ARTICLE IN PRESS

PERGAMON

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Automatica ■■■ (■■■■) ■■■–■■■

automatica

www.elsevier.com/locate/automatica

Optimal hysteresis for a class of deterministic deteriorating two-armed Bandit problem with switching costs[☆]

F. Dusonchet*,[1] , M.-O. Hongler[2]

EPFL-DMT-IPM, Laboratoire de Production Microtechnique (LPM), Institut de Production et Robotique (IPR), CH-1015 Lausanne, Switzerland

Received 21 September 2001; received in revised form 6 May 2003; accepted 4 June 2003

Abstract

We derive the optimal policy for the dynamic scheduling of a class of deterministic, deteriorating, continuous time and continuous state two-armed Bandit problems with switching costs. Due to the presence of switching costs, the scheduling policy exhibits an hysteretic character. Using this exactly solvable class of models, we are able to explicitly observe the performance of a sub-optimal policy derived from a set of generalized priority indices (generalized Gittins' indices) similar to those first introduced in a contribution of Asawa and Teneketzis (IEE Trans. Automat. Control 41 (1996) 328).

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Multi-armed Bandit process; Switching costs; Optimal switching curves; Hysteretic policy; Priority index policy

UNCORRECTED PROOF

1. Introduction

In the vast domain of sequential decision problems, the class of multi-armed Bandits processes (MABP) does play a privileged role as it can be solved optimally. The MABP consists in sequentially selecting one among a class of N parallel payoff projects in order to maximize a global reward on an infinite horizon. After the seminal and pioneering work of Gittins and Jones (1974), we know that the optimal strategy can be fully characterized by priority indices (the Gittins' indices), provided that *no setup cost and/or time* is incurred when switching from one project to another. It is however very common to observe in actual situations, that switchings generate costs and often cannot be instantaneous (for example when non-preemptive constraints are taken into account).

In presence of switching costs and/or time delays, it is no more possible to characterize an optimal strategy by using priority indices. A counterexample has been constructed by Banks and Sundaram (1994) to illustrate this point. In addition, numerical experiments such as those performed for instance in Ha (1997) and Van Oyen and Teneketzis (1994) show that, in presence of switching costs, the optimal strategy exhibits a highly complex structure. While the complete and analytical characterization of the optimal strategy for MABP with switching costs, remains a mathematical challenge, it is not clear that overcoming this difficulty will be of great benefit for applications. Indeed, optimal strategy imply often complex implementations, a drawback that will drive most practitioners to prefer efficient (though sub-optimal) rules which are more easy to use. In particular, strategies based on generalized priority indices potentially remain, due to there simplicity, very appealing.

How far from optimality can we expect to be when using generalized priority indices in MABP with switching costs? We will approach this question in the present paper by studying a class of models involving MABP for which it is possible to exactly determine the optimal strategy by direct calculation. The model we consider belongs to the class of deteriorating MABP (DMABP), for which the reward is monotonously decreasing. For these DMABP with switching costs, we show in Section 3 that when two projects are considered, the optimal policy exhibits an hysteretic shape.

[☆] This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associated Editor ■ ■ ■ under the direction of Editor ■ ■ ■.

* Corresponding author.

E-mail addresses: fabrice.dusonchet@epfl.ch (F. Dusonchet), max.hongler@epfl.ch (M.-O. Hongler).

[1] In part supported by the "Fonds National Suissee pour la Recherche Scientifique".

[2] In part supported by the "Fundação para a Ciência e a Tecnologia, MCT, Portugal".

0005-1098/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0005-1098(03)00203-6

UNCORRECTED PROOF

The hysteresis reflects the intuitive fact that not only the present state but also the history of the process are to be taken into account in order to decide which is the optimal scheduling. In Section 5, we introduce a possible generalization of the priority indices (along the same lines as those proposed in Banks and Sundaram (1994) and Asawa and Teneketzis (1996)) and we compare, for this two-armed process, the sub-optimal strategy resulting by the use of these indices, with the optimal scheduling previously derived.

2. Multi-armed Bandit problem with switching costs—general and deteriorating case

The multi-armed Bandit problem (MABP) consists in deriving an optimal scheduling of N parallel projects (i.e. the arms) in order to maximize a global reward. We shall write $X_j(t)\in\mathscr{X}_j$, $j=1,\ldots,N$ for the state at time t and $\mathscr{X}_j$ is the state space of the project j. In the following we will consider continuous time MABP and the state space will be the real line (i.e. $\mathscr{X}_j=\mathbb{R}$). The time evolutions $X_j(t)$ follow in general stochastic processes and we assume the statistical independence of these processes. At any particular time, only one project is engaged, the other $(N-1)$ disengaged projects remain dynamically “frozen”. The state of the engaged project evolves with time while the “frozen” projects stay fixed in their positions. The engaged project j earns rewards at rate $h_j(X_j(t))$. Disengaged projects bring no reward. We write $\{t_i,\ i=0,1,\ldots\}$, with $0\leqslant t_1<\cdots<t_i<t_{i+1}<\cdots, i=1,2,\ldots$, the sequence of ordered switching times occurring when it is decided to stop a project and to engage another one. We assume that, each time a switching is operated, a fixed switching cost $C>0$ is incurred. Note that C does neither depend on the project we leave nor on the project we engage. The switching decision at time t_i is based on the observation of $X_j(t)$, $j=1,\ldots,N$, $\forall t\leqslant t_i$.

Let us define the initial conditions: $\vec{X}(0)=(X_1(0),\ldots,X_N(0))$, and $\vec{I}^\pi(0)=(I_1^\pi(0),\ldots,I_N^\pi(0))$, where $I_j^\pi(t)$ stand for the indicator function defined by

$$I_j^\pi(t) = \begin{cases} 1 & \text{if project } j \text{ is engaged at time } t \text{ under policy } \pi,\\ 0 & \text{otherwise.}\end{cases}$$

The solution of the MABP consists in determining the optimal strategy $\pi^*\in\Pi$, where Π is the set of all admissible (i.e. non-anticipating) policies which specifies the switching time sequence $\{t_i^*, i=0,1,\ldots\}$ and for each t_i^*, it indicates which project to engage in order to maximize the global reward:

$$J^{\pi^*}(\vec{X}(0),\vec{I}^{\pi^*}(0)) = \max_{\pi\in\Pi} E_\pi\left\{\int_0^\infty e^{-\beta t}\left(\sum_{j=1}^{N} h_j(X_j(t))I_j^\pi(t) - \sum_i C\delta^\pi(t-t_i)\right)dt \,\middle|\, \vec{X}(0),\vec{I}^\pi(0)\right\} \qquad (1)$$

with $E_\pi\{\cdot\mid\vec{X}(0),\vec{I}^\pi(0)\}$ being the conditional expectation with respect to the initial conditions $\vec{X}(0)$ and $\vec{I}^\pi(0)$, $0<\beta$ is a discounting factor and $\delta^\pi(t-t_i)$ is the Dirac mass distribution.

In absence of switching cost (i.e. when $C\equiv 0$), the MABP is optimally solved by a priority index policy. This policy is based on the possibility to assign to each project an index $v_j(x_j)$ (i.e. the Gittins' index defined on the state $x_j\in\mathscr{X}_j$) depending only on the dynamic $X_j(t)$ and the reward structure $h_j(x_j)$. In terms of the $v_j(x_j)$, the optimal strategy reduces to the rule: “*At each time t engage the project exhibiting the largest index value $v_j(X_j(t))$.*”

The Gittins' index of project j can be determined by studying an associated optimal stopping problem (problem $\mathscr{SP}_j$), which consists in determining $\tau^*\geqslant 0$, that maximizes the global reward $J_j^M(X_j(0))$ gained by engaging project j until time τ^*, then stop and collect a reward $e^{-\beta\tau^*}M$:

$$J_j^M(X_j(0)) = E\left\{\int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\,dt + e^{-\beta\tau^*}\frac{M}{\beta}\,\middle|\, X_j(0)\right\}. \qquad (2)$$

Definition (Gittins' index). The Gittins' index $v_j(x_j)$ associated with a position $X_j(0)=x_j$ of the project j is defined by (Gittins & Jones, 1974; Whittle, 1982; Walrand, 1988; Gittins, 1989):

$$v_j(x_j) = \sup_{\tau\geqslant 0}\frac{E\left\{\int_0^\tau e^{-\beta t}h_j(X_j(t))\,dt\right\}}{E\left\{\int_0^\tau e^{-\beta t}\right\}}, \qquad (3)$$

where the supremum is taken over all admissible stopping time.

Definition (Deteriorating MABP (Whittle, 1982)). We say that a MABP is deteriorating, if for all $j=1,\ldots,N$, $J_j^M(X_j(t))$ is decreasing for t increasing. For future use, we shall write DMABP for the class of deteriorating MABP.

Property 1. *In* Whittle (1982) *the following results are established*:

(i) *A MABP is a DMABP if and only if for all $j=1,\ldots,N$, $h_j(X_j(t))$ is decreasing for t increasing.*

(ii) *The Gittins' index for DMABP is*

$$v_j(x_j) = h_j(x_j). \qquad (4)$$

3. Optimal hysteretic policy for a class of deterministic deteriorating DMABP with switching cost—the two-armed case

In presence of switching costs, it is obvious that when comparing two projects with identical dynamics and being in the same state, to stay on the project currently in use is necessarily more attractive than to switch to the other one

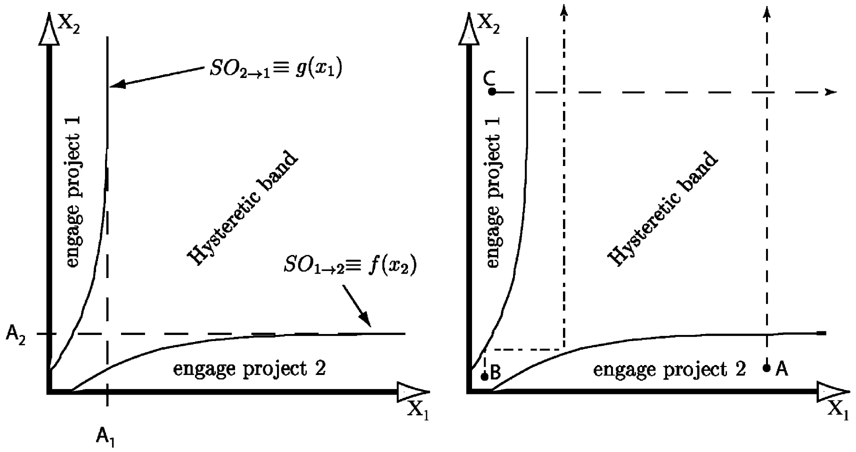

Fig. 1. (a) Typical shape of the optimal policy. (b) The dashed lines are the optimal trajectories for three different initial conditions A, B and C.

UNCORRECTED PROOF

(as no switching cost is incurred). Clearly, the past history of the system affects the *decision maker* (DM) in selecting his action. Accordingly, the scheduling policy will generically include an hysteretic buffer which will be determined by two switching curves.

Let us now focus on the optimal policy for a simple class of two-armed DMABP with switching costs, having the following properties:

$$\frac{\mathrm{d}X_j}{\mathrm{d}t} = \theta_j,$$

$$X_j(0) = x_{0_j} \quad \text{and} \quad h_j(x_j) := \Gamma_j(1 + \mathrm{e}^{-\alpha_j x_j}). \tag{5}$$

Note that:

- The dynamics of the $X_j(t)$, $j = 1, \dots, N$ is deterministic.
- The reward functions $h_j(x_j)$, $j = 1, \dots, N$ are decreasing.
- For any initial condition $X_j(0)$, the instantaneous reward $h_j(X_j(t))$ fulfills:

$$\lim_{t\to\infty} h_j(X_j(t)) = \Gamma_j \in \mathbb{R}, \quad j = 1, 2. \tag{6}$$

- $h_j(X_j(t_1)) < h_j(X_j(t_2))$, $\forall\ t_2 > t_1$ and then Property 1(i) of Section 2 holds. Therefore the problem does belong to the class of DMABP.

Claim. *For a two-armed continuous time, deterministic DMABP with switching costs, for which the dynamical processes and the reward functions are defined by Eqs.* (5) *and* (6), *the optimal policy is characterized by two non-decreasing switching curves* $\mathcal{SO}_{1\to 2}$ *and* $\mathcal{SO}_{2\to 1}$. *Moreover, given an initial condition, only a finite number of switchings occur under the optimal policy.*

Proof of the claim. We report in the appendix the essential steps of the proof. More details can be found in Dusonchet and Hongler (2002).

4. Explicit derivation of the switching curves

From the fact that the optimal switching curve $\mathcal{SO}_{1\to 2}$, [respectively, $\mathcal{SO}_{2\to 1}$], are non-decreasing and that the optimal policy involves only a finite number of switchings, it necessarily exists two values A_1 and A_2, such that for any initial condition $(X_1(0) \geqslant A_1,\ X_2(0),\ 2)$ [respectively, $(X_1(0),\ X_2(0) \geqslant A_2,\ 1)$], the optimal policy commands to engage the project 2 [respectively, the project 1], forever (i.e. the optimal switching curves exhibit the qualitative shape sketched in Fig. 1(a). We can calculate these values as follows:

Starting at the initial condition $(\infty, A_2, 1)$, [respectively, $(A_1, \infty, 2)$], it is equivalent to either engage the project 1 forever [respectively, the project 2 forever], or to switch initially from project 1 to 2, [respectively, from project 2 to 1] and then to engage it forever (i.e. the initial conditions $(\infty, A_2, 1)$ and $(A_1, \infty, 2)$ are on the switching curves). Accordingly, we can write:

$$\left[\int_0^\infty \mathrm{e}^{-\beta t} h_1(X_1(t))\,\mathrm{d}t \,|\, X_1(0) = \infty\right]$$

$$= -C + \left[\int_0^\infty \mathrm{e}^{-\beta t} h_2(X_2(t))\,\mathrm{d}t \,|\, X_2(0) = A_2\right] \tag{7}$$

which determines A_2. In Eq. (7), we have used the notation $[\,\cdot\,|\,X_i(t) = x_i]$ to indicate that the project i is in state x_i at time t. To simplify the exposition, we assume first that both projects have identical dynamics and reward characteristics (i.e. we consider symmetric DMABP). In this case, the Fig. 1(a) is symmetric and $A_1 = A_2$.

The non-decreasing property of the switching curves enables to determine them recursively. To see this, write $f(x_2)$ [respectively, $g(x_1)$] for the function which describes $\mathcal{SO}_{1\to 2}$ [respectively, $\mathcal{SO}_{2\to 1}$]. Define the sequences of

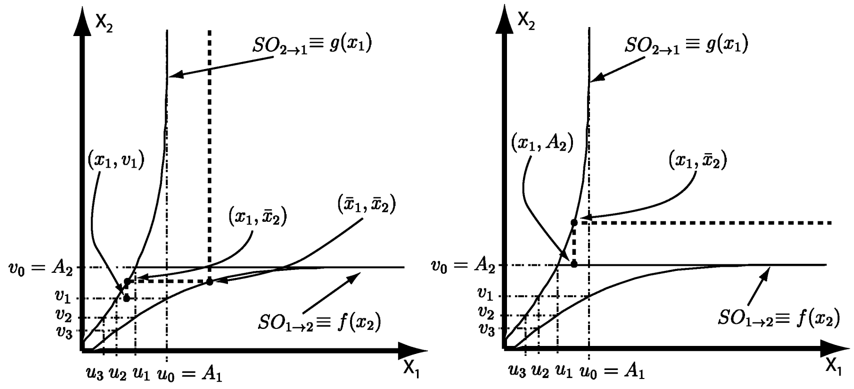
X_2
$SO_{2\to1} \equiv g(x_1)$
(x_1, v_1)
$(x_1, \bar{x}_2)$
$(\bar{x}_1, \bar{x}_2)$
$v_0 = A_2$
v_1
v_2
v_3
$SO_{1\to2} \equiv f(x_2)$
$u_3\ u_2\ u_1\ u_0 = A_1$
X_1
X_2
$SO_{2\to1} \equiv g(x_1)$
(x_1, A_2)
$(x_1, \bar{x}_2)$
$v_0 = A_2$
v_1
v_2
v_3
$SO_{1\to2} \equiv f(x_2)$
$u_3\ u_2\ u_1\ u_0 = A_1$
X_1

Fig. 2. The optimal switching curves $\mathscr{SO}_{2\to1}$ and $\mathscr{SO}_{1\to2}$.

points $(u_0, u_1, \ldots)$ and $(v_0, v_1, \ldots)$ as (see Fig. 2):

$$u_0 = A_1, \quad u_1 = g^{-1}(A_2),$$
$$u_2 = g^{-1}(v_1), \ldots, u_k = g^{-1}(v_{k-1}),$$
$$v_0 = A_2, \quad v_1 = f^{-1}(A_1),$$
$$v_2 = f^{-1}(u_1), \ldots, v_k = f^{-1}(u_{k-1}).$$

Remark. For symmetric two-armed DMABP $g(x_1) = f^{-1}(x_1)$.

Iteration (1), calculation of $\mathscr{SO}_{2\to1}$ in the interval $[u_1, A_1]$. Assume that the DM is initially engaged on project 2, and that the initial positions are $u_1 \leqslant X_1(0) = x_1 < A_1$ and $X_2(0) = A_2$ (see Fig. 2). Following the optimal policy, the DM switches only once, when the state of the system reaches the position $(X_1(t) = x_1, X_2(t) = \bar{x}_2, 2)$ (i.e. $(x_1, \bar{x}_2)$ lies on $\mathscr{SO}_{2\to1}$, see Fig. 2). Therefore the optimal reward for the initial condition $(x_1, A_2, 2)$ fulfills:

$$JO(x_1, A_2, 2; \bar{x}_2) = \left[\int_0^{\tau(\bar{x}_2)} \mathrm{e}^{-\beta t} h_2(X_2(t))\,\mathrm{d}t \,|\, X_2(0) = A_2\right] + \mathrm{e}^{-\beta\tau(\bar{x}_2)} \times \left(-C + \left[\int_0^{\infty} \mathrm{e}^{-\beta t} h_1(X_1(t))\,\mathrm{d}t \,|\, X_1(0) = x_1\right] \mathrm{d}t\right), \tag{8}$$

where $\tau(\bar{x}_2)$ is the time at which the process $X_2(\tau(\bar{x}_2)) = \bar{x}_2$. By optimality, the value of $\bar{x}_2$ must fulfill:

$$\frac{\partial}{\partial \bar{x}_2} JO(x_1, A_2, 2; \bar{x}_2) = 0.$$

For the symmetric DMABP, we directly get the switching curve $\mathscr{SO}_{1\to2}$ on the interval $[A_1, \infty]$ by symmetry. Now we can calculate the position of the switching curve $\mathscr{SO}_{2\to1}$ on the interval $[u_2, u_1]$ as follows:

Iteration (2), calculation of $\mathscr{SO}_{2\to1}$ in the interval $[u_2, u_1]$. Assume that project 2 is initially engaged and that the initial positions are $u_2 \leqslant X_1(0) = x_1 < u_1$ and $X_2(0) = v_1$. Following the optimal policy, the DM switches exactly twice, first in the interval $[u_2, u_1]$, when the state of the system reaches the position $(X_1(t) = x_1, X_2(t) = \bar{x}_2, 2)$ and a second times in the interval $[A_1, \infty]$ when the state of the system reaches the position $(X_1(t) = \bar{x}_1, X_2(t) = \bar{x}_2, 1)$ (Note that $\mathscr{SO}_{1\to2}$ for $x \in [u_1, A_1]$ has been calculated previously, (see Fig. 2)). Therefore the optimal reward for $(x_1, v_1, 2)$ is

$$\begin{aligned} &JO(x_1, v_1, 2; \bar{x}_2) \\ &= \left[\int_0^{\tau_1(\bar{x}_2)} \mathrm{e}^{-\beta t} h_2(X_2(t))\,\mathrm{d}t \,\middle|\, X_2(0) = v_1\right] \mathrm{d}t \\ &+ \mathrm{e}^{-\beta\tau_1(\bar{x}_2)} \left(-C + \left[\int_0^{\tau_2(\bar{x}_1)} \mathrm{e}^{-\beta t} h_1(X_1(t))\,\mathrm{d}t \,\middle|\, X_1(\tau_1(\bar{x}_2)) = x_1\right] \mathrm{d}t\right. \\ &+ \mathrm{e}^{-\beta(\tau_1(\bar{x}_2) + \tau_2(\bar{x}_1))} \left(-C + \left[\int_0^{\infty} \mathrm{e}^{-\beta t} h_2(X_2(t))\,\mathrm{d}t \,\middle|\right.\right. \\ &\left.\left.\left. \times\ X_2(\tau_1(\bar{x}_2) + \tau_2(\bar{x}_1)) = \bar{x}_2\right] \mathrm{d}t\right)\right), \end{aligned} \tag{9}$$

where $\tau_1(\bar{x}_2)$ is the time at which the process $X_2(\tau_1(\bar{x}_2)) = \bar{x}_2$ (i.e. is on $\mathscr{SO}_{2\to1}$) and $\tau_2(\bar{x}_1)$ is the time at which the process $X_1(\tau_2(\bar{x}_1)) = \bar{x}_1$ (i.e. is on $\mathscr{SO}_{1\to2}$). Here again, by definition of the switching curve, the value of $\bar{x}_2$ must fulfill: $(\partial/\partial\bar{x}_2) JO(x_1, v_1, 2; \bar{x}_2) = 0$. The switching curve $\mathscr{SO}_{1\to2}$ on the interval $[u_1, A_1]$ is again given by symmetry. Iteratively, we clearly can calculate the complete curve $\mathscr{SO}_{1\to2}$.

Remark. For asymmetric two-armed DMABP, the above procedure can be generalized straightforwardly. Indeed, the symmetry assumption is not required to iterate the construction of $\mathscr{SO}_{2\to1}$.

4.1. Explicitly solved example—deteriorating and deterministic MABP

To illustrate our method, let us calculate explicitly the recursion for the deterministic two-armed symmetric DMABP for which the dynamical processes and the reward functions are defined in Eqs. (5) and (6) with: $\alpha_1 = \alpha_2 = \alpha$, $\Gamma_1 = \Gamma_2 = \Gamma$.

In this case, Eq. (7) reduces to: $\int_0^\infty \mathrm{e}^{-\beta t}\Gamma(1 + \mathrm{e}^{-\alpha(\theta_1 t + \infty)})\,\mathrm{d}t = -C + \int_0^\infty \mathrm{e}^{-\beta t}\Gamma(1 + \mathrm{e}^{-\alpha(\theta_2 t + A_2)})\,\mathrm{d}t$, from

UNCORRECTED PROOF

UNCORRECTED PROOF

which we obtain: $A_2 = -(1/\alpha)\ln[C(\beta+\alpha\theta_2)/\Gamma]$. The Eq. (8) reduces to:

$$JO(x_1, A_2, 2, \bar{x}_2) = \left(\int_0^{\tau(\bar{x}_2)} e^{-\beta t}\Gamma(1+e^{-\alpha(\theta_2 t + A_2)})\,dt + e^{-\beta\tau(\bar{x}_2)}\left(-C + \int_0^{\infty} e^{-\beta t}\Gamma(1+e^{-\alpha(\theta_1 t + x_1)})\,dt\right)\right)$$

with $\tau(\bar{x}_2) = \bar{x}_2 - A_2/\theta_2$.

Eqs. (9) reduces to:

$$JO(x_1, v_1, 2, \bar{x}_2) = \left(\int_0^{\tau_1(\bar{x}_2)} e^{-\beta t}\Gamma(1+e^{-\alpha(\theta_2 t + v_1)})\,dt + e^{-\beta\tau_1(\bar{x}_2)}\left(-C + \int_0^{\tau_2(\bar{x}_1)} e^{-\beta t}\Gamma(1+e^{-\alpha(\theta_1 t + x_1)})\,dt + e^{-\beta(\tau_1(\bar{x}_2)+\tau_2(\bar{x}_1))} \times \left(-C + \int_0^{\infty} e^{-\beta t}\Gamma(1+e^{-\alpha(\theta_2 t + \bar{x}_2)})\,dt\right)\right)\right)$$

with $\tau_1(\bar{x}_2) = \bar{x}_2 - v_1/\theta_2$ and $\tau_2(\bar{x}_1) = \bar{x}_1 - x_1/\theta_1$.

These equations are transcendant for general values of $\alpha, \beta, \theta_i, i = 1, 2$. When $\alpha = \beta = \theta_1 = \theta_2 = 1$, an explicit solution can however be found. It reads:

$$A_1 = A_2 = -\ln\left[\frac{2C}{\Gamma}\right], \quad u_1 = v_1 = -\ln\left[\frac{6C}{\Gamma}\right],$$

$$u_2 = v_2 = -\ln\left[\frac{16C}{7\Gamma - \sqrt{33}\Gamma}\right], \quad \bar{x}_1 = -\ln\left[\frac{e^{-\bar{x}_2}}{2} - \frac{C}{\Gamma}\right].$$

Hence the switching curves for positive initial conditions $(X_1(0), X_2(0)) \in \mathbb{R}_+ \times \mathbb{R}_+$ read as:

$$\mathscr{SO}_{2\to 1} = \begin{cases} \infty & \text{if } x_1 > A_1, \\ -\ln\left[\dfrac{e^{-x_1}}{2} - \dfrac{C}{\Gamma}\right] & \text{if } u_1 \leqslant x_1 < A_1, \\ -\ln\left[\dfrac{2(\Gamma - e^{x_1}C)^2}{\Gamma e^{x_1}(2\Gamma + 2e^{x_1}C + \sqrt{\Gamma^2 + 14\Gamma C e^{x_1} + Q^2 e^{2x_1}}}\right] & \text{if } u_2 \leqslant x_1 < u_1 \\ \vdots & \end{cases}$$

and

$$\mathscr{SO}_{1\to 2} = \begin{cases} -\ln\left[2\left(e^{-x_1} + \dfrac{C}{\Gamma}\right)\right] \\ \text{if } x_1 \geqslant A_1, \\ -\ln\left[\dfrac{2\Gamma + 2Ce^{x_1} + \sqrt{\Gamma^2 + 16\Gamma C e^{x_1}}}{2\Gamma e^{x_1}}\right] \\ \text{if } u_1 \leqslant x_1 < A_1 \\ \vdots \end{cases}.$$

The above results are drawn in Fig. 3.

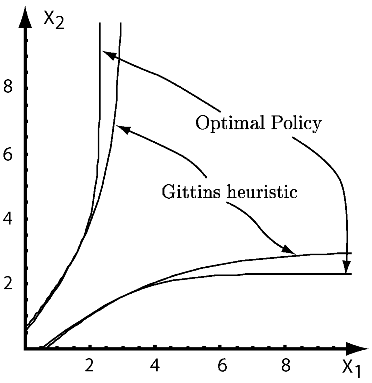

Fig. 3. Optimal policy and the GIH for the parameter value: $\alpha = \beta = \theta_i = 1$ $\Gamma = 2$, $C = 0.1$.

5. Generalized index heuristic and suboptimal hysteresis

Clearly, the hysteretic type optimal scheduling which results from the presence of switching costs, precludes a naive generalization of the Gittins' index policy. By following the idea first exposed in Banks and Sundaram (1994) and Asawa and Teneketzis (1996), let us introduce a set of two indices for each project, namely:

- a continuation index $vc_j(x_j)$,
- a switching index $vs_j(x_j)$.

This duplication of indices enables to construct a generalized priority index heuristics (GIH) which takes into account information regarding the history of the system and hence does exhibit an hysteretic shape topologically similar to the optimal solution. In terms of $vc_j(x_j)$ and $vs_j(x_j)$ a N-armed MABP will be sub-optimally solved by the policy:

Generalized index heuristics (*GIH*): For a project j initially engaged, the GIH read as: "*Continue to engage project j as long as $vc_j(X_j(t)) \geqslant vs_k(X_k(t)), \forall\, k \neq j$. If $vc_j(X_j(t))$ falls below the switching index of another project, then switch to the project having the largest switching index.*"

5.1. Construction of the continuation and the switching indices

To construct the indices on which the GIH is based, we first introduce a special two-armed MABP (denoted by problem $\mathscr{P}_j$ in the following) which is equivalent to the

UNCORRECTED PROOF

stopping problem $\mathscr{SP}_j$ introduced in Section 2. In problem $\mathscr{P}_j$, the first project is the project j itself and the second project (here denoted as project $\mathscr{T}$) follows the frozen dynamics $X_{\mathscr{T}}(t) \equiv \xi$, $\forall t \in \mathbb{R}_+$. When engaged, this second project yields a systematic reward $h_{\mathscr{T}}(\xi) \equiv M$. Assume that initially project j is engaged and note that once the optimal policy commands to switch from the project j to $\mathscr{T}$, it is never optimal to reengage project j. Indeed, if at time t_1, it is optimal to engage project $\mathscr{T}$, so it is for all times $t \geqslant t_1$, as the global evolution is "frozen". This observation establishes the equivalence between the $\mathscr{SP}_j$ and $\mathscr{P}_j$ problems. Write $\tilde{\mathscr{P}}_j$ for the problem $\mathscr{P}_j$, in which a switching cost $C > 0$ is added. Using the problem $\tilde{\mathscr{P}}_j$, we now define:

Definition (Continuation index $vc_j(x)$). The function $vc_j(x)$ is the continuation index of project j if and only if the curve $S_{j\to\mathscr{T}} = \{(x,y) \in \mathbb{R}^2 \mid vc_j(x) = vs_{\mathscr{T}}(y)\}$ is the optimal switching curve for problem $\tilde{\mathscr{P}}_j$ when the DM is initially engaged on j. The index $vs_{\mathscr{T}}(y)$ is the switching index of the frozen project $\mathscr{T}$ given in Lemma 2 below.

Definition (Switching index $vs_j(x)$). The function $vs_j(x)$ is the switching index of project j if and only if the curve $S_{\mathscr{T}\to j} = \{(x,y) \in \mathbb{R}^2 \mid vc_{\mathscr{T}}(x) = vs_s(y)\}$ is the optimal switching curve for problem $\tilde{\mathscr{P}}_j$ when the DM is initially engaged on $\mathscr{T}$. The index $vc_{\mathscr{T}}(y)$ is the continuation index of the "frozen" project $\mathscr{T}$ given in Lemma 2 below.

5.2. Derivation of the continuation and the switching index

Lemma 2. *The continuation and the switching indices for the "frozen" project $\mathscr{T}$ respectively, read as: $vc_{\mathscr{T}}(\xi) = M$ and $vs_{\mathscr{T}}(\xi) = M - C\beta$.*

Proof. Consider a two-armed MABP with both projects having the frozen dynamics as defined for the project $\mathscr{T}$. Suppose that the first project (project $\mathscr{T}_1$) generates a systematic reward of M_1 and that the second project (project $\mathscr{T}_2$) generates a systematic reward of M_2. Then the optimal policy if the DM is initially engaged on project $\mathscr{T}_1$ is to continue forever on this project if and only if $M_1 \geqslant M_2 - C\beta$, otherwise to switch to project $\mathscr{T}_2$ and stay on it forever. This policy is indeed achieved when the priority indices $vc_{\mathscr{T}_l}(\xi)$ and $vs_{\mathscr{T}_l}(\xi)$, $l = 1, 2$ are defined as: $vc_{\mathscr{T}_l}(\xi) = M_l$ and $vs_{\mathscr{T}_l}(\xi) = M_l - C\beta$. □

Theorem 3. *The switching index $vs_j(x_j)$ associate with position $X_j(0) = x_j$ read as*

$$vs_j(x_j) = \sup_{\tau \geqslant 0} \frac{E\left\{\int_0^\tau e^{-\beta t} h_j(X_j(t))\,dt - C(1+e^{-\beta\tau})\right\}}{E\left\{\int_0^\tau e^{-\beta t}\right\}} \tag{10}$$

with τ being a stopping time.

Proof. The optimal reward $J_j^{M,C}(X_j(t_0))$ for the problem $\tilde{\mathscr{P}}_j$ when the DM is initially engaged on project $\mathscr{T}$ reads as:

$$J_j^{M,C}(X_j(t_0)) = E\left\{-C + \int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\,dt \right.$$
$$\left. - e^{-\beta\tau^*}C + \int_{\tau^*}^{\infty} M e^{-\beta t}\,dt\right\}, \tag{11}$$

where τ^* is the time at which it is optimal to engage project $\mathscr{T}$. For an initial condition $(X_j(t_0), \xi)$ on the switching curve $S_{\mathscr{T}\to j}$ and when the DM is initially engaged on project $\mathscr{T}$, it is optimal to immediately engage project $\mathscr{T}$ and to stay on it forever. This yields a reward:

$$J_j^{M,C}(X_j(t_0)) = \int_0^\infty M e^{-\beta t}\,dt. \tag{12}$$

Using Eq. (12) into Eq. (11) implies:

$$\int_0^\infty M e^{-\beta t}\,dt = E\left\{-C + \int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\,dt \right.$$
$$\left. - e^{-\beta\tau^*}C + \int_{\tau^*}^{\infty} M e^{-\beta t}\,dt\right\}. \tag{13}$$

On the other hand, for an initial condition on $S_{\mathscr{T}\to j}$, the continuation index value of project $\mathscr{T}$ equals the switching index value of project j, namely:

$$vs_j(x_j) = vc_{\mathscr{T}}(\xi) = M \tag{14}$$

with M being the solution of Eq. (13), which is given in Eq. (10). □

Theorem 4. *The continuation index $vc_j(x_j)$ associated with position $X_j(0) = x_j$ is the Gittins index given by Eq.* (3).

Proof. Proceed along the same lines as for the proof of Theorem 3. □

Remarks.

- Our present definitions of $vs_j(x)$ and $vc_j(x)$ slightly differ from to those used in Asawa and Teneketzis (1996). Our definitions directly follow from the associated stopping problems used to construct the Gittins' indices (see Dusonchet and Hongler (2002) for more details).
- When $C \equiv 0$, we consistently have that $vs_j(x_j) = vc_j(x_j) = vg_j(x_j)$.

5.3. Explicitly solved example—deteriorating and deterministic two-armed MABP

For the explicit DMABP given by Eqs. (5) and (6), the optimal stopping time τ^* for problem $\tilde{\mathscr{P}}_j$ when the DM is initially engaged on project $\mathscr{T}$, read as

$$\tau^* = \begin{cases} 0 & \text{if } M \geqslant \Gamma(1+\mathrm{e}^{-x_0\alpha}) + C\beta, \\ -\dfrac{x_0\alpha + \ln\left[-\dfrac{\Gamma + C\beta - M}{\Gamma}\right]}{\alpha\theta_1} & \text{if } \Gamma + C\beta < M < \Gamma(1+\mathrm{e}^{-x_0\alpha}) + C\beta, \\ \infty & \text{if } M \leqslant \Gamma + C\beta. \end{cases} \tag{15}$$

To calculate the switching index $vs_j(x_j)$, we solve Eq. (10) where the supremum is reached for τ^* given by Eq. (15) and identify: $M = vs_j(x_j)$.

This equation is generally transcendant. For the special case $\alpha = \beta = \theta_1 = \theta_2 = 1$ however, a closed form solution exists and reads as

$$vs_1(x_0) = \Gamma(1+\mathrm{e}^{-x_0}) + C - 2\sqrt{\Gamma C}\,\mathrm{e}^{-x_0/2}. \tag{16}$$

Using this expression, we can explicitly characterize the switching curve resulting from the GIH for our symmetric two-armed DMABP. We indeed have:

$$S_{1\to 2} = \{(x_1, x_2) \in \mathbb{R}^2 \mid vc_1(x_1) = vs_2(x_2)\} \Rightarrow S_{1\to 2}$$
$$= \left\{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 = -2\ln\left[\mathrm{e}^{-x_1/2} + \frac{C}{\sqrt{\Gamma C}}\right]\right\} \tag{17}$$

and

$$S_{2\to 1} = \{(x_1, x_2) \in \mathbb{R}^2 \mid vc_2(x_1) = vs_1(x_2)\} \Rightarrow S_{2\to 1}$$
$$= \left\{(x_1, x_2) \in \mathbb{R}^2 \;\middle\|\; \left[\begin{array}{ll} x_2 = -2\ln\left[\mathrm{e}^{-x_1/2} - \dfrac{C}{\sqrt{\Gamma C}}\right] & \text{if } x_1 < -2\ln\left[\dfrac{\sqrt{\Gamma C}}{C}\right], \\ +\infty & \text{otherwise.} \end{array}\right]\right\|\right\}. \tag{18}$$

We plot simultaneously, in Fig. 3, the optimal hysteretic policy Eqs. (17) and (18) and the GIH. This picture, clearly shows that the optimal policy has a wider hysteretic gap. This behaviour is in agreement with the result expressed by Lemma 2.7 in Asawa and Teneketzis (1996).

Remarks.

- The claim and its demonstration can be generalized for DMABP when the dynamic of the project is given by random walks with no downward jumps (Kaspi, 2002).
- The sub-optimality of the GIH can be observed by the explicit calculation of the discounted reward obtained under a special initial condition. For example, chose $\Gamma = 2$, $C = 1.1$, $\alpha = \beta = \theta_1 = \theta_2 = 1$, and the initial conditions $(X_1(0) = 0, X_2(0) = 0, 1)$. With these values, the GIH commands to engage project 1 until the system reaches the position $(-2\ln[1 - (C/\sqrt{\Gamma C})], 0)$, then to switch to project 2 and engage it forever. This scheduling yields a global reward of 2988. Instead, the optimal policy commands to engage project 1 for ever and yields a global reward of 3.
- For large values of β, the reward gained in the close future is dominant. Hence, when β is large enough, the reward realized after the first switching tends to be negligible and the GIH is expected to bring results closer to the optimal one. We observe this fact for the class of symmetric bandit given by Eq. (5) by calculating numerically the value $A_2 \equiv A_1$ and comparing it with the optimal one. Both values indeed converge as β is increased. A numerical example is given in the following table where we calculate A_2 for $C = 0.1$, $\theta_i = \alpha = 1$, $\Gamma = 2$ and for three different values of β

β	A_2 GIH	A_2 optimal
1	2.996	2.302
5	1.386	1.203
10	0.571	0.597

Acknowledgements

We thank H. Kaspi for extensive and very stimulating discussions and the referees constructive comments.

Appendix A

The proof of the claim lies on the three following propositions:

Proposition 1. *For any given initial condition, the optimal policy commands to switch only a finite number of times.*

Proposition 2. *The optimal policy is characterized by two switching curves $\mathscr{SO}_{1\to 2}$ and $\mathscr{SO}_{2\to 1}$ which can be, respectively, described by two functions, $\tilde{y} : x_1 \mapsto \tilde{y}(x_1)$ and $\tilde{x} : x_2 \mapsto \tilde{x}(x_2)$.*

UNCORRECTED PROOF

Proposition 3. *The optimal switching curves $\mathscr{S}\mathscr{O}_{1\rightarrow 2}$ and $\mathscr{S}\mathscr{O}_{2\rightarrow 1}$ are non-decreasing.*

As the aim of this paper is to focus on a simple soluble example, we only sketch the proof of these propositions.

Sketch of the proof of Proposition 1. The space of initial conditions $(x_1, x_2, i) \in \mathbb{R}^2 \times \{1, 2\}$, where $i \in \{1, 2\}$ corresponds to the project initially engaged, can be splitted into two disjoint subsets:

(a) A set $(x_1, x_2, i) \in \Lambda$ such that when starting on Λ, the optimal policy commands to engage the project i forever.
(b) Its complementary set $\Lambda' = \{\mathbb{R}^2 \times \{1, 2\}\} \setminus \Lambda$.

Let us define the cumulate sojourn times T_1 and T_2, respectively, spent on projects 1 and 2, under the optimal policy. As we consider infinite time horizon problems, we have that $T_1 + T_2 = \infty$. By definition, for any initial condition $(x_1, x_2, i) \in \Lambda'$, the sojourn times T_1 and T_2 necessarily fulfill one of the following alternatives:

(i) $T_1 = \infty$ and $T_2 = \infty$, (ii) $T_1 < \infty$ and $T_2 = \infty$,

(iii) $T_1 = \infty$ and $T_2 < \infty$.

- It is possible to show that, for an initial condition $(x_1, x_2, i) \in \Lambda'$, if alternative (i) holds then, it exists a finite time $T < \infty$, such that: $(X_1(T), X_2(T), i(T)) \in \Lambda$. This rules out the possible occurrence of alternative (i) in the optimal policy.
- We can prove that the alternatives (ii) and (iii) both imply that $\exists T < \infty$ after which, the optimal policy does not command to switch anymore. To complete the proof, we invoke the fact that: "*Any policy that switches an infinite number of times on a finite horizon incurs an infinite cost, which cannot be possibly optimal.*"

Sketch of the proof of Proposition 2. Introduce the following definitions:

- $\Omega_n^1 = \{(x_1, x_2, 1) \in \mathbb{R}^2 \times \{1, 2\} \,|\,$ the optimal policy commands to switch immediately from project 1 to 2 and then commands to switch exactly n times$\}$, $n = 0, 1, 2, \ldots$ (Fig. 4).
- $\Omega_n^2 = \{(x_1, x_2, 2) \in \mathbb{R}^2 \times \{1, 2\} \,|\,$ the optimal policy commands to switch immediately from project 2 to 1 and then commands to switch exactly n times$\}$, $n = 0, 1, 2, \ldots$ (Fig. 5).
- Write i for the project initially engaged and $\bar{i}$ for the disengaged project.

To prove Proposition 2, we can construct the two functions $\tilde{y}(x_1)$ and $\tilde{x}(x_2)$ first on Ω_0^i with $i = 1, 2$, then iteratively on Ω_n^i, $n = 1, 2, \ldots$ as follows:

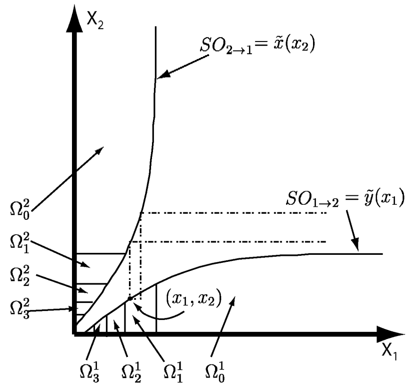

Fig. 4. In dashed lines, two different policy starting at initial condition $(x_1, x_2, 1)$.

Calculate the difference of the global reward expected when one among the two following alternative policies is used:

(i) Switch initially from project i to project $\bar{i}$ and then continue optimally.
(ii) Continue to engage project i during a time $\tau > 0$, then switch from project i to project $\bar{i}$ and finally continue optimally.

This shows that:

(a) If the point $(x_1, x_2, 1)$ belongs to Ω_n^1. Then, $\exists \tilde{y}(x_1)$ such that, $\forall z \in]-\infty, \tilde{y}(x_1)]$, we have $(x_1, z, 1) \in \Omega_n^1$. Moreover, $\forall (x_1, z', 1)$ with $z' > \tilde{y}(x_1)$, we have $(x_1, z', 1) \notin \Omega_n^1$.
(b) If the point $(x_1, x_2, 2)$ belongs to Ω_n^2. Then, $\exists \tilde{x}(x_2)$ such that, $\forall z \in]-\infty, \tilde{x}(x_2)]$, we have $(x_1, z, 2) \in \Omega_n^2$. Moreover, $\forall (x_1, z', 2)$ with $z' > \tilde{x}(x_2)$, we have $(x_1, z', 2) \notin \Omega_n^2$.

The assertion (a) and (b) imply the existence of the function $\tilde{y}(x_1)$ and $\tilde{x}(x_2)$.

Sketch of the proof of Proposition 3. Remember that only the engaged project yields a reward. In addition, the disengaged one remains "frozen" and does not yield any reward. When starting at $A = (x_1, x_2, 1)$, the optimal policy commands to immediately switch from project 1 to 2. That is to say, when starting at A, the expected reward gained by engaging project 1, is less attractive that the expected reward gained by engaging project 2.

With $h_i(x)$ $i = 1, 2$ decreasing (see Eq. (5)), it follows that the expected reward gained by engaging project 1, prior to any switch, at $B = (x_1', x_2, 1)$ with $x_1' > x_1$ is smaller than the expected reward given by engaging project 1 at $A = (x_1, x_2, 1)$. On the other hand, as x_2 is common to both A and B, the expected reward gained by engaging project 2, prior to any switch, is identical for both A and B. Hence, if the decision is to switch from project 1 to 2 at position A, the same switching decision has to be taken when starting at position B.

UNCORRECTED PROOF

Remark. Note from Fig. 3 that the optimal switching curves $\mathscr{S}\mathscr{O}_{1\to 2}$ and $\mathscr{S}\mathscr{O}_{2\to 1}$ are indeed monotonously increasing.

References

Asawa, M., & Teneketzis, D. (1996). Multi-armed bandits with switching penalties. *IEEE Transactions on Automatic Control*, *41*, 328–348.

Banks, J. S., & Sundaram, R. K. (1994). Switching cost and Gittin's index. *Econometrica*, *62*, 687–694.

Dusonchet, F., & Hongler, M.-O. (2002). *Multi-armed Bandits with switching costs and the Gittins index*. Preprint EPFL-IPR-LPM.

Gittins, J. C. (1989). *Multi-armed Bandits allocation indices*. New York: Wiley.

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

Ha, A. (1997). Optimal dynamic scheduling policy for a make-to-stock production system. *Operation Research*, *45*, 42–54.

Kaspi, H. (2002). *Two-armed Bandits with switching costs*. Preprint Technion.

Van Oyen, M. P., & Teneketzis, D. (1994). Optimal stochastic scheduling of forest networks with switching penalities. *Adv. Appl. Probab.*, *26*, 474–497.

Walrand, J. (1988). *An Introduction to Queueing Network*. Englewood Cliffs, NJ: Prentice-Hall.

Whittle, P. (1982). *Optimization over time. Dynamic Programming and Stochastic Control*. New York: Wiley.

Dusonchet Fabrice received in 1999 a Master degree in Mathematics and a Master in Informatics both from the University of Geneva, Switzerland. He spent the Academic year 1998 at the Mathematics Department of the University of Warwick, UK, where he attended lectures devoted to the optimal control theory. In 1999, he joined the Ecole Polytechnique Federale de Lausanne in the Micro-Engineering Department, where he worked on the present Thesis.

Max-Olivier Hongler received in 1981, a Doctoral degree in theoretical physics (in statistical physics), from the University of Geneva, Switzerland. He held several research positions at the Theoretical Physics Department of the University of Texas at Austin, USA, at the University of Toronto, Canada, at the University of Geneva and at the University of Lisboa, Portugal. In 1991, he joined the Ecole Polytechnique Federale de Lausanne, Switzerland, where he presently is Professor in the Micro-Engineering Department. His present research interests are production flows and stochastic models of manufacturing systems.

UNCORRECTED PROOF