armed bandit problems with *independent* arms (devices) and *full* observation or information, various recent extensions in this framework are discussed by Weiss (1984), Varaiya *et al.* (1985), Antharam *et al.* (1986*a,b*), Katehakis & Veinott (1987), Glazebrook & Fay (1987, 1988) and Lai & Ying (1988). Further results, for example by Gittins, Glazebrook and Whittle, can be found in Dempster *et al.* (1982), but few works treat the case of independent arms with *partial* observation (information), whose full information equivalent involving the *a posteriori* distribution—of central interest in this book—has *dependent* arms (see, however, Glazebrook 1985).

Several of the works cited above note that the multi-armed bandit problem is a version of the *single machine* stochastic *scheduling* problem in which arms are *jobs* and the resource to be allocated at each moment of time is a *unit* of *processing capacity* of a machine (e.g. a computer)—often with a discounted cost criterion. In this context, successes correspond to *job completions*, sharable resources to *time-sharing*, dynamic allocation index policies to stochastic versions of *Smith's rule* and Whittle's "arm-acquiring bandits" to problems with a *job input* process for the machine. Thus natural extensions of multi-bandit problems are *multi-processor* stochastic *scheduling* problems in which jobs (possibly arriving in a generalized Poisson stream) may be processed on any one of several identical or similar machines (see Dempster *et al.* 1982). Following the original works of Sevcik (1974) and Bruno (1976), more recent contributions on these problems with *independent* job processing times are given by Glazebrook & Nash (1976), Weiss & Pinedo (1980), Weber (1982), Weber *et al.* (1986), Rademacher (1986) and Kämpke (1987). *Dependent* job processing times are treated in Gittins & Glazebrook (1977), Möhring *et al.* (1984*a,b*) and Möhring & Rademacher (1985, 1989).

Multi-processor stochastic scheduling problems are special cases of a general class of controlled Markov processes termed *piecewise deterministic processes* (PDPs) whose trajectories are generated by ordinary dynamical systems punctuated by random jumps (see Davis 1984*a*, Dempster & Solel 1987). These processes model almost all stochastic systems not involving diffusions (Vermes 1980, Davis 1984*a*). They are related to several classes of similar processes (see e.g. Yushkevich 1987), but arose in the context of capacity expansion (Davis *et al.* 1987). Gittins' concept of "superprocesses" (for a unit resource) allows control actions which affect rewards and transition measures and hence utilize more of the general features of PDPs. An extensive optimal control theory for PDPs is now available in Vermes (1985), Lenhardt & Liao (1985), Davis (1986), Soner (1986*a,b*), Costa & Davis (1987, 1988), Gatarek (1988*a,b*), Dempster (1989) and Dempster & Ye (1989*a–d*). Nevertheless, a number of interesting open problems remain. One of these (*cf.* Chapter 4 of this book) concerns the relation of the interjump to the local version of the Bellman(–Hamilton–Jacobi) optimality equation for these processes. Another concerns the exact relation between this equation and the maximum principle (*cf.* Chapter 6).

Some general references on stochastic control, mainly with complete information, are Kushner (1971), Bertsekas & Shreve (1978), Whittle (1982, 1983), Ross (1983), Stengel (1983) and Bertsekas (1987). Partial information is discussed in Davis (1984*b*), Davis & Vinter (1984) and Kumar & Varaiya (1986). Related concepts in the economics of information are treated in Zellner (1980), Boyer & Kihlstrom (1984) and Easley & Kiefer (1988). Klimov (1974, 1978), Stone (1975), Heynman & Sobel (1982, 1983), Whittle (1986) and Kämpke (1987) contain applications to computer science and operations research.

We take this opportunity to acknowledge the support of this project by the Natural Science and Engineering Research Council of Canada, Balliol College, Oxford, National Research Council of Italy (GNAFA) and the Steklov Mathematical Institute, Academy of Sciences of the USSR. The final editing during the 1988–89 academic year was made easier by the kind hospitality of the Department of Statistics, Probability and Applied Statistics, University of Rome "La Sapienza" and the Department of Mathematics, University of Pisa.

We are once again grateful to Leslie Stockhausen for her cheerful and meticulous typesetting of this volume with the TEX system running on the MicroVAX II/VMS computer Earth of the School of Business Administration, Dalhousie University.

A number of people have provided encouragement, support, suggestions, references, etc. In particular, we wish to thank Gianni Andreatta, Vadim Arkin, Mark Davis, Robert Erdahl, Nicholas Frey, Franco Giannessi, John Gittins, Georgio Koch, Rolf Möhring, Giorgio

Picci, Michael Pinedo, Alfredo Rizzi, Wolfgang Rungaldier, Gabriella Salinetti, Albert Shiryaev, Pravin Varaiya and Gideon Weiss.

Halifax, Nova Scotia E.A.M-D.
October 1989 M.A.H.D.

INTRODUCTION

Decision making in many areas of human activity possesses two important features: it occurs under conditions of *incomplete knowledge* and is implemented in *steps*. These two features are connected with each other. On the one hand, the continuous flow of events in time, and the nonnegligible period required for effective decision making in a complex environment, resign us as a rule to the fact that it is impossible to predict future events exactly. On the other hand, when strong uncertainty exists with respect to future events, the tendency to minimize possible losses resulting from wrong predictions necessitates the division of the problem solution into only a few steps, the introduction of preliminary tests and experiments, followed by initial decisions, and so on.

These two features are connected with the *dualistic* character of decision making: control at each step must use the information obtained during the evolution of the process to date, and the nature of this information may depend essentially on the type of control applied.

Attempts at abstraction of such situations in order to find the correspondingly "optimal" behaviour rules has led to the creation of various different mathematical models of sequential control with incomplete information. Initially, the main source of these models lay in practical problems of optimal control arising in engineering. However, in recent decades such models have appeared in medicine and biology, and have also become an important part of economic and management research. With regard to managerial economics in recent times, this phenomenon is due to the widespread application of goal-oriented, formalized planning techniques, incorporating scientific and technological progress in planning and control and the resulting increasing sophistication.

The mathematical models and methods of sequential control with incomplete information are part of the general theory of *optimal control.* The cornerstones of the deterministic part of this theory for dynamical processes are the general concepts of *dynamic programming* and the *Pontryagin maximum principle* (see Bellman 1960; Pontryagin *et al.* 1961; Boltyanski 1969).

In developing the stochastic part of this theory, a seminal rôle belongs to the work of Wald (1967) and other statistical authors in developing *sequential analysis*, which emphasizes consideration of the future evolution of a sequence of statistical tests in order to create a general theory of *statistical decisions.* The various extensions of this theory and related ideas under such titles as *stochastic dynamic programming* and *controlled stochastic* (*random*) *processes* are represented in the works of Howard (1960), Blackwell (1965), Strauch (1969), Dynkin & Yushkevitch (1976), Raiffa & Schlaifer (1960), Gikhman & Skorohod (1977) and others. Relatively recently, mathematical theories related to the same class of problems were developed, such as the *control of diffusion processes* (see Krylov 1977), the theory of *optimal stopping for Markov chains, statistical inference for random processes* (see Shiryaev 1976; Lipster & Shiryaev 1974) and the *stochastic maximum principle* (Arkin & Evstigneev 1979). A vast literature devoted to the *applied* aspects of sequential control with incomplete information exists; we mention only Krasovski (1968), Yudin (1974), Kurzanski (1977), Chernousko & Kolmanovski (1978).

Adaptive control theory, which concerns the problem of finding the control which is in some sense optimal for a whole *class* of controlled objects, is considered in Sragovich (1981). Stochastic analogues of the classical models of *economic dynamics* are studied in Arkin & Evstigneev (1979).

A general outline of the construction of a multistage controlled process may be presented in the following way. The *state* of the system (object) is described by a point in some *state space.* At each successive moment of time, a *controlling action* (*control*) belonging to some set of admissible controls must be chosen. Depending on the chosen control and the current state—more generally, depending on *all* past controls and states—the system moves to a new state. If this dependency is deterministic, we have a *controlled deterministic process.* If, on the other hand, the chosen control and the history of the system to date define only a probability distribution for the new state, we deal with a *controlled stochastic process.* If, further, only *partial information* is available about the transition law (to the new state) or about the state of the system itself, then the problem is one of control with *incomplete information.* This case, indeed, is the one of interest to us in this book.

In statistical decision theory, two main approaches can be differentiated: the *minimax* approach, in which the quality of a strategy is measured in terms of the *worst* possible value(s) of the unknown parameter with respect to a utility function, and the *Bayesian* approach, in which some *a priori* distribution is prescribed for the unknown parameter and the value of a utility function weighted with respect to this distribution is maximized (or minimized).

In this book, attention is mainly given to the relatively narrow class of optimal control problems with incomplete information in which: firstly, it is supposed that at each successive moment of time decisions are chosen from a *finite* number of controlling actions (controls) $a^1, \ldots, a^m$ or their *mixtures*; secondly, the result of the choice of a control a^j is the observation of a random real number (in many cases, a *Bernoulli number*, i.e. having two values 0 or 1, representing respectively *failure* or *success*) whose distribution depends exclusively on the chosen control; and thirdly, we have a finite number of hypotheses $H_1, ..., H_N$ regarding (parameters of) the distribution function of the observations over which there exists a known *a priori* distribution

$$\xi := (\xi_1, \ldots, \xi_N), \qquad \xi_i \geq 0, \qquad \sum_{i=1}^{N} \xi_i = 1.$$

We consider also the case of continuous time, under the requirements that a finite number of controls are available and a finite number of hypotheses are considered: only the *times* of successful realizations are observed, and the choice of a control (or mixture of controls) defines the *intensity* of realizations of subsequent successes.

The content of the book is thus somewhat narrower than its title. In this regard, however, the following points should be noted. Firstly, the problems treated here present all the principal difficulties appearing in more general problems of stochastic control with incomplete information. Moreover, these difficulties are virtually impossible to overcome in general if one is not to be excessively restricted to approximate and heuristic methods. Secondly, in spite of the fact that formally the problems treated here belong to the theory of dynamic programming and of the control of Markov chains, finding optimal strategies already presents great difficulties of a fundamental character in simple examples. These difficulties can easily be seen in the following simplest model of the type treated in this book. Let the number of controls and hypotheses both equal two and suppose that Bernoulli random variables are observed. According to the first hypothesis, H_1, a success is observed with probability λ^1 if the first control is used and probability λ^2 ($\lambda^1 < \lambda^2$) if the second control is used (a failure is observed with the *complementary* probabilities). Under the second hypothesis, H_2, the probability λ^1 corresponds to use of the second control and λ^2 to use of the first. Further, a number ξ ($0 \leq \xi \leq 1$) is given which represents the *a priori* probability of H_1, so that that of H_2 is equal to $1 - \xi$. The aim of the *decision maker* (or *statistician*) is to maximize the expected number of successes over a fixed number of observations.

This problem is presented in many books (see for example De Groot 1970; Yakowitz 1969; Dynkin & Yushkevitch 1976; Prohorov & Rozanov 1973) and in the literature it is called the *two-armed bandit* problem, by analogy with playing an automat (slot machine) with two arms. The first problems of this type were considered in Thompson (1933), Robbins (1952), Brandt *et al.* (1956), Bellman (1956). The case of an arbitrary (finite) number of controls is called the *multi-armed bandit* problem.

As in more general situations, it can be shown for the two-armed bandit problem that in order to make an optimal decision at each moment of time it suffices to know only the *number* of observations remaining and the *a posteriori* probability of the first alternative hypothesis H_1 calculated in terms of the realizations of previous observations. The optimal strategy for this problem is as follows. If the *a posteriori* probability of this first hypothesis (according to which the second control is more profitable) is more than $\frac{1}{2}$, then, independent of the number of observations remaining, it is optimal to use the second control. If this probability is less than $\frac{1}{2}$, then it is optimal to use the first control.

In spite of the seeming obviousness of this fact, its rigorous proof was obtained only in the early 1960s by the American mathematician Feldman (1962). In fact by using different controls at a given time one not only obtains different benefits, but also causes the transition to different information states at the next moment. Therefore, it is possible that control values giving less than the greatest benefit at the current time might nevertheless be useful because they provide good *discrimination* between hypotheses and this permits more effective control in the future. This is the situation in the case—unlike the one just described—when the matrix of success probabilities is *nonsymmetric.*

In spite of our perception that such a survey is necessary, we have not attempted the task of giving a complete review of the literature relating to the stochastic control problem with incomplete information as described above. The authors' main results—both published and previously unpublished—are included in this book. Special attention has been given to the continuous time case, since it allows a considerable advance in the solution of some problems and exhibits interesting effects which are absent in the discrete time case, such as, for example, the appearance of a *turnpike.*

Our main aims in the present book are, on the one hand, to give the solutions of some problems with the general structure described above and to present methods specially developed for them, and, on the other, to demonstrate the application of the general methods of stochastic control theory with incomplete information using this relatively narrow class of problems as an example.

It seems to us that the new approaches and methods used in this book may find application to more general problems than those considered here—particularly to problems of the control of *pure jump* stochastic processes (see the details in Chapters 1 and 6).

Readers with minimal mathematical background who want to gain a general idea of the nature of this extensive topic can restrict themselves to reading the first chapter, where the main results and methods are presented at the heuristic level, but in some detail; they are rigorously established and applied in the following chapters.

In Chapter 1, a description is also presented of some economic situations falling into the scheme of sequential stochastic control with incomplete information.

Chapters 2 and 3 are devoted to the discrete time case. Precise definitions of problems and details of their general solutions are given. The presentation here has a technical character, and it is assumed that the reader is familiar with the main concepts of probability and decision theory—specifically with the Bayesian approach. The reader familiar with the contents of De Groot's (1970) book *Optimal Statistical Decisions* is completely prepared for reading these chapters.

Problems in continuous time are considered in Chapters 4 and 5. The rigorous formulation of such problems requires rather advanced techniques of modern stochastic process theory (*martingales, point processes*, etc.). To facilitate reading these chapters, we provide an Appendix in which the necessary facts and results are precisely stated without proof. Where proofs are not given herein, references are given to statements and proofs in other books.

The approach connected with the application of the *Pontryagin maximum principle* is considered in Chapter 6, where it is assumed that the reader is familiar with its general formulation. The content of this chapter consists partly in the formulation of some open problems.

In Chapter 7, such questions as discrimination amongst hypotheses and non-Bayesian formulations of problems are considered, results of other authors are presented, and some unsolved problems are discussed.

E. L. Presman wrote §§3.5, 4.3–4.5, 5.2, 5.3, 7.2 and 7.4. I. M. Sonin wrote §§3.1–3.3, 5.1, 6.4–6.6 and 7.3. The rest of the book was written together.

In conclusion, the authors would like to express their gratitude to V. I. Arkin, who brought to their attention a large number of questions and formulated the two-armed bandit problem in continuous time, to Yu. M. Kabanov, for numerous useful comments, and to all our colleagues in the Laboratory for Stochastic Problems in the Control of Economic Processes, Central Economic Mathematical Institute, Academy of Sciences of the U.S.S.R., for helpful discussions.

SOME NOTATION

All vectors are considered to be row vectors of the form $x = (x_1, \ldots, x_n)$.

Scalars, vectors and random variables and vectors are not distinguished notationally.

$*$ denotes matrix (vector) transpose.

diag x denotes the diagonal matrix formed by placing the elements of the vector x on the diagonal of a square matrix of zeroes of the same dimension.

S^n denotes the $(n-1)$-dimensional simplex

$$S^n := \left\{ x : x = (x_1, \ldots, x_n),\ x_i \geq 0,\ \sum_{i=1}^{n} x_i = 1 \right\}$$

$$e_i^n := (0, \ldots, \underset{i}{1}, \ldots, 0) \qquad e_0^n := (0, \ldots, 0)$$

$$\widehat{S}^n := \{e_i^n,\ i = 1, \ldots, n\}.$$

$\Lambda := \{\lambda_i^j\}$ denotes the matrix of success probabilities under hypotheses $i = 1, \ldots, N$ for control actions $j = 1, \ldots, m$.

$\xi = (\xi_1, \ldots, \xi_N)$ denotes the vector of *a posteriori* probabilities of hypotheses (relative likelihoods).

$\tilde{\eta}(\xi)$ denotes the change of variables for the *a posteriori* probabilities which represent the transformation to logarithmic relative likelihoods $\eta_i := \ln(\xi_i/\xi_N)$, $i = 1, \ldots, N-1$, $\eta_N := 0$.

$\tilde{\xi}(\eta)$ denotes the transformation inverse to $\tilde{\eta}(\xi), \xi_i := \exp \eta_i / \sum_{i=1}^{N} \exp \eta_k$.

$\tilde{\Gamma}^j := \tilde{\Gamma}^{1j}$ $(\tilde{\Gamma}^{0j})$ denotes the transformation of the *a posteriori* (logarithmic) probabilities upon the observation of a success (failure) on the j^{th} device.

$$\tilde{\Gamma}^{1j}\eta := \eta + \gamma^{1j} \qquad \tilde{\Gamma}^{0j}\eta := \eta + \gamma^{0j}$$

$$\gamma^{1j} := (\gamma_1^{1j}, \ldots, \gamma_N^{1j}) \qquad \gamma_i^{1j} := \ln(\lambda_i^j / \lambda_N^j)$$

$$\gamma^{0j} := (\gamma_1^{0j}, \ldots, \gamma_N^{0j}) \qquad \gamma_i^{0j} := \ln[(1 - \lambda_i^j)/(1 - \lambda_N^j)].$$

$\{A\}$ denotes a random event and $P\{A\}$ denotes the probability of this event.

$E(\cdot)$ denotes mathematical expectation.

$I(D)(x)$ or simply $I(D)$ denotes the indicator function of the set D, i.e. the function which is 1 for $x \in D$ and is 0 outside D.

$V_u^{\beta}(\xi)$, $W_u^{\beta}(\xi)$ $F_u^{\pi}(\xi)$ denote the value functions corresponding to the time interval $[0, \varkappa)$ using action rule β (strategy π) and starting point ξ.

For the 2×2 square matrix $\{\lambda_i^j\}$:

$$\delta^j := \lambda_1^j - \lambda_2^j \qquad \varepsilon_i := \lambda_i^1 - \lambda_i^2$$

$$\varepsilon := \varepsilon_1 - \varepsilon_2 = \delta^1 - \delta^2$$

$$\gamma^{1j} := \ln(\lambda_1^j / \lambda_2^j) \qquad \gamma^{0j} := \ln[(1 - \lambda_1^j)/(1 - \lambda_2^j)] \qquad j = 1, 2.$$

To our fathers
Lev A. Presman and Mikhail Ya. Sonin,
who will never see this book.