DE ECONOMIST 152, NO. 4, 2004

A SURVEY ON THE BANDIT PROBLEM WITH SWITCHING COSTS**

BY

TACKSEUNG JUN*

Summary

The paper surveys the literature on the bandit problem, focusing on its recent development in the presence of switching costs. Switching costs between arms makes not only the Gittins index policy suboptimal, but also renders the search for the optimal policy computationally infeasible. This survey will first discuss the decomposability properties of the arms that make the Gittins index policy optimal, and show how these properties break down upon the introduction of costs on switching arms. Having established the failure of the simple index policy, the survey focus on the recent efforts to overcome the difficulty of finding the optimal policy in the bandit problem with switching costs: characterization of the optimal policy, exact derivation of the optimal policy in the restricted environments, and lastly approximation of optimal policy. The advantages and disadvantages of the above approaches are discussed.

Key words: decomposability, multi-armed bandits, switching costs

JEL classification: C44

1 INTRODUCTION

Consider a sequential decision problem where at each time the agent must choose one of n available actions, knowing the "state" of each action. The chosen action reveals some information about the action and the agent receives a corresponding payoff to the action. States of actions may change over time. The information received by the agent may help to decide the choice of actions in the future. The goal of the agent is to maximize the present value of the stream of payoffs he or she would receive by choosing the right sequence of actions.

The above problem is known as the "bandit" problem in the literature (Berry (1972), Whittle (1980), Berry and Fristedt (1985), Gittins (1989)). The

* School of Economics and International Trade, Department of Economics. Kyung Hee University, Seoul, 130-701, South Korea; phone: 82-2-961-0964, e-mail: tj32k@hanmail.net.

** I am indebted to Prajit K. Dutta for his guidance throughout this research. Comments from Richard Ericson, Rajiv Sethi, Lalith Munasinghe, and Levent Kockesen greatly helped me refine this paper. I also would like to thank participants of Micro Theory workshop at Columbia University. Financial support from the Kyung Hee Alumni is greatly acknowledged. All remaining errors are mine.

De Economist **152**, 513–541, 2004.

name bandit comes from an n-armed bandit, which is a slot machine with n arms. The bandit problem found applications in numerous fields: clinical trials, optimal experiments, new product development, job search, oil exploration, research & development, technology choice, and resource allocation.[1] Under certain conditions, the bandit problem can be treated as a dynamic programming problem (Bellman (1956, 1957)), which has an optimal solution given by a stationary policy. To solve for the optimal policy, one must compute every possible realization of payoffs for each arm. This makes the direct computation of the stream of payoffs for each action in every state a daunting and impractical task, since the state space of each arm is huge as it must count all possible realizations.

What makes the bandit problem tractable is Gittins' index policy. According to Gittins (1979) and Gittins and Jones (1974), the index of an arm can be defined as the maximal attainable payoff rate of the arm. Gittins (1979) showed that in the sequential problem mentioned above, it is optimal to choose an arm with the highest index. This policy is relatively simple to implement because the index of an arm depends only on the properties of the corresponding arm. This reduces the n-dimensional problem to an one-dimensional problem. This index policy greatly simplifies the search for the optimal policy in various complex decision environments (see Berninghaus et al. (1987), Banks and Sundaram (1992), Smith and Sorensen (2001) for example).

However, the optimality of the index policy does not carry over to the bandit problem where switching arms involves costs. Banks and Sundaram (1994) showed that no index policy is optimal when switching between arms incurs costs. This survey will first discuss the properties of the bandit problem that make the Gittins index policy optimal, and show how these properties break down upon the introduction of costs on switching arms. Having established the failure of the simple index policy, the survey focus on the recent efforts to overcome the difficulty of finding the optimal policy in the bandit problem in the presence of switching costs: characterization of the optimal policy, the exact derivation of the optimal policy in the restricted environments, and lastly the approximation of optimal policy.[2] This survey will also discuss the scope for the further expansions of the three approaches.

The paper is organized as follows. In the next section, some important applications of the bandit framework are briefly surveyed. In Section 3, the bandit framework is formalized. In Section 4, the optimality of the Gittins index for the bandit problem without switching costs is shown, and its inoptimality in the presence of switching cost is explained. In Section 5, the recent developments in the bandit problem in the presence of switching costs are surveyed. Section 6 concludes with a brief summary.

1 Various applications of the bandit framework are discussed in Section 2.

2 A comprehensive survey on the multi-armed bandit problem can be found at Basu et al. (1990).

2 APPLICATIONS

In the fluctuating economy with incomplete information, resources constantly move, seeking out their best use. In response to this ubiquitous phenomena, the bandit framework has been widely accepted as the basic framework of analysis. In this section, I survey the application of the bandit model in the following areas: job search, industrial policy, optimal search, experiment, and game theory.[3]

2.1 *Job Search and Labor Mobility*

The bandit framework has been applied to explain job search and labor mobility by Johnson (1978), Miller (1984), Viscusi (1980), MacDonald (1980), Waldman (1984), Jovanovic (1984), and Kennan and Walker (2003). For example, McCall and McCall (1987) applied the bandit problem to explain the migration behavior combined with job search among a set of cities, workers can observe the prevailing wage and match values only by actually moving to the city. This problem of "search-and-migrate" for worker is simply the multi-armed bandit problem with an unknown distribution of payoffs.

Moreover, the multi-armed bandit framework with a sample distribution of pay-offs for each arm[4] is identical to Jovanovic's (1979) mismatch theory. According to the theory, the productivity of the worker is match-specific. The worker faces an infinite number of identical potential employers *a priori*. Both firms and worker start with the same belief on the match quality, which is updated as the worker is employed on the job and information is obtained on the quality of match. In other words, the hypothesis for why a worker moves from one job to another is that accumulation of experience is accompanied by a sorting process in which employers and workers learn which skilled job each worker can do best. As they learn it, the worker is assigned to occupations where his kind of ability is needed. This is the model of accumulation of information about a worker's *innate* trait.[5] Casting this into the bandit framework, firms are the arms of the bandit, and the productivity of a match is the "true" distribution of payoffs from the arm. Therefore the optimization problem facing the worker is precisely a bandit problem.

3 The bandit framework has been applied to various problem in the scheduling problems. I will discuss some of related papers later in this paper, and so skip their discussion in this section.

4 See Section 3.2 for its formulation.

5 The bandit models of job search imply that worker with higher innate ability moves toward more complex and high-paying job. This hypothesis is empirically supported by studies of Wilk and Sackett (1995), Wilk et al. (1995), and Murnane et al. (1995).

2.2 *Industrial Policy*

The bandit problem is also applied to explain the industrial policy of government. Klimenko (2003) examined the problem faced by government trying to allocate country's the limited capacity among industries based on the uncertain information on foreign technologies. There is a finite set of competitive industries who try to import foreign technologies. There is uncertainty over the match quality between labor in this country and imported technology. Therefore for the policy-maker, the relevant state variable is the posterior distribution of the probability of good match between a foreign technology and domestic labor. Government can choose a finite number of firms to license to use foreign technology. This is the problem of industry targeting policy faced by many countries in the real world. The objective of the government is to maximize the overall welfare by intervening the entry of new firms with foreign technology. Casting the above problem into the bandit problem, the state is represented by the posterior distribution on match quality, and the allocation of the license to domestic firms is equivalent to choosing arms with uncertain payoffs. Since the government can allocate licences to more than one firm at a time, it is the multi-armed bandit problem with multiple plays (See Pandelis and Teneketzis (1995)).

2.3 *Optimal Search*

The bandit problem has been applied to the problem of optimal search (Weitzman (1979), Smith (1995)). Weitzman (1979) considered the model of searching for the best alternative. In the problem called Pandora's box, the agent selects a box to open at a time. The payoff from a box is probabilistically distributed. It costs agent each time he opens a box. If the agent decides to stop searching, he will get the maximum payoff that he collected so far. The Pandora's problem is a classic multi-armed bandit problem with unknown payoffs where each arm represents a box with an unknown distribution of payoffs.

The trade-off between exploitation and exploration is a classic consideration in the problem of searching for natural resources, such as oil and gas (see Benkherouf and Bather (1988), Benkherouf (1990), Benkherouf et al. (1992)). For example, an oil company has a finite set of areas for drill, and each area has an unknown amount of oil. The company may or may not explore more than one area at a time. The value of each oil field is given and drilling an oil field is costly for the company. The probability of finding a new oil field is distributed as some finite distribution. If an area is drilled and an oil field is not found, then the expected probability of finding a new oil field from this area is updated in a Bayesian fashion. The objective of the company is to find the strategy that maximizes the total expected payoffs. This

problem is the multi-armed bandit problem where oil fields represent arms with an unknown distribution of payoffs.

2.4 *Experiment and Learning*

Since the bandit problem provides the natural setting for studying the trade-off between exploitation and exploration, it has been widely applied to the problem of optimal experiments. The bandit problem was first introduced to model experiment and learning by Rothschild (1974). He modeled the market experiment problem faced by a seller who chooses in each period from finitely many prices with unknown expected returns. This model is nothing but the multi-armed bandit problem with unknown payoffs for arms. McLennan (1984) extended the model of Rothschild (1974) by allowing the seller to choose from a continuum of prices. Azoulay-Schwartz et al. (2003) and Krähmer (2003) examined the problem of experiment from the buyer's point of view where buyer can not observe the quality of products. Cowan (1991) studied the problem of experiment on adapting a technology when the true merit of the technology is unknown. Other models in the literature include Brezzi and Lai (2002), Easley and Kiefer (1988), Aghion et al. (1991), Keller and Rady (1999) and Rustichini and Wolinsky (1995). The key result of the above literature is that learning can be incomplete in a sense that there is a positive probability that the agent might settle on a suboptimal arm. This "lock-in" to an arm, regardless of the true property of the arm, is well known in the bandit literature. The intuition is as follows. If an arm, even if it is a bad one, produces a good result at each time it is played, it can advance sufficiently enough such that a good arm can not catch up.

2.5 *Game Theory*

Schlag (1998) studied decision of agents in a finite population who repeatedly choose among actions yielding uncertain payoffs. At each time, each agent is equally likely to be replaced by a new agent. The newly-born agent faces the same problem as his predecessor. On entry, each agent observes the previous choice and its payoff for his predecessor and one other agent in the population. Each agent must commit to a behavioral rule before entering the population. Clearly, the basic problem for each agent is a multi-armed bandit problem of arms with uncertain payoffs.[6]

6 Schlag (2003) studied in a similar setting to Schlag (1998) the problem where agents have a rule of minimizing regret where regret is defined as the difference between the maximal discounted expected payoff obtainable and the discounted expected payoff achieved by this rule. This model extended results obtained by Berry and Fristedt (1985) for Bernoulli two-armed bandits so that the rule of minimizing regret attains the minimax regret if and only if it is an equilibrium strategy of the agent in the zero-sum game where nature maximizes, and the agent minimizes regret.

The bandit framework is also applied to explain the behavior of individuals in ultimatum games.[7] Brenner and Vriend (2003) designed and implemented an experiment where individuals are matched to play the ultimatum game with a population of players whose behavior is dictated by some computerized algorithm. Players will offer to split a pie with his computer player. If the offer is accepted, he will get his share, and gain nothing otherwise. A player is told that (1) if the computer player accepts offer x, he will also accept any offer higher than x, and (2) if offer x is rejected by the computer player, the computer player will also reject any offer less than x. This implies that a population of computerized players is characterized by a probability density function that an offer will be accepted. Hence the problem faced by each player can be translated to the multi-armed bandit where player selects from a finite set of offers and the payoff of offers are unknown *a priori*. They showed that the predicted behavior of the model that uses the simple policy of the Gittins index approximates the actual experimental behavior data.[8] Hence they empirically showed that the Gittins index policy[9] is a rule of thumb in the experimental ultimatum games.

3 BANDIT FRAMEWORK

In the literature on bandit problems, there are two types of bandit models separate from each other by the degree of information agent possesses. In one type of models, the agent has complete knowledge on the system of the bandit problem. He knows the underlying distribution of payoffs and all other relevant statistics of the problem he faces. In the other type of models, the agent's knowledge is incomplete. This imperfection may arise from uncertainty over the underlying parameters of the payoff distribution, types of the arms, etc. The agent with incomplete information may learn the unknown part of the system by observing the consequence of his actions in order to improve the performance of his strategy in the future. In this section, I formulate the bandit problem under both complete and incomplete information.

3.1 *Complete Information Case*

Suppose that the bandit process consists of a set N of arms and $|N| = n$ is the number of arms. The discrete-time bandit problem is can be formulated as follows, following Frostig and Weiss (1999).[10] The state of the bandit system

7 For the literature on ultimatum games, see Gale et al. (1995) and Thaler (1988).

8 Theoretically, the optimal solution for player is not the Gittins index policy since offers are not independent, that is, the probability of acceptance of the offer is weakly increasing in the size of the offer.

9 See Section 4 for the formulation of the Gittins index.

10 I do not present the formulation of the continuous-time bandit problem. Interested readers may consult Karatzas (1984) and El Karoui and Karatzas (1997).

at time t is given by $S_t = (S_t^i, \ldots, S_t^n)$. S_t^i is the state of arm i at time t, and $S_t^i \in E^i$ where E^i is a countable state space. At each time t, each agent plays a subset $\bar{N}_t \subset N$ of arms and their payoffs are observed. Then the states of those arms will change according to some transition function and all other arms will remain frozen. More precisely, if arm k is chosen at time t and $s_t^k = i$, the payoff is equal to $R_t(i) \equiv R_t^k(i)$ for $k \in \bar{N}_t$. Assume that $|R_t^k(i)| \leq B$ uniformly for all i and k. The state of arm k will move from i to j according to $p^k(i, j) = \mathbf{P}(s_{t+1}^k = j | s_t^k = i)$. For all other arms, $s_{t+1}^l = s_t^l$ for all $l \neq k$ and they yield no payoffs. Let H_t denote the set of all possible histories up to time t. The strategy π of the agent specifies the subsets of arms to be played at time t given H_t. Formally, π is a sequence of measurable maps $\{\pi_t\}_{t=0}^{\infty}$ such that $\pi_t : H_t \to \mathbf{R}$. Clearly, $\bar{N}_t$ depends on π_t and thus we denote it as $\bar{N}_t(\pi_t)$. The objective of agent is to choose a policy π to maximize the infinite stream of β-discounted payoffs as follows:

$$\max_{\bar{N}_t(\pi_t) \subset N} E_{\pi_0}\left[\sum_{t=0}^{\infty} \beta^t \sum_{k \in \bar{N}_t(\pi_t)} R_t^k | \mathbf{S}_0 = \mathbf{i}\right] \qquad \text{subject to} \quad |\bar{N}_t(\pi_t)| \leq \bar{n}, \tag{1}$$

where $\mathbf{i} = (i_1, i_2, \ldots, i_n)$ and $\bar{n}$ is the maximum number of arms that can be chosen at a time. If $\bar{n} = 1$, the above problem is reduced to the bandit framework studied by Gittins (1979) and the objective function can be simply written as follows:

$$E_{\pi_0}\left\{\sum_{t=0}^{\infty} \beta^t R_t | \mathbf{S}_0 = \mathbf{i}\right\}.$$

The argument in Whittle (1982) and Ross (1983) establishes the existence of a continuous function $V(.) : \mathbf{S} \to \mathbf{R}$. $V(.)$ represents the value of the bandit problem when the states of arms are represented by $\mathbf{S}$. The Bellman equation can be written as follows:

$$V(\mathbf{i}) = \max_{k \in N}\left\{R^k(i) + \beta \sum_j p^k(i, j) V(\mathbf{j})\right\}, \tag{2}$$

where $\mathbf{j} = (i_1, i_2, \ldots, i_{k-1}, j, i_{k+1}, \ldots, i_n)$.

The bandit model can incorporate a "retirement option" as in Whittle (1980). Let $M > 0$ denote the retirement option: if agent chooses to retire, he will be given M and the game ends. Extending Eq. (2), the Bellman equation with retirement option can be written as follows:

$$V(\mathbf{i}, M) = \max_{k \in N}\left\{R^k(i) + \beta \sum_j p^k(i, j) V(\mathbf{j}, M), M\right\}. \tag{3}$$

As shown by Bellman (1957), the multi-armed bandit problem defined in (1)–(3) is a dynamic programming problem with bounded payoffs, countable state space, a finite action space, and discounted infinite horizon objectives. As the theory of dynamic programming predicts, it has an optimal solution given by a stationary policy. Also, if the transition of states is Markov decision process, the problem can be solved by linear programming (Manne (1960), Derman (1962)).

3.2 *Incomplete Information Case*

When the agent has uncertainty over the system of the bandit process, the agent faces the inherent trade-off between exploitation and exploration. Myopic objective dictates that the agent exploit the action that maximizes the immediate payoffs. On the other hand, clearing the uncertainty over the system can be valuable since it can help the agent to make a better decision in the future. However, for agent to learn about the unknown in the situation he is in, the agent may need to explore actions that may not be myopically best. When exploitation and exploration conflict, agent needs to choose the sequence of actions that maximizes the stream of payoff over the horizon. In other words, agent must take into account the quality of information the action reveals as well as the immediate payoff from it.

Uncertainty can arise from various sources. I follow the modelling strategy of Banks and Sundaram (1992) and assume that the underlying types of arms are unknown. Arm i is one of a finite number $K(i)$ of types. The true type of arm is unknown *a priori*. If the true type of arm i is $k \in \{1, \ldots, K(i)\}$, the expected payoff is R^{ik} according to the density $g^{ik}(.)$ that is, $R^{ik} = \int r g^{ik}(r)(dr)$. Without loss of generality, I assume that R^{ik} decreases in k and is bounded. The agent has a prior belief $p_t^i \in \Delta^{K(i)-1}$ on the true type distribution of arm i, with p_0^i being the initial prior.[11] Let $g_t^i(.)$ denote the expected density of payoffs from the prior distribution of p_t^i; $g_t^i(.) = \sum_k p_t^k g^{ik}(.)$, where p_t^{ik} is the k-th element of p_t^i. Then the expected payoff from playing arm i at time t is simply $R_t^i(p_t^i) = \sum_k p_t^{ik} R^{ik}$. The agent can choose one arm at a time.[12] Let r_i denote the realized payoff from arm i. Then the posterior belief for the arm i is updated, and by the independence of bandits, the posterior belief for all other arms $j \in N \backslash \{i\}$ remain unchanged. More specifically, $p_{t+1}^j = p_t^j$ for all $j \in N \backslash \{i\}$ and p_{t+1}^i is updated according to the following Bayes map: $\pi_{t+1}^i(p_t^i, R_t^i(p_t^i)) : \Delta^{K(i-1)} \times \mathbf{R} \to \Delta^{K(i-1)}$. The map is defined by $\pi_{t+1}^i(p_t^i, R_t^i, (p_t^i)) = (\pi_{t+1}^{ik}(p_t^i, R_t^i(p_t^i)))_{k=1,\ldots,K(i)}$, where $\pi_{t+1}^{ik}(p_t^i, R_t^i(p_t^i)) = p_t^{ik} R_t^{ik}(p_t^i) / \sum_{j=1}^{K(i)} p_t^{ij} r_t^{ij}(p_t^i)$. Similarly to the complete information case, the

11 $\Delta^{K(i)-1}$ is the positive unit simplex in $\mathbf{R}^n$.

12 See Pandelis and Teneketzis (1995) and Agrawal et al. (1990) for the case where agent is allowed to choose multiple arms at a time.

argument by Whittle (1982) and Ross (1983) establishes the existence of a continuous function $V(.,M):\Delta^{k-1}\to\mathbf{R}$ for each M. Hence $V(p,M)$ defines the value of the bandit system to the agent who has a prior belief p and retirement option M.[13] The Bellman equation can be written as follows:

$$V(p,M)=\max\left\{M,\,R(p)+\beta\int V(\pi(p,r),M)g(r)dr\right\}.$$

4 OPTIMAL POLICY AND THE GITTINS INDEX

4.1 *No Switching Costs*

In this section, I will develop the optimal policy for the bandit problem defined in Eq. (2). The major difficulty in obtaining the optimal policy is that the number of states becomes large very quickly as the number of arms increases: if each arm has Q possible states, the bandit problem has Q^n states. I show that the index policy developed by Gittins (1979) is optimal.[14] In the second half of this section, I will present proofs for the optimality of the Gittins index policy.

Gittins and Jones (1974) and Gittins (1979) provided a tractable solution for the optimal policy in the bandit problem of Eq. (2). They developed an index, now called the Gittins index, for each arm. In the discrete time model where agent has complete information and $\bar{n}=1$, the index for arm k is defined as follows:[15]

$$v^k(i)=\sup_{\sigma>0}v^k(i,\sigma)=\sup_{\sigma>0}\frac{\mathbf{E}\left\{\sum_{t=0}^{\sigma-1}\beta^t R_t|S_0^k=i\right\}}{\mathbf{E}\left\{\sum_{t=0}^{\sigma-1}\beta^t|S_0^k=i\right\}},\tag{4}$$

where S_0^k is the k-th element of the vector $\mathbf{S}_0$. From Eq. (4), the Gittins index is interpreted as the *maximal rate of return.* The key element for the optimality of the Gittins index is that the supremum of (4) is achieved and in fact it is achieved by the following stopping time:

$$\tau_t^k=\min\{t:v^k(\mathbf{S}_t)<v^k(i)\}.\tag{5}$$

13 $V(.,M)$ can be obtained as the unique fixed point of the contraction mapping $T:C(\Delta^{K-1})\to C(\Delta^{K-1})$ where $C(\Delta^{K-1})$ is the space of all real valued continuous functions on Δ^{K-1}. See Puterman (1994).

14 The index policy when the agent is allowed to choose more than one arm is discussed in Bergemann and Välimäki (2001). See Section 5.2.

15 Interested readers may consult Karatzas (1984) and El Karoui and Karatzas (1997) for the definition of the Gittins index in continuous time.

According to the Gittins index policy, it is optimal to choose the arm with the highest value of index. What makes the problem tractable by the Gittins index is that the computation of the index for an arm only depends the properties of the arm. This reduces the n-dimensional problem to an one-dimensional problem. Gittins (1989) also suggested computational methods to compute the index for the normal, Bernoulli and exponential distribution of payoffs. The methods involve approximating the infinite horizon problem by a finite horizon problem, using backward induction.

Theorem 1. (Gittins (1979)) *There exists functions,* $G_n(\mathbf{S}_t), n=1,2,\ldots,n$, *such that for any state* $\mathbf{S}_t$, *the policy* π^* *will activate an arm n which satisfies* $G_n(\mathbf{S})_t = \max_{1\leq m\leq N} G_m(\mathbf{S}_t)$ *is optimal. The function* $G_n(.)$ *is calculated from the dynamics of arm n alone.*

The Gittins index in the incomplete information case can be similarly defined as (4). Let z^{kt} the number of times that arm k has been played by time t. Let $p^k_{z^{kt}}$ be the posterior after arm k was played z^{kt} times which yielded the sequence of payoffs of $R^k_1, R^k_2, \ldots, R^k_{z^{kt}}$. Then the Gittins index for arm k can be defined as follows:

$$\omega^k(i) = \sup_{\sigma>0} \omega^k(i,\sigma) = \sup_{\sigma>0} \frac{\int \mathbf{E}\left\{\sum_{t=0}^{\sigma-1} \beta^t R^k_t(p^k_t) | p^k_0 = p^k_{z^{kt}}\right\} g(r)dr}{\int \mathbf{E}\left\{\sum_{t=0}^{\sigma-1} \beta^k | p^k_o = p^k_{z^{kt}}\right\} g(r)dr}. \quad (6)$$

In the rest of this section, I discuss the proofs for the optimality of the Gittins index policy. The purpose of this presentation is to clarify the underlying properties of the bandit problem that make such a simple policy optimal. Frostig and Weiss (1999) surveyed four proofs for the optimality of the Gittins index in the literature: Gittins' original proof of *interchange argument* (Gittins (1979)), Weber's *fair charge argument* (Weber (1992)), Whittle's *dual Lagrangian approach* (Whittle (1980)), and Bertsimas and Niño-Mora's *linear programming approach* of achievable region and generalized conservation laws (Bertsimas and Niño-Mora (1996)). Although these proofs are different in the details, they share a common strategy: the proof starts with a study of a single-armed bandit process, and the optimal policy is solved for in this restricted case. Then some properties of this solution are used to characterize the solution of the multi-armed bandit problem. Hence the bottom line of this methodology is that the system of the bandit process is *decomposable*, in the sense that comparison of all available arms can be based on the set of the

"representative values" of arms, each of which is computed in isolation from all other arms. This in turn implies that the decomposability of arms will fail if independence of arms breaks down.

I will discuss two of the four independent methods: the proof by Gittins (1979) and Whittle (1980).[16] Gittins' original proof of interchange argument exploits the trade-off between the "size" of the payoff and "length" of time it takes to achieve the payoff. In other words, it is certainly desirable to receive a higher payoff. However, due to discounting, it is also important to receive it as quickly as possible. The numerator of (4) captures the maximum size of the payoff, while the denominator discounts this payoff by the length of time it takes to achieve this payoff. The Gittins index achieves the optimal balance between the size and the time delay of the stream of payoffs. The proof exploits the fact that if it is optimal to choose an arm, then it is optimal to continue with the arm until its stopping time defined in (5). The reason is as follow. Given the construction in (4), as arm i is played more and more, the value of its index increases until the stopping time τ_i defined in (5). Therefore once the arm with the highest index is chosen, it is optimal to play it until its stopping time.[17]

Whittle (1980) proved the optimality of the index policy by introducing the retirement option. The Bellman equation is given in Eq. (3). The proof exploits the decomposability property of arms as follows. First, he considered the problem of arm i with a retirement option in isolation. He showed that it is optimal to continue to play arm i if $M(i) > M$ where $M(i) = \inf\{M : V(i, M) = M\}$, and retire otherwise. Then consider the case of n-armed bandit problem. If this problem were decomposable, the following policy should be optimal: retire if $M(i_j) \leq M$, and play arm j if $\max_k M(i_k) = M(i_j) > M$. He showed that this is in fact optimal.

4.2 *Switching Costs*

The original bandit problem solved by Gittins (1979) is based on the assumption that switching between arms at any time is costless. However, as Bank and Sundaram (1994) argued, it is difficult to imagine a relevant economic problem where the agent may costlessly switch between alternatives. To name a few example, a worker who switches jobs must pay non-negligible costs.[18]

16 Proof by Weber (1992) can be translated into Whittle (1980)'s proof by setting fair charge γ equal to $(1-\beta)M$.

17 This property also holds when switching arms is costly. I will discuss this in detail in Asawa and Teneketzis (1996) in Section 5.

18 Workers switching jobs entail a variety of costs: costs of learning new skills needed at the new job, or having their children adapt at a new school when the new job requires that the family relocates geographically. Mobility costs are key features of a variety of models that underpin empirical analyses of the joint determination of wages and labor mobility (e.g. Black and Loewenstein (1991), Barron et al. (1993), Kuhn (1993)).

Switching jobs from one machine to another incurs a variety of setup costs (Duenyas and Van Oyen (1996), Karaesmen and Gupta (1997), Reiman and Wein (1998), Van Oyen and Pichitlamken (1999)) and switching costs (Van Oyen and Teneketzis (1993), Koole (1997), Van Oyen et al. (1992), Kolonko and Benzing (1985)). Unfortunately, Bank and Sundaram (1994) showed that in the presence of switching costs it is not possible to define an index for each arm such that the resulting strategy produces the maximized payoffs. Below, two different arguments for the inoptimality of the index policy in the presence of switching costs are discussed. The purpose of this discussion is to clarify how the decomposability property of arms breaks down upon the introduction of switching costs, and how much complication the failure of the decomposability property would generate in computing the optimal policy.

The bandit problem in the presence of switching costs can be formulated as follows. First, Banks and Sundaram (1994) argued that any bandit problem where there are both costs of switching "away from" and "to" an arm is equivalent to another bandit problem where there is only the cost of switching to. Hence for simplicity, assume that there is only cost of switching to. Let $V(\mathbf{i}, h)$ denote the value function when the vector of states is $\mathbf{i}$ and arm $h \in N$ is the immediately-played arm. The Bellman equation when switching is costly can be obtained from (2) as follows:

$$V(\mathbf{i}, h) = \max_{k \in N} \left\{ R^k(i) - 1_{\{k \neq h\}} C^k + \beta \sum p^k(i, j) V(\mathbf{j}, k) \right\},$$

where $\mathbf{j} = (i_1, i_2, \ldots, i_{k-1}, j, i_{k+1}, \ldots, i_n)$, C^k is cost of switching to arm k, and 1_φ is equal to one iff φ holds and zero otherwise.

The inoptimality of any index policy can be understood from two different arguments. First, as Bank and Sundaram (1994), we can approach it by considering the properties that any optimal index policy, if there is any, must satisfy. If there is any possibility of switching back to the arm currently in use after abandoning it, then the index on the current arm must be *increasing* in the cost of switching to it. This is because a higher cost of switching back to the current arm makes the decision-maker more reluctant to leave the current arm. On the other hand, if switching back is a zero-probability event, the index must be independent of the cost of switching back to it. These two necessary conditions for optimality contradict each other. This implies that arms are not independent in the sense that the value of each arm is affected by the duration of plays of alternative arms, which is determined by the expected performance of the alternative arms. The independence between

arms is a necessary condition for the decomposability property of arms.[19] Bank and Sundaram (1994) showed by a numerical example that these two requirements for the optimality of the index policy are not *mutually* consistent. Later, Banks (2003) presented the nonexistence proof by explicitly showing that any index must be increasing and decreasing in switching costs at the same time.

Second, we can approach the inoptimality of index policy by translating the bandit problem in the presence of switching costs into the *restless* bandit problem. Note that in the bandit problem in the presence of switching costs, the state of the arm just abandoned changes its state from "being immediately-played arm" to "any other arm that requires switching costs if it is to be played."[20] In other words, the state of the arm that agent switches away from changes, although it is not played. Hence the bandit problem in the presence of switching costs can be translated into the restless bandit problem, introduced by Whittle (1988). The bandit problem is called restless if chosen arms as well as some or all of the unchosen arms change their states. The value function when the vector of states is $\mathbf{i}$ and retirement option is M in the restless bandit problem can be written as follows:

$$V(\mathbf{i}, M) = \max_{m \in N} \left\{ R^m(i) + \beta \sum_{k \in N} \sum_{j} p^k(i, j) V(\mathbf{j}, M), M \right\}, \tag{7}$$

where $\mathbf{j} = (j_1, j_2, \ldots, j_n)$. Whittle (1988) was not able to fully characterize the optimal policy in the restless bandit problem.[21] In fact, the restless bandit

19 The importance of independence between arms is well illustrated by Keller and Odale (2003). They considered a *branching process*, where the system of bandit problem is represented as a tree. Each node in tree represents an arm that can potentially be selected if its parent arm is selected before. A tree that starts from a node is a subtree, similar to subgame in extended games. The key feature of the model is that selection of an arm in a subtree not only yields an immediate payoff but also may deliver information on the untried arms in other subtrees. Keller and Oldale (2003) showed that the Gittins index policy remains optimal if the signal from an arm in a subtree is only informative about the payoffs of its descendent nodes in the subtree. Therefore the optimality of the Gittins index policy hinges on the independence of information structure.

20 This assumption of repeated switching cost is relaxed in Benkherouf et al. (1992) where switching cost is paid only at the first time an arm is played. The relaxation restores the optimality of the Gittins index policy in the presence of switching costs. See Section 5.2.

21 Whittle (1988) defined an index for each arm in a state as the least value of the subsidy for which it could be optimal to let the arm rest in the state. Weber and Weiss (1990) established the asymptotic optimality of Whittle's index when the differential equation describing the fluid approximation to the index policy has a globally stable equilibrium point. A different priority by index heuristic, obtained from Whittles linear programming relaxation, was developed and tested computationally by Bertsimas and Niño-Mora (2000). A polyhedral framework for analysis and computation of the Whittle index and its extensions were developed in Niño-Mora (2001).

problem belongs to the class of $\mathcal{NP}$-complete problems, which is characterized by the extreme computational load on finding the optimal policy. Therefore the decomposability property of arms and subsequent the existence of the simple index policy for the bandit problem in the presence of switching costs would contradict the difficulty of such $\mathcal{NP}$-complete problems, for which numerical solutions are usually obtained by enumerative methods such as the branch-and-bound method (Land and Doig (1960)).

5 RECENT DEVELOPMENTS

The previous section showed that the easy-to-implement tool of the index policy is not necessarily optimal in the presence of switching costs. In this section, I survey the recent efforts to overcome the difficulty of finding the optimal policy in the bandit problem with switching costs in three different ways; characterization of the optimal policy, exact derivation of the optimal policy in the restricted environments, and finally approximation of optimal policy. The advantages and disadvantages of the above approaches will be discussed.

5.1 *Characterization of the Optimal Policy*

When the exact solution of the given problem is too complex and computationally impossible, a modest approach would be to characterize the optimal policy as much as possible. This line of researches attempts to provide knowledge on the property of optimal policies as a guide for search for the optimal policy. Among such attempts, I present the work of Asawa and Teneketzis (1996) and Van Oyen and Pichitlamken (1999).

5.1.1 *Asawa and Teneketzis (1996)*

Asawa and Teneketzis (1996) attempted to determine optimal switching times. They first defined the Gittins index in presence of switching cost as follows. Suppose each arm is a deterministic sequence. Let R_l^i denote the payoff from arm i that has been used for l times. The "switching cost index" for the arm k that is not the immediately-played arm is defined as follows:

$$\tilde{v}^k(l) = \sup_{\sigma>0} \tilde{v}^k(l,\sigma) = \sup_{\sigma>0} \frac{\sum_{t=0}^{\sigma-1} \beta^t R_{t+l}^k - C}{\sum_{t=0}^{\sigma-1} \beta^t}, \tag{8}$$

where C denotes the switching cost. Let $\tilde{\tau}$, similar to the stopping time defined in (5), denote the stopping time that achieves the supremum in (8). Hence the switching cost index incorporates the cost of switching in

computing the value of the corresponding arm. The Gittins index for the immediately-played arm is defined as in (4).[22]

Asawa and Teneketzis (1996) showed that if it is optimal to play an arm at a certain time, then it is optimal to continue with it until its stopping time when the appropriate index is achieved: τ in (5) if the arm is the immediately-played arm and $\tilde{\tau}$ if it is newly-selected arm. This implies that decision about optimal choice of arms needs to be made only at times when the appropriate index of the arm is achieved. The intuition for this result can be discerned as follows. As mentioned in the original proof of the Gittins index, the value of index increases once it is selected and played more and more until its stopping time. Therefore it becomes more and more attractive to stay with the current arm as it is played given that all other arms remain frozen and yield no payoffs.

5.1.2 *Van Oyen and Pichitlamken (1999)*

Van Oyen and Pichitlamken (1999) drew upon the results from Asawa and Teneketzis (1996) and incorporated setup times for switching arms. They considered the problem of allocating a single server to a set of jobs from N families. A job in a family is designated with an instantaneous holding cost c_n and mean processing time μ_n^{-1}. There is no arrival of new jobs. Switching jobs between families incurs the random setup times D_n but switching between jobs within a family is costless. The objective is to minimize the expected costs of serving all the jobs. The special case where there is only one job in a family is completely characterized by Santos and Magazine (1985) in a deterministic model, and Van Oyen et al. (1992) in a stochastic model. Let x_n denote the number of identical jobs in family n. The index for family n is equal to

$$\chi_n = \frac{c_n \mu_n (x_n \mu_n^{-1})}{x_n \mu_n^{-1} + E(D_n)}, \tag{9}$$

and it is optimal to select the arm with the highest value of χ_n. The index given in (9) is nothing but the "$c\mu$ index" multiplied by the fraction of time over which work is performed. The corresponding $c\mu$-rule gives priority to the project with the highest delay or holding costs c_k over the expected processing time of μ_i^{-1}. Then it is optimal to select the project with the highest $c_k \mu_k^{-1}$ value.[23]

The key results of Van Oyen and Pichitlamken (1999) are as follows. First it is shown that an exhaustive policy that serves all identical jobs consecutively is optimal (Lemma 1). Second, within a family, it is optimal to process job according to *non increasing* value of $c\mu$ (Theorem 2). This is directly

22 Asawa and Teneketzis (1996) also discussed the stochastic bandit problem.

23 The optimality of the $c\mu$-rule is shown by Smith (1956) for deterministic case, and Cox and Smith (1961) for stochastic case.

analogous to the optimality of the Gittins index policy within a family, and it is proven by the interchange argument, originated by Gittins (1979). Finally, they defined the switching cost index and associated stopping time, similar to the switching cost index in (8), and showed that if it is optimal to serve a job, then it is optimal to continue until its stopping time (Theorem 3). This is qualitatively identical to the above result of Asawa and Teneketzis (1996).

5.2 *Exact Solution in Restricted Environments*

Many researches in the literature have attempted to find the exact solution of the optimal policy, but at the cost of various simplifying assumptions. Although the exact solution of the optimal policy is the merit of this approach, it often comes at the cost of limiting the scope for the further expansion. The assumptions that have been used to get the exact optimal policy include stationary distribution of arms (Bergemann and Välimäki (2001)), monotonic sequence of payoffs (Dusonchet and Hongler (2003)), first-time only switching costs (Benkherouf et al. (1992)), time-invariant payoffs (Kavadias and Loch (2000)), and two symmetric arms (Jun (2001)).

5.2.1 *Bergemann and Välimäki (2001)*

Given arms with an unknown distribution of payoffs, Bergemann and Välimäki (2001) examined the bandit problem in the presence of switching costs where more than one arm can be chosen at a time. Bergemann and Välimäki (2001) extended the Gittins index to incorporate simultaneous choice of n^* actions where $n^* < n$. However, selecting the highest Gittins index n^* arms is not necessarily optimal.[24] They constructed the optimal selection policy as follows. First, let w_0 denote the Gittins index at time 0. Let $N_0^*(\pi_0) = \{1, 2, \ldots, n^*\}$ denote the set of selected arms at time 0 under the policy π_0 and it is associated with

$$L_1^*(\pi_1) = \{i \in N_0^*(\pi_0) | w^i \geq w_0\}. \tag{10}$$

Then the pair of the sequence $\{N_t^*(\pi_t), L_{t+1}^*(\pi_{t+1})\}_{t=0}^{\infty}$ is

$$N_t^*(\pi_t) = L_{t+1}^* \pi_{(t+1)} \cup \left\{ tn^* + 1 - \sum_{j=1}^{t-1} \left|L_j^*(\pi_j)\right|, \ldots, (t+1)n^* - \sum_{j=1}^{t} \left|L_j^*(\pi_j)\right| \right\}. \tag{11}$$

24 This policy is, however, asymptotically optimal and turnpike optimal. The asymptotic optimality was proven by Ishikida and Varaiya (1994). The approach developed by Weiss (1995) for optimal scheduling of stochastic jobs on parallel server problem can be applied to prove turnpike optimality, which requires the fraction of decision times during which the prescription of the index policy contravenes the prescriptions of the optimal policy to be zero in the long-run.

In other words, the sequence of $N_t^*(\pi_t)$ operates all the arms whose Gittins indices are more than the common index w_0 of the untried arms, and leaves the arms whose indices are smaller than the value of common index. They call the sequence $\{N_t^*(\pi_t)\}_{t=o}^{\infty}$ the Gittins index n^*-policy.

For the n^*-policy defined in (10) and (11) to be optimal, the stationarity of distribution of available arms is assumed, namely the unlimited supply of untried arms. The intuition is as follows. If the distribution of arms is finite, depending on the past choices of arms, the distribution will change. Therefore the choice of arms affects the distribution of arms, and the distribution in turn affects the choice of arms. This interdependence between choice of arms and distribution of arms makes any index policy inoptimal. However, if there are unlimited supply of untried arms, the arm that is abandoned once will never be employed again. This stationarity assumption on the distribution of arms guarantees the independence of arms, which is condition we need to sustain the decomposability of the bandit problem. However, this stationarity strategy is not applicable to finite distributions where it is possible to recall once-abandoned arms. For example, it is not applicable to project development where a finite number of projects are pursued nor to the scheduling problem with finite queues. However, this restriction can be justified in job search framework where workers face a infinite number of potential employers (Mortensen (1988)).

5.2.2 *Dusonchet and Hongler (2003)*

Dusonchet and Hongler (2003) explicitly derived optimal switching thresholds in a continuous-time two-armed bandit problem. The key assumption to get the explicit solution is that the payoff is deterministic and monotonically decreases over time. The monotonically-decreasing payoffs make computation of expected payoffs very easy. Therefore the optimal switching times can be explicitly computed. The framework can be applied to a waning industry whose profitability decreases over time. However, the monotonically-decreasing payoff is a strong assumption; it is not applicable to the more general environment where payoffs are free to move upward and downward.

5.2.3 *Benkherouf et al. (1994)*

Benkherouf et al. (1994) studied the special case where switching cost is paid only at the first time the arm is played. They showed that the Gittins index is optimal (Theorem 2 of Benkherouf et al. (1994)). The intuition is as follows. As Bank and Sundaram (1994) pointed out, if there is any possibility of switching back to the arm currently in use after abandoning it, then the index on the current arm must be increasing in the cost of switching to it.

Since the concern about switching cost disappears once an arm is played, the index on the current arm does not need to increase in the cost of switching to it. Therefore this does not contradict the fact that the index on the current arm should also decrease in switching costs (Banks (2003)).

Although the one-time switching cost successfully detours the complex problem of considering potential switching costs when arms can be recalled, the model does not really deal with the problem where it is possible to recall those abandoned arms. For instance, if we interpret switching cost as a moving cost, it is incurred whenever moving happens. The model, therefore, is more suitable to the situation where switching cost is translated into (one-time) suck costs.

5.2.4 *Kavadias and Loch (2000)*

Kavadias and Loch (2000) applied the bandit framework to model the new product development problem. There are several projects, each of which must go through a fixed number of development stages to be a final product.[25] Each stage has a certain development cost. The payoff of a project only occurs at the completion of developing the project. The objective of the firm is to maximize the expected payoffs from project development by choosing the right sequence of projects. This can be modeled as a multi-armed bandit problem where each project is represented by an arm of the bandit.

More specifically, there is a set of projects $n = \{1, 2, \ldots, n\}$, time is discrete, and t_k denotes the number of periods remaining to complete the development of project k. Let x_k denote the stage of development of project k. If project k is engaged, the cost at each stage is $c_k(x_k, t_k)$ and the stage moves to y_k according to the transition probability $p(x_k, y_k)$ and the remaining time to completion becomes $t_k - 1$. The payoff at the end of completing a project depends on the state of the project at the time of completion, but it is independent of the time it takes to complete the project. Let $\pi_k(x_k)$ denote the payoff received at the time of completion.

There is a cost of switching when the working project is changed; project k has switching cost $s_k(u)$ if it is different from the immediate predecessor $u, k \neq u$. Let $V(\mathbf{x}, \mathbf{t}, M, u)$ denote the value function when the current states and remaining times until completion of each project are represented by the vectors $\mathbf{x}$ and $\mathbf{t}$, respectively, M denote retirement value, and u the immediately-processed project. The Bellman equation can be written as follows:

25 When firms attempt to develop a new product, they may pursue several potential new products simultaneously. Moreover, given the limited amount of resources, such as equipment, human capital and financial resources, how to allocate these resources to maximize the profit from successful development of new product is the central question for managers (Adler et al. (1995), Loch and Kavadias (2002)).

$$V(\mathbf{x}, \mathbf{t}, M, u) = \max_{k \in N} \left\{ -c_k(x_k, t_k) - s_k(u) + \beta \sum_{yk} p(x_k, y_k) V_k(\mathbf{x}', \mathbf{t}', M, k), M \right\},$$

where $\mathbf{x}' = (x_1, x_2, \ldots, x_{k-1}, y_k, x_{k+1}, \ldots, x_n)$ and $\mathbf{t}' = (t_1, t_2, \ldots, t_{k-1}, t_k - 1, t_{k+1}, \ldots, t_n)$ and

$$V(\mathbf{x}, \mathbf{t}, M, u) = \pi_k(x_k),$$

where $\mathbf{t}' = (t_1, t_2, \ldots, t_{k-1}, 0, t_{k+1}, \ldots, t_n)$.

Although there is a cost of switching projects, Kavadias and Loch (2000) were able to maintain the decomposability property of the Gittins index by the following two assumptions; (1) switching cost does not depend on which project is the immediate predecessor, and (2) a project's value is not affected by the delay caused by waiting for another project. The first assumption can be simply met by constant switching cost, which itself is not sufficient to sustain the optimality of the index policy. Especially, if the second assumption fails, switching cost affects the duration of plays of arms and so delays the completion of projects, and this will in turn affect the project's value. Then the argument by Bank and Sundaram (1994) can be applied and the index policy fails to achieve optimality. Consequently, the decomposability property of projects primarily hinges on the time-invariant return of a project upon its completion.

However, the time-invariant payoff is a serious drawback of the model when it is applied to the problem where timing of an event is critical for the resulting payoff of the event. In many cases, the earlier the timing of completion of a project and the following entry into the market, the higher its payoff. This is due to the first-comer advantage from having the opportunity to erect barriers to entry and discouraging potential rivals from committing the resources necessary to compete successfully (Lippman and Mamer (1993), Eswaran (1994), Bergemann and Välimäki (1999), Soberman (1999)).

5.2.5 *Jun (2001)*

Jun (2001) examined the bandit model where two arms change their states according to random walk processes, irrespective of choice of actions by agent. The model setup is as follows. There are two arms a and b and s_t^a and s_t^b, $t = 0, 1, 2, \ldots$, are the states of two arms a and b in R^1, respectively such that

$$s_t^a = \sum_{n=0}^{t} x_n^\alpha \quad \text{and} \quad s_t^b = \sum_{n=0}^{t} x_n^b,$$

where x_n^a and $x_n^b, n = 0, 1, 2, \ldots,$ are independent and identically distributed random variables taking on values $h \in R^1$and $-h$ with probabilities $p \in [0, 1]$ and $1 - p$, respectively. The agent chooses $\theta_t \in \Theta = \{a, b\}$ at each period t. If arm $i \in \{a, b\}$ is chosen at time t, the payoff is s_t^i. There is a constant switching cost C such that the (net) payoff at time t is defined as $s_t^{\theta_t}$ if $\theta_t = \theta_{t-1}$ and $s_t^{\theta_t} - C$ if $\theta_t \neq \theta_{t-1}$.

This problem is a restless bandit problem with random walk processes.[26] The argument by Whittle (1982) and Ross (1983) establishes the existence of a continuous function $V(.) : \mathbf{S} \to \mathbf{R}$; even if payoff is unbounded on the random walk without ceilings, there exists a number B_i and a constant z such that the expected payoff of arm i at time $t-1$ is bounded by $B_i n^z$ and the value function thus remains well defined. The Bellman representation can be stated as follows: for $i \in \{a, b\}$,

$$v^i(s^a, s^b) = \max_{\theta \in \{a,b\}} \left\{ s^\theta - 1_{\{\theta \neq i\}} C + \beta W^\theta \right\}, \tag{12}$$

where

$$\begin{aligned} W^i \equiv\ & p^2 v^i(s^a + h, s^b + h) + (1-p)^2 v^i(s^a - h, s^b - h) \\ & + p(1-p) v^i(s^a + h, s^b - h) + p(1-p) v^i(s^a - h, s^b + h), \end{aligned}$$

and $v^i(s^a, s^b)$ is the value function for the agent when the states of each arm are s^a and s^b and arm i is the immediately-played arm.

The search for the optimal policy can be simplified by the following argument. Consider a pair of states of two arms s^a and s^b. Consider another pair of states $s^a + l$ and $s^b + l$. If we apply whatever the optimal policy of the first pair of states to both (s^a, s^b) and $(s^a + l, s^b + l)$, the expected payoffs from

26 The restless bandit problem, which is introduced by Whittle (1988), has been studied in various directions (O'Flaherty (1987), Dusonchet and Hongler (2000), Lott and Tenekezis (2000), Niño-Mora (2000, 2004a, b), Ehsan and Liu (2004)). In particular, the restless bandit framework has been applied to model the operation of firm which involves in more than one market. He and Pindyck (1992) considered a multi-output firm whose demands of two markets evolve simultaneously. Firm must choose between specific capacity with which firm can apply to a particular market, and flexible capacity which enables the firm to supply to both markets, but it is more costly. Given that the installation of either type of capacity is irreversible, the decision problem faced by the firm is equivalent to the derivation of the option value of the flexibility. He and Pindyck (1992) did not incorporate switching costs between technologies. Similar investment models where the option value affects the decision of investment of multi-output firm includes Kulatilaka (1988), Harrison and Van Mieghem (1999), Jung (2003), and Tuluca and Stalinski (2004). Among them, Kulatilaka (1988) incorporated switching costs upon switching between modes of production, but he did not derive the optimal switching policy analytically. Overall, this literature is closely related to this model since not only alternatives (whether they are markets or arms) change their states in "restless" fashion, but also the idea of the option value is used to compute the optimal policy (McDonald and Siegel (1986), Dixit and Pindyck (1994)).

the second pair is $v^i(s^a, s^b) + l/(1-\beta)$, and so $v^i(s^a + l, s^b + l) \geq v^i(s^a, s^b) + l/(1-\beta)$ for $i \in \{a, b\}$.

In other words, this "copycat" strategy is low-bounded by $l/(1-\beta)$ because arms are symmetric. Also, if we apply whatever the optimal policy of the second pair of states to both (s^a, s^b) and $(s^a + l, s^b + l)$, the expected payoffs from the first pair is $v^i(s^a + l, s^b + l) - l/(1-\beta)$, and so $v^i(s^a, s^b) \geq v^i(s^a + l, s^b + l) - l/(1-\beta)$ for $i \in \{a, b\}$. Therefore we conclude that $v^i(s^a + l, s^b + l) - v^i(s^a, s^b) = l/(1-\beta)$, which is constant. This implies that as long as the payoff differences remain unchanged, the optimal policy must be the same, and the absolute levels of states are irrelevant to the switching decision.

Therefore, given a constant switching cost, the agent only needs to evaluate the payoff difference between the outside and the current arm to decide whether to switch or not.[27] The optimal switching rule is characterized by the minimum payoff gain, which is defined as the payoff from the outside arm minus the payoff from the current arm necessary for agent to switch arms to maximize the infinite stream of the expected payoffs in (12). Jun (2001) derived the implicit solution of this minimum payoff gain as follows.

Theorem 2. (Jun (2001)) *The agent will switch arms if the payoff from the outside arm minus the payoff from the current arm is greater than or equal to the minimum payoff gain s, which is given by*[28]

(1) $s = C(1-\beta) + 2\beta p(1-p)$ if $s < h$.

(2) $s = C(1-\beta) + 2h\rho \frac{1-\rho^{s/h}}{(1-\rho)(1+\rho^{s/h+1})}$ if $s \geq h$ and s/h is an integer.

(3) $s\left(1 - \frac{\rho^l(1+\rho)}{1+\rho^{2l+1}}\right) = C(1-\beta)\left(1 + \frac{\rho^l(1+\rho)}{1+\rho^{2l+1}}\right) + \frac{2h}{1+\rho^{2l+1}}\left(\rho\frac{1-\rho^{2l}}{(1-\rho)} - (\rho+1)k\rho^l\right)$,

if $s \geq h, s/h$ is not an integer, where $s/h < l < s/h + 1$, l is an integer, and $\rho = \frac{2p(1-p)\beta}{1-2\beta p^2-\beta+2\beta_p+\sqrt{(1-4\beta p^2-\beta+4\beta p)(1-\beta)}}$.

27 An interesting extension is to limit the number of switchings, combined with the finite horizon. One ramification of a finite number of switching opportunities is that the optimal switching policy will depend on the number of remaining switching opportunities. In fact, if other things are the same, it is expected that the minimum payoff gain for agent to switch arms is higher as the number of remaining switching opportunity diminishes. This is because the value of waiting increases as the opportunity to switch arms is smaller. However, if the horizon is finite, the optimal switching policy is expected to dictate agent to switch even at a smaller payoff gain as the horizon gets shorter. Hence the finite life of switching opportunities and the finite horizon conflict with each other in determining the optimal threshold to switch arms: the former is expected to increase the optimal threshold as the number of switching opportunities diminishes, while the latter is expected to affect the optimal threshold in the opposite way.

28 To implement the proposed policy, plug the relevant parameters of the model to the equation in Theorem 2 and solve the roots of the polynomial equation by numerical approximation methods such as Brent's method (Brent (1973)) and Secant method (Press et al. (1992)).

The model falls short on the following ground. First, although the modelling strategy here – symmetric-armed bandit – enables the derivation of the implicit solution of the underlying bandit model, it is limited in extending its result to the problem where alternatives are not symmetric. More interesting and realistic case would be that arms are heterogeneous. For example, different jobs have different wage growth paths. Projects have different dynamics of development. If the model allows such asymmetry, the minimum payoff gains will be a function of the state level of each arm.

Second, the obvious shortcoming of the model is the unrealism of the two-arm framework. For example, workers typically do not switch between the same two jobs over the life cycle of employment. Although one may argue that these two arms are the result of the agent's prior search from an outside distribution of alternatives, the question then is whether this simple two-arm framework and its theoretical implications can be interpreted within a proper search framework as in Mortensen (1988) and Jovanovic (1979). Hence an interesting extension of the model is to incorporate mismatch theory in this model. Suppose that the outside arms arises from a stationary distribution. Hence *ex ante* all arms must offer the same initial payoff. However, once an arm commences, the evolution of payoffs are determined by two distinct stochastic processes – i.e., by a random walk component, which is exogenously given, and a learning component, which arises from Bayesian updating. With the further assumption that changes of arms entail an irrecoverable switching cost, this framework clearly maintains all the elements of the mismatch theory (Jovanovic (1979)) and the exogenous evolution of payoffs in this model. Whether the prediction of this model is still maintained (and if so under what condition) is an open question worthwhile pursuing.

5.3 *Asymptotically Optimal Policy*

Finally many researches in the literature have attempted to find the nearly optimal policy by approximation. Such near-optimal solutions often exist even in the problem where finding an exact optimal solution is $\mathcal{NP}$-hard. Although the merit of such an approach is to maintain the general structure of the problem, its finite-sample performance is often not as good as its asymptotic performance. However, a simple heuristic whose performance is arbitrarily close to optimal policy has advantage in practical application.

When only one arm can be selected each time in n-armed bandit problem, Agrawal et al. (1988) constructed the policy that attains the asymptotic lower bound of the regret for every fixed unknown parameters $(\theta_1, \theta_2, \ldots, \theta_n)$ of arms so that the total switching cost up to time t is of a smaller order than the regret. The regret is nothing but the difference between the sum of expected payoffs and the sum of pay-offs from the best arm, which is unknown *a priori*. This construction was originally developed by Lai and

Robbins (1985). The switching regret is the corresponding switching costs when true parameters are unknown. More specifically, Agrawal et al. (1988) divided time into "frames" and further divided each frame f into blocks of equal length $\max\{f, 1\}$ such that $m_f - m_{f-1} = \left[\left(2^{f^2} - 2^{(f-1)^2}\right)/f\right] nf$ for $f \geq 1$ where m_f denotes the time instant at the end of frame f. The pair (f, i) denotes block i in frame f. The time instant t when (f, i) begins is a comparison instant at which upper confidence bounds U_j for the expected payoff $\mu(\theta_j)$ is computed and the arm with the largest U_j is selected for the entire block (f, i). Agrawal et al. (1988) showed that the proposed policy is asymptotically optimal.

The upper bound U_j used by Agrawal et al. (1988) is identical to that of Lai and Robbins (1985) and does not involve the horizon or discount factor. Lai (1987) improved the finite-sample performance of the policy designed by Agrawal et al. (1988) by incorporating the horizon in computing U_j. Brezzi and Lai (2002) took one step further and extended the definition of blocks used in Agrawal et al. (1988) by incorporating the basic parameters of the model to compute the upper bound. They showed these modifications of the basic structure of block allocation scheme of approximation provide a nearly optimal solution to the bandit problem in the presence of switching costs.

Agrawal et al. (1990) extended Agrawal et al. (1988) to incorporate multiple plays of arms at a time. If the underlying parameters are unknown and $\bar{n}$ of n arms, $\bar{n} < n$, are to be selected at each time, the optimal policy should play the best $\bar{n}$ arms. Since the best arms are unknown *a priori*, the sampling regret can be defined as the difference between the sum of expected payoffs and the sum of payoffs from the best $\bar{n}$ arms. Hence the criterion for the optimal policy is to minimize the "total regret", which is the sum of sampling regret and switching regret. Clearly a uniformly good policy is the one that minimizes the total regret for the entire range of unknown parameters, which is impossible to find. Agrawal et al. (1990) constructed an allocation scheme that achieves this bound asymptotically as follows. Clearly any asymptotically efficient policy must ensure that the number of plays of inferior arms must be very small. However, since the agent does not know *a priori* the time interval in which inferior arms are played, the plays of any arms are grouped together in a "block" of time interval in such a way that the contribution to switching costs by a particular arm must be very small. The block allocation scheme by Agrawal et al. (1990) chooses the interval of time in which the same arm is played such that the expected number of plays of inferior arms is controlled under a upper bound. As a consequence, the expected number of switches is controlled correspondingly and is of smaller order than the regret. Using this result, Agrawal et al. (1990), similar to Agrawal et al. (1988), showed that the proposed allocation scheme is asymptotically optimal.

6 CONCLUSIONS

This paper surveys the literature of the bandit problem and its related literature, focusing on the recent development on the bandit problem in the presence of switching costs. Switching costs between arms not only make the index policy inoptimal, but also renders the search for the optimal policy computationally infeasible. As a response to this challenge, researchers have taken three different ways to get the optimal policy as close as possible; characterization of the optimal policy, exact derivation of the optimal policy in the restricted environments, and finally approximation of optimal policy. Characterization of the optimal policy is expected to guide the future research for the optimal policy. While the exact derivation of the optimal policy is possible in restricted environments, these very restrictions limit the applicability of the models. Asymptotic approaches are expected to complement the above two approaches. Although the finite-sample performance of the proposed policy under asymptotic approaches is often not as good as its asymptotic performance, its simple structure has practical advantages.

REFERENCES

Adler, P.S., A. Mandeldaum, V. Nguyen, and R. Schwerer (1995), 'From Project to Process Management: An Empirically-Developed Framework for Analyzing Product Development Time,' *Management Science*, 41, pp. 458–484.

Aghion, P., P. Bolton, and C. Harris (1991), 'Optimal Learning by Experiment,' *Review of Economic Studies*, 58, pp. 621–654.

Agrawal, R., M. Hegde, and D. Teneketzis (1988), 'Asymptotically Efficient Allocations Rules for Multi-armed Bandit Problem with Switching Cost,' *IEEE Transactions on Automatic Control*, AC-32, pp. 968–982.

Agrawal, R., M. Hegde, and D. Teneketzis (1990), 'Multi-armed Bandit Problems with Multiple Plays and Switching Cost,' *Stochastics and Stochastic Reports*, 29, pp. 437–459.

Asawa, M. and D. Teneketzis (1996), 'Multi-Armed Bandits with Switching Penalties,' *IEEE Transactions on Automatic Control*, 41, pp. 328–348.

Azoulay-Schwartz, R., S. Kraus, and J. Wilkenfeld (2003), 'Exploitation vs. Exploration: Choosing a Supplier in an Environment of Incomplete Information,' mimeo.

Banks, J.S. (2003), 'Generalized Bandit Problems,' mimeo.

Banks, J.S. and R.K. Sundaram (1992), 'Denumerable-Armed Bandits,' *Econometrica*, 60, pp. 1071–1096.

Banks, J.S. and R.K. Sundaram (1994), 'Switching Costs and the Gittins index,' *Econometrica*, 62, pp. 687–694.

Barron, J.M., D.A. Black, and M.A. Loewenstein (1993), 'Gender Difference in Training, Capital, and Wages,' *Journal of Human Resources*, 28, pp. 343–364.

Basu, A., A. Bose and J.K. Ghosh (1990), An Expository Review of Sequential Design and Allocation Rules, Technical Report 90-08, Department of Statistics, Purdue University.

Bellman, R. (1956), 'A Problem in the Sequential Design of Experiments,' *Sankhya*, 16, pp. 221–229.

Bellman, R. (1957), *Dynamic Programming*, New Jersey, Princeton University Press.

Benkherouf, L. (1990), 'Optimal Stopping in Oil Exploration with Small and Large Oilfields,' *Probability in Engineering and Information Sciences*, 28, pp. 529–543.

Benkherouf, L. and J.A. Bather (1988), 'Oil Exploration: Sequential Decisions in the Face of Uncertainty,' *Journal of Applied Probability*, 28, pp. 529–543.

Benkherouf, L., K.D. Glazebrook, and R.W. Owen (1992), 'Gittins Indices and Oil Exploration,' *Journal of Royal Statistical Society Serial B*, 54, pp. 229–241.

Bergemann, D. and J. Välimäki (1999), 'Entry and Innovation in Vertically Differentiated Markets,' mimeo.

Bergemann, D. and J. Välimäki, (2001), 'Stationary Multi Choice Bandit Problems,' *Journal of Economic Dynamics and Control*, 25, pp. 1585–1594.

Berninghaus, S., V. Seifert, and G. Hans (1987), 'International Migration under Incomplete Information,' *Schweizerische Zeitschrift fur Volkswirtschaft und Statistik*, 123, pp. 199–218.

Bertsimas, D. and J. Niño-Mora (1996), 'Conservation Laws, Extended Polymatroids and Multi-armed Bandit Problems,' *Mathematics of Operations Research*, 21, pp. 257–306.

Bertsimas, D. and J. Niño-Mora (2000), 'Restless Bandits, Linear Programming Relaxations, and A Primal-Dual Index Heuristic,' *Operations Research*, 48, pp. 80–90.

Berry, D.A. (1972), 'A Bernoulli Two-Armed Bandit,' *Annals of Mathematical Statistics*, 43, pp. 871–897.

Berry, D.A. and B. Fristedt (1985), *Bandit Problems: Sequential Allocation of Experiments*, London, Chapman and Hall.

Black, D.A. and M.A. Loewenstein (1991), 'Self-enforcing Labor Contracts with Costly Mobility,' *Review of Labor Economics*, 12, pp. 63–83.

Brenner, T. and N.J. Vriend (2003), 'On the Behavior of Proposers in Ultimatum Games,' mimeo.

Brent, R.P. (1973), *Algorithms for Minimization Without Derivatives*, New Jersey, Prentice-Hall.

Brezzi, M. and T.L. Lai (2002), 'Optimal Learning and Experimentation in Ban- dit Problems,' *Journal of Economic Dynamics and Control*, 27, pp. 87–108.

Cvitanić, J., L. Martellini, and F. Zapatero (2002), 'Optimal Active Management Fees,' in: E. Yücesan, C.H. Chen, J.L. Snowdon and J.M. Charnes, (eds.), *Proceedings of the 2002 Winter Simulation Conference*.

Cowan, R. (1991), 'Tortoises and Hares: Choice Among Technologies of Unknown Merit,' *Economic Journal*, 407, pp. 801–814.

Cox, D.R. and W.L. Smith (1961), *Queues*, Monographs on Statistics and Applied Probability 2, New York, Chapman & Hall.

Derman, C. (1962), 'On Sequential Decision and Markov Chains,' *Management Science*, 9, pp. 16–24.

Duenyas, I. and M.P. Van Oyen (1996), 'Heuristic Scheduling of Parallel Heterogeneous Queues with Set-Ups,' *Management Science*, 42, pp. 814–829.

Dusonchet, F. and M.O. Hongler (2000), 'Continuous time Restless Bandit an Dynamic Scheduling for Make-to-Stock Production,' mimeo.

Dusonchet, F. and M.O. Hongler (2003), 'Optimal Hysteresis for a Class of Deterministic Deteriorating Two-Armed Bandit Problem with Switching Costs,' *Automatica*, 39, pp. 1947–1955.

Easley, D. and N. Kiefer (1988), 'Controlling a Stochastic Process with Unknown Parameters,' *Econometrica*, 56, pp. 1045–1064.

Ehsan, N. and M. Liu (2004), 'On the Optimality of an Index Policy for Bandwidth Allocation with Delayed State Observation and Differentiated Services,' *IEEE INFOCOM Conference*.

El Karoui, N. and I. Karatzas (1997), 'Synchronization and Optimality for Multiarmed Bandit Problems in Continuous Time,' *Computational and Applied Mathematics*, 16, pp. 117–151.

Eswaran, M. (1994), 'Licensees as Entry Barriers,' *Canadian Journal of Economics*, 27, pp. 673–688.

Frostig, E. and G. Weiss (1999), 'Four Proofs of Gittins' Multiarmed Bandit Theorem,' *Applied Probability Trust*, pp. 1–20.

Gale, J., K. Binmore and L. Samuelson (1995), 'Learning to be imperfect: The Ultimatum Games,' *Games and Economic Behavior*, 8, pp. 56–90.

Gittins, J.C. and D.M. Jones (1974), 'A Dynamic Allocation Index for the Sequential Design of Experiments,' in: European Meeting of Statisticians, J. Gani, K. Sarkadi and I. Vince, (eds.), *Progress in Statistics*, Amsterdam, North-Holland, pp. 241–266.

Gittins, J.C. (1979), 'Bandit Processes and Dynamic Allocation Indices,' *Journal of Royal Statistical Society Serial B*, 14, pp. 148–177.

Gittins, J.C. (1989), *Multi-armed Bandit Allocation Indices*, New York, Wiley.

Harrison, J.M. and J.A. Van Mieghem (1999), 'Multi-resource Investment Strategies: Operational Hedging under Demand Uncertainty,' *European Journal of Operational Research*, 113, pp. 17–29.

He, H. and R.S. Pindyck (1992), 'Investment in Flexible Production Capacity,' *Journal of Economic Dynamics and Control*, 16, p. 575–599.

Ishikida, A.T. and P. Varaiya (1994), 'Multi-armed Bandit Problem Revisited,' *Journal of Optimization Theory and Applications*, 83, pp. 113–154.

Johnson, W.R. (1978), 'A Theory of Job Shopping,' *Quarterly Journal of Economics*, 92, pp. 261–278.

Jonsson, M. and J. Večeř (2004), 'Insider Trading in Convergent Markets,' mimeo.

Jovanovic, B. (1979), 'Job Matching and the Theory of Turnover,' *Journal of Political Economy*, 87, pp. 972–990.

Jovanovic, B. (1984), 'Matching, Turnover, and Unemployment,' *Journal of Political Economy*, 92, pp. 108–122.

Jun, T. (2001), Essays on Decision Theory: Effects of Changes in Environment on Decision, Ph.D. thesis, Columbia University, New York.

Jung, A. (2003), 'Are Product Innovation and Flexible Technology Complements?,' mimeo.

Karaesmen, F. and S.M. Gupta (1997), 'Control of Arrivals in a Finite Buffered Queue with Setup Costs,' *Journal of the Operational Research Society*, 48, pp. 1113–1122.

Karatzas, I. (1984), 'Gittins Indices in the Dynamic Allocation Problem for Diffusion Processes,' *Annals of Probability*, 12, pp. 173–192.

Kast, R., A. Lapeid and S. Pardo (2003), 'Virtual Underlying Security,' mimeo.

Kavadias, S.K. and C.H. Loch (2000), Dynamic Resource Allocation Policy in Multiproject Environments, INSEAD Working Papers, 2000/10/TM.

Keller, G. and A. Oldale (2003), 'Branching Bandits: A Sequential Search Process with Correlated Pay-offs,' *Journal of Economic Theory*, 113, pp. 302–315.

Keller, G. and S. Rady (1999), 'Optimal Experimentation in a Changing Environment,' *Review of Economic Studies*, 66, pp. 475–507.

Kennan, J. and J.R. Walker (2003), The Eff1ect of Expected Income on Individual Migration Decisions, NBER Working Papers 9585.

Klimenko, M.M. (2003), 'Industrial Targeting, Experimentation and Long-run Specialization,' *Journal of Development Economics*, forthcoming.

Kolonko, M. and H. Benzing (1985), 'The Sequential Design of Bernoulli Experiments Including Switching Costs,' *Operations Research*, 2, pp. 412–426.

Koole, G. (1997), 'Assigning a Single Server to Inhomogeneous Queues with Switching Costs,' mimeo.

Krähmer, D. (2003), 'Entry and Experimentation in Oligopolistic Markets for Experience Goods,' *International Journal of Industrial Organization*, 21, pp. 1201–1213.

Kuhn, P. (1993), 'Demographic Groups and Personnel Policy,' *Labour Economics*, 1, pp. 49–70.

Kulatilaka, N. (1988), 'Valuing the Flexibility of Flexible Manufacturing Systems,' *IEEE Transactions on Engineering Management*, 35, pp. 250–257.

Lai, T.L. (1987), 'Adaptive Treatment Allocation and the Multi-Armed Bandit Problem,' *Annals of Statistics*, 15, pp. 1091–1114.

Lai, T.L. and H. Robbins (1985), 'Asymptotically Efficient Adaptive Allocation Rules,' *Advances in Applied Mathematics*, 6, pp. 4–42.

Land, A.H. and A.G. Doig (1960), 'An Automatic Method for Solving Discrete Programming Problems,' *Econometrica*, 28, pp. 497–520.

Lippman, S.A. and J.W. Mamer (1993), 'Preemptive Innovation,' *Journal of Economic Theory*, 61, pp. 104–119.

Loch, C.H. and S.K. Kavadias (2002), 'Dynamic Portfolio Selection of NPD Programs Using Marginal Returns,' *Management Science*, 48, pp. 1227–1241.

Lott, C. and D. Tenekezis (2000), 'On the Optimality of An Index Rule in Multichannel Allocation for Single-Hop Mobile Networks with Multiple Service Classes,' *Probability in the Engineering and Informational Sciences*, 14, pp. 259–197.

MacDonald, G.M. (1980), 'Person-Specific Information in the Labor Market,' *Journal of Political Economy*, 92, pp. 1086–1120.

Manne, A. (1960), 'Linear Programming and Sequential Decisions,' *Management Science*, 6, pp. 259–267.

McCall, B.P. and J.J. McCall (1987), 'A Sequential Study of Migration and Job Search,' *Journal of Labor Economics*, 5, pp. 452–476.

McDonald, R. and D. Siegel (1986), 'The Value of Waiting to Invest,' *Quarterly Journal of Economics*, 101, pp. 707–728.

McLennan, A. (1984), 'Price Dispersion and Incomplete Learning in the Long Run,' *Journal of Economic Dynamics and Control*, 7, pp. 331–347.

Miller, R.A. (1984), 'Job Matching and Occupational Choice,' *Journal of Political Economy*, 92, pp. 1086–1120.

Mortensen, D.T. (1988), 'Wages, Separations, and Job Tenure: On-the-Job Specific Training or Matching?,' *Journal of Labor Economics*, 6, pp. 445–471.

Murnane, R., F. Levy and J. Willett (1995), The Growing Importance of Cognitive Skills in Wage Determination, NBER Working Papers, pp. 50–76.

Niño-Mora, J. (2001), 'Restless Bandit, Partial Conservation Laws and Indexibility,' *Advances in Applied Probability*, 33, pp. 77–98.

Niño-Mora, J. (2004a). Restless Bandit Marginal Productivity Indices I: Single Project Case and Optimal Control of a Make-To-Stock M/G/1 Queue, Universidad Carlos III, Departamento de Estadìstica y Econometría, WS040801.

Niño-Mora, J. (2004b). Restless BanditMarginal Productivity Indices II: Rest-less Bandit Marginal Productivity Indices II: Multi-project Case and Scheduling a Multiclass Make-

To-Order/-Stock M/G/1 Queue, Universidad Carlos III, Departamento de Estadística y Econometráa, WS040902.

O'Flaherty, B. (1987). Some Results on Two-Armed Bandits When Both Projects Vary, Columbia Department of Economics Working Paper No. 359.

Pandelis, D.G. and D. Teneketzis (1995), 'On the Optimality of the Gittins Index Rule in Multi-armed Bandits with Multiple Plays,' *Proceedings of the 34th Conference on Decision & Control*, WP13 5:30, pp. 1408–1414.

Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling (1992), 'Secant Method, False Position Method, and Ridders' Method,' in: *The Art of Scientific Computing*, 2nd ed., England, Cambridge University Press. pp. 347–352.

Puterman, M.L. (1994), *Makov Decision Processes: Discrete Stochastic Dynamic Programming*, New York, Wiley.

Reiman, M.I. and L.M. Wein (1998), 'Dynamic Scheduling of a Two-Class Queue with Setups,' *Operations Research*, 46, pp. 532–547.

Ross, S. (1983), *Introduction to Stochastic Dynamic Programming*, New York, Academic Press.

Rothschild, M. (1974), 'A Two-armed Bandit Theory of Market Pricing,' *Journal of Economic Theory*, 9, pp. 185–202.

Rustichini, A. and A. Wolinsky (1995), 'Learning about Variable Demand in the Long Run,' *Journal of Economic Dynamics and Control*, 19, pp. 1283–1292.

Santos, C. and M. Magazine (1985), 'Batching in Single Operation Manufacturing System,' *Operations Research Letters*, 4, pp. 99–103.

Schlag, K.H. (1998), 'Why Imitate, and If So, How? A Bounded Rational Approach to Multi-armed Bandits,' *Journal of Economic Theory*, 78, pp. 130–156.

Schlag, K.H. (2003), 'How to Minimize Maximum Regret under Repeated Decision-Making,' mimeo.

Smith, L. (1995), 'Optimal Job Search in a Changing World,' mimeo.

Smith. L. and P. Sorensen (2001), 'Informational Herding and Optimal Experimentation, University of Michigan,' mimeo.

Smith, W.E. (1956), 'Various Optimizers for Single-State Production,' *Naval Research Logistics Quarterly*, 3, pp. 59–66.

Soberman, D.A. (1999), 'Joint Research and Development: The Lure of Dominance,' INSEAD Working Paper, 99/18/MKT.

Subramanian, A. (2002), Managerial Flexibility, Agency Costs and Optimal Capital Structure, mimeo.

Thaler, R.H. (1988), 'The Ultimatum Game,' *Journal of Economic Perspective*, 2, pp. 195–206.

Tuluca, S. and P. Stalinski (2004), 'The Manufacturing Flexibility to Switch Products: Valuation and Optimal Strategy,' mimeo.

Van Oyen, M.P. and J. Pichitlamken (1999), 'Properties of Optimal Weighted Flowtime Policies with a Makespan Constraint and Set-up Times,' mimeo.

Van Oyen, M.P. and D. Teneketzis (1993), 'Optimal Stochastic Scheduling of Forest Network with Switching Penalties,' mimeo.

Van Oyen, M.P., D.G. Pandelis and D. Teneketzis (1992), 'Optimality of Index Policies for Stochastic Scheduling with Switching Penalties,' *Journal of Applied Probability*, 29, pp. 957–966.

Van Oyen, M.P. and J. Pichitlamken (1999), 'Properties of Optimal Weighted Flowtime Policies with a Makespan Constraint and Set-up Times,' Department of Industrial Engineering and Management Science, mimeo.

Viscusi, W.K. (1980), 'A Theory of Job Shopping: A Bayesian Perspective,' *Quarterly Journal of Economics*, 94, pp. 609–614.

Waldman, M. (1984), 'Job Assignments, Signaling and Efficiency,' *Rand Journal of Economics*, 15, pp. 255–267.

Weber, R.R. (1992), 'On the Gittins index for Multiarmed Bandits,' *Annals of Probability*, 2, pp. 1024–1033.

Weber, R.R. and G. Weiss (1990), 'On an Index Policy for Restless Bandits,' *Journal of Applied Probability*, 27, pp. 637-648.

Weiss, G. (1995), 'On Almost Optimal Priority Rules for Preemptive Scheduling of Stochastic Jobs on Parallel Machines,' *Advances in Applied Probability*, 27, pp. 827–845.

Weitzman, M.L. (1979), 'Optimal Search for the Best Alternative,' *Econometrica*, 47, pp. 641–654.

Wilk, S. and P. Sackett (1995), 'A Longitudinal Analysis of Ability-Job Complexity Fit and Job Change,' mimeo.

Wilk, S., L. Desmaris, and P. Sackett (1995), 'Gravitation to Jobs Commensurate with Ability: Longitudinal and Cross-Sectional Tests,' *Journal of Applied Psychology*, 80, pp. 79–85.

Whittle, P. (1980), 'Multi-armed Bandits and the Gittins Index,' *Journal of Royal Statistical Society Serial* B, 42, pp. 143–149.

Whittle, P. (1982), *Optimization Over Time: Dynamic Programming and Stochastic Control*, Vol. 1, New York, Wiley.

Whittle, P. (1988), 'Restless Bandits: Activity Allocation in a Changing World,' *Journal of Applied Probability*, 25A, pp. 287–298.