STUDIES ON OPTIMAL STOPPING PROBLEMS FOR MULTI-ARMED BANDIT PROCESSES

吉田, 祐治

https://doi.org/10.11501/3065586

出版情報：九州大学, 1992, 博士（理学）, 論文博士
バージョン：
権利関係：

STUDIES
ON
OPTIMAL STOPPING PROBLEMS
FOR MULTI-ARMED BANDIT PROCESSES

吉　田　祐　治

STUDIES
ON
OPTIMAL STOPPING PROBLEMS
FOR MULTI-ARMED BANDIT PROCESSES

千葉大学　教養部

吉田 祐治

Yuji YOSHIDA

College of Arts and Science, Chiba University,
Yayoi-cho, Inage-ku, Chiba 263, JAPAN

1992年11月16日提出

STUDIES
ON
OPTIMAL STOPPING PROBLEMS
FOR MULTI-ARMED BANDIT PROCESSES

吉田 祐治

Contents

Chapter 1

Preface

1.1. Preface

The multi-armed bandit problem, the origin of the word is from bandit machines in gambling, is a mathematical model for optimizing in sequential manner allocating between a number of competing projects. The well-known example is :

(a) *Goldmining* : A man owns n goldmines and a gold-miming machine. Each day he must assign the machine to one of the mines. When the machine is assigned to mine i there is a probability p_i that it extracts a proportion q_i of the gold left in the mine, and a probability $1 - p_i$ that it extracts no gold and breaks down permanently. To what sequence of mines on successive days should the machine be assigned so as to maximize the expected amount of gold mined before it breaks down?

Further the multi-armed-bandit model has many examples and applications :

(b) *Scheduling* : There are n jobs which are waiting to be processed on a single industrial machine. A problem is to determine the order of the jobs to be processed so as to minimize the total costs.

(c) *Search* : A stationary object is hidden in one of n boxes. The probability that a search of i finds the object if it in box i is q_i. The probability that the object is in box is p_i. The cost of a single search of box i is c_i. A problem is to minimize the expected costs of finding the objects in a sequential search of boxes.

(d) *Industrial research* : The manager of a team of industrial scientists has n research projects which may be carried out in any order. Loss of time to switch from project to project is negligible, and a project has been successfully completed or not in some probability. The time which the team would need to spend on project i in order to complete it has a distribution function $F_i(t)$. What policy should the manager follow in order to maximize the expected total value generated by the n projects?

(e) A problem is to choose a job when a man is faced with a number of opportunities for employment which he can investigate at a rate of one per day.

(f) A problem regarding a sequence of patients and alternative treatments in a clinic.

(g) A sever with a queue of customers; and so on.

The multi-armed bandit problem has been studied by many authors. Time may be discrete or continuous and the processes themselves may be discrete or continuous. The classical type of the multi-armed bandit problem is studied in the case of several Bernoulli processes, and later the study is extended to the case of Markov processes (see [PreSon1],

[BerFri1]). Most of the literature deals with discrete time. In such a setting, each of d arms generates an infinite sequence of random variables. An observation on a particular sequence is made by selecting the corresponding arm. The tth member of a sequence is observed if the corresponding arm is selected at time t. The classical object in bandit problems is to maximize the expected value of the payoff $\sum_{t=1}^{\infty} \alpha_t Z_t$, where Z_t, the reward process, is the variable observed at time t and α_t, the *discount rates*, are non-negative numbers ($0 \leq \alpha_t \leq 1$). A strategy is called *optimal* if it yields the maximal expected payoff. The maximal expected payoff is called the *optimal* payoff.

Several methods have been studied in order to solve the multi-armed bandit problem. They are classified as follows :

(i) *Bayesian approaches* (see [BJK1], [Bel1], [Fel1]) : Using Bayes's theorem, the multi-armed bandit problem becomes a typical *dynamic programming*. Dynamic programming is introduced by [Bel2] and is a general technique devised for sequential optimization problems. The optimal strategies and the optimal payoffs are calculated by a backward induction on time t. The backward induction is called *Bellman's equation*.

(ii) *Comparison* (see [Rob1], [Isb1]) : The second approach taken in the literature is to consider particular strategies and compare their payoffs. Taking one of the strategies optimal, certain conditions for the optimality are studied.

(iii) *Dynamic allocation indices (DAI)* (see [Git1]) : The third approach is to solve by use of DAI. The DAI was introduced by [GitJon1] and is an effective method for numerical calculation regarding the problem. DAI gives a forward induction method differently from Bayesian approach. [Whi1] rewrote the proof of [GitJon1] elegantly with the dynamic programming. The early literature regarding the DAI was studied for Markov reward processes. Recently [VWB1] relaxed the Markov property.

(iv) *Minimax approach* (see [Vog1]) : This approach is a technique in non-cooperative two-person zero-sum games. Each player selects strategies so as to maximize his own payoff, however both player's payoffs are competing since the sum of both player's payoffs is assumed to be 0 in mathematical models.

This thesis deals with three kinds of themes regarding multi-armed bandit processes. One is the optimal stopping problem for discrete-time multi-armed bandit processes with independence of arms (see Chapter 2). Another is the optimal stopping problem for continuous-time multi-armed bandit processes (see Chapter 3). We deal with the problem

on the basis of [Yos3]. The other is the multi-armed bandit game (see Chapter 4). We deal with the problem in a general form, combining the results of [Yos4] and [Yos5]. In order to analyse multi-armed bandit processes, we use the theory of multi-parameter processes.

The study of multi-parameter processes are started by [McK1]. [McK1] studied Wiener sheet in the potential theory of Markov processes :

> $\{B_s\}_{s\in R_+^2}$ is a family of random variables, where R_+ is the set of all non-negative real numbers. A partial order $\geq$ is induced on the time space R_+^2 : $r \geq s$ iff $r^i \geq s^i$ $(i = 1, 2)$ for $r = (r^1, r^2), s = (s^1, s^2) \in R_+^2$. Then $\{B_s\}_{s\in R_+^2}$ is called Wiener sheet if it satisfies
>
> $$B_{(0,0)} = 0 \quad \text{and} \quad E[B_r B_s] = \frac{|r|^2 + |s|^2 - |r - s|^2}{2} \quad (r, s \in R_+^2),$$
>
> where $|r|^2 = \sum_{i=1}^2 (r^i)^2$ for $r = (r^1, r^2) \in R_+^2$. This means that maps $r^1 \mapsto B_{(r^1, r^2)}$ and $r^2 \mapsto B_{(r^1, r^2)}$ are Brownian motions for $r = (r^1, r^2) \in R_+^2$.

Recently the optimization problem for multi-parameter processes are studied by several authors. It is to find a stochastic time-sequence on the partial ordered time space so as to maximize the total expected value of the processes. [Wal1] introduced a mathematical formulation for the time-sequence, which is called an optional increasing path :

> N is the set of all non-negative integers, d is a positive integer and e_i is the i'th unit vector in N^d. $\{Z(s)\}_{s\in N^d} = \{(Z^1(s^1), \cdots, Z^d(s^d))\}_{s=(s^1,\cdots,s^d)\in N^d}$ denotes a d-parameter process with the partial order $\geq$ on N^d and $\{\mathcal{F}_s\}_{s\in N^d}$ denotes a family of sub-σ-fields. An optional increasing path $\pi = \{\pi(t)\}_{t\in N} = \{(\pi^1(t), \cdots, \pi^d(t))\}_{t\in N}$ is a N^d-valued stochastic process satisfying (d.i) — (d.iii):
>
> (d.i) $\pi(0) = (0, 0, \cdots, 0) \in N^d$.
>
> (d.ii) For all $t \in N$ it holds that $\pi(t + 1) = \pi(t) + e_i$ for some $i = 1, \cdots, d$.
>
> (d.iii) For all $t \in N$ and all $r \in N^d$ it holds that $\{\pi(t) = r\} \in \mathcal{F}_r$.

The time spaces of multi-parameter processes are called discrete or continuous if they are N^d or R_+^d respectively. [ManVan1], [LawVan1], [KreSuc1] and [DTW1] have studied the case where the time space is a more general partial ordered set. [ManVan1] has also studied the optimal stopping problem for multi-parameter processes. The problem is to decide the optimal optional increasing paths and the optimal stopping times along the paths, and then the pairs of optional increasing paths and stopping times are called tactics. [ManVan1] has studied the problem from the dynamic programming approach.

[Man1] has studied the relation between multi-armed bandit processes and multi-parameter processes, taking optional increasing paths as strategies for the bandit processes. In Chapter 2 this thesis deals with the optimal stopping problem for multi-armed bandit processes and analyses it by use of the DAI. Regarding the optimal stopping problem for d-armed bandit processes under the assumption of independence of arms, we show that the optimal strategies and the optimal stopping times are expressed by the DAI for each arm. The advantage to analyze the optimal strategies and the optimal stopping times by use of the DAI is that we can reduce the original problem to d independent classical one-parameter optimization problems. The computation efficiency of the solutions for the reduced problem, which is represented by the linear programming, is better than to solve directly Bellman's equation derived by the dynamic programming (see [CheKat1,Section 2] and [VWB1,Section 4]).

In Chapter 3 we extend the results of Chapter 2 to the case where the reward processes are one-dimensional diffusions. Then the formulation itself of multi-armed bandit processes has difficult problems. We utilize continuous multi-parameter processes in order to solve the problem. [Wal1] has defined optional increasing paths for continuous multi-parameter processes in a different form from the discrete-time. Because in the continuous-time we cannot find optimal optional increasing paths in the family of paths satisfying the condition (d.ii). [Wal1] has given the definition in the continuous-time as follows :

> In continuous-time bandit processes an optional increasing path $\pi = \{\pi(t)\}_{t\in R_+} = \{(\pi^1(t), \cdots, \pi^d(t))\}_{t\in R_+}$ is a R_+^d-valued stochastic process satisfying (c.i) — (c.iv):
>
> (c.i) $\pi(0) = (0, 0, \cdots, 0) \in R_+^d$.
>
> (c.ii) $\{\pi^i(t)\}_{t\in R_+}$ is a non-decreasing process for each $i = 1, \cdots, d$.
>
> (c.iii) $\sum_{i=1}^d \pi^i(t) = t$ for all $t \in R_+$.
>
> (c.iv) $\{\pi(t) \leq r\} \in \mathcal{F}_r$ for all $t \in R_+$ and $r \in R_+^d$.

The conditions (c.ii) and (c.iii) are weaker than (d.ii). (d.ii) models that at every time we may select one of i's, however (c.ii) means that we are allowed to select plural i's simultaneously under the condition (c.iii), which means that the total sum of time when selecting each i always increases constantly. The continuous multi-parameter process has been studied by [Mer1], [Mil1], [Maz2] and some authors. [Maz2] has formulated them as multi-parameter Markov processes, which is constructed by the product of independent usual one-parameter Markov processes. [Maz1] deals with the optimal stopping problem of multi-parameter Markov processes from the dynamic programming approach. This

thesis deals with the problem by use of DAI and obtains the extended results of Chapter 2 to the continuous-time. We show that the optimal stopping time for the original problem equals to the sum of the smallest optimal stopping times of the one-parameter optimal stopping problems for reward processes corresponding each arm. We reduce Bellman's equation, which is represented by a free boundary problem, to a fixed boundary problem when solutions of the optimal stopping problems for each arm are known.

Chapter 4 deals with zero-sum games where in every time two players alternately either select only one of arms of bandit machines or stop them. We call these games *bandit games* for abbreviation. Player A has two kinds of decisions, i.e. selecting arms and stopping the games. We represent the former with player A's strategies π_A and the latter with his stopping times τ_A. Therefore player B also has his strategies π_B and stopping times τ_B. In the game each player A (player B) alternately selects strategies π_A (π_B) or stop the game with τ_A (τ_B) so as to maximize his own payoff under the condition that the sum of both player's payoffs is 0.

The discrete-time optimal stopping games have been introduced by [Dyn2] and generalized by [Nev1,SectionVI-6] and some authors. Various types of continuous-time optimal control problems and optimal stopping problems have been developed by [BenFri1], [Ste1] and some authors. The purpose of this chapter is to formulate the bandit game with a generalized discount and to solve them as control problems. The multi-armed bandit problems with time-dependent discount rates are studied by [BerFri1]. [BerFri1] introduced the regularity condition for discount rates as one of conditions such that myopic strategies are still optimal. In this chapter we introduce a discount rate which vary together with not only time but also strategies selected by players. We introduce backward value iterations in order to analyse multi-armed bandit games and show their convergence to Bellman's equation. We construct each player's optimal Markov strategies and optimal stopping times on the basis of Bellman's equation. Finally this chapter shows that the game has a unique optimal value and that the optimal tactics are saddle points for the game.

Chapter 2

The Optimal Stopping Problem for Discrete-Time Multi-Armed Bandit Processes

2.1. Introduction

The chapter deals with the optimal stopping problem for d-armed bandit processes under the assumption of independence of arms. We analyze optimal strategies and optimal stopping times by use of the DAI and we reduce the original problem to d independent one-parameter optimization problems.

The construction and the results at each section are as follows: In Section 2.2 we describe formulations of the optimal stopping problem for d-armed bandit processes, referring [Man1]. In Section 2.3 we investigate the optimal strategies and the optimal stopping times by use of the DAI for each arm and we prove the following results (a) — (d):

(a) By the different approach from [Gla1], Theorem 2.1 shows that the DAI for each arm give the optimal strategy and the optimal stopping time. Therefore we see that in order to solve the original problem it is sufficient to calculate the DAI for each arm.

(b) Theorem 2.2 shows that the optimal stopping time given by Theorem 2.1 is expressed explicitly as the sum of d smallest optimal stopping times for one-parameter classical optimal stopping problems. In the Markov case in order to calculate the optimal stopping region it is sufficient to solve individually d one-parameter optimal stopping problems (see also Section 2.5).

(c) We give a necessary and sufficient condition for the finiteness of the optimal stopping times given by Theorem 2.1. This condition results in the finiteness of the smallest optimal stopping times of d independent one-parameter stopping problems (see Theorem 2.2(iii) and Section 2.5).

(d) Theorem 2.3 shows that the optimal stopping time given by Theorem 2.1 is the smallest optimal stopping time in the family of stopping times along the optimal strategy of Theorem 2.1.

In Section 2.4 we show that the results of Section 2.3 still hold for the extended case with constraints. In Section 2.5 we investigate the Markov case and we characterize the optimal strategies and the optimal stopping times on the basis of Theorems 2.1 and 2.2. Moreover we investigate the linear programming calculation of optimal strategies and stopping times.

2.2. Multi-armed bandit processes

We let d, the number of arms, be a positive integer. In this section we shall formulate the optimal stopping problem for d-armed bandit processes and show fundamental lemmas. This thesis deals with the case where arms are mutually independent. Therefore we regard that d-armed bandit processes consists of d mutually independent reward processes. First we shall define reward processes, following [Man1].

Let $(\Omega, \mathcal{F}, P)$ denote a probability space. Set the time space by $N = \{0, 1, 2, \cdots\}$. For each arm $i = 1, \cdots, d$, $\mathcal{F}^i = \{\mathcal{F}_t^i\}_{t \in N}$ denotes an increasing family of completed sub-σ-fields of $\mathcal{F}$ and a bounded $\mathcal{F}^i$-adapted process $Z^i = \{Z_t^i\}_{t \in N}$ means a reward process with arm i. Moreover for $i = 1, \cdots, d$ we put σ-fields $\mathcal{F}_\infty^i = \vee_{t \in N} \mathcal{F}_t^i$ * and we let $\mathcal{M}^i$ denote the family of all $\mathcal{F}^i$-adapted stopping times. Hence we assume independence of reward processes:

Assumption (F). *$\mathcal{F}_\infty^i$ $(i = 1, \cdots, d)$ are mutually independent.*

We put its time space $T = N^d$, a d-parameter process $Z(s) = (Z^1(s^1), \cdots, Z^d(s^d))$ and sub-σ-fields $\mathcal{F}_s = \mathcal{F}_{s^1}^1 \vee \cdots \vee \mathcal{F}_{s^d}^d$ for $s = (s^1, \cdots, s^d) \in T$. Let e_i denote the i'th unit vector in T. Hence we shall define strategies. For $s = (s^1, \cdots, s^d) \in T$ we define a strategy π starting from the state where each reward process with arm i has already been selected s^i times.

Such a strategy $\pi = \{\pi(t)\}_{t \in N} = \{(\pi^1(t), \cdots, \pi^d(t))\}_{t \in N}$ is a T-valued stochastic process on $(\Omega, \mathcal{F})$ satisfying (i) — (iii):

(i) $\pi(0) = s$. (2.1)

(ii) For all $t \in N$ it holds that $\pi(t+1) = \pi(t) + e_i$ for some $i = 1, \cdots, d$. (2.2)

(iii) For all $t \in N$ and all $r \in T$ it holds that $\{\pi(t) = r\} \in \mathcal{F}_r$. (2.3)

Here $\pi^i(t)$ denotes the number of selection of arm i up to time t and $\mathcal{S}(s)$ denotes the family of all the strategies starting from s. (These strategies are called optional increasing paths.) Let β, a discount rate, be a constant $(0 < \beta < 1)$. Let $\mathbf{0}$ be the zero vector in T. For a strategy $\pi \in \mathcal{S}(\mathbf{0})$, the total expected value of the (d-armed) bandit process based on the strategy π (without stopping) is defined by

$$R^\pi = E[\sum_{t \in N} \sum_{i=1}^{d} \beta^t Z^i(\pi^i(t))(\pi^i(t+1) - \pi^i(t))]. \tag{2.4}$$

Next we formulate the optimal stopping problem for d-armed bandit processes. For $s \in T$ and a strategy $\pi \in \mathcal{S}(s)$, $\{\mathcal{F}_t^\pi\}_{t \in N}$ denotes the information available at time

*This denotes the smallest sub-σ-field containing $\{\mathcal{F}_t^i \mid t \in N\}$.

t corresponding to the strategy π and $\mathcal{M}_s^\pi$ denotes the family of all $\{\mathcal{F}_t^\pi\}_{t\in N}$-stopping times along the strategy π:

$$\mathcal{F}_t^\pi = \{\Gamma \in \mathcal{F} \mid \Gamma \cap \{\pi(t) = s'\} \in \mathcal{F}_{s'} \text{ for } s' \in T\},$$

$$\mathcal{M}_s^\pi = \{\tau \mid N \cup \{\infty\}\text{-valued random variables satisfying}$$
$$\{\tau = t\} \cap \{\pi(t) = s'\} \in \mathcal{F}_{s'} \text{ for } t \in N \text{ and } s' \in T\}.$$

Then for $s = (s^1, \cdots, s^d) \in T$, a strategy $\pi \in \mathcal{S}(s)$ and a stopping time $\tau \in \mathcal{M}_s^\pi$, the expected value of the bandit process (which is starting from the state where each reward process with arm i has already been selected s^i times and which is using a strategy π and stopped at time $\tau - 1$) is denoted by

$$V^{\pi\tau}(s) = E^{\mathcal{F}_s}[\sum_{t=0}^{\tau-1}\sum_{i=1}^{d}\beta^t Z^i(\pi^i(t))(\pi^i(t+1) - \pi^i(t))].^\dagger \tag{2.5}$$

For $s \in T$ and a strategy $\pi \in \mathcal{S}(s)$ the optimal expected values of the bandit process (starting from s and using a strategy π) are defined by

$$V^{\pi *}(s) = \text{ess sup}_{\tau\in\mathcal{M}_s^\pi : P\{\tau<\infty\}=1} V^{\pi\tau}(s). \tag{2.6}$$

Then for $s \in T$ and a strategy $\pi \in \mathcal{S}(s)$ the optimal expected values of the optimal stopping problem for d-armed bandit processes (starting from s) are defined by

$$V^{**}(s) = \text{ess sup}_{\pi\in\mathcal{S}(s)} V^{\pi *}(s).$$

Here we have the following lemma regarding the finiteness of stopping times in (2.6):

Lemma 2.1. *For $s \in T$ and a strategy $\pi \in \mathcal{S}(s)$ it holds that*

$$V^{\pi *}(s) = \text{ess sup}_{\tau\in\mathcal{M}_s^\pi} V^{\pi\tau}(s). \tag{2.7}$$

Proof. Fix any $s \in T$ and any strategy $\pi \in \mathcal{S}(s)$. Then we can easily check this lemma, by noting that $\tau \wedge t \in \mathcal{M}_s^\pi$ holds for each stopping time $\tau \in \mathcal{M}_s^\pi$ and $t \in N$. □

Next we shall introduce the DAI in order to analyze the optimal stopping problem for d-armed bandit processes. For each arm $i = 1, \cdots, d$ the DAI (for the reward process) with arm i is the process $\nu^i = \{\nu^i(t)\}_{t\in N}$ defined by

$$\nu^i(t) = \text{ess sup}_{\tau\in\mathcal{M}^i : \tau\geq t+1} \frac{E^{\mathcal{F}_t^i}[\sum_{r=t}^{\tau-1}\beta^r Z^i(r)]}{E^{\mathcal{F}_t^i}[\sum_{r=t}^{\tau-1}\beta^r]}. \tag{2.8}$$

---

†We deal with the case without terminal rewards.

Hence we define the maximum index. The maximum index is the family of the largest DAI $\nu = \{\nu(s)\}_{s \in T}$ which is defined by

$$\nu(s) = \max_{i=1,\cdots,d} \nu^i(s^i) \quad \text{for } s = (s^1, \cdots, s^d) \in T. \tag{2.9}$$

Regarding DAI, the following lemma is well-known.

Lemma 2.2. ([Man1,Theorem2]) *The essential supremum in (2.8) is attained by* $\tau^i(t)$*:*

$$\tau^i(t) = \inf\{r \geq t+1 \mid \nu^i(r) \leq \nu^i(t)\}. \tag{2.10}$$

In multi-armed bandit problems, the DAI gives us an optimal strategy.

Lemma 2.3. ([Man1,Theorem1]) *For a strategy* $\pi \in \mathcal{S}(\mathbf{0})$*,* π *is optimal for the d-armed bandit problem of (2.4) if and only if* π *is an index strategy* [‡] *, i.e., for all* $t \in N$

$$\nu(\pi(t)) = \nu^i(\pi^i(t)) \text{ on } \{\pi(t+1) = \pi(t) + e_i\} \ (i = 1, \cdots, d). \tag{2.11}$$

2.3. The optimal strategies and the optimal stopping times

In this section we investigate the optimal stopping problem for d-armed bandit processes and give optimal strategies and optimal stopping times for this problem, by the method of embedding this problem into a d+1-armed bandit problem. In order to embed this problem into a d+1-armed bandit problem we shall add one more arm 0 to the d-armed bandit process defined in Section 2.2 and define an extended d+1-armed bandit process. Let $(Z^0, \mathcal{F}^0)$ denote the reward process with arm 0 satisfying (i) and (ii):

(i) $Z^0(t) = 0$ for all $t \in N$,

(ii) $\mathcal{F}^0 = \{\mathcal{F}^0_t\}_{t \in N}$ is a non-decreasing family of sub-σ-fields of $\mathcal{F}$ such that $\mathcal{F}^0_\infty$ $(= \vee_{t \in N} \mathcal{F}^0_t)$ is independent to each σ-field $\mathcal{F}^i_\infty$ $(i = 1, \cdots, d)$.

Therefore the extended d+1-armed bandit process also satisfies the mutual independence of $\mathcal{F}^i_\infty$ $(i = 0, \cdots, d)$. For the reward processes $\{(Z^i, \mathcal{F}^i) \mid i = 0, \cdots, d\}$ we shall introduce notations of the extended d+1-armed bandit problem. Take its time space N^{d+1} and let $\overline{\mathbf{0}}$ be the zero vector in N^{d+1}. We consider a $d+1$-parameter process $((Z^0(s^0), \cdots, Z^d(s^d)), \mathcal{F}^\bullet_{s^0}, \cdots, \mathcal{F}^d_{s^d})_{(s^0,\cdots,s^d) \in N^{d+1}}$. Strategies for the extended $d+1$-armed

[‡] An index strategy means that we select (the reward process corresponding to) one of the largest dynamic allocation indices at every time.

bandit problem are N^{d+1}-valued processes which is defined in the same manner as those for d-armed bandit problems in Section 2.2. Then we denote the family of all the strategies (for the extended $d+1$-armed bandit problem) starting from $\overline{\mathbf{0}}$ by $\overline{\mathcal{S}}$. For a strategy $\overline{\pi} \in \overline{\mathcal{S}}$ we express the total expected value $\overline{R}^{\overline{\pi}}$ and the optimal expected value $\overline{R}^{*}$ of the extended $d+1$-armed bandit processes by

$$\overline{R}^{\overline{\pi}} = E[\sum_{t \in N} \sum_{i=0}^{d} \beta^t Z^i(\overline{\pi}^i(t))(\overline{\pi}^i(t+1) - \overline{\pi}^i(t))], \tag{2.12}$$

and

$$\overline{R}^{*} = \sup_{\overline{\pi} \in \overline{\mathcal{S}}} \overline{R}^{\overline{\pi}}. \tag{2.13}$$

We define the DAI ν^0 for arm 0 in the same way as (2.8). Hence it is trivial that $\nu^0(t) = 0$ for all $t \in N$. Moreover we put the maximum index in arms $i = 0, \cdots, d$ by

$$\overline{\nu}((s^0, \cdots, s^d)) := \max_{i=0,1,\cdots,d} \nu^i(s^i) \quad \text{for } (s^0, \cdots, s^d) \in N^{d+1}.$$

Then the following lemma holds regarding the relation between the optimal stopping problem for d-armed bandit processes and the extended $d+1$-armed bandit problem.

Lemma 2.4. *For a strategy $\pi \in \mathcal{S}(\mathbf{0})$ and a stopping time $\tau \in \mathcal{M}_0^{\pi}$ we define a N^{d+1}-valued stochastic process $\overline{\pi}$: for $t \in N$,*

$$\overline{\pi}(t) = ((t - \tau) \vee 0, \pi(t \wedge \tau)). \tag{2.14}$$

Then (i) and (ii) hold:

(i) $\overline{\pi} \in \overline{\mathcal{S}}$,

(ii) $\overline{R}^{\overline{\pi}} = E[V^{\pi\tau}(0)]$.

Proof. (i) Fix any strategy $\pi \in \mathcal{S}(\mathbf{0})$ and any stopping time $\tau \in \mathcal{M}_0^{\pi}$. It is sufficient to show that the strategy $\overline{\pi}$, which is defined by (2.14), satisfies $\{\overline{\pi}(t) = (s^0, s)\} \in \mathcal{F}_{s^0}^0 \vee \mathcal{F}_s$ for all $t \in N$ and all $(s^0, s) \in N^{d+1}$. Fix any $t \in N$ and any $(s^0, s) \in N \times N^d$. If $s^0 > 0$, then $\{\overline{\pi}(t) = (s^0, s)\} \cap \{t < \tau\}$ is empty. While if $s^0 = 0$, then $\{\overline{\pi}(t) = (s^0, s)\} \cap \{t < \tau\} = \{\pi(t) = s\} \cap \{t < \tau\} \in \mathcal{F}_{s^0}^0 \vee \mathcal{F}_s$. And $\{\overline{\pi}(t) = (s^0, s)\} \cap \{t \geq \tau\} = \{\tau = t - s^0\} \cap \{\pi(\tau) = s\} \in \mathcal{F}_{s^0}^0 \vee \mathcal{F}_s$. Thus we obtain (i). (ii) Since $Z^0(t) = 0$ for all $t \in N$ we have

$$\begin{aligned}
\overline{R}^{\overline{\pi}} &= E[\sum_{t \in N} \sum_{i=0}^{d} \beta^t(\overline{\pi}^i(t))(\overline{\pi}^{i+1}(t) - \overline{\pi}^i(t))] \\
&= E[\sum_{t=0}^{\tau-1} \sum_{i=0}^{d} \beta^t(\pi^i(t))(\pi^{i+1}(t) - \pi^i(t))] \\
&= E[V^{\pi\tau}(0)].
\end{aligned}$$

Therefore we obtain this lemma. □

For a strategy $\pi \in \mathcal{S}(\mathbf{0})$ we define a stopping time τ^π by

$$\tau^\pi = \inf\{t \in N \mid \nu(\pi(t) \leq 0\}. \tag{2.15}$$

Then regarding the stopping times τ^π we have the following properties.

Lemma 2.5. *The following (i) and (ii) hold:*

(i) $\tau^\pi \in \mathcal{M}_0^\pi$ *for each* $\pi \in \mathcal{S}(\mathbf{0})$.

(ii) For a strategy $\pi \in \mathcal{S}(\mathbf{0})$ *we define a stopping time* τ^π *by (2.15) and we define a strategy* $\overline{\pi}$ *in the same way as (2.14), replacing* τ *with* τ^π. *Then we have* $\overline{\pi} \in \overline{\mathcal{S}}$.

Proof. (i) is trivial from the definition of τ^π. (ii) is obtained from (i) and Lemma 2.4(i). □

Now we shall construct optimal strategies for the extended $d+1$-armed bandit problem. In Chapter 2 we take π^*, τ^{π^*} and $\overline{\pi}^*$ as follows: We take an index strategy $\pi^* \in \mathcal{S}(\mathbf{0})$ (for a d- armed bandit problem) and define a stopping time τ^{π^*} by (2.15) with the index strategy π^*. Next we define a strategy $\overline{\pi}^*$ in the same way as (2.14), replacing π and τ with π^* and τ^{π^*} respectively. Then we have the following lemma.

Lemma 2.6. *For all* $t \in N$ *it holds that* $\overline{\nu}(\overline{\pi}^*(t)) = \nu(\pi^*(t)) \cdot I_{\{t<\tau^{\pi^*}\}}$, *where* I *denotes the indicator function.*

Proof. We note $\nu^0(t) = 0$ for all $t \in N$. Fix any $t \in N$. Then since $\nu(\pi^*(t)) > 0$ on $\{t < \tau^{\pi^*}\}$, we have $\overline{\nu}(\overline{\pi}^*(t)) = \nu^0(0) \vee \nu(\pi^*(t)) = \nu(\pi^*(t))$ on $\{t < \tau^{\pi^*}\}$. Next since $\nu(\pi^*(\tau^{\pi^*})) \leq 0$, the definition of τ^{π^*} implies $\overline{\nu}(\overline{\pi}^*(t)) = \nu^0(t - \tau^{\pi^*}) \vee \nu(\pi^*(\tau^{\pi^*})) = 0$ on $\{t \geq \tau^{\pi^*}\}$. Thus we obtain this lemma. □

Hence we obtain the following property of the strategy $\overline{\pi}^*$.

Proposition 2.1. *The strategy* $\overline{\pi}^*$ *is an index strategy for the extended* $d+1$*-armed bandit problem.*

Proof. From (2.14) and Lemmas 2.6 and 2.3, for all $t \in N$ and $i = 1, \cdots, d$ we obtain

$$\overline{\nu}(\overline{\pi}^*(t)) = \nu(\pi^*(t)) = \nu^i(\pi^{*i}(t)) \quad \text{on } \{\overline{\pi}^*(t+1) - \overline{\pi}^*(t) = e_i\} \cap \{t < \tau^{\pi^*}\}.$$

On the other hand for all $t \in N$ we have

$$\overline{\nu}(\overline{\pi}^*(t)) = 0 = \nu^0(\pi^{*0}(t)) \quad \text{on } \{\overline{\pi}^*(t+1) - \overline{\pi}^*(t) = e_i\} \cap \{t \geq \tau^{\pi^*}\}.$$

Consequently $\overline{\pi}^*$ is an index strategy for the extended $d+1$-armed bandit problem. □

Now we obtain the following theorem.

Theorem 2.1. *It holds that*

$$E[V^{\pi^* \tau^{\pi^*}}(0)] = E[V^{**}(0)] = \overline{R}^*.$$

Therefore if $P\{\tau^{\pi^*} < \infty\} = 1$, *then an index strategy* π^* *is an optimal strategy and* $\tau^{\pi^*} = \inf\{t \in N \mid \nu(\pi^*(t)) \leq 0\}$ *is an optimal stopping time.*

Remark. An optimal stopping time $\tau^{\pi^*} = \inf\{t \in N \mid \nu(\pi^*(t)) \leq 0\}$ means that we should continue to select on the basis of π^* and quit this game when all DAI for each arm become non-positive.

Proof. From Proposition 2.1 $\overline{\pi}^*$ is an index strategy for the extended $d+1$-armed bandit problem. Moreover, by considering Lemma 2.3 for the extended $d+1$-armed bandit problem instead of d-armed bandit problems, we obtain that $\overline{\pi}^*$ is an optimal strategy for the extended $d+1$-armed bandit problem. Therefore we obtain

$$\overline{R}^* = \overline{R}^{\overline{\pi}^*}. \tag{2.16}$$

While from Lemma 2.4 we have $\overline{R}^{\overline{\pi}^*} = E[V^{\pi^* \tau^{\pi^*}}(0)] \leq E[V^{**}(0)] \leq \overline{R}^*$. Consequently this inequality and (2.16) complete the proof of this theorem. □

Next we shall characterize the optimal stopping time $\tau^{\pi^*} = \inf\{t \in N \mid \nu(\pi^*(t)) \leq 0\}$ by classical potential theory. We would like to express the optimal stopping time τ^{π^*} by the sum of the optimal stopping times for d one-parameter optimal stopping problems for reward processes. Therefore we shall introduce one-parameter optimal stopping problems for the reward process with each arm i.

For each arm $i = 1, \cdots, d$ we consider a one-parameter optimal stopping problem for the reward process $\{Z^i(t)\}_{t \in N}$. For $t \in N$ and $\mathcal{F}^i$-adapted stopping times τ $(\tau \geq t)$ we define the expected value $V^{i\tau}(t)$ (from time t to time $\tau - 1$) of the reward process with arm i by

$$V^{i\tau}(t) = E^{\mathcal{F}_t^i}[\sum_{r=t}^{\tau-1} \beta^r Z^i(r)], \tag{2.17}$$

where in (2.17) we define that the sum takes zero if $\tau = t$. Then for $t \in N$ we define the optimal expected value $V^{i*}(t)$ of the reward process with arm i by

$$V^{i*}(t) = \text{ess sup}_{\tau \in \mathcal{M}^i : \tau \geq t} V^{i\tau}(t). \tag{2.18}$$

Hence for $i(=1,\cdots,d)$ we put an $\mathcal{F}^i$-adapted stopping time σ_*^i by

$$\sigma_*^i = \inf\{t \in N \mid V^{i*}(t) = 0\}.$$

Then the following lemma is well-known (see [Nev1]).

Lemma 2.7. *If $P\{\sigma_*^i < \infty\} = 1$, then σ_*^i is the smallest optimal stopping time for (2.18).*

Moreover we have the following relation between the optimal expected value $V^{i*}(t)$ and the DAI $\nu^i(t)$ with arm i.

Lemma 2.8. *For $t \in N$ and $i = 1,\cdots,d$ we have (i) and (ii):*

(i) $V^{i*}(t) \geq 0$,

(ii) $\{\nu^i(t) \leq 0\} = \{V^{i*}(t) = 0\}$.

Proof. (i) is trivial, since $V^{i*}(t) \geq V^{it}(t) = 0$ for every $t \in N$ and $i = 1,\cdots,d$. (ii) Fix any $t \in N$ and $i = 1,\cdots,d$. Then for all $\mathcal{F}^i$-adapted stopping times τ $(\tau \geq t+1)$ we have

$$0 \geq \nu^i(t) \geq \frac{V^{i\tau}(t)}{E^{\mathcal{F}_t^i}[\sum_{r=t}^{\tau-1} \beta^r]} \quad \text{on } \{\nu^i(t) \leq 0\}.$$

Therefore we obtain $\{V^{i*}(t) \leq 0\} \supset \{\nu^i(t) \leq 0\}$. Together with (i) this follows that $\{V^{i*}(t) = 0\} \supset \{\nu^i(t) \leq 0\}$. The reverse inclusion is obtained similarly. Therefore (ii) holds. □

Hence we obtain the following theorem.

Theorem 2.2. *Regarding the relation between the optimal stopping time τ^{π^*} of the optimal stopping problem for d-armed bandit processes and the optimal stopping times σ_*^i of independent optimal stopping problems $(i = 1,\cdots,d)$, (i) — (iii) hold:*

(i) $\sigma_*^i = \inf\{t \in N \mid \nu^i(t) \leq 0\} = \pi^{*i}(\tau^{\pi^*})$ *for all* $i = 1,\cdots,d$,

(ii) $\tau^{\pi^*} = \sum_{i=1}^d \sigma_*^i$,

(iii) $P\{\tau^{\pi^*} < \infty\} = \prod_{i=1}^d P\{\sigma_*^i < \infty\}$.

Remark. We note (a) and (b):

(a) Regarding the condition $P\{\tau^{\pi^*} < \infty\} = 1$ in Theorem 2.1, Theorem 2.2(iii) gives a necessary and sufficient condition $P\{\sigma_*^i < \infty\} = 1$ for all $i = (1,\cdots,d)$, which is from one-parameter stopping problems (2.18) for $i = (1,\cdots,d)$ (see Section 2.5).

(b) Theorem 2.2(ii) shows that in order to calculate the optimal stopping time it is also sufficient to solve individually d one-parameter optimal stopping problems (2.18) (see Section 2.5).

Proof of Theorem 2.2. (i) Fix any $i = 1, \cdots, d$. Set $\rho^i = \inf\{t \in N \mid \nu^i(\pi^*(t)) \leq 0\}$. Then we have

$$\rho^i \leq \inf\{t \in N \mid \nu(\pi^*(t)) \leq 0\} = \tau^{\pi^*}.$$

Hence for all $t \in N$, it holds that

$$\nu(\pi^*(t)) > 0 \geq \nu^i(\pi^*(\rho^i)) \quad \text{on } \{\rho^i \leq t < \tau^{\pi^*}\}.$$

This shows that the arm i is not selected at any time t on $\{\rho^i \leq t < \tau^{\pi^*}\}$. Therefore we obtain

$$\pi^{*i}(\tau^{\pi^*}) = \pi^{*i}(\rho^i). \tag{2.19}$$

On the other hand by using Assumption(F) and Lemma 2.8, we obtain

$$\begin{aligned} \pi^{*i}(\rho^i) &= \pi^{*i}(\inf\{t \in N \mid \nu^i(\pi^*(t)) \leq 0\}) \\ &= \inf\{\pi^{*i}(t) \mid \nu^i(\pi^{*i}(t)) \leq 0\} \\ &= \inf\{r \in N \mid \nu^i(r) \leq 0\} \\ &= \sigma^i_*. \end{aligned}$$

Together with (2.19) we obtain (i). (ii) and (iii) are trivial from (i). Thus this theorem holds. □

Finally we shall show that τ^{π^*} is the smallest optimal stopping time in the family $\mathcal{M}_0^{\pi^*}$ of all $\{\mathcal{F}_t^{\pi^*}\}_{t\in N}$-stopping times along π^*. For an index strategy $\pi^* \in \mathcal{S}(\mathbf{0})$ we define a one-parameter process $\{Y_t, \mathcal{F}_t^{\pi^*}\}_{t\in N}$ along the strategy π^* and its Snell's envelope by

$$Y_t = \sum_{t=0}^{\tau-1} \sum_{i=1}^{d} \beta^r Z^i(\pi^{*i}(r))(\pi^{*i}(r+1) - \pi^{*i}(r)) \quad \text{for } t \in N, \tag{2.20}$$

$$Y_t^* = \text{ess sup}_{\tau \in \mathcal{M}_0^{\pi^*}: \tau \geq t} E^{\mathcal{F}_t^{\pi^*}}[Y_\tau] \quad \text{for } t \in N. \tag{2.21}$$

Therefore we consider a one-parameter optimal stopping problem:

$$\text{To find stopping times } \tau \in \mathcal{M}_0^{\pi^*} \text{ maximizing } E[Y_\tau]. \tag{2.22}$$

Then we have the following results concerning the smallest of optimal stopping time τ^{π^*}.

Theorem 2.3. *The optimal stopping time τ^{π^*} is the smallest optimal stopping time in the family $\mathcal{M}_0^{\pi^*}$ of stopping times along an index (i.e. optimal) strategy π^*.*

Proof. It is well-known (see [Nev1]) that the smallest optimal stopping time for the optimal stopping problem (2.22) is $\inf\{t \in N \mid Y_t^* = Y_t\}$ (say ρ). Since Theorem 2.1 implies that τ^{π^*} is an optimal stopping time for the optimal stopping problem (2.22), we have

$$\rho \leq \tau^{\pi^*}. \tag{2.23}$$

On the other hand, following Lemma 2.2, for $t \in N$ and $i = 1, \cdots, d$ we define stopping times $\sigma^i(t)$ and $\tau^i(t)$ as follows: for each $s = (s^1, \cdots, s^d) \in T$

$$\sigma^i(t) = \inf\{r \geq s^i + 1 \mid \nu^i(r) \leq \nu^i(s^i)\} \quad \text{on } \{\pi^*(t) = s\}, \tag{2.24}$$

and

$$\tau^i(t) = t + \sigma^i(t) - s^i \quad \text{on } \{\pi^*(t) = s\}. \tag{2.25}$$

Hence fix any $t \in N$ and $i = 1, \cdots, d$. Since π^* is an index strategy, the arm i is selected at every time r on $\{\pi^*(t+1) - \pi^*(t) = e_i\} \cap \{\pi^{*i}(t) \leq r < \sigma^i(t)\}$. Therefore we have $\tau^i(t) \in \mathcal{M}_0^{\pi^*}$ and then together with Lemma 2.2 and Assumption(F) we obtain

$$\begin{aligned}
&\frac{E^{\mathcal{F}_s}[\sum_{r=t}^{\tau^i(t)-1} \sum_{i=1}^{d} \beta^r Z^i(\pi^{*i}(r))(\pi^{*i}(r+1) - \pi^{*i}(r))]}{E^{\mathcal{F}_s}[\sum_{r=t}^{\tau^i(t)-1} \beta^r]} \\
&= \frac{E^{\mathcal{F}_s}[\sum_{r=s^i}^{\sigma^i(t)-1} \beta^r Z^i(r)]}{E^{\mathcal{F}_s}[\sum_{r=s^i}^{\sigma^i(t)-1} \beta^r]} \\
&= \nu^i(s^i) = \nu(s) > 0
\end{aligned}$$

on $\{\pi^*(t+1) - \pi^*(t) = e_i\} \cap \{t < \tau^{\pi^*}\} \cap \{\pi^*(t) = s\}$ for all $s = (s^1, \cdots, s^d) \in T$. So we have

$$Y_t^* - Y_t \geq E^{\mathcal{F}_t^{\pi^*}}[\sum_{r=t}^{\tau^i(t)-1} \sum_{i=1}^{d} \beta^r Z^i(\pi^{*i}(r))(\pi^{*i}(r+1) - \pi^{*i}(r))] > 0$$

on $\{\pi^*(t+1) - \pi^*(t) = e_i\} \cap \{t < \tau^{\pi^*}\}$. Since this inequality holds for each $i = 1, \cdots, d$ and $t \in N$, we obtain

$$Y_t^* > Y_t \text{ on } \{t < \tau^{\pi^*}\} \quad \text{for all } t \in N.$$

Together with (2.23) and the definition of ρ, we obtain $\tau^{\pi^*} = \rho$. Thus we obtain this theorem. □

2.4. The extended case with time constraints

We shall investigate the extended case with time constraints, referring [ManVan1,Section 4]. Let C^i be a random subset of $N \cup \{\infty\}$ satisfying $\{t \in C^i\} \in \mathcal{F}_t^i$ for all $t \in N$. (This

is called a random stopping set in [ManVan1,Section 4].) Here C^i denotes a time constraint in which we must stop the reward process with arm $i(=1,\cdots,d)$. Then $\sigma_C^i(t) = \inf\{r \geq t \mid r \in C^i\}$ denotes the smallest time at which we must stop the reward process with arm i (In Markov case σ_C^i may be represented by the entry time to a state constraint in which we must stop the reward process with arm i (see [ManVan1,Section 4.3]). Hence we introduce the following time constraints:

Time constraint (C). *We can not select the arm i any more after the time $\sigma_C^i(t)$.*

Under Time constraint(C), we deal with d-armed bandit problems to maximize the values defined as (2.4). Hence in order to analyze the d-armed bandit problems with time constraints we introduce the DAI with time constraints and its maximum index:

$$\nu_C^i(t) := \begin{cases} \text{ess sup}_{\tau \in \mathcal{M}_0^i : \tau \geq t+1} \dfrac{E^{\mathcal{F}_t^i}[\sum_{r=t}^{\tau \wedge \sigma_C^i(t)-1} \beta^r Z^i(r)]}{E^{\mathcal{F}_t^i}[\sum_{r=t}^{\tau \wedge \sigma_C^i(t)-1} \beta^r]} & \text{if } t \notin C^i \\ -\infty & \text{otherwise,} \end{cases} \tag{2.26}$$

and

$$\nu_C(s) = \max_{i=1,2,\cdots,d} \nu_C^i(s^i) \quad \text{for } s = (s^1,\cdots,s^d) \in T. \tag{2.27}$$

Next we define a stopping time $\tau_C^i(t)$:

$$\tau_C^i(t) = \begin{cases} \inf\{r \geq t+1 \mid \nu_C^i(r) \geq \nu_C^i(t)\} & \text{if } t \notin C^i \\ 0 & \text{otherwise.} \end{cases} \tag{2.28}$$

Then we have the following results.

Lemma 2.9. *The following (i) and (ii) hold:*

(i) $t \leq \tau_C^i(t) \leq \sigma_C^i(t)$ *a.s. for every* $t \notin C^i$.

(ii) $\nu_C^i(t) = \dfrac{E^{\mathcal{F}_t^i}[\sum_{r=t}^{\tau_C^i(t)-1} \beta^r Z^i(r)]}{E^{\mathcal{F}_t^i}[\sum_{r=t}^{\tau_C^i(t)-1} \beta^r]}$ *for every* $t \notin C^i$.

Proof. (i) is trivial from the definitions. By considering an adapted process

$$Y^i(t) = \sum_{r=0}^{t \wedge \sigma_C^i(t)-1} \beta^r (Z^i(r) - \nu_C^i(0)) \quad \text{for } t = 1, 2\cdots,$$

we can easily check (ii) in the same line as the proof of [Man1,Section 6.3]. □

Theorem 2.4. *For a strategy $\pi \in \mathcal{S}(0)$, π is optimal for the multi-armed bandit problem with time constraints if and only if π satisfies that for all $t \in N$ it holds that*

$$\nu_C(\pi(t)) = \nu_C^i(\pi^i(t)) \text{ on } \{\pi(t+1) = \pi(t) + e_i\} \quad \text{for some } i = 1,\cdots,d. \tag{2.29}$$

Proof. For a strategy $\pi \in \mathcal{S}(\mathbf{0})$ and $i = 1, \cdots, d$ we put

$$\sigma^i(t) = \inf\{r \geq t \mid \pi^i(r) \geq \tau_C^i(t)\}.$$

Then we obtain this theorem similarly to [Man,Sections 5.4 and 5.5] by use of Lemmas 2.11 and 2.12, since $\tau_C^i(t) \leq \sigma_C^i(t)$ and $\sigma^i(t) \leq \sum_{j=1}^d \sigma_C^j(t)$ for all $t \in N$. □

Under Time constraint(C), we may also deal with the optimal stopping problem for d-armed bandit processes by similar approach to Section 2.3. Then owing to Theorem 2.4 we may develop the same arguments as Section 2.3. Consequently we see that Theorems 2.1—2.3 still hold, by replacing DAI ν^i and index strategies π^* with ν_C^i and strategies satisfying (2.29) respectively.

2.5. The Markov case and the linear programming

In this section we shall formulate and investigate the Markov case of Section 2.3. For arms $i = 1, \cdots, d$ let $(\Omega^i, \mathcal{F}^i, P^i)$ denote probability spaces and let $X^i = (X_t^i, \mathcal{F}_t^i, P^i)_{t \in N}$ denote homogeneous Markov chains, which are mutually independent, with the state space E^i. Next we introduce a d-parameter process by their products. Set its time space $T = N^d$, its path space $\Omega = \prod_{i=1}^d \Omega^i$ and its state space $E = \prod_{i=1}^d E^i$. Then we define a d-parameter Markov process X with the state space E and its σ-fields by

$$X = (X_s)_{s \in T} = (X_{s^1}^1, \cdots, X_{s^d}^d)_{s=(s^1, \cdots, s^d) \in T},$$

$$\mathcal{F}_s = \mathcal{F}_{s^1}^1 \otimes \cdots \otimes \mathcal{F}_{s^d}^d \quad \text{for } s = (s^1, \cdots, s^d) \in T.$$

Then Assumption(F) is satisfied. Hence E^x denotes the expectation induced by the probability measure $P = \prod_{i=1}^d P^i$ with an initial state $x \in E$, and for arm $i(= 1, \cdots, d)$ E^{x^i} denotes the expectation induced by the probability measure P^i with an initial state $x^i \in E^i$. For arm $i(= 1, \cdots, d)$ let f^i be a bounded measurable function on E^i. Then a reward process with arm i is given by

$$Z^i = \{Z^i(t)\}_{t \in N} = \{f^i(X_t^i)\}_{t \in N}.$$

Moreover we express strategies and stopping times in the same manner as in Section 2.2. Now the expected value function on E (for a strategy $\pi \in \mathcal{S}(\mathbf{0})$ and a stopping time $\tau \in \mathcal{M}_0^{\pi^*}$) and the optimal value function are denoted by

$$V^{\pi\tau}(x) = E^x[\sum_{t=0}^{\tau-1} \sum_{i=1}^d \beta^t f^i(X_{\pi^i(t)}^i)(\pi^i(t+1) - \pi^i(t))],$$

and

$$V^{**}(x) = \sup_{\pi \in \mathcal{S}(\mathbf{0}), \tau \in \mathcal{M}_0^{\pi^*} : P^x\{\tau < \infty\} = 1} V^{\pi\tau}(x) \quad \text{for } x = (x^1, \cdots, x^d) \in E,$$

Next the DAI function with arm $i (= 1, \cdots, d)$ and the maximum index function are expressed by

$$\nu^i(x^i) = \sup_{\tau \in \mathcal{M}^i : \tau \geq 1} \frac{E^{x^i}[\sum_{r=0}^{\tau-1} \beta^r f^i(X_r^i)]}{E^{x^i}[\sum_{r=0}^{\tau-1} \beta^r]} \quad \text{for } x^i \in E^i, \tag{2.30}$$

and

$$\nu(x) = \max_{i=1,2,\cdots,d} \nu^i(x^i) \quad \text{for } x = (x^1, \cdots, x^d) \in E. \tag{2.31}$$

Hence for arm $i = (1, \cdots, d)$ the optimal value function on E^i of the reward process with arm i and its optimal stopping time are

$$V^{i*}(x^i) = \sup_{\tau \in \mathcal{M}^i} E^{x^i}[\sum_{r=0}^{\tau-1} \beta^r f^i(X_r^i)] \quad \text{for } x^i \in E^i, \tag{2.32}$$

$$\sigma_*^i = \inf\{t \in N \mid V^{i*}(X_t^i) = 0\}. \tag{2.33}$$

Now an index strategy $\pi^* \in \mathcal{S}(\mathbf{0})$ is represented by : For each $t \in N$, π^* satisfies

$$\nu(X_{\pi^*(t)}) = \nu^i(X^i_{\pi^{*i}(t)}) \text{ on } \{\pi^*(t+1) = \pi^*(t) + e_i\} \quad \text{for some } i = 1, \cdots, d. \tag{2.34}$$

Then the smallest optimal stopping time τ^{π^*} given at the beginning of Section 2.3 is

$$\tau^{\pi^*} = \inf\{t \in N \mid \nu(X_{\pi^*(t)}) \leq 0\}. \tag{2.35}$$

Remark. In the extended case of time constraints, the time $\sigma_C^i(0)$ (of Section 2.4) which is expressed by the entry time to a state constraint to stop the reward process with arm i (see [ManVan1,Section 4.3]). Therefore we put $\sigma_C^i = \inf\{t \in N \mid X_t^i \in C^i\}$, where a Borel subset C^i of E^i denotes a stop constraint. Then by replacing (2.30) and (2.31) respectively with the following (i) and (ii): if $x \notin C^i$, then we put

(i) $\nu_C^i(x^i) = \sup_{\tau \in \mathcal{M}^i : \tau \geq 1} \frac{E^{x^i}[\sum_{r=0}^{\tau \wedge \sigma_C^i - 1} \beta^r f^i(X_r^i)]}{E^{x^i}[\sum_{r=0}^{\tau \wedge \sigma_C^i - 1} \beta^r]}$,

(ii) $\nu_C(x) = \max_{i=1,2,\cdots,d} \nu_C^i(x^i) \quad \text{for } x = (x^1, \cdots, x^d) \in E.$

We may represent an index strategy π of (2.34) and an optimal stopping time τ^{π^*} of (2.35) similarly. Then the optimal value function (2.32) of the reward process with arm i is given by

$$\sup_{\tau \in \mathcal{M}^i} E^{x^i}[\sum_{r=0}^{\tau \wedge \sigma_C^i - 1} \beta^r f^i(X_r^i)] \quad \text{for } x^i \in E^i.$$

We shall investigate the results of Section 2.3 in the Markov case and illustrate the optimal strategies π^* and the optimal stopping times τ^{π^*} more explicitly. Set a Borel subset $B^i = \{x^i \in E^i \mid V^{i*}(x^i) = 0\}$ for arm $i = 1, \cdots, d$. From Lemma 2.8 we obtain

$$B^i = \{x^i \in E^i \mid V^{i*}(x^i) = 0\} = \{x^i \in E^i \mid \nu^i(x^i) \leq 0\} \quad \text{for arm } i = 1, \cdots, d. \tag{2.36}$$

We call B^i the optimal stopping region for one-parameter stopping problem (2.32) for $i = (1, \cdots, d)$, since $\sigma_*^i = \inf\{t \in N \mid X_t^i \in B^i\}$. Hence we put $B = \prod_{i=1}^d B^i$ and then we obtain

$$B = \{x \in E \mid V^{**}(x) = 0\} = \{x \in E \mid \nu(x) \leq 0\}. \tag{2.37}$$

We call B the optimal stopping region for the optimal stopping problem for d-armed bandit processes. Then Theorem 2.2 are described as follows:

$$\sigma_*^i = \inf\{t \in N \mid X_t^i \in B^i\} = \pi^{*i}(\tau^{\pi^*}) \quad \text{for every arm } i = 1, \cdots, d. \tag{2.38}$$

$$\tau^{\pi^*} = \textstyle\sum_{i=1}^d \sigma_*^i = \inf\{t \in N \mid X_{\pi^*(t)} \in B\}. \tag{2.39}$$

$$P\{\tau^{\pi^*} < \infty\} = \textstyle\prod_{i=1}^d P\{\sigma_*^i < \infty\}. \tag{2.40}$$

Hence on the basis of the results of Theorems 2.1 and 2.2, we obtain the following characterization of optimal strategies and stopping times:

Characterization (C). *We should continue to select one of the largest DAI in all arms at every time (i.e. on every state, since index strategies are stationary in Markov case) (see (2.34) and Theorem 2.1). If the reward process X^i with an arm i entries the optimal stopping region B^i of (2.36), then we should not select the arm i any more (see (2.38)). Finally we should quit this game when all reward processes X entry the optimal stopping region B of (2.37) (see (2.39)). Moreover τ^{π^*} is the smallest optimal stopping time in the family $\mathcal{M}_0^{\pi^*}$ of stopping times along the optimal strategy π^* (see Theorem 2.3). The condition $P\{\tau^{\pi^*} < \infty\} = 1$ is equivalent to the condition $P\{\sigma_*^i < \infty\} = 1$ for all $i(= 1, \cdots, d)$. Regarding this condition we may refer to [Shi1,Theorem23 in p.94] or [Yos2], since it is not essential that reward processes (2.32) are bounded from below.*

We shall investigate the *linear programming* (LP) calculation of optimal strategies and optimal stopping times. From Characterization(C) we see that in order to solve the optimal stopping problem for d-armed bandit processes it is sufficient to calculate the DAI ν^i $(i = 1, \cdots, d)$ and the optimal stopping regions B^i $(i = 1, \cdots, d)$. The LP calculation of the optimal stopping regions B^i is well-known in one-parameter optimal stopping problems (see [Der1,pp.109-116]). Next we shall investigate the LP calculation of the DAI ν_C^i with boundary constraints more generally than the DAI ν^i. Following

[CheKat1], we shall investigate LP to calculate the DAI with time constraints when the state space is finite. Here we deal with the time σ_C^i (of Section 2.4) which is expressed by the entry time to a state constraint to stop the reward process with arm i (see [ManVan1,Section 4.3]). Moreover we fix an arm number i and we concentrate on only the reward process with arm i Therefore in the rest of this section we shall omit the arm number i for simplicity.

For one reward process, we let the finite state space $E = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$ and we put a state constraint $C = \{x^{(m+1)}, x^{(m+2)}, \cdots, x^{(n)}\}$ $(m \leq n)$. $p(i \mid j)$ denotes the transition probability from a state $x^{(i)}$ to a state $x^{(j)}$ for $i, j = 1, \cdots, n$. We put $f^{(j)} = f(x^{(j)})$ for $j = 1, \cdots, n$. Then we have

$$Pf(x^{(i)}) = E^{x^{(i)}}[f(X_1)] = \sum_{j=1}^{n} p(i \mid j) f^{(j)} \quad \text{for } i = 1, \cdots, n.$$

Hence for $i = 1, \cdots, n.$ and real numbers M we set optimal values $V_M^{(i)}$ by (2.41) and we set the DAI with a state constraint by $\nu_C^{(i)}$:

$$V_M^{(i)} = \sup_{\tau \in \mathcal{M}} E^{x^{(i)}}[\sum_{r=0}^{\tau \wedge \sigma_C - 1} \beta^r f(X_r) + \beta^{\tau \wedge \sigma_C} M], \tag{2.41}$$

Then we can easily check that for any constant M and any $i = 1, \cdots, n$, the following (2.42) and (2.43) are equivalent:

$$\nu_C^{(i)} \leq (1 - \beta)M. \tag{2.42}$$

$$V_M^{(i)} \leq M. \tag{2.43}$$

Now in order to calculate the optimal values $V_M^{(i)}$ of one-parameter optimal stopping problem (2.41) we consider the following LP.

LP $P(M)$. *Minimize* $\sum_{j=1}^{n} U^{(j)}$ *such that*

(i) $U^{(i)} - \beta \sum_{j=1}^{n} p(i \mid j) U^{(j)} \geq f^{(i)}$ *for all* $i = 1, \cdots, m$;

(ii) $U^{(i)} \geq M$ *for all* $i = 1, \cdots, m$;

(iii) $U^{(i)} = M$ *for all* $i = m + 1, \cdots, n$.

We can easily check the following lemma, by noting that $\{V_M^{(j)} \mid j = 1, \cdots, m\}$ is the smallest β-superharmonic majorant of a constant function M.

Lemma 2.10. *LP $P(M)$ has solutions* $\{V_M^{(j)} \mid j = 1, \cdots, n\}$.

Therefore we consider the following LP in order to calculate the DAI.

LP $P^{(k)}$ $(k = 1, \cdots, m)$. *Minimize* $\sum_{j=1}^{m} U^{(j)} + mM$ *such that*

(i) $U^{(i)} + M - \beta \sum_{j=1}^{n} p(i \mid j)(U^{(j)} + M) \geq f^{(i)}$ *for all* $i \in \{1, \cdots, m\} - \{k\}$;

(ii) $M - \beta \sum_{j=1}^{n} p(k \mid j)(U^{(j)} + M) \geq f^{(k)}$;

(iii) $U^{(i)} = 0$ *for all* $i = m+1, \cdots, n$.

Theorem 2.5. *For each* $k = 1, \cdots, m$, *LP* $P^{(k)}$ *has optimal solutions* $\{V_M^{(j)} \mid j = 1, \cdots, n; M\}$ *satisfying the following (a) and (b):*

(a) $U^{(k)} = 0$;

(b) $M = V_M^{(k)}$.

Then we have:

(i) *The dynamic allocation index at is* $\nu_C^{(k)} = (1 - \beta)M$ *for* $k = 1, \cdots, m$.

(ii) *The optimal value is* $V_M^{(k)}$.

Proof. By modifying LP $P(M)$, we obtain $LP^(k)$. Then we obtain this theorem in the similar way to [CheKat1], by using Lemma 2.4 and the equivalent relation between (2.42) and (2.43). □

2.6. Appendix for Section 2.4

The following lemmas are used in Theorem 2.4 of Section 2.4. Let $(\Lambda, \mathcal{G}, P)$ denote a probability space and let $\{\mathcal{G}_t\}_{t \in N}$ be an increasing family of sub-σ-fields of $\mathcal{G}$. Let $\mathcal{T}$ be the family of all $\{\mathcal{G}_t\}_{t \in N}$-stopping times. Let $\{Y(t)\}_{t \in N}$ be a bounded $\{\mathcal{G}_t\}_{t \in N}$-adapted process satisfying $E[\sum_{r \in N} |Y(r)|] < \infty$. By considering sets $\Gamma(t) = \{\sigma \geq t\} \cap \{E^{\mathcal{G}_t}[\sum_{r=t}^{\sigma-1} \alpha(r)Y(r)] > 0\}$ ($t \in N$), we can easily check the following lemmas in the same line as [VWB1,Appendix B].

Lemma 2.11. *Let* $\{\alpha(t)\}_{t \in N}$ *be an* $\{\mathcal{G}_t\}_{t \in N}$*-adapted process satisfying* $1 \geq \alpha(t) \geq \alpha(t+1) \geq 0$ *a.s. for all* $t \in N$. *For* $\sigma \in \mathcal{T}$, *it holds that*

$$E^{\mathcal{G}_0}[\sum_{r=0}^{\sigma-1} \alpha(r)Y(r)] \leq \alpha(0) \text{ ess } \sup_{\tau \in \mathcal{T}} E^{\mathcal{G}_0}[\sum_{r=0}^{\tau \wedge \sigma - 1} Y(r)]. \tag{2.44}$$

Lemma 2.12. *Let* $\{\beta(t)\}_{t \in N}$ *be an* $\{\mathcal{G}_t\}_{t \in N}$*-adapted process satisfying* $0 \leq \beta(t) \leq \beta(t+1) \leq 1$ *a.s. for all* $t \in N$ *and let* $\sigma \in \mathcal{T}$. *If there exists a stopping time* $\tau^* \in \mathcal{T}$

satisfying that $\tau^* \leq \sigma$ *a.s. and that*

$$E^{\mathcal{G}_0}[\sum_{r=0}^{\tau^*-1} Y(r)] = \text{ess sup}_{\tau \in \mathcal{T}} E^{\mathcal{G}_0}[\sum_{r=0}^{\tau \wedge \sigma - 1} Y(r)]. \tag{2.45}$$

Then it holds that

$$\beta(0) E^{\mathcal{G}_0}[\sum_{r=0}^{\tau^*-1} Y(r)] \leq E^{\mathcal{G}_0}[\sum_{r=0}^{\tau^*-1} \beta(r) Y(r)]. \tag{2.46}$$

Chapter 3

The Optimal Stopping Problem for Multi-Armed Diffusion Bandit Processes

3.1. Introduction

The purpose of this chapter is to extend the bandit processes to multi-armed Markov processes which is constructed by the product of mutually independent one-parameter diffusion processes, which is given by a solution of stochastic differential equation (3.1) in Section 3.2, using the results of Chapter 2. A difficult problem in the continuous-time case is that in general we cannot find the optimal strategies such that player selects one of arms at every time. Therefore we should find the optimal strategies in the class where we are allowed to move plural arms simultaneously. The definition is given in Section 3.2.

This chapter shows the existence of the optimal tactic, the pair of the optimal strategy and the optimal stopping time, such that the tactic expressed by the DAI for reward processes. We give the representation of the optimal stopping time for the original problem by the smallest optimal stopping times of the one-parameter optimal stopping problems for reward processes corresponding each arm. On the basis of this fact we give a certain necessary and sufficient condition concerning the finiteness of the smallest optimal stopping time for the original problem. By deriving that the optimal stopping region of the original problem is equal to the Cartesian product of the optimal stopping regions for each arm, we reduce Bellman's equation, which is represented by a free boundary problem, of the original problem to a fixed boundary problem when solutions of the optimal stopping problems for each arm are known

Regarding the optimal stopping problem for a d-parameter Markov processes, [Maz1] has studied the case of $d = 2$ and $f^i = 0$ in (3.13) of Section 3.2. However we investigate the case of $g = 0$ in (3.13). Referring [Maz1,Section2], in Section 3.2 we formulate the optimal stopping problem for multi-armed Markov processes and optimal stopping problems for the reward process X^i for each arm i. Section 3.3 extends the results of Chapter 2 to the continuous-time case. In Section 3.4 we discuss the optimal stopping region and Bellman's equation.

3.2. Multi-armed diffusion bandit processes

We let d be a positive integer and we set $R_+ = [0, \infty)$. In Section 3.2 we shall formulate three kinds of optimization problems, namely, the optimal control problem for d-parameter Markov processes, the optimal stopping problem for d-parameter Markov processes, and the optimal stopping problem for the reward process X^i. The first problem and the third problem will be utilized in order to analyse the second problem in Section 3.3.

We shall formulate d-parameter difuusion processes and their optimal control problems, referring [Maz1,Section 2]. For $i = (1, \cdots, d)$ $(\Omega^i, \mathcal{F}^i, P^i)$ denote probability spaces and $X^i = (X^i_t, \mathcal{F}^i_t, P^{x^i})_{t \in R_+}$, which are called reward processes, denote one-parameter mu-

tually independent diffusion processes with their state spaces $E^i := R$, which is a unique solution to the stochastic differential equation (3.1) on R:

$$\mathrm{d}X_t^i = \sigma^i(X_t^i)\mathrm{d}W_t^i + b^i(X_t^i)\mathrm{d}t \quad \text{for } t \in R_+, \tag{3.1}$$

where b^i is a bounded Lipschitz continuous real-valued function on R and σ^i is a bounded Lipschitz continuous real-valued function on R such that $a^i = \sigma^i\sigma^i$ is uniformly positive; and for each $i = 1, \cdots, d$ we let $W^i = \{W_t^i\}_{t\in R_+}$ be a one-dimensional Brownian motion such that $W^1, \cdots, W^d$ are independent.

Hence $\{\mathcal{F}_t^i\}_{t\in R_+}$ is an increasing right continuous family of completed sub-σ-fields of $\mathcal{F}^i$ and P^{x^i} is a probability measure on $(\Omega^i, \mathcal{F}^i)$ with an initial state $x^i \in E^i$. For $i = (1, \cdots, d)$ we set

$$\mathcal{F}_\infty^i = \vee_{t\in R_+}\mathcal{F}_t^i. \tag{3.2}$$

Next we shall introduce a d-parameter Markov process by their products as follows. We set its time space $T = R_+^d$ and introduce the partial order in T by

$$r \leq s \quad \text{iff} \quad r^i \leq s^i \ (i = 1, \cdots, d) \quad \text{for } r = (r^1, \cdots, r^d),\ s = (s^1, \cdots, s^d) \in T. \tag{3.3}$$

And we set its path space $\Omega = \prod_{i=1}^d \Omega^i$ and its state space $E = \prod_{i=1}^d E^i$. Then we define a d-parameter Markov process X with the state space E by

$$X = (X_s)_{s\in T} = (X_{s^1}^1, \cdots, X_{s^d}^d)_{s=(s^1,\cdots,s^d)\in T}. \tag{3.4}$$

Moreover we put a right continuous family $\{\mathcal{F}_s\}_{s\in T}$ of d-parameter sub-σ-fields, which are increasing with respect to the partial order (3.3), by

$$\mathcal{F}_s = \mathcal{F}_{s^1}^1 \otimes \cdots \otimes \mathcal{F}_{s^d}^d \quad \text{for } s = (s^1, \cdots, s^d) \in T \text{ and } \mathcal{F} = \mathcal{F}^1 \otimes \cdots \otimes \mathcal{F}^d. \tag{3.5}$$

Especially E^x denotes the expectation induced by the probability measure $P = \prod_{i=1}^d P^i$ with an initial state $x \in E$, and for $i = (1, \cdots, d)$ E^{x^i} denotes the expectation by the probability measure P^i with an initial state $x^i \in E^i$. Hence a strategy $\pi = \{\pi(t)\}_{t\in R_+} = \{(\pi^1(t), \cdots, \pi^d(t))\}_{t\in R_+}$ is a T-valued stochastic process on $(\Omega, \mathcal{F})$ satisfying (i) — (iv):

(i) $\pi(0) = \mathbf{0}$. (3.6)

(ii) $\{\pi^i(t)\}_{t\in R_+}$ is a non-decreasing process for each $i = 1, \cdots, d$. (3.7)

(iii) $\sum_{i=1}^d \pi^i(t) = t \quad$ for all $t \in R_+$. (3.8)

(iv) $\{\pi(t) \leq r\} \in \mathcal{F}_r \quad$ for all $t \in R_+$ and $r \in T$. (3.9)

Moreover we denote $\Pi(0)$ the family of all strategies.

Remark. We note that

(a) These strategies are called optional increasing paths in [Wal1].

(b) Note that $0 \leq \pi^i(t') - \pi^i(t) \leq t' - t$ for all $t, t' \in R_+$ satisfying $t \leq t'$.

For $i(= 1, \cdots, d)$ we let f^i be a fixed bounded continuous, strictly increasing function on E^i and we define an expected value function R^π of total rewards over the infinite horizon R_+ associated with a strategy $\pi \in \Pi(0)$ by

$$R^\pi(x) = E^x[\int_0^\infty e^{-\alpha t} \sum_{i=1}^d f^i(X^i_{\pi^i(t)}) \mathrm{d}\pi^i(t)] \quad \text{for } x \in E, \tag{3.10}$$

where $\alpha(> 0)$ is a discount factor, $i(= 1, \cdots, d)$ is a reward process number, for each $i(= 1, \cdots, d)$ f^i is a running reward function on the current state. Hence we define the optimal value function R^* of total rewards of the optimal control problem of d-parameter Markov process X by

$$R^*(x) = \sup_{\pi \in \Pi(0)} R^\pi(x) \quad \text{for } x \in E. \tag{3.11}$$

Then the optimal control problem for d-parameter Markov process X is to find a strategy $\pi^* \in \Pi(0)$ satisfying

$$R^{\pi^*}(x) = R^*(x) \quad \text{for all } x \in E. \tag{3.12}$$

Lemma 3.1. ([Kar1,Theorem6.1]) *There exist an optimal strategy π^* of the optimal control problem (3.12) of d-parameter Markov process X. Then for the optimal strategy π^*, $(X^i_{\pi^*(t)}, \mathcal{F}^i_{\pi^*(t)}, P^{x^i})_{t \in R_+}$ becomes a standard Markov process.*

We shall formulate the optimal stopping problem for d-parameter Markov processes. For a strategy $\pi \in \Pi(0)$, $\mathcal{F}^\pi_t$ denotes the information which is available at time $t(\in R_+)$ along the strategy π and $\mathcal{M}^\pi_0$ denotes the family of all $\{\mathcal{F}^\pi_t\}_{t \in R_+}$- stopping times:

$$\mathcal{F}^\pi_t = \{\Gamma \in \mathcal{F} \mid \Gamma \cap \{\pi(t) \leq r\} \in \mathcal{F}_r \quad \text{for all } t \in R_+ \text{ and } r \in T\}; \text{ and}$$

$$\mathcal{M}^\pi_0 = \{\tau \mid [0, \infty] - \text{valued random variables satisfying}$$
$$\{\tau \leq t\} \cap \{\pi(t) \leq r\} \in \mathcal{F}_r \quad \text{for all } t \in R_+ \text{ and } r \in T\}.$$

Hence the following lemma is trivial from the continuity of π and the right-continuity of $\{\mathcal{F}_s\}_{s \in T}$.

Lemma 3.2. *For a strategy $\pi \in \Pi(0)$, $\{\mathcal{F}^\pi_t\}_{t \in R_+}$ is a non-decreasing and right continuous family of sub-σ-fields of $\mathcal{F}$.*

For strategies $\pi \in \Pi(0)$ and stopping times $\tau \in \mathcal{M}_0^\pi$, the pair (π, τ) are called tactics. Further for $s \in T$ we denote by $\mathcal{T}(0)$ the family of all tactics (see [Maz1,Section1]). Next for a tactic $(\pi, \tau) \in \mathcal{T}(0)$ we define a value function $V^{\pi\tau}$ of the optimal stopping problem for d-parameter Markov process X by

$$V^{\pi\tau}(x) = E^x[\int_0^\tau e^{-\alpha t} \sum_{i=1}^d f^i(X^i_{\pi^i(t)}) \mathrm{d}\pi^i(t)] \quad \text{for } x \in E. \tag{3.13}$$

We define functions $V^{\pi *}$ $(\pi \in \Pi(0))$ and V^{**} by

$$V^{\pi *}(x) = \sup_{\tau \in \mathcal{M}_0^\pi : P^x\{\tau < \infty\} = 1} V^{\pi\tau}(x) \quad \text{for } x \in E, \tag{3.14}$$

and

$$V^{**}(x) = \sup_{\pi \in \Pi(0)} V^{\pi *}(x) \quad \text{for } x \in E. \tag{3.15}$$

Hence we have the following lemma, which is proved in similar way as Lemma 2.1, regarding the finiteness of stopping times in (3.14).

Lemma 3.3. *The following (i) and (ii) hold:*

(i) $V^{\pi *}(x) = \sup_{\tau \in \mathcal{M}_0^\pi} V^{\pi\tau}(x)$ *for* $x \in E$ *and* $\pi \in \Pi(0)$;

(ii) $V^{**}(x) = \sup_{(\pi,\tau) \in \mathcal{T}(0)} V^{\pi\tau}(x)$ *for* $x \in E$.

The optimal stopping problem for d-parameter Markov process X is to find a tactic $(\pi, \tau) \in \mathcal{T}(0)$ attaining the supremum of Lemma 2.3(ii). Then V^{**} is called the optimal value function for the problem.

We shall formulate the optimal stopping problem for each reward process X^i $(i = 1, \cdots, d)$. For each $i(= 1, \cdots, d)$ we put an optimal value function V^{i*} of a one-parameter optimal stopping problem for the reward process X^i by

$$V^{i*}(x^i) = \sup_{\tau \in \mathcal{M}^i : P^{x^i}\{\tau < \infty\} = 1} E^{x^i}[\int_0^\tau e^{-\alpha t} \sum_{i=1}^d f^i(X^i_t) \mathrm{d}t] \quad \text{for } x^i \in E^i. \tag{3.16}$$

Then the one-parameter optimal stopping problem for the reward process X^i is to find a finite stopping time $\tau (\in \mathcal{M}^i)$ which attains the supremum in (3.16), where $\mathcal{M}^i$ denotes the family of all $\{\mathcal{F}^i_t\}_{t \in R_+}$-stopping times. Hence the following lemma is well-known (see [Shi1]).

Lemma 3.4. *We put a stopping time* σ^i_0 *by*

$$\sigma^i_0 = \inf\{t \in R_+ \mid V^{i*}(X^i_t) = 0\}. \tag{3.17}$$

If $P^{x^i}\{\sigma_0^i < \infty\} = 1$ *for all* $x^i \in E^i$, *then* σ_0^i *is the smallest optimal stopping time for (3.16).*

We introduce dynamic allocation index functions in continuous time bandit processes. For each $i = 1, \cdots, d$ the following dynamic allocation index function for the reward process X^i is given by

$$\nu^i(x^i) = \sup_{\tau \in \mathcal{M}^i : \tau > 0} \frac{E^{x^i}[\int_0^\tau e^{-\alpha t} f^i(X_t^i) \mathrm{d}t]}{E^{x^i}[\int_0^\tau e^{-\alpha t} \mathrm{d}t]} \quad \text{for } x^i \in E^i. \tag{3.18}$$

By utilizing dynamic allocation indices and the related results, we analyse the optimal stopping problem for d-parameter Markov processes X.

3.3. The optimal tactics

In Section 3.4 we shall investigate the optimal tactics for the optimal stopping problem for d-parameter Markov processes, by the method of embedding this problem into an optimal control problem of $d+1$-parameter Markov processes in similar way as the arguments in Section 2.3.

We shall add one more reward process $X^\bullet$ to a d-parameter Markov process X in the original optimal stopping problem for d-parameter Markov process X defined in Section 3.2. Hence we define an optimal control problem of an extended $d+1$-parameter Markov process. Following Section 2.3, we shall define notations of an optimal control problem of the extended $d+1$-parameter Markov process by using the signature *bar* as follows. Then strategies for the optimal control problem of the extended $d+1$-parameter Markov process $\overline{X}$ are $\overline{T}$-valued processes which are defined in the same manner as those for the optimal control problem for d-parameter Markov process X in Section 2.3. We let $\overline{\Pi}(\overline{\mathbf{0}})$ denote the family of all strategies for the optimal control problem of the extended $d+1$-parameter Markov process $\overline{X}$. Then the following lemma, the proof is similar to Lemma 2.4, implies the relation between the original optimal stopping problem for d-parameter Markov process X and the optimal control problem of the extended $d+1$-parameter Markov process $\overline{X}$.

Lemma 3.5. *For a tactic* $(\pi, \tau) \in \mathcal{T}(0)$ *we define a* $\overline{T}$*-valued processes* $\overline{\pi}$ *by*

$$\overline{\pi}(t) := \begin{cases} (0, \pi(t)) & \text{on } \{t < \tau\} \\ (t - \tau, \pi(\tau)) & \text{otherwise} \end{cases} \quad \text{for } t \in R_+. \tag{3.19}$$

Then (i) and (ii) hold:

(i) $\overline{\pi} \in \overline{\Pi}(\overline{\mathbf{0}})$,

(ii) $\overline{R}^{\overline{\pi}}(\overline{x}) = V^{\pi\tau}(x)$ *for every* $\overline{x} = (x^0, x) \in \overline{E} := E^0 \times E$.

Now we shall construct optimal tactics for the optimal stopping problem for the d-parameter Markov process X.

We take $\overline{\pi}^*$, π^*, τ^{π^*} and $\tilde{\pi}^*$ as follows: By applying Lemmas 3.1 and 3.5 to the extended $d+1$-parameter Markov process $\overline{X}$, we may take a strategy $\overline{\pi}^* \in \overline{\Pi}(\overline{\mathbf{0}})$ which has the maximum index property (3.19) and is optimal for the optimal control problem of the extended $d+1$-parameter Markov process $\overline{X}$. Hence we consider a stopping time $\overline{\tau} = \inf\{t \in R_+ \mid \overline{\nu}(\overline{X}_{\overline{\pi}^*(t)}) \leq 0\}$. Then since $\overline{\pi}^*$ has the maximum index property for Markov process $\overline{X}$ (see [Man2,Theorem15]) and $\nu^0 = 0$, we may put a strategy π^* by

$$\pi^*(t) := \begin{cases} (\overline{\pi}^{*1}(t), \cdots, \overline{\pi}^{*d}(t)) & \text{on } \{t < \overline{\tau}\} \\ (\overline{\pi}^{*1}(\overline{\tau}) + t - \overline{\tau}, \overline{\pi}^{*2}(\overline{\tau}), \cdots, \overline{\pi}^{*d}(\overline{\tau})) & \text{otherwise} \end{cases} \quad \text{for } t \in R_+. \tag{3.20}$$

Then we have

$$\overline{\tau} = \inf\{t \in R_+ \mid \nu(X_{\pi^*(t)}) \leq 0\} \text{ (therefore we represent it by } \tau^{\pi^*}). \tag{3.21}$$

Moreover we define a strategy $\tilde{\pi}^*$ in the same way as (3.19), replacing π and τ with π^* and τ^{π^*} respectively. Hence we have the following lemma, which is proved in similar to Lemma 2.6:

Lemma 3.6. *The following (i) — (iii) hold:*

(i) $\overline{\nu}(\overline{X}_{\overline{\pi}^*(t)}) := \begin{cases} \nu(X_{\pi^*(t)}) & \text{on } \{t < \tau^{\pi^*}\} \\ 0 & \text{otherwise} \end{cases}$ *for every* $t \in R_+$;

(ii) $(\pi^*, \tau^{\pi^*}) \in \mathcal{T}(0)$;

(iii) $\tilde{\pi}^* \in \Pi(\overline{\mathbf{0}})$.

Now we shall introduce a few tools in order to analyse the local time behavior of standard Markov process $\{\overline{X}_{\overline{\pi}^*(t)}\}_{t \in R_+}$, referring [Man1]. For a maximum index strategy $\overline{\pi}^* = \{\overline{\pi}^*(t)\}_{t \in R_+} = \{(\overline{\pi}^{*0}(t), \cdots, \overline{\pi}^{*d}(t))\}_{t \in R_+}$ $(\in \overline{\Pi}(\overline{\mathbf{0}}))$ we represent the inverses of processes $\{\overline{\pi}^{*j}(t)\}_{t \in R_+}$ $(j = 0, \cdots, d)$ by

$$\lambda^j(t) = \inf\{t \in R_+ \mid \overline{\pi}^{*j}(t) > r\} \quad \text{for } r \in R_+ \text{ and } j = 0, \cdots, d. \tag{3.22}$$

Next for $r \in R_+$ and $j = 0, \cdots, d$ we put

$$\xi^j(r) = E^{\mathcal{F}_0^0 \otimes \cdots \otimes \mathcal{F}_\infty^j \otimes \cdots \otimes \mathcal{F}_0^d}[e^{-\alpha(\tilde{\lambda}^j(r) - r)} - e^{-\alpha(\lambda^j(r) - r)}], \tag{3.23}$$

where $\tilde{\lambda}^j(t)$ denotes the inverse of $\{\tilde{\pi}^{*j}(t)\}_{t\in R_+}$ $(j = 0, \cdots, d)$. Then we have the following lemmas.

Lemma 3.7. *The following (i) — (v) hold:*

(i) $\tilde{\lambda}^0(s) = \tau^{\pi^*} + s \leq \lambda^0(s)$ *for* $s \in R_+$.

(ii) $\tilde{\lambda}^j(s) = \infty \geq \lambda^j(s)$ *for* $s > \overline{\pi}^{*j}(\tau^{\pi^*})$ *and* $j = 1, \cdots, d$;
$\tilde{\lambda}^j(s) = \lambda^j(s)$ *for* $s \leq \overline{\pi}^{*j}(\tau^{\pi^*})$ *and* $j = 1, \cdots, d$.

(iii) $\lambda^j(s') - \lambda^j(s) \geq s' - s$ *for* $s, s' \in R_+$ $(s \leq s')$ *and* $j = 0, \cdots, d$.

(iv) $-\xi^j(r) = E^{\mathcal{F}_0^0 \otimes \cdots \otimes \mathcal{F}_\infty^j \otimes \cdots \otimes \mathcal{F}_0^d}[e^{-\alpha(\lambda^j(r)-r)} \cdot I_{\{r \geq \overline{\pi}^{*j}(\tau^{\pi^*})\}}]$ *for* $r \in R_+$ *and* $j = 1, \cdots, d$.

(v) $\xi^0(r) = E^{\mathcal{F}_\infty^0 \otimes \mathcal{F}_0^1 \otimes \cdots \otimes \mathcal{F}_0^d}[e^{-\alpha\tau^{\pi^*}} - e^{-\alpha(\lambda^0(r)-r)}]$ *for* $r \in R_+$.

Proof. (i) and (ii) are trivial from the definitions. (iv) and (v) are trivial from (i) and (ii). (iii) Fix any $s, s' \in R_+$ $(s \leq s')$ and $j = 0, \cdots, d$. Due to the definition of π we have

$$\overline{\pi}^{*j}(t') - \overline{\pi}^{*j}(t) \leq t' - t \quad \text{for } t, t' \in R_+ \ (t \leq t'). \tag{3.24}$$

Hence for $s, s' \in R_+$ $(s \leq s')$ we put $t = \lambda^j(s)$ and $t' = \lambda^j(s')$, then we have $s = \overline{\pi}^{*j}(t)$ and $s' = \overline{\pi}^{*j}(t')$. By substituting these in (3.24) we obtain (iii). □

Lemma 3.8. *The following (i) and (ii) hold:*

(i) For $j = 1, \cdots, d$, $\{-\xi^j(r)\}_{r\in R_+}$ *is a non-increasing,* $\{\mathcal{F}_\bullet^0 \otimes \cdots \otimes \mathcal{F}_r^j \otimes \cdots \otimes \mathcal{F}_0^d\}_{r\in R_+}$*-adapted and right continuous process for all* $r > E^{\mathcal{F}_0^0 \otimes \cdots \otimes \mathcal{F}_r^j \otimes \cdots \otimes \mathcal{F}_0^d}[\overline{\pi}^{*j}(\tau^{\pi^*})]$ *and satisfies* $0 \leq -\xi^j(r) \leq 1$.

(ii) $\{\xi^0(r)\}_{r\in R_+}$ *is a non-decreasing,* $\{\mathcal{F}_r^0 \otimes \mathcal{F}_0^1 \otimes \cdots \otimes \mathcal{F}_0^d\}_{r\in R_+}$*- adapted and right continuous process satisfying* $0 \leq \xi^0(r) \leq 1$ *for all* $r \in R_+$.

Proof. The right-continuity of $\{\xi^j(r)\}_{r\in R_+}$ $(j = 1, \cdots, d)$ is trivial from the definitions. The measurability of $\{\xi^j(r)\}_{r\in R_+}$ $(j = 1, \cdots, d)$ is due to (3.8). Finally we shall show the monotony of processes $\{\xi^j(r)\}_{r\in R_+}$. From Lemma 3.7 (v) for fixed any $r, r' \in R_+$ $(r \leq r')$ we have

$$\xi^0(r') - \xi^0(r) = E^{\mathcal{F}_\infty^0 \otimes \mathcal{F}_0^1 \otimes \cdots \otimes \mathcal{F}_0^d}[e^{-\alpha(\lambda^0(r)-r)} - e^{-\alpha(\lambda^0(r')-r')}]. \tag{3.25}$$

From Lemma 3.7(iii) we have $\lambda^j(r') - r' \geq \lambda^j(r) - r \geq 0$ and $0 \leq e^{-\alpha(\lambda^j(r)-r)} - e^{-\alpha(\lambda^j(r')-r')}$ for $r \in R_+$ and $j = 0, \cdots, d$. Together with (3.25) we obtain $\xi^0(r') \leq \xi^0(r)$. Similarly from Lemma 3.7 (iii), (iv), for $r' > r > E^{\mathcal{F}_0^0 \otimes \cdots \otimes \mathcal{F}_r^j \otimes \cdots \otimes \mathcal{F}_0^d}[\overline{\pi}^{*j}(\tau^{\pi^*})]$ we obtain

$$(-\xi^j(r')) - (-\xi^j(r)) = E^{\mathcal{F}_0^0 \otimes \cdots \otimes \mathcal{F}_\infty^j \otimes \cdots \otimes \mathcal{F}_0^d}[e^{-\alpha(\lambda^j(r')-r')} - e^{-\alpha(\lambda^j(r)-r)}] \leq 0.$$

Thus we obtain this lemma. □

Proposition 3.1. *It holds that*

$$\overline{R}^{\hat{\overline{\pi}}^*} \geq \overline{R}^{\overline{\pi}^*} \quad \text{on } \{y \in \overline{E} \mid \overline{\nu}(y) = 0\}.$$

Proof. Fix any $y \in \overline{E}$ satisfying $\overline{\nu}(y) = 0$. Then since $f^0 = 0$ and $\nu^0 = 0$, we have

$$\begin{aligned}
&\overline{R}^{\hat{\overline{\pi}}^*}(y) - \overline{R}^{\overline{\pi}^*}(y) \\
&= \sum_{j=1}^{d} E^y[\int_0^\infty e^{-\alpha t} f^j(X^j_{\hat{\overline{\pi}}^{*j}(t)}) \mathrm{d}\hat{\overline{\pi}}^{*j}(t)] - \int_0^\infty e^{-\alpha t} f^j(X^j_{\overline{\pi}^{*j}(t)}) \mathrm{d}\overline{\pi}^{*j}(t)] \\
&= \sum_{j=1}^{d} E^y[\int_0^\infty \xi^j(r) e^{-\alpha r} f^j(X^j_r) \mathrm{d}r].
\end{aligned}$$

Next, by using Lemma 3.8, similarly to [VWB1,Appendix] we have that

$$\text{the previous term} \geq \sum_{j=1}^{d} \sup_{\tau \in \mathcal{M}^j} E^y[\xi^j(0) \int_0^\tau e^{-\alpha r} f^j(X^j_r) \mathrm{d}r] \geq 0.$$

Therefore we obtain this proposition. □

Now the following theorem implies the existence of an optimal tactic, which is defined on the basis of dynamic allocation indices, of the optimal stopping problem for a d-parameter Markov process in Section 3.2.

Theorem 3.1. *It holds that*

$$V^{\pi^* \tau^{\pi^*}}(x) = V^{**}(x) = \overline{R}^*(\overline{x}) \quad \textit{for every } \overline{x} = (x^0, x) \in \overline{E} := E^0 \times E.$$

Therefore if $P^x[\tau^{\pi^*} < \infty] = 1$ for all $x \in E$, then (π^*, τ^{π^*}) is an optimal tactic of the optimal stopping problem for d-parameter Markov process X.

Proof. Fix any $\overline{x} = (x^0, x) \in \overline{E}$. Since $\overline{\pi}^*$ is an optimal strategy of the optimal control problem for the extended $d+1$-parameter Markov process $\overline{X}$, we have

$$\overline{R}^*(\overline{x}) = \overline{R}^{\overline{\pi}^*}(\overline{x}). \tag{3.26}$$

While from Lemma 3.5 we have

$$\overline{R}^*(\overline{x}) \geq V^{**}(\overline{x}). \tag{3.27}$$

Moreover since $\overline{\nu}(X_0^0, X_{\pi^*(\tau^{\pi^*})}) = 0$ a.s., due to Proposition 3.1 we have

$$\begin{aligned}\overline{R}^{\overline{\pi}^*}(\overline{x}) &= \sum_{j=1}^{d} E^x[\int_0^{\tau^{\pi^*}} e^{-\alpha t} f^j(X^j_{\pi^{*j}(t)}) \mathrm{d}\pi^{*j}(t)] + E^{\overline{x}}[e^{-\alpha\tau^{\pi^*}} \overline{R}^{\overline{\pi}^*}(X_0^0, X_{\pi^*(\tau^{\pi^*})})] \\ &\leq \overline{R}^{\tilde{\pi}^*}(\overline{x}) = V^{\pi^*\tau^{\pi^*}}(x) \leq V^{**}(x).\end{aligned}$$

Together with (3.26) and (3.27), this inequalty completes the proof of this theorem. □

Next we shall characterize the optimal stopping time $\tau^{\pi^*} = \inf\{t \in R_+ \mid \nu(\pi^*(t)) \leq 0\}$ of the optimal stopping problem for d-parameter Markov process X by Markov potential theory. We shall express the optimal stopping point $\pi^*(\tau^{\pi^*})$ by the optimal stopping times for d one-parameter optimal stopping problems for reward processes X^i. Therefore according to Section 3.2, for $i = 1, \cdots, d$ we shall utilize one-parameter optimal stopping problems for reward processes X^i. Hence similarly to Lemma 2.8 we have the following relation between the optimal expected value V^{i*} and the dynamic allocation index ν^i:

Lemma 3.9. *For $i = 1, \cdots, d$ we have (i) and (ii):*

(i) $V^{i*} \geq 0$.

(ii) $\{x^i \in E^i \mid \nu^i(x^i) \leq 0\} = \{x^i \in E^i \mid V^{i*}(x^i) = 0\}$.

For each $i = 1, \cdots, d$ we put a subset $B^i = \{x^i \in E^i \mid V^{i*}(x^i) = 0\}$, which is a closed set for process X^i. Then from Lemma 3.9 we obtain

$$B^i = \{x^i \in E^i \mid V^{i*}(x^i) = 0\} = \{x^i \in E^i \mid \nu^i(x^i) \leq 0\} \quad \text{for } i = 1, \cdots, d. \tag{3.28}$$

Here we call B^i an optimal stopping region for one-parameter stopping problem (3.21), because $\sigma_0^i = \inf\{t \in R_+ \mid X_t^i \in B^i\}$ (see (3.22)). Hence we define $B = \prod_{i=1}^d B^i$ and then

$$B = \{x \in E \mid \nu(x) \leq 0\}. \tag{3.29}$$

We call B an optimal stopping region for the optimal stopping problem for d-parameter Markov process X, because $\tau^{\pi^*} = \inf\{t \in R_+ \mid X_{\pi^*(t)} = 0\}$. Then we obtain the following theorem.

Theorem 3.2. *Regarding the relation between the optimal tactics (π^*, τ^{π^*}) of the optimal stopping problem for d-parameter Markov process X (in Theorem 3.1) and the optimal stopping times σ_0^i of independent one-parameter optimal stopping problems of reward processes X^i, (i) — (iv) hold:*

(i) $\sigma_0^i = \inf\{t \in R_+ \mid X_t^i \in B^i\} = \pi^{*i}(\tau^{\pi^*})$ *for every* $i = 1, \cdots, d$;

(ii) $\tau^{\pi^*} = \sum_{i=1}^{d} \sigma_0^i = \inf\{t \in R_+ \mid X_{\pi^*(t)} \in B\}$;

(iii) $P^x\{\tau^{\pi^*} < \infty\} = \prod_{i=1}^{d} P^{x^i}\{\sigma_0^i < \infty\}$ *for every* $x = (x^1, \cdots, x^d) \in E$;

(iv) $B = \{x \in E \mid V^{**}(x) = 0\}$.

Proof. (i) Fix any $i = 1, \cdots, d$. Set $\overline{\tau}^i = \inf\{t \in R_+ \mid \nu^i(X^i_{\pi^{*i}(t)}) \leq 0\}$. Then we have

$$\overline{\tau}^i \leq \inf\{t \in R_+ \mid \nu(X_{\pi^*(t)}) \leq 0\} = \tau^{\pi^*}. \tag{3.30}$$

Hence for all $t \in R_+$, it holds that

$$\nu(X_{\pi^*(t)}) > 0 \geq \nu^i(X^i_{\pi^{*i}(\overline{\tau}^i)}) \quad \text{on } \{\overline{\tau}^i \leq t < \tau^{\pi^*}\}. \tag{3.31}$$

Hence the strategy π^* does not move X^i at any time t on $\{\overline{\tau}^i \leq t < \tau^{\pi^*}\}$ due to [Man2,Theorem15]. Therefore we obtain

$$\pi^{*i}(\tau^{\pi^*}) = \pi^{*i}(\overline{\tau}^i). \tag{3.32}$$

On the other hand due to the continuity of π^* we obtain

$$\pi^{*i}(\overline{\tau}^i) = \pi^{*i}(\inf\{t \in R_+ \mid \nu^i(X^i_{\pi^{*i}(t)}) \leq 0\}) = \inf\{r \in R_+ \mid \nu^i(X^i_r) \leq 0\} = \sigma_0^i. \tag{3.33}$$

Together with (3.33) and (3.34) we obtain (i). (ii) and (iii) are trivial from (i) since X^i $(i = 1, \cdots, d)$ are independent. (iv) We put $D = \{x \in E \mid V^{**}(x) = 0\}$ and $\tau = \inf\{t \in R_+ \mid X_{\pi^*(t)} \in D\}$. Then τ is the smallest optimal stopping time of the classical one-parameter optimal stopping problem (3.17) concerning the standard Markov process $\{X_{\pi^*(t)}, \mathcal{F}_{\pi^*(t)}\}_{t \in R_+}$ (for example, see [Shi1]). Therefore since τ^{π^*} is an optimal stopping time in $\mathcal{M}_0^{\pi^*}$, from Theorem 3.1 we have $\tau \leq \tau^{\pi^*}$. On the other hand for arbitrary but fixed $i(= 1, \cdots, d)$, we define a strategy $\pi(t) = e_i \cdot t$ for $t \in R_+$. Then due to Lemma 3.9(i) we have $V^{**}(x) \geq V^{\pi*}(x) = V^{i*}(x^i) \geq 0$ for every $x = (x^1, \cdots, x^d) \in E$. Therefore

$$D = \{x \in E \mid V^{**}(x) = 0\} \subset \prod_{i=1}^{d} \{x^i \in E^i \mid V^{i*}(x^i) = 0\}. \tag{3.34}$$

So due to Lemma 3.9(ii) we obtain

$$D \subset \prod_{i=1}^{d} \{x^i \in E^i \mid \nu^i(x^i) \leq 0\} = \{x \in E \mid \nu(x) \leq 0\} = B. \tag{3.35}$$

Thus we obtain $\tau = \inf\{t \in R_+ \mid X_{\pi^*(t)} \in D\} \geq \inf\{t \in R_+ \mid X_{\pi^*(t)} \in B\} = \tau^{\pi^*}$. Consequently we obtain $\tau = \tau^{\pi^*}$. Especially we fix any $x \in B - D$ and any $x^0 \in E^0$.

Then we have $\tau^{\pi^*} = 0$ a.s. P^x. Since $\tilde{\pi}^*$ does not move X^0 on the set $\{(x^0, x) \in E^0 \times E \mid V^{**}(x) > 0\} = \{(x^0, x) \in E^0 \times E \mid \overline{R}^*(x^0, x) > 0\}$, it is an open set for standard Markov process $(\overline{X}_{\tilde{\pi}^*(t)}, \overline{\mathcal{F}}_{\tilde{\pi}^*(t)})_{t \in R_+}$. Therefore we obtain $P^x[\tau > 0] = 1$. This contradicts $\tau = \tau^{\pi^*} = 0$ a.s. P^x. Thus we obtain (iv). □

3.4. The Bellman's equation

[Maz1,Section3] has studied the Bellman's equation to the optimal stopping problem for a two-parameter Markov process in the case of $f^i = 0$ for $i = 1, 2$ in (3.13). Here we shall consider it in the case of $g = 0$ in (3.13). For each $i(= 1, 2)$ $\mathcal{L}^i$ denotes an infinitesimal generator of a diffusion process $X^i = \{X_t^i\}_{t \in R_+}$, which is a unique solution to the stochastic differential equation (3.1):

$$\mathcal{L}^i = \frac{1}{2} a^i(x^i) \frac{\mathrm{d}^2}{\mathrm{d}x^{i2}} + b^i(x^i) \frac{\mathrm{d}}{\mathrm{d}x^i}. \tag{3.36}$$

Moreover for each $i(= 1, 2)$ $\mathcal{D}^i$ denotes the domain of the generator $\mathcal{L}^i$:

$$\mathcal{D}^i = \{h \text{ are bounded twice continuously differentiable functions on } E\}. \tag{3.37}$$

Then the following theorem implies Bellman's equation for the optimal value function V^{**}.

Theorem 3.3. *Suppose $V^{**} \in \mathcal{D}^1 \cap \mathcal{D}^2$. Then (i) — (iii) hold:*

(i) $V^{**} \geq 0$ *on* E;

(ii) $\max_{i=1,2}\{\mathcal{L}^i V^{**} - \alpha V^{**} + f^i\} \leq 0$ *on* E;

(iii) $\max_{i=1,2}\{\mathcal{L}^i V^{**} - \alpha V^{**} + f^i\} = 0$ *on* $\{V^{**} > 0\}$.

Proof. (i) is trivial. (ii) Fix any $i = 1, 2$ and $\overline{x} \in \overline{E}$. For $\epsilon > 0$ we define a strategy $\tilde{\pi}$ by

$$\tilde{\pi}(t) := \begin{cases} e_i \cdot t & \text{for } t \in [0, \epsilon) \\ \overline{\pi}^{*i}(t - \epsilon) & \text{for } t \in [\epsilon, \infty). \end{cases} \tag{3.38}$$

Then we have $\tilde{\pi} \in \Pi(0)$ and

$$\overline{R}^*(\overline{x}) = \overline{R}^{\overline{\pi}}(\overline{x}) \geq \overline{R}^{\tilde{\pi}}(\overline{x}) = E^{\overline{x}}[\int_0^\epsilon e^{-\alpha t} f^i(X_t^i) \mathrm{d}t + e^{-\alpha\epsilon} \overline{R}^*(x^0, \cdots, X_\epsilon^i, \cdots, x^d)]. \tag{3.39}$$

Namely

$$E^{\overline{x}}[e^{-\alpha\epsilon} \overline{R}^*(x^0, \cdots, X_\epsilon^i, \cdots, x^d)] - \overline{R}^*(\overline{x}) \leq -E^{\overline{x}}[\int_0^\epsilon e^{-\alpha t} f^i(X_t^i) \mathrm{d}t] \quad \text{for } \epsilon > 0. \tag{3.40}$$

Due to the relation between the generators and the infinitesimal operators of diffusions in [Dyn1,Chapter5] we obtain

$$\max_{i=0,1,2}\{\mathcal{L}^i\overline{R}^* - \alpha\overline{R}^* + f^i\} \leq 0 \quad \text{on } \overline{E}.$$

Together with Theorem 3.1 this inequality implies (ii). (iii) Fix any $x \in E$ satisfying $V^{**}(x) > 0$. Hence since $\{X_{\pi^*(t)}\}_{t\in R_+}$ has the strong Markov property (see Lemma 3.1), we have

$$V^{**}(x) = V^{\pi^*\tau^{\pi^*}}(x) = E^x[\int_0^\tau e^{-\alpha t}\sum_{i=1}^2 f^i(X^i_{\pi^{*i}(t)})\mathrm{d}\pi^{*i}(t) + e^{-\alpha\tau}V^{**}(X_{\pi^*(\tau)})] \qquad (3.41)$$

for all stopping times $\tau(\in \mathcal{M}_0^{\pi^*})$ satisfying $0 < \tau \leq \tau^{\pi^*}$ (a.s. P^x). Hence Dynkin's formula for two-parameter processes holds for all functions of $\mathcal{D}^1 \cap \mathcal{D}^2$ since in the proof of [Maz1, Proposition 2.2.4] he does not use p-biexcessivity itself. In the case where X^i are one-dimensional diffusions, we refer to [Kar1,Theorem6.1]. By applying the formula to (3.41), there exist one-parameter $\{\mathcal{F}_0^{\pi^*}\}_{t\in R_+}$- adapted processes $\{\lambda^{*1}(t)\}_{t\in R_+}$ and $\{\lambda^{*2}(t)\}_{t\in R_+}$, non-vanishing simultaneously and taking values [0,1], such that for all stopping times $\tau(\in \mathcal{M}_0^{\pi^*})$ satisfying $0 < \tau \leq \tau^{\pi^*}$ (a.s. P^x) it holds that

$$\begin{aligned} -\frac{\sum_{i=1}^2 E^x[\int_0^\tau e^{-\alpha t}f^i(X^i_{\pi^{*i}(t)})\mathrm{d}\pi^{*i}(t)]}{E^x[\tau]} &= \frac{E^x[e^{-\alpha\tau}V^{**}(X_{\pi^*(\tau)})] - V^{**}(x)}{E^x[\tau]} \\ &= \frac{\sum_{i=1}^2 E^x[\int_0^\tau(\mathcal{L}^iV^{**} - \alpha V^{**})(X^i_{\pi^{*i}(t)})\lambda^{*i}(t)\mathrm{d}t]}{E^x[\tau]}. \end{aligned}$$

Due to the way to construct $\{\lambda^{*1}(t)\}_{t\in R_+}$ and $\{\lambda^{*2}(t)\}_{t\in R_+}$ in the proof of [Maz1, Proposition2.2.4], we have $\{\frac{\mathrm{d}\pi^{*i}(t)}{\mathrm{d}t} > 0\} = \{\lambda^{*i}(t) > 0\}$ a.s. P^x for almost all $t \in R_+$ and $i = 1, 2$. Therefore as letting $\tau \downarrow 0$, the previous eouality follows (iii) (see [Dyn2]). □

Regarding the optimal stopping region B we have the following proposition.

Proposition 3.2. *The optimal stopping region $B = \{V^{**} = 0\}$, which is a free boundary of Bellman's equation (3.41), has the following representations:*

$$\begin{aligned} B = \{x \in E \mid V^{**}(x) = 0\} &= \prod_{i=1}^2\{x^i \in E^i \mid V^{i*}(x^i) = 0\} \\ &= \prod_{i=1}^2\{x^i \in E^i \mid \nu^i(x^i) \leq 0\} \\ &= \{x \in E \mid \nu(x) \leq 0\}. \end{aligned}$$

Proof. It is trivial from (3.29), (3.35) and Theorem 3.2(iv). □

Chapter 4

The Multi-Armed Bandit Game

4.1. Introduction

This chapter deals with two-person zero-sum games where in every time players alternately either select only one of arms of bandit machines or stop them. The purpose of this chapter is to formulate bandit games and solve them as control problems.

Now we shall sketch bandit games, referring to Mandelbaum[7]. We regard that a discrete-time d-armed bandit process consists of d independent arms, which evolve according to $\{\mathcal{F}_t^i\}_{t\in N}$-adapted Markov chains $X^i = \{X_t^i\}_{t\in N}$ $(i = 1, \cdots, d)$, where N is the set of all non-negative integers. If one player A selects arm i, then he obtains some rewards and arm i evolves one-step according to the transition probability of Markov chain for arm i and the next is another player B's turn. However if one of players stops arms, then both players must stop selecting arms and settle accounts. Both players alternately continue to select arms until either player stops the games. Here we assume that player A may either select or stop at even time and at odd time may do player B (This case is called first-type in Section 4.2). Let $\mathbf{0}$ and e_i denote the zero vector and the i'th unit vector in N^d respectively. Put $\mathcal{F}_s = \mathcal{F}_{s^1}^1 \otimes \cdots \otimes \mathcal{F}_{s^d}^d$ for $s = (s^1, \cdots, s^d) \in N^d$, and $N(e, r) = \{\text{even } t \mid 0 \leq t < r\}$ and $N(o, r) = \{\text{odd } t \mid 0 \leq t < r\}$ for $r \in N \cup \{+\infty\}$. Player A has two kinds of decisions, i.e. selecting arms and stopping the games. We represent the former with player A's strategies π_A and the latter with his stopping times τ_A. Therefore player B also has his strategies π_B and stopping times τ_B. Both players' strategies $(\pi_A; \pi_B)$[§] are defined as follows. $\pi_A = \{\pi_A(t)\}_{t\in N(o,\infty)}$ and $\pi_B = \{\pi_B(t)\}_{t\in N(e,\infty)}$ are N^d-valued stochastic sequences on $(\Omega, \mathcal{F})$ satisfying the following (i) — (iii):

(i) $\pi_A(0) = \mathbf{0}$ and $\pi_B(0) = \mathbf{0}$.

(ii) Players alternately select only one of arms. Namely,
for all $t \in N(e, \infty)$ it holds that $\pi_A(t+1) = \pi_B(t) + e_i$ for some $i = 1, \cdots, d$, and
for all $t \in N(o, \infty)$ it holds that $\pi_B(t+1) = \pi_A(t) + e_i$ for some $i = 1, \cdots, d$.

(iii) Players' strategies are adapted to the imformation until the present time. Namely,
for all $t \in N(o, \infty)$ and all $s' \in N^d$ it holds that $\{\pi_A(t) = s'\} \in \mathcal{F}_{s'}$, and
for all $t \in N(e, \infty)$ and all $s' \in N^d$ it holds that $\{\pi_B(t) = s'\} \in \mathcal{F}_{s'}$.

Then player A's stopping times τ_A (player B's τ_B) are $N(e, \infty)$ $(N(o, \infty))$-valued random variables on $(\Omega, \mathcal{F})$ satisfying the adaptation :

(iv) For $t \in N(e, \infty)$ it holds that $\{\tau_A = t\} \in \mathcal{F}_{\pi_B(t)}$, and
for $t \in N(o, \infty)$ it holds that $\{\tau_B = t\} \in \mathcal{F}_{\pi_A(t)}$,

[§]The definition of strategies is referred from Mandelbaum[7,2.2.A 2.2.C].

where $\{\mathcal{F}_{\pi_A(t)}\}_{t\in N(o,\infty)}$ and $\{\mathcal{F}_{\pi_B(t)}\}_{t\in N(e,\infty)}$ denote informations until time t:

$$\mathcal{F}_{\pi_A(t)} = \{\Gamma \in \mathcal{F} \mid \Gamma \cap \pi_A(t) = s \in \mathcal{F}_s \text{ for } s \in N^d\} \quad \text{for } t \in N(o,\infty), \text{ and}$$
$$\mathcal{F}_{\pi_B(t)} = \{\Gamma \in \mathcal{F} \mid \Gamma \cap \pi_B(t) = s \in \mathcal{F}_s \text{ for } s \in N^d\} \quad \text{for } t \in N(e,\infty).$$

Here we need not to assume that strategies $(\pi_A;\pi_B)$ are stopped by times τ_A and τ_B, differently from the definition in Lawler-Vanderbei[6,p.643(b)]. (The reason is trivial from representations of expected rewards in Section 4.2.) Then player A's expected gain[¶] to be paid from player B at an initial state x is represented as sums of gains when player A stops the games and gains when does player B :

$$V_F^{\pi_A\tau_A\pi_B\tau_B}(x) = E^x[\sum_{t\in N(e,\tau_A)} \beta^t \sum_{i=1}^{d} f_A^i(X^i_{\pi^i_A(t+1)})(\pi_A^i(t+1) - \pi_B^i(t)) + \beta^{\tau_A} h_A(X_{\pi_B(\tau_A)})$$

$$- \sum_{t\in N(o,\tau_A)} \beta^t \sum_{i=1}^{d} f_B^i(X^i_{\pi^i_B(t+1)})(\pi_B^i(t+1) - \pi_A^i(t)) - \beta^{\tau_A} h_B(X_{\pi_B(\tau_A)}) : \tau_A < \tau_B]$$

$$+E^x[\sum_{t\in N(e,\tau_B)} \beta^t \sum_{i=1}^{d} f_A^i(X^i_{\pi^i_A(t+1)})(\pi_A^i(t+1) - \pi_B^i(t)) + \beta^{\tau_B} h_A(X_{\pi_A(\tau_B)})$$

$$- \sum_{t\in N(o,\tau_B)} \beta^t \sum_{i=1}^{d} f_B^i(X^i_{\pi^i_B(t+1)})(\pi_B^i(t+1) - \pi_A^i(t)) - \beta^{\tau_B} h_B(X_{\pi_A(\tau_B)}) : \tau_B \le \tau_A],$$

where β is a constant discount rate $(0 < \beta < 1)$, i $(= 1, \cdots, d)$ are arm numbers, f_A^i (f_B^i) models player A's (player B's) running rewards obtained at current states when he selects arm i, and h_A (h_B) models player A's (player B's resp.) rewards obtained at states where he stops. Further E^x denotes the expectation with an initial state x. Hence player A's aim is to maximize his gains $V_F^{\pi_A\tau_A\pi_B\tau_B}$, by controlling his strategies and stopping times, however player B's is to minimize $V_F^{\pi_A\tau_A\pi_B\tau_B}$. However one player's admissible strategies and admissible stopping times generally depend on another player's option of strategies and stopping times. In order to solve this problem we introduce one-step Markov strategies and *two-steps Markov* stopping times. Next by the use of them we show existence of the optimal Markov strategies and the optimal stopping times under the assumption of independence of arms. While we present a certain value iteration (see Iteration 4.2) and show the iteration converges to Bellman's equation. By using Bellman's equation, the present thesis gives the optimal values. Further we classify the state space into selection regions for each arm and a stopping region on the basis of the derived Bellman's equations. Finally this chapter gives optimal Markov strategies and

[¶]These descriptions are referred from the value of the reward process in Mandelbaum [8,(2.2)], by shifting time.

optimal stopping times, by constructing concatenations of one-step Markov strategies and concatenations of *two-steps Markov* stopping times, which are defined on the selection regions and the stopping region. We also show the uniqueness of optimal values of bandit games. This chapter is structured as follows.

In Section 4.2 we reformulate multi-armed Markov processes, strategies and stopping times for bandit games. We introduce Markov strategies and *Markov* stopping times, referring to [LawVan1,p.645(3.1)], and show a few fundamental lemmas regarding to their concatenations. We formulate players' expected rewards and bandit games. We provide a proposition to guarantee existence of the optimal Markov strategies and the optimal Markov stopping times. In Section 4.3 we give a backward value iteration and demonstrate its convergence and construct optimal Markov strategies and optimal Markov stopping times on the basis of Bellman's equation. Finally the remainder of this chapter is devoted to show the uniqueness of the optimal values.

4.2. Strategies and stopping times for bandit processes

In this section we shall formulate zero-sum bandit games. Let N be the set of all non-negative integers. Let d, the number of arms, be a positive integer. Let $\mathbf{0}$ and e_i denote the zero vector and the i'th unit vector in N^d respectively. Put $N(e,r) = \{\text{even } t \mid 0 \le t < r\}$ and $N(o,r) = \{\text{odd } t \mid 0 \le t < r\}$ for $r \in N \cup \{+\infty\}$. We deal with the case where arms are mutually independent. Therefore we regard that d-armed bandit processes consist of d mutually independent reward processes. First we shall define bandit processes, referring to [Man1].

For arms i $(= 1, \cdots, d)$, let $(\Omega^i, \mathcal{F}^i, P^i)$ denote probability spaces and let $X^i = (X^i_t, \mathcal{F}^i_t, \theta^i_t, P^i)_{t\in N}$ denote $(\mathcal{F}^i_t)_{t\in N}$-adapted time-homogeneous Markov chains, which are mutually independent, with Borel state spaces E^i. Here $(\mathcal{F}^i_t)_{t\in N}$ is an increasing family of completed sub-σ-fields of $\mathcal{F}^i$ and θ^i_t is the time-shift operator on Ω^i. Next we define a d-parameter process by their products. Set its time space $T = N^d$, its path space $\Omega = \prod_{i=1}^d \Omega^i$ and its state space $E = \prod_{i=1}^d E^i$. Hence we introduce the usual partial order into T. For $r = (r^1, \cdots, r^d), s = (s^1, \cdots, s^d) \in T$, $r \le s$ means that $r^i \le s^i$ for all $i = 1, \cdots, d$. Then we define a d-parameter process X with the state space E, its σ-fields $\mathcal{F}_s$ and its time-shift operators θ_s by

$$X = (X_s)_{s\in T} = (X^1_{s^1}, \cdots, X^d_{s^d})_{s=(s^1,\cdots,s^d)\in T},$$

$$\mathcal{F}_s = \mathcal{F}^1_{s^1} \otimes \cdots \otimes \mathcal{F}^d_{s^d} \quad \text{for } s = (s^1, \cdots, s^d) \in T, \text{ and}$$

$$\theta_s\omega = (\theta^1_{s^1}\omega^1, \cdots, \theta^d_{s^d}\omega^d) \quad \text{for } s = (s^1, \cdots, s^d) \in T \text{ and } \omega = (\omega^1, \cdots, \omega^d) \in \Omega.$$

Further E^x denotes the expectation induced by probability measure $P = \prod_{i=1}^{d} P^i$ with an initial state $x \in E$. We also use notations $|s| = \sum_{i=1}^{d} s^i$ for $s = (s^1, \cdots, s^d) \in T$.

As for games in this chapter, strategies with stopping times are called tactics. Here we shall define tactics when player A moves *first* and second does player B. We shall call them first-type tactics. We give the definition in the more general form than in Section 4.1. Such a tactic is constructed from player A's strategy π_A and stopping time τ_A and player B's strategy π_B and stopping time τ_B. First-type strategies $(\pi_A; \pi_B)$ are defined as follows. For $s = (s^1, \cdots, s^d) \in T$,

$$\pi_A = \{\pi_A(|s| + t)\}_{t \in N(o,\infty)} = (\pi_A^1(|s| + t), \cdots, \pi_A^d(|s| + t))_{t \in N(o,\infty)}$$

and

$$\pi_B = \{\pi_B(|s| + t)\}_{t \in N(e,\infty)} = (\pi_B^1(|s| + t), \cdots, \pi_B^d(|s| + t))_{t \in N(e,\infty)}$$

are T-valued stochastic sequences on $(\Omega, \mathcal{F})$ satisfying the following (i) — (iii) and player A's stopping times τ_A (player B's τ_B) are $N(e,\infty)$ ($N(o,\infty)$)-valued random variables on $(\Omega, \mathcal{F})$ satisfying the adaptation (iv):

(i) $\pi_A(|s|) = s$ and $\pi_B(|s|) = s$.

(ii) For all $t \in N(e,\infty)$ it holds that $\pi_A(|s|+t+1) = \pi_B(|s|+t)+e_i$ for some $i = 1, \cdots, d$, and
for all $t \in N(o,\infty)$ it holds that $\pi_B(|s|+t+1) = \pi_A(|s|+t)+e_i$ for some $i = 1, \cdots, d$.

(iii) For all $t \in N(o,\infty)$ ($N(e,\infty)$) and all $s' \in T$ it holds that

$$\{\pi_A(|s| + t) = s'\} \in \mathcal{F}_{s'} \ (\{\pi_B(|s| + t) = s'\} \in \mathcal{F}_{s'} \text{ resp.}).$$

(iv) For $t \in N(e,\infty)$ it holds that $\{\tau_A = t\} \in \mathcal{F}_{\pi_B(t)}$, and
for $t \in N(o,\infty)$ it holds that $\{\tau_B = t\} \in \mathcal{F}_{\pi_A(t)}$,

where $\{\mathcal{F}_{\pi_A(t)}\}_{t \in N(o,\infty)}$ and $\{\mathcal{F}_{\pi_B(t)}\}_{t \in N(e,\infty)}$ denote informations until time t:

$$\mathcal{F}_{\pi_A(t)} = \{\Gamma \in \mathcal{F} \mid \Gamma \cap \pi_A(t) = s \in \mathcal{F}_s \text{ for } s \in N^d\} \quad \text{for } t \in N(o,\infty);$$
$$\mathcal{F}_{\pi_B(t)} = \{\Gamma \in \mathcal{F} \mid \Gamma \cap \pi_B(t) = s \in \mathcal{F}_s \text{ for } s \in N^d\} \quad \text{for } t \in N(e,\infty).$$

We similarly define strategies and stopping times when player B moves first and *second* does player A, which will be called second-type, by exchanging $N(e,\infty)$ with $N(o,\infty)$. Then we define families of first-type (second-type resp.) strategies and tactics. For $s \in T$,

$$\mathcal{S}(F; s) \ (\mathcal{S}(S; s)) = \{ \text{ first (second)-type strategies } (\pi_A; \pi_B) \text{ starting at } s\};$$

$$\mathcal{T}(F;s)\ (\mathcal{T}(S;s)) = \{\text{ first (second)-type tactics } (\pi_A,\tau_A;\pi_B,\tau_B) \text{ starting at } s\};$$

$$\mathcal{S}(F)\ (\mathcal{S}(S)) = \mathcal{S}(F;\mathbf{0})\ (\mathcal{S}(S;\mathbf{0})); \text{ and } \mathcal{T}(F)\ (\mathcal{T}(S)) = \mathcal{T}(F;\mathbf{0})\ (\mathcal{T}(S;\mathbf{0})).$$

Hence when one player's tactic is fixed, the other player's admissible tactics are denoted as follows. For $s \in T$, we respectively define

$$\mathcal{D}(F;s;\pi_B,\tau_B)\ (\mathcal{S}(S;s;\pi_B,\tau_B)) = \{(\pi_A,\tau_A) \mid (\pi_A,\tau_A;\pi_B,\tau_B) \in \mathcal{T}(F;s)\ (\mathcal{T}(S;s))\};$$

$$\mathcal{D}(F;s;\pi_A,\tau_A)\ (\mathcal{S}(S;s;\pi_A,\tau_A)) = \{(\pi_B,\tau_B) \mid (\pi_A,\tau_A;\pi_B,\tau_B) \in \mathcal{T}(F;s)\ (\mathcal{T}(S;s))\};$$

$$\mathcal{D}(F;\pi_B,\tau_B)\ (\mathcal{S}(S;\pi_B,\tau_B)) = \{(\pi_A,\tau_A) \mid (\pi_A,\tau_A;\pi_B,\tau_B) \in \mathcal{T}(F)\ (\mathcal{T}(S))\};$$

$$\mathcal{D}(F;\pi_A,\tau_A)\ (\mathcal{S}(S;\pi_A,\tau_A)) = \{(\pi_B,\tau_B) \mid (\pi_A,\tau_A;\pi_B,\tau_B) \in \mathcal{T}(F)\ (\mathcal{T}(S))\}.$$

We shall introduce the definition of Markov strategies, referring to [LawVan1]. For a first-type strategy π_A (π_B) is called Markov if it satisfies the following (i) ((ii) resp.):

(i) $\pi_A(t+1)$ are $\mathcal{G}_{\pi_B(t)}$-measurable for all $t \in N(e,\infty)$;

(ii) $\pi_B(t+1)$ are $\mathcal{G}_{\pi_A(t)}$-measurable for all $t \in N(o,\infty)$;

where $\mathcal{G}_{\pi_B(t)} = \sigma\{X_{\pi_B(t)}, \pi_B(t)\}$ [‖] and $\mathcal{G}_{\pi_A(t)} = \sigma\{X_{\pi_A(t)}, \pi_A(t)\}$. Further a first-type stopping time τ_A (τ_B) is called a *Markov* stopping time if it satisfies (a) ((b) resp.):

(a) $\tau_A \wedge (t+2)$ are $\mathcal{G}_{\pi_B(\tau_A \wedge t)}$-measurable for all $t \in N(e,\infty)$.

(b) $\tau_B \wedge (t+2)$ are $\mathcal{G}_{\pi_A(\tau_B \wedge t)}$-measurable for all $t \in N(o,\infty)$.

Hence it is known from [LawVan1,p.645] that (a) ((b)) is equivalent to the following conditions (a') ((b') resp.):

(a') For any $t \in N(e,\infty)$, there exists $\Gamma_t \in \mathcal{G}_{\pi_B(t)}$ such that $\{\tau_A = t\} = \{\tau_A \geq t\} \cap \Gamma_t$.

(b') For any $t \in N(o,\infty)$, there exists $\Gamma'_t \in \mathcal{G}_{\pi_A(t)}$ such that $\{\tau_B = t\} = \{\tau_B \geq t\} \cap \Gamma'_t$.

Regarding second-type strategies and stopping times, we similarly define Markov strategies and Markov stopping times, by exchanging $N(e,\infty)$ with $N(o,\infty)$. Hence we put families of first-type (second-type resp.) Markov strategies and Markov tactics by

$$\mathcal{MS}(F)\ (\mathcal{MS}(S)) = \{\text{ Markov strategies } (\pi_A;\pi_B) \in \mathcal{S}(F)\ (\mathcal{S}(S))\};$$

[‖]This denotes the minimum completed sub-σ-field generated by the random variables $X_{\pi_B(t)}$ and $\pi_B(t)$.

$\mathcal{MT}(F)$ $(\mathcal{MT}(S))$ = { Markov tactics $(\pi_A, \tau_A; \pi_B, \tau_B) \in \mathcal{T}(F)$ $(\mathcal{T}(S))$,
i.e. $(\pi_A; \pi_B)$ are Markov strategies, and τ_A and τ_B are Markov stopping times }.

Regarding Markov strategies (Markov tactics, Markov stopping times), when we focus on only the options from time 0 to time $r(\in N)$, we shall call them r-steps Markov strategies (r-steps Markov tactics, Markov stopping times resp.). Hence we put families of first-type (second-type resp.) r-steps Markov strategies and r-steps Markov tactics by

$$\mathcal{MS}(F; r)\ (\mathcal{MS}(S; r)) = \{r\text{-steps Markov strategies } (\pi_A; \pi_B) \in \mathcal{S}(F)\ (\mathcal{S}(S))\};$$

$$\mathcal{MT}(F; r)\ (\mathcal{MT}(S; r)) = \{r\text{-steps Markov tactics } (\pi_A, \tau_A; \pi_B, \tau_B) \in \mathcal{T}(F)\ (\mathcal{T}(S))\}$$

for $r \in N$. Especially since $(\pi_A; \pi_B) \in \mathcal{MS}(F; 1)$ $(\mathcal{MS}(S; 1))$ does not depend on π_B (π_A), we shall represent it only $\pi_A \in \mathcal{MS}(F; 1)$ $(\pi_B \in \mathcal{MS}(S; 1)$ resp.). From the same reason we also write $(\pi_A, \tau_A) \in \mathcal{MT}(F; 1)$ $((\pi_B, \tau_B) \in \mathcal{MT}(S; 1)$ resp.).

Hence we shall prepare fundamental lemmas concerning concatenations of Markov strategies and concatenations of Markov stopping times. In the rest of Section 4.2 we shall deal with the first-type game. Regarding second-type cases, similar results hold, by exchanging $N(e, \infty)$ with $N(o, \infty)$.

Lemma 4.1. *The following (i) and (ii) hold:*

(i) For $r \in N(e, \infty)(N(o, \infty)$ resp.), $(\pi_A; \pi_B) \in \mathcal{MS}(F; r)$ and $\pi'_A \in \mathcal{MS}(F; 1)$ $(\pi'_B \in \mathcal{MS}(S; 1)$ resp.), we define a concatenated strategy $(\pi''_A; \pi''_B)$ of $(\pi_A; \pi_B)$ and π'_A (π'_B):

$$\pi''_A(t, \omega) = \pi_A(t, \omega) \quad \text{for } t \in N(o, r+1) \text{ and } \omega \in \Omega;$$
$$\pi''_B(t, \omega) = \pi_B(t, \omega) \quad \text{for } t \in N(e, r+1) \text{ and } \omega \in \Omega; \text{ and}$$
$$\pi''_A(r+1, \omega) = \pi_B(t, \omega) + \pi'_A(1, \theta_{\pi_B(r)}\omega) \quad \text{for } \omega \in \Omega$$
$$(\pi''_B(r+1, \omega) = \pi_A(t, \omega) + \pi'_B(1, \theta_{\pi_A(r)}\omega) \quad \text{for } \omega \in \Omega).$$

Then $(\pi''_A; \pi''_B) \in \mathcal{MS}(F; r+1)$.

(ii) For $r \in N(e, \infty)(N(o, \infty)$ resp.), $(\pi_A; \pi_B) \in \mathcal{MS}(F; r)$ and a non-increasing sequence $\{\tau_{A,t}\}_{t \in N(e,r)}(\{\tau_{B,t}\}_{t \in N(o,r)})$ of first-type 2-steps Markov stopping times, we inductively define a stopping time $\tau'_{A,r}(\tau'_{B,r})$ by

$$\tau'_{A,0}(\omega) = \tau_{A,0}(\omega) \quad \text{for } \omega \in \Omega; \text{ and}$$
$$\tau'_{A,t+2}(\omega) = \tau'_{A,t}(\omega) + \tau_{A,t+2}(\theta_{\pi_B(\tau'_{A,t})}\omega) \quad \text{for } t \in N(e, r) \text{ and } \omega \in \Omega$$
$$(\tau'_{B,1}(\omega) = 1 + \tau_{B,1}(\theta_{\pi_A(1)}\omega) \quad \text{for } \omega \in \Omega; \text{ and}$$

$$\tau'_{B,t+2}(\omega) = \tau'_{B,t}(\omega) + \tau_{B,t+2}(\theta_{\pi_A(\tau'_{B,t})}\omega) \quad \text{for } t \in N(o,r) \text{ and } \omega \in \Omega).$$

Then $\tau'_{A,r}$ ($\tau'_{B,r}$) s a first-type r-steps Markov stopping time and further $\lim_{r\to\infty} \tau'_{A,r}$ ($\lim_{r\to\infty} \tau'_{B,r}$) becomes player A's (player B's) first-type Markov stopping time.

Proof. (i) are trivial due to the definitions of Markov strategies. (ii) Fix any $r \in N(e,\infty)$, $(\pi_A;\pi_B) \in \mathcal{MS}(F;r)$ and a non-increasing sequence $\{\tau_{A,t}\}_{t\in N(e,r)}$ of first-type 2-steps Markov stopping times. Since $\{\tau_{A,t}\}_{t\in N(e,r)}$ is non-increasing, we have

$$\tau'_{A,r} = \inf\{t \in N(e,\infty) \mid \tau_{A,t}(\theta_{\pi_B(t)}) = 0\} \wedge (r+2). \tag{4.1}$$

Hence since $\tau'_{A,t}$ is $\mathcal{G}_{\pi_B(0)}$-measurable, we can easily check (a'), by taking the measurable sets $\Gamma_t = \{\tau_{A,t}(\theta_{\pi_B(t)}) = 0\} \in \mathcal{G}_{\pi_B(t)}$ for $t \in N(e,r)$. Thus we obtain (ii). The proofs of the other cases in (ii) is similar. □

A typical example of Markov stopping times is as follows. Let $(\pi_A;\pi_B) \in \mathcal{MS}(F;r)$ and let a sequence $\{D_{A,t}\}_{t\in N(e,\infty)}$ of Borel subsets of E such that $D_{A,t+2} \supset D_{A,t}$ for each $t \in N(e,\infty)$. Hence we define the following stopping times (4.2):

$$\tau_{A,t} = \begin{cases} 2 & \text{on } \{X_0 \notin D_{A,t}\} \\ 0 & \text{on } \{X_0 \in D_{A,t}\} \end{cases} \text{ for } t \in N(e,\infty). \tag{4.2}$$

Then we have the following result. Similar results also hold for player B.

Lemma 4.2. *In Lemma 4.1(ii) if we give $\tau_{A,t}$ by (4.2), then we have*

$$\tau'_{A,r} = \inf\{t \in N(e,\infty) \mid X_{\pi_B(t)} \in D_{A,t}\} \wedge (r+2) \quad \text{for } r \in N(e,\infty), \tag{4.3}$$

and

$$\tau'_A = \inf\{t \in N(e,\infty) \mid X_{\pi_B(t)} \in D_{A,t}\}. \tag{4.4}$$

Proof. This lemma is trivial from the definitions. □

From Lemma 4.1, we have the following representations for Markov tactics: For player A's (player B's) first-type Markov tactics $(\pi_A,\tau_A)((\pi_B,\tau_B)$ resp.) we write them as

$$(\pi_A,\tau_A) = [\pi_{A,1},\tau_{A,1};\pi_{A,3},\tau_{A,3};\pi_{A,5},\tau_{A,5};\cdots]; \text{ and } ((\pi_B,\tau_B) = [\pi_{B,2},\tau_{B,2};\pi_{B,4},\tau_{B,4};\cdots]).$$

4.3. Expected rewards and bandit games

First we shall define player A's expected values and player B's when player A moves *first*. For arm $i(=1,\cdots,d)$ let f^i_A (f^i_B), player A's (player B's resp.) running rewards

for arm i, be bounded measurable functions on E^i and let h_A (h_B), player A's (player B's resp.) terminal rewards, be bounded measurable functions on E. Put $h = h_A - h_B$. Hence for the sake of a convenient representation we shall introduce the following notation $\langle \cdot, \cdot \rangle$, referring to the inner product of d-dimensional real vector spaces: For example, we describe

$$\langle f_A(X_{\pi_A(1)}), \pi_A(1) - \pi_B(0) \rangle = \sum_{i=1}^{d} f_A^i(X^i_{\pi^i_A(1)})(\pi^i_A(1) - \pi^i_B(0)). \tag{4.5}$$

Let β_0 be a constant satisfying $0 < \beta_0 < 1$. For arm $i(= 1, \cdots, d)$ let β^i, a instantanous discount rate for arm i, be a bounded measurable function on E^i satisfying that

$$0 < \beta^i(x^i) \leq \beta_0 \quad \text{for all } i = 1, \cdots, d \text{ and all } x^i \in E^i. \tag{4.6}$$

Then for a strategy $(\pi_A; \pi_B)(\in \mathcal{S}(F))$, we define a discount rate at odd (even resp.) time $t+1$, which depends on a state of arms selected by player A's strategy $\pi_A(t+1)$ (player B's $\pi_B(t+1)$):

$$\beta^{\pi_A \pi_B}(t+1) = \langle \beta(X_{\pi_A(t+1)}), \pi_A(t+1) - \pi_B(t) \rangle \quad \text{for } t \in N(e, \infty); \tag{4.7}$$

and

$$\beta^{\pi_A \pi_B}(t+1) = \langle \beta(X_{\pi_B(t+1)}), \pi_B(t+1) - \pi_A(t) \rangle \quad \text{for } t \in N(o, \infty). \tag{4.8}$$

Using these, a discount rate at time t is given by their product:

$$\alpha^{\pi_A \pi_B}(0) = 1, \quad \text{and} \quad \alpha^{\pi_A \pi_B}(t) = \prod_{r=1}^{t} \beta^{\pi_A \pi_B}(r) \quad \text{for } t = 1, 2, \cdots. \tag{4.9}$$

Expecially since $\alpha^{\pi_A \pi_B}(1) = \beta^{\pi_A \pi_B}(1) = \langle \beta(X_{\pi_A(1)}), \pi_A(1) \rangle$ does not include π_B, we write it simply as $\beta^{\pi_A}(1)$. When a first-type tactic $(\pi_A, \tau_A; \pi_B, \tau_B)(\in \mathcal{T}(F))$ is taken, player A's expected gain to be paid from player B at an initial state x is represented as sums of gains when player A stops the games and gains when does player B :

$$\begin{aligned} & V_F^{\pi_A \tau_A \pi_B \tau_B}(x) \\ &= E^x[\sum_{t \in N(e, \tau_A \wedge \tau_B)} \alpha^{\pi_A \pi_B}(t) \langle f_A(X_{\pi_A(t+1)}), \pi_A(t+1) - \pi_B(t) \rangle \\ &\quad - \sum_{t \in N(o, \tau_A \wedge \tau_B)} \alpha^{\pi_A \pi_B}(t) \langle f_B(X_{\pi_B(t+1)}), \pi_B(t+1) - \pi_A(t) \rangle \\ &\quad + \alpha^{\pi_A \pi_B}(\tau_A \wedge \tau_B) h(X_{\pi_B(\tau_A) \wedge \pi_A(\tau_B)})], \end{aligned}$$

where we define $a \wedge b = \min\{a, b\}$ for $a, b \in N$ and $a \wedge b = (a^1 \wedge b^1, a^2 \wedge b^2, \cdots, a^d \wedge b^d)$ for $a = (a^1, a^2, \cdots, a^d), b = (b^1, b^2, \cdots, b^d) \in T$. Hence player A's aim is to maximize his gains $V_F^{\pi_A \tau_A \pi_B \tau_B}(x)$, by controlling his strategies and stopping times, however player B's is to

minimize it. Therefore when one player's tactic is fixed, the values optimized by another player are as follows:

$$V_F^{*\pi_B\tau_B}(x) = \sup_{(\pi_A,\tau_A)\in\mathcal{D}(F;\pi_B,\tau_B)} V_F^{\pi_A\tau_A\pi_B\tau_B}(x) \quad \text{for } x \in E; \tag{4.10}$$

$$V_F^{\pi_A\tau_A*}(x) = \inf_{(\pi_B,\tau_B)\in\mathcal{D}(F;\tau_A,\pi_A)} V_F^{\pi_A\tau_A\pi_B\tau_B}(x) \quad \text{for } x \in E. \tag{4.11}$$

When player A moves *first*, we shall call the following game first-type bandit game: To find tactics

$$(\pi_A^*, \tau_A^*; \pi_B^*, \tau_B^*) \in \mathcal{T}(F) \text{ such that } V_F^{\pi_A^*\tau_A^*\pi_B^*\tau_B^*} = V_F^{*\pi_B\tau_B} = V_F^{\pi_A\tau_A*}. \tag{4.12}$$

Next we shall similarly define values of games when player A moves second. We put, for a second-type tactic $(\pi_A, \tau_A; \pi_B, \tau_B)(\in \mathcal{T}(S))$ and $x(\in E)$,

$$\begin{aligned} &V_S^{\pi_A\tau_A\pi_B\tau_B}(x) \\ &= E^x[\sum_{t\in N(o,\tau_A\wedge\tau_B)} \alpha^{\pi_A\pi_B}(t)\langle f_A(X_{\pi_A(t+1)}), \pi_A(t+1) - \pi_B(t)\rangle \\ &\quad - \sum_{t\in N(e,\tau_A\wedge\tau_B)} \alpha^{\pi_A\pi_B}(t)\langle f_B(X_{\pi_B(t+1)}), \pi_B(t+1) - \pi_A(t)\rangle \\ &\quad + \alpha^{\pi_A\pi_B}(\tau_A\wedge\tau_B)h(X_{\pi_B(\tau_A)\wedge\pi_A(\tau_B)})]; \end{aligned}$$

$$V_S^{*\pi_B\tau_B}(x) = \sup_{(\pi_A,\tau_A)\in\mathcal{D}(S;\pi_B,\tau_B)} V_S^{\pi_A\tau_A\pi_B\tau_B}(x) \quad \text{for } x \in E; \tag{4.13}$$

$$V_S^{\pi_A\tau_A*}(x) = \inf_{(\pi_B,\tau_B)\in\mathcal{D}(S;\tau_A,\pi_A)} V_S^{\pi_A\tau_A\pi_B\tau_B}(x) \quad \text{for } x \in E. \tag{4.14}$$

Then second-type bandit games are to find

$$(\pi_A^*, \tau_A^*; \pi_B^*, \tau_B^*) \in \mathcal{T}(S) \text{ such that } V_S^{\pi_A^*\tau_A^*\pi_B^*\tau_B^*} = V_S^{*\pi_B\tau_B} = V_S^{\pi_A\tau_A*}. \tag{4.15}$$

Finally we note that

$$V_F^{\pi_A\tau_A\pi_B\tau_B} = h \text{ if } \tau_A = 0; \quad \text{and} \quad V_S^{\pi_A\tau_A\pi_B\tau_B} = h \text{ if } \tau_B = 0. \tag{4.16}$$

We need some more notations in order to prove existence of Markov tactics attaining the supremum (the infremum) in (4.10) and (4.11) ((4.13) and (4.14) resp.). Set $s = (s^1, \cdots, s^d) \in T$ such that $|s|$ is even (odd). After observations that each arm i has already been selected s^i times and that we adopt a tactic $(\pi_A, \tau_A; \pi_B, \tau_B) \in \mathcal{T}(F; s)$, values of first-type bandit games are denoted by

$$\begin{aligned} &Z_F^{\pi_A\tau_A\pi_B\tau_B}(s) \\ &= E^{\mathcal{F}_s}[\sum_{t\in N(e,\tau_A\wedge\tau_B-|s|)} \alpha^{\pi_A\pi_B}(|s|,|s|+t)\langle f_A(X_{\pi_A(|s|+t+1)}), \pi_A(|s|+t+1) - \pi_B(|s|+t)\rangle \\ &\quad - \sum_{t\in N(o,\tau_A\wedge\tau_B-|s|)} \alpha^{\pi_A\pi_B}(|s|,|s|+t)\langle f_B(X_{\pi_B(|s|+t+1)}), \pi_B(|s|+t+1) - \pi_A(|s|+t)\rangle \\ &\quad + \alpha^{\pi_A\pi_B}(0,\tau_A\wedge\tau_B)h(X_{\pi_B(\tau_A)\wedge\pi_A(\tau_B)}) : \ |s| \le \tau_A\wedge\tau_B], \end{aligned}$$

where $\alpha^{\pi_A\pi_B}(|s|, |s|+t)$ is a product of discount rates from time $|s|$ to time $|s|+t$:

$$\alpha^{\pi_A\pi_B}(r,r) = 1, \quad \text{and} \quad \alpha^{\pi_A\pi_B}(r,t) = \prod_{r'=r+1}^{t} \beta^{\pi_A\pi_B}(r') \quad \text{for } r,t = 1,2,\cdots(r<t). \tag{4.17}$$

Hence referring to (4.10) and (4.11), we put

$$Z_F^{*\pi_B\tau_B}(s) = \text{ess sup}_{(\pi_A,\tau_A)\in\mathcal{D}(F;s;\pi_B,\tau_B)} Z_F^{\pi_A\tau_A\pi_B\tau_B}(s) \quad \text{for } s \in T; \tag{4.18}$$

and

$$Z_F^{\pi_A\tau_A*}(s) = \text{ess inf}_{(\pi_B,\tau_B)\in\mathcal{D}(F;s;\pi_A,\tau_A)} Z_F^{\pi_A\tau_A\pi_B\tau_B}(s) \quad \text{for } s \in T. \tag{4.19}$$

Regarding the second-type bandit games, we define $Z_S^{\pi_A\tau_A\pi_B\tau_B}(s)$, $Z_S^{*\pi_B\tau_B}(s)$ and $Z_S^{\pi_A\tau_A*}(s)$ similarly. Now we obtain the following fundamental lemmas.

Lemma 4.3. *The following (i) and (ii) hold:*

(i) For $(\pi_A,\tau_A;\pi_B,\tau_B) \in \mathcal{T}(F)$ and $r \in N(e,\infty)$, it holds that

$$\begin{aligned} & Z_F^{\pi_A\tau_A\pi_B\tau_B}(\pi_B(r)) \\ &= E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_A(r+1)}), \pi_A(r+1) - \pi_B(r)\rangle \\ &\quad +\beta^{\pi_A\pi_B}(r+1)\ Z_S^{\pi_A\tau_A\pi_B\tau_B}(\pi_A(r+1)) : r+1 \le \tau_A \wedge \tau_B] \\ &\quad +E^{\mathcal{F}_{\pi_B(r)}}[h(X_{\pi_B(r)}) : r = \tau_A \wedge \tau_B]. \end{aligned}$$

(ii) For $(\pi_A,\tau_A;\pi_B,\tau_B) \in \mathcal{T}(F)$ and $r \in N(o,\infty)$, it holds that

$$\begin{aligned} & Z_S^{\pi_A\tau_A\pi_B\tau_B}(\pi_A(r)) \\ &= E^{\mathcal{F}_{\pi_A(r)}}[\langle -f_B(X_{\pi_B(r+1)}), \pi_B(r+1) - \pi_A(r)\rangle \\ &\quad +\beta^{\pi_A\pi_B}(r+1)\ Z_F^{\pi_A\tau_A\pi_B\tau_B}(\pi_B(r+1)) : r+1 \le \tau_A \wedge \tau_B] \\ &\quad +E^{\mathcal{F}_{\pi_A(r)}}[h(X_{\pi_A(r)}) : r = \tau_A \wedge \tau_B]. \end{aligned}$$

Proof. We can easily check it from the definitions. □

Lemma 4.4. *The following (i) and (ii) hold:*

(i) For player B's first-type Markov tactics (π_B,τ_B) and $r \in N(e,\infty)$, it holds that

$$\begin{aligned} & Z_F^{*\pi_B\tau_B}(\pi_B(r)) \\ &= \max\{\text{ess sup}_{(\pi_A,\tau_A)\in\mathcal{D}(F;\pi_B(r);\pi_B,\tau_B)} E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_A(r+1)}), \pi_A(r+1) - \pi_B(r)\rangle \\ &\quad +\beta^{\pi_A\pi_B}(r+1)\ Z_S^{*\pi_B\tau_B}(\pi_A(r+1)) : r+1 \le \tau_A \wedge \tau_B], h(X_{\pi_B(r)})\}. \end{aligned}$$

(ii) For player A's first-type Markov tactics (π_A, τ_A) and $r \in N(o, \infty)$, it holds that

$$
\begin{aligned}
& Z_S^{\pi_A \tau_A *}(\pi_A(r)) \\
&= \min\{\text{ess inf}_{(\pi_B,\tau_B)\in\mathcal{D}(F;\pi_A(r);\pi_A,\tau_A)} E^{\mathcal{F}_{\pi_A(r)}}[\langle -f_B(X_{\pi_B(r+1)}), \pi_B(r+1) - \pi_A(r)\rangle \\
&\quad + \beta^{\pi_A \pi_B}(r+1)\ Z_F^{\pi_A \tau_A *}(\pi_B(r+1)) : r+1 \le \tau_A \wedge \tau_B], h(X_{\pi_A(r)})\}.
\end{aligned}
$$

Proof. Fix any Markov tactics (π_B, τ_B). We shall show only this case of (i), because the other case is similar. Let any $r \in N(e, \infty)$ and any $(\pi_A, \tau_A) \in \mathcal{D}(F; \pi_B(r); \pi_B, \tau_B)$. The definition of the essential supremum (see [Nev1,p.121]) implies that there exists a sequence $\{(\pi_{A,n}, \tau_{A,n}; \pi_B, \tau_B) \in \mathcal{T}(F)\}_{n\in N}$ of tactics satisfying the following conditions which they are equal to (π_A, τ_A) until time $r+1$:

$$\pi_{A,n}(t) = \pi_A(t) \text{ for all odd t satisfying } 0 \le t \le r+1; \tag{4.20}$$

$$\tau_{A,n} \wedge (r+1) = \tau_A \wedge (r+1); \text{and} \tag{4.21}$$

$$
\begin{aligned}
& E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_A(r+1)}), \pi_A(r+1) - \pi_B(r)\rangle \\
&= \lim_{n\to\infty}\{E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_{A,n}(r+1)}), \pi_{A,n}(r+1) - \pi_B(r)\rangle.
\end{aligned} \tag{4.22}
$$

Then we have the previous term $= \lim_{n\to\infty} Z_F^{\pi_{A,n}\tau_{A,n}\pi_B\tau_B}(\pi_B(r)) \le Z_F^{*\pi_B\tau_B}(\pi_B(r))$. However since $\{r+1 \le \tau_A \wedge \tau_B\} \cap \{r = \tau_A \wedge \tau_B\}$ is empty, we obtain

$$
\begin{aligned}
& Z_F^{*\pi_B\tau_B}(\pi_B(r)) \\
&\ge \max\{\text{ess sup}_{(\pi_A,\tau_A)\in\mathcal{D}(F;\pi_B(r);\pi_B,\tau_B)} E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_A(r+1)}), \pi_A(r+1) - \pi_B(r)\rangle \\
&\quad + \beta^{\pi_A\pi_B}(r+1)\ Z_S^{*\pi_B\tau_B}(\pi_A(r+1)) : r+1 \le \tau_A \wedge \tau_B], h(X_{\pi_B(r)})\}.
\end{aligned}
$$

On the other hand Lemma 4.3 implies

$$
\begin{aligned}
& Z_F^{\pi_A\tau_A\pi_B\tau_B}(\pi_B(r)) \\
&= E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_A(r+1)}), \pi_A(r+1) - \pi_B(r)\rangle \\
&\quad + \beta^{\pi_A\pi_B}(r+1)\ Z_S^{\pi_A\tau_A\pi_B\tau_B}(\pi_A(r+1)) : r+1 \le \tau_A \wedge \tau_B] + E^{\mathcal{F}_{\pi_B(r)}}[h(X_{\pi_B(r)}) : r = \tau_A \wedge \tau_B] \\
&\le \max\{\text{ess sup}_{(\pi_A,\tau_A)\in\mathcal{D}(F;\pi_B(r);\pi_B,\tau_B)} E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_A(r+1)}), \pi_A(r+1) - \pi_B(r)\rangle \\
&\quad + \beta^{\pi_A\pi_B}(r+1)\ Z_S^{*\pi_B\tau_B}(\pi_A(r+1)) : r+1 \le \tau_A \wedge \tau_B], h(X_{\pi_B(r)})\}.
\end{aligned}
$$

Therefore we obtain the equality of (i). □

Hence in order to check the measurability of $Z_F^{*\pi_B\tau_B}$ and $Z_S^{*\pi_B\tau_B}$, we define, for $m = 1, 2, \cdots$ and fixed player B's first-type Markov tactics (π_B, τ_B), $\{Y_{F,m}(s)\}_{s\in T:|s|\in N(e,m+1)}$ and $\{Y_{S,m}(s)\}_{s\in T:|s|\in N(o,m+1)}$ in m-step first-type bandit game successively:

Iteration 4.1.

$(m.0)$ *For* $s \in T$ *satisfying* $|s| = m$, *put* $Y_{F,m}(s) = Y_{S,m}(s) = 0$.

$(m.F.s)$ *For* $s \in T$ *satisfying* $|s| \in N(e,m)$, *put*

$$\begin{aligned} Y_{F,m}(s) \;=\; & \max\{ess\ \sup_{(\pi_A,\tau_A)\in\mathcal{D}(F;s;\pi_B,\tau_B)} E^{\mathcal{F}_s}[\langle f_A(X_{\pi_A(|s|+1)}), \pi_A(|s|+1) - s\rangle \\ & +\langle \beta(X_{\pi_A(|s|+1)}), \pi_A(|s|+1) - s\rangle\ Y_{S,m}(\pi_A(|s|+1)) : |s|+1 \le \tau_A \wedge \tau_B], h(X_s)\}. \end{aligned}$$

$(m.S.s)$ *For* $s \in T$ *satisfying* $|s| \in N(o,m)$, *put*

$$\begin{aligned} Y_{S,m}(s) \;=\; & \sum_{i=1}^{d} E^{\mathcal{F}_s}[\langle -f_B(X_{s+e_i}), e_i\rangle \\ & +\langle \beta(X_{s+e_i}), e_i\rangle\ Y_{F,m}(s+e_i) : s+e_i = \pi_B(|s|+1), |s|+1 \le \tau_B] \\ & +h(X_s)\cdot I_{\{|s|=\tau_B\}}. \end{aligned}$$

Then we have the following lemma.

Lemma 4.5. *The following (i) and (ii) hold:*

(i) For player B's first-type Markov tactics (π_B, τ_B) *and* $r \in N(e,\infty)$, *it holds that*

$$\begin{aligned} & Z_F^{*\pi_B\tau_B}(\pi_B(r)) \\ & = \max\{ \sup_{\pi'_A\in\mathcal{MS}(F;1)} E^{X_{\pi_B(r)}}[\langle f_A(X_{\pi'_A(1)}), \pi'_A(1)\rangle \\ & \quad +\beta^{\pi'_A}(1)\ Z_S^{*\pi'_B\tau'_B}(\pi'_A(1)) : 1 \le \tau'_B], h(X_{\pi_B(r)})\}, \end{aligned}$$

where we take (π'_B, τ'_B) *by* $(\pi'_B, \tau'_B) = [\pi_{B,r+2}, \tau_{B,r+2}; \pi_{B,r+4}, \tau_{B,r+4}; \pi_{B,r+6}, \tau_{B,r+6}; \cdots]$ *for* $(\pi_B, \tau_B) = [\pi_{B,2}, \tau_{B,2}; \pi_{B,4}, \tau_{B,4}; \pi_{B,6}, \tau_{B,6}; \cdots]$.

(ii) For player A's first-type Markov tactics (π_A, τ_A) *and* $r \in N(o,\infty)$, *it holds that*

$$\begin{aligned} & Z_S^{\pi_A\tau_A*}(\pi_A(r)) \\ & = \min\{ \inf_{\pi'_B\in\mathcal{MS}(S;1)} E^{X_{\pi_A(r)}}[\langle -f_B(X_{\pi'_B(1)}), \pi'_B(1)\rangle \\ & \quad +\beta^{\pi'_B}(1)\ Z_F^{\pi'_A\tau'_A*}(\pi'_B(1)) : 1 \le \tau'_A], h(X_{\pi_A(r)})\}, \end{aligned}$$

where we take (π'_A, τ'_A) *by* $(\pi'_A, \tau'_A) = [\pi_{A,r+2}, \tau_{A,r+2}; \pi_{A,r+4}, \tau_{A,r+4}; \pi_{A,r+6}, \tau_{A,r+6}; \cdots]$ *for* $(\pi_A, \tau_A) = [\pi_{A,1}, \tau_{A,1}; \pi_{A,3}, \tau_{A,3}; \pi_{A,5}, \tau_{A,5}; \cdots]$.

Proof. We shall show (i), because the other cases are similar. Fix any Markov tactics (π_B, τ_B). First we shall show $Y_{F,m}(s) \in \mathcal{G}_s$ ** for all $s \in T$ satisfying $|s| \in N(e, m+$

**$\mathcal{G}_s$-measurable functions

1) ($Y_{S,m}(s) \in \mathcal{G}_s$ for all $s \in T$ satisfying $|s| \in N(o, m+1)$). Put $\mathcal{H}_{\pi_B(r)} = \sigma\{X_s : s \geq \pi_B(r), \pi_B(r)\}$ for $r \in N(e, \infty)$. Then from the independency of Markov chains X^i ($i = 1, \cdots, d$) (c.f. [LawVan1,Theorem 5(b)]), $\sigma\{X_{s'} : s' \geq s\}$ and $\mathcal{F}_s$ are conditionally independent, given $\mathcal{G}_s$ for $s \in T$. So from the definition of strategies, we can easily check that the future $\mathcal{H}_{\pi_B(r)}$ and the past $\mathcal{F}_{\pi_B(r)}$ are conditionally independent, given $\mathcal{G}_{\pi_B(r)}$ for $r \in N(e, \infty)$. Hence from Iteration1($m.0$) we have that $Y_{F,m}(s) = Y_{S,m}(s) = 0$ for $s \in T$ ($|s| = m$). Further in Iteration1($m.F.s$)(($m.S.s$) resp.) if $Y_{S,m}(s + e_j) \in \mathcal{G}_{s+e_j}$ for some $s \in T$ satisfying $|s| \in N(e, m)$ and all $j(= 1, \cdots, d)$ ($Y_{F,m}(s + e_j) \in \mathcal{G}_{s+e_j}$ for some $s \in T$ satisfying $|s| \in N(o, m)$ and all $j(= 1, \cdots, d)$), then the term

$$\langle f_A(X_{\pi_A(|s|+1)}), \pi_A(|s|+1) - s\rangle + \langle \beta(X_{\pi_A(|s|+1)}), \pi_A(|s|+1) - s\rangle \; Y_{S,m}(\pi_A(|s|+1)) \in \mathcal{H}_{\pi_B(|s|)}$$

$$(\langle -f_B(X_{s+e_i}), e_i\rangle + \langle \beta(X_{s+e_i}), e_i\rangle \; Y_{F,m}(s + e_i) \in \mathcal{H}_{\pi_A(|s|)}).$$

Further since (π_B, τ_B) is Markov, we have $\{s + e_i = \pi_B(|s|+1)\} \in \mathcal{G}_{\pi_A(|s|)}$ $(i = 1, \cdots, d)$, $\{|s| = \tau_B\} \in \mathcal{G}_{\pi_A(|s|)}$, $\{|s|+1 \leq \tau_A \wedge \tau_B\} \in \mathcal{H}_{\pi_A(|s|)}$ and $\{|s|+1 \leq \tau_B\} \in \mathcal{H}_{\pi_A(|s|)}$. Therefore from Iteration1($m.F.s$), we obtain $Y_{F,m}(s) \in \mathcal{G}_s$ for $s \in T$ satisfying $|s| \in N(e, m)$ (from Iteration1($m.S.s$), $Y_{S,m}(s) \in \mathcal{G}_s$ for $s \in T$ satisfying $|s| \in N(o, m)$). Thus inductively we can check backward that $Y_{F,m}(s) \in \mathcal{G}_s$ for all $s \in T$ satisfying $|s| \in N(e, m+1)$ ($Y_{S,m}(s) \in \mathcal{G}_s$ for all $s \in T$ satisfying $|s| \in N(o, m+1)$).

Next define a norm $|||\cdot|||$ on the space of bounded d-parameter processes on E: $|||W||| = \sup_{s \in T}$ ess $\sup_{\omega \in \Omega} |W_s(s, \omega)|$. for bounded d-parameter processes $W = \{W_s\}_{s \in T}$ on $\mathcal{F}$. By using the norm $|||\cdot|||$ instead of the norm $\|\cdot\|$ of Section 4.3, in similar line as Lemmas 4.5 and 4.6 and Theorem 4.1 (see Section 4.3), Iteration 1 converges:

$$Y_F(s) = \lim_{m \to \infty} Y_{F,m}(s) \in \mathcal{G}_s \text{ for all } s \in T \text{ satisfying } |s| \in N(e, \infty), \tag{4.23}$$

and

$$Y_S(s) = \lim_{m \to \infty} Y_{S,m}(s) \in \mathcal{G}_s \text{ for all } s \in T \text{ satisfying } |s| \in N(o, \infty). \tag{4.24}$$

Further the pair of Y_F and Y_S is a unique solution of the following (4.25) and (4.26):

$$\begin{aligned} Y_F(s) \;=\; & \max\{\text{ess } \sup_{(\pi_A, \tau_A) \in \mathcal{D}(F; s; \pi_B, \tau_B)} E^{\mathcal{F}_s}[\langle f_A(X_{\pi_A(|s|+1)}), \pi_A(|s|+1) - s\rangle \\ & + \langle \beta(X_{\pi_A(|s|+1)}), \pi_A(|s|+1) - s\rangle \; Y_S(\pi_A(|s|+1)) : |s|+1 \leq \tau_A \wedge \tau_B], h(X_s)\}. \end{aligned} \tag{4.25}$$

for all $s \in T$ satisfying $|s| \in N(e, \infty)$, and

$$\begin{aligned} Y_S(s) \;=\; & \sum_{i=1}^{d} E^{\mathcal{F}_s}[\langle -f_B(X_{s+e_i}), e_i\rangle \\ & + \langle \beta(X_{s+e_i}), e_i\rangle \; Y_F(s + e_i) : s + e_i = \pi_B(|s|+1), |s|+1 \leq \tau_B] + h(X_s) I_{\{|s| = \tau_B\}} \end{aligned} \tag{4.26}$$

for all $s \in T$ satisfying $|s| \in N(\bullet, \infty)$. Hence since player B's tactic (π_B, τ_B) is fixed, by using Lemmas 4.3(ii) and 4.4(i), $Z_F^{*\pi_B,\tau_B}$ and $Z_S^{*\pi_B,\tau_B}$ also satisfies (4.25) and (4.26). From the uniqueness of solutions, we obtain that $Y_F = Z_F^{*\pi_B,\tau_B}$ and $Y_S = Z_S^{*\pi_B,\tau_B}$. Therefore in (4.25) by using the Markov property and the measurability of (4.23), we obtain that for each $r \in N(e, \infty)$,

$$
\begin{aligned}
&Z_F^{*\pi_B\tau_B}(\pi_B(r)) \\
&= \max\{\text{ess sup}_{(\pi_A,\tau_A)\in\mathcal{D}(F;\pi_B(r);\pi_B,\tau_B)} E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_A(r+1)}), \pi_A(r+1) - \pi_B(r)\rangle \\
&\quad + \beta^{\pi_A\pi_B}(r+1)\ Z_S^{*\pi_B\tau_B}(\pi_A(r+1)) : r+1 \le \tau_A \wedge \tau_B], h(X_{\pi_B(r)})\}, \\
&= \max\{\sup_{\pi'_A\in\mathcal{MS}(F;1)} E^{X_{\pi_B(r)}}[\langle f_A(X_{\pi'_A(1)}), \pi'_A(1)\rangle + \beta^{\pi'_A}(1)\ Z_S^{*\pi'_B\tau'_B}(\pi'_A(1)) : 1 \le \tau'_B], h(X_{\pi_B(r)})\},
\end{aligned}
$$

where we take (π'_B, τ'_B) by $(\pi'_B, \tau'_B) = [\pi_{B,r+2}, \tau_{B,r+2}; \pi_{B,r+4}, \tau_{B,r+4}; \pi_{B,r+6}, \tau_{B,r+6}; \cdots]$ for $(\pi_B, \tau_B) = [\pi_{B,2}, \tau_{B,2}; \pi_{B,4}, \tau_{B,4}; \pi_{B,6}, \tau_{B,6}; \cdots]$. Thus we conclude this lemma. □

Hence we obtain the following results regarding (4.10), (4.11), (4.13) and (4.14).

Proposition 4.1. *For player B's (player A's) first-type Markov tactic (π_B, τ_B) $((\pi_A, \tau_A))$, there exist Markov tactics $(\pi_{A,M}, \tau_{A,M}) \in \mathcal{D}(F; \pi_B, \tau_B)$ $((\pi_{B,M}, \tau_{B,M}) \in \mathcal{D}(F; \pi_A, \tau_A))$ satisfying the following (i) ((ii) resp.):*

(i) $V_F^{\pi_{A,M}\tau_{A,M}\pi_B\tau_B} = V_F^{*\pi_B\tau_B}$.

(ii) $V_F^{\pi_A\tau_A\pi_{B,M}\tau_{B,M}} = V_F^{\pi_A\tau_A *}$.

Remark. Similar facts holds for the second-type tactics.

Proof. (i) Fix any player B's first-type Markov tactic (π_B, τ_B). Lemma 4.5 implies that for each $r \in N(e, \infty)$

$$
\begin{aligned}
&Z_F^{*\pi_B\tau_B}(\pi_B(r)) \\
&= \max\{\sup_{\pi'_A\in\mathcal{MS}(F;1)} E^{X_{\pi_B(r)}}[\langle f_A(X_{\pi'_A(1)}), \pi'_A(1)\rangle + \beta^{\pi'_A}(1)\ Z_S^{*\pi'_B\tau'_B}(\pi'_A(1)) : 1 \le \tau'_B], h(X_{\pi_B(r)})\},
\end{aligned}
$$

where we take (π'_B, τ'_B) by $(\pi'_B, \tau'_B) = [\pi_{B,r+2}, \tau_{B,r+2}; \pi_{B,r+4}, \tau_{B,r+4}; \pi_{B,r+6}, \tau_{B,r+6}; \cdots]$ for $(\pi_B, \tau_B) = [\pi_{B,2}, \tau_{B,2}; \pi_{B,4}, \tau_{B,4}; \pi_{B,6}, \tau_{B,6}; \cdots]$. Hence it holds that

$$
\begin{aligned}
&\max\{\sup_{\pi'_A\in\mathcal{MS}(F;1)} E^{X_{\pi_B(r)}}[\langle f_A(X_{\pi'_A(1)}), \pi'_A(1)\rangle + \beta^{\pi'_A}(1)\ Z_S^{*\pi'_B\tau'_B}(\pi'_A(1)) : 1 \le \tau'_B] \\
&\qquad = \max_{j=1,\cdots,d} E^{X_{\pi_B(r)}}[f_A^j(X_1^j) + \beta^j(X_1^j)\ Z_S^{*\pi'_B\tau'_B}(e_j) : 1 \le \tau'_B]. \qquad (4.27)
\end{aligned}
$$

Here for $i = 1, \cdots, d$ we define

$$\begin{aligned}\Gamma_i &= \{\max_{j=1,\cdots,d} E^{X_{\pi_B(r)}}[f_A^j(X_1^j) + \beta^j(X_1^j) Z_S^{*\pi'_B\tau'_B}(e_j) : 1 \le \tau'_B] \\ &\quad = E^{X_{\pi_B(r)}}[f_A^i(X_1^i) + \beta^i(X_1^i) Z_S^{*\pi'_B\tau'_B}(e_i) : 1 \le \tau'_B]\}.\end{aligned}$$

Further we set $\Gamma'_1 = \Gamma_1$ and $\Gamma'_{i+1} = \Gamma_{i+1} - (\Gamma_1 \cup \cdots \cup \Gamma_i)$ for $i = 1, \cdots, d-1$. By putting $\pi''_{A,M}(1, \theta_{\pi_B(r)}\omega) = e_i$ for $r \in N(e, \infty)$, $i = 1, \cdots, d$ and $\omega \in \Gamma'_i$, we have $\pi''_{A,M} \in \mathcal{MS}(F; 1)$ and then the supremum of (4.27) is attained by $\pi''_{A,M}$:

$$\begin{aligned}&Z_F^{*\pi_B\tau_B}(\pi_B(r)) \\ &= \max\{E^{X_{\pi_B(r)}}[\langle f_A(X_{\pi''_{A,M}(1)}), \pi''_{A,M}(1)\rangle + \beta^{\pi''_{A,M}(1)}\ Z_S^{*\pi'_B\tau'_B}(\pi''_{A,M}(1)) : 1 \le \tau'_B], h(X_{\pi_B(r)})\}\end{aligned}$$

for each $r \in N(e, \infty)$. Hence from Lemmas 4.1 and 4.2, we may inductively define a Markov tactic $(\pi_{A,M}, \tau_{A,M})$ by

$$\pi_{A,M}(r+1, \omega) = \pi_B(r, \omega) + \pi''_{A,M}(1, \theta_{\pi_B(r)}\omega) \text{ for } \omega \in \Omega \text{ and each } r \in N(e, \infty), \quad (4.28)$$

and

$$\tau_{A,M} = \inf\{r \in N(e, \infty) : \Lambda_r\}, \quad (4.29)$$

where for $r \in N(e, \infty)$ we define $\Lambda_r =$
$\{E^{X_{\pi_B(r)}}[\langle f_A(X_{\pi''_{A,M}(1)}), \pi''_{A,M}(1)\rangle + \beta^{\pi''_{A,M}(1)}\ Z_S^{*\pi'_B\tau'_B}(\pi''_{A,M}(1)) : 1 \le \tau'_B] \le h(X_{\pi_B(r)})\}$.
Then we obtain

$$\begin{aligned}&\max\{E^{X_{\pi_B(r)}}[\langle f_A(X_{\pi''_{A,M}(1)}), \pi''_{A,M}(1)\rangle + \beta^{\pi''_{A,M}(1)}\ Z_S^{*\pi'_B\tau'_B}(\pi''_{A,M}(1)) : 1 \le \tau'_B], h(X_{\pi_B(r)})\} \\ &= E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_{A,M}(r+1)}), \pi_{A,M}(r+1) - \pi_B(r)\rangle \\ &\quad + \beta^{\pi_{A,M}\pi_B}(r+1)\ Z_S^{*\pi_B\tau_B}(\pi_{A,M}(r+1)) : r+1 \le \tau_{A,M} \wedge \tau_B] + h(X_{\pi_B(r)}) \cdot I_{\{r=\tau_{A,M}\wedge\tau_B\}}.\end{aligned}$$

Therefore we conclude that for all $r \in N(e, \infty)$

$$\begin{aligned}&Z_F^{*\pi_B\tau_B}(\pi_B(r)) \\ &= E^{\mathcal{F}_{\pi_B(r)}}[\langle f_A(X_{\pi_{A,M}(r+1)}), \pi_{A,M}(r+1) - \pi_B(r)\rangle \\ &\quad + \beta^{\pi_{A,M}\pi_B}(r+1)\ Z_S^{*\pi_B\tau_B}(\pi_{A,M}(r+1)) : r+1 \le \tau_{A,M} \wedge \tau_B] + h(X_{\pi_B(r)}) \cdot I_{\{r=\tau_{A,M}\wedge\tau_B\}}.\end{aligned}$$

On the other hand from Lemmas 4.5 and 4.3(ii), we have that for all $r \in N(e, \infty)$

$$\begin{aligned}&Z_S^{*\pi_B\tau_B}(\pi_{A,M}(r+1)) \\ &= E^{\mathcal{F}_{\pi_{A,M}(r+1)}}[\langle -f_B(X_{\pi_B(r+2)}), \pi_B(r+2) - \pi_{A,M}(r+1)\rangle \\ &\quad + \beta^{\pi_{A,M}\pi_B}(r+2)\ Z_F^{*\pi_B\tau_B}(\pi_B(r+2)) : r+2 \le \tau_{A,M} \wedge \tau_B] + h(X_{\pi_{A,M}(r+1)}) \cdot I_{\{r+1=\tau_{A,M}\wedge\tau_B\}}.\end{aligned}$$

Hence from these two equations, we conclude the results that $V_F^{\pi_{A,M}\tau_{A,M}\pi_B\tau_B} = V_F^{*\pi_B\tau_B}$ and $(\pi_{A,M}, \tau_{A,M}; \pi_B, \tau_B) \in \mathcal{F}(F)$. We can also check the other equations similarly. □

4.4. The optimal values and the optimal tactics

Now from Proposition 4.1, the families

$$\mathcal{T}(F; lower) := \{(\pi_A, \tau_A) \mid V_F^{\pi_A\tau_A *} = V_F^{\pi_A\tau_A\pi_B\tau_B} \text{ for some } (\pi_B, \tau_B)\} \text{ and}$$

$$\mathcal{T}(F; upper) := \{(\pi_B, \tau_B) \mid V_F^{*\pi_B\tau_B} = V_F^{\pi_A\tau_A\pi_B\tau_B} \text{ for some } (\pi_A, \tau_A)\}$$

are non-empty, therefore we may respectively define the lower bound $\underline{V}_F$ and the upper bound $\overline{V}_F$ of values in the first-type bandit games by

$$\underline{V}_F = \sup_{(\pi_A,\tau_A)\in\mathcal{T}(F;lower)} V_F^{\pi_A\tau_A *} \quad \text{and} \quad \overline{V}_F = \inf_{(\pi_B,\tau_B)\in\mathcal{T}(F;upper)} V_F^{*\pi_B\tau_B}.$$

In the second-type we similarly define $\mathcal{T}(S; lower)$, $\mathcal{T}(S; upper)$, $\underline{V}_S$ and $\overline{V}_S$. In this section we investigate the following backward iteration in order to find the optimal values in both type bandit games. Further we shall show that the lower bounds and the upper bounds of values coincide and that the iteration converges to the unique optimal value.

Let us consider the following value iteration. For $r \in N$ we define successively as follows.

Iteration 4.2.

(0) *Put* $U_{F,0} = U_{S,0} = h$.

$(F.r)$ *For* $x = (x^1, \cdots, x^d) \in E$, *put*

$$U_{F,r+1}(x) = \max\{\max_{i=1,\cdots,d} E^{x^i}[f_A^i(X_1^i) + \beta^i(X_1^i)\, U_{S,r}(x^1, \cdots, X_1^i, \cdots, x^d)], h(x)\}.$$

$(S.r)$ *For* $x = (x^1, \cdots, x^d) \in E$, *put*

$$U_{S,r+1}(x) = \min\{\min_{i=1,\cdots,d} E^{x^i}[-f_B^i(X_1^i) + \beta^i(X_1^i)\, U_{F,r}(x^1, \cdots, X_1^i, \cdots, x^d)], h(x)\}.$$

First we shall prove convergence of sequences $\{U_{F,r}\}_{r\in N}$ and $\{U_{S,r}\}_{r\in N}$ in Iteration 2. Let $\|\cdot\|$ denote the supremum norm on the space of bounded measurable functions on E. For $i = 1, \cdots, d$ we shall introduce the following semi-linear operators S_A^i and S_B^i on the space of all bounded measurable functions on E:

$$S_A^i\phi(x) = E^{x^i}[f_A^i(X_1^i) + \beta^i(X_1^i)\, \phi(x^1, \cdots, X_1^i, \cdots, x^d)] \quad (x = (x^1, \cdots, x^d) \in E), \text{ and}$$

$$S_B^i\phi(x) = E^{x^i}[-f_B^i(X_1^i) + \beta^i(X_1^i)\ \phi(x^1, \cdots, X_1^i, \cdots, x^d)] \quad (x = (x^1, \cdots, x^d) \in E)$$

for bounded measurable functions ϕ on E. Then we have the following lemmas.

Lemma 4.6. *Let u_1, u_2, v_1 and v_2 be bounded measurable functions on E such that*

$$v_j = \max\{\max_{i=1,\cdots,d} S_A^i u_j, h\} \text{ for } j = 1, 2,$$

where $\max\{\phi, \psi\}$ denotes $\max\{\phi, \psi\}(x) = \max\{\phi(x), \psi(x)\}$ for functions ϕ and ψ on E. Then it holds that

$$\|v_1 - v_2\| \leq \beta_0\ \|u_1 - u_2\|.$$

Proof. We can easily check this lemma. □

Lemma 4.7. *For each $r, r' \in N$, the following (i) and (ii) hold:*

(i) $\|U_{F,r+r'+1} - U_{F,r+1}\| \leq \beta_0 \|U_{S,r+r'} - U_{S,r}\|$.

(ii) $\|U_{S,r+r'+1} - U_{S,r+1}\| \leq \beta_0 \|U_{F,r+r'} - U_{F,r}\|$.

Proof. It is trivial from Lemma 4.6 and Iteration 4.2. □

Then we obtain the following results regarding Iteration 2.

Theorem 4.1. *Iteration 2 converges:*

$$U_F(x) = \lim_{r\to\infty} U_{F,r}(x) \quad \text{and } U_S(x) = \lim_{r\to\infty} U_{S,r}(x) \quad \text{for } x \in E.$$

Further the pair of U_F and U_S is a unique solution of the following equations (4.30) and (4.31):

$$U_F = \max\{\max_{i=1,\cdots,d} S_A^i U_S, h\}; \tag{4.30}$$

$$U_S = \min\{\min_{i=1,\cdots,d} S_B^i U_F, h\}. \tag{4.31}$$

Proof. From Lemma 4.7, we have for each $r, r' \in N$

$$\|U_{F,r+r'+2} - U_{F,r+2}\| \leq \beta_0 \|U_{S,r+r'+1} - U_{S,r+1}\| \leq {\beta_0}^2 \|U_{F,r+r'} - U_{F,r}\|.$$

We inductively obtain

$$\|U_{F,r+r'} - U_{F,r}\| \leq {\beta_0}^r \|U_{F,r'} - U_{F,0}\|$$

for all $r' \in N$ and all even r. As letting r and r' infinite, we obtain the existence of $\lim_{r\to\infty} U_{F,r}$. Similarly $\lim_{r\to\infty} U_{S,r}$ exists. We obtain (4.30) and (4.31), by applying the bounded convergence theorem to Iteration 2. Finally the uniqueness of solutions U_F and U_S is easily checked, by using Lemma 4.6. □

For Markov tactics $(\pi_A, \tau_A) \in \mathcal{MT}(F;1)$, $(\pi_B, \tau_B) \in \mathcal{MT}(S;1)$ we shall introduce the following semi-linear operators $S_A^{\pi_A\tau_A}$ ($S_B^{\pi_B\tau_B}$, $Q_A^{\pi_A\tau_A}$, $Q_B^{\pi_B\tau_B}$ resp.) on the space of all bounded measurable functions on E:

$$S_A^{\pi_A\tau_A}\phi(x) = E^x[\langle f_A(X_{\pi_A(1)}), \pi_A(1)\rangle + \beta^{\pi_A}(1)\ \phi(X_{\pi_A(1)}) : \tau_A > 0]\ (x \in E);$$

$$S_B^{\pi_B\tau_B}\phi(x) = E^x[\langle -f_B(X_{\pi_B(1)}), \pi_B(1)\rangle + \beta^{\pi_B}(1)\ \phi(X_{\pi_B(1)}) : \tau_B > 0]\ (x \in E);$$

$$Q_A^{\pi_A\tau_A}\phi(x) = S_A^{\pi_A\tau_A}\phi(x) + E^x[h(X_0) : \tau_A = 0]\ (x \in E);$$

$$Q_B^{\pi_B\tau_B}\phi(x) = S_B^{\pi_B\tau_B}\phi(x) + E^x[h(X_0) : \tau_B = 0]\ (x \in E)$$

for bounded measurable functions ϕ on E. Then we obtain the following results.

Corollary 4.1.

(i) $U_F = \max\{\sup_{(\pi_A,\tau_A)\in\mathcal{MT}(F;1)} S_A^{(\pi_A,\tau_A)}U_S, h\} = \sup_{(\pi_A,\tau_A)\in\mathcal{MT}(F;1)} Q_A^{(\pi_A,\tau_A)}U_S.$

(ii) $U_S = \min\{\inf_{(\pi_B,\tau_B)\in\mathcal{MT}(S;1)} S_B^{(\pi_B,\tau_B)}U_F, h\} = \inf_{(\pi_B,\tau_B)\in\mathcal{MT}(S;1)} Q_B^{(\pi_B,\tau_B)}U_F.$

Proof. They are trivial from (4.30) and (4.31), by considering the definition of one-step Markov strategies and one-step Markov tactics. □

4.5. Construction of the optimal tactics and the uniqueness of the optimal values

Now we shall construct the optimal tactics. First we define subsets of E as follows.

$$D_A^0 = \{U_F = h\};\quad D'^j_A = \{\max_{i=1,\cdots,d} S_A^i U_S = S_A^j U_S\} \text{ for } j = 1,\cdots,d;$$

$$D_B^0 = \{U_S = h\};\quad D'^j_B = \{\min_{i=1,\cdots,d} S_B^i U_F = S_B^j U_F\} \text{ for } j = 1,\cdots,d.$$

Further we let $\{D_A^i \mid i = 1,\cdots,d\}$ and $\{D_B^i \mid i = 1,\cdots,d\}$ be partitions of E by

$$D_A^1 = D'^1_A;\quad D_A^{i+1} = D'^{i+1}_A - (D'^1_A \cup \cdots \cup D'^i_A) \text{ for } i = 1,\cdots,d-1;$$

$$D_B^1 = D'^1_B;\quad D_B^{i+1} = D'^{i+1}_B - (D'^1_B \cup \cdots \cup D'^i_B) \text{ for } i = 1,\cdots,d-1.$$

Hence D_A^i (D_B^i) mean player A's (player B's) selection regions for arms i and D_A^0 (D_B^0) mean his stopping regions. Then by putting

$$\pi_A^\circ = e_i \ (\pi_B^\circ = e_i) \text{ on } \{X_\mathbf{0} \in D_A^i \ (D_B^i)\} \quad \text{for } i = 1, \cdots, d; \tag{4.32}$$

we have Markov strategies $\pi_A^\circ \in \mathcal{MS}(F;1)$ ($\pi_B^\circ \in \mathcal{MS}(S;1)$ resp.). Further by setting

$$\tau_A^\circ = \begin{cases} 2 & \text{on } \{X_\mathbf{0} \notin D_A^0\} \\ 0 & \text{on } \{X_\mathbf{0} \in D_A^0\} \end{cases} \quad \text{and} \quad \tau_B^\circ = \begin{cases} 2 & \text{on } \{X_\mathbf{0} \notin D_B^0\} \\ 0 & \text{on } \{X_\mathbf{0} \in D_B^0\}, \end{cases} \tag{4.33}$$

we have Markov tactics $(\pi_A^\circ, \tau_A^\circ) \in \mathcal{MS}(F;1)$ ($(\pi_B^\circ, \tau_B^\circ) \in \mathcal{MS}(S;1)$ resp.). From Lemma 4.1 we may give another representation of Markov strategies. For positive even (odd) number r and $(\pi_A; \pi_B) \in \mathcal{MS}(F;r)$ we describe

$$(\pi_A; \pi_B) = [\pi_{A,1}; \pi_{B,2}; \pi_{A,3}; \pi_{B,4}; \cdots; \pi_{B,r}]. \tag{4.34}$$

where $\{\pi_{A,t} \mid t \in N(o, r+1)\}$ ($\{\pi_{B,t} \mid t \in N(e, r+1)\}$) are player A's (player B's resp.) one step Markov strategies. Hence the meaning of (4.34) is as follows. Player A selects an arm, by using Markov strategy $\pi_{A,1}$. Next player B selects, by using Markov strategy $\pi_{B,2}$. Further player A does, by using Markov strategy $\pi_{A,3}$. The game continues in this way, and finally player B (player A) selects, by using Markov strategy $\pi_{B,r}$ ($\pi_{A,r}$). Moreover we have similar representations concerning second-type Markov strategies: For positive even (odd) r and $(\pi_A; \pi_B) \in \mathcal{MS}(S;r)$ we write

$$(\pi_A; \pi_B) = [\pi_{B,1}; \pi_{A,2}; \pi_{B,3}; \pi_{A,4}; \cdots; \pi_{A,r}].$$

Hence by using these representations, we give the following Markov strategies $(\pi_A^*; \pi_B^*) \in \mathcal{MS}(F)$ and $(\pi'^*_A; \pi'^*_B) \in \mathcal{MS}(S)$ by

$$(\pi_A^*; \pi_B^*) = [\pi_A^\circ; \pi_B^\circ; \pi_A^\circ; \pi_B^\circ; \cdots] \text{ and } (\pi'^*_A; \pi'^*_B) = [\pi_B^\circ; \pi_A^\circ; \pi_B^\circ; \pi_A^\circ; \cdots].$$

Further we define Markov stopping times τ_A^* and τ_B^* (τ'^*_A and τ'^*_B) in the same line as Lemma 4.1(iii), (iv), by using Markov strategies $(\pi_A^*; \pi_B^*)$ ($(\pi'^*_A; \pi'^*_B)$) and 2-steps Markov stopping times τ_A° and τ_B° respectively. Then from Lemma 4.2, we obtain

$$\tau_A^* = \inf\{t \in N(e, \infty) \mid X_{\pi_B^*(t)} \in D_A^0\}; \quad \tau_B^* = \inf\{t \in N(o, \infty) \mid X_{\pi_A^*(t)} \in D_B^0\};$$

$$\tau'^*_A = \inf\{t \in N(o, \infty) \mid X_{\pi'^*_B(t)} \in D_A^0\}; \quad \tau'^*_B = \inf\{t \in N(e, \infty) \mid X_{\pi'^*_A(t)} \in D_B^0\}.$$

Then we obtain the following results.

Theorem 4.2. *$(\pi_A^*, \tau_A^*; \pi_B^*, \tau_B^*) \in \mathcal{T}(F)$ ($(\pi'^*_A, \tau'^*_A; \pi'^*_B, \tau'^*_B) \in \mathcal{T}(S)$) is an optimal tactic and U_F (U_S) is an optimal value for the first-type (second-type resp.) bandit game:*

(i) $V_F^{\pi_A\tau_A\pi_B^*\tau_B^*} \le U_F = V_F^{\pi_A^*\tau_A^*\pi_B^*\tau_B^*} \le V_F^{\pi_A^*\tau_A^*\pi_B\tau_B}$
for every $(\pi_A, \tau_A) \in \mathcal{D}(F; \pi_B^*, \tau_B^*)$ and $(\pi_B, \tau_B) \in \mathcal{D}(F; \pi_A^*, \tau_A^*)$.

(ii) $V_S^{\pi'_A\tau'_A\pi'^*_B\tau'^*_B} \le U_S = V_S^{\pi'^*_A\tau'^*_A\pi'^*_B\tau'^*_B} \le V_S^{\pi'^*_A\tau'^*_A\pi'_B\tau'_B}$
for every $(\pi'_A, \tau'_A) \in \mathcal{D}(S; \pi'^*_B, \tau'^*_B)$ and $(\pi'_B, \tau'_B) \in \mathcal{D}(S; \pi'^*_A, \tau'^*_A)$.

Proof. First we shall show that the inequality of (i) holds for Markov tactics. From Corollary 4.1(i) we have

$$U_F = Q_A^{(\pi_A^\circ, \tau_A^\circ)} U_S \ge Q_A^{(\pi_A, \tau_A)} U_S \tag{4.35}$$

for every Markov tactic $(\pi_A, \tau_A) \in \mathcal{MT}(F; 1)$. Hence from Corollary 4.1(ii) we have $U_S = Q_B^{(\pi_B^\circ, \tau_B^\circ)} U_F$. Together with (4.35) we obtain $U_F = Q_A^{(\pi_A^\circ, \tau_A^\circ)} Q_B^{(\pi_B^\circ, \tau_B^\circ)} U_F \ge Q_A^{(\pi_A, \tau_A)} Q_B^{(\pi_B^\circ, \tau_B^\circ)} U_F$ for every Markov $(\pi_A, \tau_A) \in \mathcal{MT}(F; 1)$. Therefore we inductively obtain

$$\begin{aligned} U_F &= Q_A^{(\pi_A^\circ, \tau_A^\circ)} Q_B^{(\pi_B^\circ, \tau_B^\circ)} Q_A^{(\pi_A^\circ, \tau_A^\circ)} Q_B^{(\pi_B^\circ, \tau_B^\circ)} \cdots Q_A^{(\pi_A^\circ, \tau_A^\circ)} Q_B^{(\pi_B^\circ, \tau_B^\circ)} U_F \\ &\ge Q_A^{(\pi_{A,1}, \tau_{A,1})} Q_B^{(\pi_B^\circ, \tau_B^\circ)} Q_{A,3}^{(\pi_A, \tau_{A,3})} Q_B^{(\pi_B^\circ, \tau_B^\circ)} \cdots Q_A^{(\pi_{A,2r-1}, \tau_{A,2r-1})} Q_B^{(\pi_B^\circ, \tau_B^\circ)} U_F \end{aligned} \tag{4.36}$$

for every $r \in N$ and every Markov $(\pi_{A,t}, \tau_{A,t}) \in \mathcal{MT}(F; 1)$ $(t \in N(o, 2r))$. Hence from the definitions of Q_A and Q_B we have

$$\|Q_A^{(\pi_A, \tau_A)}\phi_1 - Q_A^{(\pi_A, \tau_A)}\phi_2\| \le \beta_0\|\phi_1 - \phi_2\| \text{ and } \|Q_B^{(\pi_B, \tau_B)}\phi_1 - Q_B^{(\pi_B, \tau_B)}\phi_2\| \le \beta_0\|\phi_1 - \phi_2\|$$

for $(\pi_A, \tau_A) \in \mathcal{MT}(F; 1)$, $(\pi_B, \tau_B) \in \mathcal{MT}(S; 1)$ and bounded measurable functions ϕ_1, ϕ_2 on E. By letting r infinite in (4.36), from the definitions π_A^*, π_B^* and Lemmas 4.1 and 4.2 we obtain

$$U_F = V_F^{\pi_A^*\tau_A^*\pi_B^*\tau_B^*} \ge V_F^{\pi_A\tau_A\pi_B^*\tau_B^*}, \tag{4.37}$$

where $(\pi_A^*; \pi_B^*) = [\pi_A^\circ; \pi_B^\circ; \pi_A^\circ; \pi_B^\circ; \pi_A^\circ; \pi_B^\circ; \cdots] \in \mathcal{MS}(F)$, τ_A^* and τ_B^* are given by (3.8), $(\pi_A; \pi_B^*) = [\pi_{A,1}; \pi_B^\circ; \pi_{A,3}; \pi_B^\circ; \pi_{A,5}; \pi_B^\circ; \cdots] \in \mathcal{MS}(F)$ and Markov stopping times τ_A are defined by in the same line as Lemma 4.1(iii), by using Markov tactics $(\pi_A; \pi_B^*)$ and non-increasing sequences $\{\tau_{A,t}\}_{t \in N(o,\infty)}$ of 2-steps Markov stopping times. Therefore (4.37) hold for every Markov tactic $(\pi_A, \tau_A) \in \mathcal{D}(F; \pi_B^*, \tau_B^*)$ defined in the type of Lemma 4.1(iii). Since the other Markov cases can be proved similarly, we obtain the inequalities (i) for every Markov tactics $(\pi_A, \tau_A) \in \mathcal{D}(F; \pi_B^*, \tau_B^*)$ and $(\pi_B, \tau_B) \in \mathcal{D}(S; \pi_A^*, \tau_A^*)$ which are defined in the type of Lemma 4.1(iii). Next we shall show the non-Markov case. Hence by the use of Proposition 4.1, there exists a Markov tactic $(\pi_{A,M}^*, \tau_{A,M}^*) \in \mathcal{D}(F; \pi_B^*, \tau_B^*)$ which satisfies that $V_F^{\pi_{A,M}^*\tau_{A,M}^*\pi_B^*\tau_B^*} = V_F^{*\pi_B^*\tau_B^*}$. Then we have

$$V_F^{\pi_{A,M}^*\tau_{A,M}^*\pi_B^*\tau_B^*} = V_F^{*\pi_B^*\tau_B^*} \ge V_F^{\pi_A^*\tau_A^*\pi_B^*\tau_B^*}.$$

On the other hand from the definitions (4.20) and (4.29) and Lemma 4.2, $(\pi^*_{A,M}, \tau^*_{A,M}) \in \mathcal{D}(F; \pi^*_B, \tau^*_B)$ is a Markov tactic which are defined in the type of Lemma 4.1(i), (iii). Therefore from the first part of this proof we obtain

$$U_F = V_F^{\pi^*_A \tau^*_A \pi^*_B \tau^*_B} \geq V_F^{\pi^*_{A,M} \tau^*_{A,M} \pi^*_B \tau^*_B}.$$

Thus we conclude

$$U_F = V_F^{\pi^*_A \tau^*_A \pi^*_B \tau^*_B} = V_F^{\pi^*_{A,M} \tau^*_{A,M} \pi^*_B \tau^*_B} = V_F^{*\pi^*_B \tau^*_B}.$$

Since the other inequalities can be proved similarly, the proof of this theorem is completed. □

Finally we obtain the following results concerning the optimal values.

Corollary 4.2. *The bandit games have the unique optimal value:*

$$U_F = \underline{V}_F = \overline{V}_F \quad \text{and} \quad U_S = \underline{V}_S = \overline{V}_S.$$

Proof. From Theorem 4.2 we have

$$\overline{V}_F \leq \sup_{(\pi_A \tau_A) \in \mathcal{D}(F; \pi^*_B, \tau^*_B)} V_F^{\pi_A \tau_A \pi^*_B \tau^*_B} = U_F = \inf_{(\pi_B \tau_B) \in \mathcal{D}(F; \pi^*_A, \tau^*_A)} V_F^{\pi^*_A \tau^*_A \pi_B \tau_B} \leq \underline{V}_F.$$

Since $\underline{V}_F \leq \overline{V}_F$ is trivial, we obtain $U_F = \underline{V}_F = \overline{V}_F$. The other is similar. □

Acknowledgement. The author is sincerely grateful to Professor N. Furukawa, Kyusyu University, for his helpful advises and courageous suggestions. The author also would like to thank Professor S. Iwamoto, Kyusyu University, for his courageous advises and Professor S. Arikawa, Kyusyu University, for providing the author a comfortable situation during the preparation of this thesis. Further the author wishes to Professor M. Yasuda, M. Kurano and J. Nakagami, Chiba University, and Y. Ohtsubo, Kyusyu Institute of Technology, for their encouragements.

5. References

5.1. References for Chapter 2

[**Bel1**] Bellman,R. *A problem in the sequential design of experiments*, *Sankhya*, *A***16**, 1956, pp. 221-229.

[**Bel2**] Bellman,R.E., *Dynamic Programming*, Princeton University Press, 1957.

[**BerFri1**] Berry,D.A. and Fristedt,B., *Bandit Problems*, (*Sequential Allocation of Experiments*), Chapman and Hall, London, 1985.

[**BJK1**] Brad,R.N., Johnson,S.M. and Karlin,S., *On sequential designs for maximizing the sum of n observations*, *Annals Math. Stat.*, **27**, 1956, pp. 1060-1074.

[**DTW1**] Dalang,R.C., Trotter Jr.,L.E. and de Werra,D., *On Randomized Stopping Points and Perfect Graphs*, *Probab. Th. Rel. Fields*, **78**, 1988, pp. 357-378.

[**Der1**] Derman,C., *Finite state Markovian decision processes*, Academic New York, 1970.

[**Eic1**] Eick,S.G., *Gittins Procedures for Bandits with Delayed Responses*, *J. Royal Stat. Soc.*, *B***50**, 1988, pp. 125-132.

[**Fel1**] Feldman,D. *Contributions to the 'two-armed bandit' problem*, *Annals Math. Stat.*, **33**, 1962, pp. 847-856.

[**Git1**] Gittins,J.C., *Bandit processes and dynamic allocation indices*, *J. Royal Stat. Soc.*, *B***41**, 1979, pp. 148-177.

[**Git2**] Gittins,J.C., *Multi-Armed Bandit Allocation indices*, John Wiley and Sons Ltd., England, 1989.

[**GitJon1**] Gittins,J.C. and Jones,D.M., *A dynamic allocation indices for the sequential design of experiments*, *In J.Gani, K.Sarkadi and I.Vince (Eds.), Progress in Statistics. European Meeting of Statisticians*, *J. Royal Stat. Soc.*, *B***41**, 1972, North Holland, Amsterdam, pp. 241-266.

[**Gla1**] Glazebrook,K.D., *Optimal Strategies for Families of Alternative Bandit*, *IEEE Trans. Auto. Control*, *AC*-**28**, 1983, pp. 858-860.

[**Gla2**] Glazebrook,K.D., *Stoppable families of alternative bandit processes*, *J. Appl. Probab.*, **16**, 1979, pp. 843-854.

[**Hag1**] Haggstrom.G.W., *Optimal Stopping and Experimental Design*, *Annals Math. Stat.*, **37**, 1966, pp. 7-29.

[**Isb1**] Isbell,J.R., *On a problem of Robbins*, *Annals Math. Stat.*, **30**, 1959, pp. 606-610.

[**KatVei1**] Katehakis,M.N. and Veinott Jr.,A.F., *Multi-armed bandit problem: Decomposition and computation*, *Math. Oper. Res.*, **12** 1987, pp. 262-268.

[**Kal1**] Kallenberg,L.C.M., *A note on M. N. Katehakis' and Y.-R. Chen's computation of the Gittins Index*, *Math. Oper. Res.*, **11** 1986, pp. 184-186.

[**KreSuc1**] Krengel,U. and Sucheston,L., *Stopping Rules and Tactics for Processes indexed by a Directed Set*, *J. Multi. Analysis*, **11**, 1981, pp. 199-229.

[**Kur2**] Kurtz,T.G., *The Optional Sampling Theorem for Martingales indexed by Directed Sets*, *Annals Probab.*, **8**, 1980, pp. 675-681.

[**LawVan1**] Lawler,G.F. and Vanderbei,R.J., *Markov Strategies for Optimal Control Problems indexed by a Partially Ordered Set*, *Annals Probab.*, **11**, 1983, pp. 642-647.

[**Man1**] Mandelbaum,A., *Discrete multi-armed bandits and multi-parameter processes*, *J. Probab. Th. Rel. Fields*, **71**, 1986, pp. 129-147.

[**ManVan1**] Mandelbaum,A. and Vanderbei,R.J., *Optimal Stopping and Supermartingales over Partially Ordered Sets*, *Z. Wahr. verw. Geb.*, **57**, 1981, pp. 253-264.

[**Nev1**] Neveu,J., *Discrete-Parameter Martingales*, North Holland, Amsterdam, 1975.

[**PreSon1**] Presman,E.L. and Sonin,I.M., *Sequential Control with Partial Information*, Nauka, Moscow, 1982.

[**Rob1**] Robbins,H., *Some aspects of the sequential design of experiments*, *Bull. Amer. Math. Soc.*, **58**, 1986, pp. 527-535.

[**Shi1**] Shiryayev,A.N., *Optimal Stopping Rules*, Springer, New York, 1979.

[**Tan1**] Tanaka,T., *Two-parameter Optimal Stopping Problem with Switching Costs*, *Stoch. Proc. Appl.*, **36**, 1990, pp. 153-163.

[**VWB1**] Varaiya,P., Walrand,J. and Buyukkoc,C., *Extensions of the multiarmed bandit problem : The discounted case*, *IEEE Trans. Auto. Control*, *AC*-**30**, 1985, pp. 426-439.

[**Vog1**] Vogel,W., *A sequential design for the two-armed bandit*, *Annals Math. Stat.*, **31** 1960, pp. 430-443.

[WasWil1] Washburn Jr.,R.B. and Willsky,S., *Optional Sampling of Submartingales indexed by Partially Ordered Sets*, *Annals Probab.*, **9**, 1981, pp. 957-970.

[Whi1] Whittle,P., *Multi-armed bandits and the Gittins index*, *J. Royal Stat. Soc.*, *B***41**, 1980, pp. 148-164.

[Whi2] Whittle,P., *Arm-acquiring bandits*, *J. Annals Probab.*, **9**, 1981, pp. 284-292.

[Whi3] Whittle,P., *Optimization over Time, Vol. 1*, Whiley, New York, 1982.

[Yos3] Yoshida,Y., *On an Optimal Stopping Problem for Discrete-Time Multi-Parameter Diffusion Processes*, *J. Infor. Optoi. Sci.*, **11** 1990, pp. 473-492.

5.2. References for Chapter 3

[Bak1] Bakry,D., *Theoremes de Section et de Projection pour les Processus a Deux Indices*, *Z. Wahr. verw. Geb.*, **55**, 1981, pp. 55-71.

[BaxCha1] Baxter,J.R. and Chacon,R.V., *Compactness of stopping times*, *Z. Wahr. verw. Geb.*, **40**, 1977, pp. 169-181.

[Bis1] Bismut,J.-M., *Sur un Probleme de Dynkin*, *Z. Wahr. Verw. Geb.*, **39** 1977, pp. 31-53.

[BluGet1] Blumenthal,R.M. and Getoor,R.K., *Markov processes and potential theory*, Academic Press, New York, 1968.

[CaiWal1] Cairoli,R. and Walsh,J.B., *Stochastic Integrals in the Plane*, *Acta Math.*, **134**, 1975, pp. 111-183.

[Cai1] Cairoli,R., *Enveloppe de Snell d'un processus a parametre bidimensionnel*, *Annals Inst. Henri Poincare*, **18**, 1988, pp. 47-53.

[ChaLai1] Chang,F. and Lai,T.L., *Optimal stopping and dynamic allocation*, *Adv. Appl. Probab.*, **19**, 1987, pp. 829-853.

[Dal1] Dalang,R.C., *Randomization in the two-armed bandit problem*, *Annals Probab.*, **18**, 1990, pp. 218-225.

[Dal2] Dalang,R.C., *On Stopping Points in the Plane that Lie on a Unique Optional Increasing Path*, *Stochastics*, **24**, 1988, pp. 245-268.

[DolMey1] Doleans-Dade,C. and Meyer,P.A., *Un Petit Theoreme de Projection pour Processus a Deux Indices*, *Lecture Notes Math.*, **721**, 1977/78, Springer, pp. 204-215.

[DozPur1] Dozzi,M. and Puri,M.L., *Strong Solutions of Stochastic Differential Equations for Multiparameter Processes*, *Stochastics*, **17**, 1986, pp. 19-41.

[Dyn1] Dynkin,E.B., *Markov processes*, Springer, Berlin, 1965.

[Dyn3] Dynkin,E.B., *Additive Functionals of Several Time-Reversible Markov Processes*, *J. Func. Analysis*, **42**, 1985, pp. 64-101.

[Dyn4] Dynkin,E.B., *Harmonic Functions Associated with Several Markov Processes*, *Adv. Appl. Math.*, **2**, 1985, pp. 260-283.

[Dyn5] Dynkin,E.B., *Multiple Path Integrals*, *Adv. Appl. Math.*, **7**, 1986, pp. 205-219.

[Dyn6] Dynkin,E.B., *Self-Intersection Gauge for Random Walks and for Brownian Motion*, *Annals Probab.*, **16**, 1988, pp. 1-57.

[Epl1] Eplett,W.J.R., *Continuous-time allocation indices and their discrete-time approximation*, *Adv. Appl. Probab.*, **18**, 1986, pp. 724-746.

[FraSuc1] Frangos,N.E. and Sucheston,L., *On Multiparameter Ergodic and Martingale Theorems in Infinite Measure Spaces*, *Probab. Th. Rel. Fields*, **71**, 1986, pp. 477-490.

[Hur1] Hurzeler,H.E., *On The Optional Sampling Theorem for Processes indexed by A Partially Ordered Set*, *Annals Probab.*, **13**, 1985, pp. 1224-1235.

[Hur2] Hurzeler,H.E., *Stochastic Integration on Partially Ordered Sets*, *J. Multi. Analysis*, **17**, 1985, pp. 279-303.

[Imk1] Imkeller,P., *Ito's Formula for Continuous (N,d)-Processes*, *Z. Wahr. verw. Geb.*, **65**, 1984, pp. 535-562.

[Imk2] Imkeller,P., *A Stochastic Calculus for Continuous N-Parameter Strong Martingales*, *Stoch. Proc. Appl.*, **20**, 1985, pp. 1-40.

[Imk3] Imkeller,P., *A Note on the Localization of Two-Parameter Processes*, *Probab. Th. Rel. Fields*, **73**, 1986, pp. 119-125.

[Imk4] Imkeller,P., *Local Times of Continuous N-Parameter Strong Martingales*, *J. Multi. Analysis*, **19**, 1986, pp. 348-365.

[Imk5] Imkeller,P., *On Changing Time for Two-Parameter Strong Martingales : A Counterexample*, *J. Annals Probab.*, **14**, 1986, pp. 1080-1084.

[Imk6] Imkeller,P., *Stochastic Integrals of Point Processes and the Decomposition of Two-Parameter Martingales*, *J. Multi. Analysis*, **30**, 1989, pp. 98-123.

[Kar1] Karatzas,I., *Gittins indices in the dynamic allocation problem for diffusion processes*, *Annals Probab.*, **12**, 1984, pp. 173-192.

[Kur1] Kurtz,T.G., *Representations of Markov Processes as Multiparameter Time Changes*, *J. Annals Probab.*, **8**, 1980, pp. 682-715.

[Maz1] Mazziotto,G., *Two parameter optimal stopping and bi-Markov processes*, *Z. Wahr. verw. Geb.*, **69**, 1985, pp. 99-135.

[Man2] Mandelbaum,A., *Continuous multi-armed bandits and multi-parameter processes*, *Annals Probab.*, **14**, 1987, pp. 1527-1556.

[Maz2] Mazziotto,G., *Two-Parameter Hunt Processes and a Potential Theory*, *Annals Probab.*, **16**, 1988, pp. 600-619.

[MazMer1] Mazziotto,G. and Merzbach,E., *Regularity and Decomposition of Two-Parameter Supermartingales*, *J. Multi. Analysis*, **17**, 1985, pp. 38-55.

[MazMer2] Mazziotto,G. and Merzbach,E., *Point Processes indexed by Directed Sets*, *Stoch. Proc. Appl.*, **30**, 1988, pp. 105-119.

[MazMil1] Mazziotto,G. and Millet,A., *Stochastic Control of Two-parameter Processes Application: The Two-Armed Bandit Problem*, *Stochastics*, **22**, 1987, pp. 251-288.

[MenRob1] Menaldi,J.L. and Robin,M., *On the optimal reward function of the continuous time multiarmed bandit problem*, *SIAM J. Control and Optimi.*, **28**, 1990, pp. 97-112.

[MazSzp1] Mazziotto,G. and Szpirglas,J., *Arret Optimal sur le Plan*, *Z. Wahr. verw. Geb.*, **62**, 1988, pp. 215-233.

[Mer1] Merzbach,E., *Stopping for two-dimensional stochastic processes*, *Stoch. Proc. Appl.*, **10**, 1980, pp. 49-63.

[MerZak1] Merzbach,E. and Zakai,M., *Stopping a Two Parameter Weak Martingale*, *Probab. Th. Rel. Fields*, **76**, 1987, pp. 499-507.

[Mey1] Meyer,P.A., *Theorie Elementaire des Processus a Deux Indices*, *Lecture Notes Math.*, **863**, 1981, Springer, pp. 1-39.

[Mil1] Millet,A., *On randomized tactics and optimal stopping in the plane*, *Annals Probab.*, **13**, 1985, pp. 946-965.

[MilSuc1] Millet,A. and Sucheston,L., *On Regularity of Multiparameter Amarts and Martingales*, *Z. Wahr. verw. Geb.*, **56**, 1988, pp. 21-45.

[McK1] McKean Jr.,H.P., *Brownian Motion with a Several-Dimensional Time*, *Th. Probab. Appl.*, **8**, 1963, pp. 335-354.

[NuaSan1] Nualart,D. and Sanz,M., *Malliavin Calculus for Two-Parameter Wiener Functionals*, *Z. Wahr. verw. Geb.*, **70**, 1985, pp. 573-590.

[NuaYeh1] Nualart,D. and Yeh,J., *Existence and Uniqueness of a Strong Solution to Stochastic Differential Equations in the Plane with Stochastic Boundary Process*, *J. Mult. Analysis*, **28**, 1989, pp. 149-171.

[Raz1] Razanov,Y.U., *Markov Random Fields and Boundary Problems for Stochastic Partial Differential Equations*, *Theory Probab. Appl.*, **32**, 1986, pp. 1-29.

[San1] Sanz,M., *Local Time for Two-Parameter Continuous Martingales with Respect to the Quadratic Variation*, *Annals Probab.*, **16**, 1988, pp. 778-792.

[StrVar1] Stroock,D.W. and Varadhan,S.R.S, *Multidimensional diffusion processes*, Springer, New York, 1979.

[Van1] Vanderbei,R.J., *Local Time for Two-Parameter Continuous Martingales with Respect to the Quadratic Variation*, *Adv. Appl. Math.*, **4**, 1983, pp. 125-144.

[Wal1] Walsh,J.B., *Optional increasing paths*, *Lecture Notes Math.*, **863**, 1981, Springer, pp. 172-201.

[WonZak1] Wong,E. andM. Zakai, *Martingales and Stochastic Integrals for Processes with a Multi-Dimensional Parameter*, *Z. Wahr. verw. Geb.*, **29**, 1974, pp. 109-122.

[WonZak2] Wong,E. and Zakai,M., *Weak Martingale and Sstochastic Integrals in the Plane*, *Annals Probab.*, **4**, 1976, pp. 570-586.

[WonZak3] Wong,E. and Zakai,M., *Differentiation Formulas for Sstochastic Integrals in the Plane*, *Stoch. Proc. Appl.*, **6**, 1978, pp. 339-349.

[**WonZak4**] Wong,E. and Zakai,M., *Multiparameter Martingale Differential Forms*, *Probab. Th. Rel. Fields*, **74**, 1987, pp. 429-453.

[**Yus1**] Yushkevich,A., *On the Two-Armed Bandit Problem with Continuous Time Parameter and Discounted Rewards*, *Stochastics*, **23**, 1988, pp. 299-310.

[**Yos2**] Yoshida,Y., *An Optimal Stopping Problem in the Presence of Costs of Observations*, *Bull. Infor. Cyb. Res. Ass. Stat. Sci.*, **23** 1989, pp. 155-162.

5.3. References for Chapter 4

[**BenFri1**] Bensoussan,A. and Friedman,A., *Nonlinear Variational Inequalities and Differential Games with Stopping Times*, *Funct. Analysis*, **16** 1974, pp. 305-352.

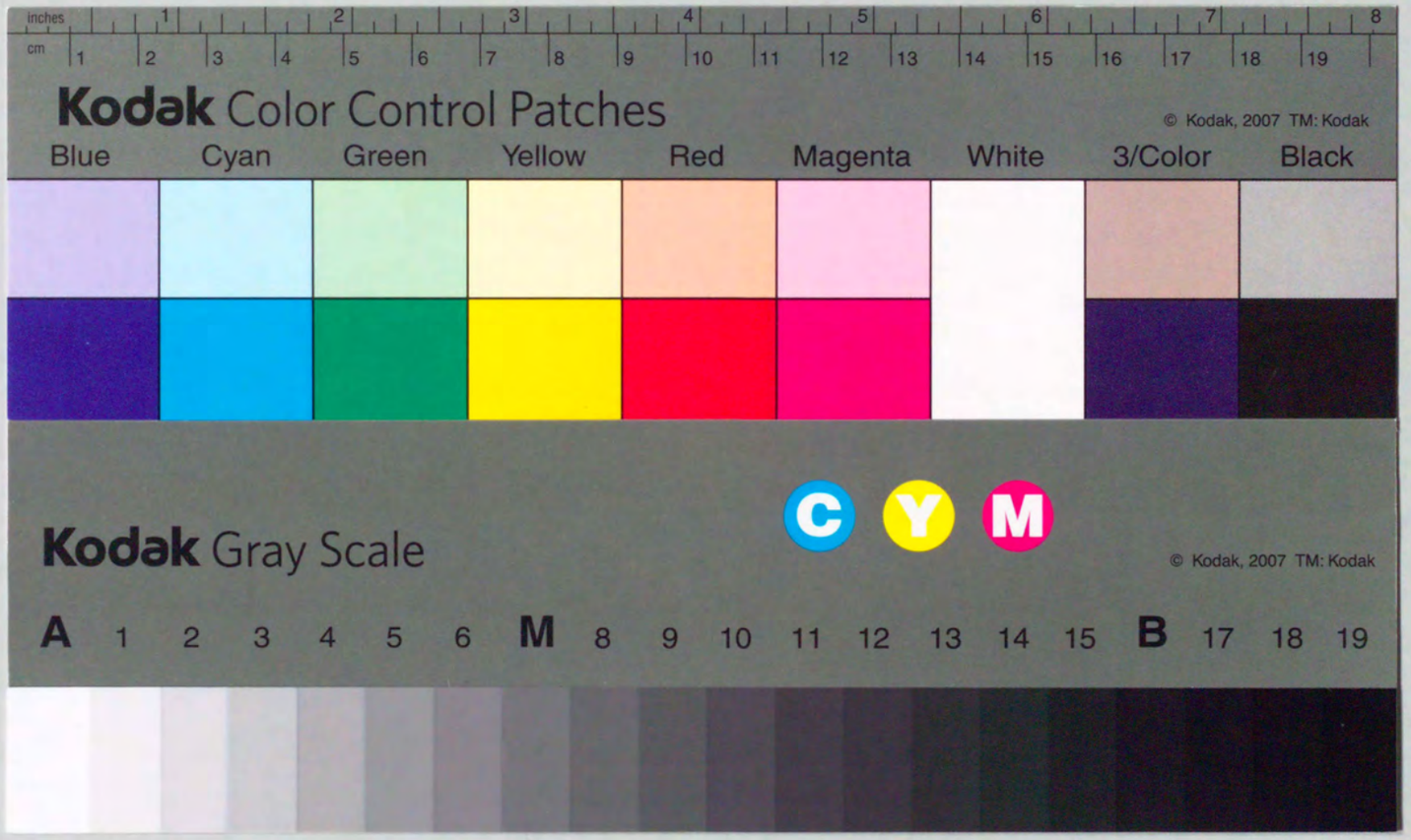
[**Dyn2**] Dynkin,E.B., *Game Variant of a Problem on Optimal Stopping*, *Soviet Math. Dokl.*, **10** 1965, pp. 270-274.

[**Ste1**] Stettner,L., *Zero-sum Games with Stopping Times and Impulsive Strategies*, *Appl. Math. Optimi.*, **9** 1982, pp. 1-24.

[**Yos1**] Yoshida,Y., *Zero-sum Games with Stopping Times and Variational Inequalities*, *Bull. Infor. Cyb. Res. Ass. Stat. Sci.*, **23** 1989, pp. 141-154.

[**Yos4**] Yoshida,Y., *Zero-sum Games for Discrete-Time Multi-Parameter Processes*, *Bull. Infor. Cyb. Res. Ass. Stat. Sci.*, **24** 1991, pp. 165-176.

[**Yos5**] Yoshida,Y., *Zero-sum Games for Discrete-Time Multi-Parameter Processes with a Generalized Discount*, *J. Infor. Optoi. Sci.*, **13** 1992, pp. 231-255.


inches
cm
Kodak Color Control Patches
© Kodak, 2007 TM: Kodak
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
Kodak Gray Scale
C
Y
M
© Kodak, 2007 TM: Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19