The Annals of Probability
1987, Vol. 15, No. 4, 1527–1556

CONTINUOUS MULTI-ARMED BANDITS AND MULTIPARAMETER PROCESSES

BY AVI MANDELBAUM

Stanford University

A general framework is proposed for continuous time dynamic allocation models of a scarce resource among competing projects. The allocation model is formulated as a multi-armed bandit model and solved as a control problem of a multiparameter process. In contrast to discrete time bandits, where only one arm can be pulled at a time, the continuous time bandit must allow simultaneous pulls. The multiparameter approach allows a strong solution of diffusion-type bandits. Here the main problem is to define precisely how to switch among arms and the solution involves local times.

Table of contents

1. Introduction.

1.1. Models of dynamic allocation of a scarce resource to competing projects have been widely used and are of great importance. The main objective of the present work is to propose a general framework for dynamic allocation models in continuous time. The problem of finding the best allocation strategies in such models is naturally formulated as a control problem of multiparameter processes. These are stochastic processes evolving in "time" that is only *partially* ordered. The focus here is on processes with continuous sample paths. We believe, however, that the formulation is appropriate for Poisson models [22], Lévy models [3] and many other scheduling models as well. As observed by Berry and Fristedt ([3], Chapter 8), the main challenge in the formulation is the appropriate definition of an allocation strategy. We resolve this difficulty by identifying an allocation strategy with the multiparameter concept of an optional increasing path [24] or, equivalently, a multiparameter random time change [13]. A byproduct of our approach is a new approximation scheme of

Received August 1985; revised July 1986.
AMS 1980 *subject classifications.* Primary 62L99, 93E20; secondary 60J60, 60K10, 60G17, 60J55.
Key words and phrases. Multi-armed bandits, dynamic allocation, Gittins' index, multiparameter processes, diffusions, local time, optional increasing path, stochastic control.

continuous optional increasing paths by discrete ones (Theorem 7) and a nontrivial unique strong solution to a multiparameter random time change problem (at the end of Section 4.2). The reader is referred to [21] for a comprehensive list of references to the multiparameter theory relevant to the present work (see, however, our remark at the end of Section 7.8).

1.2. The nature of dynamic allocation in continuous time is well demonstrated by the following example. A firm is producing a single product. A customer is expected to arrive at some random time in the future and buy all the available quantity. At any time, *one* out of d production processes can be used and the costs of switching from one process to another are negligible. There are d inventories of raw material and process i uses only inventory i, $i = 1, \dots, d$. The technology involved in process i is such that if i is activated during the kth period, a fraction $r_i\Delta_k$ of the ith inventory is depleted, where Δ_k is the duration of period k. The probability that the customer arrives during the kth period is $\Delta_k\beta$. Assuming that one unit of raw material turns into one unit of final product, the firm's problem is to sequence the processes to maximize the expected quantity produced before the customer arrives.

1.3. If switching among processes is allowed to become more and more frequent ($\Delta_k \downarrow 0$), the firm's model converges to the following continuous time model. The arrival time τ of the customer is exponentially distributed with mean $1/\beta$. Let $X_i(u)$ be the amount of raw material in inventory i after process i has been activated for u time units. Then X_i evolves according to the differential equation $dX_i(u) = -r_iX_i(u)\,du$, $X_i(0)$ given. A strategy is modelled by an R_+^d-valued process $T(t) = (T_1(t), \dots, T_d(t))$, where $T_i(t)$ is the time allocated to process i during the interval $[0, t]$, $\sum_{i=1}^d T_i(t) = t$ for $t \geq 0$. The firm's problem is to maximize the expected cumulative production $R(T) = E\int_0^\tau[-\sum_{i=1}^d dX_i(T_i(t))]$ over all strategies T. By the exponential distribution of τ,

$$R(T) = \int_0^\infty e^{-\beta t} \sum_{i=1}^d Z_i(T_i(t))\, dT_i(t), \tag{1.1}$$

where $Z_i(u) = X_i(0)r_ie^{-r_iu}$, $u \geq 0$, and the firm's problem has been reduced to that of "sequencing continuously" the processes $Z_1, \dots, Z_d$ so as to maximize (1.1). It is rather clear that because all the Z_i decrease in time, the greedy strategy which activates the process with the largest Z_i is optimal (see Section 3.1 if not convinced). Formally, T is optimal if each

1.3.A. $T_i(t)$ increases only at times t such that $Z_i(T_i(t)) = \bigvee_{j=1}^d Z_j(T_j(t))$.

It is obvious how to implement 1.3.A when there is a unique largest Z_i. Eventually, however, there must be a time in which several of the Z_i's are maximal simultaneously. For simplicity, assume that this occurs at time $t = 0$ and $Z_i(0) = Z_j(0)$ for all i, j. Then the unique strategy that satisfies 1.3.A is the one which maintains equal Z_i at *all* times, namely,

$$Z_i(T_i(t)) = Z_j(T_j(t)) \quad \text{for all } i, j, t \geq 0. \tag{1.2}$$

It follows that $r_iT_i(t) = r_jT_j(t)$ for all i, j, $t \geq 0$, which together with $\sum_{i=1}^{d} T_i(t) = t$, implies that for $i = 1, \ldots, d$, the optimal strategy is

$$T_i(t) = \frac{1/r_i}{1/r_1 + \cdots + 1/r_d} t, \qquad t \geq 0. \tag{1.3}$$

1.4. In the discrete model of Section 1.2, only one process was activated at a time. In contrast, the optimal solution (1.3) activates several processes simultaneously. Our formulation of a strategy resolves the difficulty of modelling a continuous-switching mechanism. The phenomenon of simultaneous processing is unavoidable if "continuous sequencing" is allowed. Bellman ([2], Chapter 8) recognized this fact long ago when he solved a continuous version of his gold-mining problem, which is essentially the firm's model of Section 1.3. The firm's model is a very special case of the stochastic dynamic allocation model we now describe and later solve. For historical reasons we have chosen to describe the allocation problem as that of a gambler facing a multi-armed bandit.

1.5. A *discrete* d-armed bandit consists of d statistically independent arms which may be pulled in any order and *one at a time*. The duration of a pull is one unit of time and each pull results in a reward. The problem is to find a strategy which maximizes the expected present value of rewards over an infinite horizon. The bandit model in which arms evolve like Markov chains was formulated and solved by Gittins and his collaborators. Their solution is described in [8]. Alternative solutions have been proposed by Whittle [25] and Katehakis and Veinott [12]. Recently, Varaiya, Walrand and Buyukkoc [23] proposed and solved a bandit model in which the evolution of the arms is described by arbitrary independent processes. This general model was reformulated in [18] using concepts from the theory of multiparameter processes (see Sections 5.2–5.3 of the present paper for a description). The multiparameter approach is especially useful for solving the *continuous* d-armed bandit which is the limit of discrete bandits as the durations of pulls decrease to zero. Specifically, arm i of the continuous bandit is modelled by a continuous time stochastic process Z_i: $Z_i(u)$ is the instantaneous reward from arm i after u time units have been allocated to it. The problem of controlling the d stochastic processes $Z_1, \ldots, Z_d$ is viewed as a control problem of the multiparameter process $Z(s) = (Z_1(s_1), \ldots, Z_d(s_d))$ with parameter $s = (s_1, \ldots, s_d)$ taking values in the d-dimensional nonnegative orthant R_+^d. Similarly, the information accumulated from pulling the arms can be modelled by a filtration indexed by s in R_+^d. With the multiparameter formulation, it is easy to add to the description of a strategy T, given in Section 1.3, the requirement that T should be nonanticipating (see Section 2.2). The continuous d-armed bandit problem is to find a strategy which maximizes the expected cumulative discounted reward $ER(T)$ over T, where $R(T)$ is defined in (1.1).

1.6. The model of the firm in Section 1.3 is an example of a d-armed bandit with decreasing reward processes. We call such bandits *deteriorating bandits*.

The optimal strategies for deteriorating bandits simply pull the arms which yield maximal immediate rewards. (The existence of such strategies is proved in Theorem 3. Necessary and sufficient conditions for the uniqueness of such strategies are described in Proposition 2.) Deteriorating bandits play an important role in our theory because of the following remarkable fact. To any general bandit there corresponds a deteriorating bandit so that a solution to the general bandit is obtained from the solution to the deteriorating one. Moreover, the values of these two solutions coincide. The solution to the deteriorating bandit is described in detail in Section 3. The reduction of the general bandit to the deteriorating one is described in Section 7.1. The value of the general bandit is related to the value of the corresponding deteriorating bandit via formula (7.3). Formula (7.3) is a useful relation that is especially interesting in the context of Markovian bandits.

1.7. The solution to the Markovian bandit in discrete time ([8], [25] and [12]) associates with each arm a numerical function of its possible states. This function is called a priority *index* function. Optimal strategies for the discrete Markovian bandit are exactly those strategies which always pull the arm with the highest index value. A continuous time version of the Markovian bandit was formulated and solved by Karatzas [11]. In Karatzas' model, arms evolve like solutions to Itô stochastic differential equations. We call such bandits *diffusion bandits*. The solution to the diffusion bandit also involves index functions associated with the arms. Again, the optimal strategy pulls the arm with the highest index. However, this is no longer a simple matter because of the sample path behavior of diffusion processes. Karatzas overcame this difficulty by providing a weak solution. The multiparameter approach allows a strong solution which is always unique due to the nature of the diffusion paths. An explicit formula for the index function, which was derived analytically in [11], is rederived probabilistically in Section 4.3 using formula (7.3). The strong solution of diffusion bandits is intimately related to the excursion theory of Markov processes. In fact, formula (7.3) can probably be verified directly using exit systems (see the end of Section 4.6 for more details). The solution to the diffusion bandit is described in Section 4. For the *two*-armed diffusion bandits, the solution involves a partition of R^2 into two parts, say, A_1 and A_2. Suppose that arm 1 is in the state x_1 and 2 in state x_2. Then arm i is pulled if $(x_1, x_2) \in A_i$, $i = 1, 2$. The intersection $\overline{A_1} \cap \overline{A_2}$ is a switching curve on which the controlled process exhibits a local time behaviour. A simple example is described in Section 4.5. More details can be found in [17].

1.8. The diffusion bandit problem can be viewed as a control problem of degenerate diffusion processes in R^d. The related Bellman equations were investigated by Lions [14] as totally degenerate elliptic nonlinear equations. A similar model with general Markov processes was formulated in Grigelionis and Shiryayev [9] and partially solved as a Stefan problem. Both [14] and [9] (see also [1] and [5]) are typical examples of the way dynamic allocation models in continuous time have been treated in the literature: The precise description of

the continuous time problem that is solved is missing. We believe that the multiparameter approach provides a natural and convenient framework that will enable one to fill such gaps in the future.

1.9. The paper starts by describing the continuous bandit model in Section 2. Solutions to the deteriorating and diffusion bandits are given in Sections 3 and 4, respectively. The solution to the discrete bandit is the subject of Section 5, and Section 6 deals with the approximation of continuous strategies by discrete ones. The general model is solved in Section 7 under appropriate conditions and we end in Section 8 with suggestions for future research. A recommended first pass through the paper is to read Section 2, proceed with Sections 3.1, 3.2, 4.1, 4.2, 4.5, 4.6, 5.1, 6.1 and 7.1 and conclude with Section 8.

2. The continuous bandit model.

2.1. Let (Ω, B, P) be a probability space. A continuous *d-armed bandit* is a collection of pairs $\{(Z_i, F_i), i = 1, \ldots, d)\}$, where $Z_i = \{Z_i(t), t \geq 0\}$, the *reward process* associated with arm i, is a bounded real-valued stochastic process on (Ω, B) with continuous sample paths and $F_i = \{F_i(t), t \geq 0\}$, the *information process* associated with arm i, is a complete right-continuous filtration in B. We assume that the arms are independent, meaning that for $i = 1, \ldots, d$, Z_i is adapted to F_i and the σ-fields $F_i(\infty)$ are independent.

2.2. Denote by S the d-dimensional nonnegative orthant $(S = R_+^d)$. An allocation *strategy* is an S-valued stochastic process $T = \{T(t), t \geq 0\}$ on (Ω, B) which satisfies properties 2.2.A–2.2.C. The ith coordinate of T at t, $T_i(t)$, models the total amount of time allocated to arm i over the interval $[0, t]$. In accordance with this interpretation,

2.2.A. $T_i = \{T_i(t), t \geq 0\}$ increases from 0 for all $i = 1, \ldots, d$,

and the following two properties should hold for all $t \geq 0$:

2.2.B. $$T_1(t) + \cdots + T_d(t) = t,$$

2.2.C. $$\{T_1(t) \leq s_1, \ldots, T_d(t) \leq s_d\} \in F_1(s_1) \vee \cdots \vee F_d(s_d) \quad \text{for all } (s_1, \ldots, s_d) \in S.$$

Property 2.2.C is a mathematical formulation of the nonanticipative nature of an allocation strategy: For all i, the event "no more than s_i time units have been allocated to arm i" does not depend on information beyond $F_i(s_i)$.

REMARK. For all i, $0 \leq T_i(u) - T_i(t) \leq u - t$ when $u \geq t \geq 0$. Hence, the sample paths of T_i are absolutely continuous and those of T are continuous in S.

2.3. Introduce in $S = R_+^d$ the partial order

$$r = (r_1, \ldots, r_d) \leq (s_1, \ldots, s_d) = s \quad \text{iff } r_i \leq s_i, \ i = 1, \ldots, d.$$

Properties 2.2.A and 2.2.C can be expressed more compactly as:

2.3.A. T has sample paths that increase (from 0) continuously with respect to the partial order in S.

2.3.B. $\{T(t) \leq s\} \in F(s)$ for all $s \in S$, where

$$F(s) = F_1(s_1) \vee \cdots \vee F_d(s_d), \qquad s = (s_1, \ldots, s_d) \tag{2.1}$$

is the σ-field that models the information available after s_i units of time have been allocated to arm i, $i = 1, \ldots, d$.

In the usual theory of stochastic processes ($d = 1$), properties 2.3.A and 2.3.B are the defining properties of a continuous *random time change* T with respect to the filtration $F = \{F(s), s \in S\}$. In particular, the random variable $T(t)$ is a stopping time with respect to F. In the language of multiparameter processes ($d \geq 2$), $T(t)$ is a *stopping point* [19] and the *multiparameter random time change* T [13] is an *optional increasing path* [24] with respect to F.

REMARK 1. In the sequel, stopping points, pre-$T(t)$ σ-fields and multiparameter martingales will be used. These concepts are obvious generalizations from the usual theory of stochastic processes to processes that "evolve" in "time" that is only partially ordered. For precise definitions and related theory, the reader is referred to [19] (discrete time) and [24] (continuous time).

REMARK 2. The filtrations F_i are right continuous. Hence,

$$F(s) = \bigcap_{\substack{r \geq s \\ r \neq s}} F(r),$$

which is taken as the definition of a *right-continuous multiparameter filtration.*

2.4. The present value $R(T)$ of future rewards associated with a strategy T is the random variable

$$R(T) = \int_0^\infty e^{-\beta t} Z(T(t))\, dT(t), \tag{2.2}$$

where the discount factor β is a positive real number and $Z(T(t))\, dT(t)$ is an abbreviation for $\sum_{i=1}^d Z_i(T_i(t))\, dT_i(t)$. The continuous *bandit problem* is to find *optimal strategies* which maximize the *value function* $V(T) = ER(T)$ over all strategies T.

Since T_i has absolutely continuous sample paths, (2.2) can be rewritten as

$$R(T) = \int_0^\infty e^{-\beta t} Z(T(t)) \dot{T}(t)\, dt,$$

where $Z(T(t))\dot{T}(t) = \sum_{i=1}^d Z_i(T_i(t))\dot{T}_i(t)$ and $\dot{T}_i(t)$ is the derivative of T_i with respect to t. Thus, deciding on a strategy is equivalent to a decision on the rates at which arms are pulled and rewards accumulated.

2.5. By increasing several of the component processes T_i simultaneously, our formulation *does* allow the gambler to pull more than one arm at a time. Still, the total amount of time allocated to all arms over $[0, t]$ must be t.

The option of simultaneous pulls of arms is the main qualitative difference between the discrete and continuous bandit. However, its role in the solution of the continuous bandit varies according to the nature of the arms. The two extremes, in this regard, are the deteriorating bandits and the diffusion bandits which will now be described.

2.6. An arm (Z_i, F_i) of a continuous bandit is *deteriorating* if the reward process Z_i has nonincreasing sample paths. We shall call a d-armed bandit a *deteriorating bandit* if all of its arms are deteriorating. The solution to the deteriorating bandit problem is described in Section 3. Regarding the option of simultaneously pulling arms, a deteriorating bandit is an extreme because its solution typically involves simultaneous pulls at *all* times. The other extreme is a bandit model with a solution that involves simultaneous pulls at essentially *no* time. A nontrivial such model will now be described.

2.7. An arm (Z_i, F_i) is a *diffusion arm* if it evolves like a diffusion process. Formally,

$$Z_i(t) = r_i(X_i(t)),$$

where the *reward function* $r_i(x)$ is a real-valued *increasing* bounded smooth function; X_i is the one-dimensional diffusion process that solves the Itô stochastic differential equation

$$dX_i(t) = \mu_i(X_i(t))\,dt + \sigma_i(X_i(t))\,dW_i(t), \tag{2.3}$$

where W_i is a standard Brownian motion, $\mu_i(x)$ is a smooth local drift coefficient and $\sigma_i(x) > 0$ is a smooth local diffusion coefficient; F_i is the standard complete filtration generated by X_i. The precise meaning of "smooth" for $r_i(x)$, $\mu_i(x)$ and $\sigma_i(x)$ is explained in [11], (2.2) and (3.2), respectively. We have omitted the details because they are never used as far as our results are concerned. A d-armed bandit is called a *diffusion bandit* if all its arms are diffusion arms and the Brownian motions $W_1, \ldots, W_d$ are independent. The solution to the diffusion bandit is outlined in Section 4. Due to the "wild" nature of the diffusion sample paths, the solution to the diffusion bandit involves, vaguely speaking, an uncountable "number of switches" among arms. However, simultaneous pulls *essentially never* take place.

3. Solution to the deteriorating bandit problem.

3.1. In view of the performance measure (2.2) used, an obvious question is: Why not allocate *all* time to an arm which yields *maximal immediate* reward? The answer is that there may be other arms which promise large benefits in the future and, because of the discounting, getting to these future benefits as soon as possible may turn out to be more attractive. The future, however, is never more attractive than the present when the bandit is a deteriorating bandit (as defined in Section 2.6). One expects, therefore, that optimal strategies for deteriorating bandits *do* pull the arms which yield maximal immediate reward. We shall now describe formally the solution to the deteriorating bandit problem. Even though

the solution is intuitively obvious, it illustrates best the differences between discrete and continuous time bandits. Moreover, the solution to the *general* continuous bandit problem can be viewed as a reduction to the *deteriorating* one (see Section 7.1 and Theorem 12). Hence, the present section is an essential part of the solution to the general continuous bandit.

3.2. When time is discrete, it is obvious how to pull an arm which yields maximal immediate reward. For continuous time, we propose the following

DEFINITION. Given d processes $Z_1, \ldots, Z_d$, a strategy $T = \{T(t),\ t \geq 0\}$ *follows the leader among the* Z_i*'s* if for $i = 1, \ldots, d$, T_i *increases at time* t only when Z_i is maximal at that time, namely,

$$(3.1) \quad T_i(u) > T_i(t) \quad \forall\, u > t \text{ only when } Z_i(T_i(t)) = \bigvee_{j=1}^{d} Z_j(T_j(t)).$$

The performance measure (2.2) of such a strategy is given, pathwise, by

$$(3.2) \qquad R(T) = \int_0^\infty e^{-\beta t} \bigvee_j Z_j(T_j(t))\, dt.$$

Note that (3.1) is a statement about *sample paths*, which has nothing to do with either their random or deteriorating nature. Similarly, the results of the present section are results about *functions* rather than stochastic processes and the concept of "following the leader" is applicable to arbitrary functions, not necessarily decreasing ones. A strategy that follows the leader among continuous processes always exists. Its existence is established for deteriorating arms in Theorem 3. The general case is reduced to the deteriorating one in Theorem 12.

3.3. As anticipated by the heuristics outlined in Section 3.1, any strategy that follows the leader among the arms of a deteriorating bandit is optimal (see Section 7.12 for the proof). The nature of such a strategy becomes apparent from

PROPOSITION 1. *Fix* $i \neq j$. *Suppose that both arm* i *and* j *deteriorate and* $Z_i(0) = Z_j(0)$. *If* T *follows the leader among* Z_i *and* Z_j, *then* $Z_i(T_i(t)) = Z_j(T_j(t))$ *for all* $t \geq 0$.

PROOF. Suppose first that $Z_i(T_i(t)) > Z_j(T_j(t))$ for some t. By continuity, there exists an ε in $[0, t)$ such that

$$(3.3) \qquad Z_i(T_i(\varepsilon)) = Z_j(T_j(\varepsilon))$$

and

$$Z_i(T_i(u)) > Z_j(T_j(u)) \quad \text{for } \varepsilon < u \leq t.$$

Hence, T_j does not increase over (ε, t). Since T_j is continuous, $T_j(\varepsilon) = T_j(t)$,

implying that

$$Z_i(T_i(\varepsilon)) \geq Z_i(T_i(t)) > Z_j(T_j(t)) = Z_j(T_j(\varepsilon)),$$

which contradicts (3.3). Reversing the roles of i and j establishes Proposition 1. □

From Proposition 1, one concludes that following the leader among deteriorating arms amounts to pulling simultaneously all "current leaders" at rates which are determined by the following procedure:

3.3.A. Let L be the set of "leaders at time 0," namely, $i \in L$ if and only if

$$Z_i(0) = \bigvee_{j=1}^{d} Z_j(0).$$

3.3.B. Arms out of L are *not* pulled while arms in L are pulled at rates which maintain the relations

$$Z_i(T_i(t)) = Z_j(T_j(t)), \qquad t \geq 0, \tag{3.4}$$

for all i, j in L.

3.3.C. Proceed with 3.3.B until at some time t, (3.4) holds for all i in L *and* some j *not* in L. All j not in L for which (3.4) holds become leaders as well: Add these j to L and return to 3.3.B.

REMARK. A strategy that satisfies 3.2.B and 3.2.C clearly satisfies (3.1).

3.4. We now address the *uniqueness* of strategies that follow the leader among *deteriorating* arms. As mentioned before, the analysis is that of functions (individual sample paths) rather than stochastic processes. Consequently, the emphasis will be on the ideas rather than their (cumbersome) proofs.

The simplest situation in which uniqueness holds is when the reward processes Z_i *strictly* decrease. An illuminating example is the deterministic case $Z_i(t) = -r_i t$, $i = 1, \ldots, d$, with r_i strictly positive. The only $T = \{T_i,\ i = 1, \ldots, d\}$ that satisfies (3.4) is given in (1.3) (the firm's problem). Thus, for nonuniqueness it is *necessary* that the Z_i's have "flat" parts. The sufficient condition for nonuniqueness is described in the following

DEFINITION. The two deteriorating arms i and j are *simultaneously flat* if there exists a level at which *both* arms do not decrease, i.e., there are times $t_i < u_i$ and $t_j < u_j$ such that

$$Z_i(t_i) = Z_i(u_i), \qquad Z_j(t_j) = Z_j(u_j)$$

and

$$Z_i(t_i) = Z_j(t_j).$$

PROPOSITION 2. *A strategy that follows the leader among deteriorating arms is unique if and only if among the arms that are ever pulled no two are simultaneously flat.*

PROOF. Suppose that arms i and j are simultaneously flat at time 0 ($t_i = t_j = 0$) and both are the only "leaders at time 0" in the sense of 3.3.A.

One can start by pulling *only* arm i until the first time it decreases and then switch to arm j. One can also pull only arm j first, hence, nonuniqueness. To prove the "if" direction assume that T and U are strategies that follows the leader among the Z_i's. The first step is to show that

$$\bigvee_{j=1}^{d} Z_j(T_j(t)) = \bigvee_{j=1}^{d} Z_j(U_j(t)) \quad \text{for all } t \geq 0. \tag{3.5}$$

Now assume for simplicity that $Z_i(0) = Z_j(0)$ for all $i \neq j$. By Proposition 1, $Z_i(T_i(t)) = Z_j(U_j(t))$ for all i, j and $t \geq 0$. If $T_i(t) > U_i(t)$ for some i and some t, then $T_j(t) < U_j(t)$ for some $j \neq i$, Z_i is flat over $(U_i(t), T_i(t))$, Z_j is flat over $(T_j(t), U_j(t))$ and we are done. □

An alternative approach to nonuniqueness, which is also helpful for establishing existence, is to consider strategies that apply a priority scheme whenever nonuniqueness arises. A priority scheme is described by a permutation $(i_1, \ldots, i_d)$ of $(1, \ldots, d)$: Arm i_m is pulled only when arms i_l, $l < m$, cannot be pulled. Formally, we give the

DEFINITION. A strategy T follows the leader among the *deteriorating* arms Z_i *according to the priority scheme* $(i_1, \ldots, i_d)$ if T satisfied (3.1) and, in addition, whenever T_{i_m} increases at t and $Z_{i_l}(T_{i_l}(t)) = \bigvee_j Z_j(T_j(t))$ for some $l < m$, then Z_{i_l} decreases at time $T_{i_l}(t)$, that is,

$$Z_{i_l}(u) < Z_{i_l}(T_{i_l}(t)) \quad \text{for all } u > T_{i_l}(t). \tag{3.6}$$

THEOREM 3. *A strategy that follows the leader among deteriorating arms according to a fixed priority scheme exists and is unique.*

PROOF. Rather than building on previous results, we present an alternate approach which will be useful later as well. For simplicity, assume that

$$Z_i(0) = 0, \qquad Z_i(\infty) = -\infty \quad \text{for } i = 1, \ldots, d.$$

For $x \geq 0$ and $i = 1, \ldots, d$, define

$$l_i(x) = \inf\{t \geq 0 \colon Z_i(t) = -x\} \tag{3.7}$$

to be the left-continuous inverse of $-Z_i$. Note that arms i and j are simultaneously flat if and only if l_i and l_j have a common point of increase. The key observation is:

3.4.A. The sample path of *any* strategy T which follows the leader among the Z_i's *must* pass through *all* the points $l(x) = (l_1(x), \ldots, l_d(x))$, $x \geq 0$.

The time when T crosses $l(x)$ is uniquely determined by 2.2.B: For all i,

$$T_i(\sigma(x)) = l_i(x), \qquad x \geq 0, \tag{3.8}$$

where

$$\sigma(x) = l_1(x) + \cdots + l_d(x), \qquad x \geq 0.$$

Indeed,

$$\sigma(x) = \inf\{t \geq 0\colon C(t) = -x\}, \qquad x \geq 0,$$

where $C(t) = Z_i(T_i(t)) = Z_j(T_j(t))$ for all $i, j, t \geq 0$ by Proposition 1. From (3.8), the value of all the strategies that follow the leader is uniquely determined at all times $\{\sigma(x), x \geq 0\}$. Now consider the discontinuity intervals $(\sigma(x), \sigma(x+))$, with $\sigma(x) < \sigma(x+)$. The set $L = \{i\colon l_i(x+) - l_i(x) > 0\}$ determines the arms that can be pulled during that interval. The order in which arms are pulled is uniquely determined if T follows the leader according to a priority scheme, hence, T is unique.

The preceding reasoning also suggests how to construct a strategy T that follows the leader among $Z_1, \ldots, Z_d$ according to a given priority scheme, say $(1, 2, \ldots, d)$: The values at times $\sigma(x)$ are given by (3.8), the values on $(\sigma(x), \sigma(x+))$ are determined by the priority scheme and the values at $\sigma(x+)$ are determined by continuity. Formally, for $i = 1, \ldots, d$ and t in the interval $[\Sigma_{j=1}^{i-1}(l_j(x+) - l_j(x)), \Sigma_{j=1}^{i}(l_j(x+) - l_j(x))]$,

$$\begin{aligned} T(\sigma(x) + t) = {} & l(x) + [l_1(x+) - l_1(x)]e_1 + \cdots \\ & + [l_{i-1}(x+) - l_{i-1}(x)]e_{i-1} + \left[t - \sum_{j=1}^{i-1}\left(l_j(x+) - l_j(x+)\right)\right]e_i, \end{aligned} \tag{3.9}$$

where e_i is the ith unit vector. By the construction, T satisfies 2.2.A and 2.2.B. Property 2.2.C is verified by showing that T is the limit of strategies that pull only one arm at a time. To this end, fix an $\varepsilon > 0$. Let the strategy $T^\varepsilon = [T^\varepsilon(t), t \geq 0]$ operate as follows: Start by pulling only arm 1 until Z_1 reaches level $-\varepsilon$. Proceed sequentially with each arm $i = 2, \ldots, d$, until arm i reaches level $-\varepsilon$ as well. When arm d is at level $-\varepsilon$, return to arm 1 and let it reach -2ε, proceed with $2, \ldots, d$ sequentially until they are all at level -2ε, and so on. Arm i reaches level $-k\varepsilon$ after $l_i(k\varepsilon)$ units of time have been allocated to it. In other words, T^ε and the strategy T in (3.9) coincide at the time $\{\sigma(k\varepsilon), k = 0, 1, \ldots\}$. It follows that with $\varepsilon = 2^{-n}$, $T^\varepsilon(t)$ converges to $T(t)$ as $n \uparrow \infty$. This is sufficient for 2.2.C because $T^\varepsilon(t)$ are stopping *points* with respect to a complete right-continuous filtration and a limit of such a sequence must be a stopping point as well. □

4. Solution to the diffusion bandit problem.

4.1. We now describe the solution to the diffusion bandit problem, as defined in Section 2.7. The main optimality results follow from the analysis in [11] combined with the solution to the general continuous bandit (see Section 7.11). In analogy to the discrete time Markovian bandit, an index function is associated with each arm and an optimal strategy for the diffusion bandit pulls the arm with the highest index. The rigorous description is based on the ideas developed

in Section 3.2 in order to "follow the leader" among deteriorating arms. The focus here is on the elements of the solution that are characteristic of diffusion bandits.

4.2. The *index function* $M_i(x_i)$ associated with arm i is the $C^2(R)$ strictly *increasing* function

$$M_i(x_i) = \sup_{\tau>0} \frac{E_{x_i}\int_0^\tau e^{-\beta u} r_i(X_i(u))\,du}{E_{x_i}\int_0^\tau e^{-\beta u}\,du}, \tag{4.1}$$

where the supremum is over all F^i-stopping times that are positive a.s. and are not necessarily finite. Here E_{x_i} denotes the fact that X_i solves (2.3) with initial condition $X_i(0) = x_i$.

THEOREM 4. *The optimal strategy for the d-armed diffusion bandit is the unique strategy* $I = \{I(t),\ t \geq 0\}$, *which follows the leader among the index processes* $M_i(X_i) = \{M_i(X_i(t)),\ t \geq 0\}$, $i = 1, \ldots, d$. *That is,* I_i *increases at time* t *only when*

$$M_i(X_i(I_i(t))) = \bigvee_{j=1}^{d} M_j(X_j(I_j(t))). \tag{4.2}$$

The expected present value of future rewards associated with I is

$$V(I) = E\int_0^\infty e^{-\beta t} \bigvee_j M_j(\underline{X}_j(I_j(t)))\,dt. \tag{4.3}$$

Here, the process $\underline{X}_i = \{\underline{X}_i(t),\ t \geq 0\}$ is the *lower envelope* of the process X_i,

$$\underline{X}_i(t) = \min_{0 \leq u \leq t} X_i(u).$$

When comparing Theorem 4 with the solution in [11], the reader will note the following improvements. Our solution is a pathwise solution in terms of the sample paths of the X_i's (a *strong* solution) and the solution is unique for any number of arms. We also provide an explicit representation (4.3) of the value of the optimal strategy in terms of the index functions.

REMARK. The strategy I can be viewed as the unique strong solution to a nontrivial multiparameter time change problem as formulated by Kurtz ([13], especially Section 5).

4.3. (4.3) is a consequence of the more general (7.3). To see that, note that the lower envelope of $M_i(X_i)$ is $M_i(\underline{X}_i)$ by the monotonicity of the index function M_i. Both (4.3) and (7.3) differ from their deteriorating bandit analog (3.2) in that the latter is a pathwise identity, while the former holds only in expectation. (4.3) is also valid if applied to a single arm, namely,

$$E_{x_i}\int_0^\infty e^{-\beta t} r_i(X_i(t))\,dt = E_{x_i}\int_0^\infty e^{-\beta t} M_i(\underline{X}_i(t))\,dt, \tag{4.4}$$

which is an interesting identity for the resolvent of the function r_i. We now use (4.4) to derive probabilistically an explicit formula for the index function $M_i(x_i)$. For notational convenience, the subscript i will be omitted. First, use the strong Markov property to replace the upper limits ∞ in (4.4) by a stopping time. While *not* all stopping times qualify, first hitting times of points below x do, and we get for all $\varepsilon > 0$,

$$E_x\int_0^{\tau(x-\varepsilon)} e^{-\beta t} r(X(t))\,dt = E_x\int_0^{\tau(x-\varepsilon)} e^{-\beta t} M(\underline{X}(t))\,dt, \tag{4.5}$$

where $\tau(x-\varepsilon)$ is the first hitting time of $x-\varepsilon$. Now divide both sides of (4.5) by $E_x\int_0^{\tau(x-\varepsilon)} e^{-\beta t}\,dt$ and let $\varepsilon \downarrow 0$. Since M is continuous and $\underline{X}(t)$ is "trapped" between x and $x-\varepsilon$ at all times before $\tau(x-\varepsilon)$, we get

$$M(x) = \lim_{\varepsilon \downarrow 0} \frac{E_x\int_0^{\tau(x-\varepsilon)} e^{-\beta t} r(X(t))\,dt}{E_x\int_0^{\tau(x-\varepsilon)} e^{-\beta t}\,dt}, \tag{4.6}$$

which identifies a sequence of stopping times that converge to the supremum in (4.1). By standard Markov process calculations [10],

$$E_x\int_0^{\tau(x-\varepsilon)} e^{-\beta t} r(X(t))\,dt = R(x) - R(x-\varepsilon)\frac{G(x)}{G(x-\varepsilon)}, \tag{4.7}$$

where $R(x)$ is the resolvent of r and $G(x)$ is determined by the relation $E_x e^{-\beta\tau(y)} = G(x)/G(y)$ for $x > y$. Finally, (4.6) and (4.7) yield the formula

$$M(x) = \beta\left\{R(x) - R'(x)\frac{G(x)}{G'(x)}\right\}, \tag{4.8}$$

due to Karatzas [11], who derived it analytically.

4.4. The existence of the strategy I in Theorem 4 is verified in Theorem 12. An important step is the observation that a strategy that follows the leader among continuous processes also follows the leader among their lower envelopes. We now combine this observation with Proposition 2 to show that I is *unique*. By Itô's formula, the process $Y_i = M_i(X_i)$ is a solution to a stochastic differential equation of the form (2.3) with diffusion coefficient $M_i'(y_i)\sigma(y_i) > 0$. Uniqueness of I follows then from Proposition 2 and

PROPOSITION 5. *Let* $Y_i = \{Y_i(t),\ t \geq 0\}$, $i = 1,\ldots,d$, *be a solution to the stochastic differential equation* (2.3) *with initial condition* $Y_i(0) = 0$ *and independent* W_i. *Then the lower envelopes of the* Y_i*'s are simultaneously flat with probability* 0.

PROOF. By a change of measure ([15], Theorem 7.19) one may assume, without loss of generality, that the drift coefficients of the Y_i's are all 0. By a random time change ([7], Section 2.11), it can be further assumed that the diffusion coefficients are all 1. In other words, Proposition 5 has been reduced to the case where the Y_i's *themselves* are independent standard Brownian motions.

Let τ_i be the right-continuous modification of the process l_i defined in (3.7) with Z_i replaced by the lower envelope of the Brownian motion Y_i. The process τ_i, $i = 1, \ldots, d$, are independent, increasing and have stationary independent increments (see, for example, [26], Section 64). The process Y_i are simultaneously flat with probability 0 if and only if any two of the processes τ_i increase at the same time with probability 0. For $i = 1, \ldots, d$ and $\varepsilon > 0$, let

$$N_i^{\varepsilon}(t) = \#\{u \leq t\colon \tau_i(u) - \tau_i(u-) \geq \varepsilon\}.$$

Then N_i^{ε}, $i = 1, \ldots, d$, are independent homogeneous Poisson processes implying that $\sum_{i=1}^{d} N_i^{\varepsilon}$ is Poisson as well. In particular, *with probability* 0, there is a time in which any two of the N_i^{ε} increase simultaneously. But the last statement holds for all $\varepsilon > 0$, which establishes Proposition 5. □

4.5. We conclude the section with an explicit example that illustrates the nature of the "switchings" among diffusion arms. Assume that there are only two arms ($d = 2$) characterized by X_1, X_2 independent standard Brownian motions *starting at* 0 and $r(x_i) = x_i - 1/(2\beta)^{1/2}$. Note that, being unbounded, linear rewards are formally not covered by our analysis. Nevertheless, as demonstrated in [17], Theorem 4 still applies. Using (4.8) with

$$G_i(x_i) = e^{-(2\beta)^{1/2}x_i}, \qquad R_i(x_i) = \frac{1}{\beta}\left(x_i - \frac{1}{(2\beta)^{1/2}}\right),$$

the index functions are simply $M_i(x_i) = x_i$. The optimal strategy $I = (I_1, I_2)$ is uniquely determined by the requirement that I_1 increase at t only when $B(t) \geq 0$ and I_2 increase at t only when $B(t) \leq 0$, where

$$B(t) = X_1(I_1(t)) - X_2(I_2(t)). \tag{4.9}$$

LEMMA 6. *Let $T = (T_1, T_2)$ be any strategy. The process $X(T) = \{X(T(t)), t \geq 0\}$, given by*

$$X(T(t)) = X_1(T_1(t)) - X_2(T_2(t)),$$

is a standard Brownian motion starting at 0.

PROOF. The basic idea is that a multiparameter random time change of a multiparameter martingale is typically a martingale. Specifically, let $F = \{F(s), s \in R_+^2\}$ be the two-parameter filtration

$$F(s) = \sigma\{X_1(t), t \leq s_1;\ X_2(t), t \leq s_2\}, \qquad s = (s_1, s_2) \in R_+^2. \tag{4.10}$$

The following six processes are all two-parameter martingales with respect to F:

$$\{X_i(s_i), s \in R_+^2\}, \qquad i = 1, 2,$$
$$\{X_i^2(s_i) - s_i, s \in R_+^2\}, \qquad i = 1, 2,$$
$$\{X_1(s_1)X_2(s_2), s \in R_+^2\},$$
$$\{X_1(s_1) - X_2(s_2), s \in R_+^2\}.$$

By Propositions 2.3 and 3.2 in [24], the time changed processes

4.5.A. $\{X_i(T_i(t)), t \geq 0\}, \quad i = 1,2,$

4.5.B. $\{X_i^2(T_i(t)) - T_i(t), t \geq 0\}, \quad i = 1,2,$

4.5.C. $\{X_1(T_1(t))X_2(T_2(t)), t \geq 0\},$

4.5.D. $\{X(T(t)) = X_1(T_1(t)) - X_2(T_2(t)), t \geq 0\}$

are all martingales with respect to the filtration $F^T = \{F(T(t)), t \geq 0\}$, where $F(T(t))$ is the pre-$T(t)$ σ-field obtained from F in (4.10). Consequently, $X(T)$ is an F^T-martingale with continuous sample paths and quadratic variation process

$$\langle X(T)\rangle_t = T_1(t) + T_2(t) = t.$$

By Lévy's characterization of the Brownian motion, Lemma 6 follows. □

In particular, B defined in (4.9) is a Brownian motion; the set of times when I_1 and I_2 increase simultaneously is a subset of the zero-set of a Brownian motion (which has a.s. Lebesgue measure 0); the positive (negative) excursions of B away from 0 are the periods when *only* arm 1 (arm 2) is pulled. Moreover, note that for $t \geq 0$, the positive and negative parts of B are

$$B^+(t) = X_1(I_1(t)) + \tfrac{1}{2}L^B(t), \quad t \geq 0, \tag{4.11}$$

$$B^-(t) = X_2(I_2(t)) + \tfrac{1}{2}L^B(t), \quad t \geq 0, \tag{4.12}$$

where

$$L^B(t) = -2X_1(I_1(t)) \wedge X_2(I_2(t)) = -2\underline{X}_i(I_i(t)), \quad i = 1,2,$$

which increases in t. By 4.5.A, the representations (4.11) and (4.12) are actually the unique Doob–Meyer decompositions of the supermartingales B^+ and B^-. Finally, Tanaka's formula identifies L^B as the local time of B at 0 and (4.3) provides an explicit expression for the value of I, namely,

$$\begin{aligned} ER(I) &= -\frac{1}{2}E\int_0^\infty e^{-\beta t}L^B(t)\,dt \\ &= -\frac{1}{2}E\int_0^\infty e^{-\beta t}\left(\frac{2t}{\pi}\right)^{1/2} dt \\ &= -\frac{1}{2\beta(2\beta)^{1/2}}. \end{aligned}$$

4.6. The processes τ_i and N_i^ε used in the proof of Proposition 5 provide quantitative information about the "number of switches" that take place among arms. While it is impossible to count the total number of switches, $N_i^\varepsilon(a)$, $a \geq 0$, represents the number of times arm i has been pulled *individually* for time periods that exceed ε, by the time it reaches level $-a$. For the example in Section 4.5, $\{N_i^\varepsilon(a), a \geq 0\}$, $i = 1,2$, are independent Poisson processes with parameter $(2/\pi\varepsilon)^{1/2}$ ([26], page 95). Other quantities of interest could be

calculated using the excursion theory of Markov processes. In fact, Haya Kaspi used exits systems [16] to verify (4.4) directly for any nice strong Markov process X_i.

5. Discrete time bandits.

5.1. Both the deteriorating and diffusion bandits are special cases of the continuous bandit. To solve this last model, the continuous bandit will be approximated by a family of discrete bandits in which *only one* arm may be pulled at a time. The discrete bandit will now be described in a form suitable for that approximation. The model is due to [23] and complete proofs can be found in [18]. The notation used in the current section is similar to that of Section 2. However, no confusion should arise since Section 5 is the only section where discrete bandits are dealt with explicitly.

5.2. The primitives for the discrete d-armed bandit model are identical to those of the continuous one (Section 2.1) except that time is now discrete. Thus the ith reward process $Z_i = \{Z_i(k),\ k = 0,1,\dots\}$ is adapted to the ith information process $F_i = \{F_i(k),\ k = 0,1,\dots\}$ and $F_i(\infty)$, $i = 1,\dots,d$, are independent complete σ-fields. Let N be the set of nonnegative integers and endow $S = N^d$ with the partial order of Section 2.3. An allocation *strategy* $T = \{T(k),\ k = 0,1,\dots\}$ is an S-valued stochastic process with $T(0) = 0$ such that for all $k = 0,1,\dots,$

5.2.A. $T(k+1)$ is a direct successor of $T(k)$,
5.2.B. $T(k)$ is a stopping point with respect to $F = \{F(s),\ s \in S\}$, and
5.2.C. $T(k+1)$ is measurable with respect to $F(T(k))$, the pre-$T(k)$ σ-field.

The filtration F in 5.2.B is given by

$$F(s) = F_1(s_1) \vee \cdots \vee F_d(s_d), \qquad s = (s_1,\dots,s_d) \in N^d,$$

but 5.2.A–5.2.C are equally applicable for an arbitrary discrete partially ordered set S.

REMARK 1. Property 5.2.C is a mathematical formulation of the nonanticipative nature of an allocation strategy in discrete time. Property 5.2.B justifies the use of the pre-$T(k)$ σ-field in 5.2.C. The descriptions 5.2.A–5.2.C were developed in [19] in order to solve the optimal stopping problem over discrete partial ordered sets.

REMARK 2. Using induction on k, one can prove that 5.2.B is actually a consequence of 5.2.C for any discrete partially ordered S. However, for filtrations $F = F_1 \vee \cdots \vee F_d$ with independent F_i's (and more generally, $F4$-filtrations), 5.2.B and 5.2.C are *equivalent* in the presence of 5.2.A. For more details, see Section 3 in [17].

5.3. The present value of an allocation strategy T is the random variable

$$R(T) = \sum_{k=0}^{\infty} \alpha^k Z(T(k))\,\Delta T(k), \tag{5.1}$$

where

$$Z(T(k))\,\Delta T(k) = \sum_{i=1}^{d} Z_i(T_i(k))\big[T_i(k+1) - T_i(k)\big]$$

and $0 < \alpha < 1$. The objective is to maximize over T the *value* $V(T) = ER(T)$. Optimal strategies are described in terms of index processes. The *index process* $M_i = \{M_i(k),\ k = 0, 1, \dots\}$ associated with arm i given by

$$M_i(k) = \operatorname*{ess\,max}_{\tau \geq k+1} \frac{E^{F_i(k)} \sum_{l=k}^{\tau-1} \alpha^l Z_i(l)}{E^{F_i(k)} \sum_{l=k}^{\tau-1} \alpha^l}, \tag{5.2}$$

where τ in (5.2) is an F_i-stopping time that is not necessarily finite and $\tau \geq k+1$ a.s. The ess max notation in (5.2) indicates that the essential supremum of the ratios is actually attained.

5.4. The *index field* $M = \{M(s),\ s \in N^d\}$ is the multiparameter process

$$M(s) = \bigvee_{j=1}^{d} M_j(s_j), \qquad s = (s_1, \dots, s_d) \in N^d. \tag{5.3}$$

Let $\underline{M} = \{\underline{M}(s),\ s \in N^d\}$ be the lower envelope of M, that is,

$$\underline{M}(s) = \min_{0 \leq r \leq s} M(r), \qquad r, s \in N^d.$$

The process $\underline{M}$ is related to the lower envelopes $\underline{M}_i = \{\underline{M}_i(k),\ k = 0, 1, \dots\}$ by

$$\underline{M}(s) = \bigvee_{j=1}^{d} \underline{M}_j(s_j). \tag{5.4}$$

The main results in [18] will now be summarized.

5.4.A. For any two strategies T and U over N^d,

$$V(T) \leq E \sum_{k=0}^{\infty} \alpha^k \underline{M}(U(k)). \tag{5.5}$$

DEFINITIONS. A strategy I is an *index strategy* if it *follows the leader* among the index processes M_i's. Formally, for $k = 0, 1, \dots$,

$$I(k+1) = I(k) + e_i \quad \text{on} \quad M_i(I_i(k)) = \bigvee_{j=1}^{d} M_j\big(I_j(k)\big), \tag{5.6}$$

where e_i is the ith unit vector in N^d. Let $(i_1, \dots, i_d)$ be a permutation of $(1, \dots, d)$. A strategy I follows the leader among the M_i's *according to the priority scheme* $(i_1, \dots, i_d)$ if in addition to (5.6), I pulls arm i_m only when arms i_l, $l < m$, cannot be pulled. An index strategy is called a *priority index* strategy

if it follows the leader among the M_i's according to some fixed priority scheme. While there may be many index strategies, a priority index strategy is uniquely determined by its priority scheme (compare with Proposition 3).

5.4.B. If I is an index strategy then

$$V(I) = E\sum_{k=0}^{\infty} \alpha^k \underline{M}(I(k)). \tag{5.7}$$

5.4.C. If I and J are index strategies, then

$$\underline{M}(I(k)) = \underline{M}(J(k)) \quad \text{for } k = 0, 1, \ldots . \tag{5.8}$$

By 5.4.A–5.4.C, index strategies attain highest values. Moreover:

5.4.D. The class of optimal strategies *coincides* with the class of index strategies.

REMARK. When all reward processes Z_i have decreasing sample path (a deteriorating bandit), the index process M_i coincides with Z_i for all i. If I is an index strategy for the deteriorating bandit, then pathwise

$$R(I) = \sum_{k=0}^{\infty} \alpha^k \underline{M}(I(k)) = \sum_{k}^{\infty} \alpha^k \bigvee_j \underline{M}_j(I_j(K))$$

[compare with (3.2)]. In general, however, only equality of expectations holds [compare with (4.3)].

5.5. Suppose that the reward processes are given for $i = 1, \ldots, d$ by

$$Z_i(k) = r_i(X_i(k)), \qquad k = 0, 1, \ldots,$$

where $r_i(x)$ is a bounded measurable real-valued function and $X_i = \{X_i(k), k = 0, 1, \ldots\}$ is a real-valued homogeneous Markov chain with respect to the filtration F_i.

5.5.A. The index processes M_i are of the form $M_i(k) = M_i(X_i(k))$, where

$$M_i(x) = \max_{\tau \geq 1} \frac{E_x \sum_{l=0}^{\tau-1} \alpha^l r_i(X_i(l))}{E_x \sum_{l=0}^{\tau-1} \alpha^l}, \qquad x \in R^1.$$

5.5.B. When the function $r_i(x)$ is continuous and the transition operator for X_i has the Feller property, the index function $M_i(x)$ is lower semicontinuous.

6. Approximating continuous strategies and their values.

6.1. The key to the approximation of continuous bandits by a family of discrete bandits is the ability to approximate the continuous strategies described in Section 2, by the discrete strategies of Section 5. Strategies in the discrete bandit model pull only *one* arm at a time and the *duration of a pull* is a fixed positive number. Strategies that pull *several* arms *simultaneously* are *limits* of discrete strategies as the duration of pulls converges to 0. We now describe the

approximation scheme of continuous strategies by discrete ones. The procedure is of independent interest in the pure context of multiparameter processes. This section ends with a proof that the value function is a continuous function under uniform convergence of strategies.

6.2. Let $T = \{T(t),\ t \geq 0\}$ be a continuous allocation strategy. We say that T is *discrete of order* ε if over each period of time $[k\varepsilon, (k+1)\varepsilon)$, only *one* arm is pulled. Formally, for $k = 0, 1, \ldots,$

$$T(t) = T(k\varepsilon) + (t - k\varepsilon)e_i, \qquad k\varepsilon \leq t \leq (k+1)\varepsilon, \tag{6.1}$$

for some $i = 1, \ldots, d$ (i random).

A strategy T that is discrete of order ε is determined by its values at times $\{k\varepsilon,\ k = 0, 1, \ldots\}$. Moreover, the right continuity of F in (2.1) together with (6.1) imply that

$$T((k+1)\varepsilon) \in F(T(k\varepsilon)), \qquad k = 0, 1, \ldots. \tag{6.2}$$

Thus, given a right-continuous filtration $F = \{F(s),\ s \in R_+^d\}$, the class of discrete strategies of order ε with respect to F can be identified with the class of discrete strategies in the sense of 5.2.A–5.2.C, with $S = \varepsilon N^d$. (The discrete filtration is the restriction of the continuous F to εN^d.)

6.3. For $\varepsilon > 0$, denote by Π^ε the class of strategies that are discrete of order ε.

THEOREM 7. *Let $T = \{T(t),\ t \geq 0\}$ be a strategy in the sense of* 2.2.A–2.2.C. *There exists a family of strategies $\{T^\varepsilon,\ \varepsilon > 0\}$ such that $T^\varepsilon \in \Pi^\varepsilon$ and for all $\omega \in \Omega$, T^ε converges to T uniformly in t as ε decreases to* 0, *that is,*

$$\lim_{\varepsilon \downarrow 0} \sup_{t \geq 0} |T^\varepsilon(t) - T(t)| = 0, \tag{6.3}$$

where $|\cdot|$ is any norm in R^d.

REMARK 1. Theorem 7 holds for arbitrary optional increasing path T with respect to a right-continuous filtration over R_+^d.

REMARK 2. When $d = 2$, the proof of Theorem 7 is captured by Figure 1. The strategy T^ε starts at 0 by pulling sequentially both arms for ε units of time, in an arbitrary order. Thus, $T^\varepsilon(2\varepsilon) = (\varepsilon, \varepsilon)$. Proceed by always pulling an arm which is known to be "lagging behind T." In the picture, $T_1^\varepsilon(2\varepsilon) < T_1(2\varepsilon)$, hence, 1 is pulled. The key observation is that the information available at $s = (\varepsilon, \varepsilon)$ (i.e., time $t = 2\varepsilon$) suffices to determine in which direction to proceed. If at any point T and T^ε meet [as in $s = (2\varepsilon, 2\varepsilon)$, i.e., time $t = 4\varepsilon$], again pull sequentially both arms for ε units of time each.

PROOF OF THEOREM 7. Generalizing to arbitrary $d \geq 2$, T^ε is constructed as follows. Start with $T^\varepsilon(0) = 0$ and suppose that $T^\varepsilon(k\varepsilon)$ has been constructed. To

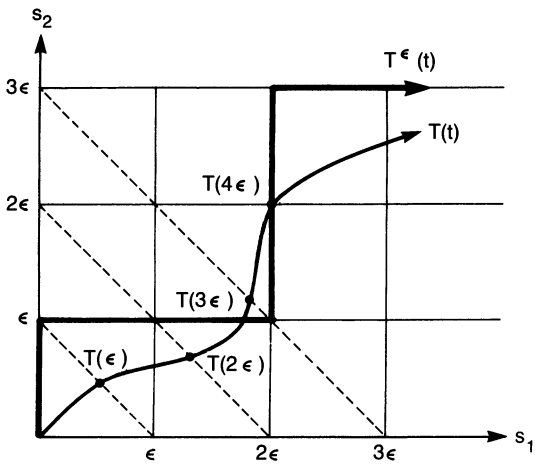

FIG. 1. *Approximating a continuous strategy by a discrete strategy of order* ε.

define $T^\varepsilon((k+1)\varepsilon)$, let

$$\tau_i = \inf\{t\colon T_i(t) > T_i^\varepsilon(k\varepsilon)\}, \qquad i = 1, \ldots, d,$$

and define

$$\tau = \min_{1 \le i \le d} \tau_i.$$

Note that τ is the first exit time of T from the cube

$$[0, T^\varepsilon(k\varepsilon)] = \{s \in R_+^d\colon 0 \le s \le T^\varepsilon(k\varepsilon)\}.$$

By 2.2.B, $\tau \le k\varepsilon$. Let $H = \{i\colon \tau_i = \tau\}$. Choose an i in H and pull arm i, that is,

$$T^\varepsilon(t) = T^\varepsilon(k\varepsilon) + (t - k\varepsilon)e_i, \qquad k\varepsilon \le t \le (k+1)\varepsilon.$$

REMARK. If $H = \{i_1, \ldots, i_h\}$, the procedure will "exhaust" arms $i_1, \ldots, i_h$ before moving forward. Also, for $i \in H$, $T_i^\varepsilon(k\varepsilon) = T_i(\tau) \le T_i(k\varepsilon)$, verifying that, indeed, only arms that are "lagging behind T" are pulled.

To show that T^ε is a strategy, it suffices to check that

$$T^\varepsilon((k+1)\varepsilon) \in F(T^\varepsilon(k\varepsilon)). \tag{6.4}$$

The random variable τ is a stopping time with respect to the right-continuous filtration $\{F(T(t)),\ t \ge 0\}$. Consequently, $T(\tau)$ is an F-stopping point and $F(T(\tau))$ is contained in $F(T^\varepsilon(k\varepsilon))$. By the definition of $T^\varepsilon((k+1)\varepsilon)$,

$$T^\varepsilon((k+1)\varepsilon) - T^\varepsilon(k\varepsilon) \in F(T(\tau)),$$

which verifies (6.4). To prove (6.3), observe first that

$$T(\tau) \le T^\varepsilon((k+1)\varepsilon) \le T(\tau) + (\varepsilon, \ldots, \varepsilon).$$

Hence,

$$|T^\varepsilon((k+1)\varepsilon) - T(\tau)|^2 \le d\varepsilon^2, \tag{6.5}$$

where $|\cdot|$ is the Euclidean norm. Also, since the vectors $[T^\varepsilon((k+1) - T^\varepsilon(k\varepsilon)]$ and $[T^\varepsilon(k\varepsilon) - T(\tau)]$ are orthogonal in R^d,

$$|T^\varepsilon((k+1)\varepsilon) - T(\tau)|^2 = \varepsilon^2 + |T^\varepsilon(k\varepsilon) - T(\tau)|^2. \tag{6.6}$$

By (6.5) and (6.6), we have

$$|T^\varepsilon(k\varepsilon) - T(\tau)|^2 \le (d-1)\varepsilon^2.$$

We conclude that the set of points $\{T^\varepsilon(t),\ t \ge 0\}$ is contained in the union $\bigcup_{t\ge 0} B(T(t), \varepsilon(d)^{1/2})$, where $B(c, a)$ is the Euclidean ball with center c and radius a. Finally, property 2.2.B guarantees that strategies close in space (R_+^d) must be close at all times, implying (6.3). □

6.4. Consider the setup of Section 2 (bandits and strategies).

A sequence of strategies T^n is said to converge to T uniformly on compacts (u.o.c.) if for all $\omega \in \Omega$, $\lim_{n\to\infty} T^n(t) = T(t)$ in R_+^d for all $t \ge 0$ and the convergence is uniform on compact t-sets. By the right continuity of the filtrations involved, the limit T must be a strategy as well.

THEOREM 8. *Suppose that $T^n \to T$ u.o.c. as $n \to \infty$. Then the present values* (2.2) *of T^n converge to that of T. Moreover,*

$$\lim_{n\to\infty} V(T^n) = V(T). \tag{6.7}$$

PROOF. By the boundedness of Z_i's, it suffices to show that

$$\int_0^x e^{-\beta t} Z_i(T_i^n(t))\, dT_i^n(t) \to \int_0^x e^{-\beta t} Z_i(T_i(t))\, dT_i(t) \tag{6.8}$$

for all positive x and $i = 1, \ldots, d$. Now (6.8) is a consequence of the continuity of Z_i, the fact that a.s. $Z_i(T_i^n(t))$ converges to $Z_i(T_i(t))$ uniformly on $[0, x]$ and that the variation of T_i^n over $[0, x]$ does not exceed x uniformly in n. □

An immediate implication of Theorems 7 and 8 is

COROLLARY 9.

$$\sup_{T\in\Pi^0} V(T) = \sup\left\{V(T),\ T \in \bigcup_{\varepsilon>0} \Pi^\varepsilon\right\},$$

where Π^0 is the class of all continuous strategies over R_+^d.

7. Solution to the continuous bandit problem.

7.1. The solution to the discrete d-armed bandit was outlined in 5.4.A–5.4.D. Our goal now is to establish analogous results for the continuous d-armed bandit $\{(Z_i, F_i),\ i = 1, \ldots, d\}$ described in Sections 2.1–2.5. To this end, associate with

arm i the *index process* $M_i = \{M_i(t),\ t \geq 0\}$ defined by

$$M_i(t) = \operatorname*{ess\,sup}_{\tau > t} \frac{E^{F_i(t)} \int_t^\tau e^{-\beta u} Z_i(u)\, du}{E^{F_i(t)} \int_t^\tau e^{-\beta u}\, du}, \tag{7.1}$$

where the essential supremum is over all F_i-stopping times τ, not necessarily finite, with $\tau > t$ a.s. A strategy $I = \{I(t),\ t \geq 0\}$ is an *index strategy* if it follows the leader among the index process, namely,

$$\begin{gathered} I_i(t) \text{ increases at } t \text{ only when} \\ M_i(I_i(t)) = \bigvee_{j=1}^{d} M_j(I_j(t)). \end{gathered} \tag{7.2}$$

The continuous bandit problem of Section 2.4 will be solved if we prove

7.1.A. An index strategy maximizes the value function $V(T)$ over all strategies T.

To prove 7.1.A, the value of an index strategy will be related to the index processes via the *index field* $M = \{M(s),\ s \in S\}$, which is the multiparameter process over $S = R_+^d$ defined by

$$M(s) = \bigvee_{j=1}^{d} M_j(s_j), \qquad s = (s_1, \ldots, s_d) \in S.$$

7.1.B. The value of an index strategy I satisfies

$$E \int_0^\infty e^{-\beta t} Z(I(t))\, dI(t) = E \int_0^\infty e^{-\beta t} \underline{M}(I(t))\, dt. \tag{7.3}$$

Here $\underline{M} = \{\underline{M}(s),\ s \in S\}$ is the lower envelope of M,

$$\underline{M}(s) = \inf_{0 \leq r \leq s} M(r) = \bigvee_{j=1}^{d} \underline{M}_j(s_j), \qquad r, s \in S,$$

and $\underline{M}_i = \{\underline{M}_i(t),\ t \geq 0\}$ is the lower envelope of M_i, $\underline{M}_i(t) = \inf_{0 \leq u \leq t} M_i(u)$.

To prove the optimality of an index strategy, we show that for *any* two strategies T and U,

$$V(T) \leq E \int_0^\infty e^{-\beta t} \underline{M}(U(t))\, dt. \tag{7.4}$$

Thus, any strategy that satisfies (7.3) is optimal.

For deteriorating bandits, 7.1.B holds trivially since the index processes will turn out to coincide with the reward process. In general, we are able to establish 7.1.B only under some additional assumptions, of which the most restrictive is the uniqueness of the index strategy. We also assume that the index processes have continuous sample paths in order to be able to apply the results of Section 3. It is clear, however, that 7.1.A and 7.1.B apply in much greater generality. They do hold for the deteriorating and diffusion bandits. The general solution will be specialized to these two models in Sections 7.11 and 7.12, completing details that were left out in Sections 4 and 3, respectively.

Index strategies always exist (see Theorem 12). We have shown that when the index strategy is unique,

$$\sup_T E\int_0^\infty e^{-\beta t}\sum_{i=1}^d Z_i(T_i(t))\,dT_i(t) = \sup_T E\int_0^\infty e^{-\beta t}\sum_{i=1}^d \underline{M}_i(T_i(t))\,dT_i(t)$$

and the supremum in both sides is attained by the strategy that follows the leader among the $\underline{M}_i$'s. The *general* d-armed bandit $\{(Z_i, F_i),\ i = 1,\dots,d\}$ is, thus, equivalent to the *deteriorating* d-armed bandit $\{(\underline{M}_i, F_i\},\ i = 1,\dots,d\}$. It must be the case that the equivalence holds in almost full generality.

7.2. Recall that Π^ε is the class of strategies that are discrete of order ε and define for $t \geq 0$,

$$\begin{aligned} Z_i^\varepsilon(t) &= E^{F_i(t)}\frac{1}{\varepsilon}\int_t^{t+\varepsilon} e^{-\beta(u-t)}Z_i(u)\,du \\ &= \frac{1}{\varepsilon}\int_0^\varepsilon e^{-\beta u}E^{F_i(t)}Z_i(t+u)\,du. \end{aligned} \tag{7.5}$$

PROPOSITION 10. *For* $T \in \Pi^\varepsilon$,

$$V(T) = E\sum_{k=0}^\infty e^{-\beta\varepsilon k}\sum_{i=1}^d Z_i^\varepsilon(T_i(k\varepsilon))\big[T_i((k+1)\varepsilon) - T_i(k\varepsilon)\big].$$

PROOF. By (6.1),

$$\begin{aligned} &\int_{k\varepsilon}^{(k+1)\varepsilon} e^{-\beta u}Z(T(u))\,dT(u) \\ &\quad = e^{-\beta\varepsilon k}\sum_{i=1}^d\left[\frac{1}{\varepsilon}\int_0^\varepsilon e^{-\beta u}Z_i(T_i(k\varepsilon)+u)\,du\right]\big[T_i((k+1)\varepsilon) - T_i(k\varepsilon)\big]. \end{aligned} \tag{7.6}$$

Conditioning both sides of (7.6) on the σ-field $F(T(k\varepsilon))$ and using (6.2), Proposition 10 will follow if for $i = 1,\dots,d$,

$$Z_i^\varepsilon(T_i) = \frac{1}{\varepsilon}\int_0^\varepsilon e^{-\beta u}E^{F(T)}Z_i(T_i+u)\,du, \tag{7.7}$$

whenever $T = (T_1,\dots,T_d)$ is an F-stopping *point* that takes on a finite number of values in R_+^d [just apply (7.7) to $T(k\varepsilon)$]. Using the fact that $E^{F(T)}$ coincides with $E^{F(s)}$ on the set $\{T = s\}$, it suffices to prove (7.7) for T nonrandom, which amounts to

$$E^{F(s)}Z_i(s_i+u) = E^{F_i(s_i)}Z_i(s_i+u). \tag{7.8}$$

Now (7.8) holds since Z_i is independent of $F_j(s_j)$, $j \neq i$. □

7.3. Comparing (5.1) with Proposition 10 suggests that the εth d-armed bandit (ε-bandit), which approximates the continuous bandit, be defined by

$\{(Z_i^\varepsilon, F_i^\varepsilon),\ i = 1, \dots, d\}$ with

$$Z_i^\varepsilon(k) = Z_i^\varepsilon(k\varepsilon), \qquad k = 0, 1, \dots, \tag{7.9}$$

$$F_i^\varepsilon(k) = F_i(k\varepsilon), \qquad k = 0, 1, \dots, \tag{7.10}$$

and that a discount factor $\alpha(\varepsilon) = e^{-\beta\varepsilon}$ be used.

An important remark about notation. The reader is asked to tolerate the ambiguous notation in (7.9). On the right-hand side, Z_i^ε is the continuous time process defined in (7.5), while on the left, Z_i^ε is a discrete time process. It should be clear from the context and the argument of Z_i^ε (k is an integer while $k\varepsilon$ need not be) which interpretation applies. That type of ambiguity will be used in the sequel without additional comments.

7.4. Strategies for the ε-bandit are strategies over $S = N^d$ in the sense of 5.2.A–5.2.C. Let $T^\varepsilon = \{T^\varepsilon(k),\ k = 0, 1, \dots\}$ be such a strategy. By linearly interpolating the sequence $\{\varepsilon T^\varepsilon(k),\ k = 0, 1, \dots\}$ in R_+^d, one obtains a strategy $T \in \Pi^\varepsilon$ given by

$$T(t) = \varepsilon T^\varepsilon(k) + (t - \varepsilon k)[T^\varepsilon(k+1) - T^\varepsilon(k)], \qquad k\varepsilon \le t \le (k+1)\varepsilon. \tag{7.11}$$

Conversely, for $T \in \Pi^\varepsilon$, define

$$T^\varepsilon(k) = \frac{1}{\varepsilon} T(k\varepsilon), \qquad k = 0, 1, \dots. \tag{7.12}$$

As discussed in Section 6.2, T^ε is a strategy for the ε-bandit with respect to the filtration $F^\varepsilon = \{F^\varepsilon(s),\ s \in N^d\}$ given by

$$F^\varepsilon(s) = F_1^\varepsilon(s_1) \vee \cdots \vee F_d^\varepsilon(s_d), \qquad s = (s_1, \dots, s_d) \in N^d,$$

with F_i^ε as in (7.10)

Proposition 10 relates the value $V(T)$ of a continuous strategy $T \in \Pi^\varepsilon$ with the value $V_\varepsilon(T^\varepsilon)$ of the strategy T^ε for the ε-bandit by

$$V(T) = \varepsilon V_\varepsilon(T^\varepsilon). \tag{7.13}$$

Here, T and T^ε are related by (7.11) or (7.12).

7.5. The ε-bandit is a discrete bandit as described in Section 5. Its solution involves the index processes $M_i^\varepsilon = \{M_i^\varepsilon(k),\ k = 0, 1, \dots\}$ calculated via (5.2) from (7.9) and (7.10). We now relate M_i^ε to the continuous bandit in a form analogous to (7.13). Let τ be an F_i^ε-stopping time. Short calculations lead to the relations

$$E^{F_i^\varepsilon(k)} \sum_{l=k}^{\tau-1} \alpha(\varepsilon)^l Z_i^\varepsilon(l) = \frac{1}{\varepsilon} E^{F_i(k\varepsilon)} \int_{k\varepsilon}^{\tau\varepsilon} e^{-\beta u} Z_i(u)\, du \tag{7.14}$$

and

$$E^{F_i^\varepsilon(k)} \sum_{l=k}^{\tau-1} \alpha(\varepsilon)^l = \frac{\beta}{1 - \alpha(\varepsilon)} E^{F_i(k\varepsilon)} \int_{k\varepsilon}^{\tau\varepsilon} e^{-\beta u}\, du. \tag{7.15}$$

If $k+1 \le \tau \le \infty$ a.s., the random variable $\tau\varepsilon$ is an F_i-stopping time taking values in $\{k+l\varepsilon,\ l=1,2,\dots,\infty\}$. For $t \ge 0$, define $\Gamma_i^\varepsilon(t)$ to be the set of all F_i-stopping times with values in $\{t+l\varepsilon,\ l=1,2,\dots,\infty\}$ and let

$$M_i^\varepsilon(t) = \operatorname*{ess\,max}_{\tau \in \Gamma_i^\varepsilon(t)} \frac{e^{F_i(t)}\int_t^\tau e^{-\beta u} Z_i(u)\,du}{E^{F_i(t)}\int_t^\tau e^{-\beta u}\,du}. \tag{7.16}$$

Using (7.14) and (7.15), one obtains a relation between the index $M_i^\varepsilon(k)$ of the ε-bandit and $M_i^\varepsilon(t)$ defined in (7.11), namely,

$$M_i^\varepsilon(k) = \frac{1-e^{-\beta\varepsilon}}{\beta\varepsilon} M_i^\varepsilon(k\varepsilon), \qquad k=0,1,\dots. \tag{7.17}$$

7.6. The optimal strategies for the ε-bandit are index strategies with respect to $M_i^\varepsilon = \{M_i^\varepsilon(k),\ k=0,1,\dots\}$. The relations (7.16) and (7.17) suggest that as $\varepsilon \downarrow 0$, these index strategies, properly interpolated, converge to an index strategy that satisfies (7.2). The first step is to verify that the definition (7.1) of the index processes M_i is, indeed, the proper one.

PROPOSITION 11. *Let $M_i^n(t) = M_i^\varepsilon(t)$ with $\varepsilon = 2^{-n}$. Then, for all $t \ge 0$,*

$$M_i^n(t) \le M_i^{n+1}(t) \le M_i(t) \quad a.s. \tag{7.18}$$

and

$$M_i(t) = \lim_{n\to\infty} M_i^n(t) \quad a.s. \tag{7.19}$$

PROOF. The relations (7.18) are obvious. Hence, one can define $M_i^\infty(t) = \lim_{n\to\infty} M_i^n(t)$ with $M_i^\infty(t) \le M_i(t)$ a.s. For the converse equality, let $\tau > t$ be an F_i-stopping time. Then there exists a sequence $\tau^n \in \Gamma_i^\varepsilon(t)$ decreasing to τ as $n \uparrow \infty$ and

$$M_i^n(t) \ge \frac{E^{F_i(t)}\int_t^{\tau^n} e^{-\beta u} Z_i(u)\,du}{E^{F_i(t)}\int_t^{\tau^n} e^{-\beta u}\,du} \quad \text{a.s.} \tag{7.20}$$

Applying the dominated convergence theorem to the right side of (7.20) as $n \uparrow \infty$ yields $M_i^\infty(t) \ge M_i(t)$ a.s. □

Another important remark about notation. The restriction of ε to $\varepsilon = 2^{-n}$ was convenient technically. However, using plain ε is convenient notationwise. Hence, the ε-notation will be maintained throughout with the understanding that whenever $\varepsilon \downarrow 0$, the convergence is along the sequence $\varepsilon = 2^{-n}$, $n \uparrow \infty$.

7.7. The following assumption will be enforced from now on: Versions of the index process M_i defined in (7.1) and the processes M_i^ε defined in (7.16) can be chosen so that for all $\omega \in \Omega$:

7.7.A. The sample paths of M_i are continuous.
7.7.B. The convergence in (7.19) is uniform on compact t-sets.

7.8. Let $I^\varepsilon = \{I^\varepsilon(k),\ k = 0,1,\dots\}$ be an index strategy for the ε-bandit. There is in Π^ε a strategy that relates to I^ε via (7.11). Without loss of clarity, denote that strategy by I^ε as well. One would like to conclude, according to the definition in Section 6.4 and the remark at the end of Section 7.6, that:

7.8.A. As $\varepsilon \downarrow 0$, the sequence $I^\varepsilon = \{I^\varepsilon(t),\ t \geq 0\}$ converges u.o.c. to an index strategy $I = \{I(t),\ t \geq 0\}$ satisfying (7.2).

Now fix an $\omega \in \Omega$. The family of R^d-valued continuos functions $\{I^\varepsilon(t),\ a \leq t \leq b\}$ is uniformly bounded and equicontinuous in the sup norm, for any $0 \leq a < b < \infty$. By the theorem of Arzelá–Ascoli, there exists a function $I = \{I(t),\ t \geq 0\}$ and a subsequence $\{I^\delta(t),\ t \geq 0\}$ converging to I uniformly on compact sets. Moreover, using 7.7.A and 7.7.B, one can show that I satisfies (7.2). Unfortunately, the outlined procedure does not establish 7.8.A because the definition of I depends on the sequence $\{\delta\}$, which, in turn, depends on $\omega \in \Omega$. Specifically, the $I(t)$, $t \geq 0$, that are defined for each ω individually, according to the preceding procedure, need not even be measurable.

REMARK. The proof of an important and widely used result in the optimal stopping theory of continuous time two-parameter processes (Proposition 2.1 in [24]) has the flaw of ignoring the ω-dependence described earlier. Hence, its validity without further conditions is questionable.

7.9. The approximation scheme outlined in the previous subsection is valid if it is guaranteed that (7.2) can be satisfied by *at most* one I (Corollary 13). Proposition 2 gives necessary and sufficient conditions for uniqueness when the arms deteriorate. However, we have

THEOREM 12. *If I follows the leader among the index processes M_i, then I also follows the leader among their (deteriorating) lower envelopes $\underline{M}_i$. The converse holds when I is the unique strategy that follows the leader among $\underline{M}_1, \dots, \underline{M}_d$ according to a fixed priority scheme. In particular, a strategy that follows the leader among continuous processes always exists.*

PROOF. The first part follows from the observation that if I satisfies (7.2), then

$$(7.21)\quad M_i(I_i(t)) \geq M_j(I_j(t)) \quad \text{implies that} \quad \underline{M}_i(I_i(t)) \geq \underline{M}_j(I_j(t)).$$

The second part is a consequence of the observation that if I follows the leader among the $\underline{M}_i$'s according to a fixed priority scheme, then the converse to (7.21) holds as well. The argument is based on the description of I in the proof of Proposition 3. In particular, Proposition 3 itself establishes the existence of a strategy that follows the leader among continuous processes. □

COROLLARY 13. *Suppose that*

7.9.A. *the lower envelopes $\underline{M}_1, \dots, \underline{M}_d$ are simultaneously flat with probability* 0.

Then there exists a unique strategy I which follows the leader among the M_i's and 7.8.A *holds.*

PROOF. Fix $\omega \in \Omega$. Consider the sample paths of I^ε as elements in the R^d-valued continuous functions metrized by uniform convergence on compact t-sets. By 7.7.A and 7.7.B, any limit point of the sequence I^ε must satisfy (7.2). By the arguments in Section 7.8, there is a limit point. Hence, 7.8.A holds. The uniqueness follows from the first part of Theorem 12. □

In analogy to (3.5), we state

COROLLARY 14. *If I and J are index strategies, then*

$$\bigvee_{j=1}^{d} \underline{M}_j(I_j(t)) = \bigvee_{j=1}^{d} \underline{M}_j(J_j(t)) \quad \text{for all } t \geq 0.$$

PROOF. Immediate from the first part of Theorem 12 and from (3.5). □

7.10. We are now ready to prove the optimality of the index strategy.

THEOREM 15. *If* 7.7.A, 7.7.B *and* 7.9.A *hold, then there exists a unique index strategy I which satisfies* (7.2), (7.3) *and* 7.1.A. *In other words, I is a solution to the continuous bandit problem.*

PROOF. The first step is to verify (7.4). Given any two strategies T, U, let $T^\varepsilon, U^\varepsilon \in \Pi^\varepsilon$ be strategies that approximate them according to Theorem 7. To $T^\varepsilon, U^\varepsilon$ there correspond via (7.12) strategies for the ε-bandit which we denote by T^ε and U^ε as well. By (5.5),

$$V_\varepsilon(T^\varepsilon) \leq E \sum_{k=0}^{\infty} \alpha(\varepsilon)^k \underline{M}^\varepsilon(U^\varepsilon(k)).$$

Using (7.12) and (7.17), one obtains

$$V_\varepsilon(T^\varepsilon) \leq E \sum_{k=0}^{\infty} \alpha(\varepsilon)^k \frac{1-\alpha(\varepsilon)}{\beta\varepsilon} \underline{M}^\varepsilon(U^\varepsilon(k\varepsilon))$$

or

$$\varepsilon V_\varepsilon(T^\varepsilon) \leq \frac{1-e^{-\beta\varepsilon}}{\beta\varepsilon} E \sum_{k=0}^{\infty} \varepsilon e^{-\varepsilon k} \underline{M}^\varepsilon(U^\varepsilon(k\varepsilon)). \tag{7.22}$$

As $\varepsilon \downarrow 0$, the left side of (7.22) converges to $V(T)$ by (7.13) and Theorem 8. The transformation of taking lower envelopes is continuous in the topology of uniform convergence on compact sets. Using 7.7.A, 7.7.B and (6.3), the functions $\underline{M}^\varepsilon(U^\varepsilon(t))$ converge to $\underline{M}(U(t))$ for all $t \geq 0$, as $\varepsilon \downarrow 0$. Since all functions involved are bounded, the right side of (7.22) converges to $E\int_0^\infty e^{-\beta t} \underline{M}(U(t))\, dt$, which establishes (7.4). Now, let I^ε be an index strategy for the ε-bandit. By (5.7),

$$V_\varepsilon(I^\varepsilon) = E \sum_{k=0}^{\infty} \alpha(\varepsilon)^k \underline{M}^\varepsilon(I^\varepsilon(k))$$

and from (7.17),

$$V_\varepsilon(I^\varepsilon) = \frac{1-e^{\beta\varepsilon}}{\beta\varepsilon} E \sum_{k=0}^{\infty} e^{-\varepsilon k} \underline{M}^\varepsilon(I^\varepsilon(k\varepsilon)), \tag{7.23}$$

where I^ε in (7.23) are given in 7.8.A. Using (7.8) and repeating the arguments in the first part of the proof, let $\varepsilon \downarrow 0$ in (7.23) to get (7.3) with I that is the unique index strategy from Corollary 13. Since I satisfies (7.3) and (7.4) holds, the proof is complete. □

7.11. Theorem 15 applies to the diffusion bandit. Condition 7.7.A holds: $M_i(t) = M_i(X_i(t))$ is continuous since $M_i(x)$ is. For any $\varepsilon > 0$ and $t \geq 0$, consider the sequence $X_i^{\varepsilon, t} = \{X_i(t + \varepsilon k),\ k = 0, 1, \dots\}$, which is a Markov chain with respect to the filtration $\{F_i(t + \varepsilon k),\ k = 0, 1, \dots\}$. According to 5.5.A and 5.5.B, the random variable $M_i^\varepsilon(t)$ in (7.16) has the form

$$M_i^\varepsilon(t) = M_i^\varepsilon(X_i(t))$$

and the function $M_i^\varepsilon(x)$ is lower semicontinuous. By Dini's theorem, 7.7.B holds as well. Finally, 7.9.A was proved in Proposition 5, which completes the verification of Theorem 4.

7.12. We now prove that all index strategies are optimal for the deteriorating bandit. For deteriorating bandits, the index processes coincide with the reward processes. Indeed, $M_i(t) \leq Z_i(t)$ by the monotonicity of Z_i and the converse will follow from

$$\lim_{\varepsilon \downarrow 0} \frac{E^{F_i(t)} \int_t^{t+\varepsilon} e^{-\beta u} Z_i(u)\, du}{E_i^{F_i(t)} \int_t^{t+\varepsilon} e^{-\beta u}\, du} = Z_i(t) \quad \text{a.s.} \tag{7.24}$$

Dividing both numerator and denominator by $\varepsilon \downarrow 0$, (7.24) is an immediate consequence of the right continuity of Z_i and the dominated convergence theorem.

Let I be any index strategy. Assume first that the Z_i are *deterministic*. Then 7.7.A holds trivially. As for 7.7.B,

$$M_i(t) = Z_i(t) \geq M_i^\varepsilon(t) \geq \frac{\int_t^{t+\varepsilon} e^{-\beta u} Z_i(u)\, du}{\int_t^{t+\varepsilon} e^{-\beta u}\, du} \geq Z_i(t + \varepsilon)$$

and, by Dini's theorem, 7.7.B follows. Conditions 7.7.A and 7.7.B are sufficient for (7.4) to hold. We conclude from (3.2) that when the reward processes are deterministic,

$$\int_0^\infty e^{-\beta t} Z(T(t))\, dT(t) \leq \int_0^\infty e^{-\beta t} Z(I(t))\, dI(t) \tag{7.25}$$

for every strategy T and every index strategy I. Now, if Z_i are stochastic deteriorating arms, (7.25) holds for each $\omega \in \Omega$, implying that any index strategy is optimal for the deteriorating bandit.

8. Future research.

8.1. The present work is, hopefully, a first step toward a formulation of more *general* bandit models. We envision at least three possible directions. The first is

to work within a Markovian model that will include [11] and be similar to [9]. Excursion theory should play an important role in that direction. A second direction is to work within semimartingale models. Preliminary results have been obtained for *two* arms. An interesting related process is the zigzag martingale introduced in [4], Sections 3 and 4. The third direction is a general model that will unify the discrete model [18], the continuous model of the present paper, the Poisson model [22] and the Lévy models [3], [1] and [5]. This direction will probably require a general theory of multiparameter processes analogous to the one described in [6] for usual processes. Such a multiparameter theory has been developed during recent years. However, results are mainly proved for two-parameter processes and the extension to higher dimensions is not always immediate (an example of "surprises" that can occur is described in Section 5 of [19]).

8.2. The concepts from multiparameter theory that were most useful for the present work were developed originally in order to formulate an optimal stopping problem over partially ordered sets. The multiparameter optimal stopping problem has been thoroughly investigated. Representative papers are [20] and [21], which can be consulted for further references as well. As far as I know, not even a single concrete example of an optimal stopping problem has been proposed where nontrivial switching between processes is optimal. In a private communication, Bob Vanderbei suggested a modification of the boundary value problem for zigzag martingales [4] that would produce a rather complicated such example. Joint work of the author with Larry Shepp provides a host of rather simple such examples. In these examples, which hopefully will be published in the near future, the local time behavior described in Section 4.5 plays a natural role in the description of optimal solutions.

Acknowledgments. The work presented here was initiated by Mike Harrison who has provided continuous help and support. In particular, the strong solution to the diffusion bandit was Mike's idea. I am also grateful to David Kreps and Evan Porteus for their patience and help.

REFERENCES

[1] BATHER, J. (1983). Optimal stopping of Brownian motion: A comparison technique. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday* (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.) 19–49. Academic, New York.

[2] BELLMAN, R. (1957). *Dynamic Programming*. Princeton Univ. Press, Princeton, N.J.

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

[4] BURKHOLDER, D. L. (1984). Boundary value problems and sharp inequalities for martingale transforms. *Ann. Probab.* **12** 647–702.

[5] CHERNOFF, H. (1972). *Sequential Analysis and Optimal Design*. SIAM, Philadelphia.

[6] DELLACHERIE, C. and MEYER, P. -A. (1975). *Probabilités et Potential*, 2nd ed., Chapters 1–4. Hermann, Paris.

[7] DURRETT, R. (1984). *Brownian Motion and Martingales in Analysis*. Wadsworth, Belmont, Calif.

[8] GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41** 148–177.

[9] GRIGELIONIS, B. I. and SHIRYAYEV, A. N. (1968). Controllable Markov processes and Stefan's problem. *Problems Inform. Transmission* **4** 47–57.

[10] ITÔ, K. and MCKEAN, H. P. (1965). *Diffusion Processes and Their Sample Paths*. Springer, New York.

[11] KARATZAS, I. (1984). Gittins indices in the dynamic allocation problem for diffusion processes. *Ann. Probab.* **12** 173–192.

[12] KATEHAKIS, M. N. and VEINOTT, A. F., JR. (1985). The multi-armed bandit problem: Decomposition and computation. Technical Report, Stanford Univ.

[13] KURTZ, T. G. (1980). Representations of Markov processes as multiparameter time changes. *Ann. Probab.* **8** 682–715.

[14] LIONS, P. L. (1981). Control of diffusion processes in R^N. *Comm. Pure Appl. Math.* **34** 121–147.

[15] LIPTSER, R. S. and SHIRYAYEV, A. N. (1977). *Statistics of Random Processes* **1**. Springer, New York.

[16] MAISONNEUVE, B. (1975). Exit systems. *Ann. Probab.* **3** 399–411.

[17] MANDELBAUM, A. (1985). A dynamic allocation problem between two Brownian motions with linear rewards. Unpublished.

[18] MANDELBAUM, A. (1986). Discrete multi-armed bandits and multi-parameter processes. *Probab. Theory Related Fields* **71** 129–147.

[19] MANDELBAUM, A. and VANDERBEI, R. J. (1981). Optimal stopping and supermartingales over partially ordered sets. *Z. Wahrsch. verw. Gebiete* **57** 253–264.

[20] MAZZIOTTO, G. (1982). Bi-Brownien et arret optimal sur R_+^2. *Lecture Notes in Control and Inform. Sci.* **44** 215–229. Springer, New York.

[21] MAZZIOTTO, G. (1985). Two parameter optimal stopping and bi-Markov processes. *Z. Wahrsch. verw. Gebiete* **69** 99–135.

[22] PRESMAN, É. L. and SONIN, I. M. (1983). Two and many-armed bandit problems with infinite horizon. *Lecture Notes in Math.* **1021** 526–540. Springer. New York.

[23] VARAIYA, P. P., WALRAND, J. C. and BUYUKKOC, C. (1985). Extensions of the multi-armed bandit problem. The discounted case. *IEEE Trans. Automat. Control* **AC-30** 426–439.

[24] WALSH, J. B. (1981). Optional increasing paths. *Lecture Notes in Math.* **863** 172–201. Springer, New York.

[25] WHITTLE, P. (1982). *Optimization Over Time: Dynamic Programming and Stochastic Control* **1**. Wiley, New York.

[26] WILLIAMS, D. (1979). *Diffusions, Markov Processes, and Martingales*. Wiley, New York.

GRADUATE SCHOOL OF BUSINESS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-5015