Optimal Hysteresis for a Class of Deterministic Deteriorating Two-armed Bandit Problem with Switching Costs.

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

[a] *EPFL*
Laboratoire de Production Microtechnique (LPM)
Institut de Production et Robotique (IPR)
CH-1015 Lausanne
Switzerland

[b] *EPFL*
Laboratoire de Production Microtechnique (LPM)
Institut de Production et Robotique (IPR)
CH-1015 Lausanne
Switzerland

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) first introduced in a contribution of M. Asawa and D. Teneketzis.

Key words: Multi-armed Bandit process, Switching costs, Optimal switching curves, Hysteretic policy, Priority index policy.

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 J.C. Gittins [6], 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 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 J. Banks [2] to illustrate this point. In addition, numerical experiments such as those performed for instance in [7] and [9] show that, in presence of switching costs, the optimal

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

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

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

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 arms are considered, the optimal policy exhibits an hysteretic shape. 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 [2] and [1]) 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 $X_j(t) \in \mathcal{X}_j,\ j = 1, \ldots, N$ for the state at time t and $\mathcal{X}_j$ 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. $\mathcal{X}_j = \mathbb{R}$). 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 gives an instantaneous reward $h_j(X_j(t))$. Disengaged projects bring no reward. We write $\{t_i,\ i = 0,\ 1, \ldots\}$, with $0 \leq t_1 \leq \cdots \leq t_i \leq t_{i+1} \leq \cdots,\ i = 1, 2, \ldots$, 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 $X_j(t),\ j = 1, \ldots, N,\ \forall t \leq t_i$.

Let us define the initial conditions:

$$\vec{X}(0) = (X_1(0), \ldots, X_N(0)),$$
$$\vec{I}^{\pi}(0) = (I_1^{\pi}(0), \ldots, I_N^{\pi}(0)),$$

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:

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

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 $\nu_j(X_j(t))$ (i.e. Gittins' index) depending only on the dynamic $X_j(t)$ and the reward structure $h_j(x_j)$. In terms of the $\nu_j(X_j(t))$, the optimal strategy reduces to the rule: *At each time t engage the project exhibiting the largest index value $\nu_j(X_j(t))$.*

The Gittins' index of project j can be determined by studying an associated optimal stopping problem (problem $\mathcal{SP}_j$), which consists in determining $\tau^* \geq 0$, 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 $e^{-\beta\tau^*}M$:

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

Definition (Gittins' index): The Gittins' index $\nu_j(X_j(0))$ associated with a position $X_j(0)$ of the project j is defined by ([6], [11], [10], [5]):

$$\nu_j(X_j(0)) = \frac{E\left\{\int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt\right\}}{E\left\{\int_0^{\tau^*} e^{-\beta t}\right\}}. \tag{3}$$

Definition (Deteriorating MABP [11]): 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 [11] the following result 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:

$$\nu_j(X_j(0)) = h_j(X_j(0)) \tag{4}$$

3 Optimal hysteretic policy for 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 (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:

$$\frac{dX_j}{dt} = \theta_j \quad ; \quad X_j(0) = x_{0_j} \tag{5}$$

and

$$h_j(x_j) := \Gamma_j(1 + e^{-\alpha_j x_j}). \tag{6}$$

Note that:

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

$$\lim_{t\to\infty} h_j(X_j(t)) = \Gamma_j \in \mathbb{R}, \;\; j = 1, 2. \tag{7}$$

- $h_j(X_j(t_1)) < h_j(X_j(t_2))$, $\forall\ t_2 > t_1$ and then Property 1 $i)$ of section 2 holds. Therefore this 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 switching occur under the optimal policy.

Proof of the claim: We report in the appendix the essential steps of the proof. The complete details can be found in [4].

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 $(X_1(0) \geq A_1, X_2(0), 2)$ [respectively $(X_1(0), X_2(0) \geq A_2, 1)$], 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.1a). We can calculate these values as follows:

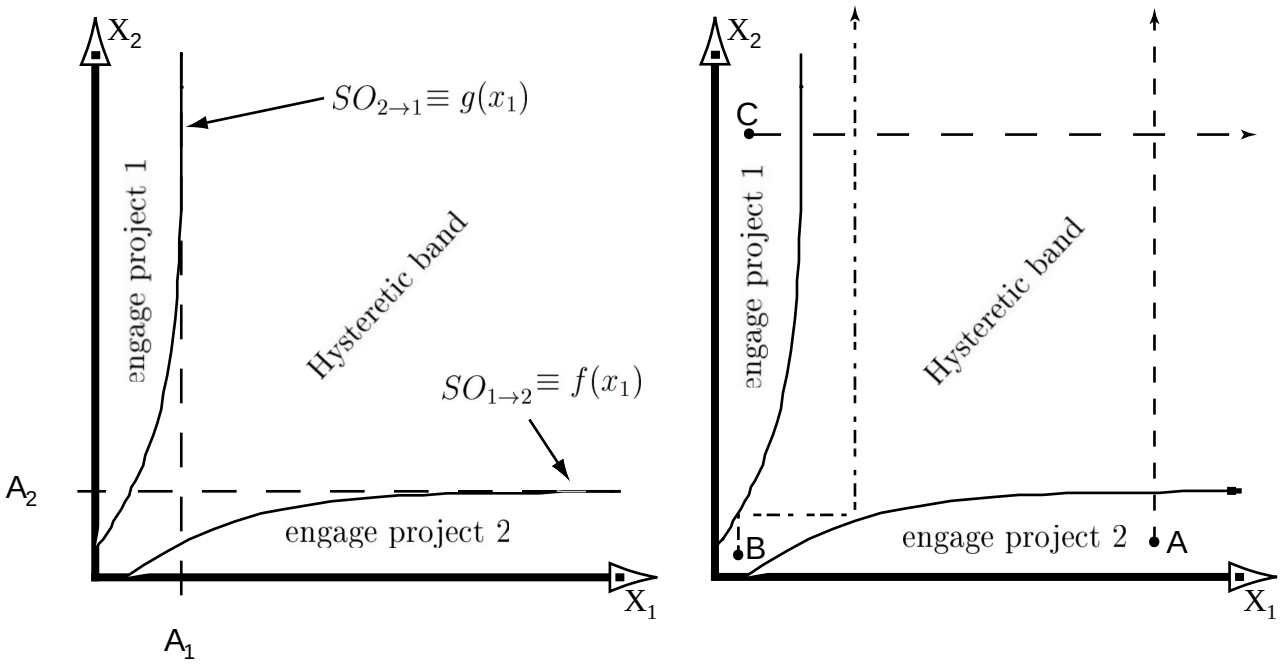

Fig. 1. a) Typical shape of the optimal policy. b) The dashed lines are the optimal trajectories for three different initial conditions A, B and C.

Starting at the initial condition $(\infty, A_2, 1)$, [respectively $(A_1, \infty, 2)$], it is equivalent to either engage the project 1 [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 $(\infty, A_2, 1)$, [respectively $(A_1, \infty, 2)$], is on the switching curve). Accordingly, we can write:

$$\left[\int_0^\infty e^{-\beta t} h_1(X_1(t))\, dt \mid X_1(0) = \infty\right] = -C + \left[\int_0^\infty e^{-\beta t} h_2(X_2(t))\, dt \mid X_2(0) = A_2\right] \tag{8}$$

which determines A_2. In Eq.(8), we have used the notation $[\,\cdot \mid X_i(t) = x_i]$ 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. 1a) is symmetric and $A_1 = A_2$.

The non-decreasing property of the switching curves enables to determine them recursively. To see this, write $f(x_1)$ [respectively $g(x_1)$] for the function which describes $\mathcal{SO}_{1\to 2}$ [respectively $\mathcal{SO}_{2\to 1}$]. Define the sequences of points $(u_0, u_1, \ldots)$ and $(v_0, v_1, \ldots)$ as: (see Fig. 2)

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

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

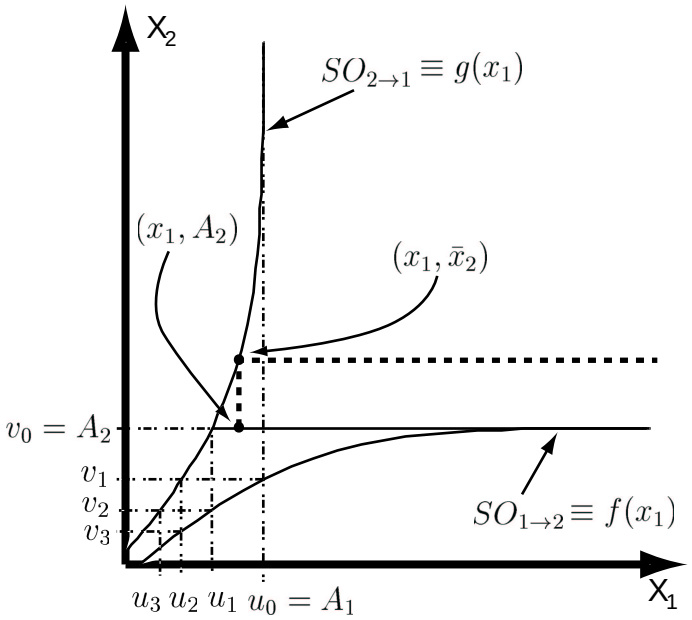

Fig. 2. The optimal switching curves $\mathcal{SO}_{2\to 1}$ and $\mathcal{SO}_{1\to 2}$.

Iteration 1), calculation of $\mathcal{SO}_{2\to 1}$ in the interval $[u_1, A_1]$:
Assume that the DM is initially engaged on project 2, and that the initial positions are $u_1 \le X_1(0) = x_1 < A_1$ 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 $\mathcal{SO}_{2\to 1}$, see Fig. 2). Therefore the optimal reward for the initial condition $(x_1, A_2, 2)$ fulfills:

$$\begin{aligned} JO(x_1, A_2, 2; \bar{x}_2) = &\left[\int_0^{\tau(\bar{x}_2)} e^{-\beta t} h_2(X_2(t))\, dt \mid X_2(0) = A_2\right] + \\ &e^{-\beta\tau(\bar{x}_2)} \left(-C + \left[\int_0^{\infty} e^{-\beta t} h_1(X_1(t))\, dt \mid X_1(0) = x_1\right] dt\right), \end{aligned} \tag{9}$$

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:

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

$\equiv g(x_1)$

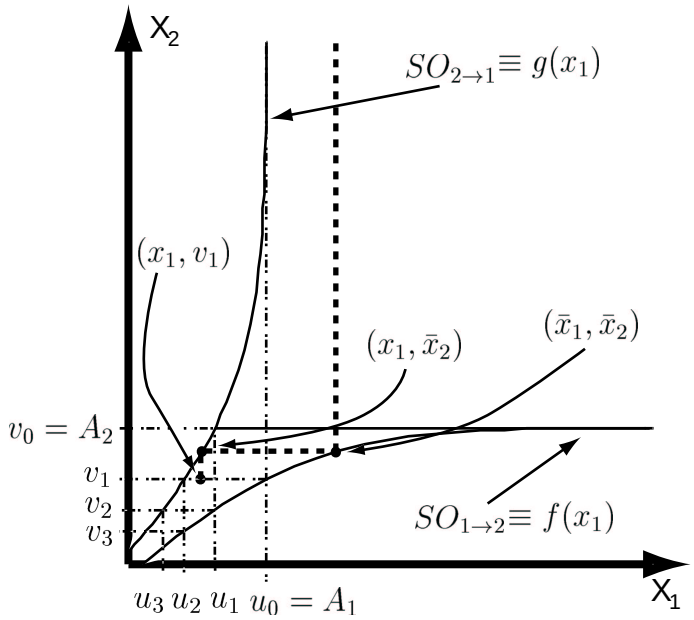

Fig. 3. The optimal switching curves $\mathcal{SO}_{2\to1}$ and $\mathcal{SO}_{1\to2}$.

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

Iteration 2), calculation of $\mathcal{SO}_{2\to1}$ in the interval $[u_2, u_1]$:

Assume that project 2 is initially engaged and that the initial positions are $u_2 \le X_1(0) = x_1 < u_1$ and $X_2(0) = v_1$. Following the optimal policy, the DM will switch 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 $\mathcal{SO}_{1\to2}$ for $x \in [u_1, A_1]$ has been calculated previously, see Fig.3). Therefore the optimal reward for $(x_1, v_1, 2)$ is:

$$JO(x_1, v_1, 2; \bar{x}_2) = \Big(\Big[\int_0^{\tau_1(\bar{x}_2)} e^{-\beta t} h_2(X_2(t))\, dt \;\Big|\; X_2(0) = v_1\Big]\, dt +$$

$$e^{-\beta\tau_1(\bar{x}_2)}\Big(- C + \Big[\int_0^{\tau_2(\bar{x}_1)} e^{-\beta t} h_1(X_1(t))\, dt \;\Big|\; X_1(\tau_1(\bar{x}_2)) = x_1\Big]\, dt + \tag{10}$$

$$e^{-\beta(\tau_1(\bar{x}_2)+\tau_2(\bar{x}_1))}\Big(- C + \Big[\int_0^{\infty} e^{-\beta t} h_2(X_2(t))\, dt \;\Big|\; X_2(\tau_1(\bar{x}_2) + \tau_2(\bar{x}_1)) = \bar{x}_2\Big] dt\Big)\Big),$$

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 $\mathcal{SO}_{2\to1}$) 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 $\mathcal{SO}_{1\to2}$). Here again, by definition of the switching curve, the value of $\bar{x}_2$ must fulfill:

$$\frac{\partial}{\partial \bar{x}_2} JO(x_1, v_1, 2; \bar{x}_2) = 0.$$

The switching curve $\mathcal{SO}_{1\to2}$ on the interval $[u_1, A_1]$ is again given by symmetry. Iteratively, we clearly can calculate the complete curve $\mathcal{SO}_{1\to2}$.

Remark: For non-symmetric two-armed DMABP, the above procedure can be generalized straightforwardly. Indeed, the symetry assumption is not requiered to iterate the construction of $\mathcal{SO}_{2\to 1}$.

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:

$$\alpha_1 = \alpha_2 = \alpha, \quad \Gamma_1 = \Gamma_2 = \Gamma$$

In this case, Eq.(8) reduces to:

$$\int_0^\infty e^{-\beta t}\Gamma(1+e^{-\alpha(\theta_1 t+\infty)})dt = -C + \int_0^\infty e^{-\beta t}\Gamma(1+e^{-\alpha(\theta_2 t+A_2)})dt,$$

from which we obtain:

$$A_2 = -\frac{1}{\alpha}\ln\left[\frac{C(\beta+\alpha\theta_2)}{\Gamma}\right].$$

Eq.(9) reduces to:

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

with

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

Eqs.(10) reduces to:

$$JO(x_1, v_1, 2, \bar{x}_2) = (\textstyle\int_0^{\tau_1(\bar{x}_2)} e^{-\beta t}\Gamma(1+e^{-\alpha(\theta_2 t+v_1)})dt +$$

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

with

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

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

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

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

$$u_2 = v_2 = -\ln\left[\frac{16C}{7\Gamma - \sqrt{33}\Gamma}\right],$$

$$\bar{x}_1 = -\ln\left[\frac{e^{-\bar{x}_2}}{2} - \frac{C}{\Gamma}\right].$$

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

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

and

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

The above results are drawn in Fig. 4.

5 Generalized Index heuristic (GIH) 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 [2] and [1], let us introduce a set of two indices for each project, namely:

- a continuation index $\nu c_j(X_j(0))$,
- a switching index $\nu s_j(X_j(0))$.

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

Definition (Continuation index $\nu c_j(x)$): The function $\nu c_j(x)$ is the continuation index of project j if and only if the curve

$$S_{j\to\mathcal{T}} = \Big\{(x,y) \in \mathbb{R}^2 \mid \nu c_j(x) = \nu s_{\mathcal{T}}(y)\Big\}$$

is the optimal switching curve for problem $\tilde{\mathcal{P}}_j$ when the DM is initially engaged on j. The index $\nu s_{\mathcal{T}}(y)$ is the switching index of the frozen project $\mathcal{T}$ given in the lemma 2 below.

Definition (Switching index $\nu s_j(x)$): The function $\nu s_j(x)$ is the switching index of project j if and only if the curve

$$S_{\mathcal{T}\to j} = \Big\{(x,y) \in \mathbb{R}^2 \mid \nu c_{\mathcal{T}}(x) = \nu s_s(y)\Big\}$$

is the optimal switching curve for problem $\tilde{\mathcal{P}}_j$ when the DM is initially engaged on $\mathcal{T}$. The index $\nu c_{\mathcal{T}}(y)$ is the continuation index of the trivial project $\mathcal{T}$ given in the lemma 2 below.

5.2 Derivation of the continuation and the switching index

Lemma 2: The continuation and the switching indices for the trivial project $\mathcal{T}$ read as:

$$\nu c_{\mathcal{T}}(\xi) = M$$

and

$$\nu s_{\mathcal{T}}(\xi) = M - C\beta.$$

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

$$\nu c_{\mathcal{T}_l}(\xi) = M_l$$

and

$$\nu s_{\mathcal{T}_l}(\xi) = M_l - C\beta.$$

□

Theorem 3: The continuation index $\nu c_j(X_j(t_0))$ read as:

$$\nu c_j(X_j(t_0)) = \frac{E\left\{\int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt\right\}}{E\left\{\int_0^{\tau^*} e^{-\beta t}\right\}}$$

with τ^* the optimal stopping time.

Proof: The optimal reward $J_j^{M,C}(X_j(t_0))$ for the problem $\tilde{\mathcal{P}}_j$ when the DM is initially engaged on arm j read as:

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

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

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

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

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

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

$$\nu c_j(X_j(t_0)) = \nu s_{\mathcal{T}}(\xi) = M - C\beta, \tag{14}$$

with M being the solution of Eq.(13), namely:

$$M = \frac{E\left\{ \int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt + C(1 - e^{-\beta \tau^*}) \right\}}{E\left\{ \int_0^{\tau^*} e^{-\beta t}\, dt \right\}}. \tag{15}$$

Introducing Eq.(15) into Eq.(14) we obtain:

$$\nu c_j(X_j(t_0)) = \frac{E\left\{ \int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt \right\}}{E\left\{ \int_0^{\tau^*} e^{-\beta t}\, dt \right\}}, \tag{16}$$

which ends the proof.

□

Theorem 4: The switching index $\nu s_j(X_j(t_0))$ read as:

$$\nu s_j(X_j(t_0)) = \frac{E\left\{ \int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt - C(1 + e^{-\beta \tau^*}) \right\}}{E\left\{ \int_0^{\tau^*} e^{-\beta t} \right\}} \tag{17}$$

with τ^* the optimal stopping time.

Proof: Proceed along the same lines as in the proof of Theorem 3.

Remarks:

- Our present definitions of $\nu s_j(x)$ and $\nu c_j(x)$ are slightly different to those used in [1]. Our definitions are those which directly follow from the associated stopping problems used to construct the Gittins' indices (see [3] for more details).
- Note that the continuation index $\nu c_j(X_j(t_0))$ is equivalent to the Gittins' index $\nu g_j(X_j(t_0))$ (i.e. Eq.(16) is equivalent to Eq.(3)). In particular for DMABP, we have $\nu c_j(x) = h_j(x)$ (see Eq.(4)).
- When $C \equiv 0$, we consistently have that $\nu s_j(X_j(t_0)) = \nu c_j(X_j(t_0)) = \nu g_j(X_j(t_0))$.

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 $\tilde{\mathcal{P}}_j$ when the DM is initially engaged on project $\mathcal{T}$, read as:

$$\tau^* = \begin{cases} 0 & \text{if } M \geq \Gamma(1 + e^{-x_0 \alpha}) + C\beta \\ -\frac{x_0 \alpha + \ln\left[-\frac{\Gamma + C\beta - M}{\Gamma}\right]}{\alpha \theta_1} & \text{if } \Gamma + C\beta < M < \Gamma(1 + e^{-x_0 \alpha}) + C\beta \\ \infty & \text{if } M \leq \Gamma + C\beta \end{cases}. \tag{18}$$

To calculate the switching index $\nu s_j(X_j(0))$ we solve Eq.(17) where τ^* is given by Eq.(18) and with the identification:

$$M = \nu s_j(X_j(0)).$$

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

$$\nu s_1(x_0) = \Gamma(1 + e^{-x_0}) + C - 2\sqrt{\Gamma C}\, e^{-\frac{x_0}{2}}, \tag{19}$$

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

$$S_{1\to 2} = \left\{(x_1, x_2) \in \mathbb{R}^2 \mid \nu c_1(x_1) = \nu s_2(x_2)\right\} \Rightarrow$$

$$S_{1\to 2} = \left\{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 = -2\ln\left[e^{-\frac{x_1}{2}} + \frac{C}{\sqrt{\Gamma C}}\right]\right\} \tag{20}$$

and

$$S_{2\to 1} = \left\{(x_1, x_2) \in \mathbb{R}^2 \mid \nu c_2(x_1) = \nu s_1(x_2)\right\} \Rightarrow$$

$$S_{2\to 1} = \left\{(x_1, x_2) \in \mathbb{R}^2 \;\middle|\; \begin{cases} x_2 = -2\ln\left[e^{-\frac{x_1}{2}} - \frac{C}{\sqrt{\Gamma C}}\right] & \text{if } x_1 < -2\ln\left[\frac{\sqrt{\Gamma C}}{C}\right] \\ +\infty & \text{otherwise} \end{cases} \right\}. \tag{21}$$

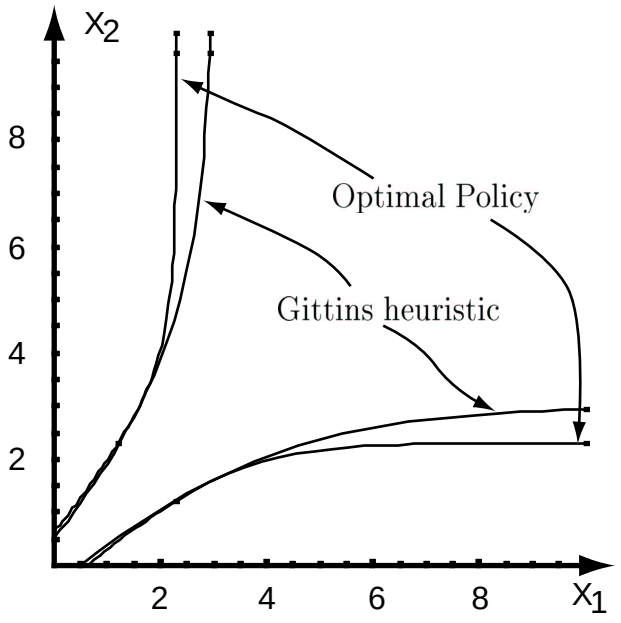

Fig. 4. Optimal policy and the GIH for the parameter value: $\alpha = \beta = \theta_i = 1$ $\Gamma = 2$, $C = 0.1$.

We plot simultaneously, in Fig. 4, the optimal hysteretic policy Eqs.(20) and (21) 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 [1].

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 (see [8] for the details).

- 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 $(-2\ln\left[1 - \frac{C}{\sqrt{\Gamma C}}\right], 0)$, then to switch to project 2 and engage it forever. This scheduling yields a global reward of 2,988. 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.(6) 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

Acknowledgment: We thank H. Kaspi for extensive and very stimulating communications and the referees constructive comments.

A Appendix

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 $\mathcal{SO}_{1\to 2}$ and $\mathcal{SO}_{2\to 1}$ which can be respectively described by two functions, $\tilde{y}:\ x_1 \mapsto \tilde{y}(x_1)$ and $\tilde{x}:\ x_2 \mapsto \tilde{x}(x_2)$.

Proposition 3: The optimal switching curves $\mathcal{SO}_{1\to 2}$ and $\mathcal{SO}_{2\to 1}$ are non-decreasing.

As the aim of this paper is to focus on a soluble example, we give here only the sketch of the proof.

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 $\Lambda' = \left\{\mathbb{R}^2 \times \{1, 2\}\right\} \setminus \Lambda$.

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) for 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 use the property: "*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:

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

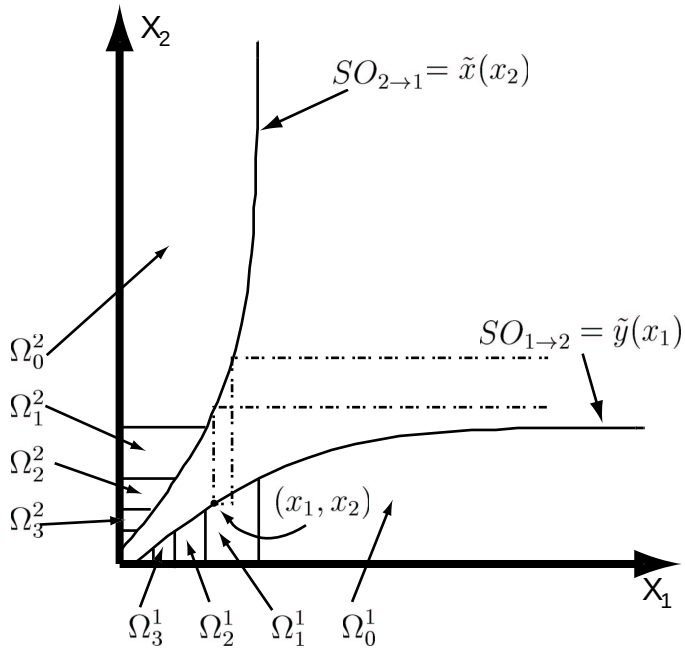

Fig. A.1. Two different policy starting at initial condition $(x_1, x_2, 1)$.

To prove proposition 2, we can construct the two functions $\tilde{y}(x_1)$ and $\tilde{x}(x_2)$ first on $\Omega_0^i,\ i = 1, 2$, then iteratively on Ω_n^i, $n = 1, 2, \ldots$ as follows:

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

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

We can show that:

a) If the point $(x_1, x_2, 1)$ belongs to Ω_n^1. Then, $\exists\ \tilde{y}(x_1)$ such that,

$$\forall z \in]-\infty, \tilde{y}(x_1)],\ \text{we have } (x_1, z, 1) \in \Omega_n^1.$$

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,

$$\forall z \in]-\infty, \tilde{x}(x_2)],\ \text{we have } (x_1, z, 2) \in \Omega_n^2.$$

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*) directly lead to 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 brings a reward and that the disengaged one remains dynamically frozen and does not bring any reward. Assume that 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 given by engaging the project 1, is less attractive that the expected reward given by engaging the project 2.
With $h_i(x)$ $i = 1, 2$ decreasing (see Eq.6)), it follows that the expected reward given 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 given 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.

Remark: In section 5.3, we plot in Fig. 4 the optimal switching curves for our class of DMABP. The increasing property of the switching curves can be seen explicitly.

References

[1] M. Asawa and D. Teneketzis. Multi-armed bandits with switching penalties. *IEEE Trans, on Aut. Cont.*, 41:328–348, 1996.

[2] J.S. Banks and R.K. Sundaram. Switching cost and the Gittins' index. *Econometrica*, 62:687–694, 1994.

[3] F. Dusonchet and M.-O. Hongler. Multi-armed bandits with switching costs and the Gittins index. *Preprint EPFL-IPR-LPM*, 2002.

[4] M.-O. Hongler F. Dusonchet and H. Kaspi. Optimal policy for deteriorating two-armed bandit problems with switching costs. *In preparation*, 2002.

[5] J. C. Gittins. *Multi-Armed Bandits Allocation Indices.* J. Wiley, New-York, 1989.

[6] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani Ed., editor, *Progress in Statistics*, pages 241–266. North Holland., 1974.

[7] A. Ha. Optimal dynamic scheduling policy for a make-to-stock production system. *Operation Res.*, 45:42–54, 1997.

[8] H. Kaspi. Two-armed bandits with switching costs. *Preprint, Technion*, 2002.

[9] M.P. Van Oyen and D. Teneketzis. Optimal stochastic scheduling of forest networks with switching penalties. *Adv. Appl. Probab.*, 26:474–497, 1994.

[10] J. Walrand. *An introduction to Queueing Network.* Prentice-Hall International, 1988.

[11] P. Whittle. *Optimization over Time. Dynamic Programming and Stochastic Control.* J. Wiley, New-York, 1982.