Bandit Processes with Covariates

by

You Liang

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

Master of Science

Department of Statistics

University of Manitoba

Winnipeg, Manitoba, Canada

Bandit Processes with Covariates

By

You Liang

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

Abstract

As in Woodroofe (1982) and Sarkar (1991), we investigate the problem of optimal Bayesian sequential allocation between two treatments incorporating a covariate. The covariate-adjusted response model is determined by a linear regression with either known or unknown σ^2. The goal of our design is to maximize the total discounted expected response from a finite population of patients. This treatment allocation problem is formulated as a two-armed bandit model and the optimal strategy is characterized by means of stochastic dynamic programming. Our model assumption is more general than that in Woodroofe (1982) and Sarkar (1991). We prove that under the general setting, the myopic strategy is not optimal. When one of the two treatments is known, the optimal strategy is characterized by an optimal stopping solution for the linear regression models with either known or unknown σ^2. On the other hand, when both treatments are characterized by linear regression models with unknown parameters, a version of the play-the-winner rule is shown to be optimal for the linear regression models with either known or unknown σ^2.

Acknowledgements

First and foremost, I must graciously thank my thesis advisor Dr. Xikui Wang for his guidance and constant help.

I would thank my committee members for their time spent in reading my thesis. All suggestions and comments are greatly appreciated. Many thanks to Dr. Saumen Mandal of the Department of Statistics and Dr. Jeffrey Pai of Warren Centre for Actuarial Studies and Research at the University of Manitoba.

I also appreciate the Faculty of Graduate Studies for awarding me the Manitoba Graduate Scholarship. I greatly thank Dr. Xikui Wang, the Faculty of Graduate Studies, the Faculty of Science, and the Department of Statistics, for their financial supports for my study and research.

Finally, I would like to extend my gratitude to my family and friends for their constant encouragement and help.

To all the people who helped me and care about me: many thanks!

Contents

1 Introduction **1**

1.1 Bandit Problems 1

1.2 Applications of Bandit Processes 3

1.3 Motivation and Summary 4

1.3.1 Motivation 4

1.3.2 Summary of the Results 6

2 Mathematical Formulation **9**

2.1 Regression Models for Unknown Treatments 9

2.1.1 Bayesian Inference for Known σ^2 10

2.1.2 Bayesian Inference for Unknown σ^2 12

2.2 Bandit Model Overview 14

3 Linear Regression Model with known σ^2 **18**

3.1 One-armed Bandit with a Covariate 18

3.1.1 Bayesian Method and Optimal Selection 20

3.1.2 Main Results: Optimal Strategy and Monotonicity 23

3.2 Two-armed Bandit Model with a Covariate 31

4 Linear Regression Model with unknown σ^2 **37**

4.1 One-armed Bandit with a Covariate 37

4.1.1 Bayesian Method and Optimal Selection 38

4.1.2 Main Results: Optimal Strategy and Monotonicity 41

4.2 Two-armed Bandit Model with a Covariate 49

5 Conclusion and Discussion **57**

Chapter 1

Introduction

1.1 Bandit Problems

In many real situations, sequential decisions are made to maximize some expected reward. But decisions, or the actions they generate, do not just bring in maximum immediate reward; they can help discover new information in order to improve future decisions. Such situations are exemplified by clinical trials where available treatments are experimented to minimize patients' losses or maximize patients' survival times. The general problem arising from these situations is to discover an allocation rule to balance reward maximization based on the information already achieved and information-gathering for better decisions in the future.

The multi-armed bandit model, originally developed by Robins (1952), is a suit-

able way to solve this general problem. Multi-armed bandit processes are sequential decision problems with successive selections from several stochastic processes (or arms, populations, treatments). Time may be discrete or continuous and the processes themselves may also be discrete or continuous. These processes are typically characterized by distributions which are unknown or have unknown parameters. The process selected for each stage depends on the previous selections and observed responses. The goal of bandit problems is to determine a strategy to maximize certain objective function of responses from all selections. This strategy specifies which of the stochastic processes to select for every set of partial history of selections together with their responses.

The majority of the bandit literature takes the Bayesian approach. In this approach, the utility of a strategy is averaged over the parameters with respect to some measure. With a Bayesian approach, a bandit is a typical sequential problem solved by the stochastic dynamic programming method. This is the major reason that much of the recent bandit literature prefers this approach.

The second approach taken in the literature is to consider particular strategies and compare their utilities as a function of the parameters. When the utility of one strategy dominates that of others, this strategy is of course the best one in the class of strategies under consideration. Otherwise, when there does not exist such a dominating strategy, various strategies can be compared using tables.

A third alternative is the minimax approach in which nature is regarded as an opponent in a two-person, zero-sum game. Nature chooses the parameters in the unit square, or in a subset of it, according to some restriction. The decision maker's goal is to minimize the expected difference between what is achieved and what could be achieved if the parameters were known. Nature's goal is to maximize the expected difference.

1.2 Applications of Bandit Processes

First posted in 1930's, bandit processes have been studied by many authors and applied to different areas such as clinical trials in medicine, optimal pricing in finance, job search in economics, and many aspects of optimization.

Since bandit processes take advantage of accruing information to optimize experimental objectives, they have long been proposed as models for clinical trials. A thorough introduction and discussion of bandit models appears in Berry and Fristedt (1985). Hardwick (1995) provides a bandit model for ethical sequential allocation in a clinical trial with immediate dichotomous responses. Eick (1988) introduces a bandit process with geometrically distributed survival times which may be censored. He characterizes optimal strategies by break-even values of the parameters and proposes an optimal stopping solution in the case of infinite horizon. Some of these results have been extended and generalized by Wang (2000) and Wang (2002); these gen-

eralizations may provide ideas for efficient computations and simulations. Hardwick (2006) utilizes a delayed response bandit to allocate treatments in a clinical trial in which patients arrive according to a Poisson process and their response times are exponential.

In the field of optimal pricing, Rothschild (1974) demonstrates that the problem of dynamically pricing a product with an unknown demand function can be formulated as a two-armed bandit model with Bernoulli arms of sequential buyers and an infinite horizon geometric discount sequence. Wang (2007) studies the extension from Bernoulli arms to more general compound Poisson processes, and from infinite horizon geometric discounting to the more realistic finite horizon general discounting.

Many articles have applied bandit models in other research fields, such as the problem of job search and match (McCall (1987), Banks (1992) and Bergemann (2001)), and effective algorithms for general online optimization problems in the bandit setting (McMahan (2004) and Dani (2008)).

1.3 Motivation and Summary

1.3.1 Motivation

Consider a response adaptive design of clinical trial with two treatments. For each patient recruited in the clinical trial, the response depends on the treatment allocated

and a common covariate. Covariates of particular interest in most clinical trials include clinic effects (in multicenter studies), demographic subgroups (such as age, gender, and race) and time trends (a drift in patient characteristics over time). We apply a response adaptive design of clinical trial so that the treatment allocated to the current patient depends on the previous treatment allocations as well as previously observed values of the covariate and response variables. Our objective is to maximize a certain measure of optimality which is defined as a function of the expected responses and the patient-specific covariates from all patients in the trial. However a drawback of this design is the lack of randomization, as explained in Berry and Cheng (2007).

Similarly we can think of the problem of dynamically pricing a product when the demand function is unknown. The profit of selling a product depends on the price posted and a covariate such as the customer's age. The objective is to post alternative prices sequentially in order to maximize the expected value of the total revenue after sells.

These two examples are typical applications of bandit processes for modeling sequential decision problems. An important characteristic of these examples is that we have to effectively deal with the conflict between information gathering (such as learning the effectiveness of the medical treatment or the demand function) and immediate payoff (such as treating the current patient effectively). Information gathering is crucial for understanding the unknown statistical characteristics of the arms or

treatments, and its benefit is in the long run. The sooner we reduce the uncertainty of the unknown treatments, the sooner we can make better informed decisions which will in the long run bring higher overall benefits.

1.3.2 Summary of the Results

Our research in this paper is focused on bandit problems as related to response adaptive designs of clinical trials. We discuss adaptive allocation strategies that adapt on the basis of patient response and observed covariate.

The first work considering covariate models in bandit problems is done by Woodroofe (1974) who investigates a one-armed bandit model with geometrically discounted responses from an infinite population. He established the asymptotic optimality of the myopic strategy. Woodroofe (1982) discusses the optimal treatment allocation policy of a bandit model where the responses of patients depending on a covariate model come from a finite population and the discount sequence is assumed to be uniform. Sarkar (1991) extends Woodroofe's model and describes the difference between the responses from the new and the standard treatment to follow a one-parameter exponential family. Her main result is that the myopic strategy is optimal under several conditions. The major restriction of the above research in bandit models with covariates is that the results depend greatly on model assumptions. Actually, myopic strategies are not always optimal in general settings.

To solve this problem, we extend Woodroofe's and Sarkar's models and formulate treatments to incorporate a linear normal regression model without any restriction. Besides, the discount sequence is extended from infinite horizon geometric discounting to more realistic finite horizon general discounting.

We begin in Chapter 2 with an basic introduction of the theoretical and methodological framework of our bandit problems. A detailed Bayesian analysis of the normal linear regression model is provided. We also explain the general model of bandit processes. In Chapter 3, the bandit problem for modeling treatments characterized by a normal linear regression model with unknown regression parameter and known error distribution is studied. We separately discuss the one-armed bandit model consisting of a new and a standard treatment and the two-armed bandit model consisting of two unknown treatments. When only one treatment is unknown, the treatment allocation is characterized by a sequence of break-even index values, which allows us to define the optimal stopping solution. Moreover, the limiting property of this sequence is discussed in detail. This limiting property provides asymptotic boundary conditions for the index values. When both treatments are unknown, a version of the play-with-winner allocation rule is developed. In Chapter 4, we further generalize the results in Chapter 3 to the more complicated case where both the regression parameter and the error distribution are unknown. Again we determine the optimal strategies for both one-armed and two-armed bandit models. Similar results demonstrate that there ex-

ists an optimal stopping solution when only one treatment is unknown. Moreover, in the case of two unknown treatments, a play-the-winner strategy is applied again to achieve the maximum of the total expected response of all patients. We conclude the thesis in Chapter 5 with a brief summary of achievements and a discussion of future research problems.

Chapter 2

Mathematical Formulation

2.1 Regression Models for Unknown Treatments

In a clinical trial, let X denote the patient-specific covariate of interest and let X_i be the covariate corresponding to the i^{th} selection. The covariates $X_i, i = 1, 2, \ldots,$ are assumed to be independent random variables with a common density function $f(x)$, a domain Ω, and a finite mean μ. Without loss of generality, assume $\mu \geq 0$.

If the i^{th} patient is assigned to an unknown treatment, then the random response Y_i of this patient given $X_i = x_i$ is determined by a regression model

$$Y_i = \beta x_i + \varepsilon_i, \ i = 1, 2, \ldots \tag{2.1.1}$$

where β is the unknown regression parameter describing the effectiveness of the unknown treatment, and ε_i is the random error. We assume that $\varepsilon_i, i = 1, 2, \ldots,$ are

independent Gaussian random variables with mean 0 and variance σ^2. We also assume that the sequence $Y_i, i = 1, 2, \ldots,$ of random variables are independent and identically distributed given $x_i = x$.

The likelihood function based on n observations $\mathscr{O} = \{(x_i, y_i), i = 1, 2, \ldots, n\}$ from the regression model (2.1.1) is

$$\begin{aligned} \ell(\beta, \sigma^2 |\ \mathscr{O}) &= \prod_{i=1}^{n} f(y_i |\ \beta, \sigma^2) \\ &\propto \sigma^{-n} \exp\left[-\left\{ (n-1)\hat{\sigma^2} + (\beta - \hat{\beta})^2 \sum_{i=1}^{n} x_i^2 \right\} / 2\sigma^2 \right], \end{aligned} \quad (2.1.2)$$

where $\propto$ means "proportional to", $\hat{\beta} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$ is the ordinary least squares estimate (OLSE) of β and $(n-1)\hat{\sigma^2} = \sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2$.

2.1.1 Bayesian Inference for Known σ^2

When σ^2 is known, the likelihood function (2.1.2) is reduced to

$$\ell(\beta, \sigma^2 |\ \mathscr{O}) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 (\beta - \hat{\beta})^2 \right].$$

We take the conjugate prior $N(\beta_0, \sigma^2/m)$ for β, then the posterior distribution of β works out to be

$$N\left(\beta^*, \sigma^2 \left(m + \sum_{i=1}^{n} x_i^2 \right)^{-1} \right),$$

where

$$\beta^* = \frac{m\beta_0 + \hat{\beta} \sum_{i=1}^{n} x_i^2}{m + \sum_{i=1}^{n} x_i^2}.$$

It is evident to see that β^* is a compromise between the prior mean β_0 and the least squares estimator $\hat{\beta}$ of β if we rewrite

$$\beta^* = \frac{\sum_{i=1}^n x_i^2}{m + \sum_{i=1}^n x_i^2}\hat{\beta} + \frac{m}{m + \sum_{i=1}^n x_i^2}\beta_0,$$

which is a weighted average of $\hat{\beta}$ and β_0.

The predictive distribution of a future observation Y_{n+1}, given $X_{n+1} = x$ and $\mathscr{O}$ is $g(y|\ x, \mathscr{O})$

$$\begin{aligned}
&= \int_{-\infty}^{+\infty} f(y|\ \beta) g(\beta|\ \mathscr{O}) d\beta \\
&\propto \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2\sigma^2}\left\{x^2\left(\beta - \frac{y}{x}\right)^2 + \left(m + \sum_{i=1}^n x_i^2\right)(\beta - \beta^*)^2\right\}\right] d\beta \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left(\frac{m + \sum_{i=1}^n x_i^2}{m + \sum_{i=1}^n x_i^2 + x^2}\right)(y - \beta^* x)^2\right],
\end{aligned}$$

which is the density function of the normal distribution

$$N\left(\beta^* x, \sigma^2\left(\frac{m + \sum_{i=1}^n x_i^2 + x^2}{m + \sum_{i=1}^n x_i^2}\right)\right).$$

We may use the method of iterative expectations to find the mean and variance of the predictive distribution. The predictive mean of a future observation Y_{n+1} is

$$E(Y_{n+1}|\ \mathscr{O}) = E(E(Y_{n+1}|\ \beta, \mathscr{O})|\ \mathscr{O}) = xE(\beta|\ \mathscr{O}) = \beta^* x,$$

and the predictive variance is given by

$$\begin{aligned}
Var(Y_{n+1}|\mathscr{O}) &= Var(E(Y_{n+1}|\ \beta, \mathscr{O})|\ \mathscr{O})) + E(Var(Y_{n+1}|\ \beta, \mathscr{O})|\ \mathscr{O}) \\
&= \sigma^2\left(\frac{m + \sum_{i=1}^n x_i^2 + x^2}{m + \sum_{i=1}^n x_i^2}\right).
\end{aligned}$$

2.1.2 Bayesian Inference for Unknown σ^2

In this section, we consider a general model where both the regression parameter β and the variance σ^2 are unknown.

Sometimes it is mathematically convenient and instructive to work with the precision $r = 1/\sigma^2$ instead of the variance. Let us assume that the sample $\mathscr{O}$ of size n is drawn from the regression model (2.1.1) where both β and σ^2 are unknown. The likelihood function of β and r given in (2.1.2) is modified to

$$\ell(\beta, r|\ \mathscr{O}) \propto r^{\frac{n}{2}} \exp\left[-\frac{r}{2}\left\{(n-1)\hat{\sigma^2} + (\beta - \hat{\beta})^2 \sum_{i=1}^{n} x_i^2\right\}\right].$$

The natural conjugate prior for (β, r) is such that

$$g(\beta, r) = g(\beta|\ r)g(r),$$

where $g(\beta|\ r)$ is the normal prior $N(\beta_0, mr)$ and $g(r)$ is the Gamma prior $G(u, v)$ so that

$$g(\beta, r) \propto r^{u+\frac{1}{2}-1} \exp\left[-r\left\{v + \frac{m}{2}(\beta - \beta_0)^2\right\}\right].$$

The posterior density is then derived as

$$g(\beta, r|\ \mathscr{O}) \propto \left[r^{1/2} \exp\left\{-\frac{r}{2}(m + \sum_{i=1}^{n} x_i^2)(\beta - \beta^*)^2\right\}\right] \left[r^{\frac{n}{2}+u-1} \exp(-rv^*)\right],$$

where

$$\beta^* = \frac{m\beta_0 + \hat{\beta}\sum_{i=1}^{n} x_i^2}{m + \sum_{i=1}^{n} x_i^2}, \tag{2.1.3}$$

and

$$v^* = v + \frac{(n-1)\hat{\sigma^2}}{2} + \frac{m\sum_{i=1}^n x_i^2(\hat{\beta}-\beta_0)^2}{2(m+\sum_{i=1}^n x_i^2)}. \tag{2.1.4}$$

Hence, the joint posterior distribution of β and r is a product of the conditional posterior distribution $N(\beta^*, (m+\sum_{i=1}^n x_i^2)r)$ of β given r, and the marginal posterior distribution $G\left(\frac{n+2u}{2}, v^*\right)$ of r. The marginal posterior distribution of β can now be obtained by integrating out r from $g(\beta, r|\ \mathscr{O})$, so that we have

$$\begin{aligned} g(\beta|\ \mathscr{O}) &= \int_{-\infty}^{+\infty} g(\beta, r|\ \mathscr{O})dr \\ &\propto \left[1+(\beta-\beta^*)^2(m+\sum_{i=1}^n x_i^2)/2v^*\right]^{-\left(\frac{n+2u+1}{2}\right)} \end{aligned}$$

which is a kernel of a 3-parameter t-distribution with $(n+2u)$ degrees of freedom, location parameter β^*, and scale parameter $\frac{(k+2u)}{2v^*}(m+\sum_{i=1}^n x_i^2)$. It is noted that the posterior variance of β is finite only when $n > 3$.

Moreover, the predictive density of a future observation Y_{n+1}, given $X_{n+1} = x$ and $\mathscr{O}$, is

$$\begin{aligned} g(y|\ x, \mathscr{O}) &= \int_{-\infty}^{+\infty}\int_0^{+\infty} g(\beta, r|\ \mathscr{O}) f(y|\ x, \beta, r)d\beta dr \\ &\propto \int_0^{+\infty} r^{\frac{n+2u+1}{2}-1}\exp\left[-r\left\{v^* + \frac{1}{2}\frac{m+\sum_{i=1}^n x_i^2}{m+\sum_{i=1}^n x_i^2+x^2}(y-\beta^* x)^2\right\}\right]dr \\ &\propto \left[1+\left(\frac{m+\sum_{i=1}^n x_i^2}{m+\sum_{i=1}^n x_i^2+x^2}\right)\left(\frac{n+2u}{2v^*}\right)\left(\frac{(y-\beta^* x)^2}{n+2u}\right)\right]^{\left(\frac{n+2u+1}{2}\right)} \end{aligned}$$

where β^* and v^* are defined in equations (2.1.3) and (2.1.4). It is obvious to see that the predictive distribution of Y_{n+1} is a 3-parameter t-distribution with $(n+2u)$ degrees

of freedom, location parameter $\beta^* x$, and scale parameter $\left(\frac{m+\sum_{i=1}^{n} x_i^2}{m+\sum_{i=1}^{n} x_i^2+x^2}\right)\left(\frac{n+2u}{2v^*}\right)$.

2.2 Bandit Model Overview

In the general setting, an arm of a k-armed bandit problem will be characterized by a probability measure F on the Borel field of subsets of $\mathscr{D}$, the space of probability distributions on $\mathbb{R}$ with the topology of convergence in distribution. The space $\mathscr{D}^k$ of ordered k-tuples of members of $\mathscr{D}$ will be considered to have the product topology arising from the above defined topology on $\mathscr{D}$. The Borel field generated by this product topology is the only σ-field of subsets of $\mathscr{D}^k$ that will be considered; it is the the product σ-field of k copies of the Borel σ-field of $\mathscr{D}$. The component Q_i of $(Q_1, Q_2, \ldots, Q_k) \in \mathscr{D}^k$ governs observations on arm i. Since $(Q_1, Q_2, \ldots, Q_k)$ is random, the probability distribution G of $(Q_1, Q_2, \ldots, Q_k)$ and the space $\mathscr{D}(\mathscr{D}^k)$ play a central role in the decision problem. A member G of $\mathscr{D}(\mathscr{D}^k)$ represents the decision maker's prior information concerning the k arms.

Now we turn to the discussion of one special case of the k-armed bandit problem, the two-armed bandit process for modeling the sequential treatment allocation problem in clinical trials consisting of two treatments.

Suppose that in a clinical trial, there are two independent treatments, treatment 1 and treatment 2, available for a common disease. Patients arrive sequentially and

are treated immediately at times $1, 2, \ldots, N$, one at a time. Let X_i, Y_{1i}, Y_{2i} denote the covariate and the responses of patients from treatment 1 and treatment 2 at time i, respectively. Suppose the distribution of X_i and the linear regression models for Y_{1i} and Y_{2i} are defined in the same way in section 2.1. Further suppose that covariates $X_1, X_2, \ldots, X_N$ are observed sequentially and for each time i, we may observe either Y_{1i} or Y_{2i}, but not both. Our objective is to sequentially allocate treatments to patients in order to maximize the total discounted expected responses from all patients. Given this, it is reasonable to model the above treatment allocation problem as a two-armed bandit consisting of two arms, a discount sequence, a set of strategies, and an objective functions as the optimality criterion for selecting an optimal strategy.

The two arms of this bandit model are the sequences of conditionally independent and identically distributed random responses $\{Y_{k1}, Y_{k2}, \ldots, Y_{kN}\}$ given the distribution of responses from treatment $k, k = 1, 2$.

From the mathematical perspective, it is necessary to add a discount factor α_i for each response in our bandit model. We assume that the discount sequence $A^N = (\alpha_1, \alpha_2, \ldots, \alpha_N, 0, \ldots)$ is nonincreasing and $\sum_{i=1}^{N} \alpha_i < \infty$. The most commonly used discount sequences include the uniform discount sequence $(1, 1, \ldots, 1, 0, \ldots)$ and the finite geometric discount sequence $(1, \alpha, \ldots, \alpha^{N-1}, 0, \ldots), 0 < \alpha < 1$. At time n, the truncated discount sequence $(\alpha_n, \alpha_{n+1}, \ldots, \alpha_N, 0, \ldots)$ is denoted by A_n^N.

By a sequential allocation, we define a strategy $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ in which each π_i takes the value 1 or 2 to indicate that we observe Y_{1i} or Y_{2i}. For the convenience of the proceeding discussion, the i^{th} patient's response under the strategy π is characterized as

$$Z_i = \begin{cases} Y_{1i}, & \text{if treatment 1 is selected,} \\ Y_{2i}, & \text{if treatment 2 is selected.} \end{cases}$$

The response Z_i is a function of X_i and $\mathscr{H}_i^\pi$, where $\mathscr{H}_i^\pi$ denote the σ-field generated by the relevant data available at time i, that is

$$\mathscr{H}_i^\pi = \sigma\{X_1, \ldots, X_{i-1}, \pi_1, \ldots, \pi_{i-1}, Z_1, \ldots, Z_{i-1}\}.$$

$\mathscr{H}_i^\pi$ may be denoted by $\mathscr{H}_i$ if the dependence on π is clear from the context. Since our bandit model is a finite horizon Markov decision process, only deterministic strategies need to be considered (Puterman (1994)). Therefore the strategy π is a sequence of measurable function $\pi_i : \mathscr{H}_i \to \{1, 2\}$ indicating treatment k to be selected at time i, where $k = \pi_i(h_i)$ and h_i is the observed history of the past selections.

Let G be the initial state of our bandit model, then the worth of a strategy π, given the discount sequence $A^N = (\alpha_1, \alpha_2, \ldots, \alpha_N, 0, \ldots)$, is defined as the expected total discounted responses

$$W(G; A^N; \pi) = E_\pi \left(\sum_{i=1}^{N} \alpha_i Z_i \right).$$

The value of the $(G; A^N)$-bandit is the maximum worth

$$V(G; A^N) = \max_{\pi} W(G; A^N; \pi) = \max_{\pi} E_{\pi} \left(\sum_{i=1}^{N} \alpha_i Z_i \right).$$

The objective of our treatment allocation problem is to find an optimal strategy π^* such that

$$W(G; A^N; \pi^*) = V(G; A^N) = \max_{\pi} W(G; A; \pi).$$

At each state $(G; A^N)$, let $V^{(1)}(G; A^N)$ and $V^{(2)}(G; A^N)$ be the worths of the strategies that allocate initially the treatment 1 and treatment 2, respectively, and follow an optimal strategy afterward. Then the dynamic programming equation becomes

$$V(G; A^N) = \max\{V^{(1)}(G; A^N), V^{(2)}(G; A^N)\}.$$

Moreover, we define the advantage of the treatment 1 over the treatment 2 as

$$\Delta(G; A^N) = V^{(1)}(G; A^N) - V^{(2)}(G; A^N),$$

which characterizes the initially optimal selection of treatment. Treatment 1 is optimal if and only if $\Delta(G; A^N) \geq 0$ while treatment 2 is optimal if and only if $\Delta(G; A^N) \leq 0$. Both treatments are optimal when $\Delta(G; A) = 0$, and there is no unique optimal selection.

However, this equation is formidable to solve in general. We will prove in Chapters 3 and 4 that there exists a sequence of break-even index values to describe the optimal selection in our bandit models. The limiting property of this sequence will also be discussed.

Chapter 3

Linear Regression Model with known σ^2

3.1 One-armed Bandit with a Covariate

Let's consider a one-armed bandit, or equivalently a two-armed bandit with one arm known. On the known (or standard) arm (or treatment), the response is random but its mean is a known linear function of the covariate. On the unknown arm (or treatment), the random response depends on the covariate and the relationship is determined by a regression model. However the coefficient of the regression model is unknown. Therefore we face the tradeoff between information gathering (in order to learn the unknown parameter characterizing the unknown treatment) and immediate

payoff (so as to maximize the objective function).

If the i^{th} selection is made with the unknown arm or treatment, the random response Y_{1i} is determined by a regression model

$$Y_{1i} = \beta x_i + \varepsilon_{1i},\ i = 1, 2, \ldots \tag{3.1.1}$$

or

$$E(Y_{1i}|\ x_i) = \beta x_i,\ i = 1, 2, \ldots, \tag{3.1.2}$$

where x_i is the observed covariate, β is the unknown regression parameter and ε_{1i} is the random error. We assume that $\varepsilon_{1i}, i = 1, 2, \ldots,$ are independent Gaussian random variables with mean 0 and variance σ_0^2. We also assume that the sequence $Y_{1i}, i = 1, 2, \ldots,$ of random variables are independent and identically distributed given $x_i = x$.

If the i^{th} selection is made with the standard arm or treatment, the expected value of the response is given by

$$E(Y_{2i}|\ x_i) = \lambda x_i, i = 1, 2, \ldots, \tag{3.1.3}$$

where λ is given. We assume that the sequence $Y_{2i}, i = 1, 2, \ldots,$ of random variables are independent and identically distributed given $x_i = x$.

3.1.1 Bayesian Method and Optimal Selection

Suppose that by time $n = 2, \ldots, N$, some patients are assigned to the new treatment at times $n_1, n_2, \ldots, n_k$. We assume the observations from this new treatment are described as $\mathscr{O}_n = \{x_i, y_{1i}, i = n_1, n_2, \ldots, n_k\}$. Let $\gamma_n = x_{n_1}^2 + \ldots + x_{n_k}^2$ and $\eta_n = x_{n_1} y_{1n_1} + \ldots + x_{n_k} y_{1n_k}$, the OLSE of β can be written as

$$\hat{\beta}_n = \frac{\eta_n}{\gamma_n}. \tag{3.1.4}$$

Assume the prior distribution for β to be $N(\beta_0, \sigma_0^2/m)$, then the posterior distribution of β is again a normal distribution

$$N(\beta_n, \sigma_n^2),$$

where

$$\beta_n = \frac{m\beta_0 + \eta_n}{m + \gamma_n}, \tag{3.1.5}$$

and

$$\sigma_n^2 = \frac{\sigma_0^2}{m + \gamma_n}. \tag{3.1.6}$$

This sequence $\{N(\beta_n, \sigma_n^2), n = 1, 2, \ldots\}$ of distributions forms a process of information-gathering and can be viewed as states of an underlying Markov process. The decision at each time can be determined by state transition and expected immediate response.

Under the normal distribution $N(\beta_n, \sigma_n^2), n = 1, 2, \ldots$, the predictive distribution of a future observation Y_{1n}, given $X_n = x$, is

$$N(\beta_n x, (\sigma_0^2 + \sigma_n^2 x^2))$$

with the density function

$$g(y|\ x) = \frac{1}{\sqrt{2\pi(\sigma_0^2 + \sigma_n^2 x^2)}} \exp\left(-\frac{1}{2(\sigma_0^2 + \sigma_n^2 x^2)}(y - \beta_n x)^2\right).$$

Hence the posterior expected response is

$$E(Y_{1n}|\ x, N(\beta_n, \sigma_n^2)) = \beta_n x.$$

Now the worth of a strategy π for the one-armed bandit with posterior distribution $N(\beta_n, \sigma_n^2)$ of β is defined as

$$W(N(\beta_n, \sigma_n^2), \lambda; A_n^N; \pi) = E_\pi\left(\sum_{i=n}^{N} \alpha_i Z_i|\ N(\beta_n, \sigma_n^2)\right),$$

where

$$Z_i = \begin{cases} Y_{1i}, & \text{if the new treatment is selected,} \\ Y_{2i}, & \text{if the standard treatment is selected.} \end{cases}$$

The optimal value of this bandit model at stage n is

$$V(N(\beta_n, \sigma_n^2), \lambda; A_n^N) = \max_{i=1,2} V^{(i)}(N(\beta_n, \sigma_n^2), \lambda; A_n^N),$$

where $V^{(i)}(N(\beta_n, \sigma_n^2), \lambda; A_n^N), i = 1, 2$ are the optimal values of allocating the new and standard treatment at stage n respectively and then following an optimal strategy. Moreover, the optimal selection of treatment at stage n is described by the advantage function of the new treatment over the standard one, which is

$$\Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^N) = V^{(1)}(N(\beta_n, \sigma_n^2), \lambda; A_n^N) - V^{(2)}(N(\beta_n, \sigma_n^2), \lambda; A_n^N).$$

Then the new (standard) treatment is optimal at stage n if and only if

$$\Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^N) \geq (\leq) 0.$$

By the principle of backward induction,

$$V^{(1)}(N(\beta_n, \sigma_n^2), \lambda; A_n^N) = \alpha_n \beta_n \mu + E(V(N(\beta_{n+1}, \sigma_{n+1}^2), \lambda; A_{n+1}^N) | \ N(\beta_n, \sigma_n^2)), \quad (3.1.7)$$

where

$$\beta_{n+1} = \frac{m\beta_0 + \eta_n + X_n Y_{1n}}{m + \gamma_n + X_n^2}, \quad (3.1.8)$$

and

$$\sigma_{n+1}^2 = \frac{\sigma_0^2}{m + \gamma_n + X_n^2}. \quad (3.1.9)$$

On the other hand,

$$V^{(2)}(N(\beta_n, \sigma_n^2), \lambda; A_n^N) = \alpha_n \lambda \mu + V(N(\beta_n, \sigma_n^2), \lambda; A_{n+1}^N).$$

Simple calculations from (3.1.7) and (3.1.8) by using equations (3.1.5) and (3.1.6) give

$$\beta_{n+1} = \frac{\sigma_0^2 \beta_n + \sigma_n^2 X_n Y_{1n}}{\sigma_0^2 + \sigma_n^2 X_n^2},$$

and

$$\sigma_{n+1}^2 = \frac{\sigma_0^2 \sigma_n^2}{\sigma_0^2 + \sigma_n^2 X_n^2}.$$

So rewriting (3.1.7),

$$\begin{aligned} & V^{(1)}(N(\beta_n, \sigma_n^2), \lambda; A_n^N) \\ = \ & \alpha_n \beta_n \mu + E\left(V\left(N\left(\frac{\sigma_0^2 \beta_n + \sigma_n^2 X_n Y_{1n}}{\sigma_0^2 + \sigma_n^2 X_n^2}, \frac{\sigma_0^2 \sigma_n^2}{\sigma_0^2 + \sigma_n^2 X_n^2}\right), \lambda; A_{n+1}^N\right) | \ N(\beta_n, \sigma_n^2)\right). \end{aligned}$$

3.1.2 Main Results: Optimal Strategy and Monotonicity

We prove the main results in this section, which concern about the existence and structure of optimal strategies.

Lemma 3.1.1. *At each stage $n, n = 1, \ldots, N$, all functions $V(N(\beta_n, \sigma_n^2), \lambda; A_n^N)$ and $V^{(i)}(N(\beta_n, \sigma_n^2), \lambda; A_n^N), i = 1, 2$ are continuous and increasing in λ. Therefore the function $\Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^N)$ is also continuous in λ.*

Proof. This result can be clearly proved by the method of induction on N. □

Lemma 3.1.2. *At each stage $n, n = 1, \ldots, N$, let $A_n^N = (\alpha_n, \alpha_{n+1}, \ldots, \alpha_N, 0, \ldots)$ denote the truncated discount sequence, then the function $\Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^N)$ is nonincreasing in λ.*

Proof. Consider the induction on the horizon N. This lemma is evidently established when $N = n$ since $\Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^n) = \alpha_n \mu(\beta_n - \lambda)$ is nonincreasing in λ.

Suppose this result is true for the horizon N. For any function $f(x)$, we define $f^+(x) = \max\{0, f(x)\}$ and $f^-(x) = \max\{0, -f(x)\}$. Hence,

$$\begin{aligned}
&\Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^{N+1}) \\
&= V^{(1)}(N(\beta_n, \sigma_n^2), \lambda; A_n^{N+1}) - V^{(2)}(N(\beta_n, \sigma_n^2), \lambda; A_n^{N+1}) \\
&= \alpha_n \beta_n \mu \\
&\quad + \int_\Omega \int_{-\infty}^{+\infty} V\left(N\left(\frac{\sigma_0^2 \beta_n + \sigma_n^2 xy}{\sigma_0^2 + \sigma_n^2 x^2}, \frac{\sigma_0^2 \sigma_n^2}{\sigma_0^2 + \sigma_n^2 x^2}\right), \lambda; A_{n+1}^{N+1}\right) g(y|\, x) f(x) dx dy
\end{aligned}$$

$$
\begin{aligned}
& -\alpha_n\lambda\mu - V(N(\beta_n,\sigma_n^2),\lambda;A_{n+1}^{N+1}) \\
= \;& \alpha_n\beta_n\mu \\
& + \int_\Omega\int_{-\infty}^{+\infty} V^{(2)}\left(N\left(\frac{\sigma_0^2\beta_n+\sigma_n^2xy}{\sigma_0^2+\sigma_n^2x^2},\frac{\sigma_0^2\sigma_n^2}{\sigma_0^2+\sigma_n^2x^2}\right),\lambda;A_{n+1}^{N+1}\right)g(y|\,x)f(x)dxdy \\
& + \int_\Omega\int_{-\infty}^{+\infty} \Delta^+\left(N\left(\frac{\sigma_0^2\beta_n+\sigma_n^2xy}{\sigma_0^2+\sigma_n^2x^2},\frac{\sigma_0^2\sigma_n^2}{\sigma_0^2+\sigma_n^2x^2}\right),\lambda;A_{n+1}^{N+1}\right)g(y|\,x)f(x)dxdy \\
& -\alpha_n\lambda\mu - V^{(1)}(N(\beta_n,\sigma_n^2),\lambda;A_{n+1}^{N+1}) - \Delta^-(N(\beta_n,\sigma_n^2),\lambda;A_{n+1}^{N+1}) \\
= \;& \alpha_n\beta_n\mu + \alpha_{n+1}\lambda\mu \\
& + \int_\Omega\int_{-\infty}^{+\infty} V\left(N\left(\frac{\sigma_0^2\beta_n+\sigma_n^2xy}{\sigma_0^2+\sigma_n^2x^2},\frac{\sigma_0^2\sigma_n^2}{\sigma_0^2+\sigma_n^2x^2}\right),\lambda;A_{n+2}^{N+1}\right)g(y|\,x)f(x)dxdy \\
& + \int_\Omega\int_{-\infty}^{+\infty} \Delta^+\left(N\left(\frac{\sigma_0^2\beta_n+\sigma_n^2xy}{\sigma_0^2+\sigma_n^2x^2},\frac{\sigma_0^2\sigma_n^2}{\sigma_0^2+\sigma_n^2x^2}\right),\lambda;A_{n+1}^{N+1}\right)g(y|\,x)f(x)dxdy \\
& -\alpha_n\lambda\mu - \alpha_{n+1}\beta_n\mu \\
& - \int_\Omega\int_{-\infty}^{+\infty} V\left(N\left(\frac{\sigma_0^2\beta_n+\sigma_n^2xy}{\sigma_0^2+\sigma_n^2x^2},\frac{\sigma_0^2\sigma_n^2}{\sigma_0^2+\sigma_n^2x^2}\right),\lambda;A_{n+2}^{N+1}\right)g(y|\,x)f(x)dxdy \\
& -\Delta^-(N(\beta_n,\sigma_n^2),\lambda;A_{n+1}^{N+1}) \\
= \;& (\alpha_n-\alpha_{n+1})\mu(\beta_n-\lambda) \\
& + \int_\Omega\int_{-\infty}^{+\infty} \Delta^+\left(N\left(\frac{\sigma_0^2\beta_n+\sigma_n^2xy}{\sigma_0^2+\sigma_n^2x^2},\frac{\sigma_0^2\sigma_n^2}{\sigma_0^2+\sigma_n^2x^2}\right),\lambda;A_{n+1}^{N+1}\right)g(y|\,x)f(x)dxdy \\
& -\Delta^-(N(\beta_n,\sigma_n^2),\lambda;A_{n+1}^{N+1}) \qquad (3.1.10)
\end{aligned}
$$

The first term in equation (3.1.10) is nonincreasing in λ since $\alpha_n \geq \alpha_{n+1}$ and $\mu \geq 0$. In addition, by the induction hypothesis,

$$\Delta^+\left(N\left(\frac{\sigma_0^2\beta_n+\sigma_n^2xy}{\sigma_0^2+\sigma_n^2x^2},\frac{\sigma_0^2\sigma_n^2}{\sigma_0^2+\sigma_n^2x^2}\right),\lambda;A_{n+1}^{N+1}\right)$$

is nonincreasing and

$$\Delta^-(N(\beta_n, \sigma_n^2), \lambda; A_{n+1}^{N+1})$$

is nondecreasing in λ. Therefore $\Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^N)$ is nonincreasing in λ. □

The existence of an index value characterizing the optimal decision at each stage is proved in the next theorem. In principle, we calculate the index value by the method of backward induction and determine the optimal decision by comparing the index value with the actual regression parameter of the standard treatment.

Theorem 3.1.3. *At each stage $n, n = 1, \ldots, N$ for any β_n and σ_n^2, there exists an index value $\lambda^* = \lambda^*(\beta_n, \sigma_n^2, A_n^N)$ such that $\Delta(N(\beta_n, \sigma_n^2), \lambda^*; A_n^N) = 0$. The set of all such index values forms an interval. Moreover, the new treatment is optimal if and only if $\lambda \leq \lambda^*(\beta_n, \sigma_n^2, A_n^N)$ while the standard one is optimal if and only if $\lambda \geq \lambda^*(\beta_n, \sigma_n^2, A_n^N)$.*

Proof. The existence of the index value follows from the continuity and monotonicity of $\Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^N)$ in λ, and the facts that $\Delta(N(\beta_n, \sigma_n^2), 0; A_n^N) > 0$ as well as $\lim_{\lambda \to \infty} \Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^N) < 0$. Furthermore, the set of all such index values forms an interval since $\Delta(N(\beta_n, \sigma_n^2), \lambda; A_n^N)$ is nonincreasing in λ. Finally, the optimal decision is made based on λ^* according to the advantage function Δ. □

Although the existence of the index has been proved, there is no close form solution of $\lambda^*(\beta_n, \sigma_n^2, A_n^N)$. In the next major theorem, the monotonicity and limiting property

of this decision index is discussed. We assume the special case of a finite horizon geometric discount sequence, in which case the index $\lambda^*(\beta_n, \sigma_n^2, A_n^N)$ is unique.

Lemma 3.1.4. *Assume that* $A_1^N = (1, \alpha, \alpha^2, \ldots, \alpha^{N-1}, 0, \ldots), 0 < \alpha < 1$. *For any given* β *and* σ^2, *if* $\Delta(N(\beta, \sigma^2), \lambda; A_1^N) = 0$, *then* $\Delta(N(\beta, \sigma^2), \lambda; A_1^{N+1}) \geq 0$.

Proof. The equation $\Delta(N(\beta, \sigma^2), \lambda; A_1^N) = 0$ implies both

$$\begin{aligned} \beta\mu - \lambda\mu &= \alpha V(N(\beta, \sigma^2), \lambda; A_1^{N-1}) \\ &\quad -\alpha E\left(V\left(N\left(\frac{\sigma_0^2\beta + \sigma^2 XY}{\sigma_0^2 + \sigma^2 X^2}, \frac{\sigma_0^2\sigma^2}{\sigma_0^2 + \sigma_2 X^2}\right), \lambda; A_1^{N-1}\right) \mid N(\beta, \sigma^2)\right) \end{aligned}$$

and

$$V(N(\beta, \sigma^2), \lambda; A_1^N) = \lambda\mu + \alpha V(N(\beta, \sigma^2), \lambda; A_1^{N-1}).$$

It follows from the above two equations that

$$\begin{aligned} &\Delta(N(\beta, \sigma^2), \lambda; A_1^{N+1}) \\ &= \beta\mu + \alpha E\left(V\left(N\left(\frac{\sigma_0^2\beta + \sigma^2 XY}{\sigma_0^2 + \sigma^2 X^2}, \frac{\sigma_0^2\sigma^2}{\sigma_0^2 + \sigma_2 X^2}\right), \lambda; A_1^N\right) \mid N(\beta, \sigma^2)\right) \\ &\quad -\lambda\mu - \alpha V(N(\beta, \sigma^2), \lambda; A_1^N)) \\ &= \alpha V(N(\beta, \sigma^2), \lambda; A_1^{N-1})) \\ &\quad -\alpha E\left(V\left(N\left(\frac{\sigma_0^2\beta + \sigma^2 XY}{\sigma_0^2 + \sigma^2 X^2}, \frac{\sigma_0^2\sigma^2}{\sigma_0^2 + \sigma_2 X^2}\right), \lambda; A_1^{N-1}\right) \mid N(\beta, \sigma^2)\right) \\ &\quad +\alpha E\left(V\left(N\left(\frac{\sigma_0^2\beta + \sigma^2 XY}{\sigma_0^2 + \sigma^2 X^2}, \frac{\sigma_0^2\sigma^2}{\sigma_0^2 + \sigma_2 X^2}\right), \lambda; A_1^N\right) \mid N(\beta, \sigma^2)\right) \\ &\quad -\alpha\lambda\mu - \alpha^2 V(N(\beta, \sigma^2), \lambda; A_1^{N-1}) \\ &= \alpha \int_\Omega \int_{-\infty}^{+\infty} \left[V\left(N\left(\frac{\sigma_0^2\beta + \sigma^2 xy}{\sigma_0^2 + \sigma^2 x^2}, \frac{\sigma_0^2\sigma^2}{\sigma_0^2 + \sigma^2 x^2}\right), \lambda; A_1^N\right)\right. \end{aligned}$$

$$-V\left(N\left(\frac{\sigma_0^2\beta+\sigma^2xy}{\sigma_0^2+\sigma^2x^2},\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2x^2}\right),\lambda;A_1^{N-1}\right)\Bigg]\,g(y|\,x)f(x)dxdy$$

$$-\alpha\lambda\mu+\alpha(1-\alpha)V(N(\beta,\sigma^2),\lambda;A_1^{N-1}). \tag{3.1.11}$$

Let π^* be the optimal strategy for

$$V\left(N\left(\frac{\sigma_0^2\beta+\sigma^2xy}{\sigma_0^2+\sigma^2x^2},\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2x^2}\right),\lambda;A_1^{N-1}\right).$$

We follow π^* for

$$V\left(N\left(\frac{\sigma_0^2\beta+\sigma^2xy}{\sigma_0^2+\sigma^2x^2},\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2x^2}\right),\lambda;A_1^{N}\right)$$

during the first $N-1$ stages and choose the standard treatment at stage N, then

$$V\left(N\left(\frac{\sigma_0^2\beta+\sigma^2xy}{\sigma_0^2+\sigma^2x^2},\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2x^2}\right),\lambda;A_1^{N}\right)$$
$$-V\left(N\left(\frac{\sigma_0^2\beta+\sigma^2xy}{\sigma_0^2+\sigma^2x^2},\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2x^2}\right),\lambda;A_1^{N-1}\right)\geq\alpha^{N-1}\lambda\mu \tag{3.1.12}$$

for all x and y. On the other hand,

$$V\left(N\left(\frac{\sigma_0^2\beta+\sigma^2xy}{\sigma_0^2+\sigma^2x^2},\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2x^2}\right),\lambda;A_1^{N-1}\right)\geq(1+\alpha+\ldots+\alpha^{N-2})\lambda\mu \tag{3.1.13}$$

if we select the standard treatment all the time. Therefore, a straight calculation from (3.1.11) by using equations (3.1.12) and (3.1.13) gives

$$\Delta(N(\beta,\sigma^2),\lambda;A_1^{N+1})\geq\alpha^N\lambda\mu-\alpha\lambda\mu+\alpha(1-\alpha)((1+\alpha+\ldots+\alpha^{N-2})\lambda\mu)=0,$$

as desired. □

Theorem 3.1.5. *For fixed β and σ^2 but changing $N = 1, 2, \ldots$, let $\lambda^*(\beta, \sigma^2, A_1^N)$ be the index value such that $\Delta(N(\beta, \sigma^2), \lambda^*(\beta, \sigma^2, A_1^N); A_1^N) = 0$, where $A_1^N = (1, \alpha, \ldots, \alpha^{N-1}, 0, \ldots)$. Then*

$$\beta = \lambda^*(\beta, \sigma^2, A_1^1) \leq \lambda^*(\beta, \sigma^2, A_1^2) \leq \ldots \leq \lambda^*(\beta, \sigma^2, A_1^N) \leq \ldots.$$

Moreover, the limit $\lambda^*(\beta, \sigma^2) = \lim_{N\to\infty} \lambda^*(\beta, \sigma^2, A_1^N)$ exists such that $\beta < \lambda^*(\beta, \sigma^2) < \infty$ and $\Delta(N(\beta, \sigma^2), \lambda^*(\beta, \sigma^2); A) = 0$, where $A = (1, \alpha, \alpha^2, \ldots)$.

Proof. Based on the monotonicity of

$$\Delta(N(\beta, \sigma^2), \lambda, A_1^{N+1})$$

in λ and the results that

$$\Delta(N(\beta, \sigma^2), \lambda^*(\beta, \sigma^2, A_1^{N+1}); A_1^{N+1}) = 0$$

and

$$\Delta(N(\beta, \sigma^2), \lambda^*(\beta, \sigma^2, A_1^N); A_1^{N+1}) \geq 0,$$

we conclude that

$$\lambda^*(\beta, \sigma^2, A_1^N) \leq \lambda^*(\beta, \sigma^2, A_1^{N+1})$$

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

Furthermore, the limit of the nondecreasing sequence $\lambda^*(\beta, \sigma^2, A_1^N)$ exists. Let $\lambda^*(\beta, \sigma^2) = \lim_{N\to\infty} \lambda^*(\beta, \sigma^2, A_1^N)$, then $\lambda^*(\beta, \sigma^2)$ satisfies the equation

$\Delta(N(\beta,\sigma^2),\lambda^*(\beta,\sigma^2),A)=0$ due to the fact that $\Delta(N(\beta,\sigma^2),\lambda,A_1^N)$ is continuous in N. If $\lambda^*(\beta,\sigma^2)=\infty$, then $V(N(\beta,\sigma^2),\lambda^*(\beta,\sigma^2),A)=\infty$, which contradicts the finiteness of the optimal value function.

Now we finish the proof by showing that $\beta<\lambda^*(\beta,\sigma^2,A_1^2)$. By contradiction, suppose that $\beta=\lambda^*(\beta,\sigma^2,A_1^2)$, then

$$\begin{aligned}
0 &= \Delta(N(\beta,\sigma^2),\lambda^*(\beta,\sigma^2,A_1^2),A_1^2)\\
&= \beta\mu+\alpha E\left(V\left(N\left(\frac{\sigma_0^2\beta+\sigma^2XY}{\sigma_0^2+\sigma^2X^2},\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma_2X^2}\right),\lambda^*(\beta,\sigma^2,A_1^2);A_1^1\right)\mid N(\beta,\sigma^2)\right)\\
&\quad -\lambda^*(\beta,\sigma^2,A_1^2)\mu-\alpha V(N(\beta,\sigma^2),\lambda^*(\beta,\sigma^2,A_1^2);A_1^1)\\
&= \alpha\int_\Omega\int_{-\infty}^{+\infty}\left[V\left(N\left(\frac{\sigma_0^2\beta+\sigma^2xy}{\sigma_0^2+\sigma^2x^2},\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2x^2}\right),\lambda^*(\beta,\sigma^2,A_1^2);A_1^1\right)\right.\\
&\quad \left.-V(N(\beta,\sigma^2),\lambda^*(\beta,\sigma^2,A_1^2);A_1^1)\right]g(y|\,x)f(x)dxdy\\
&= \alpha\int_\Omega\int_{-\infty}^{+\infty}\left[\max\{\frac{\sigma_0^2\beta+\sigma^2xy}{\sigma_0^2+\sigma^2x^2}\mu,\beta\mu\}-\beta\mu\right]g(y|\,x)f(x)dxdy. \qquad (3.1.14)
\end{aligned}$$

Since $Y|x$ is normally distributed as $N(\beta x,(\sigma_0^2+\sigma^2x^2))$, let y^* be the smallest y larger than βx when x is given. Then the right side of equation (3.1.14) is positive, which is a contradiction. □

This fundamental result reveals the important idea of balancing the immediate payoff and information gathering during the decision process. Moreover, it provides a monotonic approximation for the Gittins index $\lambda^*(\beta,\sigma^2)$ which is formidably computed in practice. There are two interesting corollaries in the proceeding context. The first corollary shows the non-optimality of the myopic strategy. The second one

illustrates the optimal stopping solution.

In bandit problems generally, it may be wise to sacrifice some potential early payoff for the prospect of gaining information that will allow for more informed choices later. In our model, a strategy is myopic for normal distribution $N(\beta, \sigma^2)$ when the new treatment (standard treatment, respectively) is chosen if and only if $\beta \geq (\leq)\lambda$. This strategy is not optimal unless $N = 1$.

Corollary 3.1.6. *The myopic strategy is not optimal in general.*

Proof. For a two-stage bandit problem, let λ be such that $\beta < \lambda < \lambda^*(\beta, \sigma^2, A_1^2)$. The new treatment is uniquely optimal at the first stage while the myopic strategy selects the standard treatment. □

Corollary 3.1.7. *If the standard treatment is uniquely optimal at stage n, then it is optimal for the rest of the decision horizon.*

Proof. If the standard treatment is uniquely optimal at stage n, then $\lambda > \lambda^*(\beta, \sigma^2, A_1^n)$, which indicates that $\lambda > \lambda^*(\beta, \sigma^2, A_1^{n-1})$. Clearly, the standard treatment is uniquely optimal again. □

In view of the above corollary, if the standard treatment is uniquely optimal, no information is gathered on the new treatment and the state of the bandit model is not changed. Hence the standard treatment is selected again by the monotonicity of the decision index. Based on this optimal stopping solution, we start by the initial

selection of the new treatment and continue until it is optimal to permanently switch to the standard one. By Theorem 3.1.3 and Theorem 3.1.5, the optimal time to switch to the standard treatment is the smallest n such that $\lambda^*(\beta_n, \sigma_n^2, A_n^N) < \lambda$. Furthermore, the above results illustrate that the longer the decision horizon N, the more profitable the information gathering, and the more chance to choose the new treatment. The sequence of index values approaches the Gittins index when N gets sufficiently large.

3.2 Two-armed Bandit Model with a Covariate

In this section, we extend our results to the case of two unknown arms, that is we examine the two-armed bandit processes. Again the variance σ^2 is assumed to be known for both arms.

We discuss the two-armed bandit problem consisting of two unknown treatments with regression parameters $\beta_i, i = 1, 2$, following the prior distributions $N(\beta_{i0}, \sigma_0^2/m)$, $i = 1, 2$, respectively. The first treatment is characterized as

$$Y_{1i} = \beta_1 x_i + \varepsilon_{1i}, i = 1, 2, \ldots, \tag{3.2.1}$$

and the second one is described as

$$Y_{2i} = \beta_2 x_i + \varepsilon_{2i}, i = 1, 2, \ldots, \tag{3.2.2}$$

where x_i is the observed covariate, and ε_{1i} and ε_{2i} are independent Gaussian random variables with mean 0 and known variance σ_0^2.

Similar to the framework in the previous section, we define $\gamma_{1n}, \eta_{1n}, \beta_{1n}, \sigma_{1n}^2$ for the unknown treatment 1 and $\gamma_{2n}, \eta_{2n}, \beta_{2n}, \sigma_{2n}^2$ for the unknown treatment 2. Further calculations show that at stage n, the predictive distribution of a future observation Y_{1n} from treatment 1, given $X_n = x$, is

$$N(\beta_{1n}x, (\sigma_0^2 + \sigma_{1n}^2 x^2)).$$

and the predictive distribution of Y_{2n} from treatment 2 is

$$N(\beta_{2n}x, \sigma_0^2 + \sigma_{2n}^2 x^2)).$$

Hence the predictive densities of Y_{1n} and Y_{2n}, given $X_n = x$, respectively, are

$$g_1(y|\ X_n = x) = \frac{1}{\sqrt{2\pi(\sigma_0^2 + \sigma_{1n}^2 x^2)}} \exp\left(-\frac{1}{2(\sigma_0^2 + \sigma_{1n}^2 x_n^2)}(y - \beta_{1n}x)^2\right),$$

and

$$g_2(y|\ X_n = x) = \frac{1}{\sqrt{2\pi(\sigma_0^2 + \sigma_{2n}^2 x^2)}} \exp\left(-\frac{1}{2(\sigma_0^2 + \sigma_{2n}^2 x_n^2)}(y - \beta_{2n}x)^2\right).$$

Furthermore, in the similar way as before, we characterize the worth function

$$W(N(\beta_{1n}, \sigma_{1n}^2), N(\beta_{2n}, \sigma_{2n}^2); A_n^N; \pi),$$

the optimal value function

$$V(N(\beta_{1n}, \sigma_{1n}^2), N(\beta_{2n}, \sigma_{2n}^2); A_n^N),$$

the optimal value function for each treatment

$$V^{(i)}(N(\beta_{1n}, \sigma_{1n}^2), N(\beta_{2n}, \sigma_{2n}^2); A_n^N), i = 1, 2,$$

and the advantage function

$$\Delta(N(\beta_{1n}, \sigma_{1n}^2), N(\beta_{2n}, \sigma_{2n}^2); A_n^N),$$

at each stage $n, n = 1, 2, \ldots, N$.

Since it is necessary to understand the unknown parameters from both treatments, the myopic strategy is reasonable on the first stage. For continuing decisions we will prove the optimality of a version of the play-the-winner strategy. The play-the-winner strategy was first studied by Zelen (1969).

Lemma 3.2.1. *For any truncated discount sequence* $A_n^N = (\alpha_n, \alpha_{n+1}, \ldots, \alpha_N, 0, \ldots)$, $n = 1, 2, \ldots, N$, *and* $N(\beta_{in}, \sigma_{in}^2), i = 1, 2$, *we have*

$\Delta(N(\beta_{1n}, \sigma_{2n}^2), N(\beta_{2n}, \sigma_{2n}^2); A_n^N)$

$$\begin{aligned}
= \; & (\alpha_n - \alpha_{n+1})(\beta_{1n} - \beta_{2n})\mu \\
& + \int_\Omega \int_{-\infty}^{+\infty} \Delta^+ \left(N\left(\frac{\sigma_0^2 \beta_{1n} + \sigma_{1n}^2 x_1 y_1}{\sigma_0^2 + \sigma_{1n}^2 x_1^2}, \frac{\sigma_0^2 \sigma_{1n}^2}{\sigma_0^2 + \sigma_{1n}^2 x_1^2} \right), N(\beta_{2n}, \sigma_{2n}^2); A_{n+1}^N \right) \\
& \times g_1(y_1 |\, x_1) f(x_1) dx_1 dy_1 \\
& - \int_\Omega \int_{-\infty}^{+\infty} \Delta^- \left(N(\beta_{1n}, \sigma_{1n}^2), N\left(\frac{\sigma_0^2 \beta_{2n} + \sigma_{2n}^2 x_2 y_2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2}, \frac{\sigma_0^2 \sigma_{2n}^2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2} \right); A_{n+1}^N \right) \\
& \times g_2(y_2 |\, x_2) f(x_2) dx_2 dy_2.
\end{aligned}$$

Proof. Using equations $V = V^{(2)} + \Delta^+$ and $V = V^{(1)} + \Delta^-$, the advantage function is expanded as

$$
\begin{aligned}
&\Delta(N(\beta_{1n}, \sigma_{2n}^2), N(\beta_{2n}, \sigma_{2n}^2); A_n^N) \\
&= \alpha_n \beta_{1n} \mu + \int_\Omega \int_{-\infty}^{+\infty} V\left(N\left(\frac{\sigma_0^2 \beta_{1n} + \sigma_{1n}^2 x_1 y_1}{\sigma_0^2 + \sigma_{1n}^2 x_1^2}, \frac{\sigma_0^2 \sigma_{1n}^2}{\sigma_0^2 + \sigma_{1n}^2 x_1^2}\right), N(\beta_{2n}, \sigma_{2n}^2); A_{n+1}^N\right) \\
&\quad \times g_1(y_1|\ x_1) f(x_1) dx_1 dy_1 - \alpha_n \beta_{2n} \mu \\
&\quad - \int_\Omega \int_{-\infty}^{+\infty} V\left(N(\beta_{1n}, \sigma_{1n}^2), N\left(\frac{\sigma_0^2 \beta_{2n} + \sigma_{2n}^2 x_2 y_2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2}, \frac{\sigma_0^2 \sigma_{2n}^2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2}\right); A_{n+1}^N\right) \\
&\quad \times g_2(y_2|\ x_2) f(x_2) dx_2 dy_2 \\
&= \alpha_n \beta_{1n} \mu + \int_\Omega \int_{-\infty}^{+\infty} V^{(2)}\left(N\left(\frac{\sigma_0^2 \beta_{1n} + \sigma_{1n}^2 x_1 y_1}{\sigma_0^2 + \sigma_{1n}^2 x_1^2}, \frac{\sigma_0^2 \sigma_{1n}^2}{\sigma_0^2 + \sigma_{1n}^2 x_1^2}\right), N(\beta_{2n}, \sigma_{2n}^2); A_{n+1}^N\right) \\
&\quad \times g_1(y_1|\ x_1) f(x_1) dx_1 dy_1 \\
&\quad + \int_\Omega \int_{-\infty}^{+\infty} \Delta^+\left(N\left(\frac{\sigma_0^2 \beta_{1n} + \sigma_{1n}^2 x_1 y_1}{\sigma_0^2 + \sigma_{1n}^2 x_1^2}, \frac{\sigma_0^2 \sigma_{1n}^2}{\sigma_0^2 + \sigma_{1n}^2 x_1^2}\right), N(\beta_{2n}, \sigma_{2n}^2); A_{n+1}^N\right) \\
&\quad \times g_1(y_1|\ x_1) f(x_1) dx_1 dy_1 - \alpha_n \beta_{2n} \mu \\
&\quad - \int_\Omega \int_{-\infty}^{+\infty} V^{(1)}\left(N(\beta_{1n}, \sigma_{1n}^2), N\left(\frac{\sigma_0^2 \beta_{2n} + \sigma_{2n}^2 x_2 y_2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2}, \frac{\sigma_0^2 \sigma_{2n}^2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2}\right); A_{n+1}^N\right) \\
&\quad \times g_2(y_2|\ x_2) f(x_2) dx_2 dy_2 \\
&\quad - \int_\Omega \int_{-\infty}^{+\infty} \Delta^-\left(N(\beta_{1n}, \sigma_{1n}^2), N\left(\frac{\sigma_0^2 \beta_{2n} + \sigma_{2n}^2 x_2 y_2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2}, \frac{\sigma_0^2 \sigma_{2n}^2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2}\right); A_{n+1}^N\right) \\
&\quad \times g_2(y_2|\ x_2) f(x_2) dx_2 dy_2 \\
&= \alpha_n \beta_{1n} \mu + \alpha_{n+1} \beta_{2n} \mu \\
&\quad + \int_\Omega \int_{-\infty}^{+\infty} \int_\Omega \int_{-\infty}^{+\infty} V\left(N\left(\frac{\sigma_0^2 \beta_{1n} + \sigma_{1n}^2 x_1 y_1}{\sigma_0^2 + \sigma_{1n}^2 x_1^2}, \frac{\sigma_0^2 \sigma_{1n}^2}{\sigma_0^2 + \sigma_{1n}^2 x_1^2}\right),\right. \\
&\quad \left. N\left(\frac{\sigma_0^2 \beta_{2n} + \sigma_{2n}^2 x_2 y_2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2}, \frac{\sigma_0^2 \sigma_{2n}^2}{\sigma_0^2 + \sigma_{2n}^2 x_2^2}\right); A_{n+2}^N\right) g_1(y_1|\ x_1) f(x_1) dx_1 dy_1
\end{aligned}
$$

$$\times g_2(y_2|\ x_2)f(x_2)dx_2dy_2$$

$$+\int_\Omega\int_{-\infty}^{+\infty}\Delta^+\left(N\left(\frac{\sigma_0^2\beta_{1n}+\sigma_{1n}^2x_1y_1}{\sigma_0^2+\sigma_{1n}^2x_1^2},\frac{\sigma_0^2\sigma_{1n}^2}{\sigma_0^2+\sigma_{1n}^2x_1^2}\right),N(\beta_{2n},\sigma_{2n}^2);A_{n+1}^N\right)$$

$$\times g_1(y_1|\ x_1)f(x_1)dx_1dy_1-\alpha_n\beta_{2n}\mu-\alpha_{n+1}\beta_{1n}\mu$$

$$-\int_\Omega\int_{-\infty}^{+\infty}\int_\Omega\int_{-\infty}^{+\infty}V\left(N\left(\frac{\sigma_0^2\beta_{1n}+\sigma_{1n}^2x_1y_1}{\sigma_0^2+\sigma_{1n}^2x_1^2},\frac{\sigma_0^2\sigma_{1n}^2}{\sigma_0^2+\sigma_{1n}^2x_1^2}\right),\right.$$

$$\left.N\left(\frac{\sigma_0^2\beta_{2n}+\sigma_{2n}^2x_2y_2}{\sigma_0^2+\sigma_{2n}^2x_2^2},\frac{\sigma_0^2\sigma_{2n}^2}{\sigma_0^2+\sigma_{2n}^2x_2^2}\right);A_{n+2}^N\right)g_2(y_2|\ x_2)f(x_2)dx_2dy_2$$

$$\times g_1(y_1|\ x_1)f(x_1)dx_1dy_1$$

$$-\int_\Omega\int_{-\infty}^{+\infty}\Delta^-\left(N(\beta_{1n},\sigma_{1n}^2),N\left(\frac{\sigma_0^2\beta_{2n}+\sigma_{2n}^2x_2y_2}{\sigma_0^2+\sigma_{2n}^2x_2^2},\frac{\sigma_0^2\sigma_{2n}^2}{\sigma_0^2+\sigma_{2n}^2x_2^2}\right);A_{n+1}^N\right)$$

$$\times g_2(y_2|\ x_2)f(x_2)dx_2dy_2. \tag{3.2.3}$$

This lemma is completed after canceling the two forth integrals in (3.2.3) by changing the order of integration. □

Theorem 3.2.2. *Let $A_1^{N+1}=(1,\dots,1,0,\dots)$ be a uniform discount sequence. For given β_i and $\sigma_i^2, i=1,2$, if $\Delta(N(\beta_1,\sigma_1^2),N(\beta_2,\sigma_2^2);A_1^{N+1})>0$, then there exist some x^* and y^* such that*

$$\Delta\left(N\left(\frac{\sigma_0^2\beta_1+\sigma_1^2x^*y^*}{\sigma_0^2+\sigma_1^2x^{*2}},\frac{\sigma_0^2\sigma_1^2}{\sigma_0^2+\sigma_1^2x^{*2}}\right),N((\beta_2,\sigma_2^2),A_1^N)\right)>0.$$

Proof. From Lemma 3.2.1 we obtain

$$\Delta(N(\beta_1,\sigma_1^2),N(\beta_2,\sigma_2^2);A_1^{N+1})$$

$$=\int_\Omega\int_{-\infty}^{+\infty}\Delta^+\left(N\left(\frac{\sigma_0^2\beta_1+\sigma_1^2x_1y_1}{\sigma_0^2+\sigma_1^2x_1^2},\frac{\sigma_0^2\sigma_1^2}{\sigma_0^2+\sigma_1^2x_1^2}\right),N(\beta_2,\sigma_2^2);A_1^N\right)$$

$$\times g_1(y_1|\ x_1)f(x_1)dx_1dy_1$$

$$-\int_{\Omega}\int_{-\infty}^{+\infty}\Delta^{-}\left(N(\beta_1,\sigma_1^2),N\left(\frac{\sigma_0^2\beta_2+\sigma_2^2x_2y_2}{\sigma_0^2+\sigma_2^2x_2^2},\frac{\sigma_0^2\sigma_2^2}{\sigma_0^2+\sigma_2^2x_2^2}\right);A_1^N\right)$$
$$\times g_2(y_2|\ x_2)f(x_2)dx_2dy_2$$

If no such x^* and y^* exist, then

$$\Delta^{+}\left(N\left(\frac{\sigma_0^2\beta_1+\sigma_1^2x_1y_1}{\sigma_0^2+\sigma_1^2x_1^2},\frac{\sigma_0^2\sigma_1^2}{\sigma_0^2+\sigma_1^2x_1^2}\right),N(\beta_2,\sigma_2^2);A_1^N\right)=0$$

for all x and y. Therefore $\Delta(N(\beta_1,\sigma_1^2),N(\beta_2,\sigma_2^2);A_1^{N+1})\leq 0$, which is a contradiction.

It is evident to prove that $V(N(\beta_1,\sigma_1^2),N(\beta_2,\sigma_2^2);A_n^{N+1})$ is increasing in β_1. In view of Theorem 3.2.2, if treatment 1 is optimal at one stage, then it should be selected again as long as y goes beyond a critical value y^* given x^*. Similar results hold for treatment 2.

□

Chapter 4

Linear Regression Model with unknown σ^2

4.1 One-armed Bandit with a Covariate

We will investigate a more complicated one-armed bandit model in this section where both the parameters β and σ^2 of the new treatment are unknown. The new treatment is defined as

$$Y_{1i} = \beta x_i + \varepsilon_i, i = 1, 2, \ldots, \tag{4.1.1}$$

where ε_i are independent Gaussian random variables with mean 0 and unknown variance σ^2. The responses from the standard treatment have constant expectation value λx_i as before if the covariate x_i are given for each time i.

4.1.1 Bayesian Method and Optimal Selection

Let us assume the new treatment is selected k times at the decision times $n_1, n_2, \ldots, n_k$ by the time n, and the observations from this treatment are characterized as $\mathscr{O}_n = \{(x_i, y_{1i}), i = n_1, n_2, \ldots, n_k\}$. Write $\gamma_n = x_{n_1}^2 + x_{n_2}^2 + \ldots + x_{n_k}^2$, $\tau_n = y_{1n_1}^2 + y_{1n_2}^2 + \ldots + y_{1n_k}^2$, and $\eta_n = x_{n_1}y_{1n_1} + x_{n_2}y_{1n_2} + \ldots + x_{n_k}y_{1n_k}$, then γ_n, τ_n and η_n are the sufficient statistics of β and r.

Now the likelihood function of β and r becomes

$$\ell(\beta, \sigma^2 |\ \mathscr{O}_n) \propto r^{k/2} \exp\left[-\frac{1}{2}\left((n-1)\hat{\sigma_n^2} + (\beta - \hat{\beta}_n)^2\gamma_n\right)/2\sigma^2\right],$$

where the OLSE $\hat{\beta}_n$ and $\hat{\sigma_n^2}$ can be calculated from

$$\hat{\beta}_n = \frac{\eta_n}{\gamma_n},$$

and

$$(k-1)\hat{\sigma_n^2} = \frac{\tau_n\gamma_n - \eta_n^2}{\gamma_n}.$$

Further assume the conjugate prior for (β, r) to be

$$g(\beta, r) = g(\beta |\ r)g(r),$$

where $g(\beta |\ r)$ is $N(\beta_0, mr)$ and $g(r)$ is Gamma density $G(u, v)$. Thus the posterior distribution can be derived as

$$g(\beta, r |\ \mathscr{O}_n) \propto \left[r^{1/2}\exp\left(-\frac{r}{2}(m+\gamma_n)(\beta - \beta_n)^2\right)\right]\left[r^{\frac{k}{2}+u-1}\exp(-rv_n)\right],$$

where

$$\beta_n = \frac{m\beta_0 + \eta_n}{m + \gamma_n},$$

and

$$v_n^2 = v + \frac{\gamma_n \tau_n - \eta_n^2}{2\gamma_n} + \frac{m(\eta_n - \beta_0\gamma_n)}{2\gamma_n(m + \gamma_n)}.$$

Hence, the joint posterior distribution of β and r is a product of the conditional posterior distribution $N(\beta_n, r(m + \gamma_n))$ of β given r, and the marginal posterior distribution $G\left(\frac{k+2u}{2}, v_n\right)$ of r. The marginal posterior distribution of β can now be obtained by integrating out r from $g(\beta, r|\ \mathscr{O}_n)$, so that we have

$$g(\beta|\ \mathscr{O}_n)) \propto \left[1 + (\beta - \beta_n)^2(m + \gamma_n)/2v_n\right]^{-\left(\frac{n+2u+1}{2}\right)}$$

which is a kernel of a 3-parameter t-distribution with $(k + 2u)$ degrees of freedom, location parameter β_n, and scale parameter $\frac{(k+2u)}{2v_n}(m + \gamma_n)$. Denote $k + 2u$ as k_1 for convenience, then the sequence

$$\left\{t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m + \gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right)\right\}, n = 1, 2, \ldots$$

constitutes a process of information gathering and can be considered as the states of an underlying Markov process of our one-armed model. Furthermore, the predictive density of a future observation Y_n at stage n, given $X_n = x$ and $\mathscr{O}_n$, is

$$g(y|\ \mathscr{O}_n, x) \propto \left[1 + \left(\frac{m + \gamma_n}{m + \gamma_n + x^2}\right)\left(\frac{k_1}{2v_n}\right)\left(\frac{y - \beta_n x}{k_1}\right)\right]^{-\left(\frac{k_1+1}{2}\right)}.$$

Thus, the predictive distribution of Y_n is a 3-parameter t-distribution with k_1 degrees of freedom, location parameter $\beta_n x$, and scale parameter $\left(\frac{m+\gamma_n}{m+\gamma_n+x^2}\right)\left(\frac{k_1}{2v_n}\right)$.

By the principle of backward induction, given $t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right)$ and $G\left(\frac{k_1}{2}, v_n\right)$, the optimal values from treatments 1 and 2 at stage n, respectively, are given as

$$\begin{aligned} & V^{(1)}\left(t\left(k_1, \beta_n, \tfrac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\tfrac{k_1}{2}, v_n\right), \lambda; A_n^N\right) \\ &= \alpha_n\beta_n\mu + E\left[V\left(t\left(k_1+1, \beta_{n+1}, \frac{k_1+1}{2v_{n+1}}(m+\gamma_n+X_n^2)\right),\right.\right. \\ &\quad \left.\left. G\left(\frac{k_1+1}{2}, v_{n+1}\right), \lambda; A_{n+1}^N\right) \mid t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right)\right)\right], \end{aligned} \tag{4.1.2}$$

where

$$\beta_{n+1} = \frac{m\beta_0+\eta_{n+1}}{m+\gamma_{n+1}} = \frac{m\beta_0+\eta_n+X_nY_n}{m+\gamma_n+X_n^2} \triangleq \beta_{n+1}(X_n, Y_n),$$

$$\begin{aligned} v_{n+1} &= v + \frac{\gamma_{n+1}\tau_{n+1}-\eta_{n+1}^2}{2\gamma_{n+1}} + \frac{m(\eta_{n+1}-\beta_0\gamma_{n+1})^2}{2\gamma_{n+1}(m+\gamma_{n+1})} \\ &= v + \frac{(\gamma_n+X_n^2)(\tau_n+Y_n^2)-(\eta_n+X_nY_n)^2}{2(\gamma_n+X_n^2)} \\ &\quad + \frac{m(\eta_n+X_nY_n-\beta_0(\gamma_n+X_n^2))^2}{2(\gamma_n+X_n^2)(m+\gamma_n+X_n^2)} \triangleq v_{n+1}(X_n, Y_n), \end{aligned}$$

and

$$\begin{aligned} & V^{(2)}\left(t\left(k_1, \beta_n, \tfrac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\tfrac{k_1}{2}, v_n\right), \lambda; A_n^N\right) \\ &= \alpha_n\lambda\mu + V(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), \lambda; A_n^{N+1}). \end{aligned} \tag{4.1.3}$$

By writing β_{n+1} and v_{n+1} as functions of the random variables X_n and Y_n, equation (4.1.2) becomes

$$V^{(1)}\left(t\left(k_1,\beta_n,\tfrac{k_1}{2v_n}(m+\gamma_n)\right),G\left(\tfrac{k_1}{2},v_n\right),\lambda;A_n^N\right)$$

$$= \alpha_n\beta_n\mu+\int_\Omega\int_{-\infty}^{+\infty}V\left(t\left(k_1+1,\beta_{n+1}(x,y),\frac{k_1+1}{2v_{n+1}(x,y)}(m+\gamma_n+x^2)\right),\right.$$
$$\left.G\left(\frac{k_1+1}{2},v_{n+1}(x,y)\right),\lambda;A_{n+1}^N\right)g(y|\ x)f(x)dxdy. \tag{4.1.4}$$

4.1.2 Main Results: Optimal Strategy and Monotonicity

The main results in this section concern optimal strategies.

Lemma 4.1.1. *At each stage $n, n=1,\ldots,N$, all functions*

$V\left(t\left(k_1,\beta_n,\tfrac{k_1}{2v_n}(m+\gamma_n)\right),G\left(\tfrac{k_1}{2},v_n\right),\lambda;A_n^N\right)$ *and*

$V^{(i)}\left(t\left(k_1,\beta_n,\tfrac{k_1}{2v_n}(m+\gamma_n)\right),G\left(\tfrac{k_1}{2},v_n\right),\lambda;A_n^N\right), i=1,2$ *are continuous and increasing in λ. Therefore the advantage function*

$\Delta\left(t\left(k_1,\beta_n,\tfrac{k_1}{2v_n}(m+\gamma_n)\right),G\left(\tfrac{k_1}{2},v_n\right),\lambda;A_n^N\right)$ *is also continuous in λ.*

Proof. This result can be clearly proved by the method of induction on N. □

Lemma 4.1.2. *At each stage $n, n=1,\ldots,N$, let $A_n^N=(\alpha_n,\alpha_{n+1},\ldots,\alpha_N,0,\ldots)$ denote the truncated discount sequence. Then the function*

$$\Delta\left(t\left(k_1,\beta_n,\frac{k_1}{2v_n}(m+\gamma_n)\right),G\left(\frac{k_1}{2},v_n\right),\lambda;A_n^N\right)$$

is nonincreasing in λ.

Proof. The proof is by induction. The conclusion of this lemma clearly holds for $N=n$. Assume it holds for horizon N and fix the horizon $N+1$, then equations

(4.1.3) and (4.1.4) give

$$
\begin{aligned}
&\Delta\left(t\left(k_1,\beta_n,\tfrac{k_1}{2v_n}(m+\gamma_n)\right),G\left(\tfrac{k_1}{2},v_n\right),\lambda;A_n^{N+1}\right)\\
&= \alpha_n\beta_n\mu+\int_\Omega\int_{-\infty}^{+\infty}V\left(t\left(k_1+1,\beta_{n+1}(x,y),\frac{k_1+1}{2v_{n+1}(x,y)}(m+\gamma_n+x^2)\right),\right.\\
&\quad \left.G\left(\frac{k_1+1}{2},v_{n+1}(x,y)\right),\lambda;A_{n+1}^{N+1}\right)g(y|\ x)f(x)dxdy\\
&\quad -\alpha_n\lambda\mu-V(t\left(k_1,\beta_n,\frac{k_1}{2v_n}(m+\gamma_n)\right),G\left(\frac{k_1}{2},v_n\right),\lambda;A_{n+1}^{N+1})\\
&= \alpha_n\beta_n\mu+\int_\Omega\int_{-\infty}^{+\infty}V^{(2)}\left(t\left(k_1+1,\beta_{n+1}(x,y),\frac{k_1+1}{2v_{n+1}(x,y)}(m+\gamma_n+x^2)\right),\right.\\
&\quad \left.G\left(\frac{k_1+1}{2},v_{n+1}(x,y)\right),\lambda;A_{n+1}^{N+1}\right)g(y|\ x)f(x)dxdy\\
&\quad +\int_\Omega\int_{-\infty}^{+\infty}\Delta^+\left(t\left(k_1+1,\beta_{n+1}(x,y),\frac{k_1+1}{2v_{n+1}(x,y)}(m+\gamma_n+x^2)\right),\right.\\
&\quad \left.G\left(\frac{k_1+1}{2},v_{n+1}(x,y)\right),\lambda;A_{n+1}^{N+1}\right)g(y|\ x)f(x)dxdy\\
&\quad -\alpha_n\lambda\mu-V^{(1)}(t\left(k_1,\beta_n,\frac{k_1}{2v_n}(m+\gamma_n)\right),G\left(\frac{k_1}{2},v_n\right),\lambda;A_{n+1}^{N+1})\\
&\quad -\Delta^-(t\left(k_1,\beta_n,\frac{k_1}{2v_n}(m+\gamma_n)\right),G\left(\frac{k_1}{2},v_n\right),\lambda;A_{n+1}^{N+1})\\
&= \alpha_n\beta_n\mu+\alpha_{n+1}\lambda\mu\\
&\quad +\int_\Omega\int_{-\infty}^{+\infty}V\left(t\left(k_1+1,\beta_{n+1}(x,y),\frac{k_1+1}{2v_{n+1}(x,y)}(m+\gamma_n+x^2)\right),\right.\\
&\quad \left.G\left(\frac{k_1+1}{2},v_{n+1}(x,y)\right),\lambda;A_{n+2}^{N+1}\right)g(y|\ x)f(x)dxdy\\
&\quad +\int_\Omega\int_{-\infty}^{+\infty}\Delta^+\left(t\left(k_1+1,\beta_{n+1}(x,y),\frac{k_1+1}{2v_{n+1}(x,y)}(m+\gamma_n+x^2)\right),\right.\\
&\quad \left.G\left(\frac{k_1+1}{2},v_{n+1}(x,y)\right),\lambda;A_{n+1}^{N+1}\right)g(y|\ x)f(x)dxdy\\
&\quad -\alpha_n\lambda\mu-\alpha_{n+1}\beta_n\mu\\
&\quad -\int_\Omega\int_{-\infty}^{+\infty}V\left(t\left(k_1+1,\beta_{n+1}(x,y),\frac{k_1+1}{2v_{n+1}(x,y)}(m+\gamma_n+x^2)\right),\right.
\end{aligned}
$$

$$
\begin{aligned}
& G\left(\frac{k_1+1}{2}, v_{n+1}(x,y)\right), \lambda; A_{n+2}^{N+1}\Big) g(y|\ x) f(x) dxdy \\
& -\Delta^-(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), \lambda; A_{n+1}^{N+1}) \\
= \ & (\alpha_n - \alpha_{n+1})\mu(\beta_n - \lambda) \\
& + \int_\Omega \int_{-\infty}^{+\infty} \Delta^+\left(t\left(k_1+1, \beta_{n+1}(x,y), \frac{k_1+1}{2v_{n+1}(x,y)}(m+\gamma_n+x^2)\right)\right), \\
& G\left(\frac{k_1+1}{2}, v_{n+1}(x,y)\right), \lambda; A_{n+1}^{N+1}\Big) g(y|\ x) f(x) dxdy \\
& -\Delta^-(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), \lambda; A_{n+1}^{N+1}). \qquad (4.1.5)
\end{aligned}
$$

The first term in equation (4.1.5) is nonincreasing in λ since $\alpha_n \geq \alpha_{n+1}$ and $\mu \geq 0$. Moreover, the induction hypothesis implies that

$$\Delta^+\left(t\left(k_1+1, \beta_{n+1}, \frac{k_1+1}{2v_{n+1}(x,y)}(m+\gamma_n+x^2)\right), G\left(\frac{k_1+1}{2}, v_{n+1}(x,y)\right), \lambda; A_{n+1}^{N+1}\right)$$

is nonincreasing and

$$\Delta^-\left(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), \lambda; A_{n+1}^{N+1}\right)$$

is nondecreasing in λ. Therefore,

$$\Delta\left(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), \lambda; A_n^{N+1}\right)$$

is nonincreasing in λ. □

Theorem 4.1.3. *At each stage $n, n = 1, \ldots, N$ for any β_n and v_n, there exists an index value $\lambda^* = \lambda^*(\beta_n, v_n, A_n^N)$ such that*

$$\Delta\left(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), \lambda; A_n^{N+1}\right) = 0.$$

The set of all such index values forms an interval. Moreover, the new treatment is optimal if and only if $\lambda \leq \lambda^*(\beta_n, v_n, A_n^N)$ *while the standard one is optimal if and only if* $\lambda \geq \lambda^*(\beta_n, v_n, A_n^N)$.

Proof. The existence of the index value is obvious from the continuity and monotonicity of

$$\Delta\left(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), \lambda; A_n^{N+1}\right)$$

in λ, and the facts that

$$\Delta\left(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), 0; A_n^{N+1}\right) > 0,$$

and

$$\lim_{\lambda\to\infty} \Delta\left(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), \lambda; A_n^{N+1}\right) < 0.$$

Furthermore, the set of all such index values forms an interval since

$$\Delta\left(t\left(k_1, \beta_n, \frac{k_1}{2v_n}(m+\gamma_n)\right), G\left(\frac{k_1}{2}, v_n\right), \lambda; A_n^{N+1}\right)$$

is nonincreasing in λ. Finally, the optimal decision is made based on λ^* according to the advantage function Δ. □

Lemma 4.1.4. *Assume that* $A_1^N = (1, \alpha, \alpha^2, \ldots, \alpha^{N-1}, 0, \ldots), 0 < \alpha < 1$. *For any given* β *and* v, *if*

$$\Delta\left(t\left(k_1, \beta, \frac{k_1}{2v}(m+\gamma)\right), G\left(\frac{k_1}{2}, v\right), \lambda; A_1^N\right) = 0,$$

then

$$\Delta\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda;A_1^{N+1}\right)\geq 0.$$

Proof. The equation

$$\Delta\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda;A_1^{N}\right)=0$$

implies

$$\begin{aligned}\beta\mu-\lambda\mu &= \alpha V\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda;A_1^{N-1}\right)\\ &\quad -\alpha\int_\Omega\int_{-\infty}^{+\infty}V\left(t\left(k_1+1,\beta(x,y),\frac{k_1+1}{2v(x,y)}(m+\gamma+x^2)\right),\right.\\ &\quad \left.G\left(\frac{k_1+1}{2},v(x,y)\right),\lambda;A_1^N\right)g(y|x)f(x)dxdy \qquad (4.1.6)\end{aligned}$$

and

$$\begin{aligned}&V\left(t\left(k_1,\beta,\tfrac{k_1}{2v}(m+\gamma)\right),G\left(\tfrac{k_1}{2},v\right),\lambda;A_1^N\right)\\ &\quad=\lambda\mu+V\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda;A_1^{N-1}\right). \qquad (4.1.7)\end{aligned}$$

It follows from equations (4.1.6) and (4.1.7) that

$$\begin{aligned}&\Delta\left(t\left(k_1,\beta,\tfrac{k_1}{2v}(m+\gamma)\right),G\left(\tfrac{k_1}{2},v\right),\lambda;A_1^{N+1}\right)\\ &= \alpha\int_\Omega\int_{-\infty}^{+\infty}\left[V\left(t\left(k_1+1,\beta(x,y),\frac{k_1+1}{2v(x,y)}(m+\gamma+x^2)\right),\right.\right.\\ &\quad G\left(\frac{k_1+1}{2},v(x,y)\right),\lambda;A_1^N\Big)-V\left(t\left(k_1+1,\beta(x,y),\frac{k_1+1}{2v(x,y)}(m+\gamma+x^2)\right),\\ &\quad \left.\left.G\left(\frac{k_1+1}{2},v(x,y)\right),\lambda;A_1^{N-1}\right)\right]g(y|x)f(x)dxdy\\ &\quad -\alpha\lambda\mu+\alpha(1-\alpha)V\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda;A_1^N\right). \qquad (4.1.8)\end{aligned}$$

Let π^* be the optimal strategy for

$$V\left(t\left(k_1+1,\beta(x,y),\frac{k_1+1}{2v(x,y)}(m+\gamma+x^2)\right),G\left(\frac{k_1+1}{2},v(x,y)\right),\lambda;A_1^{N-1}\right),$$

we define a strategy for

$$V\left(t\left(k_1+1,\beta(x,y),\frac{k_1+1}{2v(x,y)}(m+\gamma+x^2)\right),G\left(\frac{k_1+1}{2},v(x,y)\right),\lambda;A_1^{N}\right)$$

that allocates π^* during the first $N-1$ decision times and selects the standard treatment at time N. Thus,

$$\begin{aligned}&V\left(t\left(k_1+1,\beta(x,y),\frac{k_1+1}{2v(x,y)}(m+\gamma+x^2)\right),G\left(\frac{k_1+1}{2},v(x,y)\right),\lambda;A_1^{N}\right)\\&-V\left(t\left(k_1+1,\beta(x,y),\frac{k_1+1}{2v(x,y)}(m+\gamma+x^2)\right),G\left(\frac{k_1+1}{2},v(x,y)\right),\lambda;\\&A_1^{N-1}\right)\geq\alpha^{N-1}\lambda\mu\end{aligned}$$

for all x and y. Besides, if we choose the standard treatment all the time, then

$$V\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k+2u}{2},v\right),\lambda;A_1^{N}\right)\geq(1+\alpha+\ldots+\alpha^{N-2})\lambda\mu.$$

Hence, it follows from the above inequalities and equation (4.1.8) that

$$\begin{aligned}&\Delta\left(t\left(k_1,\beta,\tfrac{k_1}{2v}(m+\gamma)\right),G\left(\tfrac{k_1}{2},v\right),\lambda;A_1^{N+1}\right)\\&\geq\alpha^N\lambda\mu-\alpha\lambda\mu+\alpha(1-\alpha)((1+\alpha+\ldots+\alpha^{N-2})\lambda\mu)=0.\end{aligned}$$

□

Theorem 4.1.5. *For fixed β and v but changing $N=1,2,\ldots,$ let $\lambda^*(\beta,v,A_1^N)$ be the index value such that*

$$\Delta\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda^*(\beta,v,A_1^N);A_1^N\right)=0,$$

where $A_1^N = (1, \alpha, \ldots, \alpha^{N-1}, 0, \ldots)$. *Then*

$$\beta = \lambda^*(\beta, v, A_1^1) \leq \lambda^*(\beta, v, A_1^2) \leq \ldots \leq \lambda^*(\beta, v, A_1^N) \leq \ldots.$$

Moreover, the limit $\lambda^*(\beta, v) = \lim_{n\to\infty} \lambda^*(\beta, v, A_1^N)$ *exists such that* $\beta < \lambda^*(\beta, v) < \infty$ *and*

$$\Delta\left(t\left(k_1, \beta, \frac{k_1}{2v}(m+\gamma)\right), G\left(\frac{k_1}{2}, v\right), \lambda^*(\beta, v); A\right) = 0,$$

where $A = (1, \alpha, \alpha^2, \ldots)$.

Proof. Based on the monotonicity of

$$\Delta\left(t\left(k_1, \beta, \frac{k_1}{2v}(m+\gamma)\right), G\left(\frac{k_1}{2}, v\right), \lambda; A_1^{N+1}\right)$$

in λ and the results

$$\Delta\left(t\left(k_1, \beta, \frac{k_1}{2v}(m+\gamma)\right), G\left(\frac{k_1}{2}, v\right), \lambda^*(\beta, v, A_1^{N+1}); A_1^{N+1}\right) = 0$$

and

$$\Delta\left(t\left(k_1, \beta, \frac{k_1}{2v}(m+\gamma)\right), G\left(\frac{k_1}{2}, v\right), \lambda^*(\beta, v, A_1^N); A_1^{N+1}\right) \geq 0,$$

we conclude that $\lambda^*(\beta, v, A_1^N) \leq \lambda^*(\beta, \sigma^2, A_1^{N+1})$ for $N = 1, 2, \ldots$.

Moreover, the limit of the nondecreasing sequence $\lambda^*(\beta, v, A_1^N)$ exists. Let $\lambda^*(\beta, v) = \lim_{N\to\infty} \lambda^*(\beta, v, A_1^N)$, then $\lambda^*(\beta, v)$ satisfies the equation

$$\Delta\left(t\left(k_1, \beta, \frac{k_1}{2v}(m+\gamma)\right), G\left(\frac{k_1}{2}, v\right), \lambda^*(\beta, v); A\right) = 0$$

since

$$\Delta\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda;A_1^{N+1}\right)$$

is continuous in N. If $\lambda^*(\beta,v)=\infty$, then

$$V\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda^*(\beta,v);A\right)=\infty,$$

which contradicts the finiteness of the optimal value function.

We now turn to prove $\beta<\lambda^*(\beta,v,A_1^2)$ by contradiction. Suppose that $\beta=\lambda^*(\beta,v,A_1^2)$, then

$$\begin{aligned}
0 &= \Delta\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda^*(\beta,v,A_1^2);A_1^2\right)\\
&= \beta\mu+\alpha\int_\Omega\int_{-\infty}^{+\infty}V\left(t\left(k_1+1,\beta(x,y),\frac{k_1+1}{2v(x,y)}(m+\gamma+x^2)\right),\right.\\
&\quad \left.G\left(\frac{k_1+1}{2},v(x,y)\right),\lambda^*(\beta,v,A_1^2);A_1^1\right)g(y|\,x)f(x)dxdy\\
&\quad -\lambda^*(\beta,v,A_1^2)\mu-V\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),G\left(\frac{k_1}{2},v\right),\lambda^*(\beta,v,A_1^2);A_1^1\right)\\
&= \alpha\int_\Omega\int_{-\infty}^{+\infty}\left[V\left(t\left(k_1+1,\beta(x,y),\frac{k_1+1}{2v(x,y)}(m+\gamma+x^2)\right),\right.\right.\\
&\quad G\left(\frac{k_1+1}{2},v(x,y)\right),\lambda^*(\beta,v,A_1^2);A_1^1\Big)-V\left(t\left(k_1,\beta,\frac{k_1}{2v}(m+\gamma)\right),\\
&\quad \left.\left.G\left(\frac{k_1}{2},v\right),\lambda^*(\beta,v,A_1^2);A_1^1\right)\right]g(y|\,x)f(x)dxdy\\
&= \alpha\int_\Omega\int_{-\infty}^{+\infty}\left[\max\{\frac{(m+\gamma)\beta+xy}{m+\gamma+x^2}\mu,\beta\mu\}-\beta\mu\right]g(y|\,x)f(x)dxdy. \qquad (4.1.9)
\end{aligned}$$

If x is given, the predictive distribution of Y is a 3-parameter t-distribution with location parameter βx. Let y^* be the smallest y larger than βx, then the right side of equation (4.1.9) is positive, which is a contradiction. □

Corollary 4.1.6. *The myopic strategy is not optimal in general.*

Proof. In our model with unknown β and σ^2, a strategy is myopic for t-distribution

$$t\left(k_1, \beta, \frac{k_1}{2v}(m+\gamma)\right)$$

when the new treatment(standard treatment, respectively) is chosen if and only if $\beta \geq (\leq)\lambda$. This strategy is not optimal unless $N = 1$. Consider a two-stage bandit problem and let λ be such that $\beta < \lambda < \lambda^*(\beta, v, A_1^2)$. The new treatment is uniquely optimal at the first stage while the myopic strategy selects the standard treatment. □

Corollary 4.1.7. *If the standard treatment is uniquely optimal at stage n, then it is optimal for the rest of the decision horizon.*

Proof. If the standard treatment is uniquely optimal at stage n, then $\lambda > \lambda^*(\beta, v, A_1^n)$, which indicates that $\lambda > \lambda^*(\beta, v, A_1^{n-1})$. Clearly, the standard treatment is uniquely optimal again. □

4.2 Two-armed Bandit Model with a Covariate

In this section we assume both treatments are unknown and characterized by regression models with unknown parameters β_i and σ^2, $i = 1, 2$, respectively. Similar to Chapter 2, the conjugate prior for (β_1, r) is

$$g(\beta_1, r) = g(\beta_1 |\ r)g(r),$$

where $g(\beta_1|r)$ is normal density $N(\beta_{10}, mr)$ and $g(r)$ is the Gamma density $G(u, v)$, while the conjugate prior for (β_2, r) is

$$g(\beta_2, r) = g(\beta_2|\ r)g(r),$$

where $g(\beta_2|r)$ is $N(\beta_{20}, mr)$ and $g(r)$ is $G(u, v)$.

Assume that treatment 1 is allocated k times to patients by time n, we similarly define $\gamma_{1n}, \tau_{1n}, \eta_{1n}$ from treatment 1 and $\gamma_{2n}, \tau_{2n}, \eta_{2n}$ from treatment 2. With the above conjugate priors it can be derived that at time n, the marginal posterior distribution of β_1 and r from treatment 1 are the t-distribution

$$t(k + 2u, \beta_{1n}, \frac{k + 2u}{2v_{1n}})$$

and the Gamma distribution

$$G(\frac{k + 2u}{2}, v_{1n}),$$

where

$$\beta_{1n} = \frac{m\beta_{10} + \eta_{1n}}{m + \gamma_{1n}},$$

and

$$v_{1n} = v + \frac{\gamma_{1n}\tau_{1n} - \eta_{1n}^2}{2\gamma_{1n}} + \frac{m(\eta_{1n} - \beta_{10}\gamma_{1n})}{2\gamma_{1n}(m + \gamma_{1n})}.$$

On the other hand, the marginal posterior distribution of β_2 and r from treatment 2 are the t-distribution

$$t(n - 1 - k + 2u, \beta_{2n}, \frac{n - 1 - k + 2u}{2v_{2n}})$$

and the Gamma distribution

$$G(\frac{n-1-k+2u}{2}, v_{2n}),$$

where

$$\beta_{2n} = \frac{m\beta_{20} + \eta_{2n}}{m + \gamma_{2n}},$$

and

$$v_{2n} = v + \frac{\gamma_{2n}\tau_{2n} - \eta_{2n}^2}{2\gamma_{2n}} + \frac{m(\eta_{2n} - \beta_{20}\gamma_{2n})}{2\gamma_{2n}(m + \gamma_{2n})}.$$

In order to simplify calculations and derivations, we denote $k+2u$ and $n-1-k+2u$ as k_1 and k_2, respectively. Then the predictive distribution of a future observation Y_{1n} from treatment 1, given $X_n = x$, is a 3-parameter t-distribution with the density

$$g_1(y|\ x) \propto \left[1 + \left(\frac{m+\gamma_{1n}}{m+\gamma_{1n}+x^2}\right)\left(\frac{k_1}{2v_{1n}}\right)\left(\frac{y-\beta_{1n}x}{k_1}\right)\right]^{-\left(\frac{k_1}{2}\right)}.$$

The predictive distribution of a future observation Y_{2n} from treatment 2, given $X_n = x$, is again a 3-parameter t-distribution with the density

$$g_2(y|\ x) \propto \left[1 + \left(\frac{m+\gamma_{2n}}{m+\gamma_{2n}+x^2}\right)\left(\frac{k_2}{2v_{2n}}\right)\left(\frac{y-\beta_{2n}x}{k_2}\right)\right]^{-\left(\frac{k_2}{2}\right)}.$$

Now we define the worth function

$$W\left(t\left(k_1, \beta_{1n}, \frac{k_1}{2v_{1n}}\right), G\left(\frac{k_1}{2}, v_{1n}\right), t\left(k_2, \beta_{2n}, \frac{k_2}{2v_{2n}}\right), G\left(\frac{k_2}{2}, v_{2n}\right); A_n^N; \pi\right),$$

the optimal value function

$$V\left(t\left(k_1, \beta_{1n}, \frac{k_1}{2v_{1n}}\right), G\left(\frac{k_1}{2}, v_{1n}\right), t\left(k_2, \beta_{2n}, \frac{k_2}{2v_{2n}}\right), G\left(\frac{k_2}{2}, v_{2n}\right); A_n^N\right),$$

the optimal value function for each treatment

$$V^{(i)}\left(t\left(k_1,\beta_{1n},\frac{k_1}{2v_{1n}}\right),G\left(\frac{k_1}{2},v_{1n}\right),t\left(k_2,\beta_{2n},\frac{k_2}{2v_{2n}}\right),G\left(\frac{k_2}{2},v_{2n}\right);A_n^N\right),\ i=1,2$$

and the advantage function

$$\Delta\left(t\left(k_1,\beta_{1n},\frac{k_1}{2v_{1n}}\right),G\left(\frac{k_1}{2},v_{1n}\right),t\left(k_2,\beta_{2n},\frac{k_2}{2v_{2n}}\right),G\left(\frac{k_2}{2},v_{2n}\right);A_n^N\right),$$

at each time $n, n=1,2,\ldots,N$.

In the next context, we give a complete characterization of the optimal strategy. First, the myopic strategy is reasonable on the first stage for the reason that it is necessary to understand the unknown parameters for both treatments. Then a play-the-winner strategy is applied for continuing decisions.

Lemma 4.2.1. *For any truncated discount sequence* $A_n^N=(\alpha_n,\alpha_{n+1},\ldots,\alpha_N,0,\ldots)$, $n=1,2,\ldots,N$, *and* $t(k_i,\beta_{in},\frac{k_i}{2v_{in}}),G(\frac{k_i}{2},v_{in}),i=1,2$, *we have*

$\Delta(t(k_1,\beta_{1n},\frac{k_1}{2v_{1n}}),G(\frac{k_1}{2},v_{1n}),t(k_2,\beta_{2n},\frac{k_2}{2v_{2n}}),G(\frac{k_2}{2},v_{2n});A_n^N)$

$$\begin{aligned}
&=(\alpha_n-\alpha_{n+1})(\beta_{1n}-\beta_{2n})\mu\\
&+\int_\Omega\int_{-\infty}^{+\infty}\Delta^+\left(t\left(k_1+1,\beta_{1(n+1)}(x_1,y_1),\frac{k_1+1}{2v_{1(n+1)}(x_1,y_1)}\right),\right.\\
&\left.G\left(\frac{k_1+1}{2},v_{1(n+1)}(x_1,y_1)\right),t\left(k_2,\beta_{2n},\frac{k_2}{2v_{2n}}\right),G\left(\frac{k_2}{2},v_{2n}\right);A_n^N\right)\\
&\times g_1(y_1|\ x_1)f(x_1)dx_1dy_1\\
&-\int_\Omega\int_{-\infty}^{+\infty}\Delta^-\left(t\left(k_1,\beta_{1n},\frac{k_1}{2v_{1n}}\right),G\left(\frac{k_1}{2},v_{1n}\right),\right.\\
&\left.t\left(k_2+1,\beta_{2(n+1)}(x_2,y_2),\frac{k_2+1}{2v_{2(n+1)}(x_2,y_2)}\right),G\left(\frac{k_2+1}{2},v_{2(n+1)}(x_2,y_2)\right);A_n^N\right)
\end{aligned}$$

$$\times g_2(y_2|\ x_2)f(x_2)dx_2dy_2,$$

where

$$\beta_{i(n+1)}(x_i,y_i)=\frac{m\beta_0+\eta_{in}+x_iy_i}{m+\gamma_{in}+x_i^2}, i=1,2,$$

and

$$\begin{aligned} v_{i(n+1)}(x_i,y_i) &= v+\frac{(\gamma_{in}+x_i^2)(\tau_{in}+y_i^2)-(\eta_{in}+x_iy_i)^2}{2(\gamma_{in}+x_i^2)} \\ &\quad +\frac{m(\eta in+x_iy_i-\beta_0(\gamma_{in}+x_i^2))^2}{2(\gamma_{in}+x_i^2)(m+\gamma_{in}+x_i^2)}, i=1,2. \end{aligned}$$

Proof. Using equations $V=V^{(2)}+\Delta^+$ and $V=V^{(1)}+\Delta^-$, the advantage function is expanded as

$$\begin{aligned} &\Delta(t(k_1,\beta_{1n},\tfrac{k_1}{2v_{1n}}),G(\tfrac{k_1}{2},v_{1n}),t(k_2,\beta_{2n},\tfrac{k_2}{2v_{2n}}),G(\tfrac{k_2}{2},v_{2n});A_n^N) \\ &= \alpha_n\beta_{1n}\mu+\int_\Omega\int_{-\infty}^{+\infty}\Delta^+\left(t\left(k_1+1,\beta_{1(n+1)}(x_1,y_1),\frac{k_1+1}{2v_{1(n+1)}(x_1,y_1)}\right),\right. \\ &\quad \left. G\left(\frac{k_1+1}{2},v_{1(n+1)}(x_1,y_1)\right),t\left(k_2,\beta_{2n},\frac{k_2}{2v_{2n}}\right),G\left(\frac{k_2}{2},v_{2n}\right);A_n^N\right) \\ &\quad \times g_1(y_1|\ x_1)f(x_1)dx_1dy_1 \\ &\quad -\alpha_n\beta_{2n}\mu-\int_\Omega\int_{-\infty}^{+\infty}\Delta^-\left(t\left(k_1,\beta_{1n},\frac{k_1}{2v_{1n}}\right),G\left(\frac{k_1}{2},v_{1n}\right),\right. \\ &\quad \left. t\left(k_2+1,\beta_{2(n+1)}(x_2,y_2),\frac{k_2+1}{2v_{2(n+1)}(x_2,y_2)}\right),G\left(\frac{k_2+1}{2},v_{2(n+1)}(x_2,y_2)\right);A_n^N\right) \\ &\quad g_2(y_2|\ x_2)f(x_2)dx_2dy_2 \\ &= \alpha_n\beta_{1n}\mu+\int_\Omega\int_{-\infty}^{+\infty}V^{(2)}\left(t\left(k_1+1,\beta_{1(n+1)}(x_1,y_1),\frac{k_1+1}{2v_{1(n+1)}(x_1,y_1)}\right),\right. \\ &\quad \left. G\left(\frac{k_1+1}{2},v_{1(n+1)}(x_1,y_1)\right),t\left(k_2,\beta_{2n},\frac{k_2}{2v_{2n}}\right),G\left(\frac{k_2}{2},v_{2n}\right);A_n^N\right) \\ &\quad \times g_1(y_1|\ x_1)f(x_1)dx_1dy_1 \end{aligned}$$

$$
\begin{aligned}
&+\int_{\Omega}\int_{-\infty}^{+\infty}\Delta^{+}\left(t\left(k_1+1,\beta_{1(n+1)}(x_1,y_1),\frac{k_1+1}{2v_{1(n+1)}(x_1,y_1)}\right),\right.\\
&\left.G\left(\frac{k_1+1}{2},v_{1(n+1)}(x_1,y_1)\right),t\left(k_2,\beta_{2n},\frac{k_2}{2v_{2n}}\right),G\left(\frac{k_2}{2},v_{2n}\right);A_n^N\right)\\
&\times g_1(y_1|\ x_1)f(x_1)dx_1dy_1\\
&-\alpha_n\beta_{2n}\mu-\int_{\Omega}\int_{-\infty}^{+\infty}V^{(1)}\left(t\left(k_1,\beta_{1n},\frac{k_1}{2v_{1n}}\right),G\left(\frac{k_1}{2},v_{1n}\right),\right.\\
&\left.t\left(k_2+1,\beta_{2(n+1)}(x_2,y_2),\frac{k_2+1}{2v_{2(n+1)}(x_2,y_2)}\right),G\left(\frac{k_2+1}{2},v_{2(n+1)}(x_2,y_2)\right);A_n^N\right)\\
&g_2(y_2|\ x_2)f(x_2)dx_2dy_2\\
&-\int_{\Omega}\int_{-\infty}^{+\infty}\Delta^{-}\left(t\left(k_1,\beta_{1n},\frac{k_1}{2v_{1n}}\right),G\left(\frac{k_1}{2},v_{1n}\right),\right.\\
&\left.t\left(k_2+1,\beta_{2(n+1)}(x_2,y_2),\frac{k_2+1}{2v_{2(n+1)}(x_2,y_2)}\right),G\left(\frac{k_2+1}{2},v_{2(n+1)}(x_2,y_2)\right);A_n^N\right)\\
&g_2(y_2|\ x_2)f(x_2)dx_2dy_2. \qquad (4.2.1)
\end{aligned}
$$

This lemma is proved after canceling the second part of both $V^{(1)}$ and $V^{(2)}$ in (4.2.1), which are two forth integrals, by changing the order of integration.

□

Theorem 4.2.2. *Let $A_1^{N+1}=(1,\dots,1,0,\dots)$ be a uniform discount sequence. For given β_i and $v_i, i=1,2$, if*

$$\Delta\left(t\left(k_1,\beta_1,\frac{k_1}{2v_1}\right),G\left(\frac{k_1}{2},v_1\right),t\left(k_2,\beta_2,\frac{k_2}{2v_2}\right),G\left(\frac{k_2}{2},v_2\right);A_n^N\right)>0,$$

then there exist some x^* and y^* such that

$$\Delta\left(t\left(k_1+1,\beta_1(x^*,y^*),\frac{k_1+1}{2v_1(x^*,y^*)}\right),G\left(\frac{k_1+1}{2},v_1(x^*,y^*)\right),\right.$$

$$t\left(k_2, \beta_2, \frac{k_2}{2v_2}\right), G\left(\frac{k_2}{2}, v_2\right); A_n^N\Big) > 0,$$

where

$$\beta_1(x^*, y^*) = \frac{m\beta_0 + \eta_1 + x^* y^*}{m + \gamma_1 + x^{*2}},$$

and

$$\begin{aligned} v_1(x^*, y^*) &= v + \frac{(\gamma_1 + x^{*2})(\tau_1 + y^{*2}) - (\eta_1 + x^* y^*)^2}{2(\gamma_1 + x^{*2})} \\ &\quad + \frac{m(\eta_1 + x^* y^* - \beta_0(\gamma_1 + x^{*2}))^2}{2(\gamma_1 + x^{*2})(m + \gamma_1 + x^{*2})}. \end{aligned}$$

Proof. From Lemma 4.2.1 we obtain

$\Delta(t(k_1, \beta_1, \frac{k_1}{2v_1}), G(\frac{k_1}{2}, v_1), t(k_2, \beta_2, \frac{k_2}{2v_2}), G(\frac{k_2}{2}, v_2); A_1^{N+1})$

$$\begin{aligned} = &\int_\Omega \int_{-\infty}^{+\infty} \Delta^+\left(t\left(k_1 + 1, \beta_1(x_1, y_1), \frac{k_1 + 1}{2v_1(x_1, y_1)}\right), G\left(\frac{k_1 + 1}{2}, v_1(x_1, y_1)\right),\right. \\ &\left. t\left(k_2, \beta_2, \frac{k_2}{2v_2}\right), G\left(\frac{k_2}{2}, v_2\right); A_1^N\right) g_1(y_1|\ x_1) f(x_1) dx_1 dy_1 \\ &- \int_\Omega \int_{-\infty}^{+\infty} \Delta^-\left(t\left(k_1, \beta_1, \frac{k_1}{2v_1}\right), G\left(\frac{k_1}{2}, v_1\right), t\left(k_2 + 1, \beta_2(x_2, y_2), \frac{k_2 + 1}{2v_2(x_2, y_2)}\right),\right. \\ &\left. G\left(\frac{k_2 + 1}{2}, v_2(x_2, y_2)\right); A_n^N\right) g_2(y_2|\ x_2) f(x_2) dx_2 dy_2 \end{aligned}$$

If no such x^* and y^* exist, then

$$\begin{aligned} &\Delta^+\left(t\left(k_1 + 1, \beta_1(x_1, y_1), \frac{k_1 + 1}{2v_1(x_1, y_1)}\right), G\left(\frac{k_1 + 1}{2}, v_1(x_1, y_1)\right),\right. \\ &\left. t\left(k_2, \beta_2, \frac{k_2}{2v_2}\right), G\left(\frac{k_2}{2}, v_2\right); A_1^N\right) = 0 \end{aligned}$$

for all x_1 and y_1. Therefore,

$$\Delta\left(t\left(k_1, \beta_1, \frac{k_1}{2v_1}\right), G\left(\frac{k_1}{2}, v_1\right), t\left(k_2, \beta_2, \frac{k_2}{2v_2}\right), G\left(\frac{k_2}{2}, v_2\right); A_1^{N+1}\right) \leq 0,$$

which is a contradiction. □

It is evident to prove that

$$\Delta\left(t\left(k_1,\beta_1,\frac{k_1}{2v_1}\right),G\left(\frac{k_1}{2},v_1\right),t\left(k_2,\beta_2,\frac{k_2}{2v_2}\right),G\left(\frac{k_2}{2},v_2\right);A_n^{N+1}\right)$$

is increasing in β_1. In view of Theorem 4.2.2, if treatment 1 is optimal at one stage, then it should be selected again as long as y goes beyond a critical value y^* given x^*. Similar results can be established for treatment 2.

Chapter 5

Conclusion and Discussion

This thesis presents new allocation rules between two treatments that incorporate a covariate. The goal is to maximize the total discounted expected reward from an finite population of patients. Patient's response is determined from a general linear regression model without any restriction. We develop the optimal strategy for various cases when the variance σ^2 from the regression model is known or unknown. When one of the two treatments is known, the optimal strategy is characterized by an optimal stopping solution for both known and unknown σ^2. When both treatments are unknown, a version of the play-the-winner rule is optimal for both known and unknown σ^2. We also prove that the myopic strategy is not optimal in general settings.

Since there has been so little research addressing optimal adaptive designs with covariate-adjusted responses, or addressing exact evaluations of general designs with

covariate-adjusted responses, there are numerous outstanding problems in this area.

First, one might argue that exact optimal designs are not necessary in practice, especially when good options are available. However, without a basis of comparison it is difficult to assess how good the options are. Actually this may be a future research direction since optimal strategies could be computationally formidable. Examining the properties of optimal designs and the options can lead to the development and selection of superior sub-optimal alternatives.

Another concern is a design's robustness and how to apply it flexibly. In the future work, we will adjust the parameters in the conjugate priors or use other prior distributions to examine the robustness and operating characteristics of our bandit models.

Bibliography

[1] Banks, J. S. and Sundaram, R. K. (1992), Denumberable-armed Bandits, *Econometrica*, 60:1071–1096.

[2] Bergemann, D. and Valimaki, J. (2001), Stationary multi-choice bandit problems, *Journal of Economic Dynamics and Control*, 25:1585–1594.

[3] Berry, D. A. and Fristedt, B. (1985), *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall, London.

[4] Cheng, Y. and Berry, D. A. (2007), Optimal adaptive randomized designs for clinical trials, *Biometrika*, 94:673-689.

[5] Dani, V., Hayes, T. P. and Kakade, S. M. (2008), The price of bandit information for online optimization, In *Advances in Neural Information Processing Systems 20 (NIPS 2007).*

[6] Eick, S. G. (1988), The two-armed bandit with delayed responses, *Ann. Statist.*, 16:254–265.

[7] Hardwick, J. (1995), A Modified Bandit as an Approach to Ethical Allocation in Clinical Trials, *Adaptive designs: IMS Lecture Notes - Monograph Series*, 25:223–237.

[8] Hardwick, J., Oehmke, R. and Stout, Q. F. (2006), New adaptive designs for delayed response models, *Journal of Statistical Planning and Inference*, 136:1940–1955.

[9] McCall, B. P. (1987), A Sequential Study of Migration and Job Search, *Journal of Label Economics*, 5:452–476.

[10] McMahan, H. B. and Blum, A. (2004), Online Geometric Optimization in the Bandit setting Against an Adaptive Adversary, In *proceedings of the 17th Annual Conference on Learing Theory (COLT)*, 109–123.

[11] Puterman, M. L. (1994), *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley and Sons, Inc. New York.

[12] Robbins, H. (1952), Some Aspects of the Sequential Designs of Experiments, In *Bulletin of the American Mathematical Society*, 55:527–535.

[13] Rothschild, M. (1974), A Two-armed Bandit Theory of Market Pricing, *Journal of Economic Theory*, 9:185–202.

[14] Sarkar, J. (1991), One-armed bandit with covariates, *Ann. Stat.*, 19:1978–2002.

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

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

[17] Wang, X. (2007), Dynamic Pricing with a Poisson Bandit Model, *Sequential Analysis*, 26:355–365.

[18] Woodroofe, M. B. (1979), A one-armed bandit problem with a concomitant variable, *J. Am. Statist. Assoc.*, 74:799–806.

[19] Woodroofe, M. B. (1982), Sequential allocation with covariates, *Sankhya A*, 44:403–414.

[20] Zelen, M. (1969), Play the winner rule and the controlled clinical trial, *Journal of American Statistical Association*, 64:131–146.