*The Annals of Applied Probability*
1998, Vol. 8, No. 4, 1270–1290

# MULTI-ARMED BANDITS IN DISCRETE AND CONTINUOUS TIME[1]

BY HAYA KASPI AND AVISHAI MANDELBAUM

*Technion–Israel Institute of Technology*

We analyze Gittins' Markovian model, as generalized by Varaiya, Walrand and Buyukkoc, in discrete and continuous time. The approach resembles Weber's modification of Whittle's, within the framework of both multiparameter processes and excursion theory. It is shown that index-priority strategies are optimal, in concert with all the special cases that have been treated previously.

**1. Introduction.** A multi-armed bandit is a control model that supports dynamic allocation of scarce resources in the face of uncertainty [2, 3, 8, 15]. Each arm of the bandit represents an ongoing project and pulling arms corresponds to allocating resources among the projects. In a discrete-time model, arms are pulled one at a time and each pull results in a reward. This is in contrast to continuous time, where a more appropriate view is that of a resource (time, effort) which is to be allocated simultaneously among the arms while accruing rewards continuously. The goal is to identify optimal allocation strategies, and in this paper it is achieved for bandits with independent arms and random rewards, discounted over an infinite horizon.

Specifically, we analyze Gittins's Markovian model [9, 8] in discrete and continuous time, as generalized by [18] and [13] in the spirit of [17]. The approach resembles Weber's [20] modification of [21] (see also [5–7]), within the framework of both multiparameter processes [12, 13] and excursion theory [11]. It differs from [6, 7], which take a martingale-based approach. The outcome is a rigorous proof that is shorter and, in our opinion, conceptually clearer than its predecessors, both in discrete time [18, 5, 3, 12] but especially in continuous time [6, 7, 13, 14]. Of interest also is the connection with general excursion theory [1]; see, for example, the index representation (33), which generalizes (4.3) in [11] from a Markovian setting.

The continuous-time model is formulated and its solution presented in Section 2. One could view discrete-time bandits as a special case of continuous time, where rewards and information change only on a discrete set of predictable epochs. Nevertheless, Section 3 constitutes a self-contained treatment in discrete time: being short and accessible, it provides an introduction to the solution in continuous time, by highlighting main ideas that are not obscured by (unavoidable) technicalities. Properties of the index process are developed in Section 4 and used, in Section 5, to solve the multi-armed bandit problem.

---

Received June 1997; revised December 1997.

[1]Supported by the fund for the promotion of research at the Technion.

*AMS* 1991 *subject classifications*. Primary 60G40; secondary 60J55, 60G44.

*Key words and phrases*. Multi-armed bandits, optional increasing paths, multiparameter processes, excursions, local times, dual predictable projection.

## 2. Problem formulation and solution.

2.1. *Primitives.* Let $\mathscr{T} = [0, \infty)$, $D = \{1, \ldots, d\}$. A $d$-armed bandit model is constructed in terms of adapted stochastic processes $(Z^k, \mathscr{F}^k)$, $k \in D$, on a common probability space $(\Omega, \mathscr{F}, P)$: $Z^k = \{Z^k(t),\ t \in \mathscr{T}\}$, where the random variable $Z^k(t)$ is the reward *rate* obtained after pulling arm $k$ for $t$ units of time; $\mathscr{F}^k = \{\mathscr{F}^k(t),\ t \in \mathscr{T}\}$ is a filtration in $\mathscr{F}$ and the $\sigma$-field $\mathscr{F}^k(t)$ models the information accumulated from pulling arm $k$ for $t$ units of time; $(Z^k(0), \mathscr{F}^k(0))$ is the initial state of affairs associated with time 0; finally, adaptedness means that, for each $k \in D$, $Z^k(t) \in \mathscr{F}^k(t)$, for all $t \in \mathscr{T}$. We assume further:

1. Integrability: for a given $\beta > 0$, $P\int_0^\infty e^{-\beta t}|Z^k(t)|\,dt < \infty$, for all $k \in D$. ($Pf$ denotes the integral of measurable function $f$ with respect to the measure $P$.)
2. Independence: the filtrations $\mathscr{F}^k$, $k \in D$, are independent.
3. Regularity: the filtrations $\mathscr{F}^k$, $k \in D$, satisfy the usual hypotheses of right-continuity and completeness [4].

2.2. *Strategies.* Put $S = \mathscr{T}^d$. An allocation *strategy* $T = \{T(t),\ t \in \mathscr{T}\}$ is an $S$-valued stochastic process: $T(t) = (T^1(t), \ldots, T^d(t))$, where $T^k(t)$ is the total amount of time that $T$ allocates to arm $k$ during the first $t$ units of time. Formally,

$$T(0) = 0 \text{ and } T(t) \text{ is nondecreasing in } t \geq 0, \tag{1}$$

$$T^1(t) + \cdots + T^d(t) = t, \qquad t \in \mathscr{T}, \tag{2}$$

$$\{T(t) \leq s\} \in \mathscr{F}(s), \qquad t \in \mathscr{T}, \quad s \in S, \tag{3}$$

where

$$\mathscr{F}(s) = \mathscr{F}^1(s_1) \vee \cdots \vee \mathscr{F}^d(s_d).$$

Property (3) captures the nonclairvoyant nature of $T$ [13]. By (1) and (2), each $T^k$ is Lipschitz and thus absolutely continuous. Hence, one can talk about rates of increase of each $T^k$. (In the theory of multiparameter processes, $T(t)$ is a stopping point in $S$; an allocation strategy is called an *optional increasing path* [19]; being a nondecreasing family of stopping points, $T$ is also referred to as a multiparameter random time change.)

Under a strategy $T$, the reward rates of the bandit at actual time $t \geq 0$ are given by the random vector $(Z^1[T^1(t)], \ldots, Z^d[T^d(t)])$, and the information available then is the $\sigma$-field

$$\big\{B \in \mathscr{F}\colon B \cap \{T(t) \leq s\} \in \mathscr{F}(s),\ \forall\, s \geq 0\big\}.$$

The present value of $T$ is taken to be

$$v(T) = P\int_0^\infty e^{-\beta t} \sum_{k=1}^{d} Z^k[T^k(t)]\,dT^k(t), \tag{4}$$

where $\beta$ is the discount factor from the integrability condition. Also, in view of integrability, the *value function* $v(T)$ is bounded, as a function of $T$.

2.3. *Solution.* The $d$-armed bandit *problem* is to identify the *optimal* strategies that attain the optimal *value* $V = \sup_T v(T)$. We describe such strategies in terms of adapted *index* processes $(\Gamma^k, \mathscr{F}^k)$, $k \in D$, given by

$$\Gamma^k(t) = \operatorname*{ess\,sup}_{\tau > t} \frac{P_t^k \int_t^\tau e^{-\beta u} Z^k(u)\, du}{P_t^k \int_t^\tau e^{-\beta u}\, du}. \tag{5}$$

Here $\tau$ is a stopping time with respect to $\mathscr{F}^k$ and $P_t^k$ is an abbreviation for the conditional expectation given $\mathscr{F}^k(t)$. ($\Gamma^k$ is finite, as a consequence of the integrability condition; it is also progressively measurable, see [6].) Note also that the $\Gamma^k$'s are independent since the definition of each $\Gamma^k$ entails data of only arm $k$, that is, $\Gamma^k \in \mathscr{F}^k$. Introduce the *lower envelope* $\underline{\Gamma}^k$ of each $\Gamma^k$ by

$$\underline{\Gamma}^k(t) = \inf_{0 \le u \le t} \Gamma^k(u), \qquad t \in \mathscr{T}. \tag{6}$$

Let $M^k$ be the closure of the set $\{t \ge 0\colon \Gamma^k(t) = \underline{\Gamma}^k(t)\}$. Elements in the complement of $M^k$ are *excursion* times of $\Gamma^k$ from its lower envelope $\underline{\Gamma}^k$. They constitute an open set, which is a countable union of disjoint open *excursion intervals*.

A strategy $\hat{T} = (\hat{T}^1, \dots, \hat{T})^d$ is an *index* strategy if each $\hat{T}^k = \{\hat{T}^k(t) \ge 0\}$ right increases at a time $t \ge 0$ only when

$$\Gamma^k\big[\hat{T}^k(t)\big] = \bigvee_{j=1}^{d} \Gamma^j\big[\hat{T}^j(t)\big], \tag{7}$$

and $M^k$ includes all the times that are either left increase of $\hat{T}^k$ but not right increase or on which $\hat{T}^k$ increases at a rate smaller than 1. This definition mathematically articulates two properties: first, $\hat{T}$ *follows the leader* (largest) among the index processes; second, over an excursion interval, time must be allocated exclusively to a single arm, without switching. Index strategies need not be unique, as discussed in [13]. Nonuniqueness arises when time is allocated simultaneously to two arms, say arms $j \ne k$ (this must happen during epochs that are in both $M^j$ and $M^k$), and they start an excursion interval simultaneously (necessarily from the same level). In such circumstances, an index strategy must allocate time only to a single arm, at least until the end of its present excursion, and the prescription of an index strategy leaves the choice of this single arm unresolved. (To the best of our understanding, the "synchronization identity" of [7] does not quite address this problem.) Our solution of the bandit problem requires that this resolution must depend on the path strictly before the excursion starts, as made precise at the beginning of Section 5. One way of ensuring this is through the enforcement of priorities [13]: a *static* priority scheme is a permutation $(i_1, \dots, i_d)$ of $D$; $\hat{T}$ is an *index-priority* strategy if it *adheres to such a scheme*, namely, it allocates time at $t$

to arm $i_m$, $m > l$, only when

$$\underline{\Gamma}^{i_l}\big[\hat{T}^{i_l}(t)\big] > \underline{\Gamma}^{i_l}(u) \quad \text{for all } u > \hat{T}^{i_l}(t).$$

THEOREM 1. *Index-priority strategies exist and are optimal. Furthermore, the optimal value is expressed in terms of the lower envelopes of the indices as*

$$V = P\int_0^\infty e^{-\beta t} \bigvee_{k=1}^{d} \underline{\Gamma}^k\big[\hat{T}^k(t)\big]\,dt, \tag{8}$$

*where* $\hat{T} = (\hat{T}^1, \ldots, \hat{T}^d)$ *is any index-priority strategy.*

Index-priority strategies were constructed in Section 5.1 of [11] (see (3.29) of [7] as well). We also recommend Section 5.2 in [11] for an illuminating sample-path decomposition of the *switched* process $(\Gamma^1[\hat{T}^1], \ldots, \Gamma^d[\hat{T}^d])$, which demonstrates the role that local time plays in quantifying switching in continuous time.

Theorem 1 provides sufficient conditions for optimality. As for necessity, it will become apparent from our proof that there exist optimal strategies that do not adhere to static priorities (assigning priorities predictably suffices). Since we have been unable to completely characterize the class of optimal strategies, we have decided not to elaborate any further on priority rules which, even when treated in full generality, will still provide only a partial answer to the above sought-after characterization.

2.4. *Deteriorating bandits.* The proof of Theorem 1 entails a reduction of the general bandit problem to that of a deteriorating bandit, for which the solution is immediate. A $d$-armed bandit is *deteriorating* if its rewards do not improve with time, that is,

$$Z^k(t) \geq Z^k(u) \quad \text{for all } t \leq u \text{ and } k \in D.$$

It is directly verifiable that the optimal strategies $\hat{T}$ for deteriorating bandits are myopic. Indeed, when present rewards always dominate future ones, the discounting implies that $\hat{T}$ is optimal if and only if it prescribes pulling arms with maximal immediate rewards [$\Gamma^k = Z^k$ in (7)]; its value is then given by

$$V = v(\hat{T}) = P\int_0^\infty e^{-\beta t} \bigvee_{k=1}^{d} Z^k\big[\hat{T}^k(t)\big]\,dt. \tag{9}$$

2.5. *Reduction.* A comparison between (8) and (9) reveals a connection between the bandit $(Z^k, \mathscr{F}^k)$ and the deteriorating bandit $(\underline{\Gamma}^k, \mathscr{F}^k)$, $k \in D$. The proof of Theorem 1 is based on this connection, which is further articulated in Theorem 2 below. Specifically, we prove in Section 5 that, for any strategy $T$,

$$v(T) \leq P\int_0^\infty e^{-\beta t} \sum_{k=1}^{d} \underline{\Gamma}^k[T^k(t)]\,dT^k(t), \tag{10}$$

while equality prevails for index-priority strategies $\hat{T}$, namely,

$$v(\hat{T}) = P\int_0^\infty e^{-\beta t}\sum_{k=1}^{d}\underline{\Gamma}^k\big[\hat{T}^k(t)\big]\,d\hat{T}^k(t). \tag{11}$$

The relations (10) and (11) yield Theorem 1. This is verified first by observing that any strategy that follows the leader, also follows the leader with $\underline{\Gamma}^k$ replacing $\Gamma^k$, $k \in D$. Thus, any index-priority strategy $\hat{T}$ is optimal for the deteriorating $(\underline{\Gamma}^k, \mathscr{F}^k)$, $k \in D$. As such, $\hat{T}$ maximizes the right-hand side of (10) over $T$, its value dominates $V$ and is equal to (11). This establishes that index-priority strategies are optimal and their value, as follows from (9), coincides with the right-hand side of (8).

Our optimality result can be summarized by the following.

THEOREM 2. *Let* $(Z^k, \mathscr{F}^k)$, $k \in D$, *be a* $d$*-armed bandit that satisfies the integrability, independence and regularity conditions. Then there exists a deteriorating bandit* $(\underline{\Gamma}^k, \mathscr{F}^k)$, $k \in D$, *with* $\underline{\Gamma}^k$ *given in* (6), *that is equivalent to the original bandit in the following sense*:

(i) *the value of both bandits is equal to* (8);

(ii) *the classes of optimal strategies for both bandits coincide except that, for the original bandit, if* $T$ *is optimal and* $\Gamma^k[T^k(t)] > \underline{\Gamma}^k[T^k(t)]$ *with* $\Gamma^k$ *in* (5), *then, at that time* $t$, $T^k(t)$ *right increases at rate* 1.

REMARK. Following the leader while adhering to a static priority could, conceivably, contradict if the indices of two arms that are pulled simultaneously jump upwards at the same time and the new index of the arm with the lower priority is higher than that of the arm with the higher priority. As will be shown in Section 5, it is impossible for two arms, pulled together by an index-priority strategy, to leave their respective $M$'s at the same time due to a jump of one or two of the indices upwards. This follows from the fact that such jumps of the indices occur at totally inaccessible stopping times for both arms. Thus, if two arms start an excursion from their set of index minima at the same time (and necessarily from the same level), they start from a continuity point of their indices. Therefore, static priorities and following the leader among the indices do not contradict.

**3. Bandits in discrete time.** The multi-armed bandit problem will be now formulated and solved in discrete time. The continuous-time bandit can be viewed as a limit of discrete-time models, taken as the durations of periods between pulls converge to zero [13]. This helps explain the main difference between discrete and continuous time: in the former, time is allocated only to a single arm at a time; as durations shrink, the limit is such that time can be allocated simultaneously and continuously among the arms.

3.1. *Primitives.* Let $\mathscr{N} = \{0, 1, 2, \ldots\}$. The primitives are now adapted stochastic *sequences* $(Z^k, \mathscr{F}^k)$, $k \in D$: $Z^k = \{Z^k(n),\ n \in \mathscr{N}\}$, where $Z^k(n)$ is the reward obtained from pull $n$ of arm $k$; $\mathscr{F}^k = \{\mathscr{F}^k(n),\ n \in \mathscr{N}\}$, where

$\mathscr{F}^k(n)$ is the information accumulated during the first $n$ pulls of arm $k$. The integrability condition now takes the form $E\sum_{n=0}^{\infty}\alpha^n|Z^k(n)|<\infty$, for a given discount factor $0<\alpha<1$ and all $k\in D$, and the independence and regularity conditions are unaltered.

3.2. *Strategies.* Put $S=\mathscr{N}^d$. An allocation *strategy* $T=\{T(t),\ t\in\mathscr{N}\}$ is an $S$-valued stochastic sequence: $T(t)=(T^1(t),\dots,T^d(t))$, where $T^k(t)$ is the number of pulls of arm $k$ during the first $t$ pulls of the bandit's arms. Formally, $T$ satisfies (1)–(3), with $n\in\mathscr{N}$ replacing $t\in\mathscr{T}$. Property (2) is equivalent to the fact that *only one arm is pulled each time*. The value of a strategy $T$ is

$$v(T)=P\sum_{t=0}^{\infty}\alpha^t\sum_{k=1}^{d}Z^k[T^k(t)][T^k(t+1)-T^k(t)].$$

3.3. *Solution.* The solution to the $d$-armed bandit problem uses the *index sequences* $(\Gamma^k,\mathscr{F}^k)$, $k\in D$, where

$$\Gamma^k(n)=\operatorname*{ess\,sup}_{\tau\geq n+1}\frac{P_n^k\sum_{m=n}^{\tau-1}\alpha^m Z^k(m)}{P_n^k\sum_{m=n}^{\tau-1}\alpha^m},\qquad n\in\mathscr{N},\tag{12}$$

and it is given by the following.

THEOREM 1D. *The class of optimal strategies coincides with the class of index strategies $\hat{T}$, which pull arms with the highest index. Formally, an index strategy is a strategy $\hat{T}$ for which*

$$\hat{T}^k(t+1)=\hat{T}^k(t)+1\quad\textit{only when }\Gamma^k\big[\hat{T}^k(t)\big]=\bigvee_{j=1}^{d}\Gamma^j\big[\hat{T}^j(t)\big],\tag{13}$$

*for all $t\in\mathscr{N}$ and $k\in D$. Furthermore, the optimal value is given by*

$$V=P\sum_{t=0}^{\infty}\alpha^t\bigvee_{k=1}^{d}\underline{\Gamma}^k\big[\hat{T}^k(t)\big].\tag{14}$$

3.4. *Reduction to deteriorating bandits.* In a deteriorating bandit, $Z^k(n)\geq Z^k(n+1)$, for all $n\in\mathscr{N}$ and $k\in D$. A strategy $\hat{T}$ is optimal for a deteriorating bandit if and only if it always pulls arms with maximal immediate rewards [$\Gamma^k=Z^k$ in (13)]; its value is then given by

$$V=v(\hat{T})=P\sum_{t=0}^{\infty}\alpha^t\bigvee_{k=1}^{d}Z^k\big[\hat{T}^k(t)\big].\tag{15}$$

To establish Theorem 1D, one proves that, for every strategy $T$,

$$v(T)\leq P\sum_{t=0}^{\infty}\alpha^t\sum_{k=1}^{d}\underline{\Gamma}^k[T^k(t)][T^k(t+1)-T^k(t)],\tag{16}$$

with equality for index strategies $\hat{T}$ as in (13), namely,

$$v(\hat{T})=P\sum_{t=0}^{\infty}\alpha^t\sum_{k=1}^{d}\underline{\Gamma}^k\big[\hat{T}^k(t)\big]\big[\hat{T}^k(t+1)-\hat{T}^k(t)\big].\tag{17}$$

The relations (16) and (17) yield Theorem 1D. This is verified exactly as in continuous time, with the arguments in the paragraph that follows (11): simply replace (10), (11), (9) and (8) there by (16), (17), (15) and (14), respectively. For the converse, it is easy to show that a strategy that starts with arm $k$ with $\Gamma^k(0) < \bigvee_{j=1}^d \Gamma^j(0)$ is not optimal.

3.5. *Summary.* An index strategy that satisfies (13) *follows the leader* among $\Gamma^1, \ldots, \Gamma^d$. Equivalently, for each $k \in D$,

$$\hat{T}^k(t+1) = \hat{T}^k(t) + 1 \quad \text{only when } \underline{\Gamma}^k[T^k(t)] = \bigvee_{j=1}^d \underline{\Gamma}^j[T^j(t)] \tag{18}$$

together with

$$\Gamma^k[T^k(t)] > \underline{\Gamma}^k[T^k(t)] \quad \text{implies } \hat{T}^k(t+1) = \hat{T}^k(t) + 1, \tag{19}$$

for all $t \in \mathcal{N}$. Hence, $\hat{T}$ is optimal for the $d$-armed bandit $(Z^k, \mathcal{F}^k)$, $k \in D$, if and only if it is optimal for the deteriorating $(\underline{\Gamma}^k, \mathcal{F}^k)$, $k \in D$, and it switches arms only at times when the active index process coincides with its lower envelope (e.g., by enforcing a static priority among the arms [13]). Our optimality result can thus be summarized by the following theorem.

THEOREM 2D. *Let $(Z^k, \mathcal{F}^k)$, $k \in D$, be a discrete-time $d$-armed bandit that satisfies the integrability, independence and regularity conditions. Then there exists a deteriorating bandit $(\underline{\Gamma}^k, \mathcal{F}^k)$, $k \in D$, with $\underline{\Gamma}^k$ being the lower envelope of* (12), *which is equivalent to the original bandit in the following sense*:

(i) *the value of both bandits equals* (14);

(ii) *the classes of optimal strategies for both bandits coincide* (*up to* (19)). *Specifically, $\hat{T}$ is optimal if and only if it adheres to* (13) *or*, *equivalently*, *to both* (18) *and* (19).

Theorem 2D was anticipated in [12], Theorem 1′ there and its succeeding Remark. (Note, however, that our relation (16) is the "right" articulation of (2.8) in [12].)

3.6. *The index sequence.* We now focus on a single arm; hence its identifier $k$ will be suppressed as long as no confusion arises. Fix an arm $(Z, \mathcal{F})$, a time $n \in \mathcal{N}$ and a scalar $\gamma$. Associate with the arm a value $v_n(\gamma)$, for playing it optimally after time $n$ until stopping, while paying $\gamma$ for each pull. Formally,

$$v_n(\gamma) = \operatorname*{ess\,sup}_{\tau \geq n+1} P_n \sum_{m=n}^{\tau-1} \alpha^m [Z(m) - \gamma], \tag{20}$$

where $P_n$ is the conditional expectation with respect to $\mathcal{F}(n)$ and $\tau$ is a stopping time with respect to $\mathcal{F}$. As a function of $\gamma$, $v_n(\cdot)$ is nonincreasing and convex, being an upper envelope of $\gamma$-affine decreasing functions. By the integrability condition for $Z$, it is also finite hence continuous, with $v_n(-\infty) = \infty$

and $v_n(\infty) = -\infty$. One concludes that $v_n(\cdot)$ has a unique zero. Denote it by

$$\Gamma(n) = v_n^{-1}(0) \tag{21}$$

and call $\Gamma = \{\Gamma(n),\ n \in \mathcal{N}\}$ the *index sequence* associated with arm $(Z, \mathcal{F})$. We now verify the equivalence of (21) and (12).

To solve (20), apply Snell's optimal-stopping theory [17] to $X(m) = \sum_{j=n}^{m-1} \alpha^j [Z(j) - \gamma]$, $m \geq n+1$, as in Section 6.3 of [12]: $\tau_n(\gamma) = \inf\{m \geq n+1\colon v_m(\gamma) \leq 0\}$ is the stopping time that attains $v_n(\gamma)$ in (20). By (21) and properties of $v_n(\cdot)$, also

$$\tau_n(\gamma) = \inf\bigl\{m \geq n+1\colon \Gamma(m) \leq \gamma\bigr\}, \qquad n \in \mathcal{N}. \tag{22}$$

The equivalence of (21) and (12) is essentially the relation $v_n[\Gamma(n)] = 0$. Indeed,

$$v_n[\Gamma(n)] = 0 \geq P_n \sum_{m=n}^{\tau-1} \alpha^m [Z(m) - \Gamma(n)],$$

for all stopping times $\tau \geq n+1$, with equality to 0 attained by

$$\tau_n[\Gamma(n)] = \inf\bigl\{m \geq n+1\colon \Gamma(m) \leq \Gamma(n)\bigr\}. \tag{23}$$

One deduces for $\Gamma(n)$ in (21) that

$$\Gamma(n) = \operatorname*{ess\,sup}_{\tau \geq n+1} \frac{P_n \sum_{m=n}^{\tau-1} \alpha^m Z(m)}{P_n \sum_{m=n}^{\tau-1} \alpha^m} = \frac{P_n \sum_{m=n}^{\tau_n[\Gamma(n)]-1} \alpha^m Z(m)}{P_n \sum_{m=n}^{\tau_n[\Gamma(n)]-1} \alpha^m}, \tag{24}$$

namely, (12). Our proofs require a generalization of (24) in three aspects: an augmentation of $\mathcal{F}$, a randomization of time $n$ and a manifestation that dilating the discounting $\alpha^m$ can only hurt. Such a generalization is the following.

PROPOSITION 3. *Let $\mathcal{G}$ be a $\sigma$-field that is independent of the filtration $\mathcal{F}$. Then, for every stopping time $\varepsilon$ with respect to $\mathcal{F}(\cdot) \vee \mathcal{G}$,*

$$\Gamma(\varepsilon) = \operatorname*{ess\,sup}_{\zeta(\cdot)} \frac{\tilde{P}_\varepsilon \sum_{m=\varepsilon}^{\infty} \alpha^{\zeta(m)} Z(m)}{\tilde{P}_\varepsilon \sum_{m=\varepsilon}^{\infty} \alpha^{\zeta(m)}} = \frac{\tilde{P}_\varepsilon \sum_{m=\varepsilon}^{\tau_\varepsilon[\Gamma(\varepsilon)]-1} \alpha^m Z(m)}{\tilde{P}_\varepsilon \sum_{m=\varepsilon}^{\tau_\varepsilon[\Gamma(\varepsilon)]-1} \alpha^m}.$$

*Here $\zeta(\cdot)$ stands for random sequences with the properties that $\zeta(n) = \infty$ is allowed,*

$$\zeta(n) \in \mathcal{F}(n) \vee \mathcal{G} \quad \textit{and} \quad \zeta(n) - n \textit{ is nondecreasing}, \qquad n \in \mathcal{N}; \tag{25}$$

*$\tilde{P}_\varepsilon$ denotes conditional expectation with respect to the pre-$\varepsilon$ $\sigma$-field $\{B\colon B \cap \{\varepsilon = n\} \in \mathcal{F}(n) \vee \mathcal{G},\ \forall n \in \mathcal{N}\}$; finally*

$$\tau_\varepsilon[\Gamma(\varepsilon)] = \inf\bigl\{m \geq \varepsilon + 1\colon \Gamma(m) \leq \Gamma(\varepsilon)\bigr\}.$$

PROOF. Augmentation and randomization are straightforward to accommodate. The $\zeta$-discounting amounts to the following randomized version of (20) (see, e.g., (7.5) in [12]):

$$v_n(\gamma) = \operatorname*{ess\,sup}_{\zeta(\cdot)} \tilde{P}_n \sum_{m=n}^{\infty} \alpha^{\zeta(m)}[Z(m) - \gamma]. \qquad \square$$

3.7. *Excursions.* For later use, we decompose the evolution of the index sequence into excursions from its lower envelope. To this end, introduce a sequence of stopping times

$$\varepsilon(0) = 0; \qquad \varepsilon(l+1) = \inf\bigl\{m \geq \varepsilon(l) + 1\colon \Gamma(m) \leq \Gamma[\varepsilon(l)]\bigr\}, \quad l \in \mathscr{N}.$$

Then $\{\varepsilon(l)\colon l \in \mathscr{N}\} = \{n \in \mathscr{N}\colon \Gamma(n) = \underline{\Gamma}(n)\}$, and each $\varepsilon(l)$ is the start of an excursion interval out of the latter set. Write these intervals as $\mathscr{I}(l) = [\varepsilon(l),\ \varepsilon(l+1))$, $l \in \mathscr{N}$ and observe the relations

$$\begin{aligned} \Gamma[\varepsilon(l)] = \underline{\Gamma}[\varepsilon(l)] = \underline{\Gamma}[m], \qquad m \in \mathscr{I}(l); \\ \tau_{\varepsilon(l)}[\Gamma(\varepsilon(l))] = \varepsilon(l+1). \end{aligned} \tag{26}$$

3.8. *Proof of* (16) *and* (17). We now add again a superscript $k$ to indicate an affiliation with arm $k$. Denote by $\tilde{P}_n^k$ the conditional expectation with respect to

$$\tilde{\mathscr{F}}^k(n) = \mathscr{F}^k(n) \bigvee_{j \neq k} \mathscr{F}^j(\infty), \qquad n \in \mathscr{N},$$

which is an augmentation of $\mathscr{F}^k$ by an independent $\sigma$-field. Fix a strategy $T$ and let

$$\zeta^k(n) = \inf\bigl\{t\colon T_k(t+1) > n\bigr\}, \qquad n \in \mathscr{N}.$$

Under $T$, the $n$th pull of arm $k$ is pull number $\zeta^k(n)$ of the bandit. Starting with (16), the value of $T$ has the representations

$$\begin{aligned} v(T) &= P \sum_{k=1}^{d} \sum_{t=0}^{\infty} \alpha^t Z^k[T^k(t)][T^k(t+1) - T^k(t)] \\ &= P \sum_{k=1}^{d} \sum_{n=0}^{\infty} \alpha^{\zeta^k(n)} Z^k(n) \\ &= P \sum_{k=1}^{d} \sum_{l=0}^{\infty} \tilde{P}^k_{\varepsilon^k(l)} \sum_{m \in \mathscr{I}^k(l)} \alpha^{\zeta^k(m)} Z^k(m). \end{aligned}$$

The sequence $\zeta$, given by $\zeta(m) = \zeta^k(m)$, $m < \varepsilon^k(l+1)$, and $\zeta(m) = \infty$ otherwise, has the properties in (25). It follows from Proposition 3 that

$$\frac{\tilde{P}^k_{\varepsilon^k(l)} \sum_{m \in \mathscr{I}^k(l)} \alpha^{\zeta^k(m)} Z^k(m)}{\tilde{P}^k_{\varepsilon^k(l)} \sum_{m \in \mathscr{I}^k(l)} \alpha^{\zeta^k(m)}} \leq \Gamma^k[\varepsilon^k(l)], \tag{27}$$

for all $k$ and $l$. The observation (26) now yields

$$\begin{aligned} v(T) &\le P \sum_{k=1}^{d} \sum_{l=0}^{\infty} \Gamma^k[\varepsilon^k(l)] \tilde{P}^k_{\varepsilon^k(l)} \sum_{m \in \mathscr{I}^k(l)} \alpha^{\zeta^k(m)} \\ &= P \sum_{k=1}^{d} \sum_{l=0}^{\infty} \tilde{P}^k_{\varepsilon^k(l)} \sum_{m \in \mathscr{I}^k(l)} \alpha^{\zeta^k(m)} \underline{\Gamma}^k(m) \\ &= P \sum_{k=1}^{d} \sum_{t=0}^{\infty} \alpha^t \underline{\Gamma}^k[T^k(t)][T^k(t+1) - T^k(t)], \end{aligned}$$

which is precisely (16).

As for (17), it is a consequence of the fact that any index strategy $\hat{T}$ satisfies (27) with an equality. Indeed, under $\hat{T}$, pull $\zeta[\varepsilon^k(l)]$ of the bandit is the $\varepsilon^k(l)$th pull of arm $k$. By (13), $\hat{T}$ stays with arm $k$ also for the next $\varepsilon^k(l+1) - \varepsilon^k(l) - 1$ pulls or, formally,

$$\zeta^k(m) = \zeta[\varepsilon^k(l)] - \varepsilon^k(l) + m, \qquad m \in \mathscr{I}^k(l).$$

Substituting these relations into (27), cancelling out $\alpha^{\zeta[\varepsilon^k(l)] - \varepsilon^k(l)} \in \tilde{\mathscr{F}}^k[\varepsilon^k(l)]$, and recalling that $\varepsilon^k(l+1)$ attains the ess sup in Proposition 3, applied with $\varepsilon = \varepsilon^k(l)$, establishes (17). □

3.9. *The fundamental identity.* The formula for the value in Theorem 1D, when specialized to a single arm, yields

$$P \sum_{n=0}^{\infty} \alpha^n Z(n) = P \sum_{n=0}^{\infty} \alpha^n \underline{\Gamma}(n). \tag{28}$$

Viewing this identity along excursion intervals amounts to sweeping out of the discounted rewards during those intervals to the beginning of the excursions.

**4. The index process—continuous time.** We are now going to extend the approach of the previous section to continuous time. In this section, we treat the index process of a single arm in analogy to Sections 3.6–3.7. The required excursion theory is conceptually similar but technically demanding. In particular, formula (33) below, which extends formula (24), is an excursion-representation of the index. Both (24) and (33) are building blocks for the proof of optimality via excursion theory (in Section 3.8 for the discrete model, and in Section 5 for the continuous one).

Fix an arm $(Z, \mathscr{F})$, a time $t \in \mathscr{T}$ and a nonnegative scalar $\gamma$. (Again, the identifier $k$ will be suppressed.) Associate with the arm a value $v_t(\gamma)$, for playing it optimally after time $t$ until stopping, while paying at a rate of $\gamma$ for participating in the game. Formally,

$$v_t(\gamma) = \operatorname*{ess\,sup}_{\tau > t} P_t \int_t^{\tau} e^{-\beta u} (Z(u) - \gamma)\, du, \tag{29}$$

where $P_t$ is the conditional expectation with respect to $\mathscr{F}_t$ and $\tau$ is a stopping time with respect to $\mathscr{F}$. As a function of $\gamma$, $v_t(\cdot)$ is nonincreasing, convex and finite hence continuous; it is also nonnegative with $v_t(\gamma) = 0$ for all $\gamma > P_t \int_0^\infty e^{-\beta u}|Z(u)|\,du$. (Note that $P_t \int_0^\infty e^{-\beta u}|Z(u)|\,du$ is a.s. finite by integrability.) One concludes that there exists a minimal $\hat{\gamma}(t)$ such that $v_t(\gamma) = 0$, for all $\gamma \geq \hat{\gamma}(t)$. Denote this $\hat{\gamma}(t)$ by

$$\Gamma(t) = \inf\{\gamma\colon v_t(\gamma) = 0\}, \tag{30}$$

and call $\Gamma = \{\Gamma(t)\colon t \geq 0\}$ the *index process* associated with $(Z, \mathscr{F})$. The equivalence of (5) and (30) is well known (see, e.g., [6], Proposition 3.4). Our proofs require the following generalization of (5), in analogy to Proposition 3.

PROPOSITION 4. *Let $\mathscr{G}$ be a $\sigma$-field independent of the filtration $\mathscr{F}$. Then, for every stopping time $\varepsilon$ with respect to $\mathscr{F}(\cdot) \vee \mathscr{G}$,*

$$\Gamma(\varepsilon) = \operatorname*{ess\,sup}_{\zeta(\cdot)} \frac{\tilde{P}_\varepsilon \int_\varepsilon^\infty e^{-\beta\zeta(u)} Z(u)\,du}{\tilde{P}_\varepsilon \int_\varepsilon^\infty e^{-\beta\zeta(u)}\,du}.$$

*Here $\zeta(\cdot)$ stands for any optional process with respect to $\mathscr{F}(\cdot) \vee \mathscr{G}$ such that $\zeta(t) = \infty$ is allowed and $\zeta(t) - t$, $t \geq 0$, is nondecreasing; $\tilde{P}_\varepsilon$ denotes conditional expectation with respect to the pre-$\varepsilon$ $\sigma$-field $\{B \in \mathscr{F}(\cdot) \vee \mathscr{G}\colon B \cap \{\varepsilon \leq t\} \in \mathscr{F}(t) \vee \mathscr{G},\ \forall t \geq 0\}$.*

PROOF. Again, augmentation and randomization are straightforward. The $\zeta$-discounting is a consequence of the following lemma with $X = Z - \gamma$.

LEMMA 5. *Let $(X, \mathscr{F})$ be a progressively measurable process satisfying the integrability and regularity conditions. If*

$$\sup_\tau P \int_0^\tau e^{-\beta t} X(t)\,dt = 0,$$

*where $\tau$ runs over all $\mathscr{F}$ stopping times, then for every optional process $q$ which is nonnegative decreasing and bounded by* 1,

$$P \int_0^\infty e^{-\beta t} q(t) X(t)\,dt \leq 0.$$

PROOF. For $q$ continuous and decreasing, define the stopping times

$$\tau_l^n = \inf\left\{t\colon q(t) = 1 - \frac{l}{n}\right\}.$$

Then

$$\begin{aligned} P \int_0^\infty e^{-\beta t} q(t) X(t)\,dt &= \lim_{n\to\infty} P \sum_{l=0}^{n-1} \left(1 - \frac{l}{n}\right) \int_{[\tau_l^n,\, \tau_{l+1}^n)} e^{-\beta t} X(t)\,dt \\ &= \lim_{n\to\infty} \sum_{l=1}^{n} \frac{1}{n} P \int_0^{\tau_l^n} e^{-\beta t} X(t)\,dt \leq 0. \end{aligned}$$

The rest now follows from approximating $q$ by a sequence $\{q^n\}$ of continuous adapted functions that converge to $q$ a.e. Lebesgue. Indeed, let

$$q^n(t) = q(0) \quad \text{for } 0 \le t \le \frac{1}{n},$$
$$q^n(t) = q^n\left(\frac{k}{n}\right) + \left(q\left(\frac{k}{n}\right) - q^n\left(\frac{k}{n}\right)\right)\left(t - \frac{k}{n}\right)n, \qquad \frac{k}{n} < t \le \frac{k+1}{n}.$$

Then $q^n$ is continuous, optional and satisfies

$$q(t) \le q^n(t) \le q(t - 2/n), \qquad t \ge 0.$$

It follows that

$$P\int_0^\infty e^{-\beta t}|q^n(t) - q(t)||X(t)|\,dt \le P\int_0^\infty e^{-\beta t}(q(t - 2/n) - q(t))|X(t)|\,dt.$$

We now recall that $X$ satisfies the integrability condition, $q$ is decreasing, and thus $q(t - 2/n) - q(t))$ converges a.e. to 0. Since $q$ is bounded by 1, the Dominated Convergence Theorem implies that the last integral converges to 0, and we are done. □

Our objective now is to represent $\Gamma$ in terms of excursions from the set $M$ of its minima. This is carried out via a successive generalization of (28), culminating in (33). (In [11], the analogue of (33) provided a tool for explicit computations of $\Gamma$, as well as to the proof of Theorem 1 for Lévy processes.) To this end, define the filtration $\tilde{\mathscr{F}}^k$ by $\tilde{\mathscr{F}}^k(t) = \mathscr{F}^k(t) \vee \bigvee_{j \ne k} \mathscr{F}^j(\infty)$, then let $\tilde{P}^k_t$ ($\tilde{P}^k$) be the conditional expectation with respect to $\tilde{\mathscr{F}}^k(t)$ ($\tilde{\mathscr{F}}^k(0)$).

Continuing to suppress the identifier $k$, we first quote the following extension of (28) to continuous time:

$$\tilde{P}\int_0^\infty e^{-\beta t} Z(t)\,dt = \tilde{P}\int_0^\infty e^{-\beta t}\underline{\Gamma}(t)\,dt. \tag{31}$$

It can be obtained either by discrete approximation [12] or by the beautiful change-of-variables argument of [6]. The same argument leads to the following.

PROPOSITION 6. *Let $q$ be a positive process adapted to $\tilde{\mathscr{F}}$. Then*

$$\tilde{P}\int_0^\infty e^{-\beta t} q(t)Z(t)\,dt = \tilde{P}\int_0^\infty e^{-\beta t} q(t)\underline{\Gamma}^q(t)\,dt,$$

*where $\Gamma^q$ is the lower envelope of*

$$\Gamma^q(t) = \operatorname*{ess\,sup}_{\tau > t} \frac{\tilde{P}_t \int_t^\tau e^{-\beta u} q(u)Z(u)\,du}{\tilde{P}_t \int_t^\tau e^{-\beta u} q(u)\,du}, \qquad t \ge 0.$$

With Lemma 5 at hand, we deduce the following.

COROLLARY 7. *If $q$ is nonnegative, optional decreasing and bounded by* 1, *then $\Gamma^q(t) \le \Gamma(t)$ a.s. for all $t \ge 0$.*

The set $M = \text{closure}\{t\colon \Gamma(t) = \underline{\Gamma}(t)\}$ is a closed random set as defined in [1]. We shall use the terminology and notation of [1] and [10] with the filtration $\tilde{\mathscr{F}}$. Define

$$\begin{aligned} D_t &= \inf\{u > t\colon u \in M\}, \\ R_t &= D_t - t, \\ g_t &= \sup\{u < t\colon u \in M\}, \\ G &= \{t > 0\colon R_{t-} = 0,\ R_t > 0\}. \end{aligned}$$

The following extensions of (31) enable us to identify the index in terms of the $(\tilde{\mathscr{F}}(D_t))$ predictable exit system from $M$.

PROPOSITION 8. *For any $(\tilde{\mathscr{F}}(D_t))$ predictable $H$,*

$$\tilde{P}\int_0^\infty e^{-\beta t} H(g_t) Z(t)\, dt = \tilde{P}\int_0^\infty e^{-\beta t} H(g_t)\underline{\Gamma}(t)\, dt.$$

PROOF. By the Monotone Class Theorem, it is enough to prove the result for the generators of the predictable $\sigma$-field, that is, for $H(t) = 1_{[|0,\,\sigma|]}(t)$, where $\sigma$ is an $(\tilde{\mathscr{F}}(D_t))$ stopping time and $H_t = 1_{[|0_A|]}$ where $A \in \mathscr{F}_{D_0}$, and $0_A$ is the stopping time that is equal to 0 on $A$ and $\infty$ on $A^c$. For $H_t = 1_{[|0,\,\sigma|]}(t)$, we note that $D_\sigma$ is an $\tilde{\mathscr{F}}$ stopping time with values in $M$, and

$$\{0 \le g_t \le \sigma\} = \{0 \le t \le D_\sigma\}.$$

The result now follows from (31) and from the fact that

$$\begin{aligned} \tilde{P}\textstyle\int_{D_\sigma}^\infty e^{-\beta t} Z(t)\, dt &= \tilde{P}\tilde{P}_{D_\sigma}\displaystyle\int_{D_\sigma}^\infty e^{-\beta t} Z_t\, dt \\ &= \tilde{P}\tilde{P}_{D_\sigma}\displaystyle\int_{D_\sigma}^\infty e^{-\beta t}\underline{\Gamma}(D_\sigma, t)\, dt \\ &= \tilde{P}\displaystyle\int_{D_\sigma}^\infty e^{-\beta t}\underline{\Gamma}(D_\sigma, t)\, dt \\ &= \tilde{P}\displaystyle\int_{D_\sigma}^\infty e^{-\beta t}\underline{\Gamma}(t)\, dt, \end{aligned} \tag{32}$$

where $\Gamma(u, t)$ is defined as $\Gamma(t)$ with $\mathscr{F}_0$ replaced by $\mathscr{F}(u)$, $\tau > 0$ replaced by $\tau > u$, $\underline{\Gamma}(u, t) = \inf_{u \le v \le t}\Gamma(u, v)$ and $\tilde{P}_{D_\sigma}$ is the conditional expectation with respect to the pre-$D_\sigma$ $\sigma$-field $\tilde{\mathscr{F}}(D_\sigma)$. The last equality is a consequence of $D_\sigma \in M$ a.s. on $\{D_\sigma < \infty\}$, and $\underline{\Gamma}(D_\sigma, t) = \underline{\Gamma}(t)$ a.s. on $\{D_\sigma \le t\}$. Subtracting (32) from (31) yields the result for $H = 1_{[|0,\,\sigma|]}$. For $H_t = 1_{[|0_A|]}(t)$ with $A \in \mathscr{F}_{D_0}$, we note that

$$1_{[|0_A|]}(g_t) = 1_A 1_{\{0 \le t \le D_0\}}.$$

The rest of the proof is identical to the above argument and is therefore omitted. □

The next result deals with the downwards jumps of $\underline{\Gamma}$. In nature, this part is close to the discrete case because it is composed of a countable union of graphs of stopping times.

PROPOSITION 9. *Let $H$ be an $(\tilde{\mathscr{F}}(D_t))$ predictable process. Then*

$$\tilde{P}\int_0^\infty e^{-\beta t}H(g_t)1_{\{\Gamma(g_t)<\underline{\Gamma}(g_{t^-})\}}Z_t\,dt = \tilde{P}\int_0^\infty e^{-\beta t}H(g_t)1_{\{\Gamma(g_t)<\underline{\Gamma}(g_{t^-})\}}\Gamma(g_t)\,dt,$$

*where we note that $\Gamma(g_t)$ entering the integrand on the right instead of $Z(t)$ on the left, may be replaced by $\underline{\Gamma}(t)$.*

REMARK. Note that $\{t\colon \Gamma(g_t) < \underline{\Gamma}(g_{t-})\}$ may contain complete excursion intervals from $M$, and thus we cannot assume that this set has countable sections.

PROOF. As before, it is enough to prove the result for $H = 1_{[|0,\,T|]}$ where $T$ is an $(\tilde{\mathscr{F}}(D_t))$ stopping time. Then, as in Proposition 8, it is enough to show

$$\tilde{P}\int_0^\infty e^{-\beta t}1_{\{\Gamma(g_t)<\underline{\Gamma}(g_{t^-})\}}Z(t)\,dt = \tilde{P}\int_0^\infty e^{-\beta t}1_{\{\Gamma(g_t)<\underline{\Gamma}(g_{t^-})\}}\Gamma(g_t)\,dt.$$

This result follows from the fact that $\{t\colon \Gamma(t) < \underline{\Gamma}(t-)\}$ is contained in $M$ and is covered by a countable sequence of $\tilde{\mathscr{F}}$ stopping times $\{\tau_n\}$. Let $\tau_n$ be such a stopping time; set

$$\tilde{\tau}_n = \begin{cases} \tau_n, & \text{if } R_{\tau_n} > 0, \\ \infty, & \text{if } R_{\tau_n} = 0. \end{cases}$$

Then

$$\Gamma(\tilde{\tau}_n) = \frac{\tilde{P}_{\tilde{\tau}_n}\int_{\tilde{\tau}_n}^{D_{\tilde{\tau}_n}} e^{-\beta t}Z(t)\,dt}{\tilde{P}_{\tilde{\tau}_n}\int_{\tilde{\tau}_n}^{D_{\tilde{\tau}_n}} e^{-\beta t}\,dt}.$$

Thus, as in the discrete case,

$$\begin{aligned}
\tilde{P}\int_0^\infty e^{-\beta t}1_{\{\Gamma(g_t)<\underline{\Gamma}(g_{t^-})\}}Z(t)\,dt &= \tilde{P}\sum_{n=0}^{\infty}1_{\{\tilde{\tau}_n<\infty\}}P_{\tilde{\tau}_n}\int_{\tilde{\tau}_n}^{D_{\tilde{\tau}_n}} e^{-\beta t}Z(t)\,dt \\
&= \tilde{P}\sum_{n=0}^{\infty}1_{\{\tilde{\tau}_n<\infty\}}\Gamma(\tilde{\tau}_n)\tilde{P}_{\tilde{\tau}_n}\int_{\tilde{\tau}_n}^{D_{\tilde{\tau}_n}} e^{-\beta t}\,dt \\
&= \tilde{P}\sum_{n=0}^{\infty}1_{\{\tilde{\tau}_n<\infty\}}\int_{\tilde{\tau}_n}^{D_{\tilde{\tau}_n}}\Gamma(\tilde{\tau}_n)e^{-\beta t}\,dt \\
&= \tilde{P}\int_0^\infty e^{-\beta t}1_{\{\Gamma(g_t)<\underline{\Gamma}(g_{t^-})\}}\Gamma(g_t)\,dt \\
&= \tilde{P}\int_0^\infty e^{-\beta t}1_{\{\Gamma(g_t)<\underline{\Gamma}(g_{t^-})\}}\underline{\Gamma}(t)\,dt. \qquad \square
\end{aligned}$$

We now show that positive jumps of $\Gamma$ from its lower envelope are totally inaccessible.

PROPOSITION 10. *Let* $N = \{t \in G: \Gamma(t) > \underline{\Gamma}(t-)\}$. *Then, for every* $(\tilde{\mathscr{F}}(D_t))$ *predictable stopping time* $\tau$, $\tilde{P}\{\tau \in N\} = 0$.

PROOF. Let $\tau$ be a $(\mathscr{F}(D_t))$ predictable stopping time that, with a positive probability, is contained in $N$. Let $(\tau_n)$ be its announcing sequence and $B = \bigcap_n \{D_{\tau_n} < \tau\}$. Note that since $N \subset M$, and $\tau_n < \tau$ a.s. on $\{\tau < \infty\}$,

$$\begin{aligned}\{\tau \in N\} \cap B^c &\subset \bigcup_n \{\tau \in N,\ D_{\tau_n} = \tau\} \\ &\subset \{\tau \text{ is isolated in } M\} \\ &\subset \{\Gamma(\tau) \le \underline{\Gamma}(\tau-)\} \subset \{\tau \in N\}^c.\end{aligned}$$

Thus, $\{\tau \in N\} \cap B^c$ is empty. Denote by $\tau_B$ the stopping time that is equal to $\tau$ on $B$ and infinity otherwise. By the above argument, we may replace $\tau$ by the predictable stopping time $\tau_B$ and assume that $\{D_{\tau_n} < \tau_B\}$ for all $n$ on $\{\tau_B < \infty\}$, and that $\Gamma(D_{\tau_n})$ decreases to a random variable $\tilde{\Gamma}$ satisfying $P\{\tilde{\Gamma} < \Gamma(\tau_B)\} > 0$. Note that $\tilde{\Gamma} \in \bigvee_n \tilde{\mathscr{F}}(D_{\tau_n}) \subset \tilde{\mathscr{F}}_{\tau_B}$, and there exists a stopping time $\tau^+ > \tau_B$ on $\{\tau_B < \infty\}$ and a number $b > 0$ so that

$$\tilde{P}_{\tau_B} \int_{\tau_B}^{\tau^+} e^{-\beta u}(Z(u) - \tilde{\Gamma})\, du \ge b$$

on a set $A \in \tilde{\mathscr{F}}_{\tau_B}$ of a positive probability, say $p$.

Let $\varepsilon > 0$ and define

$$A_n = \left\{\tilde{P}_{D_{\tau_n}} \int_{D_{\tau_n}}^{\tau_B} e^{-\beta u}|Z(u)|\, du < b \cdot \varepsilon,\ \Gamma(D(\tau_n))\tilde{P}_{D_{\tau_n}} \int_{D_{\tau_n}}^{\tau_B} e^{-\beta u}\, du < b \cdot \varepsilon\right\},$$

$$B_n = \left\{\tilde{P}_{D_{\tau_n}} \int_{\tau_B}^{\tau^+} e^{-\beta u}(\Gamma(D_{\tau_n}) - \tilde{\Gamma}) < b \cdot \varepsilon\right\}.$$

Then

$$\begin{aligned} A_n &\in \tilde{\mathscr{F}}(D_{\tau_n}), \qquad P(A_n^c \cap \{\tau_B < \infty\}) \to 0, \\ B_n &\in \tilde{\mathscr{F}}(D_{\tau_n}), \qquad P(B_n^c \cap \{\tau_B < \infty\}) \to 0. \end{aligned}$$

Set

$$\tilde{\tau} = \begin{cases} \tau_B, & \text{on } A^c, \\ \tau^+, & \text{on } A. \end{cases}$$

On $A_n \cap B_n$,

$$\begin{aligned} 0 &\ge \tilde{P}_{D_{\tau_n}} \int_{D_{\tau_n}}^{\tilde{\tau}} e^{-\beta u}(Z(u) - \Gamma(D_{\tau_n}))\, du \\ &= \tilde{P}_{D_{\tau_n}} \int_{D_{\tau_n}}^{\tau_B} e^{-\beta u}(Z(u) - \Gamma(D_{\tau_n}))\, du \\ &\quad + \tilde{P}_{D_{\tau_n}} \left(1_A \int_{\tau_B}^{\tau^+} e^{-\beta u}(Z(u) - \Gamma(D_{\tau_n}))\, du\right) \end{aligned}$$

$$\begin{aligned}
&\geq -2b\varepsilon + \tilde{P}_{D_{\tau_n}}\left(1_A \int_{\tau_B}^{\tau^+} e^{-\beta u}(Z(u) - \tilde{\Gamma})\, du\right) \\
&\quad - \tilde{P}_{D_{\tau_n}}\left(1_A \int_{\tau_B}^{\tau^+} e^{-\beta u}(\Gamma(D_{\tau_n}) - \tilde{\Gamma})\, du\right) \\
&\geq \tilde{P}_{D_{\tau_n}}\left(1_A \int_{\tau_B}^{\tau^+} e^{-\beta u}(Z(u) - \tilde{\Gamma})\, du\right) - 3b\varepsilon \\
&\geq b\tilde{P}_{D_{\tau_n}}(1_A) - 3b\varepsilon.
\end{aligned}$$

Now $H_{\tau_n} = \tilde{P}(1_A|\tilde{\mathscr{F}}(D_{\tau_n}))$ is a bounded martingale that converges to $H_{\tau_{B-}} = \tilde{P}(1_A|\tilde{\mathscr{F}}_{\tau_{B-}})$ and, furthermore,

$$\tilde{P}(H_{\tau_{B-}}) = P(A) = p.$$

Let

$$C_n = \left\{H_{\tau_n} > \frac{p}{4}\right\},$$

$$C = \left\{H_{\tau_{B-}} > \frac{p}{2}\right\}.$$

Then $P(C_n) \geq P(C) > 0$ for $n$ large enough and on $A_n \cap B_n \cap C_n$ (which has a positive probability for $n$ large enough),

$$b\tilde{P}_{D_{\tau_n}}(1_A) - 3b\varepsilon > \frac{b \cdot p}{4} - 3b\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, the above can be made positive, which is a contradiction. □

We are now in a position to establish the representation of the index in terms of $(l, {}^*P)$, the $(\tilde{\mathscr{F}}(D_t))$ predictable exit system from $M$. The process $(l(t))$ is an $(\tilde{\mathscr{F}}(D_t))$ predictable local time at $M$ and ${}^*P$ is a kernel from $(\mathbb{R} \times \Omega, \tilde{\mathscr{P}})$ into $(\Omega, \tilde{\mathscr{F}})$, where $\tilde{\mathscr{P}}$ is the $(\tilde{\mathscr{F}}(D_t))$ predictable $\sigma$-field. For the construction of exit systems from closed sets, we refer the reader to [1], for $(\tilde{\mathscr{F}}(D_t))$ predictable exit systems in the Hunt case to [10], and for the general theory of kernel constructions to [16], Appendix 3. Roughly, the local time $l$ is the $(\tilde{\mathscr{F}}(D_t))$ dual predictable projection of the random measure

$$\Pi(\omega, dt) = \sum_{u \in G} \delta_u(dt)(1 - \exp(-R_u)) + 1_M(t)\lambda(dt),$$

where $\delta_u$ is the Dirac measure with mass at $u$ and $\lambda$ is the Lebesgue measure. The kernel ${}^*P$ is obtained by considering the $(\tilde{\mathscr{F}}(D_t))$ dual predictable projection $l^f$ of

$$\Pi^f(\omega, dt) = \sum_{u \in G} \delta_u(dt)(1 - \exp(-R_u))f(\theta_t\omega),$$

where $(\theta_t)$ are the usual shift operators on $\Omega$ [for which $Z(u, \theta_t\omega) = Z(u + t, \omega)$] and $f$ is an $(\Omega, \tilde{\mathscr{F}})$ measurable function. Here $l^f$ is absolutely continuous

with respect to $l$, and its Radon–Nikodym derivative $({}^*Pf)(\omega)$ is an $(\tilde{\mathscr{F}}(D_t))$ predictable process. The construction of ${}^*P$ now follows from A3.3 of [16], for example. Indeed, ${}^*P$ is characterized as the unique kernel that satisfies the following: for any $(\tilde{\mathscr{F}}(D_t))$ predictable $U$ and any $(\Omega, \tilde{\mathscr{F}})$ positive measurable $f$:

1. $\tilde{P}\sum_{u\in G} e^{-u}U(u)f(\theta_u) = \tilde{P}\int_0^\infty e^{-u}U(u)\,{}^*P_u(f)\,dl(u)$;
2. ${}^*P_u(R=0)=0$, $l$-a.e., almost surely;
3. let $m^Z(t)$ be the $(\tilde{\mathscr{F}}(D_t))$ predictable Radon–Nikodym derivative of $1_M(t)Z(t)\lambda(dt)$ with respect to $l$, and let $m = m^1$. Then
$$ {}^*P_u(1-e^{-R}) + m(u) = 1, \qquad l\text{-a.e., almost surely.}$$

THEOREM 11. *Let $(l, {}^*P)$ be the $(\tilde{\mathscr{F}}(D_t))$ predictable exit system and $m^Z$ as defined above. Introduce*

$$\tilde{\Gamma}(t) = \frac{e^{-\beta t}m^Z(t) + {}^*P_t(1_{\Gamma(t)\geq \underline{\Gamma}(t-)}\int_t^{D_t} e^{-\beta u}Z(u)\,du)}{e^{-\beta t}m(t) + {}^*P_t(1_{\Gamma(t)\geq \underline{\Gamma}(t-)}\int_t^{D_t} e^{-\beta u}\,du)}, \qquad t\geq 0, \tag{33}$$

*(with the convention that $0/0=0$). Then, for a.e. $t$, with respect to the Lebesgue measure,*

$$1_{\{\Gamma(g_t)\geq\underline{\Gamma}(g_t-)\}}\tilde{\Gamma}(g_t) = 1_{\{\Gamma(g_t)\geq\underline{\Gamma}(g_t-)\}}\underline{\Gamma}(g_t-).$$

PROOF. It follows from Propositions 8 and 9 that, for any $(\tilde{\mathscr{F}}(D_t))$ predictable $H$,

$$\begin{aligned}
&\tilde{P}\int_0^\infty e^{-\beta t}1_{\{\Gamma(g_t)\geq\underline{\Gamma}(g_t-)\}}H(g_t)Z(t)\,dt\\
&\quad= \tilde{P}\int_0^\infty e^{-\beta t}1_{\{\Gamma(g_t)\geq\underline{\Gamma}(g_t-)\}}H(g_t)\underline{\Gamma}(t)\,dt\\
&\quad= \tilde{P}\int_0^\infty e^{-\beta t}1_{\{\Gamma(g_t)\geq\underline{\Gamma}(g_t-)\}}H(g_t)\underline{\Gamma}(g_t-)\,dt\\
&\quad= \tilde{P}\sum_{u\in G}1_{\{\Gamma(u)\geq\underline{\Gamma}(u-)\}}H(u)\underline{\Gamma}(u-)\int_u^{D_u} e^{-\beta t}\,dt\\
&\qquad+ \tilde{P}\int_0^\infty e^{-\beta t}1_M(t)H(t)\underline{\Gamma}(t-)\,dt.
\end{aligned} \tag{34}$$

Applying now the exit system results first to the right-hand side of (34), and then to the left-hand side, we get

$$\begin{aligned}
&\tilde{P}\int_0^\infty H(u)\underline{\Gamma}(u-)\left({}^*P_u\left(1_{\{\Gamma(u)>\underline{\Gamma}(u-)\}}\int_u^{D_u} e^{-\beta t}\,dt\right) + e^{-\beta u}m(u)\right)dl(u)\\
&\quad= \tilde{P}\int_0^\infty H(u)\left({}^*P_u\left(1_{\{\Gamma(u)>\underline{\Gamma}(u-)\}}\int_u^{D_u} e^{-\beta t}Z(t)\,dt\right) + e^{-\beta u}m^Z(u)\right)dl(u)\\
&\quad= \tilde{P}\int_0^\infty H(u)\tilde{\Gamma}(u)\left({}^*P_u\left(1_{\{\Gamma(u)>\underline{\Gamma}(u-)\}}\int_u^{D_u} e^{-\beta t}\right) + e^{-\beta u}m(u)\right)dl(u)\\
&\quad= \tilde{P}\int_0^\infty e^{-\beta t}H(g_t)1_{\{\Gamma(g_t)\geq\underline{\Gamma}(g_t-)\}}\tilde{\Gamma}(g_t)\,dt,
\end{aligned} \tag{35}$$

where for the last equality we have applied the exit system identity following the steps from (34) to (35), this time backward.

The result follows now from the fact that both $(\underline{\Gamma}(u-))$ and $(\tilde{\Gamma}(u))$ are $(\tilde{\mathscr{F}}(D_t))$ predictable processes and the following identity that summarizes the first equality of (34) and the first and last equalities of (35):

$$\begin{aligned}\tilde{P}\int_0^\infty e^{-\beta t}H(g_t)\underline{\Gamma}(g_t-)1_{\{\Gamma(g_t)\geq\underline{\Gamma}(g_t-)\}}\,dt&\\ =\tilde{P}\int_0^\infty e^{-\beta t}H(g_t)1_{\{\Gamma(g_t)\geq\underline{\Gamma}(g_t-)\}}Z(t)\,dt&\\ =\tilde{P}\int_0^\infty e^{-\beta t}H(g_t)1_{\{\Gamma(g_t)\geq\underline{\Gamma}(g_t-)\}}\tilde{\Gamma}(g_t)\,dt,&\end{aligned}$$

for all $(\tilde{\mathscr{F}}(D_t))$ predictable $H$. □

**5. Proof of Theorem 1.** Let $T$ be any strategy. As in the discrete case, let

$$\tilde{\mathscr{F}}^k(t)=\mathscr{F}^k\vee\bigvee_{j\neq k}\mathscr{F}^j(\infty)$$

and $\zeta^k(t)=\inf\{u\colon T^k(t)>u\}$. Let $\hat{T}$ be an index-priority strategy. Since the sets $N^k$, $k\in D$, defined in Proposition 10 do not contain predictable stopping times, it follows that it is not possible for two arms pulled simultaneously by $\hat{T}$ to start an excursion from their respective $M$'s by upward jumps of at least one of their indices (necessarily from the same value of their respective minima). To see this, suppose that arms $j$ and $k$ are pulled simultaneously by $\hat{T}$ at time $t$; then this can occur only if $\hat{T}^j(t)$ and $\hat{T}^k(t)$ are in $M^j$ and $M^k$, respectively, and both $\underline{\Gamma}^j$ and $\underline{\Gamma}^k$ are strictly decreasing at $\hat{T}^j(t)$ and $\hat{T}^k(t)$, respectively. If $x$ is a point from where $\Gamma^j$ starts an excursion from its minimum, then for $\Gamma^k$ to jump upwards from its minimum $x$, $\mathscr{P}^k=\{\underline{\Gamma}^k(t-)=x,\ \underline{\Gamma}^k(t-\varepsilon)>x$ for all $\varepsilon>0\}$ must intersect $N^k$. Now $\mathscr{P}^k$ is a predictable set whose $\omega$-sections consist of a single time point and is, therefore, contained in the graph of a predictable stopping time which cannot intersect $N^k$. Thus, if the indices of two arms start excursions from a level set $x$ in their respective minima at time $t$, that time has to be a point of continuity for both $\Gamma^j[\hat{T}^j(t)]$ and $\Gamma^k[\hat{T}^k(t)]$, and then the priority rule will determine which of the two arms will be pulled first and continue to be pulled until the end of the excursion from its $M$. Thus, with $I_j(u)=D^j_{\hat{T}^j(\zeta^i(u))}-\hat{T}^j(\zeta^i(u))$, we have

$$\zeta^i(g^i_t)=\zeta^i(g^i_t-)+\sum_{j<i}I_j(g^i_t-).$$

Let

$$H^i(t)=\zeta^i(t-)+\sum_{j<i}I_j(t-).$$

Then $H^i$ is $(\tilde{\mathscr{F}}^i(D^i(t)))$ predictable. (As will be seen shortly, this predictability plays an important role in our proof. Indeed, any index strategy with this predictability property is optimal.)

With the above at hand, for any strategy $T$,

$$\begin{aligned} P\int_0^\infty e^{-\beta t} Z^i[T^i(t)]dT^i(t) &= P\int_0^\infty \exp(-\beta\zeta^i(t))Z^i(t)\,dt \\ &= P\int_0^\infty e^{-\beta t} q_i(t) Z^i(t)\,dt \\ &= P\int_0^\infty e^{-\beta t} q_i(t)\underline{\Gamma}^{i,q}(t)\,dt \\ &\le P\int_0^\infty e^{-\beta t} q_i(t)\underline{\Gamma}^{i}(t)\,dt \\ &= P\int_0^\infty \exp(-\beta\zeta^i(t))\underline{\Gamma}^{i}(t)\,dt, \end{aligned}$$

where $q_i(t) = \exp[-\beta(\zeta_i(t) - t)]$ satisfies the assumption of the remark following Proposition 6 and $\Gamma^q$ is as defined in that proposition.

If $\tilde{T}$ is an index-priority strategy, with its corresponding $\zeta^i$'s,

$$\begin{aligned} &P\int_0^\infty \exp(-\beta\zeta^i(t))Z^i(t)\,dt \\ &\quad = P\int_0^\infty \exp(-\beta\zeta^i(t))1_{M^i}(t)Z^i(t)\,dt \\ &\qquad + P\sum_{u\in G}\exp[-\beta(\zeta^i(u)-u)]1_{\{\Gamma^i(u)\ge\underline{\Gamma}^i(u-)\}}\int_u^{D^i_u} e^{-\beta t}Z^i(t)\,dt \\ &\qquad + P\sum_{u\in G}\exp[-\beta(\zeta^i(u)-u)]1_{\{\Gamma^i(u)<\underline{\Gamma}^i(u-)\}}\int_u^{D^i_u} e^{-\beta t}Z^i(t)\,dt \\ &\quad = P\int_0^\infty \exp\Big[-\beta\Big(\zeta^i(u-)+\sum_{j<i}I^j(u-)-u\Big)\Big] \\ &\qquad\qquad \times\left({}^*P^i_u\left(1_{\{\Gamma^i(u)\ge\underline{\Gamma}^i(u-)\}}\int_u^{D_u} e^{-\beta t}Z^i(t)\,dt\right)+e^{-\beta u}(m^Z)^i(u)\right)dl^i(u) \\ &\qquad + P\sum_{n=0}^\infty \exp[-\beta(\zeta^i(\tilde{\tau}^i_n)-\tilde{\tau}^i_n)]\tilde{P}^i_{\tilde{\tau}^i_n}\int_{\tilde{\tau}^i_n}^{D_{\tilde{\tau}^i_n}} e^{-\beta t}Z^i(t)\,dt, \end{aligned}$$

where $\tilde{\tau}^i_n$ was defined in the proof of Proposition 9. We recall that, for each $\tilde{\tau}^i_n$,

$$\Gamma^i(\tilde{\tau}^i_n) = \frac{\tilde{P}_{\tilde{\tau}^i_n}\int_{\tilde{\tau}^i_n}^{D_{\tilde{\tau}^i_n}} e^{-\beta t}Z^i(t)\,dt}{\tilde{P}_{\tilde{\tau}^i_n}\int_{\tilde{\tau}^i_n}^{D_{\tilde{\tau}^i_n}} e^{-\beta t}\,dt}.$$

Therefore, the above sum is equal to

$$\begin{aligned} &P\int_0^\infty \exp[-\beta(\zeta^i(u)-u)]\tilde{\Gamma}^i(u) \\ &\qquad\times\left(e^{-\beta u}m^i(u)+{}^*P^i_u\left(1_{\{\Gamma^i(u)\ge\underline{\Gamma}^i(u-)\}}\int_u^{D^i_u} e^{-\beta t}\,dt\right)\right)dl^i(u) \\ &\quad + P\sum_{n=0}^\infty \exp[-\beta(\zeta^i(\tilde{\tau}^i_n)-\tilde{\tau}^i_n)]\Gamma^i(\tilde{\tau}^i_n)\tilde{P}^i_{\tilde{\tau}^i_n}\int_{\tilde{\tau}^i_n}^{D_{\tilde{\tau}^i_n}} e^{-\beta t}\,dt \end{aligned}$$

$$
\begin{aligned}
&= P\int_0^\infty \exp(-\beta\zeta^i(u))1_{\{\Gamma^i(g_u^i)\geq \underline{\Gamma}^i(g_u^i-)\}}\underline{\Gamma}^i(g_u^i-)\,du \\
&\quad + P\int_0^\infty \exp(-\beta\zeta^i(u))1_{\{\Gamma^i(g_u^i)< \underline{\Gamma}^i(g_u^i-)\}}\Gamma^i(g_u^i)\,du \\
&= P\int_0^\infty \exp(-\beta\zeta^i(u))\underline{\Gamma}^i(u)\,du \\
&= P\int_0^\infty e^{-\beta t}\underline{\Gamma}^i\big[\hat{T}^i(t)\big]\,d\hat{T}^i(t),
\end{aligned}
$$

where the equality before last follows from Theorem 11. Summing over all arms, we get

$$
P\sum_{i=1}^d \int_0^\infty e^{-\beta t}Z^i[T^i(t)]\,dT^i(t) \leq P\sum_{i=1}^d \int_0^\infty e^{-\beta t}\underline{\Gamma}^i[T^i(t)]\,dT^i(t),
$$

with equality for index strategies. This last expression is smaller than

$$
P\int_0^\infty e^{-\beta t}\bigvee_{j=1}^d \underline{\Gamma}^j\big[\hat{T}^j(t)\big]\,dt,
$$

where $\hat{T}$ is an index strategy, and therefore follows the leader among $\underline{\Gamma}^j$, $j=1,\ldots,d$. Our proof is now complete. □

## REFERENCES


[1] AZÉMA, J. (1985). Sur les fermes aleatoires. *Séminaire de Probabilités XIX. Lecture Notes in Math.* **1123** 397–495. Springer, Berlin.

[2] BERRY, D. A. and FRISTEDT, D. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London.

[3] CAIROLI, R. and DALANG, R. C. (1996). Sequential stochastic optimization. In *Probability and Statistics*. Wiley, New York.

[4] DELLACHERIE, C. and MEYER, P. A. (1978). *Probabilities and Potentials.* North-Holland, Amsterdam.

[5] EL KAROUI, N. and KARATZAS, I. (1993). General Gittins index processes in discrete time. *Proc. Nat. Acad. Sci. U.S.A.* **90** 1232–1236.

[6] EL KAROUI, N. and KARATZAS, I. (1994). Dynamic allocation problems in continuous time. *Ann. Appl. Probab.* **4** 255–286.

[7] EL KAROUI, N. and KARATZAS, I. (1996). Synchronization and optimality for multi-armed bandit problems in continuous time. Unpublished manuscript.

[8] GITTINS, J. C. (1989). *Multi-armed Bandit Allocation Indices*. Wiley, New York.

[9] GITTINS, J. C. and JONES, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (J. Gani et al., eds.) 241–266. North-Holland, Amsterdam.

[10] KASPI, H. and MAISONNEURVE, B. (1984/85). Predictable local times and exit systems. *Séminaire de Probabilités XX. Lecture Notes in Math.* **1204** 95–100. Springer, Berlin.

[11] KASPI, H. and MANDELBAUM, A. (1994). Lévy bandits: multi-armed bandits driven by Lévy processes. *Ann. Appl. Probab.* **5** 541–565.

[12] MANDELBAUM, A. (1986). Discrete multiarmed bandits and multiparameter processes. *Probab. Theory Related Fields* **71** 129–147.

[13] MANDELBAUM, A. (1987). Continuous multi-armed bandits and multi-parameter processes. *Ann. Probab.* **15** 1527–1556.

[14] MANDELBAUM, A. and VANDERBEI, R. J. (1981). Optimal stopping and supermartingales over partially ordered sets. *Z. Wahrsch. Verw. Gebiete* **57** 253–264.
[15] PRESMAN, E. L. and SONIN, I. N. (1990). *Sequential Control with Incomplete Information: The Bayesian Approach to Multi-armed Bandit Problems*. Academic Press, New York.
[16] SHARPE, M. (1988). *General Theory of Markov Processes*. Academic Press, New York.
[17] SNELL, L. (1952). Applications of martingale systems theorems. *Trans. Amer. Math. Soc.* **73** 293–312.
[18] VARAIYA, P., WALRAND, J. and BUYUKKOC, C. (1985). Extensions of the multi-armed bandit problem. The discounted case. *IEEE Trans. Automat. Control* **AC-30** 426–439.
[19] WALSH, J. B. (1981). Optional increasing paths. *Colloque ENST-CNET: Lecture Notes in Math.* **863** 172–201. Springer, Berlin.
[20] WEBER, R. (1992). On the Gittins index for multi-armed bandits. *Ann. Appl. Probab.* **2** 1024–1035.
[21] WHITTLE, P. (1980). Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. Ser. B* **42** 143–149.

DAVIDSON FACULTY OF INDUSTRIAL
ENGINEERING AND MANAGEMENT
TECHNION–ISRAEL INSTITUTE OF TECHNOLOGY
HAIFA 32000
ISRAEL
E-MAIL: iehaya@tx.technion.ac.il
avim@tx.technion.ac.il