ABSTRACT

Title of dissertation: Optimal Learning with Non-Gaussian Rewards

Zi Ding, Doctor of Philosophy, 2014

Dissertation directed by: Professor Ilya O. Ryzhov
Robert. H. Smith School of Business

In this disseration, the author studies sequential Bayesian learning problems modeled under non-Gaussian distributions. We focus on a class of problems called the multi-armed bandit problem, and studies its optimal learning strategy, the Gittins index policy. The Gittins index is computationally intractable and approximation methods have been developed for Gaussian reward problems. We construct a novel theoretical and computational framework for the Gittins index under non-Gaussian rewards. By interpolating the rewards using continuous-time conditional Lévy processes, we recast the optimal stopping problems that characterize Gittins indices into free-boundary partial integro-differential equations (PIDEs). We also provide additional structural properties and numerical illustrations on how our approach can be used to approximate the Gittins index.

Optimal Learning with Non-Gaussian Rewards

by

Zi Ding

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Ilya Ryzhov, Chair/Advisor
Professor Michael Fu
Professor Dilip Madan
Professor John Millson
Professor Yuan Liao

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost, I'd like to thank my advisor, Professor Ilya Ryzhov for guiding me to work on the challenging and extremely interesting topics over the past three years. He has an extraordinary vision and ability to identify valuable research topics and steer our research in the right direction. It's always been of great help discussing and brain storming on research studies with him. I accredit one of our major contributions in this disseration, the continuous time interpolation of non-Gaussian rewards under conditional Lévy processes, to him. We are also grateful to Kazutoshi Yamazaki for several helpful discussions in the early stages of this work, although I haven't met him in person yet.

I would also like to thank Professor Dilip Madan. His research seminar has always been a great place where I find fresh ideas and methods. Thanks are due to Professor Michael Fu, Professor John Millson, and Professor Yuan Liao for agreeing to serve on my thesis committee and for sparing their invaluable time reviewing the manuscript.

My colleagues at the AMSC department have enriched my graduate life in many ways and deserve a special mention. Changhui Tan and Tong Meng provided me help with the characterization and solution of PIDEs. My interaction with Wenqing Hu, Huashui Qu, Bin Han, Chen Dong, and Lucas Tcheuko has been very fruitful.

I owe my deepest thanks to my family - my parents who have always been supporting me thoughout my studies. My housemates at my place of residence have been a crucial factor in my finishing smoothly. I'd like to express my gratitude to Lemeng Pan for his friendship and support.

It is impossible to remember all, and I apologize to those I've inadvertently left out. Thank you all!

Table of Contents

List of Abbreviations

KG	Knowledge gradient
UCB	Upper confidence bound
M	Number of alternatives
N	Total number of measurements
x	Alternative/action
$\mathcal{X}$	Decision space
π	Measurement policy
Π	Space of policies
k	Knowledge state
$\mathcal{K}$	Space of knowledge states
$\mathcal{F}$	σ-Algebra
τ	Stopping time
W_n	Discrete-time measurement
X_t	Continuous-time measurement
γ	Discrete-time discount factor
c	Continuous-time discount factor
r	Constant reward
R	Gittins index
μ	Random measure
$\bar{\nu}$	Mean measure
$\leq_{st}$	Stochastic dominance
$\leq_{icx}$	Increasing convex dominance
$\leq_{cx}$	Convex dominance

Chapter 1: Introduction to Optimal Learning

The field of *optimal learning* [1] studies the efficient collection of information in stochastic optimization problems subject to environmental uncertainty. We are surrounded by situations where we need to make a decision while we do not know some or all of the relevant information needed. A few examples are given below.

- Business - We need to identify the best set of features to include on a new smartphone to be released, e.g. iPhone 6. We can run market tests to collect consumer response, but these tests are time-consuming and costly. While information is not free and time is limited, how do we balance time, cost, and the need for learning consumer demand when performing market tests?

- Energy - Finding optimal place for wind farms is no easy task. Wind conditions can depend on microgeography - cliffs, valley, waters, and so forth. To find the best locations, teams with sensors must be sent to make measurements. Given the vastness of lands, how do we optimize the search process so as to minimize the cost and labor? Moreover, wind conditions may variate across seasons, making it necessary to visit a same place multiple times, which brings more challenge to the job.

- Healthcare - The first step in curing a disease usually involves finding a small family of effective base molecules and testing the family of their variations. Each test on one variation can take a day at high expense, while the performance is uncertain. How do we design an efficient and economic testing plan?

In such applications, we are facing problems where uncertainty is driven by unknown probability distributions. While we learn about these unknown distributions by making measurements/observations/tests, we have an overall objective to fulfill at the same time.

1.1 Problem Classification

Given the diversity of optimal learning problems, they can be classified based on the following problem features.

- Online versus offline - Online problems are problems in which we learn form experiences as they occur. For example, we might adjust the price of a product on the Internet and observe the revenue. Every decision or move we make incurs a payoff or cost, and therefore there is a balance between the cost of learning and future benefits. In offline problems, we might be working in the lab under a budget for making measurements and learning from unsuccessful experiments at no cost. After this stage of learning is completed, we make a decision to choose a design or a process that will be put into production.

- Objectives - Problems differ in terms of what we are trying to achieve. Most problems fit well into some minimization (on costs or losses) or maximization (on revenue or payoff) problems. Sometimes, we may also be interested in finding the best design, or finding something that is within some margin of error around the best.

- Measurement decision - In some problems, we face a small number of choices, e.g. drilling test wells to learn about the potential for oil or natural gas. The number of candidate drilling places may be small, but each test can cost millions. Alternatively, we may face big data problems, e.g. choosing 20 proposals out of 100 that have been submitted. Each of these problems introduce different computational challenges because of the size of the search space.

- Implementation decision - The ability to collect the best information depends on what we do with the information once we have it. What to observe (measurement decision) is closely related to what to implement (implementation decision). In many cases, they are the same, e.g. finding the best alternative and exploit it. Sometimes, they are different, e.g. we might measure a link in a graph in order to choose the best path.

- What we believe - In many applications in business, medicine, and various branches of engineering, the decision-maker is able to formulate a belief about the unknown distributions, and gradually improve it using information collected from expensive simulations or field experiments. We may start with some knowledge about the system we study, which allows us to make reason-

able assumptions about different choices. For example, we can put a normal distribution of belief on an unknown quantity. This part will be specifically covered in details in Section 1.3.

- Nature of a measurement - This is a part closely related to what we believe in learning. Is the measurement observed with perfect accuracy? If not, do we know the distribution of the error in taking the measurement?

In the field of optimal learning, assorted problems and applications share similar features, meaning that the general ideas and model frameworks behind them are usually the same. In this dissertation, we focus on one particular family of problems, the multi-armed bandit problem, but the implication of our study is not restricted to this type.

1.2 The Multi-armed Bandit Problem

The *multi-armed bandit problem* [2] or the bandit problem is a fundamental class of optimal learning problems that has inspired some of the pioneering work in the field of optimal learning. Rather than being an important application itself, the bandit problem is useful for helping us understand the basic idea behind optimal learning problems.

Motivated by its name, the multi-armed bandit problem refers to pulling the levers of a collection of slot machines, each with with a different winning probability. Suppose that we face M slot machines and have money enough to play N rounds. We denote our finite set of choices of machines by $\mathcal{X} = \{1, 2, ..., M\}$, which is also

called the decision space in optimal learning. If we choose to play machine $x_n \in \mathcal{X}$ at stage $n = 0, 1, 2...$, we collect single-period random winnings $W_{n+1}^{x_n}$ at the beginning of stage $n + 1$, and the iteration goes on. For each machine x, the winnings from playing it are generated from some common distribution f^x, whose expected value μ^x is unknown. We hope to maximize our total winnings at the end of the game. Therefore, naturally we would like to estimate μ^x for each machine so that we can invest our money on the best machine, while the only way to do this is by paying to play the machines and learning from the outcomes. For this reason, we have to balance the desire to search for the best machine with the overall objective to accumulate as much wealth as possible. In practical use of the bandit model, the competing "arms" or "alternatives" (slot machines in the original story) can be different system designs, pricing strategies, or hiring policies.

To achieve our objective, we need a playing strategy that tells us which machine to play at each stage based on the information we collect up to that time. Such a decision rule is called a *measurement policy* in optimal learning, denoted by π. Let $\mathcal{F}_n$ be the sigma algebra generated by the first n decisions $x_0, x_1, ..., x_{n-1}$ as well as the resulting rewards $W_1^{x_0}, ..., W_n^{x_{n-1}}$, then a policy π is a sequence of functions mapping $\mathcal{F}_n$ into $\mathcal{X}$ for each n. Let Π be the set of all such decision rules, the objective function for the bandit problem can be written as

$$\max_{\pi \in \Pi} \mathbb{E}^{\pi} \sum_{n=0}^{N} \gamma^n \mu^{x_n} \tag{1.1}$$

where $0 < \gamma < 1$ is a pre-specified discount factor (like interest rate). In words, we are looking for a playing strategy that yields the highest expected total discounted

payoff. Throughout the entire dissertation, quantities are subscripted by the time at which they become known; thus in the discrete time above, x_n is chosen at time n, but the output $W_{n+1}^{x_n}$ only becomes known at the beginning of next period. Everything else, including policy information and choice of alternative, is put into the superscript.

With the bandit problem formulated as an optimization problem, there are two key issues that need to be addressed before it is solved. First, the mean rewards μ_x are unknown. Therefore, we need a framework under which the mean rewards can be evaluated or estimated. There are two different types of philosophy on looking at it, through the *frequentist perspective* and the *Bayesian perspective* of statistics. In this thesis, we apply the Bayesian perspective to learning problems, which is described in Section 1.3. Secondly, the challenge of solving the bandit problem is that the expected total discounted payoff cannot be calculated unless a decision rule is stated first so that we know what to measure at each step. To this end, the optimal control problem (1.1) is virtually impossible to solve. As a result, most bandit literature starts from defining a measurement policy first, and then aims at proving its optimality if possible. Many such heuristic policies will be introduced in Section 1.4.

1.3 The Bayesian Perspective

Generally speaking, the core of every learning problem is a probabilistic statement of what we believe about parameters that characterize the uncertainty in the

problems we study. Such beliefs are influenced by observations, like in a bandit problem we learn about the mean rewards μ^x from the outcomes of each machine. There are two types of philosophy on how we charaterize such probabilistic beliefs, the frequentist perspective and the Bayesian perspective [3].

The *frequentist perpective* models things under classic statistical framework, and it is an approach that is most familiar to people with a background in introduction level statistics courses. For example in the bandit problem in Section 1.2, under the frequentist view each machine is assumed to generate rewards from some fixed underlying reward distribution with mean μ^x, which is seen as an unknown constant. We can estimate the value of μ^x using classic statistical methods. For example, method of moments or maximum likelihood estimates would be good choices.

The *Bayesian perspective* casts a different interpretation on the estimates of parameters, which is particularly useful in the context of sequential studies in optimal learning. Under the Bayesian perspective, unknown parameters are characterized as random variables, which are believed to follow some *prior distribution* under our initial beliefs. The initial belief describes our knowledge or subjective view on the parameters before we make any observations. After a measurement is taken, we update the prior distribution with the observation to form a *posterior distribution*, which becomes the next prior distribution to the next measurement iteratively. This dissertation is based entirely on the Bayesian perspective.

In the bandit problem, for each fixed arm x, the rewards $W_1^x, W_2^x, ...$ are drawn from a common sampling distribution with density $f^x(\cdot; \lambda^x)$, where λ^x is an unknown

parameter (or vector of parameters). The rewards are conditionally independent given λ^x. Under the Bayesian perspective, the unknown parameter λ^x is modeled as a random variable, and our beliefs about the possible values of the parameter at time n are represented by the conditional distribution of λ^x given $\mathcal{F}_n$.

For example, assume that the rewards are characterized by normal distributions. On one particular machine, while we drop the machine superscript x for simplicity, the payoffs from playing it are generated by $W \sim \mathrm{N}(\lambda, \sigma_W^2)$. The distribution mean λ is unknown and what we are interested in learning, while the variance σ_W^2, which captures the variation in observation, can either be known or unknown to us (it is actually not unrealistic to assume a known variance in many practical applications, e.g. in finance practioners frequently use constant volatility models) and in this example we assume it is known. Under the Bayesian perspective, we start with a prior belief on λ by assuming that $\lambda \sim \mathrm{N}(\theta_0, \sigma_0^2)$, which characterizes our subjective belief on the mean. In practice, we usually approach problems with some sort of prior knowledge, and when we do not, an uninformative prior can be used, so the Bayesian prior requirement is quite adaptive.

After making the first observation W_1, we can calculate by the Bayes formula

that

$$\begin{aligned} P\left(\lambda \in dx | W_1 = y\right) &= \frac{P\left(W_1 = y | \lambda \in dx\right) P\left(\lambda \in dx\right)}{P\left(W_1 = y\right)} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(y-x)^2}{2\sigma_W^2}} dy \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-\theta_0)^2}{2\sigma_0^2}} dx}{\left(\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(y-x)^2}{2\sigma_W^2}} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-\theta_0)^2}{2\sigma_0^2}} dx \right) dy} \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_W^2}} e^{-\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_W^2}\right)\left(x - \frac{\frac{\theta_0}{\sigma_0^2} + \frac{y}{\sigma_W^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_W^2}}\right)^2} \end{aligned} \tag{1.2}$$

We observe from (1.2) that the posterior distribution of λ is still normal with mean $\left(\frac{\theta_0}{\sigma_0^2} + \frac{y}{\sigma_W^2}\right) / \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_W^2}\right)$ and variance $1/\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_W^2}\right)$. This relationship can be written more concisely under the reciprocal of variance, which we define as the precision β. Precision has an intuitive meaning, as smaller variance means less uncertainty in the outcome and thereby more precise. Accordingly, we denote precisions by $\beta_W = 1/\sigma_W^2$ and $\beta_0 = 1/\sigma_0^2$. Then, (1.2) can be written as

$$P\left(\lambda \in dx | W_1\right) = \frac{\beta_1}{\sqrt{2\pi}} e^{-\frac{1}{2}\beta_1 (x - \theta_1)^2} \tag{1.3}$$

where $\beta_1 = \beta_0 + \beta_W$ and $\theta_1 = \left(\beta_0 \theta_0 + \beta_W W_1\right) / \left(\beta_0 + \beta_W\right)$. Similarly, after observing W_{n+1} in the $(n+1)$th iteration, the updated mean and precision of our belief on λ are given recursively by

$$\theta_{n+1} = \frac{\beta_n \theta_n + \beta_W W_{n+1}}{\beta_n + \beta_W} \tag{1.4}$$

$$\beta_{n+1} = \beta_n + \beta_W \tag{1.5}$$

Relationship like this is called *conjugacy* [4], meaning that prior and posterior distributions of parameters have the same type of distribution at all stages. In words, equation (1.5) states that under the Bayesian perspective we become more

certain of μ as more measurements are taken, and (1.4) characterizes the mean of our belief as a weighted average of all observations and our initial belief. This is why the sequential process of sampling and updating is also called learning. Since (θ_n, β_n) is a pair of sufficient statistics for the normal prior distributions, they fully characterize our beliefs about λ. In Bayesian conjugate models, if there is a set of sufficient statistics k_n^x for the conditional distribution of λ^x given $\mathcal{F}_n$, like (θ_n, β_n) in this example, we call them the *knowledge states*. In the classic multi-armed bandit model, parameters λ^x and λ^y are independent for any $x \neq y$, and likewise the single-period rewards are independent across alternatives. Thus, our beliefs about all the alternatives can be completely characterized by $k_n = \left\{k_n^1, ..., k_n^M\right\}$.

In this example, the sampling distribution of rewards and belief distribution of parameters are both normal. Therefore it is called the normal-normal model. There are not many conjugate pairs like this in Bayesian analysis. We will introduce mainstream non-Gaussian conjugate models in Chapter 2, and they are the focus of this dissertation. Beliefs based on conjugate priors are easy to store and update, making such models useful for practitioners.

1.4 Measurement Policy

Central to the concept of optimal learning is the measurement policy. In general, there are two types of policies, the deterministic ones and the sequential ones. In a deterministic policy, we determine what to do before making any measurment, whereas in sequential policies the decision on what to measure next may depend on

past observations.

It's possible to characterize the optimal measurement policy in the bandit problem mathematically. Assume that we are in knowledge state k_n after making n measurements, the next observation W_{n+1} will be used to update k_n to k_{n+1}. This updating process can be seen as a transition function T

$$k_{n+1} = T\left(k_n, x_n, W_{n+1}^{x_n}\right) \tag{1.6}$$

Let $V(k_n)$ be the value of being in state k_n, which is the objective function (1.1) in the maximization problem. On an infinite time horizon, i.e. $N = \infty$, Bellman's equation characterizes the optimal decision by

$$V(k_n) = \max_{x \in \mathcal{X}} \left(C(k_n, x) + \gamma \mathbb{E} V\left(k_{n+1}(x)\right) | k_n\right) \tag{1.7}$$

where the quantity $C(k_n, x)$ captures the expected gain from playing arm x in state k_n. We let the solution to (1.7) be x_n, and $X^*(\mathcal{K})$ be the complete mapping from the state space of all knowledge states $\mathcal{K}$ to the decision space of all alternatives/actions $\mathcal{X}$. We refer to the function $X^*(\mathcal{K})$ as the optimal policy if it describes the solution to (1.7) for all states $k_n \in \mathcal{K}$.

As mentioned before, the equation (1.7) is virtually impossible to solve, even for very small problems. Not surprisingly, the field of optimal learning consists primarily of finding good heuristics. There are pros and cons of every measurement policy, and thereby different policies are recommended for different problems, learning contexts, as well as objectives. We introduce some of the most popular policies under the context of bandit problems, while their use is not restricted to bandits and also generally apply to other optimal learning problems.

- Pure Exploration - A typical pure exploration strategy samples each alternative with equal probability. In a bandit problem with M arms, each alternative is sampled with probability $1/M$. Pure exploration is not recommended for bandit problems because it does not exploit the economic value of best alternatives currently available. Instead, it focus purely on estimating the value of each choice. Therefore, pure exploration can be effective for offline learning problems, especially when the number of alternatives is extremely large.

- Pure Exploitation - As opposed to pure exploration, pure exploitation exploits the best alternative given current knowledge about our choices. For example in the normal-normal bandit problem, after n iterations we would choose to measure

$$x_n = \arg\max_{x \in \mathcal{X}} \theta_n^x.$$

  The pure exploitation policy is a natural fit for online problems. However, while it seems to focus on the options that appear to be the best, it is very easy to get stuck on some sub-optimal choices, while there might exist better alternatives but we have little information about them.

- Epsilon-Greedy Exploration - This is a simple strategy introduced in [5] that mixes the pure exploration and pure exploitation, so as to avoid the limitations of them. When using this strategy, we explore with probability ϵ and exploit with probability $1 - \epsilon$, where ϵ is a tunable probability value. The problem with this mixed exploration/exploitation strategy is that ϵ must be set subjectively for each application, and when we explore we sample from

the entire population including the clearly suboptimal ones. This defect can be mitigated by using a sequence of decreasing ϵ_n values rather than a fixed one. At the begining of the experiement when we are lacking information, it is better to explore. Therefore it is reasonable to use exploration probility ϵ_n that declines with time, but not too quickly, otherwise we would easily get stuck on one alternative. One way to implement this idea is by setting

$$\epsilon_n = c/n$$

where $0 < c < 1$. In this case, we would measure x conditionally with probability $1/M$ if we explore, and in the limit the number of times x will be measured is

$$\sum_{n=1}^{\infty} \frac{c}{nM} = \infty$$

This assures that each alternative is measured infinitely often and learned perfectly so that we won't leave out an optimal choice by mistake. At the same time, we also spend more time on the alternatives that we think are the best.

- Knowledge Gradient - The name *knowledge gradient* [6] comes from the simple idea of measuring the alternative that produces the greatest value of information, i.e. "maximize the increment of our knowledge". In an M arm bandit problem, suppose we are in knowledge state k_n after n iterations and the value of being in k_n is $V_n(k_n)$. The next choice of measurement x_{n+1} will not only yield an immediate economic reward, but also updates our beliefs and generates k_{n+1}. We define the increment of knowledge, by choosing to measure x

at time $n+1$, as

$$\nu_n^{KG,x} = \mathbb{E}\left[V_{n+1}\left(k_{n+1}(x)\right) - V_n(k_n)|k_n\right] \tag{1.8}$$

and the knowledge gradient policy chooses to measure

$$X_n^{KG} = \underset{x\in\mathcal{X}}{\arg}\nu_n^{KG,x} \tag{1.9}$$

The knowledge gradient or KG policy adapts widely to many offline and online problems [7, 8, 9, 10], and it is known to be empirically competitive in performance against other policies.

- Gittins Index - The crowning result of the bandit literature was created by J.C.Gittins [11]. Gittins found a clever shortcut to solve the bandit problem (1.1) on infinite time horizon. Instead of solving the dynamic control problem in a multidimensional state space, it was possible to characterize the optimal solution using an index policy. An index policy "rates" each alternative by some index score at each iteration, and the alternative with highest index score is played in each measurement. The Gittins index $I_n^{Gittins,x}$ is computed by solving M single-dimensional problems. Gittins showed that in bandit problems on infinite time horizon, i.e. when $n \to \infty$, it is optimal to always play the alternative with the highest Gittins index. However, the Gittins index is difficult to compute. Fortunately, it has convenient scaling properties (see Section 2.3) that allow its users to obtain the index values from a table of standardized index values. Gittins have provided the tables for several most commonly used reward types in [11]. When the total number of plays is finite

in a bandit problem, the optimality proof of the Gittins policy breaks down. This motivates other index policies like the upper confidence bounding policy introduced below.

- Upper Confidence Bound - The upper confidence bounding or UCB policies are a class of policies that has received considerable interest, especially in bandit problems [12, 13, 14]. The UCB family works for many sampling distribution types and is easy to implement. For example in normal-normal bandits, UCB defines the index of alternative x to be

$$I_n^{UCB,x} = \theta_n^x + 4\sigma_W \sqrt{\frac{\log n}{N_n^x}}$$

where N_n^x is the number of times arm x has been played up to and including time n. The main reason that UCB policies became popular is because they offer some sort of optimality property. If N measurements have been made following the UCB policy, it is guaranteed that the average number of times a suboptimal arm is played will be bounded below $C \log N$ for some constant C. This order of number of times, $O(\log N)$, that we spent on suboptimal alternatives, is known as a regret bound, namely the plays that we regret not playing the optimal choice. It has been proved that this is the best possible bound up to the choice of C on the regret bound.

1.5 The Challenge of Learning

Given the richness and diversity of problem classes and learning policies, the major challege in solving optimal learning problems comes from the following few

aspects.

First, the primary challenge that almost every learning problem, especially online ones, faces is the *exploration vs. exploitation* dilemma. For example in the multi-armed bandit problem, each individual reward plays two roles: it contributes immediate economic value, and it also provides information about the alternative with the potential to improve future decisions. This trade-off between information and reward arises in many optimal learning problems and applications where decisions are made in real time, such as dynamic pricing or advertising placement in e-commerce [15] or clinical drug trials with human patients [16]. The exploration vs. exploitation issue has to be treated carefully while choosing or designing measurement policies. Practically speaking, there are many policies, like UCB, that were created with tunable parameters that control how much the policies lean toward exploration or exploitation. While offering some power in control of the "style", it also makes the use of such policies more challenging since the tuning of parameters is played by the user of the policy. It is common that the parameter needs to be changed every time in solving a new problem, while the performance of the policy can depend largely on the choice and is hence not robust.

Secondly, one important theoretical property in optimal learning is the *consistency* of learning policies. If one specific alternative is measured infinitely often, we will learn its mean perfectly by the law of large numbers, that is, the uncertainty in our belief is gone and the underlying true mean value is known almost surely. Consistency refers to a policy's ability to measure the optimal alternative infinitely many times over an infinite time horizon, so that we are guaranteed to find the truly

optimal alternative eventually. To be consistent, it means that we won't get stuck on a suboptimal choice when using a policy. As more observations are made, the policy will eventually guide us to the truly best choice in the decision space. Consistency is not a property that only belongs to a policy itself, but it also depends on the learning problem and model context where it is implemented. For example, the KG policy is known to be consistent in Gaussian reward problems, while it is not under some non-Gaussian models (see Section 2.2).

Thirdly, a concept that immediately accompanies consistency is the *convergece rate*, a measurement of efficiency on consistent learning policies. Provided that a measurement policy is consistent in some problem, we would be interested in the speed that it finds the optimal alternative. This is evaluated by the number of times a policy spends measuring suboptimal alternatives, and a most commonly used measurement of such kind is the *regret*. In the bandit problem, we define the regret value of a policy π as

$$R^{\pi}(n) = \mathbb{E}^{\pi}\left[\sum_{k=1}^{n}\left(\mu_{x^*} - \mu_{x_k^{\pi}}\right)\right]$$

where $x^* = \arg\max_{x\in\mathcal{X}}\mu_x$ is the true best alternative. So R is the expected loss incurred while not choosing the optimal alternative. This regret function grows with the number of iterations N, and if there is an upper bound controlling its growth speed, we call it the regret bound. The regret bound is a most straight-forward way of measuring the covergence rate of learning policies. For bandit problems, it's been proved that the best regret bound a measurement policy can have is of order $O(\log N)$, and one policy family that offers this order of regret bound is UCB. On

the other hand, the KG policy doesn't have a convergence rate result of any kind so far, even though it's been shown to perform very well empirically. Bull [17] made an interesting first step to show KG convergence rate recently.

Fourthly, computational tractability is an important concern in using a learning policy. For example in the bandit problem, the optimal policy on infinite time horizon is provided by the Gittins index, but it is computationally intractable as we will see from Chapter 2. Fortunately, the Gittins index has scaling properties that allow its users to easily solve the problem by referring to a table of index values to standardized versions of the problems. On the other hand, the UCB policies use indices that can be easily calculated. This is also partially why it received lots of interest in this field.

Last but most importantly, as the motivation of this dissertation, more generalized models and distribution types other than Gaussian are needed in modern studies of optimal learning. There has been a well-established pool of research available on models that are based on Gaussian distributions. While the non-Gaussian models haven't received as much attention, many applications require non-Gaussian distributions to be used. For instance, the exponential and the gamma distributions are popular for their positivity, and are thereby frequently used to model waiting time, production level, price volatility, etc. The Bernoulli distribution is a natural choice for modelling binary variables, while the beta distribution with a support between zero and one fits well in modelling unknown probability values.

1.6 Goal of this Dissertation

The main objective of this dissertation is to present our studies on optimal learning problems under non-Gaussian distributions. We chooses to focus on the bandit problem since it is a classic problem that has a clean characterization of the optimal playing policy (see Chapter 2). What we learn from the bandit problem also helps us understand other non-Gaussian learning problems.

Throughout this dissertation, we apply a Bayesian perspective and rely heavily on conjugacy. Chapter 2 first summarizes the non-Gaussian conjugate models in Bayesian analysis, and then we use one example to show that some mainstream policies can have problems when applied to non-Gaussian distributions. Therefore at the end of the chaper, we review a policy that is still optimal for non-Gaussian bandit problems over infinite time horizon, the Gittins index policy. Although under a clean characterization, the Gittins index is computationally intractable, which motivates our research to develop a new theoretical and computational framework for it.

In Chapter 3, we start by reviewing literature on Gittins index approximation for Gaussian problems. We are inspired to develop a novel framework for non-Gaussian problems. The foundation of our approach consists of constructing continuous-time, conditional Lévy processes that serve as probabilistic interpolations of the discrete-time reward processes in the bandit problem. When this idea was used in the Gaussian setting, the properties of Brownian motion allow for easy standardization and numerical solution of stopping problems in continuous-time, un-

der which the Gittins index can be approximated easily. Although these techniques are not available in the non-Gaussian setting, we have shown that the analogous stopping problems can be represented as free-boundary problems on PIDEs that equate the characteristic and infinitesimal operators of the relevant value function.

In Chapter 4, we continue to apply our Chapter 3 results on two major non-Gaussian models, the gamma-exponential and gamma-Poisson problems, and derive the PIDEs that can be solved to approximate the Gittins indices in closed form. Continued into Chapter 5, we prove more structural properties of the value functions in these free-boundary problems, as well as the Gittins indices in continuous time. Theses properties match with the discrete-time results and show that our continuous-time results are consistent with existing discre-time results. At the end of the chaper, we also present numerical illustrations showing the intuitive implications on how the free-boundary PIDE connects to the original Gittins index problem.

The framework we present in this dissertation can be intuitively extended and incorporated into more general reward processes and stopping problems, such as those in [18]. The value functions can easily accommodate different uses, while the novel continuous-time interpolation and optimal stopping to free-boundary transition techniques remain the same in solving other problems.

Chapter 2: Learning with Non-Gaussian Rewards

In the optimal learning literature, Gaussian assumptions are standard due to advantages such as the ability to concisely model correlations between estimated values [19, 20, 21, 22]. More recently, however, numerous applications have emerged where observations are clearly non-Gaussian. The operations management literature has recently studied applications in assortment planning [23, 24] where the observed demand comes from a Poisson distribution with unknown rate. The challenge of learning Poisson distributions also arises in dynamic pricing [25], optimal investment and consumption [26], models for household purchasing decisions [27], and online advertising and publishing [28]. The work by [29] studies a newsvendor problem where a Bayesian gamma prior is used to model beliefs about an exponentially distributed demand. The gamma-exponential model is also used by [30] in the problem of learning signal-to-noise ratios in channel selection, and would also be appropriate for learning service times or network latencies.

Motivated by applications like the above, we study non-Gaussian learning problems in this disseration. Section 2.1 sets up the notation for our analysis and introduces four major classes of non-Gaussian conjugate learning models. Among them, the gamma-exponential model and the gamma-Poisson model will be used

to demonstrate our research findings in Chapter 4. Section 2.2 provides additional motivation for our study by showing that non-Gaussian problems can cause inconsistent behavior in prominent learning policies. Section 2.3 reviews the Gittins index policy for bandit problems. When the reward distributions are non-Gaussian, it still solves the problem optimally on infinite time horizon, which is why we start with this policy in our studies on non-Gaussian problems. The Gittins index has its characterization as the solution to an optimal stopping problem, however, it is computational intractable. This motivates one of the major contributions of this dissertation, a novel thereotical and computational framework under continuous-time interpolation presented in Chapter 3.

2.1 Non-Gaussian Learning Models

From [31], one can see that there are relatively few conjugate models that are non-Gaussian, and of these, we present four classic Bayesian learning models where the sampling densities are non-Gaussian.

Recall that in the bandit problem we consider, the sequence of conditional distributions of λ is characterized by the knowledge states $(k_n^x)_{n=0}^{\infty}$, which is a sequence of random vectors adapted to $(\mathcal{F}_n)$. We write

$$\mathbb{E}\left(W_{n+1}^x \,|\, \mathcal{F}_n\right) = m\left(k_n^x\right) \tag{2.1}$$

for some appropriately chosen function m, so that m is the mean of the reward based on our current belief. For convenience, we also let $m_\infty^x = \mathbb{E}\left(W_{n+1}^x \middle| \lambda^x\right)$ be the "true mean" of the single-period reward, which is the mean of the reward distribution

when the true value of λ^x is known.

- Gamma-Exponential - In the *gamma-exponential* model, f^x is (conditionally) exponential with unknown rate λ^x. Under the assumption that $\lambda^x \sim Gamma\,(a_0^x, b_0^x)$, the conditional distribution of λ^x, given $\mathcal{F}_n$, is also gamma with parameters a_n^x and b_n^x. From [31], we can obtain simple recursive relationships for the parameters, given by

$$a_{n+1}^x = \begin{cases} a_n^x + 1 & \text{if } x_n = x \\ a_n^x & \text{if } x_n \neq x, \end{cases} \tag{2.2}$$

$$b_{n+1}^x = \begin{cases} b_n^x + W_{n+1}^x & \text{if } x_n = x \\ b_n^x & \text{if } x_n \neq x. \end{cases} \tag{2.3}$$

  In the gamma-exponential model, $k_n^x = (a_n^x, b_n^x)$, and the mean function m is given by $m\,(k_n^x) = \mathbb{E}\left(\frac{1}{\lambda^x}\middle|\mathcal{F}_n\right) = \frac{b_n^x}{a_n^x - 1}$.

- Gamma-Poisson - In the *gamma-Poisson* model, the sampling distribution f^x is conditionally Poisson with unknown rate λ^x. Again, we start with $\lambda^x \sim Gamma\,(a_0^x, b_0^x)$, whence the posterior distribution of λ^x at time n is again gamma with parameters a_n^x and b_n^x, and the Bayesian updating equations are now given by

$$a_{n+1}^x = \begin{cases} a_n^x + W_{n+1}^x & \text{if } x_n = x \\ a_n^x & \text{if } x_n \neq x, \end{cases} \tag{2.4}$$

$$b_{n+1}^x = \begin{cases} b_n^x + 1 & \text{if } x_n = x \\ b_n^x & \text{if } x_n \neq x. \end{cases} \tag{2.5}$$

Again, the decision-maker's knowledge about λ^x at time n is represented by $k_n^x = (a_n^x, b_n^x)$ with mean function $m\left(k_n^x\right) = \mathbb{E}\left(\lambda^x | \mathcal{F}_n\right) = \frac{a_n^x}{b_n^x}$.

- Pareto-Uniform - In the *Pareto-uniform* model, the sampling distribution f^x is conditionally uniform on the interval $[0, \lambda^x]$. We start with $\lambda^x \sim Pareto\left(a_0^x, b_0^x\right)$ with parameters $a_0 > 1$ and $b_0 > 0$, whence the posterior distribution of λ^x at time n is again Pareto with parameters a_n^x and b_n^x, and the Bayesian updating equations are now given by

$$a_{n+1}^x = \begin{cases} a_n^x + 1 & \text{if } x_n = x \\ a_n^x & \text{if } x_n \neq x, \end{cases}$$

$$b_{n+1}^x = \begin{cases} \max\left(b_n^x, W_{n+1}^x\right) & \text{if } x_n = x \\ b_n^x & \text{if } x_n \neq x. \end{cases}$$

The decision-maker's knowledge about λ^x at time n is represented by $k_n^x = (a_n^x, b_n^x)$ with mean function $m\left(k_n^x\right) = \mathbb{E}\left(\frac{1}{2}\lambda^x \middle| \mathcal{F}_n\right) = \frac{a_n^x b_n^x}{2a_n^x - 2}$.

- Beta-Benoulli - In the *beta-Bernoulli* model, the sampling distribution f^x is conditionally Bernoulli with unknown success probability λ^x. We start with $\lambda^x \sim Beta\left(a_0^x, b_0^x\right)$, whence the posterior distribution of λ^x at time n is again beta with parameters a_n^x and b_n^x, and the Bayesian updating equations are now given by

$$a_{n+1}^x = \begin{cases} a_n^x + W_{n+1}^x & \text{if } x_n = x \\ a_n^x & \text{if } x_n \neq x, \end{cases}$$

$$b_{n+1}^x = \begin{cases} b_n^x + \left(1 - W_{n+1}^x\right) & \text{if } x_n = x \\ b_n^x & \text{if } x_n \neq x. \end{cases}$$

Again, the decision-maker's knowledge about λ^x at time n is represented by $k_n^x = (a_n^x, b_n^x)$ with mean function $m\left(k_n^x\right) = \mathbb{E}\left(\lambda^x | \mathcal{F}_n\right) = \frac{a_n^x}{a_n^x + b_n^x}$.

Note that, if x is observed infinitely often, we have $m\left(k_n^x\right) \to m_\infty^x$ a.s. by martingale convergence, as proved in Lemma 2.1.1 below.

Lemma 2.1.1. *If x is measured infinitely often, then $m\left(k_n^x\right) \to m_\infty^x$ almost surely.*

Proof: According to the fundamental assumptions under Bayesian perspective, λ is an integrable random variable. By 4.7 in [32], the sequence $\left(m\left(k_n^x\right)\right)$ is a uniformly integrable martingale, whence the lemma is proved. □

The four non-Gaussian models presented in this section appears very similar to the normal-normal model, while the beliefs are all fully characterized by a 2-dimensional knowledge state vector. However, when implemented in optimal learning problems, non-Gaussian models can behave quite differently from Gaussian and cause troubles, as we will see in the following section.

2.2 Difficulty with Non-Gaussian Rewards

We use one example to show that non-Gaussian problems create unexpected theoretical challenges for mainstream policies like the knowledge gradient method.

In a multi-armed bandit problem, the KG policy [21] considers all the alternatives together and calculates the expected improvement

$$R_n^{KG,x} = \mathbb{E}\left[\max_y m\left(k_{n+1}^y\right) - \max_y m\left(k_n^y\right) \middle| \mathcal{F}_n, x_n = x\right] \tag{2.6}$$

that a single implementation contributes to an estimate of the value of the best

alternative. This method (also known by the name "value of information") has received attention in the simulation community [see e.g. 33, for an overview], because it is computationally efficient and often performs near-optimally in experiments.

If the rewards are Gaussian, the policy $X^{KG}(k_n) = \arg\max_x R_n^{KG,x}$ is statistically consistent [34], meaning that $m(k_n^x) \to m_\infty^x$ a.s. for every x. In other words, the policy is guaranteed to discover the best alternative over an infinite horizon; this property is leveraged by [21] to show the asymptotic completeness of information in the bandit setting as $\gamma \nearrow 1$. The consistency of knowledge gradient policies has been shown in numerous settings where Gaussian rewards appear [35, 36, 37]. However, as we now show, this property does *not* hold for the gamma-exponential learning model.

The work by [38] provides a closed-form solution of (2.6) for the gamma-exponential problem, given by

$$R_n^{KG,x} = \begin{cases} \frac{1}{(a_n^x-1)(C_n^x)^{a_n^x-1}} \left(\frac{b_n^x}{a_n^x}\right)^{a_n^x} & \text{if } \frac{b_n^x}{a_n^x-1} \le C_n^x \\ \frac{1}{(a_n^x-1)(C_n^x)^{a_n^x-1}} \left(\frac{b_n^x}{a_n^x}\right)^{a_n^x} - \left(\frac{b_n^x}{a_n^x-1} - C_n^x\right) & \text{if } \frac{b_n^x}{a_n^x} \le C_n^x < \frac{b_n^x}{a_n^x-1} \\ 0 & \text{if } C_n^x < \frac{b_n^x}{a_n^x}, \end{cases} \tag{2.7}$$

where $C_n^x = \max_{y \neq x} m(k_n^y)$. We see immediately that it is possible to have $R_n^{KG,x} = 0$ even though $Var(\lambda^x | \mathcal{F}_n) > 0$. That is, the marginal value of a single observation of x may be zero even when we are uncertain about λ^x. This behavior, which does not arise in the Gaussian setting, is the cause of the inconsistency result. Essentially, if the policy places zero value on alternative x, there may be a non-zero probability that the policy will never measure x, which means that $m(k_n^x)$ will not converge to the true mean reward. The proof below uses the technique of continuous interpolation,

used later in Section 3 to approximate Gittins indices.

Theorem 2.2.1. *There exists a gamma-exponential problem for which the KG policy has a non-zero probability of never measuring a particular alternative.*

Proof: Consider a problem with two alternatives. For simplicity, let $a_0^1 = a_0^2 = 2$, and choose b_0^1, b_0^2 such that $b_0^2 < \frac{b_0^1}{2}$. By (2.7), the KG policy will measure alternative 2. Our beliefs about alternative 1 will thus remain unchanged. Let E be the event that

$$\frac{b_n^2}{a_n^2 - 1} < \frac{b_0^1}{2} \qquad \text{for all } n \geq 0.$$

Clearly, this is the event that we will *never* measure alternative 1. We show that $P(E) > 0$. For notational convenience, let λ refer to the rate λ^2 of alternative 2, and let $c = \frac{b_0^1}{2}$.

Define a continuous-time stochastic process $(X_t)_{t \geq 0}$ as follows. Given λ, (X_t) is a gamma process with shape parameter 1 and scale parameter λ. The increments $X_{n+1} - X_n$, $n = 0, 1, ...$ are i.i.d. exponential with rate λ, the same as the random rewards we collect when we measure alternative 2. The initial value of the process is $X_0 = b_0^2$. Then, X_n has the same conditional distribution as b_n^2, given λ. We now observe that

$$P(E \mid \lambda) \geq P\left(\frac{X_t}{t+1} < c \text{ for all } t \geq 0 \middle| \lambda\right) = P(X_t < c(t+1) \text{ for all } t \geq 0 | \lambda). \tag{2.8}$$

Given λ, E is the event that (X_t) satisfies a certain condition at discrete points in time, which contains the event that the condition is satisfied at all continuous times.

If we can show that (2.8) is strictly positive when λ takes values in a non-negligible set, applying the tower property will show that $P(E) > 0$.

Consider the case where $\lambda > \frac{1}{c}$. Now the last expression in (2.8) equals

$$P\left(X_t < c\left(t+1\right) \text{ for all } t \geq 0 | \lambda\right) = P\left(\inf_{t \geq 0} Y_t > -c \middle| \lambda\right)$$

where $Y_t = ct - X_t$ and $Y_0 = -b_0^2$. Because (X_t) is a pure jump process that increases a.s., (Y_t) is a spectrally negative Lévy process (or a Lévy process whose jumps are always negative). Because (X_t) is conditionally a gamma process, we must have $\mathbb{E}(X_1 - X_0) = \frac{1}{\lambda}$ and hence $\mathbb{E}(Y_1 - Y_0) = c - \frac{1}{\lambda} > 0$. In this case,

$$P\left(\inf_{t \geq 0} Y_t > -c \middle| \lambda\right) = \mathbb{E}(Y_1 - Y_0) w(c + Y_0) = \mathbb{E}(Y_1 - Y_0) w(c - b_0^2) \tag{2.9}$$

where w is called the scale function of the spectrally negative Lévy process (Y_t); see [39], p. 215. The expression $\mathbb{E}(Y_1 - Y_0)$ in (2.9) is due to the fact that $\psi'(0+) = \mathbb{E}(Y_1 - Y_0)$ by the property of the moment-generating function, where ψ is the Laplace exponent $\psi(s) = \log \mathbb{E} e^{s(Y_1 - Y_0)}$. Because $w(x) > 0$ for any $x > 0$, we have shown that the conditional probability is strictly positive given values of λ in a non-negligible set. Thus, there is a strictly positive probability that we will be stuck on alternative 2 forever, and this alternative will always look worse than alternative 1. □

The main value of this result is the insight that it provides into learning with non-Gaussian rewards. In this setting, the theoretical guarantees of a well-studied class of heuristics (also used to establish some asymptotic results in the online setting) unexpectedly break down, suggesting that the only way to reliably gauge

the potential of an alternative is by looking over an infinite horizon, as the Gittins policy does in the following section.

2.3 An Optimal Policy in Non-Gaussian Bandits

In the classical multi-armed bandit setting where the decision-maker's beliefs about the alternatives are mutually independent, the work by [11] shows that the optimal strategy takes the form of an index policy. At each time stage, an index is computed for each alternative independently of our knowledge about the others, and the alternative with the highest index is implemented. The index can be expressed as the solution to an optimal stopping problem [40]. Nonetheless, despite this considerable structure [see e.g. 41, for additional scaling properties], which continues to inspire new theoretical research on Gittins-like policies [42, 43], even the stopping problem for a single arm is computationally intractable. This challenge has given rise to a large body of work on heuristic methods, which typically make additional assumptions on the reward distribution. It is especially common to require the reward distributions to be Gaussian, or to have bounded support [see e.g. 14, for examples of both].

We briefly summarize the characterization of the Gittins index policy, known to optimally solve (1.1) when $N = \infty$. For a more detailed introduction, the reader is referred to Ch. 6 of [1]. Furthermore, [2] provides a deeper theoretical treatment with several equivalent proofs of optimality for the policy.

The Gittins method considers each alternative separately from the others. Let

k denote our beliefs about an arbitrary alternative, dropping the superscript x for notational convenience. Consider a situation where, in every time stage, we have a choice between implementing this alternative and receiving a known, deterministic "retirement reward" r. The optimal decision (implement vs. retire) can be characterized using Bellman's equation for dynamic programming. We write

$$V(k,r) = \max\left\{r + \gamma V(k,r), \mathbb{E}\left[W + \gamma V(k',r)|k\right]\right\}, \tag{2.10}$$

where W is the reward obtained from the implementation, and k' is the future knowledge state arising due to the new information provided by W. In our non-Gaussian setting, k' would be computed using equations like (2.2)-(2.3) or (2.4)-(2.5). Because we do not update our beliefs when we collect the fixed reward, it follows that, if we prefer the fixed reward given the knowledge state k, we will continue to prefer it for all future time periods. Then, (2.10) becomes

$$V(k,r) = \max\left\{\frac{r}{1-\gamma}, m(k) + \gamma\mathbb{E}\left[V(k',r)|k\right]\right\}. \tag{2.11}$$

The Gittins index is a particular retirement reward value $R(k) := r^*(k)$ that makes us indifferent between the two quantities inside the maximum in (2.11). In the special case where the parameter λ^x is known, this retirement reward is equal to the mean single-period reward, as shown in the following lemma. Although this result is not directly of help in computing Gittins indices for unknown λ^x, we use it in conjunction with our continuity analysis in Section 5.2 to establish initial conditions for the numerical procedures developed later on.

Lemma 2.3.1. *If the parameter λ^x is a known constant, the Gittins index of arm x is m^x_∞.*

Proof: If λ^x is known, there is no knowledge state, and (2.10) becomes

$$V(r) = \max\left\{r + \gamma V(r), m_\infty^x + \gamma V(r)\right\},$$

whence the result follows immediately. □

Once Gittins indices have been computed, the policy

$$X_n^*(k_n) = \arg\max_x R(k_n^x)$$

can be shown to be optimal for the objective in (1.1). Thus, the Gittins method decomposes an M-dimensional problem into M one-dimensional problems, each of which can be solved independently of the others. Furthermore, in the gamma-exponential version of the problem (that is, where $k = (a, b)$ and (2.2)-(2.3) are used to update k), it has also been shown [41] that

$$R(a, b) = bR(a, 1), \tag{2.12}$$

meaning that Gittins indices only have to be computed for a restricted class of knowledge states. Equivalently, if we can find $\tilde{b}(a)$ such that $R\left(a, \tilde{b}(a)\right) = 1$, we can use (2.12) to write

$$\tilde{b}(a) R(a, 1) = R\left(a, \tilde{b}(a)\right) = 1, \tag{2.13}$$

whence $R(a, b) = \frac{b}{\tilde{b}(a)}$.

Yet, even with this structure, the problem remains computationally intractable: it is difficult to compute $R(a, 1)$ or $\tilde{b}(a)$ for arbitrary a. Efficient approximation methods have been developed for the Gittins indices under Gaussian rewards, and

they inspired us to develop a continuous-time interpolation for non-Gaussian rewards, under which the Gittins index has a novel characterization and can be solved under PIDEs.

Chapter 3: A Novel Framework for Non-Gaussian Bandits

In the Gaussian setting, a recent stream of work by [44], [45], and [46] has approximated the Gittins index for one arm by formulating an optimal stopping problem on a Brownian motion with unknown drift, a continuous-time process that serves as a probabilistic interpolation of the sequence of Gaussian rewards collected from the arm. By making the connection between Brownian motion and the heat equation [47], one can formulate and numerically solve a free-boundary problem [48] to approximate the Gittins index. On the other hand in the non-Gaussian setting, there has been a stream of research on multi-armed bandit problems driven by Lévy processes [49, 50, 51, 52]. They consider bandit models where rewards are generated in continuous time from Lévy processes, and constructed alternative characterization of the Gittins index under Wiener-Hopf decomposition of Lévy processes. However, under the Bayesian perspective in our research, the reward distributions are characterized by unknown random parameters and our beliefs on them evolve over time. The dependency of belief of sampling distribution on the information collected from a path of rewards requires the interpolation process to have increments that are generally non-stationary and non-independent, hencefully not Lévy.

These studies inspired one of our major contributions in this dissertation. In the non-Gaussian bandit problem, we interpolate the reward sequence in the non-Gaussian problems with conditional Lévy processes. Under this continuous-time interpolation, the relevant optimal stopping problems for characterizing Gittins indices can be recast into free-boundary partial integro-differential equations. Section 3.1 first reviews the continuous-time interpolation technique developed for Gaussian bandits. Section 3.2 summarizes the theory of conditional Lévy processes, and uses it to formulate the continuous time interpolation of rewards. We derive the continuous-time analog to the Gittins index equation (2.11) under this interpolation, similar to that used by [44, 45] in studying Gaussian problems. Section 3.3 uses methods in [53] to recasts it into PIDEs, which can be used to solve for the Gittins indices. The Gittins indices obtained from such a continuous-time interpolation can be used to approximate the Gittins indices in the original discrete-time problems.

3.1 Continuous-time Interpolation of Gaussian Rewards

The stream of research approximating Gittins indices for Gaussian rewards begins with the work by [44], which proposed the following idea. For arbitrary x (in the following, we again drop the superscript x for convenience as in Section 2.3), the discrete-time process $(W_n)_{n=1}^{\infty}$ of single-period rewards with unknown mean μ and known variance σ^2 is replaced by a continuous-time process $(X_t)_{t\geq 0}$. The process (X_t) is constructed in such a way that, for integer t, the increment $X_{t+1} - X_t$ has the same distribution as the single-period reward W_{t+1}. Therefore, (X_t) can be viewed

as a probabilistic interpolation of (W_n). In the Gaussian setting, (X_t) is conditionally a Brownian motion with unknown drift μ and known volatility σ. Under the Bayesian perspective, we start with prior belief $\mu \sim N(\theta_0, \beta_0)$, and while given $\mathcal{F}_t$ the conditional distribution of μ is $N(\theta_t, \beta_t)$, where $\theta_t = (\beta_0\theta_0 + W_t) / (\beta_0 + t)$ and $\beta_t = \beta_0 + t$. For integer values of t, the beliefs on μ match with (1.4)-(1.5) in the discrete-time normal-normal model.

Let c be a continuous-time discount factor (lower c corresponds to higher γ in discrete time). The formulation of the Gittins index in (2.11) can be extended to continuous time and written as the solution R to the optimal stopping problem

$$
\begin{aligned}
R\int_0^\infty e^{-ct}dt &= \sup_{\tau\geq 0}\mathbb{E}^\pi\left[\int_0^\tau e^{-ct}dW_t + R\int_\tau^\infty e^{-ct}dt\right] && (3.1)\\
&= \sup_{\tau\geq 0}\mathbb{E}^\pi\left\{\mathbb{E}\left[\int_0^\tau e^{-ct}dW_t + R\int_\tau^\infty e^{-ct}dt|\mathcal{F}_t\right]\right\} && (3.2)\\
&= \sup_{\tau\geq 0}\mathbb{E}^\pi\left[\int_0^\tau \theta_t e^{-ct}dt + R\int_\tau^\infty e^{-ct}dt\right] && (3.3)\\
&= \sup_{\tau\geq 0}\mathbb{E}^\pi\left[\frac{1}{c}\theta_0 - \frac{1}{c}\left(\theta_\tau - R\right)e^{-c\tau}\right]
\end{aligned}
$$

The equation (3.1) states that the Gittins index in continuous time is a retirement cash flow that makes the player indifferent between receiving it constantly forever and receiving it after playing the machine optimally up to some stopping time. In the next line (3.2), under taking iterated expectation (also known as the tower property of expectation) conditioned on $\mathcal{F}_t$, we are left with the conditional mean values θ_t. Then, we perform integration by parts in (3.3).

Under a time change, by making $s = (c\beta_t)^{-1}$ and $Z(s) = \left(\theta_0 - \theta_{(cs)^{-1}-\beta_0}\right)/\sqrt{c}$, it can be easily shown that $Z(s), 0 < s \leq s_0$ is a standard Brownian motion in the $-s$ scale, where $s_0 = (c\beta_0)^{-1}$ and $Z(s_0) = 0$. By letting $z_0 = (R - \theta_0)/\sqrt{c}$, the

stopping problem on R is written as

$$z_0 e^{-1/s_0} = \sup_{0<s\leq s_0} \mathbb{E}\left[e^{-\frac{1}{s}}\left(Z(s)+z_0\right)\right] \tag{3.4}$$

Therefore, R is a solution to the original stopping problem if and only if z_0 is a solution to the stopping problem (3.4) on standard Brownian motion. In [44], a corrected binomial method due to Chernoff and Petkau [54] was used to simulate standard Brownian paths to solve (3.4), with a representation of the optimal value function given in [55] to initialize the the algorithm. Then, a closed form function can be fitted to approximate the stopping boundary obtained from the solution, which provides a formula for calculating the Gittins index by using the knowledge state vector (θ, β) as inputs. These continuous-time index values serve as an approximation to the Gittins index in the original discrete-time problem.

3.2 Continuous-time Interpolation of Non-Gaussian Rewards

In non-Gaussian reward problems, we construct a continuous-time process (X_t) in the same way that, for integer t, the increment $X_{t+1} - X_t$ has the same distribution as the single-period reward W_{t+1}. Given λ, the discrete time rewards are i.i.d, so we use a conditional Lévy process with increments that are conditionally independent and stationary. For the gamma-exponential problem, (X_t) is conditionally a gamma process given λ [see e.g. 32, for a definition], with exponentially distributed increments over unit intervals. Similarly in the gamma-Poisson problem, (X_t) is conditionally a Poisson process.

The major challenge in studying non-Gaussian reward problems under this in-

terpolation technique is that we cannot exploit the time-change properties of Brownian motion to "standardize" the problem, as was done before in the Gaussian reward case. Therefore, we develop an alternate method based on equating the infinitesimal and characteristic operators [53] of value functions in an optimal stopping problem. We then obtain free-boundary problems on partial integro-differential equations (PIDEs), which can be solved numerically to approximate the Gittins index. The solutions to these equations are shown to possess intuitive properties, such as continuity and monotonicity, that are known to hold for classic bandit problems in discrete time.

By a conditional Lévy process, we mean a process (X_t) whose conditional law, given some random variable λ, is that of a process with stationary and independent increments. In this section, we follow the theoretical characterization of conditional Lévy processes introduced in [56].

Let (X_t) be a real-valued stochastic process that will later serve as the continuous-time interpolation of cumulative rewards without discounting. The parameter λ is a random variable (or random vector) whose conditional distribution given $\mathcal{F}_t$ characterizes our belief on λ at time t. While conditioned on λ, the process (X_t) has stationary and indepedent increments. Such a process is a conditional Lévy process.

Under the context of non-Gaussian conjugate models, we further restrict (X_t) to increasing and right-continuous pure jump processes, mainly because the gamma-exponential and gamma-Poisson problems fall into this category. The method below does apply to general stochastic processes, but this is not as useful for interpolation purposes in Bayesian bandit problems.

The dependence of X_t on λ is described as

$$X_t = X_0 + \int_{[0,t]\times R^+} z\mu(ds, dz) \tag{3.5}$$

where μ is conditionally (given λ) a random measure on $R^+ \times R^+$ with mean measure $\nu(\lambda, dz)ds$, satisfying $\int_{R^+} \nu(\lambda, dz)(z \wedge 1) < \infty$ for all λ [for details on random measure and mean measure, see Ch. 6 in 32]. The unconditional intensity measure of μ at time t, that is, the intensity given $\mathcal{F}_t$ but not given λ, is written as $\bar{\nu}_t(dz)ds = \mathbb{E}\left[\nu(\lambda, dz)|\mathcal{F}_t\right] ds$.

Intuitively, this definition states that a conditional Lévy process has "two layers of randomness". The conditional mean measure ν characterizes the infinitesimal behavior of the process, given a value of λ, as that of a Lévy process. The unconditional intensity measure $\bar{\nu}$ further removes the randomness in the belief on λ by integrating over its distribution. Thus, $\bar{\nu}$ can be described as "the mean of the conditional mean measure".

With the reward process (X_t) set up as a continuous-time conditional Lévy process, the Gittins logic can be extended to the continuous-time setting similarly as in (3.1). The Gittins index R is the particular value of r such that

$$r \int_0^\infty e^{-cs} ds = \sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs} dX_s + r \int_\tau^\infty e^{-cs} ds\right], \tag{3.6}$$

where τ denotes a stopping time and c is the continuous-time discount factor. This expectation is evaluated given some initial state k at time 0, which is the starting time when the index is calculated. We omit this dependence from the notation and take the present time to be zero without loss of generality. This formulation is equivalent to the one in (2.11); see e.g. [40] or [45] for more details. As before,

discounted rewards are collected from the process (X_t) until time τ, at which point we collect the fixed retirement reward R until the end of time. If (3.6) holds, we are indifferent between stopping immediately and running the process until the optimal stopping time τ.

3.3 Optimal Stopping Problems to Free-boundary Problems

Technically speaking, when the rewards are non-Gaussian, the process can still be embedded in Brownian motion [57], but the time change is random and computationally intractable. Instead, we will apply a new approach based on [53]. We manipulate the continuous-time Gittins index set up in (3.6) as follows:

$$
\begin{aligned}
0 &= \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs} dX_s - \int_0^{\tau} e^{-cs} r ds\right] \\
&= \sup_{\tau} \mathbb{E}\left[\int_{[0,\tau]\times R^+} e^{-cs} z\mu(dz, ds) - \int_0^{\tau} e^{-cs} r ds\right] \\
&= \sup_{\tau} \mathbb{E}\Bigg[\int_0^{\tau} e^{-cs}\left(\int_{R^+} z\nu(\lambda, dz) - r\right) ds \\
&\quad + \int_{[0,\tau]\times R^+} e^{-cs} z\left[\mu(dz, ds) - \nu(\lambda, dz) ds\right]\Bigg] \qquad (3.7) \\
&= \sup_{\tau} \mathbb{E}\Bigg[\int_0^{\tau} e^{-cs}\left(\int_{R^+} z\nu(\lambda, dz) - r\right) ds \\
&\quad + \int_{[0,\tau]\times R^+} e^{-cs} z\mathbb{E}\left[\mu(dz, ds) - \nu(\lambda, dz) ds | \mathcal{F}_s\right]\Bigg] \qquad (3.8) \\
&= \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\left(\int_{R^+} z\nu(\lambda, dz) - r\right) ds\right] \\
&= \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\left(\int_{R^+} z\mathbb{E}\left(\nu(\lambda, dz)|\mathcal{F}_s\right) - r\right) ds\right] \qquad (3.9) \\
&= \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\left(\int_{R^+} z\bar{\nu}_s(dz) - r\right) ds\right] \qquad (3.10)
\end{aligned}
$$

In (3.7) we use a compensating technique by adding and subtracting $\nu(\lambda, dz)$. Then. the random measure μ is cancelled in (3.8), by taking iterated conditional expected values (also called the tower property). We then use the tower property again in (3.9). Notice that $\int_{R^+} z\bar{\nu}_s(dz)$ is the mean of the infinitesimal increment of the process. For a classic Lévy process, where λ is a known constant and has no dependency on $\mathcal{F}_s$, the term $\int_{R^+} z\nu(dz)$ is called the compensator of the process [58].

We denote $\int_{R^+} z\bar{\nu}_t(dz)$ by m_t to emphasize that this quantity serves the same role as $m(k_n)$ in (2.1). Then, (3.6) in continuous time can be written as,

$$\sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\,(m_s - r)\,ds\right] = 0 \tag{3.11}$$

which we will refer to as the "calibration equation" throughout this dissertation.

We also introduce notation for the LHS of (3.11),

$$V(t, m) := \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\,(m_s - r)\,ds\right] \tag{3.12}$$

Recall that the expectation in (3.12) is evaluated given some initial state at time 0, and we dropped it for simplicity. As we observe from models such as (2.2)-(2.5), the pair (t, m), representing a time parameter and a mean parameter, is a set of sufficient statistics for the distribution of λ given $\mathcal{F}_t$ (these statistics are standard for all the models described in [2]). Therefore, $V(t, m)$ indicates that our initial knowledge is characterized by (t, m). In this value function, r is a fixed constant value and the Gittins index R is the particular value of r that makes $V = 0$. If we are currently in a knowledge state (t, m), it suffices to solve for the value r such that $V(t, m) = 0$ to obtain the Gittins index value. On the other hand, if we fix r, the

set of (t, m) for which $V(t, m) = 0$ is precisely the set of states that have r as the Gittins index.

We now construct a free-boundary problem for V by equating the characteristic and infinitesimal operators of V. In this approach, we rely on the mild condition that (m_t) is a càdlàg strong Markov process. As we will see from calculating m_t explicitly for gamma-exponential and gamma-Poisson problems in Chapter 4, this assumption is reasonable and generally satisfied in the context of Bayesian conjugate models.

The *characteristic operator* of V is defined as

$$L^{char}V(t, m) = \lim_{U \downarrow \{m\}} \frac{\mathbb{E}V\left(t_{\tau_{U^c}}, m_{\tau_{U^c}}\right) - V(t, m)}{\mathbb{E}\left(\tau_{U^c}\right)}, \tag{3.13}$$

where U is an open set that contains m, and τ_{U^c} is the hitting time of the set U^c for the process (m_t). That is,

$$\tau_{U^c} = \inf\left\{t \geq 0 \,:\, m_t \in U^c\right\}$$

is the first time at which (m_t) leaves the set U. First we consider the value function at the moment when τ_{U^c} occurs, and then we shrink U down to the singleton $\{m\}$.

The concept of the characteristic operator dates back to [59]. The value function $V = \mathbb{E}\left[\int_0^\tau e^{-cs}\left(m_s - r\right) ds\right]$ is analogous to a killed version of the Lagrange problem introduced in Ch. 6 of [53], where the value function is of the form $\int_0^\tau e^{-\Lambda(s)} L\left(m_s\right) ds$. The characteristic operator in the Lagrange problem can be explicitly calculated, and (3.13) is given in closed form in the following result.

Lemma 3.3.1. *If (m_t) is a càdlàg strong Markov process, the characteristic operator*

of V is given by

$$L^{char}V(t,m) = cV(t,m) - (m-r). \tag{3.14}$$

Proof: The lemma follows from (7.2.8) in [53]. In a killed Lagrange problem on the value function $\int_0^\tau e^{-\Lambda(s)}L(m_s)\,ds$, by inserting $\Lambda(s) = cs$ and $L(m_s) = m_s - r$, we obtain the desired results in the lemma. □

The *infinitesimal operator* L^{inf} (also called the generator of V) is derived using Itô's lemma. Our goal is to obtain an expression

$$V(t,m_t) = V(0,m_0) + \int_0^t L^{inf}V(s,m_s)\,ds + Y_t, \tag{3.15}$$

where (Y_t) is a martingale formed by adding and subtracting a continuous compensator to the jump component of V [see 58, for an exposition of this idea].

Recall that in discrete time models such as (2.2)-(2.5), the mean process (m_t) is expressed by t and X_t in simple closed form. We now assume that m_t can be written as $g(t, X_t)$ for some continuous function g with first-order derivatives. In Chapter 4, we will explicitly derive g and show that this assumption is satisfied in gamma-exponential and gamma-Poisson problems.

Lemma 3.3.2. *If (m_t) can be written into the form $m_t = g(t, X_t)$ for some continuous function g with first-order derivatives, the infinitesimal operator of V is given by*

$$L^{inf}(t,m) = V_t + g_t V_m + \int_{R^+} [V(t, g(t, X_t + z)) - V(t, g(t, X_t))]\bar{\nu}_t(dz). \tag{3.16}$$

Proof: The proposition follows from the calculation

$$
\begin{aligned}
& V(t, m_t) \\
= \; & V(0, m_0) + \int_0^t \frac{\partial V}{\partial s}(s, m_s) ds + \int_0^t \frac{\partial V}{\partial m}(s, m_s) dm_s^c + \sum_{0<s\leq t} \left[V(s, m_s) - V(s, m_{s-}) \right] \qquad (3.17) \\
= \; & V(0, m_0) + \int_0^t \frac{\partial V}{\partial s}(s, m_s) ds + \int_0^t \frac{\partial V}{\partial m}(s, m_s) dm_s^c \\
& + \int_{[0,t]\times R^+} \left[V(s, g(s, X_s + z)) - V(s, g(s, X_s)) \right] \mu(ds, dz) \\
= \; & V(0, m_0) + \int_0^t \frac{\partial V}{\partial s}(s, m_s) ds + \int_0^t \frac{\partial V}{\partial m}(s, m_s) \frac{\partial g}{\partial s} ds \qquad (3.18) \\
& + \int_{[0,t]\times R^+} \left[V(s, g(s, X_s + z)) - V(s, g(s, X_s)) \right] \bar{\nu}_s(dz) ds \\
& + \int_{[0,t]\times R^+} \left[V(s, g(s, X_s + z)) - V(s, g(s, X_s)) \right] \left(\mu(ds, dz) - \nu(\lambda, dz) ds + \nu(\lambda, dz) ds - \bar{\nu}_s(dz) ds \right) \\
= \; & V(0, m_0) + \int_0^t L^{inf} V(s, m_s) \, ds + Y_t
\end{aligned}
$$

In (3.17), we use Itô's lemma for jump-diffusion processes [see 60, Ch. 6, Theorem 31.5], and m_s^c denotes the continuous part of m_s, after removing all jumps. Since $m_s = g(s, X_s)$ and X_s is a pure jump process, we have $dm_s^c = \frac{\partial g}{\partial s} ds$. In (3.18), we applied a compensator technique and put the component with respect to the random measure μ into an $\mathcal{F}_t$-martingale Y_t. Note that, for $T \geq t$, the tower property implies that

$$
\begin{aligned}
& \mathbb{E}\left[Y_T | \mathcal{F}_t \right] \\
= \; & Y_t + \mathbb{E}\left[\int_{[t,T]\times R^+} \left[V(s, g(s, X_s + z)) - V(s, g(s, X_s)) \right] \left[\mu(ds, dz) - \nu(\lambda, dz) ds + \nu(\lambda, dz) ds - \bar{\nu}_s(dz) ds \right] \middle| \mathcal{F}_t \right] \\
= \; & Y_t + \mathbb{E}\left[\int_{[t,T]\times R^+} \left[V(s, g(s, X_s + z)) - V(s, g(s, X_s)) \right] \mathbb{E}\left[\mu(ds, dz) - \nu(\lambda, dz) ds | \mathcal{F}_s \right] \middle| \mathcal{F}_t \right] \\
& + \mathbb{E}\left[\int_{[t,T]\times R^+} \left[V(s, g(s, X_s + z)) - V(s, g(s, X_s)) \right] \mathbb{E}\left[\nu(\lambda, dz) ds - \bar{\nu}_s(dz) ds | \mathcal{F}_s \right] \middle| \mathcal{F}_t \right] \\
= \; & Y_t
\end{aligned}
$$

Therefore, the infinitesimal operator $L^{inf} V$ is,

$$
L^{inf} V(t, m) = \frac{\partial V}{\partial t}(t, m) + \frac{\partial g}{\partial t}(t, X_t) \frac{\partial V}{\partial m}(t, m) + \int_{R^+} \left[V(t, g(t, X_t + z)) - V(t, m) \right] \bar{\nu}_s(dz)
$$

which yields the result in the lemma. □

Essentially, the characteristic and infinitesimal operators are two different expressions for the derivative of V based on Kolmogorov theory and Itô calculus. Under general arguments [53], the two operators exist and coincide. By matching them, we obtain a free-boundary problem on a PIDE as a consequence of the above derivations. For notational convenience, we denote the partial derivative $\frac{\partial V}{\partial t}$ by V_t and similarly with respect to other variables.

Theorem 3.3.3. *If (m_t) is a strong Markov càdlàg process and can be written as $g(t, X_t)$ for some continuous function g with first order derivatives, the value function $V(t, m)$ solves the free-boundary problem*

$$\begin{aligned} V_t(t,m) + g_t(t,X_t) V_m(t,m) + \int_0^\infty \left[V(t, g(X_t+y)) - V(t,m)\right] \bar{\nu}_t(dy) &= cV(t,m) - (m-r) \\ V(t, m^*(t)) &= 0 \end{aligned}$$

where $m^*(t)$ is an unknown stopping boundary curve. For every point on the stopping boundary, the Gittins index $R(t, m^*(t))$ is equal to r.

Proof: According to the characteristic operator and infinitesimal operator shown in Lemmas 3.3.1 and 3.3.2, the theorem follows from Ch. 7.2 in [53]. □

We have assumed a fixed retirement reward r in this PIDE. Thus, the solution of the PIDE does not immediately yield a Gittins index $R(t, m)$ for an arbitrary knowledge state (t, m). However, the stopping boundary curve m^* describes the set of all knowledge states whose Gittins index is exactly equal to r. In general, out method applies to all infinitely divisible reward distributions, by their connection to Lévy processes [60], while shifting and scaling properties of Gittins indicies are only offered in distributions with location and scale parameters respectively [11].

Therefore, it is sufficient to compute any Gittins index for the gamma-exponential problem via (2.13) by solving only one PIDE under $r = 1$, while for other non-Gaussian models that do not equip with scale parameters, a family of PIDEs for different r values would need to be solved.

Chapter 4: Gittins Indices in Major Non-Gaussian Models

Among the non-Gaussian conjugate models introduced in Section 2.1, three are most commonly studied in the bandit literature: the gamma-exponential model, the gamma-Poisson model, and the beta-Bernoulli model. The setting of beta-Bernoulli has been relatively well-studied [16, 61], and we do not explore it here. By contrast, the gamma-exponential and gamma-Poisson models have received the least amount of theoretical attention in the bandit literature. Therefore, we now apply the general result of Theorem 3.3.3 to characterize Gittins indices for exponential and Poisson rewards. Section 4.1 covers the gamma-exponential problem, whereas Section 4.2 covers the gamma-Poisson problem.

4.1 Exponential Reward Problems

In the gamma-exponential problem, our continuous-time interpolation (X_t) is a gamma process with shape parameter 1 and unknown scale parameter λ. We begin by assuming $\lambda \sim Gamma\,(a_0, b_0)$, reflecting the decision-maker's prior beliefs. Letting $\mathcal{F}_t$ be the σ-algebra generated by the path of (X_t) up to time t, we find that the conditional distribution of λ given $\mathcal{F}_t$ is still gamma with posterior parameters

$$a_t = a_0 + t, \qquad b_t = b_0 + X_t,$$

as in (2.2)-(2.3). For convenience, we may also use the notation $k_t = (a_t, b_t)$. The value function $V(t, m)$ for the gamma-exponential problem can also be written as $V(a, m)$ under a shift of variable for simplicity as $a_t = a_0 + t$.

Theorem 4.1.1. *The value function $V(a, m)$ in gamma-exponential problem solves the free-boundary problem,*

$$\begin{aligned} V_a(a,m) - \frac{m}{a-1} V_m(a,m) + \int_0^\infty [V(a,m+z) - V(a,m)] \frac{1}{z} \left(\frac{m}{m+z}\right)^a dz &= cV(a,m) - (m-r) \\ V(a, m^*(a)) &= 0 \end{aligned}$$

where $m^*(a)$ is an unknown stopping boundary curve. For every point (a, m) on this stopping boundary, the Gittins index $R(a, m)$ is equal to r.

Proof: This can be shown through explicit calculation based on the PIDE in Theorem 3.3.3. In the conditional Lévy process we use to model exponential rewards, the conditional mean measure given λ is $\nu(\lambda, dy) = e^{-\lambda y}/y$, the same as in a gamma process, and the distribution of λ given $\mathcal{F}_t$ is $Gamma(a_t, b_t)$. Therefore, the unconditional mean measure $\bar{\nu}_t(dy)$ is calculated as

$$\begin{aligned} \bar{\nu}_t(dy) &= \int_0^\infty \frac{e^{-\lambda y}}{y} \cdot \frac{b_t^{a_t} \lambda^{a_t - 1} e^{-b_t \lambda}}{\Gamma(a_t)} d\lambda dy \\ &= \left(\frac{b_t}{b_t + y}\right)^{a_t} \frac{1}{y} dy \end{aligned}$$

and

$$\begin{aligned} m_t &= \int_0^\infty y \bar{\nu}_t(dy) \\ &= \int_0^\infty \left(\frac{b_t}{b_t + y}\right)^{a_t} dy \\ &= \frac{b_t}{a_t - 1} \end{aligned}$$

Therefore, $m_t = g(t, X_t) = \frac{b_0+X_t}{a_0+t-1}$ and $\frac{\partial g}{\partial t} = -\frac{b_0+X_t}{(a_0+t-1)^2} = -\frac{m_t}{a_t-1}$. Also,

$$\begin{aligned}
&\int_{R^+} [V(t, g(t, X_t + y)) - V(t, g(t, X_t))]\bar{\nu}_t(dy) \\
= &\int_{R^+} \left[V\left(t, \frac{b_t + y}{a_t - 1}\right) - V\left(t, \frac{b_t}{a_t - 1}\right)\right] \left(\frac{b_t}{b_t + y}\right)^{a_t} \frac{1}{y} dy \\
= &\int_{R^+} \left[V\left(t, m_t + z\right) - V\left(t, m_t\right)\right] \left(\frac{m_t}{m_t + z}\right)^{a_t} \frac{1}{z} dz
\end{aligned}$$

where the last equality is obtained by using a change of variable $z = \frac{y}{a_t-1}$. □

In the exponential reward problem, as well as the Poisson reward problem to be discussed in Section 4.2, we use integration by parts to reduce the free boundary partial integro-differential equation results from Theorem 3.3.3 to a simpler version. First, we simplify the value function,

$$\begin{aligned}
V(a, m) &= \sup_\tau \mathbb{E} \int_0^\tau e^{-cs} \left(m_s - r\right) ds \\
&= \sup_\tau \frac{1}{c} \mathbb{E} \left[(m - r) - e^{-c\tau} \left(m_\tau - r\right) + \int_0^\tau e^{-cs} \frac{d}{ds} m_s\right],
\end{aligned}$$

Observe that

$$\frac{d}{ds} m_s = \frac{d}{ds} \frac{b_0 + X_s}{a_0 + s - 1} = -\frac{b + X_s}{\left(a + s - 1\right)^2} ds + \frac{1}{a + s - 1} dX_s.$$

We take the expectation of this quantity, whence

$$\begin{aligned}
&\mathbb{E} \int_0^\tau e^{-cs} \frac{d}{ds} m\left(a_s, b_s\right) \\
= &-\mathbb{E} \int_0^\tau e^{-cs} \left(\frac{b_0 + X_s}{\left(a_0 + s - 1\right)^2}\right) ds + \mathbb{E} \int_0^\tau e^{-cs} \mathbb{E} \left[\frac{1}{a_0 + s - 1} dX_s \middle| \mathcal{F}_s\right] \\
= &-\mathbb{E} \int_0^\tau e^{-cs} \left(\frac{b_0 + X_s}{\left(a_0 + s - 1\right)^2}\right) ds + \mathbb{E} \int_0^\tau e^{-cs} \left(\frac{b_0 + X_s}{\left(a_0 + s - 1\right)^2}\right) ds \\
= &\ 0.
\end{aligned}$$

Consequently, (3.11) can be rewritten as

$$\frac{1}{c} \left[\sup_\tau \mathbb{E} \left[e^{-c\tau} \left(r - m_\tau\right)\right] + m - r\right] = 0. \tag{4.1}$$

We define a new value function $G(a,m) := \sup_{\tau} \mathbb{E}\left[e^{-c\tau}\left(r - m_\tau\right)\right] = cV(a,m) - m + r$ for fixed r, and plug it into Theorem 4.1.1 to obtain the following equivalent free boundary problem.

Proposition 4.1.2. *The value function $G\left(a,m\right)$ in the gamma-exponential problem solves the free-boundary problem*

$$\begin{aligned} G_a\left(a,m\right) - \frac{m}{a-1}G_m\left(a,m\right) + \int_0^\infty \left[G\left(a,m+z\right) - G\left(a,m\right)\right]\frac{1}{z}\left(\frac{m}{m+z}\right)^a dz &= cG\left(a,m\right) \\ G\left(a,m^*\left(a\right)\right) &= r - m^*\left(a\right) \end{aligned}$$

where $m^*\left(a\right)$ is an unknown stopping boundary curve. For every point (a,m) on this stopping boundary, the Gittins index $R(a,m)$ is equal to r.

Proof: By substituting $V(a,m) = \frac{1}{c}\left[G(a,m) + m - r\right]$ in Theorem 4.1.1, we get

$$\begin{aligned} & V_a\left(a,m\right) - \frac{m}{a-1}V_m\left(a,m\right) + \int_0^\infty \left[V\left(a,m+z\right) - V\left(a,m\right)\right]\frac{1}{z}\left(\frac{m}{m+z}\right)^a dz \\ = \; & \frac{1}{c}G_a\left(a,m\right) - \frac{1}{c}\frac{m}{a-1}\left[G_m\left(a,m\right) + 1\right] + \frac{1}{c}\frac{m}{a-1} \\ & + \frac{1}{c}\int_0^\infty \left[G\left(a,m+z\right) - G\left(a,m\right)\right]\frac{1}{z}\left(\frac{m}{m+z}\right)^a dz \\ = \; & \frac{1}{c}\left[G_a\left(a,m\right) - \frac{m}{a-1}G_m\left(a,m\right) + \int_0^\infty \left[G\left(a,m+z\right) - G\left(a,m\right)\right]\frac{1}{z}\left(\frac{m}{m+z}\right)^a dz\right] \end{aligned}$$

and

$$cV(a,m) - (m-r) = G(a,m)$$

while on the stopping boundary, $\left[G(a,m) + m - r\right]/c = 0$. □

The formulation in Proposition 4.1.2 is equivalent to the more intuitive one in Theorem 4.1.1 where $L^{inf}V = L^{char}V$ and the value function equals zero on the stopping boundary. We will use it for proving structural properties and for numerical convenience in Chapter 5.

4.2 Poisson Reward Problems

In the gamma-Poisson problem, the continuous-time interpolation (X_t) is a Poisson process with unknown rate λ. Again, we assume $\lambda \sim Gamma\,(a_0, b_0)$, let $\mathcal{F}_t$ be the σ-algebra generated by the path of (X_t) up to time t, and update the posterior parameters using

$$a_t = a_0 + X_t, \qquad b_t = b_0 + t,$$

as in (2.4)-(2.5).

Similar to in Section 4.1, we get the following free boundary PIDE through calculating m_t explicitly.

Theorem 4.2.1. *The value function $V\,(b, m)$ in the gamma-Poisson problem solves the free-boundary problem,*

$$\begin{aligned} V_b\,(b,m) - \frac{m}{b} V_m\,(b,m) + \left[V\left(b, m + \frac{1}{b}\right) - V\,(b,m)\right] m &= cV\,(b,m) - (m - r) \\ V\,(b, m^*\,(b)) &= 0 \end{aligned}$$

where $m^*\,(b)$ is an unknown stopping boundary curve. And, for every point (b, m) on this stopping boundary, the Gittins index $R(b, m)$ is equal to r.

Proof: It suffices to show through explicit calculation based on the PIDE in Theorem 3.3.3. In the conditional Lévy process we use to model exponential rewards, the conditional mean measure given λ is identical to that of a Poisson process $\nu(\lambda, dy) = \lambda\delta_1$, where δ_1 is the Dirac delta function, and the distribution of λ given

$\mathcal{F}_t$ is $Gamma(a_t, b_t)$. Therefore, the mean measure intensity $\bar{\nu}_t(dy)$ is calculated as

$$\begin{aligned}\bar{\nu}_t(dy) &= \int_0^\infty \lambda \cdot \frac{b_t^{a_t}\lambda^{a_t-1}e^{-b_t\lambda}}{\Gamma(a_t)} d\lambda\delta_1 dy \\ &= \frac{a_t}{b_t}\delta_1 dy\end{aligned}$$

and

$$\begin{aligned}m_t &= \int_0^\infty y\bar{\nu}_t(dy) \\ &= \frac{a_t}{b_t} = \frac{a_0 + X_t}{b_0 + t}\end{aligned}$$

Therefore, $m_t = g(t, X_t) = \frac{a_0+X_t}{b_0+t}$ and $\frac{\partial g}{\partial t} = -\frac{a_0+X_t}{(b_0+t)^2} = -\frac{m_t}{b_t}$. Also,

$$\begin{aligned}&\int_{R^+} [V(t, g(t, X_t + y)) - V(t, g(t, X_t))]\bar{\nu}_t(dy) \\ =& \int_{R^+} \left[V\left(t, \frac{a_t + y}{b_t}\right) - V\left(t, \frac{a_t}{b_t}\right)\right] \frac{a_t}{b_t}\delta_1 dy \\ =& \left[V\left(t, m_t + \frac{1}{b_t}\right) - V\left(t, m_t\right)\right] m_t\end{aligned}$$

whence the theorem is proved. □

Again, we use integration by parts to simplify the value function to get

$$V(b, m) = \frac{1}{c}\left[\sup_\tau \mathbb{E}\left[e^{-c\tau}\left(r - m_\tau\right)\right] + m - r\right] = 0. \tag{4.2}$$

By defining $G(b, m) := \sup_\tau \mathbb{E}\left[e^{-c\tau}\left(r - m_\tau\right)\right] = cV(b, m) - m + r$ for fixed r and replacing V in Theorem 4.2.1, we obtain the equivalent free boundary problem for the gamma-Poisson problem.

Proposition 4.2.2. *The value function $G(b, m)$ in the gamma-Poisson problem solves the free-boundary problem,*

$$\begin{aligned}G_b(b, m) - \frac{m}{b}G_m(b, m) + \left[V\left(b, m + \frac{1}{b}\right) - V(b, m)\right] m &= cG(b, m) \quad (4.3) \\ G(b, m^*(b)) &= r - m^*(b)\end{aligned}$$

where $m^*(b)$ *is an unknown stopping boundary curve. For every point* (b, m) *on this stopping boundary, the Gittins index* $R(b, m)$ *is equal to* r*.*

The proof is same as that of Proposition 4.1.2 and we omit it here. Unfortunately, the scaling properties of the Gittins index are not as straightforward in the gamma-Poisson problem as they are in the gamma-exponential problem. However, in the following section, we derive new scaling properties for the gamma-Poisson problem.

Chapter 5: Structural Properties of the New Approach

Continued from Chapter 4, in this chapter we provide more theoretical results on the structure of the Gittins index for the two non-Gaussian problems in continuous time. We focus on two major aspects. Section 5.1 considers scaling properties, primarily for the gamma-Poisson problem. Section 5.2 investigates the continuity and monotonicity of the Gittins index and value function, and concludes the theoretical analysis with an asymptotic convergence result. These theoretical properties provide us more insights on how the Gittins index, an indifference indexing rule, prices the value of a state via looking at its intrinsic value and uncertainty. Also, these properties match with the discrete-time results shown in [2, 62], showing that our results exhibit the correct structure established in the theory. At the end of this chapter, we also provide a numerical example to show the intuition behind our approach. Then, we conclude the dissertation with a brief discussion of the value and implications of our studies.

Throughout this chapter, we abuse notation slightly by writing value functions V, G, and the Gittins index R as a function of (t, m), (a, m), or (b, m), as is convenient. Most results apply to both gamma-exponential and gamma-Poisson problems, and therefore we use (t, m) most of the time. We will specifically use

(a, m) or (b, m) in proofs when needed. When we hold a parameter constant, we omit it in the argument, e.g. $V(m)$ denotes the value function $V(t, m)$ while we hold t constant. We use the subscript r for value functions, e.g. $V_r(t, m)$, to denote that they are calculated given that fixed r value. This notation facilitates writing our proofs in this chapter.

5.1 Distributional and scaling properties

We begin with two computational results on the predictive distributions appearing in the gamma-exponential and gamma-Poisson problems. These results are used in the proofs of some structural properties in this section, and later on help us to create initial conditions for PIDE solution procedures. The proofs can be found in the Appendix.

Lemma 5.1.1. *In the gamma-exponential model, the predictive distribution of* $\frac{X_t}{b_0}$*, given* $\mathcal{F}_0$*, is the beta-prime distribution with parameters* t *and* a_0*.*

Lemma 5.1.2. *In the gamma-Poisson model, the predictive distribution of* X_t*, given* $\mathcal{F}_0$*, is the generalized negative binomial distribution with parameters* a_0 *and* $\frac{t}{b_0+t}$*.*

Next, we establish scaling properties of the Gittins index for both non-Gaussian problems. Theorem 5.1.3 extends the result of (2.12) to the continuous-time setting, where the Gittins index is defined to be the value of r that solves (3.6); we include this proof for completeness. We then derive two different scaling properties for the gamma-Poisson problem. For the gamma-Poisson problem, we also emphasize the dependence of R on the discount factor c, as this plays a role in the scaling prop-

erties. To our knowledge, Theorem 5.1.4 is the first known scaling result for the gamma-Poisson problem.

Theorem 5.1.3. *In the gamma-exponential problem, the Gittins index satisfies* $R(a,b) = bR(a,1)$.

Proof: We factor b_0 out of the calibration equation (3.11) of the Gittins index by b_0 to obtain

$$\begin{aligned} & \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\left(m_s - r\right) ds\right] \\ = & \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\left(\frac{b_0 + X_s}{a_0 + s - 1} - r\right) ds\right] \\ = & \; b_0 \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\left(\frac{1 + \frac{X_s}{b_0}}{a_0 + s - 1} - \frac{r}{b_0}\right) ds\right] \qquad (5.1) \\ = & \; 0 \end{aligned}$$

The factor b_0 in (5.1) can be dropped since (5.1) equals zero. By applying the scaling properties of the gamma process and gamma distribution, we see that the process $\left(\frac{X_t}{b_0}\right)$ has the same law as a conditional gamma process with the prior $\lambda \sim Gamma\left(a_0, 1\right)$. Then, if R balances (3.11), it follows that the index $\frac{R}{b_0}$ balances the calibration equation for a gamma-exponential problem starting from the knowledge state $(a_0, 1)$. Thus, $R(a,b) = bR(a,1)$, as required. □

Theorem 5.1.4. *In the gamma-Poisson problem, the Gittins index satisfies the scaling property*

$$R(m,b,c) = \frac{1}{\sigma} R\left(\sigma m, \frac{b}{\sigma}, \sigma c\right)$$

for all $\sigma > 0$.

Proof: We consider the calibration equation for the gamma-Poisson problem and write

$$\begin{aligned} & \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\left(m_s - r\right) ds\right] \\ = & \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\left(\frac{a_0 + X_s}{b_0 + s} - r\right) ds\right] \\ = & \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}\left(\frac{a_0 + X_s}{\frac{b_0}{\sigma} + \frac{s}{\sigma}} - r\sigma\right) d\left(\frac{s}{\sigma}\right)\right]. \end{aligned} \tag{5.2}$$

Letting $t = \frac{s}{\sigma}$ and $Y_t = X_{\sigma t}$, we rewrite (5.2) as

$$\begin{aligned} & \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}(m_s - r) ds\right] \\ = & \sup_{\tau} \mathbb{E}\left[\int_0^{\frac{\tau}{\sigma}} e^{-c\sigma t}\left(\frac{a_0 + X_{\sigma t}}{\frac{b_0}{\sigma} + t} - r\sigma\right) dt\right] \\ = & \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-c\sigma t}\left(\frac{a_0 + Y_t}{\frac{b_0}{\sigma} + t} - r\sigma\right) dt\right] \end{aligned}$$

Observe that $\frac{\tau}{\sigma}$, where τ is a stopping time for X_t, is a stopping time for Y_t, and Y_t defines a conditional Poisson process with rate $\sigma\lambda$, which is equivalent to a conditional Poisson process with the prior $\lambda \sim Gamma\left(a_0, \frac{b_0}{\sigma}\right)$. This suggests a comparison with the calibration equation under discount factor $c\sigma$ and prior $\lambda \sim Gamma\left(a_0, \frac{b_0}{\sigma}\right)$, which yields the desired scaling property $R\left(m, b, c\right) = \frac{1}{\sigma} R\left(\sigma m, \frac{b}{\sigma}, \sigma c\right)$. □

Corollary 5.1.5. *From Theorem 5.1.4, it follows that*

$$R\left(m, b, c\right) = \frac{1}{b} R\left(mb, 1, bc\right) = cR\left(\frac{m}{c}, bc, 1\right).$$

Thus, we can scale either b or c to 1, but the other parameter will also be changed.

For the gamma-exponential problem, any Gittins index can be obtained by computing a family of stopping boundaries corresponding to $r = 1$ for each value of

c. In the gamma-Poisson problem, we can standardize the discount factor, but it is necessary to construct a family of curves indexed by b and m. Since the value of c is fixed throughout a given bandit problem, while the values of a and b change in each time step, the gamma-exponential problem will be less computationally intensive.

5.2 Continuity and monotonicity

In this section, we study various continuity and monotonicity properties of the value functions $V_r(t,m)$, $G_r(t,m)$, $R(t,m)$.

We are interested in continuity and monotonicity mainly for providing more intuitions on the Gittins index as an indifference pricing rule. By showing that the value functions and index values are monotonic on time parameter t and mean parameter m, we provide more insights on the fact that the index policy assigns higher value to states with higher belief mean m (intrinsic value or immediate payoff that can be exploited) and more uncertainty under smaller t (potential gain from exploration). Even though the non-Gaussian problems we study involve processes with jumps, the feature that, the Gittins index analyzes the expected payoff when an alternative is optimally played, suggests that the index change continuously with respect to the state parameters t and m (a continuous mean field).

Another use of results in this section has to do with the numerical solution of the PIDEs in Section 5.3. Recall from Lemma 2.3.1 that $R(\infty,m)=m$. That is, $m_t \to m_\infty$ a.s. by the law of large numbers, where m_∞ is the true mean reward (e.g. for the gamma-exponential problem we have $m_\infty = \frac{1}{\lambda}$ where λ is the true

rate value). Consequently, if our estimate of the mean at time infinity is m, this means that the true mean reward is equal to m, and the Gittins index must also equal this value. The monotinicity and continuity properties of r eventually imply that $R(a_t, m_t)$ converges to the true mean m_∞. It follows that, for a fixed r, the stopping boundary curves $m^*(a)$ and $m^*(b)$ in Theorems 4.1.1-4.2.2 will converge to r as a and b grow large, providing us with initial estimates of values at the stopping boundary for large time values. The continuity of the Gittins index for discrete-time problems has been studied e.g. by [63]. Such proofs typically use induction arguments that do not apply in continuous time. We mostly consider continuous-time problems, but we also provide new discrete-time results.

First, we recall that the predictive distributions of X_t have been given in Lemmas 5.1.1 and 5.1.2. In computing $V_r(t, m)$, $G_r(t, m)$, $R(t, m)$, only information in the current knowledge state (t, m) is used, so when we write m_t we are implying the predictive distribution of (m_t) conditioned on $\mathcal{F}_0$. We adopt this viewpoint throughout the following analysis. Starting with the next result, we will repeatedly compare two arbitrary prior knowledge states. Let (m_t) denote the process starting with the prior parameters (t_0, m_0), and let (m'_t) denote the process starting with (t'_0, m'_0).

Our proofs in this section are heavily based on stochastic dominance theory [64, 65], also used by [62] to establish discrete-time results. We shall follow the notation used in [64] and will use the usual stochastic order $\leq_{st}$, the convex order $\leq_{cx}$, and the increasing convex order $\leq_{icx}$. For random variables X and Y, $X \leq_{st} Y$ if $f_X(c)/f_Y(c)$ is decreasing in c. Also, $X \leq_{cx} Y$ (respectively, $X \leq_{icx} Y$) if

$\mathbb{E}\phi(X) \leq \mathbb{E}\phi(Y)$ for all convex functions (respectively, convex and increasing) ϕ. Useful properties that we will use include the equivalent definition $X \leq_{st} Y$ if $F_X \geq F_Y$ [1.2.1 in 64], the implication $\leq_{icx} \Rightarrow \leq_{cx}$ when $\mathbb{E}X = \mathbb{E}Y$, and the coupling techniques that will be re-stated into Lemmas 5.2.2 and 5.2.3 in this section.

Lemma 5.2.1. *The two following stochastic order properties hold for predictive mean processes, for every t:*

$$m_t \geq_{st} m'_t \text{ if } m_0 \geq m'_0 \text{ and } t_0 = t'_0 \tag{5.3}$$

$$m_t \geq_{cx} m'_t \text{ if } m_0 = m'_0 \text{ and } t_0 \leq t'_0 \tag{5.4}$$

Proof: We prove (5.3) first. It suffices to show that, when $m_0 \geq m'_0$ and $t_0 = t'_0$, we have $F_{m_t} \leq F_{m'_t}$.

For the gamma-exponential problem, $t_0 = t'_0$ implies $a_0 = a'_0$, and we denote this common value by a. By Lemma 5.1.1 we have

$$\begin{aligned} P(m_t \geq m) &= P\left(\frac{b_0 + X_t}{a + t - 1} \geq m \middle| a, m_0\right) \\ &= 1 - F\left(\frac{m \cdot (a + t - 1)}{m_0 \cdot (a - 1)} - 1\right) \end{aligned}$$

and

$$\begin{aligned} P(m'_t \geq m) &= P\left(\frac{b'_0 + X_t}{a + t - 1} \geq m \middle| a, m'_0\right) \\ &= 1 - F\left(\frac{m \cdot (a + t - 1)}{m'_0 \cdot (a - 1)} - 1\right) \end{aligned}$$

where F is the cdf of the $Beta'(t, a)$ distribution. When $m_0 \geq m'_0$,

$$\frac{m \cdot (a + t - 1)}{m_0 \cdot (a - 1)} \leq \frac{m \cdot (a + t - 1)}{m'_0 \cdot (a - 1)}.$$

and therefore $P(m_t \geq m) \geq P(m'_t \geq m)$, i.e. $F_{m_t} \leq F_{m'_t}$.

For the gamma-Poisson problem, $t_0 = t'_0$ implies $b_0 = b'_0$, and we denote this common value by b. By Lemma 5.1.2 we have

$$\begin{aligned} P(m_t \geq m) &= P\left(\frac{a_0 + X_t}{b+t} \geq m \middle| b, m_0\right) \\ &= 1 - F\left(m \cdot (b+t) - m_0 b\right) \end{aligned}$$

and

$$\begin{aligned} P(m'_t \geq m) &= P\left(\frac{a'_0 + X_t}{b+t} \geq m \middle| b, m'_0\right) \\ &= 1 - F'\left(m \cdot (b+t) - m'_0 b\right) \end{aligned}$$

where F is the cdf of the generalized negative binomial (GNB) distribution with parameters $m_0 b$ and $\frac{t}{b+t}$, and F' is the cdf of GNB $\left(m'_0 b, \frac{t}{b+t}\right)$. When $m_0 \geq m'_0$,

$$\begin{aligned} F(m \cdot (b+t) - m_0 b) &\leq F'(m \cdot (b+t) - m_0 b) \\ &\leq F'(m \cdot (b+t) - m'_0 b), \end{aligned}$$

whence $P(m_t \geq m) \geq P(m'_t \geq m)$ as required.

Secondly, we prove (5.4). We first consider the gamma-exponential case; the gamma-Poisson version can be shown in exactly the same way.

In the gamma-exponential case, (5.4) assumed that $m_0 = \frac{b_0}{a_0 - 1} = \frac{b'_0}{a'_0 - 1} = m'_0$, which we denote by m, and $a_0 \leq a'_0$. We prove convex dominance by showing

$$\begin{aligned} m_t &= \left(\frac{b_0 + X_t}{a_0 + t - 1} \middle| \lambda \sim Gamma(a_0, b_0)\right) \\ &\geq_{cx} \left(\frac{b'_0 + X_t}{a'_0 + t - 1} \middle| \lambda \sim Gamma(a_0, b_0)\right) \qquad (5.5) \\ &\geq_{cx} \left(\frac{b'_0 + X_t}{a'_0 + t - 1} \middle| \lambda \sim Gamma(a'_0, b'_0)\right) \qquad (5.6) \\ &= m'_t \end{aligned}$$

We observe that

$$\begin{aligned} & \left(\frac{b_0+X_t}{a_0+t-1}\middle|\lambda\sim Gamma(a_0,b_0)\right) \\ = & \left(\frac{b_0+mt+(X_t-mt)}{a_0+t-1}\middle|\lambda\sim Gamma(a_0,b_0)\right) \\ = & \ m+\frac{1}{a_0+t-1}\left(X_t-mt|\lambda\sim Gamma(a_0,b_0)\right) \end{aligned}$$

and similarly

$$\left(\frac{b_0'+X_t}{a_0'+t-1}\middle|\lambda\sim Gamma(a_0,b_0)\right)=m+\frac{1}{a_0'+t-1}\left(X_t-mt|\lambda\sim Gamma(a_0,b_0)\right)$$

where $(X_t-mt|\lambda\sim Gamma(a_0,b_0))$ is a random variable with zero mean. If we write $Y_t:=(X_t-mt|\lambda\sim Gamma(a_0,b_0))$, then to prove (5.5) it suffices to show

$$m+\frac{1}{a_0+t-1}Y_t\geq_{cx} m+\frac{1}{a_0'+t-1}Y_t.$$

By Theorem 1.5.18 in [64], for a zero mean random variable X, $aX+b\leq_{icx} cX+d$, when $0\leq a\leq c$ and $b\leq d$. Since $\frac{1}{a_0+t-1}\geq\frac{1}{a_0'+t-1}$, we have

$$\left(\frac{b_0+X_t}{a_0+t-1}\middle|\lambda\sim Gamma(a_0,b_0)\right)\geq_{icx}\left(\frac{b_0'+X_t}{a_0'+t-1}\middle|\lambda\sim Gamma(a_0,b_0)\right),$$

and then $\geq_{cx}$ follows from the fact that they have equal means, whence (5.5) is proved.

Next, (5.6) follows from Theorem 3.A.21 in [65]. It suffices to prove the condition of the theorem that, for every convex function ϕ, $\mathbb{E}\left[\phi(X_t|\frac{1}{\lambda})\right]$ is convex in $\frac{1}{\lambda}$ for gamma-exponential problem, and $\mathbb{E}\left[\phi(X_t|\lambda)\right]$ is convex in λ for gamma-Poisson.

In the gamma-exponential case, for all $\theta \geq \theta'$ and $\alpha \in (0,1)$,

$$\begin{aligned}
& \mathbb{E}\left[\phi\left(X_t \middle| \frac{1}{\lambda} = \alpha\theta + (1-\alpha)\theta'\right)\right] \\
= \; & \mathbb{E}\left\{\phi\left[\left(X_t \middle| \frac{1}{\lambda} = \alpha\theta\right) + \left(X_t \middle| \frac{1}{\lambda} = (1-\alpha)\theta'\right)\right]\right\} && (5.7) \\
= \; & \mathbb{E}\left\{\phi\left[\alpha\left(X_t \middle| \frac{1}{\lambda} = \theta\right) + (1-\alpha)\left(X_t \middle| \frac{1}{\lambda} = \theta'\right)\right]\right\} && (5.8) \\
\leq \; & \mathbb{E}\left[\alpha\phi\left(X_t \middle| \frac{1}{\lambda} = \theta\right) + (1-\alpha)\phi\left(X_t \middle| \frac{1}{\lambda} = \theta'\right)\right] && (5.9) \\
= \; & \alpha\mathbb{E}\left[\phi\left(X_t \middle| \frac{1}{\lambda} = \theta\right)\right] + (1-\alpha)\mathbb{E}\left[\phi\left(X_t \middle| \frac{1}{\lambda} = \theta'\right)\right]
\end{aligned}$$

in which (5.7) and (5.8) are due to scaling properties of the gamma distribution, and (5.9) is due to ϕ being convex. Therefore, Theorem 3.A.21 of [65] holds, whence (5.6) is proved.

With (5.5) and (5.6) shown, (5.4) is proved (the gamma-Poisson case is proved in exactly the same way and we omit it). □

Before we provide our first monotonicity result, we restate two results from [64] for completeness. These results are also known as the "coupling" techniques.

Lemma 5.2.2. *If $(m_t) \geq_{st} (m'_t)$ for all t, there exist two processes $(\hat{m}_t)$ and $(\hat{m}'_t)$ defined on the same filtration $\mathcal{F}_t$ that are identical in distribution to (m_t) and (m'_t), and $\hat{m}_t \geq \hat{m}'_t$ almost surely [64, Theorem 1.2.4].*

Lemma 5.2.3. *If $(m_t) \geq_{cx} (m'_t)$ for all t, there exist two processes $(\hat{m}_t)$ and $(\hat{m}'_t)$ defined on the same filtration $\mathcal{F}_t$ that are identical in distribution to (m_t) and (m'_t), and $\mathbb{E}(\hat{m}_t|\hat{m}'_t) = \hat{m}'_t$ [64, Theorem 3.4.2].*

We will repeatedly use these coupling techniques when proving monotonicity and continuity properties in the rest of this section. When we take two initial

states (t_0, m_0) and (t_0', m_0') that generate two predictive mean processes m_t and m_t' satisfying stochastic dominance $\leq_{st}$ or convex dominance $\leq_{cx}$, Lemmas 5.2.2 and 5.2.3 give us two processes defined on the same filtration with a.s. dominance or the conditional expectation property, respectively. We will denote these two coupled processes by $\hat{m}_t$ and $\hat{m}_t'$. When we take a sequence of states (t_1, m_1), (t_2, m_2),..., that all dominate or are dominated by some state (t, m), each state in the sequence can be coupled with (t, m), and we denote the coupled process of (t_k, m_k) by $\hat{m}_t^k$.

Theorem 5.2.4. *$V(m)$ is increasing in m, and $G(m)$ is decreasing in m.*

Proof: Assume that $m_0 \geq m_0'$ and $t_0 = t_0'$. Then, Lemma 5.2.2 gives us two processes defined on the same filtration with a.s. dominance. The processes (m_t) and $(\hat{m}_t)$ are identically distributed, as are (m_t') and $(\hat{m}_t')$. Using the arguments of [66], the values of V and G, as well as the optimal stopping time τ, depend only on the law of m_t. This result is also given in [67]. Therefore,

$$\begin{aligned} \sup_\tau \mathbb{E}\left[e^{-c\tau}(r - m_\tau)\right] &= \sup_\tau \mathbb{E}\left[e^{-c\tau}(r - \hat{m}_\tau)\right], \\ \sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s - r)ds\right] &= \sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs}(\hat{m}_s - r)ds\right]. \end{aligned}$$

This allows us to write V and G using $\hat{m}_t$ and $\hat{m}_t'$, which provides the almost sure dominance necessary to complete the proof, that is, $\hat{m}_t(\omega) \geq \hat{m}_t'(\omega)$ for a.e. ω. We

calculate

$$\begin{aligned} V_r(m_0') &= \sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s' - r)ds\right] \\ &= \sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs}(\hat{m}_s' - r)ds\right] \\ &= \sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs}(\hat{m}_s - r)ds + \int_0^\tau e^{-cs}(\hat{m}_s' - \hat{m}_s)ds\right] \\ &\leq \sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs}(\hat{m}_s - r)ds\right] \\ &= V_r(m_0) \end{aligned}$$

and

$$\begin{aligned} G_r(m_0) &= \sup_\tau \mathbb{E}\left[e^{-c\tau}(r - \hat{m}_\tau)\right] \\ &= \sup_\tau \mathbb{E}\left[e^{-c\tau}(r - \hat{m}_\tau') + e^{-c\tau}(\hat{m}_\tau' - \hat{m}_\tau)\right] \\ &\leq \sup_\tau \mathbb{E}\left[e^{-c\tau}(r - m_\tau')\right] \\ &= G_r(m_0'), \end{aligned}$$

as required. □

The monotonicity results for V and G can be used to obtain similar results for the stopping boundaries of the PIDEs, as well as the Gittins indices. Below, we find that the Gittins index is increasing in the mean parameter m, matching the result of [62] for discrete time.

Proposition 5.2.5. *The stopping boundaries $m_r^*(t)$, indexed by the retirement reward r, are ordered and do not cross. That is, $m_r^* \geq m_{r'}^*$ for $r \geq r'$.*

Proof: Let m_r^* be the stopping boundary corresponding to r and take $r' \leq r$. Then,

$$\begin{aligned}\sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s - r')ds\right] &= \sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s - r) + e^{-cs}(r - r')ds\right] \\ &\geq \sup_\tau \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s - r)\right] \\ &= 0.\end{aligned}$$

Therefore, $V_{r'}(m_r^*) \geq 0$. By monotonicity in Theorem 5.2.4, we get $m_r^* \geq m_{r'}^*$. □

Corollary 5.2.6. *From Proposition 5.2.5, it follows that the Gittins index R is increasing in m.*

With the monotonicity in m proved, we are now able to show that R is continuous in m.

Theorem 5.2.7. *$R(m)$ is continuous in m.*

Proof: Monotonicity in Corollary 5.2.6 guarantees the existence of $\lim_{\epsilon\to 0^-} R(m+\epsilon)$ and $\lim_{\epsilon\to 0^+} R(m+\epsilon)$ provided $R(m)$ is finite, and it suffices to show that they are all equal, i.e. $\lim_{\epsilon\to 0^-} R(m+\epsilon) = \lim_{\epsilon\to 0^+} R(m+\epsilon) = R(m)$.

First, we prove left-continuity. For any fixed t, we take an infinite increasing sequence of values $\{m_k\}$ converging to m from the left, and denote the corresponding Gittins indices $R(t, m_k)$ by R_k. We also denote the Gittins index corresponding to (t, m) by R. Then, taking the limit of both sides of (4.2) yields $\lim_{k\to\infty} R_k = \lim_{k\to\infty} m_k + \lim_{k\to\infty} G_{R_k}(m_k)$. We denote $\lim_{k\to\infty} R_k$ by $\bar{R}$. By Proposition 5.2.5, $\bar{R} \leq R$.

Now we show that $\bar{R} \geq R$.

$$\begin{aligned}
\bar{R} &= \lim_{k\to\infty} m_k + \lim_{k\to\infty} G_{R_k}(m_k) \\
&= m + \lim_{k\to\infty} G_{R_k}(m_k) \\
&= m + \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - m_\tau^k)\right] \\
&= m + \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau + \hat{m}_\tau - \hat{m}_\tau^k)\right] \qquad (5.10) \\
&\geq m + \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau)\right] \qquad (5.11) \\
&\geq m + \sup_{\tau} \lim_{k\to\infty} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau)\right] \qquad (5.12) \\
&= m + \sup_{\tau} \left[\mathbb{E}\left(e^{-c\tau}(\bar{R} - \hat{m}_\tau)\right) + \lim_{k\to\infty} \mathbb{E}\left(e^{-c\tau}(R_k - \bar{R})\right)\right] \\
&= m + \sup_{\tau} \mathbb{E}[e^{-c\tau}(\bar{R} - \hat{m}_\tau)]
\end{aligned}$$

In (5.10), we used the coupling technique in Lemma 5.2.2 to map the predictive processes m_t and m_t^k onto the same filtration and obtain almost sure dominance, which provides the inequality (5.11). Equation (5.12) is due to

$$\sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau)\right] \geq \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau)\right]$$

for every k and therefore

$$\lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau)\right] \geq \lim_{k\to\infty} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau)\right]$$

for each τ. This yields

$$\lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau)\right] \geq \sup_{\tau} \lim_{k\to\infty} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau)\right]$$

Therefore, we have

$$m - \bar{R} + \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(\bar{R} - \hat{m}_\tau)\right] = \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s - \bar{R})ds\right] \leq 0 \qquad (5.13)$$

By

$$V_R(m) = \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}(m_s - R)ds\right] = 0$$

we have

$$\begin{aligned} 0 &= \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}(m_s - R)ds\right] \\ &= \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}(m_s - \bar{R} + \bar{R} - R)ds\right] \\ &\leq \sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}(m_s - \bar{R})ds\right] + (\bar{R} - R)\sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}ds\right] \end{aligned}$$

which by (5.13) leads to

$$\begin{aligned} (\bar{R} - R)\sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}ds\right] &\geq -\sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}(m_s - \bar{R})ds\right] \\ &\geq 0 \end{aligned}$$

Since $\sup_{\tau} \mathbb{E}\left[\int_0^{\tau} e^{-cs}ds\right] \geq 0$, we have $(\bar{R} - R) \geq 0$ and therefore $\bar{R} \geq R$, whence left-continuity is proved.

Right-continuity can be proved in a similar way. For any m and t fixed, take an infinite increasing sequence of values $\{m_k\}$ converging to m from the right, and

under the same notation we show $\bar{R} \leq R$.

$$\begin{aligned}
\bar{R} &= \lim_{k\to\infty} m_k + \lim_{k\to\infty} G_{R_k}(m_k) \\
&= m + \lim_{k\to\infty} G_{R_k}(m_k) \\
&= m + \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - m_\tau^k)\right] \\
&= m + \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau + \hat{m}_\tau - \hat{m}_\tau^k)\right] \\
&\leq m + \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - \hat{m}_\tau)\right] \\
&= m + \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(\bar{R} - \hat{m}_\tau + R_k - \bar{R})\right] \\
&\leq m + \lim_{k\to\infty} \{\sup_{\tau} \mathbb{E}\left[e^{-c\tau}(\bar{R} - \hat{m}_\tau)\right] + \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(R_k - \bar{R})\right]\} \\
&= m + \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(\bar{R} - \hat{m}_\tau)\right] + \lim_{k\to\infty}\left[(R_k - \bar{R}) \sup_{\tau} \mathbb{E}\left(e^{-c\tau}\right)\right] \\
&= m + \sup_{\tau} \mathbb{E}\left[e^{-c\tau}(\bar{R} - \hat{m}_\tau)\right]
\end{aligned}$$

This shows that $V_{\bar{R}}(m) = \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s - \bar{R})ds\right] \geq 0$, and therefore

$$\begin{aligned}
0 &\leq \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s - \bar{R})ds\right] \\
&= \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s - R + R - \bar{R})ds\right] \\
&\leq \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}(m_s - R)ds\right] + (R - \bar{R}) \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}ds\right]
\end{aligned}$$

which leads to

$$(R - \bar{R}) \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}ds\right] \geq 0$$

whence right continuity is proved. □

Lemma 5.2.8. *The Gittins index $R(t, m)$ is monotonically decreasing in t, while holding m fixed.*

Proof: Under the convex order provided in (5.4), the proposition follows directly from Theorem 5.4 in [68]. □

Theorem 5.2.9. *$R(t)$ is continuous in t.*

Proof: Monotonicity in Lemma 5.2.8 guarantees the existence of $\lim_{\epsilon\to 0^-} R(t+\epsilon)$ and $\lim_{\epsilon\to 0^+} R(t+\epsilon)$ provided $R(t)$ is finite, and it suffices to show that they are all equal, i.e. $\lim_{\epsilon\to 0^-} R(t+\epsilon) = \lim_{\epsilon\to 0^+} R(t+\epsilon) = R(t)$.

First, we prove left-continutity. For any m fixed, take an infinite increasing sequence of values $\{t_k\}$ converging to t from the left, and denote corresponding Gittins indices $R(t_k, m)$ by R_k. We denote $\lim_{k\to\infty} R_k$ by $\bar{R}$. By Proposition 5.2.5, we have $\bar{R} \leq R$. Now we show that $\bar{R} \geq R$.

By taking limit on both sides of the calibration equation yields

$$
\begin{aligned}
0 &= \lim_{k\to\infty} V_{R_k}(a_k, m) \\
&= \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(m_s^k - R_k\right) ds\right] \\
&= \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(\hat{m}_s^k - R_k\right) ds\right] \\
&= \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(\hat{m}_s^k - \hat{m}_s + \hat{m}_s - R + R - R_k\right) ds\right] \\
&\geq \sup_{\tau} \lim_{k\to\infty} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(\hat{m}_s^k - \hat{m}_s + \hat{m}_s - R + R - R_k\right) ds\right] \\
&= \sup_{\tau}\Bigg\{ \lim_{k\to\infty} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(\hat{m}_s^k - \hat{m}_s\right) ds\right] + \mathbb{E}\left[\int_0^\tau e^{-cs}\left(\hat{m}_s - R\right) ds\right] \\
&\qquad + \left(R - \bar{R}\right) \lim_{k\to\infty} \mathbb{E}\int_0^\tau e^{-cs} ds \Bigg\} \\
&= \sup_{\tau}\left\{ \lim_{k\to\infty} \mathbb{E}\left[\int_0^\tau e^{-cs}\mathbb{E}\left(\hat{m}_s^k - \hat{m}_s | \hat{m}_s\right) ds\right] + \left(R - \bar{R}\right) \lim_{k\to\infty} \mathbb{E}\int_0^\tau e^{-cs} ds \right\} \\
&= \left(R - \bar{R}\right) \lim_{k\to\infty} \mathbb{E}\int_0^\tau e^{-cs} ds \qquad (5.14)
\end{aligned}
$$

By Lemma 5.2.1, $\hat{m}_t \leq_{cx} \hat{m}_t^k$ for every t, and hence $\mathbb{E}\left(\hat{m}_s^k - \hat{m}_s | \hat{m}_s\right) = 0$ by Lemma 5.2.3, leading to equation (5.14). Therefore $R - \bar{R} \leq 0$, whence left continuity is proved.

Right-continuity can be proved in a similar way. For any m and t fixed, take an infinite increasing sequence of values $\{t_k\}$ converging to t from the right, and under the same notation we show $\bar{R} \leq R$. By taking limit on both sides of the calibration equation yields

$$
\begin{aligned}
0 &= \lim_{k\to\infty} V_{R_k}(a_k, m) \\
&= \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(m_s^k - R_k\right) ds\right] \\
&= \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(\hat{m}_s^k - R_k\right) ds\right] \\
&= \lim_{k\to\infty} \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(\hat{m}_s^k - \hat{m}_s + \hat{m}_s - R + R - R_k\right) ds\right] \\
&\leq \lim_{k\to\infty} \left\{ \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(\hat{m}_s^k - \hat{m}_s\right) ds\right] + \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs}\left(\hat{m}_s - R\right) ds\right] \right. \\
&\qquad \left. + (R - R_k) \sup_{\tau} \mathbb{E} \int_0^\tau e^{-cs} ds \right\} \\
&= \lim_{k\to\infty} \left\{ \sup_{\tau} \mathbb{E}\left[\int_0^\tau e^{-cs} \mathbb{E}\left(\hat{m}_s^k - \hat{m}_s | \hat{m}_s^k\right) ds\right] + \left(R - \bar{R}\right) \sup_{\tau} \mathbb{E} \int_0^\tau e^{-cs} ds \right\} \\
&= \left(R - \bar{R}\right) \sup_{\tau} \mathbb{E} \int_0^\tau e^{-cs} ds
\end{aligned}
$$

whence right-continuity is proved. □

Theorem 5.2.10. *The Gittins index* $\lim_{t\to\infty} R(t, m) = m$ *for each* m *fixed, and* $R(t, m_t)$ *converges to* m_∞ *as* $t \to \infty$ *almost surely.*

Proof: By Theorem 5.2.7 and 5.2.9, $R(t, m)$ is continuous in (t, m). As $t \to \infty$, $m_t \to m_\infty$ almost surely. Therefore, the theorem follows from Lemma 2.3.1. □

5.3 Numerical Illustrations

Recall from Chapter 3 that, it suffices to solve only one PIDE in 4.1.2 under $r = 1$ for the gamma-exponential problem to obtain all Gittins indices under its scaling property. In this section, we use it as a numerical example to illustrate the main insights of the theoretical properties of the framework we developed throughout this dissertation. We will observe intuitively how the Gittins index is found when the solution surface of a PIDE hits the stopping boundary.

Solving the problems in Theorems 4.1.2 and 4.2.2 numerically poses a substantial challenge, because we do not know the stopping boundary or even the exact value of V at any point, making it difficult to define suitable initial conditions. We implement an approximation that gives a lower bound on the value function, based on a one-stage stopping rule (also used by [46]). For deterministic $B \geq 0$, define the stopping time τ_B as follows. Starting from an initial set of parameters at time 0, we observe the process (X_t) until time B. If $m_B < r$, we retire, and if $m_B \geq r$, we continue running the process until infinity. We then calculate the value achieved by τ_B, given by the quantity

$$\bar{G}_B = \mathbb{E}[e^{-cB}(r - m_B)^+], \tag{5.15}$$

and use $\sup_B \bar{G}_B$ to approximate the value of G for the prior parameters. For both gamma-exponential and gamma-Poisson models, (5.15) can be computed in closed form, and $\sup_B \bar{G}_B$ is relatively easy to calculate numerically. The proofs of the following results are given in the Appendix.

Proposition 5.3.1. *In the gamma-exponential model,*

$$\bar{G}_B = e^{-cB} \frac{b_0}{A+1} \int_0^A F(s)ds$$

where $A = \frac{r(a_0+B-1)}{b_0} - 1$*, and* F *is the cdf of a Beta prime distribution with parameters* B *and* a_0*.*

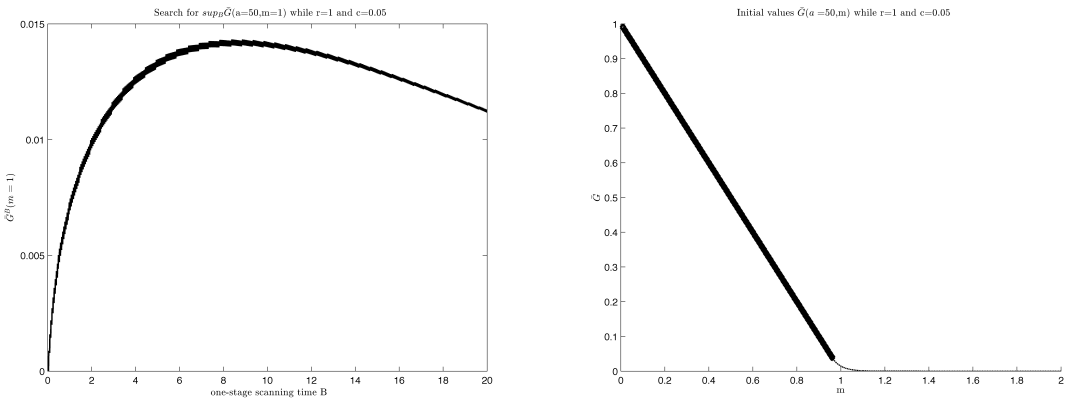
Proposition 5.3.2. *In the gamma-Poisson model,*

$$\bar{G}_B = \frac{e^{-cB}}{b_0+B} \left[\sum_{k \leq A} F(K) - (\lceil A \rceil - A) F(\lfloor A \rfloor) \right]$$

where $A = rb_0 + rB - m_0 b_0$*, and* F *is the cdf of a generalized negative binomial distribution with parameters* a_0 *and* $\frac{B}{b_0+B}$*.*

We use these results to calculate the initial conditions at (a, m) for fixed a and all $m > 0$. The following figures illustrate the one-stage stopping rule and the search for a lower bound more intuitively, through a gamma-exponential example with $r = 1$ and $c = 0.05$. First, Figure 1(a) shows that the approximation $\bar{G}_B$ is unimodal for $B \in [0, 20]$ with $a = 50$ and $m = 1$. The maximum value of this curve is then implemented as an approximation for $G(a, m)$ with $a = 50$ and $m = 1$. Figure 1(b) shows the results of this procedure for all m values, with $a = 50$ fixed. The bold line segment shows that the initial-value approximation is close to the stopping trigger value $r - m$ with high precision when m is low. The tail curve approaching zero shows where the approximation starts to deviate from $r - m$. In the stopping problem, the section in bold would correspond to the stopping region, while the other section corresponds to the continuation region.

Using the lower bound to approximate the initial value of G, we solve the PIDEs numerically using Euler's finite difference schemes. It is preferable to calculate

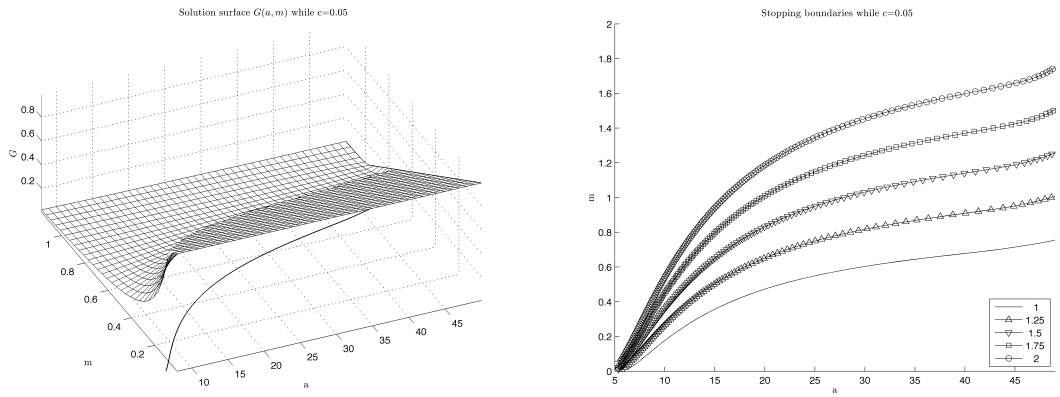


(a) Initial value

(b) One-stage searching

Figure 5.1: Demonstration of initial values obtained from the one-stage searching method

the initial value approximation for large time values, since the quality of the lower bound $\sup_B \bar{G}_B$ is much better when the relevant time parameter (a or b) is large. The PIDEs can be modified to express the dynamics for moving backward in time rather than forward. Figure 2(a) demonstrates the solution surface to the PIDE for $r = 1$, $c = 0.05$, and the initial value approximation (the right edge of the surface) with $a = 50$. The surface was created by propagating the initial value curve from Figure 1(a) from $a = 50$ backward to $a = 1$. The solution surface is stopped and cut off when it hits the tilted plane $G(a, m) = r - m$. The curve is the stopping boundary, a projection of the surface values on this hitting plane onto the (a, m) plane. Figure 2(b) shows boundary curves for several values of r, all with initial conditions set at $a = 50$. Each of these curves represents the set of all knowledge states whose Gittins index is precisely equal to the given r value; for any knowledge



(a) Solution surface

(b) Stopping Boundaries

Figure 5.2: Stopping boundaries of the value function and 2D plots for different r values

state above the curve, we prefer to continue collecting rewards from the process (X_t), whereas for any knowledge state below the curve, we prefer to stop and accrue the fixed reward r instead.

We briefly mention some properties of the solution to the PIDE. We can see that the stopping boundary $m^*(a)$ described by Theorems 3.3.3 and 4.1.2 should converge to the retirement value r as the time parameter becomes large. Therefore, the curves in Figure 2(b) behave as expected, increasing over time but remaining dominated by their r values. We also note that the boundary curves appear to be concave; the slight bumps close to $a = 50$ are due to numerical issues stemming from proximity to the initial value. It is clear that the key to such procedures is the ability to find good boundary curves. However, the results in Figure 2 demonstrate that the numerical solution behaves in accordance with our intuition about the problem.

5.4 Conclusion

We have presented a theoretical framework that can be used to approximate the computation of optimal policies for multi-armed bandit problems with non-Gaussian rewards. The foundation of our approach consists of constructing continuous-time, conditional Lévy processes that serve as probabilistic interpolations of the discrete-time reward processes in the bandit problem. This idea was previously used in the Gaussian setting, where the properties of Brownian motion allow for easy standardization and numerical solution of a stopping problem in continuous-time. Although these techniques are not available in the non-Gaussian setting, we have shown that the analogous stopping problems can be represented as free-boundary problems on PIDEs that equate the characteristic and infinitesimal operators of the relevant value function. We have also proved the structural properties of the value functions in these free-boundary problems, as well as the Gittins indices in continuous time. Theses properties match with the discrete-time results and show that our results exhibit the correct structure established in the theory. We also presented numerical illustrations showing the intuitive implications on how the free-boundary PIDE connects to the original Gittins index problem. Our approach is especially promising in the gamma-exponential case, where the Gittins index enjoys scaling properties.

While this is outside the scope of the dissertation, the framework we have presented can be intuitively extended and incorporated into more general reward processes and stopping problems, such as those in [18]. The value functions can

easily accommodate different uses, while the interpolation and optimal stopping to free-boundary transition techniques remain the same.

Chapter Appendix:

In this appendix, we include all proofs of lemmas and theorems that were not included in the main body of this paper.

Proof of Lemma 5.1.1: In the gamma-exponential model, $X_t \sim Gamma(t, \lambda)$, and $\lambda \sim Gamma\,(a_0, b_0)$ given $\mathcal{F}_0$. Therefore, the predictive distribution of X_t is

$$\begin{aligned}
P\left(X_t \in dx\right) &= \int_0^\infty \frac{b_0(b_0\lambda)^{a_0-1}e^{-b_0\lambda}}{\Gamma(a_0)}\frac{\lambda(\lambda x)^{t-1}e^{-\lambda x}}{\Gamma(t)}d\lambda dx \\
&= \frac{\Gamma(a_0+t)b^{a_0}x^{t-1}}{\Gamma(t)\Gamma(a_0)(x+b_0)^{a_0+t}}\int_0^\infty \frac{(x+b_0)^{a_0+t}\lambda^{a_0+t-1}e^{-(x+b_0)\lambda}}{\Gamma(a_0+t)}d\lambda dx \\
&= \frac{\Gamma(a_0+t)b_0^{a_0}x^{t-1}}{\Gamma(t)\Gamma(a_0)(x+b_0)^{a_0+t}}dx \\
&= \frac{\Gamma(a_0+t)b_0^{a_0}}{\Gamma(t)\Gamma(a_0)}\cdot\frac{x^{t-1}}{(x+b_0)^{a_0+t}}dx \\
&= \frac{\Gamma(a_0+t)}{\Gamma(t)\Gamma(a_0)}\cdot\frac{\left(\frac{x}{b_0}\right)^{t-1}}{\left(1+\frac{x}{b_0}\right)^{a_0+t}}d\left(\frac{x}{b_0}\right) \\
&= \frac{1}{Beta(t,a_0)}\frac{\left(\frac{x}{b_0}\right)^{t-1}}{\left(1+\frac{x}{b_0}\right)^{a_0+t}}d\left(\frac{x}{b_0}\right),
\end{aligned}$$

which is the beta-prime density with parameters t and a_0, i.e. $\frac{X_t}{b_0} \sim Beta'(t, a_0)$. □

Proof of Lemma 5.1.2: In the gamma-Poisson model, $X_t \sim Poisson(\lambda t)$, and

$\lambda \sim Gamma\,(a_0, b_0)$ given $\mathcal{F}_0$. Therefore, the predictive distribution of X_t is

$$\begin{aligned} P(X_t = k) &= \int_0^\infty \frac{b_0(b_0\lambda)^{a_0-1}e^{-b_0\lambda}}{\Gamma(a_0)} e^{-\lambda t}\frac{(\lambda t)^k}{k!} d\lambda \\ &= \frac{t^k}{k!}\int_0^\infty \frac{b_0^{a_0}\lambda^{a_0+k-1}e^{-(b_0+t)\lambda}}{\Gamma(a_0)} d\lambda \\ &= \frac{t^k}{k!}\cdot\frac{\Gamma(a_0+k)}{\Gamma(a_0)}\cdot\frac{b_0^{a_0}}{(b_0+t)^{a_0+k}} \\ &= \frac{\Gamma(a_0+k)}{k!\Gamma(a_0)}\left(\frac{b_0}{b_0+t}\right)^{a_0}\left(\frac{t}{b_0+t}\right)^k, \end{aligned}$$

which is the generalized negative binomial pmf with parameters a_0 and $\frac{t}{b_0+t}$. □

Proof of Proposition 5.3.1: Starting from initial (a_0, b_0) at time 0, we make a decision at fixed time B based on (a_B, b_B): stop if $\frac{b_B}{a_B-1} < r$ and continue sampling to infinity if $\frac{b_B}{a_B-1} \geq r$.

By Lemma 5.1.1, if we define $Y_B := \frac{X_B}{b_0}$, then $Y_B \sim Beta'(B, a_0)$. Therefore,

$$\begin{aligned} & \mathbb{E}\left[e^{-cB}\left(r - \frac{b_B}{a_B-1}\right)^+\right] \\ = \;& \mathbb{E}\left[e^{-cB}\left(r - b_0\frac{1+Y_B}{a_0+B-1}\right)^+\right] \\ = \;& e^{-cB}\mathbb{E}\left[\left(r - b_0\frac{1+Y_B}{a_0+B-1}\right)\mathbb{I}_{\left\{Y_B \leq \frac{r(a_0+B-1)}{b_0}-1\right\}}\right] \\ = \;& e^{-cB}\left\{\left(r - \frac{b_0}{a_0+B-1}\right)F\left(\frac{r(a_0+B-1)}{b_0}-1\right)\right. \\ & \left. -\frac{b_0}{a_0+B-1}\mathbb{E}\left[Y_B\mathbb{I}_{\left\{Y_B \leq \frac{r(a_0+B-1)}{b_0}-1\right\}}\right]\right\} \\ = \;& e^{-cB}\left[\left(r - \frac{b_0}{a_0+B-1}\right)F\left(\frac{r(a_0+B-1)}{b_0}-1\right)\right. \\ & \left. -\frac{b_0}{a_0+B-1}\left(\frac{r(a_0+B-1)}{b_0}-1\right)F\left(\frac{r(a_0+B-1)}{b_0}-1\right)\right] \\ + \;& e^{-cB}\frac{b_0}{a_0+B-1}\int_0^{\frac{r(a_0+B-1)}{b_0}-1} F(s)ds \\ = \;& e^{-cB}\frac{b_0}{a_0+B-1}\int_0^{\frac{r(a_0+B-1)}{b_0}-1} F(s)ds \end{aligned}$$

where $F(s)$ is the cdf of $Beta'(B, a_0)$. By denoting $A = \frac{r(a_0+B-1)}{b_0} - 1$, we obtain a simple form below,

$$\mathbb{E}\left[e^{-cB}\left(r - \frac{b_B}{a_B - 1}\right)^+\right] = e^{-cB}\frac{r_0}{A+1}\int_0^A F(s)ds$$

whence the proposition is proved. □

Proof of Proposition 5.3.2: By Lemma 5.1.2, $X_B \sim \text{NB}\left(a_0, \frac{B}{b_0+B}\right)$. Under the "one-stage" stopping rule, the value $\bar{V}_B$ is computed as

$$\begin{aligned}\mathbb{E}\left[e^{-cB}(r - m_B)^+\right] &= \mathbb{E}\left[\frac{e^{-cB}}{b_0+B}(rb_0 + rB - m_0b_0 - X_B)^+\right] \\ &= \sum_{k \le rb_0+rB-m_0b_0} \frac{e^{-cB}}{b_0+B}(rb_0 + rB - m_0b_0 - k)\cdot f(k)\end{aligned}$$

where $f(k)$ is the pmf of $\text{NB}\left(a_0, \frac{B}{b_0+B}\right)$. If we denote $A := rb_0 + rB - m_0b_0$ and apply summation by parts, we get the simple form of $\bar{V}_B$

$$\begin{aligned}\mathbb{E}\left[e^{-cB}(r - m_B)^+\right] &= \frac{Ae^{-cB}}{b_0+B}F(\lfloor A\rfloor) - \frac{e^{-cB}}{b_0+B}\sum_{k\le A} k\cdot f(k) \\ &= \frac{Ae^{-cB}}{b_0+B}F(\lfloor A\rfloor) - \frac{e^{-cB}}{b_0+B}\left[\lceil A\rceil\cdot F(\lfloor A\rfloor) - \sum_{k\le A}F(k)\right] \\ &= \frac{e^{-cB}}{b_0+B}\left[\sum_{k\le A}F(k) - (\lceil A\rceil - A)F(\lfloor A\rfloor)\right]\end{aligned}$$

whence the proposition is proved. □

1

Bibliography

[1] W. B. Powell and I. O. Ryzhov. *Optimal Learning*. John Wiley and Sons, 2012.

[2] J. C. Gittins, K. D. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices (2nd ed.)*. John Wiley and Sons, 2011.

[3] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2009.

[4] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. Bayesian data analysis. *Chapman and Hall*, page 63, 2004.

[5] R.S. Sutton and A.G. Barto. *Reinforcement Learning*, volume 35. MIT Press, 1998.

[6] P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

[7] P. I. Frazier and W. B. Powell. Convergence to Global Optimality with Sequential Bayesian Sampling Policies. *SIAM Journal on Control and Optimization (to appear)*, 2011.

[8] P. I. Frazier, W. B. Powell, and S. Dayanik. The knowledge-gradient policy for correlated normal rewards. *INFORMS Journal on Computing*, 21(4):599–613, 2009.

[9] I. O. Ryzhov, P. I. Frazier, and W. B. Powell. On the robustness of a one-period look-ahead policy in multi-armed bandit problems. In *Proceedings of the 2010 International Conference on Computational Science*, pages 1629–1638, 2010.

[10] I. O. Ryzhov, P. I. Frazier, and W. B. Powell. A new optimal stepsize for approximate dynamic programming. *Submitted for publication*, 2012.

[11] J. C. Gittins and D. M. Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, 66(3):561–565, 1979.

[12] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[13] R. Agrawal. Sample mean based index policies with O (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

[14] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[15] M. Chhabra and S. Das. Learning the Demand Curve in Posted-Price Digital Goods Auctions. In *Proceedings of the 10th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 63–70, 2011.

[16] D.A. Berry and L.M. Pearson. Optimal designs for clinical trials with dichotomous responses. *Statistics in Medicine*, 4(4):497–508, 1985.

[17] A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12:2879–2904, 2011.

[18] S. E. Chick and P. I. Frazier. Sequential sampling with economics of selection procedures. *Management Science*, 58(3):550–569, 2012.

[19] B. Nelson and F. Matejcik. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Science*, 41(12):1935–1945, 1995.

[20] S. E. Chick and K. Inoue. New procedures to select the best simulated system using common random numbers. *Management Science*, 47(8):1133–1149, 2001.

[21] I. O. Ryzhov, W. B. Powell, and P. I. Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.

[22] H. Qu, I. O. Ryzhov, and M. C. Fu. Ranking and selection with unknown correlation structures. In C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, editors, *Proceedings of the 2012 Winter Simulation Conference*, page 12, 2012.

[23] F. Caro and J. Gallien. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292, 2007.

[24] K. D. Glazebrook, J. Meissner, and J. Schurr. How big should my store be? On the interplay between shelf-space, demand learning and assortment decisions. *Working paper, Lancaster University*, 2013.

[25] V.F. Farias and B. Van Roy. Dynamic pricing with a prior on market response. *Operations Research*, 58(1):16–29, 2010.

[26] X. Wang and Y. Wang. Optimal investment and consumption with stochastic dividends. *Applied Stochastic Models in Business and Industry*, 26(6):792–808, 2010.

[27] Q. Zhang, P. B. Seetharaman, and C. Narasimhan. The indirect impact of price deals on households' purchase decisions through the formation of expected future prices. *Journal of Retailing*, 88(1):88–101, 2012.

[28] D. Agarwal, B.-C. Chen, and P. Elango. Explore/exploit schemes for web content optimization. In *Proceedings of the 9th IEEE International Conference on Data Mining*, pages 1–10, 2009.

[29] M. A. Lariviere and E. L. Porteus. Stalking information: Bayesian inventory management with unobserved lost sales. *Management Science*, 45(3):346–363, 1999.

[30] W. Jouini and C. Moy. Channel selection with Rayleigh fading: a multi-armed bandit framework. In *Proceedings of the 13th IEEE International Workshop on Signal Processing Advances in Wireless Communications*, pages 299–303, 2012.

[31] M. H. DeGroot. *Optimal Statistical Decisions*. John Wiley and Sons, 1970.

[32] E. Cinlar. *Probability and Stochastics*. Springer, 2011.

[33] S. E. Chick. Subjective Probability and Bayesian Methodology. In S. G. Henderson and B. L. Nelson, editors, *Handbooks of Operations Research and Management Science, vol. 13: Simulation*, pages 225–258. North-Holland Publishing, Amsterdam, 2006.

[34] P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

[35] E. Vazquez and J. Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and Inference*, 140(11):3088–3095, 2010.

[36] P. I. Frazier and W. B. Powell. Consistency of Sequential Bayesian Sampling Policies. *SIAM Journal on Control and Optimization*, 49(2):712–731, 2011.

[37] I. O. Ryzhov and W. B. Powell. Information collection for linear programs with uncertain objective coefficients. *SIAM Journal on Optimization*, 22(4):1344–1368, 2012.

[38] I. O. Ryzhov and W. B. Powell. The value of information in multi-armed bandits with exponentially distributed rewards. In *Proceedings of the 2011 International Conference on Computational Science*, pages 1363–1372, 2011.

[39] A. E. Kyprianou. *Introductory lectures on fluctuations of Lévy processes with applications*. Springer, 2006.

[40] M. N. Katehakis and A. F. Veinott. The multi-armed bandit problem: Decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268, 1987.

[41] J. C. Gittins and Y. G. Wang. The learning component of dynamic allocation indices. *The Annals of Statistics*, 20(3):1625–1636, 1992.

[42] R. Filliger and M.-O. Hongler. Explicit Gittins indices for a class of superdiffusive processes. *Journal of Applied Probability*, 44(2):554–559, 2007.

[43] K. D. Glazebrook and R. Minty. A generalized Gittins index for a class of multiarmed bandits with general resource requirements. *Mathematics of Operations Research*, 34(1):26–44, 2009.

[44] M. Brezzi and T.L. Lai. Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, 27(1):87–108, 2002.

[45] Y. Yao. Some results on the Gittins index for a normal reward process. In H. Ho, C. Ing, and T. Lai, editors, *Time Series and Related Topics: In Memory of Ching-Zong Wei*, pages 284–294. Institute of Mathematical Statistics, Beachwood, OH, USA, 2006.

[46] S. E. Chick and N. Gans. Economic analysis of simulation selection problems. *Management Science*, 55(3):421–437, 2009.

[47] M. J. Steele. *Stochastic Calculus and Financial Applications*. Springer, New York, 2000.

[48] P. van Moerbeke. On optimal stopping and free boundary problems. *Archive for Rational Mechanics and Analysis*, 60(2):101–148, 1976.

[49] N. El Karoui and I. Karatzas. Dynamic allocation problems in continuous time. *The Annals of Applied Probability*, 4(2):255–286, 1994.

[50] H. Kaspi and A. Mandelbaum. Lévy bandits: Multi-armed bandits driven by Lévy processes. *The Annals of Applied Probability*, 5(2):541–565, 1995.

[51] A. Mandelbaum. Discrete multiarmed bandits and multiparameter processes. *Probability Theory Related Fields*, 71:129–147, 1986.

[52] A. Mandelbaum. Continuous multiarmed bandits and multiparameter processes. *Annals of Probability*, 15:1527–1556, 1987.

[53] G. Peskir and A. N. Shiryaev. *Optimal stopping and free boundary problems*. Birkhauser Verlag, 2006.

[54] H. Chernoff and A.J. Petkau. Numerical solutions for bayes sequential decision problems. *SIAM Journal of Scientific and Statistical Computing*, pages 46–59, 1986.

[55] H. Chernoff and A.J. Petkau. Optimal stopping for brownian motion in bandit problems and sequential analysis. *Working Paper*, 1999.

[56] E. Cinlar. Conditional Lévy processes. *Computers and Mathematics with Applications*, 46(7):993–997, 2003.

[57] I. Monroe. Processes that can be embedded in Brownian motion. *The Annals of Probability*, 6(1):42–56, 1978.

[58] K. Itô, O. E. Barndorff-Nielsen, and K.-I. Sato. *Stochastic processes: lectures given at Aarhus University*. Springer Verlag, 2004.

[59] E. B. Dynkin. *Markov processes*. Springer, 1965.

[60] K.-I. Sato. *Lévy processes and infinitely divisible distributions*. Cambridge University Press, 1999.

[61] D. Bertsimas and A. J. Mersereau. A learning approach for interactive marketing to a customer segment. *Operations Research*, 55(6):1120–1135, 2007.

[62] Y. Yu. Structural properties of Bayesian bandits with exponential family distributions. *arXiv preprint arXiv:1103.3089*, 2011.

[63] S. Aalto, U. Ayesta, and R. Righter. Properties of the Gittins index with application to optimal scheduling. *Probability in the Engineering and Informational Sciences*, 25(3):269–288, 2011.

[64] A. Müller and D. Stoyan. *Comparison methods for stochastic models and risks*. John Wiley and Sons, 2002.

[65] M. Shaked and J.G. Shanthikumar. *Stochastic Orders*. Springer, 2007.

[66] D. Lamberton and G. Pagès. Sur l'approximation des réduites. *Annales de l'institut Henri Poincaré*, B26(2):331–355, 1990.

[67] F. Coquet and S. Toldo. Convergence of values in optimal stopping and convergence of optimal stopping times. *Electronic Journal of Probability*, 12(8):207–228, 2007.

[68] A. Muller. How does the value function of a markov decision process depend on the transition probabilities? *Mathematics of Operations Research*, 22(4):872–885, 1997.