Stochastic Games with Metric State Space

By H.A.M. Couwenbergh, Eindhoven

Abstract: In this paper the stochastic two-person zero-sum game of Shapley is considered, with metric state space and compact action spaces. It is proved that both players have stationary optimal strategies, under conditions which are weaker than those of Maitra/Parthasarathy (a.o. no compactness of the state space). This is done in the following way: we show the existence of optimal strategies first for the one-period game with general terminal reward, then for the n-period games \( n = 1, 2, \ldots \); further we prove that the game over the infinite horizon has a value \( v \), which is the limit of the \( n \)-period game values. Finally the stationary optimal strategies are found as optimal strategies in the one-period game with terminal reward \( v \).

1. Introduction

The stochastic games we consider are non-cooperative two-person zero-sum games with discrete time parameter [originally introduced by Shapley]. This means that a system is given with a set of states \( S \), and two so-called action spaces: \( A \) for player I, \( B \) for player II. The system is started at time \( t = 0 \) in a state \( s \in S \); both players choose an action: \( a \in A, b \in B \). As a consequence of these actions, player I receives a "reward" \( r(s, a, b) \) (this may be negative) from II, and the system moves to a new state \( s' \) according to a (sub-) probability measure \( p(\cdot|s, a, b) \) on \( S \) (in case \( p(\cdot|s, a, b) \) is defective, the game has a positive stopping probability). Then this process is repeated at time \( t = 1 \) from the new starting state \( s' \), and so forth. The reward at time \( t \) is multiplied by \( \beta^t \), where \( \beta > 0 \) is called the discount factor. The object of player I is to maximize the total expected discounted reward (over the infinite horizon); the object of player II is to minimize this same amount. Sometimes these opposite ambitions can be met simultaneously, namely if there exists an optimal pair of strategies; then for each player it is not profitable to deviate from his optimal strategy. In this case the game also has a "value" (in fact a real function on \( S \), see section 2 for a definition). In the literature several sets of conditions have been given for the existence of "stationary" optimal strategies. We mention results of Vrieze [1976] and Wessels [1977] in case of a countable state space, and of Maitra/Parthasarathy [1970] and T. Parthasarathy [1973] when \( S \) is a compact metric space.

1) The author is grateful to Professor Wessels for his encouragement and assistance during the research and the preparation of this paper.

2) H.A.M. Couwenbergh, Eindhoven University of Technology, P.O. Box 513, NL--Eindhoven.
The theorems presented in this paper are obtained by combining and generalizing the bounding function method of Wessels, and Maitra/Parthasarathy's approach for a compact state space. The first theorem (section 3) makes use of the continuity on $S$ of reward function $r$ and transition law $p$, and produces a continuous value function; the second one (section 4) uses measurability of $r$ and $p$ on $S$, causing the value function to be measurable. In the proofs of both results we do not need a contraction theorem in order to find the value. Section 5 contains an extension and some remarks; we start in section 2, with a formal model and preparations.

2. Model and Prerequisites

First we have to give a definition of (sub-) transition probability. Let $X$ and $Y$ be metric spaces; denote by $\sigma_Y$ the Borel $\sigma$-algebra on $Y$. The map $q : \sigma_Y \times X \to [0,1]$ (closed unit interval) is called a (sub-) transition probability (abbreviation: (sub-) trpr) $X \to Y$ if $q (\cdot \mid x)$ is a measure on $Y$ with $q (Y \mid x) \leq 1$ for all $x \in X$, and $q (Y_0 \mid \cdot)$ is a measurable function on $X$ for all $Y_0 \in \sigma_Y$ (by measurability we shall always mean Borel measurability). If, moreover, $q (Y \mid x) = 1, x \in X$, then $q$ is called a (nondefective) trpr. In this section we require that the state space $S$ and the action spaces $A$ and $B$ be nonempty metric spaces, that the transition law $p$ is a (sub-) trpr $S \times A \times B \to S$, and that the reward $r$ is a measurable map $S \times A \times B \to \mathbb{R}$ (set of reals); $\beta$ is a positive number.

Define for $t \geq 1$ $H_t := S \times A \times B \times S \times \ldots \times B$ ($t$ times $S \times A \times B$); an element $h_t = (s_0, a_0, \ldots, b_{t-1})$ of $H_t$ is called a history. Let $F$ be the set of all trpr's $f, S \to A$, and $G$ the set of all trpr's $g, S \to B$; for $t \geq 1, (A)$ is the set of all trpr's $\pi_t$, $H_t \times S \to A$. Now $\Pi (A) := F \times \Pi_1 (A) \times \Pi_2 (A) \times \ldots$ defines the set of general strategies $\pi$ for player I. A strategy $\pi$ is called Markov if each $\pi_t$ is independent of history, so $\pi = (f_0, f_1, \ldots)$ with $f_t \in F, t \geq 0$. The set of Markov strategies is denoted by $R (A)$. If $\pi = (f, f, \ldots)$ for some $f \in F$, then $\pi$ is said to be stationary; we write $\pi = f^\infty$. Similarly for player II the set $\Pi (B)$ of general strategies, the set $R (B)$ of Markov strategies, and the stationary strategies $g^\infty$ are defined. Take $\Pi := \Pi (A) \times \Pi (B)$, $R := R (A) \times R (B)$.

Let $s_0 \in S$ and $(\pi, \rho) \in \Pi$ be given. According to a theorem of Ionescu Tulcea (see Neveu [1965], Prop. V 1.1; in order to apply this result $p$ has to be nondefective, which may be accomplished by extending $S$ with an extra state $*$; take $p (\cdot \mid s, a, b) := 1 - p (S \mid s, a, b)$ and extend all realvalued functions $h$ on $S$ by $h (\cdot) := 0$) this starting state and pair of strategies determine a (sub-) probability measure $P_{s_0, \pi, \rho}$ with the following property. If $E = S_0 \times A_0 \times \ldots \times B_t$ is a measurable rectangle in $H_{t+1}$ then

$$P_{s_0, \pi, \rho} (E) = \int_{S_0} \left( \int_{A_0} \left( \int_{B_0} \left( \int_{S_1} \ldots \int_{S_t} \int_{A_t} \left( \int_{B_t} \left( \int_{h_t} \left( \int_{s_t} \left( \int_{S_0} \left( \int_{A_0} \left( \int_{B_0} \left( \int_{S_1} \ldots \right) \ldots \right) \right) \right) \right) \right) \right) \right) \right) \right),$$