Monographs on Statistics and Applied Probability

Donald A. Berry
Bert Fristedt *Editors*

Bandit Problems: Sequential Allocation of Experiments (Monographs on Statistics and Applied Probability)

MONOGRAPHS ON
STATISTICS AND APPLIED PROBABILITY

General Editors

D. R. Cox, D. V. Hinkley, D. Rubin and B. W. Silverman

Probability, Statistics and Time
M. S. Bartlett

The Statistical Analysis of Spatial Pattern
M. S. Bartlett

Stochastic Population Models in Ecology and Epidemiology
M. S. Bartlett

Risk Theory
R. E. Beard, T. Pentikäinen and E. Pesonen

Residuals and Influence in Regression
R. D. Cook and S. Weisberg

Point Processes
D. R. Cox and V. Isham

Analysis of Binary Data
D. R. Cox

The Statistical Analysis of Series of Events
D. R. Cox and P. A. W. Lewis

Analysis of Survival Data
D. R. Cox and D. Oakes

Queues
D. R. Cox and W. L. Smith

Stochastic Modelling and Control
M. H. A. Davis and R. Vinter

Stochastic Abundance Models
S. Engen

The Analysis of Contingency Tables
B. S. Everitt

An Introduction to Latent Variable Models
B. S. Everitt

Finite Mixture Distributions
B. S. Everitt and D. J. Hand

Population Genetics
W. J. Ewens

Classification
A. D. Gordon

Monte Carlo Methods
J. M. Hammersley and D. C. Handscomb

Identification of Outliers
D. M. Hawkins

Generalized Linear Models
P. McCullagh and J. A. Nelder

Distribution-free Statistical Methods
J. S. Maritz

Multivariate Analysis in Behavioural Research
A. E. Maxwell

Applications of Queueing Theory
G. F. Newell

Some Basic Theory for Statistical Inference
E. J. G. Pitman

Statistical Inference
S. D. Silvey

Models in Regression and Related Topics
P. Sprent

Sequential Methods in Statistics
G. B. Wetherill

An Introduction to Latent Variable Models
B. S. Everitt

(Full details concerning this series are available from the Publishers)

Bandit problems

SEQUENTIAL ALLOCATION OF EXPERIMENTS

DONALD A. BERRY
BERT FRISTEDT

Department of Theoretical Statistics
School of Mathematics
University of Minnesota

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

ISBN 978-94-015-3713-1

British Library Cataloguing in Publication Data

Berry, Donald A.
Bandit problems: sequential allocation of experiments. —·· (Monographs on statistics and applied probability)
1. Mathematical statistics
I. Title II. Fristedt, Bert III. Series
519.5 QA276
ISBN 978-94-015-3713-1 ISBN 978-94-015-3711-7 (eBook)
DOI 10.1007/978-94-015-3711-7

Library of Congress Cataloging in Publication Data

Berry, Donald A.
Bandit problems.
(Monographs on statistics and applied probability)
Bibliography: p.
Includes index.
1. Experimental design. I. Fristedt, Bert, 1937– II. Title. III. Series.
QA279, B47 1985 001.4'34 85-9696
ISBN 978-94-015-3713-1

Contents

Preface

Our purpose in writing this monograph is to give a comprehensive treatment of the subject. We define bandit problems and give the necessary foundations in Chapter 2. Many of the important results that have appeared in the literature are presented in later chapters; these are interspersed with new results. We give proofs unless they are very easy or the result is not used in the sequel. We have simplified a number of arguments so many of the proofs given tend to be conceptual rather than calculational. All results given have been incorporated into our style and notation.

The exposition is aimed at a variety of types of readers. Bandit problems and the associated mathematical and technical issues are developed from first principles. Since we have tried to be comprehensive the mathematical level is sometimes advanced; for example, we use measure-theoretic notions freely in Chapter 2. But the mathematically uninitiated reader can easily sidestep such discussion when it occurs in Chapter 2 and elsewhere. We have tried to appeal to graduate students and professionals in engineering, biometry, economics, management science, and operations research, as well as those in mathematics and statistics. The monograph could serve as a reference for professionals or as a text in a semester or year-long graduate level course.

A uniform treatment of the numerous papers that deal with bandit problems is not possible. We have tried to compensate with the inclusion of an Annotated Bibliography. In it we list about 200 papers and books that have appeared on the subject and comment on those available to us. We say what problems the authors address, indicate when we think the paper contains mistakes, and occasionally criticize the approach taken.

The individual chapters stand as much on their own as seems reasonable, but all readers should learn our notation as given in

Section 2.2. Chapter 4 should be read before Chapters 5, 6 or 7. Chapters 5 and 6 form a natural pair. Chapter 3 should be read or skimmed early, and either of Chapters 8 or 9 can be read without having learned the content of the earlier chapters.

Results of other authors are appropriately credited, so those given without reference are new. As an example of the latter, in Chapter 6 we prove a converse of the famous Gittins–Jones theorem. Most of the examples (of which there are many) and most of the tables and figures are new. The tables and figures given in Section 5.4 were made with the programming assistance of K. Samaranayake. The tables in Section 5.6 are due to M. K. Clayton. Information for Figures 8.2 and 8.3 was supplied by A. J. Petkau.

We have benefited greatly from conversations with many people concerning the preparation of this monograph. These include J. A. Bather, M. K. Clayton, D. C. Heath, N. C. Jain, R. E. McCulloch, S. Orey, A. J. Petkau, and M. Schäl. The contributions of S. G. Eick and W. D. Sudderth in this regard have been especially numerous and deep. We thank T. S. Ferguson for commenting on the manuscript. We also thank our typists, P. Linman and A. M. Ruggles, for their diligence through our numerous revisions. Finally, we thank our wives, Donna and Shirin, for their patience and support.

Minneapolis, Minnesota D. A. Berry
October 1984 B. Fristedt

CHAPTER 1

Introduction

Suppose two treatments are available for a certain disease. Patients arrive at a clinic one at a time and one of the treatments must be used on each. Information as to the effectiveness of the treatments accrues as they are used. The overall objective is to treat as many patients as effectively as possible. This seemingly innocent but important problem is surprisingly difficult, even when the responses are dichotomous, either success or failure. It is an example of a two-armed bandit problem.

A bandit problem in statistical decision theory involves sequential selections from $k \geqslant 2$ stochastic processes (or 'arms', machines, treatments, etc.). Time may be discrete or continuous and the processes themselves may be discrete or continuous. The processes are characterized by parameters which are typically unknown. The process selected for observation at any time depends on the previous selections and results. A decision procedure (or strategy) specifies which process to select at any time for every history of previous selections and observations. A utility is defined on the space of all histories. This provides a definition for the utility of a strategy in the usual way, by averaging over all possible histories resulting from the strategy.

Most of the literature, and most of this monograph, deals with discrete time. In such a setting, each of the k arms generates an infinite sequence of random variables. An observation on a particular sequence is made by selecting the corresponding arm. The mth member of a sequence is observed if the corresponding arm is 'selected' at stage m. The classical objective in bandit problems is to maximize the expected value of the payoff, $\Sigma_1^\infty \alpha_m Z_m$, where Z_m is the variable observed at stage m and the α_m are nonnegative numbers. Though our approach will be somewhat more general (see Section 3.2), the *discount factors* α_m are usually assumed to be known with $0 < \Sigma_1^\infty \alpha_m < \infty$ and,

sometimes, $\alpha_m \geqslant \alpha_{m+1}$. $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$ is called a *discount sequence.* A strategy is *optimal* if it yields the maximal expected payoff. An arm is said to be *optimal* if it is the first selection when following some optimal strategy.

The discount sequences which are most frequently considered in the literature are:

(i) finite horizon uniform: $(1, \ldots, 1, 0, \ldots)$,
(ii) geometric: $(1, \alpha, \alpha^2, \ldots)$, $0 < \alpha < 1$.

For (i) the objective is to maximize the sum of the first n observations (where n is the horizon). If the mth observation is made at time m and is paid in inflated monetary units, then (ii) may be appropriate where $\alpha^{-1} - 1$ is the inflation rate. Further discussion of discount sequences and some motivation for general discounting are given in Chapter 3.

There are two benefits derived from selecting an arm: (1) immediate payoff, and (2) information that can make for better later selections and greater future payoff. When an arm is selected, available information concerning the arms is modified and the discount sequence changes: $(\alpha_1, \alpha_2, \alpha_3, \ldots)$ becomes $(\alpha_2, \alpha_3, \ldots)$. For example, in (i) above, the horizon is decreased by 1. In (ii) above, the new sequence is proportional to the original sequence. The decision problem is unchanged if the discount sequence is multiplied by a constant, so for geometric discounting (and only for this case) the discount sequence is effectively the same throughout the trial. This characteristic can make bandit problems with geometric discounting more tractable than other problems (see Chapter 6).

1.1 Bayesian and other approaches

Consider the Bernoulli case for ease of discussion. Given a discount sequence **A**, the utility of a particular strategy can be calculated as a function of **A** and of the Bernoulli parameters θ_1 and θ_2. It is generally too much to expect that a strategy will exist that is best for all pairs (θ_1, θ_2).

The vast majority of the bandit literature takes one of two approaches. In the Bayesian approach, the utility of a strategy is averaged over (θ_1, θ_2) with respect to some measure; papers by Bradt, Johnson, and Karlin (1956), Bellman (1956), and Feldman (1962) are early examples. This measure represents information that is present

about the various processes separate from the current experiment. Many adherents to the Bayesian approach regard this measure as being subjective (Savage, 1954; Barnett, 1982) and quantifying the knowledge of the experimenter concerning the two arms.

The second approach taken in the literature is to consider particular strategies and compare their utilities as a function of (θ_1, θ_2). Papers by Robbins (1952) and Isbell (1959) are early examples. When the utility of one strategy uniformly dominates that of the others then, of course, it is best in the class of strategies under consideration. When one is not uniformly best, the various strategies can be compared using tables (e.g. Wahrenberger, Antle, and Klimko, 1977).

A third alternative is the minimax approach in which nature is regarded as an opponent in a two-person, zero-sum game (e.g. Vogel, 1960a, b). Nature chooses (θ_1, θ_2) in the unit square, or in a subset of it, according to some *a priori* restriction. The decision maker's goal is to minimize the expected difference between what is achieved and what could be achieved were (θ_1, θ_2) known. Nature's goal is to maximize this expected difference. The minimax approach is discussed in Chapter 9.

Thompson (1933, 1935) posed the first bandit problem. He considered two Bernoulli processes, uniform discounting ($\alpha_1 = \ldots = \alpha_n = 1, \alpha_{n+1} = \ldots = 0$), and took a Bayesian point of view. In this setting the objective is to maximize the expected number of successes in the first n trials. Thompson regarded the two processes as independent with their parameters having beta distributions (cf. Chapter 7).

After Thompson (1933, 1935), the bandit problem received little attention until it was studied by Robbins (1952, 1956), who also considered two arms but took the second approach mentioned above. Robbins (1952) suggested a selection strategy that depends on the history only through the last selection and the result of that selection; namely, the same arm is selected after a success and the other after a failure. Robbins's objective was to maximize the long-run proportion of successes. He showed that the 'stay on a winner, switch on a loser' strategy uniformly dominates random selection. This originated an approach called 'finite memory': the decision maker's choice at any stage can depend only on the selections and results in the previous r stages; Isbell (1959) and Smith and Pyke (1965) are examples of this approach.

Bradt, Johnson, and Karlin (1956) took a Bayesian approach for the

finite horizon, uniform case. They characterized optimal strategies when one parameter is known *a priori*.

With a Bayesian approach, a strategy requires (and 'remembers') only the sufficient statistics: the numbers of successes and failures on the two arms. Much of the recent bandit literature – and most of this monograph – takes the Bayesian approach. It is not that researchers in bandit problems tend to be 'Bayesians'; rather, Bayes's theorem provides a convenient mathematical formalism that allows for adaptive learning, and so is an ideal tool in sequential decision problems.

With a Bayesian approach, a bandit is a typical problem in dynamic programming. When the horizon is finite ($\alpha_{n+1} = \alpha_{n+2} = \ldots = 0$ for some n), backwards induction can be used to determine optimal strategies (cf. Sections 2.3 and 2.4). One first finds the maximal conditional expected payoff (together with the arm or arms that give it) at the very last stage for every possible $(n-1)$-history (sequence of selections and results), optimal and otherwise. Here, 'conditional' refers to the particular history. Proceeding to the penultimate stage, one maximizes the conditional expected payoff from the last two observations for every possible $(n-2)$-history. Continuing backwards, while remembering the optimal arms at each partial history, gives all optimal strategies. The problem is four-dimensional in the Bernoulli setting since that is the dimension of a minimal sufficient statistic. But a computer program requiring on the order of $n^3/6$ storage locations can be devised.

1.2 Myopic strategies

Feldman (1962) solved the Bernoulli two-armed bandit problem with uniform discounting for a deceptively difficult special case: both probabilities of success are known, but not which goes with which arm. Feldman showed that *myopic* strategies are optimal: at every stage, select the arm with greater expected immediate gain (the unconditional probability of success with arm j is the current mean of the Bernoulli parameter θ_j).

It is important to recognize that myopic strategies are not optimal – or even good – in general. The following is a case in point.

Example 1.2.1 Suppose **A** is uniform with horizon n. Assume θ_2 is known to be 1/2 (selecting arm 2 is like tossing a fair coin). And θ_1 is either 1 or 0 (the other coin is either two-headed or two-tailed); let r be

the initial probability that $\theta_1 = 1$ so r is the prior mean of θ_1. The fact that a single selection of arm 1 reveals complete information makes the analysis of this problem rather easy. If $r < 1/2$ then a myopic strategy indicates selections of arm 2 indefinitely, and has utility $n/2$. On the other hand, selecting arm 1 initially and then indefinitely if it is successful and never again if it is not, results in n successes with probability r and an average of $(n-1)/2$ successes with probability $1-r$. The advantage of this strategy over the myopic is

$$rn + (1-r)(n-1)/2 - n/2 = [r(n+1) - 1]/2,$$

which is positive for $r > 1/(n+1)$.

All we have shown is that the indicated strategy is better than the myopic when $r > 1/(n+1)$, but, as a consequence of Theorem 5.2.2, it is optimal. □

In this example, and in bandit problems generally, it may be wise to sacrifice some potential early payoff for the prospect of gaining information that will allow for more informed choices later. This aspect prompted Whittle (1982, p. 210) to claim that a bandit problem 'embodies in essential form a conflict evident in all human action'. The 'information versus immediate payoff' question makes the general problem difficult; the issue is seldom as clear as it is in Example 1.2.1.

In the clinical trial setting, sacrificing early payoff for information means that patients arriving later are likely to be treated better because they benefit from the responses of early patients (cf. Corollary 5.2.3). This also characterizes most designs actually used in clinical trials – but in the extreme. Patients are assigned treatments randomly in a clinical trial to gain information about the various treatments; the accumulating information is seldom used in treatment assignment during the trial. After the trial, patients can be assigned the treatment found to be most effective. While a sequential design with the exclusive purpose of gathering information is a special case of our approach (see Section 3.6), our general framework allows for weighting of future and present patients. This is done by appropriate choice of the discount sequence. An advantage of the geometric discount sequence in this regard is that it is 'democratic': the current patient is always weighted the same when compared with future patients.

1.3 Preview

Chapter 2 lays various foundations for the rest of the monograph. Firstly, it introduces the basic ideas of bandit problems to the uninitiated. In particular, Sections 2.1 and 2.4 contain many easy examples which embody the flavour of bandit problems. Secondly, the chapter constructs a technical framework for the chapters that follow. The fundamental equation of dynamic programming is developed and optimal strategies are shown to exist (Sections 2.3 and 2.5). Readers are urged not to get bogged down in the technical considerations in Sections 2.2, 2.3, and 2.5 at the expense of the various interesting aspects of bandits appearing in the rest of the chapter and in later chapters. We have written the other chapters to stand on their own as much as possible; in particular, only occasional reference to Chapter 2 is made.

Chapter 3 deals with the role of the discount sequence. Our approach to discounting is more general than that of other authors. A purpose of Chapter 3 is to motivate our approach, and to explain its applicability. In so doing we present a theory of random discounting. We also describe a setting in which the stages of a bandit problem occur in real time, and randomly. This setting will be discussed again in Chapters 5 and 8.

Most of this monograph treats independent arms. Chapter 4 develops some basic results for the case in which the arms are independent and also Bernoulli. Many of these results are applied in Chapters 5, 6, and 7.

Chapter 5 deals with the special case of two arms when the characteristics of one arm are completely known. Such bandits have been treated as stopping problems in the literature, and they are correctly regarded as stopping problems for some discount sequences. We characterize such sequences and give a variety of results for the case in which the problem is to decide when to stop selecting the unknown arm. For ease of presentation, in Sections 5.1 to 5.4 we assume that the unknown arm is Bernoulli. This assumption is dropped in Section 5.5 and a rather comprehensive setting is given in Section 5.6. In view of the main result in Chapter 6, the results of Chapter 5 take on new meaning when the discount sequence is geometric.

In Chapter 6 we give the celebrated result of Gittins and Jones (1974). This result shows that when the discounting is geometric, a

bandit problem involving k independent arms can be solved by solving k different two-armed bandits, each involving one known and one unknown arm. We prove the converse of this result by showing that a bandit involving two or more unknown arms can be solved in this way only when the discount sequence is geometric.

Chapter 7 treats the case of two independent Bernoulli arms when the horizon is finite and discounting is uniform. We give sufficient conditions for optimality, some depending on the horizon and some not.

In Chapter 8 we turn to the setting of continuous time in which both payoff and information accrue continuously. We treat Weiner processes and Lévy processes, mostly by example. We show, partly by example, the difficulties involved in treating continuous-time bandits. In particular, we argue that a satisfactory treatment of the foundations of continuous-time bandits is essentially an open problem.

Chapter 9 takes a different approach from that of the rest of the monograph. Instead of averaging over the unknown characteristics of the arms, the decision maker takes the worst-case point of view – a minimax approach.

References

Barnett, V. (1982) *Comparative Statistical Inference* (2nd edn), Wiley, New York.

Bellman, R. (1956) A problem in the sequential design of experiments. *Sankhyā A* **16**: 221–229.

Bradt, R. N., Johnson, S. M. and Karlin, S. (1956) On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.* **27**: 1060–1074.

Feldman, D. (1962) Contributions to the 'two-armed bandit' problem. *Ann. Math. Statist.* **33**: 847–856.

Gittins, J. C. and Jones, D. M. (1974) A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (eds J. Gani *et al.*), pp. 241–266, North-Holland, Amsterdam.

Isbell, J. R. (1959) On a problem of Robbins. *Ann. Math. Statist.* **30**: 606–610.

Robbins, H. (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58**: 527–535.

Robbins, H. (1956) A sequential decision problem with finite memory. *Proc. Nat. Acad. Sci. U.S.A.* **42**: 920–923.

Savage, L. J. (1954) *The Foundations of Statistics*, Wiley, New York.

Smith, C. V. and Pyke, R. (1965) The Robbins–Isbell two-armed bandit problem with finite memory. *Ann. Math. Statist.* **36**: 1375–1386.

Thompson, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**: 275–294.

Thompson, W. R. (1935) On the theory of apportionment. *Am. J. Math.* **57**: 450–456.

Vogel, W. (1960a) A sequential design for the two-armed bandit. *Ann. Math. Statist.* **31**: 430–443.

Vogel, W. (1960b) An asymptotic minimax theorem for the two-armed bandit problem. *Ann. Math. Statist.* **31**: 444–451.

Wahrenberger, D. L., Antle, C. E. and Klimko, L. A. (1977) Bayesian rules for the two-armed bandit problem. *Biometrika* **64**: 172–174.

Whittle, P. (1982) *Optimization Over Time: Dynamic Programming and Stochastic Control*, Vol. I, Wiley, New York.

CHAPTER 2

Notation and preliminaries

Bandit problems are described formally in this chapter. One purpose of the chapter is to develop some necessary notation. Another is to define strategies rigorously. It is shown that optimal strategies exist and that the 'fundamental equation of dynamic programming' is satisfied.

Some of the more important notions are developed in a simple example in Section 2.1. The general setting is presented in Section 2.2. In Section 2.3 and 2.5 the existence of optimal strategies is proved; the method of dynamic programming is used in the finite horizon case in Section 2.3. A variety of finite horizon examples is given in Section 2.4. In Section 2.6 an approximate method for the infinite horizon case is discussed and an illustrative example given.

Much of the literature that deals with bandit problems, and other stochastic decision problems, implicitly assumes that everything works the way one would like it to work. The reader will see that a careful development is quite technical. Some readers may want to skim the technical Sections 2.2, 2.3, and 2.5; others will skip them entirely.

2.1 An example: maximizing the sum of two observations

Suppose there are two opportunities to receive payoff and the objective is to maximize the sum of the two observations, so the discount sequence is $\mathbf{A} = (1, 1, 0, \ldots)$. At the first stage the decision maker has two choices: one is a random variable which may be regarded as being produced by a mechanism, arm 1, and the other is the known constant λ which, with future considerations in mind, we will regard as being produced by arm 2. Let Z_1 denote the observation at stage 1; so $Z_1 = \lambda$ in case the decision maker selects arm 2, and Z_1 is random in case arm 1 is chosen.

Taking into account the result and selection at stage 1, the decision

maker faces the same choice at stage 2. The setting is now different for two reasons. First, there is but one remaining observation. Second, there may now be different information available to the decision maker, depending on whether arm 1 was selected initially and on the resulting observation.

The notion of a strategy has been defined implicitly above. More formally, a *strategy* is a function that assigns to each (partial) history of observations the integer 1 or 2 indicating the arm to be observed at the next stage. Thus, a strategy τ assigns to the empty history the integer indicating the arm to be observed initially, and $\tau(z_1)$ indicates the arm to be observed at stage 2 when z_1 is observed at stage 1. We require

$$\{z_1 : \tau(z_1) = i\}$$

to be a Borel set for $i = 1$ and therefore also for $i = 2$; that is, we require the function $z_1 \mapsto \tau(z_1)$ to be Borel measurable. The decision maker's objective is to choose τ to maximize $E_\tau(Z_1 + Z_2)$, where Z_m indicates the observation at stage m and the subscript τ indicates the dependence on the strategy τ. The use of expectation implies the existence of an underlying probability structure; we now turn our attention to this structure.

The characteristics of arm 2 are known to the decision maker: whenever arm 2 is selected a known constant λ is observed. The characteristics of arm 1 are not completely known. Let X_m denote the outcome from arm 1 at stage m; it is useful for notational reasons to assume there is an outcome on arm 1 at each stage whether or not arm 1 is in fact selected.

We assume that the X_m are normally distributed random variables having mean θ and variance 1. The parameter θ is unknown and is assumed to have an initial distribution which is normal with mean μ and variance $\rho^2 > 0$, both of which are known. Conditional on θ, X_1 and X_2 are independent. The value of X_1 is known to the decision maker at stage 2 if and only if arm 1 was selected initially. Accordingly, the full strategy τ can be specified prior to stage 1, for a robot can carry out a decision maker's wishes by evaluating τ at the value of Z_1.

The *utility* or *worth* of a strategy τ is $E_\tau(Z_1 + Z_2)$. For this to be well-defined, Z_2 must be a random variable; that is, it must be measurable. That it is measurable follows from the measurability of $z_1 \mapsto \tau(z_1)$ and the identity

$$Z_2 = X_2 \mathbf{1}_{\{z_1 : \tau(z_1) = 1\}}(Z_1) + \lambda \mathbf{1}_{\{z_1 : \tau(z_1) = 2\}}(Z_1). \qquad (2.1.1)$$

We introduce a notation that will be adapted more generally. The worth of a strategy depends on the discount sequence and the available information concerning the arms. In the case at hand the information about arm 1 is indexed by (μ, ρ), arm 2 is specified by the constant λ, and the discount sequence is $\mathbf{A} = (\alpha_1, \alpha_2, \ldots) = (1, 1, 0, 0, 0, \ldots)$; the problem is called the $((\mu, \rho), \lambda; \mathbf{A})$-bandit. The worth of a strategy τ is

$$W((\mu, \rho), \lambda; \mathbf{A}; \tau) = E_\tau \sum_{m=1}^{\infty} \alpha_m Z_m$$
$$= E_\tau(Z_1 + Z_2).$$

The *value* of the $((\mu, \rho), \lambda; (1, 1, 0, \ldots))$-bandit is defined to be

$$V((\mu, \rho), \lambda; (1, 1, 0, \ldots)) = \sup_\tau W((\mu, \rho), \lambda; (1, 1, 0, \ldots); \tau).$$

Any τ for which the supremum is attained is an *optimal strategy*.

We shall calculate $V((\mu, \rho), \lambda; (1, 1, 0, \ldots))$ and find an optimal strategy. We begin by calculating a strategy (call it τ_2) that is best among those that begin with arm 2 at stage 1. Since

$$E(X_2) = E(X_1) = E(E(X_1|\theta)) = E(\theta) = \mu, \tag{2.1.2}$$

we can take

$$\tau_2(z_1) = \begin{cases} 2 \text{ if } \mu \leqslant \lambda \\ 1 \text{ if } \mu > \lambda. \end{cases}$$

Moreover, with '$\vee$' denoting maximum,

$$W((\mu, \rho), \lambda; (1, 1, 0, \ldots); \tau_2) = \lambda + (\lambda \vee \mu). \tag{2.1.3}$$

The next step will be to find a strategy, say τ_1, that is best among those that begin with arm 1. Accordingly, $Z_1 = X_1$ in (2.1.1) and

$$E(Z_2|X_1) = \mathbf{1}_{\{z_1:\, \tau_1(z_1) = 1\}}(X_1)E(X_2|X_1) + \mathbf{1}_{\{z_1:\, \tau_1(z_1) = 2\}}(X_1)\lambda.$$

This is maximized by τ_1 defined by $\tau_1(z_1) = 2$ if $E(X_2|X_1 = z_1) \leqslant \lambda$ and $\tau_1(z_1) = 1$ otherwise. A well-known formula (DeGroot 1970, p. 248) for the posterior mean of a normal distribution with a normally distributed mean gives

$$E(X_2|X_1) = \frac{\mu + \rho^2 X_1}{1 + \rho^2}.$$

So we can take

$$\tau_1(z_1) = \begin{cases} 2 \text{ if } & \dfrac{\mu + \rho^2 z_1}{1+\rho^2} \leqslant \lambda \\ 1 \text{ if } & \dfrac{\mu + \rho^2 z_1}{1+\rho^2} > \lambda, \end{cases}$$

and

$$W((\mu, \rho), \lambda; (1, 1, 0, \ldots); \tau_1) = \mu + E\left(\frac{\mu + \rho^2 X_1}{1+\rho^2} \vee \lambda\right). \quad (2.1.4)$$

From (2.1.3) and (2.1.4) we obtain

$$\begin{aligned} &V((\mu, \rho), \lambda; (1, 1, 0, \ldots)) \\ &\quad = W((\mu, \rho), \lambda; (1, 1, 0, \ldots); \tau_1) \vee W((\mu, \rho), \lambda; (1, 1, 0, \ldots); \tau_2) \\ &\quad = \left[\mu + E\left(\frac{\mu + \rho^2 X_1}{1+\rho^2} \vee \lambda\right)\right] \vee [\lambda + (\lambda \vee \mu)] \\ &\quad = \left[\mu + E\left(\frac{\mu + \rho^2 X_1}{1+\rho^2} \vee \lambda\right)\right] \vee 2\lambda \vee [\lambda + \mu] \\ &\quad = \left[\mu + E\left(\frac{\mu + \rho^2 X_1}{1+\rho^2} \vee \lambda\right)\right] \vee 2\lambda. \end{aligned}$$

Therefore, optimal strategies depend on the sign of

$$\mu - \lambda + E\left[\left(\frac{\mu + \rho^2 X_1}{1+\rho^2} - \lambda\right) \vee 0\right] = \frac{\rho^2}{\sqrt{(1+\rho^2)}}[\Psi(t) - t],$$

where $t = (\lambda - \mu)\rho^{-2}\sqrt{(1+\rho^2)}$, the transform

$$\begin{aligned} \Psi(t) &= \int_t^\infty (x - t)\,\varphi(x)\,dx \\ &= \varphi(t) - t[1 - \Phi(t)] \end{aligned}$$

(cf. DeGroot 1970, p. 247), and φ and Φ are the standard normal density and distribution functions, respectively. The function Ψ is positive, continuous, and strictly decreasing. So there is a unique t_0 satisfying $\Psi(t_0) = t_0$; numerically, $t_0 \approx 0.2760$.

If $t \geqslant t_0$ then τ is optimal if $\tau(\varnothing) = \tau(z_1) = 2$; that is, arm 2 is

optimal at both stages. If $t \leqslant t_0$ then τ is optimal if $\tau(\varnothing) = 1$ and

$$\tau(z_1) = \begin{cases} 1 \text{ for } z_1 \geqslant \lambda + (\lambda - \mu)/\rho^2 \\ 2 \text{ for } z_1 \leqslant \lambda + (\lambda - \mu)/\rho^2. \end{cases}$$

So arm 1 is optimal initially if $(\lambda - \mu)\rho^{-2}\sqrt{(1+\rho^2)} \leqslant t_0$ and it is optimal at stage 2 if its current mean is greater than λ.

The optimal initial selection has various properties which can aid in understanding bandit problems more generally. If the mean μ of an observation on the unknown arm is at least as large as the sure-thing λ, then the unknown arm is optimal. If, on the other hand, $\mu < \lambda$ then whether arm 1 is optimal depends on ρ. Since $\rho^{-2}\sqrt{(1+\rho^2)}$ decreases to 0 as $\rho \to \infty$, arm 1 is optimal for sufficiently large ρ for any fixed μ and λ! So if the characteristics of arm 1 are sufficiently uncertain (that is, if ρ is sufficiently large), then any expected loss will be tolerated on the initial selection for the possibility of a large gain on the second selection. (Many statisticians who take the Bayesian approach to inference recommend using improper priors and taking $\rho \to \infty$ for convenience in the case of normal sampling. This may not be grossly incorrect in a problem in which some sampling will occur in any case. But it can be very misleading and lead to inane strategies when the issue is to sample or not!)

One purpose of this example is to give a concrete setting for some of the abstract notions to be developed in later sections of this chapter. An important quantity in the example is $E(X_2|X_1)$. When the distribution of an arm is normal with unknown mean and known variance, its distribution conditional on observations from the arm is easy to calculate. This is true in large part because it has a single unknown one-dimensional parameter. No such simple parametrization exists in general.

In the general setting of the following section, an arm will be characterized by a probability measure F on the Borel field of subsets of $\mathscr{D}$, the space of probability distributions on $\mathbb{R}$ with the topology of convergence in distribution. Considering $F(\cdot|X_1)$ for a general F, there is no problem in defining, up to a set of probability 0, $F(\mathscr{C}|X_1)$ for each fixed Borel subset $\mathscr{C}$ of $\mathscr{D}$. But we also want, for each fixed ω in the underlying probability space, that the function

$$\mathscr{C} \mapsto F(\mathscr{C}|X_1)(\omega)$$

be a probability measure on $\mathscr{D}$. By a result of Parthasarathy (1967, Theorem V.8.1), we can have this and more. Namely, there exists a

function $f(\mathscr{C}, x)$ such that for each x the function $\mathscr{C} \mapsto f(\mathscr{C}, x)$ is a probability measure on $\mathscr{D}$, for each fixed $\mathscr{C}$ the function $x \mapsto f(\mathscr{C}, x)$ is a Borel measurable function on $\mathbb{R}$, and, almost surely,

$$F(\mathscr{C} \mid X_1)(\omega) = f(\mathscr{C}, X_1(\omega)). \tag{2.1.5}$$

In the example of this section, $f(\cdot, x)$ assigns probability one to the set of normal distributions having variance one; the mean of this distribution is itself normally distributed with mean $(\mu + \rho^2 x)/(1 + \rho^2)$ and variance $\rho^2/(1 + \rho^2)$.

2.2 General setting

The development in this section depends on the preceding section for some notation and concepts. However, the setting is general and calculations involving normal distributions will not play a role. There is an arbitrary finite number k of arms.

Recall from Section 2.1 that $\mathscr{D}$ denotes the space of probability distributions on $\mathbb{R}$. We use the topology of convergence in distribution on $\mathscr{D}$. Thus, with Q_l, $Q \in \mathscr{D}$ viewed as cumulative distribution functions, $Q_l \to Q$ as $l \to \infty$ if and only if $Q_l(x) \to Q(x)$ for every x at which Q is continuous. From the measure-theoretic viewpoint, $Q_l \to Q$ as $l \to \infty$ if and only if $Q_l(B) \to Q(B)$ for every Borel subset B of $\mathbb{R}$ for which $Q(\text{boundary of } B) = 0$, or equivalently, $\int_{\mathbb{R}} h \, dQ_l \to \int_{\mathbb{R}} h \, dQ$ as $l \to \infty$ for every bounded continuous function h on $\mathbb{R}$. It is shown in Parthasarathy (1967, Section II.6) that $\mathscr{D}$ inherits many properties from $\mathbb{R}$: $\mathscr{D}$ is separable, locally compact, metrizable to be complete, and its topology is determined by convergent sequences and the limits of such sequences. The Borel field of subsets of $\mathscr{D}$ is the smallest σ-field containing the open sets; it is this σ-field of subsets of $\mathscr{D}$ that is used throughout.

The space $\mathscr{D}^k$ of ordered k-tuples of members of $\mathscr{D}$ will be considered to have the product topology arising from the above-defined topology on $\mathscr{D}$. The Borel field generated by this product topology is the only σ-field of subsets of $\mathscr{D}^k$ that will be considered; it is the product σ-field of k copies of the Borel σ-field of $\mathscr{D}$. The component Q_i of $(Q_1, \ldots, Q_k) \in \mathscr{D}^k$ governs observations on arm i. Since $(Q_1, \ldots, Q_k)$ is random, the probability distribution G of $(Q_1, \ldots, Q_k)$ plays a central role.

The space $\mathscr{D}(\mathscr{D}^k)$ of probability distributions on $\mathscr{D}^k$ will therefore also play a central role. A member G of $\mathscr{D}(\mathscr{D}^k)$ represents the decision

maker's prior information concerning the k arms. The member of $\mathscr{D}(\mathscr{D}^2)$ that represents the prior information for the example of Section 2.1 is supported by a rather small subset of $\mathscr{D}^2$—namely (normal distributions with variance 1) × (a constant). We use the topology of convergence in distribution on $\mathscr{D}(\mathscr{D}^k)$. Thus, for members G_i and G of $\mathscr{D}(\mathscr{D}^k)$, $G_i \to G$ if and only if

$$\int_{\mathscr{D}^k} h \, dG_i \to \int_{\mathscr{D}^k} h \, dG$$

for every bounded continuous h on $\mathscr{D}^k$. According to Parthasarathy (1967, Section II.6) the space $\mathscr{D}(\mathscr{D}^k)$ inherits from $\mathscr{D}^k$ the properties of being separable, locally compact, metrizable to be complete, and having its topology determined by the convergent sequences and their limits. The σ-field of subsets of $\mathscr{D}(\mathscr{D}^k)$ that will be used is the Borel field.

We turn to the construction of a probability space Ω with a natural structure rich enough to reflect randomness in the structure of the arms as well as in the observations resulting from selecting arms according to an arbitrary strategy (yet to be defined). Let Ω be the product space obtained from $\mathscr{D}^k$ and infinitely many copies of the open unit interval, one for each pair (i, m), $1 \leqslant i \leqslant k$, $m = 1, 2, \ldots$; that is,

$$\Omega = \mathscr{D}^k \times \bigtimes_{i=1}^{k} \bigtimes_{m=1}^{\infty} (0, 1).$$

The probability measure P on Ω is the product of G on $\mathscr{D}^k$ and Lebesgue measure on each unit interval, and it is defined on the product σ-field $\mathscr{F}$ of the various Borel fields. A member of Ω can be written in the form

$$\omega = (Q_i, 1 \leqslant i \leqslant k; \omega_{im}, 1 \leqslant i \leqslant k, m = 1, 2, \ldots) \qquad (2.2.1)$$

where each $Q_i \in \mathscr{D}$ and each $\omega_{im} \in (0, 1)$. Denoting by Q_i^{-1} the right-continuous 'inverse' function of Q_i, regarded as a cumulative distribution function, we set

$$X_{im}(\omega) = Q_i^{-1}(\omega_{im})$$

for ω given by (2.2.1). We leave it to the reader to check that each X_{im} is measurable and thus a random variable and that, conditional on $(Q_1, \ldots, Q_k)$,

$$\{X_{im}: 1 \leqslant i \leqslant k, m = 1, 2, \ldots\}$$

is an independent family with the conditional distribution of X_{im} being Q_i. The value of X_{im} is the outcome on arm i at stage m (whether or not X_{im} is observed by the decision maker, as defined below).

A *strategy* τ assigns to each (partial) history of observations an integer from 1 to k indicating the arm to be selected at the next stage. Thus, $\tau(\varnothing)$ indicates the arm to be selected initially when following τ, $\tau(z_1)$ indicates the arm to be selected at the stage 2 given that z_1 is observed at stage 1, $\tau(z_1, z_2)$ is the arm to be selected at stage 3, etc. The value of τ indicates the arm to be selected at a particular stage. The outcome on that arm at that stage is an *observation*. Using all previous observations, the decision maker can evaluate τ for the next stage. For τ to be a strategy, we require that the set of observations for which arm i is indicated at stage m,

$$\{(z_1, \ldots, z_{m-1}): \tau(z_1, \ldots, z_{m-1}) = i\},$$

be a Borel subset of $\mathbb{R}^{m-1}$.

According to the description of a strategy in Chapter 1, the arm selected at any stage may depend on the history of observations *and* arms selected. Our notation is consistent with that definition but the dependence of the current selection on the prior selections is not as clear from the notation as is its dependence on the previous observations. To see that our notation is consistent with the earlier description, consider, for example, $\tau(z_1, z_2, z_3)$ for particular values of z_1, z_2, and z_3. Given τ, the arms associated with the three observations are determined: z_1 resulted from arm $\tau(\varnothing)$, z_2 from arm $\tau(z_1)$, and z_3 from arm $\tau(z_1, z_2)$.

The sequence of observed random variables is recursively defined via

$$Z_1 = X_{\tau(\varnothing),1},$$

$$Z_m = X_{\tau(Z_1, \ldots, Z_{m-1}),m}, \quad m > 1.$$

That each Z_m is in fact measurable (and thus a random variable) follows by induction.

Since the general problem will be to choose τ to maximize the expected value of

$$\sum_{m=1}^{\infty} \alpha_m Z_m,$$

it is natural for us to introduce the following requirements on G and the discount sequence $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$. We require that each $\alpha_m \geqslant 0$

and $\sum \alpha_m < \infty$. We also require that each component Q_i of $(Q_1, \ldots, Q_k) \in \mathscr{Q}^k$ has finite first absolute moment with G-probability one and, moreover, that this moment has finite G-expectation. We use $\mathscr{Q}^*(\mathscr{Q}^k)$ to denote the subspace of $\mathscr{Q}(\mathscr{Q}^k)$ consisting of those G's satisfying this condition. Since we sometimes write expectations as integrals, the above condition can be written in a variety of ways:

$$E(|X_{i1}|) = E(E(|X_{i1}| \,|\, Q_i)) = \int_{\mathscr{Q}} \int_{\mathbf{R}} |x| Q_i(\mathrm{d}x)\, G(\mathrm{d}(Q_1, \ldots, Q_k))$$

$$= \int_{\mathscr{Q}} \left(\int_{\mathbf{R}} |x| Q_i(\mathrm{d}x) \right) F_i(\mathrm{d}Q_i) < \infty$$

where F_i denotes the distribution of Q_i, that is, the ith marginal distribution of G. Since

$$|Z_m| \leqslant \bigvee_{i=1}^{k} |X_{im}| \leqslant \sum_{i=1}^{k} |X_{im}|,$$

Z_m is integrable and for any strategy τ,

$$\left| E_\tau \left(\sum_{m=1}^{\infty} \alpha_m Z_m \right) \right| \leqslant \left[\sum_{i=1}^{k} E(|X_{im}|) \right] \sum_{m=1}^{\infty} \alpha_m < \infty,$$

where the subscript τ indicates the dependence of the expectation on the strategy τ. The quantity $E_\tau(\Sigma_{m=1}^{\infty} \alpha_m Z_m)$, called the *worth* of the strategy τ, will be denoted by $W(G; \mathbf{A}; \tau)$. The *value* of the $(G; \mathbf{A})$-*bandit* is the maximal worth:

$$V(G; \mathbf{A}) = \sup_{\tau} W(G; \mathbf{A}; \tau) = \sup_{\tau} E_\tau \left(\sum_{m=1}^{\infty} \alpha_m Z_m \right). \qquad (2.2.2)$$

A strategy for which $V(G; \mathbf{A})$ is attained is *optimal*. In Sections 2.3 and 2.5 it will be shown that there is an optimal strategy for every G and $\mathbf{A}$.

At the first stage of any bandit problem, the decision maker is faced with an initial distribution and a discount sequence. At the second stage (after the initial selection and observation) the decision maker is faced with a new distribution and a new discount sequence, that is, with a new bandit problem. So the second stage selection can be viewed as the initial selection in this new bandit; and similar statements apply at every future stage. Therefore, if an optimal initial selection were known for all possible bandits – that is, all possible pairs $(G; \mathbf{A})$ – then an optimal strategy would be known for every bandit. This gives the initial selection special significance. However, as the

subsequent development shows, an optimal initial selection cannot be made without considering future selections.

The worth of selecting an arm can be separated into the sum of the expected payoff at stage 1 and the expected value of the best that can be subsequently achieved. Partitioning the set of strategies according to the arm selected initially (as in Section 2.1), the supremum in (2.2.2) can be represented as the maximum of k suprema:

$$V(G;\mathbf{A}) = \bigvee_{i=1}^{k} \sup_{\tau(\varnothing)=i} W(G;\mathbf{A};\tau)$$

$$= \bigvee_{i=1}^{k} \left[\alpha_1 E(X_{i1}) + \sup_{\tau(\varnothing)=i} E_\tau \left(\sum_{m=2}^{\infty} \alpha_m Z_m \right) \right]. \tag{2.2.3}$$

We proceed to introduce notation that will aid in the study of (2.2.3) and which reflects the fact that the decision maker is faced with a new bandit at every stage. For a discount sequence $\mathbf{A} = (\alpha_1, \alpha_2, \alpha_3, \ldots)$, $\mathbf{A}^{(1)}$ denotes the discount sequence $(\alpha_2, \alpha_3, \ldots)$. More generally, $\mathbf{A}^{(m)}$ denotes the discount sequence $(\alpha_{m+1}, \alpha_{m+2}, \ldots)$.

The discount sequence $\mathbf{A}^{(1)}$ and the posterior distribution of $(Q_1, \ldots, Q_k)$ after stage 1 characterize the bandit confronting the decision maker at stage 2. To indicate the dependence of this posterior distribution on the observation we use $(x)_i G$ to denote a version of the conditional distribution of $(Q_1, \ldots, Q_k)$ given observation x on arm i. Lemma 2.2.1 below asserts that $(x)_i G$ can be chosen to depend measurably on (x, G). No suitable reference for this lemma was found in the literature. (This lemma, or a variation, is fundamental in statistical contexts involving random prior distributions, but such considerations are seldom made explicit.) For the lemma recall that

$$\Omega = \mathscr{D}^k \times \bigtimes_{i=1}^{k} \bigtimes_{m=1}^{\infty} (0,1).$$

Lemma 2.2.1 For each $j = 1, \ldots, k$ there exists a measurable function $(x, G) \mapsto (x)_j G$ from $\mathbb{R} \times \mathscr{D}(\mathscr{D}^k)$ into $\mathscr{D}(\mathscr{D}^k)$ such that for every Borel subset C of $\mathscr{D}^k$,

$$\Big((X_{jm}(\omega))_j G \Big)(C) = P\left(C \times \bigtimes_{i=1}^{k} \bigtimes_{m=1}^{\infty} (0,1) \,\middle|\, X_{jm} \right)(\omega) \qquad \text{a.e. } \omega. \tag{2.2.4}$$

Remarks The following proof is essentially due to Varadhan (1983) who treats the joint measurability problem, but Varadhan was not

concerned with our special product space structure. Rhenius (1977) proves a similar result using similar methods. □

Before proving the lemma we introduce notation and two other lemmas. We use $\mathscr{D}(\Omega)$ to denote the space of probability distributions on $(\Omega, \mathscr{F})$. Again we use the topology of convergence in distribution: $P_l \to P$ if and only if

$$\int_\Omega h \, dP_l \to \int_\Omega h \, dP$$

for every bounded continuous h on Ω. The measurable subsets of $\mathscr{D}(\Omega)$ are the Borel sets. As in similar contexts discussed previously, $\mathscr{D}(\Omega)$ is separable, locally compact, metrizable to be complete, and has a topology determined by the convergent sequences and their limits. We omit the easy proof of the following result.

Lemma 2.2.2 The function which maps $P \in \mathscr{D}(\Omega)$ to its marginal on $\mathscr{D}^k$ is continuous. So is the function which maps $G \in \mathscr{D}(\mathscr{D}^k)$ to the member of $\mathscr{D}(\Omega)$ that is the product of G and Lebesgue measure on each $(0, 1)$.

Lemma 2.2.3 If Y is a bounded random variable, then $P \mapsto \int_\Omega Y \, dP$ is a measurable function from $\mathscr{D}(\Omega)$ into $\mathbb{R}$ and $P \mapsto Y \, dP$ is a measurable function from $\mathscr{D}(\Omega)$ into $\mathscr{D}(\Omega)$.

Proof (cf. Dubins and Freedman, 1966, 3.1) Suppose first that Y is continuous. The function $P \mapsto \int_\Omega Y \, dP$ is then continuous by definition. Suppose that $P_n \to P$ as $n \to \infty$ and that h is a bounded continuous function on Ω. Then hY is bounded and continuous, so

$$\lim_{n \to \infty} \int_\Omega h \, (Y \, dP_n) \to \int_\Omega h \, (Y dP).$$

Thus, by definition $Y \, dP_n \to Y \, dP$ as $n \to \infty$. Therefore $P \to Y \, dP$ is a continuous function from $\mathscr{D}(\Omega)$ into $\mathscr{D}(\Omega)$.

Next suppose that $Y = \mathbf{1}_B$ for some closed $B \subset \Omega$. Let

$$Y_q(\omega) = 0 \vee [1 - (\text{distance from } B \text{ to } \omega) q].$$

Since Y_q is continuous, it follows that the functions $P \mapsto \int_\Omega Y_q \, dP$ and $P \mapsto Y_q \, dP$ are continuous. For each P, $\int_\Omega Y_q \, dP \to \int_\Omega \mathbf{1}_B \, dP$ as $q \to \infty$ and, for each bounded continuous h, $\int_\Omega h (Y_q \, dP) \to \int_\Omega h (\mathbf{1}_B \, dP)$ as

$q \to \infty$, by Lebesgue's dominated convergence theorem. Therefore, the functions $P \mapsto \int_\Omega \mathbf{1}_B \, dP$ and $P \mapsto \mathbf{1}_B \, dP$ are the pointwise limits of sequences of continuous functions, and so are measurable.

The class of measurable B's for which the functions $P \mapsto \int_\Omega \mathbf{1}_B \, dP$ and $P \mapsto \mathbf{1}_B \, dP$ are measurable contains Ω and is closed under disjoint unions and proper differences (by addition and subtraction of measurable functions). It is also closed under increasing limits – by the monotone (or dominated) convergence theorem in the case of the first function, and by the dominated convergence theorem applied when the integrand is multiplied by an arbitrary bounded continuous h in the case of the second function. Therefore (Chow and Teicher, 1978, Theorem 1.3.2, for instance), the class of B's for which $P \mapsto \int_\Omega \mathbf{1}_B \, dP$ and $P \mapsto \mathbf{1}_B \, dP$ contains the σ-field generated by the closed sets.

For any bounded random variable Y we conclude that the functions $P \mapsto \int_B Y \, dP$ and $P \mapsto Y \, dP$ are measurable by taking limits of the two sequences of functions obtained when Y is replaced by a sequence of simple functions approaching Y everywhere. □

Remark Eick (1984) helped us with the proof of Lemma 2.2.3 and with formulating the lemma in a manner most useful for the proof of Lemma 2.2.1. □

Proof of Lemma 2.2.1 For $x \in [r2^{-l}, (r+1)2^{-l})$, define the probability measure $((x)_j G)_l$ by

$$((x)_j G)_l(C) = \begin{cases} G(C) & \text{if } P(r2^{-l} \leqslant X_{jn} < (r+1)2^{-l}) = 0 \\ P(C \times \overset{k}{\underset{i=1}{\mathsf{X}}} \overset{\infty}{\underset{m=1}{\mathsf{X}}} (0,1) \mid r2^{-l} \leqslant X_{jn} < (r+1)2^{-l}) & \text{otherwise} \end{cases}$$

for C a measurable subset of $\mathscr{D}^k$. Notice that, although n indexes the random variable, the definition is independent of n. Because of the rather simple dependence on x, the measurability of $(x, G) \mapsto ((x)_j G)_l$ will follow from the measurability of $G \mapsto ((x)_j G)_l$ for each fixed x. This measurability is an immediate consequence of Lemmas 2.2.2 and 2.2.3 with the indicator function of $\{\omega: r2^{-l} \leqslant X_{jn}(\omega) < (r+1)2^{-l}\}$ being used for Y in Lemma 2.2.3. Let

$$(x)_j G = \begin{cases} \lim_{l \to \infty} ((x)_j G)_l & \text{if the limit exists} \\ G & \text{otherwise.} \end{cases} \tag{2.2.5}$$

Clearly, the function $(x, G) \mapsto (x)_j G$ is measurable.

It remains to prove (2.2.4) for each measurable $C \subset \mathscr{D}^k$. By the martingale convergence theorem,

$$\left((X_{jn}(\omega))_j G \right)_l (C) \to P(C \times \overset{k}{\underset{i=1}{\mathsf{X}}} \overset{\infty}{\underset{m=1}{\mathsf{X}}} (0,1) \mid X_{jn})(\omega) \qquad \text{a.e. } \omega$$

as $l \to \infty$. If ω does not belong to the exceptional set and is such that $X_{jn}(\omega)$ equals an x for which the limit in (2.2.5) exists, then

$$((X_{jn}(\omega))_j G)(C) = P\left(C \times \overset{k}{\underset{i=1}{\mathsf{X}}} \overset{\infty}{\underset{m=1}{\mathsf{X}}} (0,1) \,\middle|\, X_{jn} \right)(\omega).$$

We shall complete the proof by showing that

$$\lim_{l \to \infty} ((X_{jn}(\omega))_j G)_l \tag{2.2.6}$$

exists for almost every ω. To obtain a candidate for the limit we first observe that the range of X_{jn} equals $\mathbb{R}$; any $x \in \mathbb{R}$ equals $Q_j^{-1}(\omega_{jn})$ for some Q_j and ω_{jn}. In view of this fact and Theorem V.2.2 of Parthasarathy (1967), Theorem V.8.1 of Parthasarathy (1967) is applicable. This latter theorem asserts the existence of a regular conditional probability distribution P_ω on Ω, one property of which is

$$E(h \mid X_{jn})(\omega) = \int_\Omega h \, dP_\omega \qquad \text{a.e. } \omega, \tag{2.2.7}$$

for every bounded measurable h on Ω. In case h depends on $(Q_1, \ldots, Q_k)$, (2.2.7) can be rewritten

$$E(h \mid X_{jn})(\omega) = \int_{\mathscr{D}^k} h \, dG_\omega \qquad \text{a.e. } \omega, \tag{2.2.8}$$

where G_ω denotes the projection of P_ω on $\mathscr{D}^k$. The measure G_ω is the candidate for the almost sure limit at (2.2.6).

For a bounded measurable h that depends only on $(Q_1, \ldots, Q_k)$, the martingale convergence theorem and (2.2.8) give

$$\lim_{l \to \infty} \int_{\mathscr{D}^k} h \, d((X_{jn}(\omega))_j G) = \int_{\mathscr{D}^k} h \, dG_\omega \qquad \text{a.e. } \omega; \tag{2.2.9}$$

the exceptional null set may depend on h. If we restrict to a countable determining set of continuous h's, the null set may be chosen independently of h. The equality in (2.2.9) for such h's is sufficient for concluding that the limit at (2.2.6) exists and equals G_ω. □

The existence of a regular conditional probability distribution was used in the proof of Lemma 2.2.1; there is little significance in the reverse implication. Nevertheless, the following corollary of Lemma 2.2.1 sheds further light on that lemma.

Corollary 2.2.4 For each measurable subset $C \subset \mathscr{D}^k$, each $G \in \mathscr{D}(\mathscr{D}^k)$, and each $i = 1, \ldots, k$, the function $x \mapsto ((x)_i G)(C)$ is measurable.

Proof The function of interest is the composition of $x \mapsto (x)_i G$ and $G \mapsto G(C)$. The first is measurable by Lemma 2.2.1 and the second by Lemma 2.2.3. □

We now describe two distinct ways of viewing a bandit problem. The previous development makes it clear that they are equivalent.

Suppose a decision maker can give instructions to an aide who is to carry out the decision maker's wishes. One way is to give a strategy τ to the aide. As indicated earlier in this section, the aide will evaluate $\tau(\varnothing)$, select that arm at stage 1, observe a consequent number z_1, evaluate $\tau(z_1)$, select that arm at stage 2, observe z_2, evaluate $\tau(z_1, z_2)$, select that arm at stage 3, etc.

Another way is for the decision maker to consider all pairs of distributions in $\mathscr{D}^*(\mathscr{D}^k)$ and discount sequences, $(G, \mathbf{A})$, and to specify $\tau_{G,\mathbf{A}}(\varnothing)$, the arm to be selected initially for each $(G, \mathbf{A})$-bandit. When faced with a *particular* G and $\mathbf{A}$, the aide selects $\tau_{G,\mathbf{A}}(\varnothing)$ and observes the result z_1. The aide will then (with ease) calculate $\mathbf{A}^{(1)}$ and (possibly with difficulty) calculate $(z_1)_j G$ with $j = \tau_{G,\mathbf{A}}(\varnothing)$, which distribution the aide might call $G^{(1)}$, the current information about the arms after stage 1. At stage 2 the aide will select arm $\tau_{G^{(1)},\mathbf{A}^{(1)}}(\varnothing)$ and observe z_2. Then the aide will calculate $\mathbf{A}^{(2)}$ and $G^{(2)}$ and select arm $\tau_{G^{(2)},\mathbf{A}^{(2)}}(\varnothing)$ at stage 3. And so on.

There are a number of things worth mentioning concerning the latter approach. First, since any particular outcome may have probability 0, the aide cannot be permitted to calculate conditional distributions after making observations but must choose versions in advance. Secondly, it is important for the evaluation of expectations and therefore for the assessment of strategies, that, for instance, $\{z_1 : \tau(z_1) = i\}$ be measurable. Now $\tau(z_1)$ is defined implicitly:

$$\tau(z_1) = \tau_{G^{(1)},\mathbf{A}^{(1)}}(\varnothing).$$

Lemma 2.2.1 helps by assuring that $G^{(1)}$ depends measurably on z_1 and the decision maker must guarantee that $\tau_{G,\mathbf{A}}(\varnothing)$ depends measurably on G.

The notation $G^{(1)}$ (and more generally, $G^{(m)}$) for the conditional distribution on $\mathscr{Q}^k$ after stage 1 (and m) is useful for avoiding subscripted subscripts and notations such as $(z_2)_j(z_1)_i G$. It should be emphasized that $G^{(1)}$, say, can be regarded as either a function $(z_1)_{\tau(\varnothing)}G$ of the initial observation, or as a function $(Z_1(\omega))_{\tau(\varnothing)}G$ of ω. The context will distinguish between the two interpretations. For instance, $G^{(1)}$ is a random distribution in the expression $E_\tau V(G^{(1)}; \mathbf{A}^{(1)})$.

Many discussions may involve several G's simultaneously. Rather than using corresponding subscripts and superscripts on P and E, we will use P and E without subscripts and a G to the right of a conditioning bar to indicate the G from which P is constructed. Accordingly,

$$P_\tau(Z_2 > 7 | G)$$

is the probability that the second observation is larger than 7 when strategy τ is followed in a $(G; \mathbf{A})$-bandit. Also,

$$E_\tau(X_{12}|G, Z_1) = E_\tau(X_{12}|G^{(1)})$$

is a random variable which depends on τ, while no subscript on E is necessary when writing $E(X_{12}|G, X_{31})$; this last expression cannot be simplified using the notation $G^{(1)}$.

Consistent with the above comments we can write

$$E(|X_{j2}|\,\big|\,G) = E_\tau\Big(E(|X_{j1}|\,\big|\,G^{(1)})\,\Big|\,G\Big). \tag{2.2.10}$$

When we speak of the $(G; \mathbf{A})$-bandit we assume implicitly that $G \in \mathscr{Q}^*(\mathscr{Q}^k)$ and not just $G \in \mathscr{Q}(\mathscr{Q}^k)$. From (2.2.10) we see that $G \in \mathscr{Q}^*(\mathscr{Q}^k)$ implies $G^{(1)} \in \mathscr{Q}^*(\mathscr{Q}^k)$ with probability one; so that it makes sense to speak of the $(G^{(1)}; \mathbf{A}^{(1)})$-bandit.

In case $G = F_1 \times \ldots \times F_k$, $(x)_i$ can be applied directly to F_i with the subscript dropped; so

$$(x)_i(F_1 \times \ldots \times F_k) = (F_1 \times \ldots \times (x)F_i \times \ldots \times F_k)$$

where $(x)F_i$ denotes the conditional distribution of Q_i given an observation x on arm i. In the example of Section 2.1, $(x)F_1$ is supported by normal distributions having variance 1 and distributes

the mean of such a normal distribution normally with mean $(\mu+\rho^2 x)/(1+\rho^2)$ and variance $\rho^2/(1+\rho^2)$ (F_1 is denoted by F in Section 2.1 since arm 1 is the only unknown arm in that section).

In the case of Bernoulli arms, where all outcomes are either 1 or 0, we frequently refer to the outcomes as 'success' and 'failure'. In this case we replace $(1)_i$ by σ_i and $(\dot{0})_i$ by φ_i. Accordingly, $\sigma_1\varphi_1\sigma_2 G$ is the conditional distribution of $(Q_1, \ldots, Q_k)$ given a success and a failure on arm 1 and a success on arm 2, and, with exponents in lieu of repetitions, $\sigma_1\varphi_4^3 G$ is the conditional distribution of $(Q_1, \ldots, Q_k)$ given a success on arm 1 and three failures on arm 4.

At various places, we will need a topology on the space of discount sequences as well as on the other spaces we have discussed, such as $\mathscr{D}(\mathscr{D}^k)$. We will use the l_1-metric topology arising from the l_1-norm. The norm of a discount sequence $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$ is

$$|\mathbf{A}|_1 = \sum_{m=1}^{\infty} \alpha_m$$

and the distance between it and a discount sequence $\mathbf{B} = (\beta_1, \beta_2, \ldots)$ is

$$|\mathbf{A}-\mathbf{B}|_1 = \sum_{m=1}^{\infty} |\alpha_m - \beta_m|.$$

We will use $\mathscr{A}$ to denote this metric space of discount sequences.

Notice that

$$|\mathbf{A}^{(m)}|_1 = \sum_{p=m+1}^{\infty} \alpha_p$$

for which we will also use the notation γ_{m+1} in Chapter 5 in order to facilitate the algebraic manipulations there. In particular, $\gamma_1 = |\mathbf{A}|_1$.

2.3 Dynamic programming for finite horizons

The *horizon* of a discount sequence $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$, or of a bandit with that discount sequence, equals

$$\inf\{n: \alpha_m = 0 \quad \text{for } m > n\}.$$

The horizon may be infinite, and it equals 0 if $\alpha_m = 0$ for all m. The main feature of this section is a proof of the existence of an optimal strategy for each bandit having finite horizon. Also included is the

beginning of our study of the dependence of the value function V and optimal strategies on G and $\mathbf{A}$.

The following lemma shows the existence of an optimal strategy when the horizon is finite. It will be extended to general discount sequences in Section 2.5. Recall from Section 2.2 that $\mathscr{A}$ denotes the space of all discount sequences and $\mathscr{D}^*(\mathscr{D}^k)$ denotes the space of all G such that $E(|X_{i1}|\,|G) < \infty$ for all i.

Lemma 2.3.1 There exists a strategy $\tau_{G,\mathbf{A}}$ for each $\mathbf{A} \in \mathscr{A}$ having finite horizon and each $G \in \mathscr{D}^*(\mathscr{D}^k)$ such that

$$\{(G; \mathbf{A}; z_1, \ldots, z_{m-1}) : \tau_{G,\mathbf{A}}(z_1, \ldots, z_{m-1}) = i\}$$

is a measurable subset of $\mathscr{D}^*(\mathscr{D}^k) \times \mathscr{A} \times (-\infty, \infty)^{m-1}$ for each $i = 1, \ldots, k$ and $m = 1, 2, \ldots$, and $\tau_{G,\mathbf{A}}$ is, for each G and $\mathbf{A}$, optimal for the $(G; \mathbf{A})$-bandit. Restricted to discount sequences having finite horizon, the function $(G, \mathbf{A}) \mapsto V(G; \mathbf{A})$ is measurable and satisfies

$$V(G; \mathbf{A}) = \bigvee_{i=1}^{k} E\left(\alpha_1 X_{i1} + V((X_{i1})_i G; \mathbf{A}^{(1)}) \middle| G\right). \tag{2.3.1}$$

Remark Recall that $(X_{i1})_i G$ denotes the measure obtained by conditioning G on the observation X_{i1} on arm i and $\mathbf{A}^{(1)}$ denotes the sequence obtained by dropping the first member from $\mathbf{A}$. Thus, (2.3.1) gives V recursively for finite horizon bandits, but the recursion proceeds backwards. *Dynamic programming* and *backwards induction* are terms often used for such a backward recursion. □

Proof of Lemma 2.3.1 The set of discount sequences having a particular horizon is measurable. Accordingly, we may proceed by induction on the horizon. The only discount sequence having horizon 0 is $\mathbf{0} = (0, 0, 0, \ldots)$. Clearly $V(G; \mathbf{0}) = 0$ for every G and all strategies are optimal. To be specific we set

$$\tau_{G,0}(z_1, \ldots, z_{m-1}) = 1$$

or all $m = 1, 2, \ldots$ and $z_1, z_2, \ldots, z_{m-1}$. Equation (2.3.1) obviously holds with $\mathbf{A} = \mathbf{0}$.

Suppose $\tau_{G,\mathbf{A}}$ has been defined and the assertions of the lemma established for all discount sequences $\mathbf{A}$ having horizon less than n. We proceed to define $\tau_{G,\mathbf{A}}$ for $\mathbf{A}$ having horizon n. Define a strategy $\tau^i_{G,\mathbf{A}}$

that selects arm i initially and then proceeds optimally:

$$\tau^i_{G,\mathbf{A}}(\varnothing) = i$$

$$\tau^i_{G,\mathbf{A}}(z_1, \ldots, z_{m-1}) = \tau_{(z_1)_i G, \mathbf{A}^{(1)}}(z_2, \ldots, z_{m-1}) \quad \text{if } m > 1.$$

Since the functions $\mathbf{A} \mapsto \mathbf{A}^{(1)}$ and $(z_1, G) \mapsto (z_1)_i G$ are measurable, the induction hypothesis yields the measurability of

$$\{(G; \mathbf{A}; z_1, \ldots, z_{m-1}) : \tau^i_{G,\mathbf{A}}(z_1, \ldots, z_{m-1}) = j\}$$

for each i and m and, in particular, that each $\tau^i_{G,\mathbf{A}}$ is a strategy.

To prove (2.3.1) we consider the right-hand side of (2.2.3):

$$\sup_{\tau(\varnothing)=i} E_\tau\left(\sum_{m=2}^{\infty} \alpha_m Z_m \middle| G\right) = \sup_{\tau(\varnothing)=i} E_\tau\left(E_\tau\left(\sum_{m=2}^{\infty} \alpha_m Z_m \middle| X_{i1}\right)\middle| G\right)$$

$$\leqslant E_\tau\left(\sup_{\tau(\varnothing)=i}\left(E_\tau\left(\sum_{m=2}^{\infty} \alpha_m Z_m \middle| X_{i1}\right)\middle| G\right)\right).$$

It is clear that each supremum on the right-hand side is attained for $\tau^i_{G,\mathbf{A}}$ and the corresponding supremum equals $V((X_{i1})_i G; \mathbf{A}^{(1)})$, which, by the induction hypothesis and the measurability of $(x, G) \mapsto (x)_i G$, is a measurable function of $(\omega, G, \mathbf{A})$. So, (2.3.1) follows from (2.2.3); and from (2.3.1) and the induction hypothesis we obtain the measurability of $(G, \mathbf{A}) \mapsto V(G; \mathbf{A})$.

We complete the proof by setting $\tau_{G,\mathbf{A}}$ equal to the $\tau^i_{G,\mathbf{A}}$ having the smallest i for which the maximum in (2.3.1) occurs. □

For calculational purposes, the following rewritings of (2.3.1) may be useful:

$$V(G; \mathbf{A}) = \bigvee_{i=1}^{k} \int_{\mathscr{Q}^k} \int_R \left[\alpha_1 x + V((x)_i G; \mathbf{A}^{(1)})\right] \cdot Q_i(\mathrm{d}x) G(\mathrm{d}(Q_1, \ldots, Q_k))$$

$$= \bigvee_{i=1}^{k} \int_{\mathscr{Q}} \int_R \left[\alpha_1 x + V((x)_i G; \mathbf{A}^{(1)})\right] Q_i(\mathrm{d}x) F_i(\mathrm{d}Q_i). \tag{2.3.2}$$

We now show how to use (2.3.2) to find optimal strategies. Suppose the $(G; \mathbf{A})$-bandit has horizon 1. Since the horizon of $\mathbf{A}^{(1)}$ is zero,

$V(G^{(1)}; \mathbf{A}^{(1)}) = 0$ for any $G^{(1)}$ and so (2.3.2) simplifies to

$$V(G;(\alpha_1,0,0,\ldots)) = \alpha_1 \bigvee_{i=1}^{k} \int_{\mathscr{Q}^1} \int_{\mathbf{R}} x\, Q_i(\mathrm{d}x)\, G(\mathrm{d}(Q_1,\ldots,Q_k)) \tag{2.3.3}$$

$$= \alpha_1 \bigvee_{i=1}^{k} \int_{\mathscr{Q}} \int_{\mathbf{R}} x Q_i(\mathrm{d}x) F_i(\mathrm{d}Q_i).$$

An optimal initial selection is any arm i for which the maximum in (2.3.3) is attained.

Suppose the $(G; \mathbf{A})$-bandit has horizon 2. To use (2.3.2) we require the values of bandits having $\mathbf{A}^{(1)}$, which has horizon 1, as the discount sequence. Not all distributions need be considered in conjunction with $\mathbf{A}^{(1)}$, only those of the form $(x)_i G$ for some possible observation x on some arm i. The desired values $V((x)_i G; \mathbf{A}^{(1)})$ can be obtained from (2.3.3). When (2.3.2) is used to find $V(G; \mathbf{A})$, the optimal initial selections are found as those indicated by the i's for which the maximum in (2.3.2) is attained.

Now that we know (in principle) how to find $V(G; \mathbf{A})$ for any $(G; \mathbf{A})$ with a horizon of 2, we can (again, in principle) find $V(G; \mathbf{A})$ when the horizon is 3 by applying (2.3.2). In general, a bandit with horizon n can be solved (in principle) by first solving many bandits with horizon $n-1$, etc.

This process can be carried out to solve an arbitrary $(G; \mathbf{A})$-bandit in which the horizon of $\mathbf{A}$ is $n < \infty$. The first step is to calculate, for each $m = 1, \ldots, n-1$ and each allocation of m selections on the k arms, the possible conditional distributions $G^{(m)}$ of $(Q_1, \ldots, Q_k)$ given these observations. Since $\mathbf{A}^{(n-1)}$ has horizon 1, each $V(G^{(n-1)}; \mathbf{A}^{(n-1)})$ is obtained from (2.3.3). Then each $V(G^{(n-2)}; \mathbf{A}^{(n-2)})$ is calculated from (2.3.2). This process of backward induction or dynamic programming continues until $V(G; \mathbf{A})$ is obtained. For each $(G^{(m)}; \mathbf{A}^{(m)})$-bandit that can arise starting with the $(G; \mathbf{A})$-bandit, the optimal initial selections are indicated by those i's for which the maximum in (2.3.2) is attained. These selections can be pieced together to give all optimal strategies for the $(G; \mathbf{A})$-bandit.

The method described above is illustrated with several examples in the next section.

Remark Equations such as (2.3.2) are common in the literature; the term 'fundamental equation of dynamic programming' is used. In

various probabilistic contexts, posterior distributions have to be identified with the 'states' in the basic dynamic programming formulations. We do not think it is a trivial matter to do so and then to complete the appropriate arguments; for us, Lemma 2.2.1 has played a central role. We feel that the literature contains some oversights in this regard. □

2.4 Examples having finite horizons

Various examples are provided here to illustrate the technique described in Section 2.3 and to give some flavour of the variety of issues that can arise in bandit problems. Detailed calculations are given in the first two examples. Only the most interesting aspects of the remaining examples are given. The interested reader can perform the requisite backward induction to verify our statements in these latter examples.

Only Examples 2.4.2 and 2.4.5 have more than two arms. Only Example 2.4.6 has an arm that is not Bernoulli (except that there is an arbitrary but known arm in Example 2.4.7). Example 2.4.7 is the only example concerned with an infinite subset of $\mathscr{D}(\mathscr{D}^k)$. Example 2.4.1 is the only one having dependent arms.

The first example is due to Bradt, Johnson, and Karlin (1956), who gave it as a counterexample to the 'stay-with-a-winner rule'. The setting is quite simple but the arms are dependent.

Example 2.4.1 Suppose $k = 2$ and the arms are Bernoulli with parameters θ_1 and θ_2 so that

$$P(X_{im} = 1 | \theta_i) = \theta_i = 1 - P(X_{im} = 0 | \theta_i). \tag{2.4.1}$$

The distributions Q_1 and Q_2 do not appear explicitly in (2.4.1); they are represented by the random parameters θ_1 and θ_2, and a probability distribution G on $\mathscr{D}^2$ can be represented by a distribution, that we also call G, on the unit square.

The parameters θ_1 and θ_2 are known to be either both small or both large. More precisely, G is a two-point probability measure:

$$G = \frac{4}{5}\delta_{(1/10,0)} + \frac{1}{5}\delta_{(9/10,1)},$$

where $\delta_{(x,y)}$ is the delta measure at (x, y). As in Section 2.1, assume $\mathbf{A} = (1, 1, 0, 0, \ldots)$.

The reader can easily list all possible strategies, calculate $V(G;\mathbf{A}) = 53/100$, and conclude that the only optimal strategy τ is given by $\tau(\varnothing) = 1$, $\tau(1) = 2$, $\tau(0) = 1$. Accordingly, arm 1 should be selected initially and a switch to arm 2 should be made if and only if a success is observed at stage 1. This fact gave the example its original importance. Our goal is to illustrate the dynamic programming method described in the preceding section and in so doing we will reach the above conclusions in a rather laborious manner.

As indicated in Section 2.3, we begin by listing all the $G^{(m)}$, $m = 1, \ldots, n-1$, that can arise, where n is the horizon. In this example $n = 2$ and so 1 is the only relevant value of m. The distributions $G^{(1)}$ that can arise are four in number and are given by (see Section 2.2 for notation):

$$\sigma_1 G = \frac{4}{13}\delta_{(1/10,0)} + \frac{9}{13}\delta_{(9/10,1)},$$

$$\phi_1 G = \frac{36}{37}\delta_{(1/10,0)} + \frac{1}{37}\delta_{(9/10,1)},$$

$$\sigma_2 G = \delta_{(9/10,1)},$$

$$\phi_2 G = \delta_{(1/10,0)}.$$

From (2.3.3) we obtain:

$$V(\sigma_1 G;\mathbf{A}^{(1)}) = \left(\frac{4}{13}\cdot\frac{1}{10} + \frac{9}{13}\cdot\frac{9}{10}\right) \vee \left(\frac{4}{13}\cdot 0 + \frac{9}{13}\cdot 1\right)$$

$$= \frac{17}{26} \vee \frac{9}{13} = \frac{9}{13}, \tag{2.4.2}$$

$$V(\varphi_1 G;\mathbf{A}^{(1)}) = \left(\frac{36}{37}\cdot\frac{1}{10} + \frac{1}{37}\cdot\frac{9}{10}\right) \vee \left(\frac{36}{37}\cdot 0 + \frac{1}{37}\cdot 1\right)$$

$$= \frac{9}{74} \vee \frac{1}{37} = \frac{9}{74}, \tag{2.4.3}$$

$$V(\sigma_2 G;\mathbf{A}^{(1)}) = \left(1\cdot\frac{9}{10}\right) \vee (1\cdot 1) = \frac{9}{10} \vee 1 = 1, \tag{2.4.4}$$

$$V(\varphi_2 G;\mathbf{A}^{(1)}) = \left(1\cdot\frac{1}{10}\right) \vee (1\cdot 0) = \frac{1}{10} \vee 0 = \frac{1}{10}. \tag{2.4.5}$$

To make the optimal selections clear, we have written the maxima in detail and in the order of the arm number. For instance, arm 2 is the

optimal selection in the $(\sigma_1 G; (1,0,0,\ldots))$-bandit since 9/13 is the maximum and is second in the expression (17/26) ∨ (9/13).

For the second and, in this example, last step in the recursion, we use (2.3.2) to obtain

$$V(G;\mathbf{A}) = \left(\frac{4}{5}\left[\frac{1}{10}\left(1+\frac{9}{13}\right)+\frac{9}{10}\left(0+\frac{9}{74}\right)\right]+\frac{1}{5}\left[\frac{9}{10}\left(1+\frac{9}{13}\right)\right.\right.$$
$$\left.\left.+\frac{1}{10}\left(0+\frac{9}{74}\right)\right]\right)\vee\left(\frac{4}{5}\left[0(1+1)+1\left(0+\frac{1}{10}\right)\right]\right.$$
$$\left.+\frac{1}{5}\left[1(1+1)+0\left(0+\frac{1}{10}\right)\right]\right)$$
$$=\frac{53}{100}\vee\frac{12}{25}=\frac{53}{100};$$

and so the only optimal first selection is arm 1. From (2.4.2) and (2.4.3) we see that, at stage 2, the decision maker should select arm 2 after a success with arm 1 and arm 1 after a failure with arm 1. Calculations (2.4.4) and (2.4.5) are necessary to determine that arm 1 is optimal initially, but once that determination is made they are irrelevant.

□

Although we were able to make a complete list of all distributions $G^{(1)}$ in the above example, this is not necessary for an explicit solution using dynamic programming. In particular, the example in Section 2.1 was solved with dynamic programming, although that terminology was not used.

While every possible strategy is evaluated in the previous example, this does not happen using dynamic programming when the horizon is larger than 2. The next example (in which the horizon is 3) makes this clear.

Example 2.4.2 Suppose $k = 3$ and the arms are Bernoulli with parameters θ_1, θ_2, θ_3. The distribution G can be regarded as a distribution on the cube $\{(u_1,u_2,u_3)\colon 0 \leqslant u_i \leqslant 1\}$ and the distributions F_i can be regarded as the corresponding marginal distributions of the θ_i. We suppose the distribution is supported by the plane $u_3 = 8/15$ on which it has a (two-dimensional) uniform density. So the three parameters are independent: $G = F_1 \times F_2 \times F_3$ where F_1 and F_2 are uniform distributions on $(0,1)$, written $U(0,1)$, and

$F_3 = \delta_{8/15}$. Take $\mathbf{A} = (1,1,1,0,0,\ldots)$; so the objective is to maximize the expected sum of the first three observations.

The first step in the dynamic program is to calculate all possible $G^{(m)}$ for $m = 1, 2$. Parameters θ_1 and θ_2 have beta densities and θ_3 remains known; on $(0,1) \times (0,1) \times \{8/15\}$:

$$\begin{aligned} &\mathrm{d}(\sigma_1^{s_1}\varphi_1^{f_1}\sigma_2^{s_2}\varphi_2^{f_2}\sigma_3^{s_3}\varphi_3^{f_3}G(u_1,u_2,u_3)) \\ &\propto u_1^{s_1}(1-u_1)^{f_1}u_2^{s_2}(1-u_2)^{f_2}\,\mathrm{d}u_1\,\mathrm{d}u_2. \end{aligned} \tag{2.4.6}$$

The possible $G^{(1)}$ are obtained by letting $s_1 + f_1 + s_2 + f_2 + s_3 + f_3 = 1$, namely: $\sigma_1 G$, $\varphi_1 G$, $\sigma_2 G$, $\varphi_2 G$, and G. The possible $G^{(2)}$ are obtained by letting $s_1 + f_1 + s_2 + f_2 + s_3 + f_3 = 2$. There are fifteen such distributions; these include the five mentioned above which now correspond to $s_3 + f_3 \geq 1$.

The fifteen values $V(G^{(2)}; \mathbf{A}^{(2)})$ were easily obtained using (2.3.3), multiplied by 180, and entered into Table 2.1(i). The maxima have been written explicitly in the order corresponding to the arm numbers so as to facilitate the eventual identification of optimal strategies.

For the next recursive step, (2.3.2) is to be used five times to calculate $V(\sigma_1 G; \mathbf{A}^{(1)})$, $V(\varphi_1 G; \mathbf{A}^{(1)})$, $V(\sigma_2 G; \mathbf{A}^{(1)})$, $V(\varphi_2 G; \mathbf{A}^{(1)})$, and $V(G; \mathbf{A}^{(1)})$; these are given in Table 2.1(ii). Let us, for instance, calculate $V(\sigma_1 G; \mathbf{A}^{(1)})$. To do this we need the normalizing constant in (2.4.6): it equals 2 for $s_1 = 1, f_1 = s_2 = f_2 = 0$. We may regard $\mathscr{D}$ to be $[0,1]$; so from (2.3.2) and Table 2.1(i) we obtain

$$\begin{aligned} 180V(\sigma_1 G; \mathbf{A}^{(1)}) &= \left(\int_0^1 [u_1(180+135) + (1-u_1)(0+96)]\, 2u_1\, \mathrm{d}u_1 \right) \\ &\quad \vee \left(\int_0^1 [u_2(180+120) + (1-u_2)(0+120)]\, \mathrm{d}u_2 \right) \\ &\quad \vee \left(\frac{8}{15}(180+120) + \frac{7}{15}(0+120) \right) \\ &= 242 \vee 210 \vee 216 \end{aligned}$$

which is entered in the appropriate place in Table 2.1(ii). Only one use of (2.3.2) is required at the last step in the recursion. The result of that calculation is entered in Table 2.1(iii).

From Table 2.1(iii) we immediately obtain that $V(G; \mathbf{A}) = 310/180$. In addition, we see that both arms 1 and 2 are optimal initially (they are obviously equally good by symmetry). Correspondingly, there are

Table 2.1 180 $V(G^{(m)}; \mathbf{A}^{(m)})$ *for m observations and various possible* $G^{(m)}$

(i) $m = 2$: $\mathbf{A}^{(2)} = (1,0,0,\ldots)$

		φ_1	σ_1	φ_1^2	$\varphi_1\sigma_1$	σ_1^2
	90 v 90 v 96	60 v 90 v 96	120 v 90 v 96	45 v 90 v 96	90 v 90 v 96	135 v 90 v 96
φ_2	90 v 60 v 96	60 v 60 v 96	120 v 60 v 96			
σ_2	90 v 120 v 96	60 v 120 v 96	120 v 120 v 96			
φ_2^2	90 v 45 v 96					
$\varphi_2\sigma_2$	90 v 90 v 96					
σ_2^2	90 v 135 v 96					

(ii) $m = 1$: $\mathbf{A}^{(1)} = (1,1,0,0,\ldots)$

		φ_1	σ_1
	198 v 198 v 192	156 v 198 v 192	242 v 210 v 216
φ_2	198 v 156 v 192		
σ_2	210 v 242 v 216		

(iii) $m = 0$: $\mathbf{A} = (1,1,1,0,0,\ldots)$

310 v 310 v 294

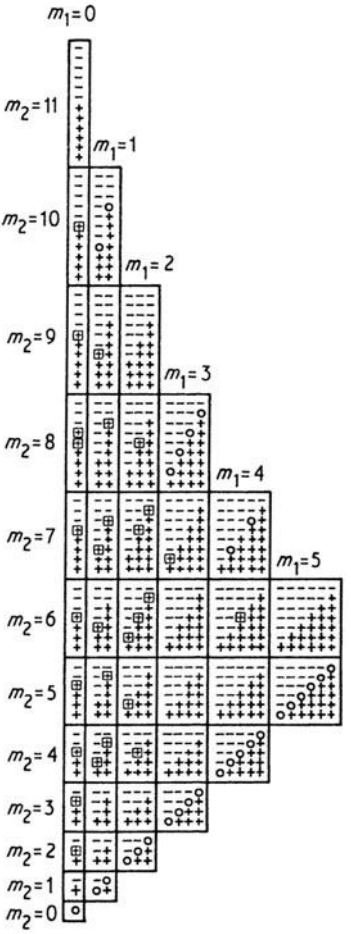
at least two optimal strategies; let τ be one with $\tau(\varnothing) = 1$. From the right-hand portion of Table 2.1(ii) we obtain $\tau(1) = 1$ and $\tau(0) = 2$. From the rightmost two entries in Table 2.1(i) we obtain $\tau(1,1) = 1$ and $\tau(1,0) = 3$. From the lower two entries in the second column of Table 2.1(i) we obtain $\tau(0,1) = 2$ and $\tau(0,0) = 3$, which, since selections after stage 3 are irrelevant, completes the description of τ. From Table 2.1 there is just one other optimal strategy and it is obtained by interchanging the roles of arms 1 and 2 in the above description.

There are a number of locations in Table 2.1 (for example, $\sigma_1\sigma_2 G$) that play no role in the eventual specification of an optimal strategy since they cannot be reached when following some optimal strategy.

But every value in the table is required to calculate $V(G;\mathbf{A})$ and, therefore, for specifying optimal strategies.

Another interesting aspect of the table is that many strategies are not evaluated; an example is the myopic strategy. There is a unique myopic strategy in this example; it selects arm 3 at all three stages and has worth $3(8/15) = 288/180$. Using backwards induction, strategies are ruled out if they do not perform well after the stage under consideration; so the myopic strategy was ruled out at stage 2. □

Table 2.2.



The next example is somewhat more interesting than the last in that the horizon is larger. It supposes there are two independent arms (arms 1 and 2 of the previous example) and a horizon of 12.

Example 2.4.3 Suppose $k = 2$ and arms 1 and 2 are independent Bernoulli distributions with uniformly distributed parameters: regarding F_1 and F_2 as distributions of θ_1 and θ_2, $F_i = U(0, 1)$. Take $\alpha_1 = \ldots = \alpha_{12} = 1, \alpha_{13} = \ldots = 0$, so the objective is to maximize the expected number of successes in the first 12 trials.

Every optimal strategy can be found from Table 2.2. Optimal initial selections are indicated in the table for every $(G^{(m)}; \mathbf{A}^{(m)})$ up to $m = 11$ that is possible when starting from $G = F_1 \times F_2$ and $\mathbf{A}$. The table gives the sign of

$$\Delta(G^{(m)}; \mathbf{A}^{(m)}) = W(G^{(m)}; \mathbf{A}^{(m)}; \tau_1) - W(G^{(m)}; \mathbf{A}^{(m)}; \tau_2), \quad (2.4.7)$$

where τ_i is a strategy for the $(G^{(m)}; \mathbf{A}^{(m)})$-bandit that begins with arm i and proceeds optimally thereafter. (The function Δ is used here for the various $(G^{(m)}; \mathbf{A}^{(m)})$ merely to display the optimal strategy. It is discussed in detail in Section 4.2 and used subsequently to show various properties of optimal strategies.) So (2.4.7) is positive when arm 1 is optimal for the $(G^{(m)}; \mathbf{A}^{(m)})$-bandit, negative when arm 2 is optimal, and zero when both are optimal.

Table 2.2, which indicates the sign of Δ, uses a scheme similar to that of Table 2.1. The distributions possible are

$$G^{(m)} = F_1^{(m_1)} \times F_2^{(m_2)} = \sigma^{s_1} \varphi^{f_1} F_1 \times \sigma^{s_2} \varphi^{f_2} F_2$$

where $s_i + f_i = m_i$, $m = m_1 + m_2$, and $F_i^{(m_i)}$ is a beta distribution; cf. (2.4.6). Within each (m_1, m_2) box, s_2 runs from 0 in the bottom row to m_2 in the top row; similarly, s_1 runs from 0 in the leftmost column to m_1 in the rightmost. Because of symmetry, the part of the table with $m_2 < m_1$ has been omitted. A square circumscribing a '+' in the table indicates a state in which the optimal selection disagrees with a myopic selection; that is, a state $(G^{(m)}; \mathbf{A}^{(m)})$ for which arm 1 is optimal but $E(X_{1,m+1} | G^{(m)}) < E(X_{2,m+1} | G^{(m)})$.

To see how Table 2.2 can be used, consider the first few selections. The 'O' in the lowermost box indicates that both arms are optimal initially (this is obvious from symmetry). Suppose arm 2 is used and is successful. The '−' in the top of the $(m_1 = 0, m_2 = 1)$ box indicates that arm 2 should be selected again. Suppose it gives a second success, then it should be used a third time. If it now gives

a failure, arm 1 should be used at stage 4 even though the probability of success on arm 2 is $(s_2 + 1)/(m_2 + 2) = 3/5$ while on arm 1 it is $(s_1 + 1)/(m_1 + 2) = 1/2$.

Except for considerations of symmetry, which are available occasionally, there is no easy calculation which indicates an optimal arm at any stage – there is no alternative to an extensive description such as that in Table 2.2. Still, there are some attributes of optimal strategies that are evinced by Table 2.2 and that generalize to uniform discounting with independent arms. These will be discussed in Chapter 7. □

The next example has the same arms as the previous one, but the discount sequence is very different.

Example 2.4.4 As in the previous example, suppose $k = 2$ with $F_i = U(0, 1)$, $i = 1, 2$. Let $\mathbf{A} = (0, 0, 0, 0, 1, 0, 0, \ldots)$; so the objective is to maximize the expected value of the fifth observation. There are four learning observations in preparation for the only observation whose payoff counts.

The value of this bandit is 37/60. It is obvious that the fifth observation will be taken on the arm with greater current expected value. One might guess in view of symmetry that any nonsequential strategy (one that ignores the observations) that selects both arms twice is optimal. There are many optimal strategies, and some are nonsequential, but this strategy is not one of them. Any choices whatever can be made at stages 1, 2, and 3. For example, arm 1 can be selected the first three times; arm 2 must then be chosen at stage 4 if $z_1 + z_2 + z_3$ is either 1 or 2 and both arms are optimal at stage 4 if this sum is 0 or 3. This makes it clear that the set of optimal strategies includes those nonsequential strategies that select, in any order, one of the arms three times and the other once. There is no optimal strategy, sequential or not, that with probability one indicates two selections on each arm. Pearson (1980, Theorem 4.2.1 and its proof) shows that if an even number n of learning observations on these two arms are available, then when restricted to nonsequential strategies, it is uniquely optimal to allocate $(n/2) - 1$ to either arm and $(n/2) + 1$ to the other. □

Bandit problems frequently involve a trade-off between immediate payoff and gaining information. Since more information is available

later in the experiment than earlier, one might expect a greater contribution to the value of a bandit from later stages than from earlier stages; at least when following an optimal strategy (cf. Section 1.2). While this is true in a variety of settings (see Corollary 5.2.3, for example), it is not true in general, as the next example shows.

Example 2.4.5 Suppose there are two independent Bernoulli arms with parameters θ_1 and θ_2. Where F_1 and F_2 are viewed as distributions of θ_1 and θ_2, let $F_1 = \frac{2}{9}\delta_1 + \frac{7}{9}\delta_0$ and $F_2 = \frac{1}{12}\delta_1 + \frac{3}{4}\delta_{1/2} + \frac{1}{6}\delta_0$. The discount sequence is $\mathbf{A} = (1,1,1,0,0,\ldots)$. An optimal strategy τ (uniquely optimal through stage 3) is as follows:

$$\tau(\varnothing) = 2,$$
$$\tau(1) = 2,\ \tau(0) = 1,$$
$$\tau(1,1) = \tau(1,0) = 2,\ \tau(0,1) = 1,\ \tau(0,0) = 2.$$

The worth of τ and the value of this bandit equals $E_\tau(Z_1 + Z_2 + Z_3|G)$ where

$$E_\tau(Z_1|G) = E(\theta_2|G) = \frac{198}{432}$$

$$E_\tau(Z_2|G) = E(\theta_2^2 + (1-\theta_2)\theta_1|G) = \frac{169}{432}$$

$$E_\tau(Z_3|G) = E(\theta_2^3 + \theta_2^2(1-\theta_2) + (1-\theta_2)\theta_1^2 + (1-\theta_2)(1-\theta_1)\theta_2|G) = \frac{232}{432}.$$

The total is 599/432; but the interesting aspect of these calculations is that the expected contribution from the second stage is less than that from the first! □

The above example indicates that $E_\tau(Z_m|G)$ is not necessarily increasing in m when τ is an optimal strategy. However, the next result shows that the last observation (at stage n) is at least as large as every other observation in expectation. Its proof is due to Eick (1984).

Theorem 2.4.1 If the horizon of $\mathbf{A}$ is n then for any G, $\mathbf{A}$, optimal strategy τ, and $m < n$,

$$E_\tau(Z_n|G) \geq E_\tau(Z_m|G).$$

Proof Suppose τ is an optimal strategy with

$$E_\tau(Z_n|G) < E_\tau(Z_m|G)$$

for some $m < n$. Let τ^* be a modification of τ that imitates τ up to stage n and at stage n selects the arm indicated by τ at stage m. That is,

$$\tau^*(z_1, \ldots, z_j) = \tau(z_1, \ldots, z_j)$$

for all $(z_1, \ldots, z_j)$ with $j = 0, 1, \ldots, n-2$, and

$$\tau^*(z_1, \ldots, z_{m-1}, \ldots, z_{n-1}) = \tau(z_1, \ldots, z_{m-1})$$

for all $(z_1, \ldots, z_{n-1})$. (Continuations beyond stage n are of course irrelevant.) Then

$$E_{\tau^*}(Z_j|G) = E_\tau(Z_j|G)$$

for $j = 1, 2, \ldots, n-1$, and

$$E_{\tau^*}(Z_n|G) = E_\tau(Z_m|G) > E_\tau(Z_n|G).$$

So τ cannot be optimal. □

The next example illustrates how the result on one arm can affect the choice among the other arms, even when the arms are independent.

Example 2.4.6 Suppose there are 3 independent arms, so $G = F_1 \times F_2 \times F_3$, and $\mathbf{A} = (1, 1, 1, 0, 0, \ldots)$. Arms 1 and 2 are Bernoulli arms (with two-point prior distributions on the parameters). Arm 3 yields only 0's, only $\frac{1}{2}$'s, or only 1's. Since there is a non-Bernoulli arm, we will retreat from the convention used in previous examples of regarding F_i to be a probability distribution on $[0, 1]$. The distribution $G = F_1 \times F_2 \times F_3$ on $\mathscr{D}^3$ is given by:

$$F_1(\{\delta_1\}) = 17/57, \qquad F_1\left(\left\{\frac{19}{20}\delta_0 + \frac{1}{20}\delta_1\right\}\right) = 40/57,$$

$$F_2(\{\delta_1\}) = 107/470, \qquad F_2\left(\left\{\frac{47}{66}\delta_0 + \frac{19}{66}\delta_1\right\}\right) = 363/470,$$

$$F_3(\{\delta_1\}) = 9/20, \qquad F_3(\{\delta_{1/2}\}) = 1/10, \qquad F_3(\{\delta_0\}) = 9/20.$$

An optimal strategy τ, uniquely optimal in the first three stages, is as follows:

$$\tau(\varnothing) = 3$$
$$\tau(1) = 3,\ \tau(\tfrac{1}{2}) = 2,\ \tau(0) = 1,$$
$$\tau(1,1) = 3,\ \tau(\tfrac{1}{2},1) = 2,\ \tau(\tfrac{1}{2},0) = 3,\ \tau(0,1) = 1,\ \tau(0,0) = 2;$$

and $V(G;\mathbf{A}) = 1153/600$. The interesting feature of this optimal strategy is that the choice between arms 1 and 2 at stage 2 depends on the result on arm 3 at stage 1, even though the arms are independent. □

The next example generalizes Example 1.2.1 in a number of ways. The setting in which only one arm has unknown characteristics is generalized in Chapter 5.

Example 2.4.7 Suppose that there are two independent arms so that $G = F_1 \times F_2$. Further, suppose as in Section 2.1 that, for some constant λ,

$$F_2\{Q_2 : \int_R x Q_2(\mathrm{d}x) = \lambda\} = 1,$$

that is, that the random distribution on arm 2 has mean λ with probability one: we frequently label such an arm as 'known'. With no loss we may assume that each observation on arm 2 equals λ. We suppose that arm 1 is Bernoulli and thus consider F_1 to be the distribution of the Bernoulli parameter θ_1. We assume $0 < \lambda < 1$; if not, the problem is trivial. We assume that F_1 is supported by $[0,\lambda] \cup \{1\}$.

Recall that $|A^{(m)}|_1 = \sum_{p=m+1}^{\infty} \alpha_p$. We assume

$$\frac{|\mathbf{A}^{(m+1)}|_1}{|\mathbf{A}^{(m)}|_1} \leqslant \frac{|\mathbf{A}^{(m)}|_1}{|\mathbf{A}^{(m-1)}|_1} \tag{2.4.8}$$

whenever $|\mathbf{A}^{(m)}|_1 \neq 0$.

The assumption in this section that the horizon is finite is not necessary for this example.

Since φF_1 is supported by $[0,\lambda]$, selections of arm 2 are always optimal following a failure on arm 1. What is not so obvious is that it is

always optimal to continue selecting arm 1 following a success if the selection of arm 1 was optimal. Also, it is always optimal to continue selecting arm 2 once it has been selected if that selection was optimal. These assertions may be incorrect if (2.4.8) is not satisfied; they are verified for general F_1 in Theorem 5.2.2 in case (2.4.8) holds.

It is easy to calculate the value and find optimal strategies, for there are only two strategies that satisfy the above assertions – namely, τ_2: select arm 2 indefinitely, and τ_1: select arm 1 until (if ever) a failure is observed and select arm 2 thereafter. From these assertions,

$$V(F_1, \lambda; \mathbf{A}) = W(F_1, \lambda; \mathbf{A}; \tau_1) \vee W(F_1, \lambda; \mathbf{A}; \tau_2)$$

$$= \left(\sum_{m=1}^{\infty} \alpha_m \int_{[0,1]} [u_1^m + (1 - u_1^{m-1})\lambda] F_1(\mathrm{d}u_1) \right) \vee \left(|\mathbf{A}|_1 \lambda \right)$$

Furthermore, arm 1 is uniquely optimal if

$$\lambda < \frac{\sum_{m=1}^{\infty} \alpha_m E(\theta_1^m | F_1)}{\sum_{m=1}^{\infty} \alpha_m E(\theta_1^{m-1} | F_1)}, \tag{2.4.9}$$

arm 2 is uniquely optimal if the reverse inequality holds, and both are optimal if (and only if) equality holds.

Table 2.3 gives the right-hand side of (2.4.9) for two particular distributions and for a family of discount sequences. Namely, assume $F_1 = pU(0, 1/2) + (1-p)\delta_1$, where $p = 1/2$ or $9/10$, $U(0, 1/2)$ is the uniform distribution on $(0, 1/2)$, and $\mathbf{A}$ is the n-horizon uniform sequence. For large n, this quantity is approximately $1 - p/[(1-p)n]$.

Table 2.3 *Right-hand side of* (2.4.9)

n	1	2	5	10	20	50	100	200	500	1000
$p = 1/2$	0.625	0.718	0.844	0.912	0.953	0.981	0.990	0.995	0.998	0.999
$p = 9/10$	0.325	0.377	0.485	0.600	0.723	0.856	0.920	0.958	0.982	0.991

□

When the horizon is finite, the proof of the existence of optimal strategies, given in Section 2.3, was explicit and allowed us to calculate

optimal strategies in the examples of this section. The general existence proof given in the next section is not so constructive.

2.5 Existence of an optimal strategy

The purpose of this section is to extend Lemma 2.3.1 by showing that there exists an optimal strategy for all G and any discount sequence $\mathbf{A}$. We begin with a theorem that asserts continuity in $\mathbf{A}$. This theorem is useful when approximating arbitrary discount sequences by those having finite horizons. (Recall that $|\mathbf{A}|_1 = \sum_{m=1}^{\infty} \alpha_m$.)

Theorem 2.5.1 The function $(G, \mathbf{A}) \mapsto V(G; \mathbf{A})$ is measurable and satisfies

$$\begin{aligned} -\infty < \bigvee_{i=1}^{k} E(X_{i1}|G)|\mathbf{A}|_1 \leqslant V(G;\mathbf{A}) \\ \leqslant E\left(\bigvee_{i=1}^{k} E(X_{i1}|Q_i)\middle|G_i\right)|\mathbf{A}|_1 \leqslant E\left(\bigvee_{i=1}^{k} X_{i1}\middle|G\right)|\mathbf{A}|_1 < \infty \end{aligned} \tag{2.5.1}$$

and

$$V(G;\mathbf{A}) = \bigvee_{i=1}^{k} V^{(i)}(G;\mathbf{A}), \tag{2.5.2}$$

where

$$V^{(i)}(G;\mathbf{A}) = E(\alpha_1 X_{i1} + V((X_{i1})_i G; \mathbf{A}^{(1)})|G).$$

For each $G \in \mathscr{D}^*(\mathscr{D}^k)$, the function $V(G;\cdot)$ is uniformly continuous.

Proof The finiteness inequalities in (2.5.1) follow from the fact that $G \in \mathscr{D}^*(D^k)$ and from the inequality

$$E\left(\bigvee_{i=1}^{k} X_{i1}\middle|G\right) \leqslant E\left(\sum_{i=1}^{k} |X_{i1}|\middle|G\right) = \sum_{i=1}^{k} E(|X_{i1}||G).$$

The second inequality follows from

$$W(G;\mathbf{A};\tau_i) = E(X_{i1}|G)|\mathbf{A}|_1,$$

where τ_i denotes the strategy: always select arm i. To prove the third

inequality in (2.5.1) we proceed as follows:

$$
\begin{aligned}
E_\tau(Z_m|G) &= E(X_{\tau(Z_1,\ldots,Z_{m-1}),m}|G) \\
&= \sum_{i=1}^{k} E\left(\mathbf{1}_{\{\tau(Z_1,\ldots,Z_{m-1})=i\}} X_{im}\middle|G\right) \\
&= \sum_{i=1}^{k} E\Big(E(\mathbf{1}_{\{\tau(Z_1,\ldots,Z_{m-1})=i\}} X_{im}|Q_1,\ldots,Q_k, \\
&\qquad\qquad Z_1,\ldots,Z_{m-1})\Big|G\Big) \\
&= \sum_{i=1}^{k} E\Big(\mathbf{1}_{\{\tau(Z_1,\ldots,Z_{m-1})=i\}} \mathbf{E}(X_{im}|Q_1,\ldots,Q_k, \\
&\qquad\qquad Z_1,\ldots,Z_{m-1})|G\Big) \\
&\leqslant \sum_{i=1}^{k} E\left(\mathbf{1}_{\{\tau(Z_1,\ldots,Z_{m-1})=i\}} \bigvee_{j=1}^{k} E(X_{jm}|Q_1,\ldots,Q_k)|G\right) \\
&= E\left(\bigvee_{j=1}^{k} E(X_{jm}|Q_j)\middle|G\right) = E\left(\bigvee_{j=1}^{k} E(X_{j1}|Q_j)\middle|G\right).
\end{aligned}
\tag{2.5.3}
$$

The fourth inequality in (2.5.1) follows from

$$
\begin{aligned}
\bigvee_{i=1}^{k} E(X_{i1}|Q_i) &= \bigvee_{i=1}^{k} E(X_{i1}|Q_1,\ldots,Q_k) \\
&\leqslant E\left(\bigvee_{i=1}^{k} X_{i1}\middle|Q_1,\ldots,Q_k\right).
\end{aligned}
$$

To show uniform continuity in $\mathbf{A}$, fix G and let $\varepsilon > 0$. Let $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$ and $\mathbf{B} = (\beta_1, \beta_2, \ldots)$ be discount sequences satisfying

$$
|\mathbf{A} - \mathbf{B}|_1 < \varepsilon / E\left(\bigvee_{i=1}^{k} |X_{i1}|\middle|G\right).
$$

For any strategy τ,

$$
\begin{aligned}
|W(G;\mathbf{A};\tau) - W(G;\mathbf{B};\tau)| &= \left|\sum_{m=1}^{\infty} (\alpha_m - \beta_m) E_\tau(Z_m|G)\right| \\
&\leqslant \sum_{m=1}^{\infty} |\alpha_m - \beta_m| E_\tau(|Z_m|\big|G) \leqslant |\mathbf{A} - \mathbf{B}|_1 E\left(\bigvee_{i=1}^{k} |X_{i1}|\middle|G\right) < \varepsilon.
\end{aligned}
$$

Hence, for fixed G, $W(G; \mathbf{A}; \tau)$ depends continuously on $\mathbf{A}$, uniformly in $\mathbf{A}$ and τ. The uniform continuity of $V(G; \cdot)$ then follows from its definition (2.2.2).

The function $(G, \mathbf{A}) \mapsto (G, \mathbf{A}_n)$, where $\mathbf{A}_n = (\alpha_1, \alpha_2, \ldots, \alpha_n, 0, 0, 0, \ldots)$ when $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$, is measurable (in fact, continuous), so, by Lemma 2.3.1, $(G, \mathbf{A}) \mapsto V(G; \mathbf{A}_n)$ is measurable. The pointwise limit as $n \to \infty$ of these functions is the function $(G, \mathbf{A}) \mapsto V(G; \mathbf{A})$ which is, therefore, measurable. Equality (2.5.2) follows from its finite horizon version (2.3.1) and Lebesgue's dominated convergence theorem, the applicability of which follows from (2.5.1). □

Remarks Suppose strategies were permitted to depend on $Q_1, \ldots, Q_k$; that is, that the decision maker is told the distributions $Q_1, \ldots, Q_k$ before stage 1. The best such 'strategy' would be the one that always selects the arm that maximizes $E(X_{i1} | Q_i)$ and its worth would equal $E\left(\bigvee_{i=1}^{k} E(X_{i1} | Q_i) \middle| G \right) |\mathbf{A}|_1$ (cf. (2.5.1)).

As in Section 2.3, (2.5.2) can be rewritten as (2.3.2). □

The following example shows that V is not continuous in G.

Example 2.5.1 Let $\mathbf{A} = (1, 1, 0, 0, 0, \ldots)$, $k = 2$, and for $l = 2, 3, \ldots$, $G_l = F_{1l} \times F_{2l}$ where, under F_{2l}, observations on arm 2 always equal 0 and F_{1l} has atoms of size 1/2 at each of the following two Q_1's:

$$Q_1(\{-l^{-1}\}) = Q_1(\{-1\}) = 1/2,$$
$$Q_1(\{l^{-1}\}) = Q_1(\{1\}) = 1/2.$$

Clearly an optimal strategy is to select arm 1 at stage 1 and then, at stage 2, select arm 1 if a positive result was observed and select arm 2 otherwise. An easy calculation gives

$$V(G_l; \mathbf{A}) = (1 + l^{-1})/4 \to 1/4 \text{ as } l \to \infty.$$

As $l \to \infty$, $G_l \to G = F_1 \times F_2$ where, under F_2, observations on arm 2 always equal 0 and F_1 has atoms of size 1/2 at each of the following two Q_1's:

$$Q_1(\{0\}) = Q_1(\{-1\}) = 1/2,$$
$$Q_1(\{0\}) = Q_1(\{+1\}) = 1/2.$$

Clearly an optimal strategy is to select arm 1 at stage 1 and then, at

stage 2, select arm 1 if $+1$ was observed at stage 1 and otherwise select arm 2. An easy calculation gives $V(G; \mathbf{A}) = 1/8 < 1/4$. □

The reason for this lack of continuity in the preceding example is that the limit G hides information much better than do the G_l, even for large l. It is difficult to imagine an example where 'better' could be replaced by 'worse'. Accordingly, for fixed $\mathbf{A}$ we conjecture that $V(\cdot; \mathbf{A})$ is a lower semicontinuous function on $\{G: P(X_{im} < -c|G) = 0 \text{ for some } c < \infty \text{ and each } i\}$.

The next theorem asserts the existence of an optimal strategy. This has already been proved (Lemma 2.3.1) in case the horizon is finite. The proof was accomplished by constructing optimal strategies recursively: first for bandits having horizon 0, then for those having horizon 1, then for those having horizon 2, etc. Then (2.5.2) was obtained for the finite horizon case. By going to the limit we have obtained (in Theorem 2.5.1) the same formula for V in the infinite horizon case. In the infinite horizon case this formula cannot be viewed as a recursion on the horizon since both $\mathbf{A}$ and $\mathbf{A}^{(1)}$ have infinite horizon, but it can be viewed as being recursive on the stages. Accordingly, we will use (2.5.2) in the proof of the next theorem to define optimal strategies stage-by-stage beginning from stage 1. This definition will not be very constructive since to carry it out in a specific example one needs to know $V(G; \mathbf{A})$ for a wide variety of bandits. This stage-by-stage definition will give an optimal strategy for every bandit; in the finite horizon case it will be the same strategy that was constructed in the proof of Lemma 2.3.1.

Theorem 2.5.2 There exists a strategy $\tau_{G,\mathbf{A}}$ for each $\mathbf{A} \in \mathscr{A}$ and each $G \in \mathscr{D}^*(\mathscr{D}^k)$ such that

$$\{(G; \mathbf{A}; z_1, \ldots, z_{m-1}): \tau_{G,\mathbf{A}}(z_1, \ldots, z_{m-1}) = i\} \tag{2.5.4}$$

is a measurable subset of $\mathscr{D}^*(\mathscr{Q}^{k_1} \times \mathscr{A} \times (-\infty, \infty)^{m-1}$ for each $i = 1, \ldots, k$ and $m = 1, 2, \ldots$, and $\tau_{G,\mathbf{A}}$ is optimal for the $(G; \mathbf{A})$-bandit.

Proof Define $\tau_{G,\mathbf{A}}$ recursively as follows: $\tau_{G,\mathbf{A}}(\varnothing)$ is the smallest i for which the maximum in (2.5.2) is attained and, for $m > 1$,

$$\tau_{G,\mathbf{A}}(z_1, \ldots, z_{m-1}) = \tau_{G^{(1)},\mathbf{A}^{(1)}}(z_2, \ldots, z_{m-1}) \tag{2.5.5}$$

where

$$G^{(1)} = (z_1)_{\tau_{G,\mathbf{A}}(\varnothing)} G.$$

The measurability of (2.5.4) follows immediately by an induction argument on m using Lemma 2.2.1 to deal with $G^{(1)}$ in (2.5.5).

For the remainder of the proof we write τ for $\tau_{G,\mathbf{A}}$. From (2.5.2) and the definition of τ we conclude that

$$V(G;\mathbf{A}) = E_\tau(\alpha_1 Z_1 + V(G^{(1)};\mathbf{A}^{(1)})|G)$$

and, by induction,

$$V(G;\mathbf{A}) = E_\tau\left(\sum_{m=1}^{n} \alpha_m Z_m + V(G^{(n)};\mathbf{A}^{(n)})\middle| G\right).$$

Since

$$W(G;\mathbf{A};\tau) = E_\tau\left(\sum_{m=1}^{\infty} \alpha_m Z_m \middle| G\right),$$

we can complete the proof by showing

$$\lim_{n\to\infty} E_\tau\left(\sum_{m=n+1}^{\infty} \alpha_m Z_m - V(G^{(n)};\mathbf{A}^{(n)})\middle| G\right) = 0. \tag{2.5.6}$$

We have

$$\begin{aligned}\left|E_\tau\left(\sum_{m=n+1}^{\infty} \alpha_m Z_m \middle| G\right)\right| &\leqslant \sum_{m=n+1}^{\infty} \alpha_m E_\tau\left(|Z_m|\middle| G\right)\\ &\leqslant \sum_{m=n+1}^{\infty} \alpha_m E\left(\bigvee_{i=1}^{k} |X_{im}|\middle| G\right)\\ &= E\left(\bigvee_{i=1}^{k} |X_{i1}|\middle| G\right)|\mathbf{A}^{(n)}|_1 \to 0.\end{aligned}$$

From (2.5.1),

$$|V(G^{(n)};\mathbf{A}^{(n)})| \leqslant E\left(\bigvee_{i=1}^{k} |X_{i1}|\middle| G^{(n)}\right)|\mathbf{A}^{(n)}|_1$$

and, hence

$$E_\tau(|V(G^{(n)};\mathbf{A}^{(n)})|\big|G) \leqslant E\left(\bigvee_{i=1}^{k} |X_{i1}|\middle| G\right)|\mathbf{A}^{(n)}|_1 \to 0.$$

Therefore, (2.5.6) holds. □

We immediately obtain

Corollary 2.5.3 Any i for which the maximum in (2.5.2) is attained is an optimal initial selection in the $(G; \mathbf{A})$-bandit, and conversely.

Remark The requirement in the preceding proof that $\tau_{G,\mathbf{A}}(\varnothing)$ be chosen to yield the maximum in (2.5.2) gives $\tau_{G,\mathbf{A}}$ the property (used in gambling theory literature; e.g. Dubins and Savage (1976)) of being *conserving* at stage 1. The recursive relation (2.5.5) makes $\tau_{G,\mathbf{A}}$ conserving at every stage – that is, *thrifty*. We could have relied on a general theorem to conclude that for our context any thrifty strategy is optimal. Instead we adapted arguments used in the proofs of such general theorems, which, incidentally, have hypotheses in addition to thriftiness. □

Although optimal strategies and V are the main objects of study, the behaviour of the function W may be of some interest.

Theorem 2.5.4 For each fixed τ, the function $(G, \mathbf{A}) \mapsto W(G; \mathbf{A}; \tau)$ is measurable. For each fixed G, the function $\mathbf{A} \mapsto W(G; \mathbf{A}; \tau)$ is continuous, uniformly in $\mathbf{A}$ and τ.

Proof The proof of uniform continuity is contained in the proof of Theorem 2.5.1.

To prove measurability it suffices to prove measurability of the function $G \mapsto E_\tau(Z_m|G)$ for each m. But this measurability follows easily from

$$Z_m = \sum_{i=1}^{k} X_{im} \mathbf{1}_{\{\tau(Z_1, \ldots, Z_{m-1}) = i\}}$$

and an application of Lemma 2.2.3 using approximations of each summand by bounded random variables. □

While optimal strategies exist generally, finding them can be difficult when the horizon is infinite. One route is to find V; we turn to approximating V in the infinite horizon case.

2.6 Approximating value functions for infinite horizons

A method for finding an upper or lower bound for the function $V(\cdot; \mathbf{A})$ is to replace $V(\cdot; \mathbf{A}^{(1)})$ with an upper or lower bound in (2.2.5). The same statement applies as well for $V(\cdot; \mathbf{A}^{(m)})$ and $V(\cdot; \mathbf{A}^{(m+1)})$ for

all integers $m \geqslant 0$. Therefore, upper and lower bounds for $V(\cdot; \mathbf{A}^{(n)})$ can be used via dynamic programming to find bounds for $V(\cdot; \mathbf{A})$.

Two bounds are given by Theorem 2.5.1. Letting $(G^{(n)}; \mathbf{A}^{(n)})$ play the role of $(G; \mathbf{A})$, we have

$$\bigvee_{i=1}^{k} E(X_{i1} \Big| G^{(n)}) |\mathbf{A}^{(n)}|_1 \leqslant V(G^{(n)}; \mathbf{A}^{(n)}) \leqslant E\left(\bigvee_{i=1}^{k} E(X_{i1} | Q_i) \Big| G^{(n)} \right) |\mathbf{A}^{(n)}|_1 \tag{2.6.1}$$

Let $L_n(G; \mathbf{A})$ denote the approximation for $V(G; \mathbf{A})$ obtained by dynamic programming using the left side of (2.6.1) to begin the backward induction; thus, $L_n(G; \mathbf{A})$ equals $V(G; \mathbf{A}^*)$ where $\mathbf{A}^* = (\alpha_1, \alpha_2, \ldots, \alpha_n, |\mathbf{A}^{(n)}|_1, 0, 0, \ldots)$. Similarly, let $U_n(G; \mathbf{A})$ denote the approximation obtained using the right side of (2.6.1) as starting values. The next theorem indicates that V is bounded by each L_n and each U_n; in addition, it gives a very crude bound for the difference between L_n and U_n.

Theorem 2.6.1 For all $G \in \mathscr{D}(\mathscr{D}^k)$ and all discount sequences $\mathbf{A}$,

$$L_n(G; \mathbf{A}) \leqslant V(G; \mathbf{A}) \leqslant U_n(G; \mathbf{A}) \tag{2.6.2}$$

and

$$U_n(G; \mathbf{A}) - L_n(G; \mathbf{A}) \leqslant \left[E\left(\bigvee_{i=1}^{k} E(X_{i1} | Q_i) \Big| G \right) - \bigvee_{i=1}^{k} E(X_{i1} | G) \right] |\mathbf{A}^{(n)}|_1. \tag{2.6.3}$$

In particular,

$$\lim_{n \to \infty} [U_n(G; \mathbf{A}) - L_n(G; \mathbf{A})] = 0.$$

Proof The inequalities in (2.6.2) hold as indicated previously. The method of dynamic programming and analogues of (2.3.1) are appropriate for the following three modifications of the $(G; \mathbf{A})$-bandit. In the first, the experiment terminates at stage n and a (random) amount

$$\sum_{m=n+1}^{\infty} \left(\alpha_m \bigvee_{i=1}^{k} E(X_{im} | Q_i) \right) \tag{2.6.4}$$

is added to $\Sigma_{m=1}^{n} \alpha_m Z_m$. The value for the modification is $U_n(G; \mathbf{A})$.

In the second, after n stages the decision maker must select one arm to be used for all stages $m > n$. The value for this modification is clearly $L_n(G; \mathbf{A})$.

Finally, consider a third modification whose value is obviously no larger than $L_n(G; \mathbf{A})$. The decision maker must, before stage 1, select one arm to be used for all stages $m > n$. Of course, that choice will be a j for which

$$E(X_{j1}|G) = \bigvee_{i=1}^{k} E(X_{i1}|G).$$

Let V_i denote the value for the ith modification. From (2.6.4) we see that

$$V_1(G; \mathbf{A}) \leqslant V_3(G; \mathbf{A})$$

$$+ |\mathbf{A}^{(n)}|_1 \left[E\left(\bigvee_{i=1}^{k} E(X_{i1}|Q_i) \middle| G \right) - \bigvee_{i=1}^{k} E(X_{i1}|G) \right]. \quad \square \quad (2.6.5)$$

Remarks This theorem can be used in two ways to decide on a truncation stage in order to approximate V with any desired accuracy. First, n can be chosen so that the right side of (2.6.3) is sufficiently small. Second, $U_n - L_n$ can be calculated for various values of n until the desired accuracy is attained. While the second can involve a number of dynamic programs, the total calculation time is usually less because the bound provided by (2.6.3) is so crude. $\square$

Example 2.6.1 Let $\mathbf{A}$ be geometric: $\mathbf{A} = (1, 0.9, (0.9)^2, \ldots)$. Suppose there are two Bernoulli arms. Assume θ_1 is uniformly distributed on $[0, 1]$ and $\theta_2 = 0.6$.

To use (2.6.3) we calculate

$$E\Big(E(X_{i1}|\theta_1) \vee E(X_{i2}|\theta_2)\Big) = \int_0^1 (u_1 \vee 0.6)\,du_1 = 0.68,$$

$$E(X_{11}|G) \vee E(X_{21}|G) = 0.5 \vee 0.6 = 0.6,$$

and

$$|\mathbf{A}^{(n)}|_1 = \sum_{m=n+1}^{\infty} (0.9)^{m-1} = 10(0.9)^n.$$

Suppose we want to approximate $V(G; \mathbf{A})$ with an error no greater than 0.005. Accordingly, we require the right-hand side of (2.6.3) to be no larger than 0.01:

$$(0.08)(10)(0.9)^n \leqslant 0.01.$$

The smallest n that suffices is 42. Dynamic programming gives $L_{42}(G; \mathbf{A}) \approx 6.38827$ and $U_{42}(G; \mathbf{A}) \approx 6.39020$. The average, 6.38923, is an approximation as desired and, in fact, is in error by no more than 0.001. In carrying out the two dynamic programs only one strategy is obtained: the one derived in calculating $L_{42}(G; \mathbf{A})$. In view of (2.6.3) we know that its worth is within 0.01 of the value. Having calculated $U_{42}(G; \mathbf{A})$ we know that $V - L_{42} \leqslant 0.002$. In view of the results described below it turns out that $V - L_{42} \approx 4.3 \times 10^{-6}$.

Proceeding without reference to (2.6.3) we can generate selected L_n and U_n as indicated in Table 2.4, stopping when desired accuracy is achieved. For example, we would stop at $n = 28$ if we want our estimate to be in error by less than 0.005.

Table 2.4

n	$L_n(G; \mathbf{A})$	$U_n(g; \mathbf{A})$	$U_n - L_n$
1	6.2	6.72	0.52
2	6.335	6.648	0.313
3	6.335	6.5832	0.2482
4	6.36524	6.53183	0.16659
5	6.3736756	6.514334	0.1406584
10	6.3855284	6.4536415	0.0681131
15	6.3874732	6.4249362	0.0374630
20	6.3880147	6.4095730	0.0215583
28	6.3882272	6.3972135	0.0089863
50	6.3882686	6.3890906	0.0008220
100	6.3882698	6.3882740	0.0000042
150	6.3882698	6.3882698	0.0000000

It is evident that neither of these approaches is as good as using L_n for modest n. For example, L_7 is about as accurate as $(L_{50} + U_{50})/2$. However, we do not have a way of assessing error using the sequence $\{L_n\}$ alone. □

References

Bradt, R. N., Johnson, S. M. and Karlin, S. (1956) On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.* **27**: 1060–1074.

Chow, Y. S. and Teicher, H. (1978) *Probability Theory*, Springer-Verlag, New York.

DeGroot, M. H. (1970) *Optimal Statistical Decisions*, McGraw-Hill, New York.

Dubins, L. and Freedman, D. (1964) Measurable sets of measures, *Pac. J. Math.* **14**: 1211–1222.

Dubins, L. E. and Savage, L. J. (1976) *Inequalities for Stochastic Processes: How to Gamble If You Must*, Dover, New York.

Eick, S. G. (1984) Personal communication.

Parthasarathy, K. R. (1967) *Probability Measures on Metric Spaces*, Academic Press, New York.

Pearson, L. M. (1980) Treatment allocation for clinical trials in stages. Ph.D. thesis, Univ. of Minnesota, USA.

Rhenius, D. (1977) Faktorisierung von übergangswahrscheinlichkeiten in Markoffschen Lernmodellen. *Arbeiten aus den Psychologischen Instituten der Univ. Hamburg Nr.* 44.

Varadhan, S. R. S. (1983) Personal communication via N. C. Jain.

CHAPTER 3

The discount sequence

The particular discount sequence plays a critical role in any bandit or other decision problem. Various interpretations of discount sequences are discussed in this chapter. One purpose of the discussion is to aid a user in choosing an appropriate sequence; another is to motivate interest in the generality of discounting allowed in this monograph.

As indicated in Chapter 1, much of the bandit literature concerns maximizing the sum of n observations. In our notation the corresponding discount sequence is the n-horizon uniform sequence: $\alpha_m = 1$ for $m \leqslant n$ and $\alpha_m = 0$ for $m > n$ (see Chapter 7 for an extensive treatment of this case). The other important discount sequence in the literature is the geometric: $\alpha_m = \alpha^{m-1}$ for some $\alpha \in (0, 1)$ (see Chapter 6). There has been very little discussion in the literature of other cases, and the question arises as to whether our more general approach has any practical relevance. This chapter gives an affirmative answer. Some of the ideas in this chapter are considered in Berry (1983).

The possibility that the discount sequence is unknown is considered in the first four sections of this chapter. Random uniform discount sequences are discussed in Section 3.1. The next three sections are devoted to more general random discount sequences. An important consideration is whether or not learning about the unknown discount sequence is possible as the experiment develops. The two cases of observing and not observing the discount factors are both considered. The orientation throughout the first four sections is to identify hypotheses under which random discount factors may, without loss, be replaced by their expectations.

Section 3.5 deals with nonrandom real-time discount sequences. Time is continuous and times at which trials (that is, stages) occur are random.

Section 3.6 motivates discount sequences that are not monotone. The important special case $(0, 0, \ldots, 0, 1, 0, 0, \ldots)$ is discussed.

3.1 Mixtures of uniform sequences

Consider Bernoulli trials and suppose a success at any stage is worth 1, as in uniform discounting. Also suppose the experiment may terminate at any stage for reasons outside the decision maker's control; let η_m be the probability of termination with the mth trial, $m = 1, 2, \ldots$. For example, in a clinical trial the number of patients in the population to be treated may not be precisely known. Or, a new arm that is clearly better than any in the trial may be discovered at any time. The true discount sequence may be an n-horizon uniform, all successes being equally valued, but n is not known.

This situation can be placed in the framework of Chapter 2 as follows. As in Chapter 2 we assume that there is an infinite sequence of trials. We account for the fact that the experiment may have terminated before a given stage by specifying that a success at that stage is worth only the probability that the experiment has not yet terminated; that is, the appropriate discount sequence $(\alpha_1, \alpha_2, \ldots)$ is given by

$$\alpha_m = \sum_{l=m}^{\infty} \eta_l. \tag{3.1.1}$$

Any nonincreasing discount sequence can arise in this manner by normalizing to $\alpha_1 = 1$. However, it may happen that $(\alpha_1, \alpha_2, \ldots)$ given by (3.1.1) is not a discount sequence because $\Sigma \alpha_m = \Sigma l \eta_l = \infty$.

Assume the length of the trial is independent of the strategy followed and responses obtained. Then the setting is precisely the one described in the first two chapters. In particular, (2.2.2) applies where α_m is defined in (3.1.1). Though the discount sequence is random since it is a mixture of uniforms, the problem is identical to one in which the discount sequence is deterministic—namely, an average of uniforms weighted by the η_m's.

For example, assume there is a constant probability of terminating at each stage given that that stage has been reached, so that

$$\eta_m = (1 - \alpha)\alpha^{m-1} \qquad \text{for some } \alpha \in (0, 1).$$

Then $\alpha_m = \alpha^{m-1}$, $m = 1, 2, \ldots$, and the appropriate discount sequence is geometric.

3.2 Random discount sequences

The previous section dealt with mixtures of uniform discount sequences. The generalization to mixtures of arbitrary sequences is considered in this section.

Results in general are considerably more complicated than in the uniform setting for an important reason: the character of the decision problem depends upon whether or not discount factors are observed at each stage. Mixtures of uniform sequences are special in this regard: conditioning on previous discount factors offers no advantage. A discount factor is 0 when it is not 1 in the uniform case, so it can be assumed to be 1 without loss – if it is 0 then any selection (and continuation) is of no consequence.

In the next section we discuss bandit problems in which the discount factors are not observed and, hence, strategies cannot depend on them. In Section 3.4, selections are allowed to depend on all previous discount factors. (A third possibility, one not discussed here, is that $\alpha_m Z_m$ is observed, but not α_m and Z_m individually.) In this section we shall discuss some facets common to both settings.

The decision maker's prior knowledge concerning the random discount sequence consists of a probability distribution $\mathbf{H}$ on the space $\mathscr{A}$ of all discount sequences. We assume

$$E\left(\sum_{m=1}^{\infty} \alpha_m \middle| \mathbf{H}\right) < \infty,$$

where $\mathbf{H}$ appears in the notation to make explicit the dependence of expected values on the distribution of the random discount sequence. Some expectations depend on both G and $\mathbf{H}$ – for instance, we write $E(\alpha_m X_{im} | G, \mathbf{H})$. We always assume the arms to be independent of the discount sequence, so we deal only with the product measure $G \times \mathbf{H}$. Accordingly,

$$E(\alpha_m X_{im} | G, \mathbf{H}) = E(\alpha_m | \mathbf{H})\, E(X_{im} | G).$$

However, in Section 3.4, $E_\tau(\alpha_m Z_m | G, \mathbf{H})$ may not factor; for τ and therefore Z_m may depend on $\alpha_1, \ldots, \alpha_{m-1}$.

For the purposes of this chapter we extend previous terminology and notation by speaking of the $(G; \mathbf{H})$-bandit and using $W(G; \mathbf{H}; \tau)$ and $V(G; \mathbf{H})$ instead of $W(G; \mathbf{A}; \tau)$ and $V(G; \mathbf{A})$. The value $V(G; \mathbf{H})$ depends on the class of strategies being considered. Therefore $V(G; \mathbf{H})$ will be no smaller in the setting of Section 3.4 than in that of

Section 3.3. For either of the two cases, the important equation (2.5.2) is valid when modified appropriately:

$$V(G; \mathbf{H}) = \bigvee_{i=1}^{k} E(\alpha_1 X_{i1} + V((X_{i1})_i G; \mathbf{H}^{(1)}) | G, \mathbf{H}) \quad (3.2.1)$$

where $\mathbf{H}^{(1)}$ denotes the conditional distribution of the discount sequence $\mathbf{A}^{(1)} = (\alpha_2, \alpha_3, \ldots)$ given what has been observed at stage 1. Formula (3.2.1) means different things depending on context: in Section 3.4 we condition on α_1 to obtain $\mathbf{H}^{(1)}$ but in Section 3.3 we will only condition on being at the second stage. In general, (3.2.1) can be rewritten:

$$V(G; \mathbf{H}) = \bigvee_{i=1}^{k} [E(\alpha_1 | \mathbf{H}) E(X_{i1} | G) + E((V(X_{i1})_i G; \mathbf{H}^{(1)}) | G, \mathbf{H})]; \quad (3.2.2)$$

although, as will be seen in the next section, in the nonobservable case the simpler relation (2.5.2) applies with α_m replaced by its expected value under $\mathbf{H}$.

The topology we use on the space of all $\mathbf{H}$'s is the topology of convergence in distribution; the measurable sets of $\mathbf{H}$'s are the Borel sets.

3.3 Nonobservable discount factors

In this section strategies τ are defined as in Chapter 2. As such they are measurable with respect to the σ-field generated by $\{X_{im}: i = 1, \ldots, k; m = 1, 2, \ldots\}$. (This requirement on strategies will be expressed by saying that the discount factors are 'nonobservable'.) Hence, each Z_m is measurable with respect to this σ-field. On the other hand, the random discount sequence is independent of it. Therefore, for each τ,

$$\begin{aligned} W(G; \mathbf{H}; \tau) &= \sum_{m=1}^{\infty} E_\tau(\alpha_m Z_m | G, \mathbf{H}) \\ &= \sum_{m=1}^{\infty} E(\alpha_m | \mathbf{H}) E_\tau(Z_m | G) \\ &= E_\tau\left(\sum_{m=1}^{\infty} E(\alpha_m | \mathbf{H}) Z_m | G \right) \end{aligned}$$

which leads to (2.2.2) with $E(\alpha_m|\mathbf{H})$ playing the role of α_m. Thus we have proved the following result.

Theorem 3.3.1 Suppose a random discount sequence **A** governed by a distribution **H** is independent of $\{X_{im}: i = 1, \ldots, k; m = 1, 2, \ldots\}$ and the discount factors are not observable. Then, for any G, the $(G; \mathbf{H})$-bandit is equivalent to the $(G; E(\mathbf{A}|\mathbf{H}))$-bandit; that is,

$$W(G; \mathbf{H}; \tau) = W(G; E(\mathbf{A}|\mathbf{H}); \tau)$$

for all τ. In particular, the two bandits have the same value and the same set of optimal strategies. In addition, $V(G; \mathbf{H})$ and $W(G; \mathbf{H}; \tau)$ depend continuously on **H** and measurably on the pair $(G, \mathbf{H})$.

Example 3.3.1 Suppose $E(\mathbf{A}|\mathbf{H}) = (1, \ldots, 1, 0, \ldots)$. Then the discount sequence relevant for choosing a strategy is the finite horizon uniform. Chapter 7 and Examples 2.4.1 to 2.4.3 and 2.4.5 to 2.4.7 apply. □

Example 3.3.2 Suppose $E(\alpha_m|\mathbf{H}) = \alpha^{m-1}$ for known $\alpha > 0$. The relevant discount sequence is geometric. Chapter 6 applies. □

Example 3.3.3 Suppose **H** assigns probability $\frac{1}{2}$ to each of two geometric discount sequences: $(3/4, (3/4)^2, \ldots)$ and $(1/4, (1/4)^2, \ldots)$. Suppose that $k = 2$ and that, under G, both arms are Bernoulli with arm 2 having a known probability λ of success and arm 1 having probability 1 or 0 of success, each with probability 1/2. By Theorem 3.3.1 the $(G; \mathbf{H})$-bandit is equivalent to the $(G; \mathbf{A})$-bandit where

$$\mathbf{A} = \left(\frac{1}{2}, \frac{5}{16}, \ldots, \frac{1}{2}\left(\frac{3}{4}\right)^m + \frac{1}{2}\left(\frac{1}{4}\right)^m, \ldots\right).$$

Notice that (2.4.8) is not satisfied for this sequence for any $m > 1$ even though each component geometric sequence does satisfy (2.4.8).

Complete information is obtained with a single observation of arm 1. Accordingly, whenever arm 1 is selected it is to be used indefinitely thereafter if successful, and never again otherwise. So the only uncertainty in describing an optimal strategy is specifying when to select arm 1. Straightforward calculations show that there is a sequence $\{\lambda_i\}$ with $0.7273 \approx 8/11 = \lambda_0 < \lambda_1 < \lambda_2 < \ldots < \lambda_\infty = \lim \lambda_r = 4/5$, with the following interpretation. If $\lambda \in [0, \lambda_0]$ then arm 1 is optimal initially. If $\lambda \in [\lambda_0, \lambda_1]$ then it is optimal to select

arm 2 initially, followed by arm 1 at stage 2. Generally, if $\lambda \in [\lambda_{r-1}, \lambda_r]$ then it is optimal to select arm 2 at stages 1 through r, followed by arm 1 at stage $r+1$. If $\lambda \in [\lambda_\infty, 1]$ then arm 2 is optimal indefinitely and arm 1 should never be used.

When a nonrandom discount sequence is geometric, Theorem 5.2.2 applies to show that there is an optimal strategy that either begins with arm 1 or else never uses arm 1. This example shows that for a mixture of geometrics, such simplicity may be lost – after observing the known arm for some time, the decision maker may want to switch to the unknown arm despite having received no additional information. □

3.4 Observable discount factors

In this and the following section, and in these two sections only, a strategy τ depends on previous discount factors as well as on previous observations. Thus $\tau(z_1, z_2; \alpha_1, \alpha_2)$ indicates the arm to be observed at stage 3 if z_1 and z_2 are the outcomes on arms $\tau(\varnothing)$ and $\tau(z_1; \alpha_1)$, respectively, at stages 1 and 2 and α_1 and α_2 are the first two discount factors.

Example 3.4.1 We use the same G and $\mathbf{H}$ as in Example 3.3.3. As in that example, if arm 1 is selected initially then optimal continuations are clear. So suppose arm 2 is selected initially. At stage 2 the discount sequence becomes known – one of two geometrics. Since (2.4.8) is satisfied for geometric sequences, Example 2.4.7 applies for the bandit problem starting at stage 2.

Accordingly, one of the following three strategies is optimal:

τ_1: select arm 1 initially, then proceed optimally;

τ_2: select arm 2 always;

τ_{2^*}: select arm 2 initially; if $\alpha_1 = 1/4$, observe arm 2 thereafter, if $\alpha_1 = 3/4$, observe arm 1 and thereafter proceed optimally.

Easy calculations show:

$$W(G; \mathbf{H}; \tau_1) = 7\lambda/12 + 5/6,$$
$$W(G; \mathbf{H}; \tau_2) = 5\lambda/3,$$
$$W(G; \mathbf{H}; \tau_{2^*}) = 185\lambda/192 + 9/16.$$

Therefore τ_1 is optimal if $\lambda \leqslant 52/73 \approx 0.7123$, τ_{2^*} is optimal if $52/73 \leqslant \lambda \leqslant 4/5$, and τ_2 is optimal for $\lambda \geqslant 4/5$.

If $52/73 < \lambda < 4/5$, then the value is larger than that for the same bandit in the nonobservable setting. If λ is not in this interval then the values $\lambda \neq 52/73$ and $\lambda \neq 4/5$ are the same; if in addition then optimal strategies are the same. □

The $(G; \mathbf{H})$-bandit treated in Examples 3.3.3 and 3.4.1 is rather unusual in that the observable version is easier to solve than is the nonobservable version. The observable version is generally very difficult. The only general results we give for the observable case are the lemma and two theorems of this section.

Minor changes in the methods leading to Lemma 2.3.1, Theorem 2.5.1, and Theorem 2.5.2 lead to the following theorem and lemma.

Theorem 3.4.1 In the observable case, (3.2.1) holds. An optimal initial selection for an arbitrary $(G; \mathbf{H})$-bandit is given by the smallest i for which the maximum in (3.2.1) is attained. These optimal initial selections fit together to constitute optimal strategies $\tau_{G,\mathbf{H}}$ whose dependence on $(G, \mathbf{H})$ is measurable. The value function $(G, \mathbf{H}) \mapsto V(G; \mathbf{H})$ is measurable.

For the lemma we need a notation. As in Section 2.5 let $\mathbf{A}_n$ denote the discount sequence obtained from $\mathbf{A}$ by replacing all terms after the nth term by zeros. The mapping $\mathbf{A} \mapsto \mathbf{A}_n$ induces a mapping, say $\mathbf{H} \mapsto \mathbf{H}_n$, of probability measures on the space of discount sequences.

Lemma 3.4.2 For $\mathbf{H}_n$ defined as above, $V(G; \mathbf{H}_n) \mapsto V(G; \mathbf{H})$ as $n \to \infty$.

Despite the preceding lemma, V does not depend continuously on $\mathbf{H}$, as the following example shows.

Example 3.4.2 Let $\mathbf{H}$ assign probability $\frac{1}{2}$ to each of the discount sequences $(1, 0, 2, 0, 0, 0, \ldots)$ and $(1, 2, 0, 0, 0, 0, \ldots)$ and let $\mathbf{H}_n^*$ assign probability $\frac{1}{2}$ to each of the discount sequences $(1 - n^{-1}, 0, 2, 0, 0, 0, \ldots)$ and $(1 + n^{-1}, 2, 0, 0, 0, 0, \ldots)$. So $\mathbf{H}_n^* \to \mathbf{H}$ as $n \to \infty$. Suppose that under G, there are two independent Bernoulli arms having parameters θ_1 and θ_2 where θ_1 is uniformly distributed on $(0, 1)$ and $\theta_2 = 2/5$ with probability one.

Consider the $(G; \mathbf{H})$-bandit. No information concerning the discount sequence is available at stage 1. The fact that the decision maker learns the discount sequence at stage 2 is irrelevant since, in any case, an optimal selection at stage 3 is the arm with the higher current mean. So the $(G; \mathbf{H})$-bandit is equivalent to one with nonrandom discounting: $(G; (1, 1, 1, 0, 0, \ldots))$. The first three selections of an optimal strategy τ are as follows:

$$\tau(\varnothing) = 1,$$
$$\tau(1; 1) = 1,$$
$$\tau(1, 1; 1, 0) = \tau(1, 1; 1, 2) = \tau(1, 0; 1, 0) = \tau(1, 0; 1, 2) = 1,$$
$$\tau(0; 1) = 2,$$
$$\tau(0, 1; 1, 0) = \tau(0, 1; 1, 2) = \tau(0, 0; 1, 0) = \tau(0, 0; 1, 2) = 2.$$

An easy calculation gives $V(G; \mathbf{H}) = 94/60$.

Now consider the $(G; \mathbf{H}_n^*)$-bandit. Complete information concerning the discount sequence is present at stage 1. It is easy to see that for all n, an optimal strategy τ^* satisfies:

$$\tau^*(\varnothing) = 1,$$
$$\tau^*(1; 1 - n^{-1}) = \tau(1; 1 + n^{-1}) = 1,$$
$$\tau^*(1, 1; 1 - n^{-1}, 0) = \tau^*(1, 0; 1 - n^{-1}, 0) = 1,$$
$$\tau^*(0; 1 - n^{-1}) = 1, \qquad \tau^*(0; 1 + n^{-1}) = 2,$$
$$\tau^*(0, 1; 1 - n^{-1}, 0) = 1, \qquad \tau^* = (0, 0; 1 - n^{-1}, 0) = 2;$$

and $V(G; \mathbf{H}_n^*) = 95/60$.

So $V(G; \mathbf{H}_n^*) \nrightarrow V(G; \mathbf{H})$. □

When the discount sequence is observable, one typically cannot replace a random discount sequence by a nonrandom sequence without changing the problem in an essential way. The next result gives a special situation where such a replacement is possible.

Theorem 3.4.3 Suppose an observable random discount sequence $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$ is independent of $\{X_{im}: i = 1, \ldots, k;\ m = 1, 2, \ldots\}$ and is given by

$$\alpha_m = \prod_{l=1}^{m} U_l, \tag{3.4.1}$$

where, under the distribution $\mathbf{H}$, $\{U_l: l = 1, 2, \ldots\}$ is an independent sequence. Then, for all G,

$$V(G; \mathbf{H}) = V(G; E(\mathbf{A}|\mathbf{H}))$$

and each optimal strategy for the $(G; E(\mathbf{A}|\mathbf{H}))$-bandit is optimal for the $(G; \mathbf{H})$-bandit.

Proof The proof is by induction followed by a limiting argument. Let

$$\mathscr{H}_n = \{\mathbf{H}: P(U_{n+1} = 0|\mathbf{H}) = 1, \{U_l: l \geqslant 1\} \text{ is independent under } \mathbf{H}\}.$$

The conclusion of the theorem obviously holds for $\mathbf{H} \in \mathscr{H}_0$. Assume it holds for all $\mathbf{H} \in \mathscr{H}_{n-1}$ and fix $\mathbf{H} \in \mathscr{H}_n$.

The distribution of $\{U_l: l = 1, 2, \ldots\}$ is governed by $\mathbf{H}$ and, hence, $\mathbf{H}$ determines the distribution $\mathbf{H}_1$ of the random discount sequence

$$(U_2(\omega), U_2(\omega)U_3(\omega), \ldots, \prod_{l=2}^{m} U_l(\omega), \ldots).$$

For a set S of discount sequences let S/u, for $u > 0$, denote the set of discount sequences obtained by dividing each member of each sequence in S by u. It is clear that the random distribution $\mathbf{H}^{(1)}$ used in (3.2.2) almost surely belongs to $\mathscr{H}_{n-1}$ and satisfies

$$\mathbf{H}^{(1)}(S, \omega) = \mathbf{H}_1(S/U_1(\omega))$$

if $U_1(\omega) \neq 0$ and, when $U_1(\omega) = 0$,

$$\mathbf{H}^{(1)}(\{0, 0, 0, \ldots)\}, \omega) = 1.$$

In the random setting, as in the nonrandom, multiplication of the discount sequence by a constant multiplies the value by that constant. Hence,

$$V(G^{(1)}(\cdot, \omega); \mathbf{H}^{(1)}(\cdot, \omega)) = U_1(\omega) V(G^{(1)}(\cdot, \omega); \mathbf{H}_1).$$

Since $\mathbf{H}_1$ is not random,

$$E_\tau(U_1(\omega) V(G^{(1)}(\cdot, \omega); \mathbf{H}_1)|G, \mathbf{H}) = E(U_1(\omega)|\mathbf{H}) E_\tau(V(G^{(1)}(\cdot, \omega); \mathbf{H}_1)|G, \mathbf{H}),$$

which, by the induction hypothesis, equals

$$E(U_1|\mathbf{H}) E_\tau(V(G^{(1)}; (E(U_2|\mathbf{H}), E(U_2U_3|\mathbf{H}), \ldots))|G).$$

On moving the constant $E(U_1|\mathbf{H})$ through the expectations and

through V, this expression becomes

$$E_\tau(V(G_i^{(1)}; (E(U_1U_2|\mathbf{H}), E(U_1U_2U_3|\mathbf{H}), \ldots))|G).$$

By comparing (3.2.2) with (2.5.2) we see that the $(G; \mathbf{H})$-bandit has the same value and optimal initial selections as does the $(G; E(\mathbf{A}|\mathbf{H}))$-bandit. This completes the induction step.

If $\mathbf{H}$ is such that $P(U_n > 0|\mathbf{H}) > 0$ for all n, a limiting argument using Lebesgue's dominated convergence theorem and Lemma 3.4.2 easily completes the proof. □

Remark Suppose the hypothesis and, therefore, the conclusion of Theorem 3.4.3 hold. Since the class of strategies is much richer when the discount sequence is random and discount factors are observable, there may be optimal strategies for the $(G; \mathbf{H})$-bandit that are meaningless for the $(G; E(\mathbf{A}|\mathbf{H}))$-bandit; for, when two or more selections are equally good for the $(G; E(\mathbf{A}|\mathbf{H}))$-bandit, the choice from among these selections in the $(G; \mathbf{H})$-bandit can depend on preceding values of the discount sequence – all resulting strategies are, of course, equally good. □

Formula (3.4.1), with a highly dependent sequence $\{U_l: l = 1, 2, \ldots\}$, is appropriate for Example 3.4.1. In that example, $U_l = 1/4$ or $3/4$ each with probability $1/2$ and $U_l = U_1$ for each l. This is an instance of a certain type of dependence of the sequence $\{U_l: l = 1, 2, \ldots\}$: a random distribution R on $\mathbb{R}$ is chosen and then $\{U_l: l = 1, 2, \ldots\}$ is a conditionally independent sequence in which the conditional distribution of each U is R. In Example 3.4.1, R is a one-point distribution concentrated at either $1/4$ or $3/4$, each with probability $1/2$. It would be interesting to pursue the study of observable sequences in this manner for an arbitrary member of $\mathscr{D}(\mathbb{R})$ governing R.

When (3.4.1) holds, an interpretation of U_l is that the payoff at stage l is discounted by the random factor U_l as compared with the previous stage: $(1 - U_l)/U_l$ is the random inflation rate.

The next section deals with the situation in which stages occur in real time and discounting is also in terms of real time.

3.5 Real-time discounting

Consider a clinical trial in which patients arrive at haphazard times. These times are not predictable in advance, and, indeed, the number of

patients to arrive in any fixed time period is unknown. Suppose that the response of a patient who arrives and is treated at time $t \geqslant 0$ is weighted by the factor $\exp(-\beta t)$, $0 < \beta < \infty$. Were the patients' arrival times known in advance to be $t_1, t_2, t_3, \ldots$, then the appropriate discount sequence would be nonrandom:

$$\mathbf{A} = (\exp(-\beta t_1), \exp(-\beta t_2), \exp(-\beta t_3), \ldots).$$

Since arrival times are not known, we let T_m denote the random time at which the mth trial (or stage, or patient) occurs. Let $Y_1 = T_1$ and $Y_m = T_m - T_{m-1}$, $m = 2, 3, \ldots$, denote the interarrival times. The appropriate discount sequence is random:

$$\begin{aligned}(\exp(-\beta T_1), \exp(-\beta T_2), \ldots, \exp(-\beta T_m), \ldots) \\ = (\exp(-\beta Y_1), \exp(-\beta(Y_1 + Y_2)), \ldots, \\ \exp\Big(-\beta \sum_{l=1}^{m} Y_l\Big), \ldots) \\ = (U_1, U_1 U_2, \ldots, \prod_{l=1}^{m} U_l, \ldots)\end{aligned}$$

where $U_l = \exp(-\beta Y_l)$, $l = 1, 2, \ldots$.

The ability to observe discount factors means in the current context that the decision maker observes $Y_1, \ldots, Y_m$ and uses this information when selecting an arm at the mth stage. If $\{Y_l: l = 1, 2, \ldots\}$ is an independent sequence and is independent of $\{X_{im}: i = 1, \ldots, k; m = 1, 2, \ldots\}$, then the same is true for $\{U_l: l = 1, 2, \ldots\}$. Theorem 3.4.3 does not apply directly because the selection at stage m in the current setting may depend on U_m as well as on $U_1, \ldots, U_{m-1}$. Nevertheless, an argument similar to the proof of Theorem 3.4.3 applies to show that the discount factors may, without loss, be replaced by their expected values. This argument uses the fact that the discounting is exponential.

The situation is more complicated for real-time *discount functions* that are not of the form $\exp(-\beta t)$. Suppose, for example, that

$$\alpha_t = \begin{cases} 1, & t \in [0, 1] \\ 0, & t \in (1, \infty); \end{cases}$$

in a clinical trial the objective is to maximize the sum of the responses of all patients arriving in $[0, 1]$. If the first stage occurs at time 0.01 the arm selected may be very different than if it occurs at time 0.99; a

'risky' arm may be appropriate in the former and an arm with large mean may be a clear choice in the latter.

Allowing strategies that depend on real time t does not fit into the framework we have developed. But, when $\{Y_l: l = 1, 2, \ldots\}$, as defined above, is independent of $\{X_{im}: i = 1, \ldots, k;\ m = 1, 2, \ldots\}$ and is a sequence of independent exponentially distributed random variables having mean $\kappa > 0$, a simple artifice gives an arbitrarily good approximation using a bandit as described in Chapter 2.

We let $\varepsilon > 0$ and construct a bandit whose mth stage occurs at real time $m\varepsilon$. If ε is small then we are introducing too many stages by having them occur at each integral multiple of ε. We would like to correct for this by concealing observations so that the lengths of intervals between successive unconcealed observations are independent random variables having mean κ and distributions that approach the exponential as $\varepsilon \downarrow 0$.

Concealing observations is not a possibility reckoned with in Chapter 2, so we need to obtain the effect of concealment in another manner. Instead of concealing an observation at a particular stage, we suppose that a constant c_ε is observed that gives no information about the character of the various arms. For this purpose we consider a number c_ε such that $P(X_{im} = c_\varepsilon | G) = 0$ for $i = 1, \ldots, k$. To allow for the observation c_ε on any arm we need to replace $G \in \mathscr{D}^*(\mathscr{D}^k)$ by an appropriate $G_\varepsilon \in \mathscr{D}^*(\mathscr{D}^k)$. Define $\psi_\varepsilon: \mathscr{D}^k \to \mathscr{D}^k$ by

$$\psi_\varepsilon(Q_1, \ldots, Q_k) = \left(\frac{\varepsilon}{\kappa} Q_1 + \left(1 - \frac{\varepsilon}{\kappa}\right)\delta_{c_\varepsilon}, \ldots, \frac{\varepsilon}{\kappa} Q_k + \left(1 - \frac{\varepsilon}{\kappa}\right)\delta_{c_\varepsilon}\right)$$

and let $G_\varepsilon = G \circ \psi_\varepsilon^{-1}$. For each i, $P(X_{im} = c_\varepsilon | G_\varepsilon) = 1 - \varepsilon/\kappa$ as desired. In addition, the conditional distribution $(c_\varepsilon)_i G_\varepsilon$ equals G_ε. This last fact is needed since we want the 'concealed observations' to contain no information about the arms.

The unwanted observations of c_ε, while not being relevant for strategies, do contribute the quantity

$$\sum_{m=1}^{\infty} c_\varepsilon(1 - \varepsilon/\kappa)\alpha_{\varepsilon m} = \varepsilon^{-1} c_\varepsilon (1 - \varepsilon/\kappa) \sum_{m=1}^{\infty} \alpha_{\varepsilon m}\varepsilon \qquad (3.5.1)$$

to the worth of any strategy. We want (3.5.1) to approach 0 as $\varepsilon \downarrow 0$ so we suppose, as is natural, that $\int_0^\infty \alpha_t\, dt < \infty$ and we choose c_ε so that $\varepsilon^{-1} c_\varepsilon \to 0$.

In (3.5.1), and implicitly elsewhere, we have been using the discount

sequence $\mathbf{A}_\varepsilon = (\alpha_\varepsilon, \alpha_{2\varepsilon}, \ldots)$ when the distribution is G_ε. Let τ_ε denote an optimal strategy for the $(G_\varepsilon; \mathbf{A}_\varepsilon)$-bandit. We would like to let $\varepsilon \downarrow 0$ to obtain an optimal strategy for the original setting, which we call the $(G; \alpha_t, \kappa)$-bandit. To accomplish this we make the following definition for the $(G; \alpha_t, \kappa)$-bandit. A *strategy* is a function that assigns an integer indicating the arm to be selected at the mth stage to each sequence $t_1 < \ldots < t_m$ of observation times and each (partial) history $z_1, \ldots, z_{m-1}$ of observations. Thus $\tau(\varnothing; t_1)$ indicates the arm to be selected at stage 1 if stage 1 occurs at time t_1; $\tau(z_1; t_1, t_2)$ indicates the arm to be selected at stage 2 if stage 2 occurs at time t_2 and z_1 was observed at time t_1 (necessarily on arm $\tau(\varnothing; t_1)$); etc. The definition of the $(G; \alpha_t, \kappa)$-bandit is now complete; we denote its value by $V(G; \alpha_t, \kappa)$.

Since the optimal strategy τ_ε for the $(G_\varepsilon; \mathbf{A}_\varepsilon)$-bandit is not a strategy for the $(G; \alpha_t, \kappa)$-bandit according to the preceding definition, we define a strategy $\tau_{\varepsilon,\kappa}$ for the $(G; \alpha_t, \kappa)$-bandit that is closely related to τ_ε. The observations of c_ε for the $(G_\varepsilon; \mathbf{A}_\varepsilon)$-bandit must not be used in the definition of $\tau_{\varepsilon,\kappa}$ except as timekeepers. Set

$$\tau_{\varepsilon,\kappa}(z_1, \ldots, z_{m-1}; t_1, \ldots, t_m) = \tau_\varepsilon(c_\varepsilon, \ldots, c_\varepsilon, z_1, c_\varepsilon, \ldots, c_\varepsilon, z_2, \ldots, z_{m-1}, c_\varepsilon, \ldots, c_\varepsilon)$$

where $z_1, \ldots, z_{m-1}$ occur at positions $[t_1/\varepsilon], \ldots, [t_{m-1}/\varepsilon]$ and the number of positions is $[t_m/\varepsilon] - 1$; here $[t]$ denotes the greatest integer no larger than t.

The preceding discussion leads to the following theorem, the formal proof of which we will omit.

Theorem 3.5.1 Suppose the discount function α_t is nonnegative, of bounded variation, and that it satisfies $\int_0^\infty \alpha_t \, dt < \infty$. For $\varepsilon > 0$ let c_ε satisfy $P(X_{il} = c_\varepsilon | G) = 0$ for $i = 1, \ldots, k$, and suppose that $\varepsilon^{-1} c_\varepsilon \to 0$ as $\varepsilon \downarrow 0$. Then

$$V(G; \alpha_t, \kappa) = \lim_{\varepsilon \downarrow 0} V(G_\varepsilon; \mathbf{A}_\varepsilon).$$

In addition, τ_κ is an optimal strategy for the $(G; \alpha_t, \kappa)$-bandit, where τ_κ is defined by

$$\tau_\kappa(z_1, \ldots, z_{m-1}; t_1, \ldots, t_m) = \limsup \tau_{\kappa,\varepsilon}(z_1, \ldots, z_{m-1}; t_1, \ldots, t_m) \quad (\varepsilon \downarrow 0 \text{ through a fixed sequence}). \tag{3.5.2}$$

Remarks If $P(X_{im} = 0|G) = 0$ for each i, then c_ε can be chosen to be 0 for each ε. The lim sup in (3.5.2) can be replaced by lim inf or any other scheme that chooses an i for which $i = \tau_{\varepsilon,\kappa}(z_1, \ldots, z_{m-1}; t_1, \ldots, t_m)$ for infinitely many ε in the sequence. □

3.6 Nonmonotone discount sequences

In many potential applications of bandit problems, future observations are worth less than the current one. The corresponding discount sequence is decreasing.

In other applications certain future observations are worth more than the current one. Future investments may necessarily be larger than they are now. Or in a clinical trial, the frequency of patient arrivals may be increasing, necessitating future treatment in larger batches.

Most of the bandit literature–and the sequential decision theory literature generally–concerns monotone discount sequences. The principal exception is the sequence given by $\alpha_n = 1, \alpha_m = 0, m \neq n$ (cf. Example 2.4.4): the only observation with payoff is the one at stage n and the first $n-1$ observations are made sequentially with the sole objective of obtaining information to aid in the nth selection. While information and payoff separate nicely in this problem, it is still quite difficult. When $\mathbf{A} = (0, \ldots, 0, 1, 0, 0, \ldots)$, the problem is similar to a conventional problem in decision theory: there is a sampling period followed by a terminal decision. The terminal decision is simply to choose the best arm. The sampling is sequential with a restriction on the total number of observations. Some appropriate references are Lindley and Barnett (1965), Ray (1965), Pratt (1966), and Clayton (1983).

References

Berry, D. A. (1983) Bandit problems with random discounting. *Mathematical Learning Models–Theory and Algorithms* (eds U. Herkenrath, D. Kalin and W. Vogel), pp. 12–25, Springer-Verlag, New York.

Clayton, M. K. (1983) Bayes sequential sampling for choosing the better of two populations. Ph.D. thesis, Univ. of Minnesota, USA.

Lindley, D. V. and Barnett, B. N. (1965) Sequential sampling: two decision problems with linear losses for binomial and normal random variables. *Biometrika* **52**: 507–532.

Pratt, J. W. (1966) The outer needle of some Bayes sequential continuation regions. *Biometrika* **53**: 455–467.

Ray, S. N. (2965) Bounds on the maximal sample size of a Bayes sequential procedure. *Ann. Math. Statist.* **36**: 859–878.

CHAPTER 4

Independent Bernoulli arms

Many of the examples in the first three chapters assume independent arms: $G = F_1 \times \ldots \times F_k$. In this chapter we consider independent Bernoulli arms. As has been our convention in the Bernoulli case, we regard F_i as a distribution on the Bernoulli parameter $\theta_i \in [0, 1]$ rather than on $Q_i \in \mathscr{Q}$; and consistent with an earlier modification of notation, we write the conditional distribution of $(\theta_1, \theta_2, \ldots, \theta_k)$ given success on arm 1, say, as

$$\sigma_1(F_1, F_2, \ldots, F_k) = (\sigma F_1, F_2, \ldots, F_k),$$

and given a failure on arm 1 and as

$$\varphi_1(F_1, F_2, \ldots, F_k) = (\varphi F_1, F_2, \ldots, F_k).$$

Some of the definitions and results in this chapter also apply when the F_i are arbitrary, with, perhaps, some modification. These will be given without announcing such generality; various extensions will be developed as needed in later chapters. Also, some of the results can be extended to certain kinds of dependence among the arms, but most do not apply more generally. One result (Corollary 4.3.10) is given for a particular class of distributions for which the arms are dependent.

Some results in this chapter have intrinsic importance; notable examples are Theorems 4.1.6, 4.3.6, 4.3.8, and 4.3.9. But the primary purpose of the chapter is to set the stage and develop technical foundations for the following three chapters, which deal exclusively with settings in which the arms are independent. In Chapter 5, **A** is arbitrary, $k = 2$, and F_2 associates all its mass with a constant. In Chapter 6, **A** is geometric and k is arbitrary. In Chapter 7, **A** is finite horizon uniform, $k = 2$, and the arms are Bernoulli.

The most efficient way to read this chapter will depend on the reader. Though we give examples during the discourse, the primary motivation is frequently provided by results in Chapters 5, 6, and 7. So

the reader may choose to skim this chapter first, paying special attention to the theorems mentioned above. Such a reader can refer back later for more serious but selective perusal.

4.1 Monotonicity of the value function

The purpose of this section is to describe the behaviour of V in the independent case. We shall repeat some of the notation and terminology of Chapter 2 as it applies to the current setting. The worth of strategy τ depends on the initial state of information $(F_1, \ldots, F_k; \mathbf{A})$ and is written $W(F_1, \ldots, F_k; \mathbf{A}; \tau)$. Recall that

$$V(F_1, \ldots, F_k; \mathbf{A}) = \sup_{\tau} W(F_1, \ldots, F_k; \mathbf{A}; \tau).$$

We saw in Example 2.5.1 that V is not always continuous in the distribution G. However, it is continuous on the restricted domain of the present setting, as we prove in the following theorem.

Theorem 4.1.1 In the independent Bernoulli case, the value V is a continuous function of $(F_1, \ldots, F_k; \mathbf{A})$.

Proof We must compare two bandits, say $(F_1, \ldots, F_k; \mathbf{A})$ and $(F_1^*, \ldots, F_k^*; \mathbf{A}^*)$, and then take the suprema of their worths over τ. We omit the details, noting only the method of proof. For any fixed $\varepsilon > 0$, n, τ and history of n observations, the probabilities that the two bandits yield this history for the first n trials (when strategy τ is used) differ by less than ε if $(F_1^*, \ldots, F_k^*)$ is sufficiently close to $(F_1, \ldots, F_k)$. (This also gives uniform continuity assuming one of the usual metrics for convergence in distribution.) □

The nonnegativity of each X_{im} immediately gives the next result.

Theorem 4.1.2 The value V is an increasing function of $\mathbf{A}$.

The following example illustrates that monotonicity of V in $(F_1, \ldots, F_k)$ can be tricky.

Example 4.1.1 Let $k = 2$ and $\mathbf{A} = (1, \alpha, \alpha^2, \ldots)$, where $\alpha = 0.95$. Suppose θ_2 is known to be 0.7, that is, $F_2 = \delta_{0.7}$. Consider two different distributions for arm 1: F_1^* assigns probability 1/2 to each of

1/4 and 3/4, that is, $F_1^* = \frac{1}{2}\delta_{1/4} + \frac{1}{2}\delta_{3/4}$; whereas $F_1 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{3/4}$. From Example 5.4.1 it will follow that arm 2 is always optimal for the $(F_1^*, F_2; \mathbf{A})$-bandit and hence, that

$$V(F_1^*, F_2; \mathbf{A}) = 0.7 \sum_{m=1}^{\infty} \alpha^{m-1} = 0.7/(1-0.95) = 14.$$

Define strategy τ as follows: $\tau(\varnothing) = 1$ and, for $m > 2$, $\tau(z_1, \ldots, z_{m-1}) = 1$ if $z_1 = 1$ and $\tau(z_1, \ldots, z_{m-1}) = 2$ if $z_1 = 0$. We have

$$W(F_1, F_2; \mathbf{A}; \tau) = \frac{449}{32} > 14$$

and, hence, $V(F_1, F_2; \mathbf{A}) > 14$. (It happens that τ is optimal but this is incidental to our purpose.)

This is a rather surprising conclusion for it indicates that the $(F_1, F_2; \mathbf{A})$-bandit is preferable to the $(F_1^*, F_2; \mathbf{A})$-bandit even though θ_1 is stochastically larger under F_1^* than under F_1. □

The first of the following two definitions is equivalent to stochastic ordering of random variables and the second is motivated by the preceding example. The importance of the second definition will be made clear in Chapter 5. The second definition will be extended in Section 4.3; this extension will play an important role in Chapter 7.

Definition 4.1.1 A one-dimensional distribution function F^* is *to the right of* a one-dimensional distribution function F if $F^*(x) \leqslant F(x)$ for every x; that is, in terms of measures, $F^*[x, \infty) \geqslant F[x, \infty)$ for every x. If, in addition, $F^* \neq F$, this relationship is *strict*.

Recall from Chapter 2 that $\sigma^s\varphi^f F$ is the current distribution of an arm that has yielded s successes and f failures and whose prior distribution is F.

Definition 4.1.2 A one-dimensional distribution F^* on $[0, 1]$ is *strongly to the right of* a one-dimensional distribution F on $[0, 1]$ if $\sigma^s\varphi^f F^*$ is to the right of $\sigma^s\varphi^f F$ for every pair (s, f) of nonnegative integers for which both $\sigma^s\varphi^f F^*$ and $\sigma^s\varphi^f F$ are defined. If, in addition, $F^* \neq F$, this relationship is *strict*.

In Example 4.1.1 the distribution F_1^* is to the right of F_1, but it is not strongly to the right:

$$\sigma F_1(\{3/4\}) = 1 > \sigma F_1^*[3/4, 1] = 3/4.$$

The following three easy lemmas are given without proof. The first says that the probability of success is at least as large under F^* as under F when F^* is to the right of F. The second says that 'strongly to the right' is preserved under identical experimental results. The third asserts that successes on arm 1 move the distribution of θ_1 strongly to the right and that the opposite happens with failures.

As a natural adjustment of notation we write $E(\theta_1|F_1)$, for example, in lieu of $E(\theta_1|G)$ when $G = (F_1, \ldots, F_k)$, for this expectation depends on G only through F_1.

Lemma 4.1.3 If F_1^* is to the right of F_1 then $E(\theta_1|F_1^*) \geqslant E(\theta_1|F_1)$ with equality if and only if $F_1^* = F_1$.

For any one-dimensional distribution F, if $F = \delta_0$ then σF is not defined, and the same is true for φF when $F = \delta_1$. The next two lemmas do not cover these possibilities for F or F^*.

Lemma 4.1.4 If F^* is strongly to the right of F then σF^* and φF^* are, respectively, strongly to the right of σF and φF (when all these distributions exist).

Lemma 4.1.5 For any F, σF is strongly to the right of F, which is strongly to the right of φF (when these distributions exist). The relationships are strict if and only if F is not concentrated at one point.

While Example 4.1.1 rules out monotonicity of V in F_i with respect to 'to the right', the following theorem gives monotonicity of V with respect to 'strongly to the right'. In view of Lemma 4.1.5 this implies that the value is not decreased if the number of successes on an arm were to be increased or the number of failures decreased. The result has inherent significance but is not of great importance in the sequel; a number of papers discussed in the Annotated Bibliography prove special cases.

Theorem 4.1.6 Suppose distribution F_i^* is strongly to the right of F_i for $i = 1, \ldots, k$. Then for any $\mathbf{A}$,

$$V(F_1^*, \ldots, F_k^*; \mathbf{A}) \geqslant V(F_1, \ldots, F_k; \mathbf{A}).$$

Remarks The following proof of this result exploits Lemma 4.1.4, which says that 'strongly to the right' is preserved in the two bandits from one stage to the next. It uses truncation and induction on the horizon of the truncated sequence. □

Proof of Theorem 4.1.6 For an inductive proof, assume first that the horizon n of $\mathbf{A}$ is 0. Then $\mathbf{A} = (0, 0, \ldots)$ and hence, $V(F_1^*, \ldots, F_k^*; \mathbf{A}) = 0 = V(F_1, \ldots, F_k; \mathbf{A})$.

Take $n \geqslant 1$ and assume that the conclusion of the theorem holds for all discount sequences having horizon less than n. Suppose the horizon of $\mathbf{A}$ is n and let τ denote an optimal strategy for the $(F_1, \ldots, F_k; \mathbf{A})$-bandit. Define τ^* to be a strategy for the $(F_1^*, \ldots, F_k^*; \mathbf{A})$-bandit that specifies the same first selection as τ and is optimal for the new bandit presenting itself at stage 2. We need only show

$$W(F_1^*, \ldots, F_k^*; \mathbf{A}; \tau^*) \geqslant V(F_1, \ldots, F_k; \mathbf{A}).$$

For notational ease assume $\tau(\varnothing) = 1$, and therefore $\tau^*(\varnothing) = 1$. Then from (2.5.2) we have

$$\begin{aligned} W(F_1^*, \ldots, F_k^*; \mathbf{A}; \tau^*) = {} & \alpha_1 E(\theta_1 | F_1^*) \\ & + E(\theta_1 | F_1^*)\, V(\sigma F_1^*, F_2^*, \ldots, F_k^*; \mathbf{A}^{(1)}) \\ & + E(1 - \theta_1 | F_1^*)\, V(\varphi F_1^*, F_2^*, \ldots, F_k^*; \mathbf{A}^{(1)}). \end{aligned}$$

Since τ is optimal in the $(F_1, \ldots, F_k; \mathbf{A})$-bandit,

$$\begin{aligned} V(F_1, \ldots, F_k; \mathbf{A}) = {} & \alpha_1 E(\theta_1 | F_1) \\ & + E(\theta_1 | F_1)\, V(\sigma F_1, F_2, \ldots, F_k; \mathbf{A}^{(1)}) \\ & + E(1 - \theta_1 | F_1)\, V(\varphi F_1, F_2, \ldots, F_k; \mathbf{A}^{(1)}). \end{aligned}$$

We want to show that the following is nonnegative:

$$\begin{aligned} & W(F_1^*, \ldots, F_k^*; \mathbf{A}; \tau^*) - V(F_1, \ldots, F_k; \mathbf{A}) \\ & \quad = \alpha_1 [E(\theta_1 | F_1^*) - E(\theta_1 | F_1)] \\ & \qquad + E(\theta_1 | F_1) [V(\sigma F_1^*, F_2^*, \ldots, F_k^*; \mathbf{A}^{(1)}) \\ & \qquad\qquad - V(\sigma F_1, F_2, \ldots, F_k; \mathbf{A}^{(1)})] \end{aligned}$$

$$+E(1-\theta_1|F_1^*)[V(\varphi F_1^*, F_2^*, \ldots, F_k^*; \mathbf{A}^{(1)}) - V(\varphi F_1, F_2, \ldots, F_k; \mathbf{A}^{(1)})]$$
$$+[E(\theta_1|F_1^*)-E(\theta_1|F_1)][V(\sigma F_1^*, F_2^*, \ldots, F_k^*; \mathbf{A}^{(1)}) - V(\varphi F_1, F_2, \ldots, F_k; \mathbf{A}^{(1)})]. \quad (4.1.1)$$

The first term on the right-hand side of (4.1.1) is nonnegative by Lemma 4.1.3. The next two terms are nonnegative by Lemma 4.1.4 and the induction hypothesis. That the last term is nonnegative follows from Lemmas 4.1.3, 4.1.4, and 4.1.5, and the induction hypothesis.

The result for a discount sequence with infinite horizon now follows from the continuity of V in $\mathbf{A}$ (Theorem 4.1.1). □

In the next section we assume $k = 2$ and introduce a notation for the difference in worths between the two initial selections. We give a recursion for this difference that is useful in later demonstrations concerning the nature of optimal strategies.

4.2 The advantage of one arm over another

In this section we specialize to $k = 2$ and define a function of $(F_1, F_2; \mathbf{A})$ which represents the difference between selecting arm 1 followed by an optimal continuation and arm 2 followed by an optimal continuation. Analogous functions can be defined when k is arbitrary – see Quisel (1965) – but we have had little success with these generalizations.

There are two possible initial selections: arm 1 and arm 2. As in Theorem 2.5.1, let $V^{(i)}$ denote the worth of selecting arm i initially and then continuing with an optimal strategy:

$$V^{(i)}(F_1, F_2; \mathbf{A}) = \sup_{\tau(\varnothing)=i} W(F_1, F_2; \mathbf{A}; \tau).$$

From (2.5.2), the value $V(F_1, F_2; \mathbf{A})$ is the maximum of these two numbers. We have

$$V^{(1)}(F_1, F_2; \mathbf{A}) = \alpha_1 E(\theta_1|F_1) + E(\theta_1|F_1)\, V(\sigma F_1, F_2; \mathbf{A}^{(1)}) + E(1-\theta_1|F_1)\, V(\varphi F_1, F_2; \mathbf{A}^{(1)}), \quad (4.2.1)$$

$$V^{(2)}(F_1, F_2; \mathbf{A}) = \alpha_1 E(\theta_2|F_2) + E(\theta_2|F_2)\, V(F_1, \sigma F_2; \mathbf{A}^{(1)}) + E(1-\theta_2|F_2)\, V(F_1, \varphi F_2; \mathbf{A}^{(1)}). \quad (4.2.2)$$

Let $\Delta(F_1, F_2; \mathbf{A})$ denote the advantage of arm 1 over arm 2

assuming optimal continuations in both cases:

$$\Delta(F_1, F_2; \mathbf{A}) = V^{(1)}(F_1, F_2; \mathbf{A}) - V^{(2)}(F_1, F_2; \mathbf{A}).$$

The sign of Δ indicates the optimal initial selection: arm 1 if $\Delta \geqslant 0$ and arm 2 if $\Delta \leqslant 0$. In fact, the sign of Δ was used in Example 2.4.3 to display an optimal strategy. But as the following development makes clear, the sign of $\Delta(F_1, F_2; \mathbf{A})$ depends on the magnitude of $\Delta(\sigma F_1, F_2; \mathbf{A}^{(1)})$, for example, and not just its sign. Therefore, we need to consider the magnitude of the Δ function when we want to find optimal strategies.

For all $\mathbf{A}$ and any (F_1, F_2),

$$V^{(1)}(F_1, F_2; \mathbf{A}) = V(F_1, F_2; \mathbf{A}) - \Delta^-(F_1, F_2; \mathbf{A}), \tag{4.2.3}$$

$$V^{(2)}(F_1, F_2; \mathbf{A}) = V(F_1, F_2; \mathbf{A}) - \Delta^+(F_1, F_2; \mathbf{A}), \tag{4.2.4}$$

where $\Delta^+ = 0 \vee \Delta$ and $\Delta^- = 0 \vee (-\Delta)$. In view of (4.2.4) and (4.2.3), respectively, (4.2.1) and (4.2.2) become

$$\begin{aligned} V^{(1)}(F_1, F_2; \mathbf{A}) &= \alpha_1 E(\theta_1 | F_1) \\ &\quad + E(\theta_1 | F_1)[V^{(2)}(\sigma F_1, F_2; \mathbf{A}^{(1)}) + \Delta^+(\sigma F_1, F_2; \mathbf{A}^{(1)})] \\ &\quad + E(1-\theta_1 | F_1)[V^{(2)}(\varphi F_1, F_2; \mathbf{A}^{(1)}) + \Delta^+(\varphi F_1, F_2; \mathbf{A}^{(1)})]. \end{aligned} \tag{4.2.5}$$

$$\begin{aligned} V^{(2)}(F_1, F_2; \mathbf{A}) &= \alpha_1 E(\theta_2 | F_2) \\ &\quad + E(\theta_2 | F_2)[V^{(1)}(F_1, \sigma F_2; \mathbf{A}^{(1)}) + \Delta^-(F_1, \sigma F_2; \mathbf{A}^{(1)})] \\ &\quad + E(1-\theta_2 | F_2)[V^{(1)}(F_1, \varphi F_2; \mathbf{A}^{(1)}) + \Delta^-(F_1, \varphi F_2; \mathbf{A}^{(1)})]. \end{aligned} \tag{4.2.6}$$

The sum

$$\begin{aligned} &\alpha_1 E(\theta_1 | F_1) + E(\theta_1 | F_1) V^{(2)}(\sigma F_1, F_2; \mathbf{A}^{(1)}) \\ &\quad + E(1-\theta_1 | F_1) V^{(2)}(\varphi F_1, F_2; \mathbf{A}^{(1)}) \end{aligned} \tag{4.2.7}$$

in (4.2.5) amounts to the worth of selecting arm 1 first and arm 2 second and then continuing optimally. Likewise,

$$\begin{aligned} &\alpha_1 E(\theta_2 | F_2) + E(\theta_2 | F_2) V^{(1)}(F_1, \sigma F_2; \mathbf{A}^{(1)}) \\ &\quad + E(1-\theta_2 | F_2) V^{(1)}(F_1, \varphi F_2; \mathbf{A}^{(1)}) \end{aligned} \tag{4.2.8}$$

in (4.2.6) is the expected worth of selecting arm 2 first and arm 1 second and then continuing optimally. Since the selections are exchanged in these two interpretations, the difference between (4.2.7) and (4.2.8) is $(\alpha_1 - \alpha_2)[E(\theta_1 | F_1) - E(\theta_2 | F_2)]$. Therefore, subtracting (4.2.6) from (4.2.5) gives the following lemma.

Lemma 4.2.1 When there are two independent Bernoulli arms,

$$\begin{aligned}\Delta(F_1, F_2; \mathbf{A}) = (\alpha_1 - \alpha_2)[E(\theta_1 \mid F_1) - E(\theta_2 \mid F_2)] \\ + E(\theta_1 \mid F_1)\Delta^+(\sigma F_1, F_2; \mathbf{A}^{(1)}) \\ + E(1-\theta_1 \mid F_1)\Delta^+(\varphi F_1, F_2; \mathbf{A}^{(1)}) \\ - E(\theta_2 \mid F_2)\Delta^-(F_1, \sigma F_2; \mathbf{A}^{(1)}) \\ - E(1-\theta_2 \mid F_2)\Delta^-(F_1, \varphi F_2; \mathbf{A}^{(1)}). \qquad (4.2.9)\end{aligned}$$

Remark In the repeated application of (4.2.9) certain Δ's will not be defined when F_1 or F_2 gives zero probability to the open interval (0,1). For example, $\varphi\sigma F_1$ is not defined when $F_1(\{1, 0\}) = 1$. But the multiplier of such a term is always 0, and our convention is that the product is 0. □

A rather easy consequence of later development is that not all terms in (4.2.9) can be zero, except when F_1 and F_2 are the same one-point distribution. In addition, it seems reasonable to expect that the vanishing of particular terms in (4.2.9) implies the vanishing of other terms. In fact, when $\mathbf{A}$ is nonincreasing, it will follow from Theorem 4.3.6 that $\Delta(\sigma F_1, F_2; \mathbf{A}) > 0$ whenever $\Delta(\varphi F_1, F_2; \mathbf{A}) > 0$ and, symmetrically, $\Delta(F_1, \sigma F_2; \mathbf{A}) < 0$ whenever $\Delta(F_1, \varphi F_2; \mathbf{A}) < 0$. So an arm that is optimal after a failure is also optimal after a success.

When the horizon n is finite, (4.2.9) defines Δ recursively. The uniform case (in which the objective is to maximize the expected successes in the first n observations) is treated in Chapter 7. In this case the right-hand side of (4.2.9) reduces to the first term when $n = 1$; when $n \geqslant 2$ the first term is zero, and the other terms apply. A detailed discussion of the application of (4.2.9) when $n = 2$ is given in Section 7.2.

When the horizon is 1, (4.2.9) makes it clear that Δ depends on F_1 and F_2 only through their first moments. Applying induction to (4.2.9) on the horizon n of $\mathbf{A}$ shows that $\Delta(F_1, F_2; \mathbf{A})$ depends on F_1 and F_2 only through their first n moments.

Proposition 4.2.2 If the horizon of $\mathbf{A}$ is n and $E(\theta_1^r \mid F_1^*) = E(\theta_1^r \mid F_1)$ for $r = 1, 2, \ldots, n$, then $\Delta(F_1^*, F_2; \mathbf{A}) = \Delta(F_1, F_2; \mathbf{A})$.

4.3 Staying with a winner

Since the arms are Bernoulli, the only possible outcomes when an arm is selected are success and failure. When the arm selected has known

probability of success, we are indifferent to these two outcomes, except for the difference in immediate income. But, in view of Theorem 4.1.6, the information contained in a success is distinctly preferred to that of failure when the arm is unknown since σF is strongly to the right of φF (Lemma 4.1.5). However, this preference may not translate into a desire to stay with the successful arm. Example 2.4.1 gives a setting with dependent arms in which it does not translate. As we will see (Examples 5.2.1 and 5.2.2), even when the arms are independent, it may not be optimal to stay with a winner.

A detailed analysis of this issue is carried out in Chapter 5 when only one arm is unknown. The current section gives a partial characterization of discount sequences for which staying with a winner is a property of optimal strategies when there are two independent Bernoulli arms. This result, and most of the other results in this section, assumes that $\mathbf{A}$ is nonincreasing, $\alpha_m \geqslant \alpha_{m+1}$ for $m \geqslant 1$. Examples 4.3.3 and 4.4.4 clarify the role of this assumption.

The 'stay-with-a-winner' rule, Theorem 4.3.8, generalizes Theorem 6.2 of Berry (1972) which is the corresponding result for finite horizon uniform discounting. The current proof is more direct and simpler. It uses the fact that $\Delta(F_1^*, F_2; \mathbf{A}) \geqslant \Delta(F_1, F_2; \mathbf{A})$ if F_1^* is strongly to the right of F_1, which is contained in Theorem 4.3.3. Because a stronger result is needed in Chapter 7, the forthcoming Theorem 4.3.3 uses a weaker notion of order among distribution measures; this notion is defined next. Since this definition applies to any arm, we drop the subscript from the Bernoulli parameter θ as well as from its distribution F.

Definition 4.3.1 Let $m \geqslant 0$. For one-dimensional distribution measures, $F^* \overset{m}{\gtrsim} F$ (read F^* *m-greater than* F) if

$$E(\theta|\sigma^s\varphi^f F^*) \geqslant E(\theta|\sigma^s\varphi^f F) \tag{4.3.1}$$

whenever $s+f \leqslant m$ and $\sigma^s\varphi^f F^*$ and $\sigma^s\varphi^f F$ are both defined. If, in addition, $E(\theta|F^*) > E(\theta|F)$, then $F^* \overset{m}{\succ} F$ (read F^* *strictly m-greater than* F).

The following two examples illustrate this notion. The first is trivial in that there is obviously no stochastic ordering even though the first several moments under F^* are larger than under F. The second has the flavour of Example 4.3.1 since one distribution measure is to the

right of the other but it is not strongly to the right. Repeated failures eventually reverse the inequality between the means in the first example, while successes do so in the second.

Example 4.3.1 Suppose F^* has density

$$\mathrm{d}F^*(u) \propto u^9(1-u)^4\,\mathrm{d}u,$$

on (0, 1) and $F = \delta_{1/2}$. Then

$$E(\theta|\sigma^s\varphi^f F^*) = \frac{s+10}{s+f+15}, \qquad E(\theta|\sigma^s\varphi^f F) = \frac{1}{2}.$$

So $F^* \overset{m}{\gtrsim} F$ for $m \leqslant 5$ and F^* and F are not comparable for $m > 5$. □

Example 4.3.2 Consider the family of two-point distributions F_x for $x \in (0, 1)$ where $F_x = \frac{1}{2}\delta_x + \frac{1}{2}\delta_1$. Then

$$E(\theta|\sigma^s\varphi^f F_x) = \begin{cases} x & \text{if } f > 0 \\ \dfrac{1+x^{s+1}}{1+x^s} & \text{if } f = 0. \end{cases}$$

So $F_y \overset{m}{\gtrsim} F_x$ provided y is sufficiently larger than x, but for fixed x and y this relation does not hold for arbitrarily large m. For example, $F_{0.9} \overset{m}{\gtrsim} F_x$ only for $x \in [x^*, 0.9]$ where x^* is given to two decimals in the following table:

m	0	1	2	3	4	10	∞
x^*	0	0.05	0.25	0.43	0.56	0.87	0.90

The reason small values of x are excluded here is the same reason the order of the V's in Example 4.1.1 is opposite from that of naive expectation: when a success is observed, θ is more likely to be 1 under F_x for smaller x (or more likely to be 3/4 in Example 4.1.1). □

The following two propositions clarify the strict version of Definition 4.3.1 and relate 'm-greater than' with 'strongly to the right' (Definition 4.1.2). The proof of the first is easy and is omitted.

Proposition 4.3.1 If $F^* \overset{m}{\gtrsim} F$ and (4.3.1) is strict for some (s,f) with $s+f \leqslant m$, then $F^* \overset{m}{>} F$.

Proposition 4.3.2 A distribution F^* is strongly to the right of F if and only if $F^* \overset{m}{\gtrsim} F$ for every m. The first relationship is strict if and only if the second is strict for every m.

In view of Proposition 4.3.2 we use '∞-greater than' synonymously with 'strongly to the right'. The proof of Proposition 4.3.2 is omitted. The 'only if' part is easy and relevant, for example, in connection with the subsequent Corollary 4.3.7 from which it follows that arm 1 is optimal if F_1 is strongly to the right of F_2. The 'if' part is more difficult to prove and will not be used in the sequel.

When the horizon n is finite it would not be surprising, in view of Propositions 4.2.2 and 4.3.2, that a general theorem with a hypothesis involving 'strongly to the right' could be strengthened by using '$(n-1)$-greater than' instead. For example, the next theorem strengthens Theorem 4.1.6 in this way. While it is true for $k \geqslant 2$, it is stated for $k = 2$ because that is the setting of this section. The proof mimics that of Theorem 4.1.6 in the obvious way and so is omitted.

Theorem 4.3.3 Suppose that **A** has horizon n and distribution F_i^* is $(n-1)$-greater than F_i for $i = 1, 2$. Then

$$V(F_1^*, F_2^*; \mathbf{A}) \geqslant V(F_1, F_2; \mathbf{A}).$$

We will require the following two lemmas that are analogous to Lemmas 4.1.4 and 4.1.5. We omit the proof of Lemma 4.3.4. Lemma 4.3.5 is a logical consequence of Lemma 4.1.5.

Lemma 4.3.4 For $m \geqslant 1$, if $F^* \overset{m}{\gtrsim} F$ then $\sigma F^* \overset{m-1}{\gtrsim} \sigma F$ and $\varphi F^* \overset{m-1}{\gtrsim} \varphi F$ (whenever these distributions exist).

Lemma 4.3.5 For any F and $m \geqslant 0$, $\sigma F \overset{m}{\gtrsim} F \overset{m}{\gtrsim} \varphi F$ (whenever these distributions exist). These relationships are strict unless F is concentrated at one point.

It is convenient to have a terminology for functions of distribution measures that are monotonic with respect to 'm-greater than':

Definition 4.3.2 Let $m \in \{1, 2, \ldots, \infty\}$. A function h on the set of distribution measures on $[0, 1]$ is *m-increasing* if $h(F^*) \geqslant h(F)$ when $F^* \overset{m}{\gtrsim} F$. It is *strictly* m-increasing if, in addition, $h(F^*) > h(F)$ whenever $F^* \overset{m}{\gtrsim} F$. Similarly, h is *m-decreasing* or *strictly* m-decreasing according as $-h$ is m-increasing or strictly m-increasing.

Using this terminology, in view of Theorem 4.1.6 the conclusion of Theorem 4.3.3 is that V is $(n-1)$-increasing in either F_1 or F_2.

The most important step in proving the stay-with-a-winner rule is the next theorem.

Theorem 4.3.6 Assume $\mathbf{A}$ is nonincreasing and has horizon $n \in \{1, 2, \ldots, \infty\}$. For fixed F_2, $\Delta(F_1, F_2; \mathbf{A})$ is a strictly $(n-1)$-increasing function of F_1.

Remark By symmetry, $\Delta(F_1, F_2; \mathbf{A})$ is strictly $(n-1)$-decreasing in F_2. □

Proof The parts of the proof are labelled from (i) to (iv). In (i) we use Lemma 4.2.1 to obtain an expression for $\Delta(F_1^*, F_2; \mathbf{A}) - \Delta(F_1, F_2; \mathbf{A})$ that we will apply when $F_1^* \overset{n-1}{\gtrsim} F_1$. In (ii) we use induction on the horizon to show that

$$\Delta(F_1^*, F_2; \mathbf{A}) \geqslant \Delta(F_1, F_2; \mathbf{A}) \tag{4.3.2}$$

when $F_1^* \overset{n-1}{\gtrsim} F_1$ and $n < \infty$. In (iii) we let the horizon approach ∞ to obtain (4.3.2) for all $\mathbf{A}$. Finally, in (iv) we use induction on the smallest m for which $\alpha_{m+1} < \alpha_1$ to prove $\Delta(F_1^*, F_2; \mathbf{A}) > \Delta(F_1, F_2; \mathbf{A})$ whenever $F_1^* \overset{n-1}{\succ} F_1$, that is, when $F_1^* \overset{n-1}{\gtrsim} F_1$ and $E(\theta_1 | F_1^*) > E(\theta_1 | F_1)$. Throughout we will assume that neither F_1^* nor F_2 is supported by $\{0\}$ and that neither F_1 nor F_2 is supported by $\{1\}$. These cases are easily treated separately without induction.

(i) Lemma 4.2.1 gives

$$\begin{aligned}
\Delta(F_1^*, F_2; \mathbf{A}) - \Delta(F_1, F_2; \mathbf{A}) &= (\alpha_1 - \alpha_2)[E(\theta_1 | F_1^*) - E(\theta_1 | F_1)] \\
&\quad + E(\theta_1 | F_1^*)[\Delta^+(\sigma F_1^*, F_2; \mathbf{A}^{(1)}) - \Delta^+(\sigma F_1, F_2; \mathbf{A}^{(1)})] \\
&\quad + E(1 - \theta_1 | F_1^*)[\Delta^+(\varphi F_1^*, F_2; \mathbf{A}^{(1)}) - \Delta^+(\varphi F_1, F_2; \mathbf{A}^{(1)})] \\
&\quad + [E(\theta_1 | F_1^*) - E(\theta_1 | F_1)][\Delta^+(\sigma F_1, F_2; \mathbf{A}^{(1)}) \\
&\qquad\qquad - \Delta^+(\varphi F_1, F_2; \mathbf{A}^{(1)})] \\
&\quad + E(\theta_2 | F_2)[\Delta^-(F_1, \sigma F_2; \mathbf{A}^{(1)}) - \Delta^-(F_1^*, \sigma F_2; \mathbf{A}^{(1)})] \\
&\quad + E(1 - \theta_2 | F_2)[\Delta^-(F_1, \varphi F_2; \mathbf{A}^{(1)}) - \Delta^-(F_1^*, \varphi F_2; \mathbf{A}^{(1)})].
\end{aligned} \tag{4.3.3}$$

As indicated following Lemma 4.2.1, some of the quantities in (4.3.3) are not defined in case F_1^* is supported by $\{0\}$ or F_1 is supported by $\{1\}$. We interpret an undefined term multiplied by 0 to equal 0. We take $\Delta^+(\sigma F_1, F_2; \mathbf{A}^{(1)})$ to equal 0 when it is undefined and observe that this interpretation makes the second term on the right-hand side of (4.3.3) positive and the fourth term zero. So the remainder of the proof can proceed without further attention to this case.

(ii) In view of (4.2.9), $\Delta(F_1^*, F_2; \mathbf{A}) \geqslant \Delta(F_1, F_2; \mathbf{A})$ when $\mathbf{A}$ has horizon 1 and $F_1^* \overset{0}{\gtrsim} F_1$. Suppose $\mathbf{A}$ has finite horizon $n > 1$ and that $\Delta(F_1^*, F_2; \mathbf{A}^{(1)}) \geqslant \Delta(F_1, F_2; \mathbf{A}^{(1)})$ whenever $F_1^* \overset{n-2}{\gtrsim} F_1$. We now show that (4.3.3) is nonnegative when $F_1^* \overset{n-1}{\gtrsim} F_1$.

The first term on the right-hand side of (4.3.3) is nonnegative since (4.3.1) applies with $s = f = 0$ and since $\alpha_1 \geqslant \alpha_2$ by hypothesis. The next two and the last two terms are nonnegative for similar reasons; consider the second term for definiteness. Since $F_1^* \overset{n-1}{\gtrsim} F_1$ it follows immediately from Lemma 4.3.4 that $\sigma F_1^* \overset{n-2}{\gtrsim} \sigma F_1$. Therefore

$$\Delta(\sigma F_1^*, F_2; \mathbf{A}^{(1)}) \geqslant \Delta(\sigma F_1, F_2; \mathbf{A}^{(1)}) \tag{4.3.4}$$

by the inductive hypothesis since the horizon of $\mathbf{A}^{(1)}$ is $n-1$. Since $\Delta^+ = \Delta \vee 0$, it follows that (4.3.4) holds as well with Δ^+ in place of Δ. (The proof that each of the last two terms in (4.3.3) is nonnegative is even easier. Lemma 4.3.4 need not be used; rather Definition 4.3.1 gives $F_1^* \overset{n-2}{\gtrsim} F_1$ as an immediate consequence of $F_1^* \overset{n-1}{\gtrsim} F_1$.)

The first factor in the fourth term is nonnegative as indicated in discussing the first term. The other factor in the fourth term is nonnegative by the inductive hypothesis since $\sigma F_1 \overset{n-2}{\gtrsim} \varphi F_1$ in view of Lemma 4.3.5.

Therefore,

$$\Delta(F_1^*, F_2; \mathbf{A}) - \Delta(F_1, F_2; \mathbf{A}) \geqslant 0.$$

(iii) Suppose $n = \infty$ and $F_1^* \overset{\infty}{\gtrsim} F_1$. Since $F_1^* \overset{n-1}{\gtrsim} F_1$ for each finite n, part (ii) gives

$$\Delta(F_1^*, F_2; \mathbf{A}_n) - \Delta(F_1, F_2; \mathbf{A}_n) \geqslant 0,$$

where $\mathbf{A}_n$ is the truncated version of $\mathbf{A}$ – it agrees with $\mathbf{A}$ through stage n and has only zeros thereafter. Let $n \to \infty$ to obtain

$$\Delta(F_1^*, F_2; \mathbf{A}) - \Delta(F_1, F_2; \mathbf{A}) \geqslant 0.$$

(iv) Let $m_* = \inf\{m: \alpha_{m+1} < \alpha_1\}$. Suppose $m_* = 1$ and that $F_1^* \overset{n-1}{\succ} F_1$, where n is the horizon **A**. By parts (ii) and (iii), every term on the right-hand side of (4.3.3) is nonnegative and it is easily seen that the first term of (4.3.3) is positive. Hence,

$$\Delta(F_1^*, F_2; \mathbf{A}) > \Delta(F_1, F_2; \mathbf{A}).$$

Suppose m_* (necessarily finite) is larger than 1 and assume that $\Delta(F_1^*, F_2; \mathbf{A}^{(1)}) > \Delta(F_1, F_2; \mathbf{A}^{(1)})$ whenever $F_1^* \overset{n-2}{\succ} F_1$. By parts (ii) and (iii) every term on the right-hand side of (4.3.3) is nonnegative; it remains to show that at least one of them is positive.

If F_1 is not supported by one point, Lemma 4.3.5 implies that $\sigma F_1 \overset{n-2}{\succ} \varphi F_1$ and so, by the induction hypothesis, the fourth term on the right-hand side of (4.3.3) is positive. If F_1 is supported by one point, then, by the induction hypothesis, the second term or the penultimate term on the right-hand side of (4.3.3) is positive because, as a consequence of Lemma 4.3.5 and the induction hypothesis, $\Delta(\sigma F_1^*, F_2; \mathbf{A}^{(1)}) > 0$ or $\Delta(F_1^*, \sigma F_2; \mathbf{A}^{(1)}) \leqslant 0$. □

Corollary 4.3.7 Suppose **A** is nonincreasing with horizon $n \in \{1, 2, \ldots, \infty\}$. If $F_1 \overset{n-1}{\succsim} F_2$ then arm 1 is optimal initially for the $(F_1, F_2; \mathbf{A})$-bandit. Moreover, if $F_1 \overset{n-1}{\succ} F_2$ then arm 1 is uniquely optimal.

Proof By symmetry, $\Delta(F_2, F_2; \mathbf{A}) = 0$. The result now follows by applying Theorem 4.3.6. □

The next result is a stay-with-a-winner rule. The bandit is trivial when both F_1 and F_2 are one-point distributions; this possibility is not considered in the theorem. The theorem says that an arm optimal initially continues to be optimal after a success when $\alpha_1 = \alpha_2$ or when the arm has smaller initial probability of success (that is, has smaller mean). So the result applies in the setting of Chapter 7 ($\alpha_1 = \ldots = \alpha_n = 1$, $\alpha_{n+1} = \ldots = 0$). It will be supplemented by results in Chapters 5 and 6.

Theorem 4.3.8 Suppose that **A** is an arbitrary nonincreasing sequence with horizon $n \in \{2, 3, \ldots, \infty\}$ and the support of either F_1 and F_2 consists of more than one point. Then $\Delta(F_1, F_2; \mathbf{A}) \geqslant 0$

implies $\Delta(\sigma F_1, F_2; \mathbf{A}^{(1)}) > 0$ provided

$$\text{(i)}\ \alpha_1 = \alpha_2$$

or

$$\text{(ii)}\ E(\theta_1|F_1) \leqslant E(\theta_2|F_2).$$

Proof From Lemma 4.2.1 and the hypotheses $\Delta(F_1, F_2; \mathbf{A}) \geqslant 0$ and (i) or (ii), we obtain

$$\begin{aligned} 0 \leqslant\ & E(\theta_1|F_1)\Delta^+(\sigma F_1, F_2; \mathbf{A}^{(1)}) + E(1-\theta_1|F_1)\Delta^+(\varphi F_1, F_2; \mathbf{A}^{(1)}) \\ & - E(\theta_2|F_2)\Delta^-(F_1, \sigma F_2; \mathbf{A}^{(1)}) - E(1-\theta_2|F_2)\Delta^-(F_1, \varphi F_2; \mathbf{A}^{(1)}). \end{aligned} \tag{4.3.5}$$

Suppose, for a proof by contradiction, that $\Delta^+(\sigma F_1, F_2; \mathbf{A}^{(1)}) = 0$. Then $\Delta^+(\varphi F_1, F_2; \mathbf{A}^{(1)}) = 0$ and so each of the four terms in (4.3.5) equals 0. Since, by Lemma 4.3.5 and Theorem 4.3.6,

$$\begin{aligned} \Delta^-(F_1, \sigma F_2; \mathbf{A}^{(1)}) &\geqslant -\Delta(F_1, \sigma F_2; \mathbf{A}^{(1)}) \\ &> -\Delta(\sigma F_1, F_2; \mathbf{A}^{(1)}) \geqslant -\Delta^+(\sigma F_1, F_2; \mathbf{A}^{(1)}) = 0, \end{aligned}$$

it must be that $E(\theta_2|F_2) = 0$ in order that the third term in (4.3.5) equals 0. Thus, the support of F_2 consists of the one point 0; so, F_1 has more than one point in its support and $\Delta(\sigma F_1, F_2; \mathbf{A}^{(1)}) > 0$, the desired contradiction. □

There is no switch-from-a-loser rule analogous to the stay-with-a-winner rule. One arm may be so much better *a priori* than the other that even many losses with it would not make the other arm worth considering. For example, F_1 may be 'wholly to the right' of F_2. The case in which one arm is known can be fruitful in understanding this asymmetry between success and failure, for in this case success and failure are equivalent except for the obvious difference in immediate income.

Robbins (1952) considered a particular stay-with-a-winner strategy: choose randomly initially and always switch on a loser. He calculated the asymptotic advantage of this 'play-the-winner rule' over random selection. This strategy has been considered extensively in the 'ranking and selection' literative (Sobel and Weiss, 1970; Nordbrock, 1976).

There are many examples where one should not select the arm having larger mean (cf. Example 1.2.1). However, myopic strategies are optimal in various special cases; one is given in the next result.

Theorem 4.3.9 Suppose that **A** is nonincreasing and F_1 and F_2 are supported by the same two points. Then an arm is optimal initially if and only if it has the larger mean. Specifically, arm 1 is optimal initially if and only if $E(\theta_1|F_1) \geq E(\theta_2|F_2)$.

Proof When both are supported by the same two points, it is easy to check that $F_1 \overset{n-1}{\geq} F_2$ for any $n \geq 1$ if and only if $E(\theta_1|F_1) \geq E(\theta_2|F_2)$. The result follows from Corollary 4.3.7. □

The following example shows that the hypothesis that **A** is nonincreasing cannot be dropped from Theorem 4.3.9, nor from either Corollary 4.3.7 or Theorem 4.3.6.

Example 4.3.3 Let $k = 2$, $\mathbf{A} = (0, 1, 0, 0, 0, \ldots)$, and suppose that $F_i = (1 - p_i)\delta_0 + p_i\delta_{1/2}$ for $i = 1, 2$ where $1 > p_1 > p_2 > 0$. It is easy to check that the only optimal initial selection is arm 2 and that it should be followed by a selection of arm 2 if and only if a success is observed at stage 1. Since $E(\theta_1|F_1) > E(\theta_2|F_2)$, the conclusion of Theorem 4.3.9 fails. The advantage gained by selecting arm 2 is due to the fact that a success with it at stage 1 indicates that it is the appropriate selection at stage 2. If p_2 is sufficiently small an initial selection of arm 1 is worthless; whatever is observed at stage 1, arm 1 will be optimal at stage 2. □

Since F_1 is strongly to the right of F_2 in the preceding example, it shows that the hypothesis on **A** cannot be dropped in either Corollary 4.3.7 or Theorem 4.3.6. The next example is a little more complicated and shows that the hypothesis on **A** cannot be dropped from Theorem 4.3.8.

Example 4.3.4 Let $k = 2$, $\mathbf{A} = (0, 0, 1, 0, 0, 0, \ldots)$, and suppose that $F_i = (1 - p_i)\delta_{1/4} + p_i\delta_{1/2}$ for $i = 1, 2$. Further suppose that $4p_1/(3 + p_1) > p_2 > p_1$. Then both (i) and (ii) of Theorem 4.3.8 are satisfied. By considering a small number of cases it is easy to check that $\Delta(F_1, F_2; \mathbf{A}) > 0$ and $\Delta(\sigma F_1, F_2; \mathbf{A}^{(1)}) < 0$; in fact, the only optimal selections (through stage 3) are given by $\tau(\varnothing) = 1$, $\tau(1) = 2$, $\tau(0) = 1$, $\tau(1, 1) = 2$, $\tau(1, 0) = 1$, $\tau(0, 1) = 1$, and $\tau(0, 0) = 2$. □

One of the early papers dealing with a two-armed bandit is Feldman (1962); see also DeGroot (1970, Section 14.7). The discount sequence

in the setting considered by Feldman is $\mathbf{A} = (1, \ldots, 1, 0, 0, \ldots)$. He considered two *dependent* arms. In particular, G is the following very special distribution: (θ_1, θ_2) is known to be either (a, b) or (b, a). So the two parameters are known but not which goes with which arm. While this problem is outside the setting of the current chapter, it is discussed here since Feldman's result is an easy consequence of Theorem 4.3.9; the basic idea for the following argument is due to Kadane (1969).

Assume that discount sequence is nonincreasing. Consider Feldman's prior and assume without loss that $a > b$. The initial probability that (θ_1, θ_2) equals (a, b) is $G(a, b)$, and $G(b, a) = 1 - G(a, b)$. Construct a new distribution G^* with support $\{(a, b), (b, a), (a, a), (b, b)\}$ so that θ_1 and θ_2 are independent and the (marginal) probability of $\{\theta_1 = a\}$ is $G(a, b)$:

$$G^*(a, b) = G^2(a, b), \qquad G^*(b, a) = G^2(b, a),$$
$$G^*(a, a) = G^*(b, b) = G(a, b)G(b, a).$$

So $F_1 = G(a, b)\delta_a + G(b, a)\delta_b$ and $F_2 = G(b, a)\delta_a + G(a, b)\delta_b$. Theorem 4.3.9 applies to G^* to show that arm 1 is optimal if and only if $G(a, b) > G(b, a)$. But if it were known in advance that the arms were identical ($\theta_1 = \theta_2 = a$ or $\theta_1 = \theta_2 = b$) then neither arm would be strictly preferred. The only possibilities that influence the preference for an arm have $\theta_1 \neq \theta_2$ (that is, either $(\theta_1, \theta_2) = (a, b)$ or (b, a)). Therefore, Theorem 4.3.9 applies to show that arm 1 is optimal when and only when it is *a priori* at least as likely that $(\theta_1, \theta_2) = (a, b)$ as that $(\theta_1, \theta_2) = (b, a)$. Therefore, we have the following as a generalization of Feldman's result.

Corollary 4.3.10 Let $0 \leqslant b \leqslant a \leqslant 1$. Suppose $k = 2$ and that the two arms are Bernoulli with parameters a for arm 1 and b for arm 2 with probability $G(a, b)$ and a for arm 2 and b for arm 1 with probability $G(b, a) = 1 - G(a, b)$. Suppose the discount sequence is nonincreasing. Then arm 1 is optimal initially if and only if $G(a, b) \geqslant 1/2$ and arm 2 is optimal initially if and only if $G(b, a) \geqslant 1/2$.

Feldman's result has been generalized in a number of other directions by Fabius and van Zwet (1970), Kelley (1974), Rodman (1978), and Zaborskis (1976). These directions are indicated in the Annotated Bibliography.

References

Berry, D. A. (1972) A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43**: 871–897.

DeGroot, M. H. (1970) *Optimal Statistical Decisions*, McGraw-Hill, New York.

Fabius, J. and van Zwet, W. R. (1970) Some remarks on the two-armed bandit. *Ann. Math. Statist.* **41**: 1906–1916.

Feldman, D. (1962) Contributions to the 'two-armed bandit' problem. *Ann. Math. Statist.* **33**: 847–856.

Kadane, J. B. (1969) Personal communication.

Kelley, T. A. (1974) A note on the Bernoulli two-armed bandit problem. *Ann. Statist.* **2**: 1056–1062.

Nordbrock, E. (1976) An improved play-the-winner sampling procedure for selecting the better of two binomial populations. *J. Amer. Statist. Assoc.* **71**: 137–139.

Quisel, K. (1965) Extensions of the two-armed bandit and related processes with on-line experimentation. Tech. Rep. No. 137, Institute for Mathematical Studies in the Social Sciences, Stanford Univ., USA.

Robbins, H. (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58**: 527–535.

Rodman, L. (1978) On the many-armed bandit problem. *Ann. Prob.* **6**: 491–498.

Sobel, M. and Weiss, G. H. (1970) Play-the-winner sampling for selecting the better of two binomial populations. *Biometrika* **57**: 357–365.

Zaborskis, A. A. (1976) Sequential Bayesian plan for choosing the best method of medical treatment. *Avtomatika i Telemekhanika* **2**: 144–153.

CHAPTER 5

Two arms, one arm known

In this chapter we assume that there are two arms ($k = 2$) and that one arm, say arm 2 for definiteness, has known mean λ. The only uncertainty is embodied in F_1, now abbreviated to F, the distribution of the random measure Q_1. For arbitrary λ we can, without loss, assume that arm 2 always produces the known observation λ. Since G is given by the pair (F, λ), we now speak of the $(F, \lambda; \mathbf{A})$-bandit.

Depending on $\mathbf{A}$, continuous-time approximations may be available. For example, if $\mathbf{A}$ is n-horizon uniform with n large and F is Bernoulli, then the example discussed in Section 8.2 applies as an approximation.

There are two main types of results in this chapter. One compares bandits that differ in some respect – say that have different λ's. The other type of result gives properties of optimal strategies. The chapter is organized partly on the basis of various rather weak restrictions on $\mathbf{A}$. We give two easy and intuitive results in the present section that apply for arbitrary discount sequences, as well as arbitrary F.

Theorem 5.0.1 For all F and $\mathbf{A}$, the value $V(F, \lambda; A)$ is a continuous nondecreasing function of λ.

Remark This result is immediate in the Bernoulli case in view of Theorems 4.1.1 and 4.1.6 since a one-point distribution on $\lambda^* > \lambda$ is strongly to the right of a one-point distribution on λ. □

Proof of Theorem 5.0.1 Suppose $\lambda^* > \lambda$ and τ is an optimal strategy in the $(F, \lambda; \mathbf{A})$-bandit; such a strategy exists in view of Theorem 2.5.2. Suppose τ is followed in the $(F, \lambda^*; \mathbf{A})$-bandit. The only change in worth as compared with the $(F, \lambda; \mathbf{A})$-bandit is when arm 2 is selected;

the corresponding observation is λ^* as opposed to λ. Therefore

$$\begin{aligned} V(F, \lambda; \mathbf{A}) &= W(F, \lambda; \mathbf{A}; \tau) \\ &\leqslant W(F, \lambda^*; \mathbf{A}; \tau) \leqslant V(F, \lambda^*; \mathbf{A}). \end{aligned}$$

To prove continuity let τ^* be optimal for the $(F, \lambda^*; \mathbf{A})$-bandit. Since the only difference between $W(F, \lambda; \mathbf{A}, \tau^*)$ and $W(F, \lambda^*; \mathbf{A}; \tau^*)$ is a result of the observations on arm 2,

$$\begin{aligned} V(F, \lambda^*; \mathbf{A}) &= W(F, \lambda^*; \mathbf{A}; \tau^*) \\ &\leqslant W(F, \lambda; \mathbf{A}; \tau^*) + (\lambda^* - \lambda)|\mathbf{A}|_1 \\ &\leqslant V(F, \lambda; \mathbf{A}) + (\lambda^* - \lambda)|\mathbf{A}|_1. \end{aligned}$$

So, not only is V a continuous function of λ, but it is absolutely continuous with a derivative bounded by $|\mathbf{A}|_1$. □

Since arm 2 is known, it is used only to achieve immediate payoff; using arm 1 can gain information as well. So arm 1 is optimal if it gives greater immediate payoff than does arm 2. (The analogous result in continuous time is illustrated in the example of Figure 8.2 by that fact that the boundary is less than 0.) This fundamental fact is intuitive and easy to show:

Theorem 5.0.2 If $E(X_{11}|F) \geqslant \lambda$ then arm 1 is optimal for any $\mathbf{A}$.

Proof Suppose τ is an optimal strategy in the $(F, \lambda; \mathbf{A}^{(1)})$-bandit. Let τ^* be the strategy in the $(F, \lambda; \mathbf{A})$-bandit which indicates arm 1 initially and then follows τ, thereby ignoring the initial result with arm 1. This strategy has worth

$$\begin{aligned} W(F, \lambda; \mathbf{A}; \tau^*) &= \alpha_1 E(X_{11}|F) + V(F, \lambda; \mathbf{A}^{(1)}) \\ &\geqslant \alpha_1 \lambda + V(F, \lambda; \mathbf{A}^{(1)}) = W(F, \lambda; \mathbf{A}; \tau_2), \end{aligned}$$

say, where τ_2 is a strategy which indicates arm 2 initially and then proceeds optimally (by following τ, for example). Since there is a strategy that starts with arm 1 and is at least as good as the best strategy that starts with arm 2, arm 1 is optimal. □

For an example of how much larger the mean of arm 2 must be to compete with an unknown arm 1, see Example 5.5.3.

The first four sections of this chapter assume arm 1 to be Bernoulli with the random parameter θ_1. As has become our convention in such a setting, we regard F as the distribution of θ_1 (on $[0, 1]$) rather than as a distribution on $\mathscr{D}$.

In Section 5.1 we give various properties of $\Delta(F, \lambda; \mathbf{A})$ when $\mathbf{A}$ is monotone. In Section 5.2 we consider discount sequences which may not be monotone and give necessary and sufficient conditions on $\mathbf{A}$ for the $(F, \lambda; \mathbf{A})$-bandit to be a stopping problem for all (F, λ). So we answer the question: for which $\mathbf{A}$ does the problem always reduce to deciding when to stop selecting arm 1?

Section 5.3 is based on the main result of Section 5.2. It contains a rather complete characterization of optimal strategies when $\mathbf{A}$ is restricted to the class determined by that result. In addition, Section 5.3 gives separate demonstrations of the results of Section 5.1 for this class, which contains some monotone and some non-monotone sequences.

Section 5.4 gives Bernoulli examples in which optimal strategies are calculated using the results of Section 5.3. (Those examples in which $\mathbf{A}$ is geometric will be given greater relevance by the Gittins–Jones result discussed in Chapter 6.) We also find bounds which provide sufficient conditions for an arm to be optimal. These bounds are compared by example with some exact derivations.

Many of the results derived in Section 5.2 and 5.3 apply also outside the Bernoulli setting. Throughout Section 5.2 and in the initial part of Section 5.3 the Bernoulli assumption is unnecessary; it is made in these sections to facilitate the presentation. The results for which we will make no general claims will be so indicated. Most such results concern bandits that arise from another bandit after a success or a failure on arm 1. In Section 5.5 we discuss general distributions for arm 1. The most important results from the earlier sections and those to be used in the later development will be repeated. A rather comprehensive example is discussed in Section 5.6; the distribution F of Q_1 is a Dirichlet process as defined by Ferguson (1973).

In Section 5.7 we turn to discounting in real time as described in Section 3.5. We describe the extent to which the results of Sections 5.2 and 5.3 apply in this setting.

5.1 Monotone discount sequences

In this section we assume arm 1 is Bernoulli.

An important class of discount sequences has $\alpha_m \geqslant \alpha_{m+1}$ for $m = 1, 2, \ldots$. With this monotonicity assumption we show that there is a 'break-even value' $\Lambda \in [0, 1]$ for which arm 2 is optimal when $\lambda \geqslant \Lambda$ and arm 1 is optimal when $\lambda \leqslant \Lambda$ (Corollary 5.1.2). We also show that when arm 1 is optimal and yields a success it is optimal again (Theorem 5.1.3). When $\mathbf{A}$ is monotone we can apply Theorem 4.3.6 to show that arm 2 becomes more desirable when λ increases.

Corollary 5.1.1 The difference $\Delta(F, \lambda; \mathbf{A})$ is strictly decreasing in λ when $\mathbf{A}$ is nonincreasing with $\mathbf{A} \neq \mathbf{0}$.

Proof Let n denote the horizon of $\mathbf{A}$. According to Theorem 4.3.6, $\Delta(F, \cdot; \mathbf{A})$ is strictly $(n-1)$-decreasing. The result follows since a one-point distribution on $\lambda^* > \lambda$ is ∞-greater than a one-point distribution on λ. □

The next result is a consequence of this fact. It states that there is a unique value of λ, say $\Lambda(F, \mathbf{A})$, such that both arms are optimal when $\lambda = \Lambda(F, \mathbf{A})$.

Corollary 5.1.2 For each nonincreasing discount sequence $\mathbf{A}$ with $\mathbf{A} \neq \mathbf{0}$ and each distribution F on $[0, 1]$, there exists a unique $\Lambda(F, \mathbf{A}) \in [0, 1]$ such that arm 1 is optimal initially in the $(F, \lambda; \mathbf{A})$-bandit if and only if $\lambda \leqslant \Lambda(F, \mathbf{A})$ and arm 2 is optimal if and only if $\lambda \geqslant \Lambda(F, \mathbf{A})$.

Remarks Corollary 5.1.2 generalizes the results of Bradt, Johnson, and Karlin (1956, Lemma 4.2) where uniform discounting was considered, and of Bellman (1956, Theorem 2) where geometric discounting was considered.

The break-even value of λ, $\Lambda(F, \mathbf{A})$, is called a 'dynamic allocation index' by Gittins and Jones (1974). It plays an especially important role in multi-armed bandits with geometric discounting (Theorem 6.1.1). □

Proof From Corollary 5.1.1, there exists a unique $\Lambda(F, \mathbf{A}) \in [0, 1]$ such that arm 1 is uniquely optimal initially in the $(F, \lambda; \mathbf{A})$-bandit if $\lambda < \Lambda(F, \mathbf{A})$ and arm 2 is uniquely optimal initially if $\lambda > \Lambda(F, \mathbf{A})$. It remains to prove that both arms are optimal initially in the

$(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit. We assume $0 < \Lambda < 1$ and leave it to the reader to consider the easier cases $\Lambda = 0$ and $\Lambda = 1$.

For $\lambda < \Lambda(F, \mathbf{A})$,

$$V^{(1)}(F, \lambda; \mathbf{A}) > V^{(2)}(F, \lambda; \mathbf{A})$$

(quantities defined in Theorem 2.5.1); while the inequality is reversed for $\lambda < \Lambda(F, \mathbf{A})$. By Theorem 5.0.1, $V^{(1)}$ and $V^{(2)}$ are continuous. Therefore,

$$V^{(1)}(F, \Lambda(F, \mathbf{A}); \mathbf{A}) = V^{(2)}(F, \Lambda(F, \mathbf{A}); \mathbf{A}),$$

which is equivalent to both arms being optimal initially for the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit.

□

It seems reasonable to expect the desirability of arm 2 to increase with λ whether or not $\mathbf{A}$ is monotone. However, the next example shows that the monotonicity hypothesis in Corollaries 5.1.1 and 5.1.2 cannot be dropped. Namely, for a particular discount sequence which is not monotonic the example shows that arm 2 can be optimal for one value of λ but not for a larger value of λ.

Example 5.1.1 Let $\mathbf{A} = (1, 0, a, 0, 0, 0, \ldots)$ and let

$$F = (5/8)\delta_0 + (1/4)\delta_{1/4} + (1/8)\delta_1;$$

so $E(\theta_1 | F) = 3/16$. We shall show that when a is sufficiently large, arm 2 is optimal initally when $\lambda = 1/4$ but not when $\lambda = 1/2$.

Let us first consider the $(F, 1/4; \mathbf{A})$-bandit. Irrespective of the selection and result at stage 1, arm 1 is clearly optimal at stage 2 since $\alpha_2 = 0$. At stage 3, arm 1 is optimal (not necessarily uniquely) if a success was obtained at stage 2 and arm 2 is optimal otherwise – this being the case irrespective of the arm selected and result obtained at stage 1. Accordingly, arm 2 is optimal at stage 1 for any $a \geqslant 0$ since $E(\theta_1 | F) = 3/16 < 1/4 = \lambda$.

Now consider the $(F, 1/2; \mathbf{A})$-bandit. Again, arm 1 can be selected without loss at stage 2. And as in the case $\lambda = 1/4$, arm 1 should not be selected at stage 3 if it has yielded a failure previously. Therefore, if arm 1 is selected initially there is only one continuation that need be considered: select arm 1 at stage 2 and again at stage 3 if and only if successes were obtained at both stages 1 and 2. The probability of the latter is $(1/8)(1)^2 + (1/4)(1/4)^2 = 9/64$. So the expected payoff using

this strategy is

$$\frac{3}{16}(1)+\left[\left(\frac{9}{64}\right)\frac{11}{12}+\left(1-\frac{9}{64}\right)\frac{1}{2}\right]a=\frac{3}{16}+\frac{143}{256}a.$$

The first term is the contribution from stage 1 and the second that of stage 3. If arm 2 is selected initially and arm 1 at stage 2 then the optimal selection at stage 3 is arm 1 in the case of a success at stage 2 ($3/4 > 1/2$) and arm 2 in case of a failure at this stage. The expected payoff of this strategy is

$$\frac{1}{2}(1)+\left[\left(\frac{3}{16}\right)\frac{3}{4}+\left(1-\frac{3}{16}\right)\frac{1}{2}\right]a=\frac{1}{2}+\frac{140}{256}a<\frac{3}{16}+\frac{143}{256}a$$

which is true when $a > 80/3$.

Hence, arm 2 is uniquely optimal when $\lambda = 1/4$ and, assuming $a > 80/3$, arm 1 is uniquely optimal when $\lambda = 1/2$! □

As a corollary of Theorem 5.0.2, and of Theorem 4.3.8, which applies when **A** is nonincreasing, it follows that arm 1 continues to be optimal if it is successful.

Theorem 5.1.3 Suppose **A** is nonincreasing. Then $\Delta(F, \lambda; \mathbf{A}) \geqslant 0$ implies $\Delta(\sigma F, \lambda; \mathbf{A}^{(1)}) \geqslant 0$. (In other words, $\Lambda(\sigma F, \mathbf{A}^{(1)}) \geqslant \Lambda(F, \mathbf{A})$.)

Proof Theorem 4.3.8 applies when $E(\theta_1 | F) \leqslant \lambda$ and Theorem 5.0.2 applies otherwise since $E(\theta_1 | \sigma F) \geqslant E(\theta_1 | F)$ by the Cauchy–Schwarz inequality. □

This result generalizes Lemma 4.6 of Bradt, Johnson, and Karlin (1956) who considered finite horizon uniform discount sequences; Theorem 2 of Bellman (1956) who considered geometric discounting; and Theorem 4.1 of Berry and Fristedt (1979) who showed it for *regular* nonincreasing discount sequences (see Definition 5.2.1).

Theorem 5.1.3 applies when **A** is nonincreasing to show that there are optimal strategies that stay with the unknown arm if it is successful. But we shall see that there are $(F, \lambda; \mathbf{A})$-bandits where all optimal strategies switch from the known arm; and since an observation on the known arm – success or not – has no effect on future payoff, this means that such strategies switch on a success. As is

shown in the next section, this cannot happen when **A** is sufficiently 'smooth'.

5.2 Regular discount sequences

In comparing the possible selections at any stage, arm 1 has two potential benefits: it can yield immediate success and it can give information to aid in future selections. On the other hand, arm 2 has no information value since its characteristics are completely known. This means that there can be no advantage in basing future selections on results from arm 2. One might therefore speculate that arm 1 can be set aside for the indefinite future once arm 2 becomes optimal. This is true in a variety of circumstances, but it is not generally true. When it is true, the problem is one of optimal stopping: when should experimentation with arm 1 cease? This section characterizes discount sequences **A** for which the $(F, \lambda; \mathbf{A})$-bandit is an optimal stopping problem for all F and λ. This characterization is of interest, but the fact that a simple characterization is possible may be more interesting.

Example 3.3.3 shows that it may be uniquely optimal to begin with arm 2 and change to arm 1 at a subsequent stage. In that example the discount sequence is a mixture of geometrics. The next example provides a much simpler setting in which this phenomenon occurs. As in Example 5.1.1, to which it is very similar, the critical aspect of the example is the discount sequence, which encourages gain rather than information-gathering at stage 1 and vice versa at stage 2.

Example 5.2.1 Let $\mathbf{A} = (1, 0, 1, 0, 0, \ldots), \lambda > 1/2$, and $F = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ (cf. Example 1.2.1). Since $\alpha_2 = 0$, complete information can be obtained on arm 1 at stage 2 without risk. The first three selections of every optimal strategy are therefore clear. Since its mean is greater than that of arm 1, arm 2 should be selected initially. Then arm 1 should be selected (at stage 2). If successful, arm 1 should be used again at stage 3 (since θ_1 is then known to be 1) and if it is not successful, arm 2 should be used at stage 3 (since θ_1 is then known to be 0). Continuations beyond stage 3 are immaterial. The maximal expected payoff is

$$V(F, \lambda; \mathbf{A}) = \lambda + (1/2 + \lambda/2). \qquad \square$$

The discount sequence in the previous example was chosen to make the issue transparent. The following example makes it clear that $\alpha_2 = 0$ and the lack of monotonicity in the discount sequence play no essential role.

Example 5.2.2 Let $\mathbf{A} = (4, 1, 1, 0, 0, 0, \ldots)$, $\lambda = 0.6$, and F be as in the previous example. It is easy to check that $V(F, 0.6; \mathbf{A}) = 3.7$ and that this expected payoff can be obtained only by the strategies indicated as optimal in the previous example. □

The phenomenon described in these examples cannot arise if the discount sequence has the regularity property that we now define; cf. (2.4.8).

Definition 5.2.1 For any discount sequence

$$\mathbf{A} = (\alpha_1, \alpha_2, \ldots) \text{ let } \gamma_m = \sum_{j=m}^{\infty} \alpha_j; \text{ } \mathbf{A} \text{ is } \textit{regular} \text{ if, for } m = 1, 2, \ldots,$$

$$\frac{\gamma_{m+2}}{\gamma_{m+1}} \leqslant \frac{\gamma_{m+1}}{\gamma_m} \tag{5.2.1}$$

provided that $\gamma_{m+1} > 0$.

The infinite sum γ_m is the maximum available worth from stage m onward, that is, it is the total worth of successes into the indefinite future.

The condition of regularity is somewhat weaker than the corresponding condition with α's in place of γ's in (5.2.1).

Proposition 5.2.1 The discount sequence $\mathbf{A}$ is regular if, for each $m = 1, 2, \ldots$, either

$$\frac{\alpha_{m+2}}{\alpha_{m+1}} \leqslant \frac{\alpha_{m+1}}{\alpha_m} \tag{5.2.2}$$

or $\alpha_{m-j} = 0$ for $j = 0, \ldots, m-1$ or $\alpha_{m+j} = 0$ for $j = 1, 2, \ldots$.

Remark Such sequences are called *superregular* in Berry and Fristedt (1983). □

Proof The hypothesis implies

$$\alpha_{m+1}\alpha_{m+j} \geqslant \alpha_m \alpha_{m+j+1}$$

for all positive integers m and j. Therefore,

$$\alpha_{m+1} \sum_{j=1}^{\infty} \alpha_{m+j} \geqslant \alpha_m \sum_{j=1}^{\infty} \alpha_{m+j+1}.$$

It follows that, for all m,

$$\alpha_{m+1}\gamma_{m+1} + \alpha_{m+1}\gamma_{m+2} + \gamma_{m+2}^2 \geqslant \alpha_m\gamma_{m+2} + \alpha_{m+1}\gamma_{m+2} + \gamma_{m+2}^2;$$

that is,

$$\gamma_{m+1}^2 \geqslant \gamma_m\gamma_{m+2}$$

and (5.2.1) follows. □

The two most important classes of regular discount sequence are the finite horizon uniform and the geometric; both have been discussed a number of times and will be considered again in this and subsequent chapters. But geometric sequences are barely regular: each inequality in (5.2.1)–and also in (5.2.2)–holds with equality for all m. Furthermore, by an application of Jensen's inequality, any nontrivial linear combination of geometrics is not regular (see $\mathbf{A}_8$ below).

Other regular sequences are:

$\mathbf{A}_1 = (1, \alpha, \alpha^2, \ldots, \alpha^{n-1}, 0, 0, \ldots), \alpha \geqslant 0,$
$\mathbf{A}_2 = (0, 0, \ldots, 0, 1, 0, 0, \ldots),$
$\mathbf{A}_3 = (0, 0, \ldots, 0, 1, 1, \ldots, 1, 0, 0, \ldots),$
$\mathbf{A}_4 = (0, 1, 2, 3, 4, 3, 2, 1, 1, 0, 0, \ldots),$
$\mathbf{A}_5 = (3, 4, 3, 4, 3, 0, 0, \ldots),$
$\mathbf{A}_6 = e^{-\alpha}(1, \alpha, \alpha^2/2, \ldots, \alpha^{m-1}/(m-1)!, \ldots),$
$\mathbf{A}_7 = (1, 1-e^{-\alpha}, 1-(1+\alpha)e^{-\alpha}, \ldots, e^{-\alpha}\Sigma_m^{\infty}\alpha^{i-1}/(i-1)!, \ldots).$

($\mathbf{A}_6$ is the sequence of Poisson probabilities and $\mathbf{A}_7$ is the cumulative Poisson sequence.)

Some nonregular sequences are:

$$\mathbf{A}_8 = \left(\frac{1}{2}, \frac{5}{16}, \frac{7}{32}, \ldots, \frac{1}{2}\left(\frac{3}{4}\right)^{m-1} + \frac{1}{2}\left(\frac{1}{4}\right)^{m-1}, \ldots\right),$$

$$\mathbf{A}_9 = (4, 1, 1, 0, 0, \ldots),$$

$$\mathbf{A}_{10} = (1, 0, a, 0, 0, \ldots), \quad a > 0.$$

($\mathbf{A}_8$ was considered in Example 3.3.3 and $\mathbf{A}_9$ in Example 5.2.2; $\mathbf{A}_{10}$ was considered in Example 5.1.1 and in Example 5.2.1 with $a = 1$.) If $\mathbf{A}$ is regular and $\alpha_m = 0$ then either $\alpha_1 = \ldots = \alpha_{m-1} = 0$ or

$\alpha_{m+1} = \alpha_{m+2} = \ldots = 0$ (compare $\mathbf{A}_2$ and $\mathbf{A}_{10}$). And if $\alpha_{m+1} \geqslant \alpha_m$ then condition (5.2.1) is always satisfied; that is, regularity cannot be violated on an increasing part of a discount sequence.

The importance of regularity of discount sequences derives from the next result. It is an extension of Theorem 2.1 of Berry and Fristedt (1979). A modification of the proof given there applies in this more general setting in which **A** need not be monotone, and is given below. (The result holds beyond the present Bernoulli context and so will be discussed again in Sections 5.5 and 5.7).

Theorem 5.2.2 Consider a discount sequence **A**. The following two statements are equivalent:

(i) For every $(F, \lambda;\ \mathbf{A})$-bandit there is an optimal strategy under which every selection of the known arm, arm 2, is followed by another selection of arm 2;
(ii) **A** is regular.

Furthermore, if $\alpha_1 > 0$ and **A** is regular, then if ever arm 2 becomes optimal, an optimal continuation is to select arm 2 exclusively and indefinitely.

The most useful half of this result is (ii) $\Rightarrow$ (i). Before proving the theorem we give an intuitive demonstration of this implication when **A** is the n-horizon uniform that is due to Bradt, Johnson, and Karlin (1956, Lemma 4.1); the reader will see a resemblance between this demonstration and the proof that is given below. (It is surprising to us that the existence of such a trivial argument in this case has not been appreciated by a number of authors.)

Suppose at some stage – there is no loss in saying the first stage – that arm 2 is uniquely optimal. It is selected for a number of stages and eventually, but before stage n, arm 1 becomes optimal. Is this possible? Suppose this latter selection of arm 1 is exchanged with the initial selection of arm 2. Nothing is lost since successes are exchangeable when the discount factors are equal. Therefore arm 1 is also optimal initially and so the assumption that arm 2 is uniquely optimal initially is contradicted. And (i) of the theorem follows in this special case.

Actually, selecting arm 1 early (but not committing to arm 2 thereafter) has some benefit over using it late; namely, there is more opportunity to use any information that is gained. If the discount

sequence is not uniform something may be lost when two selections are interchanged. It will develop in the following proof that regularity is precisely the condition on the discount sequence that assures that this loss will not outweigh the gain from obtaining early information.

Proof of Theorem 5.2.2

Part I, (ii) $\Rightarrow$ (i). Let $\mathscr{S}_n$ denote the set of all regular discount sequences with horizon n. For any $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$ the notational conventions $\Sigma_{m=\infty}^{\infty}\alpha_m = \gamma_\infty = 0$ will be useful. The proof is by induction on n.

Clearly, (i) holds for every member of $\mathscr{S}_0 \cup \mathscr{S}_1$. Suppose $n \geqslant 2$. Assume it holds for every member of $\mathscr{S}_{n-1}$. Let $\mathbf{A} = (\alpha_1, \alpha_2, \ldots) \in \mathscr{S}_n$; then $\mathbf{A}^{(1)} = (\alpha_2, \alpha_3, \ldots) \in \mathscr{S}_{n-1}$. Assume it is optimal to select arm 1 initially. The inductive hypothesis applies to show that there is an optimal continuation which never switches back to arm 1 after a switch to arm 2. If it is optimal to select arm 2 initially, then the inductive hypothesis applies immediately to show (i) unless an optimal strategy has the form τ^*: select arm 2 initially, select arm 1 at stages $2, \ldots, N$, and arm 2 subsequently. The stage N is random with $P(N > 1) = 1$; it may be infinite with positive probability and it may depend on the history of observations. We may assume that τ^* does not depend on the initial observation on arm 2. So for each m, $\{N > m\}$ is measurable with respect to the σ-field generated by the outcomes of the selections of arm 1 at stages 2 through m, that is, the σ-field generated by $(X_{12}, \ldots, X_{1m})$. By Theorem 5.0.2 we may assume with no loss of generality that if s successes and $f = m - s - 1$ failures have been obtained with arm 1 at stages 2 through m when following τ^*, then

$$N = m \Rightarrow E(\theta_1 | \sigma^s \varphi^f F) < \lambda. \tag{5.2.3}$$

We show that there is a strategy τ which starts with arm 1 and is at least as good as τ^*. We choose τ by modifying τ^* as follows: select arm 1 initially and imitate τ^* subsequently by selecting the indicated arm one stage earlier. The worth of τ^* is

$$W(F, \lambda; \mathbf{A}; \tau^*) = E_{\tau^*}(\lambda\alpha_1 + \sum_{m=2}^{N} X_{1m}\alpha_m + \lambda \sum_{m=N+1}^{\infty} \alpha_m | F), \tag{5.2.4}$$

which, since τ^* is optimal, is no smaller than $\lambda\gamma_1$. Hence,

$$\sum_{m=2}^{\infty} E_{\tau^*}[(X_{1m} - \lambda)\mathbf{I}_{\{N \geqslant m\}} | F]\alpha_m = E_{\tau^*}[\sum_{m=2}^{N} (X_{1m} - \lambda)\alpha_m | F] \geqslant 0. \tag{5.2.5}$$

The worth of τ is

$$W(F, \lambda; \mathbf{A}; \tau) = E_{\tau^*}\left[\sum_{m=2}^{N} X_{1m}\alpha_{m-1} + \lambda \sum_{m=N+1}^{\infty} \alpha_{m-1} | F\right], \quad (5.2.6)$$

which we show to be at least as large as the worth of τ^*. Subtracting (5.2.4) from (5.2.6) gives

$$\sum_{m=2}^{\infty} E_{\tau^*}[(X_{1m} - \lambda)\mathbf{I}_{\{N \geq m\}} | F](\alpha_{m-1} - \alpha_m). \quad (5.2.7)$$

That (5.2.7) is nonnegative follows from (5.2.5) by showing that

$$\sum_{m=2}^{\infty} b_m\alpha_m \geq 0 \Rightarrow \sum_{m=2}^{\infty} b_m(\alpha_{m-1} - \alpha_m) \geq 0, \quad (5.2.8)$$

where

$$b_m = E_{\tau^*}[(X_{1m} - \lambda)\mathbf{I}_{\{N \geq m\}} | F].$$

Write the first sum in (5.2.8) as follows:

$$\sum_{m=2}^{\infty} b_m\alpha_m = \sum_{m=2}^{\infty} b_m(\gamma_m - \gamma_{m+1}) = b_2\gamma_2 + \sum_{m=2}^{\infty} (b_{m+1} - b_m)\gamma_{m+1}.$$

We multiply the first inequality in (5.2.8) by α_1/γ_2 to obtain

$$b_2\alpha_1 + \sum_{m=2}^{\infty} (b_{m+1} - b_m)\gamma_{m+1}\alpha_1/\gamma_2 \geq 0. \quad (5.2.9)$$

The second sum in (5.2.8) can be written

$$\sum_{m=2}^{\infty} b_m(\alpha_{m-1} - \alpha_m) = b_2\alpha_1 + \sum_{m=2}^{\infty} (b_{m+1} - b_m)\alpha_m.$$

Accordingly, the second inequality in (5.2.8) becomes

$$b_2\alpha_1 + \sum_{m=2}^{\infty} (b_{m+1} - b_m)\alpha_m \geq 0. \quad (5.2.10)$$

In view of (5.2.9) and (5.2.10), (5.2.8) follows from two facts. The first is immediate for regular discount sequences: $\gamma_{m+1}\alpha_1/\gamma_2 \leq \alpha_m$, $m = 2, 3, \ldots$. The second is that the sequence $(b_1, b_2, \ldots)$ is nondecreasing. To show the latter, write

$$\begin{aligned} b_{m+1} - b_m &= E_{\tau^*}[(\lambda - X_{1,m+1})\mathbf{I}_{\{N=m\}} | F] \\ &\quad + E_{\tau^*}[(X_{1,m+1} - X_{1m})\mathbf{I}_{\{N \geq m\}} | F] \\ &= E_{\tau^*}[(\lambda - X_{1,m+1})\mathbf{I}_{\{N=m\}} | F] \geq 0. \end{aligned}$$

For the second equality we used the exchangeability of the X_{1j}'s and the fact that $\{N \geq m\}$ is measurable with respect to the σ-field generated by $\{X_{12}, \ldots, X_{1,m-1}\}$; for the inequality we used (5.2.3).

We have proved (i) for any regular discount sequence having finite horizon. Now assume **A** is a regular discount sequence having infinite horizon. Let τ be the optimal strategy identified in the proof of Theorem 2.5.2; τ indicates a selection of arm 2 initially if and only if the maximum in (2.5.2) is attained for $i = 2$ but not for $i = 1$. A similar statement holds at any stage for the appropriate modification of (2.5.2). Because both sides of (2.5.2) depend continuously on **A**, we see that if τ indicates a selection of arm 2 at some stage, then arm 2 is also uniquely optimal for finite horizon approximations for a sufficiently large horizon. For each such approximation arm 2 is optimal thereafter and so, again using the fact that both sides of (2.5.2) depend continuously on **A**, we conclude that arm 2 is optimal thereafter for the infinite horizon bandit.

We have shown that (ii) $\Rightarrow$ (i). Now assume (ii) and $\alpha_1 \neq 0$. We show that whenever arm 2 becomes optimal, an optimal continuation is to select arm 2 exclusively and indefinitely. The proof is by contradiction. Suppose that arm 2 is optimal at stage 1 and arm 1 is uniquely optimal at stage 2. We have proven that there is an optimal strategy that indicates arm 1 through a random stage $N \geq 2$ (possibly $+\infty$) after which arm 2 is indicated exclusively and indefinitely; call this strategy τ^*. To contradict the optimality of τ^* we imitate the above development but now we must ensure that the second inequality in (5.2.8) is strict. The unique optimality of arm 1 at stage 2 gives strict inequality at (5.2.5) and therefore in the first inequality in (5.2.8). With the additional assumption $\alpha_1 \neq 0$, the argument used above to prove (5.2.8) applies to give the desired strict inequality.

We have shown that arm 1 cannot be uniquely optimal at stage 2 when arm 2 is optimal and selected initially. We generalize by supposing that, with positive probability, it is optimal to select arm 2 at some stage m ($m > 1$) and uniquely optimal to select arm 1 at stage $m + 1$. Then $\alpha_m \neq 0$ follows from the assumptions that $\alpha_1 \neq 0$ and **A** is regular, for were $\alpha_m = 0$, then α_{m+j} would equal 0 for $j = 1, 2, \ldots$, and neither arm would be uniquely optimal at stage $m + 1$. So we can simply drop the first $m - 1$ stages, leaving the mth stage as the new first stage and a discount sequence whose first member is nonzero. The argument of the preceding paragraph now applies to give a contradiction.

Part II, not (ii) $\Rightarrow$ not (i). Suppose that

$$\gamma_M \gamma_{M+2} > \gamma_{M+1}^2 \tag{5.2.11}$$

for some M. We shall prove that (i) fails by finding a pair (F, λ) for which there is a strategy τ that follows a selection of arm 2 with one of arm 1 (on a history that has positive probability under τ) and which is strictly better than every strategy that does not.

For an $\varepsilon \in (0, 1)$ still to be specified, define F so that after $M-1$ failures on arm 1, the probabilities that $\theta_1 = 0$ and $\theta_1 = 1-\varepsilon$ both are $1/2$; that is

$$\varphi^{M-1}F = \tfrac{1}{2}\delta_0 + \tfrac{1}{2}\delta_{1-\varepsilon}.$$

This can be accomplished by setting

$$F = \frac{\varepsilon^{M-1}}{1+\varepsilon^{M-1}}\delta_0 + \frac{1}{1+\varepsilon^{M-1}}\delta_{1-\varepsilon}.$$

Since $\sigma F = \delta_{1-\varepsilon}$ we may restrict our attention to strategies under which a success with arm 1 is followed indefinitely by selections of arm 1. Among such strategies, the only ones having the property that no selection of arm 2 is followed by a selection of arm 1 are the strategies $\tau_J, J = 0, 1, \ldots, \infty$: select arm 1 at the first J stages; thereafter select arm 1 or 2 according as a success was or was not observed at one or more of the first J stages. A straightforward calculation gives:

$$W(F, \lambda; \mathbf{A}; \tau_J) = \frac{(1-\varepsilon)\gamma_1}{1+\varepsilon^{M-1}} + \lambda\frac{(\varepsilon^{M-1}+\varepsilon^J)\gamma_{J+1}}{1+\varepsilon^{M-1}} - \frac{(1-\varepsilon)\varepsilon^J\gamma_{J+1}}{1+\varepsilon^{M-1}}. \tag{5.2.12}$$

We shall define a strategy τ and show that it is better than each τ_J, though it is not necessarily optimal: select arm 1 at the first $M-1$ stages; continue indefinitely with arm 1 thereafter if a success was observed at one or more of the first $M-1$ stages, and, if not, select arm 2 at stage M, arm 1 at stage $M+1$, and thereafter select arm 1 or 2 indefinitely according as a success was or was not obtained at stage $M+1$. A straightforward calculation gives:

$$W(F, \lambda; \mathbf{A}; \tau) = \frac{(1-\varepsilon)\gamma_1}{1+\varepsilon^{M-1}} + \lambda\frac{\varepsilon^{M-1}[2\alpha_M + (1+\varepsilon)\gamma_{M+2}]}{1+\varepsilon^{M-1}} - \frac{(1-\varepsilon)(\varepsilon^{M-1}\alpha_M + \varepsilon^M\gamma_{M+2})}{1+\varepsilon^{M-1}}. \tag{5.2.13}$$

For $J \leqslant M-2$, subtraction of (5.2.12) from (5.2.13) gives the following necessary and sufficient condition for τ to be better than τ_J:

$$\lambda < \frac{(1-\varepsilon)(\gamma_{J+1} - \varepsilon^{M-1-J}\alpha_M - \varepsilon^{M-J}\gamma_{M+2})}{\gamma_{J+1} - \varepsilon^{M-1-J}[2\alpha_M + (1+\varepsilon)\gamma_{M+2} - \gamma_{J+1}]} \tag{5.2.14}$$

which is asymptotic to $1-\varepsilon$ as $\varepsilon \downarrow 0$. For $J = M-1$. (5.2.14) is again the correct condition, and it can be rewritten in this case as

$$\lambda < \frac{(1-\varepsilon)(\gamma_{M+1} - \varepsilon\gamma_{M+2})}{2\gamma_{M+1} - \gamma_{M+2} - \varepsilon\gamma_{M+2}}. \tag{5.2.15}$$

For $J \geqslant M$ and ε small, the necessary and sufficient condition for τ to be better than τ_J is

$$\lambda > \frac{(1-\varepsilon)(\alpha_M + \varepsilon\gamma_{M+2} - \varepsilon^{J-M+1}\gamma_{J+1})}{2\alpha_M - \alpha_{M+1} + (\gamma_{M+1} - \gamma_{J+1}) + \varepsilon(\gamma_{M+2} - \varepsilon^{J-M}\gamma_{J+1})} \tag{5.2.16}$$

which is asymptotically (as $\varepsilon \downarrow 0$) bounded above by

$$\frac{(1-\varepsilon)\alpha_M}{2\alpha_M - \alpha_{M+1}}$$

uniformly in $J \geqslant M$. Conditions (5.2.14), (5.2.15), and (5.2.16) can all be satisfied by an appropriately small ε provided

$$\frac{\alpha_M}{2\alpha_M - \alpha_{M+1}} < \frac{\gamma_{M+1}}{2\gamma_{M+1} - \gamma_{M+2}}.$$

This is an easy consequence of (5.2.11). □

The fact that (ii) implies (i) in Theorem 5.2.2 means that whenever the discount sequence is regular the decision maker need only decide when, if ever, to stop selecting arm 1. The worth of using arm 2 exclusively from stage m onward is $\lambda\gamma_m$. Thus $\lambda\gamma_m$ could be offered at stage m as a lump-sum alternative to selecting arm 1 and the decision problem would not be changed. So the name 'one-armed bandit' is appropriate for a two-armed bandit with one known arm in the case of regular discounting.

Example 3.3.3, in which **A** is a mixture of two geometric sequences ($\mathbf{A}_8$ above) shows how complicated the solution of a two-armed bandit with one arm known can be when the discount sequence is not regular. In that example, F is the simplest nondegenerate distribution possible in a bandit problem, and yet arm 2 can be selected an arbitrary number of times (depending on λ) before testing arm 1.

Theorem 5.2.2 does not address the question of unique optimality. It is possible for both arms to be optimal at stage 1. Suppose both arms are optimal and that the discount sequence is geometric: $\mathbf{A} = (1, \alpha, \alpha^2, \ldots)$. If arm 2 is selected then no information is gained; the new discount sequence is $(\alpha, \alpha^2, \ldots) = \alpha\mathbf{A}$ and the problem is essentially unchanged. So again both arms are optimal – and so on. There are many optimal strategies and, in particular, there is one which specifies an arbitrary fixed number of selections of arm 2 and then arm 1, regardless of the results obtained with arm 2. But, of course, the theorem is not violated: there are optimal strategies that stay with arm 2 once it is used.

There are obviously a variety of technically different optimal strategies when the horizon is finite ($\alpha_{n+1} = \ldots = 0$ for some n) since it makes no difference what decisions are made at stages later than n. On the other hand, during an initial segment of 0's in the discount sequence, arm 1 is always optimal; but even here, if there is no information on arm 1 to be obtained that will help in later decisions, or if information gathering can safely be delayed (for instance, if θ_1 is either 0 or 1 and $\alpha_2 = 0$ as well as $\alpha_1 = 0$), then arm 2 will also be optimal. In the latter instance arm 2 would not be used indefinitely.

In Section 1.2 we indicated that patients arriving late in a clinical trial were likely to be treated better than those arriving early when an optimal assignment procedure is followed. The next result verifies this characteristic when but one arm is unknown and the discount sequence is regular. It says that the expected value of an observation increases with time when an optimal strategy is followed. It should be compared with Example 5.2.2, which provides a counterexample in the nonregular case. It should also be compared with Example 2.4.5, which provides a counterexample in a case which is regular but which involves two unknown arms.

Corollary 5.2.3 If $\mathbf{A}$ is regular there exists an optimal strategy τ for the $(F, \lambda; \mathbf{A})$-bandit such that

$$E_\tau(Z_{m+1}|F, \lambda) \geqslant E_\tau(Z_m|F, \lambda) \tag{5.2.17}$$

for $m = 1, 2, \ldots$.

Proof If arm 2 is uniquely optimal initially then take $\tau(\varnothing) = 2$; otherwise take $\tau(\varnothing) = 1$. Define τ inductively to be an optimal strategy that indicates arm 1 whenever possible without requiring a

selection of arm 1 following a selection of arm 2. (The fuss in the preceding sentences is required to take care of the possibility that some discount factors may be zero.) We prove (5.2.17) only for $m = 1$; the general case is similar, but requires conditioning on the first $m-1$ stages.

If $\tau(\varnothing) = 2$, then $\tau(z_1) = 2$ for all z_1, so $E_\tau(Z_2|F, \lambda) = \lambda$ as does $E_\tau(Z_1|F, \lambda)$. If $\tau(\varnothing) = 1$, then $\tau(1) = 1$ and there are two cases. First suppose that $\tau(0) = 1$. Then $E_\tau(Z_2|F, \lambda) = E(\theta_1|F)$ as does $E_\tau(Z_1|F, \lambda)$. Next suppose $\tau(0) = 2$. Then

$$E(Z_2|F, \lambda) = E(\theta_1^2|F) + E(1-\theta_1|F)\lambda$$

which, by Theorem 5.0.2, is at least as large as

$$\begin{aligned} &E(\theta_1^2|F) + E(1-\theta_1|F)E(\theta_1|\varphi F) \\ &\quad = E(\theta_1^2|F) + E(1-\theta_1|F)E(\theta_1(1-\theta_1)|F)/E(1-\theta_1|F). \\ &\quad = E(\theta_1|F) = E_\tau(Z_1|F, \lambda). \end{aligned}$$

□

5.3 Optimal strategies – regular discounting

When **A** is regular, the major result of the previous section allows us to obtain a number of results concerning optimal strategies. Though we assume regularity throughout this section, we apply it mainly through its consequence that a bandit with a regular discount sequence is a stopping problem. Results thus obtained generalize to any $(F, \lambda; \mathbf{A})$-bandit that is a stopping problem. When **A** is not regular, deciding whether it is a stopping problem seems as difficult as actually finding an optimal strategy. So we do not incorporate this extra generality explicitly.

The present context is that arm 1 is Bernoulli. But the first two results in this section are true more generally and will be discussed again in Section 5.5.

Corollary 5.1.2 shows the existence of a unique Λ when **A** is nonincreasing. Example 5.1.1 shows that the result does not hold when **A** is not monotone. But it does hold for nonmonotone sequences that are regular. The following theorem is the same as Corollary 5.1.2 with 'regular' in place of 'nonincreasing'. The proof is essentially that of Berry and Fristedt (1979, Theorem 2.2) applied to this more general setting.

Theorem 5.3.1 For each regular discount sequence $\mathbf{A}$ with $\alpha_1 > 0$ and each distribution F on $[0, 1]$, there exists a unique $\Lambda(F, \mathbf{A}) \in [0, 1]$ such that arm 1 is optimal initially in the $(F, \lambda; \mathbf{A})$-bandit if and only if $\lambda \leqslant \Lambda(F, \mathbf{A})$ and arm 2 is optimal if and only if $\lambda \geqslant \Lambda(F, \mathbf{A})$.

Proof Let $\Lambda(F, \mathbf{A}) = \inf\{\lambda \in [0, 1]$: arm 2 is optimal for the $(F, \lambda; \mathbf{A})$-bandit$\}$. So arm 1 is uniquely optimal if $\lambda < \Lambda(F, \mathbf{A})$. An easy modification of the appropriate portion of the proof of Corollary 5.1.2 shows that both arms are optimal initially for the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit. Therefore,

$$V(F, \Lambda(F, \mathbf{A}); \mathbf{A}) = \gamma_1 \Lambda(F, \mathbf{A}), \tag{5.3.1}$$

where, as previously defined, $\gamma_1 = |\mathbf{A}|_1 = \sum_{m=1}^{\infty} \alpha_m$.

It remains to show that arm 1 is not optimal for the $(F, \lambda; \mathbf{A})$-bandit when $\lambda > \Lambda(F, \mathbf{A})$. Consider the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit and a strategy τ^* that begins with a selection of arm 1. Then

$$\mathrm{V}(\mathrm{F}, \Lambda(F, \mathbf{A}); \mathbf{A}) \geqslant W(F, \mathbf{A}); \mathbf{A}; \tau^*). \tag{5.3.2}$$

Selections of arm 2 determine the difference in payoffs for the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$- and $(F, \lambda; \mathbf{A})$-bandits when strategy τ^* is used in both. Thus

$$\begin{aligned} W(F, \lambda; \mathbf{A}; \tau^*) - W(F, \Lambda(F, \mathbf{A}); \mathbf{A}; \tau^*) &\leqslant [\lambda - \Lambda(F, \mathbf{A})]\gamma_2 \\ &< [\lambda - \Lambda(F, \mathbf{A})]\gamma_1. \end{aligned} \tag{5.3.3}$$

From (5.3.1), (5.3.2), and (5.3.3) we see that

$$W(F, \lambda; \mathbf{A}; \tau^*) < \lambda\gamma_1 \leqslant V(F, \lambda; \mathbf{A}),$$

the second inequality following because selecting arm 2 exclusively has worth $\lambda\gamma_1$. So τ^* is not optimal for the $(F, \lambda; \mathbf{A})$-bandit. □

The next result gives a method for calculating Λ, and therefore, for finding optimal strategies. It has been shown by a number of authors for particular discount sequences; notably, Bradt, Johnson, and Karlin (1956, Theorem 4.1) and Gittins and Jones (1979).

Corollary 5.3.2 For **A** regular with $\alpha_1 > 0$, the function Λ is given by

$$\Lambda(F, A) = \max_{\tau(\varnothing)=1} \frac{E_\tau\left(\sum_{m=1}^{M} \alpha_m X_{1m} | F\right)}{E_\tau\left(\sum_{m=1}^{M} \alpha_m | F\right)} \tag{5.3.4}$$

where M is the (random) stage (possibly $+\infty$) at which arm 1 is used for the last time when following strategy τ. Among those τ's that begin with arm 1 and never switch back to arm 1 after a selection of arm 2, those that are optimal for the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit are those for which the maximum in (5.3.4) is attained.

Remarks Various example calculations illustrating (5.3.4) will be carried out in the next section. The right-hand side of (2.4.9) is a special case of (5.3.4). □

Proof of Corollary 5.3.2. In view of Theorem 5.2.2, the only strategies that need be considered are those that begin with arm 1 and never return to arm 1 after switching to arm 2. For such a strategy τ,

$$\begin{aligned} E_\tau\left(\sum_{m=1}^{M} \alpha_m Z_m | F\right) + \Lambda(F, \mathbf{A}) E_\tau\left(\sum_{m>M} \alpha_m | F\right) \\ = W(F, \Lambda(F, \mathbf{A}); \mathbf{A}; \tau) \\ \leqslant V(F, \Lambda(F, \mathbf{A}); \mathbf{A}) = \Lambda(F, \mathbf{A}) \sum_{m=1}^{\infty} \alpha_m, \end{aligned}$$

where the last equality follows because, by Theorems 5.2.2 and 5.3.1, arm 2 is optimal indefinitely for the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit. Hence,

$$E_\tau\left(\sum_{m=1}^{M} \alpha_m Z_m \middle| F\right) \leqslant \Lambda(F, \mathbf{A}) E_\tau\left(\sum_{m=1}^{M} \alpha_m \middle| F\right). \tag{5.3.5}$$

Equality holds in (5.3.5) if and only if τ is an optimal strategy. □

The statement of the next result would be meaningful in the general setting and not just the Bernoulli, but an example similar to Example 2.5.1 can be constructed to show it would not hold.

Corollary 5.3.3 Assume **A** is regular and $\alpha_1 > 0$. Then $\Lambda(F, \mathbf{A})$ is a continuous function of F.

Proof We use superscripts on F to identify distributions that are terms of a sequence. Suppose that $F^n \to F$ as $n \to \infty$ and $\Lambda(F^n, \mathbf{A}) \to \lambda$. We will show that both arms are optimal initially for the $(F, \lambda; \mathbf{A})$-bandit and, therefore, $\lambda = \Lambda(F, \mathbf{A})$. By Theorem 4.1.1,

$$V^{(i)}(F^n, \Lambda(F^n, \mathbf{A}); \mathbf{A}) \to V^{(i)}(F, \lambda; \mathbf{A})$$

for $i = 1, 2$. Since

$$V^{(1)}(F^n, \Lambda(F^n, \mathbf{A}); \mathbf{A}) = V^{(2)}(F^n, \Lambda(F^n, \mathbf{A}); \mathbf{A}),$$

we obtain $V^{(1)}(F, \lambda; \mathbf{A}) = V^{(2)}(F, \lambda; \mathbf{A})$ as desired. □

We turn to results whose statements or proofs are set in the Bernoulli context. They may be true more generally, perhaps in modified form. The weak inequality in the first result follows from Theorem 4.1.6; some additional work is required to obtain the strict inequality.

Corollary 5.3.4 Assume $\mathbf{A}$ is regular, $\alpha_1 > 0$, and F^* is strongly to the right of F. Then

$$\Lambda(F^*, \mathbf{A}) \geqslant \Lambda(F, \mathbf{A}) \tag{5.3.6}$$

with equality if and only if $F^* = F$.

Proof For a proof by contradiction suppose that F^* is strongly to the right of F, $F^* \neq F$, and $\Lambda(F^*, \mathbf{A}) \leqslant \Lambda(F, \mathbf{A})$. Arm 1 is optimal initially for both the $(F^*, \Lambda(F^*, \mathbf{A}); \mathbf{A})$-bandit and the $(F, \Lambda(F^*, \mathbf{A}); \mathbf{A})$-bandit. By Corollary 2.5.3, Lemmas 4.1.3 and 4.1.4, and Theorem 4.1.6,

$$\begin{aligned}
&V(F^*, \Lambda(F^*, \mathbf{A}); \mathbf{A}) - V(F, \Lambda(F^*, \mathbf{A}); \mathbf{A}) \\
&\quad = \alpha_1 E(X_{11} | F^*) + P(X_{11} = 1 | F^*) V(\sigma F^*, \Lambda(F^*, \mathbf{A}); \mathbf{A}^{(1)}) \\
&\qquad + P(X_{11} = 0 | F^*) V(\varphi F^*, \Lambda(F^*, \mathbf{A}); \mathbf{A}^{(1)}) \\
&\qquad - \alpha_1 E(X_{11} | F) - P(X_{11} = 1 | F) V(\sigma F, \Lambda(F^*, \mathbf{A}); \mathbf{A}^{(1)}) \\
&\qquad - P(X_{11} = 0 | F) V(\varphi F, \Lambda(F^*, \mathbf{A}); \mathbf{A}^{(1)}) \\
&\quad > [P(X_{11} = 1 | F^*) - P(X_{11} = 1 | F)] V(\sigma F, \Lambda(F^*, \mathbf{A}); \mathbf{A}^{(1)}) \\
&\qquad - [P(X_{11} = 0 | F) - P(X_{11} = 0 | F^*)] V(\varphi F, \Lambda(F^*, \mathbf{A}); \mathbf{A}^{(1)}),
\end{aligned}$$

which is nonnegative since

$$V(\sigma F, \Lambda(F^*, \mathbf{A}); \mathbf{A}^{(1)}) \geqslant V(\varphi F, \Lambda(F^*, \mathbf{A}); \mathbf{A}^{(1)}),$$

a consequence of Lemma 4.1.5 and Theorem 4.1.6. On the other hand, since the strategy that indicates arm 2 at every stage is optimal for the $(F^*, \Lambda(F^*, \mathbf{A}); \mathbf{A})$-bandit,

$$V(F^*, \Lambda(F^*, \mathbf{A}); \mathbf{A}) - V(F, \Lambda(F^*, \mathbf{A}); \mathbf{A}) = \gamma_1 \Lambda(F^*, \mathbf{A}) - V(F, \Lambda(F^*, \mathbf{A}); \mathbf{A}) \leqslant 0,$$

a contradiction. □

The following characterization of optimal strategies is the analogue of Theorem 5.1.3. It says that when $\mathbf{A}$ is regular, arm 1 should be selected again if it was optimal and gave a success. Since Theorem 5.1.3 applies when $\mathbf{A}$ is nonincreasing, we now have two separate proofs of the result when $\mathbf{A}$ is nonincreasing and also regular.

Theorem 5.3.5 Let $\mathbf{A} = (\alpha_1, \alpha_2, \alpha_3, \ldots)$ be a regular discount sequence with $\alpha_1 > 0$ and $\alpha_2 > 0$. Then for all F,

$$\Lambda(F, \mathbf{A}) \leqslant \Lambda(\sigma F, \mathbf{A}^{(1)}),$$

with equality if and only if F is a one-point distribution.

Remarks While it is picturesque, this stay-with-a-winner characterization offers little help in finding an optimal strategy via dynamic programming, though some calculations can be avoided. On the other hand, as we shall see in Theorem 5.3.7, it does aid in the use of Corollary 5.3.2 since it substantially reduces the number of strategies to be considered in (5.3.4). □

Proof of Theorem 5.3.5 For a proof by contradiction we suppose that $\Lambda(F, \mathbf{A}) \geqslant \Lambda(\sigma F, \mathbf{A}^{(1)})$ and F is not a one-point distribution. By Corollary 5.3.4 $\Lambda(\sigma F, \mathbf{A}^{(1)}) \geqslant \Lambda(\varphi F, \mathbf{A}^{(1)})$. Therefore, an optimal strategy for the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit indicates arm 1 at stage 1 and arm 2 thereafter. So,

$$V(F, \Lambda(F, \mathbf{A}); \mathbf{A}) = \alpha_1 E(\theta_1 | F) + \gamma_2 \Lambda(F, \mathbf{A}),$$

which must be at least as large as $\alpha_1 \Lambda(F, \mathbf{A}) + \gamma_2 \Lambda(F, \mathbf{A})$, the worth of the strategy that indicates arm 2 at every stage. Hence,

$$\Lambda(F, \mathbf{A}) \leqslant E(\theta_1 | F) < E(\theta_1 | \sigma F),$$

where strict inequality applies since F is not a one-point distribution.

This contradicts

$$E(\theta_1 | \sigma F) \leqslant \Lambda(\sigma F, \mathbf{A}^{(1)}) \leqslant \Lambda(F, \mathbf{A}),$$

which is a consequence of Theorem 5.0.2 and our supposition. □

The next result is the analogue of Theorem 5.3.5 in case of failure. It indicates that the inclination to select arm 1 is not increased when a failure is observed. It is, however, not a 'switch-with-a-loser' rule; for, though a failure on arm 1 decreases its relative utility, it may still be preferable to arm 2. The theorem generalizes the corresponding result of Bellman (1956, Theorem 2) in the geometric case.

Theorem 5.3.6 Let $\mathbf{A}$ be a regular discount sequence with $\alpha_1 > 0$ and $\alpha_2 > 0$. Then for all F,

$$\Lambda(\varphi F, \mathbf{A}^{(1)}) \leqslant \Lambda(F, \mathbf{A}),$$

with equality if and only if F is a one-point distribution.

Proof By Corollary 5.3.4 and Lemma 4.1.5, $\Lambda(\varphi F, \mathbf{A}^{(1)}) \leqslant \Lambda(F, \mathbf{A}^{(1)})$ with equality if and only if F is a one-point distribution. By Theorem 5.2.2, $\Lambda(F, \mathbf{A}^{(1)}) \leqslant \Lambda(F, \mathbf{A})$. □

The next theorem is a corollary of the previous two theorems. It says that the search for optimal strategies can be restricted in the regular case to the one using arm 2 exclusively and those of the following form. Use arm 1 as long as it is successful. When and if it fails, continue with arm 1 only if there were a sufficient number of successes on arm 1, say r_1, before the first failure and otherwise switch permanently to arm 2. If the number of successes prior to the first failure is at least r_1 and so arm 1 is continued, then it is again used as long as it is successful. When and if arm 1 fails for the second time, continue with arm 1 if and only if the total number of successes on arm 1 is at least $r_1 + r_2$. Proceeding in this fashion, arm 1 is continued after the jth failure if and only if it has produced at least $r_1 + \ldots + r_j$ successes. Let $\tau^{(\mathbf{R})}$ denote the strategy just described that corresponds to the sequence $\mathbf{R} = (r_1, r_2, \ldots)$, $r_i \in \{0, 1, \ldots, \infty\}$.

Theorem 5.3.7 Suppose $\mathbf{A}$ is regular and $\alpha_1 > 0$. If $\lambda \leqslant \Lambda(F, \mathbf{A})$, then some $\tau^{(\mathbf{R})}$ is an optimal strategy. If $\lambda = \Lambda(F, \mathbf{A})$, r_1 can be chosen to be positive.

Remark As a consequence of this theorem, the maximum in (5.3.4) can be taken over strategies $\tau^{(\mathbf{R})}$ with $r_1 > 0$. □

Proof of Theorem 5.3.7 Since $\lambda \leqslant \Lambda(F, \mathbf{A})$, there is an optimal strategy for the $(F, \lambda; \mathbf{A})$-bandit which begins with arm 1 and, by Theorem 5.3.5, switches to arm 2, if ever, only after a failure on arm 1. Corollary 5.3.4 and Lemma 4.1.5 apply to show the existence of $j_i \in \{0, 1, \ldots, \infty\}$ such that it is optimal to switch after the ith failure if the number of successes is less than j_i and not to switch otherwise. That some $\tau^{(\mathbf{R})}$ is an optimal strategy will follow by setting $r_1 = j_1$ and $r_i = j_i - j_{i-1}$ for $i > 1$ (with $\infty - \infty$ defined arbitrarily) if we can show $j_i \geqslant j_{i-1}$.

Suppose to the contrary that $j_i < j_{i-1}$ for some $i > 1$. Then, with $\mathbf{A}^{(m)}$ denoting the discount sequence $(\alpha_{m+1}, \alpha_{m+2}, \ldots)$, arm 2 is optimal initially for the $(\sigma^{j_i}\varphi^{i-1}F, \lambda; \mathbf{A}^{(j_i+i-1)})$-bandit and arm 1 is optimal initially for the $(\sigma^{j_i}\varphi^{i}F, \lambda; \mathbf{A}^{(j_i+i)})$-bandit. Hence,

$$\Lambda(\varphi\sigma^{j_i}\varphi^{i-1}F, \mathbf{A}^{(j_i+i)}) \geqslant \Lambda(\sigma^{j_i}\varphi^{i-1}F, \mathbf{A}^{(j_i+i-1)}) \qquad (5.3.8)$$

(assuming both sides are meaningful), which contradicts Theorem 5.3.6 provided $\sigma^{j_i}\varphi^{i-1}F$ is not a one-point distribution. If it is a one-point distribution, then without loss we can redefine $j_i = \infty$ if $j_{i-1} > 0$, and $j_i = 0$ if $j_{i-1} = 0$. Suppose $\alpha_{j_i+i} = 0$, so the left-hand side of (5.3.8) is not defined. Then, since $\mathbf{A}$ is regular and $\alpha_1 \neq 0$, it follows that $\alpha_m = 0$ for $m > j_i + i$. Hence, we can redefine $j_i = \infty$ without losing optimality. All distributions $\sigma^{j_i}\varphi^{i}F$ in (5.3.8) are meaningful unless $F(\{1, 0\}) = 1$. But in this case $\tau^{(\mathbf{R})}$ with $\mathbf{R} = (\infty, \infty, \ldots)$ is optimal.

For the second part of the theorem suppose $\lambda = \Lambda(F, \mathbf{A})$. Suppose arm 1 is selected initially and yields a failure. Then by Theorem 5.3.6, arm 2 must be selected at stage 2 provided F is not a one-point distribution. Thus, r_1 must be positive for $\tau^{(\mathbf{R})}$ to be optimal in this case. If F is a one-point distribution, then $F = \delta_\lambda$ since $\lambda = \Lambda(F, \mathbf{A})$ and so all strategies are optimal; in particular, any $\tau^{(\mathbf{R})}$ with $r_1 > 0$ is optimal. □

By way of illustration, suppose the discount sequence $\mathbf{A}$ is regular with $\alpha_1 > 0$ and $\alpha_m = 0$ for $m \geqslant 6$. To find Λ via Corollary 5.3.2 we can apply Theorem 5.3.7 to see that only eight essentially different

sequences **R** need be considered:

$$(4, \ldots), \qquad (1, 2, \ldots),$$
$$(3, \ldots) \qquad (1, 1, \ldots),$$
$$(2, 0, \ldots), \qquad (1, 0, 1, \ldots),$$
$$(2, 1, \ldots), \qquad (1, 0, 0, \ldots).$$

The parts of these sequences indicated with dots are of no consequence in any calculation. The list is complete in the sense that no other sequence corresponds to a strategy essentially different from those that correspond to listed sequences. For instance, $(5, \ldots)$ is essentially the same as the listed sequence $(4, \ldots)$: both require a switch to arm 2 if a failure occurs at any of the first four stages, and subsequent selections are immaterial if the first failure occurs after stage 4.

We now have two ways of finding an optimal strategy in a particular problem. One is to use dynamic programming as discussed in Chapter 2, where an approximation is required if the horizon is infinite. Another is to use Corollary 5.3.2, with Theorem 5.3.7 eliminating the need to consider many strategies.

An advantage of the latter approach is that, for the pair $(F, \mathbf{A})$, it finds Λ and therefore indicates an optimal selection for any $\lambda \in (0, 1)$. A disadvantage is that it gives only the optimal initial selections. When arm 1 is optimal initially, deciding on a second selection requires finding $\Lambda(\sigma F, \mathbf{A}^{(1)})$ or $\Lambda(\varphi F, \mathbf{A}^{(1)})$, which entails additional calculation.

On the other hand, if $\Lambda(F, \mathbf{A})$ is desired, and not just an optimal strategy, then dynamic programming is handicapped; for successive approximations are necessary and a separate dynamic program must be carried out at each step. One way to proceed is to guess a value for Λ, decide which arm is optimal, adjusting the guess accordingly (upwards if arm 1 is optimal and downwards if it is not). Repeating this process until both arms are optimal gives Λ. Still, such a computer program is easy to write, and we found ours to be more efficient than our program that finds Λ using Corollary 5.3.2 and Theorem 5.3.7.

Calculations of Λ are given in the next section for various examples. These are compared with easy-to-compute lower and upper bounds.

5.4 Bernoulli examples and bounds – regular discounting

In view of Theorem 5.3.1, when $\mathbf{A}$ is regular the optimal initial selection is specified by $\Lambda(F, \mathbf{A})$; namely, arm 1 or arm 2 according as $\lambda \leqslant$ or $\geqslant \Lambda(F, \mathbf{A})$. In view of Theorem 5.2.2, if $\lambda \geqslant \Lambda(F, \mathbf{A})$ then, assuming arm 2 is selected initially, at least one optimal continuation is clear: select arm 2 indefinitely. When $\lambda < \Lambda(F, \mathbf{A})$ then optimal continuations can be found from $\Lambda(\sigma^s \varphi^f F, \mathbf{A}^{(s+f)})$ for nonnegative integers s and f. In any case, optimal strategies are specified by the various Λ's.

As indicated in the previous section, calculating $\Lambda(F, \mathbf{A})$ can require extensive numerical calculations. However, for certain combinations of F and $\mathbf{A}$ it is possible to give an explicit formula for $\Lambda(F, \mathbf{A})$. Our main objective in this section is to give bounds for Λ. These can be used to approximate Λ and will be most useful when Λ is not available. But we begin with an example in which explicit calculation of Λ is possible. The discount sequence in this example is geometric. So in view of the forthcoming Theorem 6.1.1, the calculation of Λ has special significance. This theorem says that when there are k independent arms, so that $G = F_1 \times \ldots \times F_k$, arm i is optimal initially if and only if

$$\Lambda(F_i, \mathbf{A}) = \max_{1 \leqslant j \leqslant k} \Lambda(F_j, \mathbf{A}).$$

Examples 6.1.1 and 6.1.2 apply Theorem 6.1.1 by exploiting the calculations given in the next example.

Example 5.4.1 Suppose $\mathbf{A} = (1, \alpha, \alpha^2, \ldots)$ and

$$F = p\delta_a + (1-p)\delta_b$$

where $0 < b < a < 1$. Berry and Fristedt (1979, Example 4.3) show that

$$\Lambda(F, \mathbf{A}) = \frac{b(1-p)g(\alpha, b, c) + apg(\alpha, a, c)}{(1-p)g(\alpha, b, c) + pg(\alpha, a, c)},$$

where

$$c = \frac{\log[(1-b)/(1-a)]}{\log[(1-b)/(1-a)] + \log[a/b]},$$

$$g(\alpha, u, c) = \exp\left[-\sum_{m=1}^{\infty} \frac{\alpha^m}{m} P(S_m < 0) \right],$$

and $(S_1, S_2, \ldots)$ is a random walk starting at 0 with individual steps of

$$\begin{aligned} 1-c &\text{ with probability } u \\ -c &\text{ with probability } 1-u. \end{aligned}$$

As a special case, when $a+b=1$ then $c=1/2$ and

$$g(\alpha, u, 1/2) = 1 - \frac{1-[1-4\alpha^2 u(1-u)]^{1/2}}{2\alpha u}$$

After considerable algebra,

$$\Lambda(F, \mathbf{A}) = \frac{[ap+b(1-p)][1+(1-4\alpha^2 ab)^{1/2}] - 2\alpha ab}{1+(1-4\alpha^2 ab)^{1/2} - 2\alpha[ap+b(1-p)]}.$$

Specializing further by taking $p = \frac{1}{2}$, we obtain

$$\Lambda(F, \mathbf{A}) = \frac{[1+(1-4\alpha^2 ab)^{1/2}] - 4\alpha ab}{2[1+(1-4\alpha^2 ab)^{1/2}] - 2\alpha}.$$

Corresponding to the intuitive notion that experimenting with arm 1 has more potential benefit for larger α, this is an increasing function of α that equals $1/2 = E(\theta_1 | F)$ at $\alpha = 0$ and approaches a as $\alpha \uparrow 1$.

□

The lower and upper bounds we derive in this section will be compared with each other and with the exact Λ in special cases. Two such cases are provided by the next two examples. Explicit formulas for Λ are not possible in these examples; of the two methods for finding Λ that were compared at the end of the previous section, we used dynamic programming since it uses computer time more efficiently.

Example 5.4.2 Suppose F has a beta density: for $a, b > 0$,

$$dF(u) \propto u^{a-1}(1-u)^{b-1}\,du. \tag{5.4.1}$$

Take $\mathbf{A}$ to be the n-horizon uniform: $\alpha_1 = \ldots = \alpha_n = 1, \alpha_{n+1} = \ldots = 0$. It is clear that $\Lambda(F, \mathbf{A}) = E(\theta_1 | F) = a/(a+b)$ if $n = 1$. From Theorem 5.4.1 it will follow that $\Lambda(F, \mathbf{A}) \to 1$ as $n \to \infty$. Figure 5.1 shows $\Lambda(F, \mathbf{A})$ up to $n = 500$ for $(a, b) = (1, 5)$ and $(1, 1)$; the latter corresponds to the uniform distribution on $(0, 1)$. □

Example 5.4.3 We return to the discounting of Example 5.4.1: $\mathbf{A} = (1, \alpha, \alpha^2, \ldots)$. As in Example 5.4.2, θ_1 has a distribution in the

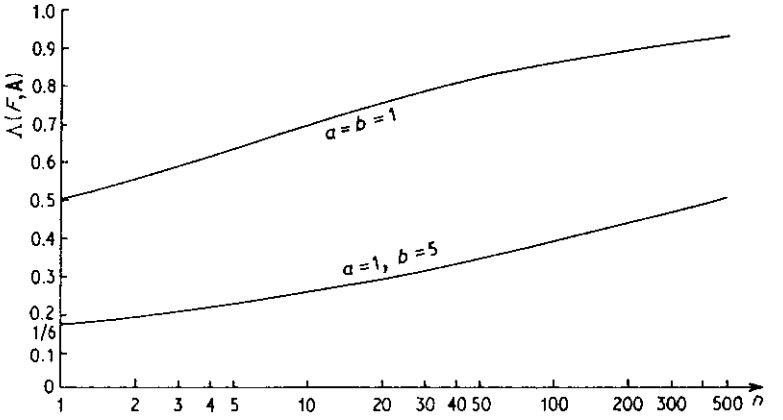


Fig. 5.1 *The break-even value of the known arm as a function of the horizon n for two beta distributions.*

beta family (5.4.1). For reasons similar to those given in that example for the finite-horizon case, $\Lambda(F, \mathbf{A}) = E(\theta_1 | F) = a/(a+b)$ when $\alpha = 0$ and $\Lambda(F, \mathbf{A}) \to 1$ as $\alpha \to 1$. Figure 5.2 shows $\Lambda(F, \mathbf{A})$ for $\alpha \in (0, 1)$

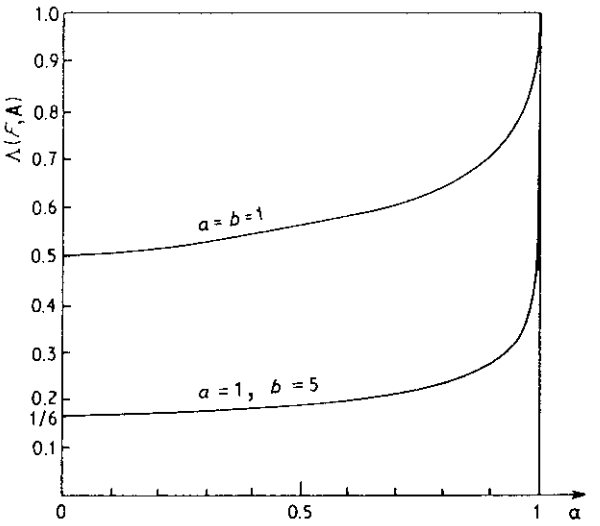


Fig. 5.2 *The break-even value of the known arm as a function of the discount factor α for two beta distributions.*

for the same distributions F considered in Example 5.4.2. In Figure 5.3, $\alpha = 0.9$ and various contours of $\Lambda(F, \mathbf{A})$ are shown in the (a, b)-plane. For example, the figure indicates that $\Lambda(F, \mathbf{A})$ is about 0.7 when F is the beta distribution given by $a = 5$, $b = 3$. □

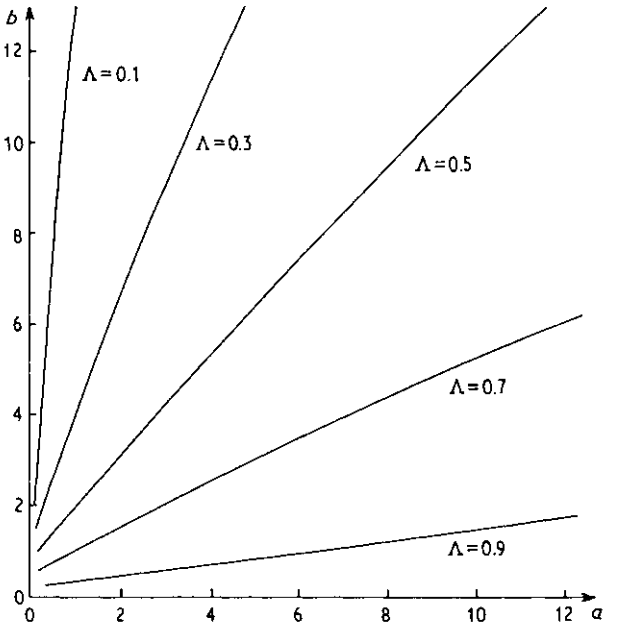


Fig. 5.3 *Contours of* $\Lambda(F, \mathbf{A})$ *when* $\alpha = 0.9$ *as a function of beta parameters a and b.*

Lower bounds for Λ are rather easy to establish by using any particular strategy $\tau^{(\mathbf{R})}$, as defined just before Theorem 5.3.8, instead of maximizing on the right-hand side of (5.3.4). The following result establishes a family of lower bounds by considering the class of $\tau^{(\mathbf{R})}$'s with $\mathbf{R} = (r, 0, 0, \ldots)$. Recall that $\tau^{(r, 0, 0, \ldots)}$ indicates arm 1 until it fails; if this occurs before the rth success (i.e., by the rth stage) then arm 2 is used indefinitely thereafter, and if it occurs later then arm 1 is continued indefinitely.

Theorem 5.4.1 Suppose that $\mathbf{A}$ is a regular discount sequence with $\alpha_1 > 0$. Then for each $r \in \{1, 2, \ldots, \infty\}$, $\Lambda(F, \mathbf{A}) \geqslant \Lambda_r(F, \mathbf{A})$, where

$$\Lambda_r(F, \mathbf{A}) = \frac{\sum_{m=1}^{\infty} \alpha_m E(\theta_1^{m \wedge (r+1)} | F)}{\sum_{m=1}^{\infty} \alpha_m E(\theta_1^{(m-1) \wedge r} | F)}. \tag{5.4.2}$$

Proof That $\Lambda \geqslant \Lambda_r$ follows by evaluating the ratio in (5.3.4) at $\tau^{(\mathbf{R})}$ where $\mathbf{R} = (r, 0, 0, \ldots)$, instead of maximizing over $\tau(\varnothing) = 1$. □

The bounds Λ_r given in Theorem 5.4.1 do not usually equal Λ. The next result lists the circumstances under which they are exact in the infinite horizon case. The result is given without proof.

Theorem 5.4.2 In addition to the assumptions of Theorem 5.4.1, assume that the horizon of $\mathbf{A}$ is infinite. Then $\Lambda(F, \mathbf{A}) = \Lambda_r(F, \mathbf{A})$ if and only if either (i) F is supported by $\{0, \Lambda_r(F, \mathbf{A}), 1\}$, or (ii) $r = 1$ and F is supported by $\{0\} \cup [\Lambda_1(F, \mathbf{A}), 1]$, or (iii) $r = \infty$ and F is supported by $[0, \Lambda(F, \mathbf{A})] \cup \{1\}$.

Let

$$\Lambda_*(F, \mathbf{A}) = \sup \{\Lambda_r(F, \mathbf{A}): r = 1, 2, \ldots, \infty\}.$$

The next result says that this supremum is attained and the sequence $(\Lambda_1, \Lambda_2, \ldots, \Lambda_\infty)$ is quite well-behaved.

Proposition 5.4.3 For any F and regular $\mathbf{A}$ with $\alpha_1 > 0$,

$$\{\Lambda_r(F, \mathbf{A}): r = 1, 2, \ldots, \infty\}$$

is unimodal and $\Lambda_\infty(F, \mathbf{A}) = \lim_{r \to \infty} \Lambda_r(F, \mathbf{A})$.

Proof Unimodality will follow from

$$\Lambda_{r-1} - \Lambda_r \geqslant 0 \Rightarrow \Lambda_r - \Lambda_{r+1} \geqslant 0,$$

or equivalently,

$$\begin{aligned} &E\left((\theta_1 - \Lambda_r) \sum_{m=1}^{\infty} \alpha_m \theta_1^{(m-1) \wedge (r-1)} \middle| F\right) \geqslant 0 \\ &\Rightarrow E\left((\Lambda_r - \theta_1) \sum_{m=1}^{\infty} \alpha_m \theta_1^{(m-1) \wedge (r+1)} \middle| F\right) \geqslant 0. \end{aligned} \tag{5.4.3}$$

In view of (5.4.2),

$$E\left((\theta_1 - \Lambda_r)\sum_{m=1}^{\infty}\alpha_m\theta_1^{(m-1)\wedge r}\,|F\right) = 0. \tag{5.4.4}$$

Subtracting (5.4.4) from the first expression in (5.4.3) and adding it to the second reduces the problem to showing

$$\left(\sum_{m=r+1}^{\infty}\alpha_m\right)E\left((\theta_1 - \Lambda_r)\theta_1^{r-1}(1-\theta_1)\,\middle|F\right) \geqslant 0$$

$$\Rightarrow\left(\sum_{m=r+2}^{\infty}\alpha_m\right)E\left((\theta_1 - \Lambda_r)\theta_1^{r}(1-\theta_1)\,\middle|F\right) \geqslant 0.$$

If $\sum_{m=r+2}^{\infty}\alpha_m = 0$ then the implication holds trivially. So we need only show that

$$E\left((\theta_1 - \Lambda_r)\theta_1^{r-1}(1-\theta_1)\,\middle|F\right) \geqslant 0 \tag{5.4.5}$$

implies

$$E\left((\theta_1 - \Lambda_r)\theta_1^{r}(1-\theta_1)\,\middle|F\right) \geqslant 0. \tag{5.4.6}$$

Subtracting Λ_r times the left-hand side of (5.4.5) from (5.4.6) gives

$$E\left((\theta_1 - \Lambda_r)^2\theta_1^{r-1}(1-\theta_1)\,\middle|F\right),$$

which is obviously nonnegative. Therefore $(\Lambda_1, \Lambda_2, \ldots, \Lambda_\infty)$ is unimodal for all $(F, \mathbf{A})$. The continuity of $\Lambda_r(F, \mathbf{A})$ at $r = \infty$ is immediate. □

In case the horizon n is finite, Λ_* can be found with a finite number of calculations since $\Lambda_r = \Lambda_n$ for $r \geqslant n$. The next result indicates that Λ_* can also be found with a finite number of calculations when n is infinite. The result gives an easy-to-check condition for the sequence $(\Lambda_1, \ldots, \Lambda_\infty)$ to be nondecreasing. The proof is not trivial but is omitted since the result is not important in the sequel.

Proposition 5.4.4 Assume $\mathbf{A}$ is regular with $\alpha_1 > 0$. When the horizon of $\mathbf{A}$ is infinite, $\Lambda_*(F, \mathbf{A}) = \Lambda_\infty(F, \mathbf{A})$ if and only if F is supported by $[0, \Lambda_\infty(F, \mathbf{A})] \cup \{1\}$.

Let r_* denote the smallest r for which (5.4.2) gives the best lower bound:

$$r_* = \min\{r: \Lambda_r(F, \mathbf{A}) = \Lambda_*(F, \mathbf{A})\}.$$

So if a decision maker were required to select arm 1 until it fails once and then commit permanently to one or the other arm, it would be optimal to commit to arm 1 if at least r_* successes had been observed and to arm 2 otherwise.

We now use Theorem 5.4.2 in case F is supported by $\{1, 0\}$. This easy example (cf. Example 1.2.1) is given as a setting in which a simple expression for $\Lambda(F, \mathbf{A})$ is possible, and is given by the bounds in Theorem 5.4.1. Theorem 5.4.2 provides generalizations of this example in two directions when the horizon is infinite.

Example 5.4.4 Suppose $\mathbf{A}$ is an arbitrary regular discount sequence with $\alpha_1 > 0$ and

$$F = p\delta_1 + (1-p)\delta_0.$$

Then every bound in (5.4.2) is exact. So

$$\Lambda(F, \mathbf{A}) = \frac{p\sum_{m=1}^{\infty}\alpha_m}{p\sum_{m=1}^{\infty}\alpha_m + (1-p)\alpha_1} = \frac{p\gamma_1}{p\gamma_2 + \alpha_1},$$

where, as defined earlier, $\gamma_2 = \sum_{m=2}^{\infty}\alpha_m$. □

The next two examples illustrate Theorem 5.4.1 with the discount sequences considered in Examples 5.4.2 and 5.4.3 and with θ_1 uniform on $(0, 1)$.

Example 5.4.5 Consider n-horizon uniform discounting: $\mathbf{A} = (1, 1, \ldots, 1, 0, \ldots)$. Then $\Lambda(F, \mathbf{A})$ is at least

$$\Lambda_r(F, \mathbf{A}) = \frac{\sum_{m=1}^{n} E(\theta_1^{m\wedge(r+1)}|F)}{\sum_{m=1}^{n} E(\theta_1^{(m-1)\wedge r}|F)}$$

for $r = 1, 2, \ldots, \infty$; these bounds were obtained by Bradt, Johnson, and Karlin (1956).

For $F = U(0, 1)$, $\Lambda(F, \mathbf{A})$ is shown in Figure 5.1 (where it is given by the curve labelled $a = b = 1$). For $r \leqslant n$,

$$\Lambda_r(F, \mathbf{A}) = \frac{\sum_{m=1}^{r} \frac{1}{m+1} + \frac{n-r}{r+2}}{\sum_{m=1}^{r} \frac{1}{m} + \frac{n-r}{r+1}},$$

and $\Lambda_{n-1} = \Lambda_n = \ldots = \Lambda_\infty$. It can be shown that $r_* \sim \sqrt{n}$ as $n \to \infty$.

Table 5.1 compares Λ with Λ_* for various values of n, and gives r_*. The values of Λ were calculated to four-decimal accuracy using the BASIC program listed below. It involves dynamic programming and iterates on λ until both arms are optimal. The example output for $n = 100$ determines Λ to be 0.8685. The nine iterations required a total of six minutes on a Commodore 64 personal computer and only 15 memory locations were used for V. Because $\Lambda(F, \mathbf{A})$ increases with n, the locations and time required increase substantially less rapidly than n and n^2, respectively. Still, computation time can be prohibitive

Table 5.1 *Lower and upper bounds for* $\Lambda(F, \mathbf{A})$ *where* $F = U(0, 1)$ *and* $\mathbf{A}$ *is the n-horizon uniform.*

n	r_*	$\Lambda_*(F, \mathbf{A})$	$\Lambda(F, \mathbf{A})$	$\Lambda^*(F, \mathbf{A})$
1	1	0.5000	0.5000	0.5000
2	1	0.5556	0.5556	0.5556
5	3	0.6357	0.6357	0.6490
10	4	0.6954	0.6981	0.7189
20	6	0.7512	0.7576	0.7806
50	8	0.8164	0.8262	0.8466
100	12	0.8575	0.8685	0.8850
200	16	0.8914	0.9024	0.9148
500	24	0.9258	0.9351	0.9436
1 000	34	0.9452	0.9536	0.9590
10 000	103	0.9811	0.9850	0.9863
100 000	321	0.9938	0.9953	0.9953
$\to \infty$	$\sim \sqrt{n}$	1.0000	1.0000	1.0000

for large n. Such limitations can be substantially alleviated using a sequence of truncations.

```
 10 DIM V(15): L = .87: N = 100: H = .0064
 20 M = N: M1 = INT ((1 - L)*N) + 1
 30 FOR I = 0 TO M1: V(I) = 0: NEXT I
 40 IF M = 0 THEN GOTO 110
 50 M = M - 1
 60 FOR I = 0 TO M1: R = M - I: P = (R + 1)/(M + 2)
 70 V1 = P*(1 + V(I)) + (1 - P)*V(I + 1): V2 = (N - M)*L
 80 IF V1 > V2 THEN V(I) = V1: GOTO 100
 90 V(I) = V2: GOTO 40
100 NEXT I
110 D = V(0) - N*L
120 PRINT "L = "L, "D = "D
130 IF H < .00004 THEN END
140 IF D = 0 THEN L = L - H: H = H/2: GOTO 20
150 L = L + H: H = H/2: GOTO 20
```

EXAMPLE OUTPUT

```
L =  .87          D = 0
L =  .8636        D =  .0587253608
L =  .8668        D =  .0198577987
L =  .8684        D = 6.55137934E - 04
L =  .8692        D = 0
L =  .8688        D = 0
L =  .8686        D = 0
L =  .8685        D = 0
L =  .86845       D = 7.6983124E - 05
```

Table 5.1 also gives an upper bound $\Lambda^*(F, \mathbf{A})$ to be discussed shortly. □

Example 5.4.6 Suppose $\mathbf{A}$ is geometric: $\mathbf{A} = (1, \alpha, \alpha^2, \ldots)$. Then $\Lambda(F, \mathbf{A})$ is at least

$$\Lambda_r(F, \mathbf{A}) = \frac{E[\theta_1(1-\alpha^r\theta_1^r)(1-\alpha\theta_1)^{-1} + \alpha^r\theta_1^{r+1}(1-\alpha)^{-1} \mid F)}{E[(1-\alpha^r\theta_1^r)(1-\alpha\theta_1)^{-1} + \alpha^r\theta_1^r(1-\alpha)^{-1} \mid F]}$$

for $r = 1, 2, \ldots, \infty$. While not generally the best lower bound, that

given by $r = \infty$ simplifies:

$$\Lambda_\infty(F, \mathbf{A}) = \frac{\psi(\alpha) - 1}{\alpha\psi(\alpha)},$$

where the generating function ψ is given by

$$\psi(\alpha) = E[(1 - \alpha\theta_1)^{-1} | F].$$

Suppose $F = U(0, 1)$, then $\Lambda(F, \mathbf{A})$ is shown in Figure 5.2 (by the curve labelled $a = b = 1$). For $r = 1, 2, \ldots, \infty$,

$$\Lambda_r(F, \mathbf{A}) = \frac{\sum_{m=1}^{r} \frac{\alpha^{m-1}}{m+1} + \frac{\alpha^r}{(r+2)(1-\alpha)}}{\sum_{m=1}^{r} \frac{\alpha^{m-1}}{m} + \frac{\alpha^r}{(r+1)(1-\alpha)}}.$$

Using

$$\psi(\alpha) = -\log(1-\alpha)/\alpha,$$

the limit of this sequence is

$$\Lambda_\infty(F, \mathbf{A}) = \frac{1}{\alpha} + \frac{1}{\log(1-\alpha)}.$$

As a corollary of the corresponding fact with $n \to \infty$ in Example 5.4.5, $r_* \sim 1/\sqrt{(1-\alpha)}$ as $\alpha \to 1$. Table 5.2 compares Λ with Λ_* and Λ_∞ for various values of α. To calculate Λ we used a modification of the program given in the previous example and the truncations discussed in Section 2.6. Table 5.2 also gives an upper bound to be discussed shortly. Note the similarity between Tables 5.1 and 5.2 with n corresponding to $1/(1-\alpha)$. □

To obtain a lower bound, one need only specify a strategy and evaluate the ratio in (5.3.4) for that strategy. An upper bound requires another method. One that applies to decision problems generally is to assume that any unknowns will be revealed at some future time, and that the decision maker behaves optimally with this new source of information taken into consideration. Such a program is carried out next. The result is not as readily applied as the lower bound given by Theorem 5.4.1 since the solution of the forthcoming equation (5.4.7) will require iteration.

Table 5.2 *Lower and upper bounds for* $\Lambda(F, \mathbf{A})$ *where* $F = U(0, 1)$ *and* $\mathbf{A}$ *is geometric with parameter* α.

α	r_*	$\Lambda_\infty(F, \mathbf{A})$	$\Lambda_*(F, \mathbf{A})$	$\Lambda(F, \mathbf{A})$	$\Lambda^*(F, \mathbf{A})$
0	1	0.5000	0.5000	0.5000	0.5000
0.1	2	0.5088	0.5088	0.5088	0.5110
0.2	2	0.5186	0.5187	0.5187	0.5234
0.3	2	0.5297	0.5299	0.5300	0.5374
0.4	2	0.5424	0.5431	0.5432	0.5536
0.5	2	0.5573	0.5588	0.5590	0.5728
0.6	2	0.5753	0.5781	0.5788	0.5962
0.7	3	0.5980	0.6028	0.6046	0.6260
0.8	3	0.6286	0.6385	0.6413	0.6667
0.9	4	0.6768	0.6974	0.7029	0.7317
0.95	6	0.7188	0.7525	0.7614	0.7888
0.97	7	0.7457	0.7900	0.8004	0.8252
0.99	12	0.7930	0.8579	0.8699	0.8874
0.999	34	0.8562	0.9452	0.9538	0.9593
0.9999	103	0.8915	0.9811	0.9851	0.9864
0.99999	321	0.9132	0.9938	0.9953	0.9953
1.0	$\sim(1-\alpha)^{-1/2}$	1.0000	1.0000	1.0000	1.0000

Theorem 5.4.5 Suppose that $\mathbf{A}$ is a regular discount sequence with $\alpha_1 > 0$. Then $\Lambda(F, \mathbf{A})$ is not greater than the unique solution in $[0, 1]$, say $\lambda = \Lambda^*(F, \mathbf{A})$, of the equation:

$$\lambda[\alpha_1 + \gamma_2 E(\theta_1 | F)] - [\alpha_1 E(\theta_1 | F) + \gamma_2 E(\theta_1^2 | F)] - \sum_{m=3}^{\infty} \alpha_m E[(\lambda - \theta_1)^+ \theta_1 (1 - \theta_1^{m-2}) | F] = 0, \tag{5.4.7}$$

where $\gamma_2 = \sum_{m=2}^{\infty} \alpha_m$.

Proof The left-hand side of (5.4.7) is nonpositive when $\lambda = 0$, nonnegative when $\lambda = 1$, and strictly increasing for $\lambda \in [0, 1]$. So we need only show that it is nonpositive when $\lambda = \Lambda(F, \mathbf{A})$.

For calculational convenience we temporarily assume that $F(\{1\}) = 0$.

Both arms are optimal initially in the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit in view

of Theorem 5.3.1. First consider an optimal strategy, say τ, that starts with arm 1. In view of Theorem 5.3.6 we can let τ indicate arm 2 indefinitely if arm 1 fails on the initial selection: $\tau(0) = 2$ and $\tau(0, z_2, \ldots, z_m) = 2$ for all $m \geqslant 2$ and all $z_2, \ldots, z_m$. And in view of Theorem 5.3.5 we can take $\tau(1) = 1$ and, more generally, $\tau(1, 1, \ldots, 1) = 1$.

We have no general results which indicate $\tau(1, 1, \ldots, 1, 0)$, the arm to select after a failure (on arm 1) that occurs later than the first stage. But we know that the value of the $(F, \lambda; \mathbf{A})$-bandit is no greater than the corresponding 'value' should θ_1 become known in such a circumstance. So, where $\lambda = \Lambda(F, \mathbf{A})$,

$$V(F, \lambda; \mathbf{A}) = E_\tau\Bigg[(1-\theta_1)\lambda\gamma_2 + \sum_{n=1}^{\infty} \theta_1^n(1-\theta_1)\Bigg(\sum_{m=1}^{n} \alpha_m + \sum_{m=n+2}^{\infty} \alpha_m Z_m\Bigg)\Bigg| F\Bigg]$$

$$\leqslant E\Bigg[(1-\theta_1)\lambda\gamma_2 + \sum_{n=1}^{\infty} \theta_1^n(1-\theta_1)\Bigg(\sum_{m=1}^{n} \alpha_m + \sum_{m=n+2}^{\infty} \alpha_m(\theta_1 \vee \lambda)\Bigg)\Bigg| F\Bigg]. \quad (5.4.8)$$

Simple calculations give

$$\sum_{n=1}^{\infty} \theta_1^n(1-\theta_1) \sum_{m=1}^{n} \alpha_m = \sum_{m=1}^{\infty} \alpha_m \sum_{n=m}^{\infty} \theta_1^n(1-\theta_1) = \sum_{m=1}^{\infty} \alpha_m \theta_1^m$$

and

$$\sum_{n=1}^{\infty} \theta_1^n(1-\theta_1)(\theta_1 \vee \lambda) \sum_{m=n+2}^{\infty} \alpha_m = \sum_{m=3}^{\infty} \alpha_m \sum_{n=1}^{m-2} \theta_1^n(1-\theta_1)(\theta_1 \vee \lambda)$$

$$= \sum_{m=3}^{\infty} \alpha_m(\theta_1 - \theta_1^{m-1})(\theta_1 \vee \lambda)$$

$$= \sum_{m=3}^{\infty} \alpha_m(\lambda-\theta_1)^+\theta_1(1-\theta_1^{m-2}) + \sum_{m=3}^{\infty} \alpha_m\theta_1^2(1-\theta_1^{m-2})$$

$$= \sum_{m=3}^{\infty} \alpha_m(\lambda-\theta_1)^+\theta_1(1-\theta_1^{m-2}) + \gamma_3\theta_1^2 - \sum_{m=3}^{\infty} \alpha_m\theta_1^m.$$

So rewriting (5.4.8),

$$V(F, \lambda; \mathbf{A}) \leqslant E\Bigg[(1-\theta_1)\lambda\gamma_2 + \alpha_1\theta_1 + \gamma_2\theta_1^2 + \sum_{m=3}^{\infty} \alpha_m(\lambda-\theta_1)^+\theta_1(1-\theta_1^{m-2}) | F \Bigg]. \quad (5.4.9)$$

Since the expression in square brackets in (5.4.9) is appropriate as well when $\theta_1 = 1$, we can lift the restriction $F(\{1\}) = 0$.

Now suppose arm 2 is selected in the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit. Theorem 5.2.2 applies to show that arm 2 is optimal indefinitely and

$$V(F, \lambda; \mathbf{A}) = \lambda\gamma_1, \quad (5.4.10)$$

where $\lambda = \Lambda(F, \mathbf{A})$. Subtracting the right-hand side of (5.4.9) from the right-hand side of (5.4.10) gives the left-hand side of (5.4.7) and shows that it is nonpositive when $\lambda = \Lambda(F, \mathbf{A})$, as desired. □

The bound given implicitly in (5.4.7) can be difficult to evaluate. A cruder but more easily computable bound is provided next. Its proof is immediate upon replacing $(\lambda-\theta_1)^+$ with the larger quantity $\lambda(1-\theta_1)$ in (5.4.7) and solving.

Corollary 5.4.6 Under the assumptions of Theorem 5.4.5,

$$\Lambda(F, \mathbf{A}) \leqslant \frac{\alpha_1 E(\theta_1 | F) + \gamma_2 E(\theta_1^2 | F)}{\alpha_1 + \gamma_2 E(\theta_1 | F) - \sum_{m=3}^{\infty} \alpha_m E[\theta_1(1-\theta_1)(1-\theta_1^{m-2}) | F]}.$$

Example 5.4.7 Suppose $\mathbf{A}$ is regular with $\alpha_1 > 0$ and the support of F is contained in $[0, \Lambda(F, \mathbf{A})] \cup \{1\}$. Theorem 5.4.2 applies to show that $\Lambda(F, \mathbf{A}) = \Lambda_\infty(F, \mathbf{A})$ which is defined at (5.4.2); cf. Example 5.4.4. Since a single failure on arm 1 reveals the better arm, the proof of Theorem 5.4.5 makes it clear that $\Lambda(F, \mathbf{A})$ also coincides with upper bound $\Lambda^*(F, \mathbf{A})$. □

Example 5.4.8 As in Example 5.4.5, assume the discount sequence is the n-horizon uniform. Then (5.4.7) becomes

$$\lambda[1 + (n-1)E(\theta_1 | F)] - [E(\theta_1 | F) + (n-1)E(\theta_1^2 | F)] - E[(\lambda-\theta_1)^+\theta_1[n-2-(n-1)\theta_1 + \theta_1^{n-1}](1-\theta_1)^{-1} | F] = 0.$$

Specializing to $F = U(0, 1)$:

$$\lambda(n+1)/2 - (2n+1)/6$$
$$-\int_0^\lambda (\lambda - u)u[n-2-(n-1)u+u^{n-1}](1-u)^{-1}\,du = 0.$$

The solution $\lambda = \Lambda^*(F, \mathbf{A})$ of this equation is the upper bound given in Table 5.1 for various values of n. □

Example 5.4.9 As in Example 5.4.6, assume the discount sequence is geometric. Then (5.4.7) becomes (after multiplying by $(1-\alpha)$),

$$\lambda[1-\alpha+\alpha E(\theta_1|F)] - [(1-\alpha)E(\theta_1|F) + \alpha E(\theta_1^2|F)]$$
$$-\alpha^2 E[(\lambda-\theta_1)^+\theta_1(1-\theta_1)(1-\alpha\theta_1)^{-1}|F] = 0.$$

And when $F = U(0, 1)$ it becomes

$$\lambda[1-\alpha/2] - [1/2-\alpha/6] - \alpha^2\int_0^\lambda (\lambda-u)\,u(1-u)(1-\alpha u)^{-1}\,du = 0.$$

The solution $\lambda = \Lambda^*(F, \mathbf{A})$ of this equation is the upper bound given in Table 5.2. □

In the next section we leave the Bernoulli setting that has occupied most of our attention so far in this chapter and throughout Chapter 4.

5.5 The non-Bernoulli case with regular discounting

In Sections 5.1 to 5.4, arm 1 has been assumed to produce only 0's and 1's. In this section we take the point of view that Q_1 has arbitrary form; the distribution measure F reflects the information available initially concerning Q_1. In the next section we give an example in which the support of F can be quite large.

We assume regular discounting and continue to assume that arm 2 is known to produce λ at every stage, now with $\lambda \in \mathbb{R}$. By subtracting the known constant produced by arm 2 from every distribution in the support of F, we could always assume $\lambda = 0$. This is especially convenient when discussing stopping problems where the alternative to making an observation from Q_1 is no observation. But we allow λ to be arbitrary since, as in the Bernoulli case, we want to discuss the value of λ, as a function of F and discount sequence $\mathbf{A}$, such that arm 1 and arm 2 are both optimal.

The advantage of arm 1 over arm 2 is easily adaptable to general distributions: for all F, λ, and $\mathbf{A}$,

$$\Delta(F, \lambda; \mathbf{A}) = V^{(1)}(F, \lambda; \mathbf{A}) - V^{(2)}(F, \lambda; \mathbf{A}), \tag{5.5.1}$$

where, as defined in Theorem 2.5.1,

$$V^{(1)}(F, \lambda; \mathbf{A}) = E[\alpha_1 X_{11} + V((X_{11})F, \lambda; \mathbf{A}^{(1)})|F], \tag{5.5.2}$$

$$V^{(2)}(F, \lambda; \mathbf{A}) = \alpha_1 \lambda + V(F, \lambda; \mathbf{A}^{(1)}), \tag{5.5.3}$$

and $V = V^{(1)} \vee V^{(2)}$.

In the Bernoulli case the decision maker has two reasons to like a success at a particular stage: it increases immediate income and, because of exchangeability, it enhances future expectations. We have seen (Theorem 5.3.5) that this preference for success can translate into a preference for the arm that was successful. But when more than two outcomes are possible it is easy to construct examples in which a large observation can be quite distasteful both in terms of value and desirability of the arm that produced it.

Example 5.5.1 Let $\mathbf{A}$ be the 2-horizon uniform and take $\lambda = 0$. Suppose

$$F = \tfrac{1}{2}\delta_Q + \tfrac{1}{2}\delta_{Q^*},$$

where

$$Q = p\delta_6 + (1-p)\delta_{-4},$$

$$Q^* = \delta_4.$$

Using (5.5.2) gives

$$V^{(1)}(F, 0; \mathbf{A}) = 5p + \tfrac{1}{2}V((4)F, 0; \mathbf{A}^{(1)}) + \tfrac{1}{2}V((6)F, 0; \mathbf{A}^{(1)}),$$

since

$$V((-4)F, 0; \mathbf{A}^{(1)}) = V((6)F, 0; \mathbf{A}^{(1)}).$$

Continuing, using $\mathbf{A}^{(1)} = (1, 0, 0, \ldots)$,

$$V^{(1)}(F, 0; \mathbf{A}) = 5p + \tfrac{1}{2}[4 \vee 0] + \tfrac{1}{2}[(10p - 4) \vee 0] = 5p + (5p \vee 2).$$

Similarly, (5.5.3) gives

$$V^{(2)}(F, 0; \mathbf{A}) = 0 + [5p \vee 0] = 5p.$$

So

$$\Delta(F, 0; \mathbf{A}) = 5p \vee 2 > 0$$

and

$$\Delta((6)F, 0; \mathbf{A}^{(1)}) = 10p - 4,$$

which is negative for $p < 2/5$. For such a value of p, arm 1 is uniquely optimal initially and switching is optimal if '6' is observed but not if '4' is observed.

In addition, $6 + V((6)F, 0; \mathbf{A}^{(1)}) = (10p + 2) \vee 6$ is less than $4 + V((4)F, 0; \mathbf{A}^{(1)}) = 8$ for $p < 3/5$; for such a value of p, the decision maker would rather observe $X_{11} = 4$ than $X_{11} = 6$. If the horizon were greater than 2 the difference between the V's would be even greater. □

While a bandit problem with general F is less wieldy than one for which F concentrates its mass on Bernoulli distributions, many of the results given in the early sections of this chapter hold more generally. Some proofs are unchanged and others can be modified easily. We will repeat those results formally in this section.

The first is Theorem 5.2.2, the proof of which applies without change.

Theorem 5.5.1 Consider a discount sequence **A**. The following two statements are equivalent:

(i) For every $(F, \lambda; \mathbf{A})$-bandit there is an optimal strategy under which every selection of the known arm, arm 2, is followed by another selection of arm 2;
(ii) **A** is regular.

Furthermore, if $\alpha_1 > 0$ and **A** is regular, then if ever arm 2 becomes optimal, an optimal continuation is to select arm 2 exclusively and indefinitely.

The modification of the proof of Corollary 5.2.3 required for the present context is sufficiently involved that we give it here. First we state the result.

Corollary 5.5.2 For each regular discount sequence **A** and each distribution F on $\mathscr{D}$, there exists an optimal strategy τ for the $(F, \lambda; \mathbf{A})$-bandit such that

$$E_\tau(Z_{m+1} | F, \lambda) \geq E_\tau(Z_m | F, \lambda) \tag{5.5.4}$$

for $m = 1, 2, \ldots$.

Proof If arm 2 is uniquely optimal initially then take $\tau(\varnothing) = 2$; otherwise take $\tau(\varnothing) = 1$. Define τ inductively to be an optimal strategy that indicates arm 1 whenever possible without requiring a selection of arm 1 following a selection of arm 2. (The fuss in the preceding sentences is required to take care of the possibility that some discount factors may be zero.) We prove (5.5.4) only for $m = 1$; the general case is similar, but requires conditioning on the first $m-1$ stages.

If $\tau(\varnothing) = 2$, then $\tau(z_1) = 2$ for all z_1 so $E_\tau(Z_2|F, \lambda) = \lambda$ as does $E_\tau(Z_1|F, \lambda)$. Suppose $\tau(\varnothing) = 1$. Then

$$E_\tau(Z_2|F, \lambda) = E(X_{12}\mathbf{I}_{\{\tau(X_{11})=1\}}|F) + E(\lambda\mathbf{I}_{\{\tau(X_{11})=2\}}|F).$$

If $\tau(X_{11}) = 2$, then Theorem 5.0.2 applies to show that $\lambda > E(X_{12}|X_{11}, F)$. Hence,

$$\begin{aligned} E(\lambda\mathbf{I}_{\tau(X_{11})=2}|F) &\geqslant E(E(X_{12}|X_{11}, F)\mathbf{I}_{\{\tau(X_{11})=2\}}|F) \\ &= E(E(X_{12}\mathbf{I}_{\{\tau(X_{11})=2\}}|X_{11}, F)|F) \\ &= E(X_{12}\mathbf{I}_{\{\tau(X_{11})=2\}}|F). \end{aligned}$$

So

$$\begin{aligned} E(Z_2|F, \lambda) &\geqslant E(X_{12}\mathbf{I}_{\{\tau(X_{11})=1\}}|F) + E(X_{12}\mathbf{I}_{\{\tau(X_{11})=2\}}|F) \\ &= E(X_{12}|F) = E(X_{11}|F) = E_\tau(Z_1|F, \lambda), \end{aligned}$$

as desired. □

Example 5.5.1 shows that there is no 'break-even' observation, an X_{11} such that arm 1 continues to be optimal for larger observations and not for smaller. But the next theorem says there is a break-even value of λ. The theorem is adapted from Theorem 5.3.1, whose proof applies in this setting as well.

Theorem 5.5.3 For each regular discount sequence $\mathbf{A}$ with $\alpha_1 > 0$ and each distribution F on $\mathscr{Y}$, there exists a unique $\Lambda(F, \mathbf{A}) \in \mathbb{R}$ such that arm 1 is optimal initially in the $(F, \lambda; \mathbf{A})$-bandit if and only if $\lambda \leqslant \Lambda(F, \mathbf{A})$ and arm 2 is optimal if and only if $\lambda \geqslant \Lambda(F, \mathbf{A})$.

Example 5.5.2 Consider the $(F, \lambda; \mathbf{A})$-bandit from Example 5.5.1, but with λ not restricted to equal 0. Then

$$\Delta(F, \lambda; \mathbf{A}) = 5p + \tfrac{1}{2}[4 \vee \lambda] + \tfrac{1}{2}[(10p - 4) \vee \lambda] - \lambda - [5p \vee \lambda].$$

When this equals 0, and so $\lambda = \Lambda(F, \mathbf{A})$, Theorem 5.5.1 applies to give $\Lambda(F, \mathbf{A}) \geqslant 5p = E(X_{11}|F)$. Solving

$$5p + \tfrac{1}{2}[4 \vee \lambda] + \tfrac{1}{2}[(10p - 4) \vee \lambda] - 2\lambda = 0$$

for λ in terms of p gives

$$\Lambda(F, \mathbf{A}) = \begin{cases} \dfrac{10p + 4}{3}, & p \in [0, \frac{4}{5}] \\ \dfrac{20p - 4}{3}, & p \in [\frac{4}{5}, 1]. \end{cases}$$

In view of Theorem 5.5.1, $V(F, \Lambda(F, \mathbf{A}); \mathbf{A}) = 2\Lambda(F, \mathbf{A})$. □

An example in which the unknown arm has a normal distribution was considered in Section 2.1. We return to that example to calculate Λ.

Example 5.5.3 Consider the $((\mu, \rho), \lambda; \mathbf{A})$-bandit discussed in Section 2.1. Observations X_{1m} on arm 1 are normally distributed with unknown mean θ_1 and variance 1. Distribution F can be regarded as a distribution on θ_1; it is itself normal with mean μ and variance $\rho^2 > 0$. The discount sequence is $\mathbf{A} = (1, 1, 0, 0, 0, \ldots)$.

In Section 2.1 we showed that arm 1 is optimal initially if and only if

$$(\lambda - \mu)\rho^{-2}\sqrt{(1 + \rho^2)} \leqslant t_0,$$

where $t_0 \approx 0.2760$. Therefore,

$$\Lambda((\mu, \rho), \lambda; \mathbf{A}) = \mu + t_0\rho^2/\sqrt{(1 + \rho^2)}. \tag{5.5.5}$$

This is another instance of the phenomenon indicated by Theorem 5.0.2: arm 2 is optimal only if it offers higher immediate payoff than does arm 1, and it must offer substantially higher payoff if arm 1 involves much uncertainty (ρ^2 large). (Cf. discussion in Section 2.1.) □

The next result gives a general method for calculating $\Lambda(F, \mathbf{A})$. Its statement and proof are identical to those of Corollary 5.3.2.

Corollary 5.5.4 For $\mathbf{A}$ regular with $\alpha_1 > 0$, the function Λ is given by

$$\Lambda(F, \mathbf{A}) = \max_{\tau(\emptyset) = 1} \frac{E_\tau\left(\sum_{m=1}^{M} \alpha_m X_{1m} \middle| F\right)}{E_\tau\left(\sum_{m=1}^{M} \alpha_m \middle| F\right)} \tag{5.5.6}$$

where M is the (random) stage (possibly $+\infty$) at which arm 1 is used for the last time when following strategy τ. Among those τ's that begin with arm 1 and never switch back to arm 1 after a selection of arm 2, those that are optimal for the $(F, \Lambda(F, \mathbf{A}); \mathbf{A})$-bandit are those for which the maximum in (5.5.6) is attained.

In applying this result it is advisable – perhaps essential – to first pare down the set of strategies to be considered in (5.5.6). For example, the formula for Λ given by (5.5.5) can be obtained using Corollary 5.5.4, but the strategies to be considered are too numerous to be manageable. However, it is possible to argue that one need only consider those strategies in (5.5.6) that indicate arm 1 at stage 2 if the initial observation (on arm 1) is sufficiently large, and arm 2 otherwise.

In the next section we turn to a rather comprehensive example that uses the results of the present section.

5.6 Arms with Dirichlet measures

In this section we give an application of the considerations in Section 5.5. The form of distribution Q_1 is not known, perhaps not even up to a parameter with a countable number of dimensions. Data can be gathered on its form by observing arm 1, but the decision maker is not willing in advance to specify that Q_1 is normal, say, as in Example 5.5.3, with only the mean of Q_1 unknown. In this section we consider a bandit problem incorporating the nonparametric approach to statistical inference of Ferguson (1973). In doing so we exploit more fully the general notation developed in Section 2.2. The section is based on Clayton and Berry (1984). Following Sethuraman and Tiwari (1982) we use the term 'Dirichlet measure' rather than 'Dirichlet process prior' used by Ferguson.

The distribution F of Q_1 is a Dirichlet measure with parameter ν, which is itself a measure. The parameter ν is a finite non-null measure on $\mathbb{R}$ (though we will consider the limit as its total measure approaches 0 or ∞). For any measurable partition $(B_1, \ldots, B_m)$ of $\mathbb{R}$ and any m, the random vector $(Q_1(B_1), Q_1(B_2), \ldots, Q_1(B_m))$ has a Dirichlet distribution with parameter $(\nu(B_1), \nu(B_2), \ldots, \nu(B_m))$. So, for example,

$$E(Q_1(B_1)|F) = \nu(B_1)/\nu(\mathbb{R}). \tag{5.6.1}$$

Let $I = \nu(\mathbb{R})$ and let $\pi(\mathrm{d}x) = \nu(\mathrm{d}x)/I$ be the normalized form of ν;

we sometimes write πI in place of ν. In view of (5.6.1), π is the prior mean of Q_1:

$$E(Q_1(\mathrm{d}x)|F) = \pi(\mathrm{d}x).$$

The total measure I (which stands for amount of 'information') can be interpreted as the weight of the prior in terms of sample number (Ferguson, 1973). The prior mean of an observation on arm 1 is the mean of π:

$$E(X_{11}|F) = \int_{\mathbf{R}} x\pi(\mathrm{d}x),$$

which we assume to be finite.

The parameter ν summarizes prior information concerning Q_1. Given $X_{11}, \ldots, X_{1m}$, the posterior distribution of Q_1 is a Dirichlet measure with parameter $\nu + \sum_{j=1}^{m} \delta_{X_{1j}}$ (Ferguson, 1973, Theorem 1).

The case $I = 0$ gives rise to an improper Dirichlet measure which we define as follows. When $\nu = \pi \cdot 0$, observations $X_{11}, \ldots, X_{1n}$ are almost surely equal and each has unconditional distribution π. As shown by Sethuraman and Tiwari (1982), as $I \to 0$ the Dirichlet measure with parameter πI tends to the improper measure with parameter $\pi \cdot 0$ defined here.

The limiting case $I \to \infty$ corresponds to knowing in advance that $Q_1 = \pi$.

Ferguson (1973, Proposition 3) shows that, with respect to the topology of convergence in distribution, the support of F is the set of all distributions whose supports are contained in the support of ν. So the Dirichlet measure allows for modelling settings in which responses can take on values in any specified set. It actually encompasses the Bernoulli model when the parameter of the Bernoulli has a beta distribution. To see this, let $\nu = a\delta_1 + b\delta_0$ (in which case $I = a + b$); then $X_{11}, X_{12}, \ldots$ are distributed as conditionally independent Bernoulli variables whose parameter has the density given in (5.4.1). Also, the improper Dirichlet measure with $\nu = (a\delta_1 + b\delta_0) \cdot 0$ corresponds to the two-point prior on the Bernoulli parameter that is assumed in Example 5.4.4 (by setting $p = a/(a+b)$).

Assume **A** is the n-horizon uniform. Since it depends only on the horizon n we shall adopt the convention (to be used again in Chapter 7) of using n where **A** normally appears. In a similar vein, we write ν in place of F. So $(F, \lambda; \mathbf{A})$ is now written $(\nu, \lambda; n)$.

For $n \geqslant 1$ we have

$$V^{(1)}(v, \lambda; n) = E(X_{11}|v) + E[V(v + \delta_{X_{11}}, \lambda; n-1)|v], \qquad (5.6.2)$$
$$V^{(2)}(v, \lambda; n) = \lambda + V(v, \lambda; n-1)$$

and, of course, $V(v, \lambda; 0) = 0$. Theorem 5.5.1 applies to show that the $(v, \lambda; n)$-bandit is a stopping problem. So

$$V(v, \lambda; n) = V^{(1)}(v, \lambda; n) \vee n\lambda.$$

Since **A** is regular, Theorem 5.5.2 indicates that there exists a $\Lambda(v, n)$ such that arm 1 is optimal if and only if $\lambda \leqslant \Lambda(v, n)$.

We shall find $\Lambda(v, n)$ in a simple example.

Example 5.6.1 Suppose $n = 2$ and v is the following discrete measure:

$$v(\{x\}) = \left(\frac{1}{2}\right)^{x+1} I, \qquad x = 0, 1, 2, \ldots$$

To obtain $\Lambda(v, 2)$ we will solve the equation

$$V^{(1)}(v, \lambda; 2) = 2\lambda. \qquad (5.6.3)$$

To find $V^{(1)}(v, \lambda; 2)$ we require

$$\begin{aligned} V(v + \delta_{X_{11}}, \lambda; 1) &= E(X_{12}|v + \delta_{X_{11}}) \vee \lambda \\ &= E\left[\left(\frac{I}{I+1} E(X_{12}|v) + \frac{1}{I+1} X_{11}\right) \vee \lambda \,\middle|\, v\right] \\ &= E\left[\frac{I + X_{11}}{I+1} \vee \lambda \,\middle|\, v\right], \end{aligned}$$

since $E(X_{12}|v) = 1$. Therefore $\Lambda(v, 2)$ is the unique solution of (5.6.3) in view of Theorems 5.5.1 and 5.5.3.

Consider equation (5.6.3) for $\lambda \in [1, (I+2)/(I+1)]$. We obtain

$$\frac{I + X_{11}}{I+1} \vee \lambda = \begin{cases} \lambda & \text{if } X_{11} \in \{1, 0\} \\ \dfrac{I + X_{11}}{I+1} & \text{if } X_{11} \in \{2, 3, \ldots\}. \end{cases}$$

Hence, for such λ, a straightforward calculation gives

$$E[V(v + \delta_{X_{11}}, \lambda; 1)|v] = \frac{3\lambda}{4} + \frac{I+3}{4(I+1)}.$$

From (5.6.2),

$$V^{(1)}(v, \lambda; 2) = \frac{3\lambda}{4} + \frac{5I+7}{4(I+1)}.$$

Setting this equal to 2λ we obtain $\lambda = (5I+7)/(5I+5)$, which does satisfy $1 \leqslant \lambda \leqslant (I+2)/(I+1)$. So we have found the unique solution of (5.6.3):

$$\Lambda(v, 2) = \frac{5I+7}{5I+5},$$

valid for $0 \leqslant I \leqslant \infty$.

This example provides another instance in which arm 1 is more attractive when less is known about it, which in this case occurs when I is smaller. □

In the above example there is a number such that arm 1 continues to be optimal if X_{11} is greater than that number and arm 2 is optimal otherwise. Example 5.5.1 shows that there may not exist such a number when X_{1m} can take on more than two values. But the Dirichlet model exhibits sufficient smoothness that such an example is not possible. Clayton and Berry (1984) give the following result; see their paper for the proof.

Theorem 5.6.1 Fix $\lambda, n \geqslant 2$, and $I < \infty$. For any $(v, \lambda; n)$-bandit there exists a unique $b^* = b^*(v, \lambda; n)$ such that if arm 1 is selected (whether or not it is optimal) then arm 1 is optimal at stage 2 if $X_{11} \geqslant b^*$ and arm 2 is optimal if $X_{11} \leqslant b^*$.

Remarks When λ is greater than the support of v, so is $b^*(v, \lambda; n)$. While the probability is 0 that X_{11} is greater than the support of v, the Dirichlet measure with parameter $v + \delta_{X_{11}}$ is well-defined nonetheless.

In view of the theorem, $b^*(v, \lambda; n)$ is the unique solution of the equation

$$\lambda = \Lambda(v + \delta_{b^*}, n-1). \qquad \square$$

The 'break-even' observation $b^*(v, \lambda; n)$ obviously depends on λ. The special case $\lambda = \Lambda(v, n)$ gives the next result. The result says there is a quantity such that the attractiveness of arm 1 (as measured by Λ) increases if X_{11} is greater than that quantity. Such an X_{11} might be

called a 'winner', so the coming result generalizes the stay-with-a-winner property of Bernoulli bandits.

Corollary 5.6.2 For $n \geqslant 2$ and all ν with $I < \infty$, there exists a unique $b(\nu, n)$ such that

$$\Lambda(\nu + \delta_{X_{11}}, n-1) \geqslant \Lambda(\nu, n) \text{ if } X_{11} \geqslant b(\nu, n)$$
$$\Lambda(\nu + \delta_{X_{11}}, n-1) \leqslant \Lambda(\nu, n) \text{ if } X_{11} \leqslant b(\nu, n).$$

Remark Clearly, $b(\nu, n) = b^*(\nu, \Lambda(\nu, n); n)$. □

The function b^* determines all optimal strategies but the function b does not.On the other hand, calculating b is easier than calculating b^* since b does not depend on λ. Still, the task of finding $b(\nu, n)$ is formidable if not impossible when n is moderate and the support of ν is at all complicated. Not only is dynamic programming required, but there are substantial computational difficulties when dealing with Dirichlet measures (see Berry and Christensen, 1979, for example).

The next two examples give some calculations of $\Lambda(\nu, n)$ and $b(\nu, n)$ for small n.

Example 5.6.2 Suppose π is the uniform distribution on (0, 1). Table 5.3 gives $\Lambda(\pi I, n)$ and $b(\pi I, n)$ for $I = 0,\ 0.1,\ 1,\ 5,\ \infty$, and $n = 2, 3, 4$. □

Example 5.6.3 Suppose π is the standard normal distribution. Table 5.4 gives $\Lambda(\pi I, n)$ and $b(\pi I, n)$ for the same combinations of I and n considered in the previous example. □

The specification of an optimal strategy is quite complicated, depending on ν and n, when $\lambda < \Lambda(\nu, n)$. Arm 1 is optimal initially. After observing X_{11}, one calculates $\Lambda(\nu + \delta_{X_{11}}, n-1)$ and compares it with λ. If arm 1 is optimal again, one calculates $\Lambda(\nu + \delta_{X_{11}} + \delta_{X_{12}}, n-2)$ and compares it with λ. And so on.

Example 5.6.4 As in Example 5.6.2 suppose $\pi = U(0, 1)$ and now assume $n = 4$. Table 5.3 indicates which arm is optimal initially. Suppose arm 1 is selected; the new bandit is $(\pi I + \delta_{X_{11}}, \lambda; 3)$. Table 5.5 gives $\Lambda(\pi I + \delta_{X_{11}}, 3)$ for various X_{11} and $I = 0.1, 1, 5$; of course the 'I' for the new bandit is $I + 1$. Table 5.5 also gives $b(\pi I + \delta_{X_{11}}, 3)$.

Table 5.3 $\Lambda(\pi I, n)$ *and* $b(\pi I, n)$ *where* $\pi = U(0, 1)$

I	$\Lambda(\pi I, 2)$	$b(\pi I, 2)$	$\Lambda(\pi I, 3)$	$b(\pi I, 3)$	$\Lambda(\pi I, 4)$	$b(\pi I, 4)$
0	0.586	0.586	0.634	0.634	0.667	0.667
0.1	0.578	0.586	0.623	0.630	0.654	0.661
1	0.543	0.586	0.570	0.610	0.590	0.630
5	0.514	0.586	0.524	0.584	0.532	0.589
∞	0.500	–	0.500	–	0.500	–

Table 5.4 $\Lambda(\pi I, n)$ *and* $b(\pi I, n)$ *where* π *is standard normal*

I	$\Lambda(\pi I, 2)$	$b(\pi I, 2)$	$\Lambda(\pi I, 3)$	$b(\pi I, 3)$	$\Lambda(\pi I, 4)$	$b(\pi I, 4)$
0	0.276	0.276	0.436	0.436	0.549	0.549
0.1	0.251	0.276	0.400	0.424	0.505	0.529
1	0.138	0.276	0.228	0.359	0.295	0.421
5	0.046	0.276	0.079	0.276	0.105	0.284
∞	0.000	–	0.000	–	0.000	–

Table 5.5 $\Lambda(\pi I + \delta_{X_{11}}, 3)$ [*and* $b(\pi I + \delta_{X_{11}}, 3)$] *where* $\pi = U(0, 1)$

X_{11}	0.5	0.6	0.7	0.8	0.9	1.0
$I = 0.1$	0.510 [0.516]	0.598 [0.602]	0.690 [0.695]	0.783 [0.790]	0.877 [0.885]	0.971 [0.981]
$I = 1$	0.529 [0.562]	0.575 [0.602]	0.630 [0.662]	0.685 [0.723]	0.741 [0.786]	0.798 [0.851]
$I = 5$	0.518 [0.572]	0.535 [0.588]	0.552 [0.608]	0.570 [0.630]	0.588 [0.652]	0.607 [0.674]

Suppose $I = 1$ and $\lambda = 0.58$. Then arm 1 is optimal initially since $\lambda < \Lambda(\pi I, 4) = 0.590$ from Table 5.3. According to Table 5.5, if $X_{11} = 0.5$ or 0.6, then arm 2 should be observed next (and for the remaining two selections as well). On the other hand, arm 1 remains optimal if $X_{11} \geqslant 0.7$. So $0.6 < b^*(\pi \cdot 1, 0.58; 4) < 0.7$. Suppose $X_{11} = 0.7$. Then the tabulated value $b(\pi \cdot 1 + \delta_{0.7}, 3) = 0.662$ in-

dicates that arm 1 continues to be optimal if $X_{12} \geqslant 0.662$ (arm 1 is also optimal for X_{12} as small as $b^*(\pi \cdot 1 + \delta_{0.7}, 0.58; 3) < 0.662$). □

Examples 5.6.2 to 5.6.4 suggest that $\Lambda(v, n) \leqslant b(v, n)$; Clayton and Berry (1984) conjecture that this holds generally. This conjecture has the following motivation. Suppose the decision maker is indifferent regarding the two arms, arm 2 producing the known constant $\lambda = \Lambda(v, n)$. If arm 1 is selected, it is selected with the hope that it is better than the alternative, which is arm 2. So if it yields less than arm 2 is known to deliver, the hope for it fades. And since it was barely optimal to start with, arm 1 must no longer be optimal if $X_{11} < \Lambda(v, n)$.

Consider the $(\pi \cdot 0, \lambda; n)$-bandit. Since arm 1 becomes known with just a single observation, any $X_{11} \geqslant \lambda$ will make arm 1 optimal again (and thereafter) and any $X_{11} \leqslant \lambda$ will make arm 2 optimal. Therefore, $b^*(\pi \cdot 0, \lambda; n) = \lambda$. On the other hand, when $I > 0$ one should not have to observe as large an X_{11} to want to stay with arm 1. So the following result of Clayton and Berry (1984) seems reasonable and is supported by Tables 5.3 and 5.4; see their paper for its proof.

Theorem 5.6.3 For all $v = \pi I$ and $n \geqslant 2$, $b(v, n) \leqslant \Lambda(\pi \cdot 0, n)$.

This theorem gives an easily computable bound for $b(v, n)$.

In the next section we turn to the setting in which the stages occur at random times.

5.7 Real-time discounting

So far in this chapter we have considered the observation at any stage to be discounted independent of the time t at which the stage occurs. In this section we suppose, as described in Section 3.5, that the observation is weighted by α_t which is unrelated to the stage number. We allow more general discounting than that of the exponential function emphasized in Section 3.5. We require the times between stages to be independent and exponentially distributed with known mean κ. In this section we shall indicate which results in the previous sections of this chapter carry over to the current setting.

There are at least two ways of thinking about a strategy as defined in Section 3.5. Selections can depend on t. We can imagine that the decision maker waits until a stage occurs (a patient arrives, a part fails,

etc.), selects an arm and instantaneously observes the result. Or, we can suppose that the decision maker keeps an arm operative at all times, changing arms depending on t as well as on any observations on the arms; when a stage occurs the operative arm is observed. This latter view is particularly handy for describing various aspects of optimal strategies.

Suppose the first stage occurs at time $s > 0$. The appropriate bandit is not $(F, \lambda; \alpha_t, \kappa)$, but rather $(F, \lambda; \alpha_t^{(s)}, \kappa)$, where $\alpha_t^{(s)}$ is α_t shifted by s units; that is, $\alpha_t^{(s)} = \alpha_{t+s}$ for $t \geqslant 0$.

Theorems 5.0.1 and 5.0.2 apply in the present setting. We do not know whether the '(i) $\Rightarrow$ (ii)' portion of Theorem 5.5.1 applies; the randomness in the present setting may enable 'regular portions' of a discount function to compensate for 'irregular portions'. The remaining results of Section 5.5 do apply, provided appropriate definitions are used. In analogy with $\gamma_m = \sum_{i=m}^{\infty} \alpha_i$ in discrete time, let

$$\gamma_t = \int_t^{\infty} \alpha_s \, ds.$$

Definition 5.7.1 A discount function α_t is *regular* if α_t/γ_t is nondecreasing at all t for which $\gamma_t > 0$.

The following proposition makes it clear that this definition is a natural continuous-time version of Definition 5.2.1.

Proposition 5.7.1 If α_t is regular, then $\gamma_{t+2s}\gamma_t \leqslant \gamma_{t+s}^2$ for every nonnegative t and s.

Proof Assume that α_t is regular and, with no loss, that $\gamma_{t+2s} > 0$. Then

$$\log(\gamma_t/\gamma_{t+s}) = \int_t^{t+s} (\alpha_u/\gamma_u)\, du \leqslant \int_{t+s}^{t+2s} (\alpha_u/\gamma_u)\, du$$
$$= \log(\gamma_{t+s}/\gamma_{t+2s}). \qquad \square$$

Remark The converse of this proposition is not true since, for instance, α_s can be changed at any particular s, leaving γ_t unchanged for all $t \geqslant 0$. $\square$

With this definition of regularity the '(ii) $\Rightarrow$ (i)' and 'furthermore' portions of Theorem 5.5.1 hold. Therefore we can view a bandit with

a regular discount function as a stopping problem. Corollary 5.5.2 also holds in the setting of real-time discounting. This means that when α_t is regular there is an optimal strategy for which later observations have no smaller expectations than earlier ones.

The analogue of Theorem 5.5.3 holds here as well. This means that there is a break-even value $\Lambda(F, \alpha_t, \kappa)$ that can be calculated for arm 1 and that indicates the optimal arm by comparing it with the constant λ yielded by arm 2. Suppose we imagine the decision maker as keeping one arm operative at all times anticipating the occurrence of a stage. When α_t is regular we have as an immediate corollary of the stopping nature of the bandit that the decision maker need never consider switching control from arm 2 to arm 1, whether or not stages occur. But it is easy to find F, λ, regular α_t, and κ such that an optimal decision maker keeps arm 1 operative and, should no stage occur within some period of time, switch and put arm 2 into operation. Summarizing, we have the following result.

Corollary 5.7.2 Suppose α_t is regular with $\alpha_0 > 0$. Then $\Lambda(F, \alpha_t^{(s)}, \kappa)$ is a nonincreasing function of s when $\alpha_s > 0$.

For the present setting the appropriate modification of the formula (5.5.6) for Λ is

$$\Lambda(F, \alpha_t, \kappa) = \max_{\tau(\varnothing;0)=1} \frac{E_\tau\left(\sum_{m=1}^{M} \alpha_{T_m} X_{1m} \middle| F, T_1 = 0\right)}{E_\tau\left(\sum_{m=1}^{M} \alpha_{T_m} \middle| F, T_1 = 0\right)}, \quad (5.7.1)$$

where T_m denotes the random time at which stage m occurs.

Example 5.7.1 Suppose that arm 1 is Bernoulli with $F = p\delta_1 + (1-p)\delta_0$. To use (5.7.1) we can assume without loss that a stage occurs at $t = 0$ and arm 1 is selected. The only strategy that need be considered in (5.7.1) is staying with arm 1 indefinitely if a success is obtained and switching permanently to arm 2 if failure is obtained. Accordingly,

$$\Lambda(F, \alpha_t, \kappa) = \frac{p \sum_{m=1}^{\infty} E(\alpha_{T_m} | T_0 = 1)}{p \sum_{m=1}^{\infty} E(\alpha_{T_m} | T_0 = 1) + (1-p)\alpha_0}. \qquad \square$$

The construction leading to Theorem 3.5.1 does not give Bernoulli discrete-time approximations to a Bernoulli real-time bandit. Nevertheless, the inductive proof of Theorem 4.1.6 can be mimicked by considering real-time bandits that terminate after a fixed number of stages, no matter when they occur. Hence, $V(F^*, \lambda; \alpha_t, \kappa) \geq V(F, \lambda; \alpha_t, \kappa)$ if F^* is strongly to the right of F. If, in addition, α_t is regular and $\alpha_0 \neq 0$, then $\Lambda(F^*, \alpha_t, \kappa) \geq \Lambda(F, \alpha_t, \kappa)$ (cf. Corollary 5.3.4).

Many of the results obtained in the Bernoulli case considered in Chapter 4 and earlier in this chapter do not carry over easily to this setting. Suppose, for example, $\alpha_t = \mathbf{1}_{[0,1]}(t)$. It is easy to choose F, λ, and κ so that the following hold. Arm 1 is optimal for the $(F, \lambda; \alpha_t, \kappa)$-bandit at $t = 0$ if stage 1 occurs then. It is not optimal at stage 2 if stage 2 occurs at time t close to 1, even if a success is observed at stage 1. As an analogue of Theorem 5.3.5, when α_t is regular it is possible to show that $\Lambda(F, \alpha_t, \kappa) \leq \Lambda(\sigma F, \alpha_t^{(s)}, \kappa)$ for all sufficiently small s, but the result is of little consequence since it is not possible to predict the timing of the stages.

However, most of the results in the Bernoulli case with geometric discounting have natural and useful analogues in the current setting if we assume an exponential discount function. For, if $\alpha_t = \exp(-\beta t)$ (cf. Section 3.5) then $(F, \lambda; \alpha_t^{(s)}, \kappa)$ is effectively the same bandit for all $s \geq 0$. So it is easy to show that the stay-with-a-winner rule (Theorem 5.3.5) holds in this case.

References

Bellman, R. (1956) A problem in the sequential design of experiments. *Sankhyā A* **16**: 221–229.

Berry, D. A. and Christensen, R. (1979) Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.* **7**: 558–568.

Berry, D. A. and Fristedt, B. (1979) Bernoulli one-armed bandits—arbitrary discount sequences. *Ann. Statist.* **7**: 1086–1105.

Berry, D. A. and Fristedt, B. (1983) Maximizing the length of a success run for many-armed bandits. *Stochastic Process. Appl.* **15**: 317–325.

Bradt, R. N., Johnson, S. M. and Karlin, S. (1956) On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.* **27**: 1060–1074.

Clayton, M. K. and Berry, D. A. (1984) Bayesian nonparametric bandits. Statistics Tech. Rep. No. 427, Univ. of Minnesota, USA.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**: 209–230.

Gittins, J. C. and Jones, D. M. (1974) A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (eds J. Gani *et al.*), pp. 241–266, North-Holland, Amsterdam.

Gittins, J. C. and Jones, D. M. (1979) A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika* **66**: 561–565.

Sethuraman, J. and Tiwari, R. C. (1982) Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical Decision Theory and Related Topics III*, Vol. 2 (eds S. Gupta and J. O. Berger), pp. 305–315, Academic Press, New York.

CHAPTER 6

Many independent arms; geometric discounting

We have indicated several times that the two most widely studied discount sequences are the n-horizon uniform and the geometric. The former applies when the problem is to maximize the sum of n observations. When n is unknown the corresponding random discount sequence can be taken to be nonrandom (see Section 3.1); it can be any nonincreasing sequence depending on the uncertainty in n. As a special case suppose n has a geometric distribution; so the opportunity for gain ceases at each stage with constant probability α. Then, and in many other circumstances as well, the appropriate discount sequence is geometric: $\mathbf{A} = (1, \alpha, \alpha^2, \alpha^3, \ldots)$.

This fact gives geometric discounting special significance and, in our view, makes the geometric sequence the single most important discount sequence. Luckily, it is also the most mathematically tractable (except, of course, for sequences with very small horizons). The fact that, alone among discount sequences, the geometric does not change from one stage to the next (after a suitable normalization at each stage) gives a decision problem with geometric discounting a special quality and makes possible the elegant result of Gittins and Jones (1974) presented in Section 6.1. This result says that for k independent arms and geometric discounting, the desirability of an arm can be completely specified by a number that depends only on that arm (and on α) and not on the other arms. Gittins and Jones (1974) call this number the arm's *dynamic allocation index*. Any arm with the largest index is optimal. In view of independence, an arm's index can change only when the arm is observed. So optimal strategies are conceptually easy to describe: always select an arm with the largest index, switching (to the arm that was second-best) only when its index drops from first place.

Suppose we define, as seems reasonable, the dynamic allocation index of an arm with known mean to be that mean. Then the index of arm i can only be $\Lambda(F_i, \mathbf{A})$ as defined in Theorem 5.5.3. So in order to decide how good an arm is, one need only compare it with known arms, defining an arm's index to be the mean of a known arm such that both arms would be optimal in a two-armed bandit problem. And the optimal initial selections in a k-armed bandit with independent arms can be found by solving k families of two-armed bandits, comparing each arm in turn with various known arms.

We show in Section 6.2 (Theorem 6.2.1) that when discounting is regular and strategies specified by dynamic allocation indices are optimal, the discount sequence can only be geometric. So the Gittins–Jones characterization is only possible in the geometric case. An equivalent and rather picturesque characterization is as follows, given in the context of $k = 3$. Consider three independent arms and suppose they are compared in pairs in three two-armed bandits. Is it possible that arm 1 is uniquely optimal in the $(F_1, F_2; \mathbf{A})$-bandit, arm 2 is uniquely optimal in the $(F_2, F_3; \mathbf{A})$-bandit, and arm 3 is uniquely optimal in the $(F_3, F_1; \mathbf{A})$-bandit? Such lack of transitivity is impossible if optimal strategies are specified by dynamic allocation indices. For $\Delta(F_1, F_2; \mathbf{A}) > 0$ would imply $\Lambda(F_1, \mathbf{A}) > \Lambda(F_2, \mathbf{A})$, and $\Delta(F_2, F_3; \mathbf{A}) > 0$ would imply $\Lambda(F_2, \mathbf{A}) > \Lambda(F_3, \mathbf{A})$; so $\Delta(F_3, F_1; \mathbf{A}) > 0$ would be impossible. On the other hand, lack of transitivity is possible when optimal strategies are not specified by dynamic allocation indices. The following example shows that such strategies are not optimal when $\mathbf{A}$ is the 2-horizon uniform. The example is actually a special case of the family of examples used to prove Theorem 6.2.1.

Example 6.0.1. Suppose $\mathbf{A} = (1, 1, 0, 0, 0, \ldots)$. Consider two Bernoulli arms with, as has been our custom in this setting, F_i taken to be the distribution of the Bernoulli parameter θ_i. Suppose

$$F_1 = \tfrac{1}{2}\delta_0 + \tfrac{1}{2}\delta_1,$$
$$F_2 = \tfrac{5}{7}\delta_{1/2} + \tfrac{2}{7}\delta_1.$$

From Example 5.4.7 or Theorem 5.4.2.

$$\Lambda(F_1, \mathbf{A}) = 2/3 < \Lambda(F_2, \mathbf{A}) = 31/46.$$

Nevertheless, as we show below, arm 1 is optimal in the $(F_1, F_2; \mathbf{A})$-bandit.

To show lack of transitivity, consider a third arm with $F_3 = \delta_c$. Any $c \in (2/3, 31/46)$ will do; we take $c = 185/276$ to be specific. Since the maximum worth when selecting arm i in the $(F_i, F_j; \mathbf{A})$-bandit is

$$V^{(i)}(F_i, F_j; \mathbf{A}) = [2E(\theta_i|F_i)] \vee [E(\theta_i|F_i) + E(\theta_i^2|F_i) + (1 - E(\theta_i|F_i))E(\theta_j|F_j)]$$

and similarly for selecting arm j (cf. (4.2.1) and (4.2.2)), it follows that

$$\Delta(F_1, F_2; \mathbf{A}) = [\tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} \cdot \tfrac{9}{14}] - [2 \cdot \tfrac{9}{14}] + \tfrac{1}{28},$$

$$\Delta(F_2, F_3; \mathbf{A}) = [\tfrac{9}{14} + \tfrac{13}{28} + \tfrac{5}{14}c] - [2c] = \tfrac{1}{168},$$

$$\Delta(F_3, F_1; \mathbf{A}) = [2c] - [\tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2}c] = \tfrac{1}{184}.$$

(Alternatively, the latter two Δ's are known to be positive in view of the choice of c as compared with $\Lambda(F_1, \mathbf{A})$ and $\Lambda(F_2, \mathbf{A})$.) Therefore arm 1 is 'better than' arm 2, as advertised above, and arm 2 is 'better than' 3, which is 'better than' arm 1!

Incidentally, arm 2 is optimal in the $(F_1, F_2, F_3; \mathbf{A})$-bandit and

$$V(F_1, F_2, F_3; \mathbf{A}) = V(F_2, F_3; \mathbf{A}) = \tfrac{9}{14} + \tfrac{13}{28} + \tfrac{5}{14}c \approx 1.347. \qquad \square$$

In Section 6.1 we prove the Gittins–Jones result and in Section 6.2 we show that geometric discounting is necessary for the result. Some readers may find it useful to read Sections 6.1 and 6.2 in reverse order since the proof of Theorem 6.2.1 can help the intuition by showing why nongeometric sequences are ruled out.

6.1 Theorem of Gittins and Jones

When $\mathbf{A} = (1, \alpha, \alpha^2, \ldots)$ we will substitute α for $\mathbf{A}$ and write, for example, $(F_1, \ldots, F_k; \alpha)$ in place of $(F_1, \ldots, F_k; \mathbf{A})$. We note that the results in this section apply with occasional evident modification if $\mathbf{A} = (b, b\alpha, b\alpha^2, \ldots)$ with $b > 0$.

As indicated previously, the dynamic allocation indices $\Lambda(F_i, \alpha)$ play a central role in the $(F_1, \ldots, F_k; \alpha)$-bandit. This role is shown in the following fundamental result of Gittins and Jones (1974), stated as it applies with our definition of bandit problems. Their definition is broader than ours, so their result is more general than the version we state.

Theorem 6.1.1 The optimal initial selections in the $(F_1, \ldots, F_k; \alpha)$-bandit are those i for which

$$\Lambda(F_i, \alpha) = \bigvee_{j=1}^{k} \Lambda(F_j, \alpha).$$

Moreover,

$$V(F_1, \ldots, F_k; \alpha) = \frac{1}{1-\alpha} \lim_{\rho \to \infty} \left\{ \rho - (1-\alpha)^k \int_{-\infty}^{\rho} \prod_{j=1}^{k} \frac{\partial}{\partial \lambda} V(F_j, \lambda; \alpha) \mathrm{d}\lambda \right\} \quad (6.1.1)$$

Remark In view of Theorems 5.2.2 and 5.3.1, when $\lambda \geqslant \Lambda(F_i, \alpha)$,

$$V(F_i, \lambda; \alpha) = \lambda/(1-\alpha). \quad (6.1.2)$$

So

$$(1-\alpha)^k \prod_{j=1}^{k} \frac{\partial}{\partial \lambda} V(F_j, \lambda; \alpha) = 1$$

for sufficiently large λ and the limit in (6.1.1) can be replaced with the expression evaluated at any ρ at least as large as $\bigvee_{j=1}^{k} \Lambda(F_j, \alpha)$. □

We will essentially use the argument of Whittle (1982, Section 14.4) to prove Theorem 6.1.1. We need a modification of $(F_1, \ldots F_k; \alpha)$; this modification is not a bandit problem because of a restriction on the available strategies. To get this quasi-bandit, adjoin a $(k+1)$st arm with $X_{(k+1),m} = \lambda^*$ for all m and consider only those strategies under which arm $k+1$ is continued indefinitely once it has been selected. An asterisk indicates that this quasi-bandit is being considered; for instance, $V^*(F_1, \ldots, F_k, \lambda^*; \alpha)$ is its value. Slight modifications of the results in Sections 2.3 and 2.5 apply; so, for instance,

$$V^*(F_1, \ldots F_k, \lambda^*; \alpha) = [\lambda^*/(1-\alpha)] \vee \bigvee_{i=1}^{k} E(X_{i1} + \alpha V^*(F_1, \ldots, (X_{i1})F_i, \ldots, F_k, \lambda^*; \alpha) | F_i). \quad (6.1.3)$$

We need the following result.

Lemma 6.1.2 For each F and each $F_1, \ldots, F_k$, the functions $\lambda \mapsto V(F, \lambda; \alpha)$ and $\lambda^* \mapsto V^*(F_1, \ldots, F_k, \lambda^*; \alpha)$ are nondecreasing and concave upwards.

Proof For every τ that stays with arm 2 once arm 2 is selected, the function $\lambda \mapsto W(F, \lambda; \alpha; \tau)$ is obviously linear and nondecreasing. Similarly for $\lambda^* \mapsto W^*(F_1, \ldots, F_k, \lambda^*; \alpha; \tau^*)$, where τ^* is any strategy for the quasi-bandit $(F_1, \ldots, F_k, \lambda^*; \alpha)$. The supremum of a collection of nondecreasing linear functions is nondecreasing and concave upwards. □

Remark From Lemma 6.1.2 we conclude that $V(F, \cdot; \alpha)$ is absolutely continuous, and a nondecreasing right-continuous version of $\partial V(F, \lambda; \alpha)/\partial\lambda$ can be chosen. This derivative is integrable at $-\infty$ since $V(F, \lambda; \alpha) \to E(X_{11}|F)/(1-\alpha)$ as $\lambda \to -\infty$. Therefore,

$$(1-\alpha)^k \int_{-\infty}^{\rho} \prod_{j=1}^{k} \frac{\partial}{\partial\lambda} V(F_j, \lambda; \alpha)\, d\lambda < \infty$$

for each finite ρ. □

Proof of Theorem 6.1.1 For the quasi-bandit $(F_1, \ldots, F_k, \lambda^*; \alpha)$ we will prove that arm $k+1$ is optimal initially if and only if

$$\lambda^* \geq \bigvee_{j=1}^{k} \Lambda(F_j, \alpha), \tag{6.1.4}$$

that for $i = 1, \ldots, k$, arm i is optimal if and only if

$$\Lambda(F_i, \alpha) = \lambda^* \vee \bigvee_{j=1}^{k} \Lambda(F_j, \alpha), \tag{6.1.5}$$

and that $V^*(F_1, \ldots, F_k, \lambda^*; \alpha)$ equals

$$\xi(F_1, \ldots, F_k, \lambda^*; \alpha) = [1/(1-\alpha)] \lim_{\rho \to \infty} \left\{ \rho - (1-\alpha)^k \int_{\lambda^*}^{\rho} \prod_{j=1}^{k} \frac{\partial}{\partial\lambda} V(F_j, \lambda; \alpha)\, d\lambda \right\}. \tag{6.1.6}$$

The theorem will then follow by letting $\lambda^* \to -\infty$.

Define

$$\zeta_i(F_1, \ldots, F_k, \lambda; \alpha) = (1-\alpha)^{k-1} \prod_{j \neq i} \frac{\partial}{\partial\lambda} V(F_j, \lambda; \alpha).$$

By (6.1.2) and Lemma 6.1.2, ζ_i is nonnegative, nondecreasing in λ, and it equals 1 when

$$\lambda \geq \bigvee_{j \neq i} \Lambda(F_j, \alpha).$$

In view of (6.1.2), integrating by parts in (6.1.6) gives

$$\xi(F_1, \ldots, F_k, \lambda^*; \alpha) = V(F_i, \lambda^*; \alpha)\zeta_i(F_1, \ldots, F_k, \lambda^*; \alpha) + \int_{\lambda^*}^{\infty} V(F_i, \lambda; \alpha)\, \mathrm{d}_\lambda \zeta_i(F_1, \ldots, F_k, \lambda; \alpha). \tag{6.1.7}$$

That ξ satisfies (6.1.3) will follow by showing

$$\xi(F_1, \ldots, F_k, \lambda^*; \alpha) \geqslant \lambda^*/(1-\alpha) \tag{6.1.8}$$

with equality if and only if (6.1.4) holds, and that

$$\xi(F_1, \ldots, F_k, \lambda^*; \alpha) \geqslant E(X_{i1} + \alpha\xi(F_1, \ldots, (X_{i1})F_i, \ldots, F_k, \lambda^*; \alpha)|F_i) \tag{6.1.9}$$

with equality if and only if (6.1.5) holds.

Suppose (6.1.4) holds. Then $V(F_i, \lambda^*; \alpha) = \lambda^*/(1-\alpha)$ and $\zeta_i(F_1, \ldots, F_k, \lambda; \alpha) = 1$ for $\lambda \geqslant \lambda^*$. Hence, (6.1.7) gives equality at (6.1.8).

Now suppose (6.1.4) fails and choose i satisfying (6.1.5). For $\lambda^* \leqslant \lambda < \Lambda(F_i, \alpha)$, $V(F_i, \lambda; \alpha) > \lambda^*/(1-\alpha)$. So from (6.1.7) we obtain

$$\begin{aligned}\xi(F_1, \ldots, F_k, \lambda^*; \alpha) &> [\lambda^*/(1-\alpha)]\left[\zeta_i(F_1, \ldots, F_k, \lambda^*; \alpha) + \int_{\lambda^*}^{\Lambda(F_i,\alpha)} \mathrm{d}_\lambda \zeta_i(F_1, \ldots, F_k, \lambda; \alpha)\right] \\ &= [\lambda^*/(1-\alpha)]\zeta_i(F_1, \ldots, F_k, \Lambda(F_i, \alpha); \alpha) \\ &= \lambda^*/(1-\alpha),\end{aligned}$$

and strict inequality at (6.1.8) follows.

Using (6.1.7), we can write the difference between the two sides of (6.1.9) as follows:

$$\begin{aligned}&\{V(F_i, \lambda^*; \alpha) - E(X_{i1} + \alpha V((X_{i1})F_i, \lambda^*; \alpha)|F_i)\} \\ &\cdot\zeta_i(F_1, \ldots, F_k, \lambda^*; \alpha) + \int_{\lambda^*}^{\infty} \{V(F_i, \lambda; \alpha) \\ &\quad - E(X_{i1} + \alpha V((X_{i1})F_i, \lambda; \alpha)|F_i)\}\, \mathrm{d}_\lambda \zeta_i(F_1, \ldots, F_k, \lambda; \alpha).\end{aligned} \tag{6.1.10}$$

The factors within braces are nonnegative. The first term equals 0 if and only if $\lambda^* \leqslant \Lambda(F_i, \alpha)$. And when $\lambda^* \leqslant \Lambda(F_i, \alpha)$, the second term equals 0 if and only if

$$\zeta_i(F_1, \ldots, F_k, \Lambda(F_i, \alpha); \alpha) = \zeta_i(F_1, \ldots, F_k, +\infty; \alpha) = 1.$$

Thus, (6.1.10) equals 0 when (6.1.5) holds, and is positive when (6.1.5) fails; correspondingly, (6.1.9) holds with equality and strict inequality.

It remains to prove that $\zeta = V^*$. They both satisfy (6.1.3); we address the question of uniqueness for solutions of (6.1.3). Consider λ^* and α to be fixed and temporarily restrict consideration to those F_j's for which

$$P(|X_{j1}| > c|F_j) = 0 \tag{6.1.11}$$

for some c. Using the supremum norm on bounded measurable functions of $F_1, \ldots, F_k$ and the fact that the operator on the right-hand side of (6.1.3) is a contraction on the space of such functions we conclude that (6.1.3) has a unique solution.

We will use a limiting argument to remove restriction (6.1.11), even though V^* is not continuous (Example 2.5.1). We leave it to the reader to check that

$$\lim_{c \to \infty} V^*(F_{1c}, F_{2c}, \ldots, F_{kc}, \lambda^*; \alpha) = V^*(F_1, F_2, \ldots, F_k, \lambda^*; \alpha)$$

where F_{jc} is the measure on $\mathscr{D}$ induced by F_j and the mapping $Q_j \mapsto Q_{jc}$ and distribution function $Q_{jc}(x)$ is

$$Q_{jc}(x) = \begin{cases} 0 & \text{if } x < -c \\ Q_j(x) & \text{if } |x| < c \\ 1 & \text{if } x > c. \end{cases}$$

The functions $F_j \mapsto V(F_j, \lambda; \alpha)$ and $F_j \mapsto \dfrac{\partial}{\partial \lambda} V(F_j, \lambda; \alpha)$ also have this 'restricted continuity' property. Hence, so does ξ and, therefore, $V^* = \xi$. □

Remarks The restriction on switching from arm $k+1$ in the quasi-bandit $(F_1, \ldots, F_k, \lambda^*; \alpha)$ was necessary in the proof of Theorem 6.1.1. But in view of the theorem it is clear that this restriction is of no consequence in determining an optimal strategy in the quasi-bandit and so, effectively, the quasi-bandit is a bandit. □

The following two results are immediate consequences of Theorem 6.1.1, Corollary 5.3.4, and Lemma 4.1.5. The first says that staying with a winner is optimal; the second says that switching from a loser is optimal when the arm selected was barely optimal. No attempt to discuss unique optimality is made in stating these results.

Corollary 6.1.3 Suppose arm i is Bernoulli and is optimal in the $(F_1, \ldots, F_k; \alpha)$-bandit. If arm i yields a success then it is optimal at stage 2 as well.

Corollary 6.1.4 Suppose arms i and j in the $(F_1, \ldots, F_k; \alpha)$-bandit are both optimal. If arm i is Bernoulli, is selected, and yields a failure, then arm j is optimal at the second stage.

At several places, particularly in Section 5.4, we have considered examples with geometric discounting and one unknown arm. Theorem 6.1.1 gives those examples added significance: namely, it applies to show how to use the $\Lambda(F, \alpha)$ calculated for the unknown arm in solving a bandit for which one of the k arms has distribution F.

We shall give two related examples that show how Theorem 6.1.1 can be applied. These examples are straightforward and easy, but are not meant to suggest any restrictions on the applicability of the theorem. The only real difficulty in applying the theorem to solve a bandit problem with independent arms and geometric discounting arises from difficulties in calculating the various Λ's. The examples avoid such difficulties by choosing distributions F_i for which $\Lambda(F_i, \alpha)$ can be evaluated explicitly.

Example 6.1.1. Consider the $(F_1, F_2; 1/2)$-bandit where the arms are Bernoulli and, given as distributions on the Bernoulli parameters,

$$F_1 = (1-p_1)\delta_{1/3} + p_1\delta_1,$$
$$F_2 = (1-p_2)\delta_{2/3} + p_2\delta_1.$$

From Theorem 5.4.2 or Example 5.4.7 we find

$$\Lambda(F_1, 1/2) = \frac{1+4p_1}{3+2p_1},$$

$$\Lambda(F_2, 1/2) = \frac{2+2p_2}{3+p_2}.$$

Theorem 6.1.1 applies to show that the choice between arms 1 and 2 can be made by deciding which of these two quantities is greater.

Suppose the means of the two arms are equal: $p_1 + (1-p_1)/3 = p_2 + 2(1-p_2)/3$, or $p_1 = (1+p_2)/2$. Then

$$\Lambda(F_1, 1/2) = \frac{1+4p_1}{3+2p_1} = \frac{3+2p_2}{4+p_2} > \frac{2+2p_2}{3+p_2} = \Lambda(F_2, 1/2)$$

and so arm 1 is uniquely optimal. This is consistent with the notion, considered in Theorem 5.0.2, for example, that the best arm to select is the one that is 'riskier' – has larger variance – when the means are the same. □

Example 6.1.2 Consider the $(F_1, \ldots, F_k; 1/2)$-bandit where the k Bernoulli arms are identical: for $i = 1, \ldots, k$,

$$F_i = (1-p)\delta_{1/2} + p\delta_1.$$

As in the previous example we use Theorem 5.4.2 or Example 5.4.7 to conclude that

$$\Lambda(F_i, 1/2) = \frac{1+2p}{2+p}.$$

Optimal strategies in each $(F_i, \lambda; 1/2)$-bandit are clear. Always select the known arm if $\lambda \geqslant (1+2p)/(2+p)$. Always select the unknown arm if $\lambda \leqslant 1/2$. For other λ's select the unknown arm until (if ever) a failure is observed and then switch to the known arm. Easy calculations give, for $i = 1, \ldots, k$.

$$V(F_i, \lambda; 1/2) = \begin{cases} 2\lambda & \text{if } \lambda \geqslant \dfrac{1+2p}{2+p} \\ 2p + 2(1-p)(1+\lambda)/3 & \text{if } \frac{1}{2} < \lambda < \dfrac{1+2p}{2+p} \\ p+1 & \text{if } \lambda \leqslant \frac{1}{2} \end{cases}$$

Clearly (even without using Theorem 6.1.1), it is optimal to select any arm and use it until it fails, switching to a second arm if and when it does; this pattern is repeated until and if each arm fails once. If that happens then it is clear that $\theta_1 = \ldots = \theta_k = 1/2$ and so all continuations are optimal.

We can use Theorem 6.1.1 to calculate the value of this bandit. Applying (6.1.1) and letting

$$\rho^* = \bigvee_{i=1}^{k} \Lambda(F_i, \alpha) = \frac{1+2p}{2+p},$$

we obtain

$$V(F_1, \ldots, F_k; \alpha) = 2\left[\rho^* - \left(\frac{1}{2}\right)^k \int_0^{\rho^*} \left(\frac{\partial}{\partial\lambda} V(F_1, \lambda; 1/2)\right)^k d\lambda\right]$$
$$= 2\left[\rho^* - \int_{1/2}^{\rho^*} \left(\frac{1-p}{3}\right)^k d\lambda\right]$$
$$= \frac{2(1+2p)}{2+p} - \frac{3p}{2+p}\left(\frac{1-p}{3}\right)^k.$$

In particular, the value approaches $2(1+2p)/(2+p)$ as $k \to \infty$. This compares with $\gamma_1 = 2$, a payoff that would be achieved using an arm that gives all successes. While, in the limit as $k \to \infty$, an arm that gives all successes will be found with probability 1 (if $p > 0$), there are losses incurred while looking for it. □

As advertised in the introduction to this chapter, we turn in the next section to a converse of Theorem 6.1.1.

6.2 Necessity of geometric discounting for the Gittins–Jones result

In the introduction to this chapter we claimed that the Gittins–Jones result (Theorem 6.1.1) requires geometric discounting. We now demonstrate this fact assuming **A** is regular (Definition 5.2.1) with $\alpha_1 > 0$. In view of Theorem 5.5.3, these assumptions guarantee the existence of the various $\Lambda(F_i, \mathbf{A})$.

Theorem 6.2.1 Assume **A** is regular with $\alpha_1 > 0$. If, for all $(F_1, \ldots, F_k)$, the optimal initial selections in the $(F_1, \ldots, F_k; \mathbf{A})$-bandit are those i for which

$$\Lambda(F_i, \mathbf{A}) = \bigvee_{j=1}^{k} \Lambda(F_j, \mathbf{A}),$$

then $\mathbf{A} \propto (1, \alpha, \alpha^2, \ldots)$ for some $\alpha \geqslant 0$.

Proof Fix **A** to be regular with $\alpha_1 > 0$. We shall consider certain distributions (F_1, F_2) on Bernoulli parameters θ_1, θ_2. We restrict consideration to pairs (F_1, F_2) for which

$$\Lambda(F_1, \mathbf{A}) = \Lambda(F_2, \mathbf{A}).$$

The hypothesis of the theorem then implies

$$V^{(1)}(F_1, F_2; \mathbf{A}) = V^{(2)}(F_1, F_2; \mathbf{A})$$

for all such (F_1, F_2). So the theorem will follow when we show that this implies $\mathbf{A}$ is geometric.

The parts of the proof are labelled from (i) to (vii). In (i) we define a set of (F_1, F_2) which is indexed by parameters $t \in (0, 1)$ and $p \in [0, 1)$. The functions $\Lambda(F_1, \mathbf{A})$ and $\Lambda(F_2, \mathbf{A})$ are calculated in (ii) and (iii). In (iv) these are set equal by fixing p to be a certain function of t (for the remainder of the proof). The value of selecting arm i and proceeding optimally thereafter, $V^{(i)}(F_1, F_2; \mathbf{A})$, is calculated for $i = 1$ in (v) and for $i = 2$ and all sufficiently small t in (vi). Finally, in (vii) we set $V^{(1)} = V^{(2)}$ which gives the generating function of $\mathbf{A}$, determining it (up to a constant multiple) as being geometric.

(i) For $t \in (0, 1)$ and $p \in [0, 1)$ let

$$F_1 = (1-t)\delta_0 + t\delta_1,$$
$$F_2 = (1-p)\delta_t + p\delta_1.$$

(ii) From Example 5.4.4,

$$\Lambda(F_1, \mathbf{A}) = \frac{t\gamma_1}{t\gamma_1 + (1-t)\alpha_1},$$

where, as previously defined, $\gamma_1 = \sum_{m=1}^{\infty} \alpha_m$.

(iii) From Theorem 5.4.2 or Example 5.4.7,

$$\Lambda(F_2, \mathbf{A}) = \Lambda_\infty(F_2, \mathbf{A}) = \frac{tp + t(1-p)\eta(t)}{tp + (1-p)\eta(t)},$$

where Λ_∞ is defined at (5.4.2) and

$$\eta(t) = \sum_{m=1}^{\infty} (\alpha_m/\gamma_1)t^m,$$

the generating function of $\mathbf{A}/\gamma_1$.

(iv) Setting $\Lambda(F_1, \mathbf{A}) = \Lambda(F_2, \mathbf{A})$ gives

$$\frac{p}{1-p} = \gamma_2\eta(t)/\alpha_1. \tag{6.2.1}$$

Hereafter, for each t we fix p to satisfy (6.2.1).

(v) If arm 1 is selected initially then an optimal continuation is clear: use arm 1 exclusively if it was successful and use arm 2 exclusively thereafter if it was not. It is simple to calculate

$$V^{(1)}(F_1, F_2; \mathbf{A}) = t\gamma_1 + (1-t)[p + (1-p)t]\gamma_2. \tag{6.2.2}$$

(vi) Calculating $V^{(2)}(F_1, F_2; \mathbf{A})$ is substantially more difficult than calculating $V^{(1)}(F_1, F_2; \mathbf{A})$. Should arm 2 fail then it is easy to see, assuming (6.2.1), that arm 1 is then optimal; but an optimal continuation is not clear when it succeeds. We will show that for all sufficiently small t arm 2 continues to be optimal as long as it is successful. Equivalently, we show that staying indefinitely with a successful arm 2 is at least as good as switching after any given number of successes when $\mathbf{A}$ is regular. (We note that staying with a winner is not generally a property of optimal strategies when there are two unknown Bernoulli arms even with regular discounting (cf. Example 4.3.4).)

Define a sequence of strategies τ_n where $\tau_n(\varnothing) = 2$ and arm 2 is continued when following τ_n until it fails or until stage n is reached; in either case arm 1 is then used once and the arm then known to be better used thereafter. Clearly,

$$V^{(2)}(F_1, F_2; \mathbf{A}) = W(F_1, F_2; \mathbf{A}; \tau_\infty) \vee \bigvee_{n=1}^{\infty} W(F_1, F_2; \mathbf{A}; \tau_n). \tag{6.2.3}$$

Straightforward calculations give

$$E_{\tau_n}(Z_m|\theta_1 = 1, \theta_2 = 1) = 1;$$

$$E_{\tau_n}(Z_m|\theta_1 = 1, \theta_2 = t) = \begin{cases} 1 - t^{m-1}(1-t) & \text{if } m \leqslant n \\ 1 & \text{if } m > n; \end{cases}$$

$$E_{\tau_n}(Z_m|\theta_1 = 0, \theta_2 = 1) = \begin{cases} 0 & \text{if } m = n+1 \\ 1 & \text{if } m \neq n+1; \end{cases}$$

$$E_{\tau_n}(Z_m|\theta_1 = 0, \theta_2 = t) = \begin{cases} t & \text{if } m = 1 \text{ or } m > n+1 \\ t - t^{m-1}(1-t) & \text{if } 1 < m < n+1 \\ t - t^n & \text{if } m = n+1. \end{cases}$$

The sum of these quantities weighted by tp, $t(1-p)$, $(1-t)p$, and

$(1-t)(1-p)$, respectively, is $E_{\tau_n}(Z_m|F_1, F_2)$. An easy calculation gives

$$W(F_1, F_2; \mathbf{A}; \tau_n) = \sum_{m=1}^{\infty} \alpha_m E_{\tau_n}(Z_m|F_1, F_2)$$

$$= \gamma_1 - (1-t)^2(1-p)\gamma_2 - (1-t)(1-p)\sum_{m=1}^{n+1} \alpha_m t^{m-1} - (1-t)p\alpha_{n+1}$$

where $\alpha_\infty = 0$. For $n < \infty$,

$$W(F_1, F_2; \mathbf{A}; \tau_\infty) - W(F_1, F_2; \mathbf{A}; \tau_n)$$
$$= -(1-t)(1-p)\sum_{m=n+2}^{\infty} \alpha_m t^{m-1} + (1-t)p\alpha_{n+1}.$$

Dividing this by $(1-t)(1-p)$ and substituting for $p/(1-p)$ using (6.2.1) gives

$$-\sum_{m=n+2}^{\infty} \alpha_m t^{m-1} + \gamma_2 \alpha_{n+1} \eta(t)/\alpha_1,$$

which is no smaller than

$$-t^2\gamma_{n+1} + t\gamma_2\alpha_{n+1}/\gamma_1.$$

That this quantity is nonnegative for all n provided t is sufficiently small follows since, when $\gamma_{n+1} \neq 0$,

$$\frac{\gamma_{n+1}}{\gamma_{n+1}} = 1 - \frac{\gamma_{n+2}}{\gamma_{n+1}}$$

is nondecreasing for regular $\mathbf{A}$.

We have shown that the supremum in (6.2.3) is achieved by $W(F_1, F_2; \mathbf{A})$ for $n = \infty$. Hence

$$V^{(2)}(F_1, F_2; \mathbf{A}) = \gamma_1 - (1-t)^2(1-p)\gamma_2 - (1-t)(1-p)t^{-1}\eta(t)\gamma_1. \tag{6.2.4}$$

(vii) We set $V^{(1)}(F_1, F_2; \mathbf{A}) = V^{(2)}(F_1, F_2; \mathbf{A})$ while continuing to assume $\Lambda(F_1, \mathbf{A}) = \Lambda(F_2, \mathbf{A})$ which is effected by guaranteeing that p and t satisfy (6.2.1). Equating (6.2.2) and (6.2.4) we find that

$$(1-p)\eta(t) = t(1 - [(1-p)(1-t) + p + (1-p)t]\gamma_2/\gamma_1)$$
$$= t\alpha_1/\gamma_1. \tag{6.2.5}$$

From (6.2.1) we find

$$1 - p = \frac{\alpha_1}{\alpha_1 + \gamma_2 \eta(t)};$$

substituting into (6.2.5) we obtain

$$\gamma_1 \eta(t) = \frac{t\alpha_1}{1 - t(\gamma_2/\gamma_1)},$$

the generating function of the geometric sequence. We conclude from the uniqueness of generating functions that $\mathbf{A} = \alpha_1(1, \alpha, \alpha^2, \ldots)$ where $\alpha = \gamma_2/\gamma_1$. □

Since $\Lambda(F_i, \mathbf{A})$ is not defined for all $\mathbf{A}$, Theorem 6.2.1 will not generalize to include all nonregular discount sequences and sequences for which $\alpha_1 = 0$. We conjecture, however, that an analogous result holds in complete generality. Such a result can be stated along the lines of transitivity of arms as discussed in the introduction to this chapter. Fix $\mathbf{A}$ and for all k-armed bandits $(F_1, \ldots, F_k; \mathbf{A})$ suppose that $\Delta(F_1, F_2; \mathbf{A}) > 0$ and $\Delta(F_2, F_3; \mathbf{A}) > 0$ imply $\Delta(F_1, F_3; \mathbf{A}) > 0$. Then, we conjecture, $\mathbf{A}$ is geometric.

References

Gittins, J. C. and Jones, D. M. (1974) A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (eds J. Gani *et al.*), pp. 241–266, North-Holland, Amsterdam.

Whittle, P. (1982) *Optimization Over Time: Dynamic Programming and Stochastic Control*, Vol. I, Wiley, New York.

CHAPTER 7

Two independent Bernoulli arms; uniform discounting

In this chapter we assume that $k = 2$ and the arms are Bernoulli. The arms are independent initially, and therefore also henceforth. The discount sequence **A** has horizon n and is uniform: $\alpha_1 = \ldots = \alpha_n = 1$ and $\alpha_{n+1} = \alpha_{n+2} = \ldots = 0$. Such uniform discounting has been considered extensively through examples in the first five chapters of this book, and in the literature generally. The objective implicit in uniform discounting is to maximize the expected sum of the first n observations.

When one of the arms has known probability of success, the arms are obviously independent. So this chapter generalizes those parts of Chapter 5 that deal with Bernoulli arms and a uniform discount sequence.

Some of the results in this chapter are in Berry (1972). Section 7.1 specializes aspects of Chapter 4 to the uniform-discounting setting. The use of the recursion formula developed in Section 4.2 to determine optimal strategies is discussed in Section 7.2; in particular, this formula is applied to the easy case $n = 2$. In Section 7.3 the horizon is general but the relationship between F_1 and F_2 is special; each is taken to be some posterior distribution resulting from a common underlying distribution F. A central feature of the discussion in Section 7.3 is a necessary and sufficient condition for $F_1 \overset{n}{\gtrsim}{}^{1} F_2$.

7.1 Preliminaries

In this chapter the discount sequence **A** is completely described by the single parameter n, the horizon. We therefore use n in place of **A** throughout. And, consistent with this convention, $\mathbf{A}^{(j)}$ is replaced by

$n-j$ since the horizon is reduced by j at stage $j+1$. For example, we write $W(F_1, F_2; n; \tau)$ for the worth of strategy τ.

The results of Chapter 4 hold in the special setting of this chapter. For example, Lemma 4.2.1 gives the following recursion: for $n = 1$,

$$\Delta(F_1, F_2; 1) = E(\theta_1 | F_1) - E(\theta_2 | F_2);$$

for $n \geqslant 2$,

$$\begin{aligned}\Delta(F_1, F_2; n) = {} & E(\theta_1 | F_1)\Delta^+(\sigma F_1, F_2; n-1) \\ & + E(1-\theta_1 | F_1)\Delta^+(\varphi F_1, F_2; n-1) \\ & - E(\theta_2 | F_2)\Delta^-(F_1, \sigma F_2; n-1) \\ & - E(1-\theta_2 | F_2)\Delta^-(F_1, \varphi F_2; n-1). \qquad (7.1.1)\end{aligned}$$

Thus, for $n \geqslant 2$,

$$\begin{aligned}\Delta(F_1, F_2; n) \leqslant {} & E(\theta_1 | F_1)\Delta^+(\sigma F_1, F_2; n-1) \\ & + E(1-\theta | F_1)\Delta^+(\varphi F_1, F_2; n-1).\end{aligned}$$

The stay-with-a-winner rule, Theorem 4.3.8, is now easily expressed:

Corollary 7.1.1 Suppose $n \geqslant 2$ and the support of either F_1 or F_2 consists of more than one point. If $\Delta(F_1, F_2; n) \geqslant 0$ then $\Delta(\sigma F_1, F_2; n-1) > 0$.

An easy consequence of this result is given next. It says that if the mean of arm 1 is sufficiently large (depending on arm 2) then arm 1 is optimal. In particular, it says that if arm 1 would have a larger probability of success than arm 2 even if arm 2 produces $n-1$ successes in $n-1$ trials, then arm 1 is optimal initially. While the bound is very crude, it can be instructive.

Corollary 7.1.2 If

$$E(\theta_1 | F_1) \geqslant E(\theta_2 | \sigma^{n-1} F_2) \qquad (7.1.2)$$

then arm 1 is optimal initially.

Proof If F_2 is supported by one point then $E(\theta_2 | \sigma^{n-1} F_2)$ equals $E(\theta_2 | F_2)$ and the result follows from Theorem 5.0.2. The result is also immediate if $n = 1$.

Suppose F_2 is not supported by one point and $n > 1$. Then if arm 2

were optimal initially it would, by Corollary 7.1.1, be uniquely optimal from stage 2 onwards as long as successes were obtained. In particular, it would be uniquely optimal at stage n if $n-1$ successes were obtained. This contradicts (7.1.2). (Moreover, this argument shows arm 1 to be uniquely optimal initially in case F_1 is not supported by one point and $n > 1$.) □

In the next section we give a complete characterization of optimal strategies when $n = 2$, carrying out the detailed calculations of $\Delta(F_1, F_2; n)$. The case $n = 2$ is relatively easy of course, but calculating Δ is still not a trivial matter.

7.2 Optimal selections when the horizon is two

The recursion given in (7.1.1) allows for explicit calculation of Δ. We will describe this procedure, specializing to the case $n = 2$ to obtain simple explicit formulas. The complexity of the calculations for large n will be apparent.

The calculation of $\Delta(F_1, F_2; n)$ requires $\Delta^+(\sigma F_1, F_2; n-1)$, $\Delta^+(\varphi F_1, F_2; n-1)$, $\Delta^-(F_1, \sigma F_2; n-1)$, and $\Delta^-(F_1, \varphi F_2; n-1)$. Since there are two possibilities – positive and zero – for each of these four quantities, there might be as many as sixteen different formulas for $\Delta(F_1, F_2; n)$ in terms of $\Delta(\cdot, \cdot; n-1)$. But eight of these cases are ruled out by Theorem 4.3.6 and Lemma 4.3.5; the remaining eight will be given below. The cases will be identified by quadruples; for example, $(+0++)$ indicates that $\Delta^+(\sigma F_1, F_2; n-1)$, $\Delta^-(F_1, \sigma F_2; n-1)$, and $\Delta^-(F_1, \varphi F_2; n-1)$ are positive and $\Delta^+(\varphi F_1, F_2; n-1) = 0$.

Since $\Delta(\cdot, \cdot; 1)$ is just the difference of expectations, the eight possible formulas can be written simply and explicitly as follows (the designations F_1 and F_2 have been suppressed when expectations involving θ_1 and θ_2 are with respect to these distributions):

$$\Delta(F_1, F_2; 2) =$$

$(+000)$ $\quad E(\theta_1^2) - E(\theta_1)E(\theta_2)$ $\quad$ if $E(\theta_1|\sigma F_1) > E(\theta_2) \geq E(\theta_1|\varphi F_1)$ & $E(\theta_1) \geq E(\theta_2|\sigma F_2)$

$(++00)$ $\quad E(\theta_1) - E(\theta_2)$ $\quad$ if $E(\theta_1|\varphi F_1) > E(\theta_2)$ & $E(\theta_1) \geq E(\theta_2|\sigma F_2)$

$$
\begin{array}{lll}
(+++0) & E(\theta_1)-E(\theta_2)-E(\theta_2^2) & \text{if } E(\theta_1|\varphi F_1)>E(\theta_2) \\
 & \quad +E(\theta_1)E(\theta_2) & \& \; E(\theta_2|\sigma F_2)>E(\theta_1) \\
(+0+0) & E(\theta_1^2)-E(\theta_2^2) & \text{if } E(\theta_1|\sigma F_1)>E(\theta_2)\geqslant E(\theta_1|\varphi F_1) \\
 & & \& \; E(\theta_2|\sigma F_2)>E(\theta_1)\geqslant E(\theta_2|\varphi F_2) \\
(+0++) & E(\theta_1)-E(\theta_2)+E(\theta_1^2) & \text{if } E(\theta_1|\sigma F_1)>E(\theta_2) \\
 & \quad -E(\theta_1)E(\theta_2) & \& \; E(\theta_2|\varphi F_2)>E(\theta_1) \\
(00++) & E(\theta_1)-E(\theta_2) & \text{if } E(\theta_2)\geqslant E(\theta_1|\sigma F_1) \\
 & & \& \; E(\theta_2|\varphi F_2)>E(\theta_1) \\
(00+0) & E(\theta_1)E(\theta_2)-E(\theta_2^2) & \text{if } E(\theta_2)\geqslant E(\theta_1|\sigma F_1) \\
 & & \& \; E(\theta_2|\sigma F_2)>E(\theta_1)\geqslant E(\theta_2|\varphi F_2) \\
(0000) & 0 & \text{if } E(\theta_2)\geqslant E(\theta_1|\sigma F_1) \\
 & & \& \; E(\theta_1)\geqslant E(\theta_2|\sigma F_2).
\end{array}
\tag{7.2.1}
$$

The expectations involving σF_i and φF_i are simply related to those involving F_i:

$$
\begin{aligned}
E(\theta_i|\sigma F_i) &= E(\theta_i^2)/E(\theta_i), \\
E(\theta_i|\varphi F_i) &= E(\theta_i(1-\theta_i))/E(1-\theta_i).
\end{aligned}
$$

From Lemma 4.3.5 and Theorem 4.3.6, we see that the case (0 0 0 0) occurs only when F_1 and F_2 are the same one-point distribution.

The case $(+0++)$ in (7.2.1) was used for the calculation of $\Delta(F_1, F_2; \mathbf{A})$ in Example 6.0.1. We now give a more comprehensive example.

Example 7.2.1 We consider $n = 2$ and the special case where F_1 and F_2 are beta distributions:

$$
F_i(\mathrm{d}u) \propto u^{a_i-1}(1-u)^{b_i-1}\,\mathrm{d}u, \tag{7.2.2}
$$

where $a_i > 0$ and $b_i > 0, i = 1,2$. We denote $a_i + b_i$ by I_i (which stands for information for the same reason that I was used to stand for information in Section 5.6). We write $\Delta(a_1, b_1, a_2, b_2; 2)$ in place of $\Delta(F_1, F_2; 2)$. To determine the sign of Δ(and thus the optimal initial selection) we shall find the set of (a_1, b_1, a_2, b_2) for which Δ is 0. When it equals 0, and so both arms are optimal initially, $\Delta^+(\sigma F_1, F_2; 2)$ and

$\Delta^-(F_1, \sigma F_2; 2)$ are positive by Corollary 7.1.1. Therefore, $(+++0)$, $(+0+0)$, and $(+0++)$ are the only cases possible in (7.2.1). Straightforward calculations show that in case $(+++0)$, Δ cannot be 0 when $I_1 \leqslant I_2$. Similarly, in case $(+0++)$, Δ cannot be 0 when $I_1 \geqslant I_2$; we assume $I_1 \geqslant I_2$.

To use (7.2.1) we need the following:

$$E(\theta_i) = a_i/I_i,$$
$$E(\theta_i|\sigma F_i) = (a_i+1)/(I_i+1),$$
$$E(\theta_i|\varphi F_i) = a_i/(I_i+1).$$

Evaluating cases $(+++0)$ and $(+0+0)$, we find that the locus of points for which

$$\Delta(a_1, b_1, a_2, b_2; 2) = 0 \tag{7.2.3}$$

is given as follows:

$$\begin{aligned} &\frac{a_1}{I_1} - \frac{a_2}{I_2}\left[1 + \frac{a_2+1}{I_2+1} - \frac{a_1}{I_1}\right] = 0 \qquad \text{for } \frac{a_2}{I_2} \leqslant \frac{a_1}{I_1+1} \\ &\frac{a_1}{I_1}\frac{a_1+1}{I_1+1} - \frac{a_2}{I_2}\frac{a_2+1}{I_2+1} = 0 \qquad \text{for } \frac{a_2}{I_2} \geqslant \frac{a_1}{I_1+1}. \end{aligned} \tag{7.2.4}$$

The solutions of the two equations in (7.2.4) intersect at the point

$$a_1^* = \frac{(I_1+1)(I_1-I_2-1)}{I_1+I_2+1}, \qquad a_2^* = \frac{I_2(I_1-I_2-1)}{I_1+I_2+1}, \tag{7.2.5}$$

with the first equation in (7.2.4) applying for values of $(a_1, a_2) \leqslant (a_1^*, a_2^*)$. We are only interested in $0 < a_i < I_i$, $i = 1,2$. The conditions $a_i < I_i$, $i = 1, 2$, are satisfied by (a_1^*, a_2^*). The conditions $a_1^* > 0$ are not satisfied if $I_1 \leqslant I_2 + 1$; in this case the second equation at (7.2.4) gives the points (a_1, a_2) for which $\Delta = 0$. This curve has a corner at (a_1^*, a_2^*) given by (7.2.5) if $I_1 > I_2 + 1$.

To show graphically the dependence of (7.2.4) on the quantities of information I_1 and I_2, we fix the ratio $I_1/I_2 = 3$. The solution is shown in Figure 7.1 for $I_2 = 1$ (and therefore, $I_1 = 3$). The corner point is $(a_1^*, a_2^*) = (4/5, 1/5)$, though it's not discernibly a corner in the figure. The dashed line in Figure 7.1 is the set of points for which the two means are equal, that is, $a_2 = a_1/3$.

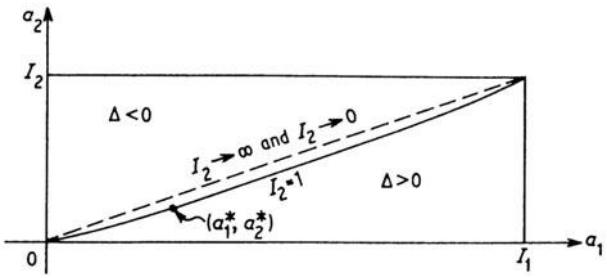


Fig. 7.1 *The sign of* $\Delta(a_1, I_1 - a_1, a_2, I_2 - a_2;\ 2)$ *for* $I_1 = 3I_2$. *The curve indicates the set for which* $\Delta = 0$ *when* $I_2 = 1$.

This dashed line has additional significance for the issue at hand: it is the set of (a_1, a_2) for which Δ is 0 for $I_i \to \infty$ and also for $I_i \to 0$ (with $I_1/I_2 = 3$). That is, it is the limit of (7.2.4) for both $I_2 \to \infty$ and $I_2 \to 0$ with $I_1/I_2 = 3$. For both I_1 and I_2 large, there is little information to be gained from an observation relative to the information already available; so only the means are relevant. On the other hand, for very small I_i, essentially complete information is gained by the first observation on *either* arm – and so, again, only the means matter.

The curve for which $\Delta = 0$ that is drawn in Figure 7.1 for $I_1 = 3$ and $I_2 = 1$ is approximately the lower envelope of such curves with $I_1/I_2 = 3$ and $0 < I_2 < \infty$. It approximates (7.2.4) for moderate values of I_2 (say between 1/10 and 10).

The region above the $\Delta = 0$ curve in Figure 7.1 is labelled '$\Delta < 0$' and indicates that arm 2 is optimal for these points. Arm 1 is optimal for points in the '$\Delta > 0$' region in the figure. □

Conditions for $\Delta \geqslant 0$ show that arm 1 is optimal initially if and only if

$$[E(\theta_1^2) - E(\theta_1)E(\theta_2)] \vee [E(\theta_1) - E(\theta_2)] \geqslant [E(\theta_2^2) - E(\theta_2)E(\theta_1)] \vee [E(\theta_2) - E(\theta_1)],$$

a result given by Bradt, Johnson, and Karlin (1956, Lemma 3.1).

The relatively easy case $n = 2$ treated in this section is actually somewhat cumbersome. As indicated in Chapter 2, complete solutions are increasingly complicated for larger horizons. In the next

section we partially determine the optimal initial selection by finding the sign of Δ for certain pairs (F_1, F_2).

7.3 Arms with identical underlying priors

Arm 1 is optimal initially in the $(F_1, F_2; n)$-bandit if F_1 is strongly to the right of F_2 or if the weaker condition $F_1 \overset{n-1}{\gtrsim} F_2$ holds (Corollary 4.3.7). In this section we find conditions for these relations when (F_1, F_2) belongs to the class of distributions in which the arms are comparable in the sense that information concerning them can be viewed as having arisen from a common, but arbitrary, distribution. That is, some time in the 'past' they had the same distribution, say F. We develop this notion next.

Let F be a distribution measure on $[0, 1]$ such that $F(\{1, 0\}) = 0$ and F is not supported by one point. Suppose that F is the prior distribution of a Bernoulli parameter θ and that a successes and b failures are observed. Then the posterior distribution of θ is

$$\frac{x^a(1-x)^b F(\mathrm{d}x)}{\int_{[0,1]} u^a(1-u)^b F(\mathrm{d}u)};$$

that is,

$$(\sigma^a \varphi^b F)(\mathrm{d}x) \propto x^a(1-x)^b F(\mathrm{d}x). \tag{7.3.1}$$

We assume that both F_1 and F_2 can be written as in (7.3.1) for some F which is common to both arms; so

$$F_i = \sigma^{a_i} \varphi^{b_i} F, \qquad i = 1, 2, \tag{7.3.2}$$

where a_i and b_i are regarded as known but arbitrary.

Two natural generalizations can be made. We permit a_i and b_i to be positive real numbers, not necessarily integers. We require only proportionality in (7.3.1). So a positive measure F that is not a probability measure may be used provided that

$$\int_{[0,1]} u^a(1-u)^b F(\mathrm{d}u) < \infty$$

for $a, b > 0$.

The most important example in this setting is given by

$F(\mathrm{d}u) = u^{-1}(1-u)^{-1}\,\mathrm{d}u$. Then F_i is a beta distribution with parameters a_i and b_i.

Theorem 7.3.2 below gives sufficient conditions for the relationship of strongly to the right to hold. For its proof we first show the corresponding result with 'to the right' in place of 'strongly to the right'.

Lemma 7.3.1 Let F_1 and F_2 be given by (7.3.2). Then F_1 is to the right of F_2 if $a_1 \geqslant a_2$ and $b_1 \leqslant b_2$. The relationship is strict if $a_1 > a_2$ or $b_1 < b_2$.

Proof Suppose that $a_1 \geqslant a_2$, $b_1 \leqslant b_2$, and at least one of these inequalities is strict. We want to show that $F_1[x, 1] \geqslant F_2[x, 1]$ for all $x \in (0, 1)$ with strict inequality for at least one x. The weak inequality is obvious if $F(0, x) = 0$, so we assume $F(0, x) > 0$ in which case $F_i(0, x) > 0$ for $i = 1, 2$. The inequality of interest can be rewritten as

$$\frac{F_1[x, 1]}{F_1[0, x)} \geqslant \frac{F_2[x, 1]}{F_2[0, x)}, \tag{7.3.3}$$

which follows from

$$\begin{aligned}
&\frac{\displaystyle\int_{[x,\,1]} u^{a_1}(1-u)^{b_1}\,F(\mathrm{d}u)}{\displaystyle\int_{[0,\,x)} u^{a_1}(1-u)^{b_1}\,F(\mathrm{d}u)} \\
&\geqslant \frac{x^{a_1-a_2}(1-x)^{b_1-b_2}\displaystyle\int_{[x,\,1]} u^{a_2}(1-u)^{b_2}\,F(\mathrm{d}u)}{x^{a_1-a_2}(1-x)^{b_1-b_2}\displaystyle\int_{[0,\,x)} u^{a_2}(1-u)^{b_2}\,F(\mathrm{d}u)}
\end{aligned} \tag{7.3.4}$$

Moreover, the inequality in (7.3.4), and therefore in (7.3.3), is strict if x is chosen so that the support of F contains a number larger than x and a number smaller than x. □

Theorem 7.3.2 Let F_1 and F_2 be given by (7.3.2). Then F_1 is strongly to the right of F_2 if $a_1 \geqslant a_2$ and $b_1 \leqslant b_2$. The relationship is strict if $a_1 > a_2$ or $b_1 < b_2$.

Proof Assume $a_1 \geqslant a_2$ and $b_1 \leqslant b_2$. For any nonnegative integers s and f, $a_1 + s \geqslant a_2 + s$ and $b_1 + f \leqslant b_2 + f$; so, by Lemma 7.3.1, $\sigma^s\varphi^f F_1$

is to the right $\sigma^s\varphi^f F_2$. Hence F_1 is strongly to the right of F_2. The strictness of this relationship in case $a_1 > a_2$ or $b_1 < b_2$ follows from Lemma 7.3.1. □

Remark In case $F(du) = u^{-1}(1-u)^{-1}\,du$, the converse of Theorem 7.3.2 holds. However, the converse does not hold in general; to see this let F be a symmetric two-point distribution. □

On the basis of Theorem 7.3.2, arm 1 is seen to be optimal for certain (a_1, b_1, a_2, b_2), arm 2 is optimal for others, and the theorem does not indicate an optimal selection for the rest.

Example 7.3.1 Let $F(du) = u^{-1}(1-u)^{-1}\,du$ and fix $a_1 + b_1 = I_1$ and $a_2 + b_2 = I_2$ with $I_1 = 3I_2$. So F_1 and F_2 are the same distributions considered in Figure 7.1. The regions in the (a_1, a_2)-plane where Theorem 7.3.2 indicates an optimal initial selection are shown in Figure 7.2.

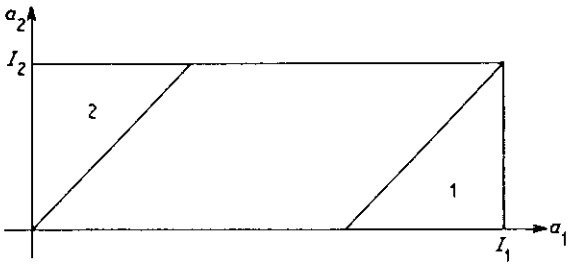


Fig. 7.2 *Regions where the indicated arms are known to be optimal on the basis of Theorem* 7.3.2.

If I_2/I_1 were closer to 1, the region of uncertainty would be smaller. If $I_2 = I_1$, then one (or both) of F_1 or F_2 is strongly to the right of the other and there is no uncertainty region. □

Theorem 7.3.2 gives conditions for optimality that do not depend on n. The condition given by the next theorem depends on n, and the result is stronger. For fixed I_1, I_2, and n the result will allow us to make the region of uncertainty in Figure 7.2 smaller by focusing on the relation $\overset{n-1}{\gtrsim}$ rather than on 'strongly to the right.'

Theorem 7.3.3 Suppose $I_1 > I_2$. For $n > 1$, a necessary and sufficient condition for $F_1 \overset{n-1}{\gtrsim} F_2$ is

$$E(\theta_1 | \sigma^{n-1} F_1) \geqslant E(\theta_2 | \sigma^{n-1} F_2). \tag{7.3.5}$$

Remark Another way to write (7.3.5) is

$$\frac{E(\theta_1^n | F_1)}{E(\theta_1^{n-1} | F_1)} \geqslant \frac{E(\theta_2^n | F_2)}{E(\theta_2^{n-1} | F_2)}. \qquad \square$$

For the proof we need the following lemma.

Lemma 7.3.4 Suppose $I_1 > I_2$ and $E(\theta_1 | F_1) = E(\theta_2 | F_2)$. Then $E(\theta_1 | \sigma F_1) < E(\theta_2 | \sigma F_2)$ and $E(\theta_1 | \varphi F_1) > E(\theta_2 | \varphi F_2)$.

Proof The second implication will follow from the first with the roles of success and failure exchanged; we will show the first.

Fix a_2 and b_2 and write

$$E(\theta_1 | F_1) = \frac{f_1(a_1, b_1)}{f_0(a_1, b_1)}$$

and

$$E(\theta_1 | \sigma F_1) = \frac{f_2(a_1, b_1)}{f_1(a_1, b_1)}, \tag{7.3.6}$$

where

$$f_j(a_1, b_1) = \int_{[0,1]} u^{a_1 + j} (1-u)^{b_1} F(\mathrm{d}u).$$

For $b_1 \geqslant b_2$, let a_1 be the function of b_1 for which the means of the two arms are equal, that is,

$$\frac{f_1(a_1, b_1)}{f_0(a_1, b_1)} = E(\theta_2 | F_2) \tag{7.3.7}$$

Clearly a_1 is a well defined strictly increasing function of b_1 that equals a_2 when $b_1 = b_2$ and approaches ∞ as $b_1 \to \infty$. Implicit differentiation in (7.3.7) gives

$$\frac{\mathrm{d}a_1}{\mathrm{d}b_1} = \frac{f_1 \mathbf{D}_2 f_0 - f_0 \mathbf{D}_2 f_1}{f_0 \mathbf{D}_1 f_1 - f_1 \mathbf{D}_1 f_0},$$

where $\mathbf{D}_i$ indicates differentiation with respect to the ith variable.

Differentiation of (7.3.6) assuming (7.3.7) gives

$$f_1[f_0\mathbf{D}_1f_1 - f_1\mathbf{D}_1f_0]\frac{\mathrm{d}}{\mathrm{d}b_1}E(\theta_1|\sigma F_1)$$
$$= f_2[\mathbf{D}_1f_0\mathbf{D}_2f_1 - \mathbf{D}_2f_0\mathbf{D}_1f_1]$$
$$+ \mathbf{D}_1f_2[f_1\mathbf{D}_2f_0 - f_0\mathbf{D}_2f_1] + \mathbf{D}_2f_2[f_0\mathbf{D}_1f_1 - f_1\mathbf{D}_1f_0]. \quad (7.3.8)$$

Obviously $f_1 > 0$, $f_2 > 0$, and easy calculations shown that $\mathbf{D}_1f_2 < 0$ and $\mathbf{D}_2f_2 < 0$. Writing

$$\pi(\mathrm{d}u) = u^{a_1}(1-u)^{b_1}F(\mathrm{d}u),$$

we have

$$f_0\mathbf{D}_1f_1 = \int_{[0,1]}\pi(\mathrm{d}u)\int_{[0,1]}v\log v\ \pi(\mathrm{d}v)$$
$$= \int_{[0,1]\times[0,1]}v\log v\,(\pi\times\pi)(\mathrm{d}(u,v))$$

and

$$f_1\mathbf{D}_1f_0 = \int_{[0,1]\times[0,1]}v\log u\,(\pi\times\pi)(\mathrm{d}(u,v)).$$

Hence,

$$f_0\mathbf{D}_1f_1 - f_1\mathbf{D}_1f_0 = \int_{[0,1]\times[0,1]}v\log(v/u)\,(\pi\times\pi)(\mathrm{d}(u,v))$$
$$= \int_{u<v}v\log(v/u)\,(\pi\times\pi))(\mathrm{d}(u,v))$$
$$+ \int_{u>v}v\log(v/u)\,(\pi\times\pi)(\mathrm{d}(u,v))$$
$$= \int_{u<v}v\log(v/u)\,(\pi\times\pi)(\mathrm{d}(u,v))$$
$$+ \int_{v>u}u\log(u/v)\,(\pi\times\pi)(\mathrm{d}(u,v))$$
$$= \int_{u<v}(v-u)\log(v/u)\,(\pi\times\pi)(\mathrm{d}(u,v)) > 0.$$

Similar calculations show

$$\mathbf{D}_1 f_0 \mathbf{D}_2 f_1 - \mathbf{D}_2 f_0 \mathbf{D}_1 f_1 < 0,$$
$$f_1 \mathbf{D}_2 f_0 - f_0 \mathbf{D}_2 f_1 > 0,$$
$$f_0 \mathbf{D}_1 f_1 - f_1 \mathbf{D}_1 f_0 > 0.$$

From (7.3.8) we conclude that $\dfrac{\mathrm{d}}{\mathrm{d}b} E(\theta_1 | \sigma F_1) < 0$. □

Proof of Theorem 7.3.3. Necessity follows by definition. We prove sufficiency by induction starting at $n = 2$. Suppose that $E(\theta_1 | \sigma F_1) \geqslant E(\theta_2 | \sigma F_2)$, but that $E(\theta_1 | F_1) \leqslant E(\theta_2 | F_2)$. By increasing a_1 (or perhaps keeping it the same) we can achieve both $E(\theta_1 | \sigma F_1) \geqslant E(\theta_2 | \sigma F_2)$ and $E(\theta_1 | F_1) = E(\theta_2 | F_2)$, contradicting Lemma 7.3.4. Therefore, $E(\theta_1 | \sigma F_1) \geqslant E(\theta_2 | \sigma F_2)$ implies $E(\theta_1 | F_1) > E(\theta_2 | F_2)$. This in turn implies $E(\theta_1 | \varphi F_1) \geqslant E(\theta_2 | \varphi F_2)$. For suppose that $E(\theta_1 | \varphi F_1) < E(\theta_2 | \varphi F_2)$. By increasing b_1 we can obtain both $E(\theta_1 | F_1) = E(\theta_2 | F_2)$ and $E(\theta_1 | \varphi F_1) < E(\theta_2 | \varphi F_2)$. This also contradicts Lemma 7.3.4.

We have shown that $E(\theta_1 | \sigma F_1) \geqslant E(\theta_2 | \sigma F_2)$ implies both $E(\theta_1 | F_1) > E(\theta_2 | F_2)$ and $E(\theta_1 | \varphi F_1) \geqslant E(\theta_2 | \varphi F_2)$ and, hence $F_1 \overset{1}{\succ} F_2$.

Let $n > 2$ and assume $E(\theta_1 | \sigma^{m-1} F_1) \geqslant E(\theta_2 | \sigma^{m-1} F_2)$ implies $F_1 \overset{m-1}{\succ} F_2$ for $m < n$. Assume $E(\theta_1 | \sigma^{n-1} F_1) \geqslant E(\theta_2 | \sigma^{n-1} F_2)$. To show that $E(\theta_1 | \sigma^s \varphi^f F_1) \geqslant E(\theta_2 | \sigma^s \varphi^f F_2)$ for all (s, f) with $s + f \leqslant n - 1$, we first prove it when $s \geqslant 1$. Then we prove it for $s = 0$, obtaining, in particular, the desired strict inequality when $s = 0 = f$.

Since $E(\theta_1 | \sigma^{n-1} F_1) \geqslant E(\theta_2 | \sigma^{n-1} F_2)$ can be written as $E(\theta_1 | \sigma^{n-2} \sigma F_1) \geqslant E(\theta_2 | \sigma^{n-2} \sigma F_2)$, application of the induction hypothesis to σF_1 and σF_2 with $m = n - 1$ gives $\sigma F_1 \overset{n-2}{\succ} \sigma F_2$. Hence, $E(\theta_1 | \sigma^s \varphi^f \sigma F_1) \geqslant E(\theta_2 | \sigma^s \varphi^f \sigma F_2)$ for $s + f \leqslant n - 2$; that is, $E(\theta_1 | \sigma^s \varphi^f F_1) \geqslant E(\theta_2 | \sigma^s \varphi^f F_2)$ when $s \geqslant 1$ and $s + f \leqslant n - 1$.

In particular, $E(\theta_1 | \sigma \varphi^f F_1) \geqslant E(\theta_2 | \sigma \varphi^f F_2)$ for $f \leqslant n - 2$. The induction hypothesis applied to $\varphi^f F_1$ and $\varphi^f F_2$ for $m = 2$ gives $E(\theta_1 | \varphi^f F_1) > E(\theta_2 | \varphi^f F_2)$ and $E(\theta_1 | \varphi^{f+1} F_1) \geqslant E(\theta_2 | \varphi^{f+1} F_2)$, for $f \leqslant n - 2$. Hence, $E(\theta_1 | \varphi^f F_1) \geqslant E(\theta_2 | \varphi^f F_2)$ for $f \leqslant n - 1$ with strict inequality if $f = 0$. □

Remark A little more care in the preceding proof gives $E(\theta_1 | \sigma^s \varphi^f F_1) > E(\theta_2 | \sigma^s \varphi^f F_2)$ whenever $s + f \leqslant n - 1$ and $s < n - 1$. □

The following complementary result follows immediately by interchanging the roles of successes and failures.

Corollary 7.3.5 Suppose $I_1 > I_2$. For $n > 1$, a necessary and sufficient condition for $F_2 \overset{n-1}{\succ} F_1$ is

$$E(\theta_2 | \varphi^{n-1} F_2) \geqslant E(\theta_1 | \varphi^{n-1} F_1).$$

We will now use this corollary and Theorem 7.3.3 to obtain a more complete picture than that provided by Figure 7.2.

Example 7.3.2 As in Example 7.3.1, suppose $I_1 = 3I_2$ and $F(\mathrm{d}u) = u^{-1}(1-u)^{-1}\,\mathrm{d}u$. To obtain formulas with simple appearance, we specialize to $I_1 = 3$, $I_2 = 1$.

From Theorem 7.3.3 we see that the set of (a_1, b_1, a_2, b_2) for which

$$\frac{a_1 + (n-1)}{a_1 + (n-1) + b_1} = \frac{a_2 + (n-1)}{a_2 + (n-1) + b_2}$$

is of interest. Using $b_1 = 3 - a_1$ and $b_2 = 1 - a_2$, we obtain

$$a_1 = \frac{n+2}{n} a_2 + \frac{2n-2}{n}, \qquad 0 \leqslant a_2 \leqslant 1.$$

For various values of n, these line segments are shown in the lower right half of Figure 7.3. For a particular value of n, arm 1 is optimal initially if (a_1, a_2) is either on or below the corresponding line segment.

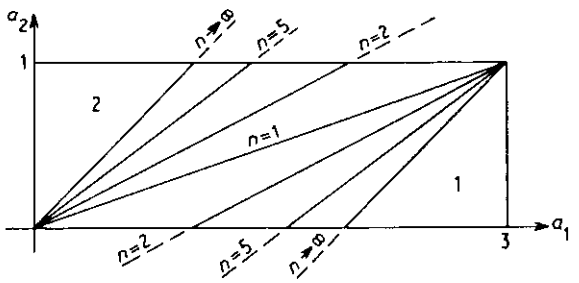


Fig. 7.3 $I_1 = 3$, $I_2 = 1$. *Regions where the indicated arms are known to be optimal on the basis of Theorem* 7.3.3 *and Corollary* 7.3.5.

Similarly, a region where arm 2 is optimal initially can be identified from Corollary 7.3.5. The boundaries are given by

$$a_1 = \frac{n+2}{n} a_2, \qquad 0 \leqslant a_2 \leqslant 1$$

and, for various values of n, these are shown in the upper left half of Figure 7.3. As expected, the picture becomes that of Figure 7.2 as $n \to \infty$.

It is somewhat surprising that the elementary Corollary 7.1.2 can be used to enlarge the region where arm 2 is known to be optimal. Setting $E(\theta_2 | F_2) = E(\theta_1 | \sigma^{n-1} F_1)$, we obtain

$$a_2 = \frac{a_1 + (n-1)}{n+2},$$

or

$$a_1 = (n+2)a_2 - (n-1), \qquad \frac{n-1}{n+2} \leqslant a_2 \leqslant 1.$$

Using this in combination with Figure 7.3, we obtain Figure 7.4. For each $n > 1$ there are two curves (one of which is a line segment and the other of which is the union of two line segments). For (a_1, a_2) on or below the lower of the two, arm 1 is optimal. For (a_1, a_2) on or above the higher of the two, arm 2 is optimal.

On the other hand, Corollary 7.1.2 cannot be used to increase the region where arm 1 is known to be optimal. Figures 7.5 and 7.6 are the analogues of Figure 7.4 for other values of I_1 and I_2 satisfying $I_2 = 3I_2$. □

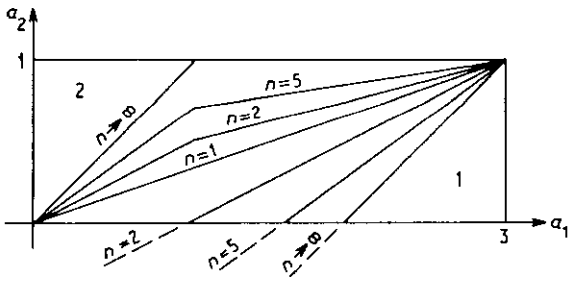


Fig. 7.4 $I_1 = 3, I_2 = 1$. *Regions where the indicated arms are known to be optimal on the basis of Theorem* 7.3.3, *Corollary* 7.3.5, *and Corollary* 7.1.2.

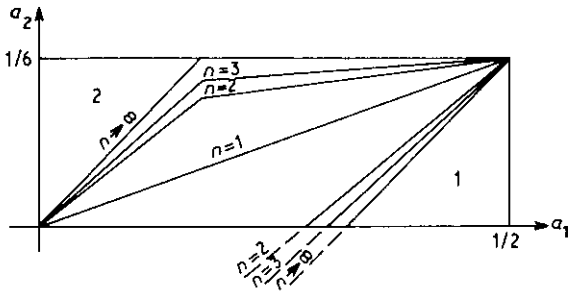


Fig. 7.5 $I_1 = 1/2, I_2 = 1/6$. *Regions where the indicated arms are known to be optimal on the basis of Theorem* 7.3.3, *Corollary* 7.3.5, *and Corollary* 7.1.2.

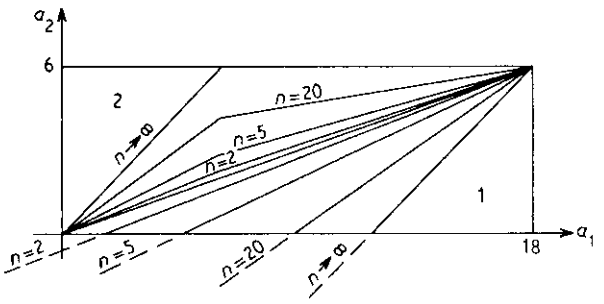


Fig. 7.6 $I_1 = 18, I_2 = 6$. *Regions where the indicated arms are known to be optimal on the basis of Theorem* 7.3.3, *Corollary* 7.3.5, *and Corollary* 7.1.2.

We would like theorems that provide better bounds than those given by Theorem 7.3.3 and Corollaries 7.1.2 and 7.3.5. We conjecture that arm 2 is optimal on or above the main diagonal, for there the arm about which there is less information also has the larger expectation. (Cf. our discussion concerning Joshi (1975) in the Annotated Bibliography.)

We have focused on the initial selection in this chapter. This is entirely appropriate since optimal strategies are composed of initial selections, but for different bandits. While the figures shown in this section may be appropriate in the original bandit, different figures will

be appropriate in the other bandits. At the second stage of the original bandit, for example, the rectangle in the appropriate figure will be lengthened by one unit in the dimension corresponding to the arm chosen, the point (a_1, a_2) will move or not one unit in the same direction according as the result obtained is success or failure, and n will be decreased by 1.

References

Berry, D. A. (1972) A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43**: 871–897.

Bradt, R. N., Johnson, S. M. and Karlin, S. (1956) On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.* **27**: 1060–1074.

Joshi, V. M. (1975) A conjecture of Berry regarding a Bernoulli two-armed bandit. *Ann. Statist.* **3**: 189–202.

CHAPTER 8

Continuous-time bandits

For processes that are evolving and can be observed continuously, we would like to formulate a bandit problem in a manner analogous to our formulation for discrete time. First, we would like a strategy to indicate a process to select and observe at time t, for all t, depending on the history of observations prior to time t. Then we would define the worth of a strategy to be a weighted sum of the increments of the processes during the time they are being observed. Unfortunately, there are serious technical difficulties with various aspects of this plan. These difficulties are frequently ignored in the literature. While this may not be unreasonable when there are good discrete-time approximations, the technical problems must be understood and overcome before progress can be made on the general problem.

In Section 8.3 we present several examples that indicate the need for care in formulating a definition of strategy. We formulate a definition of strategy using these examples as guides. But we are not sure that there do not exist other examples that would show our definition to be inappropriate.

For the examples of Sections 8.1 and 8.2 we adopt a narrower definition of strategy. This definition seems adequate for those examples and it allows us to make arguments involving discrete-time approximations. The question of how the definition given in Section 8.3 would apply to these examples remains open.

One reason that the definition of strategy used in Sections 8.1 and 8.2 seems appropriate in the examples of those sections is that the discount functions assumed are regular (cf. Definition 5.7.1) and there is only one unknown arm. This is also the case for the examples of Section 8.4, although the definition of strategy used in Section 8.1 is modified slightly in Example 8.4.1. We regard finding the most appropriate formulation of a continuous-time bandit to be an open problem.

The unknown arm in each of Sections 8.1 and 8.2 involves Brownian motion added to an unknown linear drift. In Section 8.1 the drift is known to be one of two values. This special case plays a central role in the minimax approach of Section 9.2. In Section 8.2 we discuss the work of Chernoff and Ray (1965) in which the drift is normally distributed. We hope these examples motivate the reader to consider the foundational questions that arise in Section 8.3.

The examples in Section 8.4 involve Lévy processes having jumps.

Continuous-time bandits for which there is a cost for switching arms may, from a foundational viewpoint, be more tractable than those for which there is no such cost. The cost makes it unattractive to switch back and forth instantaneously between arms. When there is no cost such switching may be attractive, say, when there are two arms that are Brownian motions added to unknown drifts. We have not considered the possibility of a cost for switching arms.

8.1 Brownian motion with unknown drift: two-point prior

In this and the next section we present two continuous-time two-armed bandit problems. The characteristics of one arm, arm 2 say, are completely known. In particular, arm 2 generates the deterministic process with constant rate λ:

$$Y_2(t) = \lambda t.$$

Arm 1, on the other hand, has unknown drift which is obscured by noise. We assume that arm 1 generates a Brownian motion

$$Y_1(t) = \theta_1 t + B(t), \tag{8.1.1}$$

where $B(t)$ is the standard Wiener process (i.e., standard Brownian motion) having mean 0 and variance 1 at $t = 1$. The drift coefficient θ_1 is itself random and, in this section, has a two-point distribution:

$$F = p\delta_a + (1-p)\delta_b$$

where $b < a$ and F is the distribution of θ_1. To avoid annoying trivialities we assume $b < \lambda < a$.

In this section the discounting is the continuous-time analogue of the geometric—namely, the discount function is $e^{-\beta t}$ for some constant β, $0 < \beta < \infty$. The *worth* of a strategy τ is

$$W(F, \lambda; e^{-\beta t}; \tau) = E_\tau \int_0^\infty e^{-\beta t} dY_{\tau(t)}(t), \tag{8.1.2}$$

where $\tau(t)$ indicates the arm being observed at time t. In continuous time some care is needed to give a precise meaning to the phrase 'the arm being observed at time t'.

Before giving a precise definition of *strategy*, we make some comparisons with the discrete-time setting. In discrete time, arm 1 is governed by a probability distribution Q_1 which is itself random. Given Q_1, the outcomes X_{1m} on arm 1 are independent and identically distributed. Were we to replace m by a real parameter t while keeping the independence requirement, we would obtain a nonmeasurable process X_{1t}. This would be quite unsatisfactory since we also need to introduce integrals as analogues of discrete-time sums. A different view is needed. For the discrete-time setting let

$$Y_{1m} = \sum_{l=1}^{m} X_{1l}, \qquad m = 1, 2, 3, \ldots; \qquad Y_{10} = 0.$$

Following the approach of previous chapters, the value is incremented by the quantity $\alpha_m X_{1m} = \alpha_m(Y_{1m} - Y_{1,m-1})$ when arm 1 is observed at stage m and where α_m is the mth term of the discount sequence. The process Y_{1m}, consisting of the partial sums of independent identically distributed random variables, does have a convenient continuous-time analogue – namely, a stochastic process with stationary independent increments (i.e., a *Lévy process*), the best known example of which is Brownian motion. So (8.1.2) is natural as a continuous-time analogue of the discrete-time definition of worth given in Chapter 2.

Since $\mathrm{d}(\lambda t) = \lambda\,\mathrm{d}t$, arm 2 in this example plays a role analogous to that of a discrete-time known arm λ. As in Example 2.4.7, nothing in this example would be changed were the deterministic process $Y_2(t) = \lambda t$ replaced by a Lévy process (or any other process) with known mean λt at time t. (The mean of a Lévy process is either a multiple of t or else does not exist for $t > 0$.)

We can rewrite (8.1.2) as follows:

$$W(F, \lambda; \mathrm{e}^{-\beta t}; \tau) = E_\tau \int_0^\infty \mathrm{e}^{-\beta t} \mathbf{1}_{\{\tau(t)=1\}} \mathrm{d}Y_1(t) + E_\tau \int_0^\infty \mathrm{e}^{-\beta t} \mathbf{1}_{\{\tau(t)=2\}} \lambda \mathrm{d}t. \tag{8.1.3}$$

The latter term is meaningful provided the random set $\{t\colon \tau(t) = 2\}$ is almost surely a measurable subset of $[0, \infty]$. In order for τ to qualify

as a strategy we impose this measurability requirement. The first term in (8.1.3) is a stochastic integral. We require $\mathbf{1}_{\{\tau(t)=1\}}$, and therefore $\mathbf{1}_{\{\tau(t)=2\}}$, to satisfy the assumption of progressive measurability that is standard in the definition of a stochastic integral with respect to continuous stochastic processes; the process $f(t, \omega)$ is *progressively measurable* with respect to an increasing family of σ-fields $\mathscr{F}_t$ if, for each t_0, $f(t, \omega)$, $0 \leqslant t \leqslant t_0$, is measurable with respect to the product of two σ-fields: the σ-field of Borel subsets of $[0, t_0]$ and $\mathscr{F}_{t_0}$. The relevant $\mathscr{F}_{t_0}$ is the σ-field generated by $\{Y_1(t)\colon 0 \leqslant t \leqslant t_0\}$.

In the discrete-time case we permit the arm indicated by a strategy τ at time m to depend only on the increments X_{1l} of Y_{1l} for those $l < m$ for which arm 1 was actually observed. We proceed to develop an analogous restriction when time is continuous. We require $\{t\colon \tau(t) = 1\} \cup [0, t_0)$ to be the union of a finite (possibly random) number of left-closed, right-open intervals for each finite t_0. Hence,

$$\{t\colon \tau(t) = 1\} = \bigcup_i [r_i, s_i),$$

where the union may be empty, finite, or infinite, r_i and s_i are random, and, when defined, $s_i < r_{i+1}$. We require that each r_{i+1} be measurable with respect to the σ-field generated by the family of increments

$$\{Y_1(t) - Y_1(r_j)\colon r_j \leqslant t \leqslant s_j, j \leqslant i\}$$

and that each event $\{s_{i+1} \leqslant t_0\}$ be measurable with respect to the σ-field generated by

$$\{Y_1(t) - Y_1(r_{i+1})\colon r_{i+1} \leqslant t \leqslant t_0\} \cup \{Y_1(t) - Y_1(r_j)\colon r_j \leqslant t \leqslant s_j, j \leqslant i\}.$$

(In general, we would permit the arm indicated by a strategy τ at time t to depend on increments observed prior to time t on the other arms, but this would make no difference in this example since the only other arm is known.)

Now that we have a precise definition of strategy, we can define V in terms of W as in the discrete-time setting:

$$V(F, \lambda; e^{-\beta t}) = \sup_{\tau} W(F, \lambda; e^{-\beta t}; \tau),$$

where the supremum is taken over the strategies τ just described.

Exponential discounting has the property possessed by the geometric discount sequence, its discrete-time analogue: except for normalization, the discount function does not change with time. Hence, once

arm 2 becomes optimal and is selected, it remains optimal. Strongly-to-the-right arguments carry over from the discrete case by approximation. Therefore, the larger the probability that $\theta_1 = a$, the stronger is the inclination to use arm 1. Accordingly, there is a constant $C \in (0, 1)$ such that arm 1 is optimal at time t if the current probability that $\theta_1 = a$ is larger than C, and arm 2 is optimal if the current probability is less than C. If this current probability equals C, only arm 2 is optimal; for if arm 1 were selected initially the wild oscillations of Brownian motion would push the new probability that $\theta_1 = a$ below C in an infinitesimal time, indicating that arm 2 should already have been in use.

Summarizing, if $p = F(\{a\}) \leqslant C$, the decision maker should select arm 2 indefinitely into the future. If $p > C$, the decision maker should select arm 1 initially and stay with arm 1 until $p(t, Y_1(t)) = C$, where $p(t, y)$ denotes the conditional probability that $\theta_1 = a$ given that $Y_1(t) = y$. At such time the decision maker should switch permanently to arm 2. Since $Y_1(t)$ and $p(t, y)$ are continuous functions, the switching time can be identified on the basis of the observations of Y occurring strictly before that time, as required in the definition of a strategy.

The current probability that $\theta_1 = a$ is

$$p(t, y) = \frac{p\frac{1}{\sqrt{(2\pi t)}}\exp\left[-(y-at)^2/(2t)\right]}{p\frac{1}{\sqrt{(2\pi t)}}\exp\left[-(y-at)^2/(2t)\right] + (1-p)\frac{1}{\sqrt{(2\pi t)}}\exp\left[-(y-bt)^2/(2t)\right]}$$

$$= \frac{p\exp(ay - a^2t/2)}{p\exp(ay - a^2t/2) + (1-p)\exp(by - b^2t/2)}. \tag{8.1.4}$$

Accordingly, the condition $p(t, y) > C$ becomes

$$\frac{p\exp(ay - a^2t/2)}{(1-p)\exp(by - b^2t/2)} > \frac{C}{1-C},$$

which is equivalent to

$$y > \frac{(a+b)}{2}t - \frac{1}{a-b}\log\frac{(1-C)p}{C(1-p)}.$$

(For the special case $t = 0 = y$ this reduces to $p > C$, as previously indicated.)

We proceed to find C. Let τ_x denote the strategy: switch to arm 2 permanently at the random time T (possibly 0 or ∞), the smallest t for which

$$Y_1(t) \leqslant \frac{(a+b)}{2}t - \frac{1}{a-b}\log\frac{(1-x)p}{x(1-p)} \tag{8.1.5}$$

or, equivalently,

$$B(t) + \frac{2\theta_1 - a - b}{2}t \leqslant -\frac{1}{a-b}\log\frac{(1-x)p}{x(1-p)}. \tag{8.1.6}$$

Then τ_C is an optimal strategy.

Let us consider an F for which $p > x$ so that the inequalities in (8.1.5) and (8.1.6) hold with equality when $t = T$. Conditioned on θ_1, the Laplace–Stieltjes transform of the distribution of T is known (for instance, Fristedt, 1974, Corollary 9.9 and the formula at top of page 351):

$$E(e^{-\beta T}|\theta_1 = b) = \left[\frac{x(1-p)}{(1-x)p}\right]^{\nu - 1/2} \tag{8.1.7}$$

and

$$E(e^{-\beta T}|\theta_1 = a) = \left[\frac{x(1-p)}{(1-x)p}\right]^{\nu + 1/2}, \tag{8.1.8}$$

where

$$\nu = \frac{\sqrt{[8\beta + (a-b)^2]}}{2(a-b)} \tag{8.1.9}$$

To calculate $W(F, \lambda; e^{-\beta t}; \tau_x)$ we will use

$$E\left(\int_0^T e^{-\beta t}\, dY_1(t)\middle|\theta_1 = a\right)$$

$$= \frac{a}{\beta}E(1 - e^{-\beta t}|\theta_1 = a) + E\left(\int_0^\infty e^{-\beta t}\, 1_{\{T > t\}}\, dB(t)\middle|\theta_1 = a\right)$$

and the similar formula for conditioning on $\theta_1 = b$. The latter term is zero as it is the expected value of a stochastic integral with respect to standard Brownian motion. (To draw this conclusion one should

observe that B is a Brownian motion adapted to the σ-fields generated by it and θ_1.) Hence,

$$
\begin{aligned}
&W(F, \lambda; e^{-\beta t}; \tau_x) \\
&= \left\{\frac{a}{\beta} E_{\tau_x}(1-e^{-\beta T}|\theta_1 = a) + E_{\tau_x}\left(\int_T^\infty e^{-\beta t}\lambda dt \,\middle|\, \theta_1 = a\right)\right\}p \\
&\quad + \left\{\frac{b}{\beta} E_{\tau_x}(1-e^{-\beta T}|\theta_1 = b) + E_{\tau_x}\left(\int_T^\infty e^{-\beta t}\lambda dt \,\middle|\, \theta_1 = b\right)\right\}(1-p) \\
&= \beta^{-1}[(E(\theta_1|F) + (\lambda - a)E_{\tau_x}(e^{-\beta T}|\theta_1 = a)p \\
&\qquad + (\lambda - b)E_{\tau_x}(e^{-\beta T}|\theta_1 = b)(1-p)]. \qquad (8.1.10)
\end{aligned}
$$

By using (8.1.7) and (8.1.8) in (8.1.10) and then differentiating with respect to x, setting equal to 0, and solving for $x = C$, we find

$$
C = \frac{(\lambda - b)(b - a + \sqrt{[8\beta + (a-b)^2]})}{(a-b)(a+b-2\lambda + \sqrt{[8\beta + (a-b)^2]})}. \qquad (8.1.11)
$$

If $p \leqslant C$ then arm 2 is optimal throughout:

$$
V(F, \lambda; e^{-\beta t}) = \beta^{-1}\lambda.
$$

If $p > C$ then (8.1.7), (8.1.8), and (8.1.9) imply

$$
\begin{aligned}
V(F, \lambda; e^{-\beta t}) &= \beta^{-1}[E(\theta_1|F) + (\lambda - (1-C)b - Ca)] \\
&\quad \times \left[\frac{C(1-p)}{(1-C)p}\right]^{\nu} \left[\frac{p(1-p)}{C(1-C)}\right]^{1/2} \qquad (8.1.12)
\end{aligned}
$$

where ν and C are defined in (8.1.9) and (8.1.11).

In the spirit of Chapter 5, we insert p for C in (8.1.11) and solve for λ. In notation consistent with Chapter 5, we obtain the solution

$$
\Lambda(F, e^{-\beta t}) = \frac{E(\theta_1^2|F) - ab + E(\theta_1|F)\sqrt{[8\beta + (a-b)^2]}}{(a-b)(2p-1) + \sqrt{[8\beta + (a-b)^2]}}. \qquad (8.1.13)
$$

If $\lambda < \Lambda(F, e^{-\beta t})$ then arm 1 is optimal initially. If $\lambda \geqslant \Lambda(F, e^{-\beta t})$ then arm 2 is optimal initially and henceforth. Figure 8.1 shows the

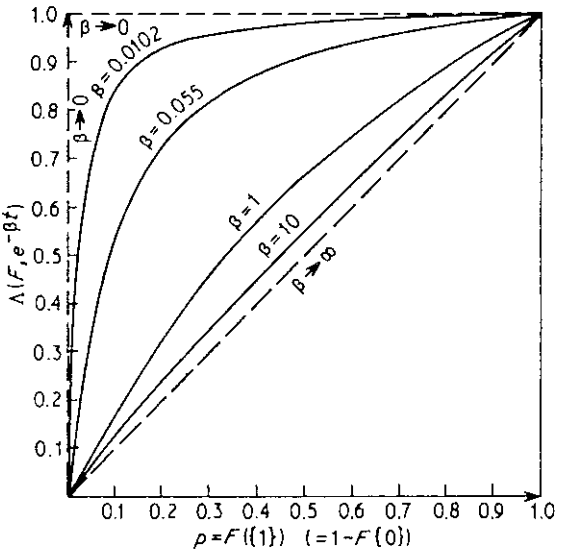


Fig. 8.1 Λ *as a function of* F *which is supported by* $\{1, 0\}$.

graph of Λ versus p for $a = 1, b = 0$, and several values of β. When β is small, arm 1 should be tested unless λ is very close to 1, or p is very close to 0. This is highlighted on the figure by two dashed line segments indicating the limits as $\beta \to 0$. When β is large, arm 1 should not be tried unless $E(\theta_1 \mid F)$ is rather close to λ. This is indicated on the figure by the dashed line $E(\theta_1 | F)$ which is the limit of the graph as $\beta \to \infty$.

Another approach to arriving at (8.1.13) or (8.1.11) is to regard $p(t, Y_1(t))$ as a diffusion process on $[0, 1]$. Letting w denote the initial value, we have, from (8.1.4),

$$p(t, Y_1(t)) = \frac{w \exp(aY_1(t) - a^2 t/2)}{w \exp(aY_1(t) - a^2 t/2) + (1 - w) \exp(bY_1(t) - b^2 t/2)}. \tag{8.1.14}$$

We would like a formula for $dp(t, y)$ in terms of a standard Brownian motion. It is tempting to try (8.1.14), (8.1.1), and Itô's formula, but this scheme fails. The process $p(t, Y_1(t))$ is not adapted to the σ-fields generated by $B(t)$ and it does not have the Markov property with respect to the σ-fields obtained by adjoining the σ-field generated by

θ_1 to those generated by $B(t)$. The σ-fields $\mathscr{F}_t^Y$ generated by Y_1 are those of interest. Theorem 7.12 of Lipster and Shiryayev (1977) gives the appropriate alternative to (8.1.1):

$$dY_1(t) = E(\theta_1 \mid \mathscr{F}_t^Y)\,dt + d\bar{B}(t)$$

where $\bar{B}(t)$ is a Brownian motion adapted to $(\mathscr{F}_t^Y: t \geq 0)$. A straightforward calculation using Itô's formula (for instance, Lipster and Shiryayev, 1977, Theorem 4.4) and

$$E(\theta_1 \mid \mathscr{F}_t^Y) = b + (a-b)p(t, Y_1(t))$$

gives

$$dp(t, Y_1(t)) = (a-b)p(t, Y_1(t))[1 - p(t, Y_1(t))]d\bar{B}(t).$$

The $(F, \lambda; e^{-\beta t})$-bandit is a stopping problem for the diffusion $p(t, Y_1(t))$. Karatzas (1984) solves this problem in terms of solutions of an ordinary differential equation involving the generator of the process. The differential equation we need to solve is

$$\frac{1}{2}[(a-b)w(1-w)]^2 u''(w) - \beta u(w) = 0.$$

Two linearly independent solutions of the form $w^q(1-w)^r$ can be obtained. According to Karatzas (1984), it is a decreasing solution that is of interest:

$$w^{1/2-\nu}(1-w)^{1/2+\nu}, \tag{8.1.15}$$

where ν is defined in (8.1.9).

The expected income from arm 1 between times t and $t + dt$ is

$$e^{-\beta t}[wa + (1-w)b]\,dt, \tag{8.1.16}$$

where w is the initial value of $p(t, Y_1(t))$ and, thus, is its expected value for all t. According to Karatzas (1984), another function of interest is the integral from 0 to ∞ of (8.1.16):

$$\beta^{-1}[wa + (1-w)b]. \tag{8.1.17}$$

Karatzas's (1984) theorem involves a quotient; the denominator is the derivative of (8.1.15) and the numerator is the product of (8.1.17) and the derivative of (8.1.15) minus the product of (8.1.15) and the derivative of (8.1.17). His theorem asserts that $\Lambda(F, e^{-\beta t})$ is obtained by replacing w by p in the quotient. The solution thus obtained agrees with (8.1.12).

8.2 Brownian motion with unknown drift: normal prior

As in Section 8.1 we consider two arms. Arm 1 is Brownian motion with unknown drift: $Y_1(t) = \theta_1 t + B(t)$, and arm 2 is deterministic: $Y_2(t) = \lambda t$. Without loss of generality we take $\lambda = 0$. So the problem is to decide when to gather information and payoff from arm 1 and when to sit idle.

This section differs from Section 8.1 in two ways. First, the prior distribution on θ_1 is normal. Second, discounting is uniform: $\alpha_t = \mathbf{1}_{[0,S]}(t)$, where S is a constant.

According to Definition 5.7.1, α_t is regular. So Proposition 5.7.1 applies to show that there are discrete-time approximations with regular discount sequences. This implies that there is an optimal strategy that indicates arm 2 permanently once it is selected. So, as in the problem considered in Section 8.1, this bandit is a stopping problem. (This conclusion depends on the fact that we are using the same concept of strategy as used in Section 8.1. As indicated in Section 8.3, it is not clear that this is the most appropriate concept.)

Since we are assuming $\lambda = 0$ and uniform discounting, the optimization problem is to choose, for $Y_1(t)$, a stopping time $T \in [0, S]$ that maximizes $E(Y_1(T)|F)$, where F denotes the distribution of θ_1. We assume F is normal with mean μ and variance ρ^2, as in Section 2.1. This problem was studied by Chernoff and Ray (1965) using some results from Chernoff (1961) and Breakwell and Chernoff (1964). We follow their method rather closely.

Consistent with our earlier notation, we use $V((\mu, \rho), 0; S)$ to denote the value of the bandit. A change of variables will prove useful:

$$V_c^*(w, s) = V((ws^{-1}, s^{-1/2}), 0; c - s), \qquad 0 < s \leqslant c,$$

where we have used

$$\mu = ws^{-1}, \quad \rho = s^{-1/2}, \quad S = c - s,$$
$$s = \rho^{-2}, \quad w = \mu\rho^{-2}, \quad c = S + \rho^{-2}.$$

A discrete-time approximation involving a modification of the proof of Theorem 4.3.6 shows that there exists a function $f(\rho, S)$ such that it is optimal, beginning at time 0, to select arm 1 if $\mu > f(\rho, S)$ and to choose arm 2 immediately if $\mu \leqslant f(\rho, S)$. The condition $\mu > f(\rho, S)$ can be written as $w > f_c^*(s)$, where $f_c^*(s) = sf(s^{-1/2}, c - s)$.

It is clear that V and the boundary function f can be recovered from V_c^* and f_c^*, $c > 0$. More is true: for an initial (μ, ρ) and S, there

exists a single fixed c – namely, $S+\rho^{-2}$ – such that f_c^* completely determines an optimal strategy for the $((\mu, \rho), 0; S)$-bandit, and not just an optimal initial selection. The reason is that after time t has elapsed the new horizon is $S-t$ and, if arm 1 is selected throughout the interval $[0, t)$, the new variance is $(\rho^{-2}+t)^{-1}$. Hence, the updated value of c is

$$(S-t)+(\rho^{-2}+t)=S+\rho^{-2},$$

and so c has not changed. Accordingly, we turn to the study of the functions V_c^* and f_c^* for fixed but arbitrary $c>0$.

Clearly, $V_c^*(w, c)=0$ for all w, and $V_c^*(w, s)=0$ for $w \leqslant f_c^*(s)$. Also, $f_c^*(s) \leqslant 0$ for all $s \in (0, c]$ and $V_c^*(w, s) \geqslant 0$ for all $w \in (-\infty, \infty)$ and $s \in (0, c]$. When w is very large the probability is close to 1 that arm 1 will be selected exclusively when following an optimal strategy. The worth of using arm 1 exclusively is $\mu S = ws^{-1}(c-s)$. Therefore,

$$\lim_{w \to \infty} V^*(w, s)/w = (c-s)/s$$

for each $s \in (0, c]$. Also, $V^*(w, s)$ is an increasing function of w for each s.

We now argue that $V_c^*(w, s)$ satisfies

$$\frac{1}{2}\frac{\partial^2}{\partial w^2} V_c^*(w, s)+\frac{w}{s}\frac{\partial}{\partial w} V_c^*(w, s)+\frac{\partial}{\partial s} V_c^*(w, s)=-\frac{w}{s} \quad (8.2.1)$$

for $w > f_c^*(s)$, $0 < s \leqslant c$. Consider a point (w_0, s_0) for which $w_0 > f_c^*(s_0)$. It is optimal to begin by selecting arm 1. A switch to arm 2 should be made when and if a (w, s) for which $w = f_c^*(s)$ is reached. (Because of continuity, hitting such a point can be anticipated, so arm 2 can be selected at the time of impact as well as thereafter.) After Y_1 has been observed up to time t, the new variance of θ_1 equals $(\rho_0^{-2}+t)^{-1}$, as described earlier. Here the subscript 0 on ρ indicates that ρ_0 is determined by (μ_0, s_0) – in fact, $\rho_0 = s_0^{-1/2}$. The posterior mean of θ_1, the mean conditioned by the observed values of Y_1, equals

$$\frac{\mu_0+\rho_0^2 Y_1(t)}{1+\rho_0^2 t},$$

where, of course, $\mu_0 = w_0 s_0^{-1}$. Thus, the point (w_t, s_t), reached at time t, is given by

$$s_t = [(\rho_0^{-2}+t)^{-1}]^{-1} = \rho_0^{-2}+t = s_0+t$$

and

$$w_t = s_t \frac{\mu_0 + \rho_0^2 Y_1(t)}{1 + \rho_0^2 t}$$

$$= \rho_0^{-2}(\mu_0 + \rho_0^2 Y_1(t)) = w_0 + Y_1(t).$$

The stochastic process (w_t, s_t) is a continuous Markov process – that is, a diffusion. For ε less than the distance from (w_0, s_0) to the graph of f_c^*, let T_ε denote the first time that this process reaches a point a distance ε from (w_0, s_0). Since $s_t = s_0 + t$, $T_\varepsilon \leqslant \varepsilon$. Writing w, s, μ, ρ, and S with corresponding subscripts throughout, we calculate

$$V_c^*(w_0, s_0) = V((\mu_0, \rho_0), 0; S_0)$$

$$= \mu_0 E(T_\varepsilon) + E\left(V((\mu_{T_\varepsilon}, \rho_{T_\varepsilon}), 0; S_{T_\varepsilon})\right)$$

$$= \mu_0 E(T_\varepsilon) + E(V_c^*(w_{T_\varepsilon}, s_{T_\varepsilon})).$$

Therefore,

$$-\mu_0 E(T) = E(V_c^*(w_{T_\varepsilon}, s_{T_\varepsilon}) - V_c^*(w_0, s_0))$$

$$= E\left((w_{T_\varepsilon} - w_0)\frac{\partial V_c^*}{\partial w}(w_0, s_0) + \frac{1}{2}(w_{T_\varepsilon} - w_0)^2 \frac{\partial^2 V_c^*}{\partial w^2}(w_0, s_0)\right.$$

$$\left. + (s_{T_\varepsilon} - s_0)\frac{\partial V_c^*}{\partial s}(w_0, s_0) + \ \ldots\right)$$

$$= E\left(Y_1(T_\varepsilon)\frac{\partial V_c^*}{\partial w}(w_0, s_0) + \frac{1}{2}[Y_1(T_\varepsilon)]^2 \frac{\partial^2 V_c^*}{\partial w^2}(w_0, s_0)\right.$$

$$\left. + T_\varepsilon \frac{\partial V_c^*}{\partial s}(w_0, s_0) + \ \ldots\right).$$

So we need to calculate

$$E(Y_1(T_\varepsilon)) = \mu_0 E(T_\varepsilon)$$

and

$$E[Y_1(T_\varepsilon)]^2 = \operatorname{var}[Y_1(T_\varepsilon)] + [E(Y_1(T_\varepsilon))]^2$$

$$= \operatorname{var}[\theta_1 T_\varepsilon + B(T_\varepsilon)] + \mu_0^2[E(T_\varepsilon)]^2$$

$$= E(T_\varepsilon) + \operatorname{var}(\theta_1 T_\varepsilon) + \mu_0^2[E(T_\varepsilon)]^2,$$

which is asymptotic to $E(T_\varepsilon)$ as $\varepsilon \downarrow 0$ since $T_\varepsilon \leqslant \varepsilon$. Keeping only terms of order $E(T_\varepsilon)$ we obtain, as $\varepsilon \downarrow 0$,

$$-\mu_0 E(T_\varepsilon) \sim \mu_0 E(T_\varepsilon)\frac{\partial V_\varepsilon^*}{\partial w}(w_0, s_0) + \frac{1}{2}E(T_\varepsilon)\frac{\partial^2 V_\varepsilon^*}{\partial w^2}(w_0, s_0)$$

$$+ E(T_\varepsilon)\frac{\partial V_\varepsilon^*}{\partial s}(w_0, s_0).$$

The desired partial differential equality (8.2.1) follows by dividing by $E(T_\varepsilon)$, using $\mu_0 = w_0 s_0^{-1}$, and then dropping the subscript 0 throughout.

It is straightforward to prove that V_ε^* is a continuous function by an argument similar to that used to prove Theorem 4.1.1. Since V_ε^* is identically 0 in the stopping region, the directional derivative of V_ε^* along the boundary of the stopping region equals 0. (The question of establishing boundary conditions for a more general reward structure is considered by Bather (1970). He studies the problem for the heat equation but his results transfer to general diffusions by scaling and using a speed measure.)

The problem of finding V_ε^* and f_ε^* satisfying the various conditions obtained above is called a 'free boundary problem' in the partial differential equations literature. This name derives from the fact that the boundary of the region where the partial differential equation is to be solved is not given, but rather obtaining the boundary is part of the problem. In fact, obtaining the boundary is for us the most important aspect of the problem, for the boundary determines optimal strategies. Chernoff and Ray (1965) indicate that the argument of Breakwell and Chernoff (1964) applies to show that the above conditions determine f_ε^* uniquely.

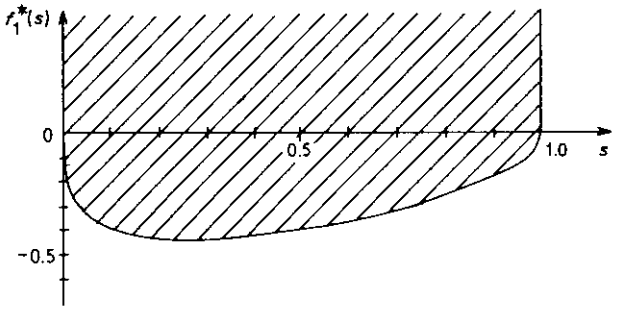


Fig. 8.2 *The continuation region in terms of* f_1^*.

The graph of f_1^* is shown in Figure 8.2, on which the continuation region is shaded. Since $w/\sqrt{s} = \mu\sqrt{s}$ is standard normal when $\theta_1 = 0$, it is helpful to view the stopping boundary in terms of $z = f_1^*(s)/\sqrt{s}$; this is shown in Figure 8.3. Both figures were constructed using Table 21 of Chernoff and Petkau (1983).

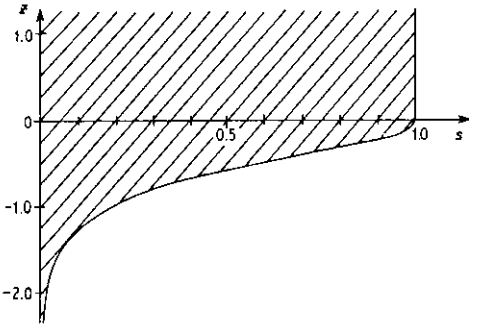


Fig. 8.3 *The continuation region in terms of* $z = f_1^*/\sqrt{s}$.

8.3 General setting

The parameters of a continuous-time bandit are similar to those of a discrete-time bandit. A distribution G describes the probabilistic character of the k arms, and a discount function α_t indicates the utility of the observation at time t. We also require definitions for strategy and for the worth of a strategy.

As in Chapter 2, G represents prior information concerning the k arms. We continue to assume $G \in \mathcal{D}^*(\mathcal{D}^k)$; that is, that each component Q_i of $(Q_1, \ldots, Q_k) \in \mathcal{D}^k$ has a finite first absolute moment with G-probability one and, moreover, that this moment has finite G-expectation. Since time is continuous we now require that G be supported by

$$\{(Q_1, \ldots, Q_k) \in \mathcal{D}^k\colon Q_i \text{ is infinitely divisible for each } i\}.$$

Given $(Q_1, \ldots, Q_k)$, the k arms produce k independent Lévy processes Y_i, $i = 1, \ldots, k$, with the distribution of Y_i at time 1 being Q_i. Lévy processes are stochastic processes with stationary in-

dependent increments and are the continuous-time analogues of sums of independent identically distributed random variables. Conditioned on Q_i, the distribution of $Y_i(t)$ is Q_i^{*t}, the tth convolution power of Q_i.

The *discount function* α_t is a nonnegative measurable function on $[0, \infty)$ that has a finite L_1 norm:

$$|\alpha_t|_1 = \int_0^\infty \alpha_t \, dt < \infty. \tag{8.3.1}$$

It is not clear how to formulate the most natural and useful definition of *strategy*. Whatever definition is used, the *worth* of a strategy τ is

$$W(G; \alpha_t; \tau) = \sum_{i=1}^{k} E_\tau \int_0^\infty \alpha_t \mathbf{1}_{\{\tau(t)=i\}} \, dY_i(t), \tag{8.3.2}$$

where $\tau(t)$ denotes the arm indicated by τ at time t. The *value* of the $(G; \alpha_t)$-bandit is

$$V(G; \alpha_t) = \sup_\tau W(G; \alpha_t; \tau).$$

A strategy τ is *optimal* if

$$W(G; \alpha_t; \tau) = V(G; \alpha_t).$$

Compound Poisson processes are among the simplest Lévy processes. They are constant between jumps. The times elapsed between successive jumps have a common exponential distribution and are independent of each other. The sizes of the jumps are identically distributed. They also are independent of each other as well as of the times at which they occur. Parts of Sections 3.5 and 5.7 involve stages occurring at random times determined by exponential waiting times. Thus the random processes of those sections can be regarded as compound Poisson processes. The random structure has the additional property that there exists a $\kappa > 0$ such that, with G-probability one, the random process associated with any arm has time lapses of mean κ between jumps. The magnitudes of the jumps for the known arm 2 in Section 5.7 all equal λ. According to (8.3.2), the intervals of constancy do not contribute to the worth of a strategy; this is consistent with Sections 3.5 and 5.7.

The meanings of *value*, *worth*, and *optimal strategy* obviously depend on the notion of strategy that is used. In particular, this notion

must be sufficiently restrictive for the stochastic integrals in (8.3.2) to be meaningful. We also want $\tau(t)$ to depend only on increments of the Lévy processes observed before time t. Whatever notion is used, the analogue of (2.5.1) should hold:

$$-\infty < \bigvee_{i=1}^{k} E(Y_i(1)|G)|\alpha_t|_1 \leqslant V(G;\alpha_t)$$
$$\leqslant E(\bigvee_{i=1}^{k} E(Y_i(1)|Q_1)|G)|\alpha_t|_1 \leqslant E(\bigvee_{i=1}^{k} Y_i(1)|G)|\alpha_t|_1 < \infty. \quad (8.3.3)$$

The strict inequalities involving $-\infty$ and $+\infty$ arise from (8.3.1) and the assumption that $G \in \mathscr{D}^*(\mathscr{D}^k)$.

Conditional on $(Q_1, \ldots, Q_k)$, $(Y_1, \ldots, Y_k)$ is a k-dimensional Lévy process having independent components and finite first moments. The observed process is

$$Z(t) = \sum_{i=1}^{k} \int_{[0,t]} \mathbf{1}_{\{\tau(s)=i\}} \, dY_i(s). \quad (8.3.4)$$

Let $\mathscr{F}_t^Z$ and $\mathscr{F}_t^\tau$ denote the σ-fields generated by $\{Z(s): s \leqslant t\}$ and $\{\tau(s): s \leqslant t\}$. Our vaguely stated requirement that $\tau(t)$ depend only on the increments of the Lévy processes observed before time t can be given a precise meaning as follows: for each t, $\tau(t)$ is $\mathscr{F}_{t-}^Z$-measurable, or, equivalently, $\mathscr{F}_t^\tau \subset \mathscr{F}_{t-}^Z$. This condition also implies that τ is predictable (cf. Dellacherie and Meyer, 1978, page 121) with respect to the σ-fields generated by $(Q_1, \ldots, Q_k)$ and

$$\{(Y_1(s), \ldots, Y_k(s)): \quad 0 \leqslant s \leqslant t\}.$$

(Predictability, rather than just progressive measurability, is the standard assumption in the theory of stochastic integrals when the stochastic processes may have jumps.) Since $Y_i(t)$ is a semimartingale with respect to these σ-fields, the stochastic integrals in (8.3.4) are defined (cf. Dellacherie and Meyer, 1980, Chapter VIII)—or, more precisely, they would be defined if (8.3.4) were merely a sum of stochastic integrals and not a stochastic differential equation. We do not know if additional conditions need be imposed on τ in order that (8.3.4) have a solution and that it be unique.

The following two examples indicate other reasons for imposing additional conditions on τ for it to be called a strategy.

Example 8.3.1 Modify the example of Section 8.1 as follows. Let

$$Y_1(t) = \begin{cases} at + \sigma B(t) \\ bt + \eta B(t) \end{cases}$$

with probability 1/2 each, where $\sigma \neq \eta$ and $b < \lambda < a$. We continue to assume exponential discounting. If arm 1 is observed initially for some (possibly random) positive amount of time, then the standard deviation can be identified with probability one according to the law of the iterated logarithm. Hence, the mean can be identified with probability one. If arm 1 is observed for a positive length of time, with probability 1/2 it will be clear that $Y_1(t) = bt + \eta B(t)$ and therefore that arm 1 has already been observed too long. On the other hand, a brief observation of arm 1 seems necessary to determine whether $Y_1(t) = at + \sigma B(t)$. The net loss in observing arm 1 when $Y_1(t) = bt + \eta B(t)$ can be made arbitrarily small by observing arm 1 a sufficiently short time. This makes it seem that there is no optimal strategy among those that stay with either particular arm for an interval of time. Sudderth (1984) has observed, however, that, for strategies as characterized above, there is an optimal strategy – namely, observe arm 1 when $t = 0$ and for $t > 0$ observe arm 1 or arm 2 according as $Y_1(t) = at + \sigma B(t)$ or $Y_1(t) = bt + \eta B(t)$. This peculiar strategy satisfies the condition that, for each t, the events $\{\tau(t) = 1\}$ and $\{\tau(t) = 2\}$ depend only on past observations since by time t, the value of the standard deviation is known from the law of the iterated logarithm. □

The following example is similar and is due to Heath (1984).

Example 8.3.2 Consider two arms:

$$Y_1(t) = \theta_1 t + \sigma B(t),$$
$$Y_2(t) = t,$$

where $B(t)$ is standard Brownian motion, σ is a large (known) constant, and θ_1 is uniformly distributed on $[0, 2]$. The discount function is $e^{-\beta t}$.

Intuitively, one expects that if β is sufficiently small (depending on how large σ is), then arm 1 should be observed initially for a considerable duration of time so that the decision maker can learn about θ_1. However, there is a better strategy: observe arm 1 at

$t = 0$, and for $t > 0$ observe arm 1 or arm 2 according as $\theta_1 > 1$ or $\theta_1 \leqslant 1$. □

The preceding two examples indicate that the condition $F_t^Y \subset F_{t-}^Z$ is not sufficiently strong to reflect our intuitive notion of strategy. We suggest adjoining the reasonable condition $\mathscr{F}_{t+}^Y \subset \mathscr{F}_t^Z$, which is strong enough to rule out the 'undesirable' strategies of the preceding two examples. The condition $\mathscr{F}_{t+}^Y \subset \mathscr{F}_t^Z$ is implied by $\mathscr{F}_u^Y \subset \mathscr{F}_{u-}^Z$ for every u in case the family $(\mathscr{F}_u^Z: u \geqslant 0)$ is right-continuous. We believe this will be the case if there is a set of Lévy processes satisfying the following two conditions: (i) the measures induced on the function space $D[0, 1)$ by any two of the processes in the set are absolutely continuous with respect to each other; (ii) with G-probability one, each Q_t determines a Lévy process belonging to this set. ($D[0, 1)$ is the space of right-continuous functions on $[0, 1)$ having left limits.) This absolute continuity property holds in the examples of Sections 8.1 and 8.2. Skorokhod (1965, Section 4.3), for example, gives conditions for mutual absolute continuity of Lévy processes.

It should be emphasized that we do not know if the additional condition $\mathscr{F}_{t+}^Y \subset \mathscr{F}_t^Z$ is either necessary or sufficient for the stochastic differential equation (8.2.4) to have a solution. Our purpose in inserting this condition is only to rule out 'strategies' that do not correspond to our intuitive notion of strategy.

The problem of defining $\tau(t, \omega)$ as a function of ω deserves some consideration. Since we want $\mathscr{F}_t^Y \subset \mathscr{F}_{t-}^Z$, we take

$$\tau(t, \omega) = \hat{\tau}(t, Z_{[0,t)}(\omega))$$

where $\hat{\tau}(t, \cdot)$ is a measurable function on $D[0, t)$ and $Z_{[0,t)}$ denotes Z restricted to $[0, t)$. For each t regard $\hat{\tau}(t, \cdot)$ as being defined on $D[0, \infty)$ as follows:

$$\hat{\tau}(t, h) = \hat{\tau}(t, h_{[0,t)})$$

for $h \in D[0, \infty)$ with $h_{[0,t)}$ denoting the restriction of h to $[0, t)$. Then $\mathscr{F}_t^{\hat{\tau}} \subset \mathscr{D}(D[0, \infty))$ for each t, so it is meaningful to speak of $\mathscr{F}_{t+}^{\hat{\tau}}$. In order that $\mathscr{F}_{t+}^Y \subset \mathscr{F}_t^Z$ we require $\mathscr{F}_{t+}^{\hat{\tau}} \subset \mathscr{D}(D[0, t])$.

If the program indicated above can be completed, it would be desirable to have a theorem saying that the value is the limit of the sequence of values of discrete approximations. One might think that a counterexample could be built along the lines of Example 8.3.1 since a

discrete approximation will not have the information implied by the law of the iterated logarithm. However, the variance of Brownian motion at some time is the square variation of Brownian motion up to that time, and with high probability this square variation can be estimated with a discrete approximation. A variational argument can also be used to show that the value of the continuous-time bandit in the next example is the limit of the sequence of values of discrete-time approximations.

Example 8.3.3 Suppose that, with probability $\frac{1}{2}$, Y_1 is standard Brownian motion and that, with probability $\frac{1}{2}$, it is a stable Lévy process of index 3/2 having mean 2 at time 1. If Y_1 is stable of index 3/2, jumps will occur with probability one at a random sequence of arbitrarily small positive times. Suppose that arm 2 is known: $Y_2(t) = t$. For this bandit, as in Example 8.3.1, the value equals the upper bound given in (8.3.3). The decision maker can come arbitrarily close to achieving the value $(3/2)|\alpha_t|_1$ by observing arm 1 for a short interval of time and then using arm 1 or arm 2 according as jumps have or have not been observed. □

The practical reader might feel that the detailed local structure of the random processes is irrelevant because it cannot be observed in practice. Thus, for instance, one would not be able to draw the conclusions indicated in Examples 8.3.1 and 8.3.3 after observing arm 1 for a very short time. Such a reader might assign probability one to a set of Lévy processes having common local structure—for example, the set of processes each of which is the sum of a compound Poisson process, a linear function of time, and standard Brownian motion. Locally, all such processes look like standard Brownian motion.

We expect there to be continuous-time bandits without optimal strategies. However, if the definition of strategy is too narrow, optimal strategies may fail to exist in cases where it is natural to expect that such strategies should exist. For instance, a right-continuity requirement would rule out the natural optimal strategy in the following example.

Example 8.3.4 For an arbitrary discount function, suppose that $k = 2, Y_2(t) \equiv 0$, and that $Y_1(t)$ is, with probability $\frac{1}{2}$ each, either the standard (right-continuous) Poisson process with mean t or its negative. An optimal strategy is to observe arm 1 up to and including

the first jump time and thereafter use arm 1 or arm 2 according as that jump is or is not positive. □

The next example also indicates the importance of the class of strategies being considered.

Example 8.3.5 Suppose discounting is exponential and there are two independent arms having the characteristics of arm 1 in Section 8.1; the two possible drifts are the same for both arms but the probabilities may be different. Consistent with the very intuitive result of Theorem 4.3.9, we would expect that it is optimal to observe arm i at time t if the conditional probability that the drift on arm i equals a_t is larger than or equal to the corresponding conditional probability for the other arm; the conditioning is on the observed process $Z(s)$ for $s < t$. (Because of continuity this conditioning is equivalent to conditioning on $\{Z(s), s \leqslant t\}$.)

We have arranged for $\mathscr{F}_t^i \subset \mathscr{F}_{t-}^Z$. The conjecture mentioned earlier that $\{F_u^Z\colon\ u \geqslant 0\}$ is right-continuous can be verified in this case; hence $\mathscr{F}_{t+}^i \subset \mathscr{F}_t^Z$. This leaves open the question of whether the stochastic differential equation (8.3.4) has a unique solution (assuming τ is specified uniquely by, say, requiring $\tau(t) = 1$ whenever the conditional probabilities for the drifts are identical for the two arms). Extrapolation from Karatzas (1984) suggests that it does have a unique solution—Karatzas works in a slightly different context in which processes are stopped during periods when they are unobserved. □

Regularity, as defined in Section 5.7, is likely to play an important role in any general theory. It is important for the examples of the next section.

8.4 Examples

In this section we present two examples in a rather informal manner. Both examples compare a known arm and an unknown arm. In the second example the unknown arm is either deterministic drift or is better than the known arm. In the first example the unknown arm is either deterministic drift or is worse than the known arm.

In the first example there would be no optimal strategy were we to restrict ourselves, as we did in Sections 8.1 and 8.2, to strategies that

select a particular arm at times in the union of intervals of the form $[r, s)$; so the optimal strategy obtained will not have this property. The strategy involves only three possibilities: select arm 2 exclusively, select arm 1 exclusively, or select arm 1 up to and *including* a certain random time and then switch to arm 2.

Example 8.4.1 Suppose there are two arms and the discount function α_t is regular. Let $Y_2(t) = \lambda t$ for some known $\lambda \geqslant 0$. Let Y_1 be the Lévy process

$$Y_1(t) = t + \sum_{s \leqslant t} [Y_1(s) - Y_1(s-)],$$

where $Y_1(s) - Y_1(s-)$ is the jump of Y_1 at time s (Lévy processes are right-continuous and so $Y_1(s) = Y_1(s+)$).

Temporarily conditional on the distribution of $Y_1(1)$. Since Y_1 is a Lévy process, for a Borel subset $C \subset \mathbb{R} - \{0\}$ the cardinality of the set

$$\{t_1 < s \leqslant t_2 : Y_1(s) - Y_1(s-) \in C\} \tag{8.4.1}$$

is a Poisson random variable with mean $(t_2 - t_1)\nu_1(C)$ for some $\nu_1(C)$; $\nu_1(C) = \infty$ is possible, in which case the random set at (8.4.1) has countably infinitely many members with probability one. Clearly, ν_1 is a measure on $\mathbb{R} - \{0\}$; it is called *Lévy measure*.

Since we regard the distribution of Y_1 as random, ν_1 is also random. So the distribution F on the space of distributions for $Y_1(1)$ can instead be regarded as a distribution on the space of measures ν_1. Assume that F is supported by

$$\{\nu_1 : \nu_1(0, \infty) = 0,\ \nu_1(-\infty, 0) < \infty, \int y\, \nu_1(\mathrm{d}y) < -1 \text{ or } \nu_1(-\infty, 0) = 0\}.$$

This assumption means that Y_1 has no positive jumps and is either deterministic drift t (if $\nu_1(-\infty, 0) = 0$) or is inferior to arm 2 (if $\nu_1(-\infty, 0) > 0$ and hence $E(Y_1(1)) = 1 + \int y \nu_1(\mathrm{d}y) < 0 \leqslant \lambda$).

Clearly, any jump in Y_1 calls for an immediate switch to arm 2. The results of Sections 4.3 and 5.2 suggest that our search for optimal strategies can be restricted to two strategies: τ_2, always use arm 2; and τ_1, use arm 1 until (if ever) a jump occurs and then switch permanently to arm 2. This assertion can be demonstrated using discrete approximations and applying the results of Chapters 4 and 5.

The worth of τ_2 is $\lambda\gamma_0$, where γ_t is defined in Section 5.7. The worth

of τ_1 is $W_1 + W_2$, where W_1 is the contribution from arm 1 until the time T of the first jump and W_2 is the contribution from arm 2 thereafter; T may be $+\infty$. Clearly, $W_2 = \lambda E(\gamma_T | F)$ where γ_∞ is defined to be 0. Also,

$$W_1 = \int E\left(\int_0^T \alpha_t \, dY_1(t) \middle| \nu_1 \right) F(d\nu_1)$$

$$= \int E(\gamma_0 - \gamma_T | \nu_1) \, E(Y_1(1) | \nu_1) \, F(d\nu_1).$$

Hence, τ_2 is optimal if

$$\lambda \geqslant \frac{\int E(\gamma_0 - \gamma_T | \nu_1) \, E(Y_1(1) | \nu_1) \, F(d\nu_1)}{E(\gamma_0 - \gamma_T | F)}; \tag{8.4.2}$$

τ_1 is optimal if the inequality is reversed. These conclusions depend on our assumption that $\lambda \geqslant 0$. But if the right-hand side of (8.4.2) is positive then τ_1 is also optimal if $\lambda < 0$.

To be specific, suppose F assigns probability $\frac{1}{2}$ to each of two Lévy processes. One process corresponds to $\nu_1 = 0$ (and is thus the function t since jumps occur with probability 0) and the Lévy measure ν_1 of the other has a single atom of size q at $-2/q$. Whatever the value of q, $E(Y_1(1)|F) = -1$. Nevertheless, the optimal strategy depends on q. The right-hand side of (8.4.2) is

$$\frac{\int_0^\infty q e^{-qt} \gamma_t \, dt}{2\gamma_0 - \int_0^\infty q e^{-qt} \gamma_t \, dt},$$

which is a nondecreasing function of q that approaches 1 as $q \to \infty$ and 0 as $q \downarrow 0$. This is intuitively appealing: when q is large, the cost of beginning with arm 1 when it is not the better arm is small since $E(T) = 1/q$. □

The unknown arm in the next example is either deterministic drift or is superior to a known arm.

Example 8.4.2 Suppose there are two arms and the discount function α_t is regular. Let $Y_2(t) = \lambda t$ for some known $\lambda \leqslant 0$. Let

$$Y_1(t) = -t + \sum_{s \leqslant t} [Y_1(s) - Y_1(s-)].$$

As in the previous example, we regard F as a measure on the space of Lévy measures ν_1, the jumps of Y_1 being governed by ν_1. We suppose that F is supported by

$$\{\nu_1: \nu_1(-\infty, 0) = 0,\ \nu_1(0, \infty) < \infty,\ \int y\nu_1\,dy > 1 \text{ or } \nu_1(0, \infty) = 0\}.$$

Assuming the results of Chapter 5 apply in this continuous-time setting, we restrict consideration to strategies that stay with arm 2 indefinitely once it is selected. It is clear that arm 2 should never be used if a jump has been observed on arm 1. Accordingly, the only strategies that need be considered are of the form τ_u: arm 1 is used until time u and arm 2 is used indefinitely thereafter unless a jump has been observed on arm 1; $u \in [0, \infty]$ and u is fixed.

Let T (possibly $+\infty$) denote the time of the first jump of Y_1. The worth of strategy τ_u is given by

$$W(F, \lambda; e^{-\beta t}; \tau_u) = \int E\left(\mathbf{1}_{\{T \leq u\}} \int_0^\infty \alpha_t\,dY_1(t) + \mathbf{1}_{\{T > u\}} \left(\int_0^u \alpha_t\,dY_1(t) + \gamma_u \lambda \right) \middle| \nu_1 \right) F(d\nu_1)$$

$$= \int \Big((1 - e^{-u\nu_1(0,\infty)})\gamma_0 E(Y_1(1)|\nu_1) + e^{-u\nu_1(0,\infty)} ([\gamma_0 - \gamma_u]\, E(Y_1(1)|\nu_1) + \gamma_u \lambda) \Big) F(d\nu_1).$$

We will rely on the local behaviour of $W(F, \lambda; e^{-\beta t}; \tau_u)$ at $u = 0$ to decide which arm is optimal initially. Its derivative evaluated at $u = 0$ equals

$$\int \Big([\gamma_0 \nu_1(0, \infty) + \alpha_0] E(Y_1(1)|\nu_1) - [\gamma_0 \nu_1(0, \infty) + \alpha_0]\lambda \Big) F(d\nu_1).$$

It follows that exclusive use of arm 2 is optimal if

$$\lambda \geq \frac{\int [\alpha_0 + \gamma_0 \nu_1(0, \infty)] E(Y_1(1)|\nu_1) F(d\nu_1)}{E(\alpha_0 + \gamma_0 \nu_1(0, \infty)|F)}; \tag{8.4.3}$$

otherwise arm 1 is optimal initially and for some (possibly infinite) period of time. Since we have assumed $\lambda \leq 0$, arm 1 is optimal initially

when the right-hand side of (8.4.3) is positive. However, the above argument also applies for $\lambda > 0$ provided λ is less than the right-hand side of (8.4.3). If the right-hand side of (8.4.3) is positive and the inequality holds, then no conclusion can be easily drawn. If the right-hand side of (8.4.3) is nonpositive, then arm 2 is optimal in case $\lambda = 0$ and thus, also, in case $\lambda > 0$.

Returning to the original assumption that $\lambda \leqslant 0$, we observe that if (8.4.3) does not hold, arm 1 should be used until (8.4.3) holds with equality, with the distribution of v_1 conditioned by the observation of arm 1 playing the role of F.

To be specific, suppose F assigns probability $1-p$ to the deterministic Lévy process $-t$ and probability p to the Lévy process with measure v_1 consisting of a single atom of size q at $2/q$. Whatever the value of q, $E(Y_1(1)|F_1) = 2p-1$. From (8.4.3), arm 1 is optimal initially when $\lambda \leqslant 0$ if

$$\lambda < \frac{\alpha_0(2p-1)+\gamma_0 pq}{\alpha_0+\gamma_0 pq}.$$

Suppose, in addition, that the discounting is exponential. Then in case $\lambda > -1$ and no jump has occurred on arm 1, a switch to arm 2 is optimal at time

$$q^{-1}\log\frac{p(\alpha_0+\gamma_0 q)(1-\lambda)}{(1-p)\alpha_0(1+\lambda)}.$$

If $\lambda < -1$, it is optimal to observe arm 1 exclusively and indefinitely. □

References

Bather, J. A. (1970) Optimal stopping problems for Brownian motion. *Adv. in Appl. Probab.* **2**: 259–286.

Breakwell, J. and Chernoff, H. (1964) Sequential tests for the mean of a normal distribution II. *Ann. Math. Statist.* **35**: 162–173.

Chernoff, H. (1961) Sequential tests for the mean of a normal distribution, *Fourth Berkeley Symp. of Math. Statist. and Prob.* **1**: 79–91.

Chernoff, H. and Petkau, A. J. (1983) Numerical methods for Bayes sequential decision problems. Univ. of British Columbia Applied Mathematics and Statistics Tech. Rep. No. 83–126.

Chernoff, H. and Ray, S. N. (1965) A Bayes sequential sampling inspection plan. *Ann. Math. Statist.* **36**: 1387–1407.

Dellacherie, C. and Meyer, P.-A. (1978) *Probabilities and Potential*, Part 1, North-Holland, Amsterdam.

Dellacherie, C. and Meyer, P.-A. (1980) *Probabilities and Potential*, Part 2, North-Holland, Amsterdam.

Fristedt, B. (1974) Sample functions of stochastic processes with stationary, independent increments. *Advances in Probability*, Vol. 3 (eds P. Ney and S. Port), pp. 241–396, Marcel-Dekker, New York.

Heath, D. C. (1984) Personal communication.

Karatzas, I. (1984) Gittins indices in the dynamic allocation problem for diffusion processes. *Ann. Prob.* **12**: 173–192.

Lipster, R. S. and Shiryayev, A. N. (1977) *Statistics of Random Processes I*, Springer-Verlag, New York.

Skorokhod, A. V. (1965) *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Massachusetts.

Sudderth, W. D. (1984) Personal communication.

CHAPTER 9

Minimax Approach

A bandit problem is interesting only if there are arms with unknown characteristics. To choose among the available arms a decision maker must first decide how to handle this uncertainty. In the first eight chapters of this monograph the approach used is to average the payoff over the unknown characteristics with respect to a specified prior distribution – a Bayesian approach, in statistical parlance.

Some people prefer an approach that does not require specifying a prior distribution. One such is the minimax or game-theoretic approach in which the decision maker uses a strategy designed to maximize the minimum payoff – a 'worst case' approach.

If the means are assumed to be as small as possible then the decision maker should simply choose the arm for which the smallest possible mean is largest, and the decision problem is trivial. In this chapter we modify the objective in a way which makes the problem nontrivial. Namely, the objective is to minimize *regret*, the expected difference between the payoff that could be obtained if the characteristics of the arms were known and the payoff actually achieved.

One may view the minimax approach as one in which a protagonist – nature – chooses the characteristics of the arms. In trying to make life difficult for the decision maker, there is no particular reason for nature to make the means small, for it may then be very easy for the decision maker to do as well as possible and thus hold the regret to zero. Indeed, for any particular choice of characteristics of the arms the decision maker always has a strategy that holds the regret to zero. But this is not possible if nature uses a random strategy. The decision maker's objective is to do as well as possible against the best random strategy of nature (this provides a connection with the Bayesian approach since a prior distribution for the decision maker is a random strategy for nature).

In Section 9.1 we discuss two Bernoulli arms in a discrete-time

setting. We study the asymptotic behaviour of the value for geometric discounting with the discount factor approaching 1 and uniform discounting with a growing horizon. In Section 9.2 a continuous-time example is treated in some detail.

9.1 Discrete time, two Bernoulli arms

In this section we consider two Bernoulli arms with success probabilities θ_1 and θ_2. The discount sequence $\mathbf{A}$ is arbitrary. Mixed (or randomized) strategies will be introduced shortly. We temporarily consider strategies as defined in Chapter 2 but we now call them 'pure strategies' to distinguish them from mixed strategies. For known θ_1 and θ_2, the worth of a pure strategy τ is

$$W(\theta_1, \theta_2; \mathbf{A}; \tau) = E_\tau \sum_{m=1}^{\infty} \alpha_m Z_m$$

$$= \sum_{m=1}^{\infty} \alpha_m (\theta_1 P_\tau\{\tau^m = 1\} + \theta_2 P_\tau\{\tau^m = 2\}),$$

where τ^m denotes the arm indicated by τ at stage m and the dependence of τ^m on the history of successes and failures preceding stage m has been suppressed in the notation.

The *regret* (sometimes called 'opportunity loss') is defined to be the expected loss (conditioned on (θ_1, θ_2)) resulting from using τ rather than the best arm at every stage:

$$R(\theta_1, \theta_2; \mathbf{A}; \tau) = \sum_{m=1}^{\infty} \alpha_m (\theta_1 \vee \theta_2) - W(\theta_1, \theta_2; \mathbf{A}; \tau)$$

$$= \sum_{m=1}^{\infty} \alpha_m (\theta_1 \vee \theta_2 - \theta_1 \wedge \theta_2) P_\tau\{\tau^m = i, \theta_i = \theta_1 \wedge \theta_2\}. \tag{9.1.1}$$

The topology on the space of pure strategies is defined as follows. The distance between two pure strategies is $1/m$, where m is the first stage for which there exists a history for which the two strategies indicate different arms. A *mixed strategy* is defined to be a probability measure S on the Borel field of subsets of this space. The *regret* for such a strategy is defined as

$$R(\theta_1, \theta_2; \mathbf{A}; S) = \int R(\theta_1, \theta_2; \mathbf{A}; \tau) S(\mathrm{d}\tau) \tag{9.1.2}$$

where the integrand is defined in (9.1.1).

For any subset Θ of $[0, 1] \times [0, 1]$, define the $(\Theta, \mathbf{A})$-*upper regret* to be

$$R_U(\Theta; \mathbf{A}) = \inf_S \sup_{(\theta_1, \theta_2) \in \Theta} R(\theta_1, \theta_2; \mathbf{A}; S). \tag{9.1.3}$$

The mixed strategy S_0 is $(\Theta; \mathbf{A})$-*minimax* if the infimum in (9.1.3) is attained for $S = S_0$.

For each fixed $\mathbf{A} \in \mathscr{A}$ the regret $R(\theta_1, \theta_2; \mathbf{A}; \tau)$ is a uniformly continuous function of $(\theta_1, \theta_2, \tau)$. For mixed strategies, $R(\theta_1, \theta_2; \mathbf{A}; S)$ inherits this property, with the convergence-in-distribution topology being used for the mixed strategies. The space of mixed strategies S also inherits compactness from the space of pure strategies τ. From these facts we obtain the following theorem.

Theorem 9.1.1 For each discount sequence $\mathbf{A}$ and each subset Θ of $[0, 1] \times [0, 1]$, there exists a $(\Theta, \mathbf{A})$-minimax mixed strategy.

A more symmetrical view of this minimax setting is sometimes useful. Suppose nature is a competitor of the decision maker and uses a mixed strategy G.

For Θ measurable, the $(\Theta, \mathbf{A})$-*lower regret* is defined by

$$R_L(\Theta; \mathbf{A}) = \sup_G \left\{ \inf_\tau \int_\Theta R(\theta_1, \theta_2; \mathbf{A}; \tau) G(\mathrm{d}(\theta_1, \theta_2)) : G(\Theta) = 1 \right\}, \tag{9.1.4}$$

and G_0, supported by Θ, is $(\Theta; \mathbf{A})$-*maximin* if the supremum in (9.1.4) is attained for $G = G_0$. In view of the compactness of $[0, 1] \times [0, 1]$ and the continuity of R, this 'two-person zero-sum game' can be approximated by a game in which each of the two players has only finitely many pure strategies. Such approximations also apply with Θ replaced by its closure $\mathrm{cl}(\Theta)$. Accordingly, from the classical theorems on finite matrix zero-sum two-person games, we obtain the following result.

Theorem 9.1.2 For each measurable $\Theta \subset [0, 1] \times [0, 1]$ and each discount sequence $\mathbf{A}$,

$$R_U(\Theta; \mathbf{A}) = R_L(\Theta; \mathbf{A}) = R_U(\mathrm{cl}(\Theta); \mathbf{A}) = R_L(\mathrm{cl}(\Theta); \mathbf{A}).$$

If Θ is closed, there exists a $(\Theta; \mathbf{A})$-maximin strategy G.

The common value of R_U and R_L is usually called the 'value of the game'. We will use *regret value* to avoid confusion with *value* as introduced in Chapter 2.

The arguments leading to Theorem 9.1.2 are standard. Indeed, rather than rely on finite matrix theorems we could have quoted theorems, for instance, Parthasarathy and Raghavan (1971, Chapter 5), about games with continuous payoff functions.

Generalizations of Theorems 9.1.1 and 9.1.2 may be valid for general bandits, not just those with Bernoulli arms. However, such extensions will have to overcome the lack of continuity mentioned in Example 2.5.1.

We now calculate the regret value and optimal strategies for a particular $(\Theta; \mathbf{A})$.

Example 9.1.1 Let $\mathbf{A} = (1, 1, 1, 0, 0, \ldots)$

and

$$\Theta = \{(\theta_1, \theta_2): 0 \leqslant \theta_1 \leqslant 1, \theta_2 = 0 \text{ or } \theta_2 = 1\};$$

so arm 1 is arbitrary but arm 2 either yields all failures or all successes. If arm 2 is selected at any stage m, then, of course, arm 1 or arm 2 should be selected subsequently according as failure or success is observed on arm 2. So there is no loss in limiting the pure strategies available to the decision maker to those having this property. We temporarily assume (and we will show that the assumption is correct) that the decision maker should use a mixture of the following three pure strategies:

τ_1: select arm 1 and switch to arm 2 at stage 2 or stay with arm 1 at stages 2 and 3 according as failure or success is observed at stage 1;

τ_2: select arm 1 until a failure is observed and then switch to arm 2;

τ_3: select arm 2 initially.

The regrets for these strategies calculated from (9.1.1) are given in Table 9.1.

Let us assume (correctly as it will develop) that τ_1, τ_2, and τ_3 need be nature's only concern. There exists a maximum G_0 in view of Theorem 9.1.2. Using the downward concavity of the six relevant functions of θ_1, that there exists an optimal G_0 for nature that is supported by just two points: one of the form $(\theta_1, 0)$ and the other of the form $(\theta_1, 1)$.

Table 9.1 *Values of* $R(\theta_1, \theta_2; (1, 1, 1, 0, 0, \ldots); \tau_i)$

(θ_1, θ_2)	τ_1	τ_2	τ_3
$(\theta_1, 0)$	$\theta_1 - \theta_1^2$	$\theta_1 - \theta_1^3$	θ_1
$(\theta_1, 1)$	$1 + \theta_1 - 2\theta_1^2$	$1 - \theta_1^3$	0

Consider a G which assigns probability p to $(a, 0)$ and probability $1-p$ to $(b, 1)$. The minimum of the three losses equals

$$ap \wedge [1 - b^3 - (1 - b^3 - a + a^3)p] \wedge [1 + b - 2b^2 - (1 + b - 2b^2 - a + a^2)p]. \tag{9.1.5}$$

In view of (9.1.4) and Theorem 9.1.2, G_0 maximizes (9.1.5).

Both $1-b^3$ and $1+b-2b^2$ are decreasing for $b > 1/4$; so restrict consideration to $b \in [0, 1/4]$. As a function of p, (9.1.5) is the minimum of three linear functions whose values at 0 are, respectively, $0 \leqslant 1 - b^3 \leqslant 1 + b - 2b^2$ and whose values at 1 are, respectively, $a \geqslant a - a^3 > a - a^2$. The first of these linear functions has a nonnegative slope, whereas the other two slopes are nonpositive. Accordingly, for fixed a and b, the p that maximizes (9.1.5) is the smaller of two values of p—one makes the first two linear functions equal and the other makes the first and third linear functions equal. When p is so chosen, (9.1.5) equals

$$\frac{a(1-b^3)}{1-b^3+a^3} \wedge \frac{a(1+b-2b^2)}{1+b-2b^2+a^2}. \tag{9.1.6}$$

For fixed b, the two quantities in (9.1.6) are equal for two values of a: 0, which need not be considered, and

$$a = \frac{1-b^3}{1+b-2b^2} = \frac{1+b+b^2}{1+2b} \geqslant 7/8 \tag{9.1.7}$$

(the inequality holding since $b \leqslant 1/4$).

As functions of a, the first quantity in (9.1.6) is decreasing for $a \geqslant 7/8$ and the second quantity is increasing on $[0, 1]$. Hence, for fixed b, the maximum of (9.1.6) is attained for a given in (9.1.7). The maximum equals $f(b)/g(b)$, where

$$f(b) = 1 + 4b + 4b^2 - b^3 - 4b^4 - 4b^5 \tag{9.1.8}$$

and

$$g(b) = 2 + 7b + 9b^2 - 2b^3 - 7b^4. \tag{9.1.9}$$

By differentiating $f(b)/g(b)$ we see that it attains its maximum value at the unique $b \in [0, 1/4]$ for which

$$1 - 2b - 8b^2 - 2b^3 - 41b^4 - 128b^5 - 107b^6 + 16b^7 + 28b^8 = 0. \tag{9.1.10}$$

We now fix b to be this solution and use the previously derived conditions on a and p to fix these quantities. When so fixed, the three quantities in (9.1.5) are equal. Numerical calculations, using (9.1.10), (9.1.7), and the equality of the (first two) quantities in (9.1.5), give $b \approx 0.218$, $a \approx 0.881$, $p \approx 0.591$, and $f(b)/g(b) \approx 0.521$, where $f(b)$ and $g(b)$ are given in (9.1.8) and (9.1.9).

It is now straightforward to calculate

$$\int_{\Theta} R(\theta_1, \theta_2; \mathbf{A}; \tau) G_0(\mathrm{d}(\theta_1, \theta_2))$$

for each of the other 15 pure strategies available to the decision maker. The result in each case is larger than 0.6, thereby affirming our earlier conjectures that the decision maker should use a mixed strategy involving only τ_1, τ_2, and τ_3 and that nature need only be concerned with these three pure strategies of the decision maker. We also conclude that the regret value equals

$$\frac{f(b)}{g(b)} = \frac{(1-b)(1+b+b^2)(1+2b)^2}{2+7b+9b^2-2b^3-7b^4} \approx 0.521. \tag{9.1.11}$$

A minimax strategy is specified by ε_1, ε_2, and ε_3, the optimal probabilities of τ_1, τ_2, and τ_3, respectively. We will calculate these probabilities by finding various linear equations that they must satisfy. The first of these is

$$\varepsilon_1 + \varepsilon_2 + \varepsilon_3 = 1. \tag{9.1.12}$$

From game theory we know that the regret value $f(b)/g(b)$ must equal the regret when the decision maker uses a minimax strategy and nature uses one of the two pure strategies having positive probability in the maximin strategy. From Table 9.1 we obtain

$$(a - a^2)\varepsilon_1 + (a - a^3)\varepsilon_2 + a\varepsilon_3 = R_U, \tag{9.1.13}$$

$$(1 + b - 2b^2)\varepsilon_1 + (1 - b^3)\varepsilon_2 = R_U. \tag{9.1.14}$$

The system consisting of (9.1.12), (9.1.13), and (9.1.14) has infinitely many solutions, but we can find another condition. Temporarily

regarding b to be a variable, the derivative of the left-hand side of (9.1.14) at the maximin b must equal 0; for otherwise $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ when inserted for S in (9.1.3) would yield something greater than R_U. Hence,

$$(1-4b)\varepsilon_1 - 3b^2\varepsilon_2 = 0. \tag{9.1.15}$$

The system consisting of (9.1.12), (9.1.13), (9.1.14), and (9.1.15) has a unique solution, which we express directly in terms of b by using the expression at (9.1.11) for the risk value R_U:

$$\varepsilon_1 = \frac{3b^2(1+b+b^2)(1+2b)^2}{(1-b)(1-2b-2b^2)(2+7b+9b^2-2b^3-7b^4)} \approx 0.257,$$

$$\varepsilon_2 = \frac{(1-4b)(1+b+b^2)(1+2b)^2}{(1-b)(1-2b-2b^2)(2+7b+9b^2-2b^3-7b^4)} \approx 0.235,$$

$$\varepsilon_3 = \frac{1+b-3b^2-15b^3-b^4+30b^5+14b^6}{(1-2b-2b^2)(2+7b+9b^2-2b^3-7b^4)} \approx 0.508. \qquad \square$$

The next two results give bounds for the regret value in case $\Theta = [0, 1] \times [0, 1]$. In the first, the discount sequence is geometric with a discount factor close to 1, and in the second it is the n-horizon uniform with large n. Such results were obtained by Vogel (1960b) for uniform discount sequences. We use ideas from there as well as from Vogel (1960a) and Bather and Simons (1985), although our treatment will be brief since we will not be concerned with obtaining 'good' values for certain constants.

Theorem 9.1.3 There exist positive finite constants c_1 and c_2 such that for every $\alpha \in (0, 1)$,

$$\begin{aligned} c_1(1-\alpha)^{-1/2} &\leqslant R_U([0, 1)] \times [0, 1]; (1, \alpha, \alpha^2, \alpha^3, \ldots)) \\ &\leqslant c_2(1-\alpha)^{-1/2}. \end{aligned} \tag{9.1.16}$$

Proof For the proof of the first equality in (9.1.16), Example 5.4.1 is relevant. We require that θ_1 be one of two values:

$$F_1 = \tfrac{1}{2}\delta_{[1-\sqrt{(1-\alpha)}]/2} + \tfrac{1}{2}\delta_{[1+\sqrt{(1-\alpha)}]/2}$$

and take arm 2 to have the known success probability $\Lambda(F_1, (1, \alpha, \alpha^2, \ldots))$, a function which is defined in Corollary 5.1.2. By Corollary 5.1.2 and Theorem 5.2.2, the value (not 'regret value') equals $(1-\alpha)^{-1}\Lambda$ which can be realized by always selecting arm 2. From

(9.1.1) and (9.1.4) the regret value is greater than or equal to

$$\left(\frac{1+\sqrt{(1-\alpha)}}{2}-\Lambda\right)\Big/[2(1-\alpha)]$$

which, by Example 5.4.1, is asymptotic to $(3-\sqrt{3})/[12\sqrt{(1-\alpha)}]$ as $\alpha \uparrow 1$. The first inequality in (9.1.16) follows.

For the proof of the second inequality in (9.1.16) there is no loss in considering only those α for which $(1-\alpha)^{-1}$ is the square of an integer K. Consider this strategy for the decision maker: select arms 1 and 2 alternatively until, after some even-numbered stage $2M$, the number of successes observed on one arm is at least K greater than that on the other arm; after that stage select the arm on which the greater number of successes has been observed. By pairing the stages we can interpret the phenomenon through the random stage $2M$ as a random walk with steps

$$\begin{aligned} &-1 \text{ with probability } \theta_1(1-\theta_2)\\ &\ \ 0 \text{ with probability } \theta_1\theta_2+(1-\theta_1)(1-\theta_2)\\ &\ \ 1 \text{ with probability } \theta_2(1-\theta_1)\end{aligned}$$

and which is stopped at time M, the first hitting time of the set $\{K, -K\}$. The regret (cf. (9.1.1)) for this strategy is no larger than

$$|\theta_2-\theta_1|E(M)+|\theta_2-\theta_1|(1-\alpha)^{-1}\sum_{m=K}^{\infty} p_m\alpha^m, \qquad (9.1.17)$$

where p_m is the probability that $M=m$ and the random walk at time m has the sign opposite to that of $\theta_2-\theta_1$. We find that $E(M) \leqslant K/|\theta_2-\theta_1|$ (cf. Feller, 1968, answer to problem 5, Chapter XIV), so the first term in (9.1.17) is bounded by $K=(1-\alpha)^{-1/2}$, as desired.

We assume $\theta_2>\theta_1$ with no loss of generality. The generating function $\Sigma p_m\alpha^m$ can be obtained via a difference equation (Feller, 1968, Section XIV.4). The generating function multiplied by $(\theta_2-\theta_1)K^2=(\theta_2-\theta_1)(1-\alpha)^{-1}$ (cf. (9.1.17)) equals

$$\begin{aligned} K^2(\theta_2-\theta_1)[2\theta_1(1-\theta_2)\alpha]^K\Big(&\{(\theta_1+\theta_2-2\theta_1\theta_2)+\\ &+[\theta_1\theta_2+(1-\theta_1)(1-\theta_2)](1-\alpha)+[(\theta_2-\theta_1)^2\\ &+2(\theta_1-\theta_1^2+\theta_2-\theta_2^2)(1-\alpha)\\ &+(\theta_1+\theta_2-1)^2(1-\alpha)^2]^{1/2}\}^K\end{aligned}$$

$$
\begin{aligned}
&+\{(\theta_1+\theta_2-2\theta_1\theta_2)+[\theta_1\theta_2+(1-\theta_1)(1-\theta_2)](1-\alpha)\\
&-[(\theta_2-\theta_1)^2+2(\theta_1-\theta_1^2+\theta_2-\theta_2^2)(1-\alpha)\\
&+(\theta_1+\theta_2-1)^2(1-\alpha)^2]^{1/2}\}^K\Bigg)^{-1}\\
&\leqslant K^2(\theta_2-\theta_1)[2\theta_1(1-\theta_2)]^K\{(\theta_1+\theta_2-2\theta_1\theta_2)+(\theta_2-\theta_1)\}^{-K}\\
&= K^2(\theta_2-\theta_1)[\theta_1(1-\theta_2)/\theta_2(1-\theta_1)]^K\\
&= K^2(\theta_2-\theta_1)\left[\left(1-\frac{\theta_2-\theta_1}{\theta_2}\right)\left(1-\frac{\theta_2-\theta_1}{1-\theta_1}\right)\right]^K\\
&\leqslant K^2(\theta_2-\theta_1)[1-(\theta_2-\theta_1]^{2K} = K^2(\theta_2-\theta_1)e^{2K\log[1-(\theta_2-\theta_1)]}\\
&\leqslant K^2(\theta_2-\theta_1)e^{-2K(\theta_2-\theta_1)}
\end{aligned}
$$

which is no larger than K multiplied by the maximum of xe^{-2x}, $0 \leqslant x \leqslant \infty$. □

Corollary 9.1.4 Suppose $\mathbf{A} = (1, \ldots, 1, 0, \ldots)$ with horizon n. There exist positive finite constants c_3 and c_4 such that

$$c_3 n^{1/2} \leqslant R_U([0,1]\times[0,1]; (1,\ldots,1,0,0,\ldots)) \leqslant c_4 n^{1/2}$$

for every n.

Proof Fix n and choose α to satisfy $1-\alpha = n^{-1}$. Consider the strategy τ (depending on n via α) constructed in the above proof to give an upper bound of (9.1.1):

$$\sum_{m=1}^{\infty} \alpha^{m-1}(\theta_1 \vee \theta_2 - \theta_1 \wedge \theta_2) P_\tau(\tau^m = i, \theta_i = \theta_1 \wedge \theta_2) \leqslant c_2 n^{-1/2}.$$

Hence

$$\sum_{m=1}^{n} (\theta_1 \vee \theta_2 - \theta_1 \wedge \theta_2) P_\tau(\tau^m = i, \theta_i = \theta_1 \wedge \theta_2) \leqslant c_2 \alpha^{-(n-1)} n^{-1/2}.$$

The upper bound of the corollary follows from

$$\alpha^{-n-1} = (1-n^{-1})^{-n-1} \to e < \infty \text{ as } n \to \infty.$$

We now turn to the lower bound. Let G_0 denote a maximin strategy for $\Theta = [0,1]\times[0,1]$ and the geometric discount sequence

$\mathbf{A} = (1, (1-n^{-1}), (1-n^{-1})^2, \ldots)$. By Theorem 9.1.3,

$$c_1 n^{1/2} \leqslant \sum_{m=1}^{\infty} (1-n^{-1})^m [E(\theta_1 \vee \theta_2 | G_0) - E_\tau(Z_m | G_0)] \qquad (9.1.18)$$

for every strategy τ and, in particular, if τ is optimal for the $(G_0; (1, \ldots, 1, 0, 0, \ldots))$-bandit with horizon n. Fix such a τ that proceeds after stage n as well as possible for the $(G_0; ((1-n^{-1})^n, (1-n^{-1})^{n+1}, (1-n^{-1})^{n+2}, \ldots))$-bandit without relying on information from the first n stages. Then, suppressing the dependence of the expectations on G_0,

$$\begin{aligned} &\sum_{m=n+1}^{\infty} (1-n^{-1})^m [E(\theta_1 \vee \theta_2 | G_0) - E_\tau(Z_m | G_0)] \\ &= (1-n^{-1})^n \sum_{m=1}^{\infty} (1-n^{-1})^m [E(\theta_1 \vee \theta_2 | G_0) - E_\tau(Z_{m+n} | G_0)] \qquad (9.1.19) \\ &\leqslant (1-n^{-1})^n \sum_{m=1}^{\infty} (1-n^{-1})^m [E(\theta_1 \vee \theta_2 | G_0) - E_\tau(Z_m | G_0)]. \end{aligned}$$

From (9.1.18) and (9.1.19) we conclude that

$$\begin{aligned} &c_1 [1-(1-n^{-1})^n] n^{1/2} \\ &\leqslant [1-(1-n^{-1})^n] \sum_{m=1}^{\infty} (1-n^{-1})^m [E(\theta_1 \vee \theta_2 | G_0) - E_\tau(Z_m | G_0)] \\ &\leqslant \sum_{m=1}^{n} (1-n^{-1})^m [E(\theta_1 \vee \theta_2 | G_0) - E_\tau(Z_m | G_0)] \\ &\leqslant \sum_{m=1}^{n} [E(\theta_1 \vee \theta_2 | G_0) - E_\tau(Z_m | G_0)]. \end{aligned}$$

This completes the proof since $c_1[1-(1-n^{-1})^n]$ is bounded below by a positive constant. □

In the proof of Theorem 9.1.3 the strategies constructed for nature were not symmetric even though Θ was symmetric. Theorem 9.1.5 below says that only symmetric strategies need to be considered when Θ is symmetric.

Definition 9.1.1 The set Θ of possible pairs (θ_1, θ_2) of Bernoulli parameters is *symmetric* if

$$(\theta_1, \theta_2) \in \Theta \Leftrightarrow (\theta_2, \theta_1) \in \Theta.$$

Definition 9.1.2 A probability measure G on a symmetric $\Theta \subset [0, 1] \times [0, 1]$ is *symmetric* if $G(\Psi) = G(\{(\theta_1, \theta_2): (\theta_2, \theta_1) \in \Psi\})$ for every measurable $\Psi \subset \Theta$.

Definition 9.1.3 A probability measure S on the space of pure strategies for the decision maker is *symmetric* if

$$S(T) = S(\{\tau: \tau_c \in T\})$$

for every measurable set T of pure strategies, where τ_c denotes the strategy obtained from τ by interchanging the roles of the two arms.

Theorem 9.1.5 Suppose that the set Θ of possible pairs (θ_1, θ_2) of Bernoulli parameters is symmetric. Then there exists a symmetric minimax strategy, and the supremum in (9.1.4) may be taken over symmetric probability measures on Θ. If, in addition, Θ is closed, there exists a symmetric maximin strategy.

Proof Let S_0 denote a $(\Theta, \mathbf{A})$-minimax strategy (cf. Theorem 9.1.1). Define S_c by $S_c(T) = S_0(\{\tau: \tau_c \in T\})$, with τ_c as in Definition 9.1.3. From (9.1.1), (9.1.2), and the symmetry of Θ it is clear that S_c is also minimax. Let $S = (S_0 + S_c)/2$. From (9.1.1) and (9.1.2), we obtain

$$R(\theta_1, \theta_2; \mathbf{A}; S) = [R(\theta_1, \theta_2; \mathbf{A}; S_0) + R(\theta_1, \theta_2; \mathbf{A}; S_c)]/2.$$

So,

$$\sup_{(\theta_1, \theta_2) \in \Theta} R(\theta_1, \theta_2; \mathbf{A}; S)$$
$$\leqslant \left[\sup_{(\theta_1, \theta_2) \in \Theta} R(\theta_1, \theta_2; \mathbf{A}; S_0) + \sup_{(\theta_1, \theta_2) \in \Theta} R(\theta_1, \theta_2; \mathbf{A}; S_c) \right] \Big/ 2$$

which, by (9.1.3) and the fact that S_0 and S_c are minimax, equals $R_U(\Theta; \mathbf{A})$. Hence, S is minimax—and it is obviously symmetric. The proof for symmetric strategies for nature is very similar and is omitted. □

9.2 A continuous-time example

The continuous-time minimax example presented here is due largely to Bather (1983). There are two arms. Arm 2 generates a process which is identically zero. The process Y_1 generated by arm 1 is given by

$$Y_1(t) = \theta_1 t + B(t),$$

where B denotes standard Brownian motion and θ_1 is an unknown real number. The discount function is $e^{-\beta t}$ for some fixed $\beta > 0$.

We will not develop a general continuous-time theory and so no general terminology or notation will be introduced. Obvious analogues of concepts and notation introduced in Section 9.1 will be used.

We first hypothesize a particular form of the solution and then verify the correctness of that form while simultaneously obtaining conditions that the parameters in that form must satisfy. We conjecture that there exist three constants $b < 0$, $a > 0$, and $p \in (0, 1)$ such that a maximin strategy assigns probability p to $at + B(t)$ and probability $1-p$ to $bt + B(t)$. In view of Section 8.1 we also conjecture that a minimax strategy is the optimal pure strategy τ for the decision maker obtained in Section 8.1 for a two-point prior distribution on θ_1; τ indicates arm 1 as long as $Y_1(t) > qt + r$ and arm 2 after any t for which $Y_1(t) \leqslant qt + r$, where

$$q = (a+b)/2 \tag{9.2.1}$$

$$r = -\frac{1}{a-b}\log\frac{(1-C)p}{C(1-p)}, \tag{9.2.2}$$

with

$$C = \frac{-b(b-a+\sqrt{[8\beta+(a-b)^2]}}{(a-b)(b+a+\sqrt{[8\beta+(a-b)^2]}}. \tag{9.2.3}$$

The regret associated with τ is

$$R(\theta_1; e^{-\beta t}; \tau) = \beta^{-1}\Big(\theta_1 \vee 0 - \theta_1 E(1 - \exp[-\beta \inf\{t: \theta_1 t + B(t) \leqslant qt + r\}])\Big).$$

To substantiate the conjectures we show that b and a can be chosen so that $R(\theta_1; e^{-\beta t}; \tau)$ takes on its maximum value at both b and a. Take $p > C$; then $r < 0$ and

$$R(\theta_1; e^{-\beta t}; \tau) = \beta^{-1}\Big(\theta_1 \vee 0 - \theta_1 E(1 - \exp[-\beta \inf\{t: B(t) = (q-\theta_1)t + r\}])\Big). \tag{9.2.4}$$

The Laplace transform in (9.2.4) is known (for instance, Fristedt, 1974,

page 351); (9.2.4) reduces to

$$R(\theta_1; e^{-\beta t}; \tau) = \beta^{-1}\Big(\theta_1 \vee 0 - \theta_1(1-\exp(r\{\theta_1 - q + \sqrt{[2\beta+(\theta_1-q)^2]}\}))\Big). \quad (9.2.5)$$

For $\theta_1 \geq 0$,

$$\frac{dR}{d\theta_1} = \beta^{-1}\left(1 + r\theta_1\left\{1 + \frac{\theta_1 - q}{\sqrt{[2\beta+(\theta_1-q)^2]}}\right\}\right) \exp(r\{\theta_1 - q + \sqrt{[2\beta+(\theta_1-q)^2]}\}).$$

Differentiation shows that

$$1 + r\theta_1\left\{1 + \frac{\theta_1 - q}{\sqrt{[2\beta+(\theta_1-q)^2]}}\right\}$$

is a decreasing function of θ_1 since $r < 0$. This function is positive at 0 and negative for large θ_1. Hence, on $[0, \infty)$, $R(\cdot; e^{-\beta t}; \tau)$ increases to a maximum and then decreases. This maximum will occur at a if and only if

$$1 + ra\left\{1 + \frac{a-q}{\sqrt{[2\beta+(a-q)^2]}}\right\} = 0. \quad (9.2.6)$$

For $\theta_1 < 0$, let

$$v_1 = -r\{[2\beta+(\theta_1-q)^2]^{1/2} + \theta_1 - q\}.$$

When $\theta_1 = 0$, $v_1 = -r\{[2\beta+q^2]^{1/2} - q\} > 0$. As $\theta_1 \to -\infty$, $v_1 \to 0$. Moreover, v_1 is an increasing function of θ_1 and so that relationship can be inverted:

$$\theta_1 = \frac{2\beta r^2 + 2qrv_1 - v_1^2}{2rv_1}. \quad (9.2.7)$$

Thus, the problem of maximizing $R(\theta_1; e^{-\beta t}; \tau)$ for $-\infty < \theta_1 \leq 0$ is equivalent to the problem of maximizing

$$f(v_1) = \frac{-2\beta r^2 - 2qrv_1 + v_1^2}{2\beta r v_1}(1 - e^{-v_1}),\ 0 < v_1 \leq -r\{[2\beta+q^2]^{1/2} - q\}.$$

Differentiating f we find

$$2\beta|r|v_1^2 e^{v_1} f'(v_1) = -(2\beta r^2 + v_1^2)(e^{v_1} - 1) + 2\beta r^2 v_1 + 2qrv_1^2 - v_1^3 = g(v_1), \tag{9.2.8}$$

say, and differentiating again,

$$g'(v_1) = -(2\beta r^2 + 2v_1 + v_1^2)(e^{v_1} - 1) + 4qrv_1 - 4v_1^2. \tag{9.2.9}$$

Suppose $g(v) = 0$. Then from (9.2.8),

$$4qrv_1 = (4\beta r^2 v_1^{-1} + 2v_1)(e^{v_1} - 1) - 4\beta r^2 + 2v_1^2,$$

which we substitute into (9.2.9) for $4qrv_1$ to obtain

$$\begin{aligned} v_1 g'(v_1) &= 4\beta r^2 (e^{v_1} - 1) - (2\beta r^2 v_1 + v_1^3)(e^{v_1} + 1) \\ &\leqslant 2\beta r^2 [2(e^{v_1} - 1) - v(e^{v} + 1)] \end{aligned}$$

which is negative for $v_1 > 0$ since all the coefficients in the Taylor's series (in powers of v_1) are nonpositive. Hence, $g(v_1) = 0 \Rightarrow g'(v_1) < 0$, so g is zero for at most one value of v_1 and, in case such a v_1 exists, g is positive for smaller values of v_1, and negative for larger values. In view of (9.2.7) and (9.2.8), the maximum of R for $\theta_1 < 0$ will occur at b if and only if

$$b = \frac{2\beta r^2 + 2qrv_1 - v_1^2}{2rv_1} \tag{9.2.10}$$

and

$$2\beta r^2 v_1 + 2qrv_1^2 - v_1^3 = (2\beta r^2 + v_1^2)(e^{v_1} - 1). \tag{9.2.11}$$

In view of (9.2.5), the requirement that $R(a; e^{-\beta t}; \tau) = R(b; e^{-\beta t}; \tau)$ becomes

$$\begin{aligned} &a \exp(r\{a - q + \sqrt{[2\beta + (a-q)^2]}\}) \\ &\quad = -b(1 - \exp(r\{b - q + \sqrt{[2\beta + (b-q)^2]}\}). \end{aligned} \tag{9.2.12}$$

According to the preceding discussion we wish to find a, b, p, q, r, v_1, and C satisfying (9.2.1), (9.2.2), (9.2.3), (9.2.6), (9.2.10), (9.2.11), and (9.2.12). In addition, these inequalities must hold: $r < 0$, $0 < C < p < 1$, $b < 0 < a$, $0 < v_1 < -r\{[2\beta + q^2]^{1/2} - q\}$. Verifying that these conditions can be satisfied simultaneously will show that a minimax strategy is to use arm 1 until $Y_1(t) = qt + r$ and thereafter use arm 2, and that a 'least favorable distribution', a maximin strategy for nature, is given by $\theta_1 = a$ with probability p and $\theta_1 = b$ with probability $1 - p$.

For any $b < 0 < a$, C as given by (9.2.3) lies in (0, 1). For $C \in (0, 1)$, $r < 0$ and $b < 0 < a$, (9.2.2) determines a unique $p \in (C, 1)$. Accordingly, we may focus our attention on finding a, b, q, r, and v_1 satisfying (9.2.1), (9.2.6), (9.2.10), (9.2.11), and (9.2.12) together with the inequalities $b < 0 < a$ and $0 < v_1 < -r\{[2\beta + q^2]^{1/2} - q\}$.

In view of (9.2.1), equation (9.2.6) can be rewritten:

$$r\{[2\beta + (b-q)^2]^{1/2} + b - q\} + r^2 a\{[2\beta + (b-q)^2]^{1/2} - b + q\} = r(b-q). \tag{9.2.13}$$

Setting $b = \theta_1$ in (9.2.2) we obtain

$$v_1 = -r\{[2\beta + (b-q)^2]^{1/2} + b - q\}.$$

Since

$$\begin{aligned}[2\beta + (b-q)^2]^{1/2} - b + q &= 2\beta\{[2\beta + (b-q)^2]^{1/2} + b - q\}^{-1} \\ &= -2\beta r v_1^{-1},\end{aligned}$$

(9.2.13) can be replaced by

$$-v_1 - 2\beta r^3 v_1^{-1} a = r(b-q). \tag{9.2.14}$$

Equations (9.2.1), (9.2.10), and (9.2.14) may be regarded as three linear equations in the unknowns a, b, and q. With k denoting $2\beta v_1^{-2} r^2$, the solution of this system of linear equations is

$$b = \sqrt{2\beta}\,(1 + k + 2kv_1 - 2k^2 v_1)/(2k^{3/2} v_1), \tag{9.2.15}$$

$$a = \sqrt{2\beta}\,(1 + k)/(2k^{3/2} v_1), \tag{9.2.16}$$

$$q = \sqrt{2\beta}\,(1 + k + kv_1 - k^2 v_1)/(2k^{3/2} v_1). \tag{9.2.17}$$

Substitution for q in (9.2.11) and replacing r by $-v_1(k/2\beta)^{1/2}$ gives a quadratic equation for k. It has one positive solution, namely

$$k = \frac{2v_1 + e^{v_1} + [(2v_1 + e^{v_1})^2 + 4(2v_1 + 1 - e^{v_1})]^{1/2}}{2(2v_1 + 1 - e^{v_1})}, \tag{9.2.18}$$

if $2v_1 + 1 - e^{v_1} > 0$, and no positive solutions otherwise. Accordingly, we want a positive v_1 small enough for $2v_1 + 1 - e^{v_1}$ to be positive.

Substitution for a, b, q, and $r = -v_1(k/2\beta)^{1/2}$ in (9.2.12) gives an equation in the single variable v_1. The intermediate value theorem applied to that equation gives the existence of a solution v_1. That solution determines k, $r < 0$, q, a, and b via (9.2.18), the relation $r = -v_1(k/2\beta)^{1/2}$, (9.2.17), (9.2.16), and (9.2.15). Moreover,

numerical answers are easy to obtain: $v_1 \approx 0.1292$, $k \approx 12.27$, $r \approx -0.3199\beta^{-1/2}$, $q \approx -0.5845\beta^{1/2}$, $a \approx 1.690\beta^{1/2}$, $b \approx -2.859\beta^{1/2}$. The reader can easily check that the desired inequalities are satisfied. From (9.2.3) and (9.2.2) we obtain $C \approx 0.1212$ and $p \approx 0.3715$.

The minimax strategy is to observe arm 1 until the process it generates meets the line given by

$$-0.3199\beta^{1/2} - 0.5845\beta^{1/2}t,$$

and then switch to arm 2. The maximin strategy has θ_1 equal to $1.690\beta^{1/2}$ with probability 0.3715, and θ_1 equal to $-2.859\beta^{1/2}$ otherwise.

If $\theta_1 > -0.5845\beta^{1/2}$, then there is positive probability that the decision maker using the minimax strategy will never use arm 2, even if $\theta_1 < 0$.

References

Bather, J. A. (1983) Personal communication.

Bather, J. A. and Simons, G. (1985) The minimax risk for clinical trials. *J.R. Statist. Soc.* B (to appear).

Feller, W. (1968) *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd edn, Wiley, New York.

Fristedt, B. (1974) Sample functions of stochastic processes with stationary, independent increments. *Advances in Probability*, Vol. 3 (eds P. Ney and S. Port), pp. 241–396, Marcel-Dekker, New York.

Parthasarathy, T. and Raghavan, T. E. S. (1971) *Some Topics in Two-Person Games*, American Elsevier, New York.

Vogel, W. (1960a) A sequential design for the two-armed bandit. *Ann. Math. Statist.* **31**: 430–443.

Vogel, W. (1960b) An asymptotic minimax theorem for the two-armed bandit problem. *Ann. Math. Statist.* **31**: 444–451.

Annotated bibliography

The papers included here deal with 'bandit problems' as described in this monograph. The essential criterion for inclusion is that the problem addressed, directly or implicitly, is one of maximizing a sum (or an integral in the continuous case), perhaps with discounting. The discount factors can be random but they cannot depend on the observations on the arms. This means that we have excluded the very large literature of 'ranking and selection' problems in which various allocation rules are proposed and data-dependent stopping rules are used.

On the other hand, we include papers that consider particular allocation rules and allow a shift in allocation to a single arm at a stage that is determined by the data, as long as the results of those stages occurring later than the shift-point are considered in the objective function. The papers in this category are listed under Colton (1963); 'See Colton (1963)' is indicated under the 'related references' for such a paper.

We have included another set of papers that does not fit in perfectly with our formulation. These are the so-called 'finite-memory' and 'finite-state' bandit problems. All are listed under Robbins (1956); and 'See Robbins (1956)' is indicated under the 'related references' for such a paper.

We make no claim of thoroughness either in the papers included or in our descriptions of the contents of the papers. Entries without annotations were not available to us. For consistency we use our own terminology and notation throughout. We point out what we perceive to be errors in the original papers, and occasionally say why we disagree with an approach or a viewpoint of the authors. Many of our criticisms apply to a number of papers even though they may be given only once.

Not all of the papers listed make contributions. A surprising number of papers duplicate results proven earlier; ofttimes the earlier proof is more elegant. Certain early papers – notably Thompson (1933), Robbins (1952), and Bradt, Johnson, and Karlin (1956) – continue to be important for their creativity and elegance. Many later papers have not made full use of the ideas in these early papers.

Abdel Hamid, A. R. (1981) Randomized sequential decision rules. D. Phil. thesis, Univ. of Sussex, England.

Anscombe, F. J. (1963) Sequential medical trials. *J. Amer. Statist. Assoc.* **58**: 365–383.

The setting is essentially the same as that of Colton (1963). Anscombe considers the sequential stopping problem during the experimental phase from a Bayesian point of view. He considers both fixed n and a fixed length for the terminal phase; only the first is consistent with our bandit approach. Stopping boundaries are provided in terms of the difference between sample means on the two arms. These have been considered by a number of authors since (e.g., Lai, Robbins, and Siegmund, 1983) and have been shown to be asymptotically optimal.

Anscombe notes that assessing the horizon n is critical and that his results are quite sensitive to n (see also Upton and Lee, 1976). While we agree, we have three reasons for being somewhat less concerned than Anscombe. Firstly, information concerning the horizon can be updated during the course of the trial. So the boundary can be changed as the first phase develops, or a second 'first phase' can be started during the 'terminal phase'! A related approach allows multiple phases in a 'two-phase look-ahead' fashion as suggested by Cornfield, Halperin, and Greenhouse (1969). Secondly, when, as is likely to be dictated in practice by logistical limitations, the length of the first phase is fixed in advance, then this length varies only as $\sqrt{n}$ (Colton, 1963; Canner 1970; and others). Finally, an unknown horizon can be reflected in a prior distribution on n in the same way that a distribution is placed on the other unknown parameters in the problem (cf. our Chapter 3). (Witmer (1983) considers a special case in which n can be replaced by its expectation.) Indeed, n may depend on these parameters – at least to an extent.

The importance of Anscombe's paper greatly transcends the eminently reasonable study described in the first paragraph above. Anscombe criticizes a book by P. Armitage [1961, *Sequential Medical Trials*, Blackwell Scientific Publications, Oxford, England] in a most elegant fashion. Anscombe's paper should be studied carefully by all biostatisticians. The sequential procedures proposed by Armitage pay heed to the possibility of obtaining data that were not obtained and to decisions that might have been made but were not. The result is that a sequential trial can actually involve a greater expected number of patients than a trial with a fixed sample size! Anscombe argues that decisions (to stop sampling, for example) should be made on the basis of the data at hand and considering the costs due to ineffective treatment. Unfortunately, it is Armitage's work and not Anscombe's that has influenced current biostatistical practice. The designs of modern clinical trials are dictated by considerations of statistical power which allow 'proper' inferences. However, classical statistical tests do not attempt to balance the welfare of the present patient against the value of eventually obtaining a 'sound' statistical conclusion. While explicit consideration of the patient

horizon may be objectionable on ethical grounds, at least it makes clear the extent to which current patients are being sacrificed for future patients. Ethical dilemmas are not solved by avoiding them.

Related references: See Colton (1963).

Bather, J. A. (1977) A simple bandit problem. *Markov Decision Theory* (eds H. C. Tijms and J. Wessels), pp. 213–220, Mathematisch Centrum, Amsterdam.

There are two Bernoulli arms: $\theta_2 = 1/2$ and θ_1 is known to be either a or $1-a$ where $a \in (1/2, 1)$. Discounting is uniform. Bather focuses on fixed strategies as the horizon approaches ∞. He shows that there is no nonrandomized stationary strategy for which the proportion of successes approaches $\max\{\theta_1, 1/2\}$ as the horizon approaches ∞ unless the prior probability of $\{\theta_1 = a\}$ is 0 or 1. He gives a class of stationary randomized strategies that do have this characteristic.

Related references: Our Example 5.4.1; Bather (1980).

Bather, J. A. (1980) Randomized allocation of treatments in sequential trials. *Adv. in Appl. Probab.* **12**: 174–182.

There are k independent Bernoulli arms with uniform discounting. Bather considers a family of strategies, one of which is the following: arm i is selected at stage m if i is such that $[s_i + Y_i(m)]/[s_i + f_i]$ is maximized, where s_i and f_i are the numbers of successes and failures on arm i in the first $m-1$ stages, and $\{Y_i(m): 1 \leqslant i \leqslant k, m \geqslant 1\}$ is a family of independent exponentially distributed random variables.

Suppose the horizon approaches ∞. Bather shows that the proportion of successes approaches the maximum of the k Bernoulli parameters. Percus and Percus (1984) consider a similar strategy in which a large constant plays the role of $Y_i(m)$; in the latter circumstance there is a positive probability of selecting an inferior arm at every stage (except for some trivial cases).

Related references: Bather (1977), Fox (1974), Percus and Percus (1984), Robbins (1952), Thompson (1933, 1935).

Bather, J. A. (1981) Randomized allocation of treatments in sequential experiments (with discussion). *J. R. Statist. Soc. B* **43**: 265–292.

This is a comprehensive article that discusses various contributions of the papers listed below as 'related references', presents some new ideas, and evaluates a large number of procedures. Also notable is the lengthy and varied discussion included with this article.

Bather's main objective is to investigate strategies for k-armed Bernoulli bandits that have good asymptotic characteristics but that also perform well in the finite-horizon case. He recommends a randomized strategy (investigated further in Bather, 1984) that selects each arm sufficiently often to be asymptotically optimal; moreover, as calculations for $k = 2$ show, it performs

reasonably well for moderate n over a variety of (θ_1, θ_2) values. (We prefer a Bayesian approach which averages over the uncertainty in (θ_1, θ_2) as well as n, and gives rise to a nonrandomized strategy. However, we admit that randomized strategies have a certain appeal to practitioners.)

Bather uses diffusion approximations as an aid to understanding the asymptotic behaviour of randomized strategies (cf. Bather, 1983a).

Related references: Bather (1980), Bellman (1956), Berry (1978), Berry and Fristedt (1979), Bradt, Johnson, and Karlin (1956), Fabius and van Zwet (1970), Feldman (1962), Fox (1974), Gittins (1979), Glazebrook (1980), Kelly (1981), Poloniecki (1978), Robbins (1952), Rodman (1978), Vogel (1960a, 1960b), Wahrenberger, Antle, and Klimko (1977), Whittle (1980), Zelen (1969).

Bather, J. A. (1983a) The minimax risk for the two-armed bandit problem. *Mathematical Learning Models–Theory and Algorithms* (eds U. Herkenrath, D. Kalin and W. Vogel), pp. 1–11, Springer-Verlag, New York.

It is shown that c_3 in our Corollary 9.1.4 may be chosen to be 0.305, provided that n is sufficiently large. To show this, Bather introduces a modified bandit in which two arms must be selected at each stage. One of the two arms selected (which one is designated by the decision maker) does *not* contribute immediately to the worth of the strategy, but does give information that may be useful in the future. The other arm selected contributes to the worth but is *not* observed. When the decision maker is constrained to select the same arm for observation and for information, then of course the modified bandit is equivalent to the original bandit. So the *regret* (see Section 9.1) is no larger in the modified problem than in the original problem.

Another technique that Bather uses is that of approximating Bernoulli processes with diffusion processes.

Related references: Our Chapter 9; Bather (1981), Bather and Simons (1985), Fabius and van Zwet (1970), Vogel (1960a, 1960b).

Bather J. A. (1983b) Optimal stopping of Brownian motion: a comparison technique. *Recent Advances in Statistics; Papers in Honor of Herman Chernoff on his Sixtieth Birthday* (eds M. H. Rizvi, J. S. Rustagi and D. Siegmund), pp. 19–49, Academic Press, New York.

Time is continuous and the discount function is $e^{-\beta t}$ (see our Chapter 8). Bather considers an arm that is the sum of Brownian motion and a linear drift whose value at $t = 1$ is normally distributed with mean μ and variance ρ^2. Bather shows that the break-even value of such an arm is $\Lambda((\mu, \rho), \beta) = \mu + \beta^{1/2}\psi(\rho/\beta)$, where ψ is a function satisfying: $\psi(v)/v \to 2^{-1/2}$ as $v \to 0$; $\psi(v)/v \leqslant 2^{-1/2}$ for $v > 0$; $\psi(v)/v^{1/2} \uparrow \infty$ as $v \uparrow \infty$. Bather gives some evidence which suggests that $\psi(v)/(2v \log v)^{1/2}$ approaches 1 as $v \to \infty$.

Bather addresses this only as a stopping problem, but one may be tempted

to apply his result in k-armed bandits as in Gittins and Jones (1974). However, as far as we know, a general continuous-time version of the Gittins–Jones theorem has not been developed. So we do not know whether the value of Λ for each of the available arms determines an optimal strategy for a multi-armed bandit.

Related references: Our Chapter 8; Breakwell and Chernoff (1964), Chernoff and Ray (1965), Karatzas (1984).

Bather, J. A. (1984) Towards a more rational allocation of treatments in medical trials. Unpublished.

A strategy for two-armed Bernoulli bandits proposed in Bather (1981) is investigated and compared with two other strategies outside a bandit setting. Namely, a conclusion concerning the sign of $\theta_1 - \theta_2$ is desired without exposing an excessive number of patients to inferior treatment.

This paper is included here for its similarity with various listed papers.

Related references: See Colton (1963).

Bather, J. A. and Simons, G. (1985) The minimax risk for clinical trials. *J. R. Statist. Soc. B* (to appear).

Numerical evidence is given supporting the conjecture that c_3 and c_4 in our Corollary 9.1.4 may be chosen to be 0.371 and 0.372, respectively, provided that n is sufficiently large. Evidence is also given that the following strategy is very close to minimax for a Bernoulli two-armed bandit with n-horizon uniform discounting: calculate the nearest integer K_n to $0.292n^{1/2}$; use the arms equally until the number of successes on one arm exceeds the number on the other arm by K_n; thereafter use the arm on which the greater number of successes has been observed.

The methods of Vogel (1960a, 1960b) are explained, used, and extended. In particular, the process of maximizing the risk over (θ_1, θ_2) in the unit square and then minimizing over K_n is described in detail.

Related references: Our Section 9.1; Bather (1983a), Fabius and van Zwet (1970); and see Colton (1963).

Beckmann, M. J. (1973) Der diskontierte Bandit. *OR-Verfahren* **XVIII**: 9–18.

Begg, C. B. and Mehta, C. R. (1979) Sequential analysis of comparative clinical trials. *Biometrika* **66**: 97–103.

The problem as posed by Anscombe (1963) is reconsidered. The authors consider the rule which terminates the experimental phase when no fixed size continuation of the phase is an improvement on stopping. Chernoff and Petkau (1981) consider continuous-time versions of this rule and the rule of Anscombe (1963) and conclude that the former tends to stop much too soon.

Related references: See Colton (1963).

Bellman, R. (1956) A problem in the sequential design of experiments. *Sankhyā A* **16**: 221–229.

There are two Bernoulli arms: θ_1 has distribution F and θ_2 is known (cf. our Chapter 5). Discounting is geometric: $\mathbf{A} = (1, \alpha, \alpha^2, \ldots)$. Bellman shows that the dynamic programming equations (our (2.5.2)) have a unique solution that is the limit of a recursion (see our Chapter 2 for a more general treatment). He shows the existence of a break-even value of θ_2, $\Lambda(F, \mathbf{A})$, such that arm 1 is optimal if and only if $\Lambda(F, \mathbf{A}) \geqslant \theta_2$ (our Corollary 5.1.2 and Theorem 5.3.1 give two different generalizations). He also shows that $\Lambda(\sigma F, \mathbf{A}) > \Lambda(F, \mathbf{A}) > \Lambda(\varphi F, \mathbf{A})$; cf. our Corollary 5.3.3 (we note that the strict inequalities mean that Bellman is implicitly assuming that F is not a one-point distribution).

Related references: Our Chapters 2, 5, and 6; Berry and Fristedt (1979), Gittins and Jones (1974), Gittins and Jones (1979), Kakigi (1983), Yakowitz (1969).

Bellman, R. (1961) *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, New Jersey.

The dynamic programming equations of Bellman (1956) are given in Section 16.15; they represent part of a more general discussion.

Related reference: Bellman (1956).

Benzing, H., Hinderer, K. and Kolonko, M. (1984) On the k-armed Bernoulli bandit: Monotonicity of the total reward under an arbitrary prior distribution. *Math. Operationsforsch. Statist. Ser. Optim.*, **15**: 583–595.

There are k Bernoulli arms with n-horizon uniform discounting. Using the result indicated in our discussion of Hengartner, Kalin, and Theodorescu (1981), the authors allow θ_1 and θ_2 to be dependent and determine conditions on G which satisfy the hypothesis of the result of that paper.

Related references: Hengartner, Kalin, and Theodorescu (1981), Kolonko and Benzing (1985).

Benzing, H. and Kolonko, M. (1984) Structured policies for a sequential design problem with general distributions. Unpublished.

There are two arms with one arm known. The unknown arm has an arbitrary distribution. Discounting is uniform. The authors prove special cases of results in Berry and Fristedt (1979); see also our Chapter 5.

Related references: Our Chapter 5; Berry and Fristedt (1979), Bradt, Johnson, and Karlin (1956), Kolonko and Benzing (1983, 1985).

Berry, D. A. (1972) A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43**: 871–897.

There are two independent Bernoulli arms. The discount sequence is the

n-horizon uniform. The stay-with-a-winner rule (our Theorem 4.3.8) is proved in this setting. This paper contains some of the results given in our Chapter 7, but the methods are somewhat different.

Related references: Our Sections 4.2 and 4.3 and Chapter 7; Berry (1978), Bradt, Johnson and Karlin (1956).

Berry, D. A. (1978) Modified two-armed bandit strategies for certain clinical trials. *J. Amer. Statist. Assoc.* **73**: 339–345.

There are two Bernoulli arms and discounting is finite-horizon uniform. Easy-to-use strategies are suggested for clinical trials. These strategies are optimal when there are two Bernoulli arms with two-point prior distributions, as in our Corollary 4.3.10. These strategies are adapted to beta distributions and their worths are compared with optimal worths for various distributions and horizons.

Related references: Our Chapter 7; Bather (1981), Berry (1972), Percus and Percus (1984), Thompson (1933, 1935).

Berry, D. A. (1983) Bandit problems with random discounting. *Mathematical Learning Models–Theory and Algorithms* (eds U. Herkenrath, D. Kalin and W. Vogel), pp. 12–25, Springer-Verlag, New York.

Many-armed bandits with random discount sequences are considered. The main results in this paper are given in our Chapter 3.

Related references: Our Chapter 3; Berry and Fristedt (1979).

Berry, D. A. (1985). One- and two-armed bandit problems. *Encyclopedia of Statistical Sciences*, Vol. VI (eds by S. Kotz and N. L. Johnson), Wiley, New York (to appear).

This is a survey article. Its perspective is similar to that of this monograph.

Berry, D. A. and Fristedt, B. (1979). Bernoulli one-armed bandits – Arbitrary discount sequences. *Ann. Statist.* **7**: 1086–1105.

Bernoulli bandits with one unknown arm are considered when discounting is arbitrary. This paper provides the basis for parts of Chapter 5 of this monograph. In particular, it gives a necessary and sufficient condition for the bandit to be a stopping problem for all θ_2 and F_1.

Related references: Our Chapter 5; Bellman (1956), Berry (1983), Bradt, Johnson and Karlin (1956), Gittins (1979), Gittins and Jones (1974).

Berry, D. A. and Fristedt, B. (1980a) Two-armed bandits with a goal, I. One arm known. *Adv. in Appl. Probab.* **12**: 775–798.

The problem considered is an adaptation of a classical bandit problem. One of two arms is selected at each of a possibly infinite number of stages and

learning takes place as the results are observed. But now the decision maker strives to turn a current fortune into a given larger fortune (the goal) before it becomes a smaller fortune (ruin). Observing an arm either increases or decreases the fortune by one unit. One of the arms has a known probability of increase and that of the other is unknown. Optimal strategies are shown to exist in great generality and are characterized for some special prior distributions on the arms. In some settings it is shown that the 'risky' arm – the unknown arm – should be selected when the decision maker is near ruin, but not otherwise.

Related references: Berry and Fristedt (1980b), Gittins (1983), Vogel (1961b).

Berry, D. A. and Fristedt, B. (1980b) Two-armed bandits with a goal, II. Dependent arms. *Adv. in Appl. Probab.* **12**: 958–971.

The setting is similar to that of Berry and Fristedt (1980a). Now both probabilities of increase are known, but not which goes with arm 1. Optimal strategies are shown to exist. Myopic strategies (cf. our Section 1.2 and Corollary 4.3.10) are shown to be optimal for various pairs of probabilities but not optimal in general.

Related references: Berry and Fristedt (1980a), Gittins (1983), Vogel (1961b).

Berry, D. A. and Fristedt, B. (1983) Maximizing the length of a success run for many-armed bandits. *Stochastic Process. Appl.* **15**: 317–325.

The same problem is considered as in Berry and Viscusi (1981). Now the discount sequence is arbitrary. It is shown that there is always an optimal strategy that uses a single arm indefinitely whenever the arms are independent and the discount sequence is *superregular* (see our Proposition 5.2.1). Optimal strategies may or may not have this characteristic when the discount sequence is not superregular.

Related references: Berry and Viscusi (1981), Viscusi (1979a, 1979b).

Berry, D. A. and Pearson, L. (1984) Optimal designs for two-stage clinical trials with dichotomous responses. Univ. of Minnesota Tech. Rep. (To be in *Statist. in Med.*).

There are two Bernoulli arms with parameters θ_1 and θ_2. Discounting is n-horizon uniform. Two-phase designs are considered in which the first is an experimental or learning phase and a single arm is selected in the second. Allocations in the first phase must be set in advance. The length of the first phase is either given or to be optimized as a function of n and the prior information (cf. Pearson, 1980; Witmer, 1983). Two forms of prior distributions are considered: (i) θ_2 is known while the distribution of θ_1 is arbitrary, and (ii) the values of θ_1 and θ_2 are known but not which is θ_1 (cf. Feldman, 1962). Graphs are provided to show optimal first-phase lengths and maximal success proportions for various n.

When θ_2 is known and θ_1 has a uniform distribution on (0, 1), the authors show that the optimal first-phase length is approximately $[(n+1)(1/\theta_2 - 1)]^{1/2} - 2$. This is similar to Canner's (1970) result which applies for both θ_1 and θ_2 uniform and equal first-phase allocations.

Related references: See Colton (1963).

Berry, D. A. and Viscusi, W. K. (1981) Bernoulli two-armed bandits with geometric termination. *Stochastic Process. Appl.* **11**: 35–45.

The standard Bernoulli bandit problem is modified so that all payoff ceases with the first failure. The discount sequence is either $\mathbf{A} = (0, \ldots, 0, 1, 0, \ldots)$ or geometric (as in our Chapter 6). It is shown that the search for an optimal strategy can be restricted without loss to 'single-arm' strategies – those which stay indefinitely with the arm selected initially. An example shows that the optimal strategies do not have this form for all discount sequences.

Related references: Berry and Fristedt (1983), Viscusi (1979a, 1979b).

Bradt, R. N., Johnson, S. M. and Karlin, S. (1956) On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.* **27**: 1060–1074.

There are two Bernoulli arms. Discounting is n-horizon uniform. The most important contributions of this paper are for the case in which one arm is known. In particular, a number of the results in our Chapter 5 are generalizations of results in this paper. One such result is worth mentioning here only because it has not been appreciated by a number of later authors: Lemma 4.1 says that if the known arm is optimal at any stage then it is also optimal thereafter (see the discussion preceding the proof of our Theorem 5.2.2).

The authors touch on a number of issues when both arms are unknown. They completely characterize optimal strategies when $n = 2$ (cf. our Section 7.2). They give an example (our Example 2.4.1) in which 'stay-with-a-loser/switch-from-a-winner' is optimal. Also, our Example 2.4.5 is similar to one of their examples. They indicate that it is possible for a one-step look-ahead strategy (i.e., a myopic strategy) to be better than a two-step look-ahead strategy. And they discuss the problem (they call it the 'classical' problem) that was eventually solved by Feldman (1962).

This paper is both innovative and comprehensive. Almost everyone writing about bandit problems refers to this paper, but not all have read it! Results proven by Bradt, Johnson, and Karlin in most elegant ways are sometimes credited to other authors who were both very late and much less elegant.

Related references: Our Chapter 5 and Section 7.2; Berry and Fristedt (1979), Feldman (1962).

Brand, H., Sakoda, J. M. and Woods, P. J. (1957) Effects of a random versus pattern instructional set in a contingent partial reinforcement situation. *Psychol. Rep.* **3**: 473–479.

The hypothesis being tested is that subjects who select the more successful arm exclusively are those convinced of randomness while those who mix selections feel that there is a pattern in the responses. We note that if the subjects were actually using optimal strategies (assuming randomness) they would not select the more successful arm exclusively but would attempt to balance information – unless available information is sufficiently one-sided. So subjects trying to anticipate patterns may be confused with subjects who use good strategies.

Some subjects were told that there was a pattern concerning which of the two arms would result in success; others were told that successes were random. (The latter were given instructions similar to those described in Brand, Woods, and Sakoda (1956) even though $\theta_1 + \theta_2$ was not always equal to 1.) The experiment was actually conducted as independent Bernoulli processes using electronic devices. The authors note a difference in behaviour between the two groups only when θ_1 or θ_2 is 1.

As a general matter, it seems to us that the problem of how the subject perceives the producing mechanism is greatly alleviated by using physical rather than electronic devices. Physical devices would seem to be beneficial to any study since conclusions would not be muddled by such questions. Thus, for example, if the subject is given a choice between two urns, each containing balls of two colours one of which is preferable, then little explanation is required to make it clear that the two arms (urns) are independent. On the other hand, if one urn is used and the subject is asked to predict the colour of the ball drawn then it is transparent that $\theta_1 = 1 - \theta_2$ even though both θ_1 and θ_2 may be unknown. Of course, sampling with replacement is necessary to preserve conditional independence (given the θ_i), but sampling without replacement would also give rise to interesting experiments.

Related references: Brand, Woods, and Sakoda (1956), Bush and Mosteller (1955), Estes (1950), Horowitz (1973), Murray (1971), Woods (1959).

Brand, H., Woods, P. J. and Sakoda, J. M. (1956) Anticipation of reward as a function of partial reinforcement. *J. Exp. Psychol.* **52**: 18–22.

Results of an experiment are reported. Subjects were allowed to choose between two arms (electronic devices) in a series of trials and were told to try to get as many successes as possible. Various pairs (θ_1, θ_2) were assigned by the experimenter. While $\theta_1 + \theta_2$ was not always equal to 1, the subjects were told, 'On a particular trial, then, one level will be correct and one incorrect.' We do not see what is to be gained from giving the subjects this misinformation.

Related references: Brand, Sakoda, and Woods (1957), Bush and Mosteller (1955), Estes (1950), Horowitz (1973), Murray (1971), Woods (1959).

Breakwell, J. and Chernoff, H. (1964) Sequential tests for the mean of a normal distribution II (large t). *Ann. Math. Statist.* **35**: 162–173.

This paper is not directly related to bandit problems. The authors consider free boundary problems for partial differential equations. In particular, they address the question: Does a solution of the free boundary problem necessarily solve the original probabilistic problem? This question is important for Chernoff and Ray (1965) (cf. our Section 8.2).

Related references: Our Section 8.2; Chernoff and Ray (1965).

Bush, R. R. and Mosteller, F. (1955) *Stochastic Models for Learning*, Wiley, New York.

Chapter 13 deals with experiments involving sequential choices by subjects in various Bernoulli two-armed bandit settings; three-armed bandits are also discussed. An *ad hoc* linear model is proposed in which the probability that a subject selects an arm increases or decreases according as the arm was successful or not. The parameters in the model are estimated for various data and the model is found to fit quite well on an overall basis; cf. Cane (1962).

Since averaging results of all players combines 'good' and 'bad' strategies, both are obscured. Of greater interest to us is the possibility that *individual* players select optimally for some prior distribution. (Horowitz (1973) addresses this issue.) For example, if a player switches on a success then there is no prior distribution for which the player is behaving optimally (unless the arms are dependent or observations on the same arm are conditionally dependent).

Related references: Brand, Sakoda, and Woods (1957), Brand, Woods, and Sakoda (1956), Cane (1962), Estes (1950), Horowitz (1973), Murray (1971), Schmalansee (1975).

Cane, V. R. (1962) Learning and inference (with discussion). *J. R. Statist. Soc. A* **125**: 183–209.

Two-armed bandit experiments involving humans and rats are described. Cane discusses models that hypothesize recursions for the probability that a subject selects arm 1, say. She cites studies which she claims invalidate linear models as discussed by Bush and Mosteller (1955).

Related references: Bush and Mosteller (1955), Schmalansee (1975).

Canner, P. L. (1970) Selecting one of two treatments when the responses are dichotomous. *J. Amer. Statist. Assoc.* **65**: 293–306.

There are two independent Bernoulli arms with parameters θ_1 and θ_2 having beta distributions (though a minimax approach is also considered). Discounting is n-horizon uniform. The two arms must be used equally in the first $2r$ stages (cf. Berry and Pearson (1984) who allow unequal allocation). Subsequent to this experimental phase, the arm with greater current mean (probability of success) is used exclusively. The only unspecified part of the strategy is the choice of r. Canner gives tables of optimal values of r. When θ_1

and θ_2 are uniformly distributed he shows that r is approximately $\sqrt{(n/2+1)} - 1$ (cf. Berry and Pearson, 1984; Kelley 1976).

Related references: See Colton (1963).

Chernoff, H. (1967) Sequential models for clinical trials. *Fifth Berkeley Symp. of Math. Statist. and Probab.* **4**: 805–812.

There are two arms with arm 2 known. Discounting is n-horizon uniform. Chernoff approximates the Bernoulli case using a continuous-time approximation from Chernoff and Ray (1965); see also our Section 8.2. As an interesting sidelight, Chernoff calculates that the number of immediate failures that the decision maker should tolerate with arm 1 has order of magnitude $\log n$ when θ_1 is uniform on (0, 1) and θ_2 is fixed and not 0 or 1.

Chernoff touches briefly on various other issues: the possibility that both arms are unknown, geometric discounting, and the two-phase design of Anscombe (1963) and Colton (1963).

Related references: Our Chapter 5 and Section 8.2; Chernoff (1968, 1972), Chernoff and Ray (1965).

Chernoff, H. (1968) Optimal stochastic control. *Sankhyā A* **30**: 221–252.

A number of sequential problems are considered, including bandits. Time is continuous and discounting is uniform. A 'one-armed bandit' of Chernoff and Ray (1965) is discussed. Chernoff considers a two-armed bandit in which the arms are independent and are Brownian motions with unknown drifts and the same (known) variance. He conjectures that the solution of a one-armed bandit provides a bound for the solution of a two-armed bandit in the following sense. Consider modifying the two-armed bandit so that whenever arm 2 is selected it must be selected exclusively and indefinitely. If arm 2 is optimal with this restriction then it must be optimal without it. But the modified problem is equivalent to a one-armed bandit. The conjecture is still open although some variations of it have been resolved by Gittins (1975).

Related references: Chernoff (1972), Chernoff and Ray (1965), Gittins (1975).

Chernoff, H. (1972) *Sequential Analysis and Optimal Design*, SIAM, J. W. Arrowsmith, Bristol, England.

This monograph surveys a broad range of sequential statistical problems, including bandit problems (in Section 18). The discussion of bandits is similar to that in Chernoff (1968).

Related references: Chernoff (1968), Chernoff and Ray (1965), Gittins (1975).

Chernoff, H. (1975) Approaches in sequential design of experiments. In *A Survey of Statistical Design and Linear Models* (ed. J. N. Srivastava), pp. 67–90, North-Holland, Amsterdam.

This is a review article which discusses a variety of problems, including bandits.

Related reference: Feldman (1962).

Chernoff, H. and Petkau, A. J. (1976) An optimal stopping problem for sums of dichotomous random variables. *Ann. Probab.* **4**: 875–889.

An optimal stopping problem is discussed and shown to be related to the problem addressed by Petkau (1978), among others.

Related references: Chernoff and Petkau (1983, 1985), Petkau (1978).

Chernoff, H. and Petkau, A. J. (1981) Sequential medical trials involving paired data. *Biometrika* **68**: 119–132.

The problem posed by Anscombe (1963) is considered for continuous time. Anscombe's procedure is compared with various other procedures (including that of Begg and Mehta (1979)) and is shown to be well-approximated by the optimal continuous-time procedure.

Related references: See Colton (1963).

Chernoff, H. and Petkau, A. J. (1983) Numerical methods for Bayes sequential decision problems. Applied Mathematics and Statistics Tech. Rep. No. 83–26, Univ. of British Columbia, Canada.

The authors discuss computational issues for a number of statistical decision problems involving sums of random variables, including bandit problems. They use continuous-time approximations for discrete-time problems. The computational techniques go in the other direction: backwards induction is used for discrete-time processes to obtain approximations for continuous-time processes.

Information for constructing our Figures 8.2 and 8.3 was taken from Table 21 of this report.

Related references: Our Section 8.2; Chernoff (1967, 1968, 1972), Chernoff and Petkau (1976, 1981, 1985a, 1985b), Chernoff and Ray (1965), Petkau (1978).

Chernoff, H. and Petkau, A. J. (1985a) Sequential medical trials with ethical cost. *Proceedings of Kiefer-Neyman Conference* (to appear).

The setting is that of Anscombe (1963) and Colton (1963) except that an 'ethical cost' is now assessed for each pair of observations made in the experimental phase. This cost is proportional to the absolute difference in the current means of the two arms. The authors give optimal stopping boundaries for various proportionality factors.

Obviously, it is better to stop experimenting sooner with such a cost than without. A similar effect can be obtained by discounting future observations

(or, approximately, by specifying a smaller horizon), a modification we prefer to the one made in this paper.

Related references: Chernoff and Petkau (1976, 1981, 1983a); and see Colton (1963).

Chernoff, H. and Petkau, A. J. (1985b) Numerical solutions for Bayes sequential decision problems. *SIAM J. Sci. Statist. Comput.* (to appear).

This is an abbreviated version of Chernoff and Petkau (1983).

Chernoff, H. and Ray, S. N. (1965). A Bayes sequential sampling inspection plan. *Ann. Math. Statist.* **36**: 1387–1407.

There are two arms with one arm known; time is continuous and discounting is uniform. The decision problem is to decide when to stop observing the unknown arm. The unknown arm is the sum of Brownian motion and a normally distributed linear drift. The boundary of the stopping region is the free boundary for a partial differential equation with certain boundary conditions. (Our Section 8.2 is based on this development.)

The authors use asymptotic methods to calculate the boundary. They discuss applying these solutions as approximations to certain discrete-time bandits.

Related references: Our Section 8.2; Bather (1983b), Breakwell and Chernoff (1964), Chernoff (1968, 1972), Chernoff and Petkau (1983, 1985b).

Chung, F. (1984) Contributions to the multiarmed bandit problem. Ph.D. thesis, Columbia Univ., USA.

Clayton, M. K. (1983) Bayes sequential sampling for choosing the better of two populations. Ph.D. thesis, Univ. of Minnesota, USA.

A number of problems are considered; these include the following bandit problem. There are two independent arms whose prior distributions are Dirichlet measures, as described in our Section 5.6. The discount sequence is $\mathbf{A} = (0, 0, \ldots, 0, 1, 0, 0, \ldots)$, which means that one of the arms is to be selected at the termination of a number of sequential 'learning observations'.

Related references: Our Sections 3.6 and 5.6.

Clayton, M. K. and Berry, D. A. (1984) Bayesian nonparametric bandits. Statistics Tech. Rep. No. 427, Univ. of Minnesota, USA. (To be published.)

The substance of this paper is treated in our Section 5.6.

Related references: Our Chapter 5; Berry and Fristedt (1979), Bradt, Johnson, and Karlin (1956).

Colton, T. (1963) A model for selecting one of two medical treatments. *J. Amer. Statist. Assoc.* **58**: 388–400.

There are two independent normal arms and discounting is n-horizon uniform. An initial experimental phase is followed by a terminal phase in which the arm with greater mean after the initial phase is used exclusively. Allocations in the initial phase are equally divided between the two arms. Two settings are considered: (i) the length r of the first phase is fixed in advance, and (ii) the first phase is stopped as a function of the accumulating data which are assumed to be available immediately.

One might view the second as the more realistic approach and the first as an approximation; however, responses are delayed in most clinical trials and so sequential assignments may not be possible. An important set of problems and a fertile area of research is provided by the case in which information becomes available gradually after an arm is selected, or the response occurs only after a number of other selections must be made.

In the first problem Colton uses a 'local maximin' argument to show that the best value of r is $n/3$. We do not feel that this approach has merit here and, for reasons indicated in our discussion of Zelen (1969), we think that r/n should tend to 0 as $n \to \infty$. Using a Bayesian approach, Colton finds that r has order of magnitude $\sqrt{n}$. In the second problem, Colton studies procedures in which the first phase is stopped when the absolute difference in the sums on the two arms is greater than a constant times $\sqrt{n}$; this is similar to boundaries calculated by Anscombe (1963) and Vogel (1960a), but see Lai, Robbins, and Siegmund (1983).

The papers listed below all consider clinical trials in stages. However, unlike Colton, those listed with an asterisk restrict consideration to rules for which the ability to make classical statistical inferences is a primary consideration. Hence, the problem (though perhaps not the solution!) is on the fringe of what we call bandit problems. Those listed with two asterisks do not consider observations at stages beyond the experimental phase as part of the objective; these are listed in this bibliography only because there are a number of ways in which they are similar to other papers that are listed.

Related references: Anscombe (1963), Bather (1984)**, Bather and Simons (1985), Begg and Mehta (1979), Berry and Pearson (1984), Canner (1970), Chernoff and Petkau (1981), Colton (1965), Cornfield, Halperin, and Greenhouse (1969), Day (1969a, 1969b), Flehinger and Louis (1971)**, Flehinger and Louis (1972)**, Flehinger, Louis, Robbins, and Singer (1972)**, Fox (1974), Goto, Sugimura, and Asano (1971)*, Kelley (1976), Lai, Levin, Robbins, and Siegmund (1980)*, Lai, Robbins, and Siegmund (1983), Langenberg and Srinivasan (1981), Louis (1975)**, Meeter (1973), Oudin and Lellouch (1972), Pearson (1980), Petkau (1978), Robbins and Siegmund (1974)**, Upton and Lee (1976), Vogel (1960a, 1960b), Witmer (1983), Zelen (1969).

Colton, T. (1965) A two-stage model for selecting one of two treatments. *Biometrics* **21**: 169–180.

The setting is that of Colton (1963) except that there is a transitional phase between the experimental and terminal phases; really a three-phase design. Two simple procedures are compared.

Related references: See Colton (1963).

Cornfield, J., Halperin, M. and Greenhouse, S. W. (1969) An adaptive procedure for sequential clinical trials. *J. Amer. Statist. Assoc.* **64**: 759–770.

The setting is that of Anscombe (1963) and Colton (1963). The approach of these latter papers is generalized by allowing arbitrary prior mean and arbitrary imbalance in the first phase. They also consider applying their two-phase design repeatedly in an r-phase trial (a kind of 'two-phase look-ahead').

The authors derive an asymptotic expression for the optimal length of the first phase that generalizes that of Colton (1963); again it is of order $\sqrt{n}$.

Related references: See Colton (1963).

Cover, T. M. (1968) A note on the two-armed bandit problem with finite memory. *Inform. and Control* **12**: 371–377.

There are two Bernoulli arms. For a memory of size 2 and allowing the choice of arms to depend on the stage at which it is to be used, Cover argues that it is possible to achieve $\max\{\theta_1, \theta_2\}$ as the limiting proportion of success.

Related references: See Robbins (1956).

Cover, T. M. and Hellman, M. E. (1970) The two-armed bandit problem with time-invariant finite memory. *IEEE Trans. Inform. Theory* **16**: 185–195.

There are two arms and two distributions (arbitrary but known)—which distribution goes with which arm is not known. One of the two distributions is regarded as better than the other. The decision maker is allowed to remember only a finite-valued statistic. The authors find the supremum of the long-run proportion of observations from the better distribution; this supremum depends on the two distributions and on the size of the memory. They argue that there are no strategies which attain the supremum but exhibit a family whose members come arbitrarily close.

B. Chandrasekaran [1970, *IEEE Trans. Inform. Theory* **16**: 494–496; and 1971, *IEEE Trans. Inform. Theory* **17**: 104–105] argues that the strategies used actually require infinite memory. The authors respond in [1970, *IEEE Trans. Inform. Theory* **16**: 496–497].

Related references: Feldman (1962); and see Robbins (1956).

Cover, T. M. and Wagner, T. J. (1976) Topics in statistical pattern recognition. *Communication and Cybernetics* 10: *Digital Pattern Recognition* (ed. K. S. Fu), pp. 15–46, Springer-Verlag, Berlin.

This is a survey article of a variety of statistical areas of which finite-memory bandit problems constitute a small part.

Related references: See Robbins (1956).

Day, N. E. (1969a) Two-stage designs for clinical trials. *Biometrics* **25**: 111–118.

Extensions of Colton's (1965) three-phase designs are considered in the normal case. Various designs are compared numerically. Day proves in an appendix that equal allocation is optimal in the first phase when the arms are exchangeable *a priori*. While we believe the result, we note that his proof seems to apply as well when the arms are Bernoulli, a case in which Pearson (1980) shows that the result is false.

Related references: See Colton (1963).

Day, N. E. (1969b) A comparison of some sequential designs. *Biometrika* **56**: 301–311.

There are two independent normal arms with n-horizon discounting. Lucid descriptions and comparisons of three designs are given: (i) optimal (unrestricted) bandit strategies, (ii) pairwise allocation in the first of two phases as in Anscombe (1963) and Colton (1963), and (iii) myopic strategies.

Related references: See Colton (1963).

DeGroot, M. H. (1970) *Optimal Statistical Decisions*, McGraw-Hill, New York.

In Sections 14.5 to 14.7 DeGroot discusses two Bernoulli bandit problems with n-horizon uniform discounting. He solves an easy example in which one arm is known and gives the problem and development of Feldman (1962).

Related references: Our Chapter 5; Berry and Fristedt (1979), Bradt, Johnson, and Karlin (1956), Feldman (1962), Kelley (1974).

Dubins, L. E. and Savage, L. J. (1976) *Inequalities for Stochastic Processes: How to Gamble If You Must*, Dover, New York.

This book provides the fundamental theory for determining optimal decisions in a wide variety of problems. Chapters 2 and 3 constitute especially relevant background for bandit problems. In particular, the theorems concerning excessivity and thriftiness are useful for finding optimal strategies. Bandit problems in discrete time are considered in Sections 12.5 and 12.6 to show that they fall within the scope of the book. In a discussion similar to that which opens our Chapter 6, the authors argue that bandit problems with geometric discounting are more realistic than those with finite horizon.

These authors credit F. Mosteller with coining the term 'two-armed bandits'.

Related reference: Our Chapter 2.

Emrich, L. J. (1983) Optimal decision making using non-expert opinions. Ph.D. thesis, State Univ. of New York at Buffalo, USA.

The standard dynamic programming equations are developed and applied to a number of problems, including bandits. Finite horizon and geometric discounting are both considered. One of the problems Emrich considers is essentially the arm-acquiring bandit of Whittle (1981); the arms are independent and responses are normal.

Related references: Our Chapter 2; Fahrenholtz (1982), Gittins (1979), Whittle (1981, 1982).

Estes, W. K. (1950) Towards a statistical theory of learning. *Psychol. Rev.* **57**: 94–107.

This is an early paper in the literature of probabilistic learning models. However, it does not deal explicitly with bandit situations.

Fabius, J. and van Zwet, W. R. (1970) Some remarks on the two-armed bandit. *Ann. Math. Statist.* **41**: 1906–1916.

There are two Bernoulli arms which may be dependent. The discount sequence is n-horizon uniform. Randomized strategies are allowed. The authors give thorough demonstrations of the relationships among Bayes, admissible, and minimax strategies. Their results concerning Bayes strategies generalize the main result of Feldman (1962).

Related references: Our Corollary 4.3.10 and Section 9.1; Feldman (1962), Vogel (1960b, 1960c, 1964).

Fahrenholtz, S. K. (1982) Normal Bayesian two-armed bandits. Ph.D. thesis, Iowa State Univ., USA.

There are two normal arms and time is discrete. Discounting is n-horizon uniform. Two families of prior distributions are considered; variances are known in both. In one, the arms are independent with normally distributed means. In the second, the arms are dependent with means that sum to zero; the difference in means is normally distributed. Various properties are shown and bounds are calculated; some numerical calculations are given. Fahrenholtz shows that myopic strategies are optimal in the dependent case (cf. Rodman, 1978).

Related references: Our Section 2.1 and Chapter 7; Berry (1972), Feldman (1962), Rodman (1978).

Feldman, D. (1962) Contributions to the 'two-armed bandit' problem. *Ann. Math. Statist.* **33**: 847–856.

This important paper solved a problem that had eluded a number of statisticians: there are two Bernoulli arms with (θ_1, θ_2) equal to either (a, b) or

(b, a) and discounting is uniform. Feldman showed that myopic strategies are optimal: always select the arm with greater mean. This result has been generalized in a number of ways (see related references listed below).

In our view there is an unfortunate aspect of the notoriety associated with Feldman's problem. Feldman poses the problem in terms of maximizing the expected number of successes. He notes correctly that this is equivalent to minimizing the expected number of selections of the inferior arm. But while they are equivalent when the support of the prior distribution contains two points, this is not true generally. Many authors have focused on minimizing the number of selections of the inferior arm. For reasons indicated in our discussion of Percus and Percus (1984), we feel that this emphasis is ill-placed.

Related references: Our Corollary 4.3.10; Berry (1972), Bradt, Johnson, and Karlin (1956), DeGroot (1970), Fabius and van Zwet (1970), Kelley (1974), Rodman (1978), Vogel (1960c, 1964), Zaborskis (1976).

Fischer, J. (1979) Der diskontierte Einarmige Bandit. *Metrika* **26**: 195–204.

There are two Bernoulli arms; $\theta_2 = 1/2$ and θ_1 is uniform on $(0, 1)$. Discounting is geometric. Fischer finds that when the discount factor α is no larger than $3 - \sqrt{5} \approx 0.7639$, then an optimal strategy is as follows: select arm 1 until (if ever) the number of failures exceeds the number of successes, then switch permanently to arm 2.

We can improve this result by showing that $3 - \sqrt{5}$ may be replaced by the solution in $(1/2, 1)$ of $4\alpha - 4\alpha^4 = 1$, which is approximately 0.8968 and that this is the maximum possible α for which the indicated strategy is optimal. The proof can be accomplished using the fundamental theorem of gambling (Dubins and Savage, 1976, Theorem 2.12.1).

More generally, Fischer shows that for each discount factor α it is optimal to select arm 2 if the number of failures on arm 1 exceeds the number of successes by $1 + \text{int}[\alpha(2-\alpha)/(4(1-\alpha))]$. (This formula does not agree with Fischer's formula (30) which we believe contains a misprint; however, Fischer's subsequent table is correct.)

Results are also given in case $\theta_2 = 1/r$ where r is a known integer.

Related references: Our Chapter 5; Beckmann (1973), Glazebrook and Jones (1983).

Flehinger, B. J. and Louis, T. A. (1971) Sequential treatment allocation in clinical trials. *Biometrika* **58**: 419–426.

There are two arms whose responses are exponential variables with parameters θ_1 and θ_2. These are observed in real time and arm selections are separated by equal intervals of time. So the event in question (a patient's death, say) may be observed after other selections have been made, however information (of a positive nature) concerning an arm accrues even though the event has not yet been observed.

There are three possibilities: $\theta_1 = \theta_2$, $\theta_1 = k\theta_2$, $\theta_1 = \theta_2/k$. A family of procedures is proposed in which allocations are made according to death rates and numbers of deaths on the two arms; generally, an arm is selected if it is performing better, except that there is a tendency toward balance. The length of the trial is determined sequentially using a likelihood ratio test. (Presumably, the arm indicated to be better in the trial is used subsequent to the trial.) For various of the allocation schemes proposed, the numbers of selections of the inferior arm are compared with that of equal allocation using simulation. While the latter tends to result in shorter trials, it results in a greater number of selections of the inferior arm.

This paper differs from most of the papers in this bibliography in that it does not expressly consider the results of selections beyond the experimental phase. Put another way, there is no discount sequence that is assumed or implicit. The paper is included because of its similarity with other listed papers that do treat sum-maximization problems.

Related references: See Colton (1963).

Flehinger, B. J. and Louis, T. A. (1972) Sequential medical trials with data dependent treatment allocation. *Sixth Berkeley Symp. of Math. Statist. and Probab.* **4**: 43–51.

There are two arms with unknown means θ_1 and θ_2. The hypotheses to be tested are: $\theta_1 - \theta_2 = 0$, $\theta_1 - \theta_2 \leqq -\delta$, $\theta_1 - \theta_2 \geqq \delta$. A sequential generalized likelihood ratio test is used to decide when to stop sampling, and the allocation schemes that are suggested are designed to yield a small number of observations on the inferior arm.

This is not a bandit problem in our sense; the paper is included for the same reason that Flehinger and Louis (1971) is included.

Related references: See Colton (1963).

Flehinger, B. J., Louis, T. A., Robbins, H. and Singer, B. H. (1972) Reducing the number of inferior treatments in clinical trials. *Proc. Nat. Acad. Sci. USA* **69**: 2993–2994.

This note provides some mathematical explanation for the results shown by simulation in Flehinger and Louis (1971, 1972).

Related references: See Colton (1963).

Fox, B. L. (1974) Finite horizon behavior of policies for two-arm bandits. *J. Amer. Statist. Assoc.* **69**: 963–965.

There are two Bernoulli arms with parameters θ_1 and θ_2. Discounting is uniform and there is an experimental phase followed by a phase in which the evidently inferior arm is used, but only occasionally.

Monte Carlo methods are used to compare various first-phase strategies for various horizons. One class of strategies is similar to that proposed by

Zelen (1969). Another is similar to that considered by Vogel (1960a). Fox exhibits worths for various lengths of the first phase. The long-run success proportion is $\max\{\theta_1, \theta_2\}$ for strategies in both classes. Fox concludes from his simulations for various pairs (θ_1, θ_2) that 'play-the-winner/switch-from-a-loser' (Zelen, 1969) is not a very good strategy. (See our discussion of Percus and Percus (1984).) Wahrenberger, Antle, and Klimko (1977) conclude that none of the strategies considered by Fox are as good as Bayes strategies, even with an incorrect prior.

Related references: See Robbins (1956); see Colton (1963).

Furukawa, N. (1964) On some properties of an optimal strategy in the 'two-armed bandit' problem. *Mem. Fac. Sci. Kyushu Univ. A* **18**: 74–88.

Gait, P. A. (1972) Optimal allocation and control under uncertainty. Ph.D. thesis, Cambridge Univ., England.

Gittins, J. C. (1975) The two-armed bandit problem: variations on a conjecture by H. Chernoff. *Sankhyā A* **37**: 287–291.

Chernoff's (1968) conjecture is resolved in the negative when time is discrete and discounting is uniform. Two related two-armed bandits are compared for the discount sequence $(1, 1, 1, 0, 0, 0, \ldots)$. The first bandit has an unknown arm with prior distribution F_1 and a known arm 2 with mean $\Lambda(F_1, (1, 1, 1, 0, 0, \ldots))$. Thus either arm – in particular the known arm – is optimal. The second bandit has independent arms. The prior distribution for arm 1 is the same F_1 mentioned above. The prior for arm 2 is F_2 whose mean is $\Lambda(F_1, (1, 1, 1, 0, 0, \ldots))$. Arm 1 is uniquely optimal in the second example. (Cf. Theorem 6.2.1.)

Gittins observes that such an example is not possible in case the discount sequence is geometric. An easy proof is given which uses the Gittins and Jones (1974) result (our Theorem 6.1.1) and a result equivalent to our Theorem 5.0.2.

Chernoff's conjecture remains open for the setting in which it was proposed: continuous time and uniform discounting.

Related references: Our Chapter 6; Chernoff (1968), Gittins and Jones (1974).

Gittins, J. C. (1979) Bandit processes and dynamic allocation indices (with discussion). *J. R. Statist. Soc. B* **41**: 148–177.

The setting is that of Gittins and Jones (1974). Methods for calculating dynamic allocation indices are explored. Applications are given.

Related references: Our Sections 5.3 and 6.1; Bellman (1956), Gittins (1982), Gittins and Jones (1974).

Gittins, J. C. (1982) Forwards induction and dynamic allocation indices. In *Deterministic and Stochastic Scheduling* (eds M. A. H. Dempster, J. K. Lenstra and A. H. G. Rinnooy Kan), pp. 125–156, Reidel, Hingham, Massachusetts, USA.

The discounting is geometric. The formula for Λ given in our Corollary 5.3.2 is rewritten to be both meaningful and true in a setting more general than that of this monograph.

Related references: Our Sections 5.3 and 6.1; Gittins and Jones (1974), Gittins (1979).

Gittins, J. C. (1983). Dynamic allocation indices for Bayesian bandits. In *Mathematical Learning Models–Theory and Algorithms* (eds U. Herkenrath, D. Kalin and W. Vogel), pp. 50–67, Springer-Verlag, New York.

The extent to which Bather's (1983b) results give information for discrete-time bandits is studied both theoretically and numerically. Gittins also considers a bandit in which the goal is to observe a value in a certain range as soon as possible.

Related references: Bather (1983b), Berry and Fristedt (1980a, 1980b), Vogel (1961b).

Gittins, J. C. and Jones, D. M. (1974) A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (eds J. Gani *et al.*), pp. 241–266, North-Holland, Amsterdam.

Discounting is geometric. The setting involves k independent arms each of which is a Markov process. In the authors' setting, our posterior distributions are 'states'. They prove that the desirability of selecting an arm can be found by finding a known arm such that both the arm under consideration and the known arm are optimal in a two-armed bandit (with the same geometric discount sequence); the arm's 'dynamic allocation index' is the mean of the known arm. This theorem is the cornerstone for most of the current work on bandit problems having geometric discounting.

(*Note:* The inequalities in Lemma 2 of this paper should be reversed.)

Related references: Our Chapter 6; Bellman (1956), Gittins (1979), Glazebrook (1983a), Roberts and Weitzman (1980), Varaiya, Walrand, and Buyukkoc (1983), Whittle (1980, 1982).

Gittins, J. C. and Jones, D. M. (1979) A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika* **66**: 561–565.

Dynamic allocation indices (Gittins and Jones, 1974) are calculated for beta distributions of Bernoulli parameters and the discount sequence $(1, 0.75, (0.75)^2, \ldots)$.

Related references: Our Sections 5.3 and 6.1; Bellman (1956), Gittins and Jones (1974).

Gittins, J. C. and Nash, P. (1977) Scheduling, queues, and dynamic allocation indices. *Proc. 1974 EMS, Prague A* 191–202, Czechoslovak Academy of Sciences, Prague.

Various sequential decision problems, including bandits, are surveyed. The authors indicate for which problems 'dynamic allocation index' strategies are optimal; such a strategy in a bandit context assigns to each arm a number that depends only on that arm and the discount sequence, and not on the other arms (cf. Gittins and Jones, 1974, or our Section 6.1). (In Section 6.2 we show that they are optimal *only* for geometric discounting, assuming regularity.)

Related references: Our Chapter 6; Gittins (1975), Gittins and Jones (1974).

Glazebrook, K. D. (1978) On the optimal allocation of two or more treatments in a controlled clinical trial. *Biometrika* **65**: 335–340.

Assuming geometric discounting, Glazebrook studies the problem of calculating the break-even value Λ for an unknown arm. He uses the approach we employ in Section 2.6 when we approximate the maximal worth of a bandit problem by beginning the backwards induction from the left side of (2.6.1). Letting Λ_n denote this approximation (n is defined in our (2.6.1)), Glazebrook observes that it is easy to show that $\Lambda_n \uparrow \Lambda$ as $n \to \infty$. The geometric discounting plays no essential role although it does simplify some formulas.

For $\mathbf{A} = (1, 0.5, (0.5)^2, (0.5)^3, \ldots)$ and $1 \leqslant a \leqslant 15$, $1 \leqslant b \leqslant 15$, Glazebrook takes $n = 30 - a - b$ and calculates an approximation of Λ in case the unknown arm is Bernoulli and its probability of success has a beta distribution with parameters a and b. The results are presented in tabular form.

Related references: Our Sections 2.6 and 6.1; Gittins (1979), Gittins and Jones (1974).

Glazebrook, K. D. (1980) On randomized dynamic allocation indices for the sequential design of experiments. *J. R. Statist. Soc. B* **42**: 342–346.

There are k independent Bernoulli arms and discounting is geometric. The randomization scheme used by Bather (1980) is applied to dynamic allocation index strategies (Gittins and Jones, 1974). Such strategies are asymptotically optimal in the sense that the long-run success proportion is $\max\{\theta_i: i = 1, \ldots, k\}$. Glazebrook shows for beta priors that there are strategies in this class that are ε-optimal (i.e., ε-Bayes).

Related references: Bather (1980, 1981), Gittins (1979), Gittins and Jones (1974, 1979), Robbins (1952).

Glazebrook, K. D. (1982) On the evaluation of suboptimal policies for families of alternative bandit processes. *J. Appl. Probab.* **19**: 716–722.

There are k independent arms. Discounting is geometric with discount factor α. Glazebrook proves that

$$V(F_1, \ldots, F_k; \alpha) - W(F_1, \ldots, F_k; \alpha; \tau)$$
$$\leqslant E_\tau \sum_{m=1}^{\infty} \alpha^{m-1} [\max\{\Lambda(F_1^{(m-1)}, \alpha), \ldots, \Lambda(F_k^{(m-1)}, \alpha)\} - \Lambda(F_{\tau_m}^{(m-1)}, \alpha)]$$

where τ_m is the arm indicated by τ at stage m and $F_i^{(m-1)}$ is the distribution of arm i at stage m. He applies this result to two Bernoulli arms; θ_1 is uniformly distributed on (0, 1) and θ_2 is known to be 1/2.

Related references: Our Sections 5.3 and 6.1; Beckmann (1973), Fischer (1979), Gittins (1979), Gittins and Jones (1974), Glazebrook and Jones (1983).

Glazebrook, K. D. (1983a) Optimal strategies for families of alternative bandit processes. *IEEE Trans. Autom. Control* **28**: 858–861.

This is an expository paper on dynamic allocation indices. It contains a proof of the fundamental theorem of Gittins and Jones (1974).

Related references: Our Section 6.1; Gittins (1979), Gittins and Jones (1974), Roberts and Weitzman (1980), Varaiya, Walrand, and Buyukkoc (1983), Whittle (1980, 1982).

Glazebrook, K. D. (1983b) The role of dynamic allocation indices in the evaluation of suboptimal strategies for families of bandit processes. *Mathematical Learning Models–Theory and Algorithms* (eds U. Herkenrath, D. Kalin and W. Vogel), pp. 68–77, Springer-Verlag, New York.

Discounting is geometric. The bound for the discrepancy between the worth of any given strategy and the optimal worth given in Glazebrook (1982) is applied to a job scheduling problem in continuous time and to a job scheduling problem in discrete time for which there are constraints on the order in which jobs must be carried out.

A number of references pertinent to job scheduling and related problems are provided. Most of these are not included in this bibliography.

Related references: Our Section 6.1; Bather (1981), Gittins and Jones (1974), Glazebrook (1982).

Glazebrook, K. D. and Cox, T. F. (1980) On the design of efficient experiments for choosing between two Bernoulli populations. *Comm. Statist. A* **9**: 255–264.

Glazebrook, K. D. and Jones, D. M. (1983) Some best possible results for a discounted one armed bandit. *Metrika* **30**: 109–115.

There are two Bernoulli arms; θ_1 is uniformly distributed on (0, 1) and θ_2 is

known to be 1/2. Discounting is geometric with discount factor α. The authors address the following question: is it optimal to select arm 1 until (if ever) more failures than successes have been observed and then to switch to arm 2? They use computer calculations to find that the answer is affirmative if and only if $\alpha \leqslant \alpha^* = 0.801$ (to 3 decimal accuracy). As indicated in our discussion of Fischer (1979), we disagree: we use analytical methods to find $\alpha^* \approx 0.8968$.

Related references: Beckmann (1973), Fischer (1979).

Goto, M., Sugimura, M. and Asano, C. (1971) Numerical tables of optimum sequential designs based on Markov chains for selecting one of two medical treatments. *Bull. Math. Statist.* **14**: 27–56.

This paper contains a large bibliography of articles by these and other authors that concern problems marginally related to bandits. We have chosen to include only this paper in our bibliography as a source for the interested reader. Generally, all the papers listed by the authors deal with two-phase trials and there is an implicit objective of selecting good treatments over the course of the trial. So they have similarities with papers listed under Colton (1963). However, the papers listed by the authors are not easily categorized. Allocations in the first phase are usually *ad hoc* and are investigated numerically. Many of the papers are closely related to the 'ranking and selection' literature that we have chosen not to review here, though a few resemble papers dealing with finite-memory bandits as listed under Robbins (1956).

Related references: See Colton (1963).

Gray, K. B., Jr (1968) Sequential selection of experiments. *Ann. Math. Statist.* **39**: 1953–1977.

As an example of a general theory, Gray shows that selections in a Bernoulli two-armed bandit can be based on the sufficient statistics.

Related references: Our Chapter 2; Fabius and van Zwet (1970), Rieder (1975).

Hamada, T. (1978) A uniform two-armed bandit problem: The parameter of one distribution is known. *J. Jap. Statist.* **8**: 29–36.

Hengartner, W., Kalin, D. and Theodorescu, R. (1981) On the Bernoulli two-armed bandit problem. *Math. Operationsforsch. Statist. Ser. Optim.* **12**: 307–316.

[This paper was not available to us. The following account is culled from Benzing, Hinderer, and Kolonko (1984).]

There are two Bernoulli arms with parameters θ_1 and θ_2; discounting is *n*-

horizon uniform. Consistent with notation introduced in our Chapter 2, write the prior distribution of (θ_1, θ_2) in terms of a joint distribution measure G as follows: $(\sigma_1^a \varphi_1^b G)(du, dv) \propto u^a (1-u)^b G(du, dv)$. The authors show that the value is increasing in a and decreasing in b whenever the same is true for $E(\theta_1 \mid \sigma_1^a \varphi_1^b G)$, which holds when θ_1 and θ_2 are independent under G. This latter implication is a special case of our Theorem 4.1.6.

Related references: Our Section 4.1, Chapters 5 and 7; Benzing, Hinderer, and Kolonko (1984).

Herkenrath, U. (1983) The N-armed bandit with unimodal structure. *Metrika* **30**: 195–210.

There are k arms which have a special kind of dependence: the arms are labelled so that the sequence of k means is unimodal, reaching a (unique) maximum at θ, which is unknown. Observations are made sequentially. There are two objectives: (i) estimate θ, and (ii) maximize the limiting proportion of selections of arm θ. Herkenrath discusses various randomized strategies with regard to the rate of convergence of error in estimating θ and the average rate of payoff.

Related reference: Robbins (1952).

Herkenrath, U. and Theodorescu, R. (1978a) On certain aspects of the two-armed-bandit problem. *Elektron. Informationsverarb. Kybernet.*

There are two Bernoulli arms. The authors consider two families of strategies. One is a class of randomized strategies in which the probability of selecting arm 1 is modified as in Bush and Mosteller (1955). The authors show that these strategies achieve a long-run success proportion of $\max\{\theta_1, \theta_2\}$. The other class is that considered by Vogel (1960a).

Related references: Bush and Mosteller (1955), Herkenrath and Theodorescu (1978b), Witten (1973); and see Colton (1963).

Herkenrath, U. and Theodorescu, R. (1978b) On a stochastic approximation procedure applied to the bandit problem. Preprint 230, Sonderforschungsbereich 72, Univ. Bonn, FRG.

In a two-armed bandit setting with arbitrary unknown distributions, the authors give a strategy which approaches an average payoff per stage that is the maximum of the means of the two distributions. They obtain similar results in a particular many armed bandit that is motivated by a problem in market pricing suggested by Rothschild (1974).

Related references: Herkenrath and Theodorescu (1978a), Rothschild (1974).

Hill, C. and Sancho-Garnier, H. (1978) The two-armed-bandit problem: a decision theory approach to clinical trials. *Biomedicine* **28 Special Issue**: 42–43.

This is an expository article which compares the use of bandit strategies with the use of completely randomized designs in clinical trials. Though the authors give little evidence for their position, they claim that a balanced randomized allocation is a 'good procedure' in a bandit setting. We disagree; randomized clinical trials can sacrifice many patients in order to gain information concerning the treatments involved. The size of a randomized clinical trial is usually determined according to considerations of statistical power. We think that statistical power is irrelevant in many cases. It would usually be better to choose trial size on the basis of a cost/benefit analysis: What is the value of the information that is to be gained? Are there many patients with the condition in question, or few? How much of a sacrifice are patients who receive treatment randomly being asked to make? Costs and benefits should be weighed in specifying the discount sequence in a bandit problem: a large horizon, or a large discount factor when the sequence is geometric, corresponds to a setting in which it is likely that many patients will be treated with one of treatments involved in the trial. Our attitude in this regard is similar to that of Anscombe (1963).

Horowitz, A. D. (1973) Experimental study of the two-armed bandit problem. Ph.D. thesis, Univ. of North Carolina at Chapel Hill, USA.

There are two independent Bernoulli arms and discounting is n-horizon uniform. Horowitz compares optimal strategies (obtained by dynamic programming – see our Chapter 2) with three suboptimal strategies: alternating choice (or an arbitrary data-independent strategy), myopic, and 'play-the-winner/switch-from-a-loser'.

Horowitz's main objective is to analyse the way in which subjects actually behave in a bandit setting. He gives a number of helpful summaries of the results. The instructions to the subjects are well conceived and well written. But we do have one complaint: the subjects should not have been offered a bonus for the best performance among those participating. The presence of a bonus (or simply a competitive atmosphere) could change a subject's behaviour – a subject trying to achieve the largest number of successes would act differently from a subject who is trying to maximize the expected number of successes.

Related references: Our Example 2.4.3 and Chapter 7; Berry (1972), Bush and Mosteller (1955), Estes (1950).

Iosifescu, M. and Theodorescu, R. (1969) *Random Processes and Learning*, Springer-Verlag, New York.

The authors devote a portion of this book to linear (cf. Bush and Mosteller, 1955) and nonlinear learning models.

Related references: Bush and Mosteller (1955).

Isbell, J. R. (1959) On a problem of Robbins. *Ann. Math. Statist.* **30**: 606–610.

There are two unknown Bernoulli arms. The objective is to maximize the long-run success proportion. The current selection can depend only on the previous r selections and observations. Isbell improves on strategies suggested by Robbins (1956).

Related references: See Robbins (1956).

Jones, D. M. (1970) A sequential method for industrial chemical research. M.Sc. thesis, Univ. of Wales, Aberystwyth.

Jones, D. M. (1974) Search procedures for industrial chemical research. Ph.D. thesis, Cambridge Univ., England.

Jones, P. W. (1975) The two-armed bandit. *Biometrika* **62**: 523–524.

There are two independent Bernoulli arms (with beta priors) and discounting is n-horizon uniform. The values (maximal expected payoffs) of two families of bandits ($n = 2$ to 15) are compared in tabular form with the worths of the myopic strategy and the 'play-the-winner/switch-from-a-loser' strategy. Not surprisingly (see our discussion of Percus and Percus (1984)), the latter strategy fares badly.

Related references: Our Example 2.4.3 and Chapter 7; Berry (1972), Jones and Kandeel (1983), Percus and Percus (1984).

Jones, P. W. (1976) Some results for the two-armed bandit problem. *Math. Operationsforsch. Statist. Ser. Optim.* 7: 471–475.

There are two independent Bernoulli arms with beta priors; discounting is n-horizon uniform. In terms of the notation used in our Example 7.2.1, Jones shows that the value of the bandit is increasing in a_i for fixed $a_i + b_i$, $i = 1, 2$; this result is an instance of our Theorem 4.1.6. He conjectures that if arm i is optimal for beta parameters a_i and b_i, it is also optimal if a_i is increased with $a_i + b_i$ fixed; that this is true is immediate from Theorem 5.2 of Berry (1972).

Related references: Our Example 2.4.3 and Chapters 4 and 7; Berry (1972), Jones and Kandeel (1983).

Jones, P. W. (1977) Some designs for the two-armed bandit with one probability known. *Biom. J.* **19**: 693–695.

There are two Bernoulli arms with one arm known; the parameter of the unknown arm has a beta distribution. Discounting is n-horizon uniform. Jones gives tables comparing the worths of the following three strategies with that of an optimal strategy: the better single-arm strategy, the myopic strategy, and a variant of the myopic strategy in which the unknown arm is also selected if its mean is slightly less than that of the known arm.

Related references: Our Chapter 5; Berry and Fristedt (1979), Jones (1978), Jones and Kandeel (1983).

Jones, P. W. (1978) On the two-armed bandit with one probability known. *Metrika* **25**: 235–239.

There are two Bernoulli arms with one arm known. Discounting is n-horizon uniform. Jones gives two theorems which have hypotheses concerning posterior means; we note that these hypotheses follow easily from the Cauchy–Schwarz inequality. The second theorem is incorrect. If this theorem were correct then it would imply, in conjunction with our Theorem 5.0.2, that the break-even value of the unknown arm is its mean; so myopic strategies would be optimal.

Related references: Our Chapter 5; Berry and Fristedt (1979), Bradt, Johnson, and Karlin (1956), Jones (1977), Jones and Kandeel (1983).

Jones, P. W. and Kandeel, H. A. (1983) Numerical investigation of the two armed bandit. *Mathematical Learning Models–Theory and Algorithms* (eds U. Herkenrath, D. Kalin and W. Vogel), pp. 101–107, Springer-Verlag, New York.

The discount sequence considered is $\mathbf{A} = (1, \alpha, \ldots, \alpha^{n-1}, 0, 0, \ldots)$ for $0 < \alpha \leqslant 1$; special emphasis is given to $\alpha = 1$, in which case $\mathbf{A}$ is the n-horizon uniform. In the first of two settings considered, there are two independent Bernoulli arms whose parameters have beta distributions. When $\alpha = 1$, optimal worths are obtained by dynamic programming and are compared in tabular form with the worths of myopic strategies. The analytical results the authors give are proved in Berry (1972) and also in our Chapter 7. The authors conjecture a special case of the conjecture we discuss in the penultimate paragraph of Chapter 7; cf. Berry (1972), Joshi (1975).

In the other setting considered by the authors, one arm is known. They briefly discuss the effect of taking $\alpha < 1$ and compare myopic and optimal strategies. They argue that the number of stages for which the unknown arm is selected is never greater for $\alpha < 1$ than it is for $\alpha = 1$. It is assumed that once the known arm is selected it should be selected thereafter. That this is correct for any α (even $\alpha > 1$!) follows from our Theorem 5.2.2, or from Berry and Fristedt (1979). However, when $\alpha = 1$ the proof is easy and is given by Bradt, Johnson, and Karlin (1956, Lemma 4.1)–see our discussion immediately preceding the proof of Theorem 5.2.2.

Related references: Our Example 2.4.3 and Chapters 5, 6, and 7; Berry (1972), Berry and Fristedt (1979), Bradt, Johnson, and Karlin (1956), Gittins and Jones (1979), Jones (1977), Jones (1978), Jones and Kandeel (1985), Joshi (1975), Kalin and Theodorescu (1982), Percus and Percus (1984).

Jones, P. W. and Kandeel, H. A. (1985) A comparison of sampling rules for a Bernoulli two armed bandit. *Comm. Statist. C* **3** (to appear).

The setting is that of Jones and Kandeel (1983). Various properties of strategies considered in that paper are examined.

Related reference: Jones and Kandeel (1983).

Joshi, V. M. (1975) A conjecture of Berry regarding a Bernoulli two-armed bandit. *Ann. Statist.* **3**: 189–202.

The setting is that of our Section 7.3; there are two independent Bernoulli arms having common underlying prior distributions and the discount sequence is the n-horizon uniform. In our notation, it is proved in Corollary 2.1 of Joshi that if $b_1 \leqslant b_2$ and $E(\theta_1 | \sigma^s \varphi^f F_1) \geqslant E(\theta_1 | F_1) \geqslant E(\theta_2 | F_2)$, then $E(\theta_1 | \sigma^s \varphi^f F_1) \geqslant E(\theta_2 | \sigma^s \varphi^f F_2)$.

Joshi claims to prove the conjecture of Berry (1972) that is discussed in the penultimate paragraph of our Section 7.3. However, there is an error in his proof that cannot be easily repaired. The proof relies on his Corollary 3.1, which is not correct. This is easily seen by taking, in Joshi's notation, μ equal to Lebesgue measure on (0, 1), $r_0 = r_0' = l_0 = l_0' = m_2 = n_2 = 0$, $m_1 = n_1 > 0$, and $n = 2$, and then applying our Section 7.2. (In response to our query regarding this matter, Professor Joshi agrees with our assessment.) Therefore, we regard the conjecture as unresolved.

Related references: Our Chapter 7; Berry (1972).

Kakigi, R. (1983) A note on discounted future two-armed bandits. *Ann. Statist.* **11**: 707–711.

There are two possibly dependent Bernoulli arms: (θ_1, θ_2) has a two-point distribution on the unit square. Discounting is geometric. Kakigi applies [Blackwell, D. (1965). *Ann. Math. Statist.* **36**: 226–235] to find the optimal strategies for some special cases.

Related references: Our Corollary 4.3.10 and Chapter 5; Berry and Fristedt (1979), Feldman (1962), Kelley (1974), Rodman (1978), Zaborskis (1976).

Kalaba, R. E. and Tesfatsion, L. (1978) Two solution techniques for adaptive reinvestment: a small sample comparison. *J. Cybernet.* **8**: 101–111.

There are two investment opportunities; arm 2 always gives 0 rate of return and arm 1 gives $+r$ and $-r$ with unknown probabilities θ_1 and $1 - \theta_1$. The decision maker has an initial capital and at each of n stages is to allocate the current capital between the two arms. The objective is to maximize the expected value of the logarithm of total capital at stage n. So there are a number of ways that this differs from the problems we consider. Foremost among these is that both arms are observed at all stages, and so payoff and information-gathering are separate considerations. The authors show that the decision problem can be decomposed into n simple maximization problems.

Related reference: Tesfatsion (1978).

Kalin, D. (1979) Über Markoffsche Entscheidungsmodelle mit halbgeordnetem Zustandsraum. *Methods of Operations Research* **33**: 233–245.

Kalin, D. (1981) Beiträge zu strukturierten Markoffschen Entscheidungsmodellen. *Habilitationsschrift*, Univ. Bonn, FRG.

Kalin, D. (1982) Zum Problem des zweiarmigen Bernoulli-Banditen mit einer bekannten Erfolgswahrscheinlichkeit und unendlich vielen Spielen. *Metrika* **29**: 261–270.

There are two Bernoulli arms, one of which is known. Discounting is geometric and truncated geometric. Kalin gives special cases of our Theorems 4.1.6 and 5.0.2 (cf. Berry and Fristedt, 1979). The material presented is similar to that of Kalin and Theodorescu (1980).

Related references: Our Chapters 4 and 5; Berry and Fristedt (1979), Jones (1978), Kalin and Theodorescu (1980).

Kalin, D. and Theodorescu, R. (1980) Sur le probleme du bandit á deux bras quand une probabilité est connue. *Publ. Inst. Statist., Univ., Paris* **XXV**: 49–60.

There are two Bernoulli arms with one arm known. Discounting is n-horizon uniform; a truncated geometric is also considered. The authors give special cases of our Theorems 4.1.6 and 5.0.2 (cf. Berry and Fristedt, 1979). The material presented is similar to that of Kalin (1982).

Related references: Our Chapters 4 and 5; Berry and Fristedt (1979), Jones (1978), Kalin (1982).

Kalin, D. and Theodorescu, R. (1982) A note on structural properties of the Bernoulli two-armed bandit problem. *Math. Operationsforsch. Statist. Ser. Optim.* **13**: 469–472.

[This paper was not available to us. We understand from other sources that it contains an independent proof of Berry's (1972) result that optimal strategies stay with a winner when there are two independent Bernoulli arms and discounting is uniform.]

Kalin, D. and Theodorescu, R. (1983) On a stopping rule for a class of sequential decision problems. *Metrika* **30**: 117–123.

There are two Bernoulli arms with one arm known. Discounting is n-horizon uniform. The authors show that once the known arm becomes optimal it remains optimal. This was proved by Bradt, Johnson, and Karlin (1956, Lemma 4.1). It is a simple instance of Theorem 2.1 of Berry and Fristedt (1979)–see our Theorem 5.2.2. The proof in this easy special case is given in the discussion immediately preceding our proof of Theorem 5.2.2.

Related references: Our Chapter 5; Berry and Fristedt (1979), Bradt, Johnson, and Karlin (1956), Jones and Kandeel (1983).

Karatzas, I. (1984) Gittins indices in the dynamic allocation problem for diffusion processes. *Ann. Probab.* **12**: 173–192.

Time is continuous and discounting is exponential. The arms are one-dimensional diffusion processes. The theorem of Gittins and Jones (1974) is extended to this setting. Explicit calculations of dynamic allocation indices are made. Karatzas's setting includes instances of our setting (for instance, our Section 8.1) in which posterior distributions can be represented by a real parameter undergoing a diffusion.

Related references: Our Section 8.1; Gittins (1979), Gittins and Jones (1974).

Keener, R. W. (1984) Further contributions to the 'two-armed bandit' problem. Statistics Tech. Rep. No. 124, Univ. of Michigan, USA.

There are two dependent arms and two states of nature: G is supported by two ordered pairs of distributions on $\mathbf{R}$. Discounting is geometric (with $\alpha = 1$ being allowed). Keener obtains optimal strategies in terms of expected values of ladder epochs of random walks.

Related references: Our Corollary 4.3.10; Feldman (1962), Kelley (1974), Quisel (1965), Rodman (1978).

Kelley, T. A. (1974) A note on the Bernoulli two-armed bandit problem. *Ann. Statist.* **2**: 1056–1062.

There are two dependent Bernoulli arms with the prior for (θ_1, θ_2) concentrated on two points: (a, b) and (c, d). The discount sequence is the n-horizon uniform. For $n \geqslant 3$ Kelley shows that, except for some simple special cases, Feldman's (1962) assumption that $a = d$ and $b = c$ is necessary for the conclusion that myopic strategies are optimal. Kelley shows that myopic strategies are optimal for $n \geqslant 3$ and two-point distributions if and only if either (i) $a \leqslant b$ and $c \leqslant d$, (ii) $a \geqslant b$ and $c \geqslant d$, (iii) $a + b = c + d = 1$, or (iv) $(a, b) = (d, c)$.

Related references: Our Corollary 4.3.10; Berry (1972), DeGroot (1970), Fabius and van Zwet (1969), Feldman (1962), Keener (1984), Rodman (1978), Zaborskis (1976).

Kelley, T. A. (1976) Two-stage procedures for the Bernoulli two-armed bandit. Statistics Tech. Rep. No. 103, Univ. of Florida, USA.

The problem studied by Canner (1970) is treated (independently). Kelley finds $(\sqrt{(2n+5)} - 1)/2$ as the (approximate) optimal first-phase length when the prior is uniform (which compares with Canner's $(\sqrt{(2n+4)} - 2)/2$).

Related references: See Colton (1963).

Kelly, F. P. (1981) Multi-armed bandits with discount factor near one: the Bernoulli case. *Ann. Statist.* **9**: 987–1001.

There are k independent Bernoulli arms with common underlying prior F (cf. our Section 7.3). Thus, $F_i = \sigma^{a_i}\phi^{b_i}F$, $1 \leqslant i \leqslant k$. Discounting is geometric. Kelly assumes that $F[x, 1]$ is regularly varying as $x \to 1$. For the discount factor α sufficiently close to 1, Kelly proves that an optimal initial arm i must satisfy $b_i = \min\{b_j\}$, and that among such i the only optimal selections are those for which a_i is maximum. (How close α must be to 1 depends on the numbers a_i and b_i, $1 \leqslant i \leqslant k$.)

Whether $F[x, 1]$ regularly varying can be replaced by $F[x, 1] > 0$ for $x < 1$ is an open problem.

Berry (1972) conjectures that this 'least failures rule' is optimal for n-horizon uniform discounting as $n \to \infty$.

Related references: Our Sections 6.1 and 7.3; Berry (1972), Gittins and Jones (1974), Woodroofe (1976).

Kolonko, M. and Benzing, H. (1983) The sequential design of Bernoulli experiments including switching costs. Unpublished.

There are two Bernoulli arms with one arm known. Discounting is n-horizon uniform (though truncated geometric is also considered). Various results shown by Bradt, Johnson, and Karlin (1956)–and incorrectly attributed by Kolonko and Benzing to other authors–are proven in the presence of fixed costs for switching arms. In particular, optimal strategies 'stay with a winner'. (This statement includes staying with the known arm once it is optimal because the state of information is the same whether the known arm is a 'winner' or 'loser'.)

When one arm is known and the discount sequence is *regular* (our Definition 5.2.1), the two-armed bandit with no switching costs is a stopping problem. It is not surprising that a decision maker would be less likely to switch when there is a fixed cost for so doing! We would like to see conceptual proofs using this notion. The following conceptual argument (cf. our discussion just before the proof of Theorem 5.2.2) shows that a bandit with the same cost for switching in either direction is a stopping problem when discounting is uniform. Suppose every optimal strategy indicates arm 2 (the known arm) initially; so arm 2 is uniquely optimal. If the problem is not a stopping problem then there is a nonrandom stage, say $r + 1$, at which some optimal strategy τ first indicates arm 1. Now consider strategy τ^* which, for stages 1 through $n - r$, indicates the arm indicated by τ at respective stages $r + 1$ through n; for stages $n - r + 1$ to n, τ^* indicates the arm indicated by τ at stages 1 through r (arm 2 in all cases). Defined thus, the expected number of successes is the same for both τ^* and τ. Moreover, the number of switches is either the same or one fewer on τ^* than on τ. Therefore, τ^* is at least as good as τ and, since τ^* indicates arm 1 initially, arm 2 cannot be uniquely optimal. It follows that arm 2 can be used exclusively once it becomes uniquely optimal.

Costs for switching are not as interesting when one arm is known as

otherwise. For in the former case there is at most one 'switch'; so the cost is really one imposed for stopping.

Related references: Our Chapter 5; Berry and Fristedt (1979), Bradt, Johnson, and Karlin (1956), Kalin and Theodorescu (1982, 1983).

Kolonko, M. and Benzing, H. (1985) On monotone optimal decision rules and the stay-on-a-winner rule for the two-armed bandit. *Metrika* (to appear).

The stay-with-a-winner rule of Berry (1972) is generalized to certain instances of dependence between the arms.

Related references: Our Examples 2.4.1 and Theorem 4.3.8; Benzing, Hinderer, and Kolonko (1984), Berry (1972), Bradt, Johnson, and Karlin (1956), Hengartner, Kalin, and Theodorescu (1981), Kalin and Theodorescu (1982).

Kôno, K. (1966) How to deal with the a priori probability on the 'two-armed' bandit problem. *Math. Rep.* **4**: 27–34.

A special case of Feldman (1962) is considered. The author treats the possibility that the prior distribution is not known; we fail to appreciate the motives or results of this discussion.

Related references: Bradt, Johnson, and Karlin (1956), Feldman (1962).

Kumar, P. R. and Seidman, T. I. (1981) On the optimal solution of the one-armed bandit adaptive control problem. *IEEE Trans. Automat. Control* **26**: 1176–1184.

There are two Bernoulli arms; θ_2 is known and θ_1 has a beta distribution with parameters a and b. Discounting is geometric with discount factor α. The authors give results that are special cases of our Theorems 5.0.2 and Corollaries 5.3.3 and 5.3.4 (cf. Berry and Fristedt, 1979). They give upper and lower bounds for $\Lambda(F, \mathbf{A})$ that are weaker than those given in our Section 5.4 and in Berry and Fristedt (1979).

The authors also consider the question studied by Fischer (1979). They obtain the weaker bound $1/2$ in lieu of Fischer's $3-\sqrt{5}$.

Related references: Our Chapter 5; Bellman (1956), Berry and Fristedt (1979), Fischer (1979), Glazebrook and Jones (1983).

Lai, T. L., Levin, B., Robbins, H. and Siegmund, D. (1980) Sequential medical trials. *Proc. Natl. Acad. Sci. USA* **77**: 3135–3138.

There are n patients (n-horizon uniform discounting) in a clinical trial. The first $2r$ patients will be allocated in pairs to arms 1 and 2; each difference (arm 1 response minus that of arm 2) is distributed with unknown mean δ and variance σ^2 (known and unknown σ^2 are dealt with). The remaining $n-2r$ patients are assigned arm 1 if the sum of the pairwise differences is positive

and arm 2 otherwise. The objective is equivalent to maximizing the sum of the n observations by choosing r sequentially. The authors consider three stopping rules (including that of Anscombe (1963)) and state theorems which indicate that the three are asymptotically optimal ($n \to \infty$). These theorems are proven by Lai, Robbins, and Siegmund (1983).

Related references: See Colton (1963).

Lai, T. L. Robbins, H. and Siegmund, D. (1983) Sequential design of comparative clinical trials. *Recent Advances in Statistics; Papers in Honor of Herman Chernoff on his Sixtieth Birthday* (eds M. H. Rizvi, J. S. Rustagi and D. Siegmund), pp. 51–68, Academic Press, New York.

The setting is that of Anscombe (1963) and Colton (1963). The authors give properties of strategies (stopping rules) discussed by Lai, Levin, Robbins, and Siegmund (1980), including that suggested by Anscombe (1963). They prove results stated in Lai, Levin, Robbins, and Siegmund (1980) and give asymptotic properties of suboptimal rules suggested by Begg and Mehta (1979) and Colton (1963).

Related references: See Colton (1963).

Lakshmanan, K. B. and Chandrasekaran, B. (1978) On finite memory solutions to two-armed bandit problem. *IEEE Trans. Inform. Theory* **24**: 244–248.

There are two unknown Bernoulli arms. The objective is to maximize the long-run proportion of selections of the better arm subject to having but m memory states available. The authors show that at most 1 bit of memory is saved (for all m) by knowing the success probability of one of the arms. They also show that optimal deterministic strategies require no more than two bits over that required for randomized strategies.

Related references: See Robbins (1956).

Langenberg, P. and Srinivasan, R. (1981) On the Colton model for clinical trials with delayed observations – normally-distributed responses. *Biometrics* **37**: 143–148.

The setting is that of Colton (1963) and Colton (1965). The authors consider a transitional phase between the experimental and terminal phase. Two simple nonsequential procedures are compared for assigning arms in this middle phase.

Related references: See Colton (1963).

Langholz, G. (1977) Interaction between stochastic automata and random environment. *Internat. J. Man-Mach. Stud.* **9**: 223–231.

Two-armed bandits with finite memory are related to stochastic automata

learning models with finite memory. Simulations are used to compare three strategies.

Related references: See Robbins (1956).

Louis, T. A. (1975) Optimal allocation in sequential tests comparing two Gaussian populations. *Biometrika* **62**: 359–370.

The problem considered is similar to that of Flehinger and Louis (1972); it is assumed that the arms are normally distributed with the same known variance. The allocation rule that is optimal (in Louis's setting) in the continuous-time analogue is shown to be asymptotically optimal when time is discrete; simulations show that it performs well in the discrete case.

Related references: See Colton (1963).

Mallows, C. L. and Robbins, H. (1964) Some problems of optimal sampling strategy. *J. Math. Anal. Appl.* **8**: 90–103.

An arbitrary number (∞ is allowed) of arms with unknown distributions (subject to mild regularity conditions – boundedness is sufficient) is considered. The authors show that there are strategies that are asymptotically optimal in the sense that, with probability one, the limiting average of the observations is the supremum of the means of the distributions. Some related problems are considered.

Related references: Bather (1981); and see Robbins (1956).

Meeter, D. A. (1975) A two-armed bandit with terminal decision (Bayes rule). *A Survey of Statistical Design and Linear Models* (eds J. N. Srivastava), pp. 419–426, North-Holland, Amsterdam.

Results from Fabius and van Zwet (1970) are generalized to the case in which the discount sequence is $\mathbf{A} = (1, 1, \ldots, 1, T, 0, 0, \ldots)$; cf. Wahrenberger, Antle, and Klimko (1977). Fabius and van Zwet (1970) treat the n-horizon uniform: $T = 1$ (or 0). An equivalent way of viewing the problem is that the decision maker's strategy is restricted to selecting the same arm for the last T stages. (We note that $\mathbf{A}$ is regular and so the appropriate results in our Chapter 5 would apply if one arm were known.)

The distinction between Meeter's problem and most of the references listed under Colton (1963) is that Meeter allows optimal sequential selections in the first phase, while most of the others place restrictions on first-phase allocations.

Related references: Fabius and van Zwet (1970); and see Colton (1963).

Meybodi, M. R. and Lakshmivarahan, S. (1983) On a class of learning algorithms with symmetric behaviour under success and failure. *Mathematical Learning Models–Theory and Algorithms.* (eds U.

Herkenrath, D. Kalin and W. Vogel), pp. 145–155, Springer-Verlag, New York.

There are k Bernoulli arms. A class of randomized strategies is proposed such that, loosely speaking, the probability of selecting an arm increases should the arm yield a success and it decreases should the arm yield a failure. Conditions are given under which such a strategy results in a long-run success proportion within ε of that of the best arm.

Related references: Bather (1981); and see Robbins (1956).

Morrison, D. F. (1967) On the two-armed bandit problem. *Psychology of Management Decision* (ed. G. Fisk), pp. 186–195, C. W. K. Gleerup, Lund, Sweden.

There are two arms and two densities f and f^*; the case of normal densities with equal variances is given special consideration. Which density goes with which arm is not known. Discounting is n-horizon uniform. (Feldman (1962) proved that myopic strategies are optimal in the Bernoulli case. Rodman (1978) removed the Bernoulli hypothesis. Morrison incorrectly attributes the general result to Feldman.) Morrison considers the n stages divided into r phases of fixed length. Allocation is balanced in the first phase (cf. Anscombe, 1963; Colton, 1963, etc.), and thereafter the arm with greater probability of having density f^* (say f^* is preferred) at the beginning of a phase is used exclusively during that phase.

An *ad hoc* strategy based on sample means is discussed.

Related references: Feldman (1962), Rodman (1978), Zaborskis (1976); and see Colton (1963).

Murray, F. S. (1971) Multiple probable situation: A study of a five one-armed bandit problem. *Psychon. Sci.* **22**: 247–249.

This reports an experimental study involving 14 subjects. There are five independent Bernoulli arms (electronic devices) with n-horizon uniform discounting, $n = 250$. The subjects were given an M&M candy for each success. (We are not convinced that this is sufficient reward to induce a subject to try to maximize the expected number of successes.) The probabilities of success on the arms were not known to the subjects, but were in fact $\theta_1 = 0.5$, $\theta_2 = 0.25$, $\theta_3 = 0.25$, $\theta_4 = 0.125$, $\theta_5 = 0$. Murray reports that the overall proportion of the stages numbered 211 to 250 in which the subjects used the arm with greatest current frequency of success (not necessarily arm 1 – he does not indicate whether it was ever a different arm) was 52%. This compares with 24% for the arm with the second highest success frequency and 7% for the lowest (presumably the latter was always arm 5).

We note that the subjects were given only 3 seconds between selections. Results from a similar experiment with unlimited time between selections would be more interesting to us.

Related references: Brand, Sakoda, and Woods (1957), Brand, Woods, and Sakoda (1956), Bush and Mosteller (1955), Estes (1950), Horowitz (1973).

Nakajima, N. and Noshi, T. (1978) On the behaviour of the shrewd automaton. Technol. Rep. No. 25, Seikei Univ., Japan.

Nash, P. (1973) Optimal allocation of resources between research projects. Ph.D. thesis, Cambridge Univ., England.

Nash, P. (1980) A generalized bandit problem. *J. R. Statist. Soc. B* **42**: 165–169.

There are k independent arms with geometric discounting. The setting is that of Gittins (1979). The payoff at any stage depends on the arm selected at that stage but also (in a multiplicative way) on the other arms. Nash shows that the theorem of Gittins and Jones (1974) applies to his problem.

This is not a bandit problem in our setting since the payoff at any stage cannot be viewed as an observation from the distribution of the arm selected.

Related references: Gittins (1979), Gittins and Jones (1974)

Obregon, I. (1968) The N-armed bandit problem and other topics in sequential decision processes. Operations Research Tech. Rep., Massachusetts Institute of Technology, USA.

Oudin, C. and Lellouch, J. (1972) La comparison de quelques stratégies dans la conduite des essais thérapeutiques *Rev. Statist. Appl.* **XX**: 5–21.

This is mainly an expository article. There are two independent Bernoulli arms with beta priors on θ_1 and θ_2. Discounting is n-horizon uniform. Maximizing the expected number of successes among the n ('ethique collective') is contrasted with maximizing the probability of immediate success ('ethique individuelle'). Two-phase and unrestricted bandit problems are put in the first category (the authors give special attention to the case in which one arm is known) and myopic strategies in the second. Geometric discounting truncated after stage n is recommended as a compromise.

For the two-phase design the approach is similar to that of Pearson (1980). When the distribution of (θ_1, θ_2) is uniform the authors conclude 'by symmetry' that the two arms should be allocated equally in the first phase. Pearson (1980) shows that equal allocation is never optimal in this case! With the condition of equal allocation the authors rederive the approximation to the optimal first-phase length of Canner (1970).

Related references: Our Chapters 5 and 7; and see Colton (1963).

Pearson, L. M. (1980) Treatment allocation for clinical trials in stages. Ph.D. thesis, Univ. of Minnesota, USA.

The setting is that of Berry and Pearson (1984). A number of extensions of the problems considered in that paper are given and numerous tables are provided.

Pearson shows that optimal first-phase allocations may not be symmetric even though the prior is symmetric in θ_1 and θ_2.

Related references: See Colton (1963).

Percus, O. E. and Percus, J. K. (1984) Modified Bayes technique in sequential clinical trials. *Comput. Biol. Med.* **14**: 127–134.

This paper deals with myopic strategies (called 'ethical' by the authors) for assigning patients to one of two medical treatments. The arms are Bernoulli. Both θ_1 and θ_2 have a beta distribution with parameters a and 1. The authors consider the myopic strategy which always uses the arm with greater $[s_1 + a]/[s_i + f_i + a + 1]$ where s_i and f_i are the current numbers of successes and failures on arm i; cf. Bather (1980). The authors use simulation to find the long-run proportion of stages for which the inferior treatment is used for various combinations of Bernoulli parameters. (The prior distribution serves only as a means of finding a strategy which pays heed to the accumulating data rather than as a reflection of initial information. In our view the resulting strategy is not 'ethical' unless this distribution corresponds to that of the clinician.) Not surprisingly, they find that this proportion is smaller when a is large.

The authors compare these 'ethical' strategies with 'play-the-winner/switch-from-a-loser'. Again not surprisingly, the former perform better according to the authors' criterion; playing a winner is reasonable, but switching on a loser can be arbitrarily bad and as a general policy has no merit whatever.

This paper is an example of a large literature (most of which is not mentioned here) which compares strategies with respect to the proportion of patients who are treated with the inferior treatment. Our approach (using $\mathbf{A} = (1, 1, \ldots, 1, 0, 0, \ldots)$) of minimizing the expected number of *failures* is quite different: a success on a 'bad' arm is as good as a success on a 'good' arm. And to us, using an arm whose success probability is substantially smaller than an alternative should be penalized more heavily than using an arm whose success probability is only slightly smaller than the alternative.

Related references: Bather (1980, 1981), Berry (1978), Feldman (1962), Fox (1974), Jones and Kandeel (1983), Oudin and Lellough (1972), Robbins (1952), Thompson (1933, 1935).

Petkau, A. J. (1978) Sequential medical trials for comparing an experimental with a standard treatment. *J. Amer. Statist. Assoc.* **73**: 328–338.

There are two Bernoulli arms; θ_1 has a beta prior distribution and θ_2 is known. Normal distributions are also considered. Discounting is n-horizon

uniform. There is a constant cost of observation for each stage until a decision is made to select a single arm for the duration, for which there are no costs for observation. The objective is to maximize the expected number of successes overall minus cost of observation in the experimental phase; the problem is to decide when to stop paying for observing arm 1 at which time either arm 1 or arm 2 is selected exclusively.

Assuming n is large, Petkau adopts the continuous-time approximation suggested by Chernoff and Ray (1965). Optimal stopping boundaries are shown for various costs of observations. When there is no such cost the problem is the same as that treated in our Section 8.2; our Figure 8.2 is essentially the same as the boundary shown in Figure A of this paper corresponding to $\gamma_1 = 0.1$, the smallest cost of observation considered (the upper boundary having disappeared), and Figure 8.3 corresponds to Figure B of this paper.

Petkau considers the sensitivity of optimal strategies to misspecification of n. He also evaluates three suboptimal strategies: (i) stop experimenting when the number of successes minus failures is sufficiently large or small – horizontal boundaries in Figure 8.2; (ii) stop experimenting when the best fixed-sample-size continuation strategy is to stop (cf. Begg and Mehta, 1979, and Chernoff and Petkau, 1981); and (iii) the best fixed-sample-size strategy – a vertical boundary in Figure 8.2.

Related references: Our Chapter 5 and Section 8.2; Chernoff (1967, 1968, 1972), Chernoff and Petkau (1976, 1983a, 1984), Chernoff and Ray (1965); and see Colton (1963).

Poloniecki, J. D. (1978) The two armed bandit and the controlled clinical trial. *Statistician* **27**: 97–102.

Presman, E. L. and Sonin, I. M. (1982) *Posledovatel'noe Upravlenie po Nepolnym Dannym*, Izdatelstvo Nauka, Moscow.

The authors devote a considerable portion at their book to the study of bandits in both discrete and continuous time. Unfortunately, we do not read Russian well enough to give a detailed review.

The title is *Sequential Control with Partial Information*. The chapter headings are: (1) Fundamental scheme for discrete and continuous time, (2) Formulation of the problem and methods of solution for discrete time, (3) The solution of certain problems in the fundamental scheme for discrete time, (4) Formulation of the problem and methods of solution for continuous time, (5) The solution of problems in the fundamental scheme for continuous time, (6) Application of the maximum principle of Pontryagin to control problems with random jumps, (7) Some other problems.

Related references: This monograph; Presman and Sonin (1983).

Presman, E. L. and Sonin, I. M. (1983) 'Two and many-armed bandit' problems with infinite horizon. *Proceedings of the Fourth USSR-Japan Symp. in Probab. Theory and Math. Statist.* (eds K. Itô and J. V. Prokhorov), pp. 526–540, Springer-Verlag, Berlin.

There are k not necessarily independent Bernoulli arms; discounting is n-horizon uniform. Randomized strategies are permitted; this is relevant since the authors are interested in limit theorems for $n \to \infty$. They assume that the distribution G of the vector $(\theta_1, \ldots, \theta_k)$ of Bernoulli parameters is supported by a finite number of atoms. Necessary and sufficient conditions are given on G for $n[E(\max(\theta_1, \ldots, \theta_k)|G] - V(G; n)$ to approach a finite limit as $n \to \infty$. In case the limit is finite, a strategy τ is indicated for which $V(G; n) - W(G; n; \tau) \to 0$ as $n \to \infty$. The authors refer the reader to Presman and Sonin (1982) for the proof.

The analogous result is obtained in a continuous-time setting. Each arm is a Poisson process with a random intensity. At each instant the decision maker is permitted to mix the intensities and observe the mixture. For example, the decision maker is permitted to observe arm 1 with weight 2/3 and arm 2 with weight 1/3.

Related references: Our Section 8.3; Bather (1981), Kelley (1974), Presman and Sonin (1982).

Quisel, K. (1965) Extensions of the two-armed bandit and related processes with on-line experimentation. Tech. Rep. No. 137, Institute for Mathematical Studies in the Social Sciences Stanford Univ., USA.

This report gives an extensive description of bandit and related problems. The point of view is the same as ours.

Quisel considers the case of two independent arms; both distributions are bounded and are in the same family, for which there is a one-dimensional sufficient statistic (for fixed sample size). The discount sequence $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$ has finite horizon and is nonincreasing, but is arbitrary otherwise. Quisel assumes that the prior distribution is in a conjugate family – an example is the beta distribution when the observations are Bernoulli. He claims to prove the 'stay-with-a-winner' rule in this context; this says that when an arm is optimal and yields the maximum possible observation ('success' or '1' in the Bernoulli setting) then it is optimal again (see our Theorem 4.3.8 and Corollary 5.6.2). But his Theorem 4.2 is not a stay-with-a-winner rule as he claims since the result applies for fixed discount sequence and does not allow for it to change from one stage to the next. It is easy to see by adapting our Examples 3.3.3 or 5.2.2 that such a result cannot hold in the generality indicated. Our Example 5.2.2 assumes $\theta_2 = 0.6$ and $\theta_1 \in \{1, 0\}$; to adapt it to Quisel's setting with conjugate priors take the beta parameters for θ_2 very large and in the ratio 6 : 4 and take those for θ_1 very small and equal.

Another setting considered by Quisel has k independent arms, each of which is governed by one of two known distributions. He shows that it is optimal at any stage to select an arm having the largest probability of having the distribution with larger mean (cf. Rodman 1978).

Related references: Our Chapters 4, 5, and 7; Berry (1972), Bradt, Johnson, and Karlin (1956), Feldman (1962), Keener (1984), Rodman (1978).

Reimnitz, P. (1978) A study of 'two-armed bandits'. Ph.D. thesis, Univ. of Rochester, Rochester, New York, USA.

The extension by Vogel (1964) of Vogel (1960c) is repeated and carried further. The approach uses an *oscillating random walk*, used also by Vogel (1960c) and named by Kemperman [(1974) The oscillating random walk. *Stochastic Process. Appl.* **2**: 1–29]. The tool is studied in detail.

Reimnitz considers the myopic strategy for two Bernoulli arms where (θ_1, θ_2) is either (a, b) or (b, a). Following such a strategy, the probability that the inferior arm is used at stage m decreases exponentially. Reimnitz considers the possibility that (θ_1, θ_2) is different from (a, b) or (b, a). He shows that the exponentially decreasing property still holds when $|\log(a/b)/\log[(1-a)/(1-b)]|$ lies between $(1-\theta_1)/\theta_1$ and $(1-\theta_2)/\theta_2$. A weaker result about the boundedness of the regret is discussed in Vogel (1964).

Related references: Our Section 9.1; Feldman (1962), Rodman (1978), Vogel (1960c, 1961a, 1964).

Rieder, U. (1975) Bayesian dynamic programming. *Adv. in Appl. Probab.* **7**: 330–348.

Rieder discusses Markov decision problems. For a given problem having random parameters he gives a method for finding an equivalent problem that has no random parameters. His approach provides an alternative to some of the technical issues we address in Chapter 2.

Related references: Our Chapter 2; Gray (1968).

Robbins, H. (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58**: 527–535.

The problem of choosing sequentially from two populations to maximize the sum of n observations is posed. Robbins considers two Bernoulli arms with parameters θ_1 and θ_2 and shows that the 'play-the-winner/switch-from-a-loser' strategy results in a greater long-run success proportion than do rules that do not depend on accumulating data – uniformly in (θ_1, θ_2). He exhibits a family of strategies, each of which achieves the long-run success proportion of $\max\{\theta_1, \theta_2\}$.

This innovative paper spawned research in a number of directions; the references listed under Robbins (1956) deal with finite memory-bandit

problems that generalize Robbins' 'play-the-winner/switch-from-a-loser' strategy.

Related references: See Robbins (1956).

Robbins, H. (1956) A sequential decision problem with a finite memory. *Proc. Nat. Acad. Sci. USA* **42**: 920–923.

There are two Bernoulli arms. Memory of the history of arm selections and results is limited to the previous r stages. Robbins calculates the long-run success proportion for the following rule: select arm 1 and switch to arm 2 if it fails; if it is successful then use it only until it gives r failures in a row; repeat this sequence alternating from one arm to the other.

Related references: Cover (1968), Cover and Hellman (1970), Cover and Wagner (1976), Fox (1974), Isbell (1959), Lakshmanan and Chandrasekaran (1978), Langholz (1977), Meybodi and Lakshmivarahan (1983), Robbins (1952), Samuels (1968), Smith and Pyke (1965), Witten (1973, 1974, 1976, 1977a, 1977b, 1983).

Robbins, H. and Siegmund, D. (1974) Sequential tests involving two populations. *J. Amer. Statist. Assoc.* **69**: 132–139.

There are two normal arms with means θ_1 and θ_2 and unit variance. The hypotheses $\theta_1 = \theta_2 \pm \delta$, where δ is known, are to be tested. A sequential likelihood ratio test is used to decide when to stop sampling, and the allocation scheme is designed to yield a small number of observations of the inferior arm.

This is not a bandit problem in our sense; the paper is included for the same reason that Flehinger and Louis (1971) is included.

Related references: See Colton (1963).

Roberts, K. W. S. and Weitzman, M. L. (1980) On a general approach to search and information gathering. Economics Working Paper 263, Massachusetts Inst. of Technology, USA.

The authors prove a somewhat more general version of the theorem of Gittins and Jones (1974). Applications to a number of problems are discussed. These include optimal search and job scheduling as well as bandits.

Related references: Our Section 6.1; Gittins (1979), Gittins and Jones (1974), Glazebrook (1983a), Varaiya, Walrand, and Buyukkoc (1983), Whittle (1980, 1982).

Robinson, D. (1983) A comparison of sequential treatment allocation rules. *Biometrika* **70**: 492–495.

There are two Bernoulli arms and discounting is n-horizon uniform. Several strategies are compared using simulation for various values of θ_1 and θ_2 and

$n = 50$ and 100. Among those considered are three index strategies of Gittins and Jones (1974) designed for geometric discounting and assuming beta prior distributions on θ_1 and θ_2 (the author does not say which betas—perhaps both are uniform on (0, 1)). These three strategies are optimal for discount factors of 0.99, 0.995, and 0.9999. (If these strategies are to be applied in this setting for which they are not designed, then the appropriate discount factor would seem to be $1 - 1/n$.)

This is a curious exercise since finding optimal strategies for these values of n and any prior distribution is quite easy—easier than finding optimal strategies for geometric discounting with factor 0.9999, for example.

Robinson also considers the possibility that half of the responses are delayed for 20 stages.

Related references: Our Chapters 3 and 6; Bather (1981, 1983), Fox (1974), Gittins (1979), Gittins and Jones (1974), Jones (1975), Percus and Percus (1984).

Rodman, L. (1978) On the many-armed bandit problem. *Ann. Probab.* **6**: 491–498.

Consider two known distributions Q^* and Q_* with different means. There are k arms with one of the arms having distribution Q^* (which one is not known) and the other $k-1$ having distribution Q_*. The discount sequence is $\mathbf{A} = (1, \alpha, \ldots, \alpha^{n-1}, 0, 0, \ldots)$ where $\alpha \in (0, 1]$ and $n \in \{1, 2, \ldots, \infty\}$. (The case $\alpha = 1$ and $n = \infty$ is not allowed in our context but makes sense in Rodman's; cf. Zaborskis (1976).) Rodman shows that myopic strategies are optimal: the decision maker can select any arm that has the highest current probability of having the distribution with larger mean. So this generalizes the result of Feldman (1962) in a number of ways. Feldman assumed $k = 2$, Q^*, and Q_* to be Bernoulli, and $\alpha = 1$ with $n < \infty$. (Our Corollary 4.3.10 generalizes Feldman's result to arbitrary nonincreasing discount sequences.)

The uniform discounting version of this result for Bernoulli arms with Q^* having a larger mean than Q_* was proven first by Zaborskis (1976), which was not known to Rodman.

Related references: Our Corollary 4.3.10; DeGroot (1970), Fabius and van Zwet (1970), Feldman (1962), Keener (1984), Kelley (1974), Quisel (1965), Vogel (1960c, 1964), Zaborskis (1976).

Ross, S. M. (1983) *Introduction to Stochastic Dynamic Programming*, Academic Press, New York.

Chapter VII treats bandit problems with geometric discounting. The Gittins–Jones (1974) result is proved using the approach of Whittle (1980, 1982), in a vein similar to our Section 6.1. Ross also considers the extension of Whittle (1981) in which new arms are generated according to a Poisson process.

Related references: Our Section 6.1; Bellman (1956), Gittins (1979), Gittins and Jones (1974), Whittle (1980, 1981, 1982).

Rothschild, M. (1974) A two-armed bandit theory of market pricing. *J. Econ. Theory* **9**: 185–202.

A theory to explain the way in which stores should set prices is constructed using bandit problems. Rothschild also gives some analytic results assuming geometric discounting and two Bernoulli arms. The arms may not be independent; the distribution on parameters θ_1 and θ_2 is arbitrary and has support $(0, 1) \times (0, 1)$. Rothschild shows that optimal strategies will, with positive probability, lead to an infinite number of selections of arm 1 and a finite number of selections of arm 2 when $0 < \theta_2 < \theta_1 < 1$. He also shows that optimal strategies will, with probability one, lead to an infinite number of selections of only one arm when $\theta_1 \neq \theta_2$ (of course, it may not be the better arm that is selected).

Related references: Bellman (1956), Percus and Percus (1984), Schmalansee (1975).

Samuels, S. M. (1968) Randomized rules for the two-armed-bandit with finite memory. *Ann. Math. Statist.* **39**: 2103–2107.

Robbins (1956), Isbell (1959), and Smith and Pyke (1965) consider strategies that stay with an arm until it gives r consecutive failures; a switch is made to the other arm if the other arm passes some predetermined test. The objective is to maximize the long-run proportion of heads. Samuels shows that all these strategies can be improved by randomizing to decide whether to perform the test or stay with the current arm; strategies which give a higher probability of staying with the current arm are better (except that this probability cannot be one).

Related references: See Robbins (1956).

Schmalansee, R. (1975) Alternative models of bandit selection. *J. Econom. Theory* **10**: 333–342.

Some 'learning models' of Bush and Mosteller (1955) are applied to the pricing problem considered by Rothschild (1974). These are *ad hoc* randomized strategies that are not optimal but are easy to apply and may reasonably fit actual behaviour (but see Cane, 1962).

Related references: Bush and Mosteller (1955), Cane (1962), Rothschild (1974), Thompson (1933, 1935).

Smith, C. V. and Pyke, R. (1965) The Robbins–Isbell two-armed bandit problem with finite memory. *Ann. Math. Statist.* **36**: 1375–1386.

There are two unknown Bernoulli arms. The objective is to maximize the long-run success proportion. The current selection can depend only on the

previous r selections and observations. The authors improve on strategies considered by Isbell (1959).

Related references: See Robbins (1956).

Sudderth, W. D. (1982) Dynamic programming. *Encyclopedia of Statistical Sciences*, Vol. II (eds S. Kotz and N. L. Johnson) Wiley, New York.

Bandits are discussed as examples of dynamic programming problems. Sudderth discusses the scope of dynamic programming problems and gives a guide to the literature. We have not discussed general theories of dynamic programming either in the body of this monograph or in this bibliography; in particular, articles that treat such general theories are cited here only if they give bandit problems as examples.

Tesfatsion, L. (1978) A new approach to filtering and adaptive control. *J. Optim. Theory Appl.* **25**: 247–261.

A two-armed bandit is posed as an example of a more general set of control problems. The author wants to bypass the use of prior probabilities and Bayes's theorem. This does not seem possible to us. We do not understand whether the author is considering only suboptimal strategies (one strategy discussed is myopic and so backwards induction is avoided) or if she is making a broader claim.

Related reference: Kalaba and Tesfatsion (1978).

Thompson, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**: 275–294.

This is the paper that started it all! There are two independent Bernoulli arms (Thompson did not use the 'arm' and 'bandit' terminology) with θ_1 and θ_2 having uniform prior distributions on (0, 1). While recognizing that his randomized strategy is 'not the best possible', Thompson proposes selecting arm 1 with probability equal to the current probability that $\theta_1 > \theta_2$. He devotes most of the paper to calculational aspects of this probability.

Even setting aside this paper's historical importance, it is our view that it should be read by all researchers in this area (though the calculational issues are not as important now as they were in 1933); many have not and have failed to improve on the original!

Related references: Our Chapter 7; Bather (1980, 1981), Berry (1978), Thompson (1935).

Thompson, W. R. (1935) On the theory of apportionment. *Amer. J. Math.* **57**: 450–456.

Further calculations are made along the line of Thompson (1933). A crude but effective simulation is carried out to evaluate the randomized strategy

suggested in that earlier paper. Even for the small horizons considered, the strategy performs quite well.

It is surprising to us that modern-day authors have not compared their strategies with Thompson's. For example, we think the strategies discussed by Fox (1974) are markedly inferior to Thompson's.

Related references: Bather (1980), Fox (1974), Percus and Percus (1984), Thompson (1933).

Upton, G. J. G. and Lee, R. D. (1976) The importance of the patient horizon in sequential analysis of binomial clinical trials. *Biometrika* **63**: 335–342.

Most of the papers in this bibliography regard the discount sequence as known. This paper makes the point that ignorance of the discount sequence, the horizon in particular, can make discussions concerning differences in allocation strategies largely irrelevant. The discussion is couched in terms of two quite special strategies but we think it applies generally. It is for this reason that we allow the horizon to be random; see our Chapter 3, especially Section 3.1, and also Witmer (1983).

Related references: Our Chapter 3; and see Colton (1963).

Varaiya, P., Walrand, J. and Buyukkoc, C. (1983) Extensions of the multi-armed bandit problem. Electrical Engineering Tech. Rep., Univ. of California, Berkeley, USA.

A generalization of the result of Gittins and Jones (1974) is proved. Various applications are discussed, including discrete- and continuous-time bandit problems.

Related references: Our Section 6.1 and Chapter 8; Gittins (1979), Gittins and Jones (1974), Roberts and Weitzman (1980), Whittle (1980, 1981, 1982).

Viscusi, W. K. (1979a) *Employment Hazards: An Investigation of Market Performance*, Harvard Univ. Press, Cambridge, Massachusetts, USA.

Viscusi, W. K. (1979b) Job hazards and worker quit rates: An analysis of adaptive worker behaviour. *Internat. Econom. Rev.* **20**: 29–58.

An interesting application of two-armed bandits to job selections is discussed. There are two arms. One arm is known and is Bernoulli. The other is unknown (a potentially hazardous job) and results in 0 or w. Discounting is truncated geometric. Various examples and calculations of optimal strategies are carried out.

Related references: Our Chapter 5; Berry and Fristedt (1979, 1983), Berry and Viscusi (1981), Viscusi (1979a).

Vogel, W. (1960a) A sequential design for the two-armed bandit. *Ann. Math. Statist.* **31**: 430–443.

There are two Bernoulli arms. The discount sequence is the n-horizon uniform. The author studies the strategy that indicates the arms alternatively until, at some even-numbered stage $2r$, the discrepancy between the numbers of successes on the two arms is some constant, and thereafter indicates a single arm: the one yielding more successes. Variants of this two-phase design have been considered by many authors – see Colton (1963). Methods of this paper are used in the proof of our Theorem 9.1.3.

Related references: Our Section 9.1; Bather and Simons (1985), Vogel (1960b); and see Colton (1963).

Vogel, W. (1960b) An asymptotic minimax theorem for the two-armed bandit problem. *Ann. Math. Statist.* **31**: 444–451.

Using the results of Vogel (1960a), it is shown that the constants c_3 and c_4 in our Corollary 9.1.4 may be chosen to be 0.187 and 0.377, respectively, provided that n is sufficiently large.

Related references: Our Section 9.1; Bather (1983a), Bather and Simons (1985), Fabius and van Zwet (1970), Vogel (1960a).

Vogel, W. (1960c) Ein Irrfahrten-Problem und seine Anwendung auf die Theorie der sequentiellen Versuchs-Plane. *Arch. Math.* **XI**: 310–320.

There are two Bernoulli arms and the discount sequence is n-horizon uniform. Vogel considers the strategy that indicates the arm for which the current number of successes minus failures is larger. The regret (cf. Section 9.1) when the Bernoulli parameters are θ_1 and θ_2 is obtained in terms of probabilities associated with a particular Markov chain that moves one step to the right or to the left with probabilities depending on the sign of the current state. Limit theorems, as $n \to \infty$, are obtained. In particular, the regret approaches a finite limit if and only if $\theta_1 = \theta_2$ or $1/2$ is between θ_1 and θ_2.

Related references: Feldman (1962), Vogel (1961a, 1964).

Vogel, W. (1961a) Bemerkungen zu einem sequentiellen Versuchsplan. *Z. Angew. Math. Mech.* **41**: 179–181.

The portion of Vogel (1960c) concerned with a bound on the risk independent of n is generalized to a non-Bernoulli setting.

Related reference: Vogel (1960c).

Vogel, W. (1961b) Sequentielle Versuchs-Pläne. *Metrika* **4**: 140–157.

There are two unknown Bernoulli arms. The problem is to minimize the number of stages required to achieve a given number of successes.

Related references: Berry and Fristedt (1980a, 1980b), Gittins (1983).

Vogel, W. (1964) Sequentielle Versuchspläne. *Unternehmensforshung, Operations Research – Recherche Opérationnelle* **8**: 65–74.

The myopic strategy is considered for two Bernoulli arms where (θ_1, θ_2) is either (a, b) or (b, a). Discounting is n-horizon uniform. As a function of $(\theta_1, \theta_2) \in [0, 1] \times [0, 1]$, the regret approaches a finite limit as $n \to \infty$ if $|\log(a/b)/\log[(1-a)/(1-b)]|$ lies between $(1-\theta_1)/\theta_1$ and $(1-\theta_2)/\theta_2$; it does, for example, when $(\theta_1, \theta_2) = (a, b)$ or (b, a).

Related references: Our Section 9.1; Feldman (1962), Reimnitz (1978), Vogel (1960c).

Wahrenberger, D. L., Antle, C. E. and Klimko, L. A. (1977) Bayesian rules for the two-armed bandit problem. *Biometrika* **64**: 172–174.

There are two independent Bernoulli arms and n stages. Allocations are sequential in an experimental phase consisting of r stages; a single arm is used thereafter. However, we cannot tell whether $\mathbf{A} = (1, 1, \ldots 1, T, 0, 0, \ldots)$ as in Meeter (1973), or if A is the r-horizon uniform with the better arm being used in a follow-on second stage not expressly considered in the objective. (Also, we cannot tell which value of r was used in constructing their tables but we think it was 50 throughout, which means $r = n$ in half the cases considered and $r = n/2$ in the other half.) Simulations are performed assuming that θ_1 and θ_2 have the same beta distribution. The effect of having the wrong prior distribution is considered by generating θ_1 and θ_2 using different beta distributions. The authors conclude that the strategies are robust. We feel that the priors considered are not sufficiently disparate to warrant this conclusion.

The strategy that was best among those strategies considered by Fox (1974) does not fare well compared with the strategies discussed above.

Related references: Our Chapter 7; Berry (1972), Fox (1974); and see Colton (1963).

Whittle, P. (1980) Multi-armed bandits and the Gittins index. *J. R. Statist. Soc. B* **42**: 143–149.

Whittle gives an alternative proof of the Gittins–Jones (1974) theorem. This result is given as Theorem 6.1.1 in our Chapter 6 and the proof we use is essentially that of Whittle.

Related references: Our Section 6.1; Gittins (1979), Gittins and Jones (1974), Glazebrook (1983a), Roberts and Weitzman (1980), Varaiya, Walrand, and Buyukkoc (1983), Whittle (1982).

Whittle, P. (1981) Arm-acquiring bandits. *Ann. Probab.* **9**: 284—292.

Each of a number of independent arms can be in one of a finite number of 'states', which are distributions in our description of the problem. (This finiteness restriction limits the paper's applicability, but Whittle claims it 'could almost certainly be relaxed'.) Time is discrete and the number of arms in each state can increase with time. Such an increase is random and independent of the number of arms in each state and of the strategy used.

Discounting is geometric. Using an approach similar to that of Whittle (1980), the author extends the Gittins–Jones (1974) result to this case: namely, he shows that 'Gittins index policies' are optimal (see Section 6.1).

Related references: Our Section 6.1; Gittins and Jones (1974), Gittins and Nash (1977), Whittle (1980, 1982).

Whittle, P. (1982) *Optimization Over Time: Dynamic Programming and Stochastic Control*, Vol I, Wiley, New York.

Chapter 14 treats bandit problems with geometric discounting. Whittle proves the Gittins–Jones (1974) result along the lines of Whittle (1980), see our Section 6.1. A number of applications are discussed.

Related references: Our Section 6.1; Gittins (1979), Gittins and Jones (1974), Glazebrook (1983a), Roberts and Weitzman (1980), Varaiya, Walrand, and Buyukkoc (1983), Whittle (1980, 1981).

Whittle, P. (1983). *Optimization Over Time: Dynamic Programming and Stochastic Control*, Vol. II, Wiley, New York.

This is included here since it is a continuation of Whittle (1982); it does not treat bandit problems.

Witmer, J. A. (1983) Bayesian multistage decision problems. Ph.D. thesis, Univ. of Minnesota, USA.

The problem is the same as the two-phase trial considered by Berry and Pearson (1984) with arm 2 known, except that the horizon n is unknown. A strategy specifies a first-phase length r, which arm to use at each stage up to r, and then, depending on results from the first phase, which arm to use for the duration of the trial (if n turns out to be larger than r). Optimal strategies depend on the distributions of n and θ_1, which are assumed to be independent. For reasons discussed in our Chapter 3, the problem is equivalent to discounting observations assuming $\mathbf{A} = (\alpha_1, \alpha_2, \ldots)$ where α_m is the probability that $n \geqslant m$. Witmer characterizes optimal strategies when $\mathbf{A}$ is *regular*, (see our Definition 5.2.1). He gives special consideration to the case in which n has a geometric distribution; he concludes in that case that the problem is then equivalent to one in which n is assumed known, and equal to the mean of the geometric.

Related references: Our Chapters 3 and 5; and see Colton (1963).

Witten, I. H. (1973) Finite-time performance of some two-armed bandit controllers. *IEEE Trans. Systems Man Cybernet*, **3**: 194–197.

A finite-memory strategy is suggested and compared with two other strategies, one of which was put forth by Cover and Hellman (1970). All three strategies achieve the maximum possible long-run success proportion (among

finite-memory strategies), but have quite different finite behaviour.
Related references: See Robbins (1956).

Witten, I. H. (1974) On the asymptotic performances of finite-state two-armed bandit controllers. *IEEE Trans. Systems Man Cybernet*, **4**: 465–467.

Witten notes that many strategies have the same long-run success proportion. As an alternative, he suggests a two-parameter asymptotic measure of the expected number of selections of the inferior arm.
Related references: See Robbins (1956).

Witten, I. H. (1976) The apparent conflict between estimation and control – a survey of the two-armed bandit problem. *J. Franklin Inst.* **301**: 161–189.

This is a survey article which gives an excellent review of the then existing literature concerning finite-memory and finite-state bandit problems, and some discussion of the Bayesian approach.
Related references: Bellman (1956), Bradt, Johnson, and Karlin (1956); and see Robbins (1956).

Witten, I. H. (1977a) An adaptive optimal controller for discrete-time Markov environments. *Informat. and Control* **34**: 286–295.

A general Markov control setting is considered; discounting is geometric. The arm selected next is allowed to depend on the previous and also the penultimate selection.
Related references: See Robbins (1956).

Witten, I. H. (1977b) Exploring, modelling, and controlling discrete sequential environments. *Internat. J. Man-Mach. Stud.* **9**: 715–735.

Witten considers a number of control problems, including bandit problems, in a manner similar to Witten (1977a).
Related references: See Robbins (1956).

Witten, I. H. (1983) Non-deterministic modelling and its application in adaptive optimal control. *Mathematical Learning Models–Theory and Algorithms* (eds U. Herkenrath, D. Kalin and W. Vogel), pp. 213–226, Springer-Verlag, New York.

This is a survey paper in which the author discusses bandit problems along the lines of Witten (1977a).
Related references: See Robbins (1956).

Woodroofe, M. (1976) On the one arm bandit problem. *Sankhyā A* **38**: 79–91.

There are two arms. Arm 2 is known to always give 0. Discounting is geometric with factor α. Woodroofe is interested in results for $\alpha \to 1$.

In one setting arm 1 is uniformly distributed on $(-1+2\theta_1, 1)$, where the conditional (conditional on Q_1 in our notation) mean θ_1 is distributed according to the distribution measure $F_1(du) = ab^a(1-u)^{-(a+1)}du$, $u < 1-b$, where $a > 1$ and $b > 0$. If $b \leqslant 1$ it is clearly optimal to always select arm 2 and $V((a, b), 0; \alpha) = 0$. For $0 < b < 1$, Woodroofe shows that $E[\max\{0, \theta_1\} | F_1]/(1-\alpha) - V((a, b), 0; \alpha) \to ab^a/(a-1)$ as $\alpha \uparrow 1$. He shows in addition that the same expression holds if the maximal worth V is replaced by the worth of the strategy that indicates arm 1 until (if ever) it gives a number $\leqslant -1$, and arm 2 thereafter.

In a second setting, the unknown distribution Q_1 on arm 1 is normal with variance 1. The mean θ_1 is unknown and its distribution F_1 is normal with mean μ and variance ρ^2 (cf. our Section 2.1). It is shown that $E[\max\{0, \theta_1\} | F_1] - V((\mu, \rho), 0; \alpha) \sim [2\rho\sqrt{(2\pi)}]^{-1} \cdot \exp[-\mu^2/(2\rho^2)] \log^2[1/(1-\alpha)]$ as $\alpha \uparrow 1$. A family of strategies is discussed in this setting.

Woodroofe makes significant use of the methods employed in [Chow, Y. S., Robbins, H., and Siegmund, D. (1970). *Great Expectations.* Houghton Mifflin, Boston].

Related references: Our Chapter 5; Kelly (1981).

Woodroofe, M. (1979) A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.* **74**: 799–806.

This paper is unique to our knowledge in that it considers the following modification of a bandit problem that has much practical relevance. The response of an arm depends on a concomitant variable (or 'covariate') which can be measured and that has a known distribution. For example, this covariate may be severity of a disease, age of a machine, difficulty of a task, etc. The objective is the same: maximize the expected discounted sum of the observations.

Discounting is geometric. There are two arms. Arm 2 is known to always give 0. The outcome X_{1m} on arm 1 at stage m can be written as $X_{1m} = R_m + Y_m + \theta_1$, where $E(Y_m) = 0$, $\{R_m : m = 1, 2, \ldots\} \cup \{Y_m : m = 1, 2, \ldots\} \cup \{\theta_1\}$ is an (unconditionally) independent family of random variables, and the vectors (R_1, Y_1), $(R_2, Y_2), \ldots$ are identically distributed. At stage m, R_m is the concomitant variable – it is observed (but is not added to the payoff) before an arm is selected for stage m.

Woodroofe makes certain assumptions concerning the various distributions; in particular, he assumes that the support of the distribution of R_m is unbounded above. He proves that the discrepancy between $(1-\alpha)^{-1} E(\max\{X_{1m}, 0\})$ and the maximal worth is asymptotic to $c \log(1/(1-\alpha))$ as $\alpha \uparrow 1$, where c is a constant that depends on the distribution of R_m and Y_m. He shows that the same is true for the worth of a myopic strategy. The reason such a result is possible is that the presence of an

unbounded R_m assures that a myopic strategy will indicate arm 1 infinitely often.

We note that the result of Gittins and Jones (1974) applies to this concomitant-variable setting.

Related references: Bather (1980, 1981), Gittins (1979), Gittins and Jones (1974), Woodroofe (1976).

Woods, P. J. (1959) The effects of motivation and probability of reward on two-choice learning. *J. Exp. Psychol.* **57**: 380–385.

The experimental procedure is similar to that of Brand, Sakoda, and Woods (1957) and Brand, Woods, and Sakoda (1956). However, an additional instruction makes that current setting different from what we term a 'bandit'. Namely, the subject is told that she is to try to outguess the experimenter. This makes the problem more like a sequential two-person game than a problem of decision under uncertainty.

Related references: Brand, Sakoda, and Woods (1957), Brand, Woods, and Sakoda (1956).

Yakowitz, S. J. (1969) *Mathematics of Adaptive Control Processes*, Prentice-Hall, Englewood Cliffs, N.J.

Chapter 4 contains an elementary account of two-armed bandits with one arm known. Examples are carried out for n-horizon uniform discounting. The development of Bellman (1956) for geometric discounting is given. Yakowitz also discusses the history of the two-armed bandit problem.

Related references: Our Chapters 1 and 5; Bellman (1956), Berry and Fristedt (1979), Bradt, Johnson, and Karlin (1956).

Zaborskis, A. A. (1976) Sequential Bayesian plan for choosing the best method of medical treatment. *Avtomatika i Telemekhanika* **II**: 144–153.

There are k Bernoulli arms. One of the arms has known success probability a (but which one is not known), and the other $k-1$ have success probability $b < a$. (Zaborskis also allows these latter success probabilities to be arbitrary but none can be greater than b.) The discount sequence is n-horizon uniform, and $n = \infty$ is allowed by considering regrets, which are shown to be finite. Zaborskis shows that myopic strategies are optimal for any n: the decision maker should always select an arm which has the largest probability of being the a-arm. This result generalizes Feldman (1962) and is itself generalized by Rodman (1978), who allows arbitrary distributions and also considers geometric discounting.

An interesting feature of this paper is that it cites an actual clinical trial that was conducted as a bandit problem (this trial is unique in this respect as far as we know); namely, the myopic strategy described above was followed.

Unfortunately, while we applaud such an application, we do not believe that the prior distribution was appropriate. The three arms involved were three different doses of the same drug (isoptin, which is used to treat cardiac arrhythmias). It does not seem reasonable to suppose that one dose is $100a$ percent effective – equally likely to be low, medium, and high – while the other two are only $100b$ percent effective. (It was assumed that $b = 0.6$ and $a = 0.8$. Predictably, after the first few patients were given the lower doses, the rest received the high dose.) Rather, this seems to be a problem in estimating a dose–response relationship: higher doses are bound to be more effective but the minimum effective dose should be used to reduce side effects.

Related references: Our Corollary 4.3.10; DeGroot (1970), Fabius and van Zwet (1970), Feldman (1962), Kelley (1974), Quisel (1965), Rodman (1978), Vogel (1960c, 1964).

Zelen, M. (1969) Play the winner rule and the controlled clinical trial. *J. Amer. Statist. Assoc.* **64**: 131–146.

There are two Bernoulli arms with n-horizon discounting. The trial is in two phases. Two strategies are considered for the first or experimental phase: (i) 'play-the-winner/switch-from-a-loser', and (ii) equal allocation. In both cases the arm yielding more successes is selected exclusively in the terminal phase.

Zelen concludes that the optimal length of the first phase should be $n/3$ for both types of allocation. For the equal allocation case he cites Colton (1963) who uses an unusual 'local maximin' argument. For the play-the-winner/switch-from-a-loser case Zelen uses a Taylor series argument on the top of page 143 that we think is incorrect. In either case we fail to see how any fixed fraction of n has merit. Suppose $r(n)$ is the first-phase length. Both allocations discussed require that a fraction of the $r(n)$ be used on the inferior arm (unless, in the case of play-the-winner/switch-from-a-loser, $\max\{\theta_1, \theta_2\} = 1$). To achieve a long-run success proportion of $\max\{\theta_1, \theta_2\}$, the design must satisfy $r(n) \to \infty$ and $r(n)/n \to 0$ as $n \to \infty$ (unless $\theta_1 = \theta_2$ or $\max\{\theta_1, \theta_2\} = 1$). Colton (1963) and Cornfield, Halperin, and Greenhouse (1969) find Bayesian solutions for which r has order of magnitude $\sqrt{n}$, as do Canner (1970) and Berry and Pearson (1984), among others.

Using tables and assuming particular values of θ_1 and θ_2, Zelen gives the overall proportion of successes achieved by using play-the-winner/switch-from-a-loser allocation in the first phase. Such a design is better than equal allocation, but, as we have indicated elsewhere (see Percus and Percus, 1984, for example), play-the-winner/switch-from-a-loser allocation has little to recommend itself in any absolute sense. In the case at hand, suppose the values of θ_1 and θ_2 are known but not which is θ_1. Then an extension of the Feldman (1962) result to the current two-phase setting applies to show that myopic strategies are optimal. Such strategies stay with a winner but there are

a great many instances in which they stay with a loser as well. One way of viewing Zelen's approach (he is far from being unique in this regard) is that he uses the two-phase design to have a way of getting out from using a bad first-phase allocation. If the decision maker's hands are not tied then there is no reason at all to want the first phase to end (so $r(n)$ would equal n). In practice, of course, there are good reasons which force one to stop experimenting eventually. But putting unnecessary restrictions on strategies to bring this about artificially gives mathematical exercises devoid of relevance beyond the mathematics.

Related references: See Colton (1963).

Name index

Subject index

Symbol index

General

Specific to discrete time

Specific to real (continuous) time