ZERO-SUM GAMES FOR DISCRETE-TIME MULTI-PARAMETER PROCESSES

Yoshida, Yuji
Statistics Laboratory, College of Arts and Sciences, Chiba University

https://doi.org/10.5109/13416

出版情報：Bulletin of informatics and cybernetics. 24 (3/4), pp.165-176, 1991-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 24, No. 3~4, 1991

ZERO-SUM GAMES FOR DISCRETE-TIME MULTI-PARAMETER PROCESSES

By

Yuji YOSHIDA*

Abstract

The present paper formulates zero-sum games for discrete-time multi-parameter processes. Under the assumption of independence of reward processes, we give the unique optimal value and the optimal Markov strategies, which are constructively provided by Bellman's equation derived from a value iteration.

1. Introduction

We treat zero-sum games where two players alternately select either one of several reward processes. The theory of multi-parameter stochastic processes has been studied by Mandelbaum [5], Mazziotto [7] and many authors. On the other hand multi-armed bandit problems have been studied by Berry-Fristedt [1], Gittins [2], Whittle [8] and many authors. Especially Mandelbaum [4] has discussed the relation between discrete-time multi-armed bandit problems and discrete-time multi-parameter processes. The purpose of this paper is to formulate the zero-sum games for multi-parameter processes, by using the theory of discrete-time multi-parameter processes in [4].

Now we shall sketch zero-sum games for multi-parameter processes. We regard that a discrete-time d-parameter process consists of d independent reward processes, which evolve according to transition laws of given Markov chains. If player A selects one of reward processes at time t, then he gets a reward at the time and the state of the process moves to a new state at time $(t+1)$ according to transition probabilities of a given Markov chain, and it is player B's turn next to select one reward process. Both players alternately continue in this way and finally settle accounts. The strategies of both players are represented by (2.1) ~ (2.3) in Section 2. Player A's aim is to maximize his gain (2.4) of Section 2 by controlling his strategy π, and player B's is to minimize (2.4) with respect to his strategy σ. Generally, admissible strategies for one player depend on another player's option of strategies.

Concerning the above-mentioned problem, we show existence of the optimal Markov strategies by using the independence of reward processes. Next we give a value iteration method and derive Bellman's equations. Finally the present paper gives the

* Statistics Laboratory, College of Arts and Sciences, Chiba University, 1–33 Yayoicho, Chiba 260, Japan

unique optimal value and the optimal Markov strategies, by constructing concatenations of one-step Markov strategies on the basis of Bellman's equations we derived in this paper.

This paper is structured as follows. In Sections 2.1 and 2.2 we formulate multi-parameter processes and strategies for zero-sum games and show a few fundamental lemmas regarding to their concatenations. In Section 2.3 players' expected rewards and zero-sum games are presented. Section 2.4 provides a proposition to guarantee existence of the optimal Markov strategies. In Section 3.1 we give a backward value iteration and demonstrate its convergence. Section 3.2 is devoted to construct the optimal Markov strategies on the basis of Bellman's equation. Finally in the remainder of this paper we demonstrate uniqueness of the optimal values.

2. Zero-sum Game for Multi-parameter Processes

2.1. Multi-parameter processes

In this section we shall formulate zero-sum games for multi-parameter processes. Let d, the number of arms, be a positive integer. We regard that d-parameter processes consist of d mutually independent reward processes. Let $\boldsymbol{N}$ be the set of non-negative integers and put

$$\boldsymbol{N}(e,r) = \{\text{even } t : 0 \le t < r\} \text{ and } \boldsymbol{N}(o,r) = \{\text{odd } t : 0 \le t < r\} \text{ for } r \in \boldsymbol{N} \cup \{+\infty\}.$$

For each $i = 1, \ldots, d$, let $(\Omega^i, \mathfrak{F}^i, \boldsymbol{P}^i)$ denote a probability space and let $X^i = (X^i_t)_{t \in N}$ denote $\{\mathfrak{F}^i_t\}_{t \in \boldsymbol{N}}$-adapted time-homogeneous Markov chain with a Borel state space E^i. Here $\{\mathfrak{F}^i_t\}_{t \in \boldsymbol{N}}$ is an increasing family of completed sub-σ-fields of $\mathfrak{F}^i$, and $\{X^i\}_{i=1,\ldots,d}$ is assumed to be mutually independent. Let θ^i_τ denote the time-shift operator on Ω^i. Next we shall define a d-parameter process by their products as follows. Set $\boldsymbol{T} = \boldsymbol{N}^d$, $\Omega = \prod_{i=1}^{d} \Omega^i$ and $E = \prod_{i=1}^{d} E^i$. $\boldsymbol{T}$, Ω and E are the time space, the path space and the state space of the d-parameter process, respectively. Hence we introduce the usual partial order into T:

For $r = (r^1, \ldots, r^d)$, $s = (s^1, \ldots, s^d) \in \boldsymbol{T}$, $r \le s$ means that $r^i \le s^i$ for all $i = 1, \ldots, d$. Then a d-parameter process X with the state space E, its σ-fields $\mathfrak{F}_s$ and its time-shift operators θ_s are defined by

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

$$\mathfrak{F}_s = \mathfrak{F}^1_{s^1} \otimes \ldots \otimes \mathfrak{F}^d_{s^d} \qquad \text{for } s = (s^1, \ldots, s^d) \in \boldsymbol{T}, \text{ and}$$

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

Further $\boldsymbol{E}^x$ denotes the expectation operator induced by a probability measure $\boldsymbol{P} = \prod_{i=1}^{d} \boldsymbol{P}^i$ with an initial state $x \in E$.

2. Strategies

Let $\mathbf{0}$ denote the zero vector in $\boldsymbol{T}$ and e_i denote the i'th unit vector in $\boldsymbol{N}^d$. Let $|s| = \sum_{i=1}^{d} s^i$ for $s = (s^1, \ldots, s^d) \in \boldsymbol{T}$. In the pair (π,σ) of A's strategy π and B's strategy σ, when player A moves 'first' and second does player B, we call the pair (π,σ) a first-type strategy. First-type strategies (π,σ) are defined as follows: For $s = (s^1, \ldots, s^d) \in \boldsymbol{T}$,

$$\pi = \{\pi(|s|+t)\}_{t\in N(o,\infty)} = \{(\pi^1(|s|+t), \ldots, \pi^d(|s|+t))\}_{t\in N(o,\infty)} \text{ and}$$

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

are T-valued stochastic sequences on $(\Omega, \mathfrak{F})$ satisfying the following (2.1) $\sim$ (2.3):

(2.1) $\pi(|s|) = \sigma(|s|) = s$.

(2.2) For all $t \in \boldsymbol{N}\ (e,\infty)$ it holds that $\pi(|s|+t+1) = \sigma(|s|+t) + e_i$ for some $i = 1, \ldots, d$,
and for all $t \in \boldsymbol{N}(o,\infty)$ it holds that $\sigma(|s|+t+1) = \pi(|s|+t) + e_i$ for some $i = 1, \ldots, d$.

(2.3) For all $t \in \boldsymbol{N}(o,\infty)$ $(\boldsymbol{N}(e,\infty))$ and all $s' \in \boldsymbol{T}$ it holds that
$\{\pi(|s|+t) = s'\} \in \mathfrak{F}_{s'}$ $(\{\sigma(|s|+t) = s'\} \in \mathfrak{F}_{s'}$ resp.)

We similarly define second-type strategies when player B moves first and "second" does player A. Namely a second-type strategy is defined by exchanging $\boldsymbol{N}(e,\infty)$ with $N(o,\infty)$ in (2.1) $\sim$ (2.3). Thus we put the families of first-type (second-type resp.) strategies and Markov strategies as follows: for $s \in \boldsymbol{T}$

$$S(F;\ s)\ (S(S;\ s)) = \{\textit{all first-type (second-type resp.) strategies } (\pi,\sigma) \textit{ starting at } s\},$$

and we put

$$S(F)\ (S(S)) = S(F;\ \mathbf{0})\ (S(S;\ \mathbf{0}))$$

and

$$MS(F)\ (MS(S)) = \{\textit{all Markov strategies } (\pi,\sigma) \in S(F)\ (S(S))\}.$$

In the Markov strategy, when we are interested in the options during the time interval $[0, r]$, $(r \in \boldsymbol{N})$, we shall call it an r-steps Markov strategy. The families of first-type (second-type resp.) r-steps Markov strategies are denoted by

$$MS(F;\ r)\ (MS(S;\ r)) = \{\textit{all r-steps Markov strategies } (\pi,\sigma) \in S(F)\ (S(S))\}$$

for $r \in \boldsymbol{N}$. Especially since $(\pi,\sigma) \in MS(F;\ 1)$ $(MS(S;\ 1))$ does not depend on σ (π), we shall represent only $\pi \in MS(F;\ 1)$ $(\sigma \in MS(S;\ 1)$ resp.).

Hence when one player's strategy is fixed, the other player's admissible strategies are denoted as follows: We respectively put

$$D(\mathfrak{F};s;\sigma)\ (D(S;s;\sigma)) = \{\pi:\ (\pi,\sigma) \in S(F;s)\ (S(S;s))\} \text{ for } s \in \boldsymbol{T},$$

$$D(\mathfrak{F};s;\pi)\ (D(S;s;\pi)) = \{\sigma:\ (\pi,\sigma) \in S(F;s)\ (S(S;s))\} \text{ for } s \in \boldsymbol{T},$$

$$D(\tilde{\mathcal{S}};\sigma)\ (D(S;\sigma)) = \{\pi\colon (\pi,\sigma) \in S(F)\ (S(S))\},\ \text{and}$$

$$D(\tilde{\mathcal{S}};\pi)\ (D(S;\pi)) = \{\sigma\colon (\pi,\sigma) \in S(F)\ (S(S))\}.$$

Finally $\{\tilde{\mathcal{S}}_{\pi(t)}\}_{t\in N(o,\infty)}$ and $\{\tilde{\mathcal{S}}_{\sigma(t)}\}_{t\in N(e,\infty)}$ denotes informations available at time t:

$\tilde{\mathcal{S}}_{\sigma(t)} = \{\boldsymbol{\Gamma} \in \tilde{\mathcal{S}}\colon \boldsymbol{\Gamma} \cap \{\sigma(t) = s\} \in \tilde{\mathcal{S}}_s \text{ for } s \in \boldsymbol{N}^d\}$ *for* $t \in \boldsymbol{N}(e,\infty)$, *and*
$\tilde{\mathcal{S}}_{\sigma(t)} = \{\boldsymbol{\Gamma} \in \tilde{\mathcal{S}}\colon \boldsymbol{\Gamma} \cap \{\sigma(t) = s\} \in \tilde{\mathcal{S}}_s \text{ for } s \in \boldsymbol{N}^d\}$ *for* $t \in \boldsymbol{N}(e,\infty)$.

Hence we shall prepare the following lemma concerning concatenations of Markov strategies.

LEMMA 1. *The following* (*i*) *and* (*ii*) *hold*:

(*i*) For $r \in \boldsymbol{N}(e,\infty)$ $(\boldsymbol{N}(o,\infty))$, $(\pi,\sigma) \in MS(F;r)$ $(MS(S;r))$ *and* $\pi' \in MS(F;1)$, *we define a concatenated strategy* (π'',σ'') *of* (π,σ) *and* π':

$$\pi''(t,\omega) = \pi(t,\omega) \qquad \text{for } t \in \boldsymbol{N}(o,r+1) \text{ and } \omega \in \Omega,$$

$$\sigma''(t,\omega) = \sigma(t,\omega) \qquad \text{for } t \in \boldsymbol{N}(e,r+1) \text{ and } \omega \in \Omega, \text{ and}$$

$$\pi''(r+1,\omega) = \sigma(r,\omega) + \pi'(1,\theta_{\sigma(r)}\omega) \qquad \text{for } \omega \in \Omega.$$

Then it holds that $(\pi'',\sigma'') \in MS(F;\ r+1)$ $(MS(S;\ r+1)$ *resp.*).

(*ii*) *For* $r \in \boldsymbol{N}(o,\infty)$ $(\boldsymbol{N}(e,\infty))$, $(\pi,\sigma) \in MS(F;r)$ $(MS(S;r))$ *and* $\sigma' \in MS(S;\ 1)$, *we define a concatenated strategy* (π'',σ'') *of* (π,σ) *and* σ':

$$\pi''(t,\omega) = \pi(t,\omega) \qquad \text{for } t \in \boldsymbol{N}(o,r+1) \text{ and } \omega \in \Omega,$$

$$\sigma''(t,\omega) = \sigma(t,\omega) \qquad \text{for } t \in \boldsymbol{N}(e,r+1) \text{ and } \omega \in \Omega, \text{ and}$$

$$\sigma''(r+1,\omega) = \pi(r,\omega) + \sigma'(1,\theta_{\pi(r)}\omega) \qquad \text{for } \omega \in \Omega.$$

Then it holds that $(\pi'',\sigma'') \in MS(F;\ r+1)$ $(MS(S;\ r+1)$ *resp.*).

PROOF. Trivial from the definitions of Markov strategies. □

2.3. Expected rewards and zero-sum games

First we shall define player A's expected values and player B's ones when player A moves 'first'. Let β, a discount rate, be a constant satisfying $0 < \beta < 1$. For $i = 1, \ldots, d$, let f^i (g^i), player A's (player B's resp.) running rewards for i, be a bounded measurable function on E^i. Hence we shall introduce the following notation $< \cdot,\cdot >$, referring to the inner product of d-dimensional real vector spaces: For example, we describe

$$< f(X_{\pi(1)}),\ \pi(1) - \sigma(0) > = \sum_{i=1}^{d} f^i(X^i_{\pi^i(1)})\ (\pi^i(1) - \sigma^i(0)).$$

When a first-type strategy (π,σ) $(\in S(F))$ is taken, player A's expected gain[1] to be paid from player B at an initial state x is

$$V_F[\pi,\sigma](x) = \boldsymbol{E}^x[\Sigma_{t\in N(e,\infty)}\ \beta^t < f(x_{\pi(t+1)}),\ \pi(t+1) - \sigma(t) > \qquad (2.4)$$

$$- \Sigma_{t\in N(o,\infty)}\ \beta^t < g(X_{\sigma(t+1)}),\ \sigma(t+1) - \pi(t) >],$$

When one player's strategy is fixed, the values optimized by another player are as follows:

$$V_F[*,\sigma](x) = \sup_{\pi\in D(F;\sigma)} V_F[\pi,\sigma](x) \text{ for } x \in E, \text{ and} \tag{2.5}$$

$$F_F[\pi,*](x) = \inf_{\sigma\in D(F;\pi)} V_F[\pi,\sigma](x) \text{ for } x \in E. \tag{2.6}$$

Then we shall call the following game when player A moves 'first' first-type zero-sum games: To find strategies $(\pi^*,\sigma^*) \in S(F)$ such that $V_F[\pi^*,\sigma^*] = V_F[*,\sigma^*] = V_F[\pi^*,*]$.

Next we shall similarly define values of games when player A moves second. For a second-type strategy (π,σ) $(\in S(S))$ and x $(\in E)$ we put

$$V_S[\pi,\sigma](x) = \boldsymbol{E}^x[\Sigma_{t\in N(o,\infty)} \beta^t < f(X_{\pi(t+1)}), \pi(t+1) - \sigma(t) > \tag{2.7}$$

$$- \Sigma_{t\in N(e,\infty)} \beta^t < g(X_{\sigma(t+1)}), \sigma(t+1) - \pi(t) >],$$

$$V_S[*,\sigma](x) = \sup_{\pi\in D(S;\sigma)} V_S[\pi,\sigma](x) \text{ for } x \in E, \text{ and} \tag{2.8}$$

$$V_S[\pi,*](x) = \inf_{\sigma\in D(S;\pi)} V_S[\pi,\sigma](x) \text{ for } x \in E. \tag{2.9}$$

Then second-type zero-sum games are as follows: To find strategies $(\pi^*,\sigma^*) \in S(S)$ such that $V_S[\pi,\sigma^*] = V_S[*,\sigma^*] = V_S[\pi^*,*]$.

2.4. Existence of optimal Markov strategies

We need some more notations in order to prove existence of optimal Markov strategies. Set $s = (s^1,\ldots,s^d) \in \boldsymbol{T}$ such that $|s|$ is even (odd). If we adopt a strategy $(\pi,\sigma) \in S(F; s)$ $(S(S; s)$ resp.) after each reward process i has been already selected s^i times, the value of first-type zero-sum game is given by

$$Z_F[\pi,\sigma](s) = \boldsymbol{E}^{\widetilde{\delta}s}[\Sigma_{t\in N(e,\infty)} \beta^t < f(x_{\pi(|s|+t+1)}), \pi(|s|+t+1) - \sigma(|s|+t) >$$

$$- \Sigma_{t\in N(o,\infty)} \beta^t < g(X_{\sigma(|s|+t+1)}), \sigma(|s|+t+1) - \pi(|s|+t) >].$$

Hence referring to (2.5) and (2.6), we put

$$Z_F[*,\sigma](s) = \text{ess sup}_{\pi\in D(F;s;\sigma)} Z_F[\pi,\sigma](s) \text{ and}$$

$$Z_F[\pi,*](s) = \text{ess inf}_{\sigma\in D(F;s;\pi)} Z_F[\pi,\sigma](s).$$

Similarly for $s = (s^1,\ldots,s^d) \in \boldsymbol{T}$ satisfying that $|s|$ is odd (even) and for $(\pi,\sigma) \in S(S; s)$ $(S(F; s)$ resp.), values of second-type zero-sum games are denoted by

$$Z_S[\pi,\sigma](s) = \boldsymbol{E}^{\widetilde{\delta}s}[\Sigma_{t\in N(o,\infty)} \beta^t < f(X_{\pi(|s|+t+1)}), \pi(|s|+t+1) - \sigma(|s|+t) >$$

$$- \Sigma_{t\in N(e,\infty)} \beta^t < g(X_{\sigma(|s|+t+1)}), \sigma(|s|+t+1) - \pi(|s|+t) >].$$

[1] This description is referred from the value of the reward process in Mandelbaum [8,(2.2)], by shifting time.

Then we define

$$Z_S[*,\sigma](s) = \text{ess sup}_{\pi\in D(S;s;\sigma)} Z_S[\pi,\sigma](s) \text{ and}$$

$$Z_S[\pi,*](s) = \text{ess inf}_{\sigma\in D(S;s;\pi)} Z_S[\pi,\sigma](s).$$

Now we obtain the following fundamental lemmas.

LEMMA 2. *The following (i) and (ii) hold:*

(i) For $(\pi,\sigma) \in S(F)$ $(S(S))$ and $r \in N(e,\infty)$ $(N(o,\infty)$ resp.) it holds that

$$Z_F[\pi,\sigma](\sigma(r)) = E^{\tilde{\delta}\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > + \beta\, Z_S[\pi,\sigma](\pi(r+1))].$$

(ii) For $(\pi,\sigma) \in S(S)$ $(S(F))$ and $r \in N(e,\infty)$ $(N(o,\infty)$ resp.) it holds that

$$Z_S[\pi,\sigma](\pi(r)) = E^{\tilde{\delta}\pi(r)}[< -g(X_{\sigma(r+1)}),\ \sigma(r+1) - \pi(r) > + \beta\, Z_F[\pi,\sigma](\sigma(r+1))].$$

PROOF. (i) Fix any $(\pi,\sigma) \in S(F)$ and $r \in N(e,\infty)$. Then we have

$$\begin{aligned} Z_F[\pi,\sigma](\sigma(r)) &= E^{\tilde{\delta}\sigma(r)}[\Sigma_{t\in N(e,\infty)}\, \beta^t < f(X_{\pi(r+t+1)}),\ \pi(r+t+1) - \sigma(r+t) > \\ &\quad - \Sigma_{t\in N(o,\infty)}\, \beta^t < g(X_{\sigma(r+t+1)}),\ \sigma(r+t+1) - \pi(r+t) >] \\ &= E^{\tilde{\delta}\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > \\ &\quad + \beta\, E^{\tilde{\delta}\pi(r+1)}[\beta < f(X_{\pi(r+3)}),\ \pi(r+3) - \sigma(r+2) > + \cdots\cdots \\ &\quad - \Sigma_{t\in N(o,\infty)}\, \beta^{t-1} < g(X_{\sigma(r+t+1)}),\ \sigma(r+t+1) - \pi(r+t) >]] \\ &= E^{\tilde{\delta}\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > + \beta\, Z_S[\pi,\sigma](\pi(r+1))]. \end{aligned}$$

Therefore we obtain (i) in the case where $(\pi,\sigma) \in S(F)$ and $r \in N(e,\infty)$. The other cases are similarly.

LEMMA 3. *The following (i) and (ii) hold:*

(i) For $(\pi,\sigma) \in S(F)$ $(S(S))$ and $r \in N(e,\infty)$ $(N(o,\infty)$ resp.) it holds that

$$\begin{aligned} &Z_F[*,\sigma](\sigma(r)) \\ &= \text{ess sup}_{\pi\in D(F;\sigma(r);\sigma)} E^{\tilde{\delta}\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > + \beta\, Z_S[*,\sigma](\pi(r+1))] \\ &= \sup_{\pi'\in MS(F;1)} E^{X\sigma(r)}[< f(X_{\pi'(1)}),\ \pi'(1) > + \beta\, Z_S[*,\sigma'](\pi'(1))], \end{aligned}$$

where we take strategies $\pi' \in MS(F;1)$ and σ' by $\pi(r+1,\omega) = \pi'(1,\theta_{\sigma(r)}\omega) + \sigma(r,\omega)$ and $\sigma(r+t,\omega) = \sigma'(t,\theta_{\sigma(r)}\omega) + \sigma(r,\omega)$ for all $t \in N(e,\infty)$ and $\omega \in \Omega$.

(ii) For $(\pi,\sigma) \in S(S)$ $(S(F))$ and $r \in N(e,\infty)$ $(N(e,\infty)$ resp.) it holds that

$$\begin{aligned} &Z_S[\pi,*](\pi(r)) \\ &= \text{ess inf}_{\sigma\in D(S;\pi(r);\pi)} E^{\tilde{\delta}\pi(r)}[< -g(X_{\pi(r+1)}),\ \sigma(r+1) - \pi(r) > + \beta\, Z_F[\pi,*](\sigma(r+1))] \\ &= \inf_{\sigma'\in MS(S;1)} E^{X\pi(r)}[< -g(X_{\sigma'(1)}),\ \sigma'(1) > + \beta\, Z_F[\pi',*](\sigma'(1))], \end{aligned}$$

where we take $\sigma' \in MS(S;1)$ and π' by $\pi(r+t,\omega) = \pi'(t,\theta_{\pi(r)}\omega) + \pi(r,\omega)$ and $\sigma(r+t,\omega) = \sigma'(1,\theta_{\pi(r)}\omega) + \pi(r,\omega)$ for all $t \in N(e,\infty)$ and $\omega \in \Omega$.

PROOF. Fix any $(\pi,\sigma) \in S(F)$ and $r \in N(e,\infty)$. We shall show only this case of

(i), because the other cases are similar. The definition of the essential supremum implies that there exists a sequence $\{(\pi_n, \sigma)\}_{n\in N}$ of strategies of $S(F)$ satisfying

$$\begin{aligned} &\pi_n(t) = \pi(t) \qquad \text{for all odd } t \text{ satisfying } 0 \le t \le r+1, \\ &\boldsymbol{E}^{\hat{\wedge}\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > + \beta\, Z_S[*,\sigma](\pi(r+1))] \text{ and} \\ &\lim_{n\to\infty} \{\boldsymbol{E}^{\hat{\wedge}\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > + \beta\, Z_S[\pi_n,\sigma](\pi(r+1))]\} \end{aligned} \tag{2.10}$$

Then we have

$$(2.10) = \lim_{n\to\infty} Z_F[\pi_n,\sigma](\sigma(r)) \le Z_F[*,\sigma](\sigma(r)).$$

Therefore we obtain

$$\begin{aligned} &\text{ess sup}_{\pi\in D(F;\sigma(r);\sigma)}\, \boldsymbol{E}^{\hat{\wedge}\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > + \beta\, Z_S[*,\sigma](\sigma(r+1))] \\ &\qquad \le Z_F[*,\sigma](\sigma(r)). \end{aligned}$$

On the other hand Lemma 2 implies

$$\begin{aligned} &Z_F[\pi,\sigma](\sigma(r)) = \boldsymbol{E}^{\hat{\wedge}\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > + \beta\, Z_S[\pi,\sigma]\ (\pi(r+1))] \\ &\le \text{ess sup}_{\pi\in D(F;\sigma(r);\sigma)}\, \boldsymbol{E}^{\hat{\wedge}\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > + \beta\, Z_S[*,\sigma](\pi(r+1))]. \end{aligned}$$

Therefore the reverse inequality of (2.11) holds. So we obtain the first equality of (i). Next by using the Markov property and the independency of Markov chains X^i (See Lawler-Vanderbei [3, Theorem 3(b)]), we obtain

$$\begin{aligned} &\text{ess sup}_{\pi\in D(F;\sigma(r);\sigma)}\, \boldsymbol{E}^{F\sigma(r)}[< f(X_{\pi(r+1)}),\ \pi(r+1) - \sigma(r) > + \beta\, Z_S[*,\sigma](\pi(r+1))] \\ &= \sup_{\pi'\in MS(F;1)} \boldsymbol{E}^{X\sigma(r)}[< f(X_{\pi'(1)}),\ \pi'(1) > + \beta\, Z_S[*,\sigma'](\pi'(1))], \end{aligned}$$

where we take strategies $\pi' \in MS(F;\ 1)$ and σ' by $\pi\ (r+1,\omega) = \pi'(1,\theta_{\sigma(r)}\omega) + \sigma(r,\omega)$ and $\sigma(r+t,\omega) = \sigma'(t,\theta_{\sigma(r)}\omega) + \sigma(r,\omega)$ for each $(\pi,\sigma) \in S(F)$, $N(e,\infty)$ and $\omega \in \Omega$. Thus we obtain (i). □

Proposition 1. *For $(\pi,\sigma) \in S(F)$ ($(\pi',\sigma') \in S(S)$), there exist Markov strategies $\pi_M \in D(F;\sigma)$ and $\sigma_M \in D(F;\pi)$ ($\pi'_M \in D(S;\sigma')$ and $\sigma'_M \in D(S;\pi')$) satisfying the following (i) ((ii) resp.):*

(i) $V_F[\pi_M,\sigma] = V_F[*,\sigma]$ *and* $V_F[\pi,\sigma_M] = V_F[\pi,*]$.

(ii) $V_S[\pi'_M,\sigma'] = V_S[*,\sigma']$ *and* $V_S[\pi',\sigma'_M] = V_S[\pi',*]$.

Proof. (i) Fix any strategy $(\pi,\sigma) \in S(F)$. Lemma 3 implies that for each $r \in N(e,\infty)$

$$Z_F[*,\sigma](\sigma(r)) = \sup_{\pi'\in MS(F;1)} \boldsymbol{E}^{X\sigma(r)}[< f(X_{\pi'(1)}),\ \pi'(1) > + \beta\, Z_S[*,\sigma'](\pi'(1))], \tag{2.12}$$

where we take strategies $\pi' \in MS(F;\ 1)$ and σ' by $\pi(r+1,\omega) = \pi'(1,\theta_{\sigma(r)}\omega) + \sigma(r,\omega)$ and $\sigma(r+t,\omega) = \sigma'(t,\theta_{\sigma(r)}\omega) + \sigma(r,\omega)$ for each $(\pi,\sigma) \in S(F)$, $t \in N(e,\infty)$ and $\omega \in \Omega$.

Hence it holds that

$$\sup_{\pi'\in MS(F;1)} \boldsymbol{E}^{X\sigma(r)}[< f(X_{\pi'(1)}),\ \pi'(1) > + \beta\, Z_S[*,\sigma'](\pi(1))]$$

$$= \max_{1\le j\le d} \boldsymbol{E}^{X\sigma(r)}[f^j(X^j_1) + \beta\, Z_S[*,\sigma'](e_j)].$$

Here we define

$$\Gamma_i = \{ \max_{1\le j\le d} \boldsymbol{E}^{X\sigma(r)}[f^j(X^i_1) + \beta\, Z_S[*,\sigma'](e_j)]$$
$$= \boldsymbol{E}^{X\sigma(r)}[f^i(X^i_1) + \beta\, Z_S[*,\sigma'](e_i)] \} \text{ for } i = 1, \ldots, d.$$

Further we set $\Gamma'_1 = \Gamma_1$ and $\Gamma'_{i+1} = \Gamma_{i+1} - (\Gamma_1 \cup \ldots \cup \Gamma_i)$ for $i = 1, \ldots, d-1$. By putting

$$\pi''_M(1,\theta_{\sigma(r)}\omega) = e_i \qquad \text{for } r \in \boldsymbol{N}(e,\infty),\ i = 1, \ldots, d \text{ and } \omega \in \Gamma'_i,$$

we have $\pi''_M \in MS(F;1)$ and then the supremum of (2.12) is attained by π''_M:

$$Z_F[*,\sigma](\sigma(r)) = \boldsymbol{E}^{X\sigma(r)}[< f(X_{\pi''_M(1)}),\ \pi''_M(1) > + \beta\, Z_S[*,\sigma'](\pi''_M(1))]$$

for each $r \in \boldsymbol{N}(e,\infty)$. Hence owing to Lemma 1, we may inductively define a Markov strategy π_M by

$$\pi_M(r+1,\omega) = \pi''_M(1,\theta_{\sigma(r)}\omega) + \sigma(r,\omega) \text{ for } \omega \in \Omega \text{ and each } r \in \boldsymbol{N}(e,\infty). \tag{2.13}$$

Then we obtain

$$\boldsymbol{E}^{X\sigma(r)}[< f(X_{\pi''_M(1)}),\ \pi''_M(1) > + \beta\, Z_S[*,\sigma'](\pi''_M(1))]$$
$$= \boldsymbol{E}^{\tilde{\pi}\sigma(r)}[< f(X_{\pi_M}(r+1)),\ \pi_M(r+1) - \sigma(r) > + \beta\, Z_S[*,\sigma](\pi_M(r+1))].$$

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

$$Z_F[*,\sigma]\ (\sigma(r)) = \boldsymbol{E}^{\tilde{\pi}\sigma(r)}[< f(X_{\pi_M(r+1)}),\pi_M(r+1)) - \sigma(r) > + \beta\, Z_S[*,\sigma](\pi_M(r+1))]. \tag{2.14}$$

On the other hand owing to Lemma 2 (ii), we have that for all $r \in N(e,\infty)$

$$Z_S[*,\sigma](\pi_M(r+1)) \tag{2.15}$$
$$= \boldsymbol{E}^{\tilde{\pi}\pi_M(r+1)}[< -\, g(X_{\sigma(r+2)}),\ \sigma(r+2) - \pi_M(r+1) > + \beta\, Z_F[*,\sigma](\sigma(r+2))].$$

Hence (2.14) and (2.15) conclude the results that $V_F[\pi_M,\sigma] = V_F[*,\sigma]$ and $(\pi_M,\sigma) \in S(F)$. We can also check the other equations similarly. □

3. The Optimal Values and the Optimal Strategies

In this section we investigate the following backward iteration in order to find the optimal values in both type zero-sum 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 values.

3.1. A value iteration and the optimal values

Let us consider the following iteration.

ITERATION 1.

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

For $r \in N$ *we define successively as follows.*

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

$$U_{F,r+1}(x) = \max_{1 \le i \le d} E^x[f^i(X_1^i) + \beta\, U_{S,r}(x^1, \ldots, X_1^i, \ldots, x^d)].$$

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

$$U_{S,r+1}(x) = \min_{1 \le i \le d} E^x[-g^i(X_1^i) + \beta\, U_{F,r}(x^1, \ldots, X_1^i, \ldots, x^d)].$$

First we shall prove convergence of sequences $\{U_{F,r}\}_{r \in N}$ and $\{U_{S,r}\}_{r \in N}$ in Iteration 1. Let $\| \cdot \|$ denote the supremum norm on the space of bounded measurable functions on E. For Markov strategies $\pi \in MS(F; 1)$ ($\sigma \in MS(S; 1)$) and for $i = 1, \ldots, d$ we shall introduce the following semi-linear operators $S_A{}^{\pi}$ ($S_B{}^{\sigma}$ resp.) and $S_A{}^i$ and $S_B{}^i$ on the space of all bounded measurable functions on E:

$$S_A{}^{\pi}\phi(x) = E^x[< f(X_{\pi(1)}), \pi(1) > + \beta\, \phi(X_{\pi(1)})] \quad (x \in E),$$

$$S_B{}^{\sigma}\phi(x) = E^x[< -g(X_{\sigma(1)}), \sigma(1) > + \beta\, \phi(x_{\sigma(1)})] \quad (x \in E),$$

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

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

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

LEMMA 4. *Let* u_1, u_2, v_1 and v_2 *be bounded measurable functions on E such that*

$$v_j = \max_{1 \le i \le d}{}^2\, S_A{}^i u_j \quad \textit{for } j = 1, 2.$$

Then it holds that $\| v_1 - v_2 \| \le \beta \| u_1 - u_2 \|$.

PROOF. We obtain this lemma, since for each $x \in E$ we have

$$|v_1(x) - v_2(x)| = |\max_{1 \le i \le d} S_A{}^i u_1(x) - \max_{1 \le i \le d} S_A{}^i u_2(x)|$$
$$\le \max_{1 \le i \le d} |S_A{}^i u_1(x) - S_A{}^i u_2(x)| \le \beta \| u_1 - u_2 \|.$$

LEMMA 5. *For each* r, $r' \in N$, *the following* (*i*) *and* (*ii*) *hold*:

(*i*) $\| U_{F,r+r'+1} - U_{F,r+1} \| \le \beta \| U_{S,r+r'} - U_{S,r} \|$.

(*ii*) $\| U_{S,r+r'+1} - U_{S,r+1} \| \le \beta \| U_{F,r+r'} - U_{F,r} \|$.

PROOF. For each r, $r' \in N$ we have

$$U_{F,r+1} = \max_{1 \le i \le d} S_A{}^i U_{S,r} \text{ and } U_{F,r+r'+1} = \max_{1 \le i \le d} S_A{}^i U_{S,r+r'}.$$

By using Lemma 4, we obtain (i). The proof of (ii) is similar.

Then we obtain the following results regarding Iteration 1.

---

[2] $\max\{\phi, \psi\}$ denotes $\max\{\phi, \psi\}(x) = \max\{\phi(x), \psi(x)\}$ for functions ϕ and ψ on E and $x \in E$.

THEOREM 1. *Iteration 1 converges*:

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

Further U_F and U_S is a unique solution of the following equations (3.1):

$$U_F = \max_{1\le i\le d} S_A{}^i U_S \text{ and } U_S = \min_{1\le i\le d} S_B{}^i U_F. \tag{3.1}$$

PROOF. From Lemma 5, we have for each r, $r' \in \boldsymbol{N}$

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

We inductively obtain $\| U_{F,r+r'} - U_{F,r} \| \le \beta^r \| U_{F,r'} - U_{F,0} \|$ for all $r' \in \boldsymbol{N}$ 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 (3.1), by applying the bounded convergence theorem to Iteration 1. Finally the uniqueness of solutions U_F and U_S is easily checked, by using Lemma 4.

COROLLARY 1.

$$U_F = \sup_{\pi\in MS(F;1)} S_A{}^\pi U_S \text{ and } U_S = \inf_{\sigma\in MS(S;1)} S_B{}^\sigma U_F.$$

PROOF. They are trivial from (3.1), by considering the definition of one-step Markov strategies.

3.2. Construction of the optimal strategies and uniqueness of the optimal values

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

$$D'_A{}^j = \{ \max_{1\le i\le d} S_A{}^i U_S = S_A{}^j U_s \} \text{ for } j = 1, \ldots, d,$$
$$D'_B{}^j = \{ \min_{1\le i\le d} S_B{}^i U_F = S_B{}^j U_F \} \text{ for } j = 1, \ldots, d.$$

Further we let

$$D_A{}^{i+1} = D'_A{}^{i+1} - (D'_A{}^1 \cup \ldots \cup D'_A{}^i) \text{ for } i = 1, \ldots, d-1,$$
$$D_B{}^{i+1} = D'_B{}^{i+1} - (D'_B{}^1 \cup \ldots \cup D'_B{}^i) \text{ for } i = 1, \ldots, d-1.$$

Then by putting

$$\pi^\circ(1) = e_i \ (\sigma^\circ(1) = e_i) \text{ on } \{X_0 \in D_A{}^i \ (D_B{}^i)\} \text{ for } i = 1, \ldots, d,$$

we have Markov strategies $\pi^\circ \in MS(F; 1)$ ($\sigma^\circ \in MS(S; 1)$ resp.).

Hence owing to Lemma 1 we may give another representation of Markov strategies. For $(\pi,\sigma) \in MS(F; r)$ we describe (π,σ) as

$$[\pi_1,\sigma_2,\pi_3,\sigma_4,\ldots], \tag{3.2}$$

where π_t and σ_t are player $A's$ (player $B's$ resp.) one step Markov strategies for t. Hence the meaning of (3.2) is as follows. Player A selects a reward process, by using Markov strategy π_1. Next player B selects, by using Markov strategy σ_2. Further player

A does, by using Markov strategy π_3. The game continues in this way. Moreover we have similar representations concerning second-type Markov strategies: For $(\pi,\sigma) \in MS(S;\ r)$ we write (π,σ) as $[\sigma_1,\pi_2,\sigma_3,\pi_4,\ldots]$. Hence by using these representations, we give the following Markov strategies $(\pi^*,\sigma^*) \in MS(F)$ and $(\pi'^*,\sigma'^*) \in MS(S)$ by

$$(\pi^*,\sigma^*) = [\pi^\circ,\sigma^\circ,\pi^\circ,\sigma^\circ,\pi^\circ,\sigma^\circ,\ldots] \text{ and } (\pi'^*,\sigma'^*) = [\sigma^\circ,\pi^\circ,\sigma^\circ,\pi^\circ,\sigma^\circ,\pi^\circ,\ \ldots]. \tag{3.3}$$

Then we obtain the following results.

THEOREM 2. *$(\pi^*,\sigma^*) \in S(F)$ ($(\pi'^*,\sigma'^*) \in S(S)$) is an optimal strategy and U_F (U_S) is an optimal value for the first-type (second-type resp.) zero-sum game:*

(i) $V_F[\pi,\sigma^*] \leq U_F = V_F[\pi^*,\sigma^*] \leq V_F[\pi^*,\sigma^*]$ *for every* $\pi \in D(F;\sigma^*)$ *and* $\sigma \in D(F;\pi^*)$.

(ii) $V_S[\pi',\sigma'^*] \leq U_S = V_S[\pi'^*,\sigma'^*] \leq V_S[\pi',{}^*\sigma']$ *for every* $\pi' \in D(S;\sigma'^*)$ *and* $\sigma' \in D(S;\pi'^*)$.

PROOF. First we shall show that the inequality of (i) holds for Markov strategies. From Corollary 1 (i) and (ii) we have

$$U_F = S_A{}^{\pi^\circ}U_S \geq S_A{}^{\pi}U_S \quad \text{and} \quad U_S = S_B{}^{\sigma^\circ}U_F \tag{3.4}$$

for every Markov strategy $\pi \in MS(F;\ 1)$. From (3.4) we obtain

$$U_F = S_A{}^{\pi^\circ}S_B{}^{\sigma^\circ}U_F \geq S_A{}^{\pi}S_B{}^{\sigma^\circ}U_F$$

for every Markov $\pi \in MS(F;\ 1)$. Therefore we inductively obtain

$$\begin{aligned} U_F &= S_A{}^{\pi^\circ}S_B{}^{\sigma^\circ}S_A{}^{\pi^\circ}S_B{}^{\sigma^\circ}\ldots S_A{}^{\pi^\circ}S_B{}^{\sigma^\circ}U_F \\ &\geq S_A{}^{\pi_1}S_B{}^{\sigma^\circ}S_A{}^{\pi_3}S_B{}^{\sigma^\circ}\ldots S_A{}^{\pi_{2r-1}}S_B{}^{\sigma^\circ}U_F \end{aligned} \tag{3.5}$$

for every $r \in N$ and every Markov $\pi_t \in MS(F;\ 1)$ $(t \in N(o,2r))$. Hence from the definitions of $S_A{}^{\pi}$ and $S_B{}^{\sigma}$ we have

$$\| S_A{}^{\pi}\phi_1 - S_A{}^{\pi}\phi_2 \| \leq \beta \| \phi_1 - \phi_2 \| \text{ and } \| S_B{}^{\sigma}\phi_1 - S_B{}^{\sigma}\phi_2 \| \leq \beta \| \phi_1 - \phi_2 \|$$

for $\pi \in MS(F;\ 1)$, $\sigma \in MS(S;\ 1)$ and bounded measurable functions ϕ_1, ϕ_2 on E. By letting r infinite in (3.5), we obtain

$$U_F = V_F[\pi^*,\sigma^*] \geq V_F[\pi,\sigma^*],$$

where $(\pi^*,\sigma^*) = [\pi^\circ,\sigma^\circ,\pi^\circ,\sigma^\circ,\pi^\circ,\sigma^\circ,\ldots] \in MS(F)$ and $(\pi,\sigma^*) = [\pi_1,\sigma^\circ,\pi_3,\sigma^\circ,\pi_5,\sigma^\circ,\ldots] \in MS(F)$. Since the other Markov cases can be proved similarly, we obtain the inequalities (i) for every Markov strategies $\pi \in D(F;\sigma^*)$ and $\sigma \in D(S;\pi^*)$.

Next we shall show the non-Markov case. Hence by the use of Proposition 1, there exists a Markov strategy $\pi_M{}^* \in D(F;\sigma^*)$ satisfying

$$V_F[\pi_M{}^*,\sigma^*] = V_F[*,\sigma^*].$$

Then we have

$$V_F[\pi_M{}^*,\sigma^*] = V_F[*,\sigma^*] \geq V_F[\pi^*,\sigma^*].$$

On the other hand from the definitions (2.13) and Lemma 1, $\pi_M^* \in D(F;\sigma^*)$ is Markov. Therefore owing to the first part of this proof we obtain

$$U_F = V_F[\pi^*,\sigma^*] \geq V_F[\pi_M^*,\sigma^*].$$

Thus we conclude

$$U_F = V_F[\pi^*,\sigma^*] = V_F[\pi_M^*,\sigma^*] = V_F[*,\sigma^*].$$

Since the other inequalities can be proved similarly, the proof is completed.

Now owing to Proposition 1 we may respectively define the lower bound $\underline{V}_F$ and the upper bound $\overline{V}_F$ of values in the first-type zero-sum games by

$$\underline{V}_F = \sup_\pi \inf_\sigma V_F[\pi,\sigma] \text{ and } \overline{V}_F = \inf_\sigma \sup_\pi V_F[\pi,\sigma].$$

In the second-type we similarly put

$$\underline{V}_S = \sup_\pi \inf_\sigma V_S[\pi,\sigma] \text{ and } \underline{V}_S = \inf_\sigma \sup_\pi V_S[\pi,\sigma].$$

Finally we obtain the following results concerning the optimal values.

COROLLARY 2. *The zero-sum games have the unique optimal values*:

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

PROOF. From Theorem 2 we have

$$\overline{V}_F \leq \sup_\pi V_F[\pi,\sigma^*] = V_F = \inf_\sigma V_F[\pi^*,\sigma] \leq \underline{V}_F.$$

Since $\underline{V}_F \leq \overline{V}_F$ is trivial, we obtain the result. The other is similar.

Acknowledgement

The author is grateful to thank Prof. N. Furukawa for his comments and suggestions.

References

[1] BERRY, D. A. and FRISTEDT, B.: *Bandit problems*, Chapman and Hall, London, (1985).
[2] GITTINS, J. C.: *Multi-armed bandit allocation indices*, John Wiley and Sons Ltd., England, (1989).
[3] LAWLER, G. F. and VANDERBEI, R. J.: *Markov strategies for optimal control problems indexed by a partially ordered set*, Ann. Prob. **11**, (1983), 642–647.
[4] MANDELBAUM, A.: *Discrete multi-armed bandits and multi-parameter processes*, Probab. Th. Rel. Fields **71**, (1986), 129–147.
[5] MANDELBAUM, A.: *Continuous multi-armed bandits and multi-parameter processes*, Ann. Prob. **14**, (1987), 1527–1556.
[6] MANDELBAUM, A. and VANDERBEI, R. J.: *Optimal stopping and supermartingales over partially ordered sets*, Z. Wahr, Verw. Geb. **57**, (1981), 253–264.
[7] MAZZIOTTO, G.: *Two parameter optimal stopping and bi-Markov processes*, Z. Wahr. verw., Gebiete **69**, (1985), 99–135.
[8] WHITTLE, P.: *Multi-armed bandits and the Gittins index*, J. Roy. Statist. Soc. Ser., **B41**, (1980), 148–164.

Received September 11, 1990
Communicated by N. Furukawa