2006

Bayesian Analysis, Endogenous Data,and Convergence of Beliefs

Andrew T. Foerster
Virginia Commonwealth University

BAYESIAN ANALYSIS, ENDOGENOUS DATA, AND CONVERGENCE OF BELIEFS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

by

ANDREW T. FOERSTER

B.A. Davidson College 2004

Director: HASSAN SEDAGHAT
ASSOCIATE PROFESSOR, DEPARTMENT OF MATHEMATICS

Director: PATRICIA PEPPLE WILLIAMSON
ASSOCIATE PROFESSOR,
DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH

Virginia Commonwealth University
Richmond, Virginia

Acknowledgements

I would like to thank my co-directors, Dr. Hassan Sedaghat and Dr. Patricia Pepple Williamson, for reading multiple drafts and providing their comments and suggestions. I would also like to thank colleagues at the Federal Reserve Bank of Richmond for informative discussion and comments, especially Huberto Ennis and Hubert Janicki. Finally, I am grateful to the Federal Reserve Bank of Richmond for financial assistance to support my studies.

Table of Contents

Abstract

BAYESIAN ANALYSIS, ENDOGENOUS DATA, AND CONVERGENCE OF BELIEFS

By Andrew T. Foerster, M.S.

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2006

Major Director: Dr. Hassan Sedaghat
Associate Professor, Department of Mathematics

Major Director: Dr. Patricia Pepple Williamson
Associate Professor, Department of Statistics and Operations Research

Problems in statistical analysis, economics, and many other disciplines often involve a trade-off between rewards and additional information that could yield higher future rewards. This thesis investigates such a trade-off, using a class of problems known as bandit problems. In these problems, a reward-seeking agent makes decisions based upon his beliefs about a parameter that controls rewards. While some choices may generate higher short-term rewards, other choices may provide information that allows the agent to learn about the parameter, thereby potentially increasing future rewards. Learning occurs if the agent's subjective beliefs about the parameter converge over time to the parameter's true value. However, depending upon the environment, learning may or may not be optimal, as in the end, the agent cares about maximizing rewards and not necessarily learning the true value of the underlying parameter

Chapter 1: Introduction

Statistical experiments and analysis often consider situations where a large quantity of data are collected and then the statistician makes a decision, such as estimation of a parameter. Another class of experiments requires witnessing only a small amount of data, making a decision, witnessing more data, making another decision, witnessing even more data, and so forth. In economics, a similar setup involves rewards generated by an unknown stochastic process, coupled with actions that yield rewards, and the actions can generate information about the stochastic process.

In general, problems of sequential analysis have one key trade-off: the decision-maker can choose between either a known outcome or an unknown outcome. Choosing the known outcome has well-defined benefits, but choosing the unknown outcome can be beneficial in that the decision-maker can gain information about the nature of the unknown outcome. Specifically, the decision-maker has a set of possible actions to choose from. One subset of those actions generates rewards from some process that is known and understood by the decision maker. A second subset of possible actions generates rewards from an unknown process, but choosing one of these actions helps reveal information about the unknown process, allowing learning to occur.

Under what circumstances is it in the decision maker's best interest to take a risk and choose an unknown with hopes of learning? Learning in sequential analysis depends highly on endogenous data, which comes from an unknown process, but it leads to learning. Using Bayesian analysis, what are the best guidelines for making decisions? Using the optimal decision-making rules, learning may or may not occur, depending upon whether it is in the best interest of the decision-maker. There are certain conditions that a sequential analysis problem must satisfy to ensure learning, it is not always guaranteed.

This paper discusses learning and endogenous data in a sequential analysis framework using

Bayesian analysis. The next two chapters provide a background: Chapter 2 discusses the measure theory definitions needed, Chapter 3 outlines other key concepts, such as optimality and martingales. The remainder of the paper discusses bandit problems, which are the typical sequential analysis problems. Chapter 4 establishes results regarding the most basic framework. Chapter 5 generalizes Chapter 4 by allowing more abstract mathematical spaces and functions. Learning, which is defined as the convergence of subjective probability beliefs, is the subject of Chapter 6. Chapter 7 presents an application of the results, and Chapter 8 concludes.

Chapter 2: Measure Theory Concepts and Definitions

While many applications of bandit problems will use familiar functions and parameter spaces, the fact that the results apply to a much larger class of objects expands their use, applicability, and generality. Chapter 4 presents a simple example that uses finite choices for parameters, and Chapter 5 expands and generalizes to include more general sets. This chapter defines key measure theory concepts needed in Chapter 5. First, the spaces for parameters may be metric spaces.

2.1 Metric spaces

In the most common space $\mathbb{R}$, the concept of distance, using the absolute value function, plays a critical role in establishment of a vast number of topics. Any topic that relies upon absolute value, such as continuity, convergence of sequences, and open or closed sets, implicitly depends upon how close two points or objects are from one another, or the distance between them.

It is easy to see that in $\mathbb{R}$, the distance between two points x and y, $d(x,y)$, is the absolute value of their difference, so $d(x,y) = |x-y|$. Similarly, this concept is extended to Euclidean space $\mathbb{R}^n$ by an expansion of the Pythagorean Theorem, so in $\mathbb{R}^3$ the distance between the points $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$ is given by

$$d(x,y) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + (y_3 - x_3)^2}. \tag{1}$$

In general the distance between points in $\mathbb{R}^n$ is

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}. \tag{2}$$

In all of these cases, the distance between two points is a positive real number, while the points are in some space. In $\mathbb{R}^n$, the distance function is a function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$.

The distance function in $\mathbb{R}^n$ has several characteristics that, when generalizing the function to more abstract spaces, it will be necessary to maintain. Specifically, for points $x, y, z \in \mathbb{R}^n$ the distance function has the properties:

1. positive: $d(x,y) \geq 0$, where $d(x,y) = 0$ if and only if $x = y$,
2. symmetric: $d(x,y) = d(y,x)$, and
3. triangle inequality: $d(x,y) + d(y,z) \geq d(x,z)$.

To generalize the concept of the distance function to spaces that may not be as easily tractable as Euclidean space, it is important to keep these properties of the distance function. Now, given any arbitrary set, a metric on that set will generalize distance.

Definition 1 *A metric on a set Ω is a function $\rho : \Omega \times \Omega \to \mathbb{R}$ that satisfies the following conditions for all $x, y, z \in \Omega$:*

1. *positive: $\rho(x,y) \geq 0$, where $\rho(x,y) = 0$ if and only if $x = y$,*
2. *symmetric: $\rho(x,y) = \rho(y,x)$, and*
3. *triangle inequality: $\rho(x,y) + \rho(y,z) \geq \rho(x,z)$.*

It is easy to see that the distance function $d(\cdot,\cdot)$ is a metric for $\mathbb{R}^n$. So the term metric applies to the extension of the distance concept, and any set that has a metric on it is a metric space.

Definition 2 *A metric space is a pair (Ω, ρ), where Ω is a set and ρ is a metric on that set. If the metric ρ is obvious or the exact function is unimportant, then the metric space is denoted by simply the set, Ω.*

From the preceding discussion on distance and Euclidean space, it is clear that $(\mathbb{R}^n, d)$ is a metric space. Since the metric d is obvious in this case, it is said that $\mathbb{R}^n$ is a metric space.

Next, the concept of complete metric spaces will be used later. The definition of completeness depends upon the concept of Cauchy sequences, which first are extended to metric spaces.

Definition 3 *A Cauchy sequence in a metric space (Ω, ρ) is a sequence $\{x_n\}_{n=1}^{\infty} \subset \Omega$ where for all $\varepsilon > 0$, there exists some $N \in \mathbb{N}$ such that if $n, m \geq N$, then $\rho(x_n, x_m) < \varepsilon$.*

Now using this concept of Cauchy sequences, the definition of complete metric spaces follows.

Definition 4 *A complete metric space is a metric space (Ω, ρ) where every Cauchy sequence in Ω converges. In other words, if $\{x_n\}_{n=1}^{\infty}$ is a Cauchy sequence, with $x_n \in \Omega$ for all $n \in \mathbb{N}$, then there exists $x \in \Omega$ such that $\lim_{n\to\infty} x_n = x$.*

Continuing with the example of $\mathbb{R}^n$, using the Euclidean distance metric $d(\cdot, \cdot)$, $\mathbb{R}^n$ is a complete metric space.

Finally, later discussion requires the notion of a separable metric space.

Definition 5 *A separable metric space is a metric space (Ω, ρ) such that there exists $E \subset \Omega$ where E is a countable, dense subset of Ω, meaning E is countable and its closure equals the space, $\overline{E} = \Omega$.*

As a quick example of a separable metric space, consider the metric space $(\mathbb{R}, |\cdot|)$. The set of rational numbers, $\mathbb{Q}$, is both countable and dense, since $\overline{\mathbb{Q}} = \mathbb{R}$. Consequently, the metric space of the real line with a metric of absolute value is separable.

2.2 Borel Sigma Algebras

Another necessary extension of Euclidean space to metric spaces is to define σ-algebras, specifically the Borel σ-algebra. First, recall the definition of a σ-algebra, which does not depend upon the space.

Definition 6 *A σ-algebra is a collection of subsets $\mathcal{A}$ of a set Ω that have the properties:*

1. *closed under complementarity: if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$, and*
2. *closed under countable union: if $\{A_n\}_{n=1}^{\infty} \subset \mathcal{A}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.*

In addition to the definition of a σ-algebra, Borel σ-algebras depend upon defining the smallest σ-algebra containing any collection of subsets. The following Theorem ensures that such a σ-algebra exists.

Theorem 7 *Let Ω be a set, and let $\mathcal{C}$ be any nonempty collection of subsets of Ω. Then there exists a smallest σ-algebra of subsets of Ω containing $\mathcal{C}$, called the σ-algebra generated by $\mathcal{C}$, and is denoted $\sigma(\mathcal{C})$.*

Now, given any set and any arbitrary collection of its subsets, it is possible to define the smallest σ-algebra containing that arbitrary collection. More specific to the purposes of this paper, if the set is a metric space, it is possible to define σ-algebras on that space using subsets of the space. In fact, the Borel σ-algebra for a metric space is the smallest σ-algebra generated by the open sets of the metric space.

Definition 8 *The Borel σ-algebra of a metric space Ω is the σ-algebra $\sigma(\mathcal{O})$, where $\mathcal{O}$ is the collection of open sets in Ω, otherwise known as the σ-algebra generated by the collection of open sets.*

The properties of σ-algebras allow the reduction of the collection of open sets to consider one specific type of open set, especially in $\mathbb{R}$. Therefore, define $\mathcal{B}$ to be the Borel σ-algebra on $\mathbb{R}$. That is,

$$\mathcal{B} = \sigma((-\infty, a)), \text{ for } a \in \mathbb{R}. \tag{3}$$

2.3 Topologies

Another type of space that will be used in later chapters is that of a topological space. First, the definition of topology follows.

Definition 9 *A topology on a set Ω is a collection $\mathcal{T} \subset \Omega$ where*

1. $\emptyset, \Omega \in \mathcal{T}$,
2. *if $S \subset \mathcal{T}$, then $\cup_{O \in S} O \in \mathcal{T}$, and*
3. *if $O_1, O_2 \in \mathcal{T}$, then $O_1 \cap O_2 \in \mathcal{T}$.*

The definition of a topology directly leads to the concept of a topological space.

Definition 10 *A topological space is a pair $(\Omega, \mathcal{T})$ where $\mathcal{T}$ is a topology on the set Ω.*

For any given set Ω, there may be multiple topologies on that set. Given two topologies on Ω, say $\mathcal{S}$ and $\mathcal{T}$, if $\mathcal{S} \subset \mathcal{T}$, then $\mathcal{S}$ is weaker than $\mathcal{T}$. Given Ω, it then makes sense to consider the weakest possible of all topologies on Ω.

Definition 11 *A weak topology for a set Ω is the topology given by $\mathcal{T}_F = \cap_{\mathcal{T} \in \mathbb{T}} \mathcal{T}$, where $\mathcal{F}$ is a set of functions such that for every $f \in \mathcal{F}$, $(\Lambda_f, \mathcal{T}_f)$ is a topological space and $f : \Omega \to \Lambda_f$, and $\mathbb{T}$ is the set of topologies in Ω for which the functions $f \in \mathcal{F}$ are continuous.*

One important characteristic of a weak topology $\mathcal{T}_F$ is that all $f \in \mathcal{F}$ are continuous, and if $\mathcal{S} \subsetneq \mathcal{T}_F$ is another topology then there is some function $f' \in \mathcal{F}$ such that $f : \Omega \to \Lambda_f$ is not continuous with respect to $\mathcal{S}$.

2.4 Probability Spaces

The use of probability distributions across potential events that comprise metric spaces makes it necessary to discuss probability spaces. However, probability and probability spaces depend on the concept of measure and measure spaces, which are now defined.

Definition 12 *A measure on a σ-algebra of subsets $\mathcal{A}$ of a set Ω is a function $\mu : \mathcal{A} \to \mathbb{R}$, that has the following properties:*

1. *non-negative:* $\mu(A) \geq 0$ *for every* $A \in \mathcal{A}$,

2. *zero for the empty set:* $\mu(\emptyset) = 0$, *and*

3. *countable additivity: if* $\{A_n\}_n \subset A$ *are pairwise disjoint, so* $A_i \cap A_j = \emptyset$ *if* $i \neq j$, *then* $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$.

If such a function exists for a Ω and an $\mathcal{A}$, then $(\Omega, \mathcal{A})$ is a measurable space. The definition of measure space immediately follows from these definitions.

Definition 13 *A measure space is a collection* $(\Omega, \mathcal{A}, \mu)$, *where* Ω *is a set,* $\mathcal{A}$ *is a* σ*-algebra of subsets of* Ω, *and* μ *is a measure on* $\mathcal{A}$.

The definition of a probability space is a measure space where the measure of the entire set equals one, and a probability measure is the measure on that space..

Definition 14 *A probability space is a measure space* $(\Omega, \mathcal{A}, \mu)$ *such that* $\mu(\Omega) = 1$. *A probability measure is the measure* μ *that satisfies this condition, and is usually denoted* P.

The fact that probability spaces are sets means that it is possible to define a probability space of a probability space. Specifically, if $P(\Omega)$ denotes the probability space $(\Omega, \mathcal{A}, \mu)$, $\mathcal{A}'$ is a σ-algebra of sets of $P(\Omega)$, and μ' is a measure on $\mathcal{A}'$ with $\mu'(P(\Omega)) = 1$, then $P(P(\Omega)) = (P(\Omega), \mathcal{A}', \mu')$ is also a probability space. This definition will be important when transition to subjective probability distributions arises in discussion of the bandit problem. But before probability distributions on probability spaces are discussed, it is necessary to define random variables on these spaces.

Definition 15 *A random variable on a probability space* $(\Omega, \mathcal{A}, \mu)$ *is a real-valued function* $X : \Omega \to \mathbb{R}$ *where* $\{\omega : X(\omega) \in B\} \in \mathcal{A}$ *for each* $B \in \mathcal{B}$.

This definition of random variables says the function X maps events onto the real line. Once random variables represent all possible events as real numbers, probability distributions represent the chance of those events occurring.

Definition 16 *A probability distribution of a random variable X on the probability space $(\Omega, \mathcal{A}, \mu)$ is the set function on $\mathcal{B}$, $\pi_X(B) = P(X \in B)$.*

Therefore, the probability distribution function gives the probability that a random variable is in a given set B. Letting B vary across all possible collections in $\mathcal{B}$ shows that the probability distribution is defined on all possible real values that the random variable X maps to. Also, it is important to note that since $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$ and $\pi_X(\mathcal{B}) = 1$, then $\pi_X(\mathcal{B})$ is a probability measure on $\mathcal{B}$.

Finally, the discussion of probability depends on almost surely convergence, which is analogous to the concept of almost everywhere in measure theory.

Definition 17 *A property holds μ almost everywhere (μ-ae) for a measurable space $(\Omega, \mathcal{A})$, if it holds except for perhaps some $D \in \mathcal{A}$, where $\mu(D) = 0$. A sequence of random variables X_n converges almost surely to X, or with probability one, if the sequence converges P-ae. Then $P(\lim_{n\to\infty} X_n = X) = 1$.*

2.5 Radon-Nikodym Derivatives

A final measure theory concept to address in order to analyze the general cases of the Bandit problem is the Radon-Nikodym derivative. As the Lebesgue integral allows integration of functions with respect to Lebesgue measure, finding a method of differentiating functions with respect to some measure, not necessarily Lebesgue, helps improve generality. The Radon-Nikodym derivative applies to σ-finite measure spaces, which are now defined.

Definition 18 *A σ-finite measure space is a measure space $(\Omega, \mathcal{A}, \mu)$ with a sequence $\{A_n\}$ of sets, where each A_n is $\mathcal{A}$-measurable, $\cup_n A_n = \Omega$, and $\mu(A_n) < \infty$ for all n.*

A measure space is therefore σ-finite if the entire set can be expressed as the union of finite-measured sets in the σ-algebra. With this concept in hand, it is time to turn to the concept of continuity in measure space.

In $\mathbb{R}$, the concept of integration and differentiation are significantly intertwined. By the Second Fundamental Theorem of Calculus, if, for $t \in [a,b]$, the function $f(t)$ is continuous, and if $F(x) = \int_a^x f(t)\,dt$, then $F'(x) = f(x)$. A similar method of inverting integration with respect to some measure gives the more general derivative.

Differentiation depends upon the concept of continuity, which is now defined for general measure.

Definition 19 *An absolutely continuous measure ν with respect to measure μ on the measurable space $(\Omega, \mathcal{A})$ is measure ν such that $\mu(A) = 0$ implies $\nu(A) = 0$ for any $A \in \mathcal{A}$.*

The definition of absolutely continuous measures leads directly to the Radon-Nikodym Theorem.

Theorem 20 *(Radon-Nikodym Theorem) Let $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space and ν a σ-finite measure on $\mathcal{A}$. If ν is absolutely continuous with respect to μ, then there exists a nonnegative extended real-valued $\mathcal{A}$-measurable function, f, on Ω such that*

$$\nu(A) = \int_A f d\mu, \text{ for all } A \in \mathcal{A}.$$

In addition, if there exists a nonnegative extended real function $\mathcal{A}$-measurable function g such that $v(A) = \int_A g d\mu$ for all $A \in \mathcal{A}$, then $g = f$, μ-ae., meaning that the function f is unique except perhaps on a set N with $\mu(N) = 0$.

As in the basic case discussed above, the integral now contains all of the information needed to define the Radon-Nikodym derivative.

Definition 21 *The Radon-Nikodym derivative of ν with respect to μ, denoted $\frac{d\nu}{d\mu}$, is the function f in the Radon-Nikodym Theorem, where f, ν and μ satisfy the conditions of that theorem.*

The second part of the Radon-Nikodym Theorem ensures that the Radon-Nikodym derivative is unique, except on sets of μ-measure zero.

Chapter 3: Other Concepts and Definitions

In addition to measure theory concepts and definitions, two other sets of concepts and definitions must be addressed before turning to the actual bandit problems. The two concepts are aspects of optimality and martingales. Discussion of converging subjective beliefs will depend upon martingales, specifically the Martingale Convergence Theorem; the necessary concepts are discussed in Section 3.2. Before turning to martingales, the concept of optimality ensures that decisions that the agent or statistician makes are the best possible decisions given the information at the time of decision-making.

3.1 Optimality

Decision theory depends upon the agent making the best possible decision to maximize gain or minimize loss. In Section 5.3, the discussion of optimal decision rules depends upon several definitions of optimality, especially to ensure existence of strategies that are optimal. In fact, producing results from the generalized bandit problem depends upon framing the problem in terms of a dynamic programming problem. Dynamic programming problems in general solve optimization problems by considering subsets of those problems. In the context of bandit problems and sequential analysis, there is a set of states of nature, a set of rewards, and probability of transitioning from one state of nature to another.

A dynamic programming problem consists of several elements. Let A be a Borel set, which denotes the choice of possible actions. Let S be another Borel set, which denotes possible states of nature. Define a function $q : S \times A \to S$, which is the transition function, mapping current state and decision onto state for next period. Let $x : S \times A \times S \to \mathbb{R}$ be a reward function, which depends upon current and future states as well as action. Future rewards are discounted by some factor.

A given dynamic programming problem can therefore be summarized by (A, S, q, x). Let $d(\cdot)$ be some decision rule that indicates what action to take in any given state; in principle, this decision rule can depend upon all previous realizations of states and actions. A stationary decision rule is a decision rule $d(s)$, where $s \in S$ denotes the current state, meaning only the current state factors into the decision.

For a certain decision rule, denote the sum of all expected future benefits from using that rule to be $U(d)$. Since there will be multiple decision rules, each yielding different future results, let V denote the supremum of the U's over all possible decision rules. That is, $V = \sup_d U(d)$.

These basic aspects of a dynamic programming problem yield the following three definitions of optimality. The first, pure optimality, is the best a decision rule can possibly achieve.

Definition 22 *An optimal decision rule d^* for the dynamic programming problem (A, S, q, x) is one in which $U(d^*) = V$.*

If a decision rule is not optimal, the next best alternative is that it can come as close as possible to achieving the value V. In the case of ε-optimality, the goal is to come within some arbitrary ε of the value achieved by optimality.

Definition 23 *An ε-optimal decision rule d^ε for the dynamic programming problem (A, S, q, x) is one in which, for $\varepsilon > 0$, $U(d^\varepsilon) \geq V - \varepsilon$.*

Finally, if a decision rule is not ε-optimal, the focus turns to whether there are situations in which a decision rule can be ε-optimal or not. That is, there may be some probability distribution on the state space where, given that distribution, a decision rule can achieve ε-optimality.

Definition 24 *A (p, ε)-optimal decision rule $d^{p,\varepsilon}$ for the dynamic programming problem (A, S, q, x) is one in which, for $\varepsilon > 0$ and $p \in P(S)$, $p(U(d^{p,\varepsilon}) \geq V - \varepsilon) = 1$, where $P(S)$ denotes the set of all probabilities on S.*

Several theorems regarding the existence of decision rules satisfying the varying definitions of optimality follow from these definitions. While the results of these theorems are important for later chapters, the proofs depend significantly upon dynamic programming concepts and are therefore beyond the scope of this paper.

From the above definitions, it is obvious that of the three optimality concepts, the (p, ε)-optimal is the weakest. Stationary decision rules significantly increase the tractability of dynamic programming problems, so ideally a decision rule will depend only upon the current state, meaning it is stationary. The next theorem, by Strauch (1966), ensures that such a decision rule exists.

Theorem 25 *Let (A, S, q, x) be a dynamic programming problem, let $\varepsilon > 0$ and $p \in P(S)$. Then there exists a stationary decision rule $d(s)$ that is (p, ε)-optimal.*

Again, since the previous theorem deals with the weakest of the three optimality definitions, existence of decision rules that satisfy the stronger optimality conditions would dominate these results. However, to ensure the existence of optimal decision rules, the conditions are stronger than those required for the weaker forms of optimality. The following two theorems provide conditions for the existence of optimal decision rules. The first is due to Blackwell (1965).

Theorem 26 *Let (A, S, q, x) be a dynamic programming problem where A is finite. Then there exists a stationary decision rule $d(s)$ that is optimal.*

The next theorem is from Maitra (1968).

Theorem 27 *Let (A, S, q, x) be a dynamic programming problem where A is compact, x is jointly continuous, and q is jointly continuous. Then there exists a stationary decision rule $d(s)$ that is optimal.*

From these theorems, if the setup of the bandit problem can be converted to the terms of dynamic programming problems, then it is possible to guarantee existence of decision rules that satisfy the three optimality definitions.

3.2 Martingales

Showing that the sequence of subjective distributions about some parameter converges over time to a fixed distribution, meaning that learning occurs, depends upon showing that this sequence is a martingale, and then showing, via the Martingale Convergence Theorem, that the martingale converges. First, given the definitions of probability spaces and random variables, the martingale definition follows.

Definition 28 *A martingale is a sequence* $\{X_n\}$*, where* X_n *is a random variable on a probability space* $(\Omega, \mathcal{A}, \mu)$*,* $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \mathcal{A}_3 \subset \cdots$ *is an increasing sequence of* σ*-algebras of* $\mathcal{A}$*, and* X_n *is* $\mathcal{A}_n$*-measurable, such that* $E[X_{n+1}|\mathcal{A}_n] = X_n$ *for all* $n \in \mathbb{N}$*.*

Closely related to the definition of martingales is the definition of supermartingales and submartingales.

Definition 29 *A submartingale is a sequence* $\{X_n\}$ *such that* $E[X_{n+1}|\mathcal{A}_n] \geq X_n$ *for all* $n \in \mathbb{N}$*. A supermartingale is a sequence* $\{X_n\}$ *such that* $E[X_{n+1}|\mathcal{A}_n] \leq X_n$ *for all* $n \in \mathbb{N}$*.*

From these definitions, it is clear that a sequence is a martingale if and only if it is a submartingale and a supermartingale, and all martingales are both submartingales and supermartingales. In addition, note that for $n, k \in \mathbb{N}$, $E[X_{n+k}|\mathcal{A}_n] = X_n$ for martingales, with the corresponding inequality substituted for super- or sub-martingales.[1]

Example 30 *The classic example of martingales is the gambler's wealth problem. Let* X_n *be the wealth of a gambler after* n *bets. Consider three betting systems:* (1) *a fair coin is tossed, the gambler receives one dollar if it lands heads, and loses a dollar if it lands tails,* (2) *a fair die is tossed, the gambler receives one dollar if it lands on a 5 or 6, and loses a dollar if it lands 1 through 4,* (3) *a fair die is tossed, the gambler receives one dollar if it lands 3 through 6, and loses one dollar if it lands 1 or 2. Given these frameworks, it is easy to see for each of the bets:*

[1] This fact is established fairly easily by the use of induction on k.

(1) X_n *is a martingale, since* $P(X_{n+1} = X_n + 1) = \frac{1}{2}$, $P(X_{n+1} = X_n - 1) = \frac{1}{2}$. *Then* $E[X_{n+1}|X_n] = \frac{1}{2}(X_n + 1) + \frac{1}{2}(X_n - 1) = X_n$.

(2) X_n *is a supermartingale, since* $P(X_{n+1} = X_n + 1) = \frac{1}{3}$, $P(X_{n+1} = X_n - 1) = \frac{2}{3}$. *Then* $E[X_{n+1}|X_n] = \frac{1}{3}(X_n + 1) + \frac{2}{3}(X_n - 1) = X_n - \frac{1}{3} \leq X_n$.

(3) X_n *is a submartingale, since* $P(X_{n+1} = X_n + 1) = \frac{2}{3}$, $P(X_{n+1} = X_n - 1) = \frac{1}{3}$. *Then* $E[X_{n+1}|X_n] = \frac{2}{3}(X_n + 1) + \frac{1}{3}(X_n - 1) = X_n + \frac{1}{3} \geq X_n$.

Several intermediate definitions and results need to be addressed in order to finish with the Martingale Convergence Theorem.[2] The first definition is one of stopping times.

Definition 31 *A stopping time for the martingale* $\{X_n\}$ *on the sequence of* σ*-algebras* $\{\mathcal{A}_n\}$ *is a random variable* $\tau \in \mathbb{N}$ *where* $\{\tau = n\} \in \mathcal{A}_n$.

Generally, a stopping time τ is a time with which an individual can make a decision for some question about a string of random variables. For example, consider an infinite string of random variables $(0, 1, 1, 0, 1, 1, 0, \ldots)$. There exists a stopping time for the consideration "two ones have been witnessed," since at any time, all the previous realizations can be examined to see if two ones have occurred. On the other hand, there is no stopping time for the consideration "the last one has been witnessed," since at any time, there may be (infinitely) more realizations of ones. In the context of the gambling problem, $\tau = n$ indicates the gambler plays n times, has reached some pre-set money winnings or losings, and decides to stop. In general, stopping times simply end the sequence of random variables $\{X_n\}$.

The first result is a Lemma that uses the definition of stopping times, where stopping times essentially put an end to a string of random variables, analyzing only the variables up to that time.

Lemma 32 *If* $\{X_k\}_{k=1}^n$ *is a submartingale on probability space* $(\Omega, \mathcal{A}, \mu)$, *and the stopping times* $\tau_1, \tau_2 \in \mathbb{N}$ *are such that* $1 \leq \tau_1 \leq \tau_2 \leq n$, *then* $E[X_{\tau_1}] \leq E[X_{\tau_2}]$.

[2] Doob (1953) was the first extensive discussion of martingale theory, and the proof of the Martingale Convergence Theorem has changed little over time. Ash (1972) and Billingsley (1986) present proofs with only slight modifications. I follow the standard methodology in these three sources.

Proof. Let $\{X_k\}_{k=1}^n$ be a submartingale, and define stopping times $\tau_1, \tau_2 \in \mathbb{N}$ such that $1 \leq \tau_1 \leq \tau_2 \leq n$. Let $\Delta_k = X_k - X_{k-1}$. Then

$$\begin{aligned} X_{\tau_2} - X_{\tau_1} &= X_{\tau_2} - X_{\tau_2-1} + X_{\tau_2-1} - X_{\tau_2-2} + \cdots + X_{\tau_1+1} - X_{\tau_1} \\ &= \Delta_{\tau_2} + \Delta_{\tau_2-1} + \cdots + \Delta_{\tau_1+1} \\ &= \sum_{k=\tau_1+1}^{\tau_2} \Delta_k. \end{aligned}$$

Next, since $E[X_k|X_1, \ldots, X_{k-1}] \geq X_{k-1}$, repeatedly applying expectation[3] yields $E[X_k] \geq E[X_{k-1}]$, so $E[\Delta_k] = E[X_k - X_{k-1}] = E[X_k] - E[X_{k-1}] \geq 0$.

Consequently,

$$\begin{aligned} E[X_{\tau_2}] - E[X_{\tau_1}] &= E[X_{\tau_2} - X_{\tau_1}] \\ &= E\left[\sum_{k=\tau_1+1}^{\tau_2} \Delta_k\right] \\ &= \sum_{k=\tau_1+1}^{\tau_2} E[\Delta_k] \geq 0, \end{aligned}$$

Since $E[X_{\tau_2}] - E[X_{\tau_1}] \geq 0$, $E[X_{\tau_2}] \geq E[X_{\tau_1}]$. ■

The next major result is a Theorem that depends upon the definition of upcrossings.

Definition 33 *The number of upcrossings of a sequence of random variables $\{X_1, \ldots, X_n\}$ for an interval $[a, b]$, is the number of times the sequence goes from below a to above b. In other words, it is the number of pairs $(c, d) \in \mathbb{N} \times \mathbb{N}$ such that $1 \leq c < d \leq n$, $X_c \leq a$, $b \leq X_d$, and $X_i \in (a, b)$ for all $i = c+1, c+2, \ldots, d-2, d-1$.*

Next, the following Theorem uses the definition of upcrossings and Lemma 32.

Theorem 34 *If $\{X_k\}_{k=1}^n$ is the initial segment of a submartingale, and if U is the number of*

[3] Note that $E[E[Y|X]] = E[Y]$.

upcrossings for the interval $[a,b]$*, then*

$$E[U] \leq \frac{E[|X_n|] + |a|}{b-a}.$$

Proof. Let $\{X_k\}_{k=1}^n$ be the initial segment of a submartingale, let $[a,b]$ be an arbitrary interval, with U being the number of upcrossings on $[a,b]$. First, normalize the martingale and the interval, letting $\kappa = b-a$, and for $k = 1,\ldots,n$, define $Y_k = \max\{X_k - a, 0\}$, so that U now equals the number of upcrossings of the interval $[0,\kappa]$ by the submartingale $\{Y_k\}_{k=1}^n$. Define a series of stopping times $\{\tau_k\}_{k=0}^\infty$, with $\tau_0 = 1$ according to:

$$\tau_k = \begin{cases} \min_j \{(\tau_{k-1} < j \leq n) \cap (j : Y_j \geq \kappa)\}, & k \text{ even and } \{(\tau_{k-1} < j \leq n) \cap (j : Y_j \geq \kappa)\} \neq \emptyset \\ n, & k \text{ even and } \{(\tau_{k-1} < j \leq n) \cap (j : Y_j \geq \kappa)\} = \emptyset \\ \min_j \{(\tau_{k-1} < j \leq n) \cap (j : Y_j = 0)\}, & k \text{ odd and } \{(\tau_{k-1} < j \leq n) \cap (j : Y_j = 0)\} \neq \emptyset \\ n, & k \text{ odd and } \{(\tau_{k-1} < j \leq n) \cap (j : Y_j = 0)\} = \emptyset \end{cases}.$$

Then for all $k = 1,\ldots,n$: (1) $\tau_k \leq n$, (2) if $\tau_{k-1} = n$, then $\tau_{k-1} = \tau_k = \tau_{k+1} = \cdots = n$, (3) if $\tau_{k-1} < n$ then $\tau_{k-1} < \tau_k$ (4) $\tau_n = \tau_{n+1} = \cdots = n$. So, $\{\tau_k\}_{k=1}^\infty$ is a nondecreasing sequence bounded above by n.

Then

$$\begin{aligned} Y_n &= Y_{\tau_n} \geq Y_{\tau_n} - Y_{\tau_0} \\ &\geq \sum_{k \text{ even}} (Y_{\tau_k} - Y_{\tau_k - 1}) + \sum_{k \text{ odd}} (Y_{\tau_k} - Y_{\tau_k - 1}). \end{aligned} \tag{4}$$

Taking expected values, and applying equation (4) to the expectation of the summation over odds:

$$E[Y_n] \geq E\left[\sum_{k \text{ even}} (Y_{\tau_k} - Y_{\tau_k - 1})\right] + E\left[\sum_{k \text{ odd}} (Y_{\tau_k} - Y_{\tau_k - 1})\right]$$

$$\geq \quad E\left[\sum_{k \text{ even}} (Y_{\tau_k} - Y_{\tau_k - 1})\right].$$

So now assume k is even (using summation over odds would require assuming k is odd, so this step is done without loss of generality). First consider $\tau_{k-1} = n$, so $Y_{\tau_{k-1}} = Y_n = Y_{\tau_k}$, meaning $Y_{\tau_k} - Y_{\tau_{k-1}} = 0$. Next consider $\tau_{k-1} < n$, which implies $Y_{\tau_{k-1}} = 0$. Since $Y_{\tau_k} \geq 0$, $Y_{\tau_k} - Y_{\tau_{k-1}} \geq 0$. Consequently, regardless of τ_{k-1}, $Y_{\tau_k} - Y_{\tau_{k-1}} \geq 0$.

Now assume (c, d) constitutes an upcrossing of $[0, \kappa]$ by $\{Y_k\}_{k=1}^n$, meaning $Y_c = 0$, $Y_d \geq \kappa$, and $Y_i \in (0, \kappa)$ for all $i \in \mathbb{N} \cap (c, d)$. Then there exists some k (k is even) such that $\tau_{k-1} \leq c < d = \tau_k$. Since $Y_c = 0$ and $Y_d \geq \kappa$, then $Y_{\tau_{k-1}} = 0$ and $Y_{\tau_k} \geq \kappa$. Therefore, $Y_{\tau_k} - Y_{\tau_{k-1}} \geq \kappa$. Since the pair (c, d) was arbitrary, and there are U upcrossings, each of which generates a different τ_k and τ_{k-1}, $\sum_{k \text{ even}} (Y_{\tau_k} - Y_{\tau_k - 1}) \geq U\kappa$. Therefore, $E[Y_n] \geq E[U\kappa]$. Then

$$E[U\kappa] = \kappa E[U] = (b - a) E[U] \leq E[Y_n] = E[\max\{X_n - a, 0\}] \leq E[|X_n|] + |a|.$$

■

Now all the tools needed to show convergence of submartingales are assembled.

Theorem 35 *(Submartingale Convergence Theorem) Let $\{X_n\}$ be a submartingale, and define $K = \sup_n E[|X_n|]$. If $K < \infty$, then there exists a random variable X with $E[|X|] \leq K$ and*

$$P\left(\lim_{n \to \infty} X_n = X\right) = 1.$$

Proof. Let $\{X_n\}$ be a submartingale, and let $[a, b]$ be an arbitrary interval. Define U_n to be the number of upcrossings of the interval $[a, b]$ by the finite sequence $X_1, \ldots, X_n$. Also, let $K = \sup_n E[|X_n|]$. By Theorem 34,

$$E[U_n] \leq \frac{E[|X_n|] + a}{b - a} \leq \frac{K + a}{b - a}.$$

Since the number of upcrossings never decreases as n increases, $\{U_n\}$ is a nondecreasing sequence. Consequently, as $\{U_n\}$ is nondecreasing and $E[U_n]$ is bounded, then $\sup_n U_n$ is finite valued and integrable almost everywhere by the Monotone Convergence Theorem.

Define $\overline{X} = \limsup_{n\to\infty} X_n$ and $\underline{X} = \liminf_{n\to\infty} X_n$. Clearly, $\underline{X} \leq \overline{X}$. Suppose that $\underline{X} < \overline{X}$. Then, by the density of the rationals, there exists $a, b \in \mathbb{Q}$ such that $\underline{X} < a < b < \overline{X}$. However, this condition would imply $\lim_{n\to\infty} U_n = \infty$, which is a contradiction, since $\sup_n U_n$ is finite. Consequently, $\underline{X} = \overline{X}$, so $\lim_{n\to\infty} X_n = \underline{X} = \overline{X} = X$. Then, by Fatou's Lemma, where μ is the probability measure,

$$E[|X|] = \int |X|\, d\mu = \int \lim_{n\to\infty} \inf_n |X_n|\, d\mu \leq \lim_{n\to\infty} \inf_n \int |X_n|\, d\mu = \lim_{n\to\infty} \inf_n E[|X_n|] \leq K.$$

Thus, $E[|X|] \leq K$, X is finite since it is integrable, and $P(\lim_{n\to\infty} X_n = X) = 1$. ■

Theorem 36 *(Martingale Convergence Theorem) Let $\{X_n\}$ be a martingale, let $M \in [0, \infty)$, and let $E[|X_n|] \leq M$ for all $n \in \mathbb{N}$. Then X_n converges to some random variable, denoted X, with probability 1, meaning*

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Proof. Let $\{X_n\}$ be a martingale. Then $\{|X_n|\}$ is a submartingale, so $\{E[|X_n|]\}$ is a nondecreasing sequence. Let $M \in [0, \infty)$ and $E[|X_n|] \leq M$ for all $n \in \mathbb{N}$. Then $\{E[|X_n|]\}_{n=1}^{\infty}$ is bounded. Then

$$\sup_n E[|X_n|] = \lim_{n\to\infty} E[|X_n|] = K \leq M < \infty.$$

Thus, by the Submartingale Convergence Theorem (Theorem 35), there exists a random variable X with $E[|X|] \leq K$ and

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

■

Chapter 4: Basic Bandit Problem

Bandit problems are a class of sequential analysis problems in decision theory that involve a choice between multiple options that have some sort of uncertainty about the parameters that control the options. This chapter discusses the most basic case of a bandit problem, strategies for a sequence of choices, how future choices are discounted, and how priors affect the outcome.[4]

The most basic bandit problem involves two choices that have the same reward; the difference between the choices is that the probability of success associated with choice 1 is known, while the probability of success associated with choice 2 is unknown. There are many examples of applied bandit problems such as this in real life. For example, a doctor may have a choice between two prescriptions to give out, the first medicine being an old one that the doctor knows has a given probability of curing an ailment, the second a new medicine that the doctor has uncertainty about its probability of curing the ailment. Under what conditions would the doctor prescribe the first or the second medicine? Several factors that might influence this result are the success rate of the old medicine (the known probability), the doctor's thoughts about the potential of the new drug (the prior about the uncertain probability), the cost of failing to cure the ailment (the reward of success), and the desire to gather information about the success rate of the new drug so that the doctor can make more informed decisions in the future (the ability to create data and discounting of the future).

Bandit problems are so named because the typical example uses slot machines. At every period $t = 0, 1, 2, 3, \ldots$ an agent approaches a two-armed machine. Arm 1 pays out one dollar with probability λ, which the agent knows, and arm 2 pays out one dollar with probability θ, which is

[4] Berry and Fristedt (1985) is a very complete source regarding bandit problems. The propositions in this chapter are based upon several results from Chapter 5.

unknown to the agent. The agent wishes to maximize the expected sum of discounted rewards, so his objective is

$$\max E_t \sum_{j=0}^{\infty} \beta_{t+j} X_{t+j} \tag{5}$$

where β_t is the discount factor at time t and X_t is the payout at time t.

4.1 Discounting

The discount factor has an important role in the agent's decision. Intuitively, if the agent discounts the future heavily, he will be less likely to sacrifice income today to possibly earn more in the future. The two typical discount factors are finite-horizon discounting

$$\beta_t = \begin{cases} 1, \text{ if } t \leq T \\ 0, \text{ if } t > T \end{cases} \tag{6}$$

for some $T \geq 0$, where T represents the horizon, and geometric discounting, where

$$\beta_t = \beta^t, \text{ for } \beta \in (0,1)\,. \tag{7}$$

The remainder of the paper will focus on geometric discounting.[5]

4.2 Subjective Probability, Actions, and Rewards

The parameter governing arm 2, θ, is fixed, but the agent does not know the value. Consequently, he will have subjective probability about the actual value of θ. Assume $\theta \in \Theta$. While it is possible to constrain θ to a smaller subset, assume $\Theta = [0,1]$. Define the subjective probability or the prior at time t to be $\pi_t(\theta)$. Note that if $P(\Theta)$ is the probability space on Θ, then $\pi_t(\theta) \in P(\Theta)$ for all

[5] The restriction $\beta \in (0,1)$ disallows trivial results. If $\beta = 0$, the agent only cares about the current period, while $\beta \geq 1$ means there is no penalty to postponing all earnings indefinitely.
The use of finite-horizon discounting yields some interesting results, since the agent does not care about anything beyond the horizon T. The existence of a finite-horizon makes it necessary to use backward induction to reach these results.

t.

At each time t, the reward depends upon the action a from a set of possible actions A. Since the action must be either choosing lever 1 or lever 2, the action set $A = \{1, 2\}$.

The reward structure is also easily derived from the structure of the problem. Clearly, $x_t \in X$, where X is the set of possible rewards, with $X = \{0, 1\}$ for all t.[6] At time t, x_t is the reward, but this reward depends upon the probabilities λ and θ, as well as the action a. Regardless of action choice, the payout occurs or it does not, so $x_t \in \{0, 1\}$ for all t. In addition, the probabilities of success and failure of action 1 are

$$p(X_t = 1|a_t = 1) = \lambda \text{ and } p(X_t = 0|a_t = 1) = 1 - \lambda \tag{8}$$

and for action 2 are

$$p(X_t = 1|a_t = 2) = \theta \text{ and } p(X_t = 0|a_t = 2) = 1 - \theta. \tag{9}$$

Straightforward calculations for the expected reward show

$$E[X_t|a_t = 1] = \lambda \tag{10}$$

and

$$E[X_t|a_t = 2] = \theta. \tag{11}$$

Depending upon the action taken and the reward received, the agent may observe a reward from arm 2, which provides additional information on the parameter θ. Using Bayes' Rule, the agent then

[6] Note that at time t, the rewards for t,$t+1$,$t+2$,...have not been observed, and are a random variable denoted $X_t, X_{t+1}, X_{t+2}, \ldots$. Previous realizations of the random variable have been observed and are denoted $x_{t-1}, x_{t-2}, \ldots$.

updates his subjective probability distribution for θ, creating the posterior $\pi_{t+1}(\theta)$. Specifically,

$$\pi_{t+1}(\theta|x_t, a_t) = \frac{\pi_t(\theta)\left[\theta^{x_t}(1-\theta)^{1-x_t}\right]}{\int_\Theta \pi_t(\theta)\left[\theta^{x_t}(1-\theta)^{1-x_t}\right]d\theta}. \tag{12}$$

The posterior for period t then becomes the prior for period $t+1$, so $\pi_{t+1}(\theta) = \pi_{t+1}(\theta|x_t, a_t)$, and the process continues.

4.3 Decision Rules

Given the setup of the problem thus far, the agent can determine a decision rule that tells him which arm to choose given his current information. The information at time t includes the prior at time t, $\pi_t(\theta)$, all past priors, and all of the past results that created potential observations regarding the true probability of θ.

At any time t, there is a t-vector of observations h_t that includes all results up to time t. Here, note that $h_t \in H_t$, where H_t is the set of all possible histories at time t. So, if $t = T$, then

$$h_T = \begin{bmatrix} x_1 & x_2 & \cdots & x_{T-1} \end{bmatrix}. \tag{13}$$

In the current example, H_T is the set of all $(T-1)$-vectors consisting of ones and zeros, or $H_t = X^t$.

The decision rule not only depends on the history of actions, but also the history of priors and the current prior. At each time t, the prior regarding the distribution of θ is $\pi_t(\theta)$. Define i_t to be the collection of all information at time t, the prior and outcome at all previous periods, and the current prior. In general, $i_t \in I_t$, the set of all information sets possible at time t, so $I_t = (P(\Theta) \times X)^t \times P(\Theta)$.

To make a decision, the agent needs to consider all information at the given point, and then make an action choice from the available options. Consequently, a decision rule is a function that maps the current information to the actions, so $\delta : I_t \to A$.

However, because of the use of Bayes' Rule, the decision framework simplifies from the entire information i_t. Because the posterior distribution $\pi_{t+1}(\theta)$ depends upon the prior $\pi_t(\theta)$ and the reward x_t, at $t = T$, the prior $\pi_T(\theta)$ reflects all elements in i_T. Consequently, for a given period t, the decision should depend solely upon the prior for that period $\pi_t(\theta)$. This discussion leads to the definition of a stationary decision rule.

Definition 37 *A stationary decision rule $\delta_t(i_t)$ is one that depends solely on the current prior $\pi_t(\theta)$ for all t. That is, $\delta_t(i_t) = \delta_t(\pi_t(\theta))$ for all t.*

In a simple, one-period framework, or in a framework where the agent knew θ, the simple decision would be to choose whichever had the highest expected value. Since the expected value of each action is given by the parameter governing that arm, the decision rule would be

$$\delta(i_t) = \begin{cases} a_t = 1, \text{ if } \lambda \geq \theta \\ a_t = 2, \text{ if } \theta > \lambda \end{cases}. \tag{14}$$

However, in the framework of the current problem, this decision rule disregards an important piece of information. Specifically, the choice $a = 2$ has additional value beyond the payout: its choice creates an observation that can help improve the agent's subjective probability about θ.

To consider the value of the additional information, the value or utility of a decision or situation must be defined.

Definition 38 *The utility of a decision rule δ, given the prior $\pi_t(\theta)$ and known probability λ, denoted $U_\delta(\pi_t(\theta), \lambda)$ is the expected discounted future rewards from the current period forward from following that decision rule. That is,*

$$U_\delta(\pi_t(\theta), \lambda) = E_\delta^{\pi_t}\left[\sum_{j=0}^{\infty} \beta^j X_{t+j}\right],$$

where $E_\delta^{\pi_t}[\cdot]$ signifies expectations given prior $\pi_t(\theta)$ and decision rule δ.

Definition 39 *The value of a situation with prior $\pi_t(\theta)$ and known probability λ, denoted $V(\pi_t(\theta), \lambda)$ is the supremum over all possible decisions δ. That is,*

$$V(\pi_t(\theta), \lambda) = \sup_{\delta} U_{\delta}(\pi_t(\theta), \lambda) = \sup_{\delta} E_{\delta}^{\pi_t}\left[\sum_{j=0}^{\infty} \beta^j X_{t+j}\right].$$

The optimal decision rule, δ^*, is a decision rule whose utility equals the situation value, meaning

$$U_{\delta^*}(\pi_t(\theta), \lambda) = V(\pi_t(\theta), \lambda). \tag{15}$$

In general, the exact decision rule depends upon the distribution of the prior $\pi_t(\theta)$, which is contained in the information set i_t. However, there are several important results that do not depend upon the specification of either.

4.4 Results

This section derives a few properties that optimal decision rules must have. In the basic framework, the value from choosing arm 1 is known, with expected value λ. The value of choosing arm 2 at time t is the expected payoff from choosing that arm, θ, plus the benefit of gaining additional information about the true value of θ. This second benefit from choosing arm 2 is very vague and potentially difficult to exactly quantify. However, the benefit is at least nonnegative, so if the expected immediate payoff from arm 2 is greater than that for arm 1, clearly it makes sense for the agent to choose arm 2. The first proposition states that the optimal decision rule must choose arm 2 if the expected value of the immediate reward is higher than if arm 1 is chosen.

Proposition 40 *If $E_{a_t=2}^{\pi_t}[X_t] \geq E_{a_t=1}^{\pi_t}[X_t]$, then $\delta^*(i_t) = 2$.*

Proof. Let δ_t^1 denote a decision rule at time t that chooses arm 1, and likewise δ_t^2 denote a decision rule that chooses arm 2. So the hypothesis states $E_{\delta_t^2}^{\pi_t}[X_t] \geq E_{\delta_t^1}^{\pi_t}[X_t]$. Let δ_{t+1}^* be an optimal decision rule at $t+1$, and assume that it is optimal regardless of the choice and result at

time t. Then

$$\begin{aligned} U_{\delta_t^2}(\pi_t(\theta),\lambda) &= E^{\pi_t}_{\delta_t^2}\left[\sum_{j=0}^{\infty}\beta^j X_{t+j}\right] = E^{\pi_t}_{\delta_t^2}[X_t] + E^{\pi_t}_{\delta^*_{t+1}}\left[\sum_{j=1}^{\infty}\beta^j X_{t+j}\right] \\ &= E^{\pi_t}_{\delta_t^2}[X_t] + E^{\pi_t}_{\delta^*_{t+1}}\left[\sum_{j=1}^{\infty}\beta^j X_{t+j}\right] \\ &\geq E^{\pi_t}_{\delta_t^1}[X_t] + E^{\pi_t}_{\delta^*_{t+1}}\left[\sum_{j=1}^{\infty}\beta^j X_{t+j}\right] \text{ (from the hypothesis)} \\ &= E^{\pi_t}_{\delta_t^1}\left[\sum_{j=0}^{\infty}\beta^j X_{t+j}\right] = U_{\delta_t^1}(\pi_t(\theta),\lambda). \end{aligned}$$

Thus, since a decision rule that chooses arm 2 is at least as good as a decision rule that chooses arm 1, then the choice of arm 2 is optimal. ■

The previous proposition discusses under what conditions it is always optimal to choose arm 2. Are there similar conditions for arm 1? Again, the fact that the benefit from additional information may be difficult to quantify – it depends upon the subjective probability distribution and the expected rewards – makes finding a rule that always holds non-trivial. However, there is a fairly trivial result that says when arm 1 is always optimal, and does so without depending upon the exact specifications of the problem.

Assume that in a given period, the agent decides to choose arm 2. Then he updates his subjective beliefs based upon the observed reward and makes another choice in the following period, which has ambiguous implications for decision rules. Now assume that in any given period, the agent optimally decides to choose arm 1. Then he receives a reward, but has gained no additional information about the unknown parameter θ, meaning that the prior for the following period, and the entire information set, is the same as for the preceding period. Since under the exact same circumstances it was optimal to choose arm 1, it must be optimal to choose arm 1 again. This intuitive result says that if arm 1 is optimal in any given period, it must be optimal to choose arm 1 for all following periods. The following Proposition demonstrates this result more rigorously.

Proposition 41 *Given some i_t, for an optimal stationary decision rule δ^*, if $\delta^*(i_t) = 1$, then $\delta^*(i_{t+1}) = 1$.*

Proof. Let δ^k be an optimal decision rule choosing arm k, and δ^{kl} be a decision rule specifying selection of arm k at time t followed by selection of l at $t+1$, regardless of the outcome at t. Since $\delta^*(i_t) = 1$, $U_{\delta_t^1}(\pi_t(\theta), \lambda) \geq U_{\delta_t^2}(\pi_t(\theta), \lambda)$, and if $a_t = 1$ then $\pi_{t+1}(\theta) = \pi_t(\theta)$. Then

$$\begin{aligned}
U_{\delta^{11}}(\pi_t(\theta), \lambda) &= E_{\delta^{11}}^{\pi_t}\left[\sum_{j=0}^{\infty}\beta^t X_{t+j}\right] = E_{\delta^1}^{\pi_t}[X_t] + E_{\delta^1}^{\pi_{t+1}}\left[\sum_{j=1}^{\infty}\beta^j X_{t+j}\right] \\
&= \lambda + E_{\delta^1}^{\pi_{t+1}}\left[\sum_{j=1}^{\infty}\beta^t X_{t+j}\right] = \lambda + \beta E_{\delta^1}^{\pi_t}\left[\sum_{j=0}^{\infty}\beta^t X_{t+j}\right] \\
&= \lambda + \beta U_{\delta_t^1}(\pi_t(\theta), \lambda) \\
&\geq \lambda + \beta U_{\delta_t^2}(\pi_t(\theta), \lambda) \\
&= \lambda + \beta E_{\delta^2}^{\pi_t}\left[\sum_{j=0}^{\infty}\beta^j X_t\right] = E_{\delta^1}^{\pi_t}[X_t] + E_{\delta^2}^{\pi_t}\left[\sum_{t=1}^{\infty}\beta^t X_t\right] \\
&= U_{\delta^{12}}(\pi_t(\theta), \lambda)
\end{aligned}$$

Therefore, any rule δ^{11} is always at least as good as δ^{12}, so if $a_t = 1$, then $a_{t+1} = 1$ must be optimal.

■

Now consider any optimal strategy. If the agent chooses arm 1 at time t, he will continue to do so in $t+1, t+2, \ldots$, as stated in Proposition 41. If the agent chooses arm 2, he does so either because the expected immediate reward is higher, or because he wants to improve his subjective beliefs by viewing more reward outcomes. If the expected immediate reward for arm 2 is greater than for arm 1, then in period t the agent believes he will choose arm 2 in period $t+1$ with the same expected reward. The same situation applies if the expected reward from arm 2 is lower than for arm 1. However, if the expected reward from arm 2 is lower than for arm 1, there is a chance that one more draw from arm 2 will sufficiently discourage the agent from choosing arm 2, making him switch to arm 1, which has higher immediate benefit. Therefore, at any given period t, the

expectation from following an optimal strategy for x_{t+1} must be at least as large as the expectation at time t for x_t. Stated more generally, because of the potential for additional information, the expected rewards from following an optimal strategy will increase. The following Proposition states these results in more concrete terms.

Proposition 42 *Given some i_t and λ, there exists some δ^* such that $E^{\pi_t}_{\delta^*}[x_{t+1}] \geq E^{\pi_t}_{\delta^*}[x_t]$.*[7]

Proof. Proposition 41 guarantees the result if $\delta^*_t(i_t) = 1$, since $\delta^*_{t+1}(i_t) = 1$ must follow, and the expected value is λ in both periods. Consequently, assume δ^2_t is the optimal decision rule that chooses arm 2 at time t. Then if success is witnessed at time t, δ^2_{t+1} is optimal, and $E^{\pi_t}_{\delta^2_t}[X_t] = E^{\pi_t}_{\delta^2_{t+1}}[x_{t+1}]$. If failure is witnessed at time t, either δ^1_{t+1} or δ^2_{t+1} is optimal. If δ^2_{t+1} is optimal after failure, then as before, $E^{\pi_t}_{\delta^2_t}[X_t] = E^{\pi_t}_{\delta^2_{t+1}}[X_{t+1}]$. Finally, consider if δ^1_{t+1} is optimal following failure. In this case, δ^* is an optimal decision rule that picks δ^2_t followed by δ^2_{t+1} if $x_t = 1$ and δ^1_{t+1} if $x_t = 0$. Then the expected value following the decision rule δ^*

$$E^{\pi_t}_{\delta^*}[X_{t+1}] = E^{\pi_t}_{\delta^2_t}[X_t] E^{\pi_t}_{\delta^2_{t+1}}[X_{t+1}] + \left(1 - E^{\pi_t}_{\delta^2_t}[X_t]\right)\lambda$$

But from Proposition 40, it follows that if the agent chooses arm 1 at $t+1$, then $\lambda \geq E^{\pi_{t+1}}[\theta]$. Then

$$E^{\pi_t}_{\delta^2_t}[X_t] E^{\pi_t}_{\delta^2_{t+1}}[X_{t+1}] + \left(1 - E^{\pi_t}_{\delta^2_t}[X_t]\right)\lambda \geq E^{\pi_t}_{\delta^2_t}[X_t] E^{\pi_t}_{\delta^2_{t+1}}[X_{t+1}] + \left(1 - E^{\pi_t}_{\delta^2_t}[X_t]\right) E^{\pi_{t+1}}[\theta | x_t = 0]$$

Next, note that, based upon expectations at t, $E^{\pi_t}_{\delta^2_t}[X_t] = E^{\pi_t}_{\delta^2_{t+1}}[X_{t+1}] = E^{\pi_t}[\theta]$. Also, from Bayes' Rule,

$$E^{\pi_{t+1}}[\theta | x_t = 0] = \left[\frac{\pi_t(\theta)(1-\theta)}{\int_0^1 \pi_t(\theta)(1-\theta)\,d\theta}\right] = \frac{E^{\pi_t}[\theta(1-\theta)]}{E^{\pi_t}[1-\theta]}.$$

[7] *Note that the expectation in both cases is conditional on $\pi_t(\theta)$, so the expectation at time t is that x_{t+1} is greater than x_t.*

Therefore,

$$\begin{aligned}
& E^{\pi_t}_{\delta^2_t}[X_t]\, E^{\pi_t}_{\delta^2_{t+1}}[X_{t+1}] + \left(1 - E^{\pi_t}_{\delta^2_t}[X_t]\right) E^{\pi_{t+1}}[\theta | x_t = 0] \\
= \; & E^{\pi_t}\left[\theta^2\right] + \left(1 - E^{\pi_t}[\theta]\right) \frac{E^{\pi_t}[\theta(1-\theta)]}{E^{\pi_t}[1-\theta]} \\
= \; & E^{\pi_t}\left[\theta^2\right] + E^{\pi_t}\left[(\theta - \theta^2)\right] \\
= \; & E^{\pi_t}\left[\theta^2\right] + E^{\pi_t}[\theta] - E^{\pi_t}\left[\theta^2\right] \\
= \; & E^{\pi_t}[\theta] = E^{\pi_t}_{\delta^*}[X_t].
\end{aligned}$$

Thus $E^{\pi_t}_{\delta^*}[X_{t+1}] \geq E^{\pi_t}_{\delta^*}[X_t]$. ■

While Proposition 42 states that expected rewards increase from following δ^*, to the agent the future reward is discounted, and each period is exactly the same, so he is indifferent between two rewards in different periods, provided their discounted values are equal. However, in certain circumstances, it is important to note that if the rewards accrue to outsiders as well, the fact that expected rewards increase means that later periods are better.

More concretely, consider the discussion at the beginning of this Chapter about the doctor and patients. Because the doctor is gaining information over the course of repeated trials of the medicine, he is essentially using the early patients as testing subjects, and perhaps even pursuing non-optimal decisions for them to gain information and help patients in the future. In the current framework, the doctor may, in acting optimally for society, give patients in earlier periods the new prescription in order to see how it works, even if he doesn't expect it to work as well. However, the doctor's choice under these circumstances is not necessarily in the best interests of the earlier patients. Proposition 42 therefore suggests that, from the doctor's viewpoint, later patients can expect better treatment, since the doctor will have information about the effectiveness of the new prescription.

These results pertaining to the basic bandit problem discuss some properties of optimal strategies, but more specific results depend upon the exact nature of the problem. In Chapter 7, I revisit this

basic framework in an application. Before the applications, I now turn to a more abstract framework that generalizes the basic problem discussed in this chapter.

Chapter 5: Generalized Bandit

While the framework discussed previously used specific beliefs, parameter spaces, reward structures, and action choices, this chapter presents a similar environment that allows for the most general spaces, and a much larger action set beyond two choices. The discussion in this chapter and Chapter 6 is based upon the work in Easley and Kiefer (1989).

5.1 General Environment

Again, at $t = 0, 1, 2, \ldots$the agent has a choice of actions whose rewards depend upon an unknown parameter θ. These actions yield a reward in every period, x_t, and the agent maximizes the expected (geometrically) discounted infinite sum of the rewards,

$$\max E_t \sum_{j=0}^{\infty} \beta^j X_{t+j}, \tag{16}$$

where, naturally, $\beta \in (0, 1)$.

The unknown parameter θ again controls the rewards, but is not necessarily the probability of success. Let $\theta \in \Theta$, where the parameter space Θ is a complete, separable metric space. Define $\mathcal{T}$ to be the σ-algebra generated by the open sets of Θ (the Borel σ-algebra). In addition, define $P(\Theta)$ to be the probability space on Θ, assume that $P(\Theta)$ has the weak topology. Let $\mathcal{F}$ be the Borel σ-algebra on $P(\Theta)$. The space $(P(\Theta), \mathcal{F})$ is a complete, separable metric space.

While the basic bandit problem and its extensions have a discrete choice set, each element of the set representing an arm of the machine, the action set now includes many other possibilities. The agent chooses action $a_t \in A$, where the action space A is a Borel subset of a complete, separable metric space. Define $\mathcal{X}$ to be the Borel σ-algebra on X.

While the reward structure in the basic example was a payoff of one under success and zero under failure, the reward structure now depends generally upon the state of nature θ and the action a_t. That is, define the conditional distribution of X_t, given a_t and θ, to be $\phi(X_t|a_t,\theta) = \eta(a_t,\theta)$. Also, assume that the reward function is continuous for all combinations of actions and states of nature, or $(a_t,\theta) \in X \times \Theta$, with respect to the measure ν on $\mathbb{R}$. Define $\mathcal{B}$ to be the Borel σ-algebra on $\mathbb{R}$. The Radon-Nikodym derivative of $\eta(a_t,\theta)$ with respect to ν defines the (unconditional) probability density function $f : X \times \Theta \times \mathbb{R} \rightarrow \mathbb{R}$ for the reward structure. That is,

$$f(X_t,\theta,a_t) = \frac{d\eta(a_t,\theta)}{d\nu}. \tag{17}$$

Consequently, given two parameter values θ and a_t, the expected reward is given by

$$E[X_t|a_t,\theta] = \int_{\mathbb{R}} X_t f(X_t,\theta,a_t)\, v(dX_t). \tag{18}$$

However, at the time of decision-making, the agent does not know the value of the parameter θ, but only has subjective beliefs about its value given by the prior $\pi_t(\theta)$. The expected reward with only knowledge of the prior is therefore the more important expected reward calculation. This expectation is given by the expectation of the reward knowing the value of θ, then considering the uncertainty about θ:

$$E[X_t|a_t,\pi_t(\theta)] = \int_{\Theta} E[X_t|a_t,\theta]\,\pi_t(d\theta) = \int_{\Theta}\int_{\mathbb{R}} X_t f(X_t,\theta,a_t)\, v(dX_t)\,\pi_t(d\theta). \tag{19}$$

Also, the rewards are bounded from above and below, meaning there exists $K \in \mathbb{R}^+$ such that $|E[X_t|a_t,\pi_t(\theta)]| \leq K$ for all possible values of $(a_t,\pi_t(\theta)) \in A \times P(\Theta)$.

In any period, the choice of action a_t and reward x_t can be used to augment the prior from that period $\pi_t(\theta)$ by using Bayes' Rule, which gives the posterior distribution for t, which is the prior for $t+1$. Since $\mathcal{F}$ is the Borel σ-algebra on the probability space $P(\Theta)$, it contains all possible sets

that θ could be a member of. Specifically, if $O \in \mathcal{F}$, then the posterior probability that $\theta \in O$ is

$$\pi_{t+1}\left(O|\pi_t\left(\theta\right), a_t, x_t\right) = \frac{\int_O f\left(X_t, \theta, a_t\right) \pi_t\left(d\theta\right)}{\int_\Theta f\left(X_t, \theta, a_t\right) \pi_t\left(d\theta\right)}, \tag{20}$$

or, if the denominator is zero,

$$\pi_{t+1}\left(O|\pi_t\left(\theta\right), a_t, x_t\right) = \pi_t\left(O\right). \tag{21}$$

Bayes' Rule takes the prior, action, and reward, and maps it to the posterior. Defining the function given by this mapping to be $B : P\left(\Theta\right) \times A \times \mathbb{R} \rightarrow P\left(\Theta\right)$. Rieder (1975) shows that this mapping is well defined and $\mathcal{F} \times \mathcal{A} \times \mathcal{B}$ measurable.

The Bayes' Rule mapping therefore takes the prior, action, and reward, and maps them to the posterior. The posterior distribution $\pi_{t+1}\left(\theta\right)$ is defined over Θ, so $\pi_{t+1}\left(\theta\right) \in P\left(\Theta\right)$. There are many such elements in $P\left(\Theta\right)$, and the exact posterior depends upon the prior and the random variables a_t and x_t. Consequently, there exists a probability space for the probability space, namely $P\left(P\left(\Theta\right)\right)$. The reward variable x_t is directly linked to the prior $\pi_t\left(\theta\right)$ and the action a_t, so it is possible to suppress consideration of that variable and determine the transition probability to an element in $P\left(P\left(\Theta\right)\right)$ given a prior and an action. Intuitively, the transition probability is a function $q : A \times P\left(\Theta\right) \rightarrow P\left(P\left(\Theta\right)\right)$ that determines the probability density function across posterior distributions, given the prior and the action. This function is

$$q\left(O|a_t, \pi_t\left(\theta\right)\right) = \int_{\Theta \times \mathbb{R}} I_O\left(B\left(\pi_t, a_t, X_t\right)\right) f\left(a_t, \theta, X_t\right) \left(\pi\left(\theta\right) \times \nu\right)\left(d\left(\theta, X_t\right)\right) \tag{22}$$

for some set $O \in \mathcal{F}$, and $I_O\left(\cdot\right)$ being the indicator function for the set O. Easley and Kiefer (1989) show that the transition probability function is $\mathcal{A} \times \mathcal{F}$ measurable.

The framework of the decision problem now depends only on the action space, the probability space, the reward structure, the transition probability, the degree of discounting, and the initial

beliefs $\pi_0(\theta)$. From the initial prior $\pi_0(\theta)$, the agent chooses an action based upon discounting the future, receives a reward, and uses the transition probability to move to a prior for the next period, which is a member of the probability space.

5.2 Decision Rules

As in the basic example, at time t, the information available to the agent is the past history of realizations, all previous priors, and the current prior. Again, let the information at time t be denoted i_t, where $i_t \in I_t = (P(\Theta) \times X)^t \times P(\Theta)$. The agent processes previous priors and past realizations of outcomes according to Bayes' Rule, meaning the current prior, $\pi_t(\theta)$, contains all relevant information at time t. As before, a stationary decision rule uses only the current prior as criteria for the decision.

A decision rule is once again a function $\delta : I_t \rightarrow A$, which maps the current information onto a decision. Because of the generality of the current problem, specifically the action space A, it is difficult to determine any other particular aspects of the decision rule. Following the notation of the basic example, however, the utility of a decision rule δ is the expected value of discounted future rewards:

$$U_\delta(\pi_t(\theta)) = E_\delta^{\pi_t}\left[\sum_{j=0}^{\infty}\beta^j X_{t+j}\right]. \tag{23}$$

In addition, following previous notation, the value of a situation is the supremum over possible decisions of all utilities:

$$V(\pi_t(\theta), \lambda) = \sup_\delta U_\delta(\pi_t(\theta)) = \sup_\delta E_\delta^{\pi_t}\left[\sum_{j=0}^{\infty}\beta^j X_{t+j}\right]. \tag{24}$$

5.3 Results

As in Section 4.4 dealing with properties of optimal decision rules for the basic bandit problem, now I turn to some results regarding decision rules for the more general framework. Because of the

nature of the basic problem, the earlier discussion took for granted the existence of optimal decision rules and instead focused on properties that optimal decision rules must have had.

A similar approach is infeasible in the current setup of the problem because of the lack of restrictions on the parameter spaces and functions. Instead, the question becomes, are there decision rules that are optimal? Further, using the three optimality definitions of Section 3.1, are there circumstances where the weaker optimality conditions hold when the stronger ones do not?

The most restricted definition discussed in Section 3.1, that of optimality, only holds under certain conditions. However, the most general of the three definitions presented, that of (p, ε)-optimality, holds under less stringent conditions.

Recall that the definition of a stationary decision rule in the current context is a decision rule that only depends upon the prior at the time of decision-making. Since the agent is assumed to process information and update his prior according to Bayes' Rule, only stationary decision rules pertain to the agent. Consequently, the existence of stationary decision rules, rather than other decision rules such as randomized rules, is the only existence question of importance.

The first Proposition of this section establishes the existence of the most general form of optimality under consideration. Specifically, the existence of (p, ε)-optimal stationary decision rules under all circumstances means that given current beliefs, there is a decision rule that the agent believes, given current beliefs, is ε-optimal.

Proposition 43 *Let $p \in P(P(\Theta))$ and let $\varepsilon > 0$. Then there exists a stationary decision rule, $\delta_t(i_t) = \delta_t(\pi_t(\theta))$, that is (p,ε)-optimal.*

Proof. Note that $(A, P(\Theta), q, X)$ form a dynamic programming problem. By Theorem 25, a (p, ε)-optimal stationary decision rule exists. ■

Now that the existence of (p, ε)-optimal decision rules under any conditions has been established, it now makes sense to consider under what circumstances can the agent do better than this basic

optimality. Two sets of conditions guarantee full optimality. The first of these has important implications for the basic bandit problem.

Proposition 44 *Let the action set A be finite. Then there exists a stationary decision rule, $\delta_t(i_t) = \delta_t(\pi_t(\theta))$, that is optimal.*

Proof. Note that $(A, P(\Theta), q, X)$ form a dynamic programming problem. By Theorem 26, an optimal stationary decision rule exists. ■

Consequently, if the agent has a finite number of actions to choose from, there will always be a stationary decision rule that satisfies the strongest optimality condition. In the basic bandit problem, the agent only has two choices. Under this setup, then, an optimal decision rule exists, and the discussion of Section 4.4 deals with properties of such a decision rule. In addition, note that extensions of the two-armed bandit to multi-armed bandits will always have existence of optimal decision rules. A simple example demonstrates the previous result.

Example 45 *First, consider the action set $A = \{a_1, a_2, \ldots, a_n\}$. Then the value function for the problem is $V(\pi_t(\theta), \lambda) = \sup\{U_1(\pi_t(\theta), \lambda), U_2(\pi_t(\theta), \lambda), \ldots, U_n(\pi_t(\theta), \lambda)\}$. Clearly the supremum can be achieved, since it is the maximum of n elements, and the optimal decision rule is the one that chooses the maximum of the utilities.*

As a counterexample, consider the action set $A = [0, 1]$ and the reward function

$$X_t = \begin{cases} a_t & \text{if } a_t \neq 1 \text{ for all } \theta \\ 0 & \text{if } a_t = 1 \text{ for all } \theta \end{cases}. \tag{25}$$

In this case, there is no optimal decision rule, since the agent will want to choose a_t as close as possible to 1 without choosing $a_t = 1$. Regardless of the choice of $a_t \in [0, 1)$, there will be an action, say $a_t + \frac{1-a_t}{2}$, that yields a greater reward. Note, however, there exist ε-optimal decision rules for all ε, in which case the ε-optimal decision rule is to choose $a_t \in (1-\varepsilon, 1)$.

While the achievement of optimality under finite decision sets deals with a large strand of the literature on bandit problems, there are many circumstances where the decision set may have either countably or uncountably infinitely many possible actions. Consider a problem similar to estimation of statistical models: the agent has data from some process depending upon θ, then in each period he estimates θ – his action, and simulates data using his estimate of θ, and earns a reward based upon the accuracy of his estimates. Under such situations, there may be infinitely many possible estimates for θ. The following Proposition deals with the cases where the action set A is infinite, but still provides for full optimality.

Proposition 46 *Let the action set A be compact, the expected reward function $X : A \times P(\Theta) \to \mathbb{R}$ be jointly continuous, and the transition probability function $q : A \times P(\Theta) \to P(P(\Theta))$ be jointly continuous. Then there exists a stationary decision rule, $\delta_t(i_t) = \delta_t(\pi_t(\theta))$, that is optimal.*

Proof. Note that $(A, P(\Theta), q, X)$ form a dynamic programming problem. By Theorem 27, an optimal stationary decision rule exists. ■

As a result of the preceding discussion, (p, ε)-optimal stationary decision rules always exist, while optimal stationary decision rules exist for some classes of decision rules. However, it remains to be seen whether the optimal decision rules lead to convergence of beliefs about the parameter θ. Learning about the parameter depends upon whether the decision rules lead the agent to generate more data. Depending upon the specifics of the problem, the goal of optimality or (p, ε)-optimality may constrain learning, that is, the agent's optimal decision may be to not learn.

Chapter 6: Convergence of Beliefs

Now that the essential aspects of the general bandit problem have been discussed, I now turn to the convergence of beliefs. That is, under what circumstances does the agent, through repeated trials and constant updating of his subjective beliefs, eventually learn the true value of the parameter θ? If the agent does learn the true value, there would be some limit prior at $t = \infty$ such that the agent puts all probability mass on the true value.

In addition to the discussion about learning, it is important to discuss optimal non-learning. In other words, times it may be optimal to not learn the true value of the parameter θ. As an example, assume that in the basic problem discussed in Chapter 4, the known parameter $\lambda = .99$, and the true value of the unknown value $\theta = .05$. At some point, the agent's subjective probability distribution may say that $P(\theta \in (0, .1)) = 1$. Then the agent knows the known parameter is much higher. Clearly the benefit of knowing that $\theta = .05$ rather than, say, $\theta = .06$, is negative, since obtaining that additional knowledge comes at the cost of passing up the action that has a much higher expected value. As a result, then, the agent may decide that the optimal action is to not pursue knowledge of the true value of the parameter.

The convergence of beliefs issue then becomes two related questions. First, do the agent's subjective beliefs converge over time to some probability distribution? Second, what are the properties of the limit beliefs if beliefs do converge? In other words, under what properties do beliefs converge to the true value of the unknown parameter θ? The goal then becomes to first show that beliefs do converge, and then show, using the three definitions of optimality discussed in Section 3.1, under what conditions those limit beliefs exactly learn the value of θ.

For beliefs to converge, the sequence of subjective probability distributions $\{\pi_t\}_{t=0}^{\infty}$ must converge almost surely to some probability distribution π_∞. To show this convergence, it suffices to show

that if $C \subset \Theta$ is any Borel-subset, then $\{\pi_t(C)\}_{t=0}^{\infty}$ converges to $\pi_\infty(C)$ almost surely.

Proposition 47 *Let $C \subset \Theta$ be an arbitrary Borel-subset. Then $\{\pi_t(C)\}_{t=0}^{\infty}$ is a martingale.*

Proof. The first step to showing that $\{\pi_t(C)\}_{t=0}^{\infty}$ is a martingale is to determine the relevant probability space. Let Z be the space of all possible rewards and actions during one period, so $Z = A \times \mathbb{R}$, let Z_∞ be the space of all possible reward and action pairs over all time, that is $Z_\infty = \times_0^\infty Z$. At a given time T, the agent has witnessed pairs of actions and rewards at $t = 0, 1, \ldots, T$, and knows that potential pairs in the form of Z exist at $t = T+1, T+2, \ldots$, so if $D \subset \times_0^T Z$ is a Borel subset, then the σ-algebra after T periods is $L_T = \left\{S \subset Z_\infty : S = D \times \left(\times_{T+1}^\infty Z\right)\right\}$. Define L_∞ to be the smallest σ-algebra such that $\cup_{t=1}^\infty L_t \subset L_\infty$. For every t, also define $L_t^* = L_t \times \{\emptyset, \Theta\}$.

Then, given a $\theta \in \Theta$ and $S \in L_t$, where $S = C \times \left(\times_{t+1}^\infty Z\right)$, the probability of A given θ is $P_\theta(A) = P_\theta\left(\left(\delta(i_k), x_k\right)_{k=1}^t \in C\right)$. The agent, knowing θ, π_0, f, q, and δ, can calculate the probability $P_\theta(A)$. Also, since f, q, and δ are Borel-measurable, so is P_θ with respect to θ. Recalling that $\mathcal{T}$ is the Borel σ-algebra on Θ, the measurable space of sample paths is $(\Theta \times Z_\infty, \sigma(\mathcal{T} \times L_\infty))$, which is the set of all pairings of the parameter with reward spaces, and the σ-algebra generated by the Borel σ-algebra on Θ and L_∞. The agent's beliefs over this space of sample paths is how the probability of $S \in L_\infty$ given θ changes with the initial prior π_0 over $C \in \mathcal{T}$; that is, $P(C \times S) = \int_C P_\theta(A)\, d\pi_0$. Denote this probability by $P_{\pi_0}(C \times S) \in P(\Theta \times Z_\infty)$. Now, the relevant distribution for almost surely convergence is therefore the initial prior π_0.

Consequently, at time t, the agent has beliefs that the parameter θ is contained in a Borel-subset $C \subset \Theta$, and this probability is $\pi_t(C) = E\left[I_{\{C \times Z_\infty\}} | L_{t-1}^*\right]$. As $t \to \infty$, meaning $\{L_t^*\}_{t=1}^\infty$ approaches L_∞^*, $E\left[I_{\{C \times Z_\infty\}} | L_t^*\right] = E\left[I_{\{C \times Z_\infty\}} | L_{t-1}^*\right]$, or $E[\pi_{t+1}(C)] = \pi_t(C)$. Thus, the sequence of beliefs for $C \subset \Theta$, $\{\pi_t(C)\}_{t=1}^\infty$, is a martingale. ■

Since the sequence of beliefs is a martingale, and since $\pi_t(C)$ must be a probability, and is therefore between zero and one, the sequence $\{\pi_t(C)\}_{t=1}^\infty$ meet the criteria of the Martingale Convergence Theorem (Theorem 36). Therefore, there must be a limit belief to which the sequence converges.

The following Proposition solidifies this idea.

Proposition 48 *There exists* $\pi_\infty \in P(\Theta)$ *such that* $\pi_t \to \pi_\infty$ *almost surely with respect to* π_0.

Proof. Let $C \subset \Theta$ be any arbitrary Borel-subset, then $\pi_t(C) = E\left[I_{\{C \times Z_\infty\}} | L_{t-1}^*\right]$. $I_{\{C \times Z_\infty\}}$ is an indicator function for the set $\{C \times Z_\infty\}$, $\pi_t(C) \in [0,1]$ almost surely for all t, so $\{\pi_t(C)\}_{t=1}^\infty$ is uniformly bounded (by ± 1), so by the Martingale Convergence Theorem (Theorem 36), $\{\pi_t(C)\}_{t=1}^\infty$ converges almost surely to some limit beliefs $\pi_\infty(C)$.

Since Θ is a complete, separable metric space, there exists a countable collection $\{C_i\}_{i=1}^\infty$ whose closure is Θ; this collection is used to determine convergence on Θ. Since the collection is countable, there exists a set of sample paths S with $P_{\pi_0}(A) = 1$ where $\lim \pi_t(C_i) = \pi_\infty(C_i)$ on S, for all i. Consequently, $\pi_t \to \pi_\infty$ almost surely with respect to P_{π_0}. Since $P(\Theta)$ is complete and $\pi_t \in P(\Theta)$ for all t, $\pi_\infty \in P(\Theta)$. ■

After verifying that the sequence of beliefs do indeed converge to some limit beliefs π_∞, it is time to consider what these beliefs may entail. The conditions on optimal decision rules will constrain the potential structure of limit beliefs. As the discussion beginning this section argued, in some circumstances it may not be optimal for beliefs about the parameter θ to converge to the actual value. In this case, the agent would want beliefs to converge enough to give him confidence that the action or actions that created endogenous data was sub-optimal relative to actions that did not provide information about θ, and then he would decide not to generate any more data, meaning that beliefs would never change, or that he had reached limit beliefs. If, on the other hand, the agent decides it is optimal (under one of the definitions of optimality) to generate data and learn the true value of θ, he would expect limit beliefs to converge so he puts probability one on the true value of θ.

For complete learning to occur, it is necessary to have some decision rule with the property that if the agent follows that decision rule, he will, in the limit, know the true value of θ.

Definition 49 *An identification decision rule* $\delta^I(i_t)$ *is a decision rule such that, given* π_0 *and* δ^I,

there exists a L_∞-measurable function g such that, for $z \in Z^t$, $g(z) = \theta$ almost surely with respect to π_0 for all $\theta \in \Theta$.

In the basic bandit problem, it is easy to see that $\delta(i_t) = 2$ is an identification decision rule, since if the agent chooses arm 2 every period, in the limit he can exactly determine θ. It is assumed that identification decision rules exist for every problem, note that they will not necessarily exist if θ changes over time. In addition, given the previous discussion on optimality and learning, it is not necessary for the identification rule to be optimal. Similarly, if the agent follows the identification rule for any problem, he will determine the true value of θ.

The following Lemma says that if the agent follows a (p,ε)-optimal decision rule, he can do so for a finite number of periods and then switch to any other decision rule and have the resulting sequence be $(p, 2\varepsilon)$-optimal.

Lemma 50 *Let $\varepsilon > 0$ and $p \in P(P(\Theta))$. Let δ' be a (p,ε)-optimal decision rule. Then there exists $T < \infty$ such that $\{\delta\}_{t=1}^{\infty} = \{\delta'_1, \delta'_2, \ldots, \delta'_T, \delta''_{T+1}, \delta''_{T+2}, \ldots\}$ is $(p, 2\varepsilon)$-optimal for any sequence of decision rules $\delta''_k : I_t \to P(A)$ for $k = T+1, T+2, \ldots$.*

Proof. Let $\{\delta'\}$ denote a sequence of δ' decision rules and $\{\delta\}$ denote a sequence of T δ' decision rules then switched to decision rule δ''. Since there exists $K \in \mathbb{R}^+$ such that $|E[X_t|a_t, \theta]| \leq K$ for all a_t and θ, then

$$\begin{aligned}
U_{\{\delta'\}}(\pi_0(\theta)) - U_{\{\delta\}}(\pi_0(\theta)) &= E^{\pi_0}_{\{\delta'\}}\left[\sum_{t=0}^{\infty}\beta^t X_t\right] - E^{\pi_0}_{\{\delta\}}\left[\sum_{t=0}^{\infty}\beta^t X_t\right] \\
&= E^{\pi_0}_{\{\delta'\}}\left[\sum_{t=0}^{T}\beta^t X_t\right] + E^{\pi_0}_{\{\delta'\}}\left[\sum_{t=T+1}^{\infty}\beta^t X_t\right] \\
&\quad - \left(E^{\pi_0}_{\{\delta'\}}\left[\sum_{t=0}^{T}\beta^t X_t\right] + E^{\pi_0}_{\{\delta''\}}\left[\sum_{t=T+1}^{\infty}\beta^t X_t\right]\right) \\
&= E^{\pi_0}_{\{\delta'\}}\left[\sum_{t=T+1}^{\infty}\beta^t X_t\right] - E^{\pi_0}_{\{\delta''\}}\left[\sum_{t=T+1}^{\infty}\beta^t X_t\right] \\
&\leq \frac{\beta^T 2K}{1-\delta} \leq \varepsilon,
\end{aligned}$$

if T is such that $\delta^T \leq \frac{\varepsilon(1-\delta)}{2K}$. ■

With this Lemma in hand, the next Proposition guarantees the existence of a (p,ε)-optimal decision rule that leads to learning the true value of θ.

Proposition 51 *Let $\varepsilon > 0$ and $p \in P(P(\Theta))$. If there exists an identification decision rule, then there exists a (p,ε)-optimal decision rule such that $\pi_t(\theta) \to \pi_\infty$ almost surely with respect to π_0, where $\pi_\infty(\theta) = 1$ at the true value of θ.*

Proof. Let δ^* be a $(p,\varepsilon/2)$-optimal decision rule, and define T as in Lemma 50. Define $z^T = \{z_t\}_{t=0}^T$ be a sequence of T observations from Z, and let π_{T+1} be the prior based upon z^T. The identification decision rule guarantees the existence of a decision rule such that $\{\pi_t\}_{t=T+1}^\infty$ converges to I_θ almost surely with respect to $P_{\pi_{T+1}}$. Lemma 50 says that following δ^* until time T and then following the identification decision rule is (p,ε)-optimal.

Define $H \subset \Theta \times Z_\infty$ to be the sequence of θ and z_t's such that $\pi_t \to I_\theta$. Then the probability of H given π_0 for a given z^T is the probability of H given π_0 and z^T multiplied the probability of z^T. Integrating over z^T gives the total probability of H given π_0. Specifically,

$$P_{\pi_0}(H) = \int_{z^T} P_{\pi_0}\left(H|z^T\right) p\left(z^T\right) dz^T.$$

But $P_{\pi_0}\left(H|z^T\right) p\left(z^T\right)$ is, by definition, $P_{\pi_{T+1}}(H)$, and integrating over all possible z^T gives

$$P_{\pi_0}(H) = \int_{z^T} P_{\pi_{T+1}}(H)\, dz^T = 1.$$

■

Note that up to this point, the results pertaining to learning have not used the concept of stationary decision rules. Unfortunately, while learning of the true value occurs given the situation discussed above, it is not necessarily a stationary decision rule that is the identification decision rule. A simple example suffices to prove that stationary decision rules may not lead to full learning.

Example 52 *Consider the case where* $\theta \in \Theta = \left[0, \frac{1}{2}\right]$, $a \in A$. *Let there be an action* a^* *that fully reveals* θ *after one period, and let the conditional probability of rewards given state* θ *and action* a_t

$$\phi\left(X_t|\theta, a_t\right) = \begin{cases} 1, & \text{for all } \theta \text{ if } a \neq a^* \\ \theta, & \text{for all } \theta \text{ if } a = a^* \end{cases}.$$

The only way to learn θ *is to choose* $a = a^*$, *which cannot be* ε*-optimal if* $\varepsilon < \frac{1}{2}$ *for any stationary policies.*

In this circumstance, and in the general setup, the agent does not care how accurate his beliefs are, he only cares about rewards, so under certain specifications of the reward function, it is very plausible that learning will not occur for stationary policies. Under the general specifications of the bandit problem discussed here, though, while beliefs will converge, the optimality concepts presented do not ensure that stationary decision rules will always lead to full information about θ. To see whether optimality encourages stationary decision rules to achieve full information, it is necessary to pin down specifics of the problem.

Chapter 7: Application: Basic Bandit Revisited

This chapter builds upon the discussion in Chapter 4 regarding the basic bandit problem. The agent approaches a two-armed machine, the first arm pays out a dollar with probability λ, the second arm pays a dollar with probability θ. The agent knows λ but not θ. Proposition 44 guarantees the existence of an optimal strategy since there are a finite number of action choices, in this case two possible actions. In addition, the discussion from Section 4.4 demonstrates that if $E^{\pi_t}[\theta] \geq \lambda$ at any time t, then arm 2 is optimal, where if $\delta^*(i_t) = 1$, then $\delta^*(i_{t+1}) = 1$.

Recall that the action space $A = \{1, 2\}$, the parameter space $\theta \in \Theta = [0, 1]$, and the reward function for arm 1 is Bernoulli(λ) and the reward function for arm 2 is Bernoulli(θ). In addition, assume the discount parameter $\beta = .9$.

Even with a simple bandit problem such as this one, computation of optimal decision rules is difficult. To provide an example that illustrates learning while allowing the problem to be solved, assume that the prior at $t = 0$ is Beta$(1, 1)$, which is uniform on $[0, 1]$. That is,

$$\pi_0(\theta) = 1. \tag{26}$$

A benefit of using the beta distribution as a prior is that the family of beta distributions is closed under Bayes' Rule when sampling from a Bernoulli distribution. Specifically, assume that $\pi_t(\theta) \sim Beta(a, b)$. Then the posterior, assuming $x_t \in \{0, 1\}$ denotes failure or success,

$$\begin{aligned} \pi_{t+1}(\theta) &= \frac{Beta(a,b)\, Bernoulli(\theta)}{\int_0^1 Beta(a,b)\, Bernoulli(\theta)\, d\theta} \\ &= \frac{\left[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right]\left[\theta^x(1-\theta)^{1-x}\right]}{\int_0^1 \left[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right]\left[\theta^x(1-\theta)^{1-x}\right] d\theta} \end{aligned}$$

Then after success ($x_t = 1$):

$$\begin{aligned}
\pi_{t+1}(\theta) &= \frac{\frac{1}{B(a,b)}\theta^a(1-\theta)^{b-1}}{\int_0^1 \frac{1}{B(a,b)}\theta^{(a+1)-1}(1-\theta)^{b-1}\,d\theta} = \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^a(1-\theta)^{b-1}}{\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{(a+1)-1}(1-\theta)^{b-1}\,d\theta} \\
&= \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^a(1-\theta)^{b-1}}{\frac{a}{(a+b)}\int_0^1 \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)}\theta^{(a+1)-1}(1-\theta)^{b-1}\,d\theta} = \frac{\frac{(a+b)\Gamma(a+b)}{a\Gamma(a)\Gamma(b)}\theta^a(1-\theta)^{b-1}}{\int_0^1 \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)}\theta^{(a+1)-1}(1-\theta)^{b-1}\,d\theta} \\
&= \frac{\frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)}\theta^{(a+1)-1}(1-\theta)^{b-1}}{1} = Beta(a+1,b)
\end{aligned}$$

and after failure ($x_t = 0$):

$$\begin{aligned}
\pi_{t+1}(\theta) &= \frac{\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b}}{\int_0^1 \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b}\,d\theta} = \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b}}{\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b}\,d\theta} \\
&= \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{(b+1)-1}}{\frac{b}{(a+b)}\int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b+1)}\theta^{a-1}(1-\theta)^{(b+1)-1}\,d\theta} = \frac{\frac{(a+b)\Gamma(a+b)}{\Gamma(a)\Gamma(b)b}\theta^{a-1}(1-\theta)^{b}}{\int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b+1)}\theta^{a-1}(1-\theta)^{(b+1)-1}\,d\theta} \\
&= \frac{\frac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b+1)}\theta^{a-1}(1-\theta)^{(b+1)-1}}{1} = Beta(a,b+1)
\end{aligned}$$

A similar argument using the Binomial distribution shows that after m successes and n failures for an agent with a $Beta(a,b)$ prior will yield a posterior that is $Beta(a+m,b+n)$. These derivations make it easy to see how the information set i_t is included in the prior at time t, as if the prior is $Beta(x,y)$, then the agent has witnessed $x-1$ successes and $y-1$ failures. In addition, the expected value of success given a $Beta(a,b)$ distribution is

$$E[X_t|Beta(a,b), a_t = 2] = \frac{a}{a+b}. \tag{27}$$

Let δ^1 be a decision rule that chooses arm 1 and δ^2 be a decision rule that chooses arm 2. Since a choice of arm 1 will be repeated indefinitely, the utility from $a_t = 1$ at t is

$$U_{\delta^1}(\pi_t(\theta), \lambda) = E_{\delta^1}^{\pi_t}\left[\sum_{j=0}^{\infty}\beta^j X_{t+j}\right] = \sum_{j=0}^{\infty}\beta^j\lambda = \frac{\lambda}{1-\beta}. \tag{28}$$

A choice of arm 2 has an expected probability of success of $\frac{a}{a+b}$, which carries a payout and then the posterior is $Beta\,(a+1,b)$. The expected probability of failure is $\frac{b}{a+b}$, which pays out nothing and the posterior is $Beta\,(a,b+1)$. Consequently, the utility from $a_t = 2$ at t is

$$\begin{aligned} U_{\delta^2}\left(Beta\,(a,b),\lambda\right) &= E_{\delta^2}^{Beta(a,b)}\left[\sum_{j=0}^{\infty}\beta^j X_{t+j}\right] \\ &= \frac{a}{a+b}\left[1+V\left(Beta\,(a+1,b),\lambda\right)\right]+\frac{b}{a+b}\left[V\left(Beta\,(a,b+1),\lambda\right)\right]. \quad (29) \end{aligned}$$

The choices $a_t = 1$ and $a_t = 2$ are the only possible choices at every time t, so the value at time t is the supremum of the utilities associated with these rules,

$$V\left(Beta\,(a,b),\lambda\right) = \max\left\{\frac{\lambda}{1-\beta}, \frac{a}{a+b}\left[1+V\left(Beta\,(a+1,b),\lambda\right)\right]+\frac{b}{a+b}\left[V\left(Beta\,(a,b+1),\lambda\right)\right]\right\}. \quad (30)$$

The optimal decision rule is the rule that maximizes equation (30). This value function can be estimated by iteration until the values for a given prior $Beta(a,b)$ converge.[8] Appendix I discusses the computation used to estimate the value function (30). The following two sections use computations to show examples of learning and optimality. In both cases, the agent has limit beliefs. The example where learning is optimal shows how beliefs converge to the true parameter, while the example where learning is not optimal shows how beliefs converge until the agent realizes learning is not optimal.

7.1 Learning is Optimal

This example will illustrate a case where learning is optimal. Since the initial prior is set to be $\pi_0(\theta) \sim Beta\,(1,1)$, $E\left[X_0|a_t = 2, Beta\,(1,1)\right] = .5$. By Proposition 40, the nontrivial cases occur when $\lambda > .5$, otherwise $a_t = 2$ is obviously the optimal decision rule. Consequently, set $\theta = .8$ and $\lambda = .7$, so the initial expectation is below λ, but the true value of θ is greater than λ. In a myopic

[8] The value function satisfies necessary criteria for dynamic programming. See Gittins (1989) for a futher discussion.

strategy (where $\beta = 0$), the individual would always choose $a_t = 1$, but it is clearly the case that the best strategy would be to learn the true value of θ.

Table 1

t	$\pi_t(\theta)$	$E[X_t \mid \pi_t(\theta), \lambda]$	$P(\pi_t(\theta) \in (\theta - .02, \theta + .02))$	$U_{\delta^2}(\pi_t(\theta), \lambda)$
0	$Beta(1,1)$	.5	.040	7.01
10	$Beta(10,2)$	.833	.118	8.35
50	$Beta(41,11)$	.788	.279	7.88
100	$Beta(86,16)$	.843	.199	8.43
500	$Beta(402,100)$	.801	.737	8.01

Figure 1: Evolution of Priors, Learning Optimal

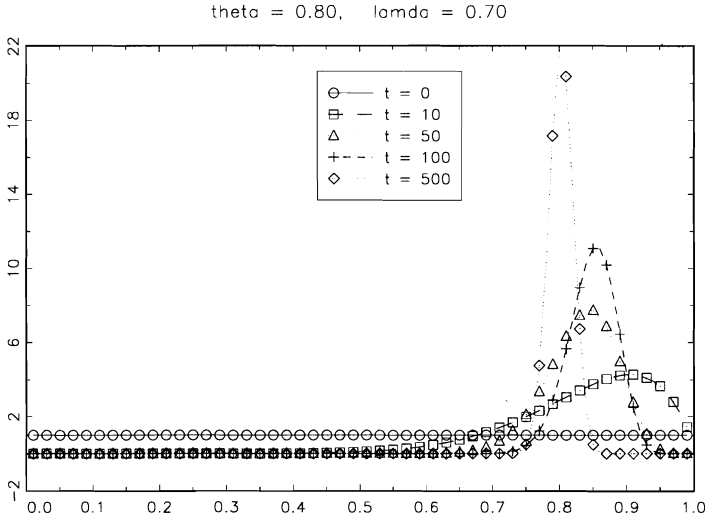

Table 1 shows the prior, expected value, accuracy of the prior, and the utility for δ^2 at various times. Since the utility for $a_t = 1$ is $\frac{\lambda}{1-\beta} = 7.0$, as long as $U_{\delta^2}(\pi_t(\theta), \lambda) \geq 7$, the agent will choose $a_t = 2$ and eventually learn the true value of θ. Also, over time, the value of information decreases,

and the entire utility is dominated by the expected value of the stream of payouts.

Figure 1 shows the evolution of the priors over time. Note that they become increasingly accurate, as the agent's optimal policy is to choose $a_t = 2$.

One disadvantage of having a uniform distribution for a prior is that it is easily shaped by early observations, making the agent susceptible to "discouragement" early in the process. For example, in the current example, since a failure at $t = 0$ after choosing arm 2 would lead to a $Beta(1,2)$ prior at $t = 1$, which has expected value $E[x_t|Beta(1,2)] = \frac{1}{3}$, the agent would immediately become discouraged and choose arm 1 after that point.

7.2 Learning is not Optimal

This example will illustrate a case where learning is not optimal. Since the initial prior is set to be $\pi_0(\theta) \sim Beta(1,1)$, $E[X_0|a_t = 2, Beta(1,1)] = .5$. As opposed to the previous section, the interesting case occurs when learning occurs for some time, then the agent realizes that it is not optimal to learn. Consequently, set $\theta = .1$ and $\lambda = .11$, so the initial expectation is above λ, but the true value of θ is less than λ. The individual would start to choose $a_t = 2$, then beliefs should stop converging when the agent realizes he should not pursue learning the true value of θ.

Table 2 shows the prior, expected value, accuracy of the prior, and the utility for δ^2 at various times. Since the utility for $a_t = 1$ is $\frac{\lambda}{1-\beta} = 1.1$, as long as $U_{\delta^2}(\pi_t(\theta), \lambda) \geq 1.1$, the agent will choose $a_t = 2$. As soon as $U_{\delta^2}(\pi_t(\theta), \lambda) < 1.1$, the agent will stop learning and terminate the problem, choosing $a_t = 1$ for the rest of the time.

Figure 2: Evolution of Priors, Learning Not Optimal

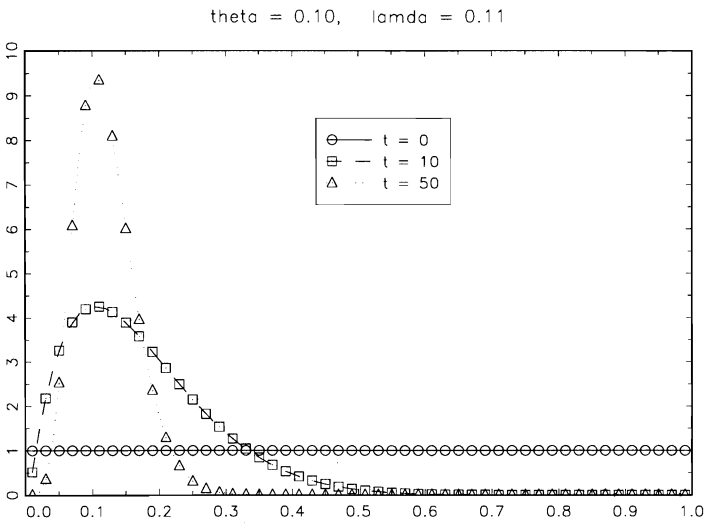

Table 2

t	$\pi_t(\theta)$	$E[X_t \mid \pi_t(\theta), \lambda]$	$P(\pi_t(\theta) \in (\theta - .02, \theta + .02))$	$U_{\delta^2}(\pi_t(\theta), \lambda)$
0	$Beta(1,1)$	.50	.04	5.01
10	$Beta(2,10)$	.167	.17	1.69
50	$Beta(6,46)$	.115	.36	1.17
57	$Beta(6,53)$	.102	.39	1.10

Figure 2 shows the evolution of the priors over time. Note that they become increasingly accurate, but the agent decides to stop learning at $t = 58$.

Unlike the previous section, where the agent can become "discouraged" and stop trying to learn when it is good to learn, the quicker the agent becomes "discouraged," the greater his overall return.

Chapter 8: Conclusion

This paper has discussed learning and endogenous data in sequential analysis or bandit problems. In both the basic and the more general bandit problems, optimal decision rules exist under certain circumstances, and the requirement of optimality may or may not allow complete learning. The sequence of subjective probability beliefs is a martingale, and consequently this sequence converges to limit beliefs. These limit beliefs, however, do not necessarily reflect complete learning, as the decision-maker's desire to maximize rewards may lead to the abandonment of the learning process. A simple example using the basic bandit framework shows that only slight manipulation of the environment can lead to learning being either optimal or not.

References

Ash, Robert B. 1972. *Real Analysis and Probability.* Orlando: Academic Press.

Banks, Jeffrey S., and Rangarajan K. Sundaram. 1992. "Denumerable-Armed Bandits." *Econometrica* Vol. 60 No. 5.

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag.

Berry, Donald, and Bert Fristedt. 1985. *Bandit Problems: Sequential Allocation of Experiments.* New York: Chapman and Hall.

Billingsley, Patrick. 1986. *Probability and Measure.* Second Edition. New York: John Wiley and Sons.

Blackwell, David. 1965. "Discounted Dynamic Programming." *Annals of Mathematical Statistics* Vol. 36 No. 1.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference.* Second Edition. Pacific Grove, CA: Duxbury.

Cyert, Richard M., and Morris H. DeGroot. 1987. *Bayesian Analysis and Uncertainty in Economic Theory.* Totowa, NJ: Rowman and Littlefield.

DeGroot, Morris H. 1970. *Optimal Statistical Decisions.* New York: McGraw-Hill.

Doob, J.L. 1953. *Stochastic Processes.* New York: John Wiley and Sons.

Easley, David, and Nicholas M. Kiefer. 1988. "Controlling a Stochastic Process with Unknown Parameters." *Econometrica* Vol. 56 No. 5.

Easley, David, and Nicholas M. Kiefer. 1989. "Optimal Learning with Endogenous Data." *International Economic Review* Vol. 30 No. 4.

Gittins, J.C. 1989. *Multi-Armed Bandit Allocation Indices.* New York: John Wiley and Sons.

Gittins, J.C., and D.M. Jones. 1974. "A Dynamic Allocation Index for the Sequential Design of Experiments." *Progress in Statistics.* Edited by J. Gani, et. al. Amsterdam: North-Holland.

Karlin, Samuel, and Howard M. Taylor. 1975. *A First Course in Stochastic Processes.* Second Edition. New York: Academic Press.

Maitra, Ashok. 1968. "Discounted Dynamic Programming in Compact Metric Spaces." *Sankhya A* Vol. 30 No. 1.

Martin, J.J. 1967. *Bayesian Decision Problems and Markov Chains.* Publications in Operations Research No. 13. New York: John Wiley and Sons.

McDonald, John N., and Neil A. Weiss. 1999. *A Course in Real Analysis.* New York: Academic Press.

Presman, E.L., and I.N. Sonin. 1990. *Sequential Control with Incomplete Information: The Bayesian Approach to Multi-Armed Bandit Problems.* Translated by: E.A. Medova-Dempster and M.A.H. Dempster. New York: Academic Press.

Rieder, Ulrich. 1975. "Bayesian Dynamic Programming." *Advanced Applied Probability* Vol. 7 No. 1.

Stokey, Nancy L., Robert E. Lucas, and Edward C. Prescott. 1989. *Recursive Methods in Economic Dynamics.* Cambridge: Harvard UP.

Strauch, Ralph E. 1966. "Negative Dynamic Programming." *Annals of Mathematical Statistics* Vol. 37 No. 4.

Sundaram, Rangarajan K. 2003. "Generalized Bandit Problems." Working Paper, Stern School of Business. Forthcoming in *Social and Strategic Behavior: Essays in Honor of Jeffrey S. Banks.*

Wahrenberger, David L., et. al. 1977. "Bayesian Rules for the Two-Armed Bandit Problem." *Biometrika* Vol. 64 No. 1.

Appendix A: Computation

This appendix describes the computation of the value function given in equation (30). Gittins (1989) notes that the value function can be estimated by recursion. See Stokey, Lucas, and Prescott (1989) for a discussion of dynamic programming methods. The steps for value function iteration are as follows:

1. Set the parameters λ, θ, β.

2. Create a $(n+1) \times (n+1)$ matrix N, initialized at any aribitrary value. The (i,j)-element of N, denoted N_{ij}, corresponds to $V(Beta(i,j),\lambda)$. Given that the intial prior is $Beta(1,1)$, the $(1,1)$-element, N_{11}, is the value at $t=0$.

3. Since the value function is

$$\begin{aligned} & V(Beta(a,b),\lambda) \\ = \; & \max\left\{\frac{\lambda}{1-\beta}, \frac{a}{a+b}\left[1+V(Beta(a+1,b),\lambda)\right]+\frac{b}{a+b}\left[V(Beta(a,b+1),\lambda)\right]\right\}, \end{aligned} \tag{31}$$

iteration loops over $i=1,\ldots,n$ and $j=1,\ldots,n$. Specifically,

$$N_{ij}^{k}=\max\left\{\frac{\lambda}{1-\beta}, \frac{i}{i+j}\left(1+N_{i+1,j}^{k}\right)+\frac{i}{i+j}N_{i,j+1}^{k}\right\}, \tag{32}$$

where k denotes the k^{th} iteration.

4. After looping over $i=1,\ldots,n$ and $j=1,\ldots,n$, the new matrix can be denoted N^1, with typical element N_{ij}^{1}. Then repeat Step 3 using N^1 to get N^2, use N^2 to get N^3, etc.

5. Iteration on N is complete when the matrix after k iterations is "close" to the matrix after $k+1$ iterations, where "close" is defined as when

$$\max_{i,j} \left| N_{ij}^{k+1} - N_{ij}^{k} \right| < 10^{-10}. \tag{33}$$

When the convergence criterion is met, the value function matrix N^* is the value function for each a and b in $V\left(Beta\left(a,b\right),\lambda\right)$. Note that in step 2, the value chosen for initialization of N will remain in row and column $n+1$, since these are never updated. The rows and columns near $n+1$ will be distorted by the chosen value. Two things can be done to remedy this distortion: (1) choosing n extremely large so the distortions only affect a small aspect of the matrix, for example, choosing $n = 1000$ when analysis will be done only on $i,j \leq 600$; (2) repeating the entire iteration using a different initialization for N and ensuring that N^* is the same for both for $i,j \leq 600$.

6. Using a $Bernoulli\left(\theta\right)$ distribution, simulate a sequence of some large number, say 600, successes $(x=1)$ or failures $(x=0)$. Put these in a vector U, so $U\left[t\right]$ is a success or failure if arm 2 is chosen at time t.

7. Run the experiment. Starting at N_{11}^* in $t=1$, see if $V\left(Beta\left(i,j\right),\lambda\right) = N_{ij}^* > \frac{\lambda}{1-\beta}$. If yes, the agent chooses arm 2, so sample from U. If $U\left[t\right]=1$, the agent moves to $V\left(Beta\left(i+1,j\right),\lambda\right)$ at $t=2$. If $U\left[t\right]=0$, the agent moves to $V\left(Beta\left(i,j+1\right),\lambda\right)$. If $N_{ij}^* = \frac{\lambda}{1-\beta}$, the agent ceases the experiment and always chooses arm 1. Keeping track of i and j allows analyzing the evolution of the prior.

VITA

Andrew T. Foerster was born May 19, 1982 in Rochester, NY and is an American citizen. He graduated from Davidson College with a Bachelor of Arts in Economics in 2004. Currently an Assistant Economist at the Federal Reserve Bank of Richmond, he will pursue a Ph.D. in Economics from Duke University starting in the fall of 2006.