Brownian Bandits

Avi Mandelbaum *   Robert J. Vanderbei †

DIMACS
Rutgers University
New Brunswick, NJ 08903

November 8, 1993

Technical Report 93-74

Abstract

We consider a multi-armed bandit whose arms are driven by Brownian motions: the state of arm i is modeled as a one-dimensional Brownian motion B^i, $i = 1, \ldots, d$. At each instant in time, a gambler must decide to pull some subset of these d arms, holding the others fixed at their current state. If arm i is pulled when B^i is in state x_i, the gambler accumulates reward at rate $r_i(x_i)$. The goal is to find a strategy that maximizes the accumulated discounted reward over an infinite horizon (with discount rate λ). Let

$$\Gamma_i(x_i) = \sup_{\tau>0} \frac{\mathbf{E}_{x_i} \int_0^\tau e^{-\lambda t} r_i(B_t^i) dt}{\mathbf{E}_{x_i} \int_0^\tau e^{-\lambda t} dt},$$

where the supremum is over all stopping times τ of B^i. Put

$$\mathcal{M}_i = \{x \in \mathbb{R}^d : \Gamma_i(x_i) = \max_j \Gamma_j(x_j)\}.$$

From prior work, one expects that an optimal control is to pull arm i when the state of the bandit belongs to $\mathcal{M}_i$. Equivalently, the optimal strategy follows the leader among the processes $\Gamma_i(B^i)$, $i = 1, \ldots, d$. Such results have been established for bounded monotone reward functions. In this paper we extend the scope to cover certain unimodal functions. At the same time, we develop a framework within which general rewards and diffusions can be studied.

Key words: Gittins indices, Brownian motion, optimal stopping, optimal switching

AMS 1991 Subject Classification: Primary 49L05, Secondary 60J65

---

*Research partially supported by The Fund for the Promotion of Research at the Technion and AT&T Bell Labs.

†Research partially supported by the Air Force through grants AFOSR-91-0359 and AFOSR F49620-93-1-0098 and AT&T Bell Labs.

1 Introduction.

We formulate the multi-armed bandit problem within the framework of multi-parameter processes. Let $(B^i, \mathcal{F}^i)$, $i = 1, \ldots, d$, be independent one-dimensional Brownian motions. The evolution of the multi-armed bandit is described by the multi-parameter process $B_s = (B^1_{s_1}, \ldots, B^d_{s_d})$, $s = (s_1, \ldots, s_d) \in [0, \infty)^d$, which is adapted to the multi-parameter filtration $\{\mathcal{F}_s = \mathcal{F}^1_{s_1} \vee \cdots \vee \mathcal{F}^d_{s_d},\ s \geq 0\}$.

A *switching strategy* T is a family of random d-tuples,

$$T = \{T(t) = (T_1(t), \ldots, T_d(t)),\ t \geq 0\},$$

satisfying

$$\begin{aligned} &T(0) = 0, \\ &T(t) \text{ is increasing in } t\ , \\ &T_1(t) + \cdots + T_d(t) = t, \quad t \geq 0 \end{aligned} \tag{1.1}$$

and

$$\{T(t) \leq s\} \in \mathcal{F}_s, \quad t \geq 0,\ s \geq 0. \tag{1.2}$$

The random variable $T_i(t)$ represents the amount of time the ith Brownian motion has been used up to time t. The interpretation of (1.1) is that, at time t, the total allocation of time between the d processes must equal t. Condition (1.2) says that the switching strategy must be non-anticipating: in other words,

$$\{T_1(t) \leq s_1, \ldots, T_d(t) \leq s_d\} \in \mathcal{F}^1_{s_1} \vee \cdots \vee \mathcal{F}^d_{s_d}$$

and in particular, the event that no more than s_i units of time have been allocated to arm i cannot depend on information about the future of arm i beyond time s_i.

Each strategy T yields a *switched process* B_T defined as

$$B_T(t) = B_{T(t)} = (B^1_{T_1(t)}, \ldots, B^d_{T_d(t)}).$$

Associated with each Brownian motion is a running reward function r_i. The *Brownian bandit* problem is to find a strategy T^* which attains the following supremum:

$$\begin{aligned} v(x) &= \mathbf{E}_x \int_0^\infty e^{-\lambda t} r(B_{T^*}(t)) \cdot dT^*(t), \\ &= \sup_T \mathbf{E}_x \int_0^\infty e^{-\lambda t} r(B_T(t)) \cdot dT(t), \quad x \in \mathbb{R}^d. \end{aligned} \tag{1.3}$$

Here λ is a fixed positive constant, $r(x) = (r_1(x_1), \ldots, r_d(x_d))$ for $x = (x_1, \ldots, x_d)$ and $r(B_T(t)) \cdot dT(t)$ represents the inner product between the vectors $r(B_T(t))$ and $dT(t)$. The function v is called the *value function* for the multi-armed bandit problem. This problem was studied by Karatzas [Kar84], Mandelbaum [Man87] and Dalang [Dal90] as a continuous time generalization of Gittins' index theorem for Markov chains (see, e.g., [Whi82] Chapter 14). Assuming each of the r_i are bounded strictly increasing functions, it was shown that there exist *index* functions Γ_i which determine the following optimal strategy; run the process with the largest $\Gamma_i(B^i)$ (this will be made precise momentarily).

We shall establish the following formula for the index function:

$$\Gamma_i(x_i) = \sup_{\tau > 0} \frac{\mathbf{E}_{x_i} \int_0^\tau e^{-\lambda t} r_i(B_t^i) dt}{\mathbf{E}_{x_i} \int_0^\tau e^{-\lambda t} dt}, \quad x_i \in \mathbb{R}, \tag{1.4}$$

where the supremum is over all stopping times τ of B^i. We also find it convenient to work with the corresponding *index process*:

$$\hat{\Gamma}^i(t) = \Gamma_i(B_t^i), \qquad t \geq 0,$$

and its lower envelope

$$\underline{\hat{\Gamma}}^i(t) = \inf_{0 \leq u \leq t} \hat{\Gamma}^i(u), \qquad t \geq 0.$$

The result to which we aspire is:

(1.5) **Theorem** *Suppose that, for each $i = 1, \ldots, d$, the function r_i is nonnegative. Then there exists a strategy T^* which follows the leader among the indices, namely,*

$$T_i^* \textit{ increases at time } t \textit{ only when } \hat{\Gamma}^i(T_i^*(t)) = \max_j \hat{\Gamma}^j(T_j^*(t))$$

and any such strategy is optimal. Furthermore,

$$v(x) = \mathbf{E}_x \int_0^\infty e^{-\lambda t} \max_j \underline{\hat{\Gamma}}^j(T_j^*(t)) dt, \quad x \in \mathbb{R}^d.$$

Put

$$\mathcal{M}_i = \{x \in \mathbb{R}^d : \Gamma_i(x_i) = \max_j \Gamma_j(x_j)\}.$$

An alternative characterization of T^* is that

$$T_i^* \text{ increases at time } t \text{ only when } B_{T^*}(t) \in \mathcal{M}_i.$$

Later in the paper we shall impose further assumptions on the shape of the reward functions r_i and, in particular, we shall assume throughout that the reward functions are bounded and continuous.

Sections 2 through 6 are devoted to an outline of the general proof of this result. Along the way certain technical assumptions will be required. In Sections 7 and 8 we analyze certain classes of monotone and unimodal reward functions and verify the technical assumptions in those cases.

As is clear from [Man87], the optimality of Gittins' index strategies holds in great generality. In this paper we analyze certain classes of monotone and unimodal functions, but our framework seems appropriate to accomodate general rewards. However, technical issues remain open and we hope that others will be inspired to join in their resolution.

Acknowledgement: This paper was mostly prepared while the authors were visiting AT&T Bell Laboratories in Murray Hill, NJ. We would like to acknowledge the generous support of Bell Labs and especially the encouragement and support of Debasis Mitra and Larry Shepp.

2 The Dynamic Programming Equation.

The principle of dynamic programming suggests that the value function satisfies the following Hamilton-Jacobi-Bellman equation:

$$\max_i \{L_i v(x) + r_i(x_i)\} = 0, \tag{2.1}$$

where

$$L_i v(x) = \frac{1}{2}\frac{\partial^2 v}{\partial x_i^2}(x) - \lambda v(x).$$

In fact, we have:

(2.2) **Theorem** *Let w be a bounded C^2 function that satisfies (2.1). Suppose that there is a switching strategy $\tilde{T}$ for which $\tilde{T}_i(t)$ increases only when $L_i v(B_{\tilde{T}}(t)) + r_i(B^i_{\tilde{T}}(t)) = 0$. Then w is the value function v, and $\tilde{T}$ is an optimal switching strategy.*

Proof. Let $Z_t = e^{-\lambda t} w(B_T(t))$. In Section 3.1 of [MSV90] it was shown that, for any switching strategy T, the quadratic variation of B_T^i is T_i,

$$\langle B_T^i \rangle_t = T_i(t),$$

and that the quadratic covariation between B_T^i and B_T^j vanishes (for $i \neq j$). Hence, applying Ito's formula to Z_t, we get

$$\begin{aligned} Z_t - Z_0 &= \int_0^t e^{-\lambda u} \sum_i \frac{\partial w}{\partial x_i}(B_T(u)) dB_T^i(u) \\ &\quad + \int_0^t e^{-\lambda u} \sum_i L_i w(B_T(u)) dT_i(u). \end{aligned}$$

Taking expectations, letting t go to infinity and using the fact that w is bounded, we see that

$$\begin{aligned} w(x) &= -\mathbf{E}_x \int_0^\infty e^{-\lambda t} \sum_i L_i w(B_T(t)) dT_i(t) \\ &= \mathbf{E} \int_0^\infty e^{-\lambda t} r(B_T(t)) \cdot dT(t) \\ &\quad -\mathbf{E}_x \int_0^\infty e^{-\lambda t} \sum_i [L_i w(B_T(t)) + r_i(B_T^i(t))] dT_i(t). \end{aligned}$$

Now, since $L_i w + r_i \leq 0$, it follows that for any switching strategy T,

$$w(x) \geq \mathbf{E} \int_0^\infty e^{-\lambda t} r(B_T(t)) \cdot dT(t).$$

However, for the specific strategy $\tilde{T}$, we have

$$w(x) = \mathbf{E} \int_0^\infty e^{-\lambda t} r(B_{\tilde{T}}(t)) \cdot d\tilde{T}(t).$$

Hence, we see that w is the value function and $\tilde{T}$ is an optimal switching strategy. □

What remains is to exhibit a smooth solution to the Hamilton-Jacobi-Bellman equation (2.1) and to establish the existence of a switching strategy with the desired property. The construction of a solution to (2.1) depends on a certain optimal stopping problem associated with each individual arm. This optimal stopping problem is studied in the next section. Then, in section 5, the solution to these optimal stopping problems are used to exhibit a smooth solution to the Hamilton-Jacobi-Bellman equation. The construction of the optimal switching strategy and the investigation of some of its properties is described in section 6.

3 An Associated Optimal Stopping Problem.

In this section, we study the following γ-parametrized family of optimal stopping problems for a single Brownian motion with a single reward function r: find a stopping time τ^* for which

$$
\begin{aligned}
v(x,\gamma) &:= \mathbf{E}_x \int_0^{\tau^*} e^{-\lambda t}(r(B_t)-\gamma)dt \\
&= \sup_{\tau} \mathbf{E}_x \int_0^{\tau} e^{-\lambda t}(r(B_t)-\gamma)dt, \quad x \in \mathbb{R}.
\end{aligned} \tag{3.1}
$$

The function v is called the *value function* for the optimal stopping problem.

Assume for the moment that γ is fixed and consider the value function as a function of only a state variable x. The principle of dynamic programming suggests that the value function satisfies the following Hamilton-Jacobi-Bellman equation:

$$
\max\{Lv + r - \gamma, -v\} = 0,
$$

where

$$
Lv(x) = \frac{1}{2}v''(x) - \lambda v(x), \quad x \in \mathbb{R}.
$$

The solution to this nonlinear equation depends on finding an open subset D of $\mathbb{R}$ such that

$$
Lv(x) = -r(x) + \gamma \quad \text{and} \quad v(x) > 0 \quad \text{for } x \in D, \tag{3.2}
$$

and

$$
Lv(x) \leq -r(x) + \gamma \quad \text{and} \quad v(x) = 0 \quad \text{for } x \notin \bar{D}, \tag{3.3}
$$

where $\bar{D}$ denotes the closure of D. The principle of smooth fit (see, e.g., [GS68]) says that D is determined by the condition that the function v be "smooth" at the boundary of D. The following theorem makes this precise.

(3.4) **Theorem** *Suppose there exists an open set D and a function w satisfying (3.2) and (3.3). Suppose further that w is differentiable and piecewise twice continuously differentiable. Then w is the value function v and the first exit time from D is the optimal stopping time.*

Proof. Let $Z_t = e^{-\lambda t} w(B_t)$. The smoothness assumptions allow us to apply Ito's formula to Z_t, and we get

$$Z_t - Z_0 = \int_0^t e^{-\lambda u} w'(B_u) dB_u + \int_0^t e^{-\lambda u} Lw(B_u) du.$$

Now letting τ be an arbitrary stopping time and taking expectations, we get Dynkin's formula

$$\mathbf{E}_x e^{-\lambda\tau} w(B_\tau) - w(x) = \mathbf{E}_x \int_0^\tau e^{-\lambda t} Lw(B_t) dt.$$

Since $w \geq 0$ and $Lw \leq -r + \gamma$,

$$w(x) \geq \mathbf{E}_x \int_0^\tau e^{-\lambda t} (r(B_t) - \gamma) dt.$$

However, for the first exit time $\tilde{\tau}$ from D, $w(B_{\tilde{\tau}}) = 0$ and $Lw(B_t) = r(B_t) - \gamma$ for $t \leq \tilde{\tau}$ and so the above inequality is actually an equality. Hence, $\tilde{\tau}$ is the optimal stopping time and $w(x)$ is the value function. □

Theorem (3.4) reduces the optimal stopping problem to a problem in analysis. Indeed, we guess a form for D depending on a few undetermined parameters, solve

$$\begin{aligned} Lw(x) &= -r(x) + \gamma \quad \text{for } x \in D, \\ w(x) &= 0 \qquad\qquad\quad \text{for } x \notin D, \end{aligned}$$

and then apply the principle of smooth fit to impose the desired level of smoothness in w (from Theorem (3.4)). This process reveals the exact values of the parameters which determine D and hence the solution to the problem. In the following sections, we carry out this program for reward functions that are monotone or have at most one critical point.

Now we consider the dependence of the value function on the parameter γ. Put

(3.5) $$\Gamma(x) = \inf\{\gamma : v(x, \gamma) = 0\}.$$

The functions v and Γ possess the following properties.

(3.6) **Theorem** *For each x,*

1. $-1/\lambda \leq \partial^+ v / \partial\gamma \leq 0$,

2. $\Gamma(x) < \infty$,

3. $v(x,0) < \infty$,

4. $\partial^+ v/\partial\gamma(x,0^-) = -1/\lambda$.

Proof. (1) Note that the "supand" in (3.1) is linear in γ:

$$\mathbf{E}_x \int_0^\tau e^{-\lambda t}(r(B_t)-\gamma)dt = a(x,\tau) - b(x,\tau)\gamma,$$

where

$$b(x,\tau) = \mathbf{E}_x \int_0^\tau e^{-\lambda t} dt.$$

Clearly, $0 \le b(x,\tau) \le 1/\lambda$. So, $v(x)$ is a supremum of linear functions each of whose slope lies in $[-1/\lambda, 0]$. Hence, $\partial^+ v/\partial\gamma$ exists and satisfies (1).

(2) Suppose that r is bounded by M. For $\gamma \ge M$, $r(B_t)-\gamma \le 0$ and so $\tau = 0$ is optimal which means that $v(x,\gamma) = 0$ for $\gamma \ge M$.

(3) Follows from (1) and (2).

(4) For $\gamma < 0$, $r(B_t) - \gamma > 0$ and so $\tau = \infty$ is optimal which means that

$$\begin{aligned} v(x,\gamma) &= \mathbf{E}_x \int_0^\infty e^{-\lambda t}(r(B_t)-\gamma)dt \\ &= \mathbf{E}_x \int_0^\infty e^{-\lambda t} r(B_t)dt - \frac{\gamma}{\lambda}. \end{aligned}$$

Hence, for $\gamma < 0$, $\partial^+ v/\partial\gamma(x,\gamma) = -1/\lambda$ and (4) follows. □

4 Alternate Formulas for the Index.

In Theorem 1.4, we aspired for a representation of the value function v in terms of the indices. The representation applies also to a single arm, in which case it takes the form

$$\mathbf{E}_{x_i} \int_0^\infty e^{-\lambda t} r_i(B_t^i)dt = \mathbf{E}_{x_i} \int_0^\infty e^{-\lambda t} \inf_{0\le u\le t} \Gamma_i(B_u^i)dt, \quad x_i \in \mathbb{R}.$$

This can be shown to imply that

$$\mathbf{E}_{x_i} \int_0^{\tau_\epsilon^i} e^{-\lambda t} r_i(B_t^i)dt = \mathbf{E}_{x_i} \int_0^{\tau_\epsilon^i} e^{-\lambda t} \inf_{0\le u\le t} \Gamma_i(B_u^i)du, \tag{4.1}$$

where

$$\tau_\epsilon^i = \inf\{t > 0 : \Gamma_i(B_t^i) \le \Gamma_i(B_0^i) - \epsilon\} \ , \qquad \epsilon > 0.$$

Dividing both sides of (4.1) by $\mathbf{E}_{x_i} \int_0^{\tau_\epsilon^i} e^{-\lambda t} dt$, letting $\epsilon \downarrow 0$, and observing that $\inf \Gamma_i(B^i)$ is trapped within $[\Gamma_i(B_0^i) - \epsilon, \, \Gamma_i(B_0^i)]$ up to time τ_ϵ^i, we get that

$$\Gamma_i(x_i) = \lim_{\epsilon \downarrow 0} \ \frac{\mathbf{E}_{x_i} \int_0^{\tau_\epsilon^i} e^{-\lambda t} r_i(B_t^i) dt}{\mathbf{E}_{x_i} \int_0^{\tau_\epsilon^i} e^{-\lambda t} dt} \ , \qquad x_i \in \mathbb{R}.$$

The relation

$$\Gamma_i(x_i) \ge \frac{\mathbf{E}_{x_i} \int_0^{\tau} e^{-\lambda t} r_i(B_t^i) dt}{\mathbf{E}_{x_i} \int_0^{\tau} e^{-\lambda t} dt} \ ,$$

for any stopping time τ of B^i, is an immediate consequence of the definition of Γ_i in (3.5). One now deduces the representation (1.4) of Γ_i.

5 Solution of the Hamilton-Jacobi-Bellman Equation.

In this section, we construct a smooth solution to the Hamilton-Jacobi-Bellman equation. For each i, consider a related optimal stopping problem:

$$v_i(x_i, \gamma) := \sup_\tau \mathbf{E}_{x_i} \int_0^\tau e^{-\lambda t}(r_i(B_t^i) - \gamma) dt.$$

Each of these value functions satisfy all the properties given in Theorem (3.6) as well as the dynamic programming equation:

$$\max\{-v_i, L_i v_i + r_i - \gamma\} = 0.$$

Associated with each of these value functions is an index function defined by (3.5). Hence, let

$$\Gamma_i(x_i) = \inf\{\gamma : v_i(x_i, \gamma) = 0\}.$$

(5.1) **Theorem** *Suppose that for each i,*

- Γ_i *is continuous and*

- *for each γ, $\partial^+ v_i / \partial \gamma$ is C^1 in x_i on $\{x_i : \Gamma_i(x_i) > \gamma\}$.*

Then the function

$$w(x) = \frac{1}{\lambda}\int_0^\infty \left[1 - \prod_i \left(\lambda \frac{\partial^+ v_i}{\partial\gamma}(x_i,\gamma) + 1\right)\right] d\gamma$$

is bounded, C^2, and satisfies the Hamilton-Jacobi-Bellman equation (2.1).

Proof. We start by showing that w is C^2. To this end, let

$$\xi_i(x_i,\gamma) = \lambda \frac{\partial^+ v_i}{\partial\gamma}(x_i,\gamma) + 1$$

and put

$$\begin{aligned} w_{ij}(x_1,\ldots,x_d) &= \int_0^{\Gamma_i(x_i)} (1 - \xi_i(x_i,\gamma)\xi_j(x_j,\gamma)\Pi_{k\neq i,j}\xi_k(x_k,\gamma))\, d\gamma \\ &+ \int_{\Gamma_i(x_i)}^{\Gamma_j(x_j)} (1 - \xi_j(x_j,\gamma)\Pi_{k\neq i,j}\xi_k(x_k,\gamma))\, d\gamma \\ &+ \int_{\Gamma_j(x_j)}^{\infty} (1 - \Pi_{k\neq i,j}\xi_k(x_k,\gamma))\, d\gamma. \end{aligned}$$

Note that $\xi_i(x_i,\Gamma_i(x_i)) = 1$ and that the function w is related to w_{ij} via

$$w(x) = \frac{1}{\lambda} w_{ij}(x) \quad \text{ on } \{x : \Gamma_i(x_i) \le \Gamma_j(x_j)\}.$$

To establish the claimed smoothness of w it suffices to show that w_{ij} and w_{ji} patch together smoothly on $\{x : \Gamma_i(x_i) = \Gamma_j(x_j)\}$. Since this set has codimension one and $w_{ij} = w_{ji}$ on this set, it suffices to check any one of the first derivatives and any one of the second derivatives (since if these agree, then normal derivatives agree and that is all that is needed). For first derivatives we have

$$\frac{\partial w_{ij}}{\partial x_i} = -\int_0^{\Gamma_i(x_i)} \xi_i'\ \xi_j\ \Pi_{k\neq i,j}\xi_k\ d\gamma$$

and

$$\frac{\partial w_{ji}}{\partial x_i} = -\int_0^{\Gamma_j(x_j)} \xi_i'\ \xi_j\ \Pi_{k\neq i,j}\xi_k\ d\gamma - \int_{\Gamma_j(x_j)}^{\Gamma_i(x_i)} \xi_i'\ \Pi_{k\neq i,j}\xi_k\ d\gamma,$$

where we have supressed the arguments of the ξ's and the primes on the ξ's denote derivatives with respect to the x variable. Evaluating at $\Gamma_i(x_i) = \Gamma_j(x_j)$, we get

$$\begin{aligned} \left.\frac{\partial w_{ij}}{\partial x_i}\right|_{\Gamma_i(x_i)=\Gamma_j(x_j)=\Gamma} &= -\int_0^{\Gamma} \xi_i'\ \xi_j\ \Pi_{k\neq i,j}\xi_k\ d\gamma \\ &= \left.\frac{\partial w_{ji}}{\partial x_i}\right|_{\Gamma_i(x_i)=\Gamma_j(x_j)=\Gamma}. \end{aligned}$$

Similarly, for the mixed partial derivative, we get

$$\frac{\partial^2 w_{ij}}{\partial x_i \partial x_j} = -\int_0^{\Gamma_i(x_i)} \xi_i' \; \xi_j' \; \Pi_{k\neq i,j}\xi_k \; d\gamma$$

and

$$\frac{\partial^2 w_{ji}}{\partial x_i \partial x_j} = -\int_0^{\Gamma_j(x_j)} \xi_i' \; \xi_j' \; \Pi_{k\neq i,j}\xi_k \; d\gamma.$$

Evaluating at $\Gamma_i(x_i) = \Gamma_j(x_j)$, we get

$$\begin{aligned} \left.\frac{\partial^2 w_{ij}}{\partial x_i^2}\right|_{\Gamma_i(x_i)=\Gamma_j(x_j)=\Gamma} &= -\int_0^{\Gamma} \xi_i' \; \xi_j' \; \Pi_{k\neq i,j}\xi_k \; d\gamma \\ &= \left.\frac{\partial^2 w_{ji}}{\partial x_i^2}\right|_{\Gamma_i(x_i)=\Gamma_j(x_j)=\Gamma}. \end{aligned}$$

Hence, w is C^2.

Let M denote a bound for all of the reward functions. For $\gamma \geq M$, the integrand vanishes and so we can write

$$w(x) = \frac{1}{\lambda}\int_0^M \left[1 - \left(\lambda \frac{\partial^+ v_j}{\partial \gamma}(x_j,\gamma) + 1\right) F_j(x,\gamma)\right] d\gamma, \tag{5.2}$$

where

$$F_j(x,\gamma) = \prod_{i\neq j}(\lambda \frac{\partial^+ v_i}{\partial \gamma}(x_i,\gamma) + 1).$$

Note that each factor in this product has all the properties of a cumulative distribution function in γ and hence so does F_j. Writing the integral of the difference in (5.2) as the difference of the integrals and integrating by parts the subtracted integral, we get

$$\begin{aligned} w(x) = \frac{1}{\lambda}\Big\{ M \quad &- \quad \Big[(\lambda v_j(x_j,\gamma) + \gamma)F_j(x,\gamma)\Big]_{\gamma=0^-}^{\gamma=M} \\ &+ \quad \int_{0^-}^M (\lambda v_j(x_j,\gamma) + \gamma)F_j(x,d\gamma) \Big\}. \end{aligned}$$

Since $F_j(x,0^-) = 0$, $F_j(x,M) = 1$, and $v_j(x_j,M) = 0$, it follows that the nonintegral terms cancel each other out and we get

$$w(x) = \frac{1}{\lambda}\int_{0^-}^{\infty} (\lambda v_j(x_j,\gamma) + \gamma)F_j(x,d\gamma).$$

From this formula, we see that w is C^1 in x_j and piecewise C^2 in x_j. We may thus apply the operator L_j to w to get

$$L_j w(x) = \int_{0^-}^{\infty} (L_j v_j(x_j,\gamma) - \gamma)F_j(x,d\gamma). \tag{5.3}$$

Now, the measure $F_j(x,\cdot)$ is supported on the interval $[0, \max_{i\neq j}\Gamma_i(x_i)]$ and the integrand in (5.3) is dominated by $-r_j(x_j)$ with the domination being an equation when $\gamma \leq \Gamma_j(x_j)$. Hence,

$$L_j w(x) \leq -r_j(x_j)$$

and the inequality is actually an equality when

$$\Gamma_j(x_j) \geq \max_{i\neq j}\Gamma_i(x_i).$$

Since, for any x, there must exist some j for which $\Gamma_j(x_j)$ dominates all the other $\Gamma_i(x_i)$'s, it follows that w indeed satisfies the dynamic programming equation. □

6 The Optimal Switching Strategy.

The final comment in the proof of Theorem (5.1) shows that

$$\mathcal{M}_i = \{x \in \mathbb{R}^d : \Gamma_i(x_i) = \max_j \Gamma_j(x_j)\} = \{x \in \mathbb{R}^d : L_i v(x) + r_i(x_i) = 0\}.$$

Hence, to construct a switching strategy, as described in Theorem (2.2), it suffices to construct a switching strategy that follows the leader among the index processes $\hat{\Gamma}_i$. The existence of such a strategy was established in [Man87], for continuous Γ_i. (An alternative construction for $d = 2$ was described in [MSV90].) An extension to $\hat{\Gamma}_i$ whose sample paths are right-continuous with left limits is carried out in [KM93].

Switching among Brownian motions is naturally quantified in terms of local times. To see that, consider for example a multi-armed bandit whose reward functions are monotone increasing and equal to each other. Then an optimal strategy T^* simply follows the leader among $B_i,\ i = 1,\ldots,d$. Assume for simplicity of exposition that $B_i(0) = B_j(0)$ for all i,j. Then, as explained in [Man87], the strategy T^* is uniquely determined by the relations

$$\underline{B}^1_{T^*_1}(t) = \cdots = \underline{B}^d_{T^*_d}(t)\ , \quad t \geq 0.$$

Introduce the *excursion* process $\xi = (\xi^1,\ldots,\xi^d)$ by

$$\xi^i_t = B^i_{T^*_i}(t) - \underline{B}^i_{T^*_i}(t)\ , \quad t \geq 0.$$

One can then show that the process B_{T^*} has the pathwise decomposition

$$B_{T^*}(t) = \xi_t - \frac{1}{d} L_t \cdot e \ , \quad t \geq 0,$$

where $L = \{L_t,\ t \geq 0\}$ is the local time of ξ at 0, and e is the d-dimensional vector of 1's. Equivalently,

$$L_t = \lim_{\epsilon \downarrow 0} \frac{1}{2\epsilon} \int_0^t 1_{[0,\epsilon e]}(\xi_u) du \ , \quad t \geq 0,$$

where the limit is taken almost surely. In other words, B_{T^*} can be described as evolving down along the diagonal $\{x \in \mathbb{R}^d : x_1 = \cdots = x_d\}$, while performing excursions in parallel to the half lines $\{x \in \mathbb{R}^d : x_i \geq 0,\ x_j = 0 \text{ for } j \neq i\},\ i = 1, \ldots, d$.

7 Monotonic Rewards.

Suppose first that r is continuous and strictly increasing. This suggests that the solution to the optimal stopping problem is the first hitting time of an interval $(-\infty, b)$. The value function then takes the form:

$$v(x,\gamma) = \begin{cases} \int_0^\infty g_\lambda(x-b,y)(r(y+b)-\gamma)dy & x \geq b \\ 0 & x \leq b \end{cases} \tag{7.1}$$

where

$$g_\lambda(x,y) = \frac{1}{\sqrt{2\lambda}}(e^{-\sqrt{2\lambda}\,|y-x|} - e^{-\sqrt{2\lambda}\,(y+x)}). \tag{7.2}$$

Note that the function g_λ is the Green's kernel for Brownian motion on $(0,\infty)$. Viewed as a function of x, it is the unique bounded solution to

$$\frac{1}{2}g'' - \lambda g = -\delta_y, \quad g(0) = 0.$$

Differentiating (7.1) with respect to x, we get

$$v'(x,\gamma) = \begin{cases} \int_0^\infty g'_\lambda(x-b,y)(r(y+b)-\gamma)dy & x \geq b \\ 0 & x \leq b. \end{cases}$$

The principle of smooth fit now says that we should equate the left and right derivatives at b:

$$\int_0^\infty e^{-\sqrt{2\lambda}\,y}(r(y+b)-\gamma)dy = 0. \tag{7.3}$$

Put

$$\Gamma(x) = \sqrt{2\lambda} \int_0^\infty e^{-\sqrt{2\lambda}\, y} r(x+y)dy. \tag{7.4}$$

Then (7.3) becomes

$$\Gamma(b) = \gamma. \tag{7.5}$$

Since r is strictly increasing, it follows that Γ is as well. Hence, we can invert to solve for b, namely

$$b = \Gamma^{-1}(\gamma).$$

We now identify the function Γ defined in (7.4) as Gittins' index (3.5). First, $v(x,\gamma) = 0$ if and only if $x \le b = \Gamma^{-1}(\gamma)$, which is equivalent to the condition that $\Gamma(x) \le \gamma$. Thus, $\inf\{\gamma : v(x,\gamma) = 0\} = \Gamma(x)$.

Finally, in order to apply Theorem (5.1), we need to show that Γ is continuous and that for each γ $\partial^+ v/\partial\gamma$ is differentiable in x on $\{x : \Gamma(x) > \gamma\}$. It follows from the continuity of r and (7.4) that Γ is continuous. In fact, it is differentiable and

$$\Gamma'(x) = \sqrt{2\lambda}\,(\Gamma(x) - r(x)).$$

It follows from (7.1) that, for $\gamma < \Gamma(x)$,

$$\begin{aligned} \frac{\partial v}{\partial \gamma} &= -\int_0^\infty g'_\lambda(x-b,y)\, b' \, (r(y+b)-\gamma)\, dy \\ &\quad + \int_0^\infty g_\lambda(x-b,y)\,(r'(y+b)\, b' - 1)\, dy \end{aligned} \tag{7.6}$$

where $b = \Gamma^{-1}(\gamma)$ and

$$b' = \Gamma^{-1\prime}(\gamma) = \frac{1}{\Gamma'(\Gamma^{-1}(\gamma))} = \frac{1}{\Gamma'(b)} = \frac{1}{\sqrt{2\lambda}\,(\gamma - r(b))}.$$

Substituting the explicit expression (7.2) for g_λ into (7.6) and integrating by parts in the integral involving r' and simplifying, we get

$$\begin{aligned} \frac{\partial v}{\partial \gamma} &= -\,2b' \int_0^\infty e^{-\sqrt{2\lambda}(y+x-b)} r(y+b)dy \\ &\quad + \gamma b' \int_0^\infty g'_\lambda(x-b,y)dy \\ &\quad - \int_0^\infty g_\lambda(x-b,y)dy. \end{aligned} \tag{7.7}$$

From (7.4) and (7.5), we see that the first integral simplifies to

$$\int_0^\infty e^{-\sqrt{2\lambda}(y+x-b)} r(y+b)dy = e^{-\sqrt{2\lambda}(x-b)} \frac{\Gamma(b)}{\sqrt{2\lambda}} = e^{-\sqrt{2\lambda}(x-b)} \frac{\gamma}{\sqrt{2\lambda}}.$$

Also, explicit calculations reveal that

$$\int_0^\infty g'_\lambda(x-b,y)dy = \sqrt{\tfrac{2}{\lambda}} e^{-\sqrt{2\lambda}(x-b)}$$

and

$$\int_0^\infty g_\lambda(x-b,y)dy = \tfrac{1}{\lambda}\left(1 - e^{-\sqrt{2\lambda}(x-b)}\right).$$

Substituting these three integration formulas into (7.7) and simplifying we finally discover that, for $\gamma \le \Gamma(x)$,

$$\frac{\partial v}{\partial \gamma} = \tfrac{1}{\lambda}\left(e^{-\sqrt{2\lambda}(x-b)} - 1\right).$$

This expression is clearly continuously differentiable in x.

We end this section by reminding the reader that the above formulas were derived under the assumption that r is strictly increasing. Had it been strictly decreasing we would have obtained similar formulas. Indeed, if we let

$$\Gamma_+(x) = \sqrt{2\lambda} \int_0^\infty e^{-\sqrt{2\lambda}\, y} r(x+y)dy \tag{7.8}$$

$$\Gamma_-(x) = \sqrt{2\lambda} \int_0^\infty e^{-\sqrt{2\lambda}\, y} r(x-y)dy, \tag{7.9}$$

then Γ_+ is the index function for strictly increasing rewards and Γ_- is the index function for strictly decreasing ones.

8 Bitonic Rewards.

In this section we consider reward functions that have one critical point. We call such functions *bitonic*. The critical point of a bitonic function can be either a maximum or a minimum point of the function. If it is a minimum, we say that it is a *DI* function (decreasing, then increasing). Otherwise, it is an *ID* function.

DI Rewards.

Suppose now that r is first strictly decreasing and then strictly increasing. As always, we assume that r is bounded and therefore has limits at plus and minus infinity. Let m denote the point at which r is minimized. To solve the optimal stopping problem, it makes sense to postulate that the optimal stopping time is the first hitting time of an interval (a, b) (typically positioned near m). The value function then takes the form:

$$v(x,\gamma) = \begin{cases} \int_0^\infty g_\lambda(x-b,y)(r(y+b)-\gamma)dy & x \geq b \\ \int_0^\infty g_\lambda(a-x,y)(r(a-y)-\gamma)dy & x \leq a \\ 0 & a \leq x \leq b, \end{cases} \tag{8.1}$$

with g_λ defined as in (7.2). Differentiating (8.1) with respect to x, we get

$$v'(x,\gamma) = \begin{cases} \int_0^\infty g'_\lambda(x-b,y)(r(y+b)-\gamma)dy & x \geq b \\ -\int_0^\infty g'_\lambda(a-x,y)(r(a-y)-\gamma)dy & x \leq a \\ 0 & a \leq x \leq b \end{cases}.$$

The principle of smooth fit now says that we should equate the left and right derivatives at points a and b:

$$\int_0^\infty e^{-\sqrt{2\lambda}\,y}(r(y+b)-\gamma)dy = 0, \tag{8.2}$$

and

$$\int_0^\infty e^{-\sqrt{2\lambda}\,y}(r(a-y)-\gamma)dy = 0. \tag{8.3}$$

In terms of the functions Γ_+ and Γ_- defined in (7.8) and (7.9), equations (8.2) and (8.3) become

$$\Gamma_+(b) = \gamma, \quad \Gamma_-(a) = \gamma.$$

If Γ_+ and Γ_- were monotone, then we could invert to solve for a and b, namely

$$b = \Gamma_+^{-1}(\gamma), \quad a = \Gamma_-^{-1}(\gamma).$$

However, they need not be monotone everywhere so we now investigate the regions in which Γ_+ is increasing and Γ_- is decreasing. Differentiating and then integrating by parts, we get

$$\begin{aligned} \Gamma'_+(x) &= \sqrt{2\lambda}\int_0^\infty e^{-\sqrt{2\lambda}\,y} r'(x+y)dy \\ &= \sqrt{2\lambda}\,(\Gamma_+(x) - r(x)) \end{aligned} \tag{8.4}$$

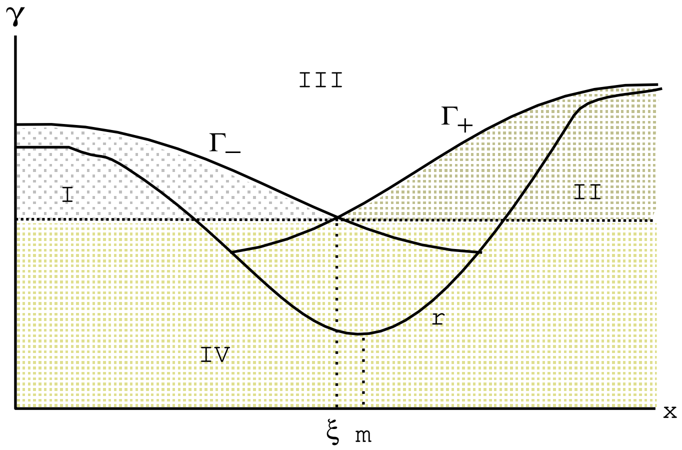

Figure 1: The functions r, Γ_+ and Γ_-.

and

$$\begin{aligned} \Gamma_-'(x) &= \sqrt{2\lambda}\int_0^\infty e^{-\sqrt{2\lambda}\,y} r'(x-y)dy \\ &= -\sqrt{2\lambda}\,(\Gamma_-(x) - r(x)) \end{aligned}$$

Without loss of generality, we will assume that the limit at plus infinity is at least as large as the limit at minus infinity (see Figure (1)). This then implies that Γ_- has at least one critical point. We denote by x^- its leftmost critical point.

There are two cases to consider depending on whether Γ_+ has critical points or not. First, suppose that it does. Let x^+ denote the rightmost critical point of Γ_+. It follows from (8.4) that $x^+ < m$ and that

$$\Gamma_+(x^+) = r(x^+) \quad \text{and} \quad \Gamma_+(x) > r(x) \quad \text{for } x > x^+.$$

Similarly, $x^- > m$ and

$$\Gamma_-(x^-) = r(x^-) \quad \text{and} \quad \Gamma_-(x) > r(x) \quad \text{for } x < x^-.$$

Hence there exists a point ξ between x^+ and x^- at which Γ_+ and Γ_- intersect (see Figure (1)).

Now suppose that Γ_+ does not have any critical points. Then for $x < x^-$ $\Gamma_-(x) \leq r(-\infty) = \Gamma_+(-\infty) < \Gamma_+(x)$. Hence, in this case Γ_+ and Γ_- have no intersection: they "intersect" at minus infinity. It will be seen that this case reduces to the monotone increasing case treated earlier.

We now identify the function given by

$$\Gamma(x) = \Gamma_-(x) \vee \Gamma_+(x)$$

as Gittins' index. First, $v(x,\gamma) = 0$ if and only if $\Gamma_-^{-1}(\gamma) = a \leq x \leq b = \Gamma_+^{-1}(\gamma)$, which is equivalent to $\Gamma_-(x) \leq \gamma$ and $\Gamma_+(x) \leq \gamma$, that is $\Gamma(x) \leq \gamma$. Thus, $\inf\{\gamma : v(x,\gamma) = 0\} = \Gamma(x)$, and we are done.

Finally, in order to apply Theorem (5.1), we need to show that Γ is continuous and that, for each γ, $\partial^+ v/\partial\gamma$ is differentiable in x on $\{x : \Gamma(x) > \gamma\}$. Since Γ_+ and Γ_- are both continuous (in fact differentiable), it follows that Γ is continuous.

Arguments analogous to those given at the end of Section 7 show that

$$\frac{\partial v}{\partial \gamma} = \begin{cases} \frac{1}{\lambda}\left(e^{-\sqrt{2\lambda}(\Gamma_-^{-1}(\gamma)-x)} - 1\right) & \text{for } (x,\gamma) \text{ in region I,} \\ \frac{1}{\lambda}\left(e^{-\sqrt{2\lambda}(x-\Gamma_+^{-1}(\gamma))} - 1\right) & \text{for } (x,\gamma) \text{ in region II.} \end{cases}$$

For (x,γ) in region IV, $\tau^* = \infty$ and so

$$\begin{aligned} v(x,\gamma) &= \mathbf{E}\int_0^\infty e^{-\lambda t}\left(r(B_t) - \gamma\right) dt \\ &= \mathbf{E}\int_0^\infty e^{-\lambda t} r(B_t) dt - \frac{\gamma}{\lambda}. \end{aligned}$$

From this it follows that, for (x,γ) in region IV,

$$\frac{\partial v}{\partial \gamma} = -\frac{1}{\lambda}.$$

Hence, from the formulas for $\partial v/\partial\gamma$ for (x,γ) in regions I, II, and IV, it is clear that $\partial v/\partial\gamma$ is C^2 in x on $\{x : \Gamma(x) > \gamma\}$.

ID Rewards.

Suppose now that r is first strictly increasing and then strictly decreasing. As before, r is bounded and so has limits at plus and minus infinity. Let m denote the point at which r is maximized. To solve the optimal stopping problem, we postulate that the optimal stopping time is the first exit time from an interval (a, b) (typically positioned near m). The value function then takes the form:

$$v(x,\gamma) = \begin{cases} \int_a^b g_\lambda(x,y)(r(y)-\gamma)dy & a \le x \le b \\ 0 & x \le a \text{ or } x \ge b \end{cases} \tag{8.5}$$

where[1]

$$g_\lambda(x,y) = \frac{\cosh\sqrt{2\lambda}\,(b-a-|x-y|) - \cosh\sqrt{2\lambda}\,(b+a-(x+y))}{\sqrt{2\lambda}\,\sinh\sqrt{2\lambda}\,(b-a)}.$$

Note that the function g_λ is the Green's kernel for Brownian motion on (a, b). Analytically, g_λ is the unique bounded solution to

$$\frac{1}{2}g'' - \lambda g = -\delta_y, \quad g(a) = g(b) = 0.$$

Differentiating (8.5) with respect to x, we get

$$v'(x,\gamma) = \begin{cases} \int_a^b g'_\lambda(x,y)(r(y)-\gamma)dy & a \le x \le b \\ 0 & x \le a \text{ or } x \ge b \end{cases}$$

The principle of smooth fit now says that we should equate the left and right derivatives at points a and b:

$$v'(a,\gamma) = 2\int_a^b \frac{\sinh\sqrt{2\lambda}\,(b-y)}{\sinh\sqrt{2\lambda}\,(b-a)}(r(y)-\gamma)dy = 0$$

and

$$v'(b,\gamma) = -2\int_a^b \frac{\sinh\sqrt{2\lambda}\,(y-a)}{\sinh\sqrt{2\lambda}\,(b-a)}(r(y)-\gamma)dy = 0.$$

Simplifying slightly, we get

$$\int_a^b \sinh\sqrt{2\lambda}\,(b-y)r(y)dy = \gamma\int_a^b \sinh\sqrt{2\lambda}\,(b-y)dy \tag{8.6}$$

and

$$\int_a^b \sinh\sqrt{2\lambda}\,(y-a)r(y)dy = \gamma\int_a^b \sinh\sqrt{2\lambda}\,(y-a)dy \tag{8.7}$$

We would like to manipulate (8.6) and (8.7) in such a way as to isolate a in one of the two equations and isolate b in the other one. Then we could write γ as a function of a from the

[1] For notational convenience, we omit the parentheses surrounding arguments of hyperbolic functions. For example, $\cosh\sqrt{2\lambda}\,(b+a-(x+y))$ stands for $\cosh(\sqrt{2\lambda}\,(b+a-(x+y)))$

one equation and γ as a function of b from the other one and putting these two formulas together appropriately would yield the Gittins index function. However, it appears that this cannot be done explicitly without making further simplifying assumptions. Hence, we assume now that r is symmetric about its minimum point which we take without loss of generality to be the origin. Then symmetry dictates that $a = b$ and both equations reduce to a single equation:

$$\int_{-b}^{b} \sinh \sqrt{2\lambda}\,(b-y)r(y)dy = \gamma \int_{-b}^{b} \sinh \sqrt{2\lambda}\,(b-y)dy.$$

Therefore, for a symmetric reward function, the index function Γ is given by

$$\Gamma(x) = \frac{\int_{-x}^{x} \sinh \sqrt{2\lambda}\,(x-y)r(y)dy}{\int_{-x}^{x} \sinh \sqrt{2\lambda}\,(x-y)dy},$$

for $x > 0$, and its value for $x < 0$ is determined by symmetry:

$$\Gamma(x) = \Gamma(-x).$$

Finally, it value at zero is determined by continuity:

$$\Gamma(0) = r(0).$$

It is tedious but straightforward to show that Γ is strictly increasing for $x \geq 0$. In order to apply Theorem (5.1) we need to show that $\partial^+ v/\partial\gamma$ is differentiable in x on $\{x : \Gamma(x) > \gamma\}$. This too is tedious but straightforward and hence omitted.

References

[Dal90] R.C. Dalang. Randomization in the two-armed bandit problem. *Ann. Prob.*, 18:218–225, 1990.

[GS68] B.I. Grigelionis and A.N. Shiryaev. Controllable Markov processes and Stefan's problem. *Problemy Peredachi Informatsii*, 4:60–72, 1968.

[Kar84] I. Karatzas. Gittins indices in the dynamic allocation problem for diffusion processes. *Ann. Prob.*, 12:173–192, 1984.

[KM93] H. Kaspi and A. Mandelbaum. Levy bandits: Multi-armed bandits driven by levy processes, 1993. In preparation.

[Man87] A. Mandelbaum. Continuous multi-armed bandits and multi-parameter processes. *Ann. Prob.*, 15:1527–1556, 1987.

[MSV90] A. Mandelbaum, L.A. Shepp, and R.J. Vanderbei. Optimal switching between a pair of Brownian motions. *Ann. Prob.*, 18:1010–1033, 1990.

[Whi82] P. Whittle. *Optimization over Time: Dynamic Programming and Stochastic Control.* Wiley, 1982.