Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

OPTIMAL LEARNING WITH NON-GAUSSIAN REWARDS

Zi Ding

Department of Mathematics
University of Maryland
College Park, MD 20742, USA

Ilya O. Ryzhov

Robert H. Smith School of Business
University of Maryland
College Park, MD 20742, USA

ABSTRACT

We propose a theoretical and computational framework for approximating the optimal policy in multi-armed bandit problems where the reward distributions are non-Gaussian. We first construct a probabilistic interpolation of the sequence of discrete-time rewards in the form of a continuous-time conditional Lévy process. In the Gaussian setting, this approach allows an easy connection to Brownian motion and its convenient time-change properties. No such device is available for non-Gaussian rewards; however, we show how optimal stopping theory can be used to characterize the value of the optimal policy, using a free-boundary partial integro-differential equation, for exponential and Poisson rewards. We then solve this problem numerically to approximate the set of belief states possessing a given optimal index value, and provide illustrations showing that the solution behaves as expected.

1 INTRODUCTION

We consider a fundamental model for learning in stochastic optimization, in which there is a finite set of design alternatives with unknown values, and a decision-maker can perform sequential experiments on individual alternatives to learn about the unknown values and eventually identify the best alternative. The simulation literature (Bechhofer et al. 1995; Kim and Nelson 2006) studies a version of this problem known as ranking and selection (R&S), where new information is collected from stochastic simulations, the total simulation budget is limited, and a single alternative will be selected for implementation after the budget has been used. The objective of the problem is to guide the allocation of simulation experiments to maximize either the probability of implementing the best alternative (Kim and Nelson 2001) or the expected value of the implemented alternative (Chick 2006). Either way, a single experiment is valuable insofar as it helps us to improve the quality of the final implementation.

The multi-armed bandit problem (Gittins et al. 2011), widely studied in applied probability and computer science, is closely related to R&S. The only difference is that the outcome of an experiment (viewed as the simulation output in R&S) now has inherent economic value, and the objective is to maximize the cumulative reward obtained across all experiments. While it is still important to identify the best alternative, the decision-maker now has to ensure that every experiment will produce a reasonably good outcome. This model is suitable for applications where decisions are implemented in real time, such as advertising placement in e-commerce or clinical drug trials with human patients, and thus each decision has economic or other consequences, in addition to providing information. In simulation, the bandit model is relevant when a single simulation experiment costs money (Chick and Gans 2009). At a high level, R&S and multi-armed bandits have many common elements, and indeed some algorithmic approaches can be easily adapted to either problem class (Ryzhov et al. 2012; Powell and Ryzhov 2012).

The vast majority of the literature on either problem typically assumes that information takes the form of samples from a Gaussian distribution centered around the true value of an alternative. The Gaussian assumption offers advantages such as the ability to incorporate correlations between estimated values

(Nelson and Matejcik 1995; Qu et al. 2012). Recently, however, the operations management literature has considered problems in pricing (Farias and Van Roy 2010) and assortment planning (Caro and Gallien 2007; Glazebrook et al. 2013) where the observed demand follows a Poisson distribution with unknown rate. In the newsvendor problem of Lariviere and Porteus (1999), a Bayesian gamma prior models beliefs about an exponentially distributed demand. The work by Jouini and Moy (2012) also applies the gamma-exponential model to learn signal-to-noise ratios in channel selection. The gamma-Poisson and gamma-exponential models are among the most intuitive non-Gaussian conjugate priors, and the best-suited for applications in pricing and assortment planning. While the Gaussian assumption may be applied in a large-sample setting, observations in bandit problems are collected individually in an online manner.

The bandit literature provides some general results for non-Gaussian rewards. For instance, the Gittins index policy (Gittins et al. 2011) is optimal for independent gamma priors on Poisson or exponential parameters; see Gittins and Wang (1992) for a discussion of scaling properties in the exponential setting. The work by Agrawal (1995) extends the upper confidence bound (UCB) approach of Lai and Robbins (1985) to the gamma-exponential model, while Ryzhov and Powell (2011) does the same for the knowledge gradient (KG) approach of Gupta and Miescke (1996). See Chapter 5 of Powell and Ryzhov (2012) for extensions of KG to other non-Gaussian problems. Nonetheless, the non-Gaussian setting still produces unexpected theoretical challenges. For example, the KG method is known to be asymptotically optimal in the Gaussian setting (Frazier and Powell 2011), meaning that it identifies the best alternative with probability 1 if given infinitely many measurements. However, this property does *not* hold when rewards are exponential (Ding and Ryzhov 2013). Conversely, the UCB method of Agrawal (1995) preserves its desirable theoretical properties, but relies on tunable parameters that are difficult to compute optimally (Liu and Zhao 2010). Perhaps for these reasons, the non-Gaussian case remains relatively unexplored.

In this paper, we summarize a new framework for optimal learning with non-Gaussian rewards. We return to the Gittins index policy, which is theoretically optimal for multi-armed bandits, but is known to be difficult to compute. In the Gaussian setting, a recent stream of work by Brezzi and Lai (2002), Yao (2006), and Chick and Gans (2009) has approximated the Gittins policy by formulating an optimal stopping problem on a Brownian motion with unknown drift, a continuous-time process that can be viewed as a probabilistic interpolation of the sequence of Gaussian rewards collected from a single alternative. By exploiting the connection between Brownian motion and the heat equation (Steele 2000), a free-boundary problem can be formulated and solved numerically to approximate the Gittins index. We use a similar approach as our foundation, and interpolate the reward sequence in the gamma-Poisson and gamma-exponential problems with conditionally Poisson and gamma processes, respectively. We then formulate stopping problems on these continuous-time processes. Although we cannot rely on the time-change properties of Brownian motion to "standardize" the problem, as in previous work, we use an alternate approach based on equating the infinitesimal and characteristic operators (Peskir and Shiryaev 2006) of the function solving the stopping problem. This leads to free boundary problems on partial integro-differential equations (PIDEs).

We describe how these problems can be solved numerically to approximate the Gittins index. In the gamma-exponential problem, the Gittins index possesses scaling properties which can be exploited to reduce the difficulty of the procedure, and, ultimately, to drive the development of computable approximations to the Gittins index. We summarize our recent and ongoing work in this regard (primarily focusing on solving the PIDEs) and point to future directions. To our knowledge, this work represents the first effort to develop approximations for optimal policies in non-Gaussian learning problems, and shows how Lévy process interpolation can play a useful role in optimal learning beyond the Gaussian model.

2 OPTIMAL LEARNING WITH NON-GAUSSIAN REWARDS

In Section 2.1, we begin with a general exposition of the multi-armed bandit problem, and then describe the specific settings considered in the rest of the paper. Section 2.2 summarizes the "Gittins index" policy, known to be optimal for the multi-armed bandit setting. We discuss the difficulty of computing this policy, motivating the need for our analysis.

2.1 Learning with Non-Gaussian Rewards

Suppose that there are M alternatives, to be considered over a large number of time periods. Let $x^n \in \{1,...,M\}$ denote the alternative chosen for simulation in the nth stage, and let W_x^{n+1} represent the output of that simulation (which becomes known only at time $n+1$). For fixed x, the outputs $W_x^1, W_x^2, ...$ are drawn from a common sampling density $f_x(\cdot;\lambda_x)$, where λ_x is a parameter or vector of parameters for alternative x. Conditional on λ_x, the outputs are independent. Let $\mathscr{F}^n$ denote the sigma-algebra generated by the first n decisions $x^0, x^1, ..., x^{n-1}$ and outputs $W_{x^0}^1, ..., W_{x^{n-1}}^n$.

However, the parameters λ_x are not known for any x, and thus are modeled as random variables. The decision-maker maintains a set of beliefs about λ_x, which can be represented by a sequence $(k_x^n)_{n=0}^{N-1}$ of random vectors, such that k^n is $\mathscr{F}^n$-measurable for all n, and we can write

$$\mathbb{E}\left(W_x^{n+1} \,|\, \mathscr{F}^n\right) = m(k_x^n)$$

for an appropriate function m. One way to interpret the *knowledge state* k_x^n is as a set of sufficient statistics for the conditional distribution of λ_x given $\mathscr{F}^n$. In the classic multi-armed bandit problem, the parameters λ_x are assumed to be independent of one another, and likewise the simulation outputs are independent across alternatives. Our beliefs about the alternatives are characterized by $k^n = \{k_1^n, ..., k_M^n\}$.

A *policy* π represents a sequence $X^{\pi,0}, X^{\pi,1}, ...$ of functions mapping knowledge states $k^0, k^1, ...$ to alternatives in $\{1,...,M\}$. In other words, a policy specifies a way to make decisions for any set of beliefs at any time stage. The decision-maker's objective can be written as

$$\sup_{\pi} \mathbb{E}^{\pi} \sum_{n=0}^{\infty} \gamma^n \mathbb{E}\left(W_{X^{\pi,n}(k^n)}^{n+1}\right). \tag{1}$$

In words, (1) chooses a policy maximizing the infinite-horizon discounted average reward obtained from the alternatives simulated by the policy. The parameter $0 < \gamma < 1$ is a pre-specified discount factor.

In this paper, we consider two classic Bayesian learning models where the rewards are non-Gaussian. The first of these is the *gamma-exponential* model. In this setting, the sampling density f_x is (conditionally) exponential with unknown rate parameter λ_x. The conditional distribution of λ_x, given $\mathscr{F}^n$, is gamma with parameters a_x^n and b_x^n. Bayesian analysis (DeGroot 1970) provides us with simple recursive relationships

$$a_x^{n+1} = \begin{cases} a_x^n + 1 & \text{if } x^n = x \\ a_x^n & \text{if } x^n \neq x, \end{cases} \tag{2}$$

$$b_x^{n+1} = \begin{cases} b_x^n + W_x^{n+1} & \text{if } x^n = x \\ b_x^n & \text{if } x^n \neq x. \end{cases} \tag{3}$$

In this case, we have $k_x^n = (a_x^n, b_x^n)$, and the mean function m is given by $m(k_x^n) = \frac{b_x^n}{a_x^n - 1}$.

The second setting we consider is the *gamma-Poisson* model, where f_x is conditionally Poisson with unknown rate λ_x. The belief distribution of λ_x at time n is again gamma with parameters a_x^n and b_x^n, and the Bayesian updating equations are now given by

$$a_x^{n+1} = \begin{cases} a_x^n + W_x^{n+1} & \text{if } x^n = x \\ a_x^n & \text{if } x^n \neq x, \end{cases} \tag{4}$$

$$b_x^{n+1} = \begin{cases} b_x^n + 1 & \text{if } x^n = x \\ b_x^n & \text{if } x^n \neq x. \end{cases} \tag{5}$$

Again, the decision-maker's knowledge about λ_x at time n is represented by $k_x^n = (a_x^n, b_x^n)$ with mean function $m(k_x^n) = \frac{a_x^n}{b_x^n}$.

We briefly note that a third well-known non-Gaussian learning model assumes Bernoulli rewards with beta belief distributions. This setting has been relatively well-represented in the literature (see e.g., Berry and Pearson (1985)), and we do not explore it here. Among the other common conjugate learning models, the gamma-exponential and gamma-Poisson settings are the most intuitive (using some of the most fundamental probability distributions) and the best-suited to applications in assortment planning and revenue management. At the same time, these learning models have received the least amount of theoretical scrutiny in the bandit literature.

2.2 Review of Gittins Index Policies

We briefly review the properties of the Gittins index policy, the theoretically optimal solution of (1). For more details, the reader is referred to Ch. 6 of Powell and Ryzhov (2012) for a more detailed introduction, and to Gittins et al. (2011) for a deeper theoretical treatment.

In the Gittins index method, each alternative is considered separately from the others. Denote by $k = (a,b)$ our beliefs about an arbitrary alternative (we drop the subscript of the alternative from these parameters for convenience). Suppose that, at each time step, we have a choice between simulating this alternative and receiving a deterministic, pre-specified retirement reward R. The optimal decision in this setting is the solution to Bellman's equation, given by

$$V(k,R) = \max\left\{R + \gamma V(k,R), \mathbb{E}\left[W + \gamma V\left(k',R\right) | k\right]\right\} \tag{6}$$

where k' is the future knowledge state obtained e.g., via (2)-(3) or (4)-(5). Of course, if the fixed reward is optimal under a set of beliefs k, it will remain optimal for all future time periods, whence (6) becomes

$$V(k,R) = \max\left\{\frac{R}{1-\gamma}, m(k) + \gamma \mathbb{E}\left[V\left(k',R\right) | k\right]\right\}. \tag{7}$$

The *Gittins index* is a value $G(k)$ that makes the decision-maker indifferent between the two quantities in (7). It has been shown that the policy

$$X^{*,n}(k^n) = \arg\max_x G(k_x^n)$$

is optimal for the objective in (1).

In this way, an M-dimensional problem can be decomposed into M one-dimensional problems, each of which is independent from the others. Furthermore, for the special case of the gamma-exponential problem, it is known that $G(a_x^n, b_x^n) = b_x^n G(a_x^n, 1)$, so potentially Gittins indices only have to be computed for a restricted class of knowledge states. Even so, (7) remains computationally intractable for most continuous reward distributions. This computational challenge serves as the motivation for our work. Currently, efficient approximations exist only for Gaussian reward distributions; we now describe a new framework for developing such approximations in non-Gaussian settings.

3 THE GITTINS INDEX AS A STOPPING BOUNDARY

The existing literature on Gittins index approximation with Gaussian rewards begins with the work by Brezzi and Lai (2002), which proposed the following idea. For arbitrary x (again, we drop the subscript x for convenience), the discrete-time process $(W^n)_{n=1}^{\infty}$ of observations with unknown mean μ and known variance σ^2 is replaced by a continuous-time process $(X_t)_{t\geq 0}$ which can be viewed as a probabilistic interpolation of the discrete-time process. That is, for integer t, the increment $X_{t+1} - X_t$ has the same distribution as the one-period reward W^{t+1}. In the Gaussian setting, (X_t) is (conditionally) a Brownian motion with unknown drift μ and known volatility σ. The formulation of the Gittins index in (7) can also be extended to the continuous-time case and written as the solution to an optimal stopping problem. The work by Chick and Gans (2009) shows how this problem can be recast as a PDE based on the heat equation.

Our approach draws inspiration from this idea. In the two non-Gaussian settings described in Section 2.1, the probabilistic interpolation of the sequence (W^n) is a conditional Lévy process. For the gamma-exponential problem, (X_t) is a gamma process (see e.g., Cinlar (2011) for a definition) with shape parameter 1 and unknown scale parameter λ, whereas in the gamma-Poisson setting, (X_t) is a Poisson process with unknown rate λ. In both cases, we begin by assuming $\lambda \sim Gamma\,(a_0, b_0)$, reflecting the decision-maker's prior beliefs. Letting $\mathscr{F}^t$ be the sigma-algebra generated by the path of X up to time t, we find that the conditional distribution of λ given $\mathscr{F}^t$ is still gamma. For the gamma-exponential setting, the posterior parameters become

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

as in (2)-(3), whereas for the gamma-Poisson problem, we have

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

as in (4)-(5).

The Gittins recursion (7) is extended to the continuous-time setting as follows. Let c be a continuous-time discount factor (lower c corresponds to higher γ in discrete time). For fixed R, we write

$$R\int_0^\infty e^{-cs}ds = \sup_\tau \mathbb{E}\int_0^\tau e^{-cs}dX_s + R\int_\tau^\infty e^{-cs}ds, \tag{8}$$

where τ denotes a stopping time. This formulation is equivalent to the one based on Bellman's equation; see e.g., Katehakis and Veinott (1987) or Yao (2006) for more details. Essentially, discounted rewards are collected from the process (X_t) until time τ, at which point we "retire" and accrue the fixed reward R until the end of time. If (8) holds, we are indifferent between collecting R for the entire time horizon and running the process until time τ, precisely the condition needed for a Gittins index.

We now show how (8) can be simplified in the specific context of the gamma-exponential problem. Recalling that (X_t) is a gamma process with shape parameter 1 and unknown scale parameter λ, we write

$$\begin{aligned}\mathbb{E}\int_0^\tau e^{-cs}dX_s &= \mathbb{E}\int_0^\tau e^{-cs}\frac{b_s}{a_s-1}ds \\ &= \mathbb{E}\left[\frac{1}{c}\left(\frac{b_0}{a_0-1} - e^{-c\tau}\frac{b_\tau}{a_\tau-1}\right) + \frac{1}{c}\int_0^\tau e^{-cs}\frac{d}{ds}\left(\frac{b_s}{a_s-1}\right)\right],\end{aligned} \tag{9}$$

where the first equality can be obtained by conditioning on λ and passing the expectation inside the integral, and the second equality follows from integration by parts. The last integral in (9) has zero expectation because the mean process $m_t = \frac{b_t}{a_t-1}$ is an $\mathscr{F}^t$-martingale. Consequently, (8) can be rewritten as

$$\sup_\tau \mathbb{E}\left[e^{-c\tau}\left(R - \frac{b_\tau}{a_\tau-1}\right)\right] = R - \frac{b_0}{a_0-1}. \tag{10}$$

For the gamma-Poisson problem, a similar analysis leads to the formulation

$$\sup_\tau \mathbb{E}\left[e^{-c\tau}\left(R - \frac{a_\tau}{b_\tau}\right)\right] = R - \frac{a_0}{b_0}. \tag{11}$$

Next, we proceed to recast (10) and (11) as free boundary problems. In the Gaussian setting, this can be done by performing a time change (Revuz and Yor 2005) on the conditional Brownian motion process to convert it into a Wiener process. The connection of the Wiener process to the heat equation (Steele 2000) is then exploited to construct a free-boundary PDE. In the non-Gaussian setting, such manipulations are not possible. We use an alternate approach inspired by Peskir and Shiryaev (2006), based on equating the characteristic and infinitesimal operators of the value function of the stopping problem.

We use the gamma-exponential problem to walk through the approach. For $m_t = \frac{b_t}{a_t - 1}$, let

$$V(a,m) = \mathbb{E}\left\{ \sup_{\tau} \mathbb{E}\left[e^{-c\tau} (R - m_\tau) \right] \mid a_0 = a, m_0 = m \right\}. \tag{12}$$

The function V in (12) is the continuous-time analog of (6). We drop the dependence on R from the notation for convenience, and replace the generic knowledge state k by (a,m), a one-to-one transformation of the gamma-exponential belief parameters (a,b).

The *characteristic operator* of this value function is defined as

$$L_m^{char} V(a,m) = \lim_{U \downarrow \{m\}} \frac{\mathbb{E} V(a_{\tau_{U^c}}, m_{\tau_{U^c}}) - V(a,m)}{\mathbb{E}(\tau_{U^c})}, \tag{13}$$

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

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

In words, we first consider the value function at the moment when the mean process (m_t) leaves U, and then shrink U down to the singleton $\{m\}$. For the gamma-exponential problem, (13) admits a closed-form solution, stated in the following result.

Proposition 1 The characteristic operator of V is given by

$$L_m^{char} V(a,m) = cV(a,m) - (m - R). \tag{14}$$

The *infinitesimal operator* L_m^{inf} is derived using Itô's formula (Sato 1999). We write

$$\begin{aligned} V(a_t, m_t) &= V(a_0, m_0) + \int_0^t \frac{\partial V(a_s, m_s)}{\partial s} ds + \int_0^t \frac{\partial V(a_s, m_s)}{\partial m} dm_s \\ &\quad + \sum_{0 < s \leq t} \left[V(a_s, m_s) - V\left(a_s, m_s^-\right) \right] \\ &= V(a_0, m_0) + \int_0^t L_m^{inf} V(a_s, m_s)\, ds + M_t, \end{aligned}$$

where M_t is a martingale formed by adding and subtracting a continuous compensator to the jump component of V. Essentially, the characteristic and infinitesimal operators are two ways to write the derivative of V, one based on Kolmogorov's theory and the other on Itô calculus. Under general arguments, the two operators are equivalent; by matching them, we arrive at a free-boundary partial integro-differential equation (PIDE), stated in the following result.

Theorem 1 Suppose that $V(a,m)$ solves the free-boundary problem

$$\begin{aligned} \frac{\partial V(a,m)}{\partial a} - \frac{m}{a-1} \frac{\partial V(a,m)}{\partial m} + \int_0^\infty \left[V(a, m+z) - V(a,m) \right] \frac{1}{z} \left(\frac{m}{m+z} \right)^a dz &= cV(a,m) \\ V(a, m^*(a)) &= R - m^*(a) \end{aligned} \tag{15}$$

where $m^*(a)$ is an unknown stopping boundary curve. Then, $V(a,m)$ is the gamma-exponential value function in (12).

For the gamma-Poisson problem, the analysis is quite similar. The value function

$$V(b,m) = \mathbb{E}\left\{ \sup_{\tau} \mathbb{E}\left[e^{-c\tau} (R - m_\tau) \right] \mid b_0 = b, m_0 = m \right\} \tag{16}$$

has the same form as in (12), though the mean process has a slightly different definition $m_t = \frac{a_t}{b_t}$. As a result, the characteristic operator is the same as in (14). Using Itô calculus to derive the infinitesimal operator, we arrive at the free-boundary problem stated below.

Theorem 2 Suppose that $V(b,m)$ solves the free-boundary problem

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

where $m^*(b)$ is an unknown stopping boundary curve. Then, $V(b,m)$ is the gamma-Poisson value function in (16).

We briefly discuss the interpretation of these free-boundary problems. Recall that both problems are derived assuming a fixed R. Thus, they do not immediately yield a Gittins index for an arbitrary knowledge state. However, the stopping boundary curve m^* describes the set of all knowledge states for which the Gittins index is exactly equal to R. For a given knowledge state, we can search over the set of boundary curves for different values of R until we find the curve where the given state belongs. We discuss some ideas for how such search procedures can be constructed, but a full implementation is outside the scope of the present paper, which focuses on the numerical implementation and solution of the free-boundary problems. These issues are discussed extensively in the following section.

Another relevant question is whether it is possible to guarantee that the PIDEs in Theorems 1 and 2 have solutions. In the Gaussian case, the approach of Brezzi and Lai (2002) is able to time-change the interpolation process (a conditional Brownian motion) into a standard Wiener process, making it possible to apply standard existence results on the Brownian PDE. For the conditional Lévy processes studied in this paper, this is much more difficult. Some limited results are available in Sections 9.1-9.2 of Peskir and Shiryaev (2006). For example, problems with jumps and no diffusion part will satisfy continuous boundary conditions, whereas continuous processes (but not necessarily those with jumps) satisfy first-order boundary conditions. Beyond these cases, the structural analysis of the PIDEs becomes much more difficult. However, our numerical results indicate that solutions do exist, and behave in the way that we would expect of the Gittins value functions.

4 NUMERICAL IMPLEMENTATION AND EXAMPLES

Solving the problems in Theorems 1 and 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 Chick and Gans 2009). 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

$$\bar{V}_B = \mathbb{E}\left[e^{-cB}(R - m_B)^+\right], \tag{17}$$

and use $\sup_B \bar{V}_B$ to approximate the value of V for the prior parameters. For both gamma-Poisson and gamma-exponential models, (17) can be computed in closed form, and $\sup_B \bar{V}_B$ is relatively easy to calculate numerically.

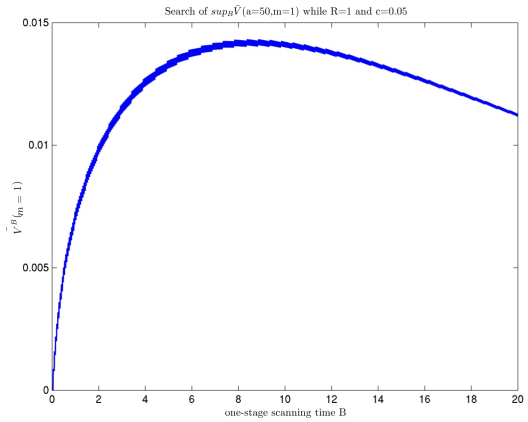
Proposition 2 In the gamma-exponential model,

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

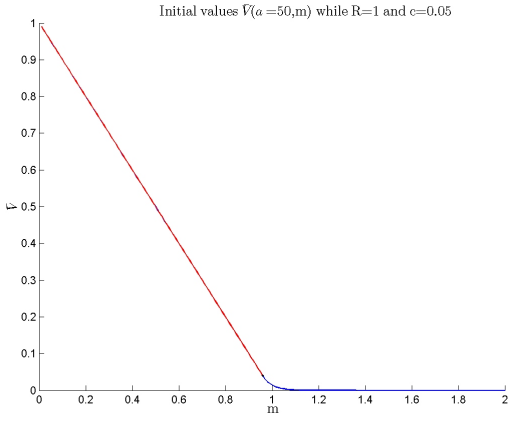
where $A = \frac{R(a_0+B-1)}{b_0} - 1$ and F is the cdf of a Beta prime distribution with parameters B and a_0.

Proposition 3 In the gamma-Poisson model,

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



(a) Initial value approximation as a function of B.



(b) Initial value approximation (optimized over B) as a function of m.

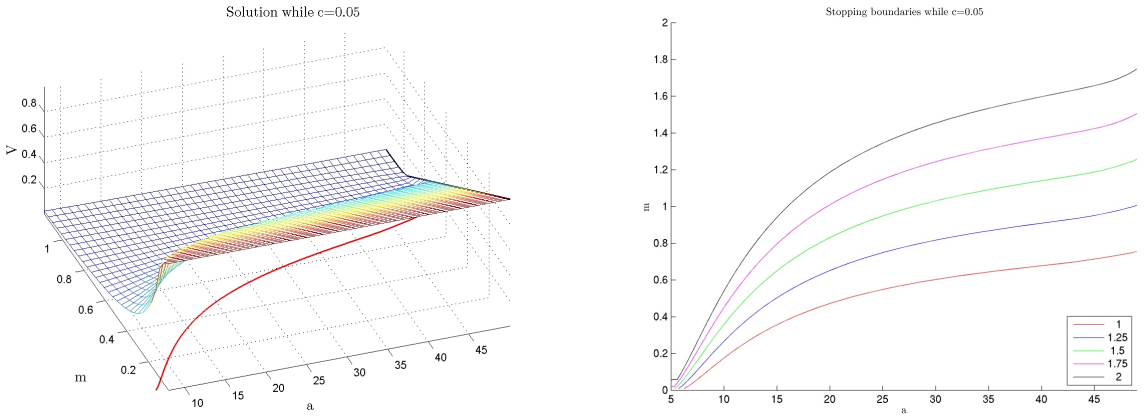
Figure 1: Illustrations of initial value approximation for PIDE solution.

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 search for lower bound more intuitively, through a gamma-exponential example with $R = 1$ and $c = 0.05$. First, Figure 1(a) shows that the approximation $\bar{V}_B$ 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 $V(a,m)$ with $a = 50$ and $m = 1$. Figure 1(b) shows the results of this procedure for all m values, with $a = 50$ fixed. The red 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 blue tail curve approaching zero shows where the approximation starts to deviate from $R - m$. In the stopping problem, the red section would correspond to the stopping region, while the blue section corresponds to the continuation region.

Using the lower bound to approximate the initial value of V, we solve the PIDEs numerically using Euler's finite difference schemes. It is preferable to calculate the initial value approximation for large time values, since the quality of the lower bound $\sup_B \bar{V}_B$ 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, highlighted in black) 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 $V(a,m) = R - m$. The red 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 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. First, it can be shown that V is decreasing in time, represented by the a parameter in the gamma-exponential model and the b parameter in the gamma-Poisson model. Second, V is decreasing in the mean parameter m. Recent work by Aalto et al. (2011) discusses the continuity of the Gittins index as a function of the continuous belief parameters. As a consequence, the stopping boundary m^* described by Theorems 1 and 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



(a) Solution surface of the PIDE, $R = 1$.

(b) Stopping boundaries for different values of R.

Figure 2: Illustrations of PIDE solutions.

curves appear to be concave; the slight bumps close to $a = 50$ are due to numerical issues stemming from proximity to the initial value. The shape of these curves suggests an intuitive search procedure to find the Gittins index for a given (a,m) pair: we first try a large value of R for which the stopping boundary is above (a,m), then apply the bisection method until we find R for which (a,m) is sufficiently close to the boundary. The implementation of this procedure is the subject of our ongoing work, and is outside the scope of this paper. However, it is clear that the key to such procedures is the ability to find good boundary curves.

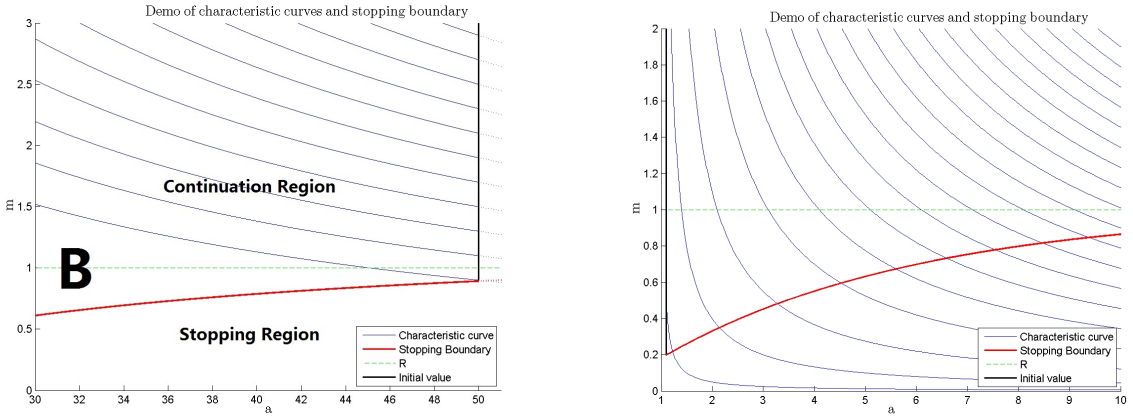
Although the backward propagation method has the benefit that the initial value approximation is much more accurate, it presents other computational challenges. Figure 3(a) demonstrates the propagation of characteristic curves in the PIDE. Intuitively, the solution at every point depends on the area above the characteristic curve that the point lies on. However, going backward in time, the stopping boundary moves in the opposite direction from the characteristic curves. Therefore, if the initial condition is given at $a = 50$ (black line), there is no way that it could propagate to the area below the lowest of the blue characteristic curves (the region marked B in the figure). Computing solutions in B requires additional techniques for building entropy solutions from PDE theory.

On the other hand, if we move forward in time, as in Figure 3(b), this issue is avoided. If we have initial conditions at small a values, it is easy to propagate the characteristic curves downward until the stopping condition is triggered. The drawback is that the initial condition is less accurate for small a. Our experience has been that backward propagation produces better solutions, despite the need for additional approximations in the B region on Figure 3(a).

The integral term in (15) must be calculated numerically (using e.g., a discrete Riemann sum), and thus presents an additional source of error. We found that applying back-propagation directly to this discretized integral produces bumpy boundary curves. This occurs because the integral term makes the information at a fixed point (a,m) non-local, so that any error at that point is passed to infinitely many points simultaneously, including points far away from (a,m). To deal with this issue, we enforce the monotonicity of the stopping boundary by simply taking the maximum of the bumpy simulated points. The results in Figure 2 demonstrate that the numerical solution behaves in accordance with our intuition about the problem.

5 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



(a) Stopping boundaries for different values of R.

(b) Solution surface of the PIDE, $R = 1$.

Figure 3: Illustrations of backward and forward PIDE solution.

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 discussed how these problems can be solved numerically, and presented illustrations showing that the results exhibit the correct structure established in the theory.

Our ongoing work concentrates on leveraging these results to obtain computationally tractable procedures for approximating Gittins indices. 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 present paper, the framework we have presented can be intuitively extended and incorporated into a search procedure to find the Gittins index for a restricted class of knowledge states. We can then fit a statistical regression model to the output; combining this with the scaling properties of the Gittins index will yield the first known computationally tractable Gittins index approximations for non-Gaussian rewards.

ACKNOWLEDGMENTS

The authors are grateful to Kazutoshi Yamazaki for several helpful discussions in the early stages of this work.

REFERENCES

Aalto, S., U. Ayesta, and R. Righter. 2011. "Properties of the Gittins Index with Application to Optimal Scheduling". *Probability in the Engineering and Informational Sciences* 25 (3): 269–288.

Agrawal, R. 1995. "Sample Mean Based Index Policies with O (log n) Regret For the Multi-Armed Bandit Problem". *Advances in Applied Probability* 27 (4): 1054–1078.

Bechhofer, R. E., T. J. Santner, and D. M. Goldsman. 1995. *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*. New York: J.Wiley & Sons.

Berry, D. A., and L. M. Pearson. 1985. "Optimal Designs for Clinical Trials with Dichotomous Responses". *Statistics in Medicine* 4 (4): 497–508.

Brezzi, M., and T. L. Lai. 2002. "Optimal Learning and Experimentation in Bandit Problems". *Journal of Economic Dynamics and Control* 27 (1): 87–108.

Caro, F., and J. Gallien. 2007. "Dynamic Assortment with Demand Learning for Seasonal Consumer Goods". *Management Science* 53 (2): 276–292.

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

Chick, S. E., and N. Gans. 2009. "Economic Analysis of Simulation Selection Problems". *Management Science* 55 (3): 421–437.

Cinlar, E. 2011. *Probability and Stochastics*. Springer.

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

Ding, Z., and I. O. Ryzhov. 2013. "Optimal Learning with Non-Gaussian Rewards". *Working paper, University of Maryland*.

Farias, V. F., and B. Van Roy. 2010. "Dynamic Pricing with a Prior on Market Response". *Operations Research* 58 (1): 16–29.

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

Gittins, J. C., K. D. Glazebrook, and R. Weber. 2011. *Multi-Armed Bandit Allocation Indices (2nd ed.)*. John Wiley and Sons.

Gittins, J. C., and Y. G. Wang. 1992. "The Learning Component of Dynamic Allocation Indices". *The Annals of Statistics* 20 (3): 1625–1636.

Glazebrook, K. D., J. Meissner, and J. Schurr. 2013. "How Big Should My Store Be? On the Interplay Between Shelf-Space, Demand Learning and Assortment Decisions". *Working paper, Lancaster University*.

Gupta, S., and K. Miescke. 1996. "Bayesian Look Ahead One-Stage Sampling Allocations for Selection of the Best Population". *Journal of Statistical Planning and Inference* 54 (2): 229–244.

Jouini, W., and C. Moy. 2012. "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*, 299–303.

Katehakis, M. N., and A. F. Veinott. 1987. "The Multi-Armed Bandit Problem: Decomposition and Computation". *Mathematics of Operations Research* 12 (2): 262–268.

Kim, S.-H., and B. L. Nelson. 2001. "A Fully Sequential Procedure for Indifference-Zone Selection in Simulation". *ACM Transactions on Modeling and Computer Simulation* 11 (3): 251–273.

Kim, S.-H., and B. L. Nelson. 2006. "Selecting the Best System". In *Handbooks of Operations Research and Management Science, vol. 13: Simulation*, edited by S. G. Henderson and B. L. Nelson, 501–534. North-Holland Publishing, Amsterdam.

Lai, T. L., and H. Robbins. 1985. "Asymptotically Efficient Adaptive Allocation Rules". *Advances in Applied Mathematics* 6:4–22.

Lariviere, M. A., and E. L. Porteus. 1999. "Stalking Information: Bayesian Inventory Management with Unobserved Lost Sales". *Management Science* 45 (3): 346–363.

Liu, K., and Q. Zhao. 2010. "Distributed Learning in Multi-Armed Bandit with Multiple Players". *IEEE Transactions on Signal Processing* 58 (11): 5667–5681.

Nelson, B., and F. Matejcik. 1995. "Using Common Random Numbers for Indifference-Zone Selection and Multiple Comparisons in Simulation". *Management Science* 41 (12): 1935–1945.

Peskir, G., and A. N. Shiryaev. 2006. *Optimal Stopping and Free Boundary Problems*. Birkhauser Verlag.

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

Qu, H., I. O. Ryzhov, and M. C. Fu. 2012. "Ranking and Selection with Unknown Correlation Structures". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. Uhrmacher. Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Revuz, D., and M. Yor. 2005. *Continuous Martingales and Brownian Motion (3rd ed.)*. Springer.

Ryzhov, I. O., and W. B. Powell. 2011. "The Value of Information in Multi-Armed Bandits with Exponentially Distributed Rewards". In *Proceedings of the 2011 International Conference on Computational Science*, 1363–1372.

Ryzhov, I. O., W. B. Powell, and P. I. Frazier. 2012. "The Knowledge Gradient Algorithm for a General Class of Online Learning Problems". *Operations Research* 60 (1): 180–195.

Sato, K.-I. 1999. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

Steele, M. J. 2000. *Stochastic Calculus and Financial Applications*. New York: Springer.

Yao, Y. 2006. "Some Results on the Gittins Index for a Normal Reward Process". In *Time Series and Related Topics: In Memory of Ching-Zong Wei*, edited by H. Ho, C. Ing, and T. Lai, 284–294. Institute of Mathematical Statistics, Beachwood, OH, USA.

AUTHOR BIOGRAPHIES

ZI DING is a Ph.D. candidate in Applied Mathematics, Statistics, and Scientific Computation at the University of Maryland. His interests lie in the areas of optimal learning, finance, and optimal stopping. His email address is zding@math.umd.edu.

ILYA O. RYZHOV is an Assistant Professor in the Robert H. Smith School of Business at the University of Maryland. He received a Ph.D. in Operations Research and Financial Engineering from Princeton University. His research deals with optimal learning and the broader area of stochastic optimization, with applications in disaster relief, energy, and revenue management. He was a recipient of WSC's Best Theoretical Paper Award in 2012. His work has appeared in *Operations Research*, and he is a co-author of the book *Optimal Learning*, published in 2012 by John Wiley & Sons. His email address is iryzhov@rhsmith.umd.edu.