Adv. Appl. Prob. **48,** 112–136 (2016)
*doi:*10.1017/apr.2015.9
© *Applied Probability Trust* 2016

OPTIMAL LEARNING WITH NON-GAUSSIAN REWARDS

ZI DING [*] AND
ILYA O. RYZHOV,[*] [**] *University of Maryland*

Abstract

We propose a novel theoretical characterization of the optimal 'Gittins index' policy in multi-armed bandit problems with non-Gaussian, infinitely divisible reward distributions. We first construct a continuous-time, conditional Lévy process which probabilistically interpolates the sequence of discrete-time rewards. When the rewards are Gaussian, this approach enables an easy connection to the convenient time-change properties of a Brownian motion. Although no such device is available in general for the non-Gaussian case, we use optimal stopping theory to characterize the value of the optimal policy as the solution to a free-boundary partial integro-differential equation (PIDE). We provide the free-boundary PIDE in explicit form under the specific settings of exponential and Poisson rewards. We also prove continuity and monotonicity properties of the Gittins index in these two problems, and discuss how the PIDE can be solved numerically to find the optimal index value of a given belief state.

Keywords: Gittins indices; optimal learning; multi-armed bandit; non-Gaussian rewards; probabilistic interpolation

2010 Mathematics Subject Classification: Primary 60G40
Secondary 60J75

1. Introduction

The field of *optimal learning* (see Powell and Ryzhov (2012)) concerns the study of the efficient collection of information in stochastic optimization problems subject to environmental uncertainty—that is, problems where uncertainty is driven by unknown probability distributions. In many applications in business, medicine, and various branches of engineering, the decision-maker is able to formulate a belief about these unknown distributions, and gradually improve it using information collected from expensive simulations or field experiments. In particular, we consider a fundamental class of optimal learning problems known as *multi-armed bandit* problems (Gittins *et al.* (2011)), in which there is a finite set of competing 'arms' or 'alternatives' (e.g. system designs, pricing strategies, or hiring policies), each with an unknown value. Alternatives are implemented sequentially in an online manner: upon choosing an alternative, we collect a reward in the form of a noisy sample centered around the unknown value. Our objective is to maximize the cumulative discounted expected reward collected over an infinite horizon.

Each individual reward thus plays two roles: it contributes immediate economic value, and it also provides information about the alternative with the potential to improve future decisions. The tradeoff between reward and information is known as the *exploration versus exploitation*

Received 22 December 2014; revision received 9 February 2015.

[*] Postal address: Robert H. Smith School of Business, University of Maryland, 4322 Van Munching Hall, College Park, MD 20742, USA.
[**] Email address: iryzhov@rhsmith.umd.edu

dilemma. This problem arises in applications where decisions are made in real time, such as dynamic pricing or advertising placement in e-commerce (Chhabra and Das (2011)) or clinical drug trials with human patients (Berry and Pearson (1985)). The bandit model is also relevant in simulation, in situations where a single simulation experiment costs money (Chick and Gans (2009)).

In the classical multi-armed bandit setting where the decision-maker's beliefs about the alternatives are mutually independent, the work by Gittins and Jones (1979) 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; see Katehakis and Veinott (1987). Nonetheless, despite this considerable structure (see, e.g. Gittins and Wang (1992) for additional scaling properties), which continues to inspire new theoretical research on Gittins-like policies (Filliger and Hongler (2007), Glazebrook and Minty (2009)), 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 often require the reward distributions to be Gaussian, or to have bounded support; see, e.g. Auer *et al*. (2002) for examples of both.

In the simulation literature, Gaussian assumptions are standard due to advantages such as the ability to concisely model correlations between estimated values (Chick and Inoue (2001), Ryzhov *et al*. (2012)). More recently, however, numerous applications have emerged where observations are clearly non-Gaussian. Recently, the operations management literature has considered (Caro and Gallien (2007) and Glazebrook *et al.* (2013)) applications where the observed demand comes from a Poisson distribution with unknown rate. The challenge of deriving Poisson distributions also arises in dynamic pricing (Farias and Van Roy (2010)), optimal investment and consumption (Wang and Wang (2010)), models for household purchasing decisions (Zhang *et al*. (2012)), and online advertising and publishing (Agarwal *et al*. (2009)). Lariviere and Porteus (1999) studied a newsvendor problem where a Bayesian gamma prior is used to model beliefs about an exponentially distributed demand. The gamma-exponential model was also used by Jouini and Moy (2012) for deriving signal-to-noise ratios in channel selection.

Motivated by applications such as the above, we consider Bayesian bandit problems under non-Gaussian, infinitely divisible reward distributions, encompassing both exponential and Poisson models. In the Gaussian setting, a recent body of work by Brezzi and Lai (2002), Yao (2006), and Chick and Gans (2009) approximates the Gittins index for an arm using an optimal stopping problem on a Brownian motion with unknown drift, a continuous-time process that probabilistically interpolates the sequence of Gaussian rewards collected from the arm. By making the connection between Brownian motion and the heat equation (Steele (2001)), one can formulate and numerically solve a free-boundary problem (Van Moerbeke (1976)) to approximate the Gittins index. Our approach uses a similar foundation: we interpolate the reward sequences in the non-Gaussian problems with conditional Lévy processes, which are generated by infinitely divisible distributions (Sato (1999)), and then derive the relevant continuous-time stopping problems. Although there has been a body of research available on multi-armed bandit problems driven by Lévy processes (El Karoui and Karatzas (1994), Kaspi and Mandelbaum (1995), Mandelbaum (1986), (1987)), this work does not consider the Bayesian perspective where the reward distributions depend on unknown (random) parameters and our beliefs about them evolve over time. This dependence leads to the use of *conditional* Lévy processes, which have previously been studied in a learning context only in the much

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

more restrictive setting of binary prior distributions; see, for example, Buonaguidi and Muliere (2013), and Cohen and Solan (2013).

The major challenge in studying non-Gaussian reward problems under this interpolation technique is that we cannot exploit the time-change properties of Brownian motion to 'standardize' the problem, as was done before in the Gaussian setting. Therefore, we develop an alternate method based on equating the infinitesimal and characteristic operators (Peskir and Shiryaev (2006)) of value functions in an optimal stopping problem. We then obtain free-boundary problems on partial integro-differential equations (PIDEs), which can potentially be used to approximate the Gittins index. The solutions to these equations are shown to possess regularity properties such as continuity and monotonicity, thus retaining the structure of the original discrete-time problems.

In this paper we make the following contributions:

1. we propose novel continuous-time approximations to Gittins indices for non-Gaussian problems, in the form of stopping problems on conditional Lévy processes that probabilistically interpolate sequences of infinitely divisible rewards;

2. we derive free-boundary PIDEs whose solutions match the value functions in these stopping problems, and illustrate how these solutions can be calculated;

3. we apply our method directly to gamma-exponential and gamma-Poisson problems and derive explicitly the PIDEs corresponding to these models;

4. we prove relevant structural properties on the monotonicity, continuity, and asymptotic behavior of the value functions and Gittins indices. We also derive a scaling property for the gamma-Poisson problem that is, to the best of the authors' knowledge, entirely new.

We view these contributions as furthering the theoretical understanding of Gittins indices for non-Gaussian problems.

2. Optimal learning with non-Gaussian rewards

In Section 2.1 we set up the notation for our analysis and describe two major classes of conjugate learning models with infinitely divisible rewards, namely gamma-exponential and gamma-Poisson. In Section 2.2 we review the Gittins index policy, known to be optimal for bandit problems. Section 2.3 provides additional motivation for our study by showing that non-Gaussian problems can cause inconsistent behavior in knowledge gradient methods, a prominent class of heuristic learning policies.

2.1. Learning models for non-Gaussian rewards

Consider a bandit problem with M alternatives, with $x_n \in \{1, \dots, M\}$ denoting the alternative chosen for implementation in stage $n = 0, 1, 2, \dots$. Let $W^{x_n}_{n+1}$ be the single-period reward observed after x_n is implemented. In the discrete-time problem, quantities are indexed by the time at which they become known; thus, x_n is chosen at time n, but $W^{x_n}_{n+1}$ becomes known one time period later.

For fixed x, the rewards $W^x_1, W^x_2, \dots$ 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. Let $\mathcal{F}_n$ be the σ-algebra generated by the first n decisions $x_0, x_1, \dots, x_{n-1}$ as well as the resulting rewards $W^{x_0}_1, \dots, W^{x_{n-1}}_n$. The unknown parameter λ^x is modeled as a random variable, and our beliefs about the possible values of the parameter

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

at time n are represented by the conditional distribution of λ^x given $\mathcal{F}_n$. In the problems we consider, this sequence of conditional distributions is characterized by a sequence $(k_n^x)_{n=0}^{\infty}$ of random vectors, where k_n^x is $\mathcal{F}_n$-measurable for all n, and we write $\mathbb{E}[W_{n+1}^x \mid \mathcal{F}_n] = m(k_n^x)$ 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 $m_\infty^x = \mathbb{E}[W_1^x \mid \lambda^x]$ 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. Note that, if x is observed infinitely often, $m(k_n^x) \to m_\infty^x$ almost surely (a.s.) by martingale convergence.

The *knowledge state* k_n^x can also be viewed as a set of sufficient statistics for the conditional distribution of λ^x given $\mathcal{F}_n$. We follow the classic multi-armed bandit model, where 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 $k_n = \{k_n^1, \dots, k_n^M\}$. Then, a *policy* π denotes a sequence $X_0^\pi, X_1^\pi, \dots$ of functions mapping knowledge states $k_0, k_1, \dots$ to elements of $\{1, \dots, M\}$. In other words, a policy is a rule for making decisions, under any possible knowledge state, at any time stage. Our objective can thus be written as

$$\sup_{\pi} \mathbb{E}^{\pi} \sum_{n=0}^{\infty} \gamma^n m(k_n^{\pi, X_n^\pi(k_n)}), \tag{2.1}$$

where $0 < \gamma < 1$ is a prespecified discount factor. In words, we maximize the expected cumulative, infinite-horizon, discounted reward obtained from alternatives implemented by our chosen policy.

We specifically highlight two classic Bayesian learning models where the sampling distributions are infinitely divisible. In the *gamma-exponential* model, f^x is (conditionally) exponential with unknown rate λ^x. Under the assumption that $\lambda^x \sim \text{gamma}(a_0^x, b_0^x)$, the conditional distribution of λ^x, given $\mathcal{F}_n$, is also gamma with parameters a_n^x and b_n^x. From DeGroot (1970), we can obtain simple recursive relationships for the parameters, given by

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

In the gamma-exponential model, $k_n^x = (a_n^x, b_n^x)$, and the mean function m is given by $m(k_n^x) = \mathbb{E}[1/\lambda^x \mid \mathcal{F}_n] = b_n^x/(a_n^x - 1)$. We also consider the *gamma-Poisson* model, where f^x is conditionally Poisson with unknown rate λ^x. Again, we start with $\lambda^x \sim \text{gamma}(a_0^x, b_0^x)$, 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

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

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 $m(k_n^x) = \mathbb{E}[\lambda^x \mid \mathcal{F}_n] = a_n^x/b_n^x$.

2.2. Review of Gittins indices

We briefly summarize the characterization of the Gittins index policy, known to optimally solve (2.1). For a more detailed introduction, we refer the reader to Powell and Ryzhov (2012, Chapter 6). Furthermore, Gittins *et al.* (2011) provides a deeper theoretical treatment with several equivalent proofs of optimality for the policy.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

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 versus retire) can be characterized using Bellman's equation for dynamic programming. We write

$$V(k,r) = \max\left\{\frac{r}{1-\gamma}, m(k) + \gamma\mathbb{E}[V(k',r) \mid k]\right\}, \tag{2.4}$$

where k' is computed using, e.g. (2.2) or (2.3).

The *Gittins index* $R(k)$ is the value of r that makes us indifferent between the two quantities inside the maximum in (2.4). When λ^x is known, this value is equal to the mean single-period reward, as shown in the following lemma. The proof is straightforward, and we omit it.

Lemma 2.1. *If the parameter λ^x is a known constant, the Gittins index of arm x is m_∞^x.*

Once the 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 (2.1). 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) is used to update k), it has also been shown by Gittins and Wang (1992) that

$$R(a,b) = bR(a,1), \tag{2.5}$$

meaning that the Gittins indices only have to be computed for a restricted class of knowledge states. Equivalently, if we can find $\tilde{b}(a)$ such that $R(a, \tilde{b}(a)) = 1$, we can use (2.5) to write $R(a, b) = b/\tilde{b}(a)$. Yet, even with this structure, it is difficult to compute $R(a, 1)$ or $\tilde{b}(a)$ for arbitrary a.

2.3. The inconsistency of knowledge gradient methods

In this section we provide additional motivation for our work by showing that non-Gaussian problems create theoretical challenges for a prominent class of suboptimal heuristics known as *knowledge gradient* (KG) methods. Such methods first calculate the expected improvement criterion

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

and then implement the alternative

$$X_n^{\mathrm{KG}}(k_n) = \arg\max_x R_n^{\mathrm{KG},x} \tag{2.7}$$

at time n. This approach has received attention in the simulation community (see, e.g. Chick (2006)), because it is computationally efficient and often performs near-optimally in experiments. Simulation optimization usually seeks to identify the alternative with the highest value, rather than to maximize the cumulative reward as in (2.1). However, these objectives are closely related, and the method can be adapted to bandit problems (Ryzhov *et al.* (2012)) with a simple modification of (2.7) known as 'online KG.'

If the rewards are Gaussian, Frazier *et al.* (2008) showed that the policy in (2.7) is statistically consistent, meaning that $m(k_n^x) \to m_\infty^x$ a.s. for every x. This is a useful regularity property for simulation optimization algorithms, and often holds when rewards are Gaussian; see Vazquez

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

and Bect (2010) and Frazier and Powell (2011). However, as we now show, this is not always true in non-Gaussian settings. Specifically, in the gamma-exponential problem, (2.6) has a closed-form solution; see Ryzhov and Powell (2011, Theorem 3.1). From that solution, it can immediately be seen that it is possible to have $R_n^{\mathrm{KG},x} = 0$ even though $\operatorname{var}(\lambda^x \mid \mathcal{F}_n) > 0$. We prove here that, if the policy places zero value on x, there may be a nonzero probability that the policy will never measure x, and $m(k_n^x)$ will not converge to the true mean reward. The proof can be found in the Appendix.

Theorem 2.1. *There exists a gamma-exponential problem for which (2.7) has a nonzero probability of never measuring a particular alternative.*

3. The Gittins index as a stopping boundary

Our analysis is based on the idea of continuous-time interpolation, first proposed by Brezzi and Lai (2002) for Gaussian rewards. If the rewards are non-Gaussian, but infinitely divisible, we construct a continuous-time process (X_t) such that, for integer t, the increment $X_{t+1} - X_t$ has the same distribution as W_{t+1}. We then formulate and study the Gittins stopping problem on this process.

3.1. Continuous-time conditional Lévy interpolation

We follow the theoretical characterization of conditional Lévy processes introduced in Çinlar (2003). Let (X_t) be a real-valued stochastic process that will later serve in Section 4 as the continuous-time interpolation of cumulative rewards without discounting. Let λ be a random variable (or random vector) such that, given λ, the process (X_t) has conditionally stationary and independent increments. We further restrict (X_t) to increasing and right-continuous pure jump processes; 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

$$X_t = X_0 + \int_{[0,t]\times\mathbb{R}^+} z\mu(\mathrm{d}s, \mathrm{d}z),$$

where μ is conditionally (given λ) a random measure on $\mathbb{R}^+ \times \mathbb{R}^+$ with mean measure $\nu(\lambda, \mathrm{d}z)\,\mathrm{d}s$, satisfying $\int_{\mathbb{R}^+} \nu(\lambda, \mathrm{d}z)(z \wedge 1) < \infty$ for all λ (for details on random measure and mean measure, see Çinlar (2011, Chapter 6)). The intensity measure of μ at time t, that is, the intensity given $\mathcal{F}_t$ but not given λ, is written as $\bar{\nu}_t(\mathrm{d}z)\,\mathrm{d}s = \mathbb{E}[\nu(\lambda, \mathrm{d}z) \mid \mathcal{F}_t]\,\mathrm{d}s$. Thus, $\bar{\nu}$ can be described as 'the mean of the conditional mean measure'.

The Gittins logic can be extended to the continuous-time setting as follows. Let c be a continuous-time discount factor (lower values of c correspond to higher values of γ in discrete time). The Gittins index R is the particular value of r such that

$$r\int_0^\infty \mathrm{e}^{-cs}\,\mathrm{d}s = \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}X_s + r\int_\tau^\infty \mathrm{e}^{-cs}\,\mathrm{d}s\biggr], \tag{3.1}$$

where τ denotes a stopping time. This expectation is evaluated given some initial state k_0; we take the starting time to be 0 without loss of generality, since the Gittins index only depends on the current time through the current state. This formulation is equivalent to the one in (2.4); see, e.g. Katehakis and Veinott (1987) or Yao (2006). As before, discounted rewards are collected from (X_t) until time τ, at which point we collect the fixed retirement reward r until the end of time. If (3.1) holds, we are indifferent between stopping immediately and running the process optimally.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

3.2. Characterization as a free-boundary problem

In the Gaussian setting, (X_t) is a conditional Brownian motion, and can be converted into a standard Wiener process via a time change; see Brezzi and Lai (2002). Then, (3.1) can be solved on the transformed process using simulation (Yao (2006)) or a free-boundary heat equation (Chick and Gans (2009), Chick and Frazier (2012)). For a more general conditional Lévy process, we can still apply a time change (Monroe (1978)), but it will be random and computationally intractable. Instead, we apply a new approach based on Peskir and Shiryaev (2006). We manipulate (3.1) as follows:

$$\begin{aligned}
0 &= \sup_{\tau} \mathbb{E}\biggl[\int_{[0,\tau]\times\mathbb{R}^+} \mathrm{e}^{-cs} z\mu(\mathrm{d}z,\mathrm{d}s) - \int_0^\tau \mathrm{e}^{-cs} r\,\mathrm{d}s\biggr] \\
&= \sup_{\tau} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\biggl(\int_{\mathbb{R}^+} z\nu(\lambda,\mathrm{d}z) - r\biggr)\mathrm{d}s \\
&\qquad + \int_{[0,\tau]\times\mathbb{R}^+} \mathrm{e}^{-cs} z[\mu(\mathrm{d}z,\mathrm{d}s) - \nu(\lambda,\mathrm{d}z)\,\mathrm{d}s]\biggr] \qquad (3.2) \\
&= \sup_{\tau} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\biggl(\int_{\mathbb{R}^+} z\nu(\lambda,\mathrm{d}z) - r\biggr)\mathrm{d}s \\
&\qquad + \int_{[0,\tau]\times\mathbb{R}^+} \mathrm{e}^{-cs} z\mathbb{E}[\mu(\mathrm{d}z,\mathrm{d}s) - \nu(\lambda,\mathrm{d}z)\,\mathrm{d}s \mid \mathcal{F}_s]\biggr] \qquad (3.3) \\
&= \sup_{\tau} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\biggl(\int_{\mathbb{R}^+} z\nu(\lambda,\mathrm{d}z) - r\biggr)\mathrm{d}s\biggr] \\
&= \sup_{\tau} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\biggl(\int_{\mathbb{R}^+} z\bar{\nu}_s(\mathrm{d}z) - r\biggr)\mathrm{d}s\biggr]. \qquad (3.4)
\end{aligned}$$

In (3.2) we use a compensating technique by adding and subtracting $\nu(\lambda,\mathrm{d}z)$. The random measure μ is cancelled in (3.3) by applying the tower property. We use the tower property again in (3.4).

We denote $\int_{\mathbb{R}^+} z\bar{\nu}_t(\mathrm{d}z)$ by m_t to emphasize that this quantity serves the same role as $m(k_n)$ in discrete time. Then, (3.1) in continuous time can be written as

$$\sup_{\tau} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(m_s - r)\,\mathrm{d}s\biggr] = 0, \qquad (3.5)$$

which we will refer to as the 'calibration equation' throughout this paper. We also write the left-hand side of (3.5) as a function of the starting state,

$$V(t,m) := \sup_{\tau} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(m_s - r)\,\mathrm{d}s\biggr]. \qquad (3.6)$$

Recall that the expectation in (3.6) is evaluated given some initial state at time 0. The pair (t,m), representing a time parameter and a mean parameter, is a set of sufficient statistics for the distribution of λ given $\mathcal{F}_t$. In this value function, r is a fixed constant value and the Gittins index $R(t,m)$ is the particular value of r that makes $V(t,m) = 0$. On the other hand, if we fix r, the set of pairs (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. We rely on the mild condition that (m_t) is a càdlàg strong Markov process; this assumption holds for the Bayesian problems considered in Section 4.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

From Dynkin (1965), the *characteristic operator* of V is defined as

$$L^{\mathrm{char}}V(t,m)=\lim_{U\downarrow\{m\}}\frac{\mathbb{E}[V[t_{\tau_{U^c}},m_{\tau_{U^c}}]]-V(t,m)}{\mathbb{E}[\tau_{U^c}]},\tag{3.7}$$

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, $\tau_{U^c}=\inf\{t\geq 0\colon m_t\in U^c\}$ is the first time at which (m_t) leaves the set U. We now show that (3.7) has a closed-form expression.

Lemma 3.1. *If (m_t) is a càdlàg strong Markov process, then $L^{\mathrm{char}}V$ is given by*

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

Proof. The lemma follows from Peskir and Shiryaev (2006, Equation (7.2.8)). In a killed Lagrange problem on the value function $\int_0^\tau \mathrm{e}^{-\Lambda(s)}L(m_s)\,\mathrm{d}s$, 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) satisfies

$$V(t,m_t)=V(0,m_0)+\int_0^t L^{\mathrm{inf}}V(s,m_s)\,\mathrm{d}s+Y_t,\tag{3.8}$$

where (Y_t) is a martingale formed by adding and subtracting a continuous compensator to the jump component of V; see Itô *et al.* (2004) for an exposition of this idea. We assume that m_t can be written as $g(t,X_t)$ for some continuous function g with first-order derivatives. In Section 4 we will explicitly derive g for gamma-exponential and gamma-Poisson problems.

Lemma 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, then the infinitesimal operator of V is given by*

$$\begin{aligned}L^{\mathrm{inf}}(t,m)&=\frac{\partial V}{\partial t}(t,m)+\frac{\partial g}{\partial t}(t,X_t)\frac{\partial V}{\partial m}(t,m)\\&\quad+\int_{\mathbb{R}^+}[V(t,g(t,X_t+z))-V(t,g(t,X_t))]\bar{\nu}_t(\mathrm{d}z).\end{aligned}$$

Proof. First, we calculate

$$\begin{aligned}V(t,m_t)&=V(0,m_0)+\int_0^t\frac{\partial V}{\partial s}(s,m_s)\,\mathrm{d}s+\int_0^t\frac{\partial V}{\partial m}(s,m_s)\,\mathrm{d}m_s^{\mathrm{c}}\\&\quad+\sum_{0<s\leq t}[V(s,m_s)-V(s,m_{s-})]\qquad(3.9)\\&=V(0,m_0)+\int_0^t\frac{\partial V}{\partial s}(s,m_s)\,\mathrm{d}s+\int_0^t\frac{\partial V}{\partial m}(s,m_s)\,\mathrm{d}m_s^{\mathrm{c}}\\&\quad+\int_{[0,t]\times\mathbb{R}^+}[V(s,g(s,X_s+z))-V(s,g(s,X_s))]\mu(\mathrm{d}s,\mathrm{d}z)\\&=V(0,m_0)+\int_0^t\frac{\partial V}{\partial s}(s,m_s)\,\mathrm{d}s+\int_0^t\frac{\partial V}{\partial m}(s,m_s)\frac{\partial g}{\partial s}\,\mathrm{d}s\\&\quad+\int_{[0,t]\times\mathbb{R}^+}[V(s,g(s,X_s+z))-V(s,g(s,X_s))]\bar{\nu}_s(\mathrm{d}z)\,\mathrm{d}s)\\&\quad+\int_{[0,t]\times\mathbb{R}^+}[V(s,g(s,X_s+z))-V(s,g(s,X_s))]\\&\qquad\times(\mu(\mathrm{d}s,\mathrm{d}z)-\nu(\lambda,\mathrm{d}z)\,\mathrm{d}s+\nu(\lambda,\mathrm{d}z)\,\mathrm{d}s-\bar{\nu}_s(\mathrm{d}z)\,\mathrm{d}s).\qquad(3.10)\end{aligned}$$

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

In (3.9), we use Itô's lemma for jump-diffusion processes (Sato (1999, Chapter 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 $\mathrm{d}m_s^{\mathrm{c}} = (\partial g/\partial s)\,\mathrm{d}s$. In (3.10), we apply a compensator technique. As a result, (3.10) has the form of (3.8) where Y_t is the last integral in the expression. It can readily be shown using the tower property that this integral is an $\mathcal{F}_t$-martingale. The desired result follows. □

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 from Peskir and Shiryaev (2006), 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.

Theorem 3.1. *Let r be fixed. If $m_t = g(t, X_t)$, where g is continuous and has first-order derivatives, the value function $V(t, m)$ solves the free-boundary problem*

$$\begin{aligned}\frac{\partial V}{\partial t}(t,m) + \frac{\partial g}{\partial t}(t,X_t)\frac{\partial V}{\partial m}(t,m) + \int_0^\infty [V(t, g(X_t + y)) - V(t,m)]\bar{\nu}_t(\mathrm{d}y)\\ = cV(t,m) - (m - r),\end{aligned}$$

$$V(t, m^*(t)) = 0,$$

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. Using the characteristic and infinitesimal operators shown in Lemmas 3.1 and 3.2, the theorem follows from Peskir and Shiryaev (2006, Chapter 7.2). □

We briefly note that the value function is not time-homogeneous, because the time index t is part of the state variable. In bandit problems on conditional Lévy processes with binary priors (see, e.g. Cohen and Solan (2013)), the binary structure leads to time-homogeneity of the optimal policy. However, for more general priors, t is usually needed (DeGroot (1970)) in order to obtain a sufficient statistic over the observed information, analogous to a sample size in discrete time.

4. Exponential and Poisson rewards

We now apply Theorem 3.1 to problems with exponential and Poisson rewards. Section 4.1 covers the gamma-exponential problem, whereas Section 4.2 covers the gamma-Poisson problem.

4.1. A free-boundary problem for gamma-exponential bandits

In the gamma-exponential problem, our continuous-time interpolation (X_t) is a conditional gamma process with shape parameter 1 and unknown scale parameter $\lambda \sim \mathrm{gamma}(a_0, b_0)$. Letting $\mathcal{F}_t$ be the σ-algebra generated by the path of (X_t) up to time t, the conditional distribution of λ given $\mathcal{F}_t$ is still gamma with posterior parameters $a_t = a_0 + t$ and $b_t = b_0 + X_t$, as in (2.2). 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 $a_t = a_0 + t$.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

Theorem 4.1. *The value function $V(a, m)$ in the gamma-exponential problem solves the free-boundary problem*

$$\frac{\partial V}{\partial a}(a, m) - \frac{m}{a-1}\frac{\partial V}{\partial m}(a, m) + \int_0^\infty [V(a, m+z) - V(a, m)]\frac{1}{z}\biggl(\frac{m}{m+z}\biggr)^a \,\mathrm{d}z$$
$$= cV(a, m) - (m-r),$$

$$V(a, m^*(a)) = 0,$$

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 3.1. In the conditional Lévy process we use to model exponential rewards, the conditional mean measure given λ is $\nu(\lambda, \mathrm{d}y) = \mathrm{e}^{-\lambda y}/y$, the same as in a gamma process, and the distribution of λ given $\mathcal{F}_t$ is $\operatorname{gamma}(a_t, b_t)$. Therefore, the unconditional mean measure $\bar{\nu}_t(\mathrm{d}y)$ is calculated as

$$\bar{\nu}_t(\mathrm{d}y) = \int_0^\infty \frac{\mathrm{e}^{-\lambda y}}{y}\frac{b_t^{a_t}\lambda^{a_t-1}\,\mathrm{e}^{-b_t\lambda}}{\Gamma(a_t)}\,\mathrm{d}\lambda\,\mathrm{d}y = \biggl(\frac{b_t}{b_t+y}\biggr)^{a_t}\frac{1}{y}\,\mathrm{d}y,$$

whence $m_t = g(t, X_t) = (b_0 + X_t)/(a_0 + t - 1)$ and $\partial g/\partial t = -(b_0 + X_t)/(a_0 + t - 1)^2 = -m_t/(a_t - 1)$. Also,

$$\int_{\mathbb{R}^+}[V(t, g(t, X_t + y)) - V(t, g(t, X_t))]\bar{\nu}_t(\mathrm{d}y)$$
$$= \int_{\mathbb{R}^+}\biggl[V\biggl(t, \frac{b_t+y}{a_t-1}\biggr) - V\biggl(t, \frac{b_t}{a_t-1}\biggr)\biggr]\biggl(\frac{b_t}{b_t+y}\biggr)^{a_t}\frac{1}{y}\,\mathrm{d}y$$
$$= \int_{\mathbb{R}^+}[V(t, m_t+z) - V(t, m_t)]\biggl(\frac{m_t}{m_t+z}\biggr)^{a_t}\frac{1}{z}\,\mathrm{d}z,$$

where the last equality is obtained by using a change of variable $z = y/(a_t - 1)$. □

We use integration by parts to simplify the free-boundary PIDE from Theorem 3.1. First, we write (3.6) as

$$V(a, m) = \sup_\tau \frac{1}{c}\mathbb{E}\biggl[(m-r) - \mathrm{e}^{-c\tau}(m_\tau - r) + \int_0^\tau \mathrm{e}^{-cs}\frac{\mathrm{d}}{\mathrm{d}s}m_s\biggr].$$

Observe that

$$\frac{\mathrm{d}}{\mathrm{d}s}m_s = \frac{\mathrm{d}}{\mathrm{d}s}\frac{b_0+X_s}{a_0+s-1} = -\frac{b+X_s}{(a+s-1)^2}\,\mathrm{d}s + \frac{1}{a+s-1}\,\mathrm{d}X_s.$$

We take the expectation of this quantity, whence

$$\mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\frac{\mathrm{d}}{\mathrm{d}s}m(a_s, b_s)\biggr]$$
$$= -\mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\biggl(\frac{b_0+X_s}{(a_0+s-1)^2}\biggr)\mathrm{d}s\biggr] + \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\mathbb{E}\biggl[\frac{1}{a_0+s-1}\,\mathrm{d}X_s \biggm| \mathcal{F}_s\biggr]\biggr]$$
$$= -\mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\biggl(\frac{b_0+X_s}{(a_0+s-1)^2}\biggr)\mathrm{d}s\biggr] + \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\biggl(\frac{b_0+X_s}{(a_0+s-1)^2}\biggr)\mathrm{d}s\biggr]$$
$$= 0.$$

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

Consequently, (3.5) can be written as $(1/c)[\sup_\tau \mathbb{E}[\mathrm{e}^{-c\tau}(r - m_\tau)] + m - r] = 0$. We define a new value function $G(a, m) := \sup_\tau \mathbb{E}[\mathrm{e}^{-c\tau}(r - m_\tau)] = cV(a, m) - m + r$ for fixed r, and substitute it into Theorem 4.1 to obtain the following equivalent free boundary problem. This equivalent formulation will be convenient in Sections 5 and 6.

Proposition 4.1. *The value function $G(a, m)$ in the gamma-exponential problem solves the free-boundary problem*

$$\frac{\partial G}{\partial a}(a, m) - \frac{m}{a-1}\frac{\partial G}{\partial m}(a, m) + \int_0^\infty [G(a, m+z) - G(a, m)]\frac{1}{z}\bigg(\frac{m}{m+z}\bigg)^a \,\mathrm{d}z = cG(a, m),$$

$$G(a, m^*(a)) = r - m^*(a),$$

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. By substituting $V(a, m) = [G(a, m) + m - r]/c$ in Theorem 4.1, we obtain

$$\begin{aligned}
&\frac{\partial V}{\partial a}(a, m) - \frac{m}{a-1}\frac{\partial V}{\partial m}(a, m) + \int_0^\infty [V(a, m+z) - V(a, m)]\frac{1}{z}\bigg(\frac{m}{m+z}\bigg)^a \,\mathrm{d}z \\
&\quad = \frac{1}{c}\frac{\partial G}{\partial a}(a, m) - \frac{1}{c}\frac{m}{a-1}\bigg[\frac{\partial G}{\partial m}(a, m) + 1\bigg] \\
&\qquad + \frac{1}{c}\int_0^\infty [G(a, m+z) - G(a, m)]\frac{1}{z}\bigg(\frac{m}{m+z}\bigg)^a \,\mathrm{d}z + \frac{1}{c}\frac{m}{a-1} \\
&\quad = \frac{1}{c}\bigg[\frac{\partial G}{\partial a}(a, m) - \frac{m}{a-1}\frac{\partial G}{\partial m}(a, m) \\
&\qquad\quad + \int_0^\infty [G(a, m+z) - G(a, m)]\frac{1}{z}\bigg(\frac{m}{m+z}\bigg)^a \,\mathrm{d}z\bigg]
\end{aligned}$$

and $cV(a, m) - (m - r) = G(a, m)$. On the stopping boundary,

$$\frac{1}{c}[G(a, m) + m - r] = 0. \qquad \square$$

4.2. A free-boundary problem for gamma-Poisson bandits

In the gamma-Poisson problem, the continuous-time interpolation (X_t) is a Poisson process with unknown rate λ. Again, we assume that $\lambda \sim \operatorname{gamma}(a_0, b_0)$, 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 $a_t = a_0 + X_t$ and $b_t = b_0 + t$, as in (2.3). We then obtain the following free-boundary PIDE through calculating m_t explicitly.

Theorem 4.2. *The value function $V(b, m)$ in the gamma-Poisson problem solves the free-boundary problem*

$$\frac{\partial V}{\partial b}(b, m) - \frac{m}{b}\frac{\partial V}{\partial m}(b, m) + \bigg[V\bigg(b, m + \frac{1}{b}\bigg) - V(b, m)\bigg]m = cV(b, m) - (m - r),$$

$$V(b, m^*(b)) = 0,$$

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.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

Proof. In the gamma-Poisson setting, the conditional mean measure is given by $\nu(\lambda, \mathrm{d}y) = \lambda\delta_1$, where δ_1 is the Dirac delta function. The distribution of λ given $\mathcal{F}_t$ is gamma(a_t, b_t). It is then straightforward to show that $\bar{\nu}_t(\mathrm{d}y) = (a_t/b_t)\delta_1\,\mathrm{d}y$, whence $m_t = (a_0 + X_t)/(b_0 + t)$. Therefore, $m_t = g(t, X_t) = (a_0 + X_t)/(b_0 + t)$ and $\partial g/\partial t = -(a_0 + X_t)/(b_0 + t)^2 = -m_t/b_t$. Finally,

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

as required. □

Again, we use integration by parts to simplify the value function in order to obtain

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

By defining $G(b, m) := \sup_\tau \mathbb{E}[\mathrm{e}^{-c\tau}(r - m_\tau)] = cV(b, m) - m + r$ for fixed r and replacing V in Theorem 4.2, we obtain the equivalent free-boundary problem for the gamma-Poisson model. The proof is the same as that of Proposition 4.1 and we omit it here.

Proposition 4.2. *The value function $G(b, m)$ in the gamma-Poisson problem solves the free-boundary problem*

$$\frac{\partial G}{\partial b}(b, m) - \frac{m}{b}\frac{\partial G}{\partial m}(b, m) + \biggl[V\biggl(b, m + \frac{1}{b}\biggr) - V(b, m)\biggr]m = cG(b, m),$$
$$G(b, m^*(b)) = r - m^*(b),$$

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.

5. Theoretical analysis

In this section we provide theoretical results on the structure of the Gittins index for non-Gaussian problems in continuous time. In Section 5.1 we consider scaling properties, more notably for the gamma-Poisson problem. In Section 5.2 we investigate the continuity and monotonicity of the Gittins index and value function. These properties match the discrete-time results shown in Gittins *et al.* (2011) and Yu (2011), supporting our framework as a generalization of the discrete-time model.

Throughout, we abuse notation slightly by writing the value functions V and G, as well as the Gittins index R, as functions 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) where possible. 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. When needed, we also 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 section.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

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. The proofs are straightforward algebraic derivations, and we omit them.

Lemma 5.1. *In the gamma-exponential model, the predictive distribution of X_t/b_0, given $\mathcal{F}_0$, is the beta-prime distribution with parameters t and a_0.*

Lemma 5.2. *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 $t/(b_0+t)$.*

Next, we establish scaling properties of the Gittins index for both non-Gaussian problems. Theorem 5.1 extends the result of (2.5) to the continuous-time setting, where the Gittins index is defined to be the value of r that solves (3.1); we include this proof for completeness.

Theorem 5.1. *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 (3.5) of the Gittins index to obtain

$$\begin{aligned}\sup_{\tau}\mathbb{E}\bigg[\int_0^{\tau}\mathrm{e}^{-cs}(m_s-r)\,\mathrm{d}s\bigg] &= \sup_{\tau}\mathbb{E}\bigg[\int_0^{\tau}\mathrm{e}^{-cs}\bigg(\frac{b_0+X_s}{a_0+s-1}-r\bigg)\,\mathrm{d}s\bigg]\\ &= b_0\sup_{\tau}\mathbb{E}\bigg[\int_0^{\tau}\mathrm{e}^{-cs}\bigg(\frac{1+X_s/b_0}{a_0+s-1}-\frac{r}{b_0}\bigg)\,\mathrm{d}s\bigg] \qquad (5.1)\\ &= 0.\end{aligned}$$

The factor b_0 in front of (5.1) can be dropped since (5.1) equals 0. By applying the scaling properties of the gamma process and gamma distribution, we see that the process (X_t/b_0) has the same law as a conditional gamma process with the prior $\lambda \sim \text{gamma}(a_0, 1)$. Then, if R balances (3.5), it follows that the index 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. □

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 property. To the best of the authors' knowledge, Theorem 5.2 is the first known scaling result for problems with Poisson rewards.

Theorem 5.2. *Let $\sigma > 0$. In the gamma-Poisson problem, the Gittins index satisfies*

$$R(b,m,c) = \frac{1}{\sigma}R\bigg(\frac{b}{\sigma}, \sigma m, \sigma c\bigg).$$

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

$$\begin{aligned}\sup_{\tau}\mathbb{E}\bigg[\int_0^{\tau}\mathrm{e}^{-cs}(m_s-r)\,\mathrm{d}s\bigg] &= \sup_{\tau}\mathbb{E}\bigg[\int_0^{\tau}\mathrm{e}^{-cs}\bigg(\frac{a_0+X_s}{b_0+s}-r\bigg)\,\mathrm{d}s\bigg]\\ &= \sup_{\tau}\mathbb{E}\bigg[\int_0^{\tau}\mathrm{e}^{-cs}\bigg(\frac{a_0+X_s}{b_0/\sigma+s/\sigma}-r\sigma\bigg)\frac{1}{\sigma}\,\mathrm{d}s\bigg]. \qquad (5.2)\end{aligned}$$

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

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

$$\sup_\tau \mathbb{E}\bigg[\int_0^\tau \mathrm{e}^{-cs}(m_s - r)\,\mathrm{d}s\bigg] = \sup_\tau \mathbb{E}\bigg[\int_0^{\tau/\sigma} \mathrm{e}^{-c\sigma t}\bigg(\frac{a_0 + X_{\sigma t}}{b_0/\sigma + t} - r\sigma\bigg)\,\mathrm{d}t\bigg]$$
$$= \sup_\tau \mathbb{E}\bigg[\int_0^{\tau} \mathrm{e}^{-c\sigma t}\bigg(\frac{a_0 + Y_t}{b_0/\sigma + t} - r\sigma\bigg)\,\mathrm{d}t\bigg].$$

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

Corollary 5.1. *From Theorem 5.2, it follows that*

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

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 from 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 is less computationally intensive.

5.2. Continuity and monotonicity

In discrete time, the Gittins index is known to possess various continuity and monotonicity properties; see Aalto *et al.* (2011) and Yu (2011). Here, we show that our continuous-time framework retains the fundamental structure of discrete-time bandit problems. Higher indices are assigned to states with higher m (higher exploitation value), and smaller t (higher exploration value). Furthermore, even though we use processes with jumps for interpolation, continuity indicates that these jumps are 'smoothed out'. Finally, we prove in Theorem 5.6 that, when $t \to \infty$, the exploration value vanishes and the Gittins index approaches the true mean value almost surely. Consequently, the stopping boundary curves from Section 4 are continuous and monotonically increase to the limit r.

Starting with the next result, we will repeatedly compare two arbitrary prior knowledge states. Let (m_t) denote the mean 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; see Müller and Stoyan (2002) and Shaked and Shanthikumar (2007). We shall follow the notation used in Müller and Stoyan (2002) and will use the usual stochastic order $\le_{\text{st}}$, the convex order $\le_{\text{cx}}$, and the increasing convex order $\le_{\text{icx}}$. For random variables X and Y, $X \le_{\text{st}} Y$ if $f_X(c)/f_Y(c)$ is decreasing in c. Also, $X \le_{\text{cx}} Y$ (respectively, $X \le_{\text{icx}} Y$) if $\mathbb{E}\phi[X] \le \mathbb{E}\phi[Y]$ for all convex functions (respectively, convex and increasing) ϕ. Useful properties that we will use include the equivalent definition $X \le_{\text{st}} Y$ if $F_X \ge F_Y$, the implication $\le_{\text{icx}} \Rightarrow \le_{\text{cx}}$ when $\mathbb{E}X = \mathbb{E}Y$, and the coupling techniques that will be restated as Lemmas 5.4 and 5.5 in this section.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

Lemma 5.3. *In the gamma-exponential and gamma-Poisson problems, the two following stochastic order properties hold for predictive mean processes, for every t:*

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

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

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 that $a_0 = a'_0$, and we denote this common value by a. By Lemma 5.1, we have

$$\begin{aligned}\mathbb{P}(m_t \geq m) &= \mathbb{P}\bigg(\frac{b_0 + X_t}{a+t-1} \geq m \;\bigg|\; a, m_0\bigg) = 1 - F\bigg(\frac{m(a+t-1)}{m_0(a-1)} - 1\bigg),\\ \mathbb{P}(m'_t \geq m) &= \mathbb{P}\bigg(\frac{b'_0 + X_t}{a+t-1} \geq m \;\bigg|\; a, m'_0\bigg) = 1 - F\bigg(\frac{m(a+t-1)}{m'_0(a-1)} - 1\bigg),\end{aligned}$$

where F is the cumulative distribution function (CDF) of the beta$'(t, a)$ distribution. When $m_0 \geq m'_0$, we have

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

whence $\mathbb{P}(m_t \geq m) \geq \mathbb{P}(m'_t \geq m)$, i.e. $F_{m_t} \leq F_{m'_t}$. The gamma-Poisson case is proved in the same way with F being the CDF of the generalized negative binomial distribution from Lemma 5.2.

Secondly, we prove (5.4). We focus on the gamma-exponential case; the gamma-Poisson version can be shown in exactly the same way. In this case, in (5.4) it was assumed that $m_0 = b_0/(a_0 - 1) = b'_0/(a'_0 - 1) = m'_0$, which we denote by m, and $a_0 \leq a'_0$. We prove convex dominance by showing that

$$\begin{aligned} m_t &= \bigg(\frac{b_0 + X_t}{a_0 + t - 1} \;\bigg|\; \lambda \sim \mathrm{gamma}(a_0, b_0)\bigg)\\ &\geq_{\mathrm{cx}} \bigg(\frac{b'_0 + X_t}{a'_0 + t - 1} \;\bigg|\; \lambda \sim \mathrm{gamma}(a_0, b_0)\bigg) \qquad (5.5)\\ &\geq_{\mathrm{cx}} \bigg(\frac{b'_0 + X_t}{a'_0 + t - 1} \;\bigg|\; \lambda \sim \mathrm{gamma}(a'_0, b'_0)\bigg) \qquad (5.6)\\ &= m'_t. \end{aligned}$$

We observe that

$$\begin{aligned}\bigg(\frac{b_0 + X_t}{a_0 + t - 1} \;\bigg|\; \lambda \sim \mathrm{gamma}(a_0, b_0)\bigg) &= \bigg(\frac{b_0 + mt + (X_t - mt)}{a_0 + t - 1} \;\bigg|\; \lambda \sim \mathrm{gamma}(a_0, b_0)\bigg)\\ &= m + \frac{1}{a_0 + t - 1}(X_t - mt \mid \lambda \sim \mathrm{gamma}(a_0, b_0))\end{aligned}$$

and, similarly,

$$\bigg(\frac{b'_0 + X_t}{a'_0 + t - 1} \;\bigg|\; \lambda \sim \mathrm{gamma}(a_0, b_0)\bigg) = m + \frac{1}{a'_0 + t - 1}(X_t - mt \mid \lambda \sim \mathrm{gamma}(a_0, b_0)),$$

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

where $(X_t - mt \mid \lambda \sim \text{gamma}(a_0, b_0))$ is a random variable with mean 0. If we write $Y_t := (X_t - mt \mid \lambda \sim \text{gamma}(a_0, b_0))$, then to prove (5.5) it suffices to show that

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

By Müller and Stoyan (2002, Theorem 1.5.18) for a random variable X with mean 0, we have $aX + b \leq_{\text{icx}} cX + d$, when $0 \leq a \leq c$ and $b \leq d$. Since $1/(a_0 + t - 1) \geq 1/(a_0' + t - 1)$, we have

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

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

Next, (5.6) follows from Shaked and Shanthikumar (2007, Theorem 3.A.21). For the gamma-exponential problem, it suffices to prove the condition of the theorem that, for every convex function ϕ, $\mathbb{E}[\phi(X_t \mid 1/\lambda)]$ is convex in $1/\lambda$. For all $\theta \geq \theta'$ and $\alpha \in (0, 1)$,

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

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, Shaked and Shanthikumar (2007, Theorem 3.A.21) 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). □

We now restate two results (known as the 'coupling' techniques) from Müller and Stoyan (2002). It is worth noting that they only require (5.3) and (5.4), meaning that the subsequent analysis will hold for any Lévy process interpolation as long as Lemma 5.3 holds.

Lemma 5.4. (Müller and Stoyan (2002, Theorem 1.2.4).) *If $(m_t) \geq_{\text{st}} (m_t')$ 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 $\hat{m}_t \geq \hat{m}_t'$ a.s.*

Lemma 5.5. (Müller and Stoyan (2002, Theorem 3.4.2).) *If $(m_t) \geq_{\text{cx}} (m_t')$ 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 $\mathbb{E}[\hat{m}_t \mid \hat{m}_t'] = \hat{m}_t'$.*

If (t_0, m_0) and (t_0', m_0') are two initial states that generate mean processes m_t and m_t' satisfying stochastic dominance $\leq_{\text{st}}$ or convex dominance $\leq_{\text{cx}}$, Lemmas 5.4 and 5.5 give us two processes $\hat{m}_t$ and $\hat{m}_t'$ defined on the same filtration with a.s. dominance or the conditional expectation property, respectively. If $(t_1, m_1), (t_2, m_2), \ldots,$ is a sequence of states that all dominate or are dominated by (t, m), each (t_k, m_k) can be coupled with (t, m). We denote the coupled process of (t_k, m_k) by $\hat{m}_t^k$.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

Theorem 5.3. *If (5.3) and (5.4) hold, then $V(m)$ 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 5.4 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 Lamberton and Pagès (1990), 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 Coquet and Toldo (2007). Therefore, the value function will be unchanged if we write V and G using $\hat{m}_t$ and $\hat{m}'_t$ instead of m_t and m'_t. This provides the almost sure dominance necessary to complete the proof, that is, $\hat{m}_t(\omega) \geq \hat{m}'_t(\omega)$ for almost every ω. We calculate

$$\begin{aligned} V_r(m'_0) &= \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}'_s - r)\,\mathrm{d}s\biggr] \\ &= \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s - r)\,\mathrm{d}s + \int_0^\tau \mathrm{e}^{-cs}(\hat{m}'_s - \hat{m}_s)\,\mathrm{d}s\biggr] \\ &\leq \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s - r)\,\mathrm{d}s\biggr] \\ &= V_r(m_0) \end{aligned}$$

and, similarly,

$$G_r(m_0) = \sup_\tau \mathbb{E}[\mathrm{e}^{-c\tau}(r - \hat{m}'_\tau) + \mathrm{e}^{-c\tau}(\hat{m}'_\tau - \hat{m}_\tau)] \leq \sup_\tau \mathbb{E}[\mathrm{e}^{-c\tau}(r - m'_\tau)] = G_r(m'_0),$$

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 Yu (2011) for discrete time.

Proposition 5.1. *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

$$\begin{aligned} \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(m_s - r')\,\mathrm{d}s\biggr] &= \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(m_s - r) + \mathrm{e}^{-cs}(r - r')\,\mathrm{d}s\biggr] \\ &\geq \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(m_s - r)\biggr]. \end{aligned}$$

Therefore, $V_{r'}(m^*_r) \geq 0$. By monotonicity in Theorem 5.3, we obtain $m^*_r \geq m^*_{r'}$. □

Corollary 5.2. *From Proposition 5.1, 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 5.4. *If (5.3) and (5.4) hold, then $R(m)$ is continuous in m.*

Proof. Monotonicity in Corollary 5.2 guarantees the existence of $\lim_{\varepsilon\to 0^-} R(m+\varepsilon)$ and $\lim_{\varepsilon\to 0^+} R(m+\varepsilon)$ provided $R(m)$ is finite, and it suffices to show that $\lim_{\varepsilon\to 0^-} R(m+\varepsilon) = \lim_{\varepsilon\to 0^+} R(m+\varepsilon) = R(m)$.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

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, recalling from Theorems 4.1 and 4.2 that $G_{R_k}(m_k) = R_k - m_k$ for all k, we obtain $\lim_{k\to\infty} R_k = \lim_{k\to\infty} m_k + \lim_{k\to\infty} G_{R_k}(m_k)$. We denote $\lim_{k\to\infty} R_k$ by $\bar{R}$. By Proposition 5.1, $\bar{R} \leq R$. Now we show that $\bar{R} \geq R$ by calculating

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

In (5.10), we use the coupling technique in Lemma 5.4 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 $\sup_\tau \mathbb{E}[\mathrm{e}^{-c\tau}(R_k - \hat{m}_\tau)] \geq \mathbb{E}[\mathrm{e}^{-c\tau}(R_k - \hat{m}_\tau)]$ for every k and thereby $\lim_{k\to\infty} \sup_\tau \mathbb{E}[\mathrm{e}^{-c\tau}(R_k - \hat{m}_\tau)] \geq \lim_{k\to\infty} \mathbb{E}[\mathrm{e}^{-c\tau}(R_k - \hat{m}_\tau)]$ for each τ. This yields $\lim_{k\to\infty} \sup_\tau \mathbb{E}[\mathrm{e}^{-c\tau}(R_k - \hat{m}_\tau)] \geq \sup_\tau \lim_{k\to\infty} \mathbb{E}[\mathrm{e}^{-c\tau}(R_k - \hat{m}_\tau)]$. Therefore, we have

$$m - \bar{R} + \sup_\tau \mathbb{E}[\mathrm{e}^{-c\tau}(\bar{R} - \hat{m}_\tau)] = \sup_\tau \mathbb{E}\bigg[\int_0^\tau \mathrm{e}^{-cs}(m_s - \bar{R})\,\mathrm{d}s\bigg] \leq 0. \tag{5.13}$$

Since $V_R(m) = \sup_\tau \mathbb{E}[\int_0^\tau \mathrm{e}^{-cs}(m_s - R)\,\mathrm{d}s] = 0$, we obtain

$$\begin{aligned}
0 &= \sup_\tau \mathbb{E}\bigg[\int_0^\tau \mathrm{e}^{-cs}(m_s - \bar{R} + \bar{R} - R)\,\mathrm{d}s\bigg] \\
&\leq \sup_\tau \mathbb{E}\bigg[\int_0^\tau \mathrm{e}^{-cs}(m_s - \bar{R})\,\mathrm{d}s\bigg] + (\bar{R} - R)\sup_\tau \mathbb{E}\bigg[\int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}s\bigg], && (5.14)
\end{aligned}$$

which leads to $(\bar{R} - R)\sup_\tau \mathbb{E}[\int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}s] \geq -\sup_\tau \mathbb{E}[\int_0^\tau \mathrm{e}^{-cs}(m_s - \bar{R})\,\mathrm{d}s] \geq 0$ by (5.13). Since $\sup_\tau \mathbb{E}[\int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}s] \geq 0$, $(\bar{R} - R) \geq 0$ 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. To show $\bar{R} \leq R$, we calculate

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

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

This shows that $V_{\bar{R}}(m) = \sup_\tau \mathbb{E}[\int_0^\tau \mathrm{e}^{-cs}(m_s - \bar{R})\,\mathrm{d}s] \geq 0$, and therefore, as in (5.14),

$$0 \leq \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(m_s - R)\,\mathrm{d}s\biggr] + (R - \bar{R}) \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}s\biggr],$$

which leads to $(R - \bar{R}) \sup_\tau \mathbb{E}[\int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}s] \geq 0$, whence right continuity is proved. □

Lemma 5.6. *Let (5.4) hold. For fixed m, the Gittins index $R(t, m)$ is decreasing in t.*

Proof. Under the convex order in (5.4), this follows from Müller (1997, Theorem 5.4). □

Theorem 5.5. *If (5.3) and (5.4) hold, then $R(t)$ is continuous in t.*

Proof. Monotonicity in Lemma 5.6 guarantees the existence of $\lim_{\varepsilon\to0^-} R(t+\varepsilon)$ and $\lim_{\varepsilon\to0^+} R(t+\varepsilon)$ provided $R(t)$ is finite, and it suffices to show that $\lim_{\varepsilon\to0^-} R(t+\varepsilon) = \lim_{\varepsilon\to0^+} R(t+\varepsilon) = R(t)$.

First, we prove left-continuity. 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 5.1, we have $\bar{R} \leq R$. Now we show that $\bar{R} \geq R$. By taking the limit of both sides of the calibration equation, we obtain

$$\begin{aligned}
0 &= \lim_{k\to\infty} \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s^k - R_k)\,\mathrm{d}s\biggr] \\
&= \lim_{k\to\infty} \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s^k - \hat{m}_s + \hat{m}_s - R + R - R_k)\,\mathrm{d}s\biggr] \\
&\geq \sup_\tau \lim_{k\to\infty} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s^k - \hat{m}_s + \hat{m}_s - R + R - R_k)\,\mathrm{d}s\biggr] \\
&= \sup_\tau \biggl\{\lim_{k\to\infty} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s^k - \hat{m}_s)\,\mathrm{d}s\biggr] + \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s - R)\,\mathrm{d}s\biggr] \\
&\qquad + (R - \bar{R}) \lim_{k\to\infty} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}s\biggr]\biggr\} \\
&= \sup_\tau \biggl\{\lim_{k\to\infty} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\mathbb{E}(\hat{m}_s^k - \hat{m}_s \mid \hat{m}_s)\,\mathrm{d}s\biggr] + (R - \bar{R}) \lim_{k\to\infty} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}s\biggr]\biggr\} \\
&= (R - \bar{R}) \lim_{k\to\infty} \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}s\biggr].
\end{aligned} \tag{5.15}$$

By Lemma 5.3, it follows that $\hat{m}_t \leq_{\mathrm{cx}} \hat{m}_t^k$ for every t, and hence in (5.15), we have $\mathbb{E}[\hat{m}_s^k - \hat{m}_s \mid \hat{m}_s] = 0$ by Lemma 5.5. 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 the limit of both sides of the calibration equation, we obtain

$$\begin{aligned}
0 &= \lim_{k\to\infty} \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s^k - \hat{m}_s + \hat{m}_s - R + R - R_k)\,\mathrm{d}s\biggr] \\
&\leq \lim_{k\to\infty} \biggl\{\sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s^k - \hat{m}_s)\,\mathrm{d}s\biggr] + \sup_\tau \mathbb{E}\biggl[\int_0^\tau \mathrm{e}^{-cs}(\hat{m}_s - R)\,\mathrm{d}s\biggr] \\
&\qquad + (R - R_k) \sup_\tau \mathbb{E} \int_0^\tau \mathrm{e}^{-cs}\,\mathrm{d}s\biggr\}
\end{aligned}$$

$$= \lim_{k\to\infty}\biggl\{\sup_{\tau}\mathbb{E}\biggl[\int_0^{\tau} \mathrm{e}^{-cs}\mathbb{E}(\hat{m}_s^k - \hat{m}_s \mid \hat{m}_s^k)\,\mathrm{d}s\biggr] + (R-\bar{R})\sup_{\tau}\mathbb{E}\biggl[\int_0^{\tau}\mathrm{e}^{-cs}\,\mathrm{d}s\biggr]\biggr\}$$
$$= (R-\bar{R})\sup_{\tau}\mathbb{E}\biggl[\int_0^{\tau}\mathrm{e}^{-cs}\,\mathrm{d}s\biggr],$$

whence right-continuity is proved. □

Theorem 5.6. *Let Theorems 5.4 and 5.5 hold. Then, the Gittins index* $\lim_{t\to\infty} R(t,m) = m$ *for each m fixed, and* $R(t,m_t)$ *converges a.s. to* m_∞ *as* $t\to\infty$.

Proof. By Theorems 5.4 and 5.5, $R(t,m)$ is continuous in (t,m). As $t\to\infty$, $m_t\to m_\infty$ a.s. The result then follows from Lemma 2.1, which is easily extended to continuous-time. □

6. Numerical illustration

Solving the problems in Theorems 4.1 and 4.2 numerically poses a substantial challenge, because we do not know the stopping boundary, making it difficult to define suitable initial conditions. This problem properly belongs to the realm of PIDE solution procedures, and thus is outside the scope of this paper. For illustration purposes, we implement an approximation that gives a lower bound on the value function, based on the following 'one-stage' stopping rule (also used by Chick and Gans (2009)). Starting from an initial set of parameters at time 0, we observe the process (X_t) until some fixed time $B \geq 0$. If $m_B < r$, we retire, and if $m_B \geq r$, we continue running the process until ∞. We then calculate the expected earnings, given by $\bar{G}_B = \mathbb{E}[\mathrm{e}^{-cB}(r-m_B)^+]$, 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, $\bar{G}_B$ can be computed in closed form, and $\sup_B \bar{G}_B$ is relatively easy to calculate numerically. The proofs are straightforward and we omit them due to space considerations.

Proposition 6.1. *In the gamma-exponential model*

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

where $A = (r(a_0+B-1)/b_0) - 1$ *and F is the CDF of a beta-prime distribution with parameters B and* a_0.

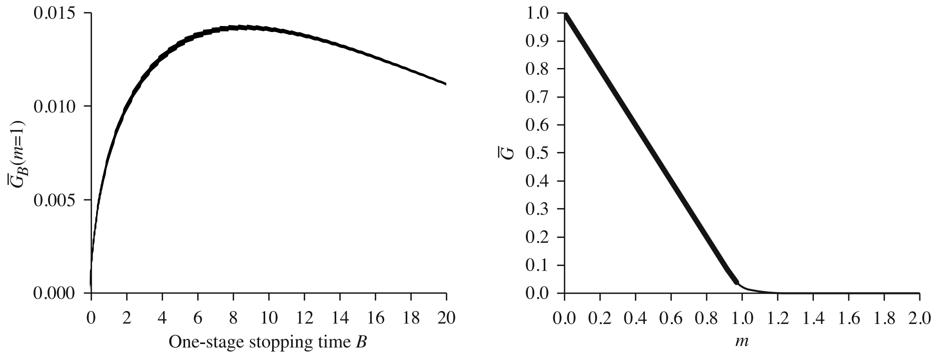
Proposition 6.2. *In the gamma-Poisson model*

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

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* $B/(b_0+B)$.

We use Propositions 6.1 and 6.2 to calculate the initial conditions at (a,m) for some large, fixed a and all $m>0$. Theorem 5.6 implies that the stopping boundary converges to r as $a\to\infty$, which suggests that the behavior of the value function is more stable (and thus easier to approximate) at large a values. The following figures illustrate the one-stage stopping rule and the search for a lower bound through a gamma-exponential example with $r=1$ and $c=0.05$. First, in Figure 1(a) we show 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

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press



(a) Initial value

(b) One-stage searching

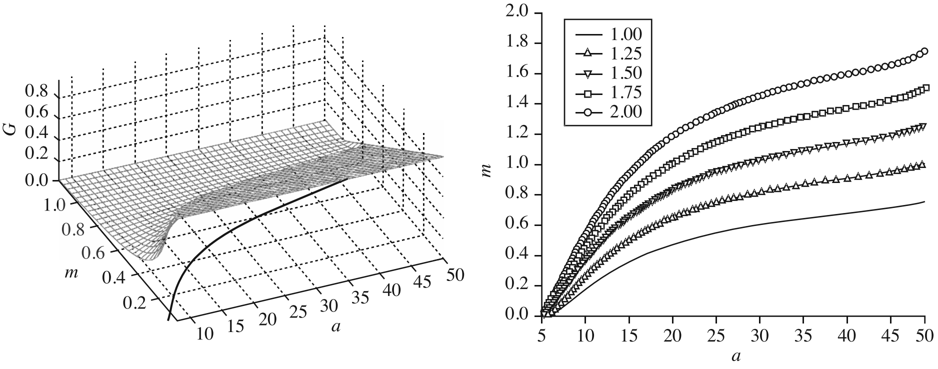
FIGURE 1: Demonstration of initial values obtained from the one-stage searching method. (a) Search for $\sup_B \overline{G}_B(a = 50, m = 1)$ while $r = 1$ and $c = 0.05$. (b) Initial values $\overline{G}_B(a = 50, m = 1)$ while $r = 1$ and $c = 0.05$.

approximation for $G(a, m)$ with $a = 50$ and $m = 1$. In Figure 1(b) we show 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 0 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.

It is preferable to calculate the initial value approximation for large time values, since the quality of the lower bound $\sup_B \bar{G}_B$ is much better when a is large, and then use the PIDEs to build the value function while moving backward in time. In Figure 2(a) we illustrate 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. In Figure 2(b) we show the 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 equal to the given r value; for any state above the curve, we prefer to continue collecting rewards from the process (X_t), whereas for any state below the curve, we prefer to stop and accrue the fixed reward r instead.

We can see that the stopping boundary $m^*(a)$ described by Theorems 3.1 and 4.1 is continuous, increasing, and bounded above by the retirement value r. The growth of m^* slows as the boundary approaches its limiting value from Theorem 5.6. 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, an issue that is outside the scope of this paper. However, the results in Figure 2 demonstrate that the numerical solution behaves in accordance with the theoretical structure of the problem.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press



(a) Solution surface

(b) Stopping boundary

FIGURE 2: Stopping boundaries of the value function and 2D plots for different r values. (a) Solution surface $G(a, m)$ while $c = 0.05$. (b) Stopping boundaries while $c = 0.05$.

7. Conclusion

We have presented a theoretical framework that generalizes multi-armed bandit problems with non-Gaussian rewards in continuous time, using conditional Lévy processes that serve as probabilistic interpolations of the discrete-time reward processes in the bandit problem. We then showed a connection between Gittins indices and free-boundary problems on PIDEs that equate the characteristic and infinitesimal operators of the relevant value function. We have also proved continuity and monotonicity properties of the value functions in these free-boundary problems, as well as the Gittins indices in continuous time. These properties match known discrete-time results, corroborating the use of a continuous-time interpolation to generalize the discrete-time problem.

Our theoretical framework can potentially be applied to any infinitely divisible reward distribution. While this is outside the scope of this paper, this approach could also be extended to more general reward processes and stopping problems, such as those in Chick and Frazier (2012). In this paper we have focused on presenting conditional Lévy interpolation and PIDE construction as a generalization of the well-known diffusion approximation for Gaussian rewards.

Appendix. Proof of Theorem 2.1

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 $b_0^2 < b_0^1/2$. By Ryzhov and Powell (2011, Theorem 3.1), the KG policy will measure alternative 2. Our beliefs about alternative 1 will thus remain unchanged. Let E be the event that $b_n^2/(a_n^2 - 1) < b_0^1/2$ for all $n \geq 0$, implying that we will *never* measure alternative 1. We show that $\mathbb{P}(E) > 0$. For notational convenience, let λ refer to the rate λ^2 of alternative 2, and let $c = b_0^1/2$.

Let (X_t) be a conditional gamma process with shape parameter 1 and scale parameter λ, and let $X_0 = b_0^2$. Then, X_n has the same conditional distribution as b_n^2, given λ. We now observe that

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

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 (A.1) is strictly

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

positive when λ takes values in a nonnegligible set, applying the tower property will show that $\mathbb{P}(E) > 0$.

Consider the case where $\lambda > 1/c$. Now the last expression in (A.1) can be expressed as

$$\mathbb{P}(X_t < c(t+1) \text{ for all } t \geq 0 \mid \lambda) = P\Big(\inf_{t\geq 0} Y_t > -c \Bigm| \lambda\Big),$$

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 (i.e. its jumps are always negative). Because (X_t) is conditionally a gamma process, we must have $\mathbb{E}[X_1 - X_0] = 1/\lambda$ and, hence, $\mathbb{E}[Y_1 - Y_0] = c - 1/\lambda > 0$. In this case

$$\mathbb{P}\Big(\inf_{t\geq 0} Y_t > -c \Bigm| \lambda\Big) = \mathbb{E}[Y_1 - Y_0]w(c + Y_0) = \mathbb{E}[Y_1 - Y_0]w(c - b_0^2), \tag{A.2}$$

where w is called the scale function of the spectrally negative Lévy process (Y_t); see Kyprianou (2006, p. 215). The expression $\mathbb{E}[Y_1 - Y_0]$ in (A.2) 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 $\psi(s) = \log \mathbb{E}\, \mathrm{e}^{s(Y_1 - Y_0)}$. Because $w(x) > 0$ for any $x > 0$, we have shown that the conditional probability is strictly positive given values of λ in a nonnegligible 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.

Acknowledgement

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

References

Aalto, S., Ayesta, U. and Righter, R. (2011). Properties of the Gittins index with application to optimal scheduling. *Prob. Eng. Inf. Sci.* **25,** 269–288.

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

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

Berry, D. A. and Pearson, L. M. (1985). Optimal designs for clinical trials with dichotomous responses. *Statist. Medicine* **4,** 497–508.

Brezzi, M. and Lai, T. L. (2002). Optimal learning and experimentation in bandit problems. *J. Econom. Dynamics Control* **27,** 87–108.

Buonaguidi, B. and Muliere, P. (2013). Sequential testing problems for Lévy processes. *Sequential Anal.* **32,** 47–70.

Caro, F. and Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Manag. Sci.* **53,** 276–292.

Chhabra, M. and Das, S. (2011). Learning the demand curve in posted-price digital goods auctions. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*, pp. 63–70.

Chick, S. E. (2006). Subjective probability and Bayesian methodology. In *Handbooks in Operations Research and Management Science*, Vol. 13, *Simulation*, North-Holland, Amsterdam, pp. 225–258.

Chick, S. E. and Frazier, P. I. (2012). Sequential sampling with economics of selection procedures. *Manag. Sci.* **58,** 550–569.

Chick, S. E. and Gans, N. (2009). Economic analysis of simulation selection problems. *Manag. Sci.* **55,** 421–437.

Chick, S. E. and Inoue, K. (2001). New procedures to select the best simulated system using common random numbers. *Manag. Sci.* **47,** 1133–1149.

Çinlar, E. (2003). Conditional Lévy processes. *Comput. Math. Appl.* **46,** 993–997.

Çinlar, E. (2011). *Probability and Stochastics*. Springer, New York.

Cohen, A. and Solan, E. (2013). Bandit problems with Lévy processes. *Math. Operat. Res.* **38,** 92–107.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

Coquet, F. and Toldo, S. (2007). Convergence of values in optimal stopping and convergence of optimal stopping times. *Electron. J. Prob.* **12,** 207–228.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Dynkin, E. B. (1965). *Markov Processes*. Academic Press, New York.

El Karoui, N. and Karatzas, I. (1994). Dynamic allocation problems in continuous time. *Ann. Appl. Prob.* **4,** 255–286.

Farias, V. F. and Van Roy, B. (2010). Dynamic pricing with a prior on market response. *Operat. Res.* **58,** 16–29.

Filliger, R. and Hongler, M.-O. (2007). Explicit Gittins indices for a class of superdiffusive processes. *J. Appl. Prob.* **44,** 554–559.

Frazier, P. I. and Powell, W. B. (2011). Consistency of sequential Bayesian sampling policies. *SIAM J. Control Optimization* **49,** 712–731.

Frazier, P. I., Powell, W. B. and Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM J. Control Optimization* **47,** 2410–2439.

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

Gittins, J. C. and Wang, Y.-G. (1992). The learning component of dynamic allocation indices. *Ann. Statist.* **20,** 1625–1636.

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

Glazebrook, K. D. and Minty, R. (2009). A generalized Gittins index for a class of multiarmed bandits with general resource requirements. *Math. Operat. Res.* **34,** 26–44.

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

Itô, K., Barndorff-Nielsen, O. E. and Sato, K.-I. (2004). *Stochastic Processes: Lectures Given at Aarhus University*. Springer, Berlin.

Jouini, W. and Moy, C. (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*, IEEE, New York, pp. 299–303.

Kaspi, H. and Mandelbaum, A. (1995). Lévy bandits: multi-armed bandits driven by Lévy processes. *Ann. Appl. Prob.* **5,** 541–565.

Katehakis, M. N. and Veinott, A. F., Jr. (1987). The multi-armed bandit problem: decomposition and computation. *Math. Operat. Res.* **12,** 262–268.

Kyprianou, A. E. (2006). *Introductory Lectures on Fluctuations of Lévy Processes with Applications*. Springer, Berlin.

Lamberton, D. and Pagès, G. (1990). Sur l'approximation des réduites. *Ann. Inst. H. Poincaré Prob. Statist.* **26,** 331–355.

Lariviere, M. A. and Porteus, E. L. (1999). Stalking information: Bayesian inventory management with unobserved lost sales. *Manag. Sci.* **45,** 346–363.

Mandelbaum, A. (1986). Discrete multiarmed bandits and multiparameter processes. *Prob. Theory Relat. Fields* **71,** 129–147.

Mandelbaum, A. (1987). Continuous multi-armed bandits and multiparameter processes. *Ann. Prob.* **15,** 1527–1556.

Monroe, I. (1978). Processes that can be embedded in Brownian motion. *Ann. Prob.* **6,** 42–56.

Müller, A. (1997). How does the value function of a Markov decision process depend on the transition probabilities? *Math. Operat. Res.* **22,** 872–885.

Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. John Wiley, Chichester.

Peskir, G. and Shiryaev, A. N. (2006). *Optimal Stopping and Free-Boundary Problems*. Birkhäuser, Basel.

Powell, W. B. and Ryzhov, I. O. (2012). *Optimal Learning*. John Wiley, Hoboken, NJ.

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

Ryzhov, I. O., Powell, W. B. and Frazier, P. I. (2012). The knowledge gradient algorithm for a general class of online learning problems. *Operat. Res.* **60,** 180–195.

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

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer, New York.

Steele, J. M. (2001). *Stochastic Calculus and Financial Applications*. Springer, New York.

Van Moerbeke, P. (1976). On optimal stopping and free boundary problems. *Arch. Rational Mech. Anal.* **60,** 101–148.

Vazquez, E. and Bect, J. (2010). Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *J. Statist. Planning Infer.* **140,** 3088–3095.

Wang, X. and Wang, Y. (2010). Optimal investment and consumption with stochastic dividends. *Appl. Stoch. Models Business Industry* **26,** 792–808.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press

YAO, Y.-C. (2006). Some results on the Gittins index for a normal reward process. In *Time Series and Related Topics*, Institute of Mathematical Statistics, Beachwood, OH, pp. 284–294.

YU, Y. (2011). Structural properties of Bayesian bandits with exponential family distributions. Preprint. Available at http://arxiv.org/abs/1103.3089.

ZHANG, Q., SEETHARAMAN, P. B. AND NARASIMHAN, C. (2012). The indirect impact of price deals on households' purchase decisions through the formation of expected future prices. *J. Retailing* **88,** 88–101.

https://doi.org/10.1017/apr.2015.9 Published online by Cambridge University Press