in Ω should be caused by n_2 and n_3. In the early stages of the process an additional decrease in Ω is caused by increasing n_s.

The results of the comparison between the observed cost and the cost using the ideal stragegy (non-realizable because it assumes prior knowledge of the estimates) is shown in fig. 4.8b. For $\Omega < 0.10$, the observed cost was

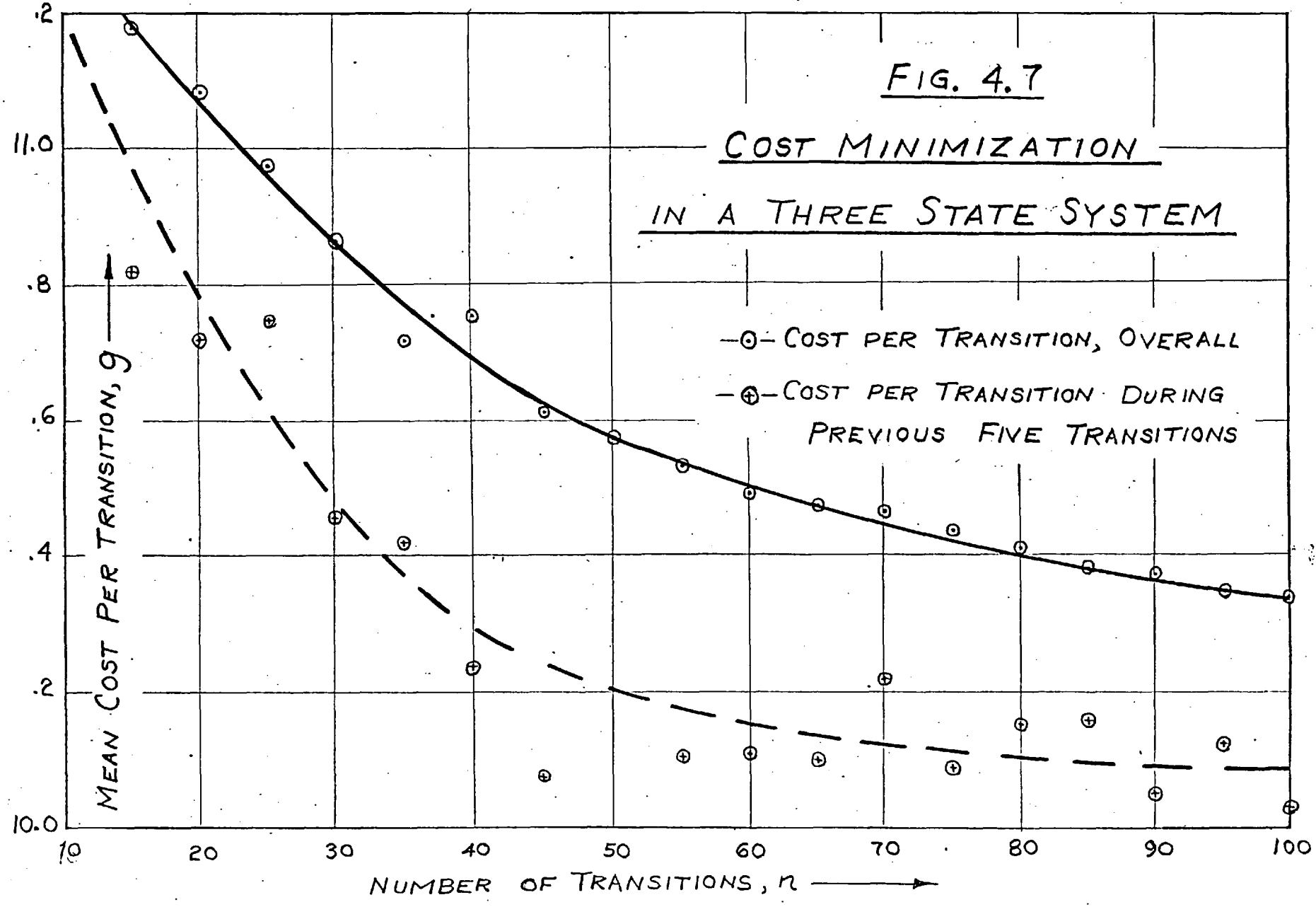


FIG. 4.7
COST MINIMIZATION
IN A THREE STATE SYSTEM
—⊙— COST PER TRANSITION, OVERALL
—⊕— COST PER TRANSITION DURING PREVIOUS FIVE TRANSITIONS
MEAN COST PER TRANSITION, g
NUMBER OF TRANSITIONS, n
·2
11.0
·8
·6
·4
·2
10.0
10
20
30
40
50
60
70
80
90
100

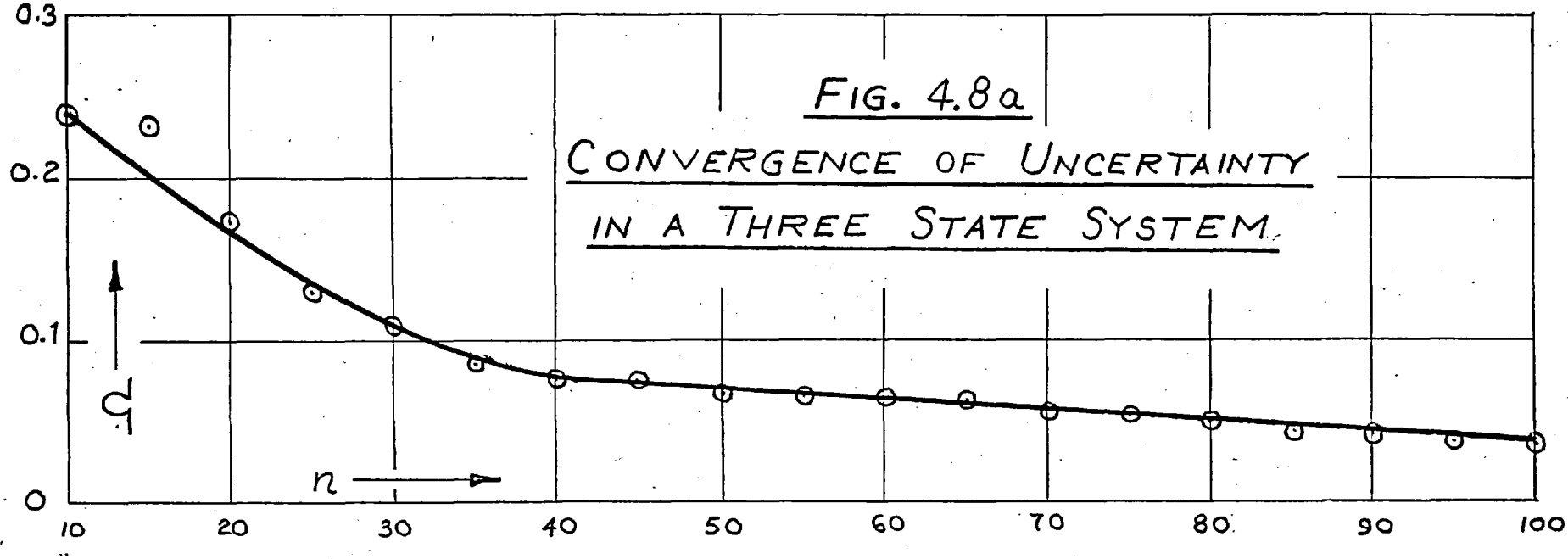


FIG. 4.8a

CONVERGENCE OF UNCERTAINTY IN A THREE STATE SYSTEM

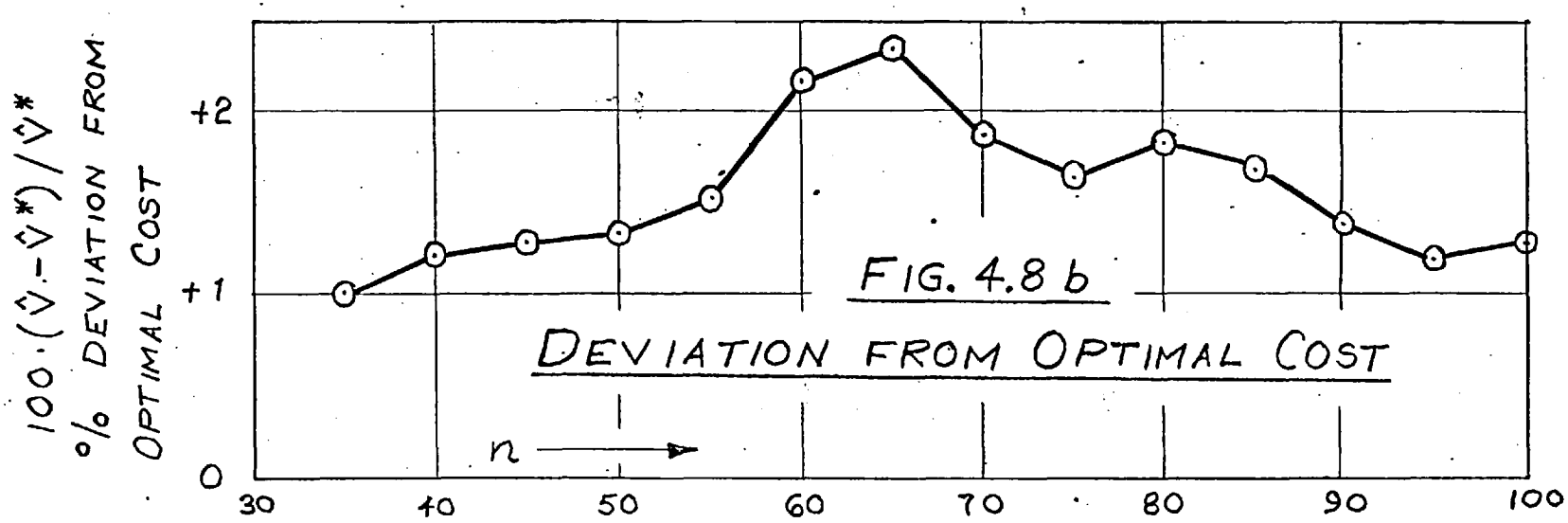


FIG. 4.8b

DEVIATION FROM OPTIMAL COST

at all times within 2.5% of the ideal cost. After 100 transitions the measured estimation cost was 36.48 for $\Omega = 0.038365$. If the posterior estimates at n = 100 had been available at the beginning of the process, the same value of Ω could have been achieved at a cost of 36.00. The solution generated by the algorithm of fig. 4.5 shows that this could have been achieved with $\underline{n}^{*} = (86.54, 10.04, 3.42)$ instead of the observed $\underline{n} = (86.72, 9.38, 3.90)$. In terms of overall cost per transition, the measured value of 10.3348 compares very well with the ideal minimum dual cost of 10.3300.

From (3.93) we see that the probability of choosing state s at stage n is given by $d_{is}(n) = 1 - \Omega$. At n = 100, this value is 0.9616. The observed frequency of choice of state s is given by $(\Delta n_s / \Delta n)$, which is plotted in fig. 4.9. At n = 100 $(\Delta n_s / \Delta n) = 0.968$ and is in close agreement with the value of d_{is}.

Table 4.1 shows the ensemble estimates $\left\{ \hat{\mu}_i \right\}$ and $\left\{ \hat{\sigma}_{oi}^2 \right\}$ at n = 100. Using these values and the theory of the characteristic vector, we can compute from (3.93) and (3.94) the decision vector $\underline{d}_i(n)$ at n = 100. It is found to be

$$\underline{d}_i(100) = (0.9616,\ 0.0304,\ 0.0080) \qquad i = 1, 2, 3$$

so that
$$g(100) = \langle \underline{d}_i, \underline{\mu} \rangle$$
$$= 0.9616(9.970) + 0.0304(12.183) + 0.0080(14.000)$$
$$= 10.070$$

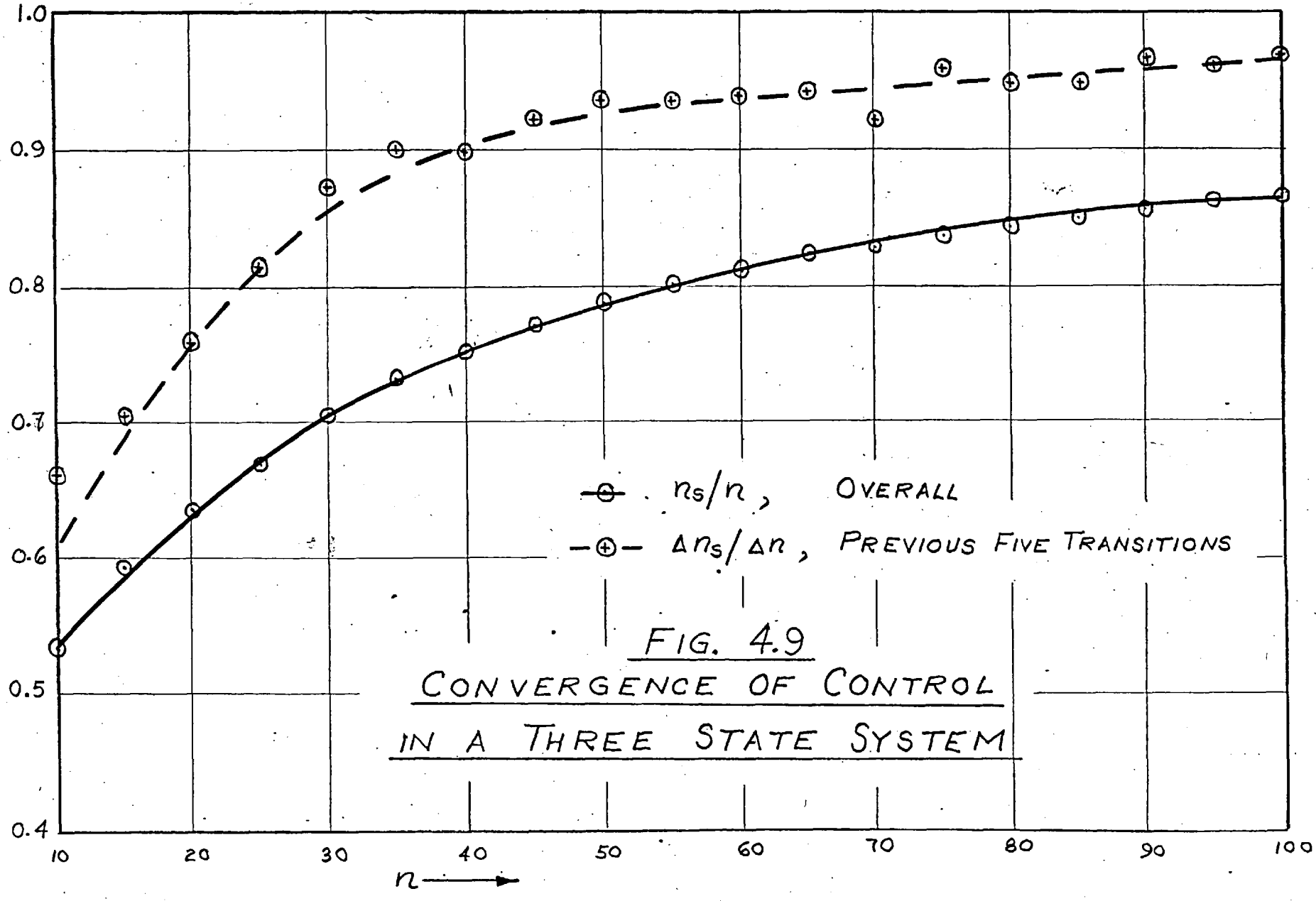


FIG. 4.9

CONVERGENCE OF CONTROL IN A THREE STATE SYSTEM

The observed cost of the last ten transitions, 10.076, is in excellent agreement with the value calculated above.

TABLE 4.1

THEORETICAL AND ESTIMATED PARAMETERS IN A THREE-STATE SYSTEM

Parameter	State 1	2	3
Theoretical mean cost	10.000	12.300	14.000
Estimated mean (n=100)	9.970	12.183	14.000
Theoretical one-stage variance	9.600	10.410	10.400
Prior estimate of variance (n=0)	8.991	22.889	10.667
Weighted estimate of variance (n=100)	9.617	12.044	10.585

Since at n = 100 Ω = 0.0384, we would expect that in an ensemble of 100 processes state s would be chosen incorrectly about three or four times. In fact an incorrect choice was made twice. In both cases state 2 was chosen as the minimum cost state. The resulting values of n_1 were 6 and 3, and the mean costs per transition were 11.95 and 11.77 respectively. Note that even in these, by far the

worst cases, the cost per transition is below the value (12.10) which would be expected if all states were chosen equally. It should be emphasized that in time, the values of $\underline{d}_i$ and g will converge correctly even in these anomalous cases.

Fig. 4.10 shows the ensemble distributions of n_s and overall cost per transition at n = 100. Because of the scales used, the two cases mentioned above are not shown; they are of course included in the computation of mean values. Fig. 4.10a shows that even though the controller has no knowledge of P at the beginning of the process, the correct state is chosen at least 90 times in the first 100 transitions in more than half of the members of the ensemble; the mean value of n_s across the ensemble is 86.72. Fig. 4.10b shows the distribution of mean cost per transition. Observed values, apart from the two cases mentioned previously, range from 9.58 to 11.20; the ensemble mean is 10.33.

4.6 Effect of Variance Estimate

In table 4.1 the prior estimate of σ_{o2}^2 is about twice its true value. Because the estimate $\hat{\sigma}_{o2}^2$ at any stage in the process is a weighted sum of the prior estimate and the

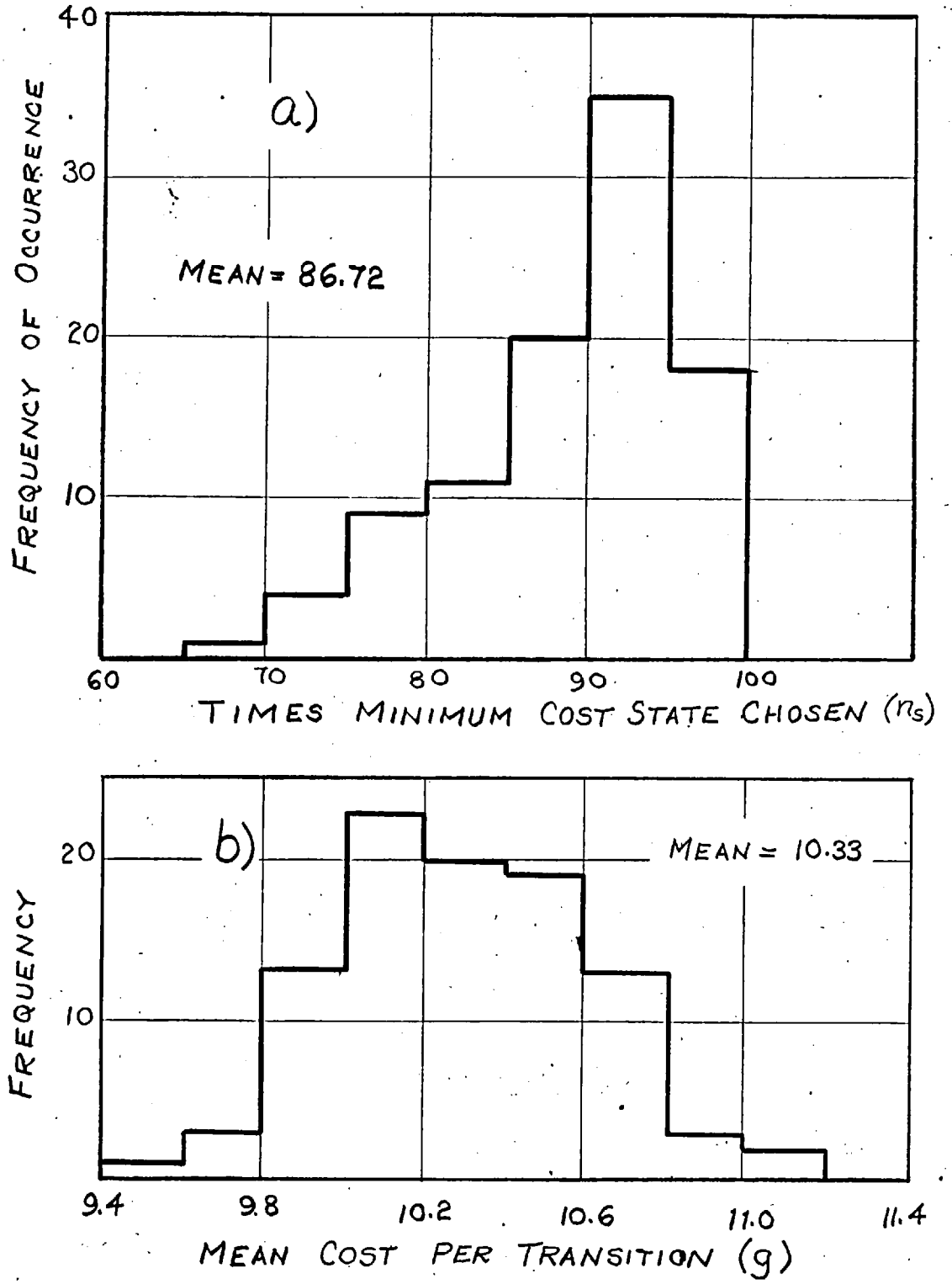


FIG. 4.10

ENSEMBLE DISTRIBUTIONS

maximum likelihood estimate, $\hat{\sigma}_{o2}^2$ is slightly high even after 100 transitions. One could of course use only the maximum likelihood estimate for $n_i \geq 2$ (at least two samples are required to estimate variance). The reason that this is not done is that there is a finite probability that the maximum likelihood estimate of variance will be exactly zero owing to the first two transitions from state i being identical. In such a case state i, unless $i = \hat{s}$, will never be chosen again by the optimal controller, since $\sigma_{oi}^2 = 0$ implies certainty concerning the estimate of μ_i. Such a situation, which could lead to a non-convergent strategy, is avoided by the weighted variance approach.

Although both the asymptotic convergence and optimality of the strategy are assured despite initially incorrect mean and variance estimates, it is beneficial to operation early in the life of the process to make these estimates as accurate as possible. Usually the behaviour of the process is known vaguely at least, and some estimate of P can be formed which is better than the automatic prior estimate of $p_{ij} = 1/N$ used in the foregoing example.

4.7 An Example: A Fluid Mixing Process

Let us consider a batch mixing process in which it is

desired to introduce a certain quantity of a solution at a fixed temperature during a batch reaction. The flow of solution is controlled by an inlet valve. Owing to pressure and flow fluctuations the concentration of the final product is distributed statistically, the distribution depending on the input valve setting. It is desired to determine, by adaptive control of a long sequence of batches, which valve setting yields a final concentration nearest to some specified value.

There are two main disturbance effects. The first is a random variation whose magnitude increases with flow, i.e. the signal-to-noise ratio is constant. The second is caused by pressure variations; pressure regulation is poor at the upper and lower extremes of the valve setting. The noise introduced at these points is caused by a switching action elsewhere in the fluid circuit. Its probability density therefore tends to be concentrated at $\pm$ x, where x is relatively large for valve settings close to the extremes, and small for intermediate settings. The numerical distribution of suitable distribution functions, together with the parameters used for this particular simulation, are considered in appendix 4.

The final product is useful if its concentration lies between 5.0% and 6.9% inclusive; the desired concentration

is 6.0%. Concentration is measured with an accuracy of ± 0.05%. We may therefore quantize the useful output concentrations into twenty states or levels, 5.0, 5.1, ... 6.9%. Two more states must also be added to represent concentrations which are outside the usable range. The cost incurred (i.e. the loss in value of the product) is deemed to increase quadratically with the difference in concentration from the desired level. In addition, an extra penalty is added if the product is outside the usable limits. The resulting costs are shown in table 4.2.

The input setting is similarly quantized into twenty states along with two dummy states added for mathematical convenience to match states 1 and 22 in the output (in general, though, there is no theoretical reason in a batch process for the input and output states to match in number). The transpose of the last column of table 4.2 is a row of the 22 x 22 cost matrix, C. Each row is identical in this case because no direct cost has been attached to input flow.

The problem of determining the correct valve setting is essentially that of climbing a discrete noisy hill in one dimension. Note that it may be multimodal or may possess multiple minima; perhaps there are two different valve settings which incur the same expected cost. The theory developed in chapter 3 assumes, in fact, no

correlation in cost between adjacent states, and is quite capable of handling such a hill.

TABLE 4.2

PRODUCT COST AS A FUNCTION OF OUTPUT CONCENTRATION

State	Concentration (%)	Cost
1	< 5.0	321
2	5.0	100
3	5.1	81
4	5.2	64
5	5.3	49
6	5.4	36
7	5.5	25
8	5.6	16
9	5.7	9
10	5.8	4
11	5.9	1
12	6.0	0
13	6.1	1
14	6.2	4
15	6.3	16
16	6.4	25
17	6.5	36
18	6.6	49
19	6.7	64
20	6.8	81
21	6.9	100
22	> 6.9	300

If the designer is unable to put forward any prior estimate of the process transition matrix P, and does not wish to make the assumption of unimodality, then the process can be optimized automatically with an equal-probability prior estimate of $\{\mu_i\}$ and $\{\sigma_{oi}^2\}$, as previously discussed. If an intelligent guess concerning P is available, though, it should be used as a prior estimate.

Suppose that unimodality can be assumed. In such a case the requirement of a global search can be dispensed with, and the following unimodal algorithm is applicable. At any stage in the process form a three-state subsystem consisting of the state whose expected cost is minimum at present, and the two adjacent states. Apply the optimal decision strategy within the subsystem and observe a further transition. Then form a new three-state subsystem in the same way (it may be the same one again) and repeat. This algorithm will invariably yield convergence to the local minimum. It is also applicable to multidimensional hills; if the number of variables is k, then the number of states in the subsystem is 3^k.

Control of the mixing process was simulated using three different initial assumptions:

1) no prior knowledge;

2) prior estimate of P for variance computation: if

the valve setting is state i, it was assumed that the output would be uniformly distributed over the seven states i - 3 to i + 3.

3) same as 2), but the cost hill was assumed to be unimodal.

In fact the hill is unimodal; its minimum is flat, though, the two control states 8 and 9 yielding equal costs. This posed no difficulty for the controller which tended asymptotically to alternate between them to the near exclusion of other states.

Fig. 4.11 shows the resulting cost convergence. Though all curves will converge to the same mean cost per transition eventually, a relatively high initial cost is incurred when no prior estimates are available and a global search is carried out. Note that even the rough approximation of assumption 2) lowers the initial cost somewhat. It can be seen that use of the unimodal algorithm is very much more efficient than either of the previous two procedures. The input control subset 7, 8, and 9 was selected as containing the minimum after only 14 transitions. The better performance with this algorithm is due both to its quick convergence and to the fact that none of the extreme valve settings, which yield a low quality expensive output, were tried at all in the local search. In a global search each one must be tried at least once.

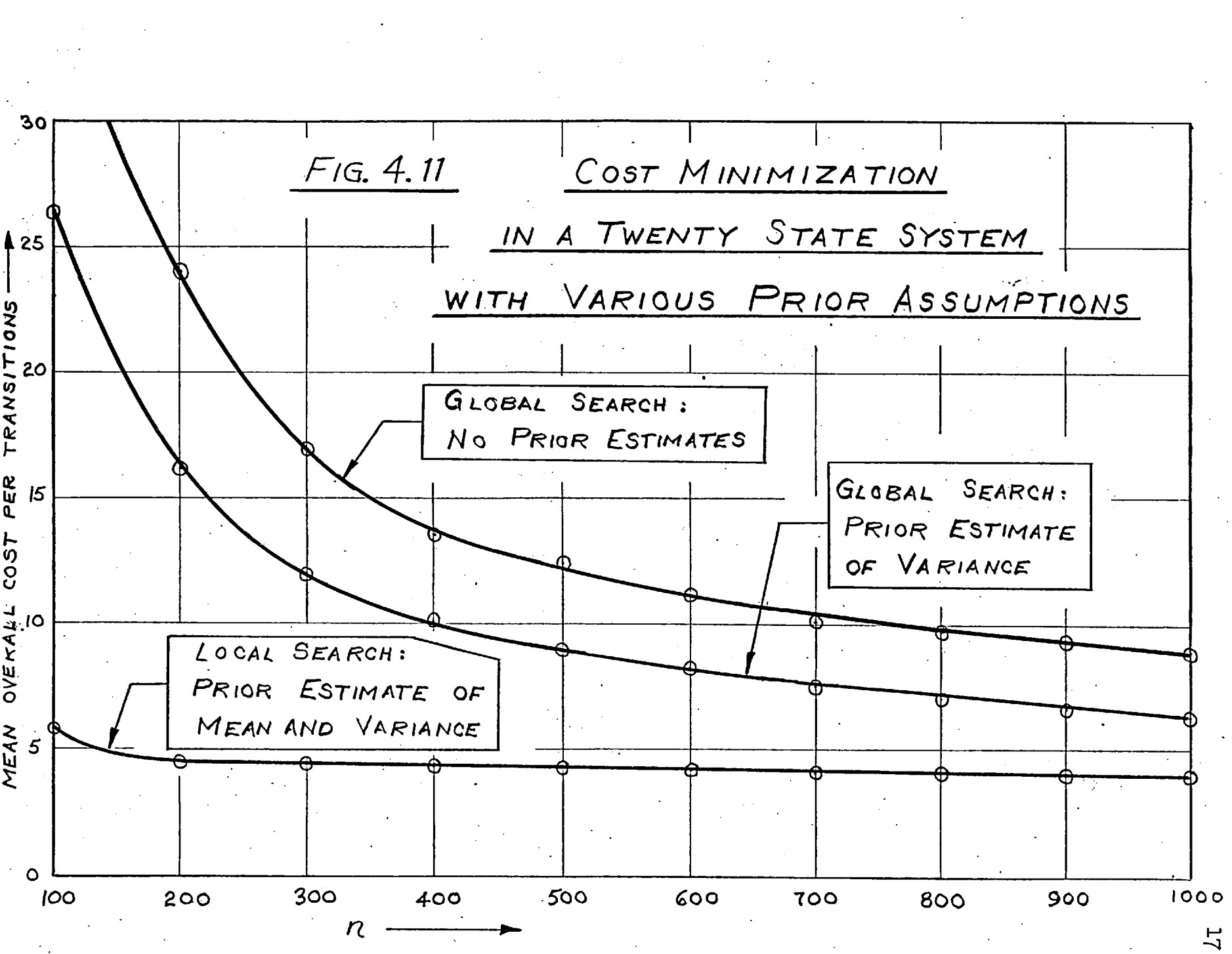


FIG. 4.11
COST MINIMIZATION
IN A TWENTY STATE SYSTEM
WITH VARIOUS PRIOR ASSUMPTIONS
GLOBAL SEARCH : NO PRIOR ESTIMATES
GLOBAL SEARCH: PRIOR ESTIMATE OF VARIANCE
LOCAL SEARCH: PRIOR ESTIMATE OF MEAN AND VARIANCE
MEAN OVERALL COST PER TRANSITIONS
n
0
5
10
15
20
25
30
100
200
300
400
500
600
700
800
900
1000

TABLE 4.3

PERFORMANCE OF OPTIMAL DECISION STRATEGY IN FLUID MIXING PROBLEM

Run number	1	2	3	4
Prior Assumption	No prior estimate	Prior estimate of variance	Local Search	No Learning
Parameter				
Mean cost/transition averaged over:				
1000 transitions	8.558	6.399	4.405	114.882
Last 900 transitions	6.606	4.180	3.853	---
Last 800 transitions	5.550	3.966	3.940	---
Last 500 transitions	4.524	3.825	3.778	---
Last 300 transitions	4.950	3.803	3.817	---
Control Effectiveness:				
Per cent of transitions using optimal settings (8 or 9)	84.4	93.7	95.7	10.0
Product Quality:				
Per cent output in range 5.6% - 6.4%	93.2	95.5	98.8	35.6
Per cent output in range 5.8% - 6.2%	75.0	78.4	78.7	19.0
Per cent reject batches	1.0	0.5	0.0	31.7

Table 4.3 gives a more detailed comparison of the three runs. Note that in run 2 convergence was nearly complete after 200 transitions (though the overall mean cost per transition decreases slowly because of high initial

costs, the incremental mean cost decreases relatively quickly; cf. fig. 4.7). The value of g^{*} appears to be about 3.8. Apart from the high initial search cost necessarily incurred in run 2, runs 2 and 3 give quite similar results. Control effectiveness and product quality (see table 4.3) are similar, and each is somewhat better than in run 1, for which no prior information was assumed.

A fourth run was made using a non-learning control in which all input valve settings were used an equal number of times. The results in table 4.3 speak for themselves.

To give a detailed idea of the step-by-step control sequence, the individual valve set points for the first 150 transitions (batches) have been plotted in fig. 4.12 from the results of run 2. The first 21 transitions are exploratory, each control state being chosen at least once. At this point the optimum setting appears to be between states 5 and 11 with high probability, and this region is searched more thoroughly. After transition 120, it appears increasingly probable that either state 8 or 9 is the optimum control setting; one or other of these is chosen in all but seven of the following 880 transitions.

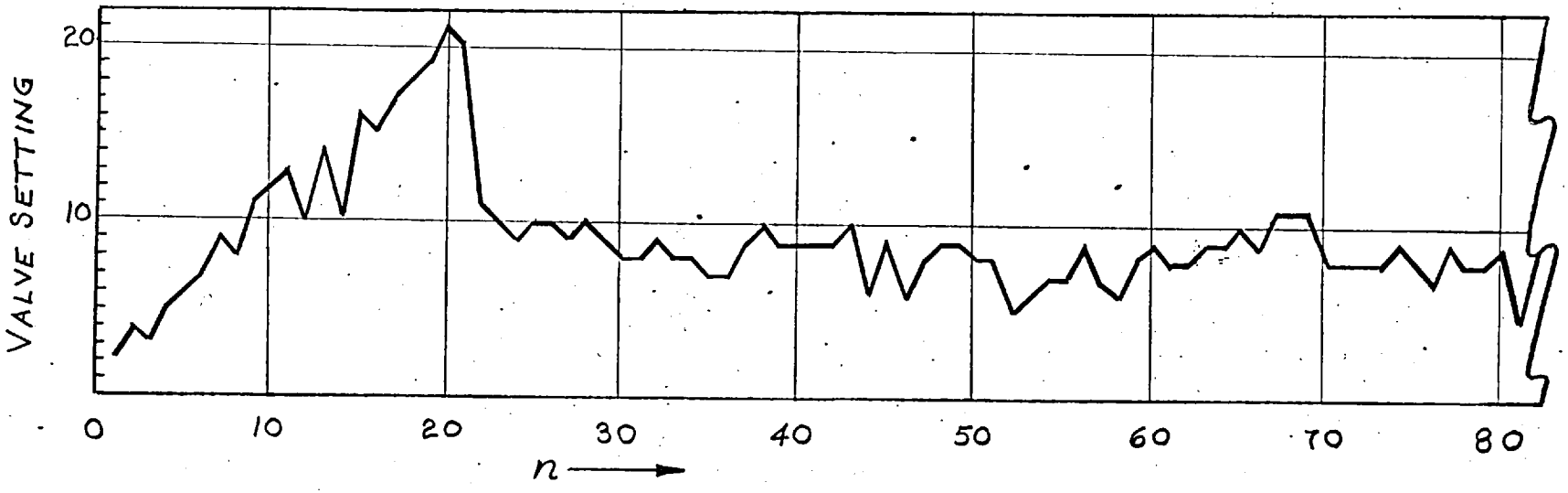
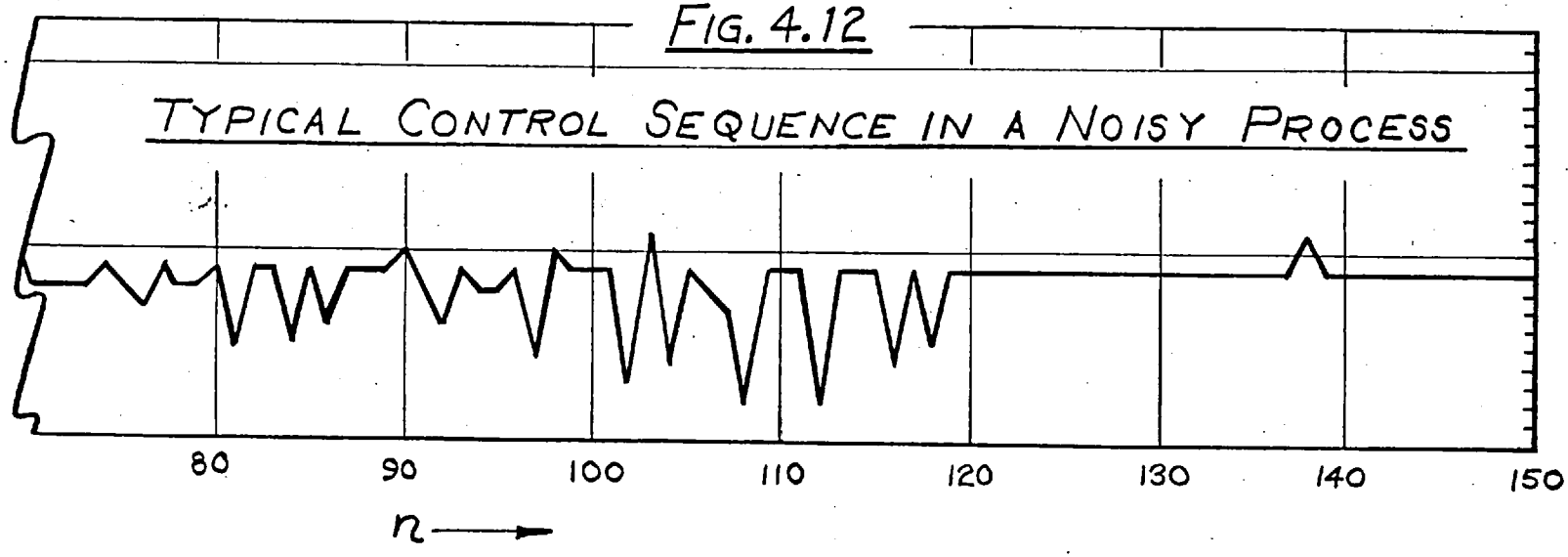


FIG. 4.12

TYPICAL CONTROL SEQUENCE IN A NOISY PROCESS



4.8 Summary

Numerical results have been presented in this chapter which illustrate and verify the theory developed in chapter 3. An ensemble of one hundred three-state systems has been simulated, and it has been seen that the (realizable) optimal dual strategy yields estimation costs which are within less than 2.5% of those associated with the (non-realizable) ideal strategy. The effects of various prior assumptions have been illustrated by the simulation of a twenty-level batch mixing process in which multiplicative non-gaussian noise is present. Even in the face of initially complete ignorance of the process and noise parameters, the optimal decision strategy converges satisfactorily. However, it has also been shown that some prior knowledge concerning the process, even if only very approximate in nature, enables the controller to reduce the estimation cost incurred. In particular, the knowledge that a given cost hill is unimodal is extremely valuable.

CHAPTER 5

ADAPTIVE ORDERING OF POWER GENERATION AS A CYCLIC MARKOV PROCESS

5.1 Introduction[‡]

The control of a discrete Markov process in which P is initially unknown normally implies the existence of a dual control problem. However this is not invariably so. Mathematically, the requirement for a dual control strategy exists because control decisions, based on the estimate, $\hat{P}$, also affect the choice of which rows of $\hat{P}$ will be subject to further estimation. In short, estimation and control are interlocked. If, though P is unknown, certain rows of P are known a priori to be completely correlated, this interdependence does not exist. As an example of such a system, we shall consider in simplified form the ordering of thermal-electric power generation. The matrix P applies to demand transitions; ordering decisions, while affecting cost, do not affect future demand. For this reason estimation can be separated from control.

[‡] Most of the results of this chapter have also been incorporated into reference 48.

5.2 The Ordering Problem

The problem of ordering, or unit commitment, of thermal-electric power generating capacity may be defined as follows:

Given: 1) t_h, the minimum time interval required to bring a generating set on line if it is at present in a no-load standby condition;

2) the predicted future power demand;

3) the cost of bringing each set on line;

4) the running cost of each set (neglecting fixed costs);

5) penalty costs incurred for failure to meet demand;

6) present operating conditions (demand, number of sets now on line, condition of standby sets).

Required: to determine the number of generating sets, if any, which should be prepared at time t so that demand will be met at time $t + t_h$ with minimum expected overall cost.

In the treatment of this problem it is customary to base calculations on an expected demand curve[45-47] which may be updated periodically. Thus once a prediction has been made, the ordering problem becomes deterministic.

In the present formulation we consider future demand to be probabilistic in nature; moreover, since the number of machines running is a member of a discrete set, we consider that demand assumes discrete levels as well.

The basic system model is shown in fig. 5.1. A central load control unit is connected to a number of generating stations. The central unit predicts the overall load[44], and allocates it to the various stations. For allocation purposes each 24 hour period is divided into T equal intervals. At the beginning of each interval the central unit makes a demand on each plant, which remains constant throughout the interval at one of N discrete levels (corresponding to the output of an integer number of generators). The generators of each plant are ordered to meet the predicted demand in the following intervals in the most economical fashion. The demand level in the next interval is assumed to depend upon present demand and time of day (interval number). Demand transitions are thus described by T stochastic transition matrices, P(1), P(2), ... P(T), where

P(t) = N x N stochastic transition matrix whose elements $p_{ij}(t)$ equal the probability that if the demand level is i at interval t, it will be j at interval t + 1.

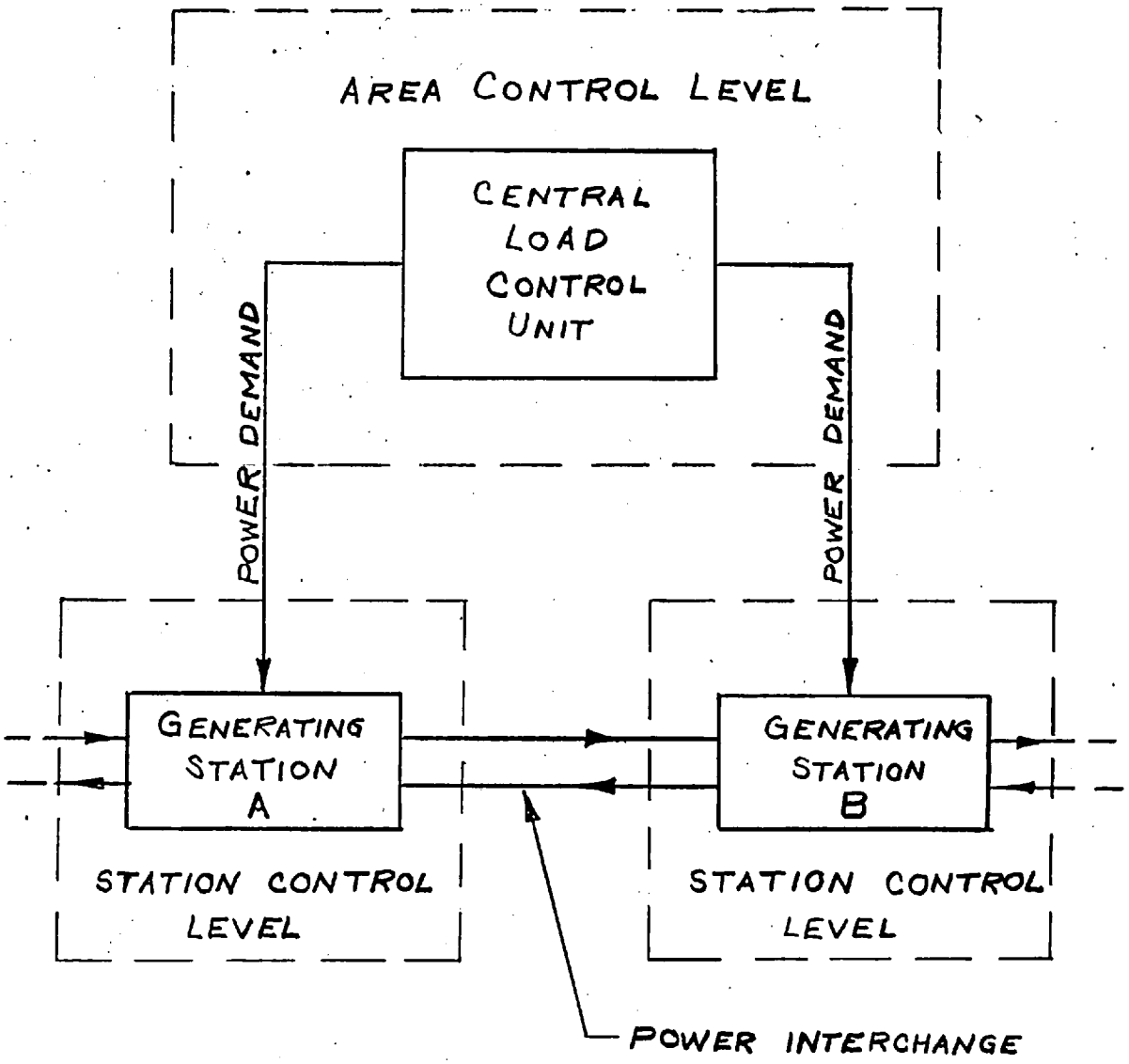


FIG. 5.1

POWER GENERATION

DECISION HIERARCHY

It should be noted that the number of levels, N, and the matrices P(t), may be different for each plant.

Generating sets may be in one of three conditions:

1) no load standby;

2) full load on-line output;

3) in transition from 1) to 2).

No cost is attached to condition 1), or the transition from 2) to 1). In condition 2) an operating cost is incurred. Completion of the transition 3) is assumed to require one time interval, and incurs a heating cost. If during any interval the available on-line generating capacity of a plant is insufficient to meet the demand, power may be purchased from an external source. Overall power costs are conveniently summarized by two cost matrices, B(t) and C(t), where

B(t) = N x N control cost matrix whose elements $b_{ij}(t)$ equal the cost of heating (j-i) additional generators for operation in interval t + 1, if "i" generators are running at the beginning of interval t. If $j \leq i$, $b_{ij}(t) = 0$.

C(t) = N x N operating cost matrix whose elements $c_{ij}(t)$ equal the cost of meeting a demand "j" in interval t + 1 if the number of sets made available for the purpose in interval t was "i".

In the present model it is assumed that if a set is ready to generate, but not required in interval t, then the cost of maintaining it in readiness for interval t + 1 is the same as the cost of bringing it from standby condition to readiness during the interval.[‡] This assumption is a drastic simplification of the real situation in which the heating history of each set for many past intervals is of importance. A study of this model serves as a convenient starting point, however, from which further development may be made. Suggested model improvements are considered in section 5.6.

5.3 Optimization Technique

To obtain an optimal ordering policy, it is necessary to consider an expanded state space. If at the beginning of an interval we wish to decide upon the ordering to meet demand in the following interval, we must consider not only the present demand level, but also the number of sets now running. Similarly, after a decision has been made, we must specify the number of sets available for the next interval. There are $\sum_{i=1}^{N} i = N(N+1)/2$ process states, the number of combinations of demand level and sets

---

[‡] The heat condition during the interval immediately preceding the present one may be accounted for simply by increasing the number of process states to N^2 (see section 5.3).

running being restricted by the fact that the latter cannot exceed the former. There are N^2 decision states since all combinations of present demand level and number of sets available for the following interval are theoretically possible.

In principle we may now derive from matrices P, B, and C new matrices P', B', and C' related to the expanded states (fortunately, these large matrices need never be stored in practice). Let us postulate a set of decision matrices D(t) defined by

$$D(t) = \left[\frac{N(N+1)}{2}\right] \times N$$ decision matrix whose

elements $d_{ij}(t)$ are the probabilities that if the process state is i at the beginning of interval t, a decision is made to go to decision state j.

The ordering of power generation using the model described is a cyclic decision process (section 2.8). Each basic interval of 24 hours is broken up into a series of probabilistic transitions (load demand changes) alternating with control decisions (generator ordering decisions). Thus one day's operation is described by the stochastic product matrix

$$\Gamma = P'(1)\, D(1)\, P'(2)\, D(2) \ldots\ldots P'(T)\, D(T) \qquad (5.1)$$

Optimum ordering is obtained by choice of the set $D^{*}(1)$, ... $D^{*}(T)$ which minimizes g, the expected daily generating cost. Because of the structure of the problem, it is more easily solved by dynamic programming than by the solution of a series of L x L matrix equations ($L = N^2$ in this case). Since the process is cyclic we consider that interval 1 follows interval T. Letting $V_i(t)$ be the cumulative cost if the state at interval t is i, we work backwards in time (beginning at $t = T$ with all $V_i(1) = 0$), adjusting successive matrices D(t) so that for i = 1, 2, ... N^2,

$$V_i(t) = \sum_{j=1}^{N(N+1)/2} p'_{ij}(t)\, c'_{ij}(t)$$

$$+ \sum_{j=1}^{N(N+1)/2} p'_{ij}(t) \cdot \underset{D(t)}{\text{Min}} \left\{ \sum_{k=1}^{N^2} d_{jk}(t)[b'_{jk}(t) + V_k(t+1)] \right\} \qquad (5.2)$$

with $T = 1 = 1$.

Equation (5.2) is equivalent to (2.9) except that in this case minimization is carried out as the recursion progresses. Provided the matrix Γ in (5.1) is ergodic (a condition which may be assumed safely in practice), continued application of (5.2) results in a stationary set of matrices $D^{*}(t)$ such that

$$g(\{D^{*}(t)\}) = \min_{\{D(t)\}} g$$

where g = expected daily power cost

$$= [V_i(t) - V_i(t+T)]_{\text{asymptotic}}$$

Note that the problem is in general a multi-stage one; physically this is so because the cost of power generated by a set depends upon the number of times it must be put on line and taken off again during the day. In deciding whether to order a set, the controller must consider its whole future operation relative to the rest of the system. An interesting feature of the present formulation, however, is that if the optimal policy specifies that no power be bought externally, single-stage optimization yields the same result as multi-stage optimization. As the amount of power bought increases (owing, for instance to a reduction of price) the relative performance of the single-stage system becomes progressively poorer.

Another point worth noting concerns the computation (5.2). The optimal matrices $D^{*}(t)$ have rows each containing one unit and N^2-1 zeroes. In practice it is necessary only to keep track of the position of the unit, so that a vector of $N(N+1)/2$ elements is sufficient to describe each $D^{*}(t)$.

5.4 A Simulation Study: Ordering in a 2000 Mw Station

To illustrate the usefulness of the discrete model, we shall consider as an example a generating plant with a maximum output of 2000 megawatts (Mw). Of this 1100 Mw is base load, being supplied by eleven 100 Mw sets. Demand may therefore assume any one of ten levels (base load plus the output of 0-9 additional sets). Table 5.7 shows the running costs associated with each of the ten sets generating the second 1000 Mw of total output (the output of set 1 actually forms part of the base load; its cost is listed for completeness, as its output represents one of the ten demand levels). Costs associated with the first 1000 Mw are not considered in any of the computations which follow; it may be assumed that the unit power cost associated with the first 1000 Mw is not greater than that of set 1, i.e. the cheapest sets are used for the base load. The heating cost for each generator is taken to be 10% of its full load running cost for one interval. Table 5.2 shows the unit cost of power bought externally. For simplicity, costs in tables 5.1 and 5.2 are assumed to be independent of time.

Demand transitions are made every two hours. The twelve transition matrices, listed in appendix 5, were

TABLE 5.1

POWER COSTS OF PEAK LOAD SETS

Set No.	Running Cost Pence/kw.hr.
1	0.510
2	0.520
3	0.520
4	0.530
5	0.540
6	0.560
7	0.580
8	0.580
9	0.590
10	0.600

TABLE 5.2

COST OF EXTERNAL POWER

Power Purchased Mw	Cost of Last 100 Mw Pence/kw.hr.
100	0.80
200	0.90
300	1.20
400	1.50
500	2.50
600	3.00
700	3.00
800	3.00
900	3.00
1000	3.00

chosen so that the resulting average demand curve, shown in fig. 5.2, is typical of a summer weekday for a station of this size[43]. The maximum variations from the mean are also shown in fig. 5.2. The extension to Saturday and Sunday operation increases the computer memory requirement, but is otherwise straightforward.

Application of equation (5.2) yields the ordering policy shown in table 5.3. Because of the simple relationship assumed between running and heating costs,

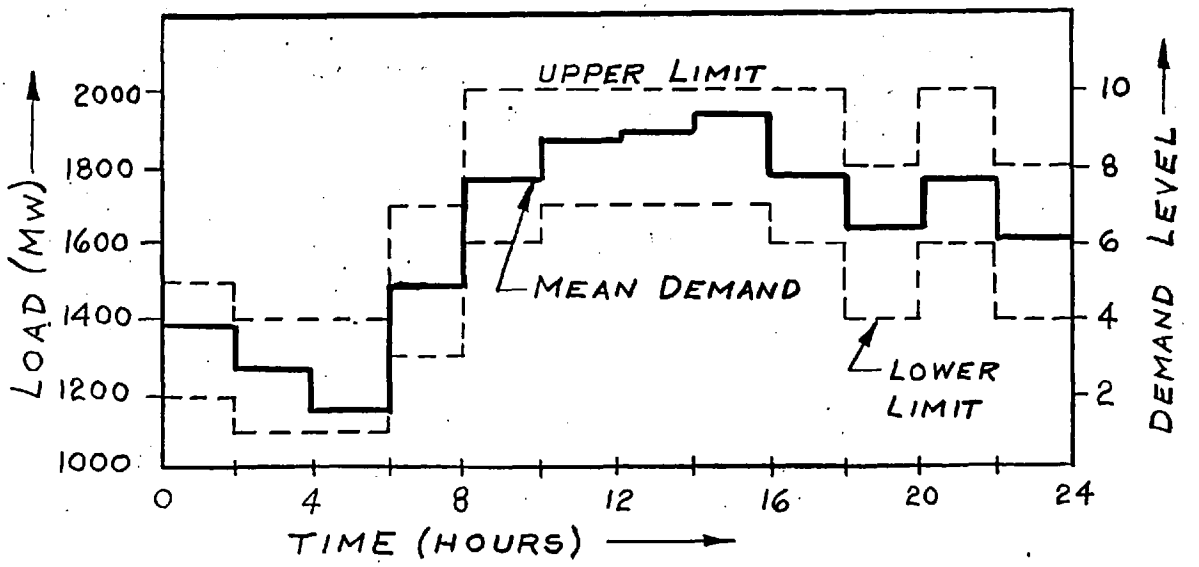


FIG. 5.2

MEAN WEEKDAY DEMAND

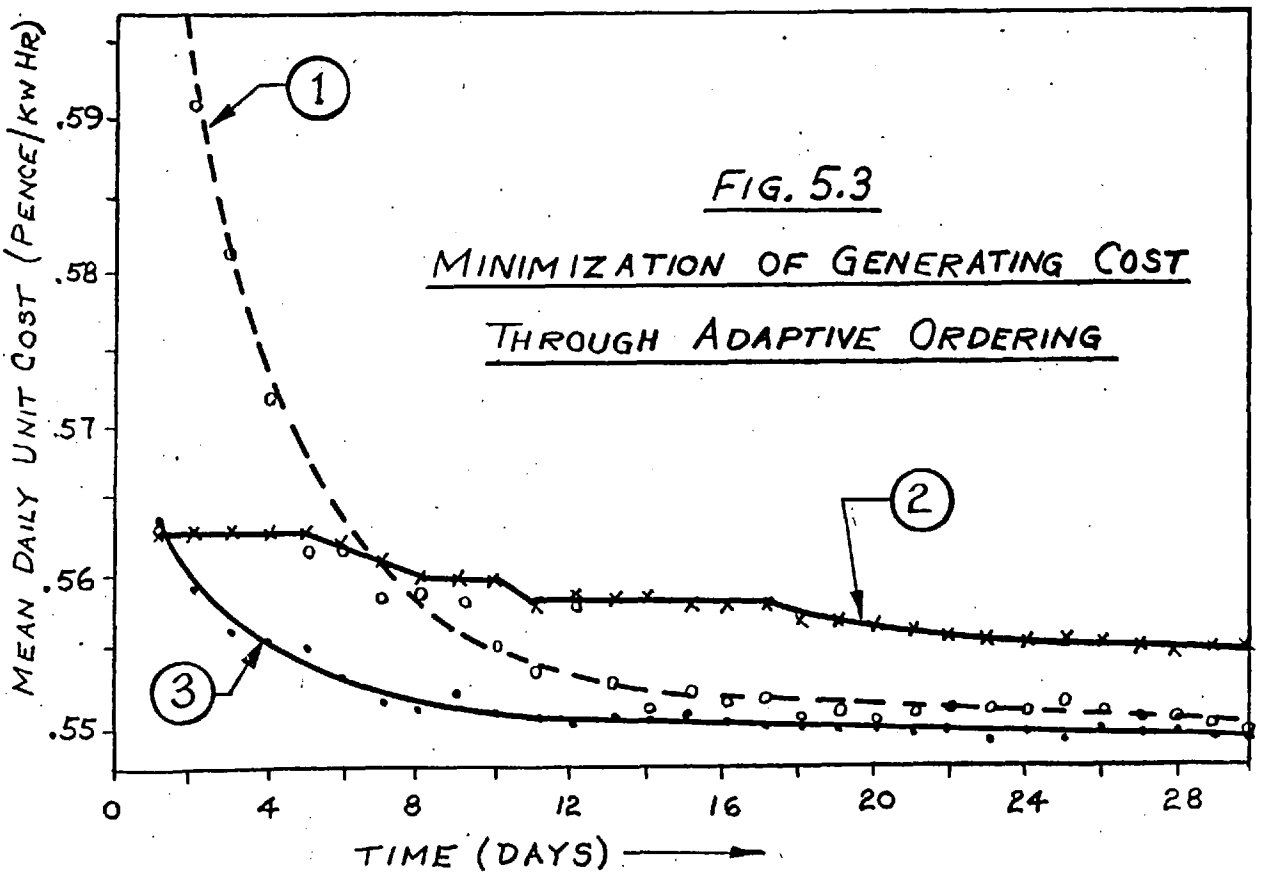


FIG. 5.3

MINIMIZATION OF GENERATING COST THROUGH ADAPTIVE ORDERING

TABLE 5.3

OPTIMAL ORDERING POLICY

Interval no.	Time	Load at Beginning of Interval	Minimum‡ Capacity Ordered for Next Interval
		Mw	Mw
1	0000-0200	200	200
		300	200
		400	300
		500	400
2	0200-0400	100	100
		200	200
		300	200
		400	300
3	0400-0600	100	500
		200	600
		300	600
		400	700
4	0600-0800	300	700
		400	800
		500	800
		600	900
		700	900
5	0800-1000	600	800
		700	900
		800	900
		900	1000
		1000	1000
6	1000-1200	700	800
		800	900
		900	1000
		1000	1000
7	1200-1400	700	800
		800	900
		900	1000
		1000	1000

8	1400-1600	700	700
		800	700
		900	800
		1000	800
9	1600-1800	600	500
		700	600
		800	700
		900	700
		1000	800
10	1800-2000	400	600
		500	700
		600	800
		700	900
		800	900
11	2000-2200	600	500
		700	600
		800	700
		900	700
		1000	800
12	2200-0000	400	300
		500	400
		600	400
		700	500
		800	500

‡ If the generating capacity actually on line during the interval is greater than that in this column, then the on-line capacity is, perforce, the one available for the next interval. If on-line capacity is less than ordered capacity, more sets must be prepared.

it was found that the total number of sets which should be made available for the next interval depends only upon present load, not upon the number of sets running (except that the number of sets in readiness at the beginning of interval t + 1 physically cannot be less than the number actually on line during interval t).

Using the tabulation of this optimum ordering policy, we can compute the approximate average amount of power which will be bought from outside the plant. Suppose that at interval t the load assumes level "i" with probability π_i, and that the power output level ordered for the next interval (obtained from table 5.3) is q = q(i,t). If the load in interval t + 1 assumes level "j", then a quantity of energy e(i,j,q) is purchased. The quantity e is given by

$$e(i,j,q) = 200 \left\{ j(t) - \text{Max}[i,q] \right\} \text{ Mw hr.}, \qquad j \geq \text{Max}[i,q]$$

$$= 0, \qquad j < \text{Max}[i,q]$$

The figure 200 arises from the fact that for each unit of positive difference between j and Max[i,q], the output of one generator (100 Mw) for one time interval (2 hours) must be purchased. The total daily energy purchase, W,

is obtained by summing e(i,j,q) over all next-interval demands "j", all present loads, "i", and all time intervals, t. Thus

$$E(W) = \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_i(t)\ p_{ij}(t)\ e(i,j,q) \qquad (5.3)$$

Inspection of matrices P(1) to P(12) (appendix 5) and of table 5.3, together with equation (5.3), allows us to construct table 5.4, which shows the distribution of power purchase probabilities throughout the day. As one would expect, purchases are most likely to be made at times of rising demand. The quantity of power purchased during any interval is never greater than 100 Mw; the expected daily energy purchase is 93.8 Mw hr., nearly one-half of the output of one machine for one interval. Therefore a purchase will be made slightly less frequently than one day in two.

The total energy demand, Q, made upon this plant during a day is

$$E(Q) = 200 \left[\sum_{t=1}^{T} \sum_{i=1}^{N} i.\pi_i(t) \right] \text{ Mw hr.} \qquad (5.4)$$

From (5.4) we find the total expected demand to be 15,459.4 Mw hr. Thus the plant when optimally ordered

TABLE 5.4

DISTRIBUTION OF EXPECTED POWER PURCHASES THROUGHOUT THE DAY

Interval t	Load Level i(t)	$\pi_i(t)$	Probability of 100 Mw Purchase in Next Interval	Mean Energy Purchased Mw hr.	Total Energy Purchased in Next Interval Mw hr.
2	1	0.0714	0.05	0.714	0.714
3	1	0.4582	0.10	9.164	15.949
	2	0.3869	0.05	3.869	
	3	0.1458	0.10	2.916	
4	3	0.0652	0.20	2.608	18.025
	5	0.5139	0.15	15.417	
5	5	0.0509	0.20	2.036	23.456
	8	0.5355	0.20	21.420	
6	7	0.0153	0.20	0.612	12.268
	8	0.2914	0.20	11.656	
7	7	0.0353	0.10	0.706	3.050
	8	0.1172	0.10	2.344	
8	7	0.0106	0.10	0.212	0.212
10	4	0.0083	0.20	0.332	20.166
	5	0.0703	0.10	1.406	
	6	0.4394	0.10	8.788	
	7	0.3766	0.10	7.532	
	8	0.1054	0.10	2.108	

Total Expected Daily Purchase, E(W) 93.840

generates $[\frac{E(Q)-E(W)}{E(Q)} \times 100]\ \% = 99.39\%$ of demanded above-base load over a long period of time when external power costs 0.80 pence/kw hr.

Recall that when no power is purchased externally, single-stage and multi-stage policies yield the same result. What happens if almost no power (0.61% in this case) is purchased? Is a single-stage policy then almost optimum? The answer seems to be affirmative. A single-stage policy was computed for the present system; it was found to be identical to that in table 5.3 except at one point: with a present demand of 900 Mw at the beginning of interval 6(10 AM) only 900 Mw capacity is to be made available for the following interval with the single-stage policy. The expected daily cost, $g(D^{*})$, was computed as £35,407 (0.549681 pence/kw hr.) with the multi-stage policy of table 5.3, and £35,409 (0.549712 pence/kw hr.) with the single-stage policy. For the system under study, multi-stage optimization requires about 20 seconds on an IBM 7090 computer, while single-stage optimization requires 1% of this time.

The danger of using a single-stage policy becomes evident when the cost of external power is reduced. Suppose that external power is available at a flat rate of 0.560 pence/kw hr. Inspection of table 5.1 shows that

it is probably economical to operate sets 1 to 5, but not sets 6 to 10. A simulation using multi-stage optimization verifies this conjecture; sets 1 to 5 run more or less normally, while all power above the 500 Mw level is supplied from the external source. Overall unit cost is then 0.537 pence per kw hr. Single-stage optimization, on the other hand, shuts down all sets above base load (set 1); the reason is that, once a set is off-line, it never appears worthwhile to re-heat it for only one interval of time. The resultant overall cost is 0.552 pence per kw hr.

5.5 Adaptive Ordering

When the demand matrices, $P(t)$, are unknown or time-varying, it is necessary to use an adaptive form of ordering. Each row of the basic matrix $P(t)$ is repeated several times throughout the expanded matrix $P'(t)$ used in (5.2); every time the estimate of a particular row of $P'(t)$ is updated, therefore, all of the other rows in $P'(t)$ known to be identical to that row are similarly updated (we note again that this operation is conceptual since only $P(t)$, and not the expanded matrix $P'(t)$ need be stored). Because of this correlation the set of rows of $P'(t)$ which is updated

is independent of the decisions made. We are therefore in the fortunate position of being able to separate the problems of estimation and control.

Suppose that we have been observing the system for n days. The asymptotic maximum likelihood estimates of the elements of P(t) are given by

$$\hat{p}_{ij}(t,n) = \frac{m_{ij}(t,n)}{\sum_{k=1}^{N} m_{ik}(t,n)} \qquad (5.5)$$

where M = N x N observation matrix whose elements $m_{ij}(t,n)$ equal the number of observed transitions from demand level "i" in interval t to demand level "j" in interval t + 1, after operation has been observed for n days.

Equation (5.5) is equivalent to (3.11).

Using the estimates P(t), we may then compute and use the estimated optimal policy $\left\{\hat{D}^{*}\right\}$, from (5.2).

To start the process initial observation matrices, M(t,o), are needed. As well as containing our best initial estimates - guesses, perhaps - of P(t), the matrices M(t,o) reflect our confidence in the estimates through the initial weighting factor

$$a_{io}(t) = \sum_{j=1}^{N} m_{ij}(t,o) \qquad (5.6)$$

a_{io} is the number of hypothetical demand transitions which are considered to have been observed from state i at time t, before physical observations have begun. Usually one sets $a_{io}(t) = a_o$ for all i and t. In such a case the weight of the prior estimates M(t,o) is equivalent to that obtained as a result of $a_o N$ days of observation. We might call a_o the "confidence factor". If it is large, many further observations are required to affect the estimates appreciably; the learning process is initially conservative. If a_o is small estimates P(t) early in the process are highly dependent upon initial observations, i.e. little confidence is placed in the prior estimates M(t,o).

Once the process has begun the matrix M is updated at each interval of operation. After observing a transition on day n from demand level k in interval t to demand level ℓ in the following interval, we update M(t,n) as follows:

$$m_{ij}(t,n) = m_{ij}(t,n-1) + \delta_{k\ell} \qquad (5.7)$$

where $\delta_{k\ell} = 1$, if $i = k$ and $j = \ell$

$= 0$, otherwise.

If the system is slowly time-varying, we may conveniently model the drift of the elements $p_{ij}(t)$ as an exponentially moving average process. We then reduce the weighting of past information at each interval by using equation (5.8) before (5.7).

$$m_{ij}(t,n) = \beta\, m_{ij}(t,n-1) \qquad (5.8)$$

$$i,j = 1, 2, \ldots N$$

where $0 \leq \beta \leq 1$

For a stationary process $\beta = 1$, and all past information is considered to have equal relevance. At the opposite extreme $\beta = 0$ describes a system in which only the last observation is significant.

To determine the duration of the transient associated with the learning process, simulation of the adaptive ordering of the generating plant described in section 5.4 was carried out for the case in which the matrices P(t) are stationary ($\beta = 1$). The learning process itself is non-stationary, so that it is necessary to consider an ensemble of plants. Thirty days' operation of an ensemble of 100 plants was simulated, and the results examined. Single-stage optimization was used, as it is faster and yields results which are very close to those with multi-stage optimization for this particular problem.

An examination of the results shows that efficient convergence is considerably affected by choice of the confidence factor, a_o. Fig. 5.3 shows the evolution of power cost with adaptive ordering. Curve 1 results when demand (above 1000 Mw) is estimated a priori to be equally distributed between 600 and 1000 Mw from 10 AM to 10 PM, and between 100 and 500 Mw during the remainder of the day; a_o is unity. Operation of the system at first results in high expense, owing chiefly to failure to meet the sharp rise in demand which occurs between 6 AM and 10 AM. However, adaptation is quite rapid; unit cost is less than 1% above its theoretical minimum value within 12 days.

Since the cost of failure to meet demand is so high, a new initial estimate was made in which demand above 1000 Mw was assumed to be 1000 Mw between 2 AM and 8 PM, and 500 Mw during the remainder of the day. This estimate results in over-ordering initially; too many generators are kept in readiness for demands which do not occur. Nevertheless, this conservative policy is much better initially than the first one; power costs about 0.564 pence/kw hr. on the first day of operation, rather than 0.631. In curve 2 the evolution of unit cost with this initial estimate is shown for $a_o = 1$. The drawback here is that with all prior "observations" concentrated

on one demand curve (instead of a distribution, as in curve 1), the controller is slow to "forget" the initial estimates, and convergence is severely hampered by them. Notice that six days elapse before any significant change in ordering policy occurs. Even at day 30, the controller obviously anticipates and prepares for such unlikely (impossible, with our process model) events as full load at 4 AM.

The step taken to combine the rapid convergence of curve 1 with the relatively low initial cost of curve 2 is the reduction of a_o. Setting $a_o = 10^{-6}$ we obtain curve 3, which begins at approximately the level of curve 2, but reduces to less than 1% above minimum cost within five days. The general rule which results from an examination of these curves coincides with our intuitive notion: if the demand transition matrices are unknown and external power cost is high, the initial estimate should ensure that virtually 100% of demand can be met at every interval; the weight given to this estimate should be small, so that as the true demand parameters are learned, the effect of the prior estimate quickly becomes negligible.

Table 5.5 compares the expected generating cost incurred in thirty days' operation using

1) theoretical optimum policy;

2) adaptive ordering (fig. 5.3, curve 3);

3) non-adaptive ordering using fixed policy to ensure that 100% of demand is met (initial policy of curve 3).

Note that while the latter policy enables the plant to meet its power demands satisfactorily, the total cost of thirty days' operation is nearly £23,000 greater than in the adaptive case. Complete a priori knowledge of demand transition matrices would have saved an additional £5000.

TABLE 5.5

COST OF THIRTY DAYS' OPERATION

Method of Operation	Mean Unit Cost Pence/kw.hr.	Cost of 30 Days' Operation £
1) All P(t) known a priori (theoretical minimum)	0.549712	1,062,280
2) Initial estimate as in fig. 5.3, curve 3, with adaptation	0.552326	1,067,331
3) Initial estimate as in fig. 5.3, curve 3, no adaptation	0.564117	1,090,117
Thirty days' saving using adaptive policy instead of fixed policy [3)-2)]		£22,786
Adaptation cost [2)-1)]		£ 5,051

5.6 Conclusion

As an example of a cyclic decision process, we have considered in this chapter a simplified version of the ordering problem in a thermal-electric power station. The advantage of the discrete state probabilistic model is that it allows us to formulate a policy which is optimal, not only for mean future demand, but for deviations from it as well. In addition, the expected effects of future operation on present decisions and vice versa are easily computed. Since the decision policy does not affect the manner in which the demand transition matrices are estimated, the problem is "neutral" in the sense of Feldbaum[19]; i.e. we may use a pure control strategy based on maximum likelihood estimates, rather than a dual control strategy. This feature facilitates the problem of adaptive operation considerably. A simulated example of adaptive ordering, in which power demand is initially unknown, has been presented. Convergence to the optimum ordering policy was found to be rapid providing that the a priori demand estimates are not heavily weighted.

In the present model, base load is considered constant; much known information concerning daily demand fluctuations is thus discarded. A well known technique[44,49] is that in

which the daily demand curve is decomposed into a sliding mean with a superimposed random component. Describing this residual random component as a Markov chain, we could handle the problem described in this chapter with a five-state, instead of a ten-state, system.

The system model used in this chapter is a greatly simplified one. In practice it is necessary to consider the past heating history of each set when making ordering decisions. This more realistic problem is therefore the next one which should be attacked if we wish to investigate further the value of the discrete Markovian model.

Two approaches suggest themselves: first, by adjoining the heating history of each set to the present state, we may retain the Markov property. Both process and decision states could then be defined by the following information:

1) present demand;
2) number of sets available to meet demand;
3) number of intervals which have elapsed for each remaining set since last it was used or available for generation.

If N generators are dealt with in any one interval and past heating history extends τ intervals in the past, the total number of states necessary is

$$L = N \sum_{i=1}^{N} \tau^{i-1} \qquad (5.9)$$

With $\tau = 5$, a three generator problem ($L = 93$) can easily be handled, and a four generator problem ($L = 625$) is feasible on a fast computer.

As a second approach, consider a state description which specifies 1) and 2) as above, together with

3) heat state of off-line sets; i.e. total heat content and apportionment of this heat amongst them.

In this case $L = \tau N^3$. A five generator problem with $\tau = 5$ ($L = 625$), or a four generator problem with $\tau = 10$ ($L = 640$) could be handled. Strictly speaking, the factors relating heat apportionment depend upon past states, it is conjectured that this is a second order effect, and that the system remains "approximately Markov".

Using either of these models, many, probably most, states in the resultant Markov chain would be inessential. If these can be isolated, the problem can be reformulated without them, with a great saving in memory requirement and computing time. The difficulty is that the inessential states do not constitute an easily identified set, as they did in chapter 2. Further research on the problem might prove interesting.

CHAPTER 6

DUAL CONTROL OF MULTI-STAGE MARKOV PROCESSES

6.1 Introduction

It has been shown in chapter 2 that in batch processes, as defined in section 2.7, the controller need consider only one stage of operation to determine an optimum policy. The dual control problem arising when the transition matrix, P, is initially unknown has been studied in some detail in chapters 3 and 4 for the single-stage process. It was believed when the study was begun that the solution to this problem would suggest a method of attacking the more difficult one which arises when control effort is costed; indeed, this has proved to be the case.

In this chapter we shall consider discretized versions of processes described by the equation

$$x_{n+1} = f(x_n, u_n, \xi_n) \qquad (6.1)$$

where x_n = output variable at time interval n

u_n = control variable at time interval n

ξ_n = disturbance at time interval n.

For the present, we assume that (6.1) is a scalar equation. The dynamics are assumed to be continuous and differentiable but may be non-linear, and either stable, conditionally stable, or unstable. Disturbances associated with any decision state ($\S_n = \S_n(x_n, u_n)$) are assumed to form independent sequences drawn from continuous distributions. However, like the process dynamics, the disturbance statistics are assumed to be unknown, and may be multiplicative and non-gaussian. Both state and control variables are constrained in a known fashion.

$$x_n \in X \tag{6.2}$$

$$u_n \in U \tag{6.3}$$

A known cost, $L(x,u)$, is associated with the operation of the system; it is assumed that the latter is stationary and will run indefinitely. Sampled data control is to be used. The object of control is to minimize the expected cost per sampling interval of process operation, and to avoid unnecessarily expensive estimation procedures while determining the optimal control policy.

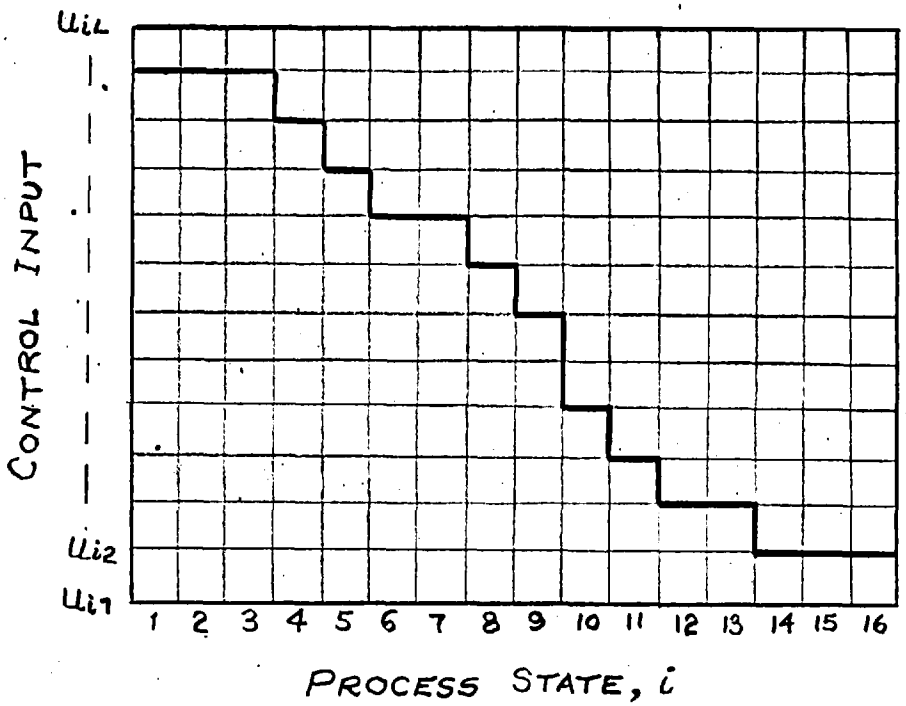
In the control of batch processes, the initial state can be chosen arbitrarily at the beginning of each new transition. In the present case, however, the control and disturbance signals act simultaneously between sampling

instants. The choice of control affects the probability distribution of process states at the following sampling instant, but cannot specify the process state deterministically. The determination of an optimal feedback policy is equivalent to the specification of a mapping of the set of process states into a subset of the decision states, as discussed in chapter 2. Corresponding to each process state is a quantized range of admissable control inputs. If we take (6.3) to mean

$$u_{min} \leq u_n \leq u_{max} \qquad (6.4)$$

then we might quantize the control variable into Q equal intervals. If the output variable is similarly quantized into N equal intervals, then there are N process states and L = NQ decision states. In effect we have placed a uniform grid over the $\{x,u\}$ space. The optimum feedback policy is thus a piecewise constant function, as shown in fig. 6.1.

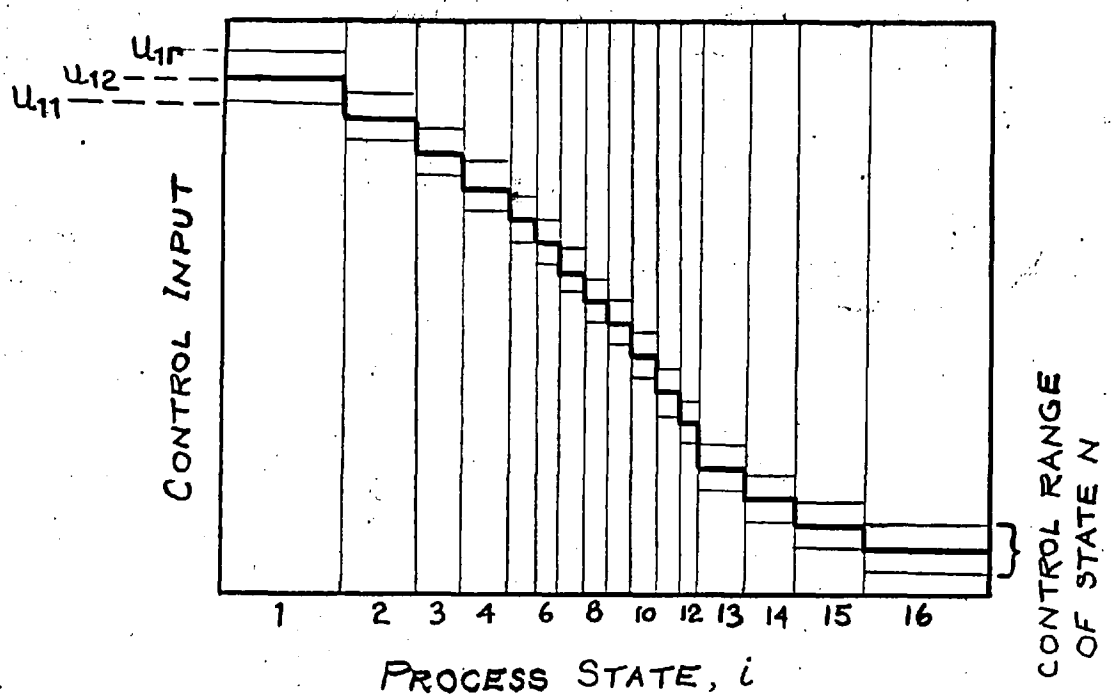
The disadvantage of this approach is that in order to obtain a reasonable approximation of the continuous feedback characteristic, it is necessary to quantize very finely, so that the number of states becomes excessive. We avoid this difficulty by quantizing non-uniformly, as shown in fig. 6.2. The most finely quantized part of the output variable range is that whose probability of occupancy is highest[32]. As



N = NO. OF PROCESS STATES = 16
L = NO. OF CONTROL ALTERNATIVES = 13
NL = NO. OF DECISION STATES = 208

FIG. 6.1

UNIFORM QUANTIZATION



N = NO. OF PROCESS STATES = 16

Γ = NO. OF CONTROL ALTERNATIVES = 3

$N\Gamma$ = NO. OF DECISION STATES = 48

FIG. 6.2

NON-UNIFORM QUANTIZATION

to the control signal range, only a small part of it need be considered for use with a particular process state. Referring to fig. 6.2, we may define the control range associated with a particular process state as the set of decision states which are reachable in one step from the process state. Typically for each process state there may be, say, five possible control alternatives; the combination of the process state with each of these defines five admissable decision states. It is assumed that the optimum control lies somewhere in the admissable range (providing the optimum is within the constraint (6.4)). We shall consider a method quantizing the output variable and specifying suitable control ranges in section 6.5.

6.2 A Multi-Stage Dual Control Algorithm

We have seen in chapter 2 that the minimization of g, the expected cost per interval of operation, is equivalent to the construction of a decision matrix D^{*}, so that for each process state, i, control j is chosen to minimize the parameter η_{ij}, $j = 1, \ldots L$. As noted previously we have restricted the control range available for each process state to, say, Γ alternatives. Thus $L = N\Gamma$ in the present case, but the minimization of η_{ij} is done only

over the Γ admissable alternatives, $j = 1, \ldots \Gamma$. If P is unknown, only the estimate $\hat{\eta}_{ij}$ is available. Even if the rows of P and C are independent, all of the parameters $\hat{\eta}_{ij}$ are highly correlated; because of this fact it is not feasible to compute an overall probability of error. Instead, we make use of the fact that equation (2.14) transforms a multi-stage problem into an equivalent single-stage one. We therefore introduce, as a measure of error probability, the uncertainty, Ω, which is defined in terms of the estimates associated with the equivalent single-stage problem.

Let
$$\Omega = \sum_{i=1}^{N} a_i \Omega_i \tag{6.5}$$

where

a_i = constant associated with process state i

Ω_i = single-stage uncertainty associated with control used for process state i

i.e.
$$\Omega_i = 1 - \int_{-\infty}^{\infty} f_{i\hat{s}}(x) \prod_{\substack{k=1 \\ k\neq\hat{s}}}^{\Gamma} [G_{ik}(x)]\, dx \tag{6.6}$$

$$f_{ik}(x) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}_{ik}} \exp\left[-\frac{1}{2}\left(\frac{x-\hat{\eta}_{ik}}{\hat{\sigma}_{ik}}\right)^2\right] \tag{6.7}$$

$$G_{ik}(x) = \int_{x}^{\infty} f_{ik}(y)\, dy \qquad (6.8)$$

$$\hat{\sigma}_{ik}^{2} = \frac{1}{n_{ik}} \Big[\sum_{j=1}^{N} \hat{p}_{ijk}(1-\hat{p}_{ijk})c_{ij}^{2} - 2 \sum_{j=1}^{N-1} \Big(\sum_{q=j+1}^{N} \hat{p}_{ijk}\, \hat{p}_{iqk}\, c_{ij}\, c_{iq} \Big) \Big] \qquad (6.9)$$

i.e. $\hat{\sigma}_{ik}^{2}$ = variance estimate of cost of one transition from process state i when control alternative k = k(i) is chosen.

n_{ik} = number of transitions observed from process state i when control alternative k is used.

$$\hat{p}_{ijk} = \frac{m_{ijk}}{n_{ik}} \qquad (6.10)$$

i.e. $\hat{p}_{ijk}$ = estimate of probability of transition from process state i to process state j when control alternative k is used.

m_{ijk} = number of transitions observed from process state i to process state j when control alternative k is used.

c_{ij} = cost of transition from process state i to process state j.

Equations (6.6) to (6.10) are analogues of equations (3.20), (3.17), (3.19), (3.16) and (3.11) respectively. We might now follow through the development of chapter 3 to obtain the following analogue of the single-stage optimal strategy:

Let the present process state be i, and suppose a control set, $\{u_{ik}\}$, $k = 1, 2, \ldots \Gamma$, is available.

Let $\hat{\eta}_{i\hat{s}} = \underset{k}{\text{Min}} \{\hat{\eta}_{ik}\}$

1) Choose $k = \hat{s}$ with probability $\theta_{i\hat{s}}$, where

$$\left.\begin{aligned} \theta_{i\hat{s}} &= 1 - \gamma\alpha, && \gamma\alpha < 0.5 \\ \theta_{i\hat{s}} &= 0.5, && \gamma\alpha \geq 0.5 \end{aligned}\right\} \qquad (6.11)$$

with $$\alpha = \sum_{\substack{k=1 \\ k\neq\hat{s}}}^{\Gamma} F_{ik}(\hat{\eta}_{i\hat{s}})$$

and $F_{ik}(\hat{\eta}_{i\hat{s}}) = 1 - G_{ik}(\hat{\eta}_{i\hat{s}})$

2) If $k = \hat{s}$ is not chosen by application of (6.11), choose alternative j from amongst the remaining $\Gamma - 1$ decision states so that

$$\frac{\exp(-\hat{\rho}_{ij}^{\,2}/2)}{\hat{\sigma}_{oij}(n_{ij})^{1/2}} = \underset{k}{\text{Max}} \left[\frac{\exp(-\hat{\rho}_{ik}^{\,2}/2)}{\hat{\sigma}_{oik}(n_{ik})^{1/2}} \right], \qquad (6.12)$$

$$k = 1, 2, \ldots \Gamma$$
$$k \neq \hat{s}$$

where

$$\rho_{ik} = \frac{\hat{\eta}_{ik} - \hat{\eta}_{i\hat{s}}}{\hat{\sigma}_{ik}}$$

$\hat{\sigma}_{oik}$ = square root of variance computed as in (6.9) with present estimates $\hat{p}_{ijk}$, but with $n_{ik} = 1$.

Each parameter Ω_i is decreased optimally with this strategy. Moreover, the theory of chapter 3 shows that each Ω_i is asymptotically driven to zero, providing the probability of occupancy of process state i is non-zero (if the latter condition is not met, process state i is inessential and can be disregarded). All variances approach zero with time according to the results of section 3.18. Thus not only Ω, but the error probability ω must approach zero with time, so that the strategy is convergent. Since both the estimation requirement (3.58) and the control requirement (3.59) are met, the algorithm defines a feasible dual strategy. While it is based on a single-stage optimal strategy, the multi-stage version is not necessarily itself an optimal strategy. Nevertheless it converges efficiently in practice to the optimal policy. A FORTRAN version of this algorithm is presented in appendix 6a.

6.3 Updating the Estimates

We have presented a dual strategy which treats the overall system as a collection of interacting single-stage systems which may be isolated for decision purposes. The interaction effects are accounted for, on-line, by updating of the η matrix after each transition has been observed. Since the matrix $\hat{P}$ changes after each observation, so does the transformation matrix Ψ^{-1}. It would appear at first sight that the matrix equation (2.14) would have to be re-solved at least once after each transition. Such a scheme seems computationally prodigal and intuitively unreasonable. After many transitions have taken place, the effect of the latest observation is small, and the estimated optimal policy is likely to remain unchanged.

We note that each observation changes only one row of $\hat{P}$, and therefore only one row of Ψ. We shall show in this section that the change in Ψ^{-1} is expressible as a dyad, and that matrix inversion may be replaced by scalar inversion for updating purposes.

Let $\hat{P}_b(n)$ = estimated stochastic transition matrix of basic chain (N states) at stage n.

The basic chain is made up of the N decision states defined by

process state is i, control used is $u_{i\hat{s}}$

where $\hat{s} = \hat{s}(i)$ is defined by

$$\hat{\eta}_{i\hat{s}} = \min_k \left\{\hat{\eta}_{ik}\right\}$$

If the process state at stage n is i, then the control u_{ik} which is applied is either the estimated optimal one, $u_{i\hat{s}}$, or one of the remaining Γ-1 alternatives for which $k \neq \hat{s}$. We shall assume for the moment that $u_{i\hat{s}}$ is applied; at the beginning of interval n + 1 we therefore observe that row i of matrix $\hat{P}$ has changed. Using the notation of chapter 2 we may write

$$\hat{P}_b(n+1) = \hat{P}_b(n) - \underline{e}_i><\underline{e}_i\ \hat{P}_b(n) + \underline{e}_i><\underline{e}_i\ \hat{P}_b(n+1)$$

$$\hat{P}_b(n+1) = \hat{P}_b(n) + \underline{e}_i><\underline{e}_i(\hat{P}_b(n+1) - \hat{P}_b(n)) \qquad (6.13)$$

Equation (6.13) expresses the fact that only one row of $\hat{P}_b$ has changed. From (2.20) and (6.13),

$$\Psi(n) = [I + \underline{w}><\underline{e}_N - \underline{e}_N><\underline{e}_N][\hat{P}_b(n)\cdot(\underline{e}_N><\underline{e}_N - I)]$$

$$\Psi(n+1) = [I + \underline{w}><\underline{e}_N - \underline{e}_N><\underline{e}_N]$$

$$+ [\hat{P}_b(n) + \underline{e}_i><\underline{e}_i\cdot(\hat{P}_b(n+1) - \hat{P}_b(n))][\underline{e}_N><\underline{e}_N - I]$$

$$\Psi(n+1) = \Psi(n) + \underline{e}_i><\underline{e}_i\cdot(\hat{P}_b(n+1) - \hat{P}_b(n))\cdot(\underline{e}_N><\underline{e}_N - I)$$

i.e. $$\Psi(n+1) = \Psi(n) + \underline{e}_i><\underline{\alpha} \qquad (6.14)$$

where $$\langle \underline{\alpha} = \langle \underline{e}_i \cdot (\hat{P}_b(n+1) - \hat{P}_b(n)) \cdot (\underline{e}_N \rangle \langle \underline{e}_N - I) \quad (6.15)$$

From (6.14) the matrix inversion lemma yields

$$\Psi^{-1}(n+1) = \Psi^{-1}(n) - \Psi^{-1}(n)\underline{e}_i \rangle [\langle \underline{\alpha} \Psi^{-1}(n)\underline{e}_i \rangle + 1]^{-1} \langle \underline{\alpha} \Psi^{-1}(n)$$

Since the inverted term is a scalar, we have

$$\Psi^{-1}(n+1) = \Psi^{-1}(n) - \frac{\Psi^{-1}(n)\ \underline{e}_i \rangle \langle \underline{\alpha} \Psi^{-1}(n)}{\langle \underline{\alpha} \Psi^{-1}(n)\underline{e}_i \rangle + 1} \quad (6.16)$$

The updated transformation matrix is easily computed with (6.16). The vector $\hat{\underline{z}}(n+1)$ is then computed from (2.14), giving us $\hat{g}$ together with the parameters $\hat{v}_{i\hat{s}}$, $i = 1, \ldots N$. The remaining parameters, $\hat{v}_{ik}$, associated with inessential decision states, are completely dependent upon the values $\hat{v}_{i\hat{s}}$.

$$\hat{v}_{ik} = \hat{\ell}_i + \sum_{j=1}^{N} \hat{p}_{ijk}(b_{j\hat{s}} + \hat{v}_{j\hat{s}}) - \hat{g}, \quad (6.17)$$

$$i = 1, \ldots N$$
$$k = 1, \ldots \Gamma$$

= estimate of relative cost of occupation of decision state (i,k), i.e. the decision state defined by the choice of control alternative k when the process state is i.

The test parameters $\hat{\eta}_{ik}$ are now given by

$$\hat{\eta}_{ik} = b_{ik} + \hat{v}_{ik} \qquad (6.18)$$

A check is then made to determine whether or not $\hat{D}^{*}(n+1)$ differs from $\hat{D}^{*}(n)$. If it does not, then updating is complete. Otherwise a new matrix $\Psi(n+1)$ must be derived from $\hat{D}^{*}(n+1)$ and inverted. The loop is then repeated as explained in section 2.4.

We have assumed that the control used was $u_{i\hat{s}}$. If u_{ik}, $k \neq \hat{s}$ is applied instead, the updating procedure is even simpler, because the decision state (i,k) is not in the basic chain. The matrix Ψ^{-1} and the parameters $\hat{v}_{i\hat{s}}$ therefore remain unchanged. We need only update $\hat{v}_{ik}$ for the particular k chosen, starting the procedure at equation (6.17). A simplified flow chart of the algorithm is shown in fig. 6.3, and a FORTRAN version is presented in appendix 6b.

As in chapter 5, we may economize considerably in the storage of matrix $\hat{D}^{*}$. Every row of $\hat{D}^{*}$ contains one unit and $\Gamma - 1$ zeros; the same information is conveyed by an N-vector whose i^{th} element is $\hat{s}(i)$.

6.4 The Optimal Feedback Transducer Characteristic

We have seen in section 6.1 that a discrete approximation

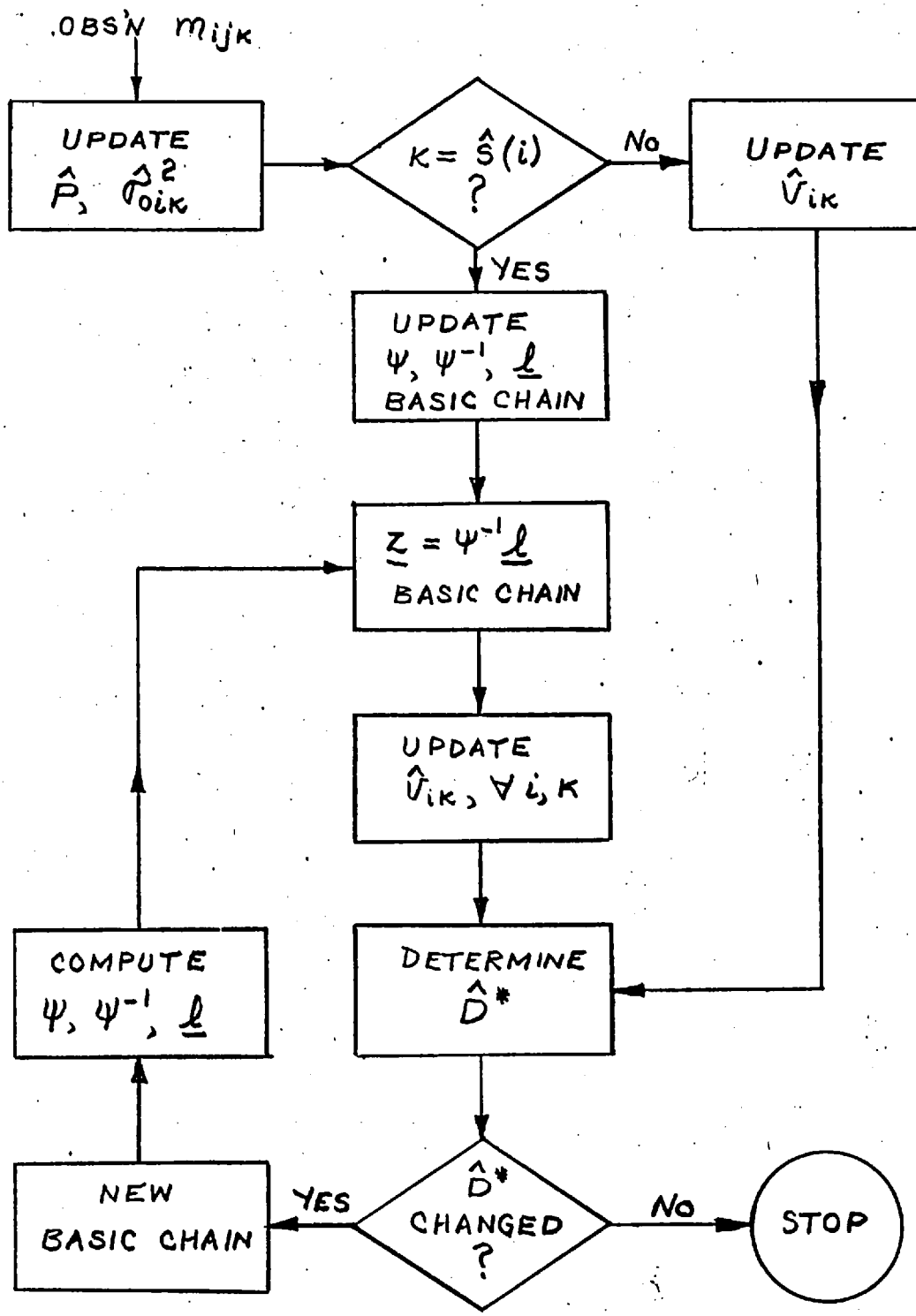


FIG. 6.3

UPDATING ROUTINE

to the optimal feedback policy can be developed from the automaton model of a continuous state process. In this section we shall present an algorithm which combines continuous and discrete optimization methods to synthesize a continuous optimal feedback transducer characteristic for a non-linear noisy system.

We shall assume that a continuous state, discrete time model of the system is known, i.e. equations (6.1) to (6.3) are given. Moreover, we shall assume that the right hand side of (6.1) is a continuous differentiable function of x and u disturbed by noise whose probability density function is continuous and differentiable. If these conditions also apply to the cost function L(x,u), then we may reasonably assume that the expected cost per transition, g, is a continuous function of the feedback transducer characteristic, u(x). Let $u^{*}(x)$ be the optimal characteristic, as shown in fig. 6.4, and let $\beta(x)$ be an arbitrary differentiable function. If ϵ is a small number, then a curve neighbouring the optimal one may be expressed as

$$\bar{u}(x) = u^{*}(x) + \epsilon\,\beta(x) \qquad (.6.19)$$

A necessary condition for g to be a minimum is that

$$\left.\frac{\partial g}{\partial \epsilon}\right|_{\epsilon=0} = 0 \qquad (6.20)$$

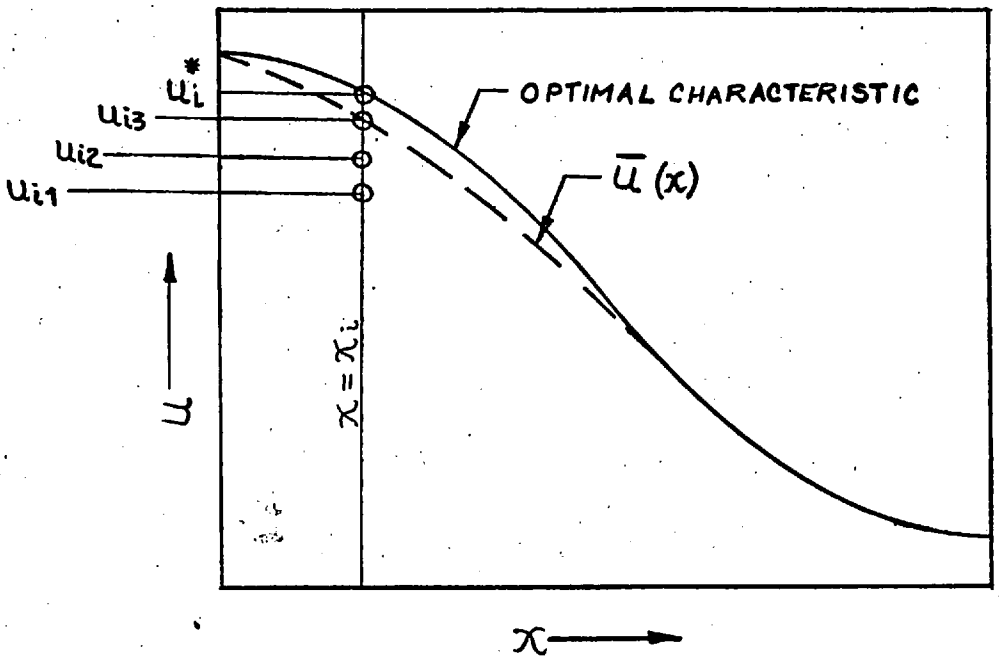


FIG. 6.4

VARIATION OF FEEDBACK CHARACTERISTIC

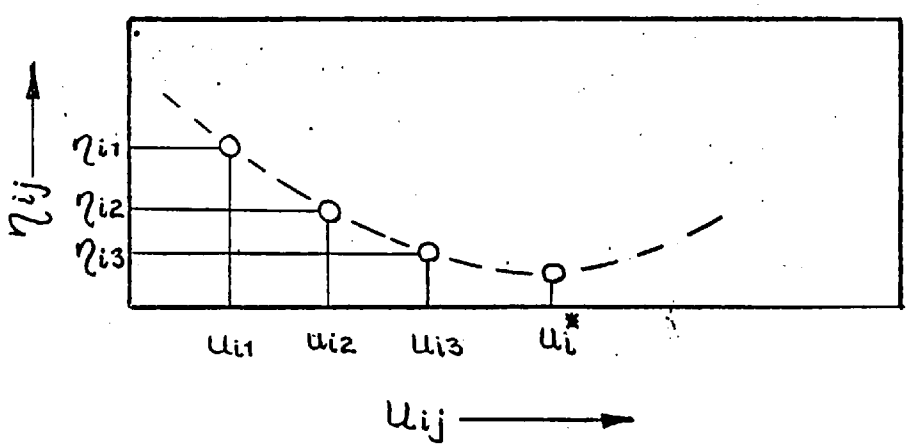


FIG. 6.5

η_{ij} VS. u_{ij}

We cannot continue this development along the lines of the Euler-Lagrange equation, since g is a non-analytic function of the parameters of (6.1). However, we may imagine the variable x to be divided into an arbitrarily large number of states i, i = 1, 2, ... N. If the optimal control for state i is u_i^*, then a condition equivalent to (6.20) is

$$\left.\frac{\partial g}{\partial u_i}\right|_{u_i=u_i^*} = 0, \qquad i = 1, \ldots\ldots N \qquad (6.21)$$

The partial differential of (6.21) is not directly obtainable, and its determination by perturbation of each u_i^* in turn would require considerable computational effort. Instead, we recall that, of two policies, the one which yields the lower value of η_{ij} will also yield the lower value of g (this fact is proved in appendix 1). If $\eta_{is} = \eta_i^*$ for state i, then (6.21) becomes

$$\left.\frac{\partial g}{\partial \eta_i}\right|_{\eta_i=\eta_i^*} = 0, \qquad i = 1, \ldots. N \qquad (6.22)$$

Thus the i^{th} row of the η matrix, η_{ij}, j = 1, ...r , is equivalent to a cross-section of the hill of g $(u_1, u_2 ..$ .... $u_N)$ taken at $u = u_i$. This relationship is illustrated in fig. 6.5. Given η_{ij}, j = 1, ...r , and assuming a continuous system, we may use second order methods to

predict the value of $u_i^* = u_{is}$. We then set up a new control set centred on u_i^* with a suitably reduced range, $|u_{ir} - u_{il}|$. This procedure is repeated for every process state i, i = 1, ... N. At this point a new transition matrix, P, and a new control cost matrix, B, are computed for the new set of control alternatives. The optimal decision matric, D^*, is re-computed, and the cycle is repeated until g is stationary. A simplified flow diagram of the algorithm is shown in fig. 6.6.

How finely x must actually be quantized in practice depends upon the problem in hand. The results of numerical examples, some of which will be presented in chapter 7, indicate that ten to twenty quantum levels is usually sufficient. Convergence of g to within one part in 10^4 typically requires six or eight iterations, using about 1.5 minutes on an IBM 7090 computer. As one would expect, computing time is not affected appreviably by non-linear dynamics, non-gaussian noise, or non-quadratic performance criteria.

6.5 Adaptive Control of Continuous State Processes

It is clear that the algorithm of section 6.4 cannot be applied directly to a process whose parameters are

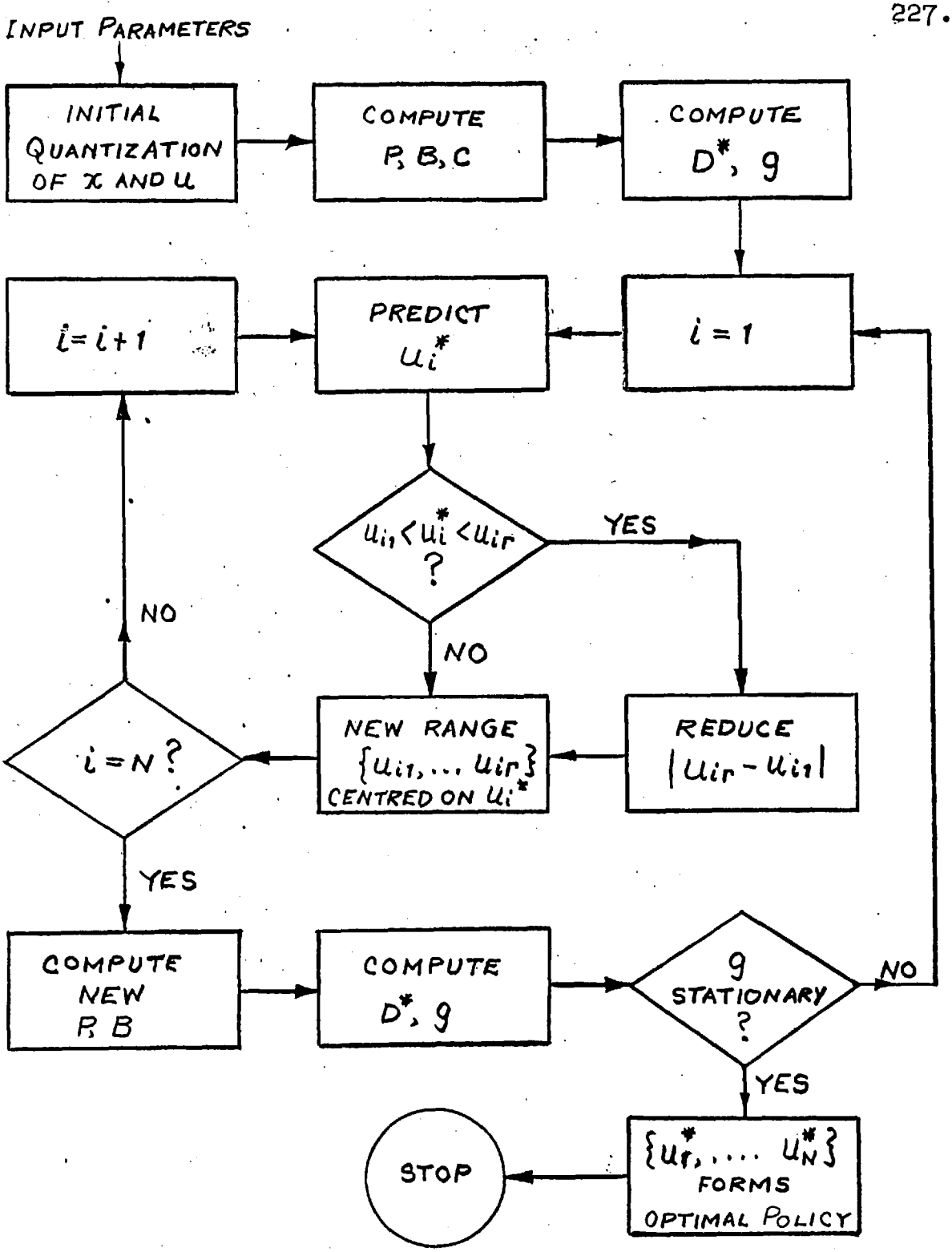


FIG. 6.6

OPTIMAL POLICY FOR CONTINUOUS STATE PROCESS

uncertain. At the beginning of the process, some a priori estimates must be made regarding the quantization of the output variable and the establishment of control ranges. The specification of these discrete states sets up a frame of reference within which we may use the techniques of sections 6.2 and 6.3 to search for an optimal control policy. The latter is optimal, however, only within the given frame of reference; it is probable that a different quantization would yield a lower cost optimal policy. Thus the quantization parameters themselves - the sort of grid we place over the $\{x,u\}$ space - must be regarded as estimates which need modification as the process continues.

We are thus led to a concept well known in adaptive systems, that of a hierarchy of adaptive loops. The innermost control loop is the adaptive controller using the strategy of section 6.2. Outside of this, as illustrated in fig. 6.7, is a slower-acting loop which adjusts the control sets $_i = \{u_{i1}, \ldots u_{iN}\}$ for each state i, so that the estimated optimal control signal, u_i^*, lies within the range of admissable controls for state i. This is done in the following fashion: if $\hat{s}(i) = 1$ and $\Omega_i < \Omega_c$ (Ω_c, chosen by the designer, is typically 0.50), the control range is extended downward by adjoining a new control $u_{io} < u_{i1}$ and deleting the control u_{ir} at the top of the

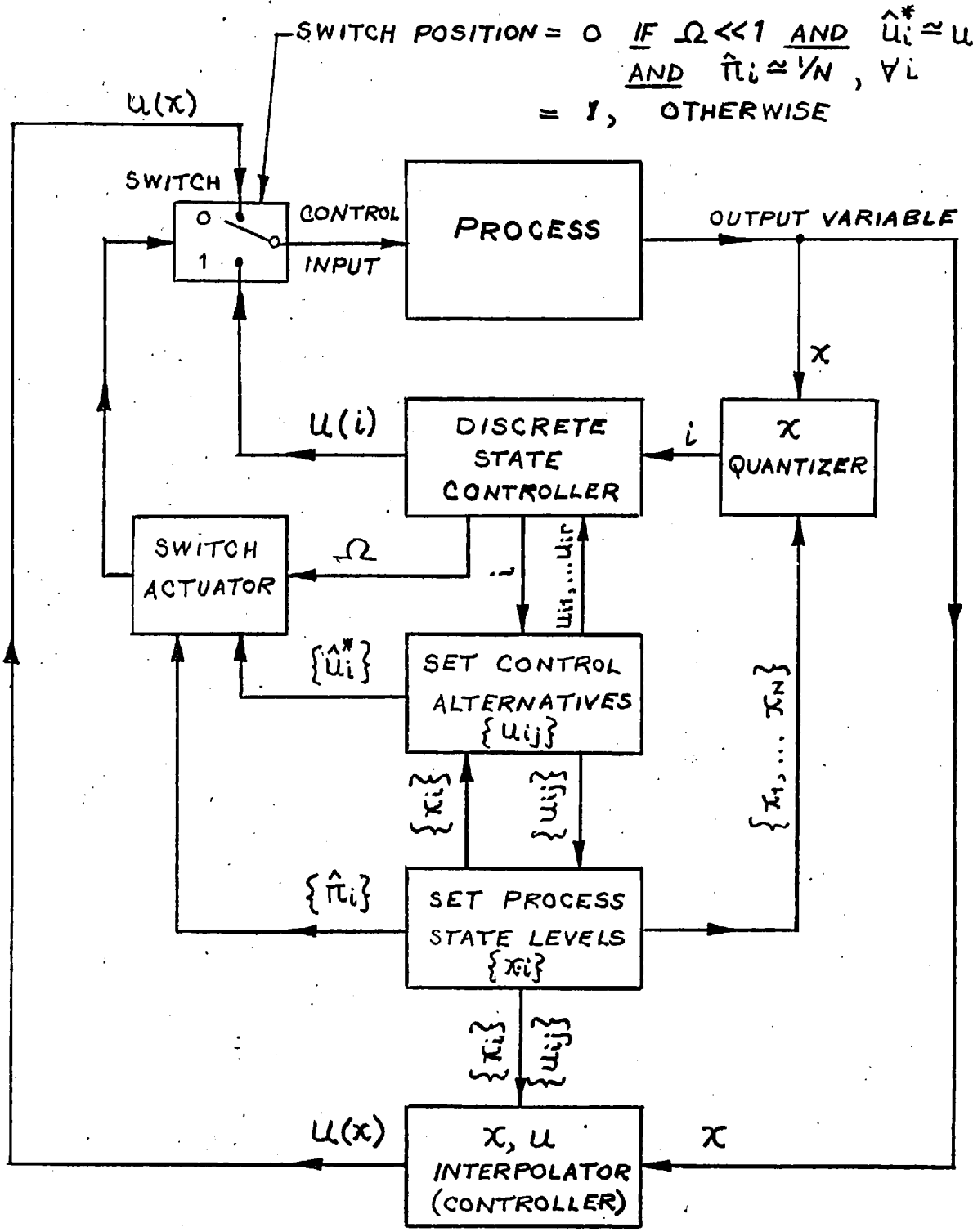
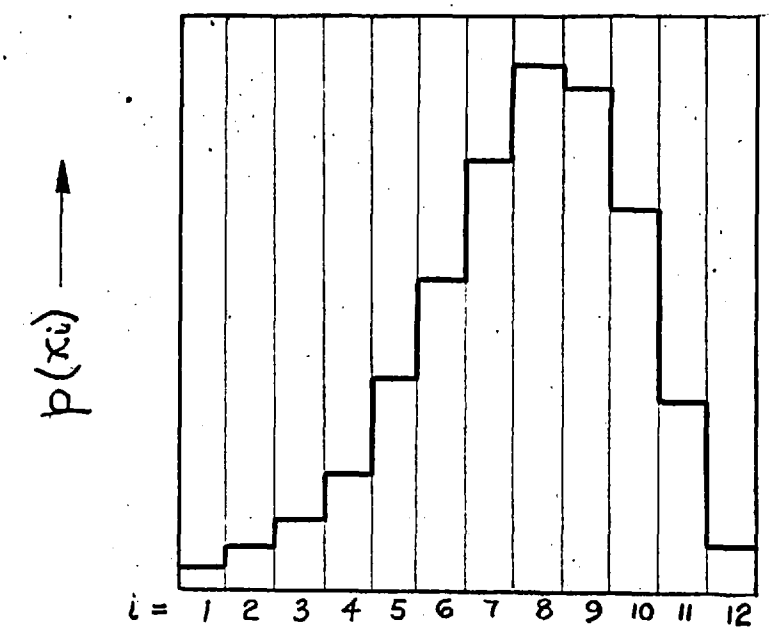


FIG. 6.7

HIERARCHY OF ADAPTIVE LOOPS

range. The new range is then re-labelled u_{i1}, ... u_{ir}. Matrices $\hat{P}$, B, and M are modified accordingly, the row of $\hat{P}$ corresponding to the decision state (i,1) being set by some a priori estimate. Similarly if the optimum control lies at the top of the range with probability greater than Ω_c, the whole range is shifted upwards one "notch". It is assumed that $u_{i1} < u_{i2} < \ldots\ldots < u_{ir}$, i.e. that U_i is an ordered set. In this way the control range for each process state "creeps" either upward or downward as the process continues, until one of the interior members of the set is the estimated optimum control input, or else a control constraint boundary is reached. A FORTRAN version of this adaptive scheme is contained in the program of appendix 7c.

Outside of the control quantizing loop is a further loop which modifies the quantization of the output variable, x. It has been noted that quantization of x should ideally be set so that each discrete state has an equal probability of occupancy. A suitable modification of quantum levels is easily arranged. From the elements of the N^{th} row of ψ^{-1} (see section 2.6) we can construct a probability density histogram, and from it a piecewise linear cumulative distribution function. We need then only shift the quantum limits so that each limit occurs at an equal interval of probability. An example is shown in fig. 6.8.



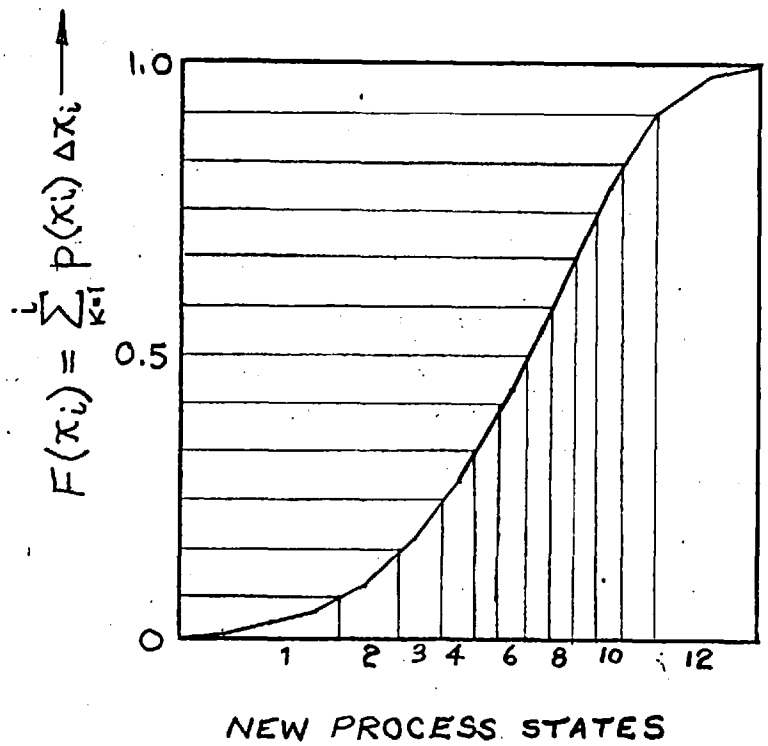


FIG. 6.8

REQUANTIZATION OF PROCESS STATES

Since the probabilities, π_i, of occupancy of each state i are functions of the control policy, it is pointless to re-quantize x before the control ranges are approximately correct. For this reason the x-quantizing loop must lie outside of the u-quantizing loop in fig. 6.7.

If the process is known to be stationary, the re-quantizing of x and u may be performed iteratively on-line, the control range being progressively decreased at each iteration. When the feedback policy becomes stationary, a continuous interpolated version of the feedback characteristic may be substituted for the stepped version, as in the off-line algorithm of section 6.4. This last adjustment, shown as the outermost loop of fig. 6.7, should remove the remaining quantization error.

6.6 Summary

When both the output and control variables are costed, and when the disturbance and control signals act simultaneously, the resulting decision process is a multi-stage one. It may usefully be regarded, though, as a set of single-stage processes with parameters interacting through, and modified by, the transformation matrix Ψ^{-1}. Dual control of each single-stage process, and hence of the

overall process, may be effected by the use of an extended version of the strategy presented in chapter 3. Updating of Ψ^{-1} at each stage is simplified through an application of the matrix inversion lemma.

If the dynamics and the cost function are continuous and differentiable, then a further refinement of the optimization technique is possible. It has been shown that the parameters η_{ik} yield gradient information which allows us to predict the control input, u_i^*, associated with each process state i, so that g is minimized. For the case in which the process dynamics are known, an extremum-seeking algorithm has been presented which optimizes alternately in discrete and continuous state space to produce the optimal non-linear feedback transducer characteristic for a continuous state, discrete time process. When the dynamics are uncertain, the same task may be performed on-line by a hierarchy of adaptive loops. In a stationary process, this method results eventually in the formulation of a continuous feedback characteristic, so that quantization errors are eliminated.

CHAPTER 7

SIMULATION RESULTS IN MULTI-STAGE MARKOV PROCESSES

7.1 Introduction

Having developed algorithms for the dual control of multi-stage Markov processes, we may now proceed to test them. A discrete time, continuous state model of a thermal process has been chosen for the test. A description of the process is given in section 7.2. In section 7.3 we consider the important question of a priori estimates. The results of the simulation of an on-line adaptive control sequence are presented in section 7.4.

7.2 The Problem: A Heat Treatment Process

It is desired to heat treat a series of long metal slabs at a temperature of $800^{\circ}A$. This cannot be done in a large oven because at about $800^{\circ}A$, an exothermic reaction begins in the metal, which causes the temperature to rise sharply. If the temperature exceeds $850^{\circ}A$, the slab affected must be melted down and re-rolled. The choice

lies between heating the whole slab in an oven with a temperature safely below 800°A, thereby avoiding the exothermic reaction and simultaneously reducing the market value of the product, or of heat treating small sections of the slab sequentially with a controlled heat unit. The advantage of this localized process is that if the temperature of the particular segment rises dangerously, cooling may be applied quickly to maintain it at a suitable level.

Localized heat treatment is preferable providing it is not too expensive. To estimate its economic feasibility, we shall consider it in more detail. The process is shown in fig. 7.1. The slab is moved longitudinally once per minute. During a given one minute interval section A is receiving heat treatment. Section B, which has emerged from an oven where it has been pre-heated to 750°, is heated by diffusion from section A. The temperature of the section under treatment is sampled at the beginning of each interval, and either heating or cooling is applied during the interval to stabilize the temperature. If the reaction in A is in its endothermic region, the process is stable; if it is exothermic, the runaway condition may spread backwards along the slab. A temperature of 850° or more in section A results in a process shutdown; the

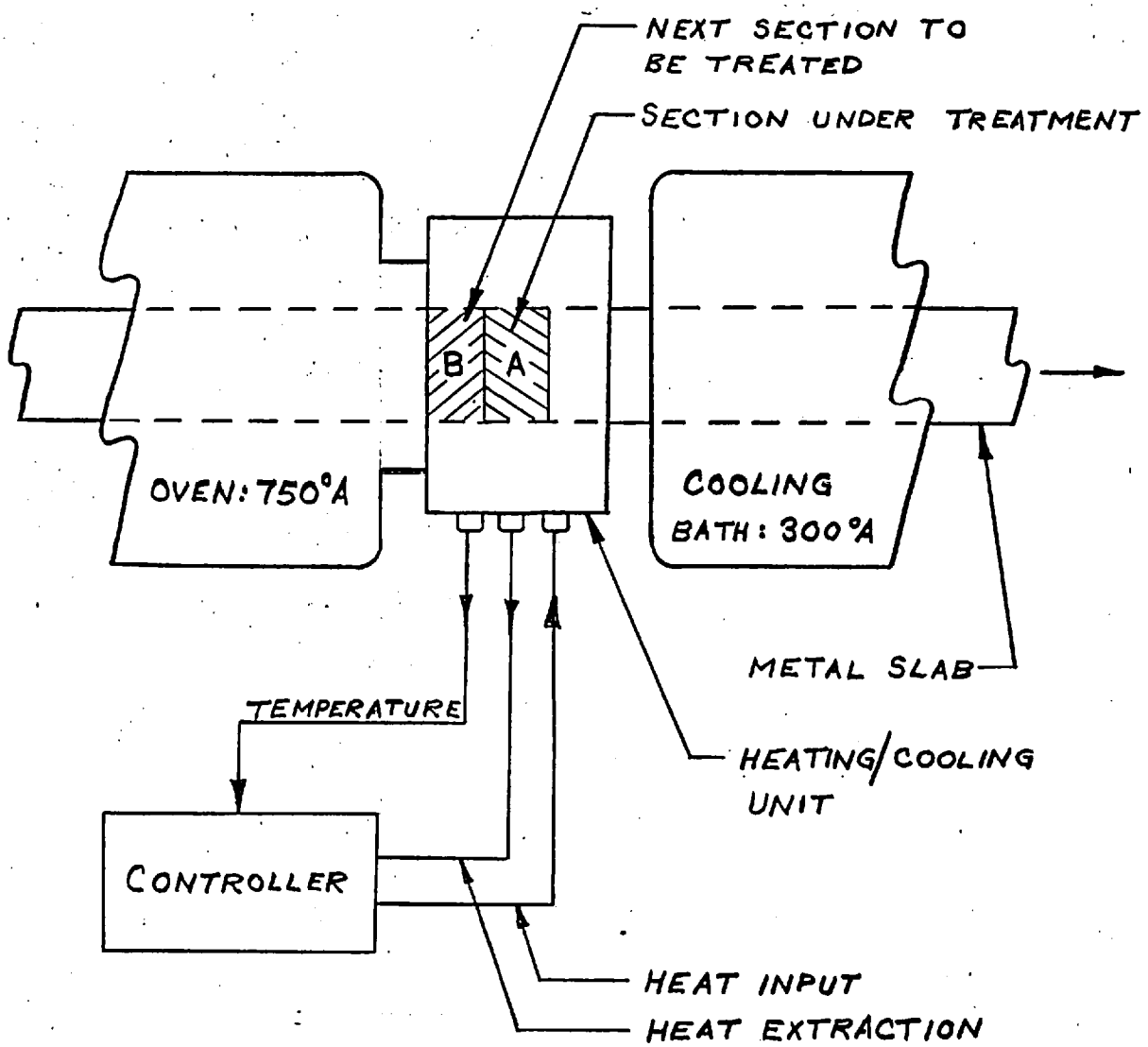


FIG. 7.1

HEAT TREATMENT PROCESS

length of slab in the pre-heat oven must then be removed and returned to an earlier stage of the manufacturing process.

Owing to small random variations in composition along the slab, the specific heat of the metal varies from one section to another. In addition, the quantity of heat actually transferred by the heat control unit for a given nominal heat setting has a random component. Finally, in the vicinity of 800°, the dynamics themselves are subject to a random effect related to the onset of the exothermic reaction. The parameters of all of these random effects are unknown.

The weight of slab under treatment in any interval is about 1000 kilograms. The specific heat of the metal is estimated to be 0.33. A maximum of 10,000 kilo-calories of heat may be transferred during an interval. The cost is a basic rate of 2d (two pence) per 1000 kilo-calories, plus an additional penalty cost if extremes (either positive or negative) of heat control are used. The reason for the latter cost is that the expected lifetime of the heating equipment is severely reduced by operation at the extremes. If u_n is the heating effort in kilocalories, then the control cost in pence during interval n is

$$b(n) = 0.002 \left| u_n \right| + 60 \left(\frac{u_n}{10^4}\right)^6 \qquad (7.1)$$

A further cost is associated with temperature. This may be regarded as a decrease in the value of the product caused by heat treatment at a non-optimum temperature. The cost is deemed to be quadratic in nature, centred upon the desired temperature of 800°, and is given (in pence) by

$$c(n) = 0.015\,[(x_n - 800)^2 + (x_{n+1} - 800)^2], \qquad x_{n+1} < 850 \tag{7.2}$$

where x_n = temperature in $^{\circ}$A at interval n. In addition, if $x_{n+1} \geq 850$, the cost of a shutdown is £12:0:0, which is added to c(n) in (7.2). If a shutdown should occur, the process is re-started with an initial temperature of 775°.

Management estimates that after amortization of the capital cost of the control hardware, but before payment of the operating cost, the increase in profit owing to heat treatment will be 2s. 6d. (two shillings, six pence) per 1000 kilograms. The installation is to be used for at least one year. The operating cost parameters ((7.1) and (7.2)) are known, but the dynamics of the process and the magnitude and nature of the disturbance are uncertain. Process estimation will have to be performed on-line, so that estimation costs, including any shutdowns which may occur, will be charged against profits.

Under these circumstances, is it profitable to install the automatic heat treating process?

7.3 A Tentative Solution: A Priori Estimates

Like many questions in engineering and economics, this one is unanswerable: insufficient data is available. However, by computing the optimal feedback policy for a set of assumed dynamics together with the known cost function, we may obtain some idea of the cost involved. As a starting point, therefore, we shall assume the following linear dynamic equation:

$$x_{n+1} = \phi x_n + \beta u_n + \xi_n \qquad (7.3)$$

where

x_n = temperature ($^{\circ}$A) at interval n

u_n = heat input (k cal.) at interval n

ξ_n = disturbance ($^{\circ}$A) at interval n

$\phi = 1.00$

$\beta = 0.0030$

The factor $\phi = 1.00$ arises because the process is assumed to be just entering its exothermic region in the vicinity of the operating point. The control multiplier, β, is the inverse of the product of specific heat and mass. Since the nature of the disturbance is unknown, ξ_n must be guessed. We have taken it to be additive gaussian noise with mean zero and standard deviation 10°A.

TABLE 7.1

TWENTY-STATE QUANTIZATION

State No.	Temperature Limits ^{o}A	Mid Point ^{o}A
1	$\leq$ 750	750.00
2	750.0-765.0	757.50
3	765.0-775.0	770.00
4	775.0-780.0	775.50
5	780.0-785.0	782.50
6	785.0-790.0	787.50
7	790.0-792.5	791.25
8	792.5-795.0	793.75
9	795.0-797.5	796.25
10	797.5-800.0	798.75
11	800.0-802.5	801.25
12	802.5-805.0	803.75
13	805.0-807.5	806.25
14	807.5-810.0	808.75
15	810.0-815.0	812.50
16	815.0-820.0	817.50
17	820.0-825.0	822.50
18	825.0.835.0	830.00
19	835.0-850.0	842.50
20	$\geq$ 850.0	850.00

The temperature range of interest is 750° to 850°, which may be quantized as shown in table 7.1. States 1 and 20 are temperature limits, while the remaining states

represent ranges. The non-uniform quantization reflects our hope that the temperature will remain near 800° most of the time with an optimal control.

A 20 x 20 transition cost matrix, C, is now computed from (7.2).

$$c_{ij} = 0.015\ [(x_i - 800)^2 + (x_j - 800^2] + \delta$$

where x_i = mid-point temperature of state i

$\delta = 2880, \quad j = 20$

$\delta = 0, \quad j \neq 20$

For each state we now choose a discrete set of five alternative control inputs ($\Gamma = 5$). Once these are chosen, the 20 x 5 control cost matrix, B, is computed using (7.1). We may then use the algorithm of section 6.4 (fig. 6.6) to determine the optimal feedback characteristic.

In fact a number of such characteristics were computed for different values of $\emptyset$, β, and $\wp_n$. The first one computed used $\emptyset = 1.00$, $\beta = 1/300$, and $\wp_n$ as in (7.3). It is perhaps worthwhile indicating how the initial control ranges were chosen. First, the extreme policy in which control is not costed was computed. Ignoring the disturbance, the control signal necessary is that which returns the temperature from its present level to 800° in one interval:

$$u_n = \frac{1}{\beta} (800 - \emptyset x_n)$$

$$u_n = 300 \ (800 - x_n) \text{ kilocalories} \qquad (7.4)$$

Second, the optimal linear control was computed for the quadratic cost function

$$L(x,u) = q(x_n - 800)^2 + r\, u_n^2 \qquad (7.5)$$

where

$q = 0.030$, equivalent to (7.2)

$r = 10^{-6}$, yielding control costs approximately equal to those of (7.1) in the region of $u_n = 2000$.

The optimal control policy, computed from the Riccati difference equation, is

$$u_n = 130 \ (800 - x_n) \qquad (7.6)$$

It was guessed that the optimal feedback characteristic would be somewhat similar to (7.6) for $x < 800$, and tend towards (7.4) for $x > 800$ because of the high cost of entering state 20. Accordingly the initial control ranges were set as shown in fig. 7.2. The final characteristic, obtained after ten iterations of the algorithm, is also shown in fig. 7.2. The expected cost per transition, g,

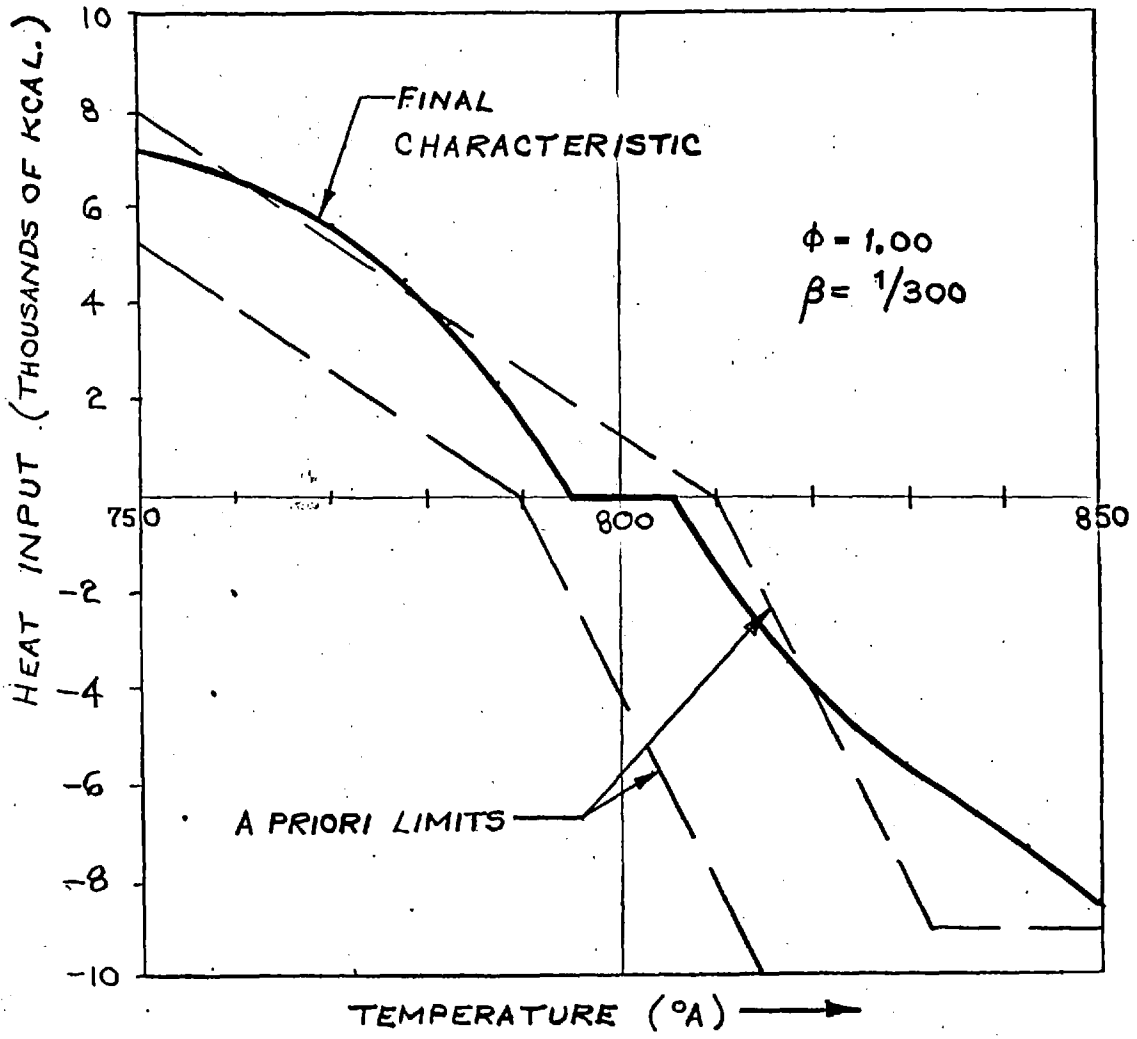


FIG. 7.2

A PRIORI OPTIMAL

FEEDBACK CHARACTERISTIC

decreased from 8.935 at the first iteration to 6.709 at the tenth (stopping criterion: change in g is less than 0.01% in last iteration).

The optimal characteristic for the dynamics as given by (7.3) was computed with the curve of fig. 7.2 as a starting condition. The result, shown in fig. 7.3, is very similar, since the only change is a 10% reduction in β. The shape of this curve is worth some comment. Its most striking feature is the discontinuity in the vicinity of $x = 800^{\circ}$. This results from the combination of the absolute value of control cost together with the quadratic function of process state. It is well known from the theory of continuous systems that cost functions of this nature lead to a relay type control system. The relay action, switching from u_{min} to u_{max}, is modified in fig. 7.3 because of the discretized time intervals, so that control operates on one of two curves separated by a dead band. Examining the characteristic, we can surmise that its flattening for low values of x is caused by the sixth power law in the control cost. For values of x in the range 780°-820°, the curve is nearly symmetrical about 800°. For higher values of x, we observe that the presence of the step function (shutdown cost) at $x = 850^{\circ}$ begins to be important; flattening of the characteristic near 850° does

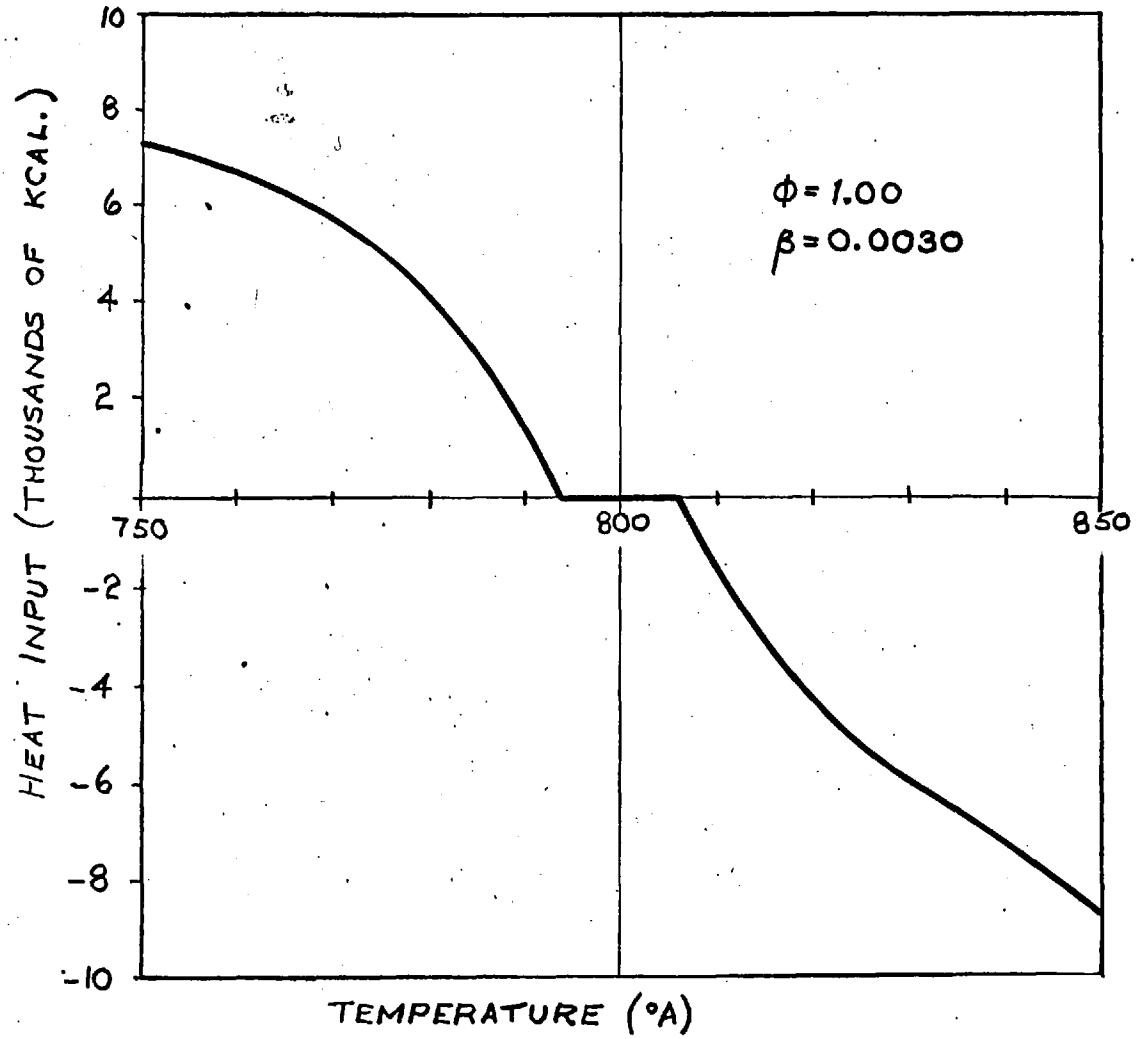


FIG. 7.3

MODIFIED A PRIORI

FEEDBACK CHARACTERISTIC

not occur since, despite the high cost of large control effort, such effort is necessary to avoid an even more expensive shutdown.

The cost associated with the curve of fig. 7.3 is 7.078; i.e. the expected cost of operation is about seven pence per interval. It appears that even if our cost assumption is in error by a factor of three or four, this process is still profitable. Our reply to the question at the end of section 7.2 is a qualified affirmative.

To implement the adaptive control strategy, it now remains only to specify the a priori quantization. We shall consider first the specification of control range for on-line adaptation. If the discrete control alternatives, u_{ik}, are spaced too widely, the probability of u_{ik}, $k \neq \hat{s}$, being chosen is extremely small, and convergence is slow. If the spacing is too close, the controller will have difficulty in deciding which is the lowest cost control. Consequently the decision to shift the control range in a favourable direction will take an inordinately large number of intervals, and convergence will again be slow. We see from inspection of the adaptive strategy of section 6.2 that a measure of the desired spacing is given by the parameter σ_{oik}, the square root of the cost variance associated with the choice of u_{ik} when the process state is i. Using the

model of equation (7.3), we search in u-space monitoring the parameters ($\eta_{ik} - \eta_{is}$) and σ_{oik} associated with a control u_{ik} in the vicinity of the optimal control u_i^*. When $\eta_{ik} - \eta_{is} = \sigma_{oik}$, the cost of using control u_i^* is one standard deviation (measured at decision state (i,k)) below the cost of using u_{ik}. Application of this technique yielded control spacings of 1000-2000 kilocalories for most process states. It was decided to set the a priori control range as follows (control is given in kilocalories)

$$u_{i1} = u_i^* - 1000$$

$$u_{i2} = u_i^* - 500$$

$$u_{i3} = u_i^*$$

$$u_{i4} = u_i^* + 500$$

$$u_{i5} = u_i^* + 1000, \qquad i = 1, \ldots\ldots N$$

It was felt that fairly coarse quantization of the temperature range might help accelerate initial convergence; an eleven-state system was chosen. The quantization of ℓ, together with the a priori estimates of optimal controls, u_i^*, is given in table 7.2.

TABLE 7.2

A PRIORI QUANTIZATION

Process State	Mid-Point Temperature $^{o}_{A}$	A Priori Estimate of u_i^{M} Kilocalories
1	750.00	7400
2	762.50	6500
3	780.00	4200
4	788.75	1900
5	795.00	0
6	800.00	0
7	805.00	0
8	811.25	- 1900
9	820.00	- 4200
10	837.50	- 6900
11	850.00	- 9000

7.4 Experimental Results: Simulated Adaptive Control

The dynamic equation (unknown to the controller) used to simulate the heat treatment process was

$$x_{n+1} = x_n \Big\{ 1.005 + 0.015 \tanh [0.1(x_n - 803.466)] + 0.0002\, q \cdot \Delta \cdot \rho_1 + 0.005\, \rho_2 [1 + (|\Delta|)^{1/2}]^{-1} \Big\} + u_n \Big\{ 0.000333 + 0.0005\, \rho_3 \Big\} + \rho_4 \qquad (7.7)$$

where

x_n = temperature (^oA) of section under treatment at the beginning of interval n.

u_n = control signal (kilocalories) applied during interval n.

$\Delta = x_n - 800$ (^oA)

$q = 1, \quad \Delta > 0$

$q = 0, \quad \Delta \leq 0$

$\mathcal{S}_1, \ldots \mathcal{S}_4$ = independent random samples from a normal distribution with mean zero, variance unity.

Equation (7.7) has the form

$$x_{n+1} = [\phi(x)]\, x_n + \beta\, u_n + \mathcal{S} \qquad (7.8)$$

We see from (7.7) that the non-linear multiplier $\phi(x)$ takes on a value of 0.99 for low values of x, so that the system is stable in this region. At $x = 800$, $\phi(x) = 1$; as x increases further (i.e. as the process enters its exothermic region) $\phi(x)$ increases towards a maximum value of 1.02. $\phi(x)$ is also disturbed by noise, one component of which is especially marked in the vicinity of $x = 800$, the other increasing linearly for $x > 800$; both of these are effects of the transition from the endothermic to the exothermic reaction. Associated with the control signal is a disturbance whose magnitude is proportional to the heating effort. Finally, a purely additive noise component ($\mathcal{S}_4$) is present.

An adaptive control run of 1500 intervals (equivalent to 25 hours operation) was simulated on an IBM 7090 computer. The a priori parameters of (7.3) were used, with a convergence factor, γ, of 10, and an initial temperature of 775°A. The results of each transition were recorded; at every hundredth transition the present estimate of the optimal transducer characteristic was printed, together with the mean and standard deviation temperature and the cost, averaged over the last one hundred transitions. The value of $\Omega(n)$, computed from (6.5) with $a_i = \pi_i(n)$, was also recorded. The FORTRAN program used is given in appendix 6c, and the principal results are given in table 7.3 and fig. 7.4. Program running time, including compilation of the main program, was 6.3 minutes.

To interpret the results we must consider the nature of the dynamics. The desired operating temperature, 800°, is a point of unstable equilibrium; the higher the temperature above 800°, the more it is likely to rise in the next interval, and conversely, the lower it is below 800°, the more it is likely to fall. Control of this system is somewhat analogous to the problem of maintaining an inverted pendulum in an upright position.

For temperatures much below 800°, the a priori estimate gives too small a heating effort to operate the actual

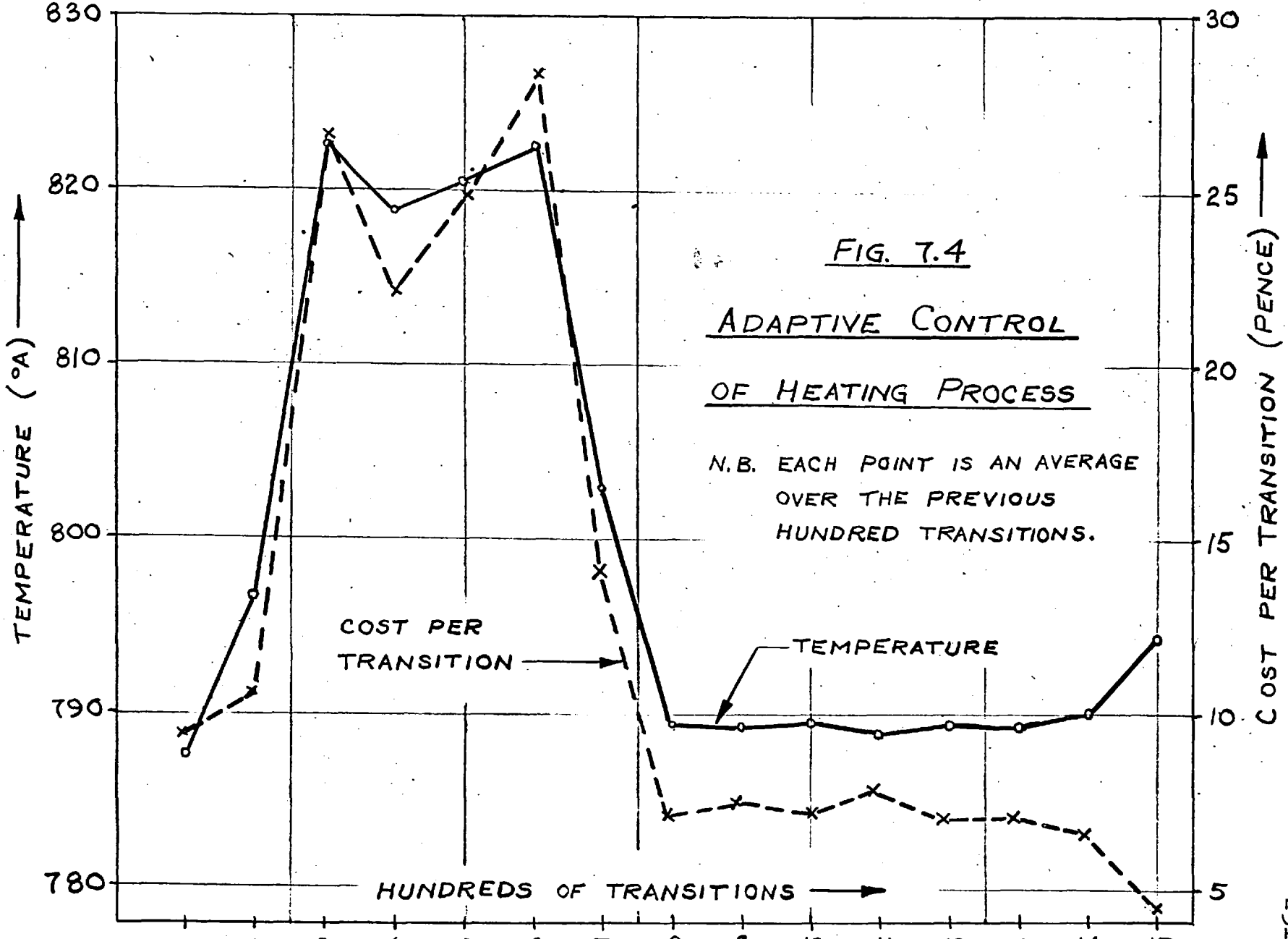

830
820
810
800
790
780
TEMPERATURE (°A)
30
25
20
15
10
5
COST PER TRANSITION (PENCE)
FIG. 7.4
ADAPTIVE CONTROL
OF HEATING PROCESS
N.B. EACH POINT IS AN AVERAGE OVER THE PREVIOUS HUNDRED TRANSITIONS.
COST PER TRANSITION
TEMPERATURE
HUNDREDS OF TRANSITIONS
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

TABLE 7.3

ADAPTIVE CONTROL OF NON-LINEAR HEAT TREATMENT PROCESS

Interval No.	Results Averaged Over Previous 100 Intervals			Mean Overall Cost/ Interval	Uncertainty	Remarks
	Mean Temp ^{o}A	Std. Dev'n ^{o}A	Cost/ Interval Pence	Pence	Ω	
100	787.50	3.60	9.32	9.32	0.3348	Temperature below 800^{o}
200	796.62	13.10	10.63	9.97	0.2263	Temperature exceeds 800^{o} at interval 181
300	822.44	6.71	26.51	15.49	0.0081	Temperature in the vicinity of 820^{o}
400	818.94	7.72	22.04	17.12	0.0131	
500	820.55	6.82	24.93	18.68	0.0891	
600	822.49	6.67	28.32	20.29	0.0763	
700	802.83	16.00	14.15	19.41	0.0047	Temperature forced below 800^{o} at interval 647
800	789.54	2.80	7.24	17.89	0.0026	Temperature below 800^{o}
900	789.40	2.84	7.69	16.76	0.0019	
1000	789.70	2.78	7.24	15.81	0.0022	
1100	788.81	3.09	8.04	15.10	0.0012	
1200	789.55	2.92	7.33	14.45	0.0011	
1300	789.32	2.81	7.43	13.91	0.0009	
1400	790.11	2.87	6.97	13.42	0.0008	
1500	794.31	3.57	4.37	12.81	0.0042	Stable operation around 800^{o} observed

process satisfactorily. The process record shows that adaptation began to occur, however, and the temperature exceeded 800° for the first time at interval 181. Almost immediately it climbed to about 820°, since the estimated cooling effort to combat the exothermic reaction was too small near 800°. Between interval 181 ($x = 804.6^{\circ}$), and interval 646 ($x = 802.8^{\circ}$) the temperature remained above 800°, reaching an uncomfortable high of 844.3° at interval 523. No plant shutdown occurred, though, and gradually the estimated feedback transducer characteristic altered to take account of the peculiar dynamics. At interval 647 the characteristic had changed sufficiently to force the temperature below 800° again. A further period of relatively low temperature operation and adaptation followed between intervals 647 and 1465. In the last 35 intervals of operation the mean temperature was 795.6°. Several temperature excursions above 800° occurred in the last hundred intervals, and each was controlled successfully. By interval 1500, therefore, operation in the vicinity of 800° was stable. Mean cost per transition dropped, by no means monotonically, by more than 50% between the first and last hundred intervals.

The a priori optimal characteristic and the estimated optimal characteristic at $n = 1500$ are shown in fig. 7.5

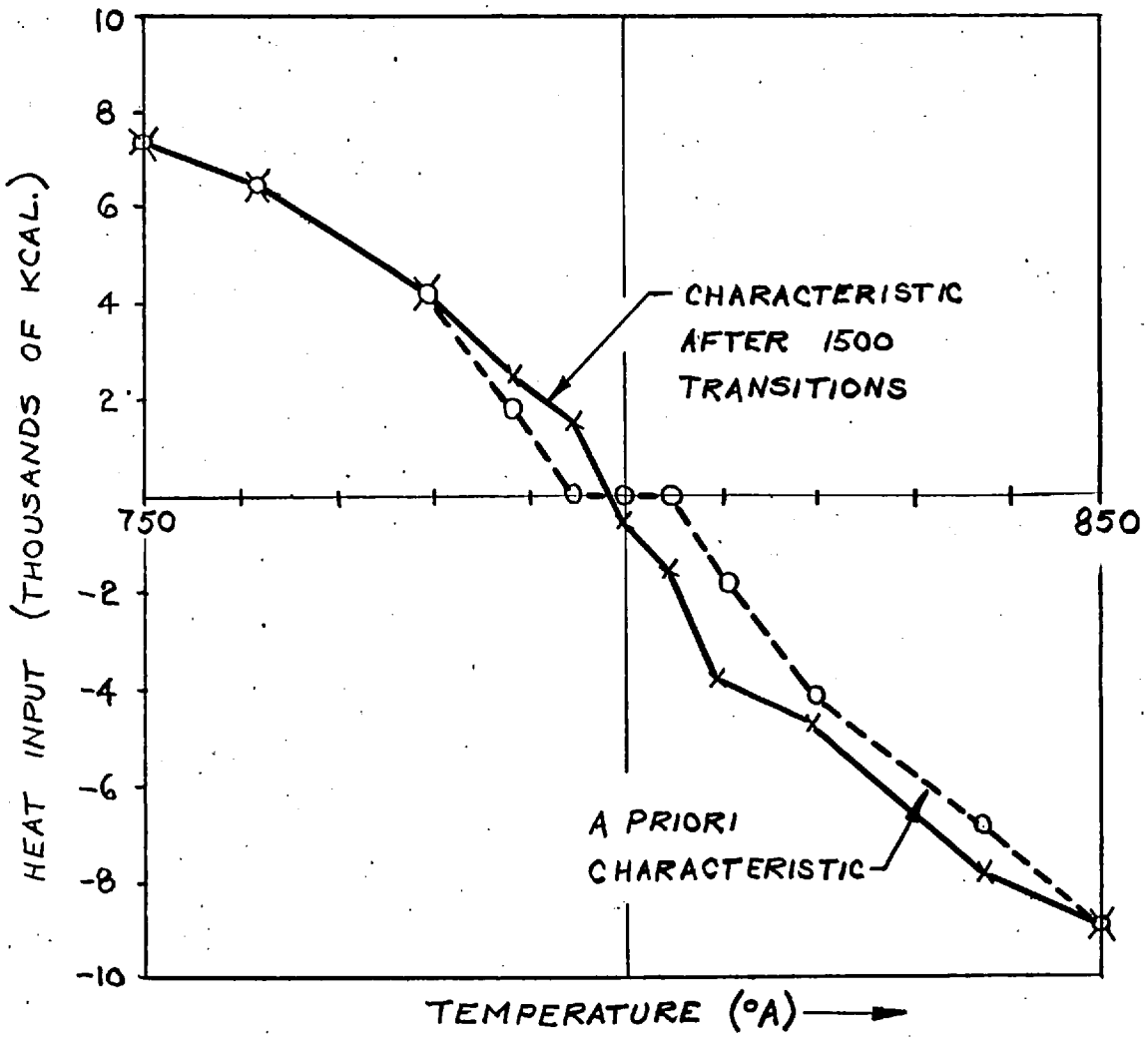


FIG. 7.5

ADAPTATION OF

FEEDBACK CHARACTERISTIC

(for diagrammatic clarity they are shown as piecewise linear; as used they were piecewise constant, i.e. stepped). To determine the theoretically optimal characteristic, the dynamics of (7.7) were used in the algorithm of section 6.4. The results show that with optimal control the temperature will remain between 792.5° and 807.5° with probability 0.997. It follows that the characteristic in the immediate vicinity of 800° is of most importance. Fig. 7. shows an enlarged version of fig. 7.5, together with the optimal characteristic, for the region $780^{\circ} \leq x \leq 820^{\circ}$. We observe that the 1500 interval estimate is considerably closer to the optimal curve than is the a priori estimate.

The theoretical minimum cost per transition is a surprisingly low 2.023 pence. This fact, together with a study of figs. 7.4 and 7.6, suggests that the controller, while operating satisfactorily, has not completely converged by interval 1500. The process seems to have reached the point where some higher level of adaptation, as outlined in section 6.5, may take place. Both x and u must be more finely quantized in the vicinity of $x = 800^{\circ}$. No provision was written into the program for this purpose, although its implementation should be straightforward.

From the maximum profit figure given in section 7.2, we can compute that the profit during the 25 hours of

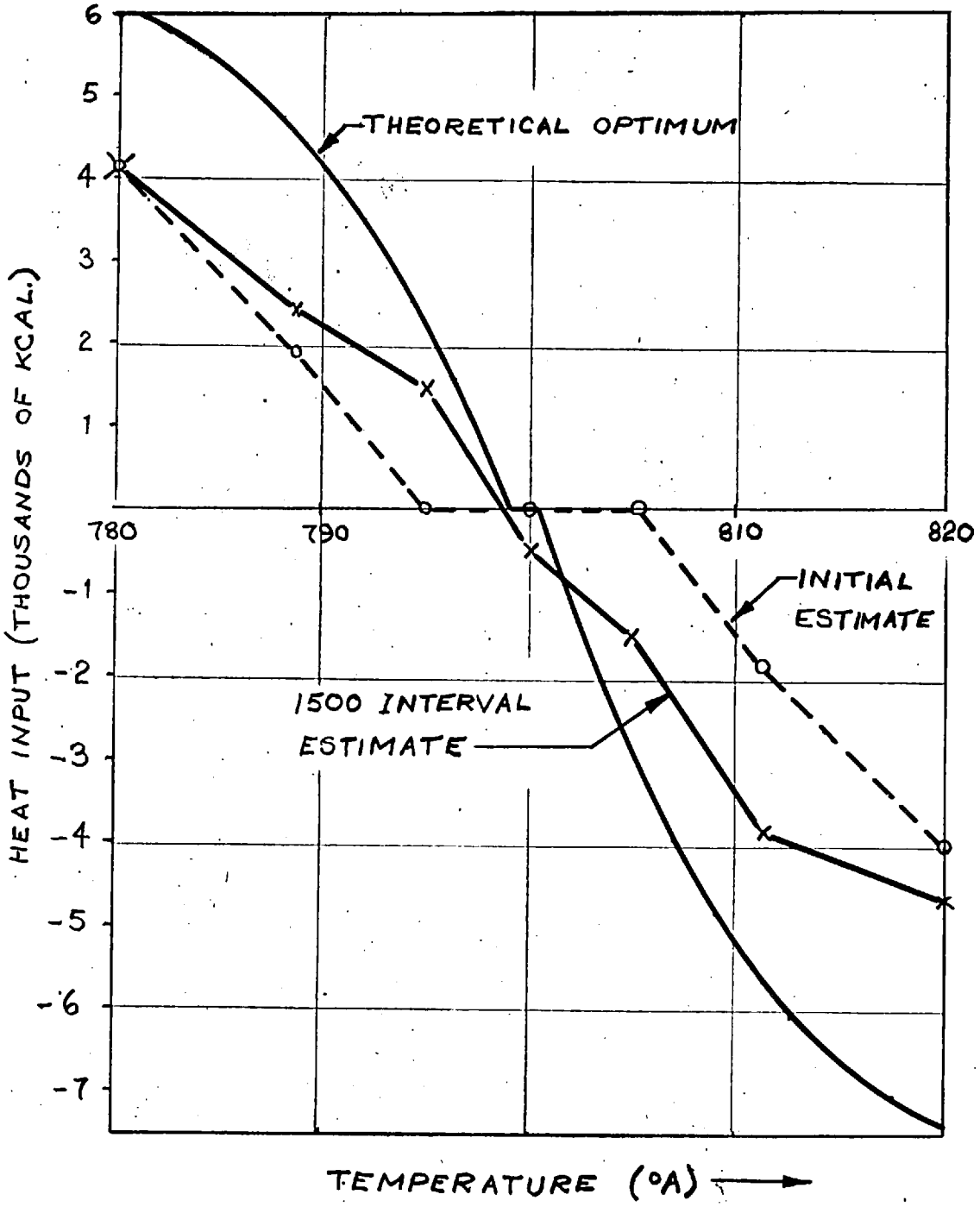


FIG. 7.6

FEEDBACK CHARACTERISTICS

NEAR 800 °A

simulated operation was £107:8:4. Had complete knowledge of dynamics been available initially the profit (with a cost of 2.023 d. per interval) would have been £174:17:0. Once again, this difference emphasizes the value of accurate initial estimates in any dual control system.

7.5 Summary

We have examined the control of a non-linear heat treatment process which is disturbed by multiplicative noise. The dynamics in the example have deliberately been made somewhat "nasty" in order to test the adaptive capabilities of the dual control strategy presented in chapter 6. The desired temperature of $800^{\circ}A$ is a point of unstable equilibrium. If the a priori feedback characteristic is used, the resulting process states divide into two sets (those with $x < 800$, and those with $x > 800$) which are almost non-communicating. With high probability the temperature is either about 789° or about 820°, depending upon the initial condition. In terms of Markov chain theory, the basic transition matrix, P_b, has amongst its eigenvalues one unit eigenvalue and two others very close to unity. The object of the adaptive strategy is to rearrange the feedback policy so that the latter two

eigenvalues are pushed towards the origin, and the probability of occupancy of states close to $x = 800^{\circ}$ becomes large.

The results presented indicate that adaptation was successful. Starting with a nominal estimate of the dynamics the controller was able, through successive adjustments of the feedback policy, to overcome the effects of non-linear dynamics and multiplicative noise, eventually transforming the region near $x = 800^{\circ}$ from an unstable operating region to a stable one.

CHAPTER 8

POINTS OF DEPARTURE

8.1 Introduction

Both theoretical and computational results pertaining to the dual control of long duration Markov processes have been presented in this thesis. A summary of these results may be found in chapter 1, and need not be repeated here. Instead it seems more profitable to discuss them in the light of the many questions they raise, since each question is a tentative signpost to new research. Some of the questions suggest fairly straightforward extensions or refinements of the work already done; others require detailed investigation, with the present results as a starting point; one or two are speculative in nature, and suggest no clear line of attack.

8.2 The Single-Stage Dual Strategy as a Multi-Modal Hill Climber

It has been pointed out in chapter 4 that dual control of a repetitive batch process is equivalent to the descent

of a noisy multimodal hill. The single-stage dual control strategy may easily be applied as an adaptive hill climber. In many cases the cost matrix, C, is not available, but this is no real hindrance. Mean and variance estimates of cost can still be determined from past results, and the uncertainty, Ω, can be computed. Moreover adaptive quantization is readily implemented by quantizing most finely those regions of the control space with largest probability of yielding a minimum cost. If the hill is known to be unimodal, the algorithm of section 4.7 is particularly useful.

One note of warning should be given concerning these and other hill-climbing techniques designed for on-line use. They are usually decidedly non-optimum as off-line hill climbers, when the additional search constraint is not that of equation (3.9), but instead specifies that computing time be minimized. In such a case resort may be made to various statistical sampling techniques[41,42] which are specifically designed to minimize the number of trials required for a decision.

8.3 Further Investigation of the Multi-Stage Dual Strategy

The results of chapter 7 demonstrate the effectiveness and feasibility of the multi-stage dual strategy presented

in chapter 6. It is clear, though, that much work remains to be done in this area. The relative merits of fine and coarse a priori quantization and the best choice of convergence factor should be considered. In addition it is worth investigating the effectiveness of the higher levels of adaptation shown in the heirarchical scheme of fig. 6.7.

While the strategy of chapter 6 yields convergence to the optimal policy, the strategy itself is not necessarily optimal. We face here the more fundamental question, "Can a better approximation than Ω (equation (6.5)) be computed for the probability of error, ω, in the multi-stage problem?"

8.4 The Economics of Generator Ordering

Some possible improvements to the Markovian model of generator ordering were indicated at the end of chapter 5. Assuming that their implementation led to a more useful model of the ordering process, we would be faced immediately with the larger question, "How do we specify an optimal ordering policy for two or more interconnected generating stations?" Treatment of several stations as an overall entity - the monolithic controller approach - yields a problem whose dimensionality is too great for practical computation. Some sort of decomposition method is therefore

desired. If each station is regarded as a semi-autonomous entity, as indeed it is in practice, the interaction variables between stations are then the (time varying) buying and , selling prices of power. A brief study of the interconnection problem, not included in this thesis, indicates that the decomposition of the overall system into a number of interacting systems may yield multiple cost minima. Further work in this direction might be useful.

8.5 Automaton Model Relevance

In presenting the algorithm of section 6.4 which determines a continuous optimal feedback characteristic, we have made the assumption that the continuous state process model could be approximated as closely as desired by a sufficiently finely quantized atucmaton model. What are the necessary and sufficient conditions which ensure this? It is thought that the conditions of continuity and differentiability given in section 6.4 are sufficient. In chapter 7, however, convergence of the algorithm was observed despite a discontinuity in the cost function at one point (step cost of shutdown). There is also the question of uniqueness: while the algorithm converges to a locally

optimal characteristic, it has not been proved to result in an absolute minimum of expected cost. On this point we have two comments: first, computing experience to date suggests that the algorithm does converge to a unique solution; second, by starting the algorithm with a very wide control range, we can approximate a global search so that, if multiple minima do exist, spurious solutions can be avoided.

8.6 Non-Stationary Processes

If the Markov process is non-stationary, the controller must filter past information so that older estimates have less weight than newer ones. The form of filter depends upon the model assumed for the non-stationarity. One approach is to estimate a regression function for the row of P corresponding to the decision state most frequently occupied. It should then be possible to relate the parameters of the regression function to those of the information filter.

8.7 Processes with Uncertainty in State Measurement

Dual control of a discrete state Markov process is made

considerably more difficult if noise is present in the measurement of state. If the parameters of both the process and the measurement noise are unknown, then the problem is insoluble since the two effects cannot be separated. Assuming that the noise parameters are known, though, it is possible in principle to answer the question, "Given a particular record of measurements, what transition matrix, $\hat{P}$, maximizes the likelihood of producing this record?" The work of Åström[38] is suggested as a starting point for this problem. Larson and Peschon[40] have also considered the effects of measurement noise on a discrete state multi-stage process, as a generalization of Kalman filtering for non-linear processes.

8.8 Finite Duration Processes

The multi-stage dual strategy can be adapted to the optimization of a stochastic process operating repetitively over a finite number of stages. In this case the operative relationship is the recursive equation (2.9); the parameters η_{ij} are time-dependent, so that the optimal policy generates a time-varying feedback characteristic.

There is one type of finite duration problem, however, which can perhaps be treated as a steady state Markovian

decision problem; that is the minimum time problem. By removing control cost, and costing process states according to their "distance" from the desired final state, we might obtain an approximately optimal feedback characteristic for minimum time operation, despite non-linearities and noise. It would be approximate because the point in state space with minimum norm is not necessarily the point from which the origin can be reached in minimum time. It seems possible that this difficulty might be overcome by adjustment of the transition cost matrix after each characteristic is computed. We might, for instance, set c_{ij} at iteration $n + 1$ equal to the value of v_j computed at interval n. Whether or not such a scheme would converge suitably is unknown.

8.9 Higher Order Processes

In this thesis the multi-stage dual strategy has been applied to a non-linear first order system. By a suitable extension of the state space it can sometimes be applied to systems with several variables; for example, three variables (time, number of sets running, present load) were handled in the power ordering problem. Even so, the discrete state approach as treated in this thesis is severely limited in

dimensionality. A second order system - modelled, say, as a one hundred state automaton - presents a large but feasible problem; a third order system would have to be quite crudely quantized to be feasible. In the next section we shall consider a simplification which might alleviate "the curse of dimensionality". Before doing so, we shall examine the automaton model of a second order system. The system of fig. 8.1 is described by the equations

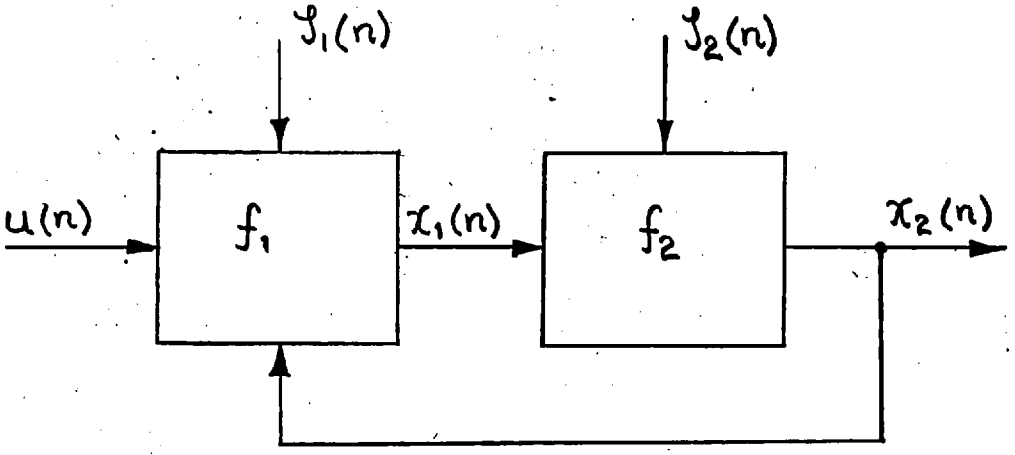
$$\begin{aligned} x_1(n+1) &= f_1(x_1(n),\ x_2(n),\ u(n),\ \vartheta_1(n)) \\ x_2(n+1) &= f_2(x_1(n),\ x_2(n),\ \vartheta_2(n)) \end{aligned} \tag{8.1}$$

where

$x_1(n)$ = value of variable x_1 at interval n.

$x_2(n)$ = value of variable x_2 at interval n.

Two points should be noted. First, we have assumed that both x_1 and x_2 can be measured. If only the output sequence $x_2(1), \ldots. x_2(n)$ is available and the system is non-linear, then the sequence cannot be described in terms of a Markov process. Second, the feedback control is a general function of x_1 and x_2. It is common in process control to feed back a separable function of the state variables, i.e. $g_1(x_1) + g_2(x_2)$ rather than $g(x_1,x_2)$. Certainly this simpler practice is to be preferred if it



$$x_1(n+1) = f_1\left(x_1(n), x_2(n), u(n), \vartheta_1(n)\right)$$

$$x_2(n+1) = f_2\left(x_1(n), x_2(n), \vartheta_2(n)\right)$$

FIG. 8.1

SECOND ORDER STOCHASTIC PROCESS

does not seriously degrade system performance. The existence of a method of computing the general optimum feedback function for a noisy non-linear system allows us to compare the performance of separable feedback functions with that of the optimum one, and so estimate the suitability of a suboptimum controller for a given process.

8.10 Theoretical Implications of the Discrete Formulation

The problem of dimensionality may be partially overcome by the specification (usually heuristically) of "sub-goals" for the controller[32]. This is equivalent to the replacement of a multi-stage optimization scheme with one which optimizes over only the next one or two stages. The advantage of such an approach is that only a limited portion of the state space need be considered in any one computation; consequently it is feasible to attempt adaptation in higher order non-linear systems. The disadvantage is that single-stage optimization may yield results which are poor from the viewpoint of overall optimization. It is hoped, of course, that a suitable choice of sub-goal can avoid such a situation.

Suppose that process dynamics are known. In such a case, single-stage optimization implies that if the present process state is i, then control alternative u_{is} is chosen,

where s is the value of k, k = 1, ... Γ , which minimizes

$$b_{ik} + \sum_{j=1}^{N} p_{ijk}\, c_{ij} \qquad (8.2)$$

On the other hand, if multi-stage optimization is used, s is the value of k which minimizes

$$b_{ik} + \sum_{j=1}^{N} p_{ijk}(c_{ij} + \eta_j^{\varkappa}) \qquad (8.3)$$

where

$\eta_j^{\varkappa} = \underset{q}{\text{Min}}\ (b_{jq} + v_{jq})$ $\quad q = 1, \ldots \Gamma$
as calculated from the algorithm of section 2.4.

b_{ik} = cost of using control alternative k when process state is i.

p_{ijk} = probability of transition from process state i to process state j when control alternative k is used.

v_{jq} = relative cost of occupation of decision state (j,q), i.e. the state defined by the choice of control q when the process state is j.

We might now specify a new cost matrix C', whose elements c'_{ij} are given by

$$c'_{ij} = c_{ij} + \eta_j^{\varkappa}$$

Single-stage optimization with the adjoined cost matrix is effected by a choice of k to minimize

$$b_{ik} + \sum_{j=1}^{N} p_{jk}\, c'_{ij} \qquad (8.4)$$

which is of course equivalent to (8.3).

Examination of functions (8.2)-(8.4) allows us to draw the following conclusion:

Any stationary multi-stage Markovian decision problem as defined in section 2.3 may be transformed into an equivalent single-stage problem by the adjoining of a state-dependent cost to the performance criterion. The two problems are equivalent in that multi-stage optimization of the first and single-stage optimization of the second yield identical control policies.

The specification of an adjoint cost function is equivalent to the choice of a set of sub-goals for the Markovian decision problem. By solving the multi-stage problem for

low order systems and comparing the resultant values of η_j^* with a given choice of sub-goals, we may evaluate the effectiveness of the choice. It is hoped that a study of typical problems may yield sufficient insight so that a rational choice of sub-goals may be made for higher order systems without recourse to a complete solution. If successful, such a study would constitute one more step towards an effective computational method of controlling general non-linear stochastic processes.

APPENDIX 1

PROOF OF CONVERGENCE OF ITERATIVE COMPUTATION OF OPTIMAL DECISION MATRIX

The iterative scheme outlined in section 2.4 is an extension of the algorithm originally proposed by Howard[14], and the following proof is also based upon his work.

Suppose we have a policy, denoted A, with cost per stage g^A. Let policy B be found by improvement upon A. We wish to prove that $g^B - g^A \leq 0$.

According to the improvement routine we have for each process state i

$$\sum_{j=1}^{L} d_{ij}^B b_{ij} + \sum_{j=1}^{L} d_{ij}^B v_j^A \leq \sum_{j=1}^{L} d_{ij}^A b_{ij} + \sum_{j=1}^{L} d_{ij}^A v_j^A \qquad \text{(A1.1)}$$

where the superscripts refer to the policy used. Let

$$\gamma_i \equiv \sum_{j=1}^{L} d_{ij}^B b_{ij} - \sum_{j=1}^{L} d_{ij}^A b_{ij} + \sum_{j=1}^{L} d_{ij}^B v_j^A - \sum_{j=1}^{L} d_{ij}^A v_j^A \qquad \text{(A1.2)}$$

Note that $\gamma_i \leq 0$ by definition.

Now from equation (2.12) we have

$$g^B + v_i^B = \sum_{j=1}^{L} r_{ij}^B v_j^B + \ell_i^B \qquad \text{(A1.3)}$$

$$g^A + v_i^A = \sum_{j=1}^{L} r_{ij}^A v_j^A + \ell_i^A \qquad (A1.4)$$

(A1.3)-(A1.4) gives

$$g_B - g^A + v_i^B - v_i^A = \sum_{j=1}^{L} r_{ij}^B v_j^B - \sum_{j=1}^{L} r_{ij}^A v_j^A + \ell_i^B - \ell_i^A \qquad (A1.5)$$

where $$r_{ij} = \sum_{k=1}^{N} p_{ik} d_{kj} \qquad (A1.6)$$

By definition,

$$\ell_i^B - \ell_i^A = \sum_{j=1}^{N} p_{ij} c_{ij} + \sum_{j=1}^{N} p_{ij} \sum_{k=1}^{L} d_{jk}^B b_{jk}$$

$$- \sum_{j=1}^{N} p_{ij} c_{ij} - \sum_{j=1}^{N} p_{ij} \sum_{k=1}^{L} d_{jk}^A b_{jk}$$

$$\ell_i^B - \ell_i^A = \sum_{j=1}^{N} p_{ij} \left(\sum_{k=1}^{L} d_{jk}^B b_{jk} - \sum_{k=1}^{L} d_{jk}^A b_{jk} \right)$$

Substituting (A1.2), we have

$$\ell_i^B - \ell_i^A = \sum_{j=1}^{N} p_{ij} \left(\gamma_j - \sum_{k=1}^{L} d_{jk}^B v_k^A + \sum_{k=1}^{L} d_{jk}^A v_k^A \right)$$

From (A1.6)

$$\ell_i^B - \ell_i^A = \sum_{j=1}^{N} p_{ij}\gamma_j - \sum_{k=1}^{L} r_{ik}^B v_k^A + \sum_{k=1}^{L} r_{ik}^A v_k^A \qquad (A1.7)$$

Let $$\delta_i \equiv \sum_{j=1}^{N} p_{ij}\,\gamma_j$$

$$\Delta g \equiv g^B - g^A$$

$$\Delta v_i \equiv v_i^B - v_i^A$$

Now substitution of (A1.7) into (A1.5) yields

$$\Delta g + \Delta v_i = \delta_i + \sum_{j=1}^{L} r_{ij}^B v_j^B - \sum_{j=1}^{L} r_{ij}^B v_j^A$$

i.e.

$$\Delta g + \Delta v_i = \delta_i + \sum_{j=1}^{L} r_{ij}^B \Delta v_j \qquad (A1.8)$$

Since the expansion of δ_i shows that it is the single-stage cost difference between policies B and A, equation (A1.8) is the analogue of (2.12), and has the same form of solution, viz.

$$\Delta g = \sum_{i=1}^{N} \pi_i \delta_i = \langle \underline{\pi} \;\; \underline{\delta} \rangle \qquad (A1.9)$$

where $\underline{\pi} = (\pi_1, \ldots \pi_N)$ = principal row eigenvector of R^B.

Since $\pi_i \geq 0 \quad \forall i$

and $\delta_i = \sum_{j=1}^{N} p_{ij}\gamma_j \leq 0$

because $p_{ij} \geq 0$ and $\gamma_j \leq 0$

it follows from (A1.9) that

$$\Delta g \leq 0$$

Q.E.D.

Thus each new decision matrix yields a cost per stage which is less than or equal to the value obtained at the last iteration

We use Howard's proof that the minimum cost policy will always be discovered.

Proof by contradiction:

Suppose that $g^B < g^A$ but the algorithm has converged on policy A.

Then for all states $\gamma_i \geq 0$

so that all $\delta_i \geq 0$

Since all $\pi_i \geq 0$, we obtain from (A1.9)

$$\Delta g = g^B - g^A \geq 0$$

i.e. $g^B \geq g^A$

which contradicts the assumption

Q.E.D.

Therefore the optimal decision matrix will always be discovered by iteration.

APPENDIX 2

"EXACT" COST ESTIMATES FROM MULTIDIMENSIONAL BETA DISTRIBUTIONS

Given a set of observations, $M = \{m_{ij}\}$, we wish to obtain the likelihood distribution of cost associated with one transition from state i in an N-state system. We have seen in section 3.3 that if the estimates of the transition probabilities, p_{ij}, are considered to form a multivariate normal distribution (this is true asymptotically), then the normalized likelihood function, $f_i(x)$, may be computed quite easily from equations (3.14), (3.16) and (3.17). Actually the estimates of p_{ij} form a multidimensional beta distribution. In this appendix we shall examine the computation of the cost likelihood function using the exact distribution instead of the approximation.

The estimates $\{\tilde{p}_{i1}, \ldots \tilde{p}_{iN}\}$ have a likelihood function

$$L_i(\tilde{p}_{i1},\ldots\tilde{p}_{iN}) = \prod_{j=1}^{N} (\tilde{p}_{ij})^{m_{ij}}, \qquad 0 \leq \tilde{p}_{ij} \leq 1$$

$$\sum_{j=1}^{N} \tilde{p}_{ij} = 1 \qquad \text{(A2.1)}$$

$$= 0, \text{ otherwise.}$$

For notational simplicity, we shall drop the bar over the symbol $\tilde{p}_{ij}$; it is understood that p_{ij} refers to the estimate in the remainder of this appendix. Since the parameters p_{ij} must always sum to unity, one of them, say p_{iN}, is dependent. The likelihood function therefore takes the form of a hypersurface in N-1 space. We now wish to express the likelihood that the mean cost, μ_i, assumes a particular value x. We begin by eliminating another p_{ij}, say $p_{i(N-1)}$, by substituting

$$x = \sum_{j=1}^{N} p_{ij}\, c_{ij} \qquad \text{(A2.2)}$$

into (A2.1) to obtain $L_i(p_{i1},\ldots p_{i(N-2)},x)$. The likelihood function of x alone is now obtained by integrating out the first N-2 variables.

$$L_i(x) = \underbrace{\int_{-\infty}^{\infty} \cdots\cdots \int_{-\infty}^{\infty}}_{N-2} L_i(p_{i1},\ldots p_{i(N-2)},x) \prod_{j=1}^{N-2} dp_{ij} \qquad \text{(A2.3)}$$

The normalized likelihood function $f_i(x)$ is given by

$$f_i(x) = \frac{L_i(x)}{\int_{-\infty}^{\infty} L_i(x)dx} \qquad \text{(A2.4)}$$

The difficulty of computing (A2.4) is caused both by the dimensionality of the function $L_i(p_{i1}, \cdots p_{i(N-2)}, x)$, and by its shape; the function is bounded by the hyperplanes specified in (A2.1), so that in practice it is necessary to vary the limits of integration according to which bounding hyperplane is intersected by a hyperplane of constant cost.

Example

As a simple example, suppose that we have made three observations of transitions from state 1 in a three state system ($N = L = 3$). The relevant parameters are

$$m_{11} = 1 \qquad m_{12} = 1 \qquad m_{13} = 1$$
$$c_{11} = 6 \qquad c_{12} = 8 \qquad c_{13} = 3$$

It is desired to determine the maximum likelihood value $\hat{\mu}_1$ of the expected cost of one transition from state 1.

We observe that, from (A2.1),

$$\begin{aligned} L_1(p_{11}, p_{12}, p_{13}) &= p_{11}\, p_{12}\, p_{13} \\ &= p_{11}\, p_{12}\, (1 - p_{11} - p_{12}) \\ &= p_{11}\, p_{12} - p_{11}^2\, p_{12} - p_{11}\, p_{12}^2 \end{aligned} \qquad \text{(A2.5)}$$

We now introduce the cost relationship

$$x = \sum_{j=1}^{3} p_{ij}c_{ij} = 6p_{11} + 8p_{12} + 3(1 - p_{11} - p_{12})$$

$$x = 3\ p_{11} + 5\ p_{12} + 3 \qquad \text{(A2.6)}$$

We may substitute

$$p_{12} = \frac{x-3-3p_{11}}{5}$$

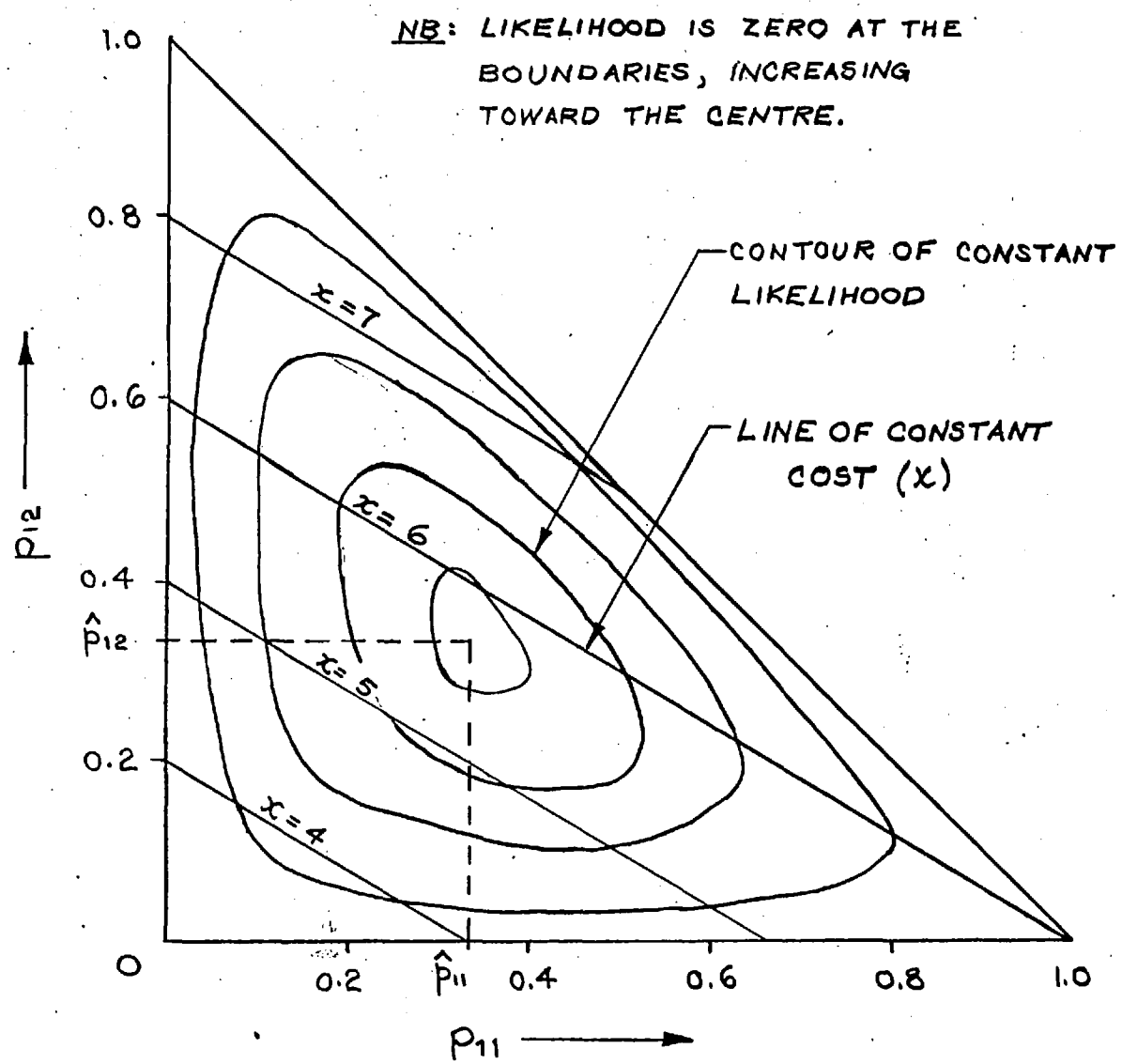
from (A2.6) into (A2.5) to obtain

$$L_i(p_{11},x) = p_{11}\left(\frac{x-3-3p_{11}}{5}\right)\left(1 - p_{11} - \frac{x-3-3p_{11}}{5}\right)$$

Only relative values of a likelihood function are of interest. To rid ourselves of the denominator in the above expression, we multiply by 25, substituting $x_1 = x-3$, to obtain

$$L_i(p_{11},x) = 6p_{11}^3 + x_1p_{11}^2 - 15p_{11}^2 - x_1^2p_{11} + 5x_1p_{11} \qquad \text{(A2.7)}$$

The two-dimensional distribution $L_i(p_{11},p_{12})$ is shown in fig. A2.1, together with lines of constant cost, x. The likelihood of a particular value of x is given by the cross-sectional area cut in the likelihood "hill" by the line of constant x.



$$\hat{p}_{11} = \hat{p}_{12} = \hat{p}_{13} = \frac{1}{3}$$

FIG. A2.1

LIKELIHOOD CONTOURS

IN A THREE-STATE SYSTEM

$$L_i(x) = \int_0^{p_{11}(max)} L_i(p_{11},x)\, d\, p_{11} \qquad (A2.8)$$

Examination of fig. A2.1 shows that $p_{11}(max)$ is a function of x. The relationship is easily seen in the present two-dimensional case, but can become quite complex for higher dimensional distributions. From fig. A2.1 we see that for $3 \leq x \leq 6$, $p_{11}(max)$ is the value of p_{11} for which x corresponds to $p_{12} = 0$.

$$\text{i.e.} \qquad p_{12} = \frac{x-3-3p_{11}}{3} = 0$$

$$p_{11}(max) = \frac{x-3}{3} = \frac{x_1}{3}$$

so that, for $3 \leq x \leq 6$, (A2.8) becomes

$$L_i'(x) = \int_0^{x_1/3} (6p_{i1}^3 + x_1 p_{11}^2 - 15p_{11}^2 - x_1^2 p_{11} + 5x_1 p_{11})dp_{11} \qquad (A2.9)$$

We observe also from fig. A2.1 that for $6 \leq x \leq 8$ the upper limit of p_{11} is defined by the relationship

$$p_{11} + p_{12} = 1$$

Thus $$p_{12} = \frac{x-3-3p_{11}}{5} = 1 - p_{11}$$

$$p_{11}(\max) = \frac{8-x}{2} = \frac{x_2}{2}$$

where $$x_2 = 8 - x = 5 - x_1 \qquad \text{(A2.10)}$$

Subsitution of (A2.10) into (A2.7) yields

$$L_1(p_{11},x) = 6p_{11}^3 - 10p_{11}^2 - x_2p_{11}^2 + 5x_2p_{11} - x_2^2p_{11}$$

and

$$L_1''(x) = \int_0^{x_2/2} L_1(p_{11},x)\, d\, p_{11} \qquad \text{(A2.11)}$$

Integration of (A2.9) and (A2.11), together with the computation of the normalizing factor

$$\left[\int_3^6 L_1'(x)\, dx + \int_6^8 L_1''(x)\, dx \right]^{-1}$$

gives the following expression for $f_1(x)$

$$f_1(x) = 0.96 \left[\frac{5}{2}\left(\frac{x-3}{3}\right)^3 - 2\left(\frac{x-3}{3}\right)^4\right], \quad 3 \le x \le 6$$

$$= 0.96 \left[\frac{5}{3}\left(\frac{8-x}{2}\right)^3 - \frac{7}{6}\left(\frac{8-x}{2}\right)^4\right], \quad 6 \le x \le 8 \qquad \text{(A2.12)}$$

We deduce from (A2.12) that the maximum likelihood value of x is $5\frac{13}{16}$.

Now suppose that we had used the gaussian approximation

$$\hat{p}_{ij} = \frac{m_{ij}}{n_i}$$

so that $\hat{p}_{11} = \hat{p}_{12} = \hat{p}_{13} = \frac{1}{3}$

and $\hat{\mu}_1 = \frac{1}{3}(6) + \frac{1}{3}(8) + \frac{1}{3}(3) = 5\frac{2}{3}$.

Note that the latter value is a good approximation of the value calculated from the multidimensional beta distribution, even when only three transitions have been observed. The computational effort associated with the "exact" case is considerably greater than that of the gaussian case; moreover the difference in computational effort increases very rapidly as the number of states increases. On the other hand the differences between the results of the two methods diminishes as the number of observations grows. For these reasons the normal approximation method of estimation seems the obvious engineering choice.

APPENDIX 3

PROPERTIES OF NORMAL LIKELIHOOD FUNCTIONS

Property 1):

If across an ensemble of statistically equivalent processes the mean value of Ω at stage n is $\Omega(n)$, and a trial is carried out in each process from a particular non-optimal state i ($i \neq s$), then at the $(n+1)^{th}$ stage

$$E[\Omega(n+1)] < E[\Omega(n)] \qquad (A3.1)$$

Proof:

Proof of (A3.1) is equivalent to showing that across the ensemble

$$\frac{\partial \Omega}{\partial n_i} < 0 , \qquad \begin{matrix} i=1, \ldots N \\ i \neq s \end{matrix} \qquad (A3.2)$$

Since σ_i decreases with increasing n_i, (A3.2) is in turn equivalent to

$$\frac{\partial \Omega}{\partial \sigma_i} > 0 \qquad (A3.3)$$

From (3.20)

$$\Omega = 1 - \phi_s$$

where $$\phi_s = \int_{-\infty}^{\infty} f_s(x) \prod_{i=1\neq s}^{N} G_i(x)\, dx \qquad (A3.4)$$

(A3.2) and (A3.3) are thus equivalent to

$$\frac{\partial \phi_s}{\partial \sigma_i} < 0 , \qquad \begin{matrix} i = 1, \ldots N \\ i \neq s \end{matrix} \qquad (A3.5)$$

which we shall proceed to prove.

From (A3.4)

$$\frac{\partial \phi_s}{\partial \sigma_i} = \frac{\partial}{\partial \sigma_i} \int_{-\infty}^{\infty} f_s(x) \prod_{k=1\neq s}^{N} G_k(x)\, dx$$

$$= \int_{-\infty}^{\infty} f_s(x) \prod_{\substack{k=1 \\ \neq i \neq s}}^{N} G_k(x) \frac{\partial G_i(x)}{\partial \sigma_i}\, dx$$

$$= \int_{-\infty}^{\mu_i} f_s(x) \prod_{\substack{k=1 \\ \neq i \neq s}}^{N} G_k(x) \frac{\partial G_i(x)}{\partial \sigma_i}\, dx$$

$$+ \int_{\mu_i}^{\infty} f_s(x) \prod_{\substack{k=1 \\ \neq i \neq s}}^{N} G_k(x) \frac{\partial G_i(x)}{\partial \sigma_i}\, dx$$

$$= \int_{\mu_i}^{\infty} [f_s(x) \prod_{\substack{k=1 \\ \neq i \neq s}}^{N} G_k(x) \frac{\partial G_i(x)}{\partial \sigma_i}$$

(Cont'd)

$$+ f_s(z) \prod_{\substack{k=1 \\ \neq i \neq s}}^{N} G_k(z) \frac{\partial G_i(z)}{\partial \sigma_i} \Big] dx \qquad (A3.6)$$

where $z = 2\mu_i - x$.

The integrand of (A3.6) will be negative for all $x > \mu_i$ providing the following sufficient conditions are met:

a) $$\frac{\partial G_i(x)}{\partial \sigma_i} = - \frac{\partial G_i(z)}{\partial \sigma_i}$$

b) $$\frac{\partial G_i(x)}{\partial \sigma_i} > 0 \qquad x > \mu_i$$

c) $$f_s(x) < f_s(z)$$

d) $$\prod_{\substack{k=1 \\ \neq i \neq s}}^{N} G_k(x) < \prod_{\substack{k=1 \\ \neq i \neq s}}^{N} G_k(z)$$

We shall start with a) and b). We note that

$$\frac{\partial G_i(x)}{\partial \sigma_i} = \frac{\partial}{\partial \sigma_i} \int_x^{\infty} f_i(y)dy = \int_x^{\infty} \frac{\partial f_i(y)}{\partial \sigma_i} dy \qquad (A3.7)$$

and from (3.17)

$$\frac{\partial f_i(y)}{\partial \sigma_i} = \frac{1}{\sqrt{2\pi}\ \sigma_i^2} \left[\left(\frac{y-\mu_i}{\sigma_i}\right)^2 - 1 \right] \exp\left[- \frac{1}{2} \left(\frac{y-\mu_i}{\sigma_i}\right)^2 \right] \qquad (A3.8)$$

Since a change in σ_i yields a net change of zero in the area under curve $f_i(y)$ integrated over the real line, we have

$$\int_{-\infty}^{\infty} \frac{\partial f_i(y)}{\partial \sigma_i} dy = 0 \qquad \text{(A3.9)}$$

Since (A3.8) is symmetrical about $x = \mu_i$ we have

$$\int_{-\infty}^{\mu_i - x} \frac{\partial f_i(y)}{\partial \sigma_i} dy = \int_{\mu_i + x}^{\infty} \frac{\partial f_i(y)}{\partial \sigma_i} dy \qquad \text{(A3.10)}$$

From (A3.7)

$$\int_{\mu_i + x}^{\infty} \frac{\partial f_i(y)}{\partial \sigma_i} dy = \frac{\partial G_i(\mu_i + x)}{\partial \sigma_i} \qquad \text{(A3.11)}$$

From (A3.8) and (A3.9)

$$\int_{-\infty}^{\mu_i - x} \frac{\partial f_i(y)}{\partial \sigma_i} dy = - \int_{\mu_i - x}^{\infty} \frac{\partial f_i(y)}{\partial \sigma_i} dy = - \frac{\partial G_i(\mu_i - x)}{\partial \sigma_i} \qquad \text{(A3.12)}$$

From (A3.10), (A3.11), and (A3.12), we obtain, with a suitable substitution of variables

$$\frac{\partial G_i(x)}{\partial \sigma_i} = - \frac{\partial G_i(2\mu_i - x)}{\partial \sigma_i}$$

and condition a) is proved.

From (A3.7) and (A3.8) it can be seen that

$$\frac{\partial G_i(x)}{\partial \sigma_i} \begin{cases} < 0, & x < \mu_i \\ = 0, & x = \mu_i \\ > 0, & x > \mu_i \end{cases}$$

and b) is proved.

Condition c) follows directly from the defining equation, (3.17), for $f_i(x)$, together with the fact that $\mu_i > \mu_s$ by definition. To prove d) we observe that the first derivative of all curves $G_i(x)$ is negative, i.e.

$$\frac{d\, G_i(x)}{dx} = - f_i(x) < 0$$

$$x$$

$$i = 1, 2, \ldots$$

We shall show that the product of any number of such curves also has a negative first derivative. We proceed by induction, assuming first that the product, $\prod_{i=1}^{M} G_i(x)$ has a negative first derivative, and then showing that multiplication by one more function $G_{M+1}(x)$ preserves this property.

$$\frac{d}{dx} \prod_{i=1}^{M+1} G_i(x) = \prod_{i=1}^{M} G_i(x) \cdot \frac{d\, G_{M+1}(x)}{dx}$$

(Cont'd)

$$+ \frac{d}{dx} \prod_{i=1}^{M} G_i(x) \cdot G_{M+1}(x)$$

Since $0 \leq G_i(x) \leq 1 \qquad \forall\ x$

$$i = 1, 2, \ldots.$$

the derivative of $\prod_{i=1}^{M+1} G_i(x)$ is the sum of two negative numbers and so is negative. Condition d) follows immediately.

In view of condition a), we may re-write (A3.6) as

$$\frac{\partial \phi}{\partial \sigma_i} = \int_{\mu_i}^{\infty} [f_s(x) \prod_{\substack{k=1 \\ \neq i \neq s}}^{N} G_k(x) - f_s(z) \prod_{\substack{k=1 \\ \neq i \neq s}}^{N} G_k(z)] \frac{\partial G_i(x)}{\partial \sigma_i} dx$$

Application of b), c), and d) shows that the integrand is negative for all $x > \mu_i$ so that

$$\frac{\partial \phi_s}{\partial \sigma_i} < 0$$

and property 1) is proved.

Property 2)

$$\lim_{n \to \infty} \Omega(n) = 0$$

if and only if

$$n \to \infty \Rightarrow n_i \to \infty \quad \forall_i$$

Proof:

It is necessary to prove that $\Omega \to 0$ only when all n_i approach infinity.

$$\Omega = 1 - \int_{-\infty}^{\infty} f_s(x) \prod_{i=1\neq s}^{N} G_i(x)\, dx$$

and $$\Omega \to 0 \Rightarrow \int_{-\infty}^{\infty} f_s(x) \prod_{i=1\neq s}^{N} G_i(x)\, dx \to 1 \quad (A3.13)$$

Since $f_s(x) > 0 \quad \forall\, x$

and $$\int_{-\infty}^{\infty} f_s(x)\, dx = 1$$

and $$0 \leq \prod_{i=1\neq s}^{N} G_i(x) \leq 1$$

the integral in (A3.13) can approach unity if and only if:

a) $f_s(x) \to 0$ for all x at which $\prod_{i=1\neq s}^{N} G_i(x) < 1$

b) $\prod_{i=1\neq s}^{N} G_i(x) \to 1$ for all x at which $f_s(x) \neq 0$.

Let μ_j be the second lowest mean cost; we know from the definition of $G_i(x)$ (equation (3.19)) that $\prod_{i=1\neq s}^{N} G_i(x) < 1$

for $x \geq \mu_j$ regardless of the values of n_i, $i \neq s$. From a) above, it is necessary that $f_s(\mu_j) \to 0$. Assuming that all mean costs μ_i are finite, this can be ensured if and only if $\sigma_s \to 0$, so that $f_s(x)$ is zero at all values of x except $x = \mu_s$. It follows from (3.16) that n_s must approach infinity.

With $n_s \to \infty$ it is necessary, from b) above, that $G_i(\mu_s) \to 1 \quad \forall_i$, $i \neq s$. The definition of $G_i(x)$ shows that it can approach unity for a finite value of $\mu_i - x$ if and only if $\sigma_i \to 0$, i.e. $n_i \to \infty \quad \forall_i$, $i \neq s$.

Thus it is both necessary and sufficient that all values n_i increase without limit in order that the uncertainty, Ω, approach zero; property 2) is proved.

APPENDIX 4

TRANSFORMATIONS OF NOISE DISTRIBUTIONS

An IBM SHARE library subroutine was available for the generation of noise with a flat and a gaussian distribution. From time to time in this work other distributions were also required. It is the purpose of this appendix to demonstrate the method used to transform a flat distribution into one of any desired shape.

Let $f(x)$ be the probability density function of the desired signal x, and $y = F(x)$ be its cumulative distribution function. We shall assume that there exists, also an inverse transformation, $x = F^{-1}(y)$. Let z be a random variable whose amplitude is distributed uniformly between zero and one with density $f(z)$. The required signal is generated by the transformation

$$x = F^{-1}(z) \qquad (A4.1)$$

To demonstrate this, we observe in fig. A4.1 that the relationship between the original signal z and the transformed signal $\mathbf{x}$ is

$$f(z)\, dz = f\,[F^{-1}(z)]\, dx \qquad (A4.2)$$

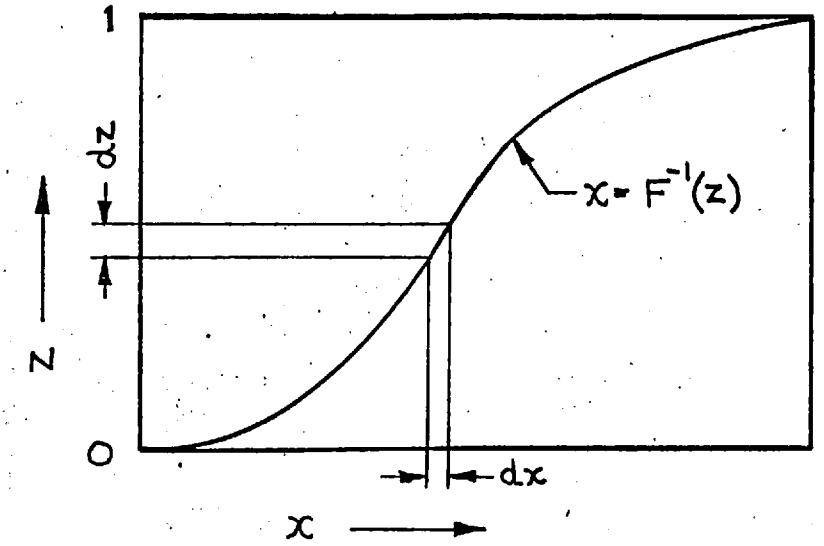


FIG. A4.1

TRANSFORMATION OF

PROBABILITY DISTRIBUTIONS

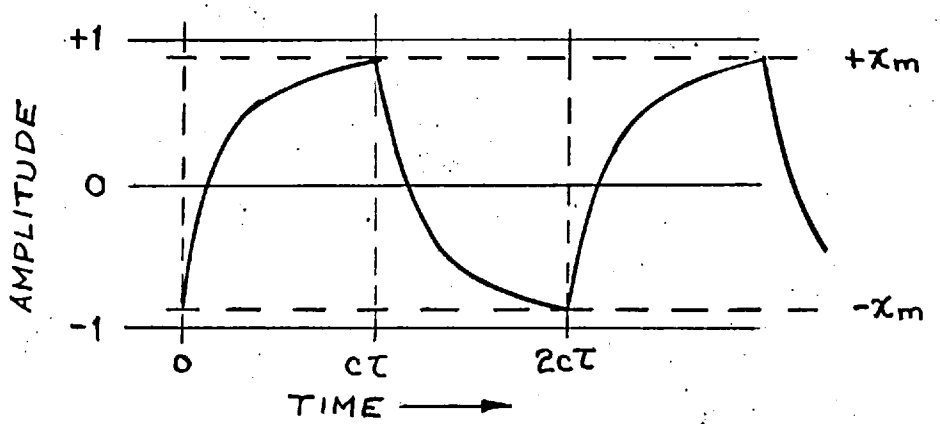


FIG. A4.2

LOW PASS FILTER

SWITCHING WAVEFORM

but $$\frac{dz}{dx} = f(x)$$

so that $$f[F^{-1}(z)] = f(z)\ f(x)$$

Since f(x) is uniform and $\int_0^1 f(x)dz = 1$, it follows that f(z) = 1 and

$$f[F^{-1}(z)] = f(x)$$

so that the transformed variable $F^{-1}(z)$ has the desired distribution, and (A4.1) is valid.

Several possibilities which may be encountered are

1) $F^{-1}(z)$ is analytic;

2) $F^{-1}(z)$ is piece-wise analytic

3) $F^{-1}(z)$ is non-analytic

As an example of 1) we may consider the generation of Rayleigh noise with a density function

$$f(x) = \frac{x}{\sigma^2}\ \exp\ [-\frac{x^2}{\sigma^2}], \quad x > 0$$

so that $$F(x) = \int_{-\infty}^{x} f(y)dy = 1 - \exp[-\frac{x^2}{2\sigma^2}]$$

Now if z is a signal whose amplitude is uniformly distributed between zero and unity, a Rayleigh distribution is generated by the transformation $x = F^{-1}(z)$, i.e.

$$x = \sqrt{2}\,\sigma\,[-\log(1-z)]^{\frac{1}{2}} \qquad (A4.3)$$

An example of 2) is the distribution of the output of a single lag low pass filter whose input is a symmetrical rectangular waveform with amplitude $\pm$ 1 unit. The output signal, shown in fig. A4.2, has limits $\pm\, x_m$; it is assumed that each half period of the input waveform is of duration $c\tau$ where τ is the filter time constant. The positive-going portion of the output is given by

$$x = 1 + (1 + x_m)\exp(-\frac{t}{\tau})$$

and the negative-going half by

$$x = (1 + x_m)\exp(-\frac{t}{\tau}) - 1$$

The respective distribution functions are

$$F_1(x) = -\frac{1}{c}\log(\frac{1-x}{1+x_m})$$

and

$$F_2(x) = 1 + \frac{1}{c}\log(\frac{1+x}{1+x_m})$$

The overall distribution function is $\frac{1}{2}(F_1+F_2)$

$$F(x) = \frac{1}{2c}[c + \log(\frac{1+x}{1-x})]$$

If z is uniformly distributed, then the desired transformation is

$$x = \frac{q-1}{q+1} \qquad (A4.4)$$

where $q = \exp[c(2z-1)]$.

The third case, in which $F^{-1}(z)$ is non-analytic, is frequently encountered. An example is the generation of noise with a beta distribution. A method which has proved successful in such circumstances is that in which a look-up table of $F^{-1}(z)$ is computed numerically for a fixed set of values of z, and second order interpolation is used to obtain x on-line.

When a program has been written to generate the required type of noise, it is a good precaution to submit some of the resultant output to statistical testing[5] to ensure that the distribution is the desired one. Either a goodness-of-fit (chi-squared) test or a likelihood ratio test is suitable. The former was applied to all noise-generating programs used in this thesis.

Parameters for Batch Process, Chapter 4

To simulate the twenty-state batch chemical process of chapter 4, a combination of Rayleigh noise and low pass filter switching noise, as previously described, was used. FORTRAN versions of (A4.3) and (A4.4) with $\sigma = 0.707$ and $c = 3$ were used to generate samples, denoted S1 and S2

respectively. In addition the noise amplitude was modulated by a state dependent function. The Rayleigh samples generated in state I were multiplied by a function RAY(I), while the switching noise was multiplied by REG(I). In FORTRAN, these functions are given by

```
RAY(I) = FAC * FLOAT(I)
```

and

```
REG(I) = 0.000775 * (FLOAT(I-12)) * * 4 + 0.25
```

where FAC = $0.8/\sqrt{\pi}$.

If the state at time n is I, then the state at time n + 1, denoted JUMP1, is given by

```
JUMP1 = I + INT(0.5 + RAY(I) * S1 + REG(I) * S2)
```

APPENDIX 5

POWER DEMAND TRANSITION MATRICES

The twelve demand transition matrices for the ordering problem of chapter 5 are listed below. It is assumed that demand lies within certain limits in any interval. Consequently only the rows representing these states need be specified. The remaining rows can be arbitrary, provided the pertinent states can be reached from them. All positions left blank in the listing are assumed to be zero.

Interval	Row	Column 1	2	3	4	5	6	7	8	9	10
1	2	0.7	0.3								
	3	0.1	0.8	0.1							
	4		0.2	0.7	0.1						
	5		0.2	0.3	0.5						
2	1	0.95	0.05								
	2	0.8	0.2								
	3	0.2	0.7	0.1							
	4	0.05	0.3	0.6	0.05						

Interval	Row	Column 1	2	3	4	5	6	7	8	9	10
3	1			0.1	0.2	0.6	0.1				
	2			0.05	0.2	0.5	0.2	0.05			
	3				0.1	0.3	0.5	0.1			
	4					0.2	0.5	0.3			
4	3						0.5	0.3	0.2		
	4						0.1	0.6	0.3		
	5							0.15	0.7	0.15	
	6							0.1	0.5	0.4	
	7								0.2	0.7	0.1
5	6							0.3	0.5	0.2	
	7								0.7	0.3	
	8								0.2	0.6	0.2
	9									0.4	0.6
	10									0.2	0.8
6	7							0.4	0.4	0.2	
	8							0.1	0.3	0.5	0.1
	9								0.05	0.7	0.25
	10									0.25	0.75
7	7							0.3	0.6	0.1	
	8								0.3	0.6	0.1
	9									0.3	0.7
	10									0.1	0.9

Interval	Row	Column 1	2	3	4	5	6	7	8	9	10
8	7						0.5	0.4	0.1		
	8						0.4	0.5	0.1		
	9							0.3	0.6	0.1	
	10							0.1	0.7	0.15	0.05
9	6				0.3	0.6	0.1				
	7					0.3	0.6	0.1			
	8						0.5	0.4	0.1		
	9						0.1	0.7	0.2		
	10							0.5	0.5		
10	4						0.8	0.2			
	5						0.5	0.4	0.1		
	6							0.4	0.5	0.1	
	7							0.2	0.5	0.2	0.1
	8								0.1	0.8	0.1
11	6				0.6	0.3	0.1				
	7					0.6	0.3	0.1			
	8						0.6	0.3	0.1		
	9						0.3	0.6	0.1		
	10						0.2	0.4	0.4		
12	4		0.8	0.2							
	5		0.1	0.6	0.3						
	6		0.05	0.3	0.5	0.15					
	7			0.2	0.2	0.6					
	8			0.05	0.25	0.7					

APPENDIX 6

FORTRAN LISTINGS

During the course of this work, several score of main programs, subroutines, and arithmetic functions were written in FORTRAN IV. Three of these which are of particular importance are presented here. Minor modifications from the original versions have been made in layout to accommodate the width of the quarto page. To enable the reader to interpret these programs, the correspondence between FORTRAN variables and those used in the text of the thesis is given in table A6.1; the meaning of the additional auxiliary functions and subroutines appearing within the listings is shown in table A6.2.

TABLE A6.1

CORRESPONDENCE OF FORTRAN AND TEXT VARIABLES

FORTRAN	Text
B(I,J)	b_{ij}
C(I,J)	c_{ij}
CHOICE(I,K)	u_{ik}
CONFAC	γ
EL(I)	ℓ_i
EN(I,K)	n_{ik}
GAIN	g
GRIN(I,J)	element of Ψ^{-1}, basic chain
KDEC(I)	$\hat{s}(i)$
M	Γ
N	N
OMEGA	Ω
P(I,J,K)	$\hat{p}_{ijk}$
QLO(I)	lower limit of output variable, state i
QMID(I)	mid point value of output variable, state i
TRANS(I,J,K)	m_{ijk}
URG(I,J)	element of Ψ, basic chain
VALUE(I)	$\hat{v}_{i\hat{s}}$, basic chain
VARO(I,K)	$\hat{\sigma}^2_{ik}$
VNU(I,K)	$\hat{v}_{ik}$
COL ETA HOLD NOBS RVEC SET	working space

TABLE A6.2

FUNCTIONS AND SUBROUTINES IN FORTRAN LISTINGS

Name	Parameter Computed
CVAR(I,K,P,C,N,M)	σ_{ik}^2
FINDG(Y,Z)	$\int_{-\sqrt{Y}.Z}^{\infty} \exp(-\frac{x^2}{2})dx$
FINPRO(Y,Z,N)	$< \underline{Y} \quad \underline{Z} >$
INVERT(A,B,N)	$B = A^{-1}$
IQUANT(TEMP,QMID,N)	Quantized state, i, of output variable TEMP.
MAVEC(A,Y,Z,N)	$\underline{Z} = A\underline{Y}$
NORMAL	Look-up table of parameters f(x) and G(x) relating to normal density function.
RNGEN(K)	Random numbers, uniformly distributed between 0 and 1.
SETB(B,CHOICE,N,M)	B
SETC(C,QMID,N)	C
SETP(P,.....)	P
TRY(P,.....)	g(B,C,P,D)

a) Function NUKR

Reference: Section 6.2

Purpose: Given present process state, IR, chooses control alternative NUKR from the set $\{1,2, \ldots \Gamma\}$.

```
£IBFTC NUKR2   DECK

      FUNCTION NUKR (IR,B,KDEC,VNU,VARO,EN,OMEGA,KODE,
     1CONFAC,N,M)

      DIMENSION B(N,M), KDEC(N), VNU(N,M), VARO(N,M),
     1EN(N,M), SET(50)

C     11 OCTOBER 1966.
C     IR IS STATE QUANTUM LEVEL, GIVEN.
C     KR IS CONTROL QUANTUM LEVEL, TO BE DETERMINED.

      KD=KDEC(IR)
      ETAMIN=B(IR,KD)+VNU(IR,KD),
      SUM=0.
      DO 4 K=1,M
      IF (K.EQ.KD) GO TO 4
      ENEST=EN(IR,K)
      SET(K)=(B(IR,K)+VNU(IR,K)-ETAMIN)/SQRT(VARO(IR,K))
      SUM=SUM+1.-FINDG(ENEST,SET(K))
    4 CONTINUE

      OMEGA=SUM
      IF (SUM.GT..5) SUM=.5
      IF (1.-CONFAC*SUM.LT.RNGEN(KODE)) GO TO 5
```

```
      KR=KD
      GO TO 7

    5 KR=1
      IF (KR.EQ.KD) KR=2
      DO 6 K=1.M
      IF (K.EQ.KD) GO TO 6
      ENEST=EN(IR,K)
      SET(K)=SET(K)**2*ENEST+ALOG(ENEST*VARO(IR,K))
      IF (SET(K).LT.SET(KR)) KR=K
    6 CONTINUE

C     ALTERNATIVE KR HAS NOW BEEN CHOSEN.

    7 NUKR=KR
      RETURN
      END
```

b) Subroutine UPDATE

Reference: Section 6.3

Purpose: Given observed transition from decision state (i,k), updates $\hat{P}$, $\hat{\sigma}_{oik}{}^2$, Ψ^{-1}, $\hat{\underline{\ell}}$, $\hat{D}^*$, and $\hat{g}$.

```
£IBFTC UPDAT5  DECK

C       29 SEPTEMBER 1966.
        SUBROUTINE UPDATE (TRANS,P,B,C,URG,GRIN,KDEC,VARO,
       1VNU,EL,VALUE,GAIN,IR,KR,N,M)
        DIMENSION TRANS(N,N,M), P(N,N,M), B(N,M), C(N,N),
       1GRIN(N,N), KDEC(N), VARO(N,M), VNU(N,M), EL(N),
       2RVEC(50), COL(50), URG(N,N), VALUE(N)

C       PRODUCES UPDATED MATRIX P, VECTORS VARO, VALUE,
C       AND VNU, AND SCALAR GAIN.

C       BEGIN BY UPDATING P AND VARO.

        SUM=0.
        DO 1 J=1,N
      1 SUM=SUM+TRANS(IR,J,KR)
        FAC=1./SUM
        DO 2 J=1,N
      2 P(IR,J,KR)=TRANS(IR,J,KR)*FAC
        VARO(IR,KR)=CVAR(IR,KR,P,C,N,M)

C       P AND VARO UPDATED.

        GLAST=1.0E30
```

```
      KOUNT=0
      NA=N-1
      IF (KR.EQ.KDEC(IR)) GO TO 5

C     IF THIS POINT IS REACHED, STATE IR IS NOT IN THE
C     BASIC CHAIN.
C     MATRIX URG THEREFORE DOES NOT REQUIRE UPDATING.
C     ONLY VNU(IR) NEED BE RECOMPUTED.

      VNU(IR,KR)=-GAIN
      DO 4 J=1,N
      K=KDEC(J)
    4 VNU(IR,KR)=VNU(IR,KR)+P(IR,J,KR)*(C(IR,J)+B(J,K)
     1+VNU(J,K))
      GO TO 14

C     MAIN DEVELOPMENT CONTINUES.

    5 DO 6 J=1,N
    6 VALUE(J)=URG(IR,J)

C     VALUE HOLDS FORMER ELEMENTS OF ROW IR.
C     NEW ROW IS NOW COMPUTED.

      DO 7 J=1,NA
      URG(IR,J)=-P(IR,J,KR)
      IF (IR.EQ.J) URG(IR,J)=URG(IR,J)+1.
    7 CONTINUE

C     NEW URG IS NOW FORMED.
C     UPDATE INVERSE (GRIN) BY HOUSEHOLDERS METHOD.
```

```
      DO 8 J=1,N
      VALUE(J)=URG(IR,J)-VALUE(J)
    8 COL(J)=GRIN(J,IR)
      FAC=1./(FINPRO(VALUE,COL,N)+1.)
      DO 9 I=1,N
      RVEC(I)=0.
      DO 9 J=1,N
    9 RVEC(I)=RVEC(I)+VALUE(J)*GRIN(J,I)
      DO 10 I=1,N
      DO 10 J=1,N
   10 GRIN(I,J)=GRIN(I,J)-FAC*COL(I)*RVEC(J)

C     NEW INVERSE=OLD INVERSE-FAC*(OUTER PRODUCT)

      EL(IR)=0.
      DO 11 J=1,N
      K=KDEC(J)
   11 EL(IR)=EL(IR)+P(IR,J,KR)*(C(IR,J)+B(J,K))
   12 CALL MAVEC(GRIN,EL,VALUE,N)
      KOUNT=KOUNT+1

C     VALUE IS Z VECTOR OF BASIC CHAIN.
C     NOW TRANSFORM VALUE INTO V VECTOR, AND COMPUTE
C     QUANTITIES VNU OF OVERALL CHAIN

      GAIN=VALUE(N)
      VALUE(N)=0.
      DO 13 I=1,N
      K=KDEC(I)
   13 VNU(I,K)=VALUE(I)

      DO 32 I=1,N
```

```
      DO 32 K=1,M
      IF (K.EQ.KDEC(I)) GO TO 32
      VNU(I,K)=-GAIN
      DO 31 J=1,N
      L=KDEC(J)
   31 VNU(I,K)=VNU(I,K)+P(I,J,K)*(C(I,J)+B(J,L)+VNU(J,L))
   32 CONTINUE
      IF (GAIN.LT.GLAST) GO TO 14

C     IF TWO STRATEGIES ARE EQUIVALENT, NUMERICAL
      INSTABILITY MAY RESULT OWING TO ROUNDOFF ERROR.
C     IF THE APPARENT COST DIFFERENCE IS EITHER ZERO OR
C     VERY SMALL AND POSITIVE FROM ONE ITERATION TO THE
C     NEXT, FURTHER ITERATION IS STOPPED.

      IF (ABS((GAIN-GLAST)/GAIN).GT.1.0E-06) GO TO 27
      RETURN

C     OPTIMALITY TEST FOLLOWS.

   14 KLAG=0
      DO 17 I=1,N
      L=1
      DO 15 K=2,M
      IF (B(I,K)+VNU(I,K).LT.B(I,L)+VNU(I,L)) L=K
   15 CONTINUE
   16 IF (KDEC(I).EQ.L) GO TO 17
      KLAG=1
      KDEC(I)=L
   17 CONTINUE
      IF (KLAG.EQ.0) RETURN
      IF (KOUNT.GT.19) GO TO 25
```

```
C     IF DECISION MATRIX IS UNCHANGED, UPDATING IS COMPLETE.
C     OTHERWISE, MATRIX URG MUST BE RECOMPUTED AND INVERTED.

      GLAST=GAIN
      DO 18 I = 1,N
      K=KDEC(I)
      DO 18 J=1,NA
      URG(I,J)=-P(I,J,K)
      IF (I.EQ.J) URG(I,J)=URG(I,J)+1.
   18 CONTINUE
      CALL INVERT (URG,GRIN,N)

      DO 19 I=1,N
      EL(I)=0.
      K=KDEC(I)
      DO 19 J=1,N
      L=KDEC(J)
   19 EL(I)=EL(I)+P(I,J,K)*(C(I,J)+B(J,L))
      GO TO 12
   25 WRITE (6,26)
   26 FORMAT (//35H NO CONVERGENCE AFTER 20 ITERATIONS)
      RETURN
   27 KX=KOUNT-1
      DEL=GAIN-GLAST
      WRITE (6,28) KX, GLAST, KOUNT, GAIN, DEL
   28 FORMAT (/28H ERROR IN UPDATING PROCEDURE/
     134H EXPECTED COST/CYCLE AT ITERATION I3,F14.6/
     234H EXPECTED COST/CYCLE AT ITERATION I3,F14.6/
     319H INCREASE IN COST=E12.4//)
      RETURN
      END
```

c) Main program AD11NL

Reference: Sections 6.5 and 7.3

Purpose: Given a priori estimate of system dynamics, simulates adaptive control of heat treatment problem, chapter 7.

```
£IBFTC AD11NL

      DIMENSION P(11,11,5), TRANS(11,11,5), C(11,11),
     1CHOICE(11,5), URG(11,11), GRIN(11,11), VARO(11,5),
     2KDEC(11), EL(11), VALUE(11), QLO(11), QMID(11),
     3HOLD(5), ETA(5), EN(11,5),
     4B(11,5), VNU(11,5), NOBS(11)

      CALL NORMAL
      READ (5,1) N, M, PHI, UFAC, SIG
    1 FORMAT (2I6/3F10.1)
      WRITE (6,2) PHI, UFAC, SIG
    2 FORMAT (25H1J S RIORDON  13 OCT 1966//
     138H ADAPTIVE CONTROL OF NON-LINEAR SYSTEM/
     226H WITH MULTIPLICATIVE NOISE//
     528H A PRIORI MODEL OF SYSTEM IS/
     220X,9H X(K+1)= F5.2,10H * X(K) + F6.4,
     217H * U(K) + ZETA(K)//
     342H WHERE ZETA IS NORMALLY DISTRIBUTED NOISE,
     422H ZERO MEAN, STD DEVN= F5.2///)
    7 READ (5,8) (QLO(I), I=1,N)
    8 FORMAT (3(5F10.1/),5F10.1)
      NA=N-1
      DO 9 I=1,NA
```

```
    9 QMID(I)=.5*(QLO(I+1)+QLO(I))
      QMID(N)=QLO(N)
      WRITE (6,10) (QLO(I), I=1,N), (QMID(I), I=1,N)
   10 FORMAT (///19H STATE QUANTIZATION//
     113H LOWER LIMITS//11F10.2//
     211H MID POINTS//11F.10.2)
      MC=(M+1)/2
      MP=M-1

C     SETUP OF MATRIX CHOICE.

      READ (5,17) (CHOICE(I,MC), I=1,N), BAND
   17 FORMAT (6F10.1/5F10.1/F10.1)
      DO 18 I=1,N
      XL=CHOICE(I,MC)-.5*BAND
      XU=XL+BAND
      IF (XL.LT.-10000.) XL=-10000.
      IF (XU.GT.10000.) XU=10000.
      RANGE=XU-XL
      DO 18 K=1,M

   18 CHOICE(I,K)=XL+RANGE/FLOAT(M-1)*FLOAT(K-1)

      WRITE (6,11) (K, K=1,M), (I, QMID(I),
     1(CHOICE(I,K), K=1,M), I=1,N)
   11 FORMAT (16H1CONTROL CHOICES//
     16H STATE,4X,6H TEMP ,5(I7,3X)//
     220(I4,F12.2,5F10.2/)///)

      CALL SETB(B,CHOICE,N,M)

      WRITE (6,12) (J, J=1,M), (I, (B(I,J), J=1,M), I=1,N)
```

```
12 FORMAT (20H1CONTROL COST MATRIX//4X,5I10//
  120(I6,5F10.4/)//)

   CALL SETC(C,QMID,N)

   WRITE (6,13) (I, I=1,N)
13 FORMAT (23H1TRANSITION COST MATRIX//5X,11I6//)
   DO 15 I=1,N
   DO 14 J=1,N
14 NOBS(J)=C(I,J)*100.+1.0E-4
15 WRITE (6,16) I, (NOBS(J), J=1,N)
16 FORMAT (21I6)

   CALL SETP (P,QLO,QMID,CHOICE,PHI,UFAC,SIG,N,M)
   DO 22 I=1,N
   NOBS(I)=0
   KDEC(I)=MC
   DO 22 K=1,M
   VARO(I,K)=CVAR(I,K,P,C,N,M)
   EN(I,K)=1.
   DO 22 J=1,N
22 TRANS(I,J,K)=P(I,J,K)
   IR=1
   KR=1
   CALL TRY(P,B,C,URG,GRIN,KDEC,VNU,EL,VALUE,GAIN,N,M)
   CALL UPDATE (TRANS,P,B,C,URG,GRIN,KDEC,VARO,VNU,EL,
  1VALUE,GAIN,IR,KR,N,M)
   WRITE (6,23) (I, QMID(I), CHOICE(I,MC), I=1,N), GAIN
23 FORMAT (41H1NOMINAL ESTIMATE OF THE OPTIMUM FEEDBACK/
  14OH TRANSDUCER CHARACTERISTIC IS AS FOLLOWS///
  26H STATE,4X,5H TEMP,4X,11H HEAT INPUT/
  38X,8H DEG ABS,6X,5H KCAL//
```

```
  411(I5,F11.2,F12.2/)//
  531H EXPECTED COST PER TRANSITION= F10.4)

   NSHUT=0
   TOTAL=0.
   TEMP=775.
   JR=IQUANT(TEMP,QMID,N)
   READ (5,60) NTRANS, KODE, NSAM, NREDO, CONFAC
60 FORMAT (4(I6/),F10.1)
   READ (5,24) NPRINT
24 FORMAT (I6)
   WRITE (6,61) NTRANS, CONFAC
61 FORMAT (23H1ADAPTIVE CONTROL OVER I4,11H INTERVALS./
  123H CONVERGENCE FACTOR IS F4.1//)
   IF (NPRINT.EQ.1) WRITE (6,25)
25 FORMAT (6H INTVL,3X,5H TEMP,3X,3H IR,2X,3H KD,4X,
  16H OMEGA,4X,
  28H CONTROL,2X,3H KR,5X,5H COST,4X,6H TOTAL,5X,
  35H MEAN//)
   TMEAN=0.
   TVAR=0.
   CLAST=0.

   DO 85 NUM=1,NTRANS

   IR=JR
   NOBS(IR)=NOBS(IR)+1
   TMEAN=TMEAN+TEMP
   TVAR=TVAR+TEMP**2
   KR=NUKR(IR,B,KDEC,VNU,VARO,EN,OMEGA,KODE,CONFAC,N,M)
   UCOST=B(IR,KR)
   TLAST=TEMP
```

```
      XDEL=TLAST-800.
      FAC=0.
      IF (XDEL.GT.0.)FAC=1.
      TEMP=TLAST*(1.005+.015*TANH(.1*(XDEL-3.466))
     1+.0002*FAC*XDEL*GAUS(KODE)
     2+(.005*GAUS(KODE+1))/(1.+SQRT(ABS(XDEL))))
     3+CHOICE(IR,KR)*(1./300.+.0005*GAUS(KODE+2))
     4+GAUS(KODE+3)
      IF (TEMP.LE.750.) TEMP=750.
      IF (TEMP.GE.850.) GO TO 62
      XCOST=.015*((TEMP-800.0**2+(TLAST-800.)**2)
      GO TO 63
   62 XCOST=.015*(2500.+(TLAST-800.)**2) +2880.
   63 JR=IQUANT(TEMP,QMID,N)
      TRANS(IR,JR,KR)=TRANS(IR,JR,KR)+1.
      EN(IR,KR)=EN(IR,KR)+1.
      COST=XCOST+UCOST
      TOTAL=TOTAL+COST
      GAVG=TOTAL/FLOAT(NUM)
      IF (TEMP.GE.850.) GO TO 65
      IF (NPRINT.EQ.1) WRITE (6,64) NUM, TLAST, IR,
     1KDEC(IR), OMEGA, CHOICE(IR,KR),KR,COST,TOTAL,GAVG
   64 FORMAT (I5,F10.2,2I5,F10.4,F12.2,I5,3F10.2)
      GO TO 67
   65 IF (NPRINT.EQ.1) WRITE (6,66) NUM, TLAST,IR,
     1KDEC(IR), OMEGA, CHOICE(IR,KR), KR, COST, TOTAL, GAVG
   66 FORMAT (I5,F10.2,2I5,F10.4,F12.2,I5,3F10.2,5X,
     115H PLANT SHUTDOWN)
      TEMP=800.
   67 CALL UPDATE(TRANS,P,B,C,URG,GRIN,KDEC,VARO,VNU,EL,
     1VALUE,GAIN,IR,KR,N,M)
      IF (NSAM*(NUM/NSAM).NE.NUM) GO TO 82
```

```
C      OUTPUT OF INTERMEDIATE RESULTS.

       WRITE (6,68) NUM
   68  FORMAT (15H1RESULTS AFTER I4,13H TRANSITIONS.///
      135H IMPROVED TRANSDUCER CHARACTERISTIC///
      26H STATE,4X,5H TEMP,5X,8H CONTROL,3X,4H NO.,5X,
      36H OMEGA,6X,3H PI,8X,3H EN,4X,4H OBS//)
       TMEAN=TMEAN/FLOAT(NSAM)
       TVAR=SQRT(TVAR/FLOAT(NSAM)-TMEAN**2)
       TOT=0.
       DO 70 I=1,N
       SUM=0.
       KD=KDEC(I)
       ETAMIN=B(I,KD)+VNU(I,KD)
       DO 69 K=1,M
       IF (K.EQ.KD) GO TO 69
       ETA(K)=B(I,K)+VNU(I,K)-ETAMIN
       SUM=SUM+1.-FINDG(EN(I,K),ETA(K)/SQRT(VARO(I,K)))
   69  CONTINUE
       TOT=TOT+SUM*GRIN(N,I)
       WRITE (6,71) I, QMID(I), CHOICE(I,KD), KD, SUM,
      1GRIN(N,I), EN(I,KD), NOBS(I)
   70  NOBS(I)=0
   71  FORMAT (I4,2F12.2,I6,F12.4,F12.6,F8.0,I8//)
       CMINT=(TOTAL-CLAST)/FLOAT(NSAM)
       CLAST=TOTAL
       WRITE (6,72) GAIN, TOT, NSAM, TMEAN, TVAR, NUM,
      1TOTAL, GAVG, NSAM, CMINT
   72  FORMAT (///31H EXPECTED COST PER TRANSITION= F10.4//
      149H UNCERTAINTY MEASURE, INNER PRODUCE (PI, OMEGA)=
      1F10.6//
      224H MEAN TEMPERATURE, LAST I4,14H TRANSITIONS= F6.2/
```

```
     319H STANDARD DEVIATION,23X,F6.2///
     424H TOTAL COST INCURRED IN I4,12H TRANSITIONS,7X,F10.2/
     534H MEAN COST PER TRANSITION, OVERALL,14X,F10.3/
     632H MEAN COST PER TRANSITION, LAST I3,
     713H TRANSITIONS F10.3)
      IF (NUM.NE.NTRANS.AND.NPRINT.EQ.1) WRITE (6,81)
  81 FORMAT (6H1INTVL,3X,5H TEMP,3X,3H IR,2X,3H KD,4X,6H
     1OMEGA,4X,
     28H CONTROL,2X,3H KR,5X,5H COST,4X,6H TOTAL,
     35X,5H MEAN//)
      TMEAN=0.
      TVAR=0.
  82 IF (NREDO*(NUM/NREDO).NE.NUM) GO TO 85

      DO 80 I=1,N

      SUM=0.
      KD=KDEC(I)
      ETAMIN=B(I,KD)+VNU(I,KD)
      DO 73 K=1,M
      IF (K.EQ.KD) GO TO 73
      ETA(K)=B(I,K)+VNU(I,K)-ETAMIN
      SUM=SUM+1.-FINDG(EN(I,K), ETA(K)/SQRT(VARO(I,K)))
  73 CONTINUE
      IF (KD.EQ.M.AND. SUM .LT..5.AND.CHOICE(I,KD).LT.9999.
     1.AND.EN(I,KD).GT.3.) GO TO 77
      IF (KD.NE.1.OR. SUM .GT..5.OR.CHOICE(I,KD).LT.-9999.
     1.OR.EN(I,KD).LT.4.) GO TO 80
      DO 74 KP=1,MP
      K=M-KP+1
      CHOICE(I,K)=CHOICE(I,K-1)
      EN(I,K)=EN(I,K-1)
```

```
      VARO(I,K)=VARO(I,K-1)
      DO 74 J=1,N
      P(I,J,K)=P(I,J,K-1)
   74 TRANS(I,J,K)=TRANS(I,J,K-1)
      CHOICE(I,1)=CHOICE(I,2)-500.
      IF (CHOICE(I,1).LT.-10000.) CHOICE(I,1)=-10000.
      EN(I,1)=1.
      DO 75 J=1,N
      P(I,J,1)=P(I,J,2)
   75 TRANS(I,J,1)=P(I,J,1)
      VARO(I,1)=CVAR(I,1,P,C,N,M)*100.
      GO TO 80
   77 DO 78 K=1,MP
      CHOICE(I,K)=CHOICE(I,K+1)
      EN(I,K)=EN(I,K+1)
      VARO(I,K)=VARO(I,K+1)
      DO 78 J=1,N
      P(I,J,K)=P(I,J,K+1)
   78 TRANS(I,J,K)=TRANS(I,J,K+1)
      CHOICE(I,M)=CHOICE(I,MP)+500.
      IF (CHOICE(I,M).GT.10000.) CHOICE(I,M)=10000.
      EN(I,M)=1.
      DO 79 J=1,N
      P(I,J,M)=P(I,J,MP)
   79 TRANS(I,J,M)=P(I,J,M)
      VARO(I,M)=CVAR(I,M,P,C,N,M)*100.
   80 CONTINUE
      CALL SETB(B,CHOICE,N,M)
      CALL TRY(P,B,C,URG,GRIN,KDEC,VNU,EL,VALUE,GAIN,N,M)
      CALL UPDATE(TRANS,P,B,C,URG,GRIN,KDEC,VARO,VNU,EL,
     1VALUE,GAIN,IR,KR,N,M)
```

REFERENCES

References are divided into sections according to subject matter. It will be appreciated that the division is in many cases somewhat arbitrary, however.

I Some Texts Giving Background Information

1. R.E. Bellman, S.E. Dreyfus, "Applied Dynamic Programming", Princeton University Press; 1962.
2. R. Deutsch, "Estimation Theory", Prentice-Hall; 1965.
3. W. Feller, "An Introduction to Probability Theory and Its Applications", vol.1, 2nd ed., Wiley; 1957.
4. E.I. Jury, "Theory and Application of the Z-Transform Method", Wiley; 1964.
5. A.M. Mood, F.A. Graybill, "Introduction to the Theory of Statistics", 2nd ed., McGraw-Hill; 1963.
6. E. Parzen, "Stochastic Processes", Holden-Day; 1962.
7. J.T. Tou, "Modern Control Theory", McGraw-Hill; 1964.
8. D.J. Wilde, "Optimum Seeking Methods", Prentice-Hall; 1964.
9. L.A. Zadeh, C.A. Desoer, "Linear System Theory", McGraw-Hill; 1963.

II Markov Chains in Operations Research and Economics

10. M.J. Beckman, "On the Theory of Stochastic Control Processes", Bull. Soc. Royale Sciences Liege, 9/10, 520-529; 1964.

11. D. Blackwell, "Discounted Dynamic Programming", Ann. Math. Statist., 36, 226-235; 1965.

12. C. Derman, "On Sequential Control Processes", Ann. Math. Statis. 35, 341-349; 1964.

13. G. Hadley, "Non-Linear and Dynamic Programming", ch.11, Addison-Wesley, 1964.

14. R.A. Howard, "Dynamic Programming and Markov Processes", Wiley, 1962.

15. W.S. Jewell, "Markov-Renewal Programming", Opns. Res., 11, 938-971; 1963.

16. R.E. Lave, "A Markov Decision Process for Economic Quality Control", I.E.E.E. Trans. Systems Science and Cybernetics, 2, 45-54; 1966.

III Dual Control

17. M. Aoki, "Optimal Control Policies for Dynamical Systems Whose Characteristics Change Randomly at Random Times", Proc. 3rd I.F.A.C. Congress, paper 29A; 1966.

18. A.A. Feldbaum, "Dual Control Theory", Automat.

Remote Control, 21, 874-880, 1033-1039; 1960. 22, 1-12, 109-121; 1961.

19. A.A. Feldbaum, "Optimal Control Systems", Academic Press, 1965.

20. H.H. Rosenbrock, "An Example of Optimal Adaptive Control", J. Electron. Control, 13, 557-567; 1964.

21. D.D. Sworder, "Control of a Linear System with a Markov Property", I.E.E.E. Trans. Automatic Control, 10, 294-300; 1965.

22. D.D. Sworder, "Optimal Adaptive Control Systems", Academic Press, 1966.

23. J.T. Tou, "Systems Optimization via Learning and Adaptation", Internat. J. of Control, 2, 21-31; 1965.

24. D.A. Xirokostas, J.G. Henderson, "Extremum Control of Dynamic Systems in the Presence of Random Disturbances and Noise", Proc. 3rd I.F.A.C. Congress, paper 7B; 1966.

IV Discrete State Learning Control Systems

25. V.Y. Krylov, "On One Automaton that is Asymptotically Optimum in a Random Medium", Automat. Remote Control, 24, 1114-1116; 1963.

26. R.W. McLaren, "A Markov Model for Learning Systems Operating in an Unknown Random Environment", Proc. Nat. Electronics Conf. 1964, 585-589.

27. G.J. McMurty, K.S. Fu, "On the Learning Behaviour of Finite State Systems in Random Environments", Proc.2nd Annual Allerton Conf. on Circuit and System Theory, 618-642; 1964.

28. G.J. McMurty, K.S. Fu, "A Variable Structure Automaton Used as a Multi-Modal Searching Technique", Proc. Nat. Electronics Conf. 1965, 494-499. Also I.E.E.E. Trans. Automatic Control, 11, 379-387; 1966.

29. J. Sklansky, "Learning Systems for Automatic Control", I.E.E.E. Trans. Automatic Control, 11, 6-19; 1966.

30. M.L. Tsetlin, "On the Behaviour of Finite Automata in Random Media", Automat. Remote Control, 22, 1210-1219; 1961.

31. V.J. Varshavskii, J.P. Vorontsova, "On the Behaviour of Stochastic Automata with a Variable Structure", Automat. Remote Control, 24, 327-333; 1963.

32. M.D. Waltz, K.S. Fu, "A Heuristic Approach to Reinforcement Learning Control Systems", I.E.E.E. Trans. Automatic Control, 10, 390-398; 1965.

V Dual Control in Discrete State Markov Processes

33. L. Meier, "Combined Optimal Control and Estimation", Proc. 3rd Annual Allerton Conf. on Circuit and Systems Theory, 109-120; 1965.

34. Z.J. Nikolic, K.S. Fu, "An Algorithm for Learning without External Supervision and Its Application to Learning Control Systems", I.E.E.E. Trans. Automatic Control, 11, 414-422; 1966.

35. S. Pashkovskii, "Adaptive Control for a System with a Finite Number of States", Proc. 2nd I.F.A.C. Congress (1963), Butterworths, 241-245.

36. J.S. Riordon, "Extremum Seeking in Discrete State Markov Processes with Uncertain Transition Matrices", Control Systems Group report no.2, Imperial College, University of London, Dec. 1965.

37. E.A. Silver, "Markovian Decision Processes with Uncertain Transition Probabilities or Rewards", Opns. Res. Centre, Massachusetts Inst. Tech., Aug. 1963.

VI Discrete State Markov Processes with Uncertainty in State Measurement

38. K.J. Åström, "Optimal Control of Markov Processes with Incomplete State Information", J. Math. Anal. Appl., 10, 174-205; 1965.

39. R.L. Kashyap, "Optimization of Stochastic Finite State Systems", I.E.E.E. Trans. Automatic Control, 11, October 1966.

40. R.E. Larson, J. Peschon, "A Dynamic Programming Approach to Trajectory Estimation", I.E.E.E. Trans. Automatic Control, 11, 537-540; 1966.

VII Sequential Estimation Procedures

41. H. Chernoff, "Sequential Tests for the Mean of a Normal Distribution", Ann. Math. Statist., 36, 28-68; 1965.

42. S.N. Ray, "Bounds on the Maximum Sample Size of a Bayes Sequential Procedure", Ann. Math. Statist. 36, 859-878; 1965.

VIII The Generator Ordering Problem

43. C.J. Baldwin et al., "A Study of the Economic Shutdown of Generating Units in Daily Dispatch", Trans. A.I.E.E., Power Apparatus and Systems, 78, 1272-1284; 1959.

44. E.D. Farmer, M.J. Potton, "The Prediction of Load on a Power System", Proc. 3rd I.F.A.C. Congress, paper 21F; 1966.

45. K. Hara et al., "A Method for Planning Economic Unit Commitment and Maintenance of Thermal Power Stations", I.E.E.E. Trans. Power Apparatus and Systems, 85, 427-436; 1966.

46. R.H. Kerr et al., "Unit Commitment", op. cit., 417-421.

47. P.G. Lowery, "Generating Unit Commitment by Dynamic Programming", op. cit. 422-426.

48. J.S. Riordon, "Adaptive Ordering of Power Generation as a Markovian Decision Process", paper to be delivered at U.K.A.C. Conference, Bristol, April 1967.

49. V. Vitek, J. Josefus, "Statistical Methods in the Automation of the Operation and Control of Power Systems", Proc. 3rd I.F.A.C. Congress, paper 21G; 1966.