J. Appl. Prob. **50,** 388–402 (2013)
Printed in England
© *Applied Probability Trust* 2013

OPEN BANDIT PROCESSES WITH UNCOUNTABLE STATES AND TIME-BACKWARD EFFECTS

XIANYI WU,[*] *East China Normal University and Macquarie University*
XIAN ZHOU,[**] *Macquarie University*

Abstract

Bandit processes and the Gittins index have provided powerful and elegant theory and tools for the optimization of allocating limited resources to competitive demands. In this paper we extend the Gittins theory to more general branching bandit processes, also referred to as *open bandit processes*, that allow uncountable states and backward times. We establish the optimality of the Gittins index policy with uncountably many states, which is useful in such problems as dynamic scheduling with continuous random processing times. We also allow negative time durations for discounting a reward to account for the present value of the reward that was received before the present time, which we refer to as *time-backward effects*. This could model the situation of offering bonus rewards for completing jobs above expectation. Moreover, we discover that a common belief on the optimality of the Gittins index in the generalized bandit problem is not always true without additional conditions, and provide a counterexample. We further apply our theory of open bandit processes with time-backward effects to prove the optimality of the Gittins index in the generalized bandit problem under a sufficient condition.

Keywords: Open bandit process; generalized bandit process; Gittins index; priority scheduling

2010 Mathematics Subject Classification: Primary 90B36; 60G40; 90C40

1. Introduction

Bandit processes and the Gittins index (see, e.g. Gittins and Jones (1974) and Gittins *et al.* (2011)) have played a crucial and prominent role in stochastic scheduling and other areas involving allocating limited resources to competitive demands. The extension to branching bandit problem (also known as the *open bandit* or *arm-acquiring bandit* problem) was first investigated in Nash (1973) and then in Whittle (1981). Thereafter, this topic was further discussed in Varaiya *et al.* (1985), Weiss (1988), Weber (1992), Tsitsiklis (1994), and Bertsimas, and Niño-Mora (1996), among others. In particular, Whittle (1981) presented an elegant and interesting proof for the optimality of Gittins index policies. Other proofs based on interchange arguments were presented in Varaiya *et al.* (1985), Weiss (1988), and Tsitsiklis (1994). Moreover, Bertsimas and Niño-Mora (1996) discussed an *achievable region* method, which is particularly useful for the algebraic computation of the Gittins indices. Another line of proof uses an intuitive deduction from the economical notion, as presented in Weber (1992)

Received 13 December 2011; revision received 12 October 2012.
This research was partially supported by the Natural Science Foundation of China, under grant number 71071056, and the Australian Research Council Discovery Project, under grant number DP1094153.

[*] Postal address: Department of Statistics and Actuarial Science, East China Normal University, Shanghai, China.
[**] Postal address: Department of Applied Finance and Actuarial Studies, Macquarie University, Sydney, Australia. Email address: xian.zhou@mq.edu.au

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

and Ishikida and Varaiya (1994). Under certain stability conditions, Lai and Ying (1988) also showed that Gittins indices for open bandit processes are equivalent to those of traditional (closed) bandit processes if the discount rate approaches 1.

A recent work by Denardo *et al.* (2007) discussed a utility-based closed multiarmed bandit process with finite states, which includes a risk-seeking model that is proved to be equivalent to the generalized bandit model of Nash (1980). Sonin (2008) investigated a multiarmed bandit problem and presented a recursive algorithm to calculate a generalized version of the Gittins index for a given bandit. This problem actually falls into the framework of closed bandit processes discussed in Gittins *et al.* (2011).

The main purpose of this paper is to extend Gittins' theory to more general branching bandit processes that allow uncountably many states and time-backward effects (negative time durations) for discounting the rewards. More specifically, the key contributions of the paper are summarized below.

(i) The theory of the Gittins index for generalized branching bandit processes in the existing literature has so far been limited to finite state spaces. In this paper we establish the optimality of the Gittins index with uncountably many states or arms, which is needed in such problems as dynamic scheduling with continuous random processing times.

(ii) We allow time-backward effects (negative durations) for discounting a reward in the sense that the reward was received sometime ago and its present value is higher than its original value (due to income earned from the reward, such as interest). This could model the situation of offering a bonus reward for completing a job above expectation (cf. Remark 2.2 below). We prove the optimality of the Gittins index with real-valued durations under certain conditions.

(iii) Following the theorem in Nash (1980) on a generalized bandit problem, it has been widely believed that the Gittins index rule remains optimal if the arms with negative Gittins indices are operated before the arms with positive Gittins indices. In this paper, however, we discover that such a common belief is not always true without additional conditions, and provide a counterexample (Example 4.1), which overturns a long-held belief in the literature.

(iv) We identify a sufficient condition for the optimality of the Gittins index in the generalized bandit problem and prove this optimality under the sufficient condition by applying our theory on open bandit processes with time-backward effects (Theorem 4.2).

The rest of the paper is organized as follows. In Section 2 we formulate the problem and provide some preliminaries. In Section 3 we prove our main theory on the optimality of the Gittins index policy for open bandit processes with time-backward effects. In Section 4 we discuss the generalized bandit problems by applying the theory of Section 3.

2. Formulation and preliminaries

In a bandit problem, there is a server and a set of arms of different types that can be operated by the server. Traditionally, each arm may have more than one state. However, we here follow the notation of Whittle (1981) to consider an arm with a different state as a different arm (or different arm type). This enables each arm to take exactly one state without loss of generality.

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

The system is modeled as the following Markov decision setting.

- *States.* There are many types of arms (possibly uncountable), labeled by the elements u of an arbitrary abstract space S, which is equipped with a certain σ-algebra to ensure the measurability of involved functions of the states, especially the rewards. The state of the process at any nonnegative integer time t is indicated by the numbers of arms at every type: $\boldsymbol{n}_t = (n_t(u) \colon u \in S)$ for $t \in \mathbb{N}^+ = \{0, 1, 2, \dots\}$, which is a collection of maps from S to $\mathbb{N}^+$, where $n_t(u)$ is a nonnegative integer indicating the number of arms of type u at time t. While each arm has one type, different arms may share a type with the same probabilistic features. On the other hand, a generic state $\boldsymbol{n} = (n(u) \colon u \in S)$ can also be considered a set of $n(u)$ arms of type u for all $u \in S$. In this sense, we refer to these arms as being *presented in* state $\boldsymbol{n}$.

  For any fixed $v \in S$, define $\boldsymbol{e}(v)$ to be the particular value of $\boldsymbol{n} = (n(u) \colon u \in S)$ such that $n(u) = \mathbf{1}_{\{u=v\}}$, where $\mathbf{1}_E$ denotes the indicator of a set E.

  At time 0, the initial state $\boldsymbol{n}_0$ is known with $n_0(u) = 0$ for all but finitely many $u \in S$, indicating a finite number of arms available at the starting point.

- *Actions.* At any time with the process in state $\boldsymbol{n}$, if an arm of type u from the *action space*

  $$A(\boldsymbol{n}) = \{u \colon n(u) \geq 1\}$$

  is operated, then the server can collect an immediate reward $R(u)$ and the operation gives rise to

  - a random variable $V(u)$, referred to as the *duration*, which may take negative values with positive probabilities, and affects the value of the discounted reward (referred to as *time-backward effects* when $V(u) < 0$); and

  - a new set of arms replacing the arm operated, referred to as the *descendants* of the replaced arm.

  The motivation for a possibly negative duration is presented in Remarks 2.1 and 2.2 below. The numbers of descendant arms at each type are represented by a random map $\boldsymbol{w}(u) = (w(u, v) \colon v \in S)$, where $w(u, v)$ indicates the number of the descendants of type v, which is also subject to the condition that $w(u, v) \geq 1$ for at least one but only for finitely many $v \in S$. Generally, for each fixed u, $\boldsymbol{w}(u)$ is actually a stochastic process with 'time parameter' u, whose distribution can be routinely identified by finite-dimensional distributions. The condition $w(u, v) \geq 1$ for at least one $v \in S$ is to ensure that n_t^π is nonempty under any policy π. This is in turn to ensure the easy exposition in the definition of Gittins indices and the later proof of the optimality of the Gittins index rule when we need an infinite time horizon to make easy exposition of stopping times. The case of $w(u, v) = 0$ for all $v \in S$ (no descendants) can be modeled by adding a dummy arm of absorbing type and sufficiently small reward on operation.

  On a selected arm of type u, the joint distribution of $V(u)$ and $\boldsymbol{w}(u)$ is assumed independent of the history of all operations and the corresponding realization of the decision process up to the current time t. Moreover, it is implicitly assumed independent of the time t so that we essentially obtain a time-homogeneous feature of $(V(u), \boldsymbol{w}(u))$.

- Idle (unoperated) arms are unaffected.

- *Policies and resulting processes.* A policy, generally denoted by π, is a decision rule governing arm selections such that the server can select one and only one available arm

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

of certain type to operate and then obtains an instant reward at any integer time t. It is based on the up-to-date information represented by a filtration (see the next item for details). More specifically, under a specified policy π, at any time t,

- the state is written as $\boldsymbol{n}_t^\pi = (n_t^\pi(u) \colon u \in S)$; the assumptions described above on the descendants ensure that $\boldsymbol{n}_t^\pi$ must satisfy $n_t^\pi(u) < \infty$ for all $u \in S$ and $n_t^\pi(u) > 0$ for only finitely many $u \in S$, so that $\sum_{u \in A(\boldsymbol{n}_t^\pi)} n_t^\pi(u) < \infty$;
- the type of the selected arm is denoted by u_t^π;
- the reward for selecting this arm to operate is $R_t^\pi = R(u_t^\pi)$;
- the duration processes are denoted by $V_t^\pi = V(u_t^\pi)$;
- the arm-acquiring process is denoted by $\boldsymbol{w}_t^\pi = \boldsymbol{w}(u_t^\pi)$; and
- the *cumulative duration* is defined by

$$D_0^\pi = 0, \qquad D_t^\pi = \sum_{j=1}^{t} V_j^\pi, \quad t = 1, 2, \ldots. \tag{2.1}$$

Then $(\boldsymbol{n}_t^\pi, R_t^\pi, D_t^\pi)$ forms a triplet stochastic process in discrete time $t = 0, 1, \ldots$.

- *Filtrations*. The natural filtrations generated by $\{(\boldsymbol{n}_t^\pi, D_t^\pi) \colon t = 0, 1, \ldots\}$ under policy π are denoted by $\mathcal{F}^\pi(\boldsymbol{n}) = \{\mathcal{F}_t^\pi(\boldsymbol{n}) \colon t = 0, 1, \ldots\}$, or simply $\mathcal{F}^\pi = \{\mathcal{F}_t^\pi : t = 0, 1, \ldots\}$ if no confusion arises, where $\boldsymbol{n}$ is the initial state of the system. Clearly, $\{R_t^\pi = R(u_t^\pi) \colon t \geq 0\}$ is $\mathcal{F}^\pi$-adapted. Conditional on the information at time t (i.e. the σ-algebra $\mathcal{F}_t^\pi$), the pair $(V(u), \boldsymbol{w}(u))$ are mutually independent between the arms presented in $\boldsymbol{n}_t^\pi$. Note that $\mathcal{F}_0^\pi = \sigma(\boldsymbol{n}_0)$ is policy independent, written $\mathcal{F}_0$. In addition, in the terms of stochastic processes with filtrations, π is a policy if and only if $\{u_t^\pi : u_t^\pi \in A(\boldsymbol{n}_t^\pi)\}_{t=0}^\infty$ is $\mathcal{F}^\pi$-adapted. For simplicity in expositions, we focus on the natural filtrations in this paper. It is not difficult to alter the framework below to other filtrations that are generally refinements of the natural filtrations and may allow for side information.
- *Final objectives*. Over the infinite time horizon, the server can finally obtain a total discounted reward $\sum_{t=0}^\infty \beta^{D_t^\pi} R_t^\pi$, where $\beta \in (0, 1)$ is the *discount factor*. Denote the expectation under policy π by $\mathbb{E}_\pi$, i.e. the expected total reward is expressed by

$$\mathbb{E}_\pi\biggl[\sum_{t=0}^\infty \beta^{D_t} R_t \biggm| \boldsymbol{n}_0\biggr] = \mathbb{E}\biggl[\sum_{t=0}^\infty \beta^{D_t^\pi} R_t^\pi \biggm| \boldsymbol{n}_0\biggr].$$

The objective is to find a policy π^* to maximize the expected total reward:

$$\mathbb{E}_{\pi^*}\biggl[\sum_{t=0}^\infty \beta^{D_t} R_t \biggm| \boldsymbol{n}_0\biggr] = \max_\pi \mathbb{E}_\pi\biggl[\sum_{t=0}^\infty \beta^{D_t} R_t \biggm| \boldsymbol{n}_0\biggr].$$

Remark 2.1. Here the integer time t indicates the action rounds such that the term 'at time $0, 1, 2, \ldots$' does mean the first, second, third, …, rounds of action, rather than a real calendar time. Therefore, if $V(u) \geq 0$ with probability 1 for all u then this model reduces to the classical branching bandit problem discussed in Weiss (1988). In particular, when the total number of arm types is finite and $V(u) = 1$ is independent of u, this model further reduces to the arm-acquiring bandit process introduced in Whittle (1981).

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

Remark 2.2. A positive $V = V(u)$ can be interpreted as an ordinary 'duration' for calculating the discounted value in the sense that β^V is the *present value* of 1 received V units of time later. In this paper, however, we allow $V < 0$ with a positive probability, and still refer to β^V as the discounted value of 1. When $V < 0$, β^V represents the present value of 1 received $-V$ units of time ago. It is in that sense that a negative V is referred to as a 'negative duration' for the purpose of discounting. Thus, if a type u arm is operated at round t under a policy π and if $V(u) < 0$, then D^π_{t+1} is less than D^π_t by $-V(u)$. Because D^π_t is in effect the true time for reward discounting at the tth round of operation of this system of arms, the reduction of D^π_t to D^π_{t+1} by $-V$ corresponds to turning the clock back by $-V$ units of the calendar time D^π_t. In this sense, a negative duration $V < 0$ indicates a time-backward effect. This may arise in the following scenario. Suppose that an operation is completed at a given time t. If the operation meets certain criteria, then a 20% bonus reward will be paid to all future operations, which has an effect of a multiplier 1.2β for all future rewards. If we write $1.2\beta = \beta^V$ and consider the situation with $\beta > 0.9$, then $V = (\log 1.2/\log\beta) + 1 \leq (\log 1.2/\log 0.9) + 1 = -0.73 < 0$. This is effectively a discounting factor with negative duration $V < 0$.

The Gittins index involves not only the data from this arm but also the data from its descendants as well. For the bandit with initial state $\boldsymbol{e}(u)$, due to the presence of descendants at the operation of this arm, from time 1 onwards, there may be more than one arm available to select and thus a certain policy π is needed to govern the selection among arms. The Gittins index of an arm at type u is defined by

$$M(u) = \sup_{\pi,\tau>0} \frac{\mathbb{E}_\pi[\sum_{t=0}^{\tau-1} \beta^{D_t} R_t \mid u]}{1 - \mathbb{E}_\pi[\beta^{D_\tau} \mid u]}, \tag{2.2}$$

which has the same form as the traditional Gittins index, where

- the u-conditioning means the total discounted reward is collected from the system starting with a single arm u;
- π is any policy governing the selections among the descendants of arm u from time 1 onwards; and
- τ is any stopping time with respect to the filtration $\mathcal{F}^\pi = \{\mathcal{F}^\pi_t(\boldsymbol{e}(u))\colon t = 0, 1, 2, \ldots\}$ in the standard sense that $\{\tau \leq t\} \in \mathcal{F}^\pi_t$ for all $t = 0, 1, 2, \ldots$.

Due to the possibility of negative $V(u)$ for arm type u, one may have $\mathbb{E}_\pi[\beta^{D_\tau} \mid u] \geq 1$, so that the denominator in (2.2) takes zero or negative values. This can be prevented by satisfying the following assumption.

Assumption 2.1. $\sup_{u\in S} \mathbb{E}[\beta^{V(u)}] < 1$.

This assumption is satisfied in two obvious cases.

(i) The number of arm types is finite and $\mathbb{E}[\beta^{V(u)} \mid u] < 1$ for all types u. This in particular covers the traditional case of a finite state space with positive durations.

(ii) The number of arm types is infinite and there exists a $\varepsilon > 0$ such that $V(u) \geq \varepsilon$ for all types u. This in particular covers the Markov bandit processes in which $V(u) \equiv 1$.

Proposition 2.1. *Under Assumption 2.1, $D^\pi_\infty = +\infty$ almost surely and $\mathbb{E}_\pi[\beta^{D_\tau} \mid u] < 1$ for any policy π and $\mathcal{F}^\pi$-stopping time τ.*

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

Proof. Under policy π, the stochastic process $\{\beta^{D_t^\pi} : t = 1, 2, \ldots\}$ is a supermartingale because $\mathbb{E}_\pi[\beta^{D_{t+1}} \mid \mathcal{F}_t^\pi] = \beta^{D_t^\pi}\mathbb{E}[\beta^V] \le \alpha\beta^{D_t^\pi}$ under the proposition assumption. Hence, the martingale convergence theorem states that $\beta^{D_\infty} = \lim_{t\to\infty}\beta^{D_t}$ almost surely. An application of Fatou's lemma shows that

$$0 \le \mathbb{E}_\pi[\beta^{D_\infty} \mid u] = \mathbb{E}_\pi\Big[\lim_{t\to\infty}\beta^{D_t} \,\Big|\, u\Big] \le \liminf_{t\to\infty}\mathbb{E}_\pi[\beta^{D_t} \mid u] \le \lim_{t\to\infty}\alpha^t = 0.$$

This implies that $\beta^{D_\infty} = 0$ almost surely, and, hence, $D_\infty = +\infty$ almost surely. Moreover, for any $\mathcal{F}^\pi$-stopping time τ, Doob's optional stopping time theorem states that

$$\mathbb{E}_\pi[\beta^{D_{\tau\wedge t}} \mid u] \le \mathbb{E}_\pi[\beta^{D_1} \mid u] = \mathbb{E}_\pi[\beta^{V(u)} \mid u] < 1.$$

Letting $t \to \infty$ yields $\mathbb{E}_\pi[\beta^{D_\tau}\mathbf{1}_{\{\tau<\infty\}} \mid u] \le \mathbb{E}_\pi[\beta^{V(u)} \mid u] < 1$. Therefore, for any $\mathcal{F}^\pi$-stopping time τ,

$$\mathbb{E}_\pi[\beta^{D_\tau} \mid u] = \mathbb{E}_\pi[\beta^{D_\tau}\mathbf{1}_{\{\tau<\infty\}} \mid u] + \mathbb{E}_\pi[\beta^{D_\infty}\mathbf{1}_{\{\tau=\infty\}} \mid u] \le \mathbb{E}_\pi[\beta^{V(u)} \mid u] < 1.$$

This completes the proof.

Having defined the Gittins indices $M(u)$ in (2.2), we define, for a generic state $\boldsymbol{n}$ of a bandit process,

$$\bar{M}(\boldsymbol{n}) = \max\{M(u) \colon u \in A(\boldsymbol{n})\}$$

for the maximum Gittins index of the currently available arms.

For a bandit with initial state $\boldsymbol{e}(u)$, let us operate the arms according to the Gittins indices rule (written as policy G): at any time activate an arm with the highest Gittins index. At any time $t \ge 0$, the running states $\boldsymbol{n}_t = (n_t(x) \colon x \in S)$ are such that $n_0(x) = \mathbf{1}_{\{x=u\}}$ and, for $t \ge 1$, $n_t(x) = 0$ for all but finitely many $x \in S$. Define $\tau(u)$ to be the first time to clear out all arms in the descendants of an arm of type u with their Gittins indices no less than $M(u)$ (so that all remaining ones have Gittins indices below $M(u)$); in other words, $\tau(u) = \min\{t \in \mathbb{N}^+ \colon \bar{M}(\boldsymbol{n}_t) < M(u)\}$, with $\tau(u) = \infty$ if $\bar{M}(\boldsymbol{n}_t) \ge M(u)$ for all $t \in \mathbb{N}^+$. Also, write $W_{\tau(u)}$ for the discounted rewards collected in the interval $[0, \tau(u))$. Then $\tau(u)$ is a stopping time and $W(u) \in \mathcal{F}^G_{\tau(u)}(u)$, where $\mathcal{F}^G_t(u)$ is the filtration corresponding to the Gittins policy applied to this bandit with initial state $\boldsymbol{e}(u)$ and $\mathcal{F}^G_{\tau(u)}(u)$ the σ-algebra at stopping time $\tau(u)$. It is not difficult to see that

$$M(u) = \frac{\mathbb{E}[W_{\tau(u)}]}{1 - \mathbb{E}[\beta^{D_{\tau(u)}}]} \tag{2.3}$$

(cf. Varaiya *et al.* (1985) for a general treatment or Bertsimas and Niño-Mora (1996) for the case of finite types). Expression (2.3) indicates that the Gittins indices are achievable in the sense that we have a policy (the Gittins index rule policy) and a stopping time $\tau(u)$ such that $M(u)$ is achieved at their combination.

3. Optimality of the Gittins index

The Gittins index rule selects at any time an arm with the highest Gittins index to operate. In Theorem 3.1 below, we prove the optimality of the Gittins index for the problem formulated in Section 2. This theorem and its proof differ from the existing literature as follows.

(i) It extends the proof invented in Whittle (1981) to the situations with infinitely many states, either countable or uncountable, whereas the results of Whittle (1981) on the optimality of the Gittins index rule are limited to finite state spaces.

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

(ii) It covers for the first time the case of allowing negative durations for discounting effects (see Remark 2.2 for more discussions).

(iii) In the case of finite states and nonnegative durations, our proof is new and simpler than previous ones. In particular, the proof of Whittle (1981) needs the assistance of Doob's optional stopping theorem (referred to as Wald's equality there). This is no longer needed in our proof, which presents a simpler and straightforward treatment by introducing the discounting function (see (3.2) below).

The complication in the following proof mainly arises from the treatment of uncountable arm types and negative time durations associated with every pull of an arm, which correspond to certain generalized bandit problems proposed in Nash (1980) and will be further discussed in Section 4.

Theorem 3.1. *For the problem formulated in Section 2, a policy π is optimal if it operates the arms according to the Gittins index rule, provided Assumption 2.1 is satisfied.*

Proof. Because the proof mainly involves the Gittins index rule, the symbol to indicate the policies as in the previous section is dropped to simplify the notation, except where confusion may arise, we will clearly specify it. More specifically, we use $\{(\boldsymbol{n}_t, R_t, D_t)\colon t = 0, 1, \ldots\}$ for the stochastic process generated under the Gittins index policy, where any tie (a type with more than one arm) can be broken arbitrarily. Denote further by $\mathcal{F}(\boldsymbol{n})$ the natural filtration generated by the process $\{(\boldsymbol{n}_t, R_t, D_t)\colon t = 0, 1, \ldots\}$. The initial state is denoted by $\boldsymbol{n}_0$.

Define $T(x, \boldsymbol{n})$ to be the smallest time (or the first time) needed to clear out all arms (including originals and descendants) with Gittins indices above x from an initial state $\boldsymbol{n}_0 = \boldsymbol{n} = (n(u)\colon u \in S)$ following the Gittins index policy. It can be expressed by

$$T(x, \boldsymbol{n}) = \min\{t \geq 0\colon \bar{M}(\boldsymbol{n}_t) \leq x; \boldsymbol{n}_0 = \boldsymbol{n}\}.$$

We can also define

$$T(x-, \boldsymbol{n}) = \min\{t \geq 0\colon \bar{M}(\boldsymbol{n}_t) < x; \boldsymbol{n}_0 = \boldsymbol{n}\}$$

to be the time at which all the arms with Gittins indices at or above x have been operated. For $T(x, \boldsymbol{n})$ and $T(x-, \boldsymbol{n})$, we have the following facts.

1. $T(x, \boldsymbol{n})$ is right continuous, $T(x-, \boldsymbol{n})$ is left continuous, and $T(x, \boldsymbol{n}) \leq T(x-, \boldsymbol{n})$.

2. Both $T(x, \boldsymbol{n})$ and $T(x-, \boldsymbol{n})$ are stopping times with respect to the filtration $\mathcal{F}(\boldsymbol{n})$, and may thus allow positive probabilities to take the value $+\infty$.

3. It is apparent that

$$T(x, \boldsymbol{n}) = \sum_{s \in A(\boldsymbol{n})} \sum_{l=1}^{n(s)} T_l(s, x, \boldsymbol{n}) = \sum_{\{s \in A(\boldsymbol{n})\colon M(s) \geq M(u)\}} \sum_{l=1}^{n(s)} T_l(s, x, \boldsymbol{n}), \tag{3.1}$$

where $T_l(s, x, \boldsymbol{n})$ indicates the smallest time needed for the lth type s arm to clean up all the arms with Gittins indices above x. For fixed s and u, the $T_l(s, x, \boldsymbol{n})$ are independent and identically distributed over $l = 1, 2, \ldots, n(s)$ as a representative $T(x, \boldsymbol{e}(s))$, provided the bandit begins with the initial state $\boldsymbol{e}(s)$. Here we take the convention that $T(x, \boldsymbol{e}(s)) = 0$ if $M(s) \leq x$. Clearly, for $T(x-, \boldsymbol{n})$, the relationship displayed in (3.1) also holds but with $u-$ in place of u.

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

Suppose that we are now at the time instant $T(x, \boldsymbol{n})$ and write $W(x, \boldsymbol{n})$ for the total discounted reward during the interval working period to clean up all arms with Gittins index x and their descendents with Gittins indices no less than x, valued at time $T(x, \boldsymbol{n})$. Then the total discounted reward valued at time 0 is $\beta^{D(x,\boldsymbol{n})} W(u, \boldsymbol{n})$, where $D(x, \boldsymbol{n}) = D_{T(x,\boldsymbol{n})}$ as defined in (2.1).

We now examine the whole process of the bandit under the Gittins index policy. Starting with the initial state $\boldsymbol{n}$, the server operates one arm at each operation and thus at most countably many types of arms can be operated over the whole time horizon.

1. First let $x_1 = \bar{M}(\boldsymbol{n})$ and so $T(x_1, \boldsymbol{n}) = 0$. The arms with Gittins indices x_1 and their descendants with indices no less than x_1 will be operated from time 0 to $T(x_1-, \boldsymbol{n})$.

2. Next take $x_2 = \bar{M}(\boldsymbol{n}_{T(x_1-,\boldsymbol{n})})$. Then $T(x_2, \boldsymbol{n}) = T(x_1-, \boldsymbol{n})$. At $T(x_2, \boldsymbol{n})$, one selects an arm of type u_2 such that $M(u_2) = x_2$ and then operates up to time $T(x_2-, \boldsymbol{n})$.

Continue this way to obtain $x_3, x_4, \ldots$. Then we generate a strictly decreasing (random) sequence of indices $\{x_i : i \geq 1\}$ such that

$$0 = T(x_1, \boldsymbol{n}) < T(x_1-, \boldsymbol{n}) = T(x_2, \boldsymbol{n}) < \cdots < T(x_{i-1}-, \boldsymbol{n}) = T(x_i, \boldsymbol{n}) < \cdots$$

for $i = 2, 3, \ldots$. Moreover, $T(x, \boldsymbol{n}) = T(x_i, \boldsymbol{n})$ for any $x \in [x_i, x_{i-1})$ and $T(x-, \boldsymbol{n}) = T(x_{i-1}, \boldsymbol{n})$ for any $x \in (x_i, x_{i-1}]$, that is, $T(x, \boldsymbol{n})$ and $T(x-, \boldsymbol{n})$ are both step-down functions in x and $T(x-, \boldsymbol{n})$ is indeed the left limit of $T(x, \boldsymbol{n})$ in the usual sense that $T(x-, \boldsymbol{n}) = \lim_{x' \uparrow x} T(x', \boldsymbol{n})$.

Given a state $\boldsymbol{n}, T(x, \boldsymbol{n})$ itself can be considered as a stochastic process with 'time parameter' x. Consequently, by (2.1),

$$D(x, \boldsymbol{n}) := D_{T(x,\boldsymbol{n})} = \sum_{j=0}^{T(x,\boldsymbol{n})} V_j$$

is also a stochastic process. Therefore, for given $\boldsymbol{n}$, any path of $\mathbf{1}_{\{T(x,\boldsymbol{n})<\infty\}} \beta^{D(x,\boldsymbol{n})}$ is a step function in x.

Under the notation just described, the total discounted reward is given by

$$\tilde{R}(\boldsymbol{n}) = \sum_{i=1}^{\infty} \beta^{D(x_i,\boldsymbol{n})} W(x_i, \boldsymbol{n}) \mathbf{1}_{\{T(x_i,\boldsymbol{n})<\infty\}} .$$

Furthermore, similar to (2.3), we have

$$\mathbf{1}_{\{T(x_i,\boldsymbol{n})<\infty\}} \mathbb{E}[W(x_i, \boldsymbol{n}) \mid \mathcal{F}_{T(x_i,\boldsymbol{n})}] = x_i \mathbf{1}_{\{T(x_i,\boldsymbol{n})<\infty\}} \mathbb{E}[1 - \beta^{D(x_i-,\boldsymbol{n})-D(x_i,\boldsymbol{n})} \mid \mathcal{F}_{T(x_i,\boldsymbol{n})}].$$

Let $\mathrm{Va}(\boldsymbol{n}) = \mathbb{E}[\tilde{R}(\boldsymbol{n})]$ denote the expected total discounted reward, also referred to as the *value function*.

Because $T(x_i-, \boldsymbol{n}) = \infty$ implies that $\beta^{D(x_i-,\boldsymbol{n})} = \beta^{\infty} = 0$ by Proposition 2.1, the step function $\mathbf{1}_{\{T(x,\boldsymbol{n})<\infty\}} \beta^{D(x,\boldsymbol{n})}$ has jumps $\mathbf{1}_{\{T(x_i,\boldsymbol{n})<\infty\}} (\beta^{D(x_i-,\boldsymbol{n})} - \beta^{D(x_i,\boldsymbol{n})})$ (may be positive or negative) at x_i, $i = 1, 2, \ldots$. Note also that $T(x, \boldsymbol{n}) = 0$ for $x \geq x_1$ (recall that $T(x_1, \boldsymbol{n}) = 0$); hence, $D(x, \boldsymbol{n}) = 0$ and so $\mathbf{1}_{\{T(x,\boldsymbol{n})<\infty\}} \beta^{D(x,\boldsymbol{n})} = 1$ for $x \geq x_1$. Combined with the fact that

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

x_i is $\mathcal{F}_{T(x_i,\boldsymbol{n})}$-measurable, it can be readily verified that

$$\begin{aligned}\mathrm{Va}(\boldsymbol{n}) &= \mathbb{E}[\tilde{R}(\boldsymbol{n})]\\ &= \mathbb{E}\biggl[\sum_{i=1}^{\infty} x_i\, \mathbf{1}_{\{T(x_i,\boldsymbol{n})<\infty\}}\, \mathbb{E}[\beta^{D(x_i,\boldsymbol{n})}\mathbb{E}[1-\beta^{D(x_i-,\boldsymbol{n})-D(x_i,\boldsymbol{n})} \mid \mathcal{F}_{T(x_i,\boldsymbol{n})}]]\biggr]\\ &= \mathbb{E}\biggl[\sum_{i=1}^{\infty} x_i\, \mathbf{1}_{\{T(x_i,\boldsymbol{n})<\infty\}}(\beta^{D(x_i,\boldsymbol{n})}-\beta^{D(x_i-,\boldsymbol{n})})\biggr]\\ &= \mathbb{E}\biggl[\int_0^{\infty} x\,\mathrm{d}(\mathbf{1}_{\{T(x,\boldsymbol{n})<\infty\}}\,\beta^{D(x,\boldsymbol{n})})\biggr]\\ &= \mathbb{E}\biggl[\int_0^{\infty}(1-\mathbf{1}_{\{T(x,\boldsymbol{n})<\infty\}}\,\beta^{D(x,\boldsymbol{n})})\,\mathrm{d}x\biggr].\end{aligned}$$

Define a function, referred to as the *discounting function* of the bandit process, by

$$g(x;\boldsymbol{n}) = \mathbb{E}[\mathbf{1}_{\{T(x,\boldsymbol{n})<\infty\}}\,\beta^{D(x,\boldsymbol{n})} \mid \boldsymbol{n}_0=\boldsymbol{n}], \qquad x\in[0,\infty). \tag{3.2}$$

Then we have

$$g(x;\boldsymbol{n}_1+\boldsymbol{n}_2) = g(x;\boldsymbol{n}_1)g(x;\boldsymbol{n}_2) \tag{3.3}$$

due to the facts that

(i) $T(x,\boldsymbol{n}_1+\boldsymbol{n}_2) = T(x,\boldsymbol{n}_1)+T(x,\boldsymbol{n}_2)$;

(ii) $\mathbf{1}_{\{T(x,\boldsymbol{n}_1+\boldsymbol{n}_2)<\infty\}} = \mathbf{1}_{\{T(x,\boldsymbol{n}_1)<\infty\}}\,\mathbf{1}_{\{T(x,\boldsymbol{n}_2)<\infty\}}$; and

(iii) the arms presented in $\boldsymbol{n}_1+\boldsymbol{n}_2$ are independent.

If $n(u)\geq 1$ then $\mathrm{Va}(\boldsymbol{n})$ can be expressed as

$$\begin{aligned}\mathrm{Va}(\boldsymbol{n}) &= \int_0^{\infty}(1-g(x;\boldsymbol{n}))\,\mathrm{d}x\\ &= \int_0^{\infty} x\,\mathrm{d}g(x;\boldsymbol{n})\\ &= \int_0^{\infty} xg(x;\boldsymbol{e}(u))\,\mathrm{d}g(x;\boldsymbol{n}-\boldsymbol{e}(u)) + \int_0^{\infty} xg(x;\boldsymbol{n}-\boldsymbol{e}(u))\,\mathrm{d}g(x;\boldsymbol{e}(u)).\end{aligned}$$

For any reward function $\psi(\boldsymbol{n})$, define

$$L_u\psi(\boldsymbol{n}) = R(u)+\mathbb{E}[\beta^{V(u)}\psi(\boldsymbol{n}-\boldsymbol{e}(u)+\boldsymbol{w}(u)) \mid u], \tag{3.4}$$

where the expectation $\mathbb{E}[\cdot \mid u]$ is with respect to the random variables $V(u)$ and $\boldsymbol{w}(u)$.

As previously mentioned, the model setting ensures that $A(\boldsymbol{n}_t)$ is a finite set. By the theory of dynamic programming, the Gittins index policy is optimal if its value function $\mathrm{Va}(\boldsymbol{n})$ satisfies the optimality equation $\mathrm{Va}(\boldsymbol{n}) = \max_{u\in A(\boldsymbol{n})} L_u\mathrm{Va}(\boldsymbol{n})$. Define

$$\Delta_u(\boldsymbol{n}) = \mathrm{Va}(\boldsymbol{n}) - L_u\mathrm{Va}(\boldsymbol{n}).$$

Then the optimality of the Gittins index is equivalent to the statement

$$\Delta_u(\boldsymbol{n}) = 0 \quad\Longleftrightarrow\quad M(u) = \bar{M}(\boldsymbol{n}), \tag{3.5}$$

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

where '⇔' represents 'if and only if'. To prove (3.5), we modify the bandit process by introducing an auxiliary arm of a specific type (say ∞) that, once operated, gives an instant reward $(1-\beta)m$, constant duration $V(\infty)=1$, and descendants of the same type ∞. We use a superscript m to indicate relevant items in this modified bandit. Then,

(i) this arm of type ∞ and all its descendants have Gittins index m; and

(ii) under the Gittins index rule, once an arm of type ∞ is operated, one inevitably keeps operating arms of type ∞ forever, with the effect of finishing with a final reward m (i.e. $T^m(x,\boldsymbol{n})=T(x,\boldsymbol{n})$ for $x\ge m$ and $T^m(x,\boldsymbol{n})=\infty$ for $x<m$).

For this new bandit, the corresponding discounting function (cf. (3.2)) is

$$g^m(x;\boldsymbol{n})=\mathbb{E}[\mathbf{1}_{\{T(x,\boldsymbol{n})^m<\infty\}}\,\beta^{D(x,\boldsymbol{n})}]=g(x;\boldsymbol{n})\,\mathbf{1}_{\{x\ge m\}}\,.$$

Hence, it is easy to verify that the value function for this new bandit under the Gittins index rule is

$$\begin{aligned}\mathrm{Va}(m;\boldsymbol{n})&=\int_0^\infty(1-g^m(x;\boldsymbol{n}))\,\mathrm{d}x\\&=m+\int_m^\infty(1-g(x;\boldsymbol{n}))\,\mathrm{d}x\\&=\int_0^\infty(1-g(x;\boldsymbol{n}))\,\mathrm{d}x+\int_0^m g(x;\boldsymbol{n})\,\mathrm{d}x\\&=\mathrm{Va}(\boldsymbol{n})+\int_0^m g(x;\boldsymbol{n})\,\mathrm{d}x.\end{aligned}\tag{3.6}$$

It follows immediately that

$$\mathrm{Va}(0;\boldsymbol{n})=\mathrm{Va}(\boldsymbol{n})\quad\text{and}\quad\frac{\partial\mathrm{Va}(m;\boldsymbol{n})}{\partial m}=g(m;\boldsymbol{n}).\tag{3.7}$$

Moreover, let $\Delta_u(m;\boldsymbol{n})=\mathrm{Va}(m;\boldsymbol{n})-L_u\mathrm{Va}(m;\boldsymbol{n})$. Then, by (3.4),

$$\Delta_u(m;\boldsymbol{n})=\mathrm{Va}(m;\boldsymbol{n})-R(u)-\mathbb{E}[\beta^{V(u)}\mathrm{Va}(m;\boldsymbol{n}-\boldsymbol{e}(u)+\boldsymbol{w}(u))\mid u],\tag{3.8}$$

and, consequently,

$$\Delta_u(m;\boldsymbol{e}(u))=\mathrm{Va}(m;\boldsymbol{e}(u))-R(u)-\mathbb{E}[\beta^{V(u)}\mathrm{Va}(m;\boldsymbol{w}(u))\mid u].\tag{3.9}$$

Applying (3.6), it follows that

$$\begin{aligned}&\Delta_u(m;\boldsymbol{n})-\Delta_u(m;\boldsymbol{e}(u))\\&\quad=\mathrm{Va}(m;\boldsymbol{n})-\mathrm{Va}(m;\boldsymbol{e}(u))+\mathbb{E}[\beta^{V(u)}[\mathrm{Va}(m;\boldsymbol{w}(u))-\mathrm{Va}(m;\boldsymbol{n}-\boldsymbol{e}(u)+\boldsymbol{w}(u))]\mid u]\\&\quad=\mathrm{Va}(m;\boldsymbol{n})-\mathrm{Va}(m;\boldsymbol{e}(u))-\mathbb{E}\bigg[\beta^{V(u)}\int_m^\infty g(x;\boldsymbol{w}(u))[1-g(x;\boldsymbol{n}-\boldsymbol{e}(u))]\,\mathrm{d}x\;\bigg|\;u\bigg].\end{aligned}$$

This gives

$$\Delta_u(m;\boldsymbol{n})-\Delta_u(m;\boldsymbol{e}(u))=0\quad\text{for } m\ge\bar{M}(\boldsymbol{n})\tag{3.10}$$

because $\mathrm{Va}(x;\boldsymbol{n})=\mathrm{Va}(x;\boldsymbol{e}(u))=x$ and $g(x;\boldsymbol{n}-\boldsymbol{e}(u))=1$ for $x\ge\bar{M}(\boldsymbol{n})$.

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

Differentiate (3.8) with respect to m and combine it with (3.3), (3.7), and (3.9) to obtain

$$\begin{aligned}\frac{\partial \Delta_u(m; \boldsymbol{n})}{\partial m} &= g(m; \boldsymbol{n}) - \mathbb{E}[\beta^{V(u)} g(m; \boldsymbol{n} - \boldsymbol{e}(u) + \boldsymbol{w}(u)) \mid u] \\ &= g(m; \boldsymbol{n} - \boldsymbol{e}(u))\{g(m; \boldsymbol{e}(u)) - \mathbb{E}[\beta^{V(u)} g(m; \boldsymbol{w}(u)) \mid u]\} \\ &= g(m; \boldsymbol{n} - \boldsymbol{e}(u)) \frac{\partial \Delta_u(m; \boldsymbol{e}(u))}{\partial m}.\end{aligned}$$

Integrating this equation over the interval $(0, \bar{M}(\boldsymbol{n}))$ gives the expression

$$\begin{aligned}\Delta_u(\bar{M}(\boldsymbol{n}); \boldsymbol{n}) - \Delta_u(0; \boldsymbol{n}) = {}& \Delta_u(\bar{M}(\boldsymbol{n}); \boldsymbol{e}(u)) g(\bar{M}(\boldsymbol{n}); \boldsymbol{n} - \boldsymbol{e}(u)) \\ & - \Delta_u(0; \boldsymbol{e}(u)) g(0; \boldsymbol{n} - \boldsymbol{e}(u)) \\ & - \int_0^{\bar{M}(\boldsymbol{n})} \Delta_u(x; \boldsymbol{e}(u)) \, \mathrm{d}g(x; \boldsymbol{n} - \boldsymbol{e}(u)). \qquad (3.11)\end{aligned}$$

By (3.10) together with $g(\bar{M}(\boldsymbol{n}); \boldsymbol{n} - \boldsymbol{e}(u)) = 1$ and $\Delta_u(0; \boldsymbol{e}(u)) = 0$, we see that (3.11) implies that

$$\begin{aligned}\Delta_u(\boldsymbol{n}) &= \Delta_u(0; \boldsymbol{n}) \\ &= \int_0^{\bar{M}(\boldsymbol{n})} \Delta_u(x; \boldsymbol{e}(u)) \, \mathrm{d}g(x; \boldsymbol{n} - \boldsymbol{e}(u)) \\ &= \int_0^{\bar{M}(\boldsymbol{n} - \boldsymbol{e}(u))} \Delta_u(x; \boldsymbol{e}(u)) \, \mathrm{d}g(x; \boldsymbol{n} - \boldsymbol{e}(u)). \qquad (3.12)\end{aligned}$$

Since $\Delta_u(x; \boldsymbol{e}(u)) = 0$ for $x \in [0, M(u)]$ and $\Delta_u(x; \boldsymbol{e}(u)) > 0$ for $x > M(u)$, (3.12) shows that

$$\begin{aligned}\Delta_u(\boldsymbol{n}) = 0 \quad &\iff \quad \Delta_u(x; \boldsymbol{e}(u)) = 0 \quad \text{for all } x \in [0, \bar{M}(\boldsymbol{n} - \boldsymbol{e}(u))] \\ &\iff \quad \bar{M}(\boldsymbol{n} - \boldsymbol{e}(u)) \leq M(u) \\ &\iff \quad M(u) = \bar{M}(\boldsymbol{n}).\end{aligned}$$

This proves (3.5) and thus the theorem.

Remark 3.1. While Theorem 3.1 has so far been proved under the setting of Markov models, it actually holds more generally under the semi-Markov setting or under even more general branching processes if we define the state of a stochastic process at any time t as its filtration history at that time.

4. Generalized branching bandit problems

The generalized bandit problem was first discussed in Nash (1980), under a discrete-time setting with closed bandit processes (i.e. fixed number of arms, $\sum_{x \in A(\boldsymbol{n}_t)} n_t(x) = d$ for some d that is independent of time t, if we use the notation in the previous section) in which the state of every arm evolves according to a Markov fashion and the immediate reward from an arm being operated is not only a function of its state but also influenced by the states of the other frozen arms. This problem under the branching bandit setting appeared to be first investigated in Crosbie and Glazebrook (2000) by means of the popular framework of achievable region methods. In their work, however, only a finite type of arms can be treated, as in all papers with achievable region methods. In this section we apply the general theory for the branching bandits

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

with time-backward effects to deduce the corresponding results of *generalized branching bandit problems* with arbitrary arm types. The deduction is based on the equivalence between the generalized bandits and the bandits with durations (which are semi-Markov bandit problems in the case of positive durations). For easy reference, we first recall the results of Nash (1980) and then apply Theorem 3.1 to the generalized branching bandits.

4.1. Nash's generalized bandit problem

We here follow the notation of Nash (1980) to present the model on a discrete-time process setting and thus the term 'states' in this subsection correspond to arm types in the previous sections. Specifically, we have fixed d arms that are modeled by d stochastic processes $\{X_t(i): t \geq 0\}$ to indicate the state evolving in discrete time on the filtered probability spaces $(\Omega, \mathcal{F}(i), \mathbb{P}(\cdot))$ with filtration $\mathcal{F}(i) = \{\mathcal{F}_t(i): t \geq 0\}$, $i = 1, 2, \dots, d$. For each $i = 1, 2, \dots, d$, $X_t(i)$ represents the state of arm i, taking values in certain abstract space, say $\Omega_i \subset \Omega$, equipped with a σ-algebra, such that $\{X_t(i): t \geq 0\}$ is $\mathcal{F}(i)$-adapted. The state space Ω_i of arm i is also assumed to be the support of $X_t(i)$ in the sense that it is the smallest set such that $\mathbb{P}(\bigcap_{t=1}^{\infty}(X_t(i) \in \Omega_i)) = 1$. The reward process of $X_t(i)$ is given by $R(X_t(i))$.

At any calendar time t, under a policy π, suppose that arm i has been operated for $T(t, i)$ times ($\sum_{i=1}^{d} T(t, i) = t$, $T(t, i) - T(t-1, i) = 0$ if arm i is idle, and $T(t, i) - T(t-1, i) = 1$ if it is operated at time $t-1$), and is thus at state $X_{T(t,i)}(i)$, $i = 1, 2, \dots, d$. Instead of $R(X_{T(t,i)}(i))$ alone, the rewards in this generalized bandit problem are $R(X_{T(t,i)}(i))$ multiplied by the factors $Q(X_{T(t,j)}(j))$ for $j \neq i$, where Q is a strictly positive function of states. We should note that Q and R may depend on the arm for which the rewards and multiplication factors are being computed. For ease of notation, however, we have implicitly used the simplified notation $Q(x)$ and $R(x)$ instead of the more precise $Q(x, i)$ and $R(x, i)$. Hence, the real reward is determined by the states of all arms, rather than just the currently activated arm i. Consequently, the value of a policy π is computed by

$$v(\pi) = \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t \sum_{i=1}^{d} \Bigg\{\prod_{j \neq i} Q(X_{T(t,j)}(j))\Bigg\} R(X_{T(t,i)}(i))\{T(t+1, i) - T(t, i)\}\Bigg]. \tag{4.1}$$

For fixed arm i and nonnegative integer s, denote by $\mathcal{T}_s(i)$ the set of all positive $\mathcal{F}(i)$-stopping times strictly larger than s and write

$$\mathcal{T}_s'(i) = \{\tau \in \mathcal{T}_s(i): \mathbb{E}[Q(X_s(i)) - \beta^\tau Q(X_\tau(i)) \mid \mathcal{F}_s(i)] < 0\}.$$

At a time instant when arm i has been operated for s times, define the index by

$$G_s(i) = \operatorname*{esssup}_{\tau \in \tilde{\mathcal{T}}_s} \frac{\mathbb{E}[\sum_{j=s}^{\tau-1} \beta^j R(X_j(i)) \mid \mathcal{F}_s(i)]}{\mathbb{E}[Q(X_s(i)) - \beta^\tau Q(X_\tau(i)) \mid \mathcal{F}_s(i)]}, \tag{4.2}$$

where $\tilde{\mathcal{T}}_s = \mathcal{T}_s(i)$ if $\mathcal{T}_s'(i) = \varnothing$ or $\tilde{\mathcal{T}}_s = \mathcal{T}_s'(i)$ otherwise. Nash (1980) claimed that at any calendar time t, the optimal policy is to play arm i other than arm j if either $\operatorname{sgn}(G_{T(t,i)}(i)) < \operatorname{sgn}(G_{T(t,j)}(j))$ or $\operatorname{sgn}(G_{T(t,i)}(i)) = \operatorname{sgn}(G_{T(t,j)}(j))$ and $G_{T(t,i)}(i) > G_{T(t,j)}(j)$, where in the case of $G_s(i) = 0$, $\operatorname{sgn}(G_s(i))$ is defined as 1 if $\mathcal{T}_s' = \varnothing$ or -1 otherwise. This result is proved by Nash using an interchange argument.

We can reformulate Nash's model as one similar to the typical closed bandit problem. To this end, define a new reward function $\tilde{R}(x) = R(x)/Q(x)$. Let

$$V_t(i) = 1 + \frac{\log Q(X_t(i)) - \log Q(X_{t-1}(i))}{\log \beta} \quad \text{for } t > 0,$$

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

and define

$$D_0(i) = 0, \qquad D_t(i) = \sum_{l=1}^{t} V_l(i), \quad \text{and} \quad D_t = \sum_{i=1}^{d} D_{T(t,i)}(i) \quad \text{for } t > 0.$$

Then (4.1) can be rewritten as

$$v(\pi) = \prod_{j=1}^{d} Q(X_0(j))\mathbb{E}\bigg[\sum_{t=0}^{\infty} \beta^{D_t} \sum_{i=1}^{d} \tilde{R}(X_{T(t,i)}(i))\{T(t+1,i) - T(t,i)\}\bigg]. \tag{4.3}$$

Since $\prod_{j=1}^{d} Q(X_0(j))$ is independent of the policy π, the performance measure $v(\pi)$ is the same as that of a Markov bandit problem with reward $\tilde{R}$ and durations $V_t(i)$, $i = 1, 2, \dots, d$, $t = 0, 1, \dots$. Recall that Ω_i is the support of $X_t(i)$. If

$$\max_{x,y \in \Omega_i} \frac{Q(x)}{Q(y)} \le \frac{1}{\beta}, \quad \text{or, equivalently,} \quad \frac{\min\{Q(x) \colon x \in \Omega_i\}}{\max\{Q(x) \colon x \in \Omega_i\}} \ge \beta, \tag{4.4}$$

then $V_t(i) \ge 0$ for all i and t, and hence the generalized bandit model can be reduced to the classical closed multiarmed bandits with a Markov structure and positive durations if the evolution of $X_t(i)$ follows a multiarmed bandit process with semi-Markovian arms. This model has been discussed in Section 3 and its Gittins indices thus coincides with (and can be deduced by) that in (4.2). This correspondence has been pointed out in Gittins *et al.* (2011, Section 3.5.1, pp. 65–66) and Glazebrook and Owen (1991).

If (4.4) fails, however, some $V_t(i)$ may take strictly negative values with a positive probability, so that in the induced model, a pull of arm i at time t has the time-backward effect as explained before. This is the most interesting part of the Nash's model. In such a case, the classical result in closed bandit problems cannot produce a solution for Nash's problem, and the solution claimed in Nash (1980) appears invalid.

We provide a counterexample below, in which (4.4) is not satisfied and the Gittins index policy claimed in Nash (1980) fails to optimally solve the bandit problem.

Example 4.1. Consider a closed bandit problem with two deterministic arms. Arm 1 always has a fixed state, say 0, and thus a fixed reward $R(0) = m$ with $Q(0) = 1$. Arm 2 is initiated at state 1 and subject to deterministic state transition $1 \to 2 \to \cdots$ with $Q(u) = \beta^{-2(u-1)}$ and the reward sequence $R(u) = \alpha^{2(u-1)}$, $u = 1, 2, \dots$, for some fixed $\alpha \in (0, \sqrt{\beta})$. Clearly, the Gittins index for arm 1 is positive and the Gittins index for all states of arm 2 are negative because, for any stopping time $\tau > 0$,

$$Q(u) - \beta^{\tau} Q(u+\tau) = \beta^{-2(u-1)} - \beta^{\tau}\beta^{-2(u+\tau-1)} = \beta^{-2(u-1)}(1 - \beta^{-\tau}) < 0.$$

According to Nash (1980), the 'optimum' is the Gittins index rule that operates arm 2 at all times $0, 1, 2, \dots$, which gives the total discounted reward as $\sum_{t=0}^{\infty} \beta^{-t}\alpha^{2t} = \beta/(\beta - \alpha^2)$. On the other hand, a policy that operates arm 1 all the time would have a total reward $m/(1-\beta)$. Therefore, if $m/(1-\beta) > \beta/(\beta - \alpha^2)$, which can be easily achieved by taking a sufficiently large m, then the Gittins index rule is not optimal.

This example can also be converted to the model with a negative constant duration $V = -1$ (cf. (4.3)), which violates the condition of Proposition 2.1 since $\mathbb{E}[\beta^V] = \beta^{-1} > 1$.

To avoid the situation in Example 4.1, Nash's theorem should be modified to the following narrower version, which is deduced from Theorem 3.1. It can be easily checked that Assumption 2.1 corresponds to the following assumption in the setting of the generalized bandit.

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

Assumption 4.1. *There exists a constant* $\delta \in (0, 1)$ *such that*

$$(1-\delta)Q(X_s(i)) - \mathbb{E}[\beta Q(X_{s+1}(i)) \mid \mathcal{F}_s(i)] > 0$$

almost surely for all s *and* $i = 1, 2, \ldots, d$.

By Proposition 2.1, Assumption 4.1 implies that $\mathbb{E}[Q(X_s(i)) - \beta^\tau Q(X_\tau(i)) \mid \mathcal{F}_s(i)] > 0$ for any integer time s and $\mathcal{F}(i)$-stopping time $\tau > s$, $i = 1, 2, \ldots, d$.

Theorem 4.1. *Under Assumption 4.1, the generalized bandit can be optimally operated under the Gittins index rule with the indices defined by*

$$G_s(i) = \operatorname*{esssup}_{\tau > s} \frac{\mathbb{E}[\sum_{m=s}^{\tau-1} \beta^m R(X_m(i)) \mid \mathcal{F}_s(i)]}{\mathbb{E}[Q(X_s(i)) - \beta^\tau Q(X_\tau(i)) \mid \mathcal{F}_s(i)]}, \qquad s = 0, 1, \ldots,\ i = 1, 2, \ldots, d, \tag{4.5}$$

where s *represents the number of times that arm* i *has been operated.*

The Gittins indices in (4.5) are expressed in terms of conditional expectations on filtration. In the special case where the state $X_t(i)$ evolves according to a time-homogeneous Markov fashion such that the time instant plays no role in the definition of Gittins indices, they can be equivalently expressed by

$$M_i(u) = \sup_{\tau > 0} \frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \beta^t R(X_t(i)) \mid X_0(i) = u]}{\mathbb{E}[Q(X_0(i)) - \beta^\tau Q(X_\tau(i)) \mid X_0(i) = u]} \tag{4.6}$$

in terms of conditional expectations on the current state of the arm and the notation introduced in the previous sections.

4.2. Generalized branching bandit problem

Now we return the meaning of 'states' back to what we used in Sections 2 and 3. Here we discuss a straightforward extension of the model in Nash (1980) to the *generalized branching bandit* problems by applying Theorem 3.1, which covers the model analyzed in Crosbie and Glazebrook (2000) for finite arm types. Compared with the branching bandits discussed in Sections 2 and 3, without loss of generality, we can take $V(u) = 1$ and treat the Markov model without extra durations. Any branching with positive durations can be easily translated to the one we are to discuss here. In this model, at any time t with state $\boldsymbol{n}_t$ governed by a policy π, if the server operates an arm $u = u_t^\pi \in A(\boldsymbol{n}_t^\pi)$, it will receive a discounted reward $\beta^t \prod_{v \in A(\boldsymbol{n}_t)} [Q(v)]^{n(v)} R(u_t^\pi)/Q(u_t^\pi)$ and the state evolves to $\boldsymbol{n}_t - \boldsymbol{e}(u) + \boldsymbol{w}(u)$. This corresponds to an instant reward $\tilde{R}(u_t^\pi) = R(u_t^\pi)/Q(u_t^\pi)$ with discounting factor $\beta^t \prod_{v \in A(\boldsymbol{n}_t)} [Q(v)]^{n(v)}$. Define $\tilde{Q}(\boldsymbol{n}) = \prod_{v \in A(\boldsymbol{n})} [Q(v)]^{n(v)}$. In particular, $\tilde{Q}(\boldsymbol{e}(u)) = Q(u)$. At the next time point $t+1$, the discount factor then changes to $\beta^{t+1} \tilde{Q}(\boldsymbol{n}_t - \boldsymbol{e}(u) + \boldsymbol{w}(u)) = \beta^{t+1} \prod_{v \in A(\boldsymbol{n}_t - \boldsymbol{e}(u) + \boldsymbol{w}(u))} [Q(v)]^{n(v)}$. Note that

$$\begin{aligned}
\frac{\beta^{t+1} \tilde{Q}(\boldsymbol{n}_t - \boldsymbol{e}(u) + \boldsymbol{w}(u))}{\beta^t \tilde{Q}(\boldsymbol{n}_t - \boldsymbol{e}(u) + \boldsymbol{w}(u))} &= \frac{\beta^{t+1} \prod_{v \in A(\boldsymbol{n}_t - \boldsymbol{e}(u) + \boldsymbol{w}(u))} [Q(v)]^{n(v)}}{\beta^t \prod_{v \in A(\boldsymbol{n}_t)} [Q(v)]^{n(v)}} \\
&= \frac{\beta}{Q(u)} \prod_{v \in A(\boldsymbol{w}(u))} [Q(v)]^{n(v)} \\
&= \beta^{1 + [\sum_{v \in A(\boldsymbol{w}(u))} n(v) \log Q(v) - \log Q(u)]/\log \beta},
\end{aligned}$$

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press

where the exponential part indicates the duration caused by operating an arm of type u. Therefore, this generalized branching bandit problem corresponds to a semi-Markov branching bandit model with instant rewards $\tilde{R}(u) = R(u)/Q(u)$ and durations

$$\tilde{V}(u) = 1 + \frac{1}{\log \beta}\bigg[\sum_{v \in A(\boldsymbol{w}(u))} n(v) \log Q(v) - \log Q(u)\bigg].$$

In this sense, Assumption 2.1 can be translated to the following.

Assumption 4.2. *There exists a constant* $\delta \in (0, 1)$ *such that*

$$(1 - \delta) Q(u) \geq \mathbb{E}\bigg[\prod_{v \in A(\boldsymbol{w}(u))} [Q(v)]^{n(v)}\bigg].$$

As previously stated, this assumption ensures that $\mathbb{E}_\pi[Q(u) - \beta^\tau Q(\boldsymbol{n}_\tau) \mid \boldsymbol{n}_0 = \boldsymbol{e}(u)] > 0$ for any branching bandit process with an initial state $\boldsymbol{e}(u)$ under any policy π. To define the Gittins indices, consider the branching bandit process initiated by a single arm of type u. Similar to (4.6), by inserting $\tilde{R}(u)$ and $\tilde{V}(u)$ into (2.2) for $R(u)$ and $V(u)$, respectively, the Gittins index for this arm can be defined by

$$M^{\mathrm{g}}(u) = \sup_{\pi, \tau > 0} \frac{\mathbb{E}_\pi[\sum_{t=0}^{\tau-1} \beta^t R_t \mid \boldsymbol{n}_0 = \boldsymbol{e}(u)]}{\mathbb{E}_\pi[Q(u) - \beta^\tau Q(\boldsymbol{n}_\tau) \mid \boldsymbol{n}_0 = \boldsymbol{e}(u)]}, \tag{4.7}$$

where the superscript 'g' stands for the 'generalized' branching bandit and the supremum is over all policies π and positive $\mathcal{F}^\pi$-stopping times τ. Thus we have the following theorem.

Theorem 4.2. *Under assumption 4.2, the Gittins index rule based on (4.7) is optimal for the generalized branching bandit problem just described.*

References

Bertsimas D. and Niño-Mora, J. (1996). Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Math. Operat. Res.* **21,** 257–306.

Crosbie, J. H. and Glazebrook, K. D. (2000). Index policies and a novel performance space structure for a class of generalised branching bandit problems. *Math. Operat. Res.* **25,** 281–297.

Denardo, E. V., Park, H. and Rothblum, U. G. (2007). Risk-sensitive and risk-neutral multiarmed bandits. *Math. Operat. Res.* **32,** 374–394.

Gittins, J. and Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics*, ed. J. Gani, North-Holland, Amsterdam, pp. 241–266.

Gittins, J., Glazebrook, K. and Weber, R. (2011). *Multi-Armed Bandit Allocation Indices*. John Wiley, Chichester.

Glazebrook, K. D. and Owen, R. W. (1991). New results for generalised bandit processes. *Internat. J. Systems Sci.* **22,** 479–494.

Ishikida, T. and Varaiya, P. (1994). Multi-armed bandit problem revisited. *J. Optimization Theory Appl.* **83,** 113–154.

Lai, T. L. and Ying, Z. (1988). Open bandit processes and optimal scheduling of queueing networks. *Adv. Appl. Prob.* **20,** 447–472.

Nash, P. (1973). Optimal allocation of resources between research projects. Doctoral Thesis, Cambridge University.

Nash, P. (1980). A generalized bandit problem. *J. R. Statist. Soc. B* **42,** 165–169.

Sonin, I. M. (2008). A generalized Gittins index for a Markov chain and its recursive calculation. *Statist. Prob. Lett.* **78,** 1526–1533.

Tsitsiklis, J. N. (1994). A short proof of the Gittins index theorem. *Ann. Appl. Prob.* **4,** 194–199.

Varaiya, P. P., Walrand, J. C. and Buyukkoc, C. (1985). Extensions of the multiarmed bandit problem: the discounted case. *IEEE Trans. Automatic Control* **30,** 426–439.

Weber, R. (1992). On the Gittins index for multiarmed bandits. *Ann. Appl. Prob.* **2,** 1024–1033.

Weiss, G. (1988). Branching bandit processes. *Prob. Eng. Inf. Sci.* **2,** 269–278.

Whittle, P. (1981). Arm-acquiring bandits. *Ann. Prob.* **9,** 284–292.

https://doi.org/10.1239/jap/1371648948 Published online by Cambridge University Press