Journal of the Operations Research
Society of Japan
Vol.21, No.1, March, 1978

ON CONTINUOUS TIME MARKOV GAMES WITH COUNTABLE STATE SPACE

Kensuke Tanaka

Niigata University

and

Kazuyoshi Wakuta

Nagaoka Technical College

(Received September 9, 1976; Revised April 26, 1977)

ABSTRACT

This paper is a continuation of our papers [6] and [7] and is concerned with a continuous time Markov game in which the state space is countable and the action spaces of player I and player II are compact metric spaces. In the game, the players continuously observe the state of the system and then choose actions. As a result, the reward is paid to player I from player II and the system moves to a new state by the known transition rates. Then we consider the optimization problem to maximize the total expected discounted gain for player I and, at the same time, to minimize the total expected loss for player II as the game proceeds to the infinite future.

We show that such a two-person zero-sum game is strictly determined and both players have optimal stationary strategies.

1. Introduction and Summary

In recent years, a considerable attention has been given to games in order to solve practical problems in management science and, at the same time, the game theory has been actively investigated. The concept of a Markov game was first formulated by Shapley in [4]. And, by introducing a discount factor and using the results in dynamic programming, Maitra and Parthasarathy investigated a Markov game with an infinite horizon in [3]. In this paper, we consider con-

tinuous time one with a discount factor. Such a game has been hardly found, as far as the present authors are aware, except this paper. Then we use the results with relation to the continuous time Markov decision process given, mainly, in [2].

We give a formulation of our game. We determine "a continuous time Markov game" by a sextuple (S, A, B, q, r, α). Here S is a countable set labeled (1, 2, 3, ...), the set of states of a system; A is a non-empty Borel subset of a complete separable metric space, the set of actions available to player I; B is a non-empty Borel subset of a complete separable metric space, the set of actions available to player II; q is a transition rate function which governs the law of motion of the system and is a bounded function $q(\cdot|i, a, b)$ on S for each triple $(i, a, b) \in S \times A \times B$; r, a reward function, is a bounded measurable function on $S \times A \times B$; α, a discount factor, is a positive number.

In this game, player I and player II continuously observe the state of the system and classify it into one of the possible states $i \in S$ and then player I and player II choose actions $a \in A$ and $b \in B$, respectively. As a consequence of the present states $i \in S$ and the actions a and b chosen by the players, player II pays player I reward r(i, a, b) unit of money and the system moves to a new state $j \in S$, which is governed by the transition rate $q(j|i, a, b)$. Then, our optimization problem is to maximize the total expected discounted gain for player I and, at the same time, to minimize the total expected discounted loss for player II as the game proceeds to the infinite future.

We assume that strategies for player I and player II are independent of the past history of the system and depend only on the present state of the system. Such a strategy $\pi = \pi(t)$ for player I is specified by a family $\{\mu_t\}$, where $\{\mu_t\}$ is a function $\mu_t(M|i)$ of i, t and M such that for each $i \in S$ and $t \in [0, \infty)$ it is a probability measure $\mu_t(\cdot|i)$ on the measurable space (A, B(A)) where B(A) is the σ-field generated by the metric on A, and that for each $i \in S$ and $M \in B(A)$ it is a Lebesgue measurable function $\mu_\cdot(M|i)$. Then we call such a strategy a Markov strategy. Moreover, such a Markov strategy $\pi = \pi(t)$ is said to be stationary if $\pi(t)$ is independent of t, that is, there exists a map μ from S into P_A such that $\mu_t = \mu$ for all $t \in [0, \infty)$, where P_A is the set of all probability measures on (A, B(A)). Π denotes the class of all Markov strategies for player I. Markov strategies and stationary strategies for player II are defined analogously. Γ denotes the class of all Markov strategies for player II.

Throughout the paper, we assume, for the transition rate matrix $Q(a, b) = \{q(j|i, a, b)\ ;\ i, j \in S\}$, the following:

Assumption 1. For each $i, j \in S$, $q(j|i, a, b)$ is a continuous function on

$A \times B$, and for all $a \in A$, $b \in B$, $q(j|i, a, b) \gtrsim 0$, $j \neq i$, $\Sigma_j q(j|i, a, b) = 0$ and $|q(i|i, a, b)| \lesssim M$ for all $i \in S$ and some positive number $M < \infty$.

When a pair of the Markov strategies, (π, σ) for player I and player II is used, transition rates are defined as follwos: for each $t \gtrsim 0$,

$$q(j|i, t, \pi, \sigma) = \int\int q(j|i, a, b) d\mu_t(a|i) d\lambda_t(b|i),$$

where the strategies π and σ are specified by the families $\{\mu_t\}$ and $\{\lambda_t\}$, respectively. Then, for all i and $j \in S$, $q(j|i, t, \pi, \sigma)$ are plainly Lebesgue measurable in t and, from Assumption 1, satisfy the following conditions: for each $t \gtrsim 0$,

$$q(j|i, t, \pi, \sigma) \gtrsim 0, \quad j \neq i, \quad \Sigma_j q(j|i, t, \pi, \sigma) = 0 \tag{1.1}$$

and

$$|q(i|i, t, \pi, \sigma)| \lesssim M. \tag{1.2}$$

We write the transition rate matrix corresponding to π and σ as $Q(t, \pi, \sigma) = \{q(j|i, \pi, \sigma) ; i, j \in S\}$ and if π and σ are stationary strategies, we write $Q(\pi, \sigma)$ instead of $Q(t, \pi, \sigma)$.

The existence of a unique transition function corresponding to a given transition rate matrix was shown by Feller for arbitrary state space, and Reuter and Ledermann gave a simpler aprroach suitable for denumerable state space. These authors considered the case where the transition rate is continuous in t. In our situation, the transition rate is only measurable in t. In this case, using the approach of Reuter and Ledermann, Kakumann showed in [1] that, under the conditions (1.1) and (1.2), there exists a unique stochastic matrix $F(s, t, \pi, \sigma) = \{f_{ij}(s, t, \pi, \sigma) ; i, j \in S\}$ corresponding to $Q(t, \pi, \sigma)$ which satisfies the Kolmogorov forward differential equations: for almost all $t \in [s, \infty)$,

$$\frac{\partial}{\partial t} F(s, t, \pi, \sigma) = F(s, t, \pi, \sigma) Q(t, \pi, \sigma) \text{ with } F(s, s, \pi, \sigma) = I, \tag{1.3}$$

where I is the infinite identity matrix.

It was also shown that a measurable Markov process $\{X(t, \pi, \sigma) ; t \gtrsim s\}$ corresponding to the stochastic matrix $F(s, t, \pi, \sigma)$ exists and is well-behaved. Then, for a pair of Markov strategies (π, σ), $Q(t, \pi, \sigma)$ which governs the law of motion of the system, is an infinitesimal generator of the process $\{X(t, \pi, \sigma) ; t \gtrsim s\}$ in the sense that

$$f_{ij}(t+\Delta t, \pi, \sigma) = P\{X(t+\Delta t, \pi, \sigma) = j \mid X(t, \pi, \sigma) = i\}$$
$$= \delta_{ij} + q(j \mid i, t, \pi, \delta)\Delta t + o(\Delta t),$$

where δ_{ij} is the Kronecker delta and $\frac{o(\Delta t)}{\Delta t} \to 0$ as $\Delta t \to 0^+$. In the game, the state of the system moves according to the Markov process $\{X(t, \pi, \sigma) ; t \geq S\}$ with initial time $s = 0$. In this view we write $F(t, \pi, \sigma)$ instead of $F(0, t, \pi, \sigma)$. Now, we define the expected discounted gain function. When a pair of the Markov strategies (π, σ) is chosen by player I and player II, at any time t the expected gain rate for player I out of state $i \in S$ is given by

$$r(i, t, \pi, \sigma) = \int\int r(i, a, b)d\mu_t(a \mid i)d\lambda_t(b \mid i). \tag{1.4}$$

It is clear that $r(i, t, \pi, \sigma)$ is a Lebesgue measurable function of t. Thus when the system starts from a state $i \in S$ and a pair of Markov strategies (π, σ) is used, the total expected discounted gain for player I is defined to be

$$\psi(i, \pi, \sigma) = \int_0^\infty e^{-\alpha t} \sum_j f_{ij}(t, \pi, \sigma) r(j, t, \pi, \sigma) dt. \tag{1.5}$$

A Markov strategy π^* is optimal for player I if, for all $\sigma' \in \Gamma$ and $i \in S$,

$$\inf_{\sigma \in \Gamma} \sup_{\pi \in \Pi} \psi(i, \pi, \sigma) \leq \psi(i, \pi^*, \sigma').$$

A Markov strategy σ^* is optimal for player II if, for all $\pi' \in \Pi$ and $i \in S$,

$$\sup_{\pi \in \Pi} \inf_{\sigma \in \Gamma} \psi(i, \pi, \sigma) \geq \psi(i, \pi', \sigma^*).$$

We say that a continuous time Markov game is strictly determined if for all initial states $i \in S$,

$$\sup_{\pi \in \Pi} \inf_{\sigma \in \Gamma} \psi(i, \pi, \sigma) = \inf_{\sigma \in \Gamma} \sup_{\pi \in \Pi} \psi(i, \pi, \sigma).$$

This common quantity as a function on S is called the value function of the game.

2. The Existence of Optimal Stationary Strategies

In this section, we shall show the existence of optimal stationary strategies. First, in order to prove main results, we assume the following:

Assumption 2. (i) A and B are compact metric spaces, (ii) $r(i, a, b)$ is a

continuous function on $A \times B$ for each $i \in S$.

Then, by Assumption 2(i), it is well known that P_A and P_B endowed with the weak topologies are compact metric spaces. Throughout the paper, we assume that P_A and P_B are endowed with the weak topologies.

Let $C(S)$ denote the family of all bounded functions on S. For $u \in C(S)$ we define $\|u\| = \sup_i |u(i)|$. $C(S, d)$ is a complete metric space, where $d(u, v) = \|u - v\|$ for each u and $v \in C(S)$.

Lemma 1. For each $i \in S$ and $u \in C(S)$, $\sum_j q(j|i, a, b)u(j)$ converges uniformly in a and b. As a result from this, $\sum_j q(j|i, a, b)u(j)$ is a bounded continuous function on $A \times B$.

Proof. From Assumption 1 and 2(i), for each $i \in S$, $\sum_{j \neq i} q(j|i, a, b) = -q(i|i, a, b)$ converges uniformly in a and b by Dini's theorem. For $\varepsilon > 0$, there exists a positive integer $N > i$ such that, for all $n \gtrsim N$, $a \in A$ and $b \in B$,

$$\sum_{j \gtrsim n} q(j|i, a, b) \lesssim \frac{\varepsilon}{\|u\|}. \tag{2.1}$$

Then, from (2.1), we obtain for all $n \gtrsim N$, $a \in A$ and $b \in B$,

$$\left|\sum_{j \gtrsim n} q(j|i, a, b)u(j)\right| \lesssim \varepsilon. \tag{2.2}$$

From (2.2), $\sum_j q(j|i, a, b)u(j)$ converges uniformly in a and b. Thus, the lemma is proved.

Now, for each $\mu \in P_A$ and $\lambda \in P_B$, we define an operator $L(\mu, \lambda): C(S) \to C(S)$ as follows: for each $i \in S$ and $u \in C(S)$,

$$L(\mu, \lambda)u(i) \equiv r(i, \mu, \lambda) + \sum_j q(j|i, \mu, \lambda)u(j),$$

where

$$r(i, \mu, \lambda) = \iint r(i, a, b)d\mu(a)d\lambda(b)$$

and

$$q(j|i, \mu, \lambda) = \iint q(j|i, a, b)\, d\mu(a)d\lambda(b).$$

Since, by Lemma 1, $\sum_j q(j|i, \mu, \lambda)u(j)$ is continuous on $P_A \times P_B$ and, by Assumption 2(H), $r(i, \mu, \lambda)$ is also continuous on $P_A \times P_B$, $L(\mu, \lambda)u(i)$ is a continuous

function on $P_A \times P_B$. $L(\mu, \lambda)u(i)$, P_A and P_B satisfy the conditions of Sion's minimax theorem (Theorem 3.4 of [5]) because of its bilinearity in (μ, λ). We have consequently for each $i \in S$ and $u \in C(S)$,

$$\sup_{\mu \in P_A} \inf_{\lambda \in P_B} L(\mu, \lambda)u(i) = \inf_{\lambda \in P_B} \sup_{\mu \in P_A} L(\mu, \lambda)u(i). \tag{2.3}$$

Moreover, since P_A and P_B are compact, it was shown in [3] that sup and inf in (2.3) can be replaced by max and min, respectively. Then, there exist maps μ^* and λ^* from S into P_A and P_B, respectively, such that for each $i \in S$ and $u \in C(S)$,

$$\begin{aligned} \min_{\lambda \in P_B} L(\mu^*, \lambda)u(i) &= \max_{\mu \in P_A} \min_{\lambda \in P_B} L(\mu, \lambda)u(i) \\ &= \min_{\lambda \in P_B} \max_{\mu \in P_A} L(\mu, \lambda)u(i) \\ &= \max_{\mu \in P_A} L(\mu, \lambda^*)u(i). \end{aligned} \tag{2.4}$$

In order to prove the main result, the following theorem is important.

Theorem 2.1. There exists a function $v \in C(S)$ such that for each $i \in S$,

$$\alpha v(i) = \max_{\mu \in P_A} \min_{\lambda \in P_B} L(\mu, \lambda)v(i), \tag{2.5}$$

where α is the discount factor.

Proof. First, for each $\mu \in P_A$ and $\lambda \in P_B$, we define a new one-step transition probability matrix $P(\mu, \lambda)$ with relation to $Q(\mu, \lambda)$ by

$$P(\mu, \lambda) = I + \frac{1}{M} Q(\mu, \lambda), \tag{2.6}$$

whose (i, j)th element is given by

$$p(j|i, \mu, \lambda) = \delta_{ij} + \frac{1}{M} q(j|i, \mu, \lambda),$$

where M is the positive number in Assumption 1. Then, since P_A and P_B are compact metric spaces and $r(i, \mu, \lambda)$ and $\sum_j p(j|i, \mu, \lambda)u(j)$ are continuous function on $P_A \times P_B$, we can define an operator $T: C(S) \to C(S)$ as follows: for each $i \in S$ and $u \in C(S)$,

$$Tu(i) = \max_{\mu \in P_A} \min_{\lambda \in P_B} \left\{ \frac{r(i, \mu, \lambda)}{\alpha + M} + \frac{M}{\alpha + M} \sum_j p(j|i, \mu, \lambda)u(j) \right\}.$$

This operator is a contraction mapping on C(S) because $0 < M(\alpha+M)^{-1} < 1$. Furthermore, since C(S) is a complete metric space, T has a unique fixed point in C(S) by the Banach's fixed point theorem. Let v be the unique fixed point of T. Then, it holds that for each $i \in S$,

$$(2.7) \qquad v(i) = \max_{\mu \in P_A} \min_{\lambda \in P_B} \{\frac{r(i, \mu, \lambda)}{\alpha + M} + \frac{M}{\alpha + M} \sum_j p(j|i, \mu, \lambda)v(j)\}.$$

Substituting $q(j|i, \mu, \lambda)$ for $p(j|i, \mu, \lambda)$ in (2.7), we get for each $i \in S$,

$$(2.8) \quad v(i) = \max_{\mu \in P_A} \min_{\lambda \in P_B} \{\frac{r(i, \mu, \lambda)}{\alpha + M} + \frac{M}{\alpha + M} \sum_j (\delta_{ij} + \frac{1}{M} q(j|i, \mu, \lambda))v(j)\}$$

$$= \max_{\mu \in P_A} \min_{\lambda \in P_B} \{\frac{r(i, \mu, \lambda)}{\alpha + M} + \frac{M}{\alpha + M} v(i) + \frac{1}{\alpha + M} \sum_j q(j|i, \mu, \lambda)v(j)\}.$$

Multiplying both sides of (2.8) by $(\alpha + M)$ and, then subtracting $Mv(i)$ from both sides, we obtain

$$\alpha v(i) = \max_{\mu \in P_A} \min_{\lambda \in P_B} \{r(i, \mu, \lambda) + \sum_j q(j|i, \mu, \lambda)v(j)\}$$

$$= \max_{\mu \in P_A} \min_{\lambda \in P_B} L(\mu, \lambda)v(i).$$

Thus, the theorem is proved.

Theorem 2.2. Under Assumptions 1 and 2, the game is strictly determined and both players have optimal stationary strategies.

Proof. From (2.4) and Theorem 2.1, there exists a map μ^* from S into P_A such that for each $i \in S$,

$$(2.9) \qquad \alpha v(i) = \max_{\mu \in P_A} \min_{\lambda \in P_B} L(\mu, \lambda)v(i)$$

$$= \min_{\lambda \in P_B} L(\mu^*, \lambda)v(i)$$

and moreover, (2.9) is written as follows: for each $i \in S$ and all $\lambda \in P_B$,

$$\alpha v(i) \lesssim r(i, \mu^*, \lambda) + \sum_j q(j|i, \mu^*, \lambda)v(j).$$

Hence, for a stationary strategy μ^* for player I and any Markov strategy σ for player II, we have for each $t \gtrsim 0$,

$$\alpha v(i) \lesssim r(i, t, \mu^*, \sigma) + \sum_j q(j|i, t, \mu^*, \sigma)v(j). \tag{2.10}$$

Multiplying both sides of (2.10) by $e^{-\alpha t} f_{1i}(t, \mu^*, \sigma)$ and summing over all $i \in S$, we have for each $1 \in S$,

$$\alpha e^{-\alpha t} \sum_i f_{1i}(t, \mu^*, \sigma)v(i) \lesssim e^{-\alpha t} \sum_i f_{1i}(t, \mu^*, \sigma)r(i, t, \mu^*, \sigma)$$
$$+ e^{-\alpha t} \sum_i \sum_j f_{1i}(t, \mu^*, \sigma)q(j|i, t, \mu^*, \sigma)v(j).$$

Since $\sum_i \sum_j |f_{1i}(t, \mu^*, \sigma)q(j|i, t, \mu^*, \sigma)v(j)| < \infty$, the order of the double sum in the second term of the right-hand side can be interchanged. Using the Kolmogorov forward differential equation (1.3), we obtain for each $1 \in S$,

$$\alpha e^{-\alpha t} \sum_i f_{1i}(t, \mu^*, \sigma)v(i) \lesssim e^{-\alpha t} \sum_i f_{1i}(t, \mu^*, \sigma)r(i, t, \mu^*, \sigma)$$
$$+ e^{-\alpha t} \sum_j \frac{\partial}{\partial t} f_{1j}(t, \mu^*, \sigma)v(j). \tag{2.11}$$

By integrating on both sides of (2.11) with respect to $t \in [0, \infty)$, we have for each $1 \in S$,

$$v(1) \lesssim \int_0^\infty e^{-\alpha t} \sum_i f_{1i}(t, \mu^*, \sigma)r(i, t, \mu^*, \sigma)dt. \tag{2.12}$$

Thus, from (2.12), it holds that for each $1 \in S$,

$$v(1) \lesssim \inf_{\sigma \in \Gamma} \psi(1, \mu^*, \sigma) \lesssim \sup_{\pi \in \Pi} \inf_{\sigma \in \Gamma} \psi(1, \pi, \sigma). \tag{2.13}$$

Similarly, it holds that for each $1 \in S$,

$$v(1) \gtrsim \sup_{\pi \in \Pi} \psi(1, \pi, \lambda^*) \gtrsim \inf_{\sigma \in \Gamma} \sup_{\pi \in \Pi} \psi(1, \pi, \sigma), \tag{2.14}$$

where λ^* is the stationary strategy defined in (2.4) for player II.

On the other hand, it is generally true that for each $1 \in S$,

$$\sup_{\pi \in \Pi} \inf_{\sigma \in \Gamma} \psi(1, \pi, \sigma) \lesssim \inf_{\sigma \in \Gamma} \sup_{\pi \in \Pi} \psi(1, \pi, \sigma). \tag{2.15}$$

By (2.13), (2.14) and (2.15), we have $v \in C(S)$ as the value function of the game and μ^* and λ^* are optimal stationary strategies for player I and player II, respectively. Thus, the theorem is proved.

Theorem 2.3 Under Assumptions 1 and 2, for a pair of the stationary strategies (μ, λ) of both players, $v(i) = \psi(i, \mu, \lambda)$ is the unique bounded solution to

$$\alpha v(i) = r(i, \mu, \lambda) + \sum_j q(j|i, \mu, \lambda)v(j). \tag{2.16}$$

Proof. Using a similar argument to the proof of Theorem 2.2, it is easy to show that for each $i \in S$, any solution v to (2.16) must be equal to $\psi(i, \mu, \lambda)$.

Now, we shall show that $\psi(i, \mu, \lambda)$ is a solution to (2.16). For each $i \in S$ and $t \geq 0$, we have

$$\psi(i, \mu, \lambda) = \int_0^t e^{-\alpha s} \sum_j f_{ij}(s, \mu. \lambda) r(j, \mu, \lambda) ds + e^{-\alpha t} \sum_j f_{ij}(t, \mu, \lambda)\psi(j, \mu, \lambda). \tag{2.17}$$

Differentiating both sides of (2.17) with respect to t and taking limit as $t \to 0^+$, we obtain

$$\lim_{t \to 0^+} \sum_j f_{ij}(t, \mu, \lambda) r(j, \mu, \lambda) - \alpha \lim_{t \to 0^+} \sum_j f_{ij}(t, \mu, \lambda)\psi(j, \mu, \lambda) + \lim_{t \to 0^+} \sum_j \frac{\partial}{\partial t} f_{ij}(t, \mu, \lambda)\psi(j, \mu, \lambda) = 0. \tag{2.18}$$

From the definition of $F(t, \mu, \lambda)$, we have for each $i, j \in S$,

$$\lim_{t \to 0^+} f_{ij}(t, \mu, \lambda) = \delta_{ij}$$

and

$$\lim_{t \to 0^+} \frac{\partial}{\partial t} f_{ij}(t, \mu, \lambda) = q(j|i, \mu, \lambda).$$

Hence, since $\| \psi(i, \mu, \lambda) \| \leq \alpha^{-1} \| r \|$, we obtain from (2.18),

$$r(i, \mu, \lambda) - \alpha\psi(i, \mu, \lambda) + \sum_j q(j|i, \mu, \lambda)\psi(j, \mu, \lambda) = 0$$

Thus, the theorem is proved.

Acknowledgment

The authors are grateful to the referees for their various comments and suggestions.

References

[1] Kakumanu, P., "Continuous time Markov decision models with applications to optimization problems", Tech. Rep. 63, Dept. O. R., Cornell University.

[2] Kakumanu, P., "Continuous discounted Markov decision model with countable state and action space", Ann. Math. Statist., 42 (1971), 919 - 926.

[3] Maitra, A. and Parthasarathy, T., "On stochastic games", Journ. Opti. Theory and Appli., 5 (1970), 289 - 300.

[4] Shapley, L. S., "Stochastic games", Proc. National Acad. Sci. U. S. A., 39 (1953), 1095 - 1100.

[5] Sion, M., "On general minimax theorems", Pacific J. Math., 8 (1958), 171 - 176.

[6] Tanaka, K., Iwase, S. and Wakuta, K., "On Markov games with the expected average reward criterion", Sci. Rep. Niigata Univ., Ser. A, No. 13 (1976), 31 - 41.

[7] Tanaka, K. and Wakuta, K., "On semi-Markov games", Sci. Rep. Niigata Univ., Ser. A, No. 13 (1976), 55 - 64.

Kensuke TANAKA: Department of Mathematics,
The Faculty of Science, Niigata
University, Ikarashi, Niigata,
950-21, Japan