A BANDIT MODEL
FOR
CONSUMPTION-INVESTMENT DECISIONS

BY

YAN WANG

A Thesis Submitted to the Faculty of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Statistics

University of Manitoba

Winnipeg, Manitoba

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

A BANDIT MODEL FOR CONSUMPTION-INVESTMENT DECISIONS

BY

YAN WANG

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

MASTER OF SCIENCE

Contents

Abstract

In this thesis, we formulate two kinds of bandit models for the investment-consumption problem. Two risky assets with uncertainty in the returns are represented by the two arms of the bandit model. Both arms follow independent Poisson distributions with intensity rates λ and μ respectively. For the first kind of bandit model, we take a fixed proportion of the total wealth to consume at each time and reinvest the remaining portion. For the second kind of bandit model, we invest a fixed amount at each time and consume all wealth after the investment. The objective is to maximize the total expected discounted consumption.

Under the Bayesian approach, the unknown λ is assumed to follow a gamma prior distribution, and its posterior distribution is updated in a discrete setting with the availability of new complete information. A key issue in making investment-consumption decisions is to balance between information gathering about the uncertainty in the returns and immediate payoff for maximizing consumption.

The value function and the advantage function of the unknown arm over the known arm are introduced and properties are examined and described in detail. An index value is introduced which acts as the benchmark to measure the performance of both arms. The myopic strategy and the play-the-winner strategy are derived and shown to be optimal in different cases of this model.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my thesis advisor, Dr. Xikui Wang for his wisdom, patience, dedication, support, countless hours of work and discussion with me throughout the development of this thesis. Without his helpful comments and motivation, this thesis would not have been possible. I am also very grateful to the faculty and staff of the Department of Statistics for their teaching and support throughout the last two years of graduate studies. I would also like to thank committee members Dr. Dean Slonowsky of the Department of Statistics, and Dr. Kiril Kopotun of the Department of Mathematics.

I wish to thank Dr. Xikui Wang for his generous financial support from his research grants. I thank the Faculty of Science for the Studentship, the Faculty of Graduate Studies for the UMGF, and UMSU for the UMSU Scholarship, whose financial support has greatly contributed to my studies.

To my family and friends, thank you for your love and care. Your support is appreciated more than you know.

Chapter 1

Introduction

1.1 Modern Portfolio Theory

Over the past 20 years, portfolio management has evolved enormously. The basic investment-choice problem for an individual is to determine the optimal allocation of his or her wealth among the available investment opportunities. The theory for solving the general problem of choosing the best investment combination is called the **portfolio-selection theory**.

The seminal paper Markowitz (1952) developed a theory for the portfolio choice in an uncertain environment. Markowitz was the first to quantify the difference between the risk of individual portfolio assets and the overall risk of the portfolio. He considered the portfolio as a whole, whereas previous studies had been focused on securities on an individual basis. Markowitz's work founded the modern portfolio theory.

Portfolio theory assumes that for a given level of risk, investors prefer

higher returns to lower returns. Similarly, for a given level of expected return, investors prefer less risk to more risk. The expected return of the portfolio is measured by its mean return. The risk corresponds to the uncertainty of obtaining the return and is measured in terms of the variance or standard deviation of the returns. The measure is under the assumption that returns are distributed normally.

The theory developed by Markowitz is based on maximizing the utility of the investor's terminal wealth. This utility function is defined according to the expected return and the standard deviation of the wealth. Using linear regression, the solution to the optimization problem of choosing the optimal portfolio can be offered to a risk-averse investor: the optimal portfolios are defined as those having the largest mean returns, subject to keeping the risks (i.e., the variances) below a specified acceptable threshold, or as those giving the lowest level of risk for each level of expected return. The complete set of these portfolios forms the efficient frontier, and the selections are limited to this efficient frontier, which constitutes the convex envelope of all the portfolios that can be obtained.

Markowitz's approach is described as a mean-variance approach because only two parameters, the mean return and the return variance, are taken into account. That is, only the first two moments of their distribution are used to characterize the investor's portfolio. The constraint is that either the returns are assumed to be normally distributed or the investor's utility function is quadratic. Markowitz also proposed semi-variance as a good measure of risk, but finally chose variance due to practical implementation

reasons.

1.1.1 Transaction Cost on Portfolio Selection

When determining the optimal portfolio, the impact of transaction costs is not taken into account in the mathematical model developed by Markowitz, although these costs have a significant impact on portfolio performance. But transaction costs are difficult to estimate because they are not fixed, and the exact values of the costs can't be possibly obtained until the security trade has taken place. The costs include the commission, which is the tax per security paid to execute the transaction; the bid/ask spread, which is the differential between the requested price and the offered price; and the liquidity or market impact cost, which is the additional cost of trading several securities compared with the cost of trading a single security. To optimize the portfolio while taking transaction costs into account, the amounts of transaction costs are introduced into the utility function. The higher the chosen value of transaction costs, the less the portfolio will evolve.

1.1.2 The Capital Asset Pricing Model (CAPM)

Initiated by Sharp (1964) and Lintner (1965), the CAPM has played an important role in finance and has been a focus point in the empirical finance literature. The CAMP is a single-period specialization of the fundamental valuation equation. The contribution of the model is that it relates the ex-

pected excess returns to the market portfolio return. However this feature of the CAMP is criticized by Fama (1976), Roll (1977), and others, who point out that the model is testable only if the market portfolio return is observable. The only empirically testable implication of the CAPM is that the market portfolio is mean-variance efficient. Then the arbitrage pricing theory (APT) is put forth by Cox and Ross (1976, 1977) to address the criticism on the observability of the market portfolio return levelled against the CAPM.

Let's discuss the CAPM in detail. Merton (1969) regards the behavior of a single agent acting as a market price-taker and seeks to maximize the expected utility of consumption. The utility function of the agent is assumed to be a power function, and the market is assumed to comprise a risk-free asset with a constant rate of return and several stocks, each with constant mean rate of return and volatility. With only information of current prices, with infinitely divisible assets, and without transaction costs, Merton was able to derive a closed-form solution to the stochastic control problem faced by the agent. Later on, by assuming nonconstant market coefficients which depend on a "state" variable, Merton addressed the issue of price formation with necessary conditions for equilibrium prices. However he didn't resolve the question of the existence of a solution to these conditions.

Based on Merton's model, several directions were generalized. The restriction to utility functions of a power form was removed in Karatzas, Lehoczky, Sethi and Shreve (1986). Market coefficients depending in an

adapted way on an underlying Brownian motion were treated in Cox and Huang (1989), Karatzas, Lehoczky,and Shreve (1987) and Pliska (1986).

Later, in an important breakthrough, Cox and Huang (1989a) and Karatzas, Lehoczky, and Shreve (1986, 1987, 1990) showed that the martingale representation theory can be applied to reduce the stochastic dynamic programming problem to a static problem in complete markets.

An important innovation of the model by Liu (1998) is that the stock returns exhibit stochastic volatility or predictability and he is able to consider incomplete markets explicitly. Wachter (1999) used martingale methods to characterize the consumption and portfolio strategies in complete markets when stock returns are predictable. Chacko and Viceira (1999) developed portfolio and consumption rules under an incomplete market setting with stochastic volatility. They relied on an approximation scheme to solve the Bellman equation in their general applications. Kogan and Uppal (1999) provided approximation methods for solving consumption and portfolio problems in a continuous-time setting.

Based on CAPM, Sharp, Lintner and Mossin (SLM) have developed mean-variance equilibrium capital asset pricing under uncertainty. The model shows that there is a linear relationship between the equilibrium expected return on an asset and its systematic risk which is measured by the covariance between the asset's return and the return on market portfolio. The SLM capital asset pricing model and many of its extended versions are formulated in nominal terms under the constraint of assuming implicitly that there are no price level changes.

1.1.3 Continuous-time method

Continuous-time methods have proved to be the most attractive way to conduct research and gain economic intuition in certain core areas in finance (such as, asset pricing, derivatives valuation, and portfolio selection). These methods can be traced backed to the seminal contributions of Merton (1969, 1971, 1973b) in the late 1969s and early 1970s. Merton (1969) initiated the study of financial markets using continuous-time stochastic models. He examined the continuous-time consumption-portfolio problem for an individual whose income is generated by capital gains on investments in assets under the "Geometric Brownian motion" hypothesis, which implies that asset prices are log-normally distributed with temporally constant parameters. He derived explicit solutions for the above problem under the additional assumption of a constant relative or constant absolute risk aversion utility function.

The estimation strategies used in continuous-time models can be categorized into the following areas:

1. Maximum likelihood method;
2. Generalized method of moments (GMM);
3. Simulated method of moments (SMM);
4. Efficient method of moments (EMM);
5. Nonparametric approaches;
6. Methods based on empirical characteristic functions;
7. Bayesian methods.

Later, the seminal contributions on options pricing by Black and Scholes (1973) and Merton (1973a) was made, which provided the first truly satisfactory model for pricing options on equity. These contributions made a strong impact in this field during the period from 1969 through 1980, and changed the way in which the practitioners viewed the finance research.

There are two key ingredients to pricing and hedging in the Black-Sholes framework. The first one is that the discounted asset prices are martingales by changing the probability measure, and the second is that the pricing formula is in the form of the discounted expected value of a claim.

The original Black-Scholes formula only applies to European call and put options. Probably the Black-Scholes formula is the most famous model used in the theory of option pricing. However the equity options pricing formulation was criticized by a number of scholars, and some formulas that have varying resemblances to the Black-Scholes model were introduced. These formulations depend on subjective discount rates or risk aversion parameters and are not fully supported by an arbitrage-free argument.

1.2 Bandit Processes

In the present thesis, We use a bandit model to formulate the optimal investment-consumption problem. Let's introduce bandit processes briefly.

Bandit processes study optimal sequential selections from several populations or stochastic processes (or arms) with unknown characteristics. The

objective is to choose an optimal strategy for making selections among the arms in order to maximize the total expected discounted reward from all selections. Authors making early significant contributions on bandit problems include Thompson (1933, 1935), Robbins (1952, 1956), Bradt, Johnson, and Karlin (1956), Bellman (1956), Feldman (1962), Gittins and Jones (1974), Rodman (1978), Bather (1981), and Berry and Fristedt (1985). Most papers in the bandit literature are applied to clinical trials with the assumption that complete information of past history is known or is observed before the next patient. This kind of model is called the bandit model with immediate responses.

The basic bandit model and many extensions are based on the following essential components: the decision times (continuous or discrete), the number of arms, the types of the arms (a variety of kinds of population distributions or stochastic processes), the discount sequence, the availability of historic information, and different estimation methods (minimax, parametric, non-parametric, Bayesian, etc.). For a detailed introduction, refer to Wang (2001).

Bandit processes concern the trade-off between high immediate expected payoff and information gathering. The benefit of information gathering is not immediate but potentially advantageous in that the uncertainty about the unknown arms (or populations) is reduced and better informed decisions with higher payoffs in the future are expected. There are two typical kinds of strategies, one is of the complete randomization among arms (which only gathers information but ignores immediate payoff), the other

is of the myopic strategy (which always selects the arm with the highest immediate payoff at each stage, but ignores information gathering). Generally speaking, both strategies are not optimal. The main goal of bandit processes is to combine the two typical strategies to obtain the overall best performance throughout the horizon of successive decision-making stages. It is a trade-off that makes bandit problems attractive in theory, useful in applications (in clinical trials generally), but difficult in obtaining explicit solutions.

For the traditional bandit problem, it is assumed that the state of the arm will change only if the arm is selected for observation. This means that we freeze the arms that are not selected for observation. But in financial markets, if each asset is regarded as an arm, then we can no longer use this traditional bandit model, because the market information on those unselected assets still evolves over time. This implies that the state of an arm changes over time no matter whether or not the arm is selected for observation. Such a bandit process is called a restless bandit model. Mathematically speaking, restless bandit models are more complicated to formulate and more difficult to solve.

Define a strategy $\pi = (\pi_1, \pi_2, \cdots)$ as a sequence of rules such that at each time $n = 1, 2, \cdots$, the population to be selected (or the arm to be pulled) is specified by π_n based on the (partial or entire) history of previous selections and observations. Denote i_n as the population selected at time n under the strategy π and Z_n as the corresponding response. Then $\sum_{n=1}^{\infty} \alpha_n Z_n$ is a utility for the sequence $(i_1, i_2, \cdots)$ resulting from the

strategy π. Define the value of the strategy π as

$$W(\pi) = E_\pi(\sum_{n=1}^{\infty} \alpha_n Z_n),$$

where$\alpha = (\alpha_1, \alpha_2, \cdots)$ is the discount sequence, $\alpha_n \geq 0$ and $\sum_{n=1}^{\infty} \alpha_n < \infty$. $\sup_\pi W(\pi)$ is called the value of the bandit problem.

Three major approaches are frequently used in the bandit literature:

(1): the minimax approach, with which important contributions include Vogel (1960a, 1960b), Fabius and van Zwet (1970), Bather (1983), Berry and Fristedt (1985), Reimnitz (1986), and Kulkarni and Lugosi (2000);

(2): the utility comparison approach, which was initiated by Robbins (1952). Strategies are restricted to be Markovian. A further contribution is given in Robbins (1956) for the class of strategies with a finite memory.

(3): the Bayesian approach, which specifies a prior distribution, then updates it to a posterior distribution with new available observations. This approach takes advantage of all the available information for decision-making and updates the states of the bandit at the decision stages. This approach will be adopted in the present thesis and will be described in the following Chapters in detail.

1.2.1 Bandit problems and Markov decision processes

Following the Bayesian approach, the basic discrete time bandit model with immediate responses can be formulated as a Markov decision process. On the other hand, bandit processes with delayed responses can only be formulated as a general controlled stochastic processes (see Wang and Bickis

(2003) for details).

Consider a bandit with an unknown parameter θ, then the state space S consists of all distributions over the parameter space Θ. The action space is $Z = \{1, 2, \cdots, k\}$, where k is the number of arms (or distributions). With the known current state s_n (which is described by the current prior) at stage n, if action i_n is taken (arm i_n is selected), an immediate expected reward is calculated as follows:

$$r(s_n, i_n) = \int_\Theta \int_{-\infty}^{\infty} x dF_{i_n}(x) ds_n(\theta),$$

where $F_{i_n}(x)$ is the conditional distribution function of X. The state transition law is determined by the Bayes's formula.

Following the theory of Markov decision processes with a Borel state space, Bickis and Wang (2004) prove the existence of an optimal strategy given that the posterior distribution is updated continuously by both the current observation and the prior distribution. In the spirit of controlled stochastic processes (Gihman and Skorohod 1979), Bickis and Wang (2004) further show the existence of an optimal deterministic strategy if the posterior distribution is updated continuously by all past observations and the current state. Bandit processes are defined as semi-Markov decision processes in Gittins (1989) and as vector-valued Markov decision processes in Glazebrook (1991, 1993).

In a multi-armed Bernoulli bandit problem, all arms yield a payoff of 0 or 1 when selected, but with different payoff probabilities. As the number of observations increases, the decision maker gets an increasingly more

accurate estimate of each arm's payoff probability. The decision maker's goal is to determine a sequence of arms so as to maximize the expected sum of discounted rewards. The multi-armed bandit problem is easily formulated as a dynamic allocation problem, and the aim is to allocate a limited amount effort to a number of independent projects, each generating a specific stochastic reward proportional to the effort spent on it.

There are several strategies that can be carried out practically to obtain the performance of the whole bandit processes, such as the myopic strategy and the play-the-winner strategy. The myopic strategy says that the arm with the highest immediate expected payoff is selected at each stage. The deterministic play-the-winner rule was initially proposed in Zelen (1969). Assuming a bandit process with two unknown and independent Bernoulli arms, the first selection is made randomly between the arms. After that, if the result is a success, then the second selection is made on the same arm. Otherwise, the other arm is selected. Zelen (1969) and Wang and Pullman (2001) prove that using this strategy, the proportion of selections of the superior arm is maximized in long run. Samaranayake (1992) studies the randomized play-the-winner rule for dependent arms, and Bandyopadhyay and Biswas (2002) make a contribution in the use of randomized play-the-winner rule with delayed responses.

1.2.2 The Gittins index strategy for bandit models

Gittins and Jones (1974), and Gittins (1979) propose the Gittins index strategy, which is one of the most significant contributions to bandit processes. The Gittins index strategy can offer a nice complete solution to the geometrically discounted bandit processes with k independent unknown arms. The theory of Gittins indices is followed by many papers concerning such dynamic allocation problems.

The crucial idea of Gittins (1979) is concerned with reducing the multi-dimensional optimization problem to a family of simpler benchmark problems. Hence the Gittins index strategy is to define a dynamic performance measure separately for each of the arms in such a way that an optimal choice can be determined by an index-rule. The dynamic performance measure is called the Gittins index, and the unknown arm with the current highest Gittins index is shown to be optimal.

The Gittins index is based on two components: the "immediate reward" component and the "learning" (or information gathering) component. Its process can be viewed as the solution to a representation problem, and its intrinsic mathematical interest and its unifying role is easily applied to a number of different applications, including the areas of economics and finance.

Gittins (1979) and Whittle (1980) consider a discrete-time Markovian setting, Karatzas (1984) and Mandelbaum (1987) extend the analysis to diffusion models. El Karoui and Karatzas (1994) develop a general martin-

gale approach in continuous time. One of their results is that the Gittins indices can be viewed as solutions to a representation problem.

1.2.3 The use of bandit models in economics and finance

There are very few papers that deal with the economic or financial applications of bandit models. Bank and Föllmer (2002) seems to be the only paper in the mathematical finance area that develops continuous-time multi-armed bandit models. Coming from the microeconomic theory of intertemporal consumption choice, the singular control problem is reduced to a stochastic representation problem in Bank and Föllmer (2002). Following the stochastic representation approach, existence and uniqueness of a solution is easily demonstrated by backwards induction in discrete time.

In the case of option pricing, the methods of bandit models are also adopted, such as the optimal stopping method. The usual approach to option pricing and to the construction of replicating strategies is combined with an optimal stopping problem in Karatzas (1988).

However, bandit models have been used in economics for several types of problems. For example, Jovanovich (1979), Miller (1984) and Mortensen (1985) use bandit models to analyze job-search problems in labor markets. Rothschild (1974) and Schmalensee (1975) use bandit models to derive dynamic pricing problems in the face of unknown demand functions. They assume that the demand is a function of the unknown probability of pur-

chasing the product. Recently Wang (2004) extends the results to a bandit model where the unknown demand function depends on both an unknown Poisson rate of buyers and the unknown probability of purchase.

1.3 About My Thesis

In this thesis, we introduce two new bandit models to study investment-consumption problems.

We formulate two kinds of bandit models for this problem. The two arms in each bandit model denote two risky assets with uncertainty in the returns. Both arms follow independent Poisson distributions, with intensity λ and μ respectively. The parameter λ is assumed to be unknown, hence the arm with intensity rate λ is named the unknown arm. The parameterμ is known, hence the arm with μ is called the known arm.

For the first kind of bandit model, at every point of time, we take a fixed proportion of the total wealth to consume, and re-invest the remaining portion. The objective is to maximize the total expected discounted consumption. For the second kind of bandit model, we invest a fixed amount at every time and consume the remaining wealth after the investment.

We wish to find the index value for the model such that at this index value both arms are equivalent. This means that if μ is the same as this index value, there is no difference between investing in this unknown arm and investing in the known arm.

We show that the investment decision rule based on the advantage

function can be simplified to be based on a break-even index value of the parameter μ. We also demonstrate some properties of the index value.

The main results in this thesis are concerned with the value of the bandit models and the total expected discounted reward obtained sequentially from the consumption at every single time, the dynamic programming solution for two risky assets with different returns, which are evaluated using the index values, and the structure of the optimal strategy for a particular class of gamma prior distributions $G(\eta_0, \tau_0)$. Chapter 2 describes the formulation of the portfolio selection problem with constant proportion of consumption in two-armed bandit processes. The uncertainty inherent in each asset is reflected by the number of payoffs during the corresponding investment period.

Chapter 3 provides the solution to the investment-consumption problem and discusses the properties of the index value, which is a useful measure to evaluate the two risky assets (two arms) and is a unique solution to the advantage equation. Chapter 4 looks at an alternative model with the assumption that the amount of investment is fixed every time, and the remaining amount of the wealth after the previous investment is taken out for consumption. Similar results are obtained, but myopic strategy is proved to be not optimal in this case. An optimal stopping solution and a version of the play-the-winner strategy are derived. Then the final chapter concludes the thesis with a discussion on possible future research directions.

Chapter 2

A bandit model for optimal portfolio problem

In this chapter, we discuss an optimization problem arising in the microeconomic theory of inter-temporal consumption decisions. We formulate such a consumption-investment problem as a restless bandit model with two Poisson arms.

In the first section, we give a general description of the investment-consumption problem. Then this problem is formulated as a restless Poisson bandit model in section 3. In the third section, we discuss the method of dynamic programming and the optimal equation, which form the foundation for solving the bandit model and discuss various properties.

2.1 The consumption-investment problem

In the mathematical finance literature, there is a variety of ways to define consumption patterns and objective functions. However in this chapter we look at a special situation in which a fixed proportion of the total wealth is taken out for consumption at the beginning of each period, and the remaining part is invested at the same time. This process is repeated over a finite period of discrete time. The time interval between two consecutive investments is assumed to be of fixed length such as a week or a month, or even a year.

We consider an investment-consumption model consisting of two risky assets with the mean rate of return higher than the risk-free asset. The number of payments from each asset during every period is assumed to be a random variable following a certain distribution. To be specific, we consider the case of independent Poisson distributions for the number of payments from two assets. For each asset, investments grow according to fixed compound interest rates but the number of interest payments are independent and identically distributed with Poisson distribution.

The variation in returns is introduced through the variation in the frequency with which cash flows are paid. This means that the uncertainty inherent in each asset is reflected by the number of payoffs during the corresponding investment period.

For one risky asset, we assume that the intensity rate of the Poisson distribution is unknown and this asset is called the unknown arm. The

intensity rate of the Poisson distribution for the other risky asset is assumed to be known. We name this asset the known arm. The Bayesian approach is followed and the unknown intensity rate of the unknown arm follows a gamma prior distribution. At each point of time of investment, we update the distribution based on the observed number of interest payments during the previous investment period on the unknown arm. Such a posterior distribution comprises all the information necessary for making the current investment decision. The state of the bandit model then consists of the posterior distribution and the amount of wealth available for investment. Taking inflation into consideration, we incorporate a discount factor into the model.

Our objective is to choose an optimal strategy for making selections between the unknown arm and known arm in order to maximize the total expected discounted consumption from all the selections. We consider an investor who is uncertainty-averse with the additional assumptions that no transaction costs exist for switching between the two arms, and the returns on the two risky assets are independent.

2.2 The restless Poisson bandit model

Suppose that selections between the unknown arm and the known arm are made at every time point $n = 0, 1, 2, \cdots, N$. Let β_1 denote the expected return rate of the unknown arm. Assume $\beta_1 > 0$ and it is also unknown. Similarly let β_2 denote the return rate of the risk-free arm. Assume $\beta_2 > 0$

and it is known. Based on all the information gathered to date, if the investor's perception about the quality of assets is that the unknown arm is inferior to the known arm over the entire interval, he (or she) would invest nothing in the unknown arm, and he would invest all of his wealth in it if the unknown arm is considered to be superior to the known arm.

Let's now introduce some notation.

Assume the frequencies of payments during each period are governed by Poisson distributions with intensity rates λ and μ respectively for the unknown and known arms. The assumption is that cash flow streams for the two arms follow two independent Poisson processes, where λ is assumed to be an unknown random variable which follows a gamma prior distribution

$$\lambda \sim G(\eta_0, \tau_0), \quad \eta_0 > 0, \tau_0 > 0.$$

The parameter λ is the jump rate of the Poisson process, which is the expected number of cash payments that accrue over one time period. On the other hand, the parameter μ is assumed to be a known jump rate of the other Poisson process.

Suppose further that the initial wealth at the starting time $n = 0$ is M_0, which would be entirely invested without consumption. Because gamma distributions for the unknown Poisson intensity rate λ form a conjugate class, the posterior distribution of λ at any time $n = 1, 2, \cdots, N$, is still a gamma distribution $G(\eta_n, \tau_n)$.

We let X_n denote the number of payments from the unknown arm

during each investment period. At any time of investment, the observation of events of payment during the previous time interval on the unknown arm updates the prior distribution. Then at the time of the n^{th} investment, the updated posterior distribution of λ is $G(\eta_n, \tau_n)$, conditional on the observed values x_n of X_n, where $\tau_n = \tau_0 + n$, $\eta_n = \eta_0 + \sum_{i=1}^{n} x_i$.

We now consider the discrete, truncated geometric discount sequence

$$A_N = (1, \alpha, \alpha^2, \cdots, \alpha^{N-1}, \cdots), 0 < \alpha \leq 1.$$

Denote that

$$A_N^n = (\alpha^n, \alpha^{n+1}, \cdots, \alpha^{N-1}, \cdots) = \alpha^n A_{N-n}$$

where n means that n investments have been made up-to-date. One advantage of using geometric discounting is that the problem at each stage essentially is the same for the decision maker except for the change in the state.

Let Y_n denote the number of payments from the known arm during the period from time $n-1$ to time n, $n = 1, 2, \cdots$. Then Y_n follows a Poisson distribution with a known intensity μ.

Denote the expected amount of wealth before reinvestment and consumption at time n as M_n. Assuming that a fixed proportion c, $0 < c < 1$, of the total wealth is taken out for consumption, the amount of investment at time n is $(1-c)M_n$. We call the two independent Poisson processes with intensity rates λ and μ respectively a (λ, μ)-bandit process. At the time of the n^{th} investment, the state of the (λ, μ)-bandit is described by $S_n = (G(\eta_n, \tau_n), M_n)$.

An investment strategy $\pi = (\pi_1, \pi_2, \cdots)$ consists of a sequence of rules such that the investment decision at time n is determined by π_n. At time n, the history of past investment decisions and observed numbers of payments from the unknown arm is denoted as $h_n = (z_0, x_1, \cdots, z_{n-1}, x_n)$. The σ-algebra generated by the set of all observed histories prior to time n is denoted as H_n. At the initial time $n = 0$ of investment, no history is available and we define H_0 as the trivial σ-algebra, $\{\Phi, \Omega\}$.

Because the bandit model is a special kind of Markov decision processes and standard results in the theory of Markov decision processes show that there is a deterministic strategy which is optimal, we restrict our discussion to deterministic strategies only. By a deterministic strategy π, we mean that for each n, the function $\pi_n : H_n \to \{1, 2\}$ is measurable. Here for any observed path of history h_n, $\pi_n(h_n) = 2$ means that the known arm is selected for investment at time n and $\pi_n(h_n) = 1$ means that the unknown arm is selected for investment. In fact, the theory of Markov decision processes further shows that there is a Markov strategy which is optimal. A Markov strategy is one in which the investment decision at time n depends only on the current state S_n.

Suppose that the investment strategy π is deterministic and Markovian and that M_n is the total wealth at time n generated by the strategy π. Then $S_n, n = 1, 2, \cdots$, forms a Markov process. The worth of the strategy π is defined as

$$W(G(\eta_0, \tau_0), M_0, \pi) = W(S_0, \pi) = E_\pi(\Sigma_{n=1}^{N} \alpha^{n-1} c_n M_n | G(\eta_0, \tau_0), M_0)$$

where $c_1 = c_2 = \cdots = c_{N-1} = c$, $c_N = 1$, and the subscript π means that the expected value depends on the sample path generated by the strategy π.

The optimal value of the consumption-investment problem is given by

$$V(G(\eta_0, \tau_0), M_0) = V(S_0) = \sup_{\pi} W(G(\eta_0, \tau_0), M_0, \pi),$$

which is the supremum of the worth over the set of all strategies. Our objective is to find the optimal value $V(G(\eta_0, \tau_0), M_0)$ and an optimal strategy π^* such that

$$W(G(\eta_0, \tau_0), M_0, \pi^*) = V(G(\eta_0, \tau_0), M_0).$$

2.3 The method of dynamic programming

Since the above formulated bandit model is essentially a Markov decision process, we have indicated that there is a deterministic Markov strategy which is optimal. To find the optimal strategy, we apply the dynamic programming backward induction method. Based on this method, the optimal investment decision at time n and the optimal value starting at time n are characterized by the so called optimality equation at n. The most crucial part of the equation is the description of the state transition at this time point.

If the state of the bandit model at time n is $S_n = (G(\eta_n, \tau_n), M_n)$, then the posterior distribution $G(\eta_{n+1}, \tau_{n+1})$ at time $n+1$ of the unknown intensity rate λ is given by $\eta_{n+1} = \eta_n + x_{n+1}$ and $\tau_{n+1} = \tau_n + 1$, where x_{n+1}

is the observed number of payments from the unknown arm no matter which arm is selected for investment at time n.

The conditional probability distribution of X_{n+1} given the state S_n is

$$\begin{aligned}
& P(X_{n+1} = x_{n+1}|G(\eta_n, \tau_n)) = P(P(X_{n+1} = x_{n+1}|\lambda)|G(\eta_n, \tau_n)) \\
= & \int_0^\infty \frac{e^{-\lambda}\lambda^{x_{n+1}}}{x_{n+1}!}\frac{1}{\Gamma(\eta_n)}\tau_n^{\eta_n}\lambda^{\eta_n-1}e^{-\lambda\tau_n}d\lambda \\
= & \frac{\tau_n^{\eta_n}\Gamma(x_{n+1}+\eta_n)}{x_{n+1}!\Gamma(\eta_n)(\tau_n+1)^{x_{n+1}+\eta_n}} \\
& \int_0^\infty \frac{1}{\Gamma(x_{n+1}+\eta_n)}(\tau_n+1)^{x_{n+1}+\eta_n}\lambda^{x_{n+1}+\eta_n-1}e^{-\lambda(\tau_n+1)}d\lambda \\
= & \frac{\tau_n^{\eta_n}\Gamma(x_{n+1}+\eta_n)}{x_{n+1}!\Gamma(\eta_n)(\tau_n+1)^{x_{n+1}+\eta_n}}
\end{aligned}$$

The decision on investing in which arm at time n, which is the beginning time of the investment period, has an impact on the wealth M_{n+1}, which is the expected wealth at the end of this period. If we invest in the known arm with the current state S_n at time n, which is $S_n = (G(\eta_n, \tau_n), M_n)$, then we have

$$\begin{aligned}
M_{n+1}^{(2)} &= E[(1-c)M_n(1+\beta_2)^{Y_{n+1}}|\pi_n(S_n) = 0] \\
&= (1-c)M_n\sum_{y_{n+1}=0}^{\infty}(1+\beta_2)^{y_{n+1}}\frac{e^{-\mu}\mu^{y_{n+1}}}{y_{n+1}!} \\
&= (1-c)M_n e^{\beta_2\mu}.
\end{aligned}$$

where the superscript (2) of $M_{n+1}^{(2)}$ indicates that the known arm is selected at the decision time n.

On the other hand, if we invest in the unknown arm when the state at time n is $S_n = (G(\eta_n, \tau_n), M_n)$, then we have

$$M_{n+1}^{(1)} = E[(1-c)M_n(1+\beta_1)^{X_{n+1}}|G(\eta_n, \tau_n), \pi_n(S_n) = 1]$$

$$
\begin{aligned}
&= E[E((1-c)M_n(1+\beta_1)^{X_{n+1}}|\lambda)|G(\eta_n,\tau_n)] \\
&= (1-c)M_n\int_0^\infty \sum_{x_{n+1}=0}^{\infty}(1+\beta_1)^{x_{n+1}}\frac{e^{-\lambda}\lambda^{x_{n+1}}}{x_{n+1}!}g(\eta_n,\tau_n)d\lambda \\
&= (1-c)M_n\int_0^\infty e^{\beta_1\lambda}\frac{1}{\Gamma(\eta_n)}\tau_n^{\eta_n}\lambda^{\eta_n-1}e^{-\lambda\tau_n}d\lambda \\
&= (1-c)M_n\frac{\tau_n^{\eta_n}}{(\tau_n-\beta_1)^{\eta_n}}
\end{aligned}
$$

where $g(\eta_n,\tau_n)=\frac{1}{\Gamma(\eta_n)}\tau_n^{\eta_n}\lambda^{\eta_n-1}e^{-\lambda\tau_n}$ is the gamma density of the unknown intensity λ. Similarly, the superscript (1) of $M_{n+1}^{(1)}$ indicates that the unknown arm is selected at the decision time n.

In essence, the optimality equation gives us a recursive relationship for the optimal value starting at each investment time point. For this purpose, we define the worth of the strategy π starting at time n with the state $S_n=(G(\eta_n,\tau_n),M_n)$ as

$$
\begin{aligned}
W_{N-n}(S_n,\pi) &= W_{N-n}(G(\eta_n,\tau_n),M_n,A_N^n,\pi) \\
&= E_\pi(\Sigma_{t=n+1}^N\alpha^{t-1}c_tM_t|G(\eta_n,\tau_n),M_n).
\end{aligned}
$$

The optimal value of the bandit problem starting at time n with the state $S_n=(G(\eta_n,\tau_n),M_n)$ is defined as

$$
V_{N-n}(S_n)=V_{N-n}(G(\eta_n,\tau_n),M_n,A_N^n)=\sup_\pi W_{N-n}(S_n,\pi).
$$

For $i=1,2$, and $n=0,1,\cdots,N$, let $\Pi_n^{(i)}$ be the set of all strategies that select arm i at time n.

Define

$$
V^{(i)}(G(\eta_n,\tau_n),M_n,A_N^n)=\sup_{\pi\in\Pi_n^{(i)}}W(G(\eta_n,\tau_n),M_n,A_N^n,\pi).
$$

Then

$$V(G(\eta_n, \tau_n), M_n, A_N^n) = \bigvee_{i=1}^{2} V^{(i)}(G(\eta_n, \tau_n), M_n, A_N^n)$$

where

$$\bigvee_{i=1}^{2} a^{(i)} = a^{(1)} \bigvee a^{(2)} = \max\{a^{(1)}, a^{(2)}\}.$$

The investor's decision problem is to choose π to maximize V_{N-n} for every $n = 0, 1, 2, \cdots, N$. Based on the principle of optimality, if the known arm is selected for investment at time n and an optimal strategy is followed starting at time $n+1$, the value of the strategy is given by

$$\begin{aligned}
& V_{N-n}^{(2)}(S_n, A_N^n) \\
= \; & V_{N-n}^{(2)}(G(\eta_n, \tau_n), M_n, A_N^n) \\
= \; & \alpha^n V_{N-n}^{(2)}(G(\eta_n, \tau_n), M_n, A_{N-n}) \\
= \; & \alpha^n c(1-c) M_n e^{\beta_2 \mu} \\
& + \Sigma_{x_{n+1}=0}^{\infty} V_{N-(n+1)}(G(\eta_n + x_{n+1}, \tau_n + 1), (1-c)^2 M_n e^{\beta_2 \mu}, A_N^{n+1}) \\
& \times \frac{\tau_n^{\eta_n} \Gamma(x_{n+1} + \eta_n)}{x_{n+1}! \Gamma(\eta_n)(\tau_n + 1)^{x_{n+1}+\eta_n}} \\
= \; & \alpha^n c M_{n+1}^{(2)} \\
& + \alpha^{n+1} \Sigma_{x_{n+1}=0}^{\infty} V_{N-(n+1)}(G(\eta_n + x_{n+1}, \tau_n + 1), (1-c) M_{n+1}^{(2)}, A_{N-(n+1)}) \\
& \times \frac{\tau_n^{\eta_n} \Gamma(x_{n+1} + \eta_n)}{x_{n+1}! \Gamma(\eta_n)(\tau_n + 1)^{x_{n+1}+\eta_n}}.
\end{aligned}$$

where $M_{n+1}^{(0)} = (1-c) M_n e^{\beta_2 \mu}$.

Similarly, if the unknown arm is selected for investment at time n and an optimal strategy is followed starting at time $n+1$, then the value of the strategy is given by

$$V_{N-n}^{(1)}(S_n, A_N^n)$$

$$
\begin{aligned}
&= V^{(1)}_{N-n}(G(\eta_n,\tau_n), M_n, A^n_N) \\
&= \alpha^n V^{(1)}_{N-n}(G(\eta_n,\tau_n), M_n, A_{N-n}) \\
&= c(1-c)M_n \frac{\tau_n^{\eta_n}}{(\tau_n-\beta_n)^{\eta_n}} \\
&\quad + \Sigma^{\infty}_{x_{n+1}=0} V_{N-(n+1)}(G(\eta_n+x_{n+1},\tau_n+1),(1-c)^2 M_n \frac{\tau_n^{\eta_n}}{(\tau_n-\beta_n)^{\eta_n}}, A^{n+1}_N) \\
&\quad \times \frac{\tau_n^{\eta_n}\Gamma(x_{n+1}+\eta_n)}{x_{n+1}!\Gamma(\eta_n)(\tau_n+1)^{x_{n+1}+\eta_n}} \\
&= cM^{(1)}_{n+1} \\
&\quad + \alpha^{n+1}\Sigma^{\infty}_{x_{n+1}=0} V_{N-(n+1)}(G(\eta_n+x_{n+1},\tau_n+1),(1-c)M^{(1)}_{n+1}, A_{N-(n+1)}) \\
&\quad \times \frac{\tau_n^{\eta_n}\Gamma(x_{n+1}+\eta_n)}{x_{n+1}!\Gamma(\eta_n)(\tau_n+1)^{x_{n+1}+\eta_n}}.
\end{aligned}
$$

where $M^{(1)}_{n+1} = (1-c)M_n \frac{\tau_n^{\eta_n}}{(\tau_n-\beta_n)^{\eta_n}}$.

The method of the dynamic programming backward induction works as follows. At the last investment time $N-1$, the investment decision is myopic because there is only one decision to make. If the updated state is S_{N-1}, the asset to be invested is arm i such that

$$M^{(i)}_N = \max(M^{(1)}_N, M^{(2)}_N).$$

Then we go backward and derive the optimal decision and the optimal value at time $N-2$ based on the optimality equation. That is, if the updated state is S_{N-2}, the asset to be invested in is arm i such that

$$
\begin{aligned}
&V^{(i)}_{N-(N-2)}(S_{N-2}, A^{N-2}_N) \\
&= \max(V^{(1)}_{N-(N-2)}(S_{N-2}, A^{N-2}_N), V^{(2)}_{N-(N-2)}(S_{N-2}, A^{N-2}_N)).
\end{aligned}
$$

This process is repeated until time $n=0$, at which time we have found the optimal value of the overall bandit problem.

2.4 Conclusion

In this chapter we formulated a restless Poisson bandit model for an investment-consumption problem. There were two risky assets in this model, and each of them was regarded as one arm. The variation in returns was reflected by the variation in the frequency with which cash flows are paid. The notation was introduced, and the value function of the strategy was established. The dynamic programming backward induction method can be applied to find an optimal strategy, which is deterministic and Markovian.

Chapter 3

The Allocation Index and Properties

From the end of the previous chapter we see that both the optimal value of the bandit model and the optimal strategy are characterized by the backward induction equations. So in principle both the optimal value and the optimal strategy may be found by recursively applying these equations with the initial value determined at the last investment time of $N-1$. Although this algorithm is more effective than the method of comparing the worths of all possible strategies, we still face the curse of dimensionality of the state space which is in fact uncountable. This means that practically the algorithm of backward induction is computationally impossible at least for moderately large investment horizon N.

In this chapter we start by defining the advantage function of the unknown arm over the known arm. This function is defined at every time

point of investment to be the difference between the best possible performances of starting the investment in the rival arms. Consequently, the unknown arm is optimal at an investment time if and only if the advantage function at this time point is nonnegative, and the known arm is optimal otherwise.

Then in the second section, we show that this investment decision rule based on the advantage function can be simplified to be based on a break-even index value of the parameter μ. In the third section, we demonstrate some properties of the index value.

3.1 The advantage function and its properties

Following the notations introduced in the previous chapter, the advantage function of the unknown arm over the known arm is defined to be

$$\Delta_{N-n}(S_n, \mu) = V^{(1)}_{N-n}(S_n, \mu) - V^{(2)}_{N-n}(S_n, \mu)$$

or with more specific notations,

$$\Delta_{N-n}(G(\eta_n, \tau_n), M_n, \mu) = V^{(1)}_{N-n}(G(\eta_n, \tau_n), M_n, \mu) - V^{(2)}_{N-n}(G(\eta_n, \tau_n), M_n, \mu).$$

The unknown arm is optimal at state S_n if and only if $\Delta_{N-n}(S_n, \mu) \geq 0$, and both arms are optimal at state S_n if $\Delta_{N-n}(S_n, \mu) = 0$.

If the unknown arm is invested initially at time $n = 0$, recall that

$$V^{(1)}_N(S_0, \mu)$$

$$
\begin{aligned}
&= V_N^{(1)}(G(\eta_0,\tau_0), M_0, A_N, \mu) \\
&= cM_0\left(\frac{\tau_0}{\tau_0-\beta_1}\right)^{\eta_0} \\
&\quad +\alpha\Sigma_{x_1=0}^{\infty} V_{N-1}(G(\eta_0+x_1,\tau_0+1),(1-c)M_1^{(1)},A_{N-1},\mu) \\
&\quad \times\frac{\tau_0^{\eta_0}\Gamma(x_1+\eta_0)}{x_1!\Gamma(\eta_0)(\tau_0+1)^{x_1+\eta_0}}
\end{aligned}
$$

where $M_1^{(1)} = M_0\left(\frac{\tau_0}{\tau_0-\beta_1}\right)^{\eta_0}$ is the total wealth available for consumption-investment at time $n=1$. Notice that there is no consumption from M_0 at the initial time 0.

On the other hand, if the known arm is invested initially at time $n=0$, we have

$$
\begin{aligned}
&V_N^{(2)}(S_0,\mu) \\
&= V_N^{(2)}(G(\eta_0,\tau_0), M_0, A_N, \mu) \\
&= cM_0e^{\beta_2\mu} + \alpha\Sigma_{x_1=0}^{\infty} V_{N-1}(G(\eta_0+x_1,\tau_0+1),(1-c)M_1^{(2)},A_{N-1},\mu) \\
&\quad \times\frac{\tau_0^{\eta_0}\Gamma(x_1+\eta_0)}{x_1!\Gamma(\eta_0)(\tau_0+1)^{x_1+\eta_0}}
\end{aligned}
$$

where $M_1^{(2)} = M_0e^{\beta_2\mu}$ is the total wealth available for consumption-investment at time $n=1$.

Hence the advantage function at time $n=0$ of the first investment is

$$
\begin{aligned}
&\Delta_N(S_0,\mu) \\
&= \Delta_N(G(\eta_0,\tau_0), M_0, A_N, \mu) \\
&= V_N^{(1)}(G(\eta_0,\tau_0), M_0, A_N, \mu) - V_N^{(2)}(G(\eta_0,\tau_0), M_0, A_N, \mu) \\
&= c(M_1^{(1)} - M_1^{(2)}) \\
&\quad +\alpha\Sigma_{x_1=0}^{\infty}[V_{N-1}(G(\eta_0+x_1,\tau_0+1),(1-c)M_1^{(1)},A_{N-1},\mu)
\end{aligned}
$$

$$-V_{N-1}(G(\eta_0+x_1,\tau_0+1),(1-c)M_1^{(2)},A_{N-1},\mu)]\frac{\tau_0^{\eta_0}\Gamma(x_1+\eta_0)}{x_1!\Gamma(\eta_0)(\tau_0+1)^{x_1+\eta_0}}$$

where $M_1^{(1)}-M_1^{(2)}=M_0\left[\left(\frac{\tau_0}{\tau_0-\beta_1}\right)^{\eta_0}-e^{\beta_2\mu}\right]$.

Similarly, the advantage function of the unknown arm over the known arm at time n of the $(n+1)^{st}$ selection is defined as

$$\begin{aligned}
&\Delta_N(V_{N-n},\mu)\\
=\;&\Delta_N(G(\eta_n,\tau_n),M_n,A_{N-n},\mu)\\
=\;&V_{N-n}^{(1)}(G(\eta_n,\tau_n),M_n,A_{N-n},\mu)-V_{N-n}^{(2)}(G(\eta_n,\tau_n),M_n,A_{N-n},\mu)\\
=\;&c(M_{n+1}^{(1)}-M_{n+1}^{(2)})\\
&+\alpha\Sigma_{x_{n+1}=0}^{\infty}[V_{N-(n+1)}(G(\eta_n+x_{n+1},\tau_n+1),(1-c)M_{n+1}^{(1)},A_{N-(n+1)},\mu)\\
&-V_{N-(n+1)}(G(\eta_n+x_{n+1},\tau_n+1),(1-c)M_{n+1}^{(2)},A_{N-(n+1)},\mu)]G_{n+1}^*
\end{aligned}$$

where

$$\begin{aligned}
G_{n+1}^* &= \frac{\tau_n^{\eta_n}\Gamma(x_{n+1}+\eta_n)}{x_{n+1}!\Gamma(\eta_n)(\tau_n+1)^{x_{n+1}+\eta_n}},\\
M_{n+1}^{(1)} &= (1-c)M_n(\frac{\tau_n}{\tau_n-\beta_1})^{\eta_n},\\
M_{n+1}^{(0)} &= (1-c)M_n e^{\beta_2\mu}.
\end{aligned}$$

Lemma 3.1.1 *For any $n=0,1,2,\cdots,N$, and any strategy π, all functions*

$$W_{N-n}(G(\eta_n,\tau_n),M_n,\mu,\pi),$$

$$V_{N-n}(G(\eta_n,\tau_n),M_n,\mu),$$

and

$$V_{N-n}^{(i)}(G(\eta_n,\tau_n),M_n,\mu),\quad i=0,1,$$

are continuous and increasing in both μ and M_n, and linear in M_n.

Note: The continuity and monotonicity of these functions in other parameters, such as η_n, τ_n, and so on, can be established in a similar manner. But our primary focus is on discussing the properties of the functions of μ in this chapter.

Proof. Because the discount sequence is geometric, it suffices to show the Lemma by induction on the investment horizon N, with the use of the backward induction equations.

When $N = 1$, we have

$$W_1(G(\eta_0, \tau_0), M_0, \mu, \pi) = M_1^{(1)} = M_0 \left(\frac{\tau_0}{\tau_0 - \beta_1} \right)^{\eta_0},$$

if the unknown arm is selected for investment, and

$$W_1(G(\eta_0, \tau_0), M_0, \mu, \pi) = M_1^{(2)} = M_0 e^{\beta_2 \mu},$$

if the known arm is selected for investment. The Lemma is clearly true in this situation.

Suppose the Lemma is true for horizon N. For the geometric discount sequence of horizon $N + 1$, we have

$$\begin{aligned}
& V_{N+1}^{(1)}(G(\eta_0, \tau_0), M_0, \mu) \\
= \; & cM_1^{(1)} \\
& + \alpha \Sigma_{x_1=0}^{\infty} V_N(G(\eta_0 + x_1, \tau_0 + 1), (1-c)M_1^{(1)}, \mu) \frac{\tau_0^{\eta_0} \Gamma(x_1 + \eta_0)}{x_1! \Gamma(\eta_0)(\tau_0 + 1)^{x_1 + \eta_0}} \\
= \; & cM_0 \left(\frac{\tau_0}{\tau_0 - \beta_1} \right)^{\eta_0} \\
& + \alpha(1-c)M_0 \left(\frac{\tau_0}{\tau_0 - \beta_1} \right)^{\eta_0} \Sigma_{x_1=0}^{\infty} V_N(G(\eta_0 + x_1, \tau_0 + 1), 1, \mu) \\
& \times \frac{\tau_0^{\eta_0} \Gamma(x_1 + \eta_0)}{x_1! \Gamma(\eta_0)(\tau_0 + 1)^{x_1 + \eta_0}}.
\end{aligned}$$

Similarly,

$$
\begin{aligned}
& V^{(2)}_{N+1}(G(\eta_0,\tau_0),M_0,\mu) \\
= \; & cM^{(2)}_1 \\
& +\alpha\Sigma^{\infty}_{x_1=0}V_N(G(\eta_0+x_1,\tau_0+1),(1-c)M^{(0)}_1,\mu)\frac{\tau_0^{\eta_0}\Gamma(x_1+\eta_0)}{x_1!\Gamma(\eta_0)(\tau_0+1)^{x_1+\eta_0}} \\
= \; & cM_0e^{\beta_2\mu} \\
& +\alpha(1-c)M_0e^{\beta_2\mu}\Sigma^{\infty}_{x_1=0}V_N(G(\eta_0+x_1,\tau_0+1),1,\mu) \\
& \times\frac{\tau_0^{\eta_0}\Gamma(x_1+\eta_0)}{x_1!\Gamma(\eta_0)(\tau_0+1)^{x_1+\eta_0}}.
\end{aligned}
$$

The convergence of the infinite sum $\Sigma^{\infty}_{x_1=0}$ warrants the inter-change of limit and summation. So the Lemma is true after applying the induction hypothesis at horizon N. $Q.E.D.$

Corollary 3.1.1 *For any $n=0,1,\cdots,N$, the function $\Delta(G(\eta_n,\tau_n),M_n,\mu)$ is continuous in μ.*

Corollary 3.1.2 *For any given $\eta,\tau,\mu,M\neq 0$, and any strategy π,*

$$W(G(\eta,\tau),M,\mu)=MW(G(\eta,\tau),1,\mu).$$

Therefore, the optimal strategy for the $(G(\eta,\tau),M,\mu,)-$bandit does not depend on the wealth M and hence

$$V(G(\eta,\tau),M,\mu)=MV(G(\eta,\tau),1,\mu).$$

Lemma 3.1.2 *For any given η, τ, and β, we have*

(1) $\Sigma^{\infty}_{k=0}\frac{1}{k!}\frac{\Gamma(\eta+k)}{\Gamma(\eta)}\frac{\tau^{\eta}}{(\tau+1)^{\eta+k}}=1$

(2) $\Sigma^{\infty}_{k=0}\frac{1}{k!}\frac{\Gamma(\eta+k)}{\Gamma(\eta)}\frac{\tau^{\eta}}{(\tau+1)^{\eta+k}}\frac{(\tau+1)^{\eta+k}}{(\tau+1-\beta)^{\eta+k}}=\left(\frac{\tau}{\tau-\beta}\right)^{\eta}$

Proof. For the first identity, we have

$$\begin{aligned}
&\Sigma_{k=0}^{\infty}\frac{1}{k!}\frac{\Gamma(\eta+k)}{\Gamma(\eta)}\frac{\tau^{\eta}}{(\tau+1)^{\eta+k}}\\
=\ &(\frac{\tau}{\tau+1})^{\eta}\Sigma_{k=0}^{\infty}\frac{1}{k!}\frac{\Gamma(\eta+k)}{\Gamma(\eta)}\frac{1}{(\tau+1)^{k}}\\
=\ &(\frac{\tau}{\tau+1})^{\eta}\left[1+\frac{\eta}{\tau+1}+\frac{1}{2!}\frac{(\eta+1)\eta}{(\tau+1)^2}+\cdots+\frac{1}{n!}\frac{(\eta+n-1)\cdots\eta}{(\tau+1)^n}+\cdots\right].
\end{aligned}$$

Notice that the Taylor expansion of $f(x)=\left(\frac{\tau}{\tau-x}\right)^{\eta}$ at $x_0=0$ is

$$\begin{aligned}
f(x)\ &=\ \left(\frac{\tau}{\tau-x}\right)^{\eta}\\
&=\ 1+\frac{\eta}{\tau}x+\frac{1}{2!}\frac{(\eta+1)\eta}{\tau^2}x^2+\cdots+\frac{1}{n!}\frac{(\eta+n+1)\cdots\eta}{\tau^n}x^n+\cdots
\end{aligned}$$

Then taking $x=\frac{\tau}{\tau+1}$, we get

$$\begin{aligned}
&1+\frac{\eta}{\tau}x+\frac{1}{2!}\frac{(\eta+1)\eta}{\tau^2}x^2+\cdots+\frac{1}{n!}\frac{(\eta+n+1)\cdots\eta}{\tau^n}x^n+\cdots\\
=\ &1+\frac{\eta}{\tau+1}+\frac{1}{2!}\frac{(\eta+1)\eta}{(\tau+1)^2}+\cdots+\frac{1}{n!}\frac{(\eta+n-1)\cdots\eta}{(\tau+1)^n}+\cdots\\
=\ &\Sigma_{k=0}^{\infty}\frac{1}{k!}\frac{(\eta+k-1)\cdots\eta}{(\tau+1)^k}.
\end{aligned}$$

On the other hand,

$$f(x)|_{x=\frac{\tau}{\tau+1}}=\left(\frac{\tau+1}{\tau}\right)^{\eta},$$

hence

$$\Sigma_{k=0}^{\infty}\frac{1}{k!}\frac{(\eta+k-1)\cdots\eta}{(\tau+1)^k}=\left(\frac{\tau+1}{\tau}\right)^{\eta}.$$

Therefore

$$\Sigma_{k=0}^{\infty}\frac{1}{k!}\frac{\Gamma(\eta+k)}{\Gamma(\eta)}\frac{\tau^{\eta}}{(\tau+1)^{\eta+k}}=(\frac{\tau}{\tau+1})^{\eta}(\frac{\tau+1}{\tau})^{\eta}=1$$

Consequently for the second equation, replace τ by $\tau-\beta$ in the first identity and we have

$$\Sigma_{k=0}^{\infty}\frac{1}{k!}\frac{\Gamma(\eta+k)}{\Gamma(\eta)}\frac{\tau^{\eta}}{(\tau+1)^{\eta+k}}\frac{(\tau+1)^{\eta+k}}{(\tau+1-\beta)^{\eta+k}}$$

$$= \frac{\tau^\eta}{(\tau-\beta)^\eta}\Sigma_{k=0}^{\infty}\frac{1}{k!}\frac{\Gamma(\eta+k)}{\Gamma(\eta)}\frac{(\tau-\beta)^\eta}{(\tau+1-\beta)^{\eta+k}}$$
$$= \frac{\tau^\eta}{(\tau-\beta)^\eta}. \quad Q.E.D.$$

Lemma 3.1.3 *For any $N = 1, 2, \cdots$, the function $\Delta_N(G(\eta,\tau), M, \mu)$ is decreasing in μ when other parameters are fixed.*

Proof. We proceed by induction. The result is true when $A_N = (1, 0, 0, \cdots)$ is of horizon $N = 1$ since

$$\Delta_1(G(\eta,\tau), M, \mu) = M\left[(\frac{\tau}{\tau-\beta_1})^\eta - e^{\beta_2\mu}\right].$$

Suppose that $\Delta_N(G(\eta,\tau), M, \mu)$ is decreasing in μ for any fixed η, τ and M. If the horizon is $N+1$, we have

$$\begin{aligned}
&\Delta_{N+1}(G(\eta,\tau), M, \mu)\\
=\ & cM\left[(\frac{\tau}{\tau-\beta_1})^\eta - e^{\beta_2\mu}\right]\\
&+\alpha\Sigma_{k=0}^{\infty}V_N(G(\eta+k,\tau+1),(1-c)M(\frac{\tau}{\tau-\beta_1})^\eta,\mu)\frac{\Gamma(k+\eta)\tau^\eta}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}\\
&-\alpha\Sigma_{k=0}^{\infty}V_N(G(\eta+k,\tau+1),(1-c)Me^{\beta_2\mu},\mu)\frac{\Gamma(k+\eta)\tau^\eta}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}\\
=\ & cM\left[(\frac{\tau}{\tau-\beta_1})^\eta - e^{\beta_2\mu}\right]\\
&+\alpha(1-c)M(\frac{\tau}{\tau-\beta_1})^\eta\Sigma_{k=0}^{\infty}V_N(G(\eta+k,\tau+1),1,\mu)\frac{\Gamma(k+\eta)\tau^\eta}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}\\
&-\alpha(1-c)Me^{\beta_2\mu}\Sigma_{k=0}^{\infty}V_N(G(\eta+k,\tau+1),1,\mu)\frac{\Gamma(k+\eta)\tau^\eta}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}
\end{aligned}$$

For any function $f(x)$, define

$$f^+(x) = \max\{0, f(x)\}, \quad f^-(x) = \max\{0, -f(x)\}.$$

Let's write

$$V_N(G(\eta+k,\tau+1),1,\mu) = V_N^{(2)}(G(\eta+k,\tau+1),1,\mu)+\Delta_N^+(G(\eta+k,\tau+1),1,\mu)$$

for the first V_N function, and

$$V_N(G(\eta+k,\tau+1),1,\mu) = V_N^{(1)}(G(\eta+k,\tau+1),1,\mu)+\Delta_N^-(G(\eta+k,\tau+1),1,\mu)$$

for the second V_N function.

Notice that for $V_N^{(1)}(G(\eta+k,\tau+1),M=1,\mu)$, we have

$$\begin{aligned} & M_2^{(1)} = E(M|G(\eta,\tau)) \\ = \; & E(E(M|G(\eta+k,\tau+1))|G(\eta,\tau)) \\ = \; & \Sigma_{k=0}^{\infty}\frac{(\tau+1)^{\eta+k}}{(\tau+1-\beta)^{\eta+k}}\frac{1}{k!}\frac{\Gamma(\eta+k)}{\Gamma(\eta)}\frac{\tau^{\eta}}{(\tau+1)^{\eta+k}} \\ = \; & \left(\frac{\tau}{\tau-\beta}\right)^{\eta}. \end{aligned}$$

Therefore,

$$\begin{aligned} & \Delta_{N+1}(G(\eta,\tau),M,\mu) \\ = \; & cM\left[(\frac{\tau}{\tau-\beta_1})^{\eta}-e^{\beta_2\mu}\right] \\ & +\alpha(1-c)M(\frac{\tau}{\tau-\beta_1})^{\eta}\Sigma_{k=0}^{\infty}[ce^{\beta_2\mu}+\alpha(1-c)e^{\beta_2\mu} \\ & \Sigma_{l=0}^{\infty}V_{N-1}(G(\eta+k+l,\tau+2),1,\mu)\frac{\Gamma(k+l+\eta)(\tau+1)^{\eta+k}}{l!\Gamma(\eta+k)(\tau+2)^{k+l+\eta}}] \\ & \times\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \\ & +\alpha(1-c)M(\frac{\tau}{\tau-\beta_1})^{\eta}\Sigma_{k=0}^{\infty}\Delta_N^+(G(\eta+k,\tau+1),1,\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \\ & -\alpha(1-c)Me^{\beta_2\mu}\Sigma_{k=0}^{\infty}[cM_2^{(1)}+\alpha(1-c)M_2^{(1)} \\ & \Sigma_{l=0}^{\infty}V_{N-1}(G(\eta+k+l,\tau+2),1,\mu)\frac{\Gamma(k+l+\eta)(\tau+1)^{\eta+k}}{l!\Gamma(\eta+k)(\tau+2)^{k+l+\eta}}] \\ & \times\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \\ & -\alpha(1-c)Me^{\beta_2\mu}\Sigma_{k=0}^{\infty}\Delta_N^-(G(\eta+k,\tau+1),1,\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \end{aligned}$$

$$= cM\left[(\frac{\tau}{\tau-\beta_1})^{\eta} - e^{\beta_2\mu}\right]$$
$$+\alpha(1-c)M(\frac{\tau}{\tau-\beta_1})^{\eta}\Sigma_{k=0}^{\infty}\Delta_N^{+}(G(\eta+k,\tau+1),1,\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}$$
$$-\alpha(1-c)Me^{\beta_2\mu}\Sigma_{k=0}^{\infty}\Delta_N^{-}(G(\eta+k,\tau+1),1,\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}$$

after cancelling all other terms.

By the induction hypothesis, $\Delta_N(G(\eta+k,\tau+1),1,\mu)$ is decreasing in μ for any k. Therefore $\Delta_N^{+}(G(\eta+k,\tau+1),1,\mu)$ is nonincreasing in μ but $\Delta_N^{-}(G(\eta+k,\tau+1),1,\mu)$ is nondecreasing in μ.

Hence both

$$\Sigma_{k=0}^{\infty}\Delta_N^{+}(G(\eta+k,\tau+1),1,\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}$$

and

$$-e^{\beta_2\mu}\Sigma_{k=0}^{\infty}\Delta_N^{-}(G(\eta+k,\tau+1),1,\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}$$

are nonincreasing in μ. Therefore $\Delta_{N+1}(G(\eta,\tau),M,\mu)$ is decreasing in μ. *Q.E.D.*

3.2 Existence and properties of index values

In this section, we show that there is a unique solution to the equation $\Delta_N(G(\eta_n,\tau_n),M_n,\mu_{n+1})=0$ in μ_{n+1} denoting the known intensity rate of Poisson process in the known arm during the interval between time point n and $n+1$, where $n=0,1,2,\cdots$.

Theorem 3.2.1 *For any $N=1,2,\cdots$, any η,τ,M, and A_N, there exists*

a unique $\mu^* = \mu^*(\eta, \tau, M, A_N)$ *such that*

$$\Delta_N(G(\eta,\tau), \mu^*, M, A_N) = 0.$$

Moreover, the unknown arm is optimal for horizon N*, if and only if* $\mu \leq \mu^*$*, and the known arm is optimal if and only if* $\mu \geq \mu^*$.

Proof. It is clear that

$$\begin{aligned} \Delta(G(\eta,\tau), 0, M, A_N) &> 0 \quad \textit{and} \\ \lim_{\mu\to\infty} \Delta(G(\eta,\tau), \mu, M, A_N) &< 0. \end{aligned}$$

The theorem is true by Corollary 3.1.1 and Lemma 3.1.3. *Q.E.D.*

Theorem 3.2.2 *Let* $\mu^*_{n+1} = \mu^*_{n+1}(G(\eta_n, \tau_n), M_n)$ *be such that*

$$\Delta_{N-n}(G(\eta_n,\tau_n), M_n, \mu^*_{n+1}) = 0.$$

Then μ^*_{n+1} *is increasing in* η_n*, but decreasing in* τ_n*, and* μ^*_{n+1} *does not depend on the wealth* M_n *at time point* n.

Proof. First of all, $\Delta_{N-n}(G(\eta_n,\tau_n), M_n, \mu^*_{n+1}) = 0$ means that

$$\begin{aligned} & c[M^{(1)}_{n+1} - M^{(2)}_{n+1}] \\ & +\alpha\Sigma^{\infty}_{x_{n+1}=0} V_{N-(n+1)}(G(\eta_n + x_{n+1}, \tau_n + 1), (1-c)M^{(1)}_{n+1}, A_{N-(n+1)})G^*_{n+1} \\ & -\alpha\Sigma^{\infty}_{x_{n+1}=0} V_{N-(n+1)}(G(\eta_n + x_{n+1}, \tau_n + 1), (1-c)M^{(2)}_{n+1}, A_{N-(n+1)})G^*_{n+1} \\ = \; & c[M^{(1)}_{n+1} - M^{(2)}_{n+1}] \\ & +\alpha\Sigma^{\infty}_{x_{n+1}=0} (1-c)M^{(1)}_{n+1} V_{N-(n+1)}(G(\eta_n + x_{n+1}, \tau_n + 1), 1, A_{N-(n+1)})G^*_{n+1} \\ & -\alpha\Sigma^{\infty}_{x_{n+1}=0} (1-c)M^{(2)}_{n+1} V_{N-(n+1)}(G(\eta_n + x_{n+1}, \tau_n + 1), 1, A_{N-(n+1)})G^*_{n+1} \end{aligned}$$

$$
\begin{aligned}
&= [M_{n+1}^{(1)} - M_{n+1}^{(2)}] \\
&\quad \times \left\{ c + (1-c)\alpha \Sigma_{x_{n+1}=0}^{\infty} V_{N-(n+1)}(G(\eta_n + x_{n+1}, \tau_n + 1), 1, A_{N-(n+1)}) G_{n+1}^{*} \right\} \\
&= 0
\end{aligned}
$$

where

$$
G_{n+1}^{*} = \frac{\tau_n^{\eta_n} \Gamma(x_{n+1} + \eta_n)}{x_{n+1}! \Gamma(\eta_n)(\tau_n + 1)^{x_{n+1} + \eta_n}},
$$

and $M_{n+1}^{(i)}$, $i = 1$ or 2, denote the total wealth at time point $n+1$, which is the gain due to the previous investment, $i = 1$ or 2 indicates the arm that is invested in, either the unknown arm or the known arm.

Notice that in the above equation the term

$$
\left\{ c + (1-c)\alpha \Sigma_{x_n=0}^{\infty} V_{N-(n+1)}(G(\eta_n + x_{n+1}, \tau_n + 1), 1, A_{N-(n+1)}) G_{n+1}^{*} \right\}
$$

is positive at any time n, because

1) $0 < c < 1$,

2) $V_{N-(n+1)}(G(\eta_n + x_{n+1}, \tau_n + 1), 1, A_{N-(n+1)}) \geq 0$,

3) $G_{n+1}^{*} > 0$,

Hence $\Delta_{N-n}(G(\eta_n, \tau_n), M_n, \mu_{n+1}) = 0$ is equivalent to $M_{n+1}^{(1)} - M_{n+1}^{(0)} = 0$,

where

$$
\begin{aligned}
M_{n+1}^{(1)} &= (1-c) M_n (\frac{\tau_n}{\tau_n - \beta_1})^{\eta_n}, \\
M_{n+1}^{(0)} &= (1-c) M_n e^{\beta_2 \mu_{n+1}}.
\end{aligned}
$$

Therefore, solve the equation

$$
(1-c) M_n (\frac{\tau_n}{\tau_n - \beta_1})^{\eta_n} = (1-c) M_n e^{\beta_2 \mu_{n+1}},
$$

we obtain

$$\mu^*_{n+1} = \frac{1}{\beta_2}\eta_n \ln(\frac{\tau_n}{\tau_n - \beta_1}) \tag{3.1}$$

$$= \frac{1}{\beta_2}(\eta_{n-1} + x_n) \ln(\frac{\tau_{n-1}+1}{\tau_{n-1} - \beta_1 + 1}) \tag{3.2}$$

$$= \frac{1}{\beta_2}(\eta_0 + \Sigma_{i=1}^n x_i) \ln(\frac{\tau_0 + n}{\tau_0 - \beta_1 + n}) \tag{3.3}$$

From the equation (3.1), μ^*_{n+1} shows to be increasing in η_n, but decreasing in τ_n, but independent on the wealth M_n at time point n. Hence Theorem 3.2.2. is proved. *Q.E.D.*

Here the index value μ^*_{n+1} is acting as a benchmark for evaluating the two arms and determining which arm should be invested in. At any beginning time n of an investment period, the known arm is regarded to be equivalently optimal as the unknown arm, if and only if the value of μ equals to the index value μ^*_{n+1}. If $\mu < \mu^*_{n+1}$, the known arm is regarded as inferior to the unknown arm, and vice versa. In other words, the investor makes her decision by comparing μ and μ^*_{n+1}, if μ is higher, the risky asset on the known arm would be considered, and if μ^*_{n+1} is higher, the other risky asset on the unknown arm would be invested. Hence the availability of the specific index values for successive investment periods provides a solution to the investment-consumption problem in the discrete setting.

As in the definition of myopic strategy, at each stage the arm with the highest immediate expected payoff is selected. Hence the solution to the above two-risky assets bandit model provides a myopic strategy, without the necessity of considering the issue of complete information gathering. Therefore the myopic strategy in this special situation is optimal, the in-

tuition is that the market information on an asset evolves over time, no matter whether this asset is invested or not.

Similarly, the index value μ_n^* during the previous period of investment with the corresponding parameters is

$$\mu_n^* = \frac{1}{\beta_2}\eta_{n-1}\ln(\frac{\tau_{n-1}}{\tau_{n-1}-\beta_1}) \tag{3.4}$$

$$= \frac{1}{\beta_2}(\eta_0+\Sigma_{i=1}^{n-1}x_i)\ln(\frac{\tau_0+(n-1)}{\tau_0-\beta_1+(n-1)}) \tag{3.5}$$

Both the parameters η_{n-1}, τ_{n-1} are known from the information gathering until the current time point $t = n-1$. x_n is regarded as a random variable at time $n-1$, which follows a Poisson distribution with unknown intensity rate λ during the next period interval from time point $n-1$ to n. It means that the detailed information on this Poisson process can be gathered at time point n, and the specific value of x_n can be obtained and hence be regarded as known at that time.

Theorem 3.2.3 *Let $G(\eta_0,\tau_0)$ be the prior distribution of λ, where λ is the unknown intensity rate of the Poisson process on the unknown arm.*

If $\mu_{n+1}^* = \mu_{n+1}^*(G(\eta_n,\tau_n),M_n)$ is the index value during the investment period from time n to $n+1$ such that

$$\Delta_{N-n}(G(\eta_n,\tau_n),M_n,\mu_{n+1}^*) = 0,$$

then

$$\mu_{n+1}^* \geq \mu_n^*$$

if and only if

$$x_n \geq \max\left\{0,\ \eta_{n-1}\left[\frac{\ln\frac{\tau_{n-1}(\tau_n-\beta_1)}{\tau_n(\tau_{n-1}-\beta_1)}}{\ln(\frac{\tau_n}{\tau_n-\beta_1})}\right]\right\}$$

Proof. If $\mu^*_{n+1} \geq \mu^*_n$, from the equations (3.2) and (3.4), we get

$$\frac{1}{\beta_2}(\eta_{n-1} + x_n) \ln(\frac{\tau_n}{\tau_n - \beta_1}) \geq \frac{1}{\beta_2}\eta_{n-1} \ln(\frac{\tau_{n-1}}{\tau_{n-1} - \beta_1}).$$

Solving it, we have

$$x_n \geq \eta_{n-1}\left[\frac{\ln \frac{\tau_{n-1}(\tau_n - \beta_1)}{\tau_n(\tau_{n-1} - \beta_1)}}{\ln(\frac{\tau_n}{\tau_n - \beta_1})}\right].$$

Hence

$$x_n \geq \max\left\{0,\ \eta_{n-1}\left[\frac{\ln \frac{\tau_{n-1}(\tau_n - \beta_1)}{\tau_n(\tau_{n-1} - \beta_1)}}{\ln(\frac{\tau_n}{\tau_n - \beta_1})}\right]\right\}. \tag{3.6}$$

Similarly equation (3.6) implies that $\mu^*_{n+1} \geq \mu^*_n$. $Q.E.D.$

If at any time n, $\mu < \mu^*_n$, the investment is made on the unknown arm. If the observed value x_n satisfies the above equation (3.6), then $\mu < \mu^*_{n+1}$. This means that for the next investment, the unknown remains to be invested again. That is called the *play-the-winner strategy* from the perspective of bandit processes.

From the equations (3.3) and (3.5), $\mu^*_{n+1} \geq \mu^*_n$ is equivalent to

$$x_n \ln \frac{\tau_0 + n}{\tau_0 + n - \beta_1} \geq (\eta_0 + \Sigma_{i=1}^{n-1} x_i) \ln\left(\frac{1 - \frac{\beta_1}{\tau_0+n}}{1 - \frac{\beta_1}{\tau_0+n-1}}\right),$$

which is further equivalent to

$$\eta_0 \leq \frac{x_n \ln \frac{\tau_0+n}{\tau_0+n-\beta_1}}{\ln\left(\frac{1-\frac{\beta_1}{\tau_0+n}}{1-\frac{\beta_1}{\tau_0+n-1}}\right)} - \Sigma_{i=1}^{n-1} x_i.$$

Defining $\gamma_n = \frac{\ln \frac{\tau_0+n}{\tau_0+n-\beta_1}}{\ln\left(\frac{1-\frac{\beta_1}{\tau_0+n}}{1-\frac{\beta_1}{\tau_0+n-1}}\right)}$, we get $\eta_0 \leq \gamma_n x_n - \Sigma_{i=1}^{n-1} x_i$. Recall that it is assumed that $\eta_0 > 0$.

On the other hand, the state $\mu_{n+1}^* \leq \mu_n^*$ is equivalent to

$$\eta_0 \geq \gamma_n x_n - \Sigma_{i=1}^{n-1} x_i.$$

Actually,

$$\eta_0 \geq \max\{0,\ \gamma_n x_n - \Sigma_{i=1}^{n-1} x_i\},$$

where γ_n is defined as above.

3.3 Conclusion

In this chapter the advantage function of the unknown arm over the known arm was established. A geometric discount sequence was introduced to make the two-armed Poisson bandit model well defined. Besides assuming that the numbers of payoff for both assets follow independent Poisson distributions, there were two key assumptions within the derivation of the advantage function. One was that the fixed proportion of wealth is taken for consumption at every beginning time of the successive investment periods, except for the initial time $n = 0$. The other was that the complete information gathered to-date is possible, therefore the state at every investment time is updated continuously. We adopted the Bayesian approach and used the conjugate prior (gamma prior and updated gamma posterior) to describe the changing of states over time.

By backward induction method, the continuity and monotonicity properties of both the value function and the advantage function were proved and discussed. An important result in this chapter was the decreasing in μ

property of the advantage function, which was proved in Lemma 3.1.3. In section 3 the concept of index value was introduced, which is a useful measure to evaluate the two risky assets (two arms), and is a unique solution to the advantage equation. The existence and properties of index values were also discussed in detail. From the index value equation, the myopic strategy was shown to be optimal in the special two-risky assets bandit model. If the number of cash flow X satisfied the inequality (3.7), the tendency of index values was proved to be increasing, hence from the perspective of bandit processes, the play-the-winner strategy would be used.

Chapter 4

An alternative bandit model

In the previous chapter, we formulated a particular bandit model for the investment-consumption problem. Because of the particular ways of consuming a fixed proportion of the wealth and re-investing the remaining amount, we ended up with a myopic optimal strategy. But we allowed for the continual movement of the asset, whether or not it was selected for investment. Hence the bandit model in Chapter 3 was of the restless type.

In this chapter, we examine an alternative bandit model for another type of consumption-investment problem. Such a bandit model is of the classical type and is not restless. This scenario of the consumption-investment problem does not seem to be realistic in the financial world, but nevertheless the statistical problem itself seems to be interesting.

In the first section, we introduce the corresponding Poisson bandit model with two arms. In the second section, we assume that one arm is known but the other is unknown, and derive the existence and proper-

ties of the optimal strategy. In particular we show that the myopic strategy is not optimal and that there is an optimal stopping solution. We then discuss the play-the-winner strategy in the third section for the case of two unknown arms.

4.1 A Poisson bandit model

Similar to the last chapter, we assume that there are two financial assets available for investment. Again, investments are made at discrete time points $n = 0, 1, 2, \cdots, N$, where we allow for $N = \infty$. However, at each time point, we invest a fixed amount in one and only one of the two assets. The whole amount of payoff of each investment is taken for consumption. Without a loss of generality, we assume one unit of investment at each time.

The uncertainty inherent in each asset is also reflected by the number of payoffs during each investment period. We still assume two independent Poisson distribution for the number of payments from the two assets. For these assets, investments grow according to fixed compound interest rates $\beta_1 > 0$, and $\beta_2 > 0$, and the numbers of interest payments are independent and identically distributed following Poisson distributions.

The intensity rate λ for the unknown arm is unknown but follows a gamma prior distribution $G(\eta_0, \tau_0), \eta_0 > 0, \tau_0 > 0$. The intensity rate μ for the known arm is a fixed constant. Moreover, $\beta_1 > 0$, and $\beta_2 > 0$ are known expected return rates on the unknown and known arms respectively.

The discount sequence is $A_N = (1, \alpha, \alpha^2, \cdots, \alpha^{N-1}, 0, 0, \cdots)$, where $0 < \alpha \leq 1$.

For any investment strategy $\pi = (\pi_1, \pi_2, \cdots)$, denote the payoff at time n as Z_n. Our objective is to find an optimal strategy π^* to maximize

$$W_N(G(\eta_0, \tau_0), \mu, \pi) = E_\pi(\Sigma_{n=1}^N \alpha^{n-1} Z_n | G(\eta_0, \tau_0))$$

and to find the optimal value

$$V_N(G(\eta_0, \tau_0), \mu) = \sup_\pi W_N(G(\eta_0, \tau_0), \mu, \pi).$$

Let X_n be the random variable representing the number of events of payments from the unknown arm during the n^{th} investment period, $n = 1, 2, \cdots, N$. Then at the time of the n^{th} investment, the updated posterior distribution of λ is again a gamma distribution $G(\eta_n, \tau_n)$, where $n = 1, 2, \cdots, N$, and $\tau_n = \tau_0 + n$, $\eta_n = \eta_0 + \Sigma_{i=1}^n x_i$.

Notice that the state of the bandit model only consists of the posterior distribution because the amount of investment is fixed.

4.2 The case of one unknown arm

At the time of the n^{th} investment, the posterior is $G(\eta_n, \tau_n)$ and the discount sequence is

$$A_N^n = (\alpha^n, \alpha^{n+1}, \cdots, \alpha^{N-1}, 0, \cdots, 0) = \alpha^n A_{N-n},$$

where A_{N-n} is defined in chapter 2.

If at time $n = 0$, the state is $G(\eta, \tau)$ and the unknown arm is selected for investment and followed by an optimal strategy, we have

$$\begin{aligned} & V_N^{(1)}(G(\eta,\tau),\mu) \\ = & \left(\frac{\tau}{\tau-\beta_1}\right)^{\eta} + \alpha\Sigma_{k=0}^{\infty} V_{N-1}(G(\eta+k,\tau+1),\mu)\frac{1}{k!}\frac{\Gamma(k+\eta)}{\Gamma(\eta)}\frac{\tau^{\eta}}{(\tau+1)^{k+\eta}} \end{aligned}$$

On the other hand, if the known arm is selected for investment and followed by an optimal strategy, we have

$$V_N^{(2)}(G(\eta,\tau),\mu) = e^{\beta_2\mu} + \alpha V_{N-1}(G(\eta,\tau),\mu).$$

Therefore the optimality equation becomes

$$V_N(G(\eta,\tau),\mu) = \max\left\{V_N^{(1)}(G(\eta,\tau),\mu),\ V_N^{(2)}(G(\eta,\tau),\mu)\right\}$$

and the advantage function is

$$\Delta_N(G(\eta,\tau),\mu) = \max\left\{V_N^{(1)}(G(\eta,\tau),\mu) - V_N^{(2)}(G(\eta,\tau),\mu)\right\}.$$

Using arguments similar to the last chapter, we have

Lemma 4.2.1 *The functions $V_N(G(\eta,\tau),\mu)$ and $\Delta_N(G(\eta,\tau),\mu)$ are continuous in μ.*

We show that the advantage function Δ is decreasing in μ.

Lemma 4.2.2 *The function $\Delta_N(G(\eta,\tau),\mu)$ is decreasing in μ for any $N = 1, 2, \cdots$.*

Proof. We proceed by induction similar to the steps in Lemma 3.1.3.

The result is true when $N = 1$ because

$$\Delta_1(G(\eta,\tau),\mu) = \left(\frac{\tau}{\tau-\beta_1}\right)^{\eta} - e^{\beta_2\mu}.$$

Suppose that the result is true for horizon N so that $\Delta_N(G(\eta,\tau),M,\mu)$ is decreasing in μ for any fixed η,τ. Then for the horizon $N+1$, we have,

$$\begin{aligned}
&\Delta_{N+1}(G(\eta,\tau),\mu) \\
= &\left[(\frac{\tau}{\tau-\beta_1})^{\eta} - e^{\beta_2\mu}\right] \\
&+\alpha\Sigma_{k=0}^{\infty} V_N(G(\eta+k,\tau+1),\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \\
&-\alpha V_N(G(\eta,\tau),\mu).
\end{aligned}$$

Write

$$V_N(G(\eta+k,\tau+1),\mu) = V_N^{(2)}(G(\eta+k,\tau+1),\mu) + \Delta_N^{+}(G(\eta+k,\tau+1),\mu)$$

and

$$V_N(G(\eta,\tau),\mu) = V_N^{(1)}(G(\eta,\tau),\mu) + \Delta_N^{-}(G(\eta,\tau),\mu).$$

Then

$$\begin{aligned}
&\Delta_{N+1}(G(\eta,\tau),\mu) \\
= &\left[(\frac{\tau}{\tau-\beta_1})^{\eta} - e^{\beta_2\mu}\right] \\
&+\alpha\Sigma_{k=0}^{\infty}\left[e^{\beta_2\mu} + \alpha V_{N-1}(G(\eta+k,\tau+1),\mu)\right]\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \\
&+\alpha\Sigma_{k=0}^{\infty}\Delta_N^{+}(G(\eta+k,\tau+1),\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \\
&-\alpha\left[(\frac{\tau}{\tau-\beta_1})^{\eta} + \alpha\Sigma_{k=0}^{\infty}V_{N-1}(G(\eta+k,\tau+1),\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}\right] \\
&-\alpha\Delta_N^{-}(G(\eta,\tau),\mu)
\end{aligned}$$

$$
\begin{aligned}
= \; & (1-\alpha)\left[(\frac{\tau}{\tau-\beta_1})^{\eta} - e^{\beta_2\mu}\right] \\
& +\alpha\Sigma_{k=0}^{\infty}\Delta_N^{+}(G(\eta+k,\tau+1),\mu)\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \\
& -\alpha\Delta_N^{-}(G(\eta,\tau),\mu).
\end{aligned}
$$

By the induction hypothesis, $\Delta_N^{+}(G(\eta+k,\tau+1),\mu)$ is nonincreasing in μ for any k, and $\Delta_N^{-}(G(\eta,\tau),\mu)$ is nondecreasing in μ. However $(1-\alpha)\left[(\frac{\tau}{\tau-\beta_1})^{\eta} - e^{\beta_2\mu}\right]$ is decreasing in μ, so is $\Delta_{N+1}(G(\eta,\tau),\mu)$. *Q.E.D.*

Theorem 4.2.1 *For any N,η, and τ, there exists a unique index value $\mu_N^* = \mu_N^*(\eta,\tau)$ such that $\Delta_N(G(\eta,\tau),\mu_N^*) = 0$.*

Moreover, the unknown arm is optimal if and only if $\mu \leq \mu_N^*$ and the known arm is optimal if and only if $\mu \geq \mu_N^*$.

Proof. Clearly $\Delta_N(G(\eta,\tau),0) > 0$ and $\lim_{\mu\to\infty}\Delta_N(G(\eta,\tau),\mu) < 0$. The existence and uniqueness of μ_N^* follows from the continuity and monotonicity of $\Delta_N(G(\eta,\tau),\mu)$ in μ.

The unknown arm is optimal for the $(G(\eta,\tau),\mu)-$bandit if and only if $\Delta_N(G(\eta,\tau),\mu) \geq 0$, which is equivalent to $\mu \leq \mu_N^*$. The known arm is optimal for the $(G(\eta,\tau),\mu)-$bandit if and only if $\Delta_N(G(\eta,\tau),\mu) \leq 0$, which is equivalent to $\mu \geq \mu_N^*$. *Q.E.D.*

Although we have demonstrated the existence of the index value μ_N^*, there is no closed form solution as opposed to the results in the last chapter.

We point out that the index value μ_n^* is not the Gittins index because we have a finite horizon model. But it is the Gittins index when $N = \infty$.

To make a connection with the Gittins index, we check the monotonicity and limit of the sequence $\mu_N^*, N = 1, 2, \cdots$, of index values.

Lemma 4.2.3 *For any η, τ and $A_N = (1, \alpha, \cdots, \alpha^{N-1}, 0, \cdots), 0 < \alpha \leq 1$, if*

$$\Delta_N(G(\eta, \tau), \mu) = 0,$$

then

$$\Delta_{N+1}(G(\eta, \tau), \mu) \geq 0.$$

Proof. The equation $\Delta_N(G(\eta, \tau), \mu) = 0$ implies both

$$\begin{aligned} & (\frac{\tau}{\tau - \beta_1})^\eta - e^{\beta_2 \mu} \\ = \; & \alpha V_{N-1}(G(\eta, \tau), \mu) \\ & -\alpha \Sigma_{k=0}^{\infty} V_{N-1}(G(\eta + k, \tau + 1), \mu) \frac{\Gamma(k + \eta)\tau^\eta}{k!\Gamma(\eta)(\tau + 1)^{k+\eta}} \end{aligned}$$

and $V_N(G(\eta, \tau), \mu) = e^{\beta_2 \mu} + \alpha V_{N-1}(G(\eta, \tau), \mu)$.

Therefore,

$$\begin{aligned} & \Delta_{N+1}(G(\eta, \tau), \mu) \\ = \; & (\frac{\tau}{\tau - \beta_1})^\eta - e^{\beta_2 \mu} \\ & +\alpha \Sigma_{k=0}^{\infty} V_N(G(\eta + k, \tau + 1), \mu) \frac{\Gamma(k + \eta)\tau^\eta}{k!\Gamma(\eta)(\tau + 1)^{k+\eta}} \\ & -\alpha V_N(G(\eta, \tau), \mu) \\ = \; & \alpha \Sigma_{k=0}^{\infty} [V_N(G(\eta + k, \tau + 1), \mu) - V_{N-1}(G(\eta + k, \tau + 1), \mu)] \\ & \times \frac{\Gamma(k + \eta)\tau^\eta}{k!\Gamma(\eta)(\tau + 1)^{k+\eta}} \\ & -\alpha e^{\beta_2 \mu} + \alpha(1 - \alpha) V_{N-1}(G(\eta, \tau), \mu). \end{aligned}$$

For any given k and any optimal strategy for $V_{N-1}(G(\eta+k,\tau+1),\mu)$, we follow the same strategy for $V_N(G(\eta+k,\tau+1),\mu)$ for the first $N-1$ investments, and then always invest in the known arm for the last investment, then

$$V_N(G(\eta+k,\tau+1),\mu) - V_{N-1}(G(\eta+k,\tau+1),\mu) \geq \alpha^{N-1}e^{\beta_2\mu}.$$

On the other hand, if we always invest in the known arm, then

$$V_{N-1}(G(\eta,\tau),\mu) \geq (1+\alpha+\cdots+\alpha^{N-2})e^{\beta_2\mu}.$$

Therefore,

$$\begin{aligned} & V_{N+1}(G(\eta,\tau),\mu) \\ \geq\ & \alpha\Sigma_{k=0}^{\infty}[\alpha^{N-1}e^{\beta_2\mu}]\frac{\Gamma(k+\eta)\tau^{\eta}}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \\ & -\alpha e^{\beta_2\mu} + \alpha(1-\alpha)(1+\alpha+\cdots+\alpha^{N-2})e^{\beta_2\mu} \\ =\ & 0. \qquad Q.E.D. \end{aligned}$$

Theorem 4.2.2 *For any given η and τ, let $\mu_N^*(\eta,\tau)$ be such that*

$$\Delta_N(G(\eta,\tau),\mu_N^*) = 0.$$

For each $N = 1, 2, \cdots$,

$$\frac{\eta}{\beta_2}\ln\left(\frac{\tau}{\tau-\beta_1}\right) = \mu_1^*(\eta,\tau) \leq \mu_2^*(\eta,\tau) \leq \cdots \leq \mu_N^*(\eta,\tau) \leq \cdots.$$

Moreover, the limit $\mu^* = \lim_{N\to\infty}\mu_N^*$ exists such that

$$\frac{\eta}{\beta_2}\ln\left(\frac{\tau}{\tau-\beta_1}\right) < \mu^* < \infty$$

and $\Delta(G(\eta,\tau),\mu^*,A) = 0$, where $A = (1,\alpha,\alpha^2,\cdots)$.

Proof. By the Lemma, we have $\Delta_{N+1}(G(\eta,\tau),\mu_N^*) \geq 0$. But

$$\Delta_{N+1}(G(\eta,\tau),\mu_N^*) = 0.$$

So by the monotonicity of $\Delta_{N+1}(G(\eta,\tau),\mu)$ in μ, we have

$$\mu_N^* \leq \mu_{N+1}^*.$$

The limit $\mu^* = \lim_{N\to\infty}\mu_N^*$ of a non-decreasing sequence of positive numbers exists. Based on the continuity of $\Delta_N(G(\eta,\tau),\mu)$ on N, we see that the limit satisfies

$$\Delta(G(\eta,\tau),\mu^*,A) = 0$$

for $A = (1,\alpha,\alpha^2,\cdots)$.

If $\mu^* = \infty$, then $V(G(\eta,\tau),\mu^*,A) = \infty$ which contradicts with the finiteness of $V(G(\eta,\tau),\mu^*,A)$.

We show that $\frac{\eta}{\beta_2}\ln\left(\frac{\tau}{\tau-\beta_1}\right) < \mu_2^*$, and hence $\frac{\eta}{\beta_2}\ln\left(\frac{\tau}{\tau-\beta_1}\right) < \mu^*$.

Suppose that $\mu_2^* = \frac{\eta}{\beta_2}\ln\left(\frac{\tau}{\tau-\beta_1}\right)$. Then

$$\begin{aligned}
& 0 = \Delta_2(G(\eta,\tau),\mu_2^*) \\
= \;& \alpha\Sigma_{k=0}^{\infty} V_1(G(\eta+k,\tau+1),\mu_2^*)\frac{\Gamma(k+\eta)\tau^\eta}{k!\Gamma(\eta)(\tau+1)^{k+\eta}} \\
& -\alpha V_1(G(\eta,\tau),\mu_2^*) \\
= \;& \alpha\Sigma_{k=0}^{\infty}\left[\max\left\{(\frac{\tau+1}{\tau+1-\beta_1})^{\eta+k},\ (\frac{\tau}{\tau-\beta_1})^{\eta}\right\} - (\frac{\tau}{\tau-\beta_1})^{\eta}\right] \\
& \frac{\Gamma(k+\eta)\tau^\eta}{k!\Gamma(\eta)(\tau+1)^{k+\eta}}
\end{aligned}$$

Let k^* be the smallest integer such that

$$(\frac{\tau+1}{\tau+1-\beta_1})^{\eta+k^*} \geq (\frac{\tau}{\tau-\beta_1})^{\eta}.$$

Then the right hand side is positive. This is a contradiction. $Q.E.D.$

The limit μ^* is the index value for the infinite horizon geometric discount sequence, and is in fact the Gittins index. Moreover, this theorem has two very interesting corollaries.

Corollary 4.2.1 *The myopic strategy is not optimal in general.*

Proof. Take $N = 2$ and μ be such that $\mu_1^*(\eta, \tau) < \mu < \mu_2^*(\eta, \tau)$. Then the unknown arm is uniquely optimal for the $(G(\eta, \tau), \mu)-$bandit, but the myopic strategy selects the known arm. $Q.E.D.$

A myopic strategy focuses only on immediate payoff and ignores information gathering. This corollary says that in order to achieve the best performance, information gathering is necessary so we can make better informed decisions in the future.

Corollary 4.2.2 *If the known arm is uniquely optimal at some investment time, then it remains optimal for the rest of the investment horizon.*

Proof. If the known arm becomes uniquely optimal, when there are n investments to be made, we have $\mu > \mu_n^*$.

But this implies that $\mu > \mu_{n-1}^*$, since there is no change in the state. Hence the known arm is again uniquely optimal at the next investment time. Repeat the argument and we finish the proof. $Q.E.D.$

This is an optimal stopping solution which says that when no further information is gained on the unknown arm and the known arm appears to be uniquely better, then the known arm remains optimal forever.

4.3 The case of two unknown arms

In this section, we assume that the intensity rates λ and μ for the two Poisson distributions are both unknown. λ is assumed to follow a gamma prior distribution $G(\eta_1, \tau_1)$ and μ follows another gamma prior distribution $G(\eta_2, \tau_2)$.

In a way similar to that in the previous section, we define the following functions:

$$W_N(G(\eta_1,\tau_1), G(\eta_2,\tau_2), \pi),$$

$$V_N^{(i)}(G(\eta_1,\tau_1), G(\eta_2,\tau_2)),\ i = 1, 2,$$

$$V_N(G(\eta_1,\tau_1), G(\eta_2,\tau_2)),$$

$$\Delta_N^{(i)}(G(\eta_1,\tau_1), G(\eta_2,\tau_2)).$$

In this section, we focus on deriving a special case of the play-the-winner strategy.

Lemma 4.3.1 *For any* $G(\eta_1,\tau_1), G(\eta_2,\tau_2)$ *and* $A_N = (1, \alpha, \cdots, \alpha^{N-1}, 0, \cdots)$, $0 < \alpha \leq 1$, *we have*

$$\begin{aligned}
& \Delta_N^{(i)}(G(\eta_1,\tau_1), G(\eta_2,\tau_2)) \\
= \ & (1-\alpha)\left[(\frac{\tau_1}{\tau_1-\beta_1})^{\eta_1} - (\frac{\tau_2}{\tau_2-\beta_2})^{\eta_2}\right] \\
& +\alpha\Sigma_{k=0}^{\infty}\Delta_{N-1}^{+}(G(\eta_1+k,\tau_1+1), G(\eta_2,\tau_2))\frac{\Gamma(k+\eta_1)\tau_1^{\eta_1}}{k!\Gamma(\eta_1)(\tau_1+1)^{k+\eta_1}} \\
& -\alpha\Sigma_{l=0}^{\infty}\Delta_{N-1}^{-}(G(\eta_1,\tau_1), G(\eta_2+l,\tau_2+1))\frac{\Gamma(l+\eta_2)\tau_2^{\eta_2}}{l!\Gamma(\eta_2)(\tau_2+1)^{l+\eta_2}}
\end{aligned}$$

Proof. Using the equation $V^{(2)} + \Delta^{+}$ and $V^{(1)} + \Delta^{-}$, we have

$$\Delta_N(G(\eta_1,\tau_1), G(\eta_2,\tau_2))$$

$$
\begin{aligned}
= \; & (\frac{\tau_1}{\tau_1-\beta_1})^{\eta_1} \\
& +\alpha\Sigma_{k=0}^{\infty}[(\frac{\tau_2}{\tau_2-\beta_2})^{\eta_2}+\alpha\Sigma_{l=0}^{\infty}V_{N-2}(G(\eta_1+k,\tau_1+1),G(\eta_2+l,\tau_2+1)) \\
& \times\frac{\Gamma(l+\eta_2)\tau_2^{\eta_2}}{l!\Gamma(\eta_2)(\tau_2+1)^{l+\eta_2}}]\frac{\Gamma(k+\eta_1)\tau_1^{\eta_1}}{k!\Gamma(\eta_1)(\tau_1+1)^{k+\eta_1}} \\
& +\alpha\Sigma_{k=0}^{\infty}\Delta_{N-1}^{+}(G(\eta_1+k,\tau_1+1),G(\eta_2,\tau_2))\frac{\Gamma(k+\eta_1)\tau_1^{\eta_1}}{k!\Gamma(\eta_1)(\tau_1+1)^{k+\eta_1}} \\
& -(\frac{\tau_2}{\tau_2-\beta_2})^{\eta_2} \\
& -\alpha\Sigma_{l=0}^{\infty}[(\frac{\tau_1}{\tau_1-\beta_1})^{\eta_1}+\alpha\Sigma_{k=0}^{\infty}V_{N-2}(G(\eta_1+k,\tau_1+1),G(\eta_2+l,\tau_2+1)) \\
& \times\frac{\Gamma(k+\eta_1)\tau_1^{\eta_1}}{k!\Gamma(\eta_1)(\tau_1+1)^{k+\eta_1}}]\frac{\Gamma(l+\eta_2)\tau_2^{\eta_2}}{l!\Gamma(\eta_2)(\tau_2+1)^{l+\eta_2}} \\
& -\alpha\Sigma_{l=0}^{\infty}\Delta_{N-1}^{-}(G(\eta_1+k,\tau_1+1),G(\eta_2,\tau_2))\frac{\Gamma(l+\eta_2)\tau_2^{\eta_2}}{l!\Gamma(\eta_2)(\tau_2+1)^{l+\eta_2}}
\end{aligned}
$$

which implies the desired equation after cancelling out all V_{N-2} functions.

Q.E.D.

Lemma 4.3.2 *Suppose that the discount sequence $A_N = (1, 1, \cdots, 1, 0, \cdots)$ is uniform (by setting $\alpha = 1$). If $\Delta_N(G(\eta_1, \tau_1), G(\eta_2, \tau_2)) > 0$, then there exists an integer $k^* \geq 0$ such that*

$$\Delta_{N-1}(G(\eta_1 + k^*, \tau_1 + 1), G(\eta_2, \tau_2)) > 0.$$

Proof. By the Lemma, we have

$$
\begin{aligned}
& \Delta_N(G(\eta_1,\tau_1),G(\eta_2,\tau_2)) \\
= \; & \alpha\Sigma_{k=0}^{\infty}\Delta_{N-1}^{+}(G(\eta_1+k,\tau_1+1),G(\eta_2,\tau_2))\frac{\Gamma(k+\eta_1)\tau_1^{\eta_1}}{k!\Gamma(\eta_1)(\tau_1+1)^{k+\eta_1}} \\
& -\alpha\Sigma_{k=0}^{\infty}\Delta_{N-1}^{-}(G(\eta_1,\tau_1),G(\eta_2+l,\tau_2+1))\frac{\Gamma(l+\eta_2)\tau_2^{\eta_2}}{l!\Gamma(\eta_2)(\tau_2+1)^{l+\eta_2}}
\end{aligned}
$$

If no such k^* exists, then

$$\Delta_{N-1}^{+}(G(\eta_1 + k, \tau_1 + 1), G(\eta_2, \tau_2)) = 0$$

for all k. This implies that

$$\Delta_N(G(\eta_1, \tau_1), G(\eta_2, \tau_2)) \leq 0,$$

which is a contradiction. $Q.E.D.$

Similarly, if $\Delta_N(G(\eta_1, \tau_1), G(\eta_2, \tau_2)) < 0$, then there exists an integer $l^* \geq 0$ such that

$$\Delta_{N-1}(G(\eta_1, \tau_1), G(\eta_2 + l^*, \tau_2 + 1)) < 0.$$

These results mean that if one unknown arm is optimal, then there is a positive probability that it remains optimal again at the next stage. But unfortunately it is impossible to derive the formulas for k^* and l^*.

4.4 Conclusion

In this chapter, an alternative bandit model for a new type of consumption-investment problem was examined in detailed. A key assumption in this model was that the amount of investment is fixed, while the remaining part is taken out for consumption. This is different from the assumption of fixed proportion consumption in the previous bandit model in Chapter 3. The consumption pattern is also different, besides the difference in the amount of investment.

Keeping other notations consistent in this new model, we derived the new value functions, and the advantage function of one arm over the other arm. In Lemma 4.2.1, the properties of continuity and monotonicity in μ were examined for these functions. With the establishment of these new

formulae, we demonstrated the existence of the index value μ_N^*. The closed form solution of μ_N^* is not available.

Extending the horizon from finite to infinite and taking a limit of the index values, we obtained the limit μ^*, which is in fact the Gittins index. Two interesting corollaries were provided in Corollary 4.2.1 and Corollary 4.2.2, which are different from the results obtained from the bandit model in the previous chapter.

In section 3, the case of two unknown arms was discussed in detail. We assumed that both arms were unknown, with unknown intensity rate λ and μ for the Poisson distributions. A version of the play-the-winner strategy was derived.

Chapter 5

Conclusions

In this thesis, I considered the investment-consumption decision problem in two different settings, Poisson bandit models with two arms. The uncertainty inherent in the investment return is reflected by the number of cash flows (payoff) from the investment. The investor is uncertainty averse, and tries to learn from historical data. The Bayesian approach was used to describe the information gathering process on these assets. A geometric discount sequence was incorporated in the bandit model.

The dynamic programming backward induction method can be applied to find an optimal strategy, which is deterministic and Markovian. An attractive feature of the problem formulation presented here was that it did not entail computing the return distribution of the assets.

We assumed that the numbers of payoff for both assets follow independent Poisson distributions. There were two additional assumptions introduced to the restless Poisson bandit model in Chapter 3. One was

that the fixed proportion of wealth is taken out for consumption at every beginning time of the successive investment periods, except for the initial time $n = 0$. The other was that it is possible to have complete information gathered to date, therefore the state at every investment time is updated continuously. In an alternative model in Chapter 4, we assumed that the amount of investment is fixed, while the remaining part is taken out for consumption.

Both in Chapter 3 and Chapter 4, we derived the similar value functions, and the advantage function of one arm over another arm. Similar results obtained were the properties of continuity and monotonicity in μ for these functions. I payed special attention to the existence and monotonicity of the index values. The existence of the index values μ_N^* was demonstrated in both models. For the restless Poisson bandit model in Chapter 3, we worked out the specific form of these index values as time evolved during a finite horizon. For the alternative model in Chapter 4, we couldn't obtain the closed form solution. But extending the horizon from finite to infinite and taking a limit of the index values, the limit μ^* was obtained, which is the Gittins index. These index values are the benchmark for an investor to evaluate and compare the different risky assets by ranking.

However, the models that were discussed in this thesis are just basic general two-armed bandit models. For a large number of risky assets, which are not independent of each other, how to approximate the optimal value function and advantage function is still an open question in bandit processes. It will be interesting in the future to establish a bandit model

for the optimal consumption-investment problems with multi-risky assets.

Bibliography

[1] Bank, P. and Föllmer, H. (2002) American options, multi-armed bandits, and optimal consumption plans: a unifying view. *Paris-princeton lectures on mathematical finance 2002* Springer.

[2] Bandyopadhyay, U. and Biswas, A. (2000)A class of adaptive designs. *Sequential Anal.* 19, 45-62.

[3] Bather, J. A. (1983) The minimax risk for the two-armed bandit problem. *Mathematical learning models-theory and algorithms* (eds. by U. Herkenrath, D. Kalin and W. Vogel), 1-11, Springer-Verlag, New York.

[4] Bellman, R. (1956) A problem in the sequential design of experiments. *Sankhyā* A16, 221-229.

[5] Berry, D. A. and Fristedt, B. (1985) Bandit problems - sequential allocation of experiments. *Chapman and Hall, London, New York.*

[6] Bickis, M. G. and Wang, X. (2004) Modelling the information gathering processes. *to be submitted*

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

[8] Black, Fischer, and Sholes, M. (1973) The pricing of options and corporate liabilites. *J. of Political Economy* 81, 637-654.

[9] Chacko, G. and Viceira, L. (1999) Dynamic consumption and portfolio choice with stochastic volatility in incomplete markets, Working paper, Harvard Business School.

[10] Cox, J. and Huang, C. (1989) Optimal consumption and portfolio policies when asset prices follow a diffusion process. *J. Econom. Theory* 49, 33-83.

[11] Cox, J.C. and Ross, S.A. (1976a) A survey of some new results in financial options processes. *J. of Finance* 31, 382-402.

[12] Cox, J.C. and Ross, S.A. (1976b) The valuation of options for alternative stochastic processes. *J. of Financial Economics* 3, 145-166.

[13] Donchev, D.S. and Rachev, S. T. and Steigerwald, D.G. (2002) Optimal policies for investment with time-varying return distributions. *J. of Analysis and Applications* 4(4), 269-312.

[14] EI Karoui, N. and Karatzas, I. (1994) Dynamic allocation problems in continuous time. *Ann. of Applied Probab.* 4, 255-286.

[15] Gittins, J.C. (1979) Bandit processes and dynamic allocation indices (with discussion) J. Roy. *Statis. Soc. Ser. B*41, 148-164.

[16] Karatzas, I., Lehoczky, J. P., Sethi, S. and Shreve, S. E. (1986) Explicit solution ofa general consumption/investment problem. *Math. Oper. Res.* 11, 261-294.

[17] Karatzas, I., Lehoczky, J. P. and Shreve, S. E. (1987) Optimal portfolio and consumption decisions for a "small investor" on a finite horizon. *SIAM J. Control Optim.* 25, 1557-1586.

[18] Karatzas, I., Lehoczky, J. P. and Shreve, S. E. (1990) Existence and uniqueness of multi-agent equilibrium in a stochastic, dynamic consumption/investment model. *Math. Oper. Res.* 125, 80-128.

[19] Karatzas, I. (1984) Gittins indices in the dynamic allocation problem for diffusion processes. *Ann. Probab.* 12(1), 173-192.

[20] Karatzas, I. (1988) On the pricing of american options.*App. Math. Optimization.*17, 37-66.

[21] Liu, J. (1998) Portfolio selection in stochastic environments, Working paper, Graduate School of Business, Stanford University.

[22] Linter, J. (1965) The valuation of risk assets and the selection of risky investments in stock protfolios and capital budgets. *Review of Economics and Statistics* 47, 13-37.

[23] Merton, R. (1969) Lifetime portfolio selection under uncertainty: the continuous-time case. *Review of Economics and Statistics* 51, 247-257.

[24] Merton, R. (1971) Optimum consumption and portfolio rules in a continuous time model. *J. of Econ. Theory* 3, 373-413.

[25] Merton, R. (1973a) Theory of rational option pricing. *Bell J. of Economics* 4, 141-183.

[26] Merton, R. (1973b) An intertemporal capital asset pricing model. *Econnometrica* 41, 8678 - 86.

[27] Merton, R. (1990) Continuous-time finance. *Oxford University Press*, New York

[28] Mandelbaum, A. (1987) Continuous multi-armed bandits and multiparameter processes. *Ann. Probab.* 15(4), 1527-1556.

[29] Pliska, S. (1986) A stochastic calculus model of continuous trading: optimal portfolio. *Math. Oper. Res.* 11, 371-382.

[30] Sharpe, W. (1964) Capital asset prices: a theory of market equilibrium under conditionals of risk. *Journal of Finance* 19, 425-442.

[31] Kogan, L. and Uppal, R. (1999) Risk aversion and optimal portfolio policies in partial and general equilibrium economics, Working paper, Sloan School of Management, MIT.

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

[33] Feldman, D. (1962) Contributions to the "two-armed bandit" problem. *Ann. Math. Statist.* 33, 847-856.

[34] Gihman and Skorohod, A. V. (1979) *Controlled stochastic processes.* Springer-Verlag, New York

[35] Gittins, J. C. (1979) Bandit processes and dynamic allocation indices (with discussions). *J. Roy. Statist. Soc. Ser.* B41, 148-177.

[36] Gittins, J. C. (1989) *Multi-armed bandit allocation indices.* (Foreword by P. Whittle) John Wiley and Sons, Chichester.

[37] Gittins, J. C. and Jones, D. M. (1974) A dynamic allocation index for the sequential design of experiments. *Progress in statistics.* (eds. J. Gani, K. Sarkadi and I. Vince), 241-266, North-Holland, Amsterdam.

[38] Glazebrook, K. D. and Owen, R. W. (1991) New results for generalized bandit processes. *Internat. J. Systems. Sci.* 22, 479-494.

[39] Kullkarni, S. R. and Lugosi, G. (2000) Finite-time lower bounds for the two-armed bandit problem. *IEEE tran. Automat. Control.* 45, 711-714.

[40] Robbins, H. (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58, 527-535.

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

[42] Rodman, L. (1978) On the many-armed bandit problem. *Ann. Probab.* 6, 491-498.

[43] Rothchild, M. (1974) A two-armed bandit theory of market pricing. *J. of Economic Theory* 9, 185-202.

[44] Samaranayake, K. (1992) Stay-with-a-winner rule for dependent Bernoulli bandits. *Ann. Statist.* 20, 2111-2123.

[45] Schmalensee, R. (1975) Alternative models of bandit selection. *J. of Economic Theory* 10, 333-342.

[46] Thompson, W. R. (1993) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 275-294.

[47] Thompson, W. R. (1935) On the theory of apportionment. *Amer, J. Math.* 57, 450-456.

[48] Vogel, W. (1960a) An asymptotic minimax theorem for the two armed bandit problem. *Ann. Math. Statist.* 31, 444-451.

[49] Vogel, W. (1960b) A sequential design for the two-armed bandit. *Ann. Math. Statist.* 31, 430-443.

[50] Wachter, J. (1999) Portfolio and consumption decisions under mean-reverting returns: an exact solution for complete markets, Working paper, Harvard University.

[51] Wang, X. (2000) A bandit process with delayed responses. *Statist. Probab. Lett.* 48, 303-307.

[52] Wang, X. and Pullman, D. (2001) Play-the-winner rule and adaptive designs of clinical trials. *Internat. J. Math. Math. Sci.* 27, 229-236.

[53] Wang, X. (2002) Asymptotic properties of bandit processes with geometric responses. *Statist. Probab. Lett.* 60, 211-217.

[54] Wang, X. (2004) Dynamic pricing with a Poisson bandit model. *submitted*

[55] Wang, X. and Bickis, M. G. (2003) One-armed Bandit Models with continuous and delayed responses. *Math Meth Oper Res* 58, 209-219.

[56] Whittle, P. (1980) Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. Ser. B.* 42(2), 143-149.

[57] Zelen, M. (1969) Play-the winner rule and the controlled clinical trials. *J. Amer. Stat. Asso.* 64, 131-146.