Operations Research Letters 43 (2015) 110–116

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Approximation of two-person zero-sum continuous-time Markov games with average payoff criterion

José María Lorenzo [a], Ismael Hernández-Noriega [b], Tomás Prieto-Rumeau [c,*]

[a] Department of Statistics and Operations Research II, Universidad Complutense de Madrid, Spain

[b] Mathematics Department, CINVESTAV-IPN, Mexico

[c] Department of Statistics and Operations Research, UNED, Madrid, Spain

ARTICLE INFO

Article history:
Received 4 September 2014
Received in revised form
4 December 2014
Accepted 5 December 2014
Available online 12 December 2014

Keywords:
Continuous-time zero-sum Markov games
Average payoff
Approximation of game models
Policy iteration algorithm

ABSTRACT

We consider a two-person zero-sum continuous-time Markov game $\mathcal{G}$ with denumerable state space, Borel action spaces, unbounded payoff and transition rates, under the long-run expected average payoff criterion. To approximate numerically the value of $\mathcal{G}$ we construct finite state and actions game models $\mathcal{G}_n$ whose value functions converge to the value of $\mathcal{G}$. Rates of convergence are given. We propose a policy iteration algorithm for the finite state and actions games $\mathcal{G}_n$. We show an application to a population system.

1. Introduction

This paper is concerned with a two-person zero-sum continuous-time Markov game model $\mathcal{G}$ under the long-run average payoff criterion. The state space is a denumerable set, the actions spaces of the players are Borel spaces, and the transition and payoff rates may be unbounded. Under adequate conditions, which include the usual drift inequalities, continuity and compactness requirements, and irreducibility hypotheses, it is known that the game model $\mathcal{G}$ has a value and that the players have optimal (randomized) stationary strategies. Moreover, the value function and the optimal strategies can be characterized by means of an "average optimality equation". The reader can consult [7] or [12, Chapter 10] for details.

The average optimality equation, however, cannot be solved in general (it requires, in particular to determine saddle points of functions defined on infinite-dimensional spaces of probability measures). Therefore, some kind of approximation technique is needed to provide computable approximations of the value of the game. Here, we propose a discretization scheme that approximates the game model $\mathcal{G}$ with finite state and actions game models $\mathcal{G}_n$, for $n \geq 1$. (Loosely, the discretization of the state and actions spaces becomes more accurate as the index n grows.) This discretization consists in truncating the state space and in choosing finite actions sets that are somehow "dense" in the Hausdorff metric. We prove that the value function of the games $\mathcal{G}_n$ converges to the value function of $\mathcal{G}$ as $n \to \infty$ and, moreover, we provide convergence rates that can be explicitly determined from the data of the original game model $\mathcal{G}$. To the best of our knowledge, such approximation techniques have not been studied for average games. A similar approach for approximating continuous-time discounted constrained and unconstrained controlled Markov chains can be found in [9,11], respectively, and in [13] for the average reward criterion. Approximations of discounted games have been studied in [14].

To solve the finite state and actions average game $\mathcal{G}_n$ we introduce a policy iteration algorithm (PIA), whose convergence is proved. A PIA for perfect information games has been proposed in [1]. Let us mention that, in [1], the PIA is of finite nature while, in our context (in which the two players choose their actions *simultaneously*) the PIA is of continuous nature because they do not necessarily exist optimal pure strategies. Our PIA combines a (control) "optimization" step with a "minimax improvement" step. As we will see, it can be reduced to solve iteratively linear programming problems, which makes it computationally tractable. Some other techniques have been proposed in the literature to solve explicitly such finite average games. We may cite, for instance, a nonlinear programming approach in [4], or the general theoretical

* Corresponding author. Tel.: +34 913 987 812.
E-mail addresses: j.lorenzo@ccee.ucm.es (J.M. Lorenzo), ihn@math.cinvestav.edu.mx (I. Hernández-Noriega), tprieto@ccia.uned.es (T. Prieto-Rumeau).

http://dx.doi.org/10.1016/j.orl.2014.12.004
0167-6377/

framework proposed in [3]; see Example 2 therein for an application to a game model.

Finally, let us introduce some notation that will be used throughout the paper. The Borel σ-algebra of a subset of a topological space C is denoted by $\mathbb{B}(C)$. In this paper, measurability is always referred to the Borel σ-algebra. $\mathbf{I}_D$ stands for the indicator function of a set D. The Hausdorff distance between two closed sets F, G of a metric space (X, d_X) is $\rho_X(F, G) = \sup_{x\in F} \inf_{y\in G}\{d_X(x, y)\} \vee \sup_{y\in G} \inf_{x\in F}\{d_X(x, y)\}$, where $\vee$ means "maximum". If μ is a measure on some measure space and f is a measurable function, we write $\mu(f) = \int f d\mu$. We recall that weak convergence of probability measures (denoted by $\xrightarrow{d}$) on a compact metric space X is metrizable with the Wasserstein distance, defined as $d_0(\mu, \nu) = \sup\{\mu(f) - \nu(f)\}$ where the supremum ranges over all 1-Lipschitz continuous functions f on X; see Theorems 8.3.2 and 8.10.45, and Section 8.10(viii) in [2].

The rest of the paper is organized as follows. In Section 2 we state our main definitions and assumptions on the game model $\mathcal{G}$, recall some known results on continuous-time Markov games, and prove some preliminary results. The finite state and actions approximations $\mathcal{G}_n$ are defined in Section 3, in which we prove convergence of the value functions. The policy iteration algorithm for solving $\mathcal{G}_n$ is introduced also in Section 3. Finally, in Section 4, we show an application to a population model and we give some numerical approximations to illustrate our results.

2. Definitions, assumptions, and preliminaries

We consider a zero-sum continuous-time Markov game model $\mathcal{G} = \{S, A, B, \mathbb{K}, Q, r\}$ with the following elements:

- The state space is $S = \{0, 1, 2, \ldots\}$.
- The action spaces for player 1 and 2 are the Borel spaces A and B, respectively, with metrics d_A and d_B. (We recall that a Borel space is a measurable subset of a complete and separable metric space.)
- The actions available for player 1 and 2 in state $i \in S$ are the measurable sets $A(i) \subseteq A$ and $B(i) \subseteq B$, respectively. Let $\mathbb{K} = \{(i, a, b) \in S \times A \times B : a \in A(i), b \in B(i)\}$.
- The system's transition rates are $Q \equiv \{q_{ij}(a, b)\}_{(i,a,b)\in\mathbb{K}, j\in S}$. For every $i, j \in S$, the function $(a, b) \mapsto q_{ij}(a, b)$ is measurable on $A(i) \times B(i)$. The transition rates verify: $q_{ij}(a, b) \geq 0$ for all $(i, a, b) \in \mathbb{K}$ and $j \in S$ such that $i \neq j$ and, in addition, they are conservative and stable, i.e., $\sum_{j\in S} q_{ij}(a, b) = 0$ for all $(i, a, b) \in \mathbb{K}$ and $q(i) := \sup_{(a,b)\in A(i)\times B(i)}\{-q_{ii}(a, b)\} < \infty$ for all $i \in S$.
- The payoff rate function is $r : \mathbb{K} \to \mathbb{R}$. For each $i \in S$, the function $(a, b) \mapsto r(i, a, b)$ is measurable on $A(i) \times B(i)$.

The game is played as follows. At each time $t \geq 0$ both players observe the state of the system $i \in S$ and then they independently and simultaneously choose two actions $a \in A(i)$ and $b \in B(i)$. On the small time interval $[t, t + dt]$, the system makes a transition to the state $j \in S$ with probability $q_{ij}(a, b)dt$ or remains in state $i \in S$ with probability $1 + q_{ii}(a, b)dt$. Player 1 receives a reward $r(i, a, b)dt$ while player 2 incurs a cost $r(i, a, b)dt$. Player 1 wants to maximize his long-run expected average payoff, while the goal of player 2 is to minimize his long-run expected average payoff. Before giving a formal definition of the strategies of the players and the average payoff criterion, we will state our assumptions on the game model $\mathcal{G}$.

Definition 2.1. A function $w : S \to [1, \infty)$ is called a Lyapunov function on S if w is monotone nondecreasing and $\lim_{i\to\infty} w(i) = \infty$.

The w-norm of a function $u : S \to \mathbb{R}$ is defined as $\|u\|_w = \sup_{i\in S}\{|u(i)|/w(i)\}$. Let $\mathcal{B}_w(S)$ be the family of functions u on S with $\|u\|_w < \infty$, which is a Banach space with the w-norm.

Assumption 2.2. There exists a Lyapunov function w on S that verifies the following conditions.

(i) There exist $c_1 > 0$ and $d_1 \geq 0$, and a finite set $D \subset S$ with $\sum_{j\in S} q_{ij}(a, b)w(j) \leq -c_1 w(i) + d_1 \mathbf{I}_D(i)$ for all $(i, a, b) \in \mathbb{K}$.
(ii) For all $i \in S$ we have $q(i) \leq w(i)$.
(iii) For some $M > 0$, $|r(i, a, b)| \leq M w(i)$ for all $(i, a, b) \in \mathbb{K}$.

We say that $\pi^1 \equiv \{\pi_t^1(C|i)\}_{t\geq 0, i\in S, C\in\mathbb{B}(A(i))}$ is a *randomized Markov strategy* for player 1 when $C \mapsto \pi_t^1(C|i)$ is a probability measure on $A(i)$ for each $i \in S$ and $t \geq 0$, and, in addition, for each $i \in S$ and $C \in \mathbb{B}(A(i))$, the function $t \mapsto \pi_t^1(C|i)$ is measurable on $[0, \infty)$. Let Π^1 be the set of all randomized Markov strategies for player 1. The set Π^2 of randomized Markov strategies for player 2 is given a similar definition.

The family of probability measures on $A(i)$ and $B(i)$ are respectively denoted by $\overline{A}(i)$ and $\overline{B}(i)$, for all $i \in S$. We say that a randomized Markov strategy $\pi^1 \in \Pi^1$ is stationary when it does not depend on $t \geq 0$. Consequently, the family $\Pi^{1,s}$ of *randomized stationary strategies* for player 1 can be identified with $\prod_{i\in S} \overline{A}(i)$. Similarly, the family of randomized stationary strategies for player 2 is $\Pi^{2,s} = \prod_{i\in S} \overline{B}(i)$.

Given $\phi \in \overline{A}(i)$, $\psi \in \overline{B}(i)$, and $i, j \in S$, let

$$q_{ij}(\phi, \psi) = \int_{A(i)} \int_{B(i)} q_{ij}(a, b)\psi(db)\phi(da)$$

$$r(i, \phi, \psi) = \int_{A(i)} \int_{B(i)} r(i, a, b)\psi(db)\phi(da).$$

If $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ we write $q_{ij}(t, \pi^1, \pi^2) = q_{ij}(\pi_t^1(\cdot|i), \pi_t^2(\cdot|i))$ and $r(i, t, \pi^1, \pi^2) = r(i, \pi_t^1(\cdot|i), \pi_t^2(\cdot|i))$ for all $i, j \in S$ and $t \geq 0$. For stationary strategies $(\pi^1, \pi^2) \in \Pi^{1,s} \times \Pi^{2,s}$, these expressions do not depend on $t \geq 0$ and they will be written as $q_{ij}(\pi^1, \pi^2)$ and $r(i, \pi^1, \pi^2)$. Notations such as, e.g., $r(i, \phi, b)$ or $q_{ij}(a, \psi)$ are given the obvious definitions.

Let $\Omega = \mathbb{K}^{[0,\infty)}$ be the space of all the sample paths of the game endowed with the product σ-algebra $\mathcal{F}$. For each $t \geq 0$, let $x(t)$, $a(t)$, and $b(t)$ be the coordinates projections from Ω to S, A, and B, respectively. We state the next theorem on a somewhat loose form; for details, the reader can consult [7].

Theorem 2.3. *Suppose that Assumptions 2.2(i)–(ii) hold and fix an arbitrary pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$.*

(i) *There is a unique regular transition function $P_{ij}^{\pi^1,\pi^2}(s, t)$, for $i, j \in S$ and $t \geq s \geq 0$, with transition rates $q_{ij}(s, \pi^1, \pi^2)$.*
(ii) *For each initial state $i \in S$ at time 0 there exists a unique probability measure on $(\Omega, \mathcal{F})$, denoted by P^{i,π^1,π^2}, that models the dynamics of the state of the system and the actions of the players.*

The expectation operator associated to P^{i,π^1,π^2} will be denoted by E^{i,π^1,π^2}. For a proof of statement (i) we refer to [8, Appendix C], while the proof of (ii) can be found in [7]. Slightly abusing the notation, we will refer to $\{x(t)\}_{t\geq 0}$, $\{a(t)\}_{t\geq 0}$, and $\{b(t)\}_{t\geq 0}$ as the state and the actions processes for the players when P^{i,π^1,π^2} is given.

Under Assumption 2.2(i), given an initial state $i \in S$ and a pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, by [8, Lemma 6.3] we have

$$E^{i,\pi^1,\pi^2}[w(x(t))] \leq e^{-c_1 t} w(i) + \frac{d_1}{c_1}(1 - e^{-c_1 t}) \quad \text{for all } t \geq 0. \tag{2.1}$$

The long-run expected average payoff (or average payoff) is

$$J(i, \pi^1, \pi^2) = \limsup_{T\to\infty} \frac{1}{T} E^{i,\pi^1,\pi^2}\left[\int_0^T r(x(t), a(t), b(t))dt\right].$$